UNIVERSITY OF GLASGOW

MODEL SELECTION AND MODEL AVERAGING IN THE PRESENCE OF MISSING VALUES

by

KHUNESWARI GOPAL PILLAY

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
SCHOOL OF MATHEMATICS AND STATISTICS

November 2015

Declaration of Authorship

I, KHUNESWARI GOPAL PILLAY, declare that this thesis titled, 'Model selection and model averaging in the presence of missing values' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

If you fail, never give up because F.A.I.L means First Attempt In Learning.

A.P.J Abdul Kalam

Abstract

Model averaging has been proposed as an alternative to model selection which is intended to overcome the underestimation of standard errors that is a consequence of model selection. Model selection and model averaging become more complicated in the presence of missing data. Three different model selection approaches (RR, STACK and M-STACK) and model averaging using three model-building strategies (non-overlapping variable sets, inclusive and restrictive strategies) were explored to combine results from multiply-imputed data sets using a Monte Carlo simulation study on some simple linear and generalized linear models. Imputation was carried out using chained equations (via the "norm" method in the R package MICE). The simulation results showed that the STACK method performs better than RR and M-STACK in terms of model selection and prediction, whereas model averaging performs slightly better than STACK in terms of prediction. The inclusive and restrictive strategies perform better in terms of prediction, but non-overlapping variable sets performs better for model selection. STACK and model averaging using all three model-building strategies were proposed to combine the results from a multiply-imputed data set from the Gateshead Millennium Study (GMS). The performance of STACK and model averaging was compared using mean square error of prediction (MSE(P)) in a 10% cross-validation test. The results showed that STACK using an inclusive strategy provided a better prediction than model averaging. This coincides with the results obtained through a mimic simulation study of GMS data. In addition, the inclusive strategy for building imputation and prediction models was better than the non-overlapping variable sets and restrictive strategy. The presence of highly correlated covariates and response is believed to have led to better prediction in this particular context. Model averaging using non-overlapping variable sets performs better only if an auxiliary variable is available. However, STACK using an inclusive strategy performs well when there is no auxiliary variable available. Therefore, it is advisable to use STACK with an inclusive model-building strategy and highly correlated covariates (where available) to make predictions in the presence of missing data. Alternatively, model averaging with non-overlapping variables sets can be used if an auxiliary variable is available.

Acknowledgements

First of all, I would like to thank God for his blessings that enabled me to complete my studies by providing opportunities and insights for conducting the research. Next, I would like to take this opportunity to express my gratitude to my supervisor Professor John H. McColl for his advice, guidance, constructive comments and immerse contribution to my final piece of work and his invaluable help and assistance in the completion of this project. All his efforts are deeply appreciated and would not be forgotten. I would also like to thank Professor Charlotte Wright and the Gateshead Millennium Study Core Team for giving approval to use the Gateshead Millennium Study data for my research.

I would like to thank my beloved husband Mohana Kanapathy for his support, guidance, love, sacrifice and motivation throughout my research duration. I would also like to thank my beloved parents Gopal Pillay and Vasantha Kumari, my brothers Thinesh and Vijayan for their support, love and motivation throughout my research duration. I would like to thank my sister Kogilavani, my friends and colleagues for their encouragement and motivation.

I would like to thank Ministry of Education Malaysia (MOE) and Universiti Tun Hussein Onn Malaysia (UTHM) for their financial support and guidance throughout my research duration. I would like to thank Professor Zainodin Hj. Jubok from University Malaysia Sabah for his support and motivation.

Lastly, I wish to thank all lecturers, staff and friends in School of Mathematic and Statistics and College of Science and Engineering who helped me in giving guidance and support throughout my years at University of Glasgow.

Contents

List of Figures

List of Tables

Abbreviations

AIC	Akaike Information criterion
AIC$_c$	Corrected Akaike Information criterion
AV	Available-case
BIC	Bayesian Information criterion
CC	Complete-case
CV	Cross validation
EM	Expectation maximization
FCS	Fully conditional specification
GLM	Generalized linear model
GMS	Gateshead Millennium Study
JM	Joint modelling
LM	Linear model
MA	Model averaging
MAR	Missing at random
MCAR	Missing completely at random
MCMC	Markov Chain Monte Carlo
MI	Multiple imputation
MICE	Multiple imputation by chained equations
ML	Maximum likelihood
MS	Model Selection
MSE	Mean square error
MSE(P)	Mean square error of prediction
NMAR	Not missing at random
RR	Rubin's rules
SDS	Standard deviation score

Symbols

k	Polynomial order
p	Number of parameters
m	Percentage of missing value
n	Number of observations
σ	Standard deviation
D	Number of multiply imputed data sets
ρ	Correlation coefficient
v	Degrees of freedom
M	Number of models
t	Number of test values
τ	Number of iterations
P	Probability of success
s	Number of simulation
ε	Error term
σ_ε	Error variance
Y	Response variable
X	Explanatory variables
h	Error term for imputation model
σ_h	Error variance for imputation model
β	Coefficients for prediction model
φ	Coefficients for imputation model
f_i	Fraction of missing data for variable X_i

Dedicated to my family

Chapter 1

Introduction

Model-building is one of the key areas of interest in the development and application of statistical modelling. One important issue in model-building is the need for researchers to clearly identify the ultimate aim of their research in order to choose an appropriate model-building approach. There are two crucial aims of a data analysis: (1) to determine which factors/variables to include when making predictions and (2) prediction. The relative importance of these aims will guide the researchers to choose a suitable model-building approach for their research and will help in determining an appropriate structure for the model of interest.

A statistical model is a simplified description of data and it is often based on some mathematically defined relationship. A model is usually constructed in order to draw conclusions and make predictions from the data. The model should be rich enough to explain the relationships in the data. In some situations there will be a lot of factors that might affect the response and therefore many possible models to consider. Model selection provides formal support to guide the user in the search for the best model and to determine which factors/variables to be included when making predictions. Model selection is an important part of the model-building process and cannot be separated from the rest of the analysis in choosing a best model [Claeskens and Hjort, 2008].

Model selection in practice requires the choice of a selection procedure, such as forward selection or backward elimination, coupled with a selection criterion, such as AIC or BIC, to select a small subset of variables to include in the model. Model selection is well-known for introducing additional uncertainty into the model-building process. The properties of standard parameter estimates obtained from the selected model do not reflect the stochastic nature of the model selection process [Burham and Anderson, 2002]. In the literature, model averaging has been proposed as an alternative to model selection which is intended to overcome the underestimation of standard errors that is

a consequence of model selection. If the focus of model selection and model averaging is good prediction, then differences in the standard errors of estimators is not directly relevant to the comparison of these methods.

Model selection and model averaging become more complicated in the presence of missing data. Missing data is a common problem in various settings, including surveys, clinical trials and longitudinal studies. Values of both outcome/response and covariates might be missing. Many researchers usually omit the variable or samples with missing data from the analysis but this can lead to bias and loss of information. The cumulative effect of a small amount of missing data in each of several variables can lead to exclude many of the potential samples, which in turn will cause loss of precision. Exploiting relationships between the variables in order to impute the missing values can be demonstrated to be a better strategy [Little and Rubin, 2002].

Although researchers have developed many imputation methods to deal with missing data, there are no agreed guidelines for model selection in the presence of missing data. Model averaging is the most relevant method to account for both uncertainty related to imputation and model selection. However, there are no proper guidelines for model averaging in the presence of missing data. Besides that, there is no proper comparison between model selection and model averaging in the presence of missing data in terms of prediction.

1.1 Research Motivations

In the analysis of statistical models, the main issues are model-building, model selection and prediction based on the best model. Model selection introduces additional uncertainty into the model-building process, but the standard errors of parameter estimates obtained from the selected model by standard statistical procedures will underestimate the true variability. Model averaging aims to incorporate the uncertainty associated with model selection into parameter estimation, by combining estimates over a set of possible models. Model selection and model averaging in the linear and generalized linear models become complicated in the presence of missing data. Model selection in the presence of missing data has been widely explored over decade. Only in the past two years has some research been carried out on how best to carry out model averaging in the presence of missing data [Schomaker and Heumann, 2014]. There are outstanding issues, such as how to combine model averaging estimators for multiply-imputed data sets, the number of multiple imputations needed and the relationship between the imputation and prediction models, which remain unclear and need proper guidelines.

Building a good imputation model is a key factor in dealing with missing data. Researcher should build a robust imputation model with sufficient amount of complete data in order to obtain good imputed values. The imputation model and the prediction model should be compatible to provide good results [Sinharay et al., 2001]. Any discrepancy between the imputation model and the prediction model will give rise to unreliable estimates. Therefore, building robust imputation and prediction models is crucial in model-building in the presence of missing data.

Another key issue is how the strength of correlation among available variables will affect imputation and prediction. Highly correlated variables are ideal for imputation, as stated for example by Hardt et al. [2012]. However, there can be negative effects of highly correlated variables in the prediction model, such as low precision for estimating parameters. This means that highly correlated variables should be handled carefully.

Moreover, the choice of model selection criterion will have an effect on both model selection and model averaging in the presence of missing data. Although AIC is widely used as a criterion for model selection and for calculating model weights in model averaging, AIC will not necessarily choose the most parsimonious model and there is a probability of over-fitting. A corrected version of AIC, known as AIC_c, has been shown to have an advantage over AIC in small to medium-sized samples [Burham and Anderson, 2002]. BIC will choose a more parsimonious model than either AIC or AIC_c because of the stronger penalty term which discourages choosing a model with many parameters. There is no proper comparison between these model selection criteria in model selection and model averaging in the presence of missing data.

Finally, although model averaging has been proposed as an alternative to model selection, there is no proper comparison between the two in the presence of missing data, in terms of prediction. Therefore, this research will involve comparing model selection and model averaging in the presence of missing data using several model-building strategies and different model selection criteria, with the specific research objectives listed in the next section.

1.2 Research Objectives

The detailed research objectives of this research are as follows:

(i) To investigate the implications of multiple imputation for selecting and fitting additive linear and generalized linear models, using common model selection criteria.

(ii) To investigate the implications of multiple imputation for model averaging.

(iii) To investigate the effects of restrictive and inclusive strategies for imputation for both model selection and model averaging.

(iv) To compare model selection and model averaging in terms of prediction in the presence of missing values.

(v) To identify the effects of highly correlated covariates on model selection and model averaging, in the absence and presence of missing values.

1.3 Thesis Outline

The structure of this thesis is explained in this sub-chapter.

Chapter 1 presents the introduction and motivations of the current study. It also identifies the aims and objectives of this work and outlines the thesis structure.

Chapter 2 explains the methodology related to this research. It covers methods relevant to this study such as statistical approaches to analyze missing data, software packages for imputation, model selection criteria and non-Bayesian model averaging.

Chapter 3 reviews previous research on model selection and model averaging in the presence of missing values. It also covers recent developments on model selection strategies and criteria in the presence of missing data and strategies for building an imputation model.

Chapter 4 presents the results of a small scale simulation study which was carried out to investigate the effects of restrictive and inclusive strategies for single imputation on both model selection and model averaging. Model selection and model averaging using all three model-building strategies (non-overlapping variable sets, restrictive and inclusive strategies) were compared to identify the best model-building strategy.

Chapter 5 extends the simulation study of Chapter 4 to multiple imputation. Three model selection methods (RR, STACK, M-STACK) and model averaging are discussed to combine results across multiply-imputed data sets and compared. These procedures were compared using mean square error of prediction to identify the best model-building approach.

Chapter 6 presents results obtained from applying the most successful model-building approaches (STACK and model averaging) and strategies (non-overlapping variable sets, restrictive and inclusive strategies) to the prediction of children's weight at school entry and weight at eight years of age based on their first year weights in the Gateshead

Millennium Study. The model-building approaches and strategies were compared using mean square error of prediction.

Chapter 7 summarizes all the conclusions that can be drawn from this thesis. Areas of further work and a summary of the research completes this chapter.

Chapter 2

Methodology

2.1 Missing Data

Missing data is a common problem in various settings, including surveys, clinical trials and longitudinal studies. Values of both outcome/response and covariates might be missing. Researchers usually omit the variable or samples with missing data from the analysis but this can lead to bias and loss of information. The cumulative effect of a small amount of missing data in each of several variables will lead to exclude many of the potential samples, which in turn will cause loss of precision.

In order to overcome the missing data issue more appropriately, researcher should understand the missing data pattern or type. Little and Rubin [2002] classified missing data into three types (also known as missing data mechanisms) which are missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). The details of these three types of missing data are as follows:

1. If the missingness of a variable X does not depend on, or is unrelated to, the value of X itself or to any other variables in the dataset, these data are called missing completely at random (MCAR). In other words, data are MCAR if the probability of being missing is the same for all cases. There are then no systematic differences between the missing values and the observed values of variable X. For example, weight values were missing because an electric scale ran out of batteries, so some of the data were missing simply because of bad luck [van Buuren, 2012].

2. If the missingness on X is related to another variable (Y) in the analysis but not to X itself, these data are called missing at random (MAR). In other words, data are MAR if the probability of being missing is the same only within groups defined by the observed data. Any systematic difference between the missing values and

the observed values of variable X can be explained by patterns in the observed data. For example, when scales are placed on a soft surface, they may produce more missing values than when placed on a hard surface. Since the surface type is known, if one assumes data are MCAR within the type of surface, then overall the data are MAR [van Buuren, 2012].

3. If missingness is related to the value of X itself, and perhaps one or more other variables in the prediction model, these data are called not missing at random (NMAR). In other words, data are NMAR if the probability of being missing varies for reasons that are not known to the researcher. Even after the observed data are taken into account, systematic differences remain between the missing values and the observed values of variable X. For example, the weighing scale will wear out over time and produce more missing data. One may fail to note this. If heavier objects are measured later in time, then a distorted distribution of measurements will be obtained. NMAR includes the possibility that the scale produces more missing values for heavier objects [van Buuren, 2012].

2.2 Statistical Approaches to Analyze Missing Data

Performing analysis for missing data problems raises several new statistical challenges, underscoring the need for methodological development. In the literature, methods commonly proposed are complete case analysis (listwise deletion), mean imputation, regression imputation, stochastic regression imputation, hot deck imputation, EM algorithm and multiple imputation.

2.2.1 Complete-case analysis

The traditional method of dealing with missing data is to delete any cases with missing values from the analysis. This is known as complete-case (CC) analysis (listwise deletion). It is a default method of handling missing data in many statistical packages. This procedure will eliminate all cases with one or more missing values on the analysis variables [van Buuren, 2012]. The main advantages of this approach are simplicity and comparability of the results with results from the analysis of dataset with no missing values. Any standard statistical analysis can be applied without modification to complete cases [Little and Rubin, 2002]. Under MCAR, CC analysis will produce unbiased estimates of means, variance and regression coefficients. Disadvantages of this method are loss of precision, and bias when the missing data is not MCAR and the complete

cases are not a random sample of all the cases. Therefore, it is not advisable to use CC analysis to deal with missing data.

A special case of CC analysis is available-case (AV) analysis (also known as pairwise deletion). AV analysis uses all the cases with complete data on selected variables for particular analysis. According to Osborne [2013], the sample included in AV analysis can change depending on which variables are in the analysis. The estimates of means and variances are not biased if data are MCAR but modifications are needed to estimate measures of covariation. This also leads to mis-estimation and errors in data that are MAR or NMAR.

2.2.2 Single imputation

Imputation is a common and flexible method to deal with missing data. According to Little and Rubin [2002], imputations are means or draws from a predictive distribution of the missing values. Imputing one value for each missing value is called single imputation. Single imputation is often utilized because it is intuitively attractive. In single imputation, one will fill in missing values by some type of predicted values. There are many single imputation methods including mean imputation, regression imputation, stochastic regression imputation and hot deck imputation.

Mean imputation is replacing missing values with a measure of central tendency, often the sample mean for continuous data and the mode for categorical data. Mean imputation is a quick and simple fix for missing data. van Buuren [2012] states that this method will underestimate the variance, disturb the relations between variables and bias estimates of the mean, even when data are MCAR. Mean imputation should be avoided in general but it can be used as a rapid fix when a handful of data are missing.

Regression imputation replaces missing values by predicted values from a regression model for the missing variable. The first step in regression imputation is building a model from observed data. Predictions for the incomplete cases are calculated from the fitted model and used as replacements for the missing data. Under MCAR, regression imputation will produce unbiased estimates of the means and regression coefficients of the imputation model if the explanatory variables used in this model are complete [van Buuren, 2012]. However, the variability of the data is systematically underestimated. Little and Rubin [2002] stated that the degree of underestimation depends on the amount of variance explained and on the proportion of missing data.

Stochastic regression imputation is an improvement on regression imputation that adds noise (or errors) to the predictions. This will have a depressing effect on correlations. van Buuren [2012] and Little and Rubin [2002] described how this method first estimates the intercept, slope and residual variance under the linear model. Then it generates imputed values according to these parameter estimates. The noise added to the predictions opens up the distribution of the imputed values. This method will preserve both regression coefficients and correlation between variables. The main idea of drawing from the distribution of residuals is very powerful and forms the basis for more advanced imputation methods.

Both regression imputation and stochastic regression imputation will yield unbiased estimates under MAR. The common problem in single imputation comes from replacing an unknown missing value by a single value and then treating it as if it is a true value [Rubin, 1987]. Single imputation ignores uncertainty so almost always underestimates the variance. Multiple imputation can be used to overcome this problem by taking into account both within-imputation and between-imputation uncertainty.

2.2.3 Hot deck imputation

Hot deck imputation is a single imputation method to deal with missing data which involves replacing each missing value with an observed response from a similar unit. Little and Rubin [2002] stated that this is a common method in survey practice and very elaborate schemes have been developed for selecting units that are similar in order to carry out the imputation. The result of hot deck imputation is a rectangular dataset which can be used in secondary data analysis. There is a consequent gain in efficiency respective to CC analysis since information present in incomplete cases will be retained. This method does not depend on modelling the variable to be imputed, therefore it is potentially less sensitive to model misspecification than imputation methods based on a parametric model such as regression imputation [Andridge and Little, 2010].

Another important feature of this method is that it can also replace missing values with observed responses from other units. There is a reduction in non-response bias where there is an association between the variables defining imputation categories [Andridge and Little, 2010]. However, according to Roth [1994], there are several disadvantages of the hot deck imputation method. First, the number of cross-classifications of variables may become unmanageable in large survey research. Researchers are encouraged to include many variables in the identification of similar units because each one has some effect on the variable to be imputed. Deleting a classification variable means that the imputed variable will lose a fraction of its variance attributed to that classification

variable. The correlations between the imputed variable and other variables will be weaker. Second, the classification of variables required for identifying similar units sacrifices information. The third disadvantage is that estimating the standard error of parameters can be difficult.

2.2.4 EM algorithm

The Expectation Maximisation (EM) algorithm is an alternative computing strategy for incomplete data. The EM algorithm is a very general algorithm for maximum likelihood (ML) estimation for incomplete data [Little and Rubin, 2002]. It is an iterative approach that involves two steps: the expectation step (E-step) and the maximisation step (M-step). In any incomplete data problem, the distribution of the complete data X can be factorised as

$$f(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\theta}) = f(\boldsymbol{Y}, \boldsymbol{X}_{obs}; \boldsymbol{\theta}) f(\boldsymbol{X}_{mis} | \boldsymbol{Y}, \boldsymbol{X}_{obs}; \boldsymbol{\theta}) \tag{2.1}$$

Considering each term in Equation (2.1) as a function of θ, it follow that

$$\ell(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\theta}) = \ell(\boldsymbol{Y}, \boldsymbol{X}_{obs}; \boldsymbol{\theta}) + \ell(\boldsymbol{X}_{mis} | \boldsymbol{Y}, \boldsymbol{X}_{obs}; \boldsymbol{\theta}) + c \tag{2.2}$$

where $\ell(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\theta})$ denotes the complete data log-likelihood, $\ell(\boldsymbol{Y}, \boldsymbol{X}_{obs}; \boldsymbol{\theta})$ denotes the observed data log-likelihood and c is an arbitrary constant. The incomplete data log-likelihood is often inconvenient to work directly and the maximisation can be difficult to accomplish [Schafer, 1997]. The E-step takes the average of the complete data log-likelihood with respect to the distribution $f(\boldsymbol{X}_{mis} | \boldsymbol{X}_{obs}; \boldsymbol{\theta}^{(\tau)})$, where $\theta^{(\tau)}$ is the current parameter estimate of θ. This log-likelihood yields

$$\begin{aligned} &\int \ell(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\theta}) f(\boldsymbol{X}_{mis} | \boldsymbol{Y}, \boldsymbol{X}_{obs}, \boldsymbol{\theta}^{(\tau)}) d\boldsymbol{X}_{mis} \\ &\qquad = \int \ell(\boldsymbol{Y}, \boldsymbol{X}_{obs}; \boldsymbol{\theta}) f(\boldsymbol{X}_{mis} | \boldsymbol{Y}, \boldsymbol{X}_{obs}, \boldsymbol{\theta}^{(\tau)}) d\boldsymbol{X}_{mis} \\ &\qquad + \int \ell(\boldsymbol{X}_{mis} | \boldsymbol{Y}, \boldsymbol{X}_{obs}, \boldsymbol{\theta}) f(\boldsymbol{X}_{mis} | \boldsymbol{Y}, \boldsymbol{X}_{obs}, \boldsymbol{\theta}^{(\tau)}) d\boldsymbol{X}_{mis} \end{aligned} \tag{2.3}$$

Equation (2.3) can be written in the form of a Q-function and H-function as follows

$$\begin{aligned} Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(\tau)}) &= \int \ell(\boldsymbol{Y}, \boldsymbol{X}_{obs}; \boldsymbol{\theta}) f(\boldsymbol{X}_{mis} | \boldsymbol{Y}, \boldsymbol{X}_{obs}, \boldsymbol{\theta}^{(\tau)}) d\boldsymbol{X}_{mis} + H(\boldsymbol{\theta} | \boldsymbol{\theta}^{(\tau)}) \\ &= \ell(\boldsymbol{Y}, \boldsymbol{X}_{obs}; \boldsymbol{\theta}) \int f(\boldsymbol{X}_{mis} | \boldsymbol{Y}, \boldsymbol{X}_{obs}, \boldsymbol{\theta}^{(\tau)}) d\boldsymbol{X}_{mis} + H(\boldsymbol{\theta} | \boldsymbol{\theta}^{(\tau)}) \\ &= \ell(\boldsymbol{Y}, \boldsymbol{X}_{obs}; \boldsymbol{\theta}) + H(\boldsymbol{\theta} | \boldsymbol{\theta}^{(\tau)}) \end{aligned} \tag{2.4}$$

where the $H-$function is

$$H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \int \ell(\boldsymbol{X}_{mis}|\boldsymbol{Y}, \boldsymbol{X}_{obs}, \boldsymbol{\theta}) f(\boldsymbol{X}_{mis}|\boldsymbol{Y}, \boldsymbol{X}_{obs}, \boldsymbol{\theta}^{(\tau)}) d\boldsymbol{X}_{mis} \tag{2.5}$$

The E-step is based on the evaluation of the $Q-$function in Equation (2.4). The M-step involves maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(\tau)})$ with respect to θ to obtain $\theta^{(\tau+1)}$. The iteration between the E-step and M-step will continue until convergence [Little and Rubin, 2002, Schafer, 1997].

Little and Rubin [2002] stated that there are two major drawbacks of EM algorithm. First, it will converge very slowly in cases with large fractions of missing data. Second, the M-step will be difficult in some cases and then the theoretical simplicity of EM will not convert to simplicity in practice. Another problem with EM is that it leads to biased parameter estimates and underestimates the standard errors. For this reason, statisticians do not recommend EM as a final solution. Multiple imputation avoids two of the difficulties associated with maximum likelihood methods using the EM algorithm. With multiple imputation, a researcher may use standard methods of analysis once imputations are computed, and can easily obtain standard errors of estimates [Pigott, 2001].

2.2.5 Multiple imputation and Rubin's Rules

Multiple imputation (MI) is an extension of single imputation for the analysis of incomplete dataset, which has become increasingly popular because of its generality and recent software developments that makes it easier to implement. It was first proposed by Rubin in the early 1970's [Little and Rubin, 2002]. MI is the procedure of substituting each missing value by $D \geq 2$ imputed values in order to create multiple completed dataset. MI involves carrying out an analysis on each completed dataset, then combining the results to reflect the variability within-imputation and between-imputation.

MI produces asymptotically unbiased estimates when it is implemented correctly and it is also asymptotically efficient. According to White et al. [2011], there are three stages in the MI process which are described below.

- *Stage 1: Generating multiply-imputed dataset*
  The unknown missing data are replaced by D independent simulated sets of values which are drawn from the distribution of the missing data conditional on the observed data.

- *Stage 2: Analyzing multiply-imputed dataset*
  Once the multiple imputations have been generated, each imputed dataset is analyzed separately as though it was a complete dataset. Parameters are estimated from each imputed dataset. The results of these D analyses differ because the missing values have been replaced by different imputations.

- *Stage 3: Combining estimates from multiply-imputed dataset*
  The D estimates are combined into an overall estimate and variance-covariance matrix using Rubin's rules (RR). The combined variance-covariance matrix incorporates both within-imputation and between-imputation variability.

Rubin's rules are as follows. The $\widehat{\boldsymbol{\theta}}_d$ is an estimate of a univariate or multivariate quantity of interest obtained from the dth imputed dataset and $\boldsymbol{W}_d$ is the estimated variance of $\widehat{\boldsymbol{\theta}}_d$. The combined estimate $\bar{\boldsymbol{\theta}}$ is the average of the individual estimates [Rubin, 1987]:

$$\bar{\boldsymbol{\theta}} = \frac{1}{D}\sum_{d=1}^{D}\widehat{\boldsymbol{\theta}}_d \tag{2.6}$$

The total variance of $\bar{\boldsymbol{\theta}}$ is formed from the within-imputation variance $\boldsymbol{W} = \frac{1}{D}\sum_{d=1}^{D}\boldsymbol{W}_d$ and the between-imputation variance $\boldsymbol{B} = \frac{1}{D-1}\sum_{d=1}^{D}(\widehat{\boldsymbol{\theta}}_d - \bar{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_d - \bar{\boldsymbol{\theta}})^T$:

$$cov(\bar{\boldsymbol{\theta}}) = \boldsymbol{W} + \left(\frac{D+1}{D}\right)\boldsymbol{B} \tag{2.7}$$

An approximate confidence interval for $\boldsymbol{\theta}_i$ is given by

$$\bar{\boldsymbol{\theta}}_i \pm t_v\sqrt{var(\bar{\boldsymbol{\theta}}_i)} \tag{2.8}$$

or

$$\bar{\boldsymbol{\theta}}_i \pm t_v\sqrt{\boldsymbol{W}_{ii} + \left(\frac{D+1}{D}\right)\boldsymbol{B}_{ii}} \tag{2.9}$$

where the degrees of freedom v are estimated by

$$v = (D-1)\left\{1 + \frac{D\boldsymbol{W}_{ii}}{(1+D)\boldsymbol{B}_{ii}}\right\}^2 \tag{2.10}$$

and t_v is the appropriate fraction of the central t-distribution on v degree of freedom. Note that both v and $cov(\bar{\boldsymbol{\theta}}_i)$ are estimated from the data and both depend on the quantity $\boldsymbol{B}$ and v also depends on $\boldsymbol{W}$. $\boldsymbol{B}$ itself is an estimated variance with $D-1$ degrees of freedom. Schafer and Olsen [1998] stated that, with an infinite number of imputations

($D = \infty$), the total variance reduces to the sum of the two variance components and the confidence interval is based on a normal distribution ($v = \infty$). Rubin's Rules should be applied to estimators which are normally distributed. For logistic regression, Rubin's Rules can be applied on the log-odds ratio scale but not on the odds-ratio scale. Rubin's Rules can be applied analogously for other generalized linear models.

According to Patrician [2002], there are advantages of using MI over single imputation. MI incorporates random error because it requires random variation in the imputation process. Since repeated estimations are used, MI gives more reasonable estimates of standard error than single imputation methods. Moreover, MI increases the efficiency of the estimates because it reduces the standard errors.

There are some disadvantages of MI compared to single imputation. MI needs more effort to create the multiple imputations, needs more time to run the analysis and needs more computer storage space for the imputation-created dataset [Patrician, 2002, Rubin, 1987]. Computer storage capacity is not an issue nowadays since more advanced hard disk storage has been produced, and the other disadvantages are also being reduced as time and technology advances.

2.2.6 Chained equations

Two general approaches for imputing multivariate data are joint modeling (JM) and fully conditional specification (FCS). Various JM techniques were developed by Schafer [1997] for imputation under the multivariate normal, the log-linear and the general location model. JM specifies a multivariate distribution for the missing data and draws imputations from their conditional distributions by using Markov Chain Monte Carlo (MCMC) techniques [van Buuren and Groothuis-Oudshoorn, 2011].

FCS specifies the multivariate imputation model on a variable-by-variable basis by a set of conditional densities. FCS draws imputations by iterating over the conditional densities and it is started from an initial imputation. FCS requires a lower number of iterations than JM. When no suitable multivariate distribution can be proposed, FCS is an alternative method to JM. Although the basic idea of FCS is quite old, it has been proposed using a variety of names which includes stochastic relaxation, variable-by-variable imputation, regression switching, sequential regressions, ordered pseudo-Gibbs sampler, partially incompatible MCMC, iterated univariate imputation, chained equations and fully conditional specification. FCS is also known as chained equations and sequential regressions. Imputations are created by drawing from iterated conditional models.

Let hypothetically complete data X be a partially observed random sample from the p-variate multivariate distribution $P(X|\theta)$. Assume that the multivariate distribution of X is completely specified by θ, a vector of unknown parameters. The chained equation method proposes to obtain a posterior distribution of θ by iterative sampling from conditional distributions of the form [van Buuren and Groothuis-Oudshoorn, 2011]

$$\begin{aligned} &P\left(X_1 \mid X_{-1}, \theta_1\right) \\ &P\left(X_2 \mid X_{-2}, \theta_2\right) \\ &\quad\vdots \\ &P\left(X_k \mid X_{-k}, \theta_k\right) \end{aligned}$$

where X_{-i} denotes the data vector X with X_i deleted. The parameters $\theta_1, \theta_2, \ldots, \theta_k$ are specific to the respective conditional densities and are not necessarily the product of a factorization of the "true" joint distribution $P\left(X \mid \theta\right)$. Starting from a simple draw from observed marginal distributions, the τth iteration of the chained equations is a Gibbs sampler that successively draws

$$\begin{aligned} \theta_1^{*(\tau)} &\sim P\left(\theta_1 \mid X_1^{obs}, X_2^{(\tau-1)}, \ldots, X_k^{(\tau-1)}\right) \\ X_1^{*(\tau)} &\sim P\left(X_1 \mid X_1^{obs}, X_2^{(\tau-1)}, \ldots, X_k^{(\tau-1)}, \theta_1^{*(\tau)}\right) \\ &\quad\vdots \\ \theta_k^{*(\tau)} &\sim P\left(\theta_k \mid X_k^{obs}, X_1^{(\tau)}, \ldots, X_{k-1}^{(\tau)}\right) \\ X_k^{*(\tau)} &\sim P\left(X_k \mid X_k^{obs}, X_1^{(\tau)}, \ldots, X_k^{(\tau)}, \theta_k^{*(\tau)}\right) \end{aligned}$$

where $X_k^{(\tau)} = \left(X_k^{obs}, X_k^{*(\tau)}\right)$ is the kth imputed variable at iteration τ. Observe that previous imputations $X_k^{*(\tau-1)}$ only enter $X_k^{*(\tau)}$ through its relation with other variables and not directly. Therefore, it will converge quite fast compared to other MCMC methods. The name chained equation refers to implementation of the Gibbs sampler as a concatenation of univariate procedures to fill out the missing data. Royston and White [2011] suggested that more than 10 cycles are needed for the convergence of the sampling distribution of imputed values, whereas the entire procedure is repeated independently D times, yielding D imputed dataset.

2.3 Software Packages for Imputation

Multiple imputation is now widely used to handle missing values by researchers. There are several software packages including R, SAS and SPlus which can be used to simplify the process for filling in missing values with multiple imputations. There are several

multiple imputation packages in R. Two of the packages are described in the next two sections:

- Multiple Imputation(mi) package by Yu et al. [2011]
- Multivariate Imputation by Chained Equations (MICE) package by van Buuren and Groothuis-Oudshoorn [2011]

2.3.1 Multiple imputation (MI)

The mi package in R was created by Yu et al. [2011]. The mi package uses a chained equation approach (see Section 2.2.6). The package has several features that allow the researcher to use the imputation process and evaluate the reasonableness of the resulting models and imputations. The features are:

1. flexible choice of predictors, model and transformations for chained imputation models
2. binned residual plots for checking the fit of the conditional distributions used for imputation
3. plots for comparing the distributions of observed and imputed data in one and two dimensions

Although the implementation of the mi package is straightforward and uses the random imputation method, it only implements the bootstrap method and the choice of imputation model is fixed based on the variable types. According to Yu et al. [2011], the mi package uses the predictive mean matching (pmm) method to impute positive-continuous and non negative variable types and uses linear regression to impute continuous variables. Besides that, the mi package uses Bayesian regression models and weakly informative prior distributions to construct estimates of imputation models. The MICE package (described in the next section) gives more options on choosing the imputation methods for numeric variables. Since this research is generally looking at numeric variables, the MICE package was chosen to use as an imputation package and it has been explored in order to choose a best imputation method for linear and generalized linear models.

2.3.2 Multivariate imputation by chained equations (MICE)

Multivariate Imputation by Chained Equations (MICE) is a package in R for imputing incomplete multivariate data by Fully Conditional Specification (FCS), developed by

van Buuren and Groothuis-Oudshoorn [2011]. Their aim is to yield imputations that are statistically correct as in Little and Rubin [2002]. It is important to observe convergence, but in the MICE package the desired number of iterations is often a small number, between 10 to 20.

The package MICE in R contains functions for three phases of multiple imputation which includes generating multiple imputations, analyzing imputed data and pooling the analysis results. The most challenging step in multiple imputation is the specification of the imputation model. According to van Buuren and Groothuis-Oudshoorn [2011], there are seven main steps in setting up multiple imputation by MICE package. These are described below.

1. The researcher should decide whether the MAR assumption is plausible. Although the MAR assumption is a suitable starting point in many practical cases, there are also cases where the assumption is suspect. Multiple imputation for NMAR data requires additional modeling assumptions which influence the generated imputations.

2. The form of the imputation model needs to be specified. The form encompasses both the structural part and the assumed error distribution. It should be specified for each incomplete column in the data.

3. The set of variables to include as predictors in the imputation model is the next concern. The general advice is to include as many as possible relevant variables, including their interactions.

4. The imputation of variables that are functions of the other (incomplete) variables is the next step. Since many dataset contain transformed variables, sum scores, interaction variables and ratios, it is useful to incorporate the transformed variables into the multiple imputation algorithm. MICE has a special mechanism called passive imputation. It maintains the consistency among different transformations of the same data. It can be used to ensure that the transform always depends on the most recently generated imputation in the original untransformed data.

5. The order in which variables should be imputed is important. The default MICE algorithm imputes incomplete columns in the data in order from left to right.

6. The number of iterations and the starting imputation has to be setup. The convergence of the Gibbs sampler can be monitored in many ways. One usual method is to plot one or more parameters against the number of iterations. The functions in MICE produce D parallel imputation streams. When convergence is achieved,

the different streams should be freely intermingled with each other and should not show any definite trends or patterns.

7. The number of multiply-imputed dataset, D, needs to be determined. If D is set too low, it will lead to under coverage and low P-values, especially if the percentage of missing data is high.

These choices are always needed but the default choices are not necessarily the best choices for all types of data. The advantage of using MICE is its ability to handle different variable types (continuous, binary, unordered categorical and ordered categorical) because each variable is imputed using its own imputation model. The MICE package has options to modify the default settings according to researcher needs and convenience, and supplies a number of built-in elementary imputation methods, listed in Table 2.1. The package distinguishes between three types of variables which are numeric, binary (factors with 2 levels) and categorical (factors with more than two levels). Table 2.1 shows the variable types and corresponding default imputation methods.

Table 2.1: Buit-in Imputation methods in MICE

Method	**Description**	**Scale type**	**Default**
pmm	Predictive mean matching	numeric	Y
norm	Bayesian linear regression	numeric	
norm.nob	Linear regression (non Bayesian)	numeric	
norm.boot	Linear regression using bootstrap	numeric	
norm.predict	Linear regression using predicted values	numeric	
mean	Unconditional mean imputation	numeric	
2l.norm	Two-level normal imputation	numeric	
2l.pan	Two-level normal imputation using pan	numeric	
2lonly.norm	Imputation at level-2 by Bayesian linear regression	numeric	
2lonly.pmm	Imputation at level-2 by Predictive mean matching	any	
quadratic	Imputation of quadratic terms	numeric	
logreg	Logistic regression	factor, 2 levels	Y
logreg.boot	Logistic regression using bootstrap	factor, 2 levels	
polyreg	Polytomous (unordered) regression	factor, >2 levels	Y
polr	Proportional odds model	ordered, $\geqslant$ 2 levels	
lda	Linear discriminant analysis	factor, $\geqslant$ 2 levels	
sample	Random sample from the observed data	any	

The predictive mean matching (pmm) method is a general semi-parametric imputation method and it is a hot deck imputation method. When imputing a variable x_1 using variables $x_2, ..., x_k$ as predictors, it imputes a value randomly from a set of observed

values whose predicted values are closest to the predicted value for the missing value from the simulated regression model. The observed value from this "match" is used as the imputed value. According to Yu et al. [2011], this method can fail when the percentage of missing is high or when the missing values fall outside the range of the observed data. Besides that, the imputed values are restricted to the observed values and it can preserve non-linear relations even if the structural part of the imputation model is wrong. The disadvantage of this method is that it may fail to produce enough between-imputation variability if the number of predictors are small.

The methods "norm" and "norm.nob" are stochastic regression imputation methods that impute according to a linear imputation model. The "norm" method imputes univariate missing data using Bayesian linear regression analysis with normal errors whereas "norm.nob" imputes univariate missing data using linear regression analysis. Both methods are fast and efficient if the residuals are close to normal. The "norm.nob" method creates an imputation using the spread around the fitted linear regression line [van Buuren and Groothuis-Oudshoorn, 2011] but does not incorporate the variability of the regression coefficients. For small samples, there are variability in the estimation of the imputed data, therefore underestimated. In an easy way, we might say that "norm" is a Bayesian method and "norm.nob" is a non Bayesian method.

The "norm.predict" method is a regression imputation method that imputes missing data using the predicted value from a linear regression. It calculates regression coefficients from the observed data and returns the predicted values as imputations. This is different from the "norm.nob" method. The "norm.nob" imputes a value using the spread around the fitted linear regression line not just the point predictor.

2.4 Model Selection Criteria

Model selection is the process of selecting a best model from a set of candidate models. Model selection provides formal support to guide the user in their search for the best model and to determine which factors/variables to be included when making predictions. Model selection is an important part of the model-building process and cannot be separated from the rest of the analysis in choosing a best model. There are a few general issues involved in model selection and model averaging which are described below [Claeskens and Hjort, 2008].

(i) *Models are approximations*: In dealing with the issues of model-building and model selection, it needs to be understood that in most situations we will not be able to guess the 'correct' or 'true' model. This true model, which generated the collected

data, might be very complex and is always unknown. G.E.P Box expressed a view that 'All models are wrong, but some are useful' and most model selection methods were derived from this perspective.

(ii) *The bias-variance trade-off*: In model fitting and model selection, the bias and variance trade-off takes the form of balancing simplicity (fewer parameters to estimate, leads to lower variability but higher modelling bias) against complexity (including more parameters which means a higher degree of variability but smaller modelling bias). Statistical model selection must strike a proper balance between over-fitting and under-fitting.

(iii) *Parsimony*: Only important parameters should be included in a selected model.

(iv) *The context*: All modelling is rooted in a suitable scientific context and is undertaken for a certain purpose which differs from researcher to researcher. Different researchers might have different preferences in aims and purposes when building a model and analysing data. Therefore, there are different model selection methods to choose a best model.

(v) *The focus*: It is important to focus model-building and model selection efforts on criteria that favour a good performance precisely and efficiently. For the same data and same list of possible models, a different aim will lead to a different selected model.

(vi) *Conflicting recommendations*: Different model selection strategies might end up offering different selected models. Therefore, it is important to learn how the selection schemes are constructed and what are their aims and properties.

(vii) *Model averaging*: In general, model selection strategies work by assigning a certain score to each candidate model. Often there is a clear best model but sometimes there will be several selected models that do almost as well as the chosen best model. In these cases, it is important to combine inference outputs across these best models.

In general, most model selection methods are defined in terms of a suitable *information criterion*, a mechanism that uses data to give each possible model a certain score. These criteria are based on some optimal principle such as minimizing the error sum of squares (SSE) or maximizing likelihood values. A common type of criterion takes the form of the error sum of squares (SSE) multiplied by a penalty factor that depends on the model complexity as measured by the number of parameters to be estimated. A more complex model will reduce the SSE but increase the penalty. A model with a lower value of the criterion is judged to be preferable. It is possible that combining two or more criteria

might produce better results than using any single criterion. Rust et al. [1995] suggested that a combination of model selection criteria can become 'more sure' of which model is correct.

2.4.1 Stepwise selection of variables

Variable selection is designed to select the best variables. The principle of Occam's Razor states that among several reasonable explanations for a phenomenon, the simplest is best. This implies that the smallest model that fits the data adequately is best. Unnecessary variables in the prediction model will add noise to the estimation of other quantities that researchers are interested in and too many variables in the model can cause multicollinearity [Davison, 2003]. In order to overcome these problems, researchers usually use variable selection to choose variables from among a set of candidate variables. Typically, variable selection will be implemented through iterative procedures like forward, backward and stepwise selection.

Forward selection is a procedure in which variables are sequentially entered into the model. The procedure takes the null model as baseline with an intercept only. Each candidate variable is added separately to this null model. The model carried forward to the next stage where the null model augmented by the variable that most reduces the SSE. Each of the remaining variables is added separately to the new base model and the process is continued [Davison, 2003]. The process is stopped at any stage when the F-statistic for the largest reduction in sum of squares is not significant.

Backward selection is a procedure which starts with all the variables entered into the equation and consecutively remove the least significant variable at each stage. The process will stop when no term can be deleted without increasing the SSE significantly [Davison, 2003]. It is just the reverse of forward selection. The backward selection method is preferable because its initial estimate of the error variance σ^2 will be better than the forward selection method. Both methods might choose different best models.

Stepwise selection is a combination of backward and forward selection. At each step, a variable will be added, removed from the model, or swapped with a variable that was not in the model or the process will be stopped [Davison, 2003]. Stepwise selection is computationally easier, easy to explain and widely used by many researchers. There are some drawbacks of using stepwise selection. Since variables are removed or added one at a time, it is possible to miss the optimal model. Stepwise selection tends to choose models that are smaller than desirable for prediction. The stepwise selection method will yield a single final model although in practice there are often several equally good models.

Moreover, Harrell [2001] identified few crucial problems of using stepwise variable selection. This method yields standard errors of regression coefficient estimates that are biased low and confidence intervals for effects and predicted values that are falsely narrow. The choice of the variables to be included depends on estimated regression coefficients rather than their true values, so X_j is more likely to be included if its regression coefficients is overestimated than if its regression coefficient is underestimated. Moreover, stepwise variable selection is made arbitrary by collinearity. The problems of $p-$value based variable selection are worsen when the analyst interprets the final model as if it were pre-specified. All subset regression was introduced to overcome some of the issues related to stepwise variable selection.

2.4.2 All subset regression

All subset model selection is designed to select the best subset of variables and it compares all possible models using a specified pool of explanatory variables. All subset regression is an alternative to the stepwise selection method. When using this approach, a researcher first decides on the range of subset sizes that could be considered to be useful. Consider p as number of parameters in a regression model. With $p-1$ explanatory variables, there are 2^{p-1} possible regression models to be fitted. For example, consider two explanatory variables, X_1 and X_2 in a linear regression analysis. There are four possible models including the null model.

There are several different criteria that can be used for ordering variable subsets or possible models in terms of goodness of fit. The commonly used criteria are multiple R^2, adjusted R^2, and Mallow's C_p. Choosing a model to maximize the multiple R^2 or adjusted R^2 was proposed in the earliest research on model-building. When all subset regression is used in parallel with stepwise selection, the multiple R^2 statistic allows direct comparison of the best possible model identified using each approach [Chatterjee and Simonoff, 2013].

Mallow's C_p criterion was designed to estimate the expected squared prediction error of a model, and in that sense a model that minimizes the C_p criterion will be chosen as the best model. A disadvantage of using the C_p criterion is that its value depends on the pool of all candidate variables, so adding variables that provide no predictive power can change the choice of best model. According to Claeskens and Hjort [2008], the adjusted R^2 and C_p criteria are only suitable in model selection for linear models with normal data. Therefore, many researchers developed other model selection criteria such as Akaike's information criterion (AIC) and Bayesian information criterion (BIC). The details of these criteria will be discussed in the next sections.

2.4.3 Kullback-Leibler distance

Kullback-Leibler distance is a way of measuring the statistical distance from one probability density to another [Claeskens and Hjort, 2008]. If data Y are realisations of independent and identically distributed random variables, the likelihood and log-likelihood functions can be written in terms of the density $f(y, \boldsymbol{\theta})$ for an individual observation as

$$\mathcal{L}_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i, \boldsymbol{\theta}) \tag{2.11}$$

and

$$\ell_n(\boldsymbol{\theta}) = log\mathcal{L}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} logf(y_i, \boldsymbol{\theta}). \tag{2.12}$$

Here θ is a vector of unknown parameters. It is important to make a distinction between the model $f(y, \boldsymbol{\theta})$ that the researcher constructs for the data and the true density $g(y)$ of the data, which is nearly always unknown. The density $g(\cdot)$ is called the data-generating density. Although there are several ways of measuring closeness of a parametric approximation $f(\cdot, \boldsymbol{\theta})$ to the true density g, the distance intimately linked to the maximum likelihood method is Kullback-Leibler (KL) distance. It can be written as

$$KL(g, f(\cdot, \boldsymbol{\theta})) = \int g(y) log \frac{g(y)}{f(y, \boldsymbol{\theta})} dy. \tag{2.13}$$

Equation(2.13) can be written equivalently as

$$KL(g, f(\cdot, \boldsymbol{\theta})) = \int g(y) log\left(g(y)\right) dy - \int g(y) log\left(f(y, \boldsymbol{\theta})\right) dy \tag{2.14}$$

where each of the two terms on the right of the Equation(2.14) is a statistical expectation with respect to $g(\cdot)$. Thus,

$$KL(g, f(\cdot, \boldsymbol{\theta})) = E_g[log\left(g(y)\right)] - E_g[log\left(f(y, \boldsymbol{\theta})\right)] \tag{2.15}$$

The first expectation $E_g[log\left(g(y)\right)]$ is a constant across all possible fitted models, thus,

$$KL(g, f(\cdot, \boldsymbol{\theta})) = constant - E_g[log\left(f(y, \boldsymbol{\theta})\right)].$$

The relative KL distance is

$$KL(g, f(\cdot, \boldsymbol{\theta})) - constant = -E_g[log\left(f(y, \boldsymbol{\theta})\right)].$$

Akaike proposed Kullback-Leibler distance as a fundamental basis for model selection procedures. However, KL distance cannot be calculated without full knowledge of both

g (full reality) and the parameters θ in each of the candidate models $f(y, \boldsymbol{\theta})$. Akaike found that the double expectation [Claeskens and Hjort, 2008]

$$E\, E_g[log\,(f(y, \boldsymbol{\theta}))]$$

can be estimated and there is a relationship between the relative KL distance and the maximized log-likelihood.

2.4.4 Akaike's information criterion (AIC) and AIC$_c$

Akaike's information criterion (AIC) is among the most popular and versatile strategies for model selection. An asymptotically unbiased estimator of the relative, expected KL distance, $log\left(\mathcal{L}\left(\hat{\boldsymbol{\theta}} \mid y\right)\right) - p$ was multiplied by 2 to become [Claeskens and Hjort, 2008]

$$AIC = 2log\left(\mathcal{L}\left(\hat{\boldsymbol{\theta}} \mid y\right)\right) - 2p.$$

where the expression $log\left(\mathcal{L}\left(\hat{\boldsymbol{\theta}} \mid y\right)\right)$ is the numerical value of the log-likelihood at its maximum point [Burham and Anderson, 2002]. AIC was designed to be an approximately unbiased estimator of the expected Kullback-Leibler distance of a fitted model. In general, AIC for each possible model $\mathscr{M}$ is

$$AIC(\mathscr{M}) = 2logL(\mathscr{M}) - 2p \tag{2.16}$$

where $L(\mathscr{M})$ is the maximized value of the likelihood function of model $\mathscr{M}$ and p is the number of parameters in model $\mathscr{M}$. The model with the highest AIC score will be selected. The direct comparison of obtained maximum log-likelihood values for different models is not good for model selection. Including more parameters in a model always increases the maximum likelihood value [Claeskens and Hjort, 2008]. AIC acts as a penalised log-likelihood criterion, affording a balance between good fit (high value of log-likelihood) and complexity (complex models are penalised more than simple ones). The penalty term punishes the models for being too complex in the sense of containing many parameters. Akaike's method aims at finding models that have few parameters but fit the data well.

An important special case is the normal linear model, defined by

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \varepsilon_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + \varepsilon_i \tag{2.17}$$

for $i = 1, 2, \cdots, n$ with $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ independently drawn from $N(0, \sigma^2)$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^t$ a vector of regression coefficients. Here p is the number of parameters in the $\boldsymbol{\beta}$ vector. Often, $x_{i,1} = 1$ for all i, making β_1 an intercept parameter. The log-likelihood function is

$$\begin{aligned} log L(\mathscr{M}) =& \ell_n(\beta, \sigma) \\ =& \sum_{i=1}^{n} \left\{ -log\,\sigma - \frac{1}{2}\frac{(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})^2}{\sigma^2} - \frac{1}{2}log(2\pi) \right\} \\ =& -n\,log\,\sigma - \frac{n}{2}\,log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})^2 \end{aligned} \tag{2.18}$$

In general, an estimator of σ^2 might be found as [Claeskens and Hjort, 2008]

$$\hat{\sigma}^2 = \frac{\|res\|^2}{n-a} = \frac{RSS}{n-a} \tag{2.19}$$

with the cases $a = 0$ and $a = p$ corresponding to maximum likelihood and unbiased estimation respectively. RSS is the residual sum of squares. When $a = 0$, plugging $\hat{\sigma}$ into Equation (2.18),

$$\begin{aligned} \ell_n(\hat{\beta}, \hat{\sigma}) &= -n\,log\,\hat{\sigma} - \frac{n}{2}\,log(2\pi) - \frac{1}{2\hat{\sigma}^2}RSS \\ &= -n\,log\,\hat{\sigma} - \frac{n}{2}\,log(2\pi) - \frac{n}{2} \end{aligned} \tag{2.20}$$

Therefore, for model (2.17)

$$\begin{aligned} AIC &= 2\left(-n\,log\,\hat{\sigma} - \frac{n}{2}\,log(2\pi) - \frac{n}{2}\right) - 2(p+1) \\ &= -2n\,log\,\hat{\sigma} - n\,log(2\pi) - n - 2(p+1) \end{aligned} \tag{2.21}$$

since $p + 1$ is the number of parameters in $(\boldsymbol{\beta}, \sigma)$.

AIC is intended to be an approximately unbiased estimator of the expected Kullback-Leibler distance of a candidate model. However, AIC suffers from a potentially high degree of negative bias when used with samples that are small in size relative to the number of parameters in the fitted model. According to Hurvich and Tsai [1989], as the number of parameters (p) increases in comparison to sample size (n), AIC becomes a strongly negatively-biased estimator. This negative bias in AIC limits its effectiveness as a model selection criterion and can lead to over-fitting (i.e. fitting a larger model than required) especially when $\frac{p}{n}$ is large for some candidate models. On the other hand, when the sample size is large and the dimension of the candidate model is small, AIC works better as an approximately unbiased estimator. Hurvich and Tsai [1989] proposed the corrected Akaike information criterion, AIC_c, to get around the problem

with small samples. AIC_c is an adjusted version of AIC that was originally proposed for linear regression with normal errors. AIC_c in general is [Claeskens and Hjort, 2008]

$$AIC_c = 2logL(\mathscr{M}) - 2p\frac{n}{n-p-1}$$

It can be written in maximised log-likelihood form for model (2.17) as

$$AIC_c = 2\ell_n(\hat{\beta}, \hat{\sigma}) - 2(p+1)\frac{n}{n-p-2} \tag{2.22}$$

where $\hat{\beta}$ and $\hat{\sigma}$ are maximum likelihood estimates of β and σ. Plugging Equation (2.18) into Equation (2.22), then

$$\begin{aligned} AIC_c &= 2\left(-n\,log\,\hat{\sigma} - \frac{n}{2}\,log(2\pi) - \frac{n}{2}\right) - 2(p+1)\frac{n}{n-p-2} \\ &= -2n\,log\,\hat{\sigma} - n\,log(2\pi) - n - 2(p+1)\frac{n}{n-p-2} \end{aligned} \tag{2.23}$$

It has been suggested that researchers should use AIC_c when the ratio $\frac{n}{p}$ (< 40) is small. If this ratio is sufficiently large, then AIC and AIC_c are similar and tend to choose the same model [Burham and Anderson, 2002]. Alternatively, σ^2 can be estimated from Equation (2.19) with $a = p + 2$,

$$(\hat{\sigma}^*)^2 = \frac{\|res\|^2}{n-p-2} \tag{2.24}$$

Plugging $\hat{\sigma}^*$ into Equation (2.18),

$$\begin{aligned} \ell_n(\hat{\beta}, \hat{\sigma}^*) &= -n\,log\,\hat{\sigma}^* - \frac{n}{2}\,log(2\pi) - \frac{1}{2(\hat{\sigma}^*)^2}RSS \\ &= -n\,log\,\hat{\sigma}^* - \frac{n}{2}\,log(2\pi) - \frac{n-p-2}{2} \end{aligned} \tag{2.25}$$

Given AIC_c^* in general is [Claeskens and Hjort, 2008]

$$AIC_c^* = 2\ell_n(\hat{\beta}, \hat{\sigma}^*) - 2(p+1) \tag{2.26}$$

Plugging Equation (2.25) into Equation (2.26),

$$\begin{aligned} AIC_c^* &= 2\left(-n\,log\,\hat{\sigma}^* - \frac{n}{2}\,log(2\pi) - \frac{n-p-2}{2}\right) - 2(p+1) \\ &= -2n\,log\,\hat{\sigma}^* - n\,log(2\pi) - (n-p-2) - 2(p+1) \\ &= -2n\,log\,\hat{\sigma}^* - n\,log(2\pi) - n - p \end{aligned} \tag{2.27}$$

For a constant p and sufficiently large sample size n, $\hat{\sigma}^2$ and $(\hat{\sigma}^*)^2$ will converge to the same value (the maximum likelihood estimate). Therefore, for a sufficiently large n, all

three criteria (AIC, AIC_c and AIC_c^*) will converge and tend to choose the same model. The advantage of AIC_c over AIC is the application in small to medium-sized samples [Burham and Anderson, 2002, Claeskens and Hjort, 2008]. Therefore, a researcher might choose always to use the AIC_c as the model selection criterion.

2.4.5 Bayesian information criterion (BIC)

The Bayesian information criterion (BIC) takes the form of a penalised log-likelihood function where the penalty is equal to the logarithm of the sample size times the number of estimated parameters in the model. The general form of BIC is [Claeskens and Hjort, 2008]

$$BIC(\mathscr{M}) = 2logL(\mathscr{M}) - (log\, n)p \tag{2.28}$$

where $L(\mathscr{M})$ is the maximized value of the likelihood function of model $\mathscr{M}$, p is the number of parameters in model $\mathscr{M}$ and n is the sample size of the data. The model with the largest BIC value will be chosen as the best model. The 'B' in BIC is for 'Bayesian' where the $L(\mathscr{M})$ is an approximation to marginal likelihood or marginal density for model $\mathscr{M}$ under certain prior. The specification of priors for all models and for all parameters in the model models are required for a practical approximation in Bayesian model comparison [Burham and Anderson, 2002]. In Bayesian model comparison, a Bayesian procedure will select a model which is a posteriori most likely when there are different possible models. This model is identified by calculating the posterior probability of each model and selecting the model with the biggest posterior probability.

Note that the BIC formula as in Equation (2.28) only uses the maximised log-likelihood function. It was derived in this way so that no prior information is needed to obtain the BIC values [Claeskens and Hjort, 2008]. Both criteria, AIC in Equation (2.16) and BIC in Equation (2.28) are constructed as twice the maximized log-likelihood value minus a penalty for the complexity of the model. The BIC's penalty is larger than AIC for all n at least 8. This shows that the BIC more strongly discourages choosing a model with many parameters.

Both AIC and BIC can be written in a general for model $\mathscr{M}$ [Burham and Anderson, 2002]

$$I_{\mathscr{M}} = 2logL(\mathscr{M}) - c_{n,p} \tag{2.29}$$

where $L(\mathscr{M})$ is the maximized value of the likelihood function of model $\mathscr{M}$
$c_{n,p}$ is the penalty term for model $\mathscr{M}$
p is the number of parameters in model $\mathscr{M}$
n is the sample size of the data

For example, $c_{n,p}$ for AIC will be $2p$. Since the BIC penalty is stricter than the AIC, bigger models (with larger numbers of parameters) will receive a heavier 'punishment'. When the sample n gets larger, the heavier the penalty used in the BIC. Especially for large sample size, a researcher can expect that there will be a difference in the ranks of models when comparing model selection by AIC and BIC.

2.4.6 Model selection criteria for missing data

The challenge in missing data problems is to obtain a suitable and accurate approximation to the observed data likelihood, which does not involve intractable multiple integration, and directly maximize it and compute AIC or BIC. A version of AIC that can deal with models with incomplete covariates was constructed based on EM algorithm. Consider a design matrix of covariate values as $\boldsymbol{X} = (\boldsymbol{X}_{obs}, \boldsymbol{X}_{mis})$, clearly separating the set of fully observed covariates $\boldsymbol{X}_{obs}$ and those that contain at least one missing observation. Assume that the response vector $\boldsymbol{Y}$ is completely observed. The model selection criterion for missing data problems is based on the observed data likelihood $L(\boldsymbol{X}_{obs}|\theta)$. The model selection criterion based on the general EM algorithm [Ibrahim et al., 2008] is

$$\begin{aligned} IC_{H,Q} =& 2\ell(\boldsymbol{Y}, \boldsymbol{X}_{obs}|\hat{\boldsymbol{\theta}}) - \hat{c}_n(\hat{\boldsymbol{\theta}}) \\ =& 2Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) - 2H(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) - \hat{c}_n(\hat{\boldsymbol{\theta}}) \end{aligned} \tag{2.30}$$

(see Equation (2.4)) and the model selection criterion based on a Hermite approximation is

$$IC_{\tilde{H}(k),Q} = 2Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) - 2\tilde{H}(k|\hat{\boldsymbol{\theta}}) - \hat{c_n}(\hat{\boldsymbol{\theta}}) \tag{2.31}$$

where $\hat{c_n}(\hat{\boldsymbol{\theta}})$ is a penalty term that is a function of the data and the fitted model, and k is a polynomial order of approximation using a truncated Hermite approximation. The $\ell(\boldsymbol{Y}, \boldsymbol{X}_{obs}|\hat{\boldsymbol{\theta}})$ will be computed from the $Q-$function $Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})$ and the $H-$function $H(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})$ at EM convergence from EM output as discussed in Section 2.2.4. The $\tilde{H}(k|\hat{\boldsymbol{\theta}})$ can be obtained from the Hermite approximation as discussed by Ibrahim et al. [2008]. Since $\tilde{H}(k|\hat{\boldsymbol{\theta}}) \leq H(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})$ according to Jensen's inequality, $IC_{\tilde{H}(k),Q} \leq IC_{H,Q}$ and $\tilde{H}(k|\hat{\boldsymbol{\theta}})$ converges to $H(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})$ as $k \to \infty$. However, it is computationally inefficient to choose a large k. When $\hat{c_n}(\hat{\boldsymbol{\theta}}) = 2p$, the model selection criteria AIC is obtained. The model selection criterion BIC is obtained when $\hat{c_n}(\hat{\boldsymbol{\beta}}) = plog(n)$.

Another version of this model selection criterion that does not involve the H-function, whose components depend only on quantities obtained directly from EM output was proposed by Claeskens and Consentino [2008]. This criterion was proposed to avoid the need for an analytic approximation to the integrand of the H-function since its

computation will be cumbersome for large k. The proposed model selection criterion is

$$AIC_1 = 2Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) - 2p \tag{2.32}$$

where p is the number of parameters in the model. The $Q-$function, $Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})$ can be obtained at EM convergence from EM output where

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(\tau)}) = \int \ell(\boldsymbol{X};\boldsymbol{\theta}) f(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs},\boldsymbol{\theta}^{(\tau)}) dX_{mis} \tag{2.33}$$

The corrected AIC_c for small sample size as in Equation (2.22) can be derived in terms of the $Q-$function for missing data problem [Claeskens and Consentino, 2008] as

$$AIC_{1,c} = 2Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) - 2p\frac{n}{n-p-1} \tag{2.34}$$

where p is the number of parameters in the model. The advantage of using these proposed AIC statistics without the H-function is that they are computationally easier than $IC_{\tilde{H}(k),Q}$ since they do not require an approximation to the integrand of the H-function. However, a model selection criterion based on the Q-function alone can overstate the amount of information in the missing data compared with the observed data log-likelihood function. Omitting the H-function can lead to a criterion with poor model selection properties in some cases, especially when the missing data fraction is high [Ibrahim et al., 2008].

Chaurasia and Harel [2012] identified that the issue in model selection with imputed data is how to combine model selection results from imputed data and also the impact of the assumed imputation model on model selection in the analysis phase. Therefore, two AIC variants were considered for multiply-imputed dataset which are based on the Arithmetic Mean (AM) and Geometric Mean (GM) of the D point estimates. The AIC variant based on AM is [Chaurasia and Harel, 2012]

$$AIC_{AM} = n\,ln\left(D^{-1}\sum_{d=1}^{D} s_d^2\right) - 2p \tag{2.35}$$

and the AIC variant based on GM is

$$AIC_{GM} = \frac{n}{m}\sum_{d=1}^{D} ln\left(s_d^2\right) - 2p \tag{2.36}$$

where s_d^2 represents the maximum likelihood estimate of σ^2 from the dth imputed dataset, $d = 1, 2, ..., D$.

2.4.7 Comparison of model selection criteria

There are three key properties of model selection criteria which are consistency, efficiency and parsimony. If there exists one true model that generated the data and this model is one of the candidate models, then researcher would expect that the model selection method would identify this true model. This is related to consistency. A model selection method is weakly consistent if, with probability tends to one as the sample size tends to infinity, the selection method is able to select the true model from the possible models. Strong consistency is obtained when the selection of the true model almost surely happens. Another property of an information criterion is that it behaves 'almost as well' as the true model in terms of mean squared error or expected squared prediction error. Such a model selection method is called efficient. Consistency and efficiency of a criterion cannot occur together since a consistent criterion can never be efficient [Claeskens and Hjort, 2008].

One underlying purpose of model selection is to use the information criterion to select the model that is closest to the true model. According to Claeskens and Hjort [2008], the Kullback-Leibler distance can be used to measure the distance from the true density to the model density. If two or more models minimize the KL distance, then the researcher will select the model with fewest parameters. This is called the most parsimonious model.

Suppose there is a single model among the candidate models which reaches the minimum KL distance, then weak consistency is achieved by any information criterion whose penalty, $c_{n,p}$ divided by n, tends to zero as the sample size increases. Both BIC, with $c_{n,p} = p(log\, n)$, and AIC, with $c_{n,p} = 2p$, are weakly consistent.

If there are two or more candidate models which reach the minimum KL distance, then parsimony means that the simplest model (the model with fewest parameters) should be chosen from among these models. This parsimony property is sometimes called consistency in the literature. Consistency is really the condition that, with probability tends to one, the model selection criterion will select the smallest model in these circumstances. In other words, a model selection criterion is consistent if it is able to determine the order of the true model with enough data. The BIC penalty satisfies this condition, but the AIC penalty fails. Note that any criterion with a penalty that does not depend on sample size cannot satisfy the consistency property. Claeskens and Hjort [2008] stated that AIC will not necessarily choose the most parsimonious model and there is a probability of overfitting. This means AIC will often choose a model with more parameters than actually needed.

According to Hurvich and Tsai [1990], although BIC is consistent, it has poor small sample performance, whereas AIC has quite satisfactory small sample performance. The convergence rate is the rate at which the number of covariates in the selected model converges to its limiting value. AIC converges quickly to an over fitted model but BIC converges at a very slow speed to the correct value.

In conclusion, both AIC and BIC have good properties, in that AIC is efficient and BIC is consistent. Note that BIC will choose a parsimonious model because of the penalty term. The BIC's penalty is more strict than AIC and it strongly discourages choosing a model with many parameters. Whereas AIC chooses a model with more parameters, so there is a chance of over fitting. As discussed in the previous section, for a sufficiently large n, both AIC and AIC_c will converge and tend to choose the same model. The advantage of AIC_c over AIC is the application in small to medium-sized samples [Burham and Anderson, 2002]. Therefore, a researcher might choose always to use AIC_c and BIC as the model selection criterion. Further details on the performance of AIC_c and BIC will be discussed in Chapter 4 based on simulation study.

2.5 Model Averaging

Model selection is well known for introducing additional uncertainty into the model-building process. The properties of standard parameter estimates obtained from the selected model do not reflect the stochastic nature of the model selection process. Model averaging is an alternative to model selection intended to overcome the under-estimation of standard errors that is a consequence of model selection. A model average estimator weighs across all possible models rather than picking a single best model. Model averaging will shrink the estimates of the weaker variables and will yield better predictions. The 'better' models will receive higher weights. Suppose that there are M candidate models. In one approach, the weight $w_{\mathscr{M}}$ for model is [Buckland et al., 1997]

$$w_{\mathscr{M}} = \frac{exp\left(\frac{I_{\mathscr{M}}}{2}\right)}{\sum_{\mathscr{M}=1}^{M} exp\left(\frac{I_{\mathscr{M}}}{2}\right)} \tag{2.37}$$

where $I_{\mathscr{M}}$ is model selection criterion for model $\mathscr{M}$ as in Equation (2.29) and $\sum_{\mathscr{M}=1}^{M} w_{\mathscr{M}} = 1$. The estimate of a parameter β_p is

$$\hat{\beta}_p = \sum_{\mathscr{M}=1}^{M} w_{\mathscr{M}} \hat{\beta}_{(p,\mathscr{M})} \tag{2.38}$$

where $\hat{\beta}_{(p,\mathscr{M})}$ is the estimate of β_p under model $\mathscr{M}$ for $\mathscr{M} = 1, 2, ..., M$. Different researchers have suggested weights based on AIC [Buckland et al., 1997], Mallows criterion [Hansen, 2007] and the Focussed Information criterion [Hjort and Claeskens, 2003]. In this research, the modified weights will be used based on model selection criteria AIC, AIC_c and BIC. A modification was carried out for calculating the weights in order to avoid numerical error. The weights $w_{\mathscr{M}}$ were calculated as

$$w_{\mathscr{M}} = \frac{exp\left(\frac{I_{\mathscr{M}} - \bar{\ell}}{2}\right)}{\sum_{\mathscr{M}=1}^{M} exp\left(\frac{I_{\mathscr{M}} - \bar{\ell}}{2}\right)} \tag{2.39}$$

where $\bar{\ell} = \frac{1}{M}\sum_{\mathscr{M}=1}^{M} \ell_{\mathscr{M}}$ with $\ell_{\mathscr{M}}$ is log-likelihood function of model $\mathscr{M}$ for $\mathscr{M} = 1, 2, ..., M$. A general model averaging estimator for linear models after multiple imputation is [Schomaker and Heumann, 2014]

$$\hat{\beta}_p^{(MI)} = \frac{1}{D}\sum_{d=1}^{D} \hat{\beta}_p^{(d)} \tag{2.40}$$

with

$$\hat{\beta}_p^{(d)} = \sum_{\mathscr{M}=1}^{M} w_{\mathscr{M}}^{(d)} \hat{\beta}_{(p,\mathscr{M})}^{(d)} \tag{2.41}$$

and a set of candidate models, $\mathscr{M} = 1, 2, ..., M$. When carrying out model averaging along with multiple imputation, the parameters of a linear model are estimated using Equation (2.40) and Equation (2.41). The estimated variance of these estimators is

$$\begin{aligned}\widehat{Var}\left(\hat{\beta}_p^{(MI)}\right) = \frac{1}{D}\sum_{d=1}^{D}\left(\sum_{\mathscr{M}=1}^{M} w_{\mathscr{M}}^{(d)}\sqrt{\widehat{Var}\hat{\beta}_{(p,\mathscr{M})}^{(d)} + (\hat{\beta}_{(p,\mathscr{M})}^{(d)} - \hat{\beta}_p^{(d)})^2}\right)^2 \\ + \frac{D+1}{D(D-1)}\sum_{d=1}^{D}\left(\hat{\beta}_p^{(d)} - \hat{\beta}_p^{(MI)}\right)^2\end{aligned} \tag{2.42}$$

When carrying out model averaging along with multiple imputation in the context of a logistic regression model, predicted probabilities must be estimated in a similar way. Letting P_t denote the probability of success at a particular set of covariate values, then P_t is estimated as follows [Schomaker and Heumann, 2014]

$$\hat{P}_t^{(MI)} = \frac{1}{D}\sum_{d=1}^{D} \hat{P}_t^{(d)} \tag{2.43}$$

with

$$\hat{P}_t^{(d)} = \sum_{\mathscr{M}=1}^{M} w_{\mathscr{M}}^{(d)} \hat{P}_{(t,\mathscr{M})}^{(d)} \tag{2.44}$$

and a set of candidate models, $\mathscr{M} = 1, 2, ..., M$.

2.6 Bias and Mean Squared Error of Prediction

An essential part of model-building is evaluation of the model and its estimators. Researchers usually will use bias and mean squared error (MSE) to measure the performance of estimators. An unbiased estimator and minimum value of MSE is desired for a good estimator. When making predictions, the performance of the prediction model should be measured. The performance of a model can be measured using mean squared error of prediction (MSE(P)). The details of each of these measures will be discussed in the following sections.

2.6.1 Bias

Bias is the difference between the expected value of an estimator and the true value of the parameter being estimated. An estimator is called an unbiased estimator if its expectation is equal to the true value, and the observed value from a particular sample is referred to as an unbiased estimate. In other words, an estimator with the bias identically equal to 0 is called an unbiased estimator and it satisfies $E(\hat{\beta}) = \beta$ [Everitt, 2006]. Mean squared error (MSE) is the average squared difference between the estimator $\hat{\beta}$ and the parameter β. MSE is a measure of performance for an estimator. MSE of an estimator in general is [Burham and Anderson, 2002]

$$MSE(\hat{\beta}) = E\left[\left(\hat{\beta} - \beta\right)^2\right] \tag{2.45}$$

MSE is also called the risk function of an estimator, with $\left(\hat{\beta} - \beta\right)^2$ called the quadratic loss function. MSE has two components: one measures the variability of the estimator (precision) and the other measures the bias (accuracy). An estimator that has good MSE properties has small combined variance and bias. The MSE can be rewritten as

$$MSE(\hat{\beta}) = E\left[\left(\hat{\beta} - \beta\right)^2\right] = Var(\hat{\beta}) + [bias(\hat{\beta})]^2 \tag{2.46}$$

and the bias of an estimator is

$$bias(\hat{\beta}) = \sqrt{E\left[\left(\hat{\beta} - \beta\right)^2\right] - Var(\hat{\beta})} \tag{2.47}$$

An estimator is unbiased if the MSE is equal to its variance, $MSE(\hat{\beta}) = E\left[\left(\hat{\beta} - \beta\right)^2\right] = Var(\hat{\beta})$. According to Claeskens and Hjort [2008], MSE of the estimators of possible models can be used as measures of quality of possible models. Lower MSE is desirable in considering a better estimator. A good estimator requires good precision as well as good accuracy.

2.6.2 Mean squared error of prediction

An essential aspect of model evaluation is accuracy of prediction, so a reasonable measure for evaluating a model is its mean squared error of prediction (MSE(P)). In general, the MSE(P) is [Mevik and Cederkvist, 2004, Wallach and Goffinet, 1989]

$$MSE\,(P) = \frac{1}{t}\sum_{i=1}^{t}(\hat{y}_t - y_t)^2 \tag{2.48}$$

where $\hat{y}_t$ is estimated Y of test values and y_t is the actual test values used for prediction. MSE(P) is usually used to assess the performance of regressions. In Logistic regression, the MSE(P) will be calculated based on predicted and actual probability values rather than using the Y values which only take the value 0 or 1. The calculation of MSE(P) based on binary values mislead the assessment of model performance, effects of simulation parameters and missing data. In order to avoid misleading information about model performance and numerical error, the MSE(P) for Logistic regression will be calculated as

$$MSE\,(P) = \frac{1}{t}\sum_{i=1}^{t}\left(\hat{P}_t - P_t\right)^2 \tag{2.49}$$

where P is the probability of success in generalized linear models.

2.7 Multicollinearity

Collinearity or multicollinearity is described as a condition where two or more predictor variables in a statistical model are linearly related. Dormann et al. [2013] stated that perfect multicollinearity occurs if covariates are exact linear function of each other and is simply a case of model misspecification. Multicollinearity increases the estimates of

parameter variance, produces high R^2 in the face of low parameter significance, and results in parameter estimates with incorrect signs and implausible magnitudes [Mela and Kopalle, 2002]. Multicollinearity will cause unstable estimates and inaccurate variances which affects confidence intervals and hypothesis tests. Multicollinearity is a common problem where there are large numbers of covariates, especially for multiple linear regression.

There are some rules of thumb or indicators that will provide some clues about the existence of multicollinearity in concrete applications. A variance inflation factor (VIF) measures multicollinearity by regressing one independent variable on all of the remaining independent variables. The VIF is

$$VIF = \frac{1}{1 - R^2} \tag{2.50}$$

where R^2 is the coefficient of multiple determination. According to Studenmund [2006], VIF is an index of how much multicollinearity has increased the variance of an estimated coefficient. A high VIF indicates that multicollinearity has increased the estimated variance of the estimated coefficient and decreased the t-statistics. Hocking [2003] suggested that useful indicators of multicollinearity are as following:

- Simple correlation $\mid \rho \mid > 0.95$
- Variance inflation factors, $VIF > 10$

In ordinary least squares, the VIF are the diagonals of the inverse of the $X^T X$ matrix scaled to have unit variance. For models fitted with maximum likelihood estimation, the information matrix is scaled to correlation form and VIF is the diagonal of the inverse of this scaled matrix. This VIF are similar to those from a weighted correlation matrix of the original columns in the design matrix [Harrell, 2001].

2.7.1 Consequences of multicollinearity

Studenmund [2006] stated that the consequences of multicollinearity are as following:

- *Estimates will remain unbiased.* The usual estimates of the β's still will be centred around the true population value if all the assumptions are met for a correctly specified model, even if a model has significant multicollinearity.
- *The variances and standard errors of the estimates will increase.* It is difficult to identify the separate effects of the multicollinearity where it lead to make larger

error in estimating the β's. So as a result, the variance and standard errors will be larger, although the estimated coefficients are still unbiased.

- *The computed t-statistics will fall.* Since the multicollinearity increases the standard error, then t-statistics of estimated coefficients will fall.
- *Estimates will become very sensitive to changes in specification.* When significant multicollinearity exists, the addition or deletion of an independent variable or a few observations will often cause dramatic changes in the values of the $\hat{\beta}$'s.
- *The overall fit of the equation and the estimation of the coefficients of non-multicollinear variables will be largely unaffected.* The overall level of significance of a model is affected far less by multicollinearity than the level of significance of the individual regression coefficients.

In stepwise variable selection and in all subset regression, multicollinearity will cause predictors to compete and make the selection of significant variables arbitrary [Harrell, 2001]. The presence of multicollinearity can lead to drop an important variable from the model because of its low t-statistic.

Multicollinearity will cause problem when attempting to use a fitted regression model for prediction. Simple models tend to predict better than more complex models. If a model with multicollinearity is used for future prediction, the relationships among the independent variables and their relationship with the response variable will remain the same in the future [Chatterjee and Simonoff, 2013]. The variances for general linear models as in model (2.17) with $p = 2$ are

$$var(\hat{\beta}_1) = \sigma^2 \left[\sum_{i=1}^{n} x_{1i}^2 (1 - \rho_{12}^2) \right]^{-1} \tag{2.51}$$

and

$$var(\hat{\beta}_2) = \sigma^2 \left[\sum_{i=1}^{n} x_{2i}^2 (1 - \rho_{12}^2) \right]^{-1} \tag{2.52}$$

where ρ_{12} is the correlation between x_1 and x_2. As correlation increases ($\rho_{12} \to \pm 1$), both variances tends to ∞. Chatterjee and Simonoff [2013] stated that for $\rho_{12} = 0.5$, variance inflation is 1.33 and for $\rho_{12} = 0.999$, variance inflation is 500. This shows how much the variances of estimated slope coefficients are inflated due to observed multicollinearity relative to when predictors are uncorrelated. It is clear that when the correlation is high, the variability of the estimated slopes can increase dramatically.

Besides that, multicollinearity will cause problems when data are missing. Hardt et al. [2012] suggested that inclusion of highly correlated auxiliary variables can improve the

imputation model (used to impute missing data) but inclusion of auxiliary variables with low correlation is not useful. When highly correlated auxiliary variables are used in both imputation and prediction models, it will cause multicollinearity in the prediction model. Multicollinearity will affect prediction when making prediction using a prediction model with multicollinearity.

Chapter 3

Review of Model Selection and Model Averaging in the Presence of Missing Values

The aim of this chapter is to review model selection and model averaging methods in the presence of missing values. Modelling in the presence of missing data raises several new statistical challenges, underscoring the need for methodological development. In the literature, various methods were proposed as discussed in Chapter 2. Therefore, in this chapter we will describe and critique some recent developments in handling model selection and model averaging in the presence of missing values.

3.1 Model Selection in the Presence of Missing Values

Model selection and assessment with incomplete data are a very challenging process in model-building. Verbeke et al. [2008] stated two particular challenges. First, many models describe characteristics of the complete data, in spite of the fact that only an incomplete subset is observed. Direct comparison between model and data is less than straightforward. Second, many commonly used models are more sensitive to assumptions in the incomplete data situation and some of their attractive properties vanish when they are fitted to incomplete, unbalanced data. Verbeke et al. [2008] argued that model assessment should always proceed in two steps. In the first step, the fit of a model to the observed data should be assessed carefully, while the second step is assessment of the sensitivity of inferences to unverifiable assumptions, that is to how a model describes the unobserved data given the observed ones.

Model selection in practice requires the choice of a selection procedure, such as forward selection or backward elimination, coupled with a selection criterion, such as AIC or BIC, to select a small subset of variables to include in the model. Such procedures can be complicated even in the absence of missing data, because of the large number of possible models. Although researchers have developed many imputation methods to deal with missing data, there are no agreed guidelines for model selection in the presence of missing data. In the literature, researchers are still exploring model selection in imputed data sets.

3.1.1 Model selection strategies

One of the classical methods for model selection with multiply-imputed dataset is repeated use of Rubin's rules (RR approach or WALD test method) which was proposed by Rubin [1987] and Little and Rubin [2002] (as discussed in Section 2.2.5). This method uses simple backward stepwise selection. The Rubin inferential framework RR provides WALD tests for average parameter estimates obtained at MI Stage 3. Each model selection step involves fitting the model under consideration to all imputed data sets (MI Stage 2) and combining estimates of all parameters and standard errors across imputed data sets (MI Stage 3), eliminating the least significant of the non-significant parameters. RR is a most popular and well established method for combining parameters and standard errors. However, it is essentially a backward stepwise selection approach, so is open to all the general criticisms of that method discussed in Section 2.4.1 (see Harrell [2001]).

A naive approach for variable selection in multiply-imputed data sets is a 'majority vote' approach. If there are D imputed data sets, the model selection procedure will be applied to each completed dataset separately, resulting in D sets of selected predictors. The final model will comprise those predictors that are selected in 50% or more of the D data sets. The "majority vote" method fails to take into account the uncertainty caused by the missing data. The "majority vote" method gives much insight into the variability between the completed data sets. This variability can be found in the predictors selected and also in the selection of powers for one particular continuous predictor, which results in different functional forms. More than 10 imputations is required to obtain stable results if predictor and transformation selection is based on the "majority vote" method [Vergouwe et al., 2010].

Brand [1999] proposed a solution in two steps. The first step involves performing stepwise model selection separately on each imputed dataset, then construct a new super-model that contains all variables present in at least half of these models. In the second step, a

special backward elimination procedure is applied to all variables present in the super-model. Each variable is removed in turn and the pooled likelihood ratio $p - value$ is calculated. If the largest $p - value$ is larger than 0.05, the corresponding variable will be removed and the procedure repeated on the smaller model. The procedure stops if all $p \leq 0.05$. Step 1 of Brand [1999] is identical to the "majority vote" method.

In line with Brand [1999], Yang et al. [2005] identified variable selection problems with missing data in a Bayesian framework. Two alternative strategies to address the problem of choosing linear regression models when there are missing covariates were proposed. The first approach was "impute, then select" (ITS) which involves initially performing multiple imputation and then applying Bayesian variable selection to multiply-imputed data sets. The second strategy was to conduct Bayesian variable selection and missing data imputation simultaneously with one Gibbs sampling process, which was called "simultaneously impute and select" (SIAS). The Bayesian procedure known as stochastic search variable selection was used in implementing and evaluating both approaches.

The results showed that SIAS slightly outperforms ITS and provides smaller standard errors. SIAS has higher signal-to-noise ratio than ITS and a lower number of incorrect variables selected. However, ITS is easier to implement in current commercial software packages and has the flexibility of allowing the imputation step and selection step to be done by different analysts at different times. This is in agreement with Schafer [1997] who envisaged distinct imputation and analysis phases which can be carried out by two different group of researchers (imputer and analyst) [Yang et al., 2005].

Besides that, the study also showed that higher correlation among covariates leads to more precise imputation of missing data. The collinearity among covariates has the effect of blurring distinctions between predictors in the variable selection process. Yang et al. [2005] also mentioned that implementing SIAS will take some effort in developing sensible specification of priors. In addition, it was stated that current software packages have added capabilities in the past decade to implement missing data procedures but very few modules are specifically oriented toward variable selection for incomplete data sets.

Wood et al. [2008] stated that there were no proper guidelines for variable selection in multiply-imputed data sets. The common approach is to perform variable selection amongst the complete cases, which is a simple but inefficient and potentially biased procedure. They also stated that variable selection performed by repeated use RR is computationally demanding. For large data sets and large D, this process may not be computationally feasible. Therefore, Wood et al. [2008] proposed a sensible alternative method to the RR approach (WALD test method) which use stacked imputed data sets with weighted regression, called the STACK method. Variable selection will be carried

out using backward stepwise selection approach in STACK method. Stacking the D imputed data sets for the n individuals yields one large dataset of length Dn. Fitting models to this single stacked dataset yields valid parameter estimates but standard errors that are too small. A fixed weight was applied to all individuals to correct the standard errors. The three possible sets of weights were as follows:

1. W1: $w_i = \frac{1}{D}$.
   These weights scale the log likelihood for the stacked data to the equivalent of a dataset of length n but ignore the proportion of missing information.

2. W2: $w_i = \frac{(1-f)}{D}$
   where $f = \frac{total\,number\,of\,missing\,values\,across\,all\,variables}{(p-1)n}$, the average fraction of missing data across all variables

3. W3: $w_i = \frac{(1-f_i)}{D}$
   where $f_i = \frac{number\,of\,missing\,values\,for\,variables\,X_i}{n}$, the fraction of missing data for variable X_i

where $p-1$ is number of explanatory variables. Backward stepwise regression was used on the stacked dataset to choose a final model. This model was then fitted to each of the D imputed datasets in turn, and final parameter estimates were obtained by use of RR.

Proposed method was compared with complete cases, single stochastic imputation and separate imputation. The single stochastic imputation method used a single imputed dataset for variable selection. The separate imputation method is performing the model selection separately in each imputed dataset. There are three proposed strategies for this separate imputation method: select predictors that appear in any model (S1), select predictors that appear in at least half of the model (S2) and select predictors that appear in all models (S3). This approach typically leads to models with different selected predictors [Wood et al., 2008].

The results showed that complete cases fail to detect important predictors due to a lack of power. When missing data are not MCAR, it may select unimportant variables due to biased regression estimates. When multiple outcomes are of interest or numerous possible interaction terms are to be assessed, it may be impractical to use RR which is a multi-stage iterative process. The STACK method is a more sensible alternative to RR approach if repeated analyses are required at the model-building stage. Their study also showed that the stacking approach for MI variable selection improves the power to detect true predictors but has a slightly inflated type 1 error compared to RR approach (WALD test method). They recommended to use weight W3 for stacked imputed data sets with

weighted regression. A possible advantage of STACK method over RR approach is that the likelihood ratio test statistics, which are usually preferred to WALD statistics for non-linear regression and small samples, are easy to obtain. Besides that, the STACK method is computationally easier compared to the RR approach [Wood et al., 2008].

In addition, Wood et al. [2008] focused on traditional non-Bayesian variable selection for multiply-imputed data because this has the greatest practical relevance to most data analysts. This is an alternative approach to the approach described by Yang et al. [2005] which draws on the Bayesian framework of MI and variable selection. Their results showed that the two-step method of imputing and then selecting variables using RR by Wood et al. [2008] has a natural Bayesian extension and they compare it with conducting Bayesian model selection and MI simultaneously within one Gibbs sampling scheme. Their simulation results shows that such methods outperform the complete-case analysis but their integrated strategy only slightly outperforms the two-step Rubin's approach. Besides that, Wood et al. [2008] proposed to develop diagnostic procedures such as detecting influential points, making prediction, performing diagnostic tests and graphical checks for model misspecification.

Following Brand [1999] and Yang et al. [2005], Heymans et al. [2007] addressed the concern that the pooling of results across imputations in order to obtain final parameter estimates introduces complexities if automatic variable selection strategies are applied. The variable selection algorithm may easily produce different models for different imputed data sets. Therefore, they developed and tested a methodology combining MI with bootstrapping techniques for studying prognostic variable selection using backward selection. This method randomly draws multiple samples with replacement from the observed samples, thus mimicking the sampling variation in the population from which the sample was drawn. The imputation is carried out on each bootstrap sample separately. Stepwise regression analyses are then performed on each bootstrap sample. Variables were selected for a final model based on the inclusion frequency of each prognostic variable, the proportion of times that the variable appeared in the model fitted to the various imputed data sets. MICE was used to perform multiple imputations. The usual MI and bootstrap method were presented separately to identify the amount of variation generated by each method and compare them with proposed methodology that combines MI with bootstrapping techniques. For MI method, backward selection method was applied to 100 imputed data sets. Whereas for bootstrap method, backward selection was applied by drawing 200 bootstrap samples from the first imputed data sets only.

Heymans et al. [2007] found that 10 imputed data sets is adequate for analysis since use of 100 imputed data sets showed similar results as 10. They also found that the effect of

imputation variation on the inclusion frequency was larger than the effect of sampling variation. The proposed method performs better than MI only and bootstrap only. The results showed that it is possible to combine multiple imputation and bootstrapping, thereby accounting for uncertainty in imputations and uncertainty in selecting models. However, it may complicate the model-building process. Moreover, it was claimed that this was the first study that addresses both multiple imputation and sampling variation on the inclusion frequency of prognostic variables.

However, Harrell [2001] recognized that there are a number of potential drawbacks of using bootstrapping for variable selection. First, the choice of an α cutoff for determining whether a variable is retained in a given bootstrap sample is arbitrary. Second, in order to include that variable in the final model, the choice of a cutoff for the inclusion frequency is arbitrary. Third, selection from among a set of correlated predictor variable is arbitrary, so all highly correlated predictors may have a low bootstrap selection frequency. It can be the case that none of them will be selected for the final model even though when considered individually each of them may be highly significant. Lastly, the researcher must use double bootstrapping to resample the entire modelling process in order to validate the final model and to derive reliable confidence intervals. This can be computationally prohibitive. Therefore, it is not advisable to use bootstrapping for variable selection.

Vergouw et al. [2010] stated that researchers frequently use a regression analysis with a backward and forward selection strategy in the development of clinical prediction models. But this strategy can result in over-optimistically estimated regression coefficients, omission of important predictors and random selection of less important predictors which means derived models can be unstable. Incorporating a bootstrap resampling procedure in model development provides information on model stability. It is expected to produce a model which represents better the underlying population, since bootstrapping mimics the sampling variation in the population from which the sample was drawn. Therefore, their research examined influence of bootstrap and MI on model composition and stability in the presence of missing data.

Vergouw et al. [2010] examined the influence of bootstrap and MI on model composition and stability in the presence of missing data. There are four methods used to compare the effect of missing data and model stability on model composition: complete case analysis, MI, bootstrapping and MI+bootstrapping. For MI, missing data was imputed using the MICE package with 'pmm' as imputation method and five imputed data sets were generated. In each of the five imputed data sets, multiple regression was applied. The predictors which appeared in at least 2 models (an inclusion fraction of $\geq 40\%$) qualified for final model from these five models. A likelihood ratio test with a critical

$p-$value=0.157 was used to test whether these predictors significantly contributed to the final model. Predictors will be dropped from final model if $p - value > 0.157$. This is similar to the "majority vote" method as discussed earlier. For MI+bootstrapping method, missing data was imputed using MICE and five imputed data sets were created. In each of the five imputed data sets, the two step bootstrap model selection procedure was applied. First step, 500 samples with replacement were taken from complete case dataset. The predictors which appeared in $\geq$ 40% of these models qualified for the second step. In second step, 500 new complete case samples were taken and in each of which a multi-variable model was built using predictors from the first step. Information on model stability was provided by studying which combination predictors occurred most frequently in 2500 data sets [Vergouw et al., 2010].

Research by Vergouw et al. [2010] showed that accounting for missing data by MI and providing information on model stability by bootstrapping are instructive methods when deriving a prognostic model. Separating strong predictors from weak predictors by bootstrapping was shown to perform well comparative to automated backward elimination in identifying the true regression model. Moreover, the study also showed that application of the two-step bootstrap model selection procedure provides valuable information on model stability. It was suggested that MI using five imputed data sets is the most optimal choice to reduce the uncertainty in model derivation caused by missing data and it is a sufficient number in order to get stable results.

However, Vergouw et al. [2010] stated that how to optimally perform variable selection in multiply-imputed data sets is still a subject of discussion. It was proposed to identify a superior methodology for model selection in multiply-imputed data sets using a simulation study, in which true predictors and noise variables are assigned.

Vergouwe et al. [2010] demonstrated the development and validation of a prediction model obtained with logistic regression in the presence of multiply-imputed data. The analysis was performed by following three steps of model development in each of the completed data sets: (1) backward elimination of predictors and fractional polynomial (FP) transformations simultaneously, (2) estimation of regression coefficients and (3) estimation of a heuristic shrinkage factor to apply to the estimates of parameters in the final model. The FP was used to study the shape of the relationship between the continuous predictors and the outcome variable. An advantage of the multi-variable fractional polynomial (MFP) procedure is the selection of predictors and transformations can be carried out simultaneously (a way to preserve the nominal type 1 error probability). A heuristic shrinkage factor can be estimated using the model chi-square and the number of degrees of freedom. Model chi-square is the difference in -2 log likelihood between a model with only an intercept and a fitted model. The number of degrees of freedom is

the total number of degrees of freedom that are considered in the process of selecting from all candidate predictors plus all considered transformations.

Backward elimination of predictors and transformations was performed with MFP and an AIC stopping rule. This rule corresponds to a $p - value = 0.157$ for predictors with one degree of freedom. Whereas to select the predictors and transformations, the WALD test method (RR), "majority vote" method and STACK method of Wood et al. [2008] were used. All three methods were applied to 10 imputed data sets. A model was fitted for each of the 10 multiply-imputed data sets for the finally selected predictors and transformations for each of the three selection methods. RR was used to combine the estimated regression coefficients and variance from the 10 different imputed data sets. Finally, a heuristic shrinkage factor was estimated for each of the 10 models and the shrinkage factors were averaged [Vergouwe et al., 2010].

The results showed that the predictors and transformations selected with the three methods were very similar. Since it was a practical case study, generalization of the results is not possible. Although the WALD method follows RR and is a well-established approach, it has recently been shown that the use of WALD statistics to select the power in a FP model can result in biased estimates [Wood et al., 2008]. The important advantage of the STACK method is that only one dataset needs to be analyzed. The analysis will lead directly to a single set of selected predictors, with corresponding regression coefficients and standard errors [Vergouwe et al., 2010]. It was suggested to formulate general guidelines for prediction modelling in the presence of missing data for further research.

White et al. [2011] discussed that perfect prediction is a potential problem in regression models for categorical outcomes, including logistic, ordered logistic and multinomial logistic regression models, and it can be a severe problem in the presence of missing data. In logistic regression, perfect prediction occurs if there is a category of any predictor variable for which the outcome is always 0 (or always 1). In other words, the two-way table of predictor variable by outcome variable contains a zero cell. Perfect prediction can lead to infinite parameter estimates (which are not in themselves a problem), but it also will lead to difficulties in estimating the variance-covariance matrix of the parameter estimates. The standard errors computed from the information matrix will be extremely large.

van Buuren [2012] recommended the WALD test method since it is a well established approach that follows RR whereas the "majority vote" method and STACK method proposed by Wood et al. [2008] fail to take into account uncertainty caused by missing data. Indeed, as mentioned earlier, only the WALD test method preserved the type I error. However, the WALD test method is computationally intensive. An advantage of

the STACK method is that only one dataset needs to be analyzed. It was suggested that it is useful to combine methods for variable selection.

Chen and Wang [2013] criticized Bayesian variable selection strategies. Bayesian variable selection methods will perform inadequately if they are directly applied to multiply-imputed data because the selection will not be consistent across the multiple dataset generated by imputation. If a variable selection method is applied to each imputed dataset separately, it will identify different important variables in each imputed data sets. It will cause difficulties in producing the overall parameter estimates across all imputed data sets and also make it difficult to interpret the model or draw scientific conclusions. There are various leading-edge model selection methods such as MI-LASSO [Chen and Wang, 2013], CART, Random Forest, LASSO and Elastic Net [Lu and Petkova, 2014] and MI-based weighted elastic net (MI-WENet) [Wan et al., 2015] were proposed over the years. Wan et al. [2015] suggested the computational cost is mainly affected by the number of predictor variables not the sample size.

Maghsoudi et al. [2014] criticized the RR approach (WALD test method) for being time demanding since it uses backward elimination variable selection. Therefore, alternative easier variable selections were proposed. The variable selection was performed in each dataset independently where, after fitting of separate regression models to each dataset, candidate variables for a multifactorial model are finalized in a screening round. Then the estimates of selected variables across 10 multiply-imputed data sets will be combined. Maghsoudi et al. [2014] identified two limitations of the study by Wood et al. [2008]. First limitation, only monotonic forms of association were studied. Second limitation, the majority of scenarios were implemented for continuous outcomes and in binary outcome cases, missing data were generated under a MCAR mechanism. It was recommended to use easier variable selection methods such as S1, S2 and S3 that provide results comparable with complicate methods.

Schomaker and Heumann [2014] critiqued model selection in general, since it introduces additional uncertainty into the process of statistical modelling. There will be many good models to describe the data, i.e. models with very similar prediction error, but in some models a specific variable will be included and in others it will not. As a result, model selection estimators are often unstable, biased and under-estimate the estimator's variance by neglecting the uncertainty associated with the model selection process. It is often argued that model averaging is appropriate to overcome this problem. Therefore, in Section 3.2, model averaging in the presence of missing values will be discussed.

3.1.2 Model selection criteria

The model selection strategies discussed in Section 3.1.1 could be implemented using various model selection criteria. Model selection criteria typically use the likelihood function based on the observed data. It is very challenging to obtain a suitable and accurate approximation to the observed full data likelihood in the presence of missing data, as this involves intractable multiple integration. For this reason, the application of classical model selection methods such as AIC becomes more problematic when observations are missing. Therefore, Ibrahim et al. [2008] considered a class of information-based model selection criteria, called $IC_{H,Q}$ (as discussed in Section 2.4.6) for missing data problems. $IC_{H,Q}$ includes AIC and BIC as special cases as well as other model selection criteria that have been proposed in the literature. The novel feature of the proposed model selection criteria is that they essentially depend only on output from the EM algorithm for their computation. Their development is based on the fact that the observed data log-likelihood in a missing data problem can be written as a difference between two functions, the $Q-$function of the EM algorithm and another quantity called the $H-$function as discussed in Section 2.2.4.

The study showed that the theory of $IC_{\tilde{H}(k),Q}$ is quite general and can be applied to various types of missing data models for which the EM algorithm is applicable. The results showed that the criteria are consistent. Although consistency is a desirable and interesting property, it does not shed light on how to penalize the observed data likelihood for model parsimony in finite samples. Ibrahim et al. [2008] recommended further research to determine the best choice of penalty in missing data problems.

According to Garcia et al. [2010], there is no general and easy way to compute penalty and variable selection procedure for missing data problems. In many missing data problems, the observed data log-likelihood does not have a closed form and is often computationally intractable because it requires evaluation of high dimensional integrals which do not have a closed form. These integrals can be approximated but the accuracy of the approximation is essentially impossible to assess in many situations. Therefore, it can be infeasible to directly maximize the observed data log-likelihood function to select important variables and calculate their estimates. Besides that, even in the absence of missing data, model selection criterion such as AIC can become infeasible for variable selection in linear regression with a large number of covariates.

Thus, a new penalty criterion and variable selection procedures were developed for a class of statistical models for missing data problems. This extended the research of Ibrahim et al. [2008]. The proposed model selection criterion, IC_Q statistics (as in Equation (2.30)), was used to select the penalty parameters. The developed procedure

is a computationally attractive algorithm for simultaneously optimizing the penalized likelihood function and estimating penalty parameters. The study showed that the proposed variable selection procedure automatically and consistently selects the important covariates and leads to efficient estimates.

One of the disadvantages of penalized likelihood methods is that they do not provide a measure of model uncertainty (i.e. the probability of selecting each model in the model space). In general, Bayesian methods provide estimates of posterior model probability, but implementing Bayesian methods in full can be difficult in many situations. It requires specifying priors for all parameters in the model, a covariate distribution which encompasses all the models in the model space as well as calculating marginal likelihoods and enumerating all the models in the model space [Garcia et al., 2010].

Besides that, Garcia et al. [2010] suggested to explore variable selection using IC_Q under different modelling situations such as generalized linear mixed models with non-ignorable missing response and covariate data, semi-parametric survival models with missing covariate data, such as Cox model, frailty models, measurement error models and partially linear models with missing response and covariates.

Moreover, Claeskens and Consentino [2008] also proposed a model selection criterion based on the EM algorithm which is readily available for EM-based estimation methods, without much additional computational effort. Their model selection method is applicable to likelihood-based models including the class of generalized linear models. The proposed AIC for missing covariates in regression modelling structure is AIC_1 (Equation (2.32)) as discussed in Section 2.4.6.

The results have confirmed the good performance of the criterion, especially its efficiency to deal with missingness. Ignoring the missing data does not work well for model selection. Since Claeskens and Consentino [2008] focused on missing covariate data with an ignorable missingness mechanism, they suggested to extend these results to include missing response data and nonignorable missingness schemes. Besides that, it was also stated that a corrected AIC based on the EM algorithm for missing covariates can be derived as Equation (2.34). Therefore, they proposed to investigate the corrected AIC ($\mathrm{AIC}_{1,c}$) for the case of missing covariates for further research.

Consentino and Claeskens [2011] discussed handling of general model selection data via an EM algorithm based AIC and by means of a non-iterative method for specific setting of logistic regression models with a monotone pattern of missingness. This version of AIC was proposed by Claeskens and Consentino [2008] for missing covariates. The EM algorithm provides an efficient way of estimation in incomplete data problems, because it

relates maximum likelihood estimation of incomplete data to maximum likelihood estimation based on the completed data. However, its main and not negligible disadvantage is that the estimation of the Q-function is computationally intensive and can be quite time consuming, especially in a bivariate or higher dimensional situation.

The simulation study showed that this non-iterative approach works well to identify an error distribution for X_{miss}. AIC was used to investigate which distribution is modelling the data better and to decide on the best distribution of missing covariates. The results showed that this method performs well for larger sample sizes and AIC is selecting the model fitted with the true distribution with higher frequency. Besides that, the model selection method is not inflating the variances. Moreover, as a distribution selection method, the AIC based on the non-iterative method performs well and is able to distinguish normal data from the low degree $t-$distributed data in the presence of covariates with missing data. This is valid for both small and large sample sizes. Besides that, Consentino and Claeskens [2011] also suggested a criterion which will compute AIC differences and this is applicable to use with the multiple imputation for likelihood models.

It is well-known that deletion of incomplete data will result in reduced estimation precision (or reduced statistical power) and biased parameter estimates. According to Nakagawa and Freckleton [2008] and Nagakawa and Freckleton [2011], model ranking according to model selection criteria such as AIC and BIC will be biased as a consequence of biased parameter estimates due to deletion of missing data which are not MCAR. Most researchers are unaware of this issue. Nagakawa and Freckleton [2011] suggested to incorporate imputation of incomplete dataset before or as a part of model selection procedures. Symonds and Moussalli [2011] also stated that AIC cannot be compared between models if there exists missing data for some covariates. Therefore, proper guidelines for using model selection criteria in the presence of missing data is required.

Chaurasia and Harel [2012] stated that the version of AIC proposed by Claeskens and Consentino [2008] leads to models that tend to overfit, i.e. models that contain the correct model plus some additional variables. This is not surprising because it is well known that AIC tends to select over-specified models (see discussion in Section 2.4.7). Chaurasia and Harel [2012] explored model selection with an incomplete response variable in multiple linear regression, in line with Yang et al. [2005] and Wood et al. [2008].

Chaurasia and Harel [2012] proposed two AIC variants for multiply imputed data sets which are based on the Arithmetic Mean (AM) and Geometric Mean (GM) as discussed in Section 2.4.6. The study showed that the pattern of model selection by AIC_{AM} and AIC_{GM} is very similar to that of AIC_{full}. The results of AIC_{AM} and AIC_{GM}

tend to over-fit which is a known natural tendency of AIC. The correlation between variables showed a negative impact on model selection in the analysis phase. BIC also was considered in this study and showed similar model selection rates as AIC.

3.1.3 Strategies for building an imputation model

In the model-building process with missing data, it is necessary to define both the imputation and analysis models. According to Schafer [1997], the imputation and prediction phases are distinct. Therefore, it is common to ask whether MI leads to valid inferences when the imputer's model and analyst's model (prediction model) differ. Rules for combining complete-data inferences were derived under some implicit assumptions of agreement between the two models. It was stated that "the validity of MI inferences when the imputer's and analyst's models differ has been the subject of recent controversy". Schafer [1997] stated that the controversy is based on understanding the effects on inference when the analyst assumes more than the imputer or vice versa. A possible inconsistency will be that the analyst's and imputer's models differ, but the analyst's model can be considered as a special case of the imputer's. Another type of inconsistency arises when the analyst's model is more general than the imputer' model where the imputer applies assumptions to the complete data that the analyst does not assume. MI created under an erroneous model will lead to erroneous conclusions, therefore it is important to specify a correct imputation model as well as prediction model.

According to Sinharay et al. [2001], the key feature of the MI approach is a separation between the model used to obtain imputation and the final model used for analysis of the dataset. The imputation model and the data analyses should be compatible to provide good results. This coincides with Schafer [1997]'s discussion of the imputation model and prediction model. Sinharay et al. [2001] recommended that in forming the imputation model, one should include as much reasonable covariate information as is available. Any discrepancy between the imputation model and the prediction model will give rise to unreliable estimates.

Moons et al. [2006] advised to use all covariates and response/outcome in the imputation model. Ignoring the relationship between covariates with missing values and outcome will introduce a bias into the estimation of parameters in the prediction model. This is true whether the missingness mechanism is MCAR or MAR. This is supported by White and Royston [2009] where it was stated that when there exist missing values in the covariates of an analysis model, the outcome of analysis model must be used in the imputation model to impute the missing covariate value. If the imputed data will be

used to fit several different analysis models, then every variable included in any of the analysis models should be included in the imputation model White et al. [2011].

Collins et al. [2001] assessed the inclusion of auxiliary variables, comparing inclusive strategies (including numerous auxiliary variables) and restrictive strategies (including few or no auxiliary variables). Auxiliary variables are defined as variables that are included in an analysis solely to improve the performance of imputation procedures. Auxiliary variables may be included for two reasons. First, researchers may want to introduce variables that are potential causes or correlates of the missingness itself. Second, researchers may want to include variables that are simply correlated with the variables that have missing values, whether or not they are related to the mechanism of missingness. Collins et al. [2001] showed that the inclusive strategy is greatly preferred. The inclusive strategy reduces the chance of inadvertently omitting an important cause of missingness and also brings the possibility of noticeable gains in terms of increased efficiency and reduced bias. It was recommended to use MI for the inclusive strategy and it is more straightforward.

Harrell [2001] suggested that, if the main interest of a researcher is prediction and not interpretation or inference about individual effects, it is worth trying a simple imputation to see if the resulting model predicts the response almost as well as one developed after using customized imputation. In developing the model for prediction, it was suggested to use multiple imputation to impute missing data since MI is more effective in improving the precision of $\hat{\beta}$.

Clark and Altman [2003] developed a prognostic model for ovarian cancer, in the presence of missing data, using Rubin's Rules as discussed earlier. Auxiliary variables were included in the imputation model. The study showed that the inclusion of auxiliary variables and using all available information will produce multiple imputations that have minimal bias and maximal certainty.

Over the years, researchers were focusing on applying various imputation methods and investigating the performance of corresponding methods. But Ambler et al. [2007] studied performance of multivariate imputation by chained equation (MICE) for clinical outcomes as well as the reliability of the predictions after imputation. Other imputation methods such as mode imputation, mean imputation, conditional mean imputation and hot decking were compared with MICE. The results showed that MICE performs better than other imputation methods, producing the lowest biases in the regression coefficients and producing confidence intervals with coverage values close to the nominal level. No variable selection strategy was used and the study only focused on a full model approach with pre-specified predictors. Therefore, it was suggested to asses how well the methods perform when p-values are used to select predictors for the model. Based on

this research, it is clear that no proper model selection or variable selection was carried out on imputed dataset since the suggestion was to use p-value for variable selection.

In addition, Ambler et al. [2007] expected MICE to perform better in the presence of stronger associations between the covariates. Since the correlations in their research were moderate, with only 4 of the 120 possible pairwise correlations exceeding 0.5, it was proposed to investigate the performance of the imputation methods in the presence of a stronger MAR mechanism.

The imputation model plays an important role in the analysis of missing data so it is essential to choose a good imputation model. In the analysis carried out by Wood et al. [2008], the imputation model and prediction model were considered separately. The imputation model and prediction model can be built simultaneously when the outcome of interest is incomplete. Omitting variables from the imputation model causes downward biases in estimates of parameters in the prediction model. Therefore the safest rule is that the imputation model include a minimum of all candidate predictors for the prediction model.

Standard software adopts one of two approaches to deal with perfect prediction. First, it might drop terms from the imputation model to avoid perfect prediction where standard imputation procedure will end up imputing using the wrong model. A second approach, might be to retain terms and estimate a singular variance-covariance matrix which will lead either to very large standard errors or an unsuccessful attempt to correct standard errors. In these cases, the Normal approximation to the log-likelihood fails and leads to very poor draws of estimates. Although the 'ice' package is 'augmenting' the data by adding a few extra observation to the dataset to avoid perfect prediction, perfect prediction still causes problems in other software in year 2011 [White et al., 2011].

White et al. [2011] argued that a rich imputation structure is desirable in principle, but in practice fitting such complex sets of imputation models can defeat the software or lead to model instability. Since it is hard to propose universal solutions, careful exploration of the data can suggest smaller imputation models that are unlikely to cause substantial bias. In practice, researchers should try to simplify the imputation structure without damaging it. For example, omit variables that seem on exploratory investigation unlikely to be required in a 'reasonable' prediction model but avoid omitting variables that are in the prediction model or variables that clearly contribute towards satisfying the MAR assumptions. This contradicts previous studies where most of the researchers include all the available variables (both response and predictor variables) in the imputation models, except some of the researchers include auxiliary variables in the imputation model. To implement this approach, further research is need to develop useful rules of thumb.

Chaurasia and Harel [2012] identified that the issues in model selection with imputed data are how to combine model selection results from imputed data and also the impact of the assumed imputation model on model selection in the analysis phase. The importance of additional variables in the imputation model is exaggerated in the analysis phase when performing model selection and it increases with the percentage of missingness. This complexity cannot be resolved by increasing the number of imputations, therefore the researcher should not assume that MI will be forgiving when interest lies in model selection in the analysis phase. For further research, it was suggested to explore the issue about generalizing model selection procedures to account for the impact of imputation model on model selection in the analysis phase.

3.2 Model Averaging in the Presence of Missing Values

Various model selection or variable selection methods in the presence of missing data were discussed in the previous section. The majority of the variable selection methods incorporate multiple imputation to overcome the variable selection problem in the presence of missing data. Researchers are proposing new methods to deal with the model selection issue in the presence of missing data in terms of frequentist and Bayesian perspective. The proposed methods sound attractive, some have proven easy to implement and are fast. However, researchers should remember that model selection introduces additional uncertainty into the process of statistical modelling, which can be more severe in the presence of missing data. In the literature, model averaging was proposed as an alternative to model selection which intended to overcome the under-estimation of standard errors that is a consequence of model selection.

Model averaging techniques from a Bayesian point of view have been developed since the late 1970s, but were not widely used until recent advances in computing power facilitated their practical usage. Contributions from a frequentist perspective have been fewer but recent studies by Buckland et al. [1997], Hjort and Claeskens [2003] and Claeskens and Hjort [2008] have made some important progress. The details of frequentist model averaging methods were discussed in Section 2.5.

In addition, many researchers explored Bayesian model averaging in the presence of missing data over the years but very limited research was conducted using frequentist model averaging in the presence of missing data. Therefore, there are no agreed guidelines for model averaging in the presence of missing data and researchers are still exploring model averaging in multiply-imputed data sets.

For Bayesian model averaging, the prior probabilities for the potential models have to be specified and computer intensive methods such as Markov Chain Monte Carlo are required for computing the posterior distribution. But Frequentist Model Averaging (FMA) can be implemented without much difficulty or protracted computation.

Schomaker et al. [2010] proposed frequentist model averaging when there exists missing data based on two distinct approaches. The first approach combines estimates from a set of appropriate models which are weighted by scores of the missing data adjusted AIC criterion (AIC_W) derived by Hens et al. [2006]. The second approach averages over estimates of a set of models with weights based on the conventional model selection criterion (AIC) derived in Buckland et al. [1997] but with the missing data replaced by imputed values prior to estimating the models. Four types of imputation methods were compared: Generalized additive model based recursive imputation (GAMRI), Generalized linear model based recursive imputation (GLMRI), k-nearest neighbours (kNN) procedure and bootstrap based version of EM algorithm. This analysis was carried out using the R package Amelia II, which allows multiple imputation. Amelia II implements a bootstrapping-based algorithm that gives essentially the same answers as the standard EM based approaches and can handle many more variables. Amelia II provides users with a simple way to create and implement an imputation model, generate imputed data sets and check its fit using diagnostics.

The results showed that the imputation based FMA method produces closer estimates to maximum likelihood estimates than do the corresponding complete case analysis and AIC_W. The imputation based method produces accuracy by combining models whereas the complete case analysis and AIC_W are better off in selecting a single model. Model averaging based on AIC_W estimators yields more accurate estimators than the corresponding complete case estimators. The GAMRI and GLMRI based model averaging estimators performs well relative to the corresponding estimators that adopt the criterion AIC_W. Whereas the performance of estimators based on kNN and Amelia II imputation methods can vary considerably across the experimental settings and performance criteria. In addition, model averaging estimators often provide better estimates than those resulting from any single model. It was recommend to use model averaging by implementing AIC_W for multiple imputation or perform model averaging on single imputed dataset. Schomaker et al. [2010] suggested to extend the research by investigating model averaging using more sophisticated imputation techniques and model averaging based on the EM-based AIC developed by Claeskens and Consentino [2008].

Nagakawa and Freckleton [2011] stated that model averaging offers more reliable and robust point and uncertainty estimation of parameters. Such robustness is even true in complex cases as well as when there is collinearity among predictors. A study was

conducted to explore model averaging using information theoretic measures based on AIC (IT-AIC) in the presence of missing data. The model averaging method as derived in Burham and Anderson [2002] was used to average the estimators and the missing data were imputed using multiple imputation (mi package in R). The model averaging procedure was run for each imputed dataset. The parameters were pooled using model averaged estimates where by final parameter estimates is a pooled estimate combining all model-averaged estimates and their unconditional standard errors using RR. The results showed that Akaike weight were incorrectly estimated in incomplete data sets. MI was efficient in recovering Akaike weights for data sets with MCAR and MAR missingness. Nagakawa and Freckleton [2011] suggested that use of a larger number of imputations will help for the incomplete MNAR dataset but increasing D will not lead dramatic improvements. However, there is no clearly illustrated guidelines for model averaging in multiply-imputed dataset.

Schomaker and Heumann [2011] explored model averaging in factor analysis to account for model selection uncertainty associated with determination of the number of latent factors. A model averaging method using AIC, as derived in Buckland et al. [1997] was used to compromise between different models that contain different numbers of factors. Since there exist missing values, the missing values were imputed based on a k-nearest-neighbour methodology [Little and Rubin, 2002]. This imputation method was used due to the small sample size. The results showed that the model averaging method performs well in determining the latent factors. However, this study is more general application of model averaging and missing data issue was not consider a serious problem in determining the latent factor.

Model averaging estimators are often call 'unconditional' in the literature since inference does not rely on a single selected model, but they are still conditional on the set of candidate models under consideration. Although model averaging aims to incorporate the uncertainty associated with the model selection process by combining estimates over a set of models, there is still some argument over appropriate interpretation and confidence interval construction. Schomaker et al. [2010] and Schomaker and Heumann [2014] recognized these problems in the presence of missing data and there is no clear guidance how to proceed up to now. Therefore, model selection and model averaging after imputation were explored using multiple imputation strategy to deal with missing data.

It is straightforward to integrate model averaging estimates into the standard MI combining rule (RR), but it is important to discuss the consequences of this. Standard errors will become large due to combination of both selection and imputation uncertainty when

point estimates shrink towards zero if a variable is not supported throughout the imputation and candidate models. Model averaging and multiple imputation can be combined by first calculating model averaging in each dataset and then combining them by RR. Schomaker and Heumann [2014] proposed a general model averaging estimator after multiple imputation as discussed in Section 2.5.

Schomaker and Heumann [2014] investigated model averaging after multiple imputation using bootstrapping. The algorithm for combining model averaging and multiple imputation using bootstrapping is as following:

1. Create B bootstrap samples of the original data (including missing observations)
2. Generate D imputed data sets for each bootstrap sample
3. Calculate model averaging estimator for each imputed set of data in each bootstrap sample
4. Create a model averaging estimators after imputation using Equation (2.40) for each bootstrap sample
5. Use the average of the B estimates calculated in step 4 as the final point estimate
6. Construct confidence intervals based on the percentiles of the empirical distribution produced by the B estimates of step 4.

Results showed that combining model averaging and multiple imputation outperforms complete case analysis. Combining model averaging with bootstrapping helped to calculate good estimates when dealing with model selection uncertainty. Model averaging induces more stable estimates than model selection, due to its inherent shrinkage properties and therefore smaller variance in exchange for some bias. It was found that model averaging estimators produced accurate standard errors after multiple imputation for all situations under consideration. To account for both uncertainty related to imputation and model selection, the incorporation of model averaging is most relevant method in the present of missing data [Schomaker and Heumann, 2014].

Since this study was restricted to certain imputation and model averaging procedures, i.e. AIC based choices and multiple imputation was solely utilized with Amelia II. Schomaker and Heumann [2014] suggested to explore other model averaging techniques and imputation methods to provide more evidence about the generalization of their finding. Another issue that deserves more in-depth research is the implementation of proper multiple imputation and the consequences of specifying wrong imputation model.

The use of an incorrect imputation model can cause improper imputation, biased model estimates and inappropriate post model averaging estimates.

Schomaker and Heumann [2014] stated that the choice of the imputation model can affect final results even if data are missing at random. If a fully conditional imputation approach is utilized (such as imputation by chained equations), convergence to the theoretical joint distribution is not always guaranteed. Whereas, if a joint modelling approach is taken (i.e. via Amelia II), the treatment of categorical variables via the multivariate normal distribution will yield reasonable results but imputation uncertainty increases and quite large standard errors will be observed. Besides that, imputing longitudinal data is complex and it is not entirely clear how misspecification of a longitudinal imputation model will affect regression modelling. These demonstrate the complexity and sensitivity of analyses dealing with missing data and it can be more complicated using model averaging, model selection and multiple imputation. Therefore, further research has to be carried out to reveal the whole complexity of modelling uncertainty in the presence of missing data.

3.3 Summary

Various researches have been carried out on model selection in the presence of missing data, and numerous methods/strategies have been proposed and examined. The proposed strategies are the "majority vote" method, backward stepwise regression using RR (WALD test method), STACK method, single stochastic imputation, separate imputation, ITS, SIAS, MI with bootstrapping and various leading-edge model selection methods. As discussed in previous sections, these methods showed some significant contributions to the development of model selection methods in the presence of missing data. However, there are well-established disadvantages of using the RR approach, single stochastic imputation, separate imputation and the MI with bootstrapping method. It is not advisable to use these methods for model selection in the presence of missing data.

Among the proposed methods, the STACK method [Wood et al., 2008] appears to be a more attractive model selection method in the presence of missing data. The big advantage of the STACK method is that only one dataset needs to be analyzed. The data analysis will directly lead to a single set of selected predictors, corresponding regression coefficients and standard errors. This method is computationally easier compared to the RR approach and also a sensible alternative to it.

Model averaging methods were proposed and examined as an alternative to model selection, intended to overcome the under-estimation of standard errors that is a consequence of model selection. Limited researches were carried out on model averaging in the presence of missing data. Model averaging is the most relevant method to account for both uncertainty related to model selection. There are no proper guidelines for model averaging in the presence of missing data. Although model averaging was proposed as an alternative to model selection, there is no proper comparison between model selection and model averaging in the presence of missing data in terms of prediction. If the aim of both model selection and model averaging is prediction, the comparison between them should be carried out in terms of prediction using a measure such as mean square error of prediction (MSE(P)).

There are a few possible suggestions from researchers to explore in model selection and model averaging in the presence of missing data. Some of the suggestions were not fully explored and can be used as guidance for other researchers to explore them. Wood et al. [2008] proposed to develop diagnostic procedures such as detecting influential points, making prediction, performing diagnostic tests and graphical checks for model misspecification. The comparison between model selection and model averaging can be explored in terms of predictions. Vergouwe et al. [2010] also suggested to formulate general guidelines for prediction modelling in the presence of missing data. In addition, proper guidelines for using model selection criteria in the presence of missing data is required [Symonds and Moussalli, 2011]. Chaurasia and Harel [2012] suggested to explore the issue about generalizing model selection procedures to account for the impact of the imputation model on model selection in the analysis phase.

White et al. [2011] suggested to develop useful rules of thumb for building a proper imputation model. The inclusive strategies [Collins et al., 2001] can be explored using MI for building an imputation model since it is more straightforward. A suitable model selection/variable selection strategy on imputed dataset is required and the performance of that method should be explored [Ambler et al., 2007].

Nakagawa and Freckleton [2008] suggested that clearly illustrated guidelines for model averaging in multiply-imputed dataset required. Since the use of an incorrect imputation model can cause improper imputation, biased model estimates and inappropriate post model averaging estimates, it was suggested to develop rules of thumbs for building a proper imputation model when using imputed dataset for model averaging. Schomaker et al. [2010] suggested to investigate model averaging using more sophisticated imputation techniques and model averaging based on the EM-based AIC developed by Claeskens and Consentino [2008]. Schomaker and Heumann [2014] suggested to carried out research to reveal the whole complexity of modelling uncertainty in the presence of missing data

since it can be more complicated to use model selection, model averaging and multiple imputation in the presence of missing data.

Table 3.1 and Table 3.2 show a summary of researches carried out on model selection and model averaging in the presence of missing data respectively. In conclusion, comparison between model selection and model averaging in the presence of missing data is worth exploring and development of a proper model-building approach is required in both model selection and model averaging.

Table 3.1: Review of Model Selection in the Presence of Missing Data

Study	**Objectives**	**Methods**	**Outcomes**	**Advantages/disadvantages**	**Recommendations**
Rubin [1987], Little and Rubin [2002]	To combine parameters and standard error across multiple imputed data sets using Rubin's rules	backward stepwise selection/ multiple imputation (WALD test method/ RR approach)	-	Disadvantage: computationally intensive	-
Brand [1999]	To treat variable selection problem when there exists missing data	two step solutions (majority method and WALD method)	more considerable improvement over complete case analysis	-	-
Collins et al. [2001]	To asses the inclusion of auxiliary variables, comparing inclusive strategies and restrictive strategies	Use ML and MI for imputation	The inclusive strategy is greatly preferred	Advantage:The inclusive strategy reduces the chance of inadvertently omitting an important cause of missingness and also brings the possibility of noticeable gains in terms of increased efficiency and reduced bias	It was proposed to explore missing data issues with small sample size, various amount amount of missingness and types of missing data mechanisms
Clark and Altman [2003]	To develop a prognostic model in the presence of missing data	RR approach	Inclusion of auxiliary variables and using all available information were generated MI that have minimal bias and maximal certainty	Disadvantage: backward elimination method is computationally intensive	-
Yang et al. [2005]	To identify the variable selection problems with missing data in Bayesian framework	ITS and SIAS	SIAS slightly outperforms ITS, but ITS is easier to implement and flexible. Higher correlation among covariates leads to precise imputation	Disadvantage: SIAS will take some effort in developing sensible prior to specification	-

Continued on Next Page...

Table 3.1 – Continued

Study	**Objectives**	**Methods**	**Outcomes**	**Advantages/disadvantages**	**Recommendations**
Ambler et al. [2007]	To investigate performance of MICE and reliability of the predictions after imputation	mean/mode imputation, mean imputation, conditional imputation, hot decking and MICE (no variable selection)	MICE performs better (produced lowest biases and CI coverage values close to nominal level)	Disadvantage: no proper variable selection	Recommended to use p-value for variable selection in assessing MICE performance
Heymans et al. [2007]	To develop methodology for combining MI with bootstrapping	bootstrapping/stepwise backward elimination (MICE)	Combined MI and bootstrapping method accounts the uncertainty in imputation and uncertainty in selecting models	Disadvantage: combining MI and bootstrapping will complicate the model building process	-
Wood et al. [2008]	To develop STACK method, alternative methods to RR approach	STACK method with weighted regression/ multiple imputation	Sensible alternative to RR approach and has more power compromising slightly type 1 error.	Advantage: computationally easy	Recommended to use W3 weights, consider imputation and prediction models separately, proposed to develop diagnostic procedure such as detecting influential points, making prediction, performing diagnostic test and graphical checks for model mis-specification.
Vergouw et al. [2010]	To examine influence of bootstrap and MI on model composition and stability in the presence of missing data	two-step bootstrap and MI	Model selection provides more valuable information on model stability and five imputed data sets is the most optimal choice to reduce the uncertainty in model derivation	-	It was proposed to identify a superior methodology for model selection in multiply imputed dataset

Continued on Next Page...

Table 3.1 – Continued

Study	**Objectives**	**Methods**	**Outcomes**	**Advantages/disadvantages**	**Recommendations**
Vergouwe et al. [2010]	To develop and validate a prediction model obtained with logistic regression in the multiply imputed data.	MFP with AIC stopping rule, STACK, WALD test	all three method showed similar results in selected predictors and transformations and	Advantage: The selection of predictors and transformations can be carried out simultaneously using MFP	It was suggested to formulate general guidelines for prediction modelling in the presence of missing data
Maghsoudi et al. [2014]	To explore model selection in incomplete dataset	RR approach, STACK, separate imputation	STACK method and seperate imputation method showed similar results as RR approach.	-	Recommended to use easier variable selection methods

Table 3.2: Review of Model Averaging in the Presence of Missing Data

Study	**Objectives**	**Methods**	**Outcomes**	**Recommendations**
Schomaker et al. [2010]	To explore frequentist model averaging (FMA) in the presence of missing data	model averaging based on AIC [Buckland et al., 1997] and based on AIC_W derived by Hens et al. [2006]	Model averaging based on AIC_W yield more accurate estimators than the complete case estimators	Suggested to investigate model averaging using more sophisticated imputation techniques and model averaging based on the EM-based AIC developed by Claeskens and Consentino [2008].
Nagakawa and Freckleton [2011]	To explore model averaging based on AIC in the presence of missing data	model averaging based on AIC [Burham and Anderson, 2002]	The results showed that Akaike weight were incorrectly estimated in incomplete data sets. MI was efficient in recovering Akaike weights for data sets with MCAR and MAR missingness	suggested that use of a larger number of imputations will help for the incomplete MNAR dataset but increasing D will not lead dramatic improvements
Schomaker and Heumann [2011]	To explore model averaging in factor analysis in the presence of missing data	model averaging based on AIC [Buckland et al., 1997]	The results showed that the model averaging method performs well in determining the latent factors.	
Schomaker and Heumann [2014]	To investigate incorporation of model selection and model averaging after multiple imputation	bootstrapping for MI, RR for combining model averaging estimators	Combining model averaging and multiple imputation outperforms complete case analysis and Combining model averaging with bootstrapping helped to calculate good estimates when dealing with model selection uncertainty. Model averaging induces more stable estimates than model selection, due to its inherent shrinkage properties and therefore smaller variance in exchange for some bias	Suggested to explore other model averaging techniques and imputation methods to provide more evidence about the generalization of findings and carried out research to reveal the whole complexity of modelling uncertainty in the presence of missing data

Chapter 4

Comparison between Model Selection and Model Averaging

The aims of this chapter are: (i) to compare model selection and model averaging in terms of imputation and prediction; (ii) to investigate the effects of restrictive and inclusive strategies for imputation for both model selection and model averaging. The restrictive strategy (where minimal use is made of auxiliary variables in both prediction and imputation models), inclusive strategy (where numerous auxiliary variables and overlapping variable sets in both imputation and prediction models) and a strategy using non-overlapping variable sets (where the auxiliary variable is only used in the imputation model) were investigated. The effects of the imputation and simulation parameters were observed and discussed in both linear model and Logistic regression. The general simulation design and the simulation parameters used in the simulation studies are discussed in this chapter. A simple simulation scenarios with three covariates (some values are missing in one of the covariate) were considered in order to identify the effects of other simulation parameters. More than three covariates and missing values in more than one covariate become more complicated and overshadow the effects of simulation parameters. This basic simulation design will be used for other simulation studies in later chapters but the parameters will be modified according to the aim of the simulation study.

4.1 Design of Simulation

Sets of simulation studies will be carried out in both linear model and Logistic regression in this research. Generally, X_1 and X_2 are covariates in a prediction model for the response Y, and some values of X_2 (but not X_1) are missing. X_3 is an auxiliary variable,

primarily intended to use in the imputation model for X_2 and might or might not also be used in the prediction model for Y. The general covariance matrix for $\mathbf{X} = (X_1, X_2, X_3)$ is therefore

$$\Sigma = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{pmatrix} \tag{4.1}$$

where $\rho_{ij} = \rho_{ji}$ denotes the correlation between X_i and X_j. In the simulations in this chapter, $\rho_{12} = \rho_{13} = 0$ so

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \rho_{23} \\ 0 & \rho_{23} & 1 \end{pmatrix} \tag{4.2}$$

where $\rho_{23} = -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75$. The number of observations was $n = 50, 100, 200, 400$. The percentage of missing values was $m = 0, 25$ and 50, $m = 0$ was chosen to investigate the effects of correlation and sample size without any imputation. The value $m = 25$ was chosen to identify the effects of a moderate amount of imputation. An extreme value of missing percentage, $m = 50$, was chosen to identify the effect of imputation when half of the data are imputed. The additional effects of imputation will be investigated by comparing results with $m = 0$ and with $m = 25$ and $m = 50$. Simulations were carried out for every combination of n, m, and covariance matrix.

4.1.1 Linear model and Logistic regression

The general multiple linear regression model considered (true model) was

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \qquad , i = 1, 2, ..., n \tag{4.3}$$

where
Y = the response variable
X's = explanatory variables
$\beta's$ = coefficients/parameters of the model
ε = error term
n= number of observations
$\mathbf{X}$ (X_1, X_2 and X_3) values were simulated from a multivariate normal distribution with fixed zero means and a specified covariance matrix. The Y values were created based on Equation (4.3), simulated X_1 and X_2 values and error terms simulated from $N\left(0, \sigma_\varepsilon^2\right)$ where $\sigma_\varepsilon^2 = \frac{1}{16}, 1, 16$. A small value of $\sigma_\varepsilon = \frac{1}{4}$ and a large value $\sigma_\varepsilon = 4$ were chosen to identify the effects of noise. In all simulations, $\beta_0 = \beta_1 = \beta_2 = 1$. Based on Equation (4.3), one can interpret the coefficients as "if X_2 is fixed, then for each change of 1 unit in X_1, Y changes 1 unit". The β's were chosen to be 1 to investigate

the effects of σ_ε on prediction. The coefficient-to-variance ratio $\left(\frac{\beta}{\sigma_\varepsilon}\right)$ will be equal to 1 when $\sigma_\varepsilon = 1$. Rather than changing the β values, the σ_ε values were changed to identify the effects of simulation parameters. The simulation study was carried out with 1000 simulations.

The logistic regression model considered (true model) was:

$$P_i = P(Y_i = 1) = \frac{exp^{(\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i})}}{1 + exp^{(\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i})}} \qquad , i = 1, 2, ..., n \tag{4.4}$$

Equation (4.4) can be re-written as:

$$logit\, P_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \tag{4.5}$$

Here:
Y = binary response variable, which can only take the value either 0 or 1
$logit\, P_i = ln\left(\frac{P_i}{1-P_i}\right)$
P_i = probability of success (in the range 0 to 1)
X's = explanatory variables
$\beta's$ = coefficients/ parameters of the model
n= number of observations
X (X_1, X_2 and X_3) values were simulated from a multivariate normal distribution with fixed zero means and a specified covariance matrix. Y values were created based on Equation (4.4), and the simulated X_1 and X_2. In all simulations, $\beta_0 = \beta_1 = \beta_2 = 1$. The simulation study was carried out for number of simulations equal to 1000.

4.1.2 Imputation and prediction models

Three model-building strategies were considered to build the imputation and prediction models. The strategies are restrictive strategy, inclusive strategy and non-overlapping variable sets. Collins et al. [2001] defined a restrictive strategy as including few or no auxiliary variables in both imputation and prediction models. An inclusive strategy is including numerous auxiliary variables and overlapping variable sets in both imputation and prediction models. A strategy of using non-overlapping variable sets (an extremely restrictive strategy) is defined as not including auxiliary variables in the prediction model, only in the imputation model, so that non-overlapping variable sets are considered for the imputation and prediction models. Auxiliary variables are variables within the original data that are not included in the main analysis, but are correlated to the covariates of interest and may be used in the imputation model [Hardt et al., 2012]. Auxiliary variables are also known as ancillary or exogenous variables. Auxiliary

variables are defined as variables that are included in an analysis solely to improve the performance of missing data procedures.

Missing observations were created completely at random on variable X_2 with percentages of missing observations as $m = 25$ and $m = 50$. The "norm" imputation method (see Section 2.3.2) was used to impute any missing observations of X_2 using the auxiliary variable X_3. The imputation model used in both restrictive strategy and non-overlapping variable sets was

$$X_{2i} = \hat{\varphi_0} + \hat{\varphi_3} X_{3i} + \hat{\varphi_4} Y + h_i \tag{4.6}$$

and the imputation model for the inclusive strategy was

$$X_{2i} = \hat{\varphi_0} + \hat{\varphi_1} X_{1i} + \hat{\varphi_3} X_{3i} + \hat{\varphi_4} Y + h_i \tag{4.7}$$

Table 4.1: All possible prediction models

Name	**Fitted Linear Models**	**Fitted Logistic regression**	**Non-overlapping**	**Restrictive**	**Inclusive**
M000	$Y = \beta_0 + \varepsilon$	$logit\, P_i = \beta_0$	✓	✓	✓
M100	$Y = \beta_0 + \beta_1 x_1 + \varepsilon$	$logit\, P_i = \beta_0 + \beta_1 x_1$	✓	✓	✓
M010	$Y = \beta_0 + \beta_2 x_2 + \varepsilon$	$logit\, P_i = \beta_0 + \beta_2 x_2$	✓	✓	✓
M001	$Y = \beta_0 + \beta_3 x_3 + \varepsilon$	$logit\, p_i = \beta_0 + \beta_3 x_3 + \varepsilon$		✓	✓
M110	$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$	$logit\, P_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2$	✓	✓	✓
M101	$Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon$	$logit\, p_i = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon$		✓	✓
M011	$Y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$	$logit\, p_i = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$		✓	✓
M111	$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$	$logit\, p_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$		✓	✓

There are two parts to the analysis: model selection and model averaging. In each simulation, after imputation if required, each model selection criterion (AIC, AIC_c and BIC) was allowed to choose an additive model based on any combination of X_1 and X_2. There were four possible models for Linear regression and Logistic regression for non-overlapping variable sets as listed in Table 4.1. Eight possible prediction models based on all possible subsets of variables X_1, X_2 and X_3 (includes one auxiliary variable) were considered for Linear regression and Logistic regression for restrictive and inclusive strategies as listed in Table 4.1. With this terminology, the true model was M110 for all three model-building strategies. The number of times each model was selected by each criterion was recorded. For model averaging, the weights for all possible models were calculated using modified weights as in Equation (2.39) and also as described in Buckland et al. [1997] using each of AIC, AIC_c and BIC. These were then applied to the estimated parameters from all possible models in order to obtain final 'weighted' parameter estimates.

4.1.3 Test values

In this research, the performance of model selection and model averaging were compared in terms of imputation and prediction. An approach is needed to compare model selection and model averaging. The only overlapping calculation step in both model selection and model averaging is prediction. Therefore, the mean square error of prediction, MSE(P), will be calculated to compare both model selection and model averaging. A modelling strategy with minimum MSE(P) is preferred. A fixed set of test values will be created and used for all the simulation studies in this research.

X test values (X_1 and X_2) were calculated based on the probability quantile function of the standard normal distribution. The test set values consisted of 100 points in a 10×10 lattice with each of X_1 and X_2 taking values equi-spaced in probability at the $5, 15, ..., 95$ percentiles of the standard normal distribution. The Y test values were created based on Equation (4.3) for the linear model and the P test values based on Equation (4.5) for the Logistic regression, with the X_1 and X_2 test values and zero error. There were 100 sets of test values. These test values were used to calculate the mean square error of prediction of the best model or all fitted models. The performance of the model selection and averaging procedures were compared using mean square error of prediction, MSE(P).

The MSE(P) for each test values in the linear model will be calculated using Equation (4.8) where

$$MSE\,(P)_t = \frac{1}{1000}\sum_{s=1}^{1000}(y_{ts} - y_t)^2 \tag{4.8}$$

The average MSE(P) across test values in the linear model is

$$MSE\,(P) = \frac{1}{100}\sum_{t=1}^{100}\left[\frac{1}{1000}\sum_{s=1}^{1000}(y_{ts} - y_t)^2\right] \tag{4.9}$$

where s indexes the simulations ($s = 1, 2, ..., 1000$) and $t = 1, 2, ..., 100$ for the test values. As discussed in Section 2.6.2, the MSE(P) for each test values in the Logistic regression will be calculated using Equation (4.10) where

$$MSE\,(P)_t = \frac{1}{1000}\sum_{s=1}^{1000}(P_{ts} - P_t)^2 \tag{4.10}$$

The average MSE(P) across test values in Logistic regression is

$$MSE\,(P) = \frac{1}{100}\sum_{t=1}^{100}\left[\frac{1}{1000}\sum_{s=1}^{1000}(P_{ts} - P_t)^2\right] \tag{4.11}$$

where P is the probability of success as in Equation (4.4) for Logistic regression. The distribution of MSE(P) for each of $m = 0$, $m = 25$ and $m = 50$ was plotted to identify the effects of simulation parameters in both model selection and model averaging.

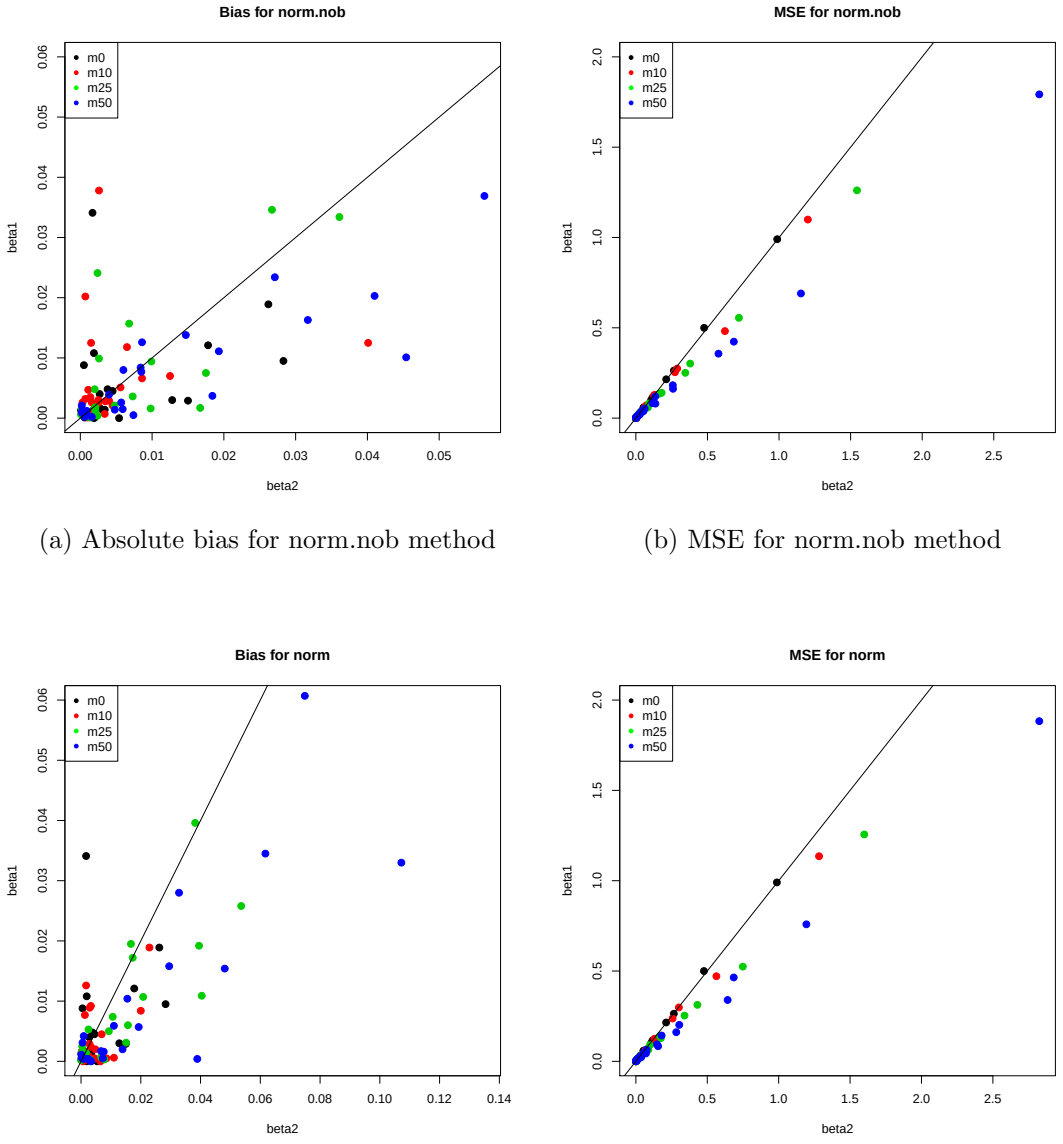
4.1.4 Choice of imputation package and method

The MICE package [van Buuren and Groothuis-Oudshoorn, 2011] was chosen to use for imputation in this research. The main advantage of MICE over the mi package by Yu et al. [2011] is the flexibility it offers for choosing an imputation method (as discussed in Section 2.3.2). There are various imputation methods in MICE, therefore a small scale simulation study was conducted in order to compare them. A few trials were carried out for both linear model and Logistic regression. The main objective of this trials is to obtain a suitable imputation method and also to understand the corresponding chosen method.

The analysis of linear model and Logistic regression showed that the best imputation methods are "norm.nob" and "norm". These methods give smaller bias and MSE values than the other methods. In Figure 4.1, the bias value for β_2 generally falls below the equality line, meaning that β_1 is less biased than β_2 for both imputation methods. The bias values are much higher for β_2 (which is estimated with imputed values) than β_1. The bias and MSE values increase as missing percentages increases for both imputation methods. In general, there are no clear differences between the two imputation methods in terms of bias and MSE values for linear regression.

There is no systematic difference between the two methods for bias in Logistic regression (see Figure 4.2) but the MSE values are slightly higher for the "norm" method. When the percentages of missing increases from 10% to 50%, the MSE values for "norm" method also increase. Therefore, there are no clear differences between the two imputation methods in terms of bias and MSE values for Logistic regression as well as linear regression.

Imputed values are affecting the estimation of β_2 where the bias and MSE values are larger for β_2 compared to β_1, see Figure 4.1 and Figure 4.2. Figure 4.1 shows that for both methods, bias and MSE values are larger for β_2. These results show that there is an effect of imputation on the parameter estimation. The imputation is not only affecting the estimation of β_2 but also affects β_1. There is a trade-off between the estimation of both coefficients in order to produce lower MSE. Therefore when β_2 is overestimated, then β_1 will be underestimated to minimize the error.



(a) Absolute bias for norm.nob method

(b) MSE for norm.nob method

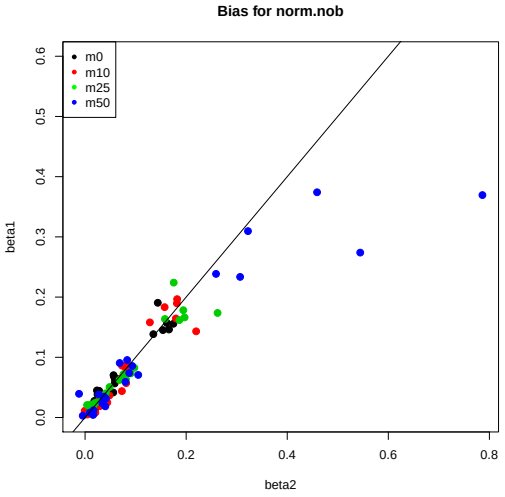
(c) Absolute bias for norm method

(d) MSE for norm method

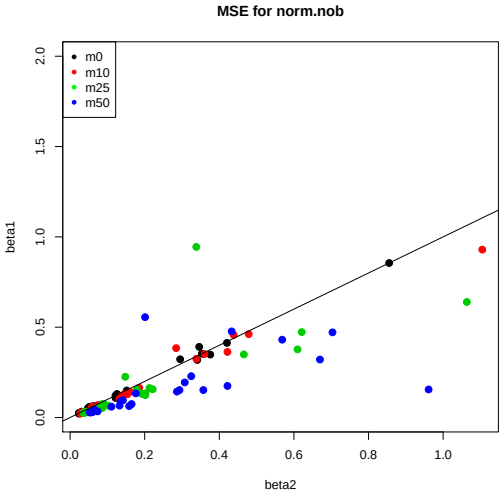
Figure 4.1: Bias and MSE for norm.nob and norm methods in package MICE for linear regression

According to Donner [1982], the linear prediction is most effective for estimating the coefficients and it has a lower MSE value compared to complete-case method, mean substitution method and piece-wise method. This is coincides with the results of this study where the "norm.nob" and "norm" methods, both regression methods, were chosen as best method. These methods were less biased and had a lower MSE. However, the β_2 in this study more biased than the β_2 in Donner [1982] since Donner [1982] did not included outcome in the imputation model. Besides that, our study showed that these two methods are better than the classical linear prediction (known as "norm.predict" in MICE package) discussed in Donner [1982]'s research. Therefore, "norm.nob" and

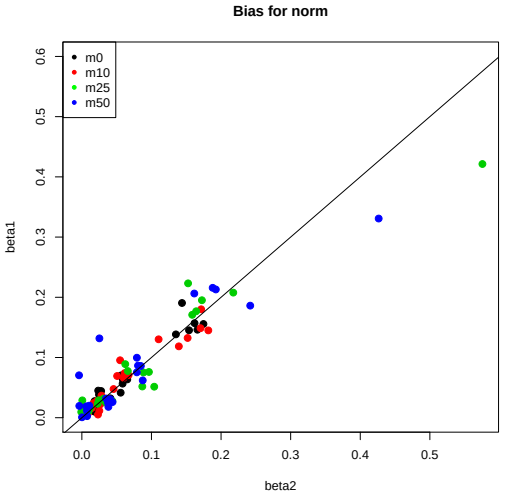
"norm" methods are best imputation methods for linear model and Logistic regression. Donner [1982] also suggested that the linear prediction method is less biased for estimating β_2 than for estimating β_1, especially when correlation coefficient is small. This coincides with the results of this study where weaker correlation produced less bias and reduced the error in the estimation of parameters.



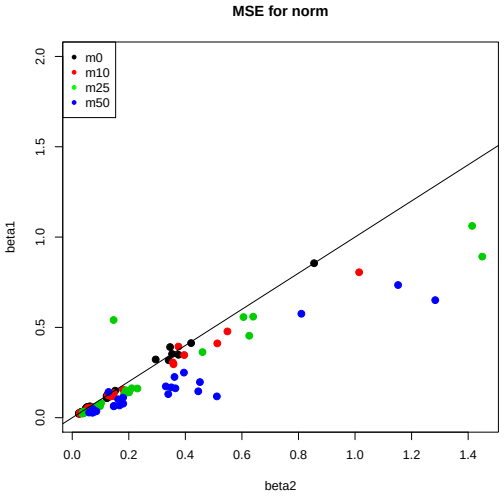
(a) Absolute bias for norm.nob method



(b) MSE for norm.nob method



(c) Absolute bias for norm method



(d) MSE for norm method

Figure 4.2: Bias and MSE for norm.nob and norm methods in package MICE for logistic regression

Schafer [1997] stated that "norm" is a proper and "norm.nob" is an improper method for multiple imputation. In conclusion, from a Bayesian perspective, the "norm" method is a proper method for multiple imputation.Therefore, in this research, the imputation

method "norm" will be used for imputing missing data in both linear regression and Logistic regression analyses.

4.2 Results

In this section, we will discuss the results for linear model and logistic regression based on the simulation design in the previous section. All three strategies of building imputation and prediction models will be compared on both linear model and logistic regression models beginning with non-overlapping variable sets. Since AIC_c converges to AIC with increases in the ratio $\frac{n}{k}$, only AIC_c and BIC results will be shown. The advantage of AIC_c over AIC is the application in small samples where it is less biased than AIC [Claeskens and Hjort, 2008]. However, all the three model selection criteria were used for simulation studies in both model selection and model averaging. The negative and positive correlations of the same magnitude showed similar results, therefore only positive correlation results will be discussed for model selection and model averaging. The performance of model selection and model averaging were compared in both linear model and logistic regression using mean square error of prediction.

4.2.1 Linear regression with non-overlapping variable sets

Model selection using non-overlapping variable sets showed similar results for negative and positive correlations of the same magnitude. When $\sigma_\varepsilon = 0.25$ (the smallest value used in these simulations), the true model M110 was chosen in each of the 1000 simulations for all combinations of ρ_{23} and for $m = 0, 25$ and 50. When $\sigma_\varepsilon = 1$, with a solitary exception, the true model M110 was chosen 100% compared to other possible models in each of 1000 simulations for $m = 0$ and all values of ρ_{23}. Also when $\sigma_\varepsilon = 1$, for $n = 100, 200$ and 400, the true model M110 was selected 100% compared to other possible models for for $m = 25$ and all values of ρ_{23}. However, for $n = 50$, the number of times the true model M110 was selected via AIC_c and BIC increased as ρ_{23} increased, the true model M110 was selected 100% for $\rho_{23} = 0.75$.

Table 4.2a and Table 4.2b show the equivalent results when $\sigma_\varepsilon = 1$ and $m = 50$. The true model M110 was selected 100% as sample size increased for all values of ρ_{23}. For $n = 50$ and $n = 100$, the number of times model M100 was selected via AIC_c and BIC decreased as the ρ_{23} increased from zero to 0.75. As n increased, the number of times model M100 was selected via AIC_c and BIC decreased for any ρ_{23}.

Table 4.2: Number of times all possible models are selected via AIC_c and BIC in each of 1000 simulations for all the combinations of ρ_{23} when $\sigma_\varepsilon = 1$ and $m = 50$ for linear regression

(a) Number of times all possible models are selected by AIC_c

$\sigma_\varepsilon = 1$ and $m = 50$																
ρ_{23}	$\rho_{23} = 0$				$\rho_{23} = 0.25$				$\rho_{23} = 0.5$				$\rho_{23} = 0.75$			
Selected model	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110
$n = 50$	0	29	1	970	0	17	0	983	0	9	0	991	0	2	0	998
$n = 100$	0	0	0	1000	0	0	0	1000	0	0	0	1000	0	0	0	1000
$n = 200$	0	0	0	1000	0	0	0	1000	0	0	0	1000	0	0	0	1000
$n = 400$	0	0	0	1000	0	0	0	1000	0	0	0	1000	0	0	0	1000

(b) Number of times all possible models are selected by BIC

$\sigma_\varepsilon = 1$ and $m = 50$																
ρ_{23}	$\rho_{23} = 0$				$\rho_{23} = 0.25$				$\rho_{23} = 0.5$				$\rho_{23} = 0.75$			
Selected model	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110
$n = 50$	0	50	5	945	0	34	5	961	0	25	5	970	0	8	0	992
$n = 100$	0	2	0	998	0	1	0	999	0	0	0	1000	0	0	0	1000
$n = 200$	0	0	0	1000	0	20	0	1000	0	0	0	1000	0	0	0	1000
$n = 400$	0	0	0	1000	0	0	0	1000	0	0	0	1000	0	0	0	1000

Table 4.3: Number of times all possible models are selected via AIC_c and BIC in each of 1000 simulations for all the combinations of ρ_{23} when $\sigma_\varepsilon = 4$ and $m = 0$ for linear regression

(a) Number of times all possible models are selected by AIC_c

$\sigma_\varepsilon = 4$ and $m = 0$																
ρ_{23}	$\rho_{23} = 0$				$\rho_{23} = 0.25$				$\rho_{23} = 0.5$				$\rho_{23} = 0.75$			
Selected model	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110
$n = 50$	143	248	213	396	133	216	231	420	149	212	242	397	140	254	236	370
$n = 100$	27	117	120	736	27	116	135	722	16	120	125	739	18	110	135	737
$n = 200$	0	15	20	965	1	20	21	958	0	15	23	962	0	19	23	958
$n = 400$	0	2	0	998	0	0	0	1000	0	0	0	1000	0	0	0	1000

(b) Number of times all possible models are selected by BIC

$\sigma_\varepsilon = 4$ and $m = 0$																
ρ_{23}	$\rho_{23} = 0$				$\rho_{23} = 0.25$				$\rho_{23} = 0.5$				$\rho_{23} = 0.75$			
Selected model	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110
$n = 50$	366	241	222	171	327	246	245	182	370	217	234	179	357	257	225	161
$n = 100$	173	218	196	413	145	212	241	402	126	205	258	411	157	207	233	403
$n = 200$	16	105	99	780	16	94	92	798	21	85	102	792	21	114	105	760
$n = 400$	0	10	3	985	0	5	10	985	0	4	2	994	0	9	7	984

When $\sigma_\varepsilon = 4$, Table 4.3a and Table 4.3b show the number of times all possible models are selected via AIC_c and BIC for all the combinations of n and ρ_{23} without any missing data in variable X_2. For a small sample size and this larger error variance, model M100 was selected more frequently compared to the true model M110. As sample size increased, the tendency to choose model M100 decreased. On the other hand, as sample size increased, the true model M110 was selected more frequently by both model selection criteria. AIC_c chose the true model M110 more often than BIC as the sample size increased. BIC tends to select a smaller model.

Table 4.4: Number of times all possible models are selected via AIC_c and BIC in each of 1000 simulations for all the combinations of ρ_{23} when $\sigma_\varepsilon = 4$ and $m = 25$ for linear regression

(a) Number of times all possible models are selected by AIC_c

$\sigma_\varepsilon = 4$ and $m = 25$																
ρ_{23}	$\rho_{23} = 0$				$\rho_{23} = 0.25$				$\rho_{23} = 0.5$				$\rho_{23} = 0.75$			
Selected model	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110
$n = 50$	154	248	211	387	145	243	232	380	144	252	210	394	140	233	231	396
$n = 100$	29	184	122	665	29	187	112	672	44	179	138	639	17	136	133	714
$n = 200$	0	56	22	922	0	63	18	919	0	54	15	931	1	33	23	943
$n = 400$	0	0	0	1000	0	4	0	996	0	1	1	998	0	0	1	999

(b) Number of times all possible models are selected by BIC

$\sigma_\varepsilon = 4$ and $m = 25$																
ρ_{23}	$\rho_{23} = 0$				$\rho_{23} = 0.25$				$\rho_{23} = 0.5$				$\rho_{23} = 0.75$			
Selected model	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110
$n = 50$	346	231	247	176	339	246	246	169	344	229	260	167	372	226	228	174
$n = 100$	173	254	219	354	152	257	231	360	185	247	234	334	160	225	235	380
$n = 200$	23	163	103	711	20	149	104	727	20	143	102	735	22	124	117	737
$n = 400$	0	26	6	968	0	30	5	965	0	24	6	970	0	8	9	983

Table 4.4a and Table 4.4b show the number of times all possible models are selected via AIC_c and BIC when $\sigma_\varepsilon = 4$ for all the combinations of n and ρ_{23} with 25% of imputed values in variable X_2. For a small sample size and this larger variance, model M100 was selected via both criteria more frequently compared to true model M110. As sample size increases, the tendency to choose model M100 decreases. Whereas true model M110 was chosen via AIC_c almost 100% as sample size increases for all the combinations. On the other hand, the chances of BIC choosing the true model M110 is almost 97% as sample size increases. For larger sample size, the chances of BIC choosing the true model M110 increases as ρ_{23} increases.

Table 4.5a and Table 4.5b show the number of times all possible models are selected via AIC_c and BIC for all the combinations of ρ_{23} and $\sigma_\varepsilon = 4$ with 50% of imputed values in variable X_2. The true model M110 was chosen almost 100% as sample size increases for all the combinations of ρ_{23} and $\sigma_\varepsilon = 4$. The chances of AIC_c and BIC choosing the true model M110 is above 90% as sample size increases. AIC_c tends to choose true model M110 more often compared to BIC for 50% imputed data.

Table 4.5: Number of times all possible models are selected via AIC_c and BIC in each of 1000 simulations for all the combinations of ρ_{23} when $\sigma_\varepsilon = 4$ and $m = 50$ for linear regression

(a) Number of times all possible models are selected by AIC_c

$\sigma_\varepsilon = 4$ and $m = 50$																
ρ_{23}	$\rho_{23} = 0$				$\rho_{23} = 0.25$				$\rho_{23} = 0.5$				$\rho_{23} = 0.75$			
Selected model	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110
$n = 50$	141	256	226	377	147	246	243	364	162	255	239	344	144	245	225	386
$n = 100$	50	200	130	620	44	221	137	598	47	230	98	625	26	184	125	665
$n = 200$	4	104	29	863	2	117	22	859	2	82	24	892	2	65	16	917
$n = 400$	0	23	0	977	0	25	0	975	0	16	0	984	0	8	1	991

(b) Number of times all possible models are selected by BIC

$\sigma_\varepsilon = 4$ and $m = 50$																
ρ_{23}	$\rho_{23} = 0$				$\rho_{23} = 0.25$				$\rho_{23} = 0.5$				$\rho_{23} = 0.75$			
Selected model	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110
$n = 50$	330	222	282	166	301	226	299	174	340	228	288	144	346	209	262	183
$n = 100$	169	256	226	349	161	277	254	308	178	268	202	352	168	248	215	369
$n = 200$	36	212	118	634	36	231	118	615	26	207	104	663	34	187	91	688
$n = 400$	0	85	9	906	0	69	8	923	0	72	7	919	0	40	7	953

Table 4.6a and Table 4.6b show the MSE(P) for the best model selected via AIC_c and BIC for all combinations of the other parameters when $m = 0$. The MSE(P) decreases as sample size increases and the decrease is proportional to sample size. As σ_ε increases, MSE(P) increases and the increase is proportional to σ_ε^2. With $m = 0$, there is no imputation so ρ_{23} should make no difference. The decreases in MSE(P) values as ρ_{23} increases is just a sampling error. Both model selection criteria show similar results.

Table 4.6: MSE(P) for best model selected via AIC_c and BIC when $m = 0$ for linear regression

(a) MSE(P) for best model selected via AIC_c

AIC_c and $m = 0$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
σ_ε	0.25	1	4	0.25	1	4	0.25	1	4	0.25	1	4
$n = 50$	0.0038	0.0586	1.3368	0.0038	0.0612	1.3306	0.0039	0.0588	1.3284	0.0036	0.0609	1.3192
$n = 100$	0.0018	0.0278	0.6152	0.0018	0.0292	0.6248	0.0017	0.0284	0.5802	0.0017	0.0288	0.5946
$n = 200$	0.0009	0.0139	0.2378	0.0009	0.0138	0.2512	0.0009	0.0143	0.2475	0.0009	0.0139	0.2418
$n = 400$	0.0004	0.0071	0.1148	0.0005	0.0067	0.1141	0.0004	0.0073	0.1043	0.0004	0.0068	0.1127

(b) MSE(P) for best model selected via BIC

BIC and $m = 0$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
σ_ε	0.25	1	4	0.25	1	4	0.25	1	4	0.25	1	4
$n = 50$	0.0038	0.0591	1.7258	0.0038	0.0612	1.7072	0.0039	0.0588	1.7072	0.0036	0.0609	1.6791
$n = 100$	0.0018	0.0278	1.0031	0.0018	0.0292	0.9845	0.0017	0.0284	0.9417	0.0017	0.0288	0.9857
$n = 200$	0.0009	0.0139	0.3812	0.0009	0.0138	0.3770	0.0009	0.0143	0.3860	0.0009	0.0139	0.3995
$n = 400$	0.0004	0.0071	0.1227	0.0005	0.0067	0.1232	0.0004	0.0073	0.1082	0.0004	0.0068	0.1225

Table 4.7: MSE(P) for best model selected via AIC_c and BIC when $m = 25$ for linear regression

(a) MSE(P) for best model selected via AIC_c

AIC_c and $m = 25$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
σ_ε	0.25	1	4	0.25	1	4	0.25	1	4	0.25	1	4
$n = 50$	0.0451	0.0999	1.4358	0.0403	0.0986	1.4339	0.0372	0.0889	1.3641	0.0242	0.0818	1.3572
$n = 100$	0.0328	0.0533	0.6758	0.0304	0.0518	0.6895	0.0255	0.0482	0.7030	0.0168	0.0386	0.6488
$n = 200$	0.0263	0.0313	0.2965	0.0249	0.0312	0.2971	0.0206	0.0268	0.2782	0.0120	0.0203	0.2784
$n = 400$	0.0232	0.0204	0.1304	0.0222	0.0196	0.1295	0.0181	0.0168	0.1304	0.0102	0.0123	0.1165

(b) MSE(P) for best model selected via BIC

BIC and $m = 25$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
σ_ε	0.25	1	4	0.25	1	4	0.25	1	4	0.25	1	4
$n = 50$	0.0451	0.1004	1.7643	0.0403	0.1003	1.7677	0.0372	0.0894	1.7214	0.0242	0.0818	1.7547
$n = 100$	0.0328	0.0533	1.0497	0.0304	0.0518	1.0459	0.0255	0.0482	1.0719	0.0168	0.0386	1.0406
$n = 200$	0.0263	0.0313	0.4615	0.0249	0.0312	0.4490	0.0206	0.0268	0.4335	0.0120	0.0203	0.4406
$n = 400$	0.0232	0.0204	0.1491	0.0222	0.0196	0.1485	0.0181	0.0168	0.1466	0.0102	0.0123	0.1255

Table 4.7a and Table 4.7b show the MSE(P) for the best model selected via AIC_c and BIC for all combinations when $m = 25$. The MSE(P) decreases as sample size increases but the decrease is not proportional to sample size. As σ_ε increases, MSE(P) increases but the increase is not proportional to σ_ε^2. The MSE(P) values were much higher after imputation. The MSE(P) values decrease as ρ_{23} increases but there are no clear effects of ρ_{23} in the variation of MSE(P) values. The decreases and increases in MSE(P) values as ρ_{23} increases is just a sampling error. For larger variance, MSE(P) for best model selected via BIC is larger compared to AIC_c.

Table 4.8a and Table 4.8b show equivalent results for $m = 50$ in terms of sample size and σ_ε. The MSE(P) values were increased as percentages of missingness increased. For larger variance, MSE(P) for best model selected via BIC is larger compared to AIC_c. For larger error variance, the MSE(P) values decrease as ρ_{23} increases and there is a very substantial decrease for small sample size.

Table 4.8: MSE(P) for best model selected via AIC_c and BIC when $m = 50$ for linear regression

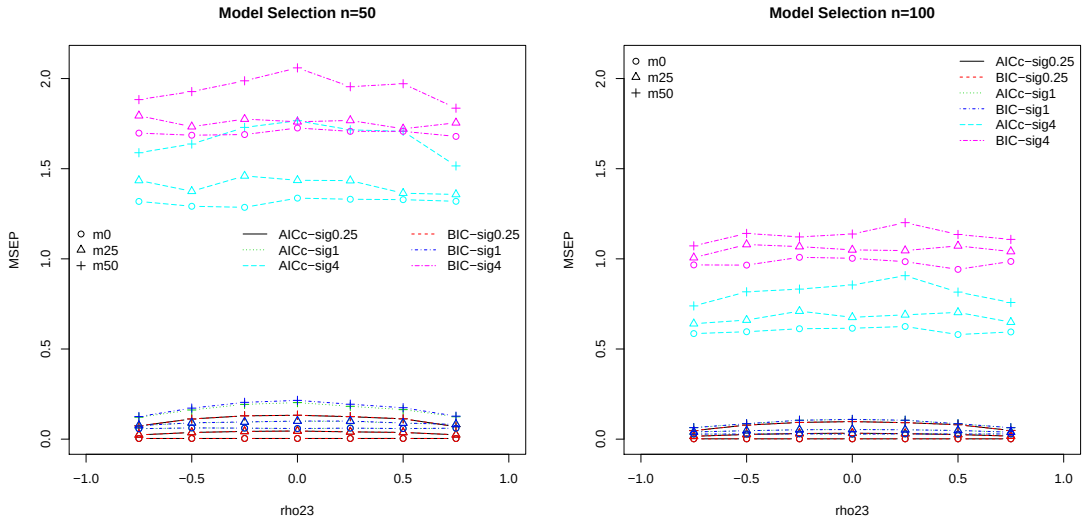
(a) MSE(P) for best model selected via AIC_c

AIC_c and $m = 50$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
σ_ε	0.25	1	4	0.25	1	4	0.25	1	4	0.25	1	4
$n = 50$	0.1324	0.2027	1.7660	0.1244	0.1821	1.7155	0.1126	0.1637	1.7072	0.0694	0.1242	1.5155
$n = 100$	0.0973	0.1088	0.8552	0.0915	0.1046	0.9066	0.0801	0.0856	0.8156	0.0482	0.0635	0.7580
$n = 200$	0.0832	0.0700	0.3946	0.0788	0.0653	0.4045	0.0647	0.0570	0.3573	0.0381	0.0370	0.3205
$n = 400$	0.0751	0.0535	0.1842	0.0714	0.049	0.1843	0.0586	0.0404	0.1693	0.0330	0.0245	0.1564

(b) MSE(P) for best model selected via BIC

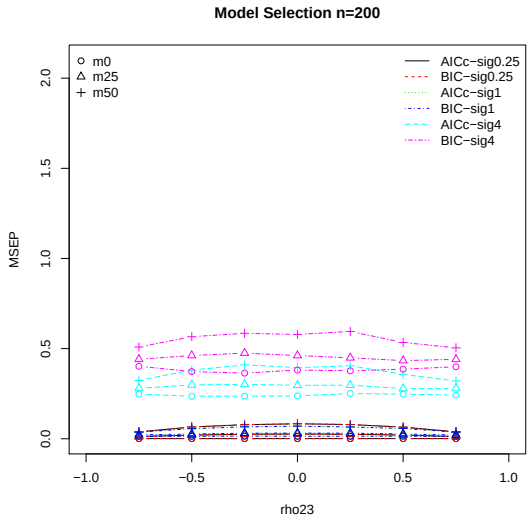
BIC and $m = 50$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
σ_ε	0.25	1	4	0.25	1	4	0.25	1	4	0.25	1	4
$n = 50$	0.1327	0.2150	2.0593	0.1257	0.1937	1.9548	0.1130	0.1743	1.9731	0.0694	0.1275	1.8357
$n = 100$	0.0973	0.1096	1.1377	0.0915	0.1050	1.2008	0.0801	0.0856	1.1351	0.0482	0.0635	1.1079
$n = 200$	0.0832	0.0700	0.5787	0.0788	0.0653	0.5952	0.0647	0.0570	0.5340	0.0381	0.0370	0.5043
$n = 400$	0.0751	0.0535	0.2251	0.0714	0.0490	0.2122	0.0586	0.0404	0.2088	0.0330	0.0245	0.1787

Figure 4.3a shows the MSE(P) for best model selected via AIC_c and BIC for each ρ_{23}, σ_ε, missing percentages and sample size, $n = 50$. The MSE(P) for best model selected via AIC_c is lower than the MSE(P) for best model selected via BIC especially for larger error variance. There is no clear difference between MSE(P) for best model selected via AIC_c and BIC for $\sigma_\varepsilon = 0.25$ and $\sigma_\varepsilon = 1$. Figure 4.3b, Figure 4.3c and Figure 4.3d show the MSE(P) for best model selected via AIC_c and BIC for each ρ_{23}, σ_ε, missing percentages and sample size, $n = 100$, $n = 200$ and $n = 400$ respectively. As sample size increases, the MSE(P) for best model selected via AIC_c and BIC decreases. However, the MSE(P) for best model selected via AIC_c is lower than the MSE(P) for best model selected via BIC for larger error variance. The MSE(P) for for best model selected via AIC_c and BIC is much higher for $\sigma_\varepsilon = 4$ and smaller sample size, however it decreases as sample size increases. Therefore, there is no clear effect of σ_ε for large sample size. In addition, there are no effects of ρ_{23} where the results showed a flat line for all combinations of σ_ε and sample size. For all combinations of σ_ε and sample size, the negative and positive correlations of the same magnitude showed similar results, with some variation for larger error variance.

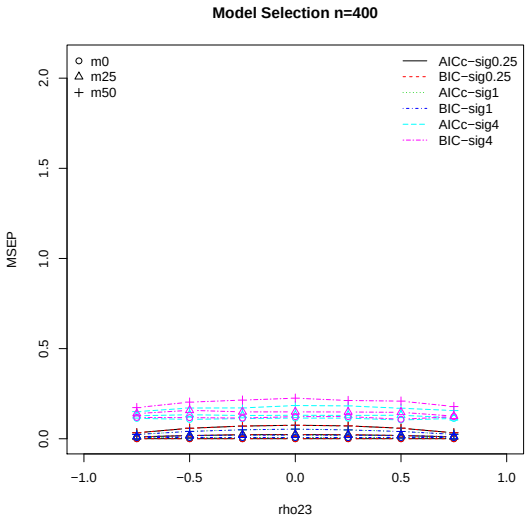


(a) $n = 50$

(b) $n = 100$



(c) $n = 200$



(d) $n = 400$

Figure 4.3: MSE(P) for best model selected via AIC_c and BIC for different sample sizes and linear regression

Table 4.9a and Table 4.9b show the MSE(P) achieved using model averaging via AIC_c and BIC for all combinations when $m = 0$. The MSE(P) decreases as sample size increases and it increases as σ_ε increases. Both model selection criteria show similar results. In general, the MSE(P) values for model averaging are lower than for the model selection procedures, especially for smaller sample sizes and higher error variances.

Table 4.9: MSE(P) for model averaging via AIC_c and BIC when $m = 0$ for linear regression

(a) MSE(P) for model averaging via AIC_c

AIC_c and $m = 0$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
σ_ε	0.25	1	4	0.25	1	4	0.25	1	4	0.25	1	4
$n = 50$	0.0037	0.0568	0.9665	0.0036	0.0614	0.9751	0.0036	0.0586	0.9314	0.0037	0.0607	0.9927
$n = 100$	0.0018	0.0302	0.4622	0.0018	0.0286	0.4426	0.0018	0.0287	0.4513	0.0018	0.0275	0.4721
$n = 200$	0.0009	0.0140	0.2327	0.0009	0.0138	0.2198	0.0009	0.0144	0.2304	0.0008	0.0140	0.2257
$n = 400$	0.0004	0.0068	0.1124	0.0004	0.0069	0.1096	0.0004	0.0072	0.1102	0.0004	0.0069	0.1040

(b) MSE(P) for model averaging via BIC

BIC and $m = 0$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
σ_ε	0.25	1	4	0.25	1	4	0.25	1	4	0.25	1	4
$n = 50$	0.0037	0.0568	0.9717	0.0036	0.0614	0.9788	0.0036	0.0586	0.9384	0.0037	0.0607	0.9988
$n = 100$	0.0018	0.0302	0.4647	0.0018	0.0286	0.4464	0.0018	0.0287	0.4530	0.0018	0.0275	0.4764
$n = 200$	0.0009	0.0140	0.2337	0.0009	0.0138	0.2211	0.0009	0.0144	0.2312	0.0008	0.0140	0.2266
$n = 400$	0.0004	0.0068	0.1124	0.0004	0.0069	0.1097	0.0004	0.0072	0.1103	0.0004	0.0069	0.1041

Table 4.10: MSE(P) for model averaging via AIC_c and BIC when $m = 25$ for linear regression

(a) MSE(P) for model averaging via AIC_c

AIC_c and $m = 25$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
σ_ε	0.25	1	4	0.25	1	4	0.25	1	4	0.25	1	4
$n = 50$	0.0444	0.0976	1.1429	0.0423	0.0975	1.0856	0.0365	0.0891	1.1720	0.0247	0.0751	1.0469
$n = 100$	0.0331	0.0512	0.5623	0.0316	0.0514	0.5568	0.0267	0.0464	0.5372	0.0154	0.0391	0.5052
$n = 200$	0.0259	0.0303	0.2696	0.0257	0.0305	0.2560	0.0206	0.0270	0.2644	0.0121	0.0209	0.2436
$n = 400$	0.0234	0.0205	0.1353	0.0220	0.0195	0.1288	0.0177	0.0175	0.1251	0.0101	0.0120	0.1195

(b) MSE(P) for model averaging via BIC

BIC and $m = 25$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
σ_ε	0.25	1	4	0.25	1	4	0.25	1	4	0.25	1	4
$n = 50$	0.0444	0.0977	1.1466	0.0423	0.0976	1.0906	0.0365	0.0891	1.1760	0.0247	0.0752	1.0505
$n = 100$	0.0331	0.0512	0.5660	0.0316	0.0514	0.5613	0.0267	0.0464	0.5417	0.0154	0.0391	0.5096
$n = 200$	0.0259	0.0303	0.2706	0.0257	0.0305	0.2573	0.0206	0.0270	0.2650	0.0121	0.0209	0.2451
$n = 400$	0.0234	0.0205	0.1353	0.0220	0.0195	0.1289	0.0177	0.0175	0.1252	0.0101	0.0120	0.1195

Table 4.10a and Table 4.10b show the MSE(P) for model averaging via AIC_c and BIC for all combinations when $m = 25$. The MSE(P) decreases as sample size increases and it increases as σ_ε increases. The MSE(P) values were higher after imputation. MSE(P) decreases as sample size increase. The MSE(P) values decrease as ρ_{23} increases but there are no clear effects of ρ_{23} in the variation of MSE(P) values. Both model selection

criteria show similar results. The MSE(P) values for model averaging are lower than for the model selection procedures after imputation.

Table 4.11: MSE(P) for model averaging via AIC_c and BIC when $m = 50$ for linear regression

(a) MSE(P) for model averaging via AIC_c

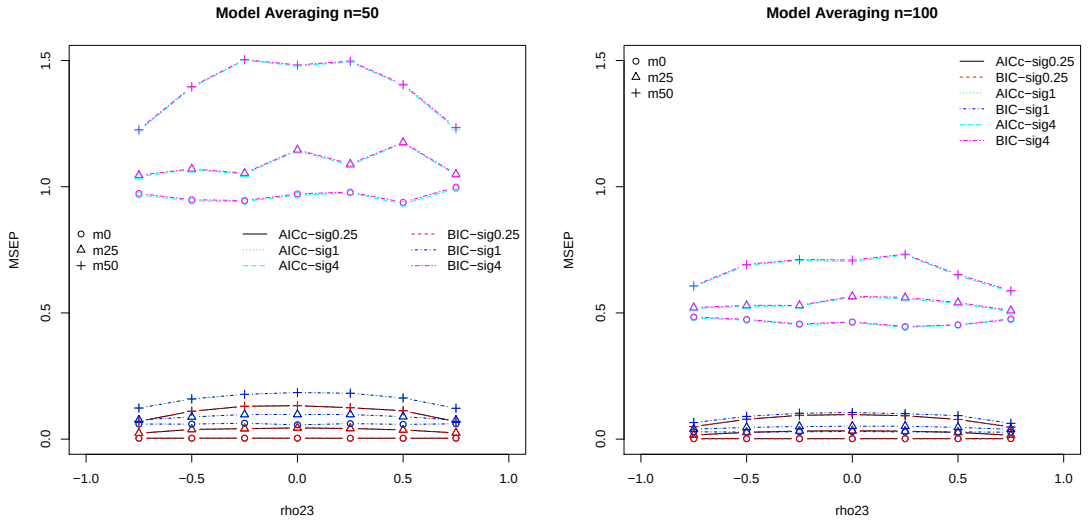
AIC_c and $m = 50$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
σ_ε	0.25	1	4	0.25	1	4	0.25	1	4	0.25	1	4
$n = 50$	0.1325	0.1845	1.4787	0.1245	0.1816	1.4938	0.1127	0.1629	1.4012	0.0695	0.1223	1.2278
$n = 100$	0.0974	0.1054	0.7041	0.0927	0.1008	0.7294	0.0781	0.0931	0.6471	0.0481	0.0626	0.5839
$n = 200$	0.0820	0.0707	0.3680	0.0780	0.0643	0.3545	0.0648	0.0559	0.3336	0.0382	0.0363	0.2863
$n = 400$	0.0745	0.0504	0.1836	0.0707	0.0499	0.1830	0.0592	0.0393	0.1563	0.0325	0.0240	0.1374

(b) MSE(P) for model averaging via BIC

BIC and $m = 50$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
σ_ε	0.25	1	4	0.25	1	4	0.25	1	4	0.25	1	4
$n = 50$	0.1325	0.1845	1.4826	0.1245	0.1819	1.4978	0.1126	0.1630	1.4060	0.0695	0.1223	1.2346
$n = 100$	0.0974	0.1054	0.7098	0.0927	0.1008	0.7327	0.0781	0.0931	0.6523	0.0481	0.0626	0.5883
$n = 200$	0.0820	0.0707	0.3704	0.0780	0.0643	0.3564	0.0648	0.0559	0.3345	0.0382	0.0363	0.2871
$n = 400$	0.0745	0.0504	0.1837	0.0707	0.0499	0.1831	0.0592	0.0393	0.1565	0.0325	0.0240	0.1375

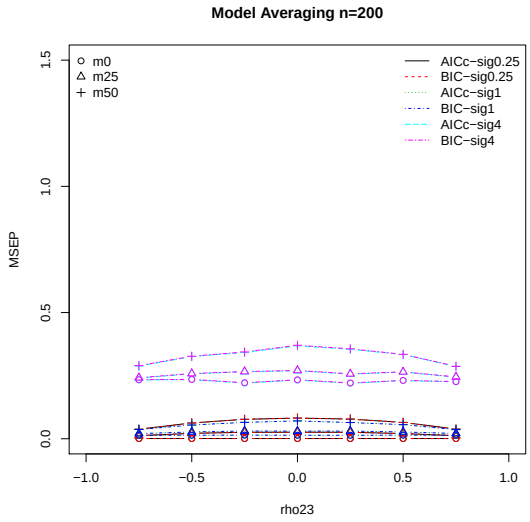
Table 4.11a and Table 4.11b show equivalent results as $m = 25$ in terms of sample size and σ_ε. The MSE(P) values were increases as percentages of missingness increases. MSE(P) decreases as ρ_{23} increases but there is no substantial decrease in the MSE(P) values. Both model selection criteria show similar results. The MSE(P) values for model averaging are lower than for the model selection procedures after imputation.

Figure 4.4a, Figure 4.4b, Figure 4.4c and Figure 4.4d show the MSE(P) for model averaging via AIC_c and BIC for each ρ_{23}, σ_ε, missing percentages and sample sizes, $n = 50$, $n = 100$, $n = 200$ and $n = 400$ respectively. There is no clear difference between MSE(P) for model averaging via AIC_c and BIC for all σ_ε and sample sizes. As sample size increases, the MSE(P) for model averaging via AIC_c and BIC decreases. The MSE(P) for model averaging via AIC_c and BIC is much higher for $\sigma_\varepsilon = 4$ and smaller sample size, however it decreases as sample size increases. Therefore, there is no clearer effect of σ_ε for large sample size. Moreover, there is no difference between MSE(P) for model averaging via AIC_c and BIC for all σ_ε in larger sample size. There are no effects of ρ_{23} where the results showed a flat line for all combinations of σ_ε and large sample size. Whereas for smaller sample size and larger error variance, MSE(P) values decreases as ρ_{23} increases. For all combinations of σ_ε and sample size, the negative and positive correlations of the same magnitude showed similar results, with some variation for larger error variance.

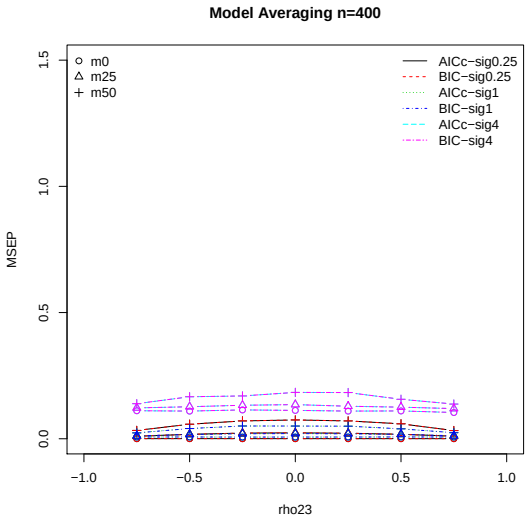


(a) $n = 50$

(b) $n = 100$



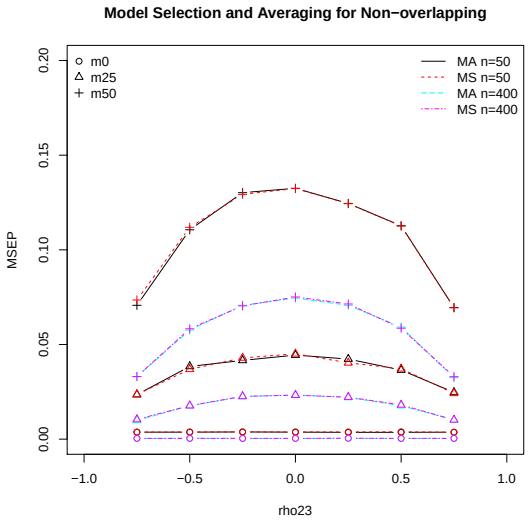
(c) $n = 200$



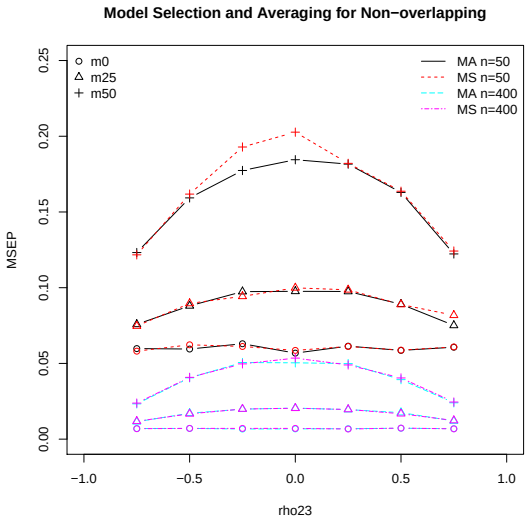
(d) $n = 400$

Figure 4.4: MSE(P) for model averaging via AIC_c and BIC for each ρ_{23}, σ_ε, missing percentages and all sample sizes for linear regression

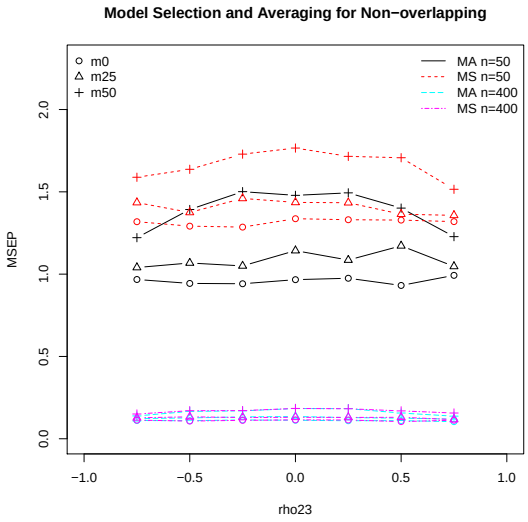
Figure 4.5a, Figure 4.5b and Figure 4.5c show comparison between model averaging and model selection for non-overlapping via AIC_c for each ρ_{23}, σ_ε, missing percentages and sample sizes, $n = 50$ and $n = 400$. The MSE(P) for model averaging via AIC_c is lower than the MSE(P) for best model selection via AIC_c for larger error variance and small sample size for each ρ_{23}. As sample size increases, the MSE(P) for model averaging and model selection via AIC_c decreases. Moreover, there is no difference between MSE(P) for model averaging and model selection via AIC_c for different values of σ_ε for larger sample size.



(a) $\sigma_\varepsilon = 0.25$



(b) $\sigma_\varepsilon = 1$



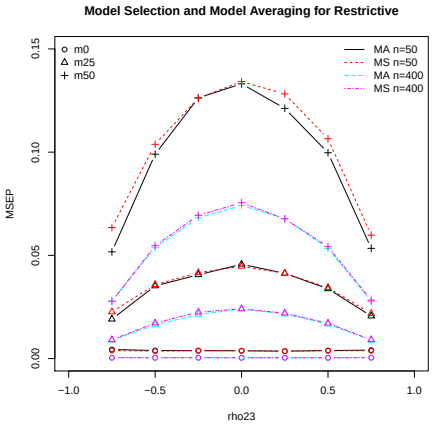
(c) $\sigma_\varepsilon = 4$

Figure 4.5: Comparison between model averaging and model selection for non-overlapping variable sets via AIC_c for each ρ_{23}, σ_ε, missing percentages and sample sizes, $n = 50$ and $n = 400$ for linear regression

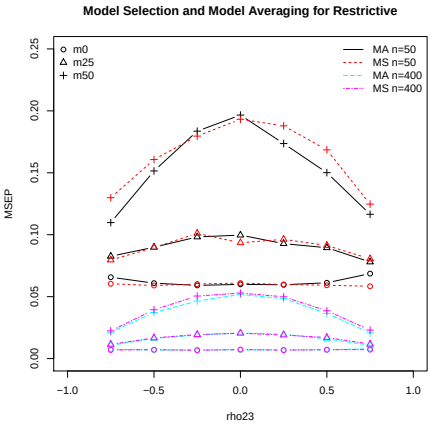
4.2.2 Linear regression with restrictive and inclusive strategies

The MSE(P) for model averaging with a restrictive strategy is lower than the MSE(P) for best model selected via AIC_c for larger error variance and small sample size for each ρ_{23}. There is no difference between model averaging and model selection via AIC_c for smaller error variance and larger sample size for each ρ_{23}. As sample size increases, the MSE(P) for model averaging and model selection via AIC_c decreases. Figure 4.6a, Figure 4.6b

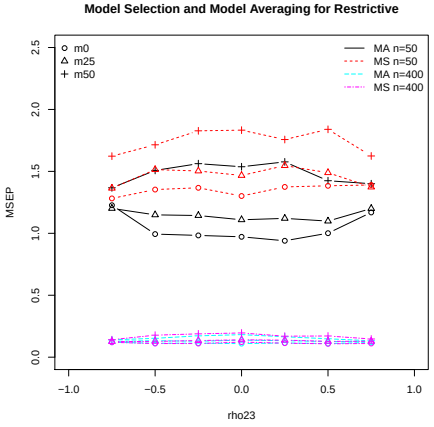
and Figure 4.6c show the comparison between model averaging and model selection for restrictive strategy via AIC_c for each ρ_{23}, σ_ε, missing percentages and sample sizes, $n = 50$ and $n = 400$. For restrictive strategy, the MSE(P) for model averaging and model selection decreases as $|\rho_{23}|$ increases for $\sigma_\varepsilon = 0.25$ and $\sigma_\varepsilon = 1$ for all sample size. For $\sigma_\varepsilon = 4$, there are no effects of $|\rho_{23}|$.



(a) $\sigma_\varepsilon = 0.25$



(b) $\sigma_\varepsilon = 1$

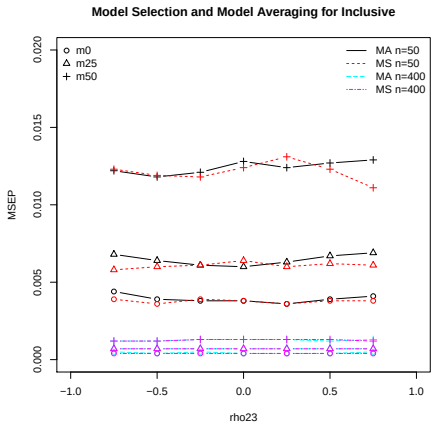


(c) $\sigma_\varepsilon = 4$

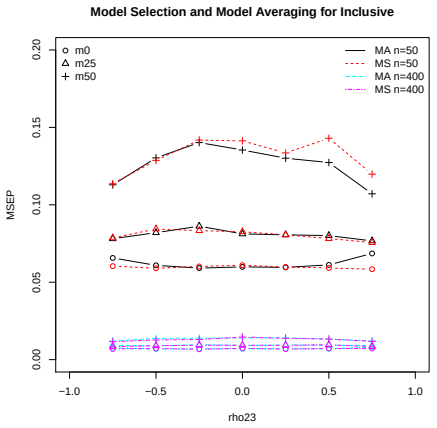
Figure 4.6: Comparison between model averaging and model selection for restrictive strategy via AIC_c for each ρ_{23}, σ_ε, missing percentages and sample sizes, $n = 50$ and $n = 400$ for linear regression

Figure 4.7a, Figure 4.7b and Figure 4.7c show comparison between model averaging and model selection for inclusive strategy via AIC_c for each ρ_{23}, σ_ε, missing percentages and sample sizes, $n = 50$ and $n = 400$. The MSE(P) for model averaging with inclusive strategy is lower than the MSE(P) for best model selected via AIC_c for larger error variance and small sample size for each ρ_{23}. The MSE(P) for model averaging and

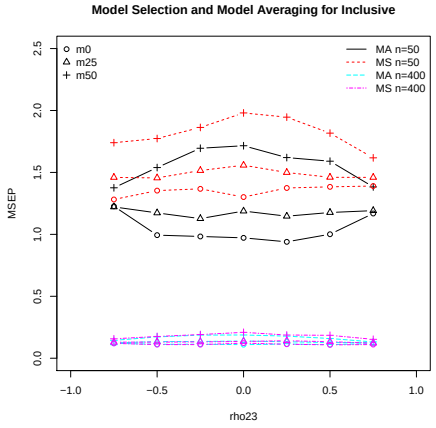
model selection decreases as $|\rho_{23}|$ increases for for $\sigma_\varepsilon = 0.25$ and $\sigma_\varepsilon = 1$ for all sample size. For $\sigma_\varepsilon = 4$, there are no effects of $|\rho_{23}|$.



(a) $\sigma_\varepsilon = 0.25$



(b) $\sigma_\varepsilon = 1$

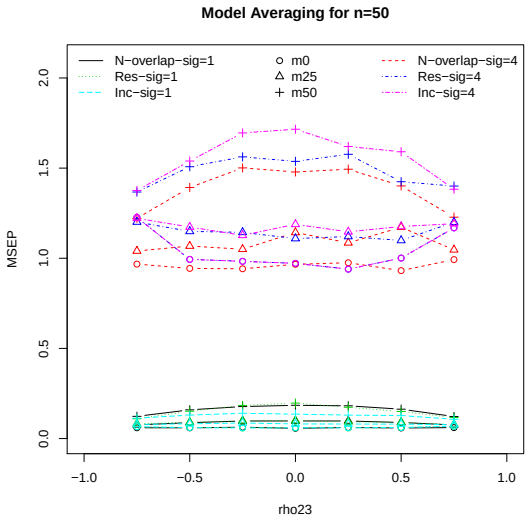


(c) $\sigma_\varepsilon = 4$

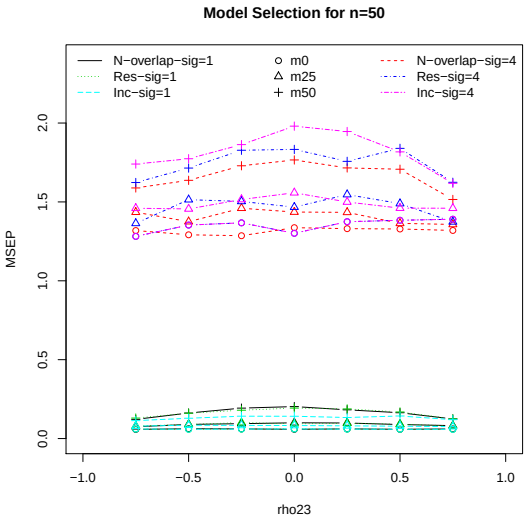
Figure 4.7: Comparison between model averaging and model selection for inclusive strategy via AIC_c for each ρ_{23}, σ_ε, missing percentages and sample sizes, $n = 50$ and $n = 400$ for linear regression

Figure 4.8 shows the comparison between all three model-building strategies (non-overlapping variable sets, restrictive and inclusive strategies) for model averaging and model selection via AIC_c for each ρ_{23}, σ_ε, missing percentages and sample sizes, $n = 50$ and $n = 400$. For $\sigma_\varepsilon = 1$, the MSE(P) for model averaging with an inclusive strategy is lower than the MSE(P) for non-overlapping variable sets and restrictive strategy for small sample size and $m = 50$. Whereas for $\sigma_\varepsilon = 4$, the MSE(P) for model averaging with non-overlapping variable sets is lower than the MSE(P) for restrictive and inclusive strategies for all sample size. There are no clear difference between the MSE(P) of all three model-building strategies for different values of ρ_{23} for all combinations of σ_ε,

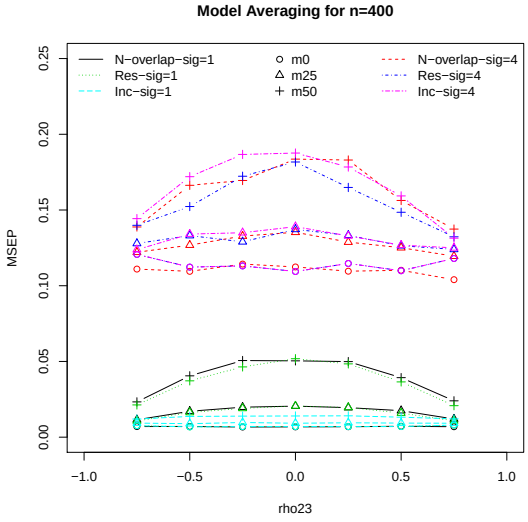
missing percentages and sample sizes. There is no effect of the negative and positive correlations of same magnitude for model averaging and model selection for all three model-building strategies.



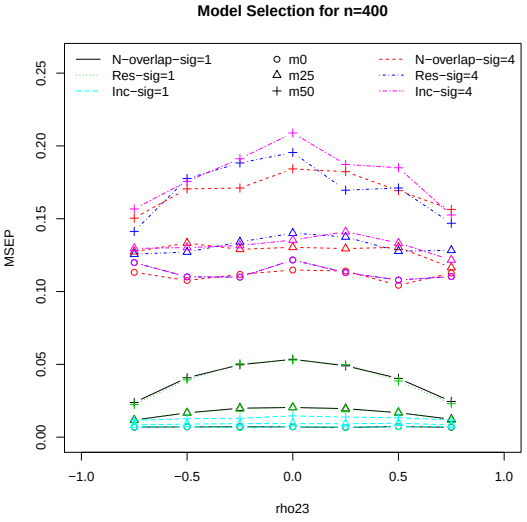
(a) Model Averaging $n = 50$



(b) Model Selection $n = 50$



(c) Model Averaging $n = 400$



(d) Model Selection $n = 400$

Figure 4.8: Comparison between all three model-building strategies for model averaging and model selection for each ρ_{23}, σ_ε, missing percentages and sample size ($n = 50$ and $n = 400$) for linear regression

4.2.3 Logistic regression with non-overlapping variable sets

This simulation study was conducted based on the simulation design discussed in Section 4.1 with a Logistic regression as in Equation (4.4). $\mathbf{X}$ (X_1, X_2 and X_3) values were simulated based on a multivariate normal distribution with fixed zero means, all variances equal to 1 and zero correlations except (generally) for ρ_{23} (the correlation between X_2 and X_3). The analysis was carried out for every combination of n, m and correlation $\rho_{23} = -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75$. Table 4.12a and Table 4.12b show the number of times all possible models were selected via AIC_c and BIC in each of 1000 simulations for all the combinations of ρ_{23} and n without any missing data in variable X_2.

Table 4.12: Number of times all possible models are selected by AIC_c and BIC in each of 1000 simulations for all the combinations of ρ_{23} when $m = 0$ for logistic regression

(a) Number of times all possible models are selected by AIC_c

$m = 0$																
ρ_{23}	$\rho_{23} = 0$				$\rho_{23} = 0.25$				$\rho_{23} = 0.5$				$\rho_{23} = 0.75$			
Selected model	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110
$n = 50$	9	82	85	824	6	96	76	822	7	79	65	849	11	99	67	823
$n = 100$	0	9	8	983	0	9	6	985	0	8	7	985	0	1	6	993
$n = 200$	0	0	0	1000	0	0	0	1000	0	0	0	1000	0	0	0	1000
$n = 400$	0	0	0	1000	0	0	0	1000	0	0	0	1000	0	0	0	1000

(b) Number of times all possible models are selected by BIC

$m = 0$																
ρ_{23}	$\rho_{23} = 0$				$\rho_{23} = 0.25$				$\rho_{23} = 0.5$				$\rho_{23} = 0.75$			
Selected model	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110
$n = 50$	42	168	174	616	50	174	172	604	54	162	160	624	48	182	146	624
$n = 100$	2	37	32	929	3	43	27	927	2	46	37	915	3	32	32	933
$n = 200$	0	1	0	999	0	1	2	997	0	1	2	997	0	1	1	998
$n = 400$	0	0	0	1000	0	0	0	1000	0	0	0	1000	0	0	0	1000

Table 4.13a and Table 4.13b show the number of times all possible models were selected via AIC_c and BIC in each 1000 simulations for all the combinations of ρ_{23} and n with imputed values of 25% in variable X_2. Both table show that, the true model M110 was selected 100% as sample size increases for all the combinations of ρ_{23}. AIC_c choose the true model M110 more frequently than BIC as the sample size increases. BIC tends to select a smaller model. BIC tends to choose model M100 more frequently for smaller sample size. For smaller sample size, the number of times the true model M110 are selected increases as correlation values increases.

Table 4.13: Number of times all possible models are selected by AIC_c and BIC in each of 1000 simulations for all the combinations of ρ_{23} when $m = 25$ for logistic regression

(a) Number of times all possible models are selected by AIC_c

$m = 25$																
ρ_{23}	$\rho_{23} = 0$				$\rho_{23} = 0.25$				$\rho_{23} = 0.5$				$\rho_{23} = 0.75$			
Selected model	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110
$n = 50$	11	155	105	729	13	135	79	773	11	139	83	767	13	105	77	805
$n = 100$	0	27	5	968	0	24	12	964	0	20	8	972	0	16	6	978
$n = 200$	0	1	0	999	0	0	0	1000	0	0	0	1000	0	0	0	1000
$n = 400$	0	0	0	1000	0	0	0	1000	0	0	0	1000	0	0	0	1000

(b) Number of times all possible models are selected by BIC

$m = 25$																
ρ_{23}	$\rho_{23} = 0$				$\rho_{23} = 0.25$				$\rho_{23} = 0.5$				$\rho_{23} = 0.75$			
Selected model	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110
$n = 50$	71	231	185	513	67	200	144	589	57	219	161	563	59	185	150	606
$n = 100$	3	96	42	859	2	79	35	884	3	67	39	891	1	56	43	900
$n = 200$	0	4	1	996	0	2	0	998	0	5	2	993	0	2	1	997
$n = 400$	0	0	0	1000	0	0	0	1000	0	0	0	1000	0	0	0	1000

Table 4.14: Number of times all possible models are selected by AIC_c and BIC in each of 1000 simulations for all the combinations of ρ_{23} when $m = 50$ for logistic regression

(a) Number of times all possible models are selected by AIC_c

$m = 50$																
ρ_{23}	$\rho_{23} = 0$				$\rho_{23} = 0.25$				$\rho_{23} = 0.5$				$\rho_{23} = 0.75$			
Selected model	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110
$n = 50$	32	176	97	695	10	193	118	679	25	179	96	700	14	173	88	725
$n = 100$	1	86	14	899	0	68	9	923	0	67	12	921	0	49	11	940
$n = 200$	0	10	0	990	0	9	0	991	0	3	0	997	0	1	0	999
$n = 400$	0	0	0	1000	0	0	0	1000	0	0	0	1000	0	0	0	1000

(b) Number of times all possible models are selected by BIC

$m = 50$																
ρ_{23}	$\rho_{23} = 0$				$\rho_{23} = 0.25$				$\rho_{23} = 0.5$				$\rho_{23} = 0.75$			
Selected model	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110	M000	M100	M010	M110
$n = 50$	88	219	150	543	78	242	175	505	91	225	158	526	67	234	160	539
$n = 100$	8	150	57	785	6	130	46	818	2	151	45	802	10	110	54	826
$n = 200$	0	34	0	966	0	35	2	963	0	27	2	971	0	13	1	986
$n = 400$	0	0	0	1000	0	0	0	1000	0	3	0	997	0	0	0	1000

Table 4.14a and Table 4.14b show the number of times all possible models were selected via AIC_c and BIC in each of 1000 simulations for all the combinations of ρ_{23} and n with imputed values of 50% in variable X_2. AIC_c choose the true model M110 more frequently compared to BIC as the sample size increases. BIC tends to choose model M100 more often compared to model M110 for smaller sample size. There are no effects of ρ_{23} in the frequency of selecting true model M110 for all missing percentages.

Table 4.15a and Table 4.15b show the MSE(P) for the best model selected via AIC_c and BIC for all combinations. The MSE(P) decreases as the sample size increases and it increases as missing percentages increases. AIC_c and BIC show similar results as

sample size increases. There are no effects of ρ_{23} in the variation of MSE(P) values. The MSE(P) values are proportional to sample sizes when there are no missing data observed in variable X_2.

Table 4.15: MSE(P) for best model selected via AIC_c and BIC for logistic regression

(a) MSE(P) for best model selected via AIC_c

AIC_c												
ρ_{23}	$\rho_{23}=0$			$\rho_{23}=0.25$			$\rho_{23}=0.5$			$\rho_{23}=0.75$		
m	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$n=50$	0.0139	0.0195	0.0246	0.0146	0.0178	0.0234	0.0139	0.0177	0.0240	0.0148	0.0162	0.0246
$n=100$	0.0058	0.0072	0.0106	0.0056	0.0074	0.0101	0.0058	0.0068	0.0098	0.0055	0.0062	0.0086
$n=200$	0.0027	0.0034	0.0050	0.0028	0.0032	0.0049	0.0028	0.0031	0.0044	0.0027	0.0030	0.0041
$n=400$	0.0013	0.0017	0.0023	0.0013	0.0017	0.0022	0.0014	0.0016	0.0022	0.0013	0.0016	0.0019

(b) MSE(P) for best model selected via BIC

BIC												
ρ_{23}	$\rho_{23}=0$			$\rho_{23}=0.25$			$\rho_{23}=0.5$			$\rho_{23}=0.75$		
m	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$n=50$	0.0198	0.0260	0.0293	0.0209	0.0237	0.0286	0.0205	0.0237	0.0251	0.0206	0.0221	0.0249
$n=100$	0.0070	0.0095	0.0130	0.0068	0.0091	0.0123	0.0073	0.0085	0.0122	0.0068	0.0079	0.0111
$n=200$	0.0027	0.0034	0.0054	0.0028	0.0032	0.0053	0.0028	0.0032	0.0048	0.0028	0.0031	0.0043
$n=400$	0.0013	0.0017	0.0023	0.0013	0.0017	0.0022	0.0014	0.0016	0.0022	0.0013	0.0016	0.0019

Table 4.16: MSE(P) for model averaging via AIC_c and BIC for logistic regression

(a) MSE(P) for model averaging via AIC_c

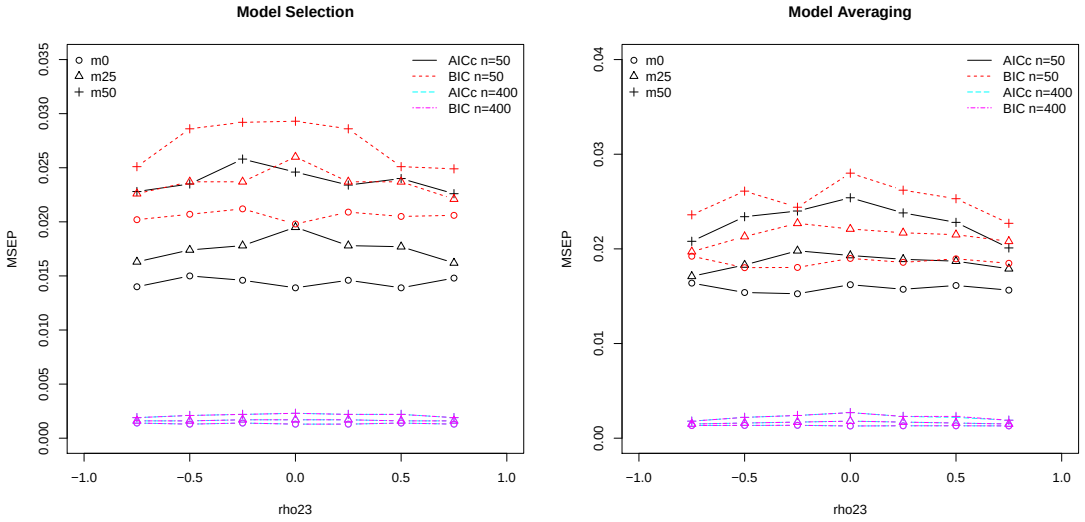
AIC_c												
ρ_{23}	$\rho_{23}=0$			$\rho_{23}=0.25$			$\rho_{23}=0.5$			$\rho_{23}=0.75$		
m	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$n=50$	0.0162	0.0193	0.0254	0.0157	0.0189	0.0238	0.0161	0.0187	0.0228	0.0156	0.0179	0.0201
$n=100$	0.0062	0.0080	0.0113	0.0062	0.0081	0.0112	0.0065	0.0077	0.0107	0.0062	0.0077	0.0092
$n=200$	0.0027	0.0036	0.0054	0.0028	0.0034	0.0052	0.0027	0.0032	0.0046	0.0028	0.0031	0.0037
$n=400$	0.0013	0.0018	0.0027	0.0013	0.0017	0.0023	0.0013	0.0016	0.0022	0.0013	0.0015	0.0019

(b) MSE(P) for model averaging via BIC

BIC												
ρ_{23}	$\rho_{23}=0$			$\rho_{23}=0.25$			$\rho_{23}=0.5$			$\rho_{23}=0.75$		
m	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$n=50$	0.0190	0.0221	0.0280	0.0186	0.0217	0.0262	0.0189	0.0215	0.0253	0.0185	0.0208	0.0227
$n=100$	0.0073	0.0097	0.0135	0.0075	0.0098	0.0133	0.0078	0.0092	0.0127	0.0074	0.0092	0.0110
$n=200$	0.0028	0.0038	0.0060	0.0029	0.0036	0.0058	0.0028	0.0033	0.0051	0.0028	0.0032	0.0039
$n=400$	0.0013	0.0018	0.0027	0.0013	0.0017	0.0023	0.0013	0.0016	0.0023	0.0013	0.0015	0.0019

Table 4.16a and Table 4.16b show the MSE(P) for model averaging via AIC_c and BIC for all combinations. The MSE(P) decreases as the sample size increases and it increases

as missing percentages increases. AIC_c and BIC show similar results as sample size increases. The variation in MSE(P) values are not significant.



(a) Model selection

(b) Model averaging

Figure 4.9: MSE(P) for best model selected and model averaging via AIC_c and BIC for each ρ_{23}, missing percentages and sample sizes (n=50 and n=400) for logistic regression

Figure 4.9a shows the MSE(P) for best model selected via AIC_c and BIC for each ρ_{23}, missing percentages and sample size, $n = 50$ and $n = 400$. It shows that there is no significant difference between MSE(P) value for negative and positive ρ_{23} values. For small sample size, there is a clear difference between the MSE(P) values for model selected via AIC_c and BIC but there is no difference as sample size increases. AIC_c performs better in terms of prediction for small sample size.

Figure 4.9b shows the MSE(P) for model averaging via AIC_c and BIC for each ρ_{23}, missing percentages and sample sizes. It shows that there is no difference between MSE(P) values for negative and positive ρ_{23} values. There is no difference between MSE(P) for model averaging via AIC_c and BIC for large sample size. It is clear that MSE(P) for model averaging via AIC_c and BIC decreases as sample size increases. There are no effects of ρ_{23} in the variation of MSE(P) values where the lines in the Figure 4.9a and Figure 4.9b are stationary as ρ_{23} increases for large sample size.

Figure 4.10 shows comparison between model averaging and model selection for non-overlapping variable sets via AIC_c for each ρ_{23}, missing percentages and sample sizes, $n = 50$ and $n = 400$. The results shows that MSE(P) for model selection are lower than model averaging for small sample sizes and $m = 50$. For large sample sizes, there

are no differences between MSE(P) values for model averaging and model selection via AIC_c for all combinations. There are no differences between MSE(P) values of model selection and model averaging via AIC_c for negative and positive ρ_{23} values of the same magnitude.

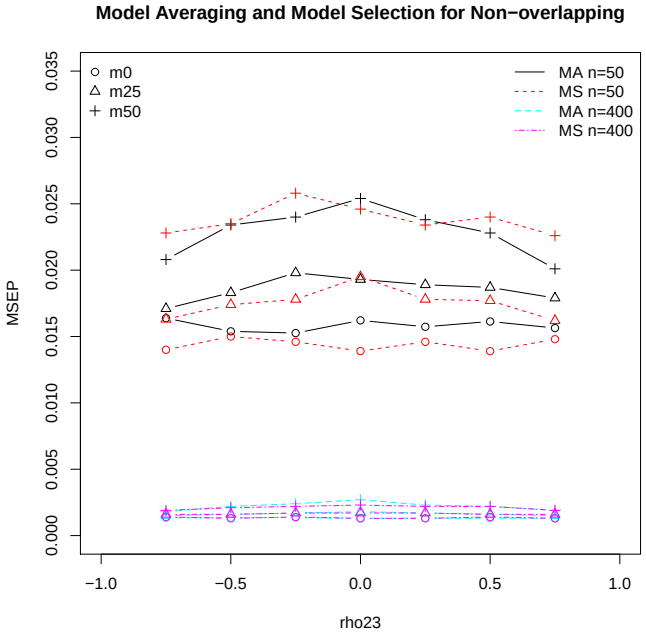
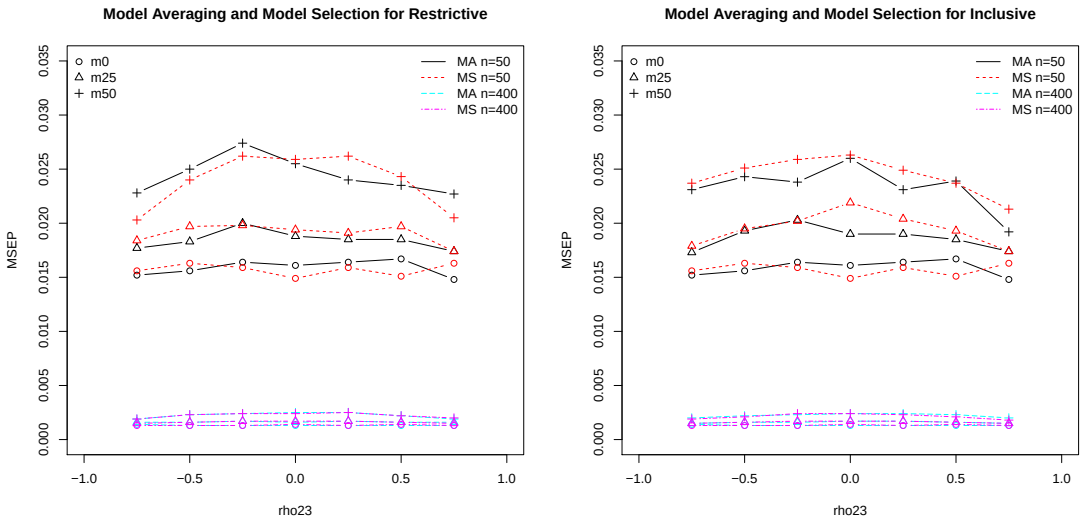


Figure 4.10: Comparison between model averaging and model selection for non-overlapping variable sets via AIC_c for each ρ_{23}, missing percentages and sample sizes (n=50 and n=400)for logistic regression

4.2.4 Logistic regression with restrictive and inclusive Strategies

The MSE(P) for model selection with restrictive strategy is lower than model averaging for all sample size. There are no differences between MSE(P) values of model averaging and model selection with restrictive strategy for negative and positive ρ_{23} values of the same magnitude. Figure 4.11a shows comparison between model averaging and model selection for restrictive strategy via AIC_c for each ρ_{23}, missing percentages and sample sizes, $n = 50$ and $n = 400$. Figure 4.11b shows comparison between model averaging and model selection for inclusive strategy via AIC_c for each ρ_{23}, missing percentages and sample sizes, $n = 50$ and $n = 400$. The MSE(P) for model selection with inclusive strategy is lower than model averaging for small sample size. There are no differences between MSE(P) values for model averaging and model selection for all combinations of $|\rho_{23}|$ and large sample size. There are no effects of $|\rho_{23}|$ in the variation of MSE(P) values for all sample size for both restrictive and inclusive strategies. There is no difference

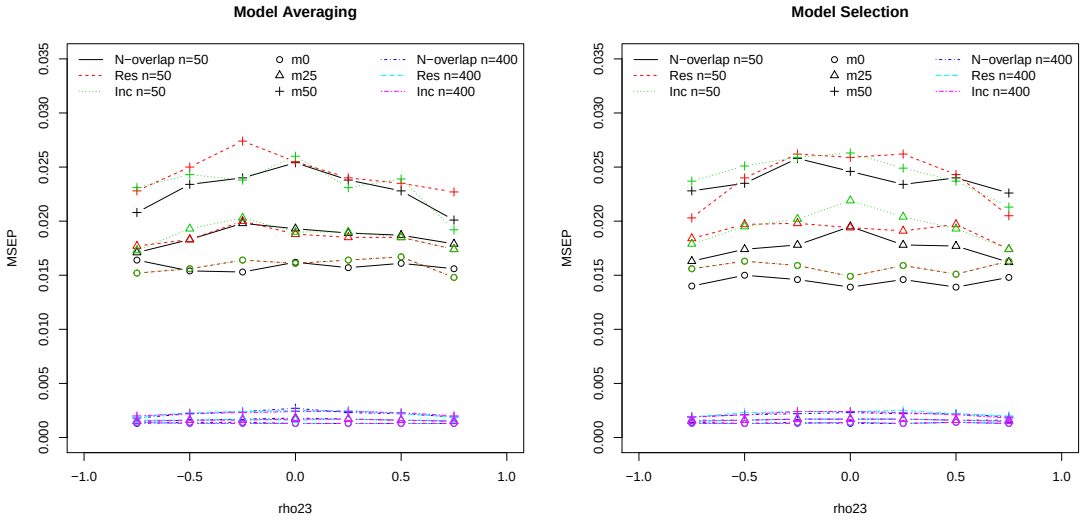
between restrictive and inclusive strategies of MSE(P) values for model averaging and model selection in terms of predictions.



(a) Restrictive strategy

(b) Inclusive strategy

Figure 4.11: Comparison between model averaging and model selection for restrictive and inclusive strategies via AIC_c for each ρ_{23}, missing percentages and sample sizes (n=50 and n=400) for logistic regression



(a) Model Averaging

(b) Model Selection

Figure 4.12: Comparison between all three model-building strategies for model averaging and model se4lection for each ρ_{23}, missing percentages and sample size (n=50 and n=400)for logistic regression

Figure 4.12 shows the all three model-building strategies (non-overlapping variable sets, restrictive and inclusive strategies) for model averaging and model selection via AIC_c for each ρ_{23}, missing percentages and sample sizes, $n = 50$ and $n = 400$. The MSE(P) values for model averaging using non-overlapping variable sets is lower than MSE(P) values for model averaging with restrictive and inclusive strategies for small sample sizes. There are no differences between the MSE(P) values for model averaging and model selection with all three model-building strategies for negative and positive correlations of same magnitude. The MSE(P) values of model selection using non-overlapping variable sets is lower than the MSE(P) values of model selection using restrictive and inclusive strategies for small sample size. There are no difference between all three strategies for large sample size.

4.3 Discussion and Conclusions

The performance of AIC_c and BIC for model selection and model averaging in linear model and Logistic regression was observed. The effects of simulation parameters (sample size (n), missing percentages (m), the correlation between X_2 and X_3 (ρ_{23}) and error variance (σ_ε^2)) on model selection and averaging also were observed.

In both linear model and Logistic regression, there are important effects of all the other simulation parameters even for complete dataset ($m = 0$). As the sample size increases, the tendency to choose true model M110 is increased in both linear model and Logistic regression. AIC_c chooses the true model more often than BIC as sample size increases. The MSE(P) for the selected best model and for model averaging decrease as sample size increases in both linear model and Logistic regression.

The error variance (σ_ε^2) has a significant effect on model selection and model averaging in complete case analysis in linear models. For larger error variance in linear models, models that are smaller than the true model (especially model M100) are selected more often. The MSE(P) for the selected best model and model averaging are increased as σ_ε increased.

Besides that, AIC_c performs better than BIC in linear models when the error variance is larger. AIC_c chooses the true model more often whereas BIC is more likely to select smaller models in both linear model and generalized models. As stated by Claeskens and Hjort [2008], the BIC penalty is stricter than the AIC so bigger models (with larger numbers of parameters) will receive a heavier 'punishment'. There are differences between the model selection criteria in terms of prediction in Logistic regression. BIC seems a bit worse than AIC_c for complete-cases when $n \leq 100$. Claeskens and Hjort

[2008] stated that there is no established theoretical reason for using AIC_c for Logistic regression, so researchers should use it with care. Our simulation studies suggest that AIC_c performs well, therefore researchers can use AIC_c for model selection and model averaging in Logistic regression.

Imputation had a substantial effect on model selection and model averaging. As sample size increases, the tendency to choose true model M110 increases in both linear model and generalized models for any missing percentages. The model M100 was chosen more often compared to true model M110 as missing percentages increased in both linear model and Logistic regression, but this happened less often as sample size increased. The MSE(P) for selected best model and model averaging decreases as sample size increases in both linear model and Logistic regression. The MSE(P) for selected best model and model averaging increases as missing percentages increases in both linear model and Logistic regression. This shows that there are joint effects of sample size and missing percentages on model selection and model averaging in both linear model and Logistic regression.

Besides that, there are no effects of $|\rho_{23}|$ in the frequency of selecting true model M110 and also in the variation of MSE(P) values in both linear model and generalized models after imputation. The variation in the MSE(P) values as $|\rho_{23}|$ increases is just a sampling error. Negative and positive correlations of the same magnitude have the same effects on prediction for model averaging and model selection.

For larger error variance, BIC selects smaller models more often compared to AIC_c for any missing percentages. The MSE(P) for selected best model and model averaging increases as σ_ε increases for linear models. The increases in MSE(P) are not proportional to σ_ε^2. In addition, AIC_c performs better than BIC in linear models when variance is large. Schomaker and Heumann [2014] showed implementation of AIC based model selection and averaging with multiple imputation is very straightforward, and those estimators perform well. Our research showed that AIC_c performs better than BIC for larger error variance. It is advisable to use AIC_c rather than using BIC or AIC (as used by [Schomaker and Heumann, 2014]) in model selection and model averaging.

Moreover, MSE(P) for model averaging is lower than for model selection in both incomplete data sets and after imputation of missing values for linear models. In Logistic regression, model selection performs better than model averaging for small sample size. There are no clear difference between model selection and model averaging for larger sample size in Logistic regression. Therefore, model averaging seem to be a better general strategy than model selection if the researcher's aim is prediction. If the researcher is interested in which variables should be included in the model-building, then model selection is preferable. Model selection and model averaging can be combined where

model selection is used to identify the variables for prediction and then predictions are made using model averaging.

MSE(P) was lowest when non-overlapping variable sets was used with model averaging and model selection for larger error variance in linear model. Whereas inclusive strategy is better for lower error variance with model averaging and model selection for linear model in terms of prediction. The non-overlapping variable sets performs significantly better than restrictive and inclusive strategies with model averaging and also for model selection with small sample size in Logistic regression. This is in agreement with Hardt et al. [2012] who stated that inclusion of auxiliary variables can improve the imputation model. Schomaker and Heumann [2014] stated that use of an incorrect imputation model can cause improper imputation, a biased model and inappropriate post model selection and model averaging estimates. Negative and positive correlations of the same magnitude have the same effect on prediction for model averaging and model selection using all three model-building strategies. There is not much difference between the restrictive and inclusive strategies in terms of prediction for model averaging and model selection in linear models.

A similar simulation study was carried out using "norm.nob" imputation method for imputing missing data and also without response variable in the imputation models for both linear models and logistic regression. The effects of simulation parameters were similar for using both "norm" and "norm.nob" methods. Moreover, the MSE(P) using "norm" and with response variable in the imputation models is slightly lower than using "norm.nob" imputation method and without response variable in the imputation model. This shows that inclusion of response variable in the imputation models improves the prediction.

In conclusion, there are important effects of all the simulation parameters on model selection and averaging in both the linear model and Logistic regression. Researchers can use model selection to identify which variables to be included when making predictions or make predictions using model averaging. Since AIC_c performs better than BIC for larger error variance and in making predictions (and is known theoretically to be less biased than AIC for small samples), AIC_c should be used as the model selection criterion of choice. Either AIC_c or BIC could be used for model averaging. Moreover, imputing missing data using a correct imputation model is essential. Since the inclusion of auxiliary variables can improve the imputation model, researchers should auxiliary variables in imputation models whenever appropriate variables are available. If the interest of the research with missing data is to identify which variables to be included when making predictions and also for making prediction in Logistic regression, researchers should use model selection with non-overlapping variable sets (use the auxiliary variable only in the

imputation model). It is advisable to use model averaging with inclusive strategies (use the auxiliary variable in both the imputation and prediction models) to make predictions in the model-building process when there exist missing data in linear models for smaller error variance. In the extreme cases, researchers can use non-overlapping variables set for model selection and model averaging.

In this chapter, we were interested in comparing the effects of simulation parameters on imputation for model selection and model averaging. All three model-building strategies (non-overlapping variable sets, inclusive and restrictive strategies) were investigated for both model selection and model averaging. In the next simulation study, we will explore three different model selection methods and model averaging. The effects of multiple imputation and simulation parameters will be observed in both linear model and Logistic regression. A best model selection method will be chosen.

Chapter 5

The Implementation of Model Selection and Model Averaging using Multiple Imputation

The aim of this chapter is to compare multiple imputation (MI) with single imputation in terms of model selection and prediction. As discussed in Section 2.2.2, single imputation does not fully account for the uncertainty at the imputation step, so almost always underestimates the variance in estimation and prediction. MI can be used to overcome this problem by taking into account both within-imputation and between-imputation uncertainty. Therefore, in this chapter, model-building approaches and model-building strategies were explored and compared for the multiply-imputed data sets.

The basic simulation design used in Chapter 4 will be used in this chapter too. Three different model selection methods (RR, STACK, M-STACK) will be investigated for combining results from multiply-imputed data sets. Model averaging using non-overlapping variable sets will be explored and compared with the best model selection method in terms of prediction. In addition, the inclusive and restrictive strategies will be compared using the best model selection method and model averaging in order to identify which model-building strategy is most suitable for multiply-imputed data sets.

As we discussed and concluded in Chapter 4, AIC_c performs better than BIC in both linear model and Logistic regression in terms of model selection and prediction. Therefore, only AIC_c will be used as a model selection criterion and weights will be based only on AIC_c for model averaging in this chapter. The results using BIC are qualitatively similar to those presented for AIC_c in this chapter. Furthermore, as concluded in Chapter 4, the non-overlapping variable set performs better for model selection, therefore it will be used

initially to investigate and compare the performance of three model selection methods (RR, STACK, M-STACK) and model averaging using mean square error of prediction in both linear model and Logistic regression. The results for the restrictive and inclusive strategies will be presented briefly. The effects of imputation and simulation parameters will be discussed for both linear model and Logistic regression.

5.1 Model Selection and Model Averaging for Multiple Imputation

The simulation design described in Section 4.1 will be used to explore three approaches to model selection based on multiple imputation methods. The three approaches are backward stepwise regression using Rubin's rules (RR), the stacked imputed dataset method (STACK) of Wood et al. [2008] and a modified stacked imputed dataset method (M-STACK). As discussed by Wood et al. [2008], the RR method is considered as gold standard approach but it is more computationally demanding when repeated analyses are required. Therefore, Wood et al. [2008] proposed the STACK method as a sensible alternative to RR method for repeated analyses. The STACK method use backward stepwise selection approach for variable selection. The backward stepwise selection approach is often criticised and its disadvantages were discussed in Section 2.4.1. A modified version of the stacked imputed data sets method (M-STACK) is proposed as an alternative to STACK and RR.

5.1.1 Rubin's rules (RR)

The first method is backward stepwise regression using repeated use of Rubin's rules (RR). The simple backward stepwise regression using Rubin's rules for four models (M000, M100, M010, M110) was carried out as follows:

Step 1: Run model M110 for each imputation, store $\hat{\beta}$ and $\hat{cov}(\hat{\beta})$ calculated in the way described in Section 2.2.5 using Equation (2.6) and Equation (2.7).

Step 2: Check $\frac{|\bar{\beta}_1|}{e.s.e(\bar{\beta}_1)} > 1.96$ and $\frac{|\bar{\beta}_2|}{e.s.e(\bar{\beta}_2)} > 1.96$.

Step 3: If both parameters are significant, record count of 1 for fitting model M110 and calculate MSE(P) using $\bar{\boldsymbol{\beta}}$.

Step 4: If β_2 is not significant, run model M100 for each imputation. Store $\hat{\beta}$ and $\hat{cov}(\hat{\beta})$.

Step 5: Check $\frac{|\bar{\beta}_1|}{e.s.e(\bar{\beta}_1)} > 1.96$. If β_1 is significant, record count of 1 for fitting model M100 and calculate MSE(P) using $\bar{\boldsymbol{\beta}}$.

Step 6: If β_1 is not significant, run model M010 for each imputation. Store $\hat{\beta}$ and $\hat{cov}(\hat{\beta})$.

Step 7: Check $\frac{|\bar{\beta}_2|}{e.s.e(\bar{\beta}_2)} > 1.96$. If β_2 is significant, record count of 1 for fitting model M010 and calculate MSE(P) using $\bar{\boldsymbol{\beta}}$.

Step 8: If β_2 is not significant, run model M000 for each imputation. Store $\hat{\beta}$ and $\hat{cov}(\hat{\beta})$. Record count of 1 for fitting model M000 and calculate MSE(P) using $\bar{\boldsymbol{\beta}}$.

5.1.2 STACK

The second method uses the stacked imputed data sets with weighted regression (STACK) [Wood et al., 2008]. In this method, D imputed data sets will be stacked for the n individuals which yields one large dataset of length Dn. A fixed weight will be applied to all individuals to correct the standard errors. Although Wood et al. [2008] proposed three possible weights, but they claimed $W3$ was the best. Therefore, weight $W3$ will be used in this research. The considered weight $W3$ is

$$w_i = \frac{(1 - f_i)}{D} \tag{5.1}$$

where f_i is the fraction of missing data for variable X_i and it is calculated as

$$f_i = \frac{number\,of\,missing\,data\,for\,variable\,X_i}{n} \tag{5.2}$$

The largest f_i will be used across all the variables in the context of more variables with missing data in a model. Weighted regression analysis will be carried out using stacked imputed data.

The essential assumption of the STACK method is that fraction of missing data equals fraction of missing information. This assumption yields the weight W3 in MCAR mechanism. Wood et al. [2008] pointed out that the W3 give solutions comparable to RR in case of MCAR. This pattern of missing data favour the STACK method and also enables a comparison between RR, STACK and M-STACK methods. The simulation settings in this research follows MCAR mechanism, therefore this is favouring the assumption of STACK method [Wood et al., 2008]. In addition, the predictors in the prediction model are uncorrelated in the setting of non-overlapping variable sets.

In this research, the model selection is carried out on stacked data using model selection criteria (AIC_c and BIC) rather than the backward stepwise selection approach. Although the original version of STACK method proposed by Wood et al. [2008] is using backward stepwise selection approach for variable selection, this research is interested in using model selection criteria for model selection. All possible models are fitted to the single stacked dataset and a best model is selected using model selection criteria. Then, the selected best model will be fitted for each imputed dataset separately and the parameter estimates will be combined using RR as in Equation (2.6). The number of times each possible model is selected via each selection criterion was calculated. The MSE(P) was calculated for the combined parameter estimates using RR.

5.1.3 M-STACK

The third method is a modified version of the stacked imputed data sets method with weighted regression (M-STACK). All possible models are fitted to the single stacked dataset and a best model is selected using model selection criteria (same as STACK). In this method, however, the final estimates of the parameters are taken to be the ones given by the analysis on the stacked dataset; this avoids the final, potentially computationally-expensive, step of STACK that involves refitting the models in each imputed dataset. This approach is justified by Appendix A of Wood et al. [2008], where it is shown that this estimator has reasonable large-sample properties. The MSE(P) was calculated using the final estimates of the parameters of the stacked dataset.

5.1.4 Model Averaging for Multiple Imputation

The model averaging estimators weigh across all possible models after imputation with any imputation method. Final model averaging parameter estimates for linear regression were obtained in two steps. First, the method outlined in Section 2.5 was used in each imputed dataset to obtain averaged parameter estimates (using either AIC_c or BIC weights). Second, the parameter estimates from the D imputed datasets were combined using RR to give the final estimates. These parameter estimates were used to predict the response for each test value. For logistic regression, the same method was applied but the estimated probabilities for each test value were calculated at each stage, as in Equation (2.43). The MSE(P) was then obtained by comparing these estimated values with the true model values.

5.2 Design of Simulation and Results

In this section, we will discuss the results for Linear regression and Logistic regression based on the simulation design in the previous chapter. The simulation study was conducted with linear model (Task LM) and Logistic regression (Task GLM) for all three model selection methods and model averaging. The error terms for Task LM were simulated from a normal distribution, $\varepsilon \sim N\left(0, \sigma_\varepsilon^2\right)$ where $\sigma_\varepsilon = 0.25$, 1 and 4. The number of observations are $n = 50, 100, 200, 400$. The missing observations were created on variable X_2 with percentage of missing observations as $m = 0, 25, 50$. As discussed in Chapter 4, the negative and positive correlations of same magnitude showed similar results, therefore only positive correlation results will be discussed. The covariance matrix as in Equation (4.2) was used with $\rho_{23} = 0, 0.25, 0.5, 0.75$. Multiple imputation was carried out with $D = 10$. The analysis was carried out for every combination of $n, m, \sigma_\varepsilon^2$ and covariance matrix. The performance of the three model selection methods (RR, STACK, M-STACK) using non-overlapping variable sets were compared in both Linear regression and Logistic regression using mean square error of prediction.

5.2.1 Linear regression

A simulation study was conducted based on simulation design as discussed earlier for linear model (Task LM). The analysis was carried out for every combination of $n, m, \sigma_\varepsilon^2$ and covariance matrix. The performance of three model selection methods and model averaging were compared using mean square error of prediction and all three model-building strategies.

5.2.1.1 Rubin's Rules (RR) using non-overlapping variable sets for Linear regression

A simulation study was conducted for simple backward stepwise regression using RR using non-overlapping variable sets. As discussed in Chapter 4, there is not much difference in the results of model selection using $\sigma_\varepsilon = 0.25$ and $\sigma_\varepsilon = 1$. When $\sigma_\varepsilon = 0.25$, the true model M110 was chosen 100% for all combinations of ρ_{23}, sample sizes and missing percentages. Therefore, there is no discussion of model selection results when $\sigma_\varepsilon = 0.25$. However, there are effects of smaller error variance in prediction (as discussed in Chapter 4).

When $\sigma_\varepsilon = 1$, Table 5.1 shows the number of times all possible models are selected in each of 1000 simulations for all the combinations of ρ_{23} with $n = 50$. The true model

M110 was chosen 100% compared to other possible models in each of 1000 simulations for all the combinations of ρ_{23} and $\sigma_\varepsilon = 1$ without any missing data in variable X_2. The chances of choosing true model M110 decreases as missing percentages increases. However, when $\sigma_\varepsilon = 1$ with $n = 100$, $n = 200$ and $n = 400$, the true model M110 was chosen 100% compared to other possible models in each of the 1000 simulations for all combinations of ρ_{23} and for $m = 0, 25$ and 50.

Table 5.1: Number of times all possible models are selected in each of 1000 simulations for all the combinations of ρ_{23} when $n = 50$ and $\sigma_\varepsilon = 1$ using RR for linear regression

$n = 50$ and $\sigma_\varepsilon = 1$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$M000$	0	0	0	0	0	0	0	0	0	0	0	0
$M100$	0	2	7	0	0	5	0	0	2	0	0	0
$M010$	0	0	0	0	0	0	0	0	0	0	0	0
$M110$	1000	998	993	1000	1000	995	1000	1000	998	1000	1000	1000

Table 5.2a and Table 5.2b show the number of times all possible models are selected in each of 1000 simulations for all the combinations of ρ_{23}, $\sigma_\varepsilon = 4$ and sample sizes, $n = 50$ and $n = 100$ respectively. For both $n = 50$ and $n = 100$, the chances of choosing true model M110 decreases as missing percentages increases. For a small sample size and this larger error variance, model M100 was selected more frequently compared to the true model M110. There are no effects of ρ_{23} in the frequency of selecting true model M110.

Table 5.2: Number of times all possible models are selected in each of 1000 simulations for all the combinations of ρ_{23}, $\sigma_\varepsilon = 4$ and sample size ($n = 50$ and $n = 100$) using RR for linear regression

(a) Number of times all possible models are selected when $n = 50$

$n = 50$ and $\sigma_\varepsilon = 4$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$M000$	309	20	13	332	21	22	344	17	22	325	15	21
$M100$	233	593	594	258	583	582	235	575	578	249	615	578
$M010$	12	1	1	15	0	0	10	0	1	10	0	0
$M110$	446	386	392	395	396	396	411	408	399	416	370	401

(b) Number of times all possible models are selected when $n = 100$

$n = 100$ and $\sigma_\varepsilon = 4$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$M000$	104	2	1	91	2	2	86	7	3	92	1	2
$M100$	231	333	416	216	310	368	178	334	363	189	322	350
$M010$	1	0	0	5	0	0	4	0	0	6	0	0
$M110$	664	665	583	688	688	630	732	659	634	713	677	648

Table 5.3a and Table 5.3b show the number of times all possible models are selected in each of 1000 simulations for all the combinations of ρ_{23}, $\sigma_\varepsilon = 4$ and sample sizes, $n = 200$ and $n = 400$ respectively. For both $n = 200$ and $n = 400$, the chances of choosing true model M110 decreases as missing percentages increases. As sample size increases, the tendency to choose model M100 decreases. Whereas true model M110 was chosen almost 100% as sample size increases for all values ρ_{23}. For large sample size, the choice of selecting model M110 decrease as missing percentages increases.

Table 5.3: Number of times all possible models are selected in each of 1000 simulations for all the combinations of ρ_{23}, $\sigma_\varepsilon = 4$ and sample size ($n = 200$ and $n = 400$) using RR for linear regression

(a) Number of times all possible models are selected when $n = 200$

$n = 200$ and $\sigma_\varepsilon = 4$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$M000$	4	0	0	4	0	0	6	0	0	8	0	0
$M100$	63	73	133	76	72	136	57	87	128	55	55	94
$M010$	0	0	0	0	0	0	1	0	0	0	0	0
$M110$	933	927	867	920	928	864	936	913	872	937	945	906

(b) Number of times all possible models are selected when $n = 400$

$n = 400$ and $\sigma_\varepsilon = 4$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$M000$	0	0	0	0	0	0	0	0	0	0	0	0
$M100$	1	0	15	4	0	15	0	0	14	1	1	4
$M010$	0	0	0	0	0	0	0	0	0	0	0	0
$M110$	999	1000	985	996	1000	985	1000	1000	986	999	999	996

Table 5.4: MSE(P) for selected best model for all the combinations of ρ_{23}, missing percentages, sample size and error variances ($\sigma_\varepsilon = 1$ and $\sigma_\varepsilon = 4$) using RR for linear regression

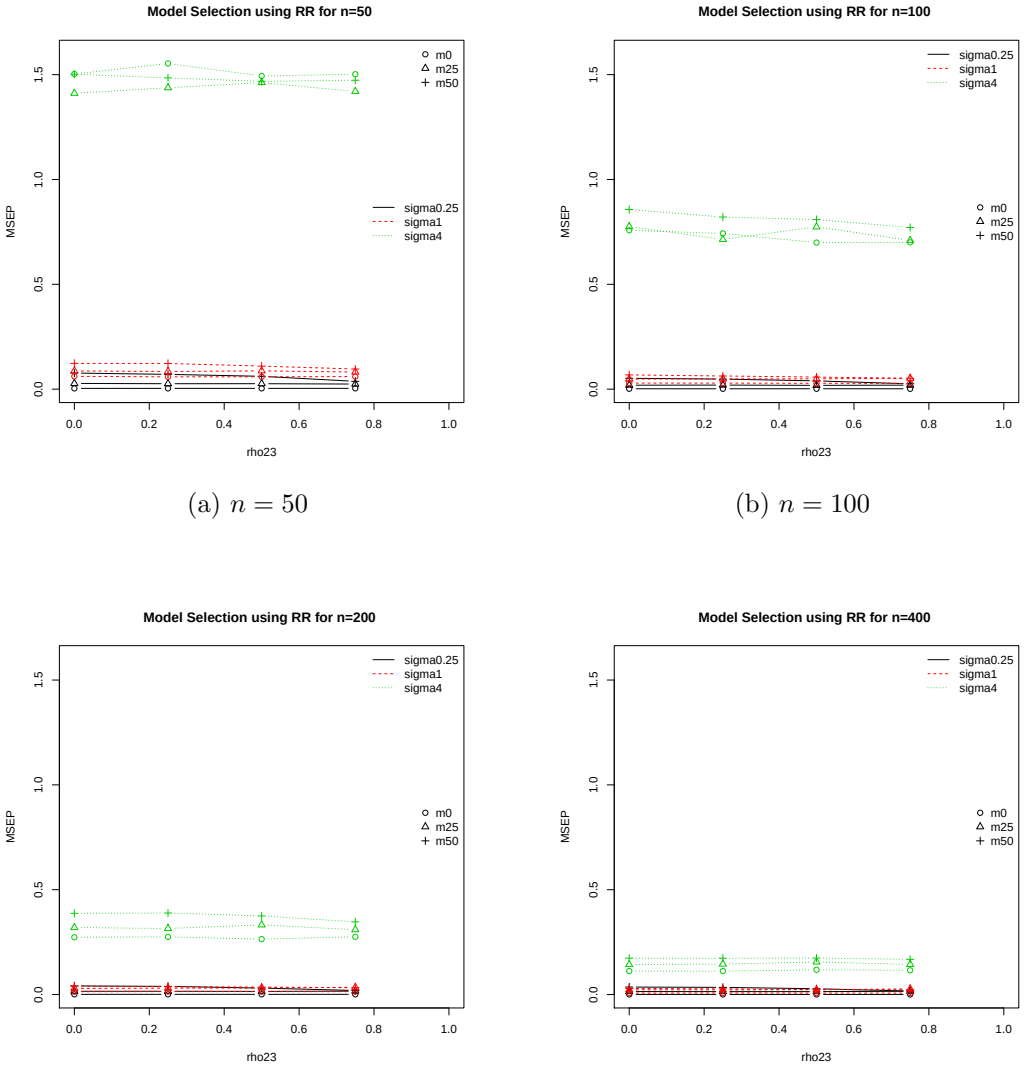
(a) MSE(P) for selected best model for $\sigma_\varepsilon = 1$

$\sigma_\varepsilon = 1$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$n = 50$	0.0612	0.0869	0.1229	0.0586	0.0844	0.1225	0.0589	0.0872	0.1098	0.0600	0.0815	0.0961
$n = 100$	0.0289	0.0480	0.0679	0.0289	0.0491	0.0628	0.0279	0.0490	0.0576	0.0275	0.0520	0.0519
$n = 200$	0.0134	0.0292	0.0409	0.0141	0.0304	0.0373	0.0140	0.0315	0.0356	0.0137	0.0346	0.0326
$n = 400$	0.0067	0.0226	0.0293	0.0073	0.0223	0.0272	0.0071	0.0232	0.0235	0.0070	0.0268	0.0227

(b) MSE(P) for selected best model for $\sigma_\varepsilon = 4$

$\sigma_\varepsilon = 4$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$n = 50$	1.5030	1.4111	1.5020	1.5531	1.4375	1.4844	1.4936	1.4622	1.4682	1.5021	1.4197	1.4733
$n = 100$	0.7577	0.7749	0.8568	0.7427	0.7139	0.8208	0.6990	0.7739	0.8088	0.7002	0.7095	0.7711
$n = 200$	0.2733	0.3194	0.3874	0.2750	0.3151	0.3888	0.2643	0.3322	0.3751	0.2754	0.3092	0.3467
$n = 400$	0.1119	0.1432	0.1735	0.1117	0.1452	0.1733	0.1180	0.1554	0.1742	0.1157	0.1438	0.1679

Table 5.4a and Table 5.4b show the MSE(P) for selected best model for all combinations of ρ_{23}, missing percentages, sample size and error variances $\sigma_\varepsilon = 1$ and $\sigma_\varepsilon = 4$ respectively. With $m = 0$, the MSE(P) decreases as sample size increases and the decrease is proportional to sample size. As σ_ε increases, MSE(P) increases and the increase is proportional to σ_ε^2 for $m = 0$. With $m = 0$, there is no imputation so ρ_{23} should make no difference. The decreases in MSE(P) values as ρ_{23} increases is just a sampling error. The MSE(P) values were increased as percentages of missingness increased for larger error variance. With $m = 25$ and $m = 50$, the MSE(P) decreases as sample size increases and the decrease is proportional to sample size. As σ_ε increases, MSE(P) increases but the increase is not proportional to σ_ε^2 for $m = 25$ and $m = 50$. There are no effects of ρ_{23} in terms of prediction.



(a) $n = 50$ (b) $n = 100$

(c) $n = 200$ (d) $n = 400$

Figure 5.1: MSE(P) for selected best model using RR for each ρ_{23}, σ_ε, missing percentages and sample sizes for linear regression

Figure 5.1a, Figure 5.1b, Figure 5.1c and Figure 5.1d show the MSE(P) for the selected best model using RR for each ρ_{23}, σ_ε, missing percentages and sample sizes, $n = 50$, $n = 100$, $n = 200$ and $n = 400$ respectively. As sample size increases, the MSE(P) for best model selected using RR decreases. For larger variance, MSE(P) for best model selected using RR decreases as sample size increases. The effects of error variance reduce as sample size increases. There are no effects of ρ_{23} in terms of prediction where the lines in the Figure 5.1 are stationary for all ρ_{23} values.

5.2.1.2 STACK using non-overlapping variable sets for Linear regression

A simulation study was conducted for stacked imputed data sets with weighted regression (STACK) using non-overlapping variable sets. When $\sigma_\varepsilon = 1$ with $n = 50$, 100, 200 and $n = 400$, the true model M110 was chosen 100% compared to other possible models in each of the 1000 simulations for all combinations of ρ_{23} and for $m = 0$, 25 and 50.

Table 5.5: Number of times all possible models are selected via AIC_c in each of 1000 simulations for all the combinations of ρ_{23}, $\sigma_\varepsilon = 4$ and sample size ($n = 50$ and $n = 100$) using STACK for linear regression

(a) Number of times all possible models are selected when $n = 50$

AIC_c $n = 50$ and $\sigma_\varepsilon = 4$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$M000$	139	14	8	140	12	11	141	14	12	153	11	17
$M100$	218	85	129	234	83	106	225	92	117	218	84	80
$M010$	245	67	80	222	82	81	249	86	85	242	81	85
$M110$	398	834	783	404	823	802	385	808	786	387	824	818

(b) Number of times all possible models are selected when $n = 100$

AIC_c $n = 100$ and $\sigma_\varepsilon = 4$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$M000$	17	0	3	24	1	2	22	0	3	22	0	2
$M100$	136	39	68	122	47	58	136	42	54	132	25	41
$M010$	109	20	22	119	28	27	118	21	25	135	17	17
$M110$	738	941	907	735	923	913	724	937	918	711	958	940

Table 5.5a and Table 5.5b show the number of times all possible models are selected via AIC_c in each of 1000 simulations for all the combinations of ρ_{23}, missing percentages, $\sigma_\varepsilon = 4$ and sample size, $n = 50$ and $n = 100$ respectively. For a small sample size and this larger error variance, the chance of choosing the true model M110 increased after imputation but it decreases as missing percentages increases. The chance of choosing the true model M110 increases as ρ_{23} increases and also after imputation for $n = 100$.

For smaller sample size, model M100 was selected more frequently compared to the true model M110. There are no effects of ρ_{23} in the frequency of selecting true model M110.

Table 5.6 shows number of times all possible models are selected via AIC_c in each of 1000 simulations for all the combinations of ρ_{23}, missing percentages and $\sigma_\varepsilon = 4$ for sample size $n = 200$. The choice of selecting true model M110 increases as missing percentages increases. The chance of selecting the true model M110 is much more better after imputation compared to without any missing data in variable X_2. For a larger variance and $n = 400$, AIC_c selects true model M110 almost 100% after imputation. As missing percentages and sample size increases, the chances of choosing the true model M110 increases. Imputation improves the choice of true model M110 as sample size increases.

Table 5.6: Number of times all possible models are selected via AIC_c in each of 1000 simulations for all the combinations of ρ_{23} when $n = 200$ and $\sigma_\varepsilon = 4$ using STACK for linear regression

AIC_c $n = 200$ and $\sigma_\varepsilon = 4$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$M000$	0	0	0	0	0	0	1	0	0	0	0	0
$M100$	9	3	23	18	3	21	22	5	11	18	3	11
$M010$	24	3	1	17	0	2	22	1	6	18	0	5
$M110$	967	994	976	965	997	977	955	994	983	964	997	984

Table 5.7: MSE(P) for selected best model via AIC_c for all the combinations of ρ_{23}, missing percentages, sample size and error variances ($\sigma_\varepsilon = 1$ and $\sigma_\varepsilon = 4$) using STACK for linear regression

(a) MSE(P) for selected best model for $\sigma_\varepsilon = 1$

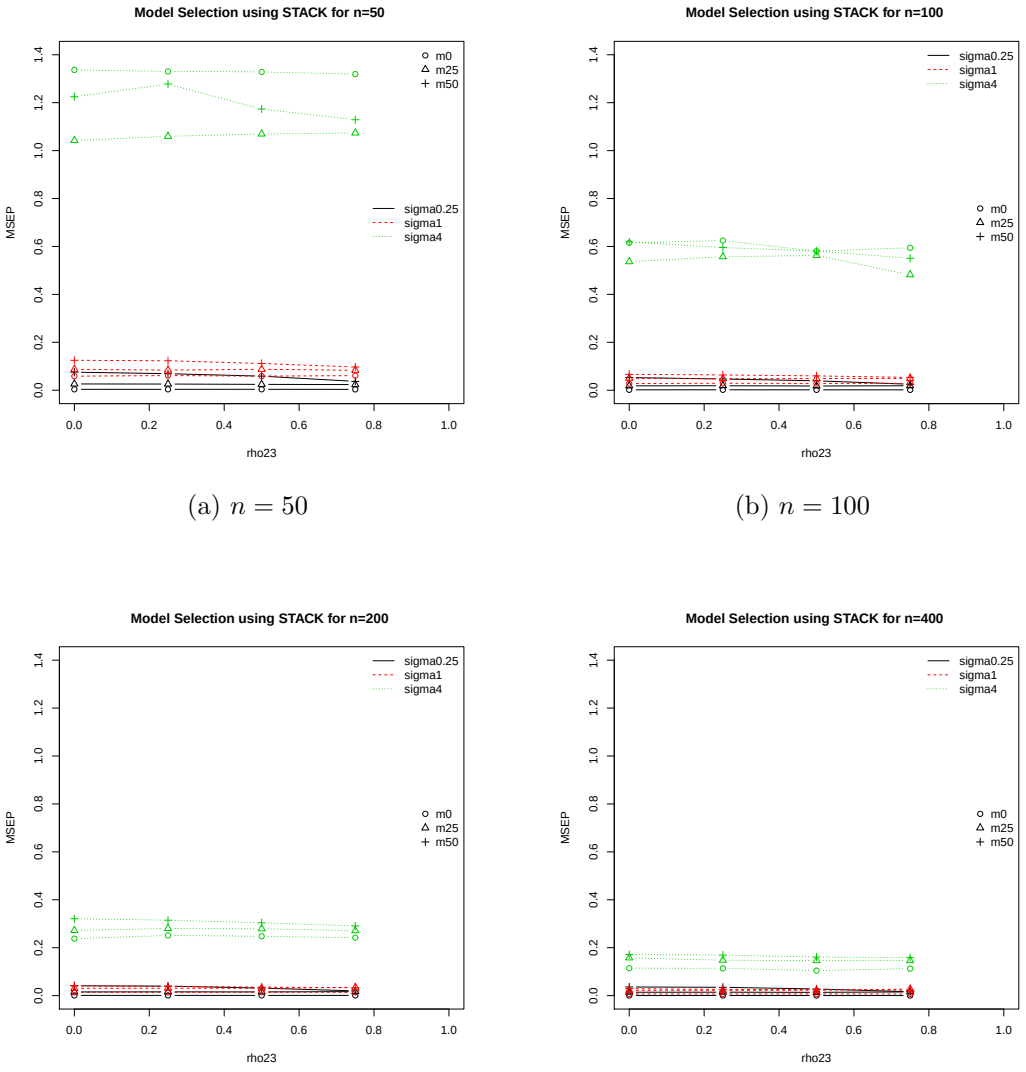
AIC_c $\sigma_\varepsilon = 1$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$n = 50$	0.0586	0.0874	0.1250	0.0612	0.0839	0.1229	0.0588	0.0875	0.1116	0.0609	0.0832	0.0971
$n = 100$	0.0278	0.0503	0.0665	0.0292	0.0481	0.0638	0.0284	0.0489	0.0599	0.0288	0.0495	0.0536
$n = 200$	0.0139	0.0307	0.0411	0.0138	0.0309	0.0391	0.0143	0.0310	0.0356	0.0139	0.0346	0.0330
$n = 400$	0.0071	0.0221	0.0290	0.0067	0.0223	0.0260	0.0073	0.0226	0.0249	0.0068	0.0274	0.0234

(b) MSE(P) for selected best model for $\sigma_\varepsilon = 4$

AIC_c $\sigma_\varepsilon = 4$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$n = 50$	1.3368	1.0426	1.2249	1.3306	1.0598	1.2777	1.3284	1.0694	1.1736	1.3192	1.0736	1.1288
$n = 100$	0.6152	0.5365	0.6188	0.6248	0.5572	0.5956	0.5802	0.5623	0.5809	0.5946	0.4820	0.5507
$n = 200$	0.2378	0.2721	0.3209	0.2512	0.2802	0.3147	0.2475	0.2792	0.3040	0.2418	0.2710	0.2909
$n = 400$	0.1148	0.1576	0.1712	0.1141	0.1478	0.1691	0.1043	0.1461	0.1610	0.1127	0.1464	0.1586

Table 5.7a shows the MSE(P) for the best model selected via AIC_c all the combinations of ρ_{23}, n and $\sigma_\varepsilon = 1$. Table 5.7b shows the MSE(P) for the best model selected via AIC_c

for all the combinations of ρ_{23}, n and $\sigma_\varepsilon = 4$. With $m = 0$, the MSE(P) decreases as sample size increases and the decrease is proportional to sample size. As σ_ε increases, MSE(P) increases and the increase is proportional to σ_ε^2 for $m = 0$. The MSE(P) values increases as percentages of missingness increases. The MSE(P) values were much higher after imputation. There are no effects of ρ_{23} in terms of prediction.



(a) $n = 50$

(b) $n = 100$

(c) $n = 200$

(d) $n = 400$

Figure 5.2: MSE(P) for best model selected via AIC_c using STACK and non-overlapping variable sets for each ρ_{23}, σ_ε, missing percentages and sample sizes for linear regression

Figure 5.2a, Figure 5.2b, Figure 5.2c and Figure 5.2d show the MSE(P) for best model selected via AIC_c using STACK for each ρ_{23}, σ_ε, missing percentages and sample sizes, $n = 50$, $n = 100$, $n = 200$ and $n = 400$ respectively. With $m = 0$, the MSE(P) decreases as sample size increases and the decrease is proportional to sample size. As σ_ε increases, MSE(P) increases and the increase is proportional to σ_ε^2 for $m = 0$. The MSE(P) values

increases as percentages of missingness increases. The MSE(P) values were much higher after imputation. As σ_ε increases, MSE(P) increases and the increase is proportional to σ_ε^2 for $m = 25$ and $m = 50$.

5.2.1.3 M-STACK using non-overlapping variable sets for Linear regression

A simulation study was conducted for modified version of stacked imputed data sets with weighted regression method (M-STACK) using non-overlapping variable sets. When $\sigma_\varepsilon = 1$ with $n = 50, 100, 200$ and 400, the true model M110 was chosen 100% compared to other possible models in each of the 1000 simulations for all combinations of ρ_{23} and for $m = 0, 25$ and 50.

Table 5.8: Number of times all possible models are selected via AIC_c in each of 1000 simulations for all the combinations of ρ_{23}, $\sigma_\varepsilon = 4$ and sample size ($n = 50$ and $n = 100$) using M-STACK for linear regression

(a) Number of times all possible models are selected when $n = 50$

AIC_c $n = 50$ and $\sigma_\varepsilon = 4$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$M000$	139	9	9	140	9	13	141	7	8	153	7	7
$M100$	218	104	1119	234	99	131	225	81	121	218	98	102
$M010$	245	78	93	222	79	77	249	73	73	242	84	93
$M110$	398	809	779	404	813	779	385	839	798	387	811	798

(b) Number of times all possible models are selected when $n = 100$

AIC_c $n = 100$ and $\sigma_\varepsilon = 4$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$M000$	17	0	1	24	2	1	22	1	2	22	0	1
$M100$	136	32	70	122	45	66	136	31	46	132	34	39
$M010$	109	25	16	119	17	17	118	18	24	135	22	23
$M110$	738	943	913	735	936	916	724	950	928	711	944	937

Table 5.8a and Table 5.8b show number of times all possible models are selected via AIC_c in each of 1000 simulations for all the combinations of ρ_{23}, missing percentages, $\sigma_\varepsilon = 4$ and sample size, $n = 50$ and $n = 100$ respectively. For a small sample size and this larger error variance, the chances of choosing true model M110 increases after imputation but it decreases as missing percentages increases. There are no effects of ρ_{23} in the frequency of selecting true model M110.

Table 5.9 shows the number of times all possible models are selected via AIC_c in each of 1000 simulations for all the combinations of ρ_{23}, missing percentages, $\sigma_\varepsilon = 4$ and sample size, $n = 200$. The choice of selecting true model M110 increases after imputation but it decreases as missing percentage increases. The chance of selecting the true model M110

is much more better after imputation compared to without any missing data in variable X_2. For a larger error variance and $n = 400$, AIC_c selects true model M110 almost 100% after imputation. Imputation improves the choice of true model M110 as sample size increases.

Table 5.9: Number of times all possible models are selected via AIC_c in each of 1000 simulations for all the combinations of ρ_{23} when $n = 200$ and $\sigma_\varepsilon = 4$ using M-STACK for linear regression

AIC_c $n = 200$ and $\sigma_\varepsilon = 4$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$M000$	0	0	0	0	0	0	1	0	0	0	0	0
$M100$	9	6	21	18	5	15	22	0	10	18	1	5
$M010$	24	1	3	17	2	0	22	4	0	18	1	2
$M110$	967	993	976	965	993	985	955	996	990	964	998	993

Table 5.10: MSE(P) for selected best model via AIC_c for all the combinations of ρ_{23}, missing percentages, sample size and error variances ($\sigma_\varepsilon = 1$ and $\sigma_\varepsilon = 4$) using M-STACK for linear regression

(a) MSE(P) for selected best model for $\sigma_\varepsilon = 1$

AIC_c $\sigma_\varepsilon = 1$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$n = 50$	0.0591	0.0887	0.1594	0.0599	0.0855	0.1567	0.0577	0.0803	0.1363	0.0624	0.0692	0.1052
$n = 100$	0.0281	0.0472	0.0987	0.0278	0.0457	0.0898	0.0283	0.0423	0.0755	0.0284	0.061	0.0570
$n = 200$	0.0141	0.0289	0.0639	0.0141	0.0276	0.0609	0.0142	0.0240	0.0522	0.0140	0.0193	0.0334
$n = 400$	0.0071	0.0195	0.0496	0.0070	0.0182	0.0462	0.0072	0.0158	0.0374	0.0070	0.0113	0.0223

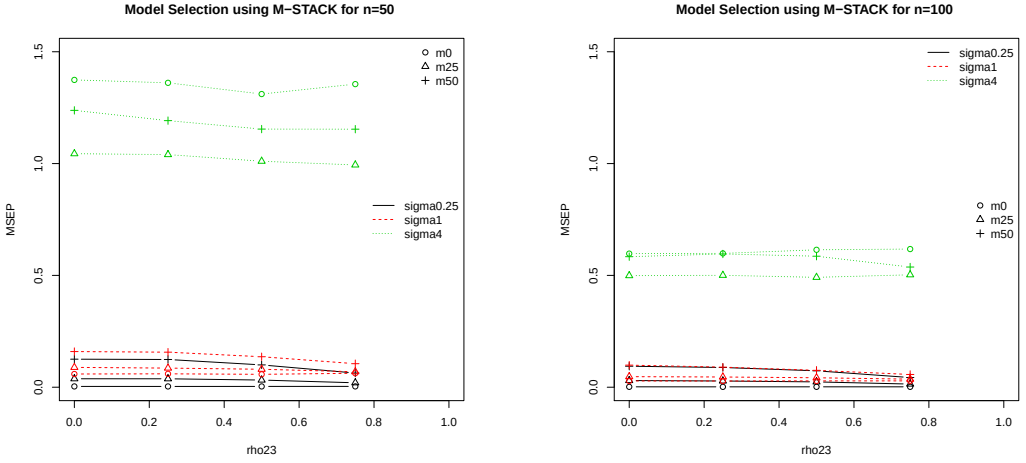
(b) MSE(P) for selected best model for $\sigma_\varepsilon = 4$

AIC_c $\sigma_\varepsilon = 4$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$n = 50$	1.3741	1.0441	1.2379	1.3611	1.0402	1.1920	1.3109	1.0105	1.1542	1.3553	0.9938	1.1538
$n = 100$	0.5974	0.4989	0.5837	0.5984	0.5004	0.5958	0.6145	0.4913	0.5858	0.6178	0.5027	0.5374
$n = 200$	0.2458	0.2458	0.2976	0.2432	0.2561	0.2973	0.2529	0.2326	0.2792	0.2366	0.2430	0.2704
$n = 400$	0.1128	0.1262	0.1370	0.1105	0.1260	0.1449	0.1118	0.1152	0.1414	0.1156	0.1145	0.1213

Table 5.10a shows the MSE(P) for the best model selected via AIC_c all the combinations of ρ_{23}, n and $\sigma_\varepsilon = 1$. Table 5.10b shows the MSE(P) for the best model selected via AIC_c all the combinations of ρ_{23}, n and $\sigma_\varepsilon = 4$. With $m = 0$, the MSE(P) decreases as sample size increases and the decrease is proportional to sample size. As σ_ε increases, MSE(P) increases and the increase is proportional to σ_ε^2 for $m = 0$. With $m = 0$, there is no imputation so ρ_{23} should make no difference. The decreases in MSE(P) values as ρ_{23} increases is just a sampling error. The MSE(P) values were increased as percentages of missingness increased. With $m = 25$ and $m = 50$, the MSE(P) decreases

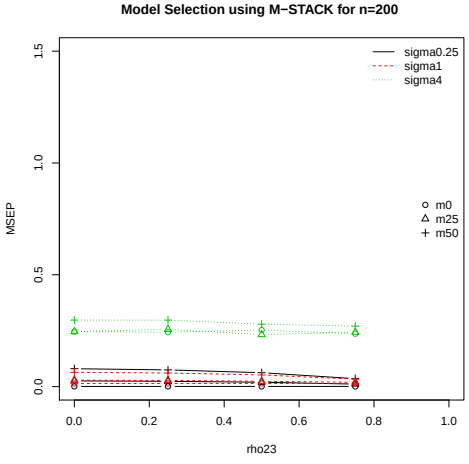
as sample size increases and the decrease is proportional to sample size. The MSE(P) values were higher after imputation. As σ_ε increases, MSE(P) increases and the increase is proportional to σ_ε^2 for $m = 25$ and $m = 50$.

Figure 5.3a, Figure 5.3b, Figure 5.3c and Figure 5.3d show the MSE(P) for best model selected via AIC_c using M-STACK for each ρ_{23}, σ_ε, missing percentages and sample sizes, $n = 50$, $n = 100$, $n = 200$ and $n = 400$ respectively. As sample size increases, the MSE(P) for best model selected using M-STACK decreases. For larger error variance, MSE(P) for best model selected using M-STACK decreases as sample size increases. The effects of error variance reduce as sample size increases. There are no effects of ρ_{23} in terms of prediction where the lines in the figures are stationary.

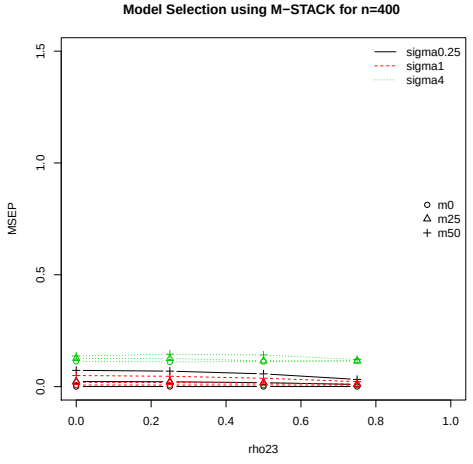


(a) $n = 50$

(b) $n = 100$



(c) $n = 200$



(d) $n = 400$

Figure 5.3: MSE(P) for best model selected via AIC_c using M-STACK for each ρ_{23}, σ_ε, missing percentages and sample for linear regression

Figure 5.4 shows the comparison between all three model selection methods (RR, M-STACK and STACK) for each ρ_{23}, σ_ε, missing percentages and $n = 100$. For larger error variance, MSE(P) for best model selected using M-STACK and STACK are lower than RR. Whereas the MSE(P) for best model selected using RR and STACK are lower than M-STACK for $\sigma_\varepsilon = 1$. STACK performs better than RR and M-STACK for all error variance, σ_ε and sample size in general. Therefore, STACK can be chosen as best model selection method.

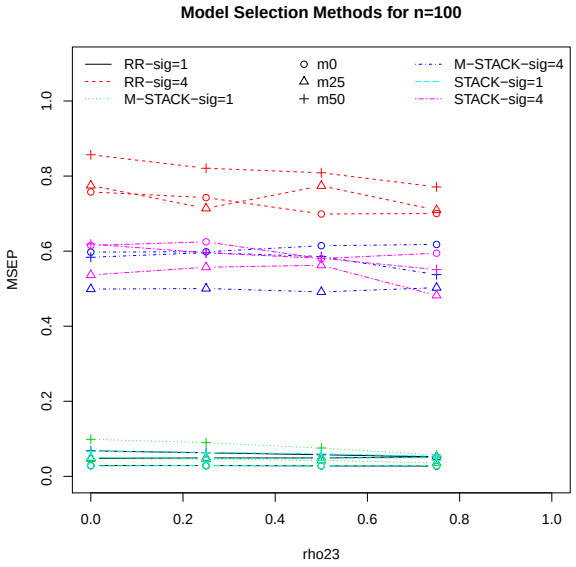


Figure 5.4: Comparison between model selection methods for each ρ_{23}, σ_ε, missing percentages and $n = 100$ for linear regression

5.2.1.4 Model averaging using non-overlapping variable sets for Linear regression

A simulation study was conducted based on simulation design as discussed earlier for linear regression using model averaging via AIC_c and BIC. The analysis was carried out for every combination of sample size, σ_ε, missing percentages and covariance matrix using non-overlapping variable sets. Table 5.11a shows the MSE(P) for model averaging via AIC_c for all the combinations of ρ_{23}, n and $\sigma_\varepsilon = 1$. Table 5.11b shows the MSE(P) for model averaging via AIC_c for all the combinations of ρ_{23}, n and $\sigma_\varepsilon = 4$. With $m = 0$ for model averaging via AIC_c, the MSE(P) decreases as sample size increases and the decrease is proportional to sample size. The MSE(P) values increases as percentages of missingness increases. With $m = 25$ and $m = 50$, the MSE(P) decreases as sample size increases and the decrease is proportional to sample size. As σ_ε increases, MSE(P) increases and the increase is proportional to σ_ε^2 for $m = 25$ and $m = 50$.

Table 5.11: MSE(P) for model averaging via AIC_c for all the combinations of ρ_{23}, missing percentages, sample size and error variances ($\sigma_\varepsilon = 1$ and $\sigma_\varepsilon = 4$) for linear regression

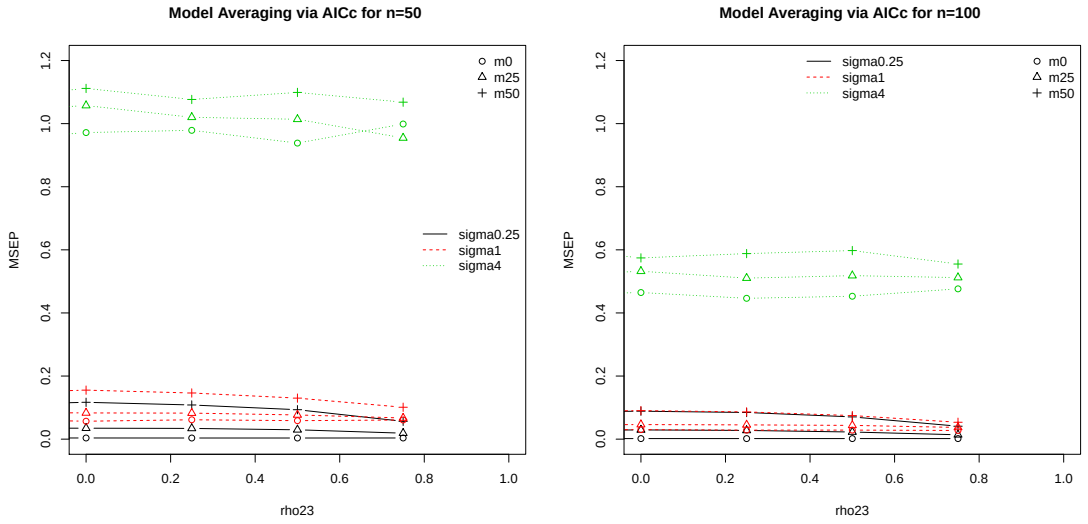
(a) MSE(P) for model averaging for $\sigma_\varepsilon = 1$

AIC_c $\sigma_\varepsilon = 1$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$n = 50$	0.0568	0.0831	0.1553	0.0614	0.0825	0.1463	0.0586	0.0767	0.1300	0.0607	0.0670	0.1011
$n = 100$	0.0302	0.0461	0.0904	0.0286	0.0452	0.0863	0.0287	0.0436	0.0747	0.0275	0.0371	0.0535
$n = 200$	0.0140	0.0291	0.0619	0.0138	0.0267	0.0584	0.0144	0.0247	0.0498	0.0140	0.0185	0.0322
$n = 400$	0.0068	0.0191	0.0500	0.0069	0.0186	0.0459	0.0072	0.0160	0.0383	0.0069	0.0112	0.0217

(b) MSE(P) for model averaging for $\sigma_\varepsilon = 4$

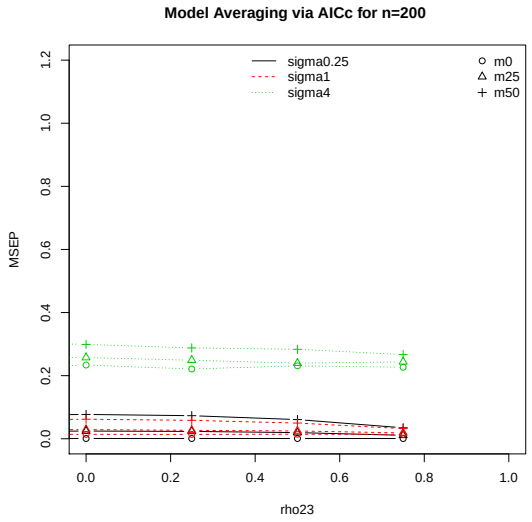
AIC_c $\sigma_\varepsilon = 4$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$n = 50$	0.9717	1.0573	1.1116	0.9788	1.0201	1.0766	0.9384	1.0138	1.0989	0.9988	0.9546	1.0680
$n = 100$	0.4647	0.5321	0.5744	0.4464	0.5104	0.5881	0.4530	0.5182	0.5977	0.4764	0.5121	0.5550
$n = 200$	0.2337	0.2573	0.2990	0.2211	0.2489	0.2879	0.2312	0.2397	0.2836	0.2266	0.2438	0.2668
$n = 400$	0.1124	0.1226	0.1471	0.1097	0.1214	0.1483	0.1103	0.1178	0.1345	0.1041	0.1173	0.1282

Figure 5.5a, Figure 5.5b, Figure 5.5c and Figure 5.5d show the MSE(P) for model averaging via AIC_c for each ρ_{23}, σ_ε, missing percentages and sample sizes, $n = 50$, $n = 100$, $n = 200$ and $n = 400$ respectively. As sample size increases, the MSE(P) for model averaging decreases. For larger error variance, MSE(P) for model averaging decreases as sample size increases. The effects of error variance reduce as sample size increases. There are no effects of ρ_{23} in terms of prediction where the lines in the figures are stationary.

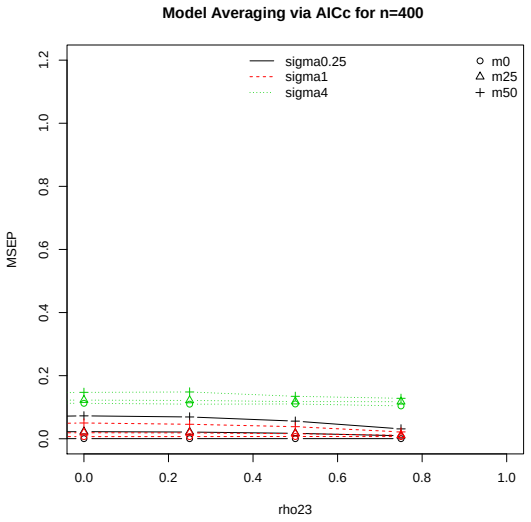


(a) $n = 50$

(b) $n = 100$



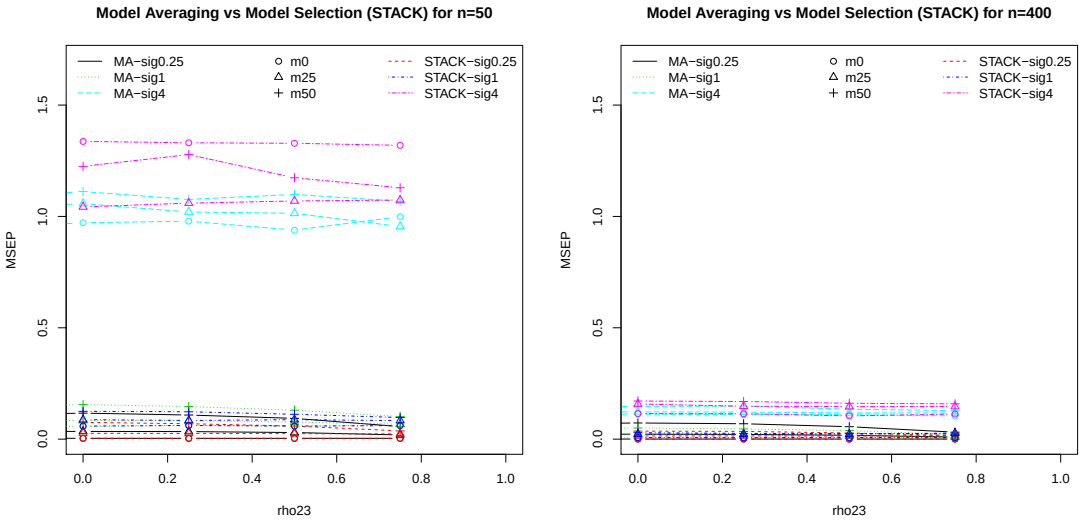
(c) $n = 200$



(d) $n = 400$

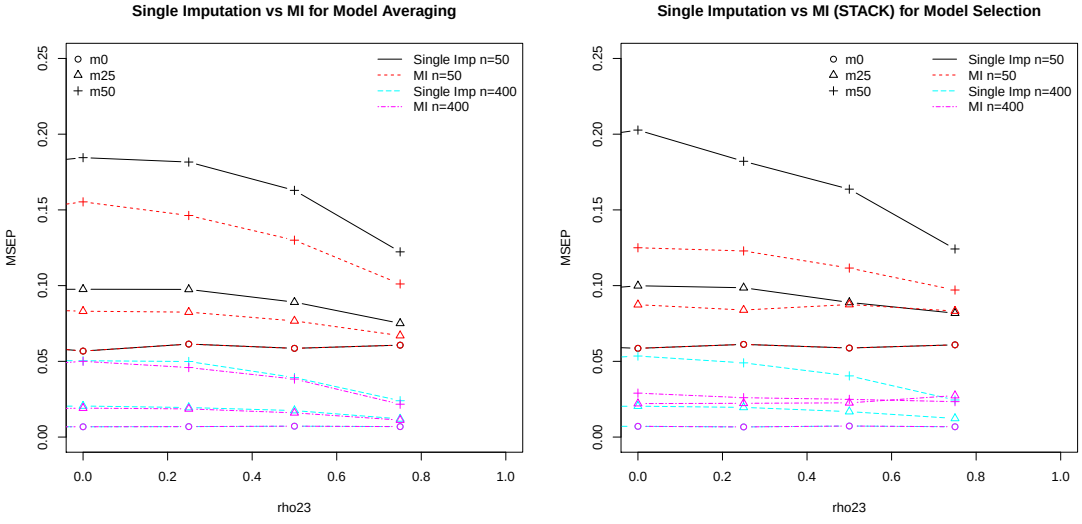
Figure 5.5: MSE(P) for model averaging via AIC_c for each ρ_{23}, σ_ε, missing percentages and sample sizes for linear regression

Figure 5.6a and Figure 5.6b show comparison between model averaging and model selection (STACK) via AIC_c for each ρ_{23}, σ_ε, missing percentages and sample sizes, $n = 50$ and $n = 400$ respectively. The results showed that for larger error variance and small sample size, model averaging is better than model selection using STACK. There are no difference between MSE(P) of model averaging and model selection using STACK for large sample size and smaller error variance.



(a) $n = 50$ (b) $n = 400$

Figure 5.6: Comparison between model averaging and model selection (STACK) via AIC_c for each ρ_{23}, σ_ε, missing percentages and sample sizes ($n = 50$ and $n = 400$) for linear regression



(a) Model Averaging (b) Model Selection

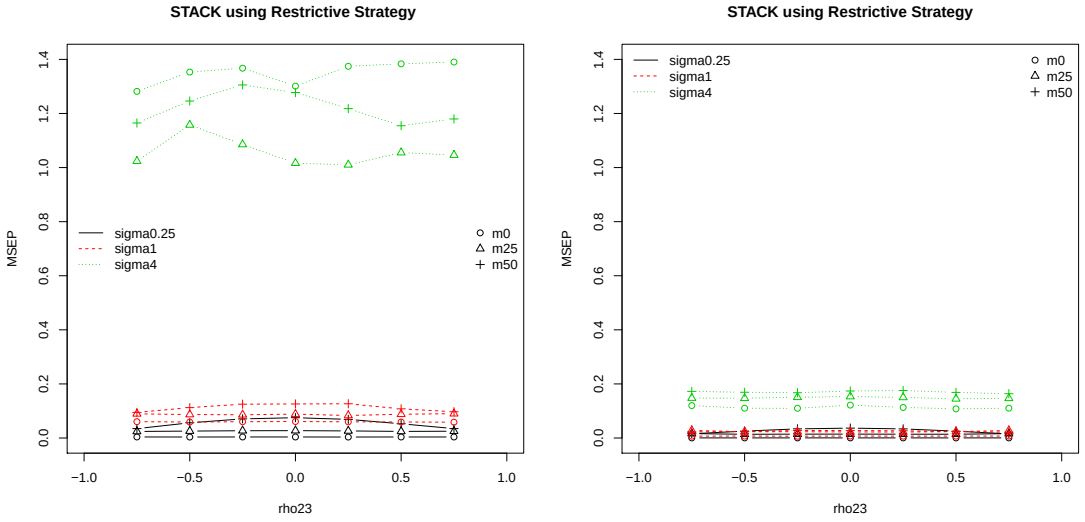
Figure 5.7: Comparison between single imputation and multiple imputation for model averaging and model selection via AIC_c for each ρ_{23}, $\sigma_\varepsilon = 1$, missing percentages and sample sizes for linear regression

Figure 5.7a shows comparison between model averaging using single imputation and multiple imputation for each ρ_{23}, $\sigma_\varepsilon = 1$, missing percentages and sample sizes ($n = 50$

and $n = 400$). It shows that MSE(P) of model averaging using multiple imputation is lower than MSE(P) of model averaging using single imputation for $\sigma_\varepsilon = 1$, missing percentages and all sample sizes. Figure 5.7b shows comparison between model selection using single imputation and multiple imputation (STACK) for each ρ_{23}, $\sigma_\varepsilon = 1$, missing percentages and sample sizes ($n = 50$ and $n = 400$). It shows that MSE(P) of model selection (STACK) using multiple imputation is lower than MSE(P) of model selection using single imputation for $\sigma_\varepsilon = 1$, missing percentages and all sample sizes.

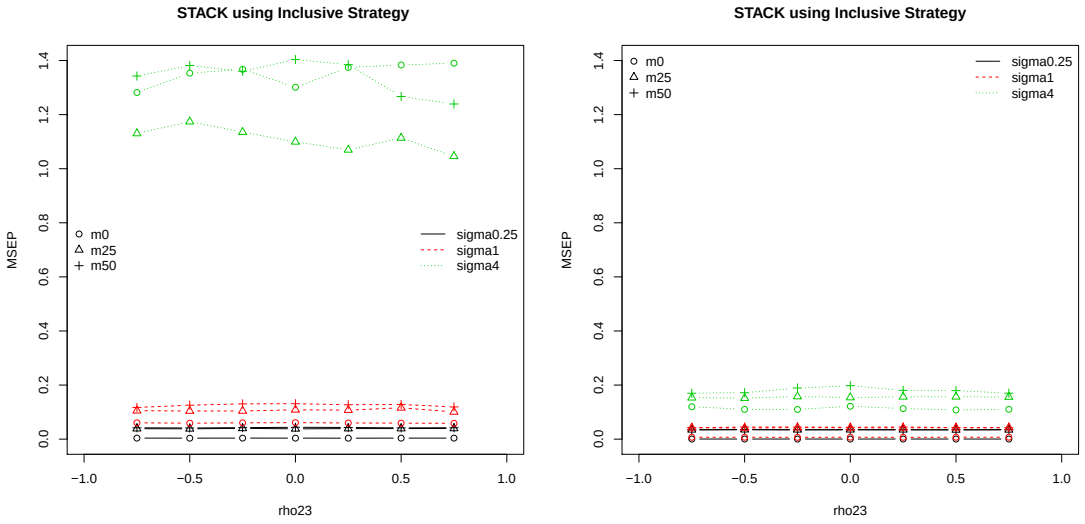
5.2.1.5 Model selection (STACK) and model averaging using restrictive and inclusive strategies for Linear regression

Figure 5.8 and Figure 5.9 show the MSE(P) for best model selected (STACK) via AIC_c using the restrictive and inclusive strategies for each ρ_{23}, σ_ε, missing percentages and sample sizes, $n = 50$ and $n = 400$ respectively. There is no effect of ρ_{23} values on model selected using STACK for the restrictive and inclusive strategies for all σ_ε. The MSE(P) for model selected (STACK) using the restrictive and inclusive strategies for the negative and positive correlations of same magnitude showed similar results.



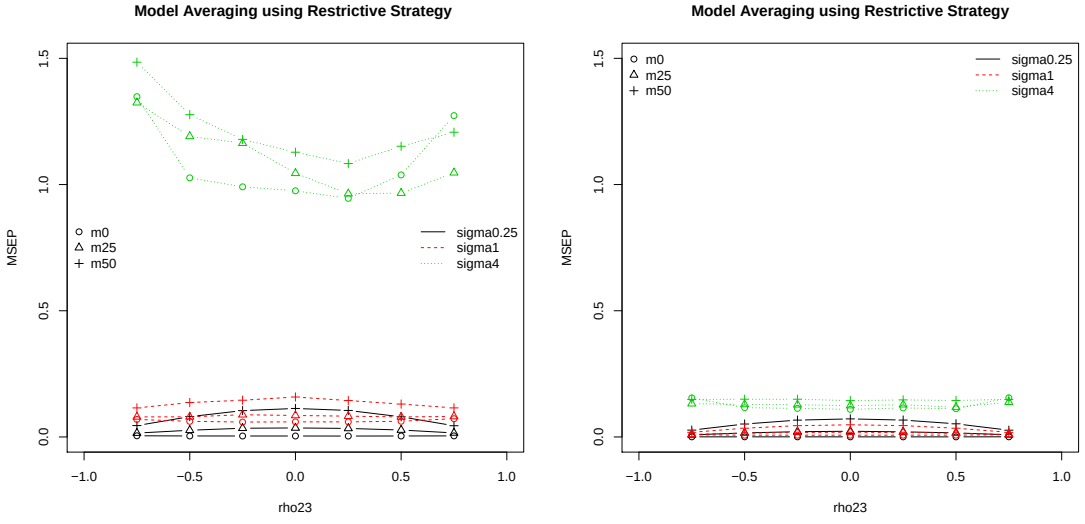
(a) $n = 50$ (b) $n = 400$

Figure 5.8: MSE(P) for best model selected via AIC_c using STACK and the restrictive strategy for each ρ_{23}, σ_ε, missing percentages and sample sizes ($n = 50$ and $n = 400$) for linear regression



(a) $n = 50$ (b) $n = 400$

Figure 5.9: MSE(P) for best model selected via AIC_c using STACK and the inclusive strategy for each ρ_{23}, σ_ε, missing percentages and sample sizes ($n = 50$ and $n = 400$) for linear regression

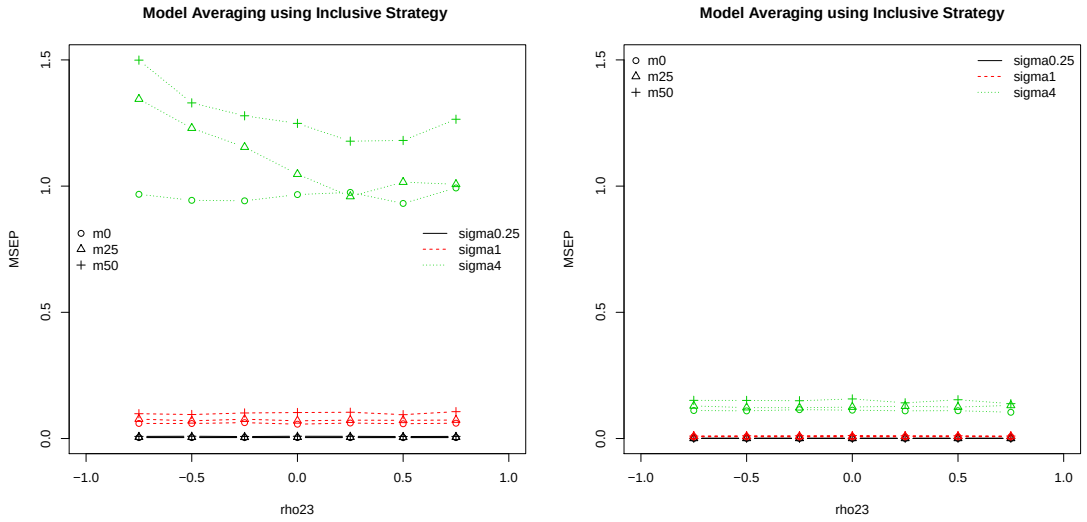


(a) $n = 50$ (b) $n = 400$

Figure 5.10: MSE(P) for model averaging via AIC_c using the restrictive strategy for each ρ_{23}, σ_ε, missing percentages and sample sizes ($n = 50$ and $n = 400$) for linear regression

Figure 5.10 and Figure 5.11 show the MSE(P) for model averaging using the restrictive and inclusive strategies for each ρ_{23}, σ_ε, missing percentages and sample sizes, $n = 50$ and $n = 400$ respectively. There is no effect of ρ_{23} values on model averaging for the restrictive and inclusive strategies with $\sigma_\varepsilon = 0.25$ and $\sigma_\varepsilon = 1$. The MSE(P) for model

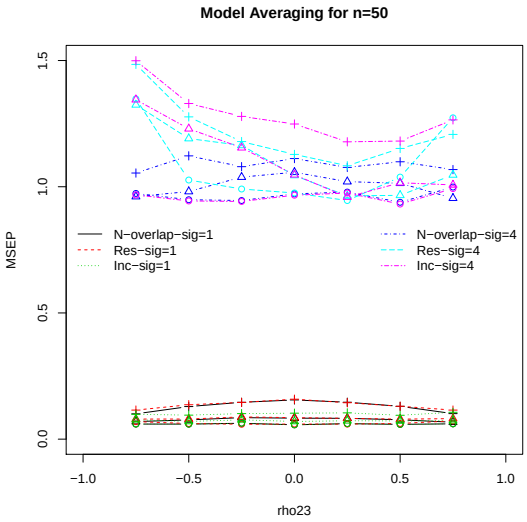
averaging using the restrictive and inclusive strategies for the negative and positive correlations of same magnitude showed similar results. For $\sigma_\varepsilon = 4$ and small sample size, there is an effect of negative ρ_{23} values on model averaging for the restrictive and inclusive strategies. The MSE(P) for model averaging using the restrictive and inclusive strategies increases as the negative ρ_{23} increases.



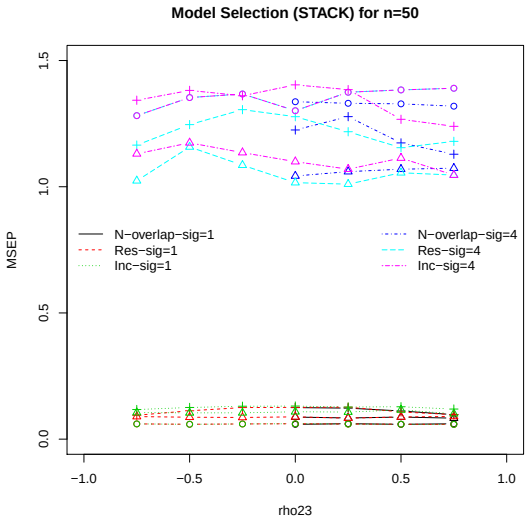
(a) $n = 50$ (b) $n = 400$

Figure 5.11: MSE(P) for model averaging via AIC_c using the inclusive strategy for each ρ_{23}, σ_ε, missing percentages and sample sizes ($n = 50$ and $n = 400$) for linear regression

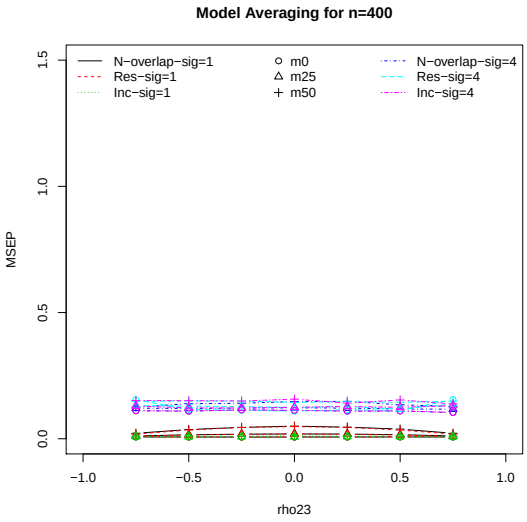
Figure 5.12 shows the comparison between all three model-building strategies (non-overlapping variable set, restrictive and inclusive strategies) for model averaging and model selection (STACK) via AIC_c for multiply-imputed data sets for each ρ_{23}, σ_ε, missing percentages and sample sizes, $n = 50$ and $n = 400$. For $\sigma_\varepsilon = 1$ and all sample sizes, there is no difference between the model-building strategies for both model selection (STACK) and model averaging. Whereas for $\sigma_\varepsilon = 4$ and large sample size, there is no difference between the MSE(P) for model averaging and model selection (STACK) using all three model-building strategies. There is no effect of the negative and positive correlations of same magnitude for model averaging and model selection (STACK) with all three model-building strategies. The MSE(P) for model averaging using all three model-building strategies increases as negative ρ_{23} increases for small sample size and $\sigma_\varepsilon = 4$.



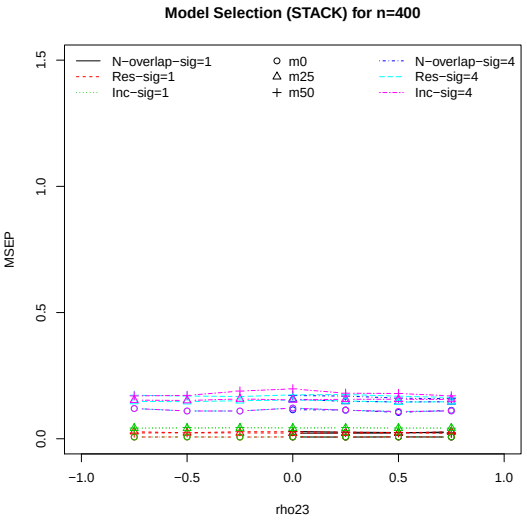
(a) Model Averaging $n = 50$



(b) Model Selection (STACK) $n = 50$



(c) Model Averaging $n = 400$

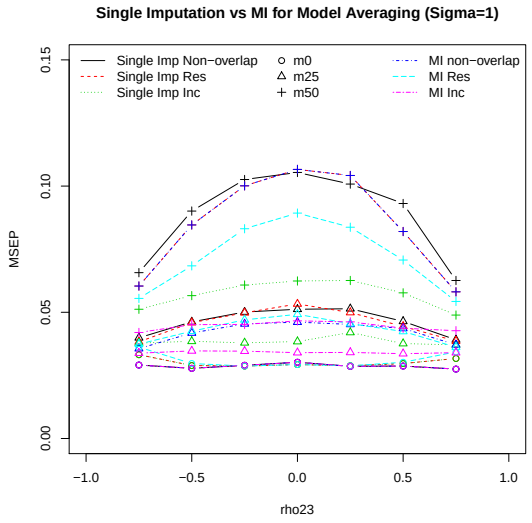


(d) Model Selection (STACK) $n = 400$

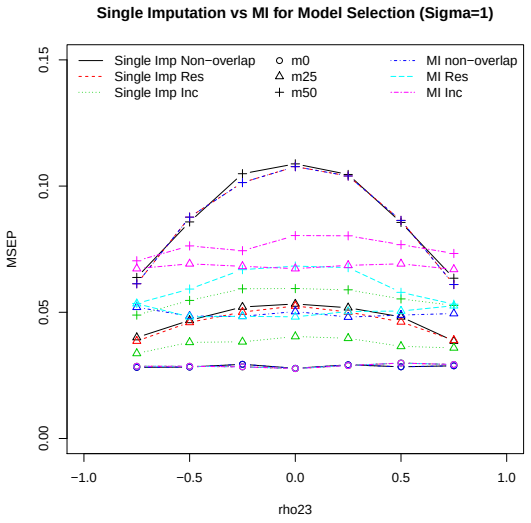
Figure 5.12: Comparison between all three model-building strategies for model averaging and model selection (STACK) for multiply-imputed data sets for linear regression

Figure 5.13 shows comparison between single imputation and multiple imputation for model averaging and model selection (STACK) using all three model-building strategies (non-overlapping variable sets, restrictive and inclusive strategies) for each ρ_{23}, missing percentages, $n = 100$ and error variances, $\sigma_\varepsilon = 1$ and $\sigma_\varepsilon = 4$ respectively. The results show that the MSE(P) of model averaging using the restrictive and inclusive strategies for multiply-imputed data sets are lower than MSE(P) of model averaging using all three model-building strategies strategies for single imputation, for all error variance and missing percentages. Moreover, MSE(P) of model averaging using the restrictive and

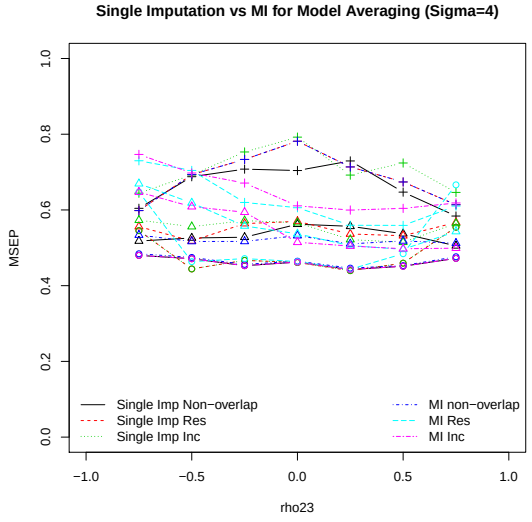
inclusive strategies are lower than the MSE(P) of model averaging using non-overlapping variable sets. Whereas, the MSE(P) of model selection using the restrictive and inclusive strategies for multiply-imputed data sets are lower than MSE(P) of model selection for single imputation using all three model-building strategies for all error variances and missing percentages. Moreover, the MSE(P) of model selection using the restrictive and inclusive strategies are lower than MSE(P) of model selection using non-overlapping variable sets for large error variance.



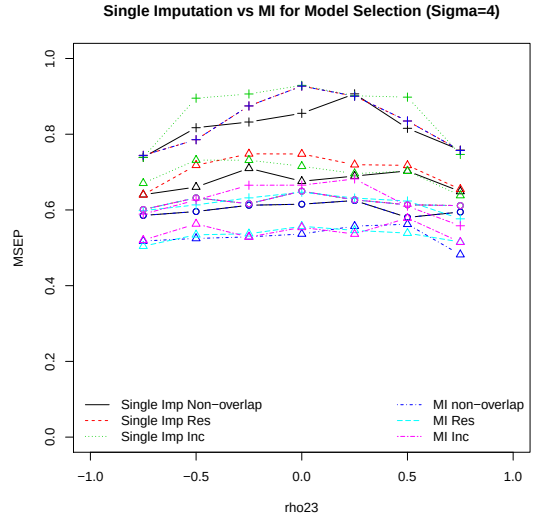
(a) Model Averaging $\sigma_\varepsilon = 1$



(b) Model Selection $\sigma_\varepsilon = 1$



(c) Model Averaging $\sigma_\varepsilon = 4$



(d) Model Selection $\sigma_\varepsilon = 4$

Figure 5.13: Comparison between single imputation and multiple imputation for model averaging and model selection for each ρ_{23}, missing percentages, $n = 100$ and error variances, $\sigma_\varepsilon = 1$ and $\sigma_\varepsilon = 4$ for linear regression

5.2.2 Logistic regression

A simulation study was conducted based on simulation design as discussed earlier for logistic regression (Task LG). The analysis was carried out for every combination of sample size, missing percentages and covariance matrix. The performance of three model selection methods and model averaging were compared using mean square error of prediction and all three model-building strategies.

5.2.2.1 Rubin's Rules (RR) using non-overlapping variable sets for Logistic regression

A simulation study was conducted for logistic regression using simple backward stepwise regression using Rubin's rule (RR) using non-overlapping variable sets. Table 5.12 shows the number of times all possible models are selected in each of 1000 simulations for all the combinations ρ_{23} and m with $n = 50$. The chances of choosing the true model M110 decreases as missing percentages increases whereas the chances of choosing model M100 increases.

Table 5.12: Number of times all possible models are selected in each of 1000 simulations for all the combinations of ρ_{23} when $n = 50$ using RR for logistic regression

$n = 50$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$M000$	64	0	7	59	2	4	72	0	6	62	2	1
$M100$	188	270	379	190	275	367	187	287	357	199	286	300
$M010$	7	0	3	1	0	18	5	0	4	2	3	0
$M110$	741	730	611	750	723	611	736	713	633	737	709	679

Table 5.13: Number of times all possible models are selected in each of 1000 simulations for all the combinations of ρ_{23} when $n = 100$

$n = 100$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$M000$	3	0	0	2	0	0	0	0	0	0	0	0
$M100$	31	47	111	27	43	101	24	34	88	31	30	52
$M010$	0	0	0	0	0	0	0	0	0	0	0	0
$M110$	966	953	889	971	957	899	976	966	912	969	970	948

Table 5.13 shows the number of times all possible models are selected in each of 1000 simulations for all the combinations ρ_{23} and missing percentages with $n = 100$. The chances of choosing the true model M110 decreases as missing percentages increases. As missing percentages increases, the chances of choosing model M100 increases.

Table 5.14: Number of times all possible models are selected in each of 1000 simulations for all the combinations of ρ_{23} when $n = 200$ using RR for logistic regression

$n = 200$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$M000$	0	0	0	0	0	0	0	0	0	0	0	0
$M100$	0	0	5	0	0	5	0	0	2	0	0	0
$M010$	0	0	0	0	0	0	0	0	0	0	0	0
$M110$	1000	1000	995	1000	1000	995	1000	1000	998	1000	1000	1000

Table 5.14 shows the number of times all possible models are selected in each of 1000 simulations for all the combinations ρ_{23} and missing percentages with $n = 200$. With $m = 0$, the true model M110 was selected 100% for all ρ_{23} values. The chances of choosing the true model M110 decreases as missing percentages increases whereas chances of choosing model M100 increases. For $n = 400$, the true model M110 was chosen 100% compared to other possible models in each of the 1000 simulations for all combinations of ρ_{23} and for $m = 0, 25$ and 50. There are no effects of ρ_{23} in the frequency of selecting true model M110 for all sample sizes.

Table 5.15: MSE(P) for best model selected for all the combinations of ρ_{23}, missing percentages and sample size using RR for logistic regression

Mean Square Error of Prediction												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$n = 50$	0.0187	0.0193	0.0256	0.0182	0.0189	0.0246	0.0198	0.0194	0.0239	0.0187	0.0190	0.0214
$n = 100$	0.0062	0.0078	0.0109	0.0060	0.0073	0.0110	0.0059	0.0072	0.0102	0.0062	0.0068	0.0085
$n = 200$	0.0027	0.0032	0.0051	0.0027	0.0034	0.0046	0.0027	0.0033	0.0044	0.0027	0.0029	0.0036
$n = 400$	0.0013	0.0017	0.0025	0.0013	0.0017	0.0024	0.0014	0.0016	0.0023	0.0014	0.0015	0.0019

Table 5.15 shows MSE(P) for best model selected for all the combinations of ρ_{23}, m and sample size. The MSE(P) decreases as sample size and ρ_{23} increases. As missing percentages increases, the MSE(P) values increases. For larger ρ_{23} values, there is no difference in MSE(P) values between $m = 0$, $m = 25$ and $m = 50$.

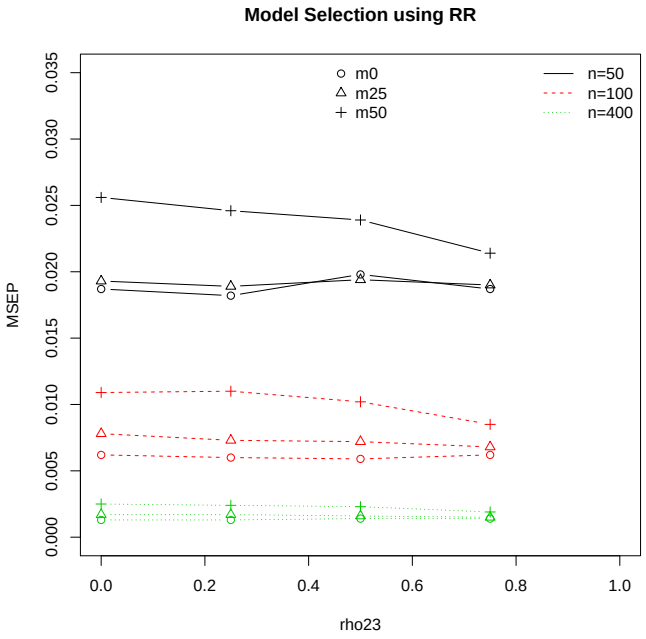


Figure 5.14: MSE(P) for best model selected using RR for each ρ_{23}, missing percentages and sample sizes for logistic regression

Figure 5.14 shows the MSE(P) for best model selected using RR for each ρ_{23}, missing percentages and sample sizes. As sample size increases, the MSE(P) for best model selected using RR decreases. The effects of missing percentages on MSE(P) for best model selected using RR reduces as sample size increases. There are no effects of ρ_{23} in term of prediction for all sample sizes using RR.

5.2.2.2 STACK using non-overlapping variable sets for Logistic regression

A simulation study was conducted for stacked imputed data sets with weighted logistic regression method (STACK) using non-overlapping variable sets. Table 5.16 shows the number of times all possible models are selected via AIC_c in each of 1000 simulations for all the combinations ρ_{23} and missing percentage with $n = 50$. For $m = 0$, the true model M110 was chosen above 80% for all ρ_{23} values and it increases as ρ_{23} increases. After imputation with $m = 25$, true model M110 was selected more often compared to without missing percentage but it decreases as missing percentages increases. After imputation, the true model M110 and model M010 are selected more often compared to without missing data, where all the model selected in different counts for $m = 0$. The chances of selecting model M010 via AIC_c increases as missing percentages increases.

Table 5.16: Number of times all possible models are selected via AIC_c in each of 1000 simulations for all the combinations of ρ_{23} and missing percentages when $n = 50$ using STACK for logistic regression

AIC_c $n = 50$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$M000$	5	0	0	6	0	0	10	0	0	9	0	0
$M100$	76	0	0	79	0	0	100	0	0	81	0	0
$M010$	72	169	266	81	143	274	79	155	274	76	115	244
$M110$	847	831	734	834	857	726	811	845	726	834	885	756

Table 5.17: Number of times all possible models are selected via AIC_c in each of 1000 simulations for all the combinations of ρ_{23} and missing percentages when $n = 100$ using STACK for logistic regression

AIC_c $n = 100$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$M000$	0	0	0	0	0	0	0	0	0	0	0	0
$M100$	3	0	0	8	0	0	3	0	0	5	0	0
$M010$	4	13	46	9	8	34	4	9	43	6	12	34
$M110$	993	987	954	983	992	966	993	991	957	989	988	966

Table 5.17 shows the number of times all possible models are selected via AIC_c in each of 1000 simulations for all the combinations ρ_{23} and missing percentages with $n = 100$. The number of times true model M110 selected reduces as missing percentages increases and it increases as value of ρ_{23} increases. The chance of selecting the true model M110 was above 95% for all ρ_{23} values and missing percentages. The chances of selecting model M010 increases as missing percentages increases. For $n = 200$ and $n = 400$, the true model M110 was chosen 100% via AIC_c compared to other possible models in each of the 1000 simulations for all combinations of ρ_{23} and for $m = 0$, $m = 25$ and $m = 50$. There are no effects of ρ_{23} in the frequency of selecting true model M110 for all sample sizes.

Table 5.18 shows the MSE(P) for the best model selected via AIC_c all the combinations of ρ_{23}, sample size and missing percentages respectively. The MSE(P) decreases as sample size increases. As missing percentages increases, the MSE(P) values increases. There are some difference on MSE(P) values as n and missing percentages increases.

Table 5.18: MSE(P) for best model selected via AIC_c for all the combinations of ρ_{23}, missing percentages and sample sizes using STACK for logistic regression

AIC_c												
ρ_{23}	$\rho_{23}=0$			$\rho_{23}=0.25$			$\rho_{23}=0.5$			$\rho_{23}=0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$n=50$	0.0141	0.0145	0.0207	0.0138	0.0147	0.0227	0.0155	0.0148	0.0202	0.0145	0.0131	0.0180
$n=100$	0.0055	0.0063	0.0077	0.0057	0.0059	0.0076	0.0054	0.0060	0.0074	0.0056	0.0058	0.0068
$n=200$	0.0027	0.0027	0.0036	0.0027	0.0029	0.0036	0.0027	0.0028	0.0035	0.0026	0.0029	0.0031
$n=400$	0.0013	0.0016	0.0018	0.0013	0.0015	0.0019	0.0013	0.0016	0.0017	0.0013	0.0015	0.0016

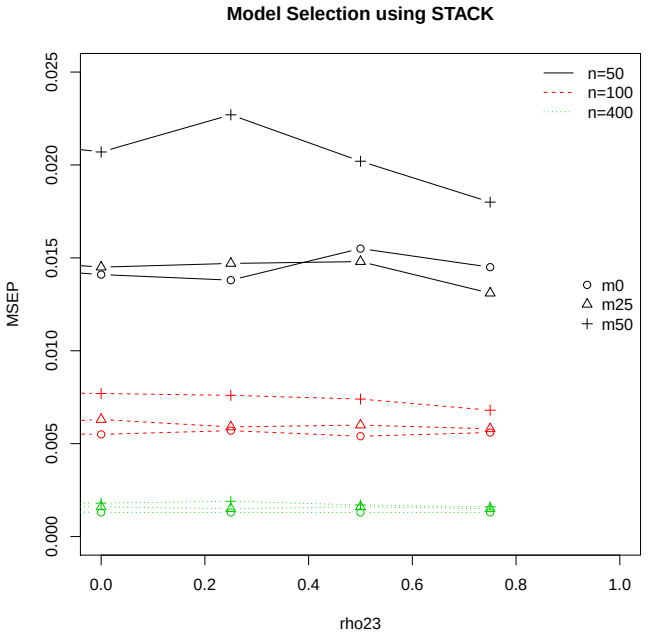


Figure 5.15: MSE(P) for for best model selected (STACK) and non-overlapping variable sets for each ρ_{23}, missing percentages and sample sizes for logistic regression

Figure 5.15 shows the MSE(P) for best model selected (STACK) using non-overlapping variable sets for each ρ_{23}, missing percentages and sample sizes. As sample size increases, the MSE(P) for best model selected using STACK decreases. The effects of missing percentages on MSE(P) for best model selected using STACK reduces as sample size increases. There are no effects of ρ_{23} in term of prediction for all sample sizes using STACK.

5.2.2.3 M-STACK using non-overlapping variable sets for Logistic regression

A simulation study was conducted for logistic regression using modified version of stacked imputed data sets with weighted logistic regression (M-STACK) using non-overlapping variable sets. Table 5.19 shows the number of times all possible models are selected via AIC_c in each of 1000 simulations for all the combinations ρ_{23} and missing percentages with $n = 50$. For $m = 0$, the true model M110 was chosen above 80% for all ρ_{23} values. After imputation with $m = 25$, true model M110 was selected more often compared to without missing percentages but it decreases as missing percentage increases. After imputation, the true model M110 and model M010 are selected more often compared to without missing data, where all the model selected in different counts for $m = 0$.

Table 5.19: Number of times all possible models are selected via AIC_c in each of 1000 simulations for all the combinations of ρ_{23} and missing percentage when $n = 50$ using M-STACK for logistic regression

AIC_c $n = 50$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$M000$	5	0	0	6	0	0	10	0	0	9	0	0
$M100$	76	0	0	79	0	0	100	0	0	81	0	0
$M010$	72	140	301	81	140	286	79	142	276	76	139	251
$M110$	847	860	699	834	860	714	811	858	724	834	861	749

Table 5.20: Number of times all possible models are selected via AIC_c in each of 1000 simulations for all the combinations of ρ_{23} and missing percentage when $n = 100$ using M-STACK for logistic regression

AIC_c $n = 100$												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$M000$	0	0	0	0	0	0	0	0	0	0	0	0
$M100$	3	0	0	8	0	0	3	0	0	5	0	0
$M010$	4	12	46	9	11	45	4	72	42	6	9	38
$M110$	993	988	954	983	989	955	993	988	958	989	991	962

Table 5.20 shows the number of times all possible models are selected via AIC_c in each of 1000 simulations for all the combinations ρ_{23} and missing percentage with $n = 100$. The number of times true model M110 selected reduces as missing percentages increases. The chance of selecting the true model M110 was above 95% for all ρ_{23} values and missing percentages. The chances of selecting model M010 increases as missing percentages

increases. For $n = 200$ and $n = 400$, the true model M110 was chosen 100% via AIC_c compared to other possible models in each of the 1000 simulations for all combinations of ρ_{23} and for $m = 0, 25$ and 50. There are no effects of ρ_{23} in the frequency of selecting true model M110 for all sample size using M-STACK.

Table 5.21: MSE(P) for best model selected via AIC_c for all the combinations of ρ_{23}, missing percentages and sample sizes using M-STACK for logistic regression

AIC_c												
ρ_{23}	$\rho_{23}=0$			$\rho_{23}=0.25$			$\rho_{23}=0.5$			$\rho_{23}=0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$n=50$	0.0141	0.0152	0.0226	0.0138	0.0157	0.0222	0.0155	0.0151	0.0207	0.0145	0.0142	0.0186
$n=100$	0.0055	0.0061	0.0085	0.0057	0.0062	0.0082	0.0054	0.0061	0.0080	0.0056	0.0060	0.0071
$n=200$	0.0027	0.0031	0.0040	0.0027	0.0030	0.0039	0.0027	0.0030	0.0035	0.0026	0.0028	0.0032
$n=400$	0.0013	0.0016	0.0021	0.0013	0.0015	0.0020	0.0013	0.0015	0.0019	0.0013	0.0014	0.0017

Table 5.21 shows the MSE(P) for the best model selected via AIC_c for all the combinations of ρ_{23}, sample size and missing percentages respectively. The MSE(P) decreases as sample size increases. As missing percentages increases, the MSE(P) values increases. For larger sample size, there is no difference in MSE(P) values between $m = 0$ and $m = 25$. There are some difference on MSE(P) values as ρ_{23}, n and missing percentages increases.

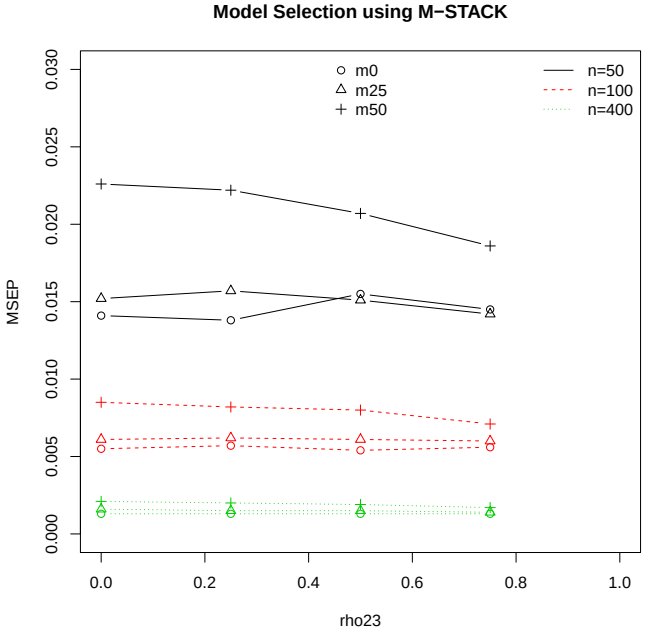


Figure 5.16: MSE(P) for for best model selected using M-STACK for each ρ_{23}, missing percentages and sample sizes for logistic regression

Figure 5.16 shows the MSE(P) for best model selected using M-STACK for each ρ_{23}, missing percentages and sample sizes. As sample size increases, the MSE(P) for best model selected using M-STACK decreases. The effects of missing percentages on MSE(P) for best model selected using M-STACK reduces as sample size increases. There are no effects of ρ_{23} in term of prediction for all sample sizes using M-STACK.

Figure 5.17 shows comparison between all three model selection methods (RR, M-STACK and STACK) via AIC_c for each ρ_{23}, missing percentages and $n = 100$. It shows that the MSE(P) for best model selected using STACK is lower than RR and M-STACK for all ρ_{23}, missing percentages and sample size.

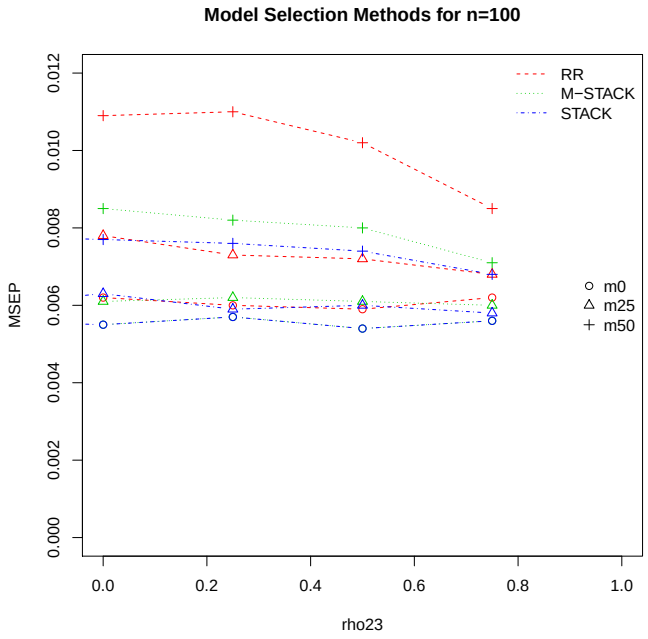


Figure 5.17: Comparison between all three model selection methods (RR, M-STACK and STACK) for each ρ_{23}, missing percentages and $n = 100$ for logistic regression

5.2.2.4 Model averaging using non-overlapping variable sets for Logistic regression

A simulation study was conducted based on simulation design as discussed earlier for logistic regression using model averaging via AIC_c and BIC. The analysis was carried out for every combination of sample size, ρ_{23} and missing percentages using non-overlapping variable sets. Table 5.22 shows the MSE(P) for model averaging via AIC_c for all the combinations of ρ_{23}, sample size and missing percentages respectively. The MSE(P) decreases as sample size and ρ_{23} increases. As missing percentages increases, the MSE(P) values increases. With $m = 0$, there is no clearer difference as ρ_{23} increases. With missing

percentages $m = 25$ and $m = 50$, the MSE(P) decreases as ρ_{23} increases. There are some difference on MSE(P) values as ρ_{23}, sample size and missing percentages increases. There are no significant increases or decreases in MSE(P) values as ρ_{23} increases.

Table 5.22: MSE(P) for model averaging via AIC_c for logistic regression

AIC_c												
ρ_{23}	$\rho_{23} = 0$			$\rho_{23} = 0.25$			$\rho_{23} = 0.5$			$\rho_{23} = 0.75$		
missing percentage	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50	m=0	m=25	m=50
$n = 50$	0.0162	0.0206	0.0256	0.0157	0.0186	0.0248	0.0161	0.0168	0.0213	0.0156	0.0171	0.0192
$n = 100$	0.0062	0.0075	0.0093	0.0062	0.0073	0.0092	0.0065	0.0070	0.0084	0.0062	0.0065	0.0076
$n = 200$	0.0027	0.0030	0.0041	0.0028	0.0031	0.0038	0.0027	0.0029	0.0038	0.0028	0.0029	0.0032
$n = 400$	0.0013	0.0015	0.0020	0.0013	0.0015	0.0019	0.0013	0.0014	0.0019	0.0013	0.0014	0.0017

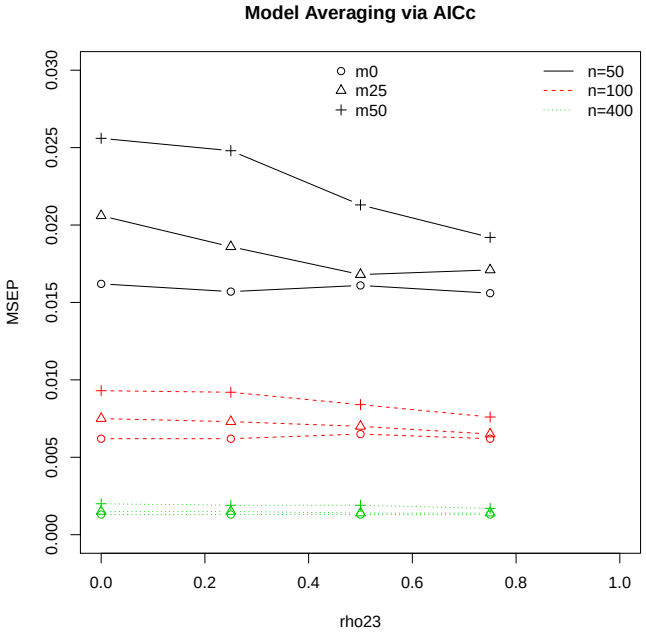
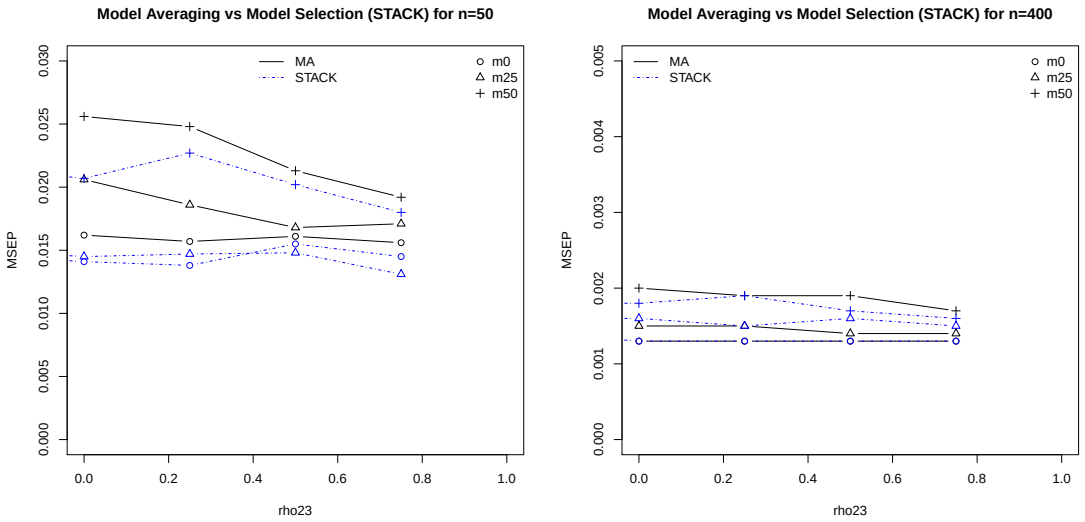


Figure 5.18: MSE(P) for model averaging via AIC_c using non-overlapping variable sets for each ρ_{23}, missing percentages and sample size for logistic regression

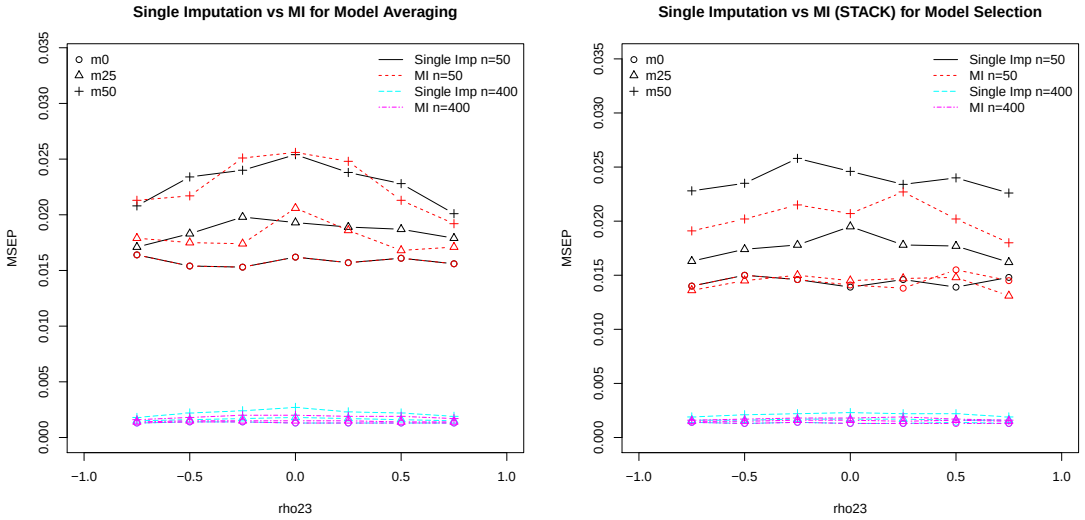
Figure 5.18 shows the MSE(P) for model averaging via AIC_c using non-overlapping variable sets for each ρ_{23}, missing percentages and sample sizes. As sample size increases, the MSE(P) for model averaging decreases. The effect of missing percentages on MSE(P) for model averaging reduces as sample size increases. Figure 5.19a shows comparison between model averaging and model selection (STACK) via AIC_c for each ρ_{23}, missing percentages and $n = 50$. Figure 5.19b shows comparison between model averaging and model selection (STACK) via AIC_c for each ρ_{23}, missing percentages and $n = 400$. The MSE(P) for model selection using STACK is lower than model averaging for small

sample sizes. It shows that model selection using STACK performs better than model averaging in terms of prediction for all sample sizes in Logistic regression.



(a) $n = 50$ (b) $n = 400$

Figure 5.19: Comparison between model averaging and model selection (STACK) via AIC_c for each ρ_{23}, missing percentages and sample sizes ($n = 50$ and $n = 400$) for logistic regression



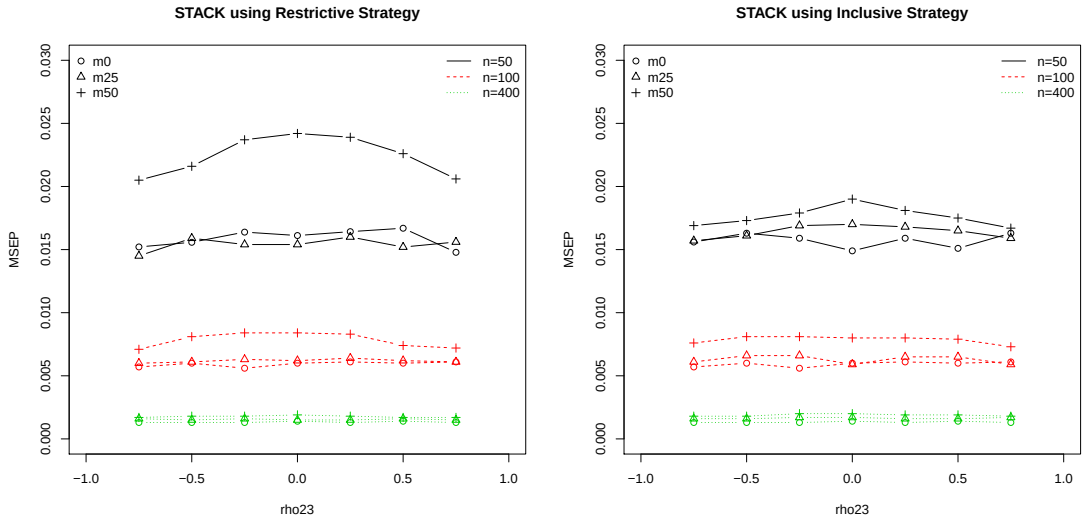
(a) Model Averaging (b) Model Selection

Figure 5.20: Comparison between single imputation and multiple imputation for model averaging and model selection via AIC_c for each ρ_{23}, $\sigma_\varepsilon = 1$, missing percentages and sample sizes for logistic regression

Figure 5.20a shows comparison between model averaging using single imputation and multiple imputation for each ρ_{23}, missing percentages and sample sizes ($n = 50$ and $n = 400$). It shows that MSE(P) of model averaging using multiple imputation is lower than MSE(P) of model averaging using single imputation for all missing percentages and sample sizes. Figure 5.20b shows comparison between model selection using single imputation and multiple imputation (STACK) for each ρ_{23}, missing percentages and sample sizes ($n = 50$ and $n = 400$). It shows that MSE(P) of model selection (STACK) using multiple imputation is lower than MSE(P) of model selection using single imputation for all missing percentages and sample sizes.

5.2.2.5 Model selection (STACK) and model averaging using restrictive and inclusive strategies for Logistic regression

Figure 5.21a and Figure 5.21b show the MSE(P) for best model selected (STACK) using the restrictive and inclusive strategies for each ρ_{23}, missing percentages and sample sizes respectively. The results shows that there are no effects of $|\rho_{23}|$ values on model selection (STACK) for logistic regression using restrictive and inclusive strategies for larger sample sizes.



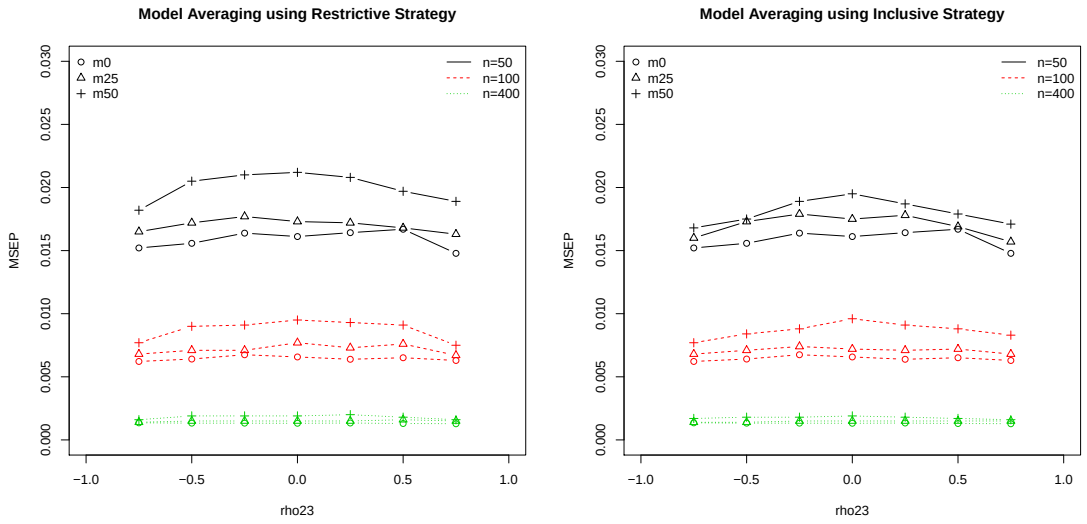
(a) Restrictive strategy

(b) Inclusive strategy

Figure 5.21: MSE(P) for for best model selected using STACK and the restrictive and inclusive strategies for each ρ_{23}, missing percentages and sample sizes for logistic regression

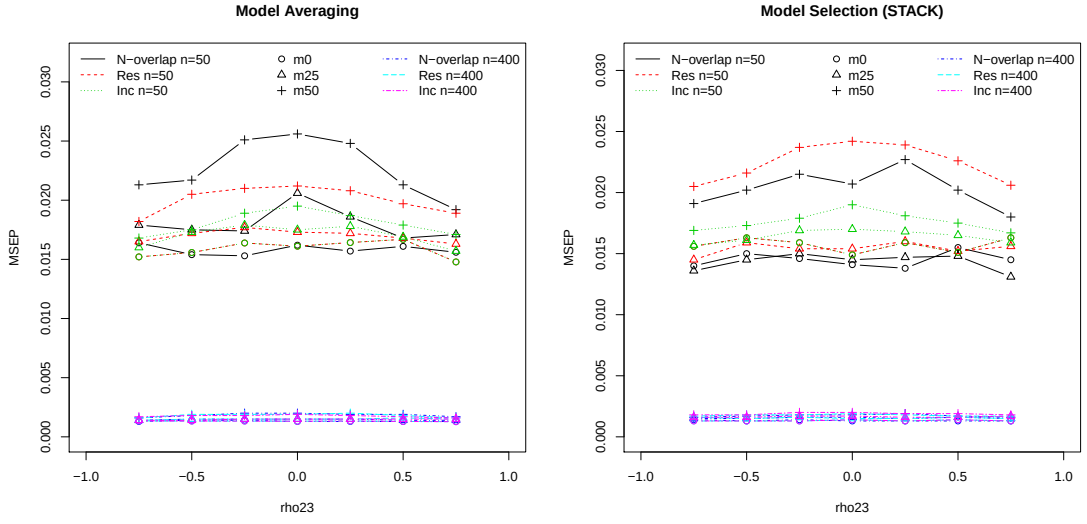
Figure 5.22a and Figure 5.22b show the MSE(P) for model averaging via AIC_c using the restrictive and inclusive strategies for each ρ_{23}, missing percentages and sample sizes respectively. The results shows that there are no effects of $|\rho_{23}|$ values on model

averaging for logistic regression using restrictive and inclusive strategies for all sample sizes.



(a) Restrictive strategy (b) Inclusive strategy

Figure 5.22: MSE(P) for model averaging via AIC_c using the restrictive and inclusive strategies for each ρ_{23}, missing percentages and sample sizes for logistic regression



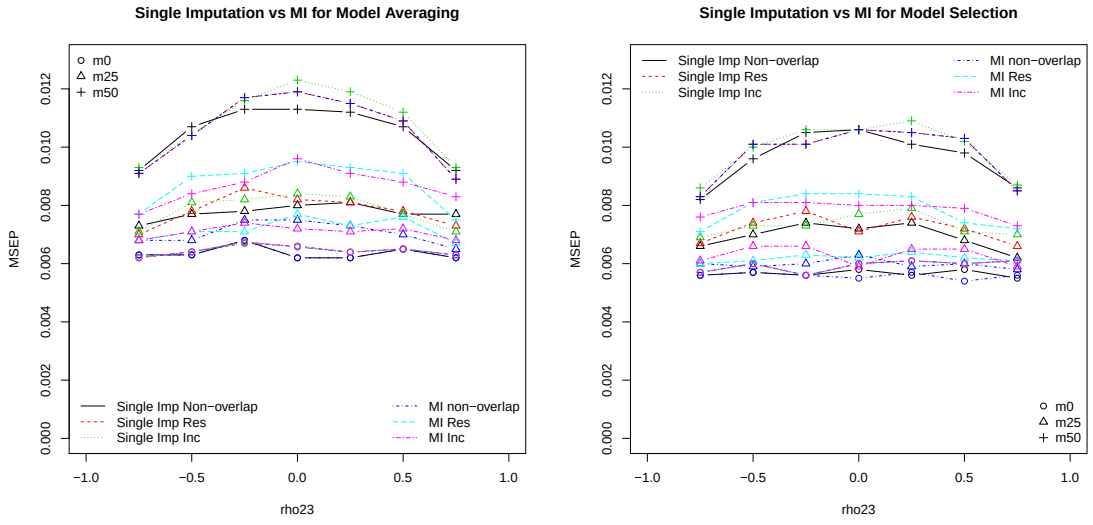
(a) Model Averaging (b) Model Selection (STACK)

Figure 5.23: Comparison between all three model-building strategies for model averaging and model selection (STACK) for multiply-imputed data sets for logistic regression

Figure 5.23 shows the comparison between all three model-building strategies (non-overlapping variable sets, restrictive and inclusive strategies) for model averaging and model selection (STACK) via AIC_c for multiply-imputed data sets for each ρ_{23}, missing

percentages and sample sizes, $n = 50$ and $n = 400$. The MSE(P) values for model averaging with an inclusive strategy is lower than MSE(P) values for model averaging with non-overlapping variable sets and restrictive strategy for small sample sizes. There are no differences between the MSE(P) values for model averaging and model selection (STACK) using all three model-building strategies for $|\rho_{23}|$. The MSE(P) values of model selection (STACK) using inclusive strategy is lower than the MSE(P) values of model selection (STACK) using non-overlapping variable sets and restrictive strategy for small sample size. However, there are no differences between the MSE(P) values for model averaging and model selection (STACK) using all three model-building strategies for large sample size.

Figure 5.24 shows comparison between single imputation and multiple imputation for model averaging and model selection using all three model-building strategies for each ρ_{23}, missing percentages and sample size, $n = 100$. The results show that the MSE(P) of model averaging using inclusive strategy for multiply-imputed data sets is better than using non-overlapping variable sets and restrictive strategy. Whereas, the MSE(P) of model selection using the inclusive strategy for multiply-imputed data sets is lower than MSE(P) of model selection for single imputation using all three model-building strategies for all missing percentages.



(a) Model Averaging

(b) Model Selection

Figure 5.24: Comparison between single imputation and multiple imputation for model averaging and model selection for each ρ_{23}, missing percentages and sample size, $n = 100$ for logistic regression

5.3 Discussion and Conclusions

The performance and effectiveness of three methods for model selection in linear model and Logistic regression were observed and compared. The effects of simulation parameters (sample size (n), missing percentages (m), the correlation between X_2 and X_3 (ρ_{23}) and error variance (σ_ε^2)) on model selection were observed. As discussed in Section 4.3, there are important effects of simulation parameters for complete data sets ($m = 0$) and imputed data sets ($m = 25$ and $m = 50$).

In linear models, σ_ε has a significant effect on model selection and prediction, both become poorer as σ_ε increases. However, STACK performs better than RR and M-STACK in terms of prediction for all values of σ_ε and sample sizes. M-STACK and STACK perform better than RR in terms of selecting the true model M110 more often. There is no difference between M-STACK and STACK in terms of selecting the true model M110. Since M-STACK and STACK perform better than RR, stacked imputed data with weighted linear regression is better than RR applied to linear regression. The performance of the three methods can be arranged in the order STACK > M-STACK > RR for linear models.

There is no difference between single imputation and multiple imputation in terms of model selection for all three methods (RR, STACK and STACK) when $\sigma_\varepsilon = 0.25$ for all sample sizes. For $\sigma_\varepsilon = 1$, even RR performs better than single imputation (as discussed in Section 4.2.1) in terms of selecting the true model M110 more frequently and giving a lower value of MSE(P), even for high missing percentage ($m = 50$). For $\sigma_\varepsilon = 4$, RR performs poorer than model selection using single imputation in model selection and prediction. However, RR performs better in terms of prediction than single imputation, for large sample size and high ρ_{23} values. This shows that model selection using single imputation is better than RR in extreme cases such as small samples and large error variances.

In addition, M-STACK performs better than model selection using single imputation for all σ_ε values in terms of model selection and prediction for linear models. This means that M-STACK is better than model selection using single imputation in all circumstances. Moreover, STACK performs better than model selection using single imputation for both $\sigma_\varepsilon = 1$ and $\sigma_\varepsilon = 4$ in terms of selecting the true model M110 more frequently and giving a lower value of MSE(P) compared to model selection using single imputation for linear models. Again, this shows that STACK is much better than model selection using single imputation in all circumstances.

Model averaging using multiple imputation for imputing missing data performs better in terms of prediction than model selection by STACK, for large error variance and small

sample size. There is no difference between model averaging and STACK in terms of prediction for small error variance and large sample size. Model averaging using multiple imputation performs better than model averaging using single imputation for all sample sizes, error variances and ρ_{23} in terms of prediction.

In the Logistic regression, model selection STACK performs better than RR and M-STACK for all sample sizes and ρ_{23} values. Model selection using single imputation is better than RR in terms of selecting the true model M110 more frequently. There is no difference in using single imputation and RR for large sample size with missing percentage $m = 25$. RR performs better than model selection using single imputation for all missing percentage, sample sizes and ρ_{23} values in terms prediction. RR showed significantly smaller MSE(P) values than model selection using single imputation. This shows that RR is better than single imputation in terms of prediction.

M-STACK performs better than model selection using single imputation for all sample sizes and ρ_{23} values in terms of selecting the true model M110 in Logistic regression after imputation. In terms of prediction, M-STACK performs better than model selection using single imputation for all sample sizes and ρ_{23} values. There are significant differences in terms of MSE(P) values between model selection using single imputation and M-STACK. Furthermore, STACK performs better than model selection using single imputation for all combinations of simulation parameters in terms of selecting the true model M110 more often after imputation. STACK performs better than model selection using single imputation for all sample sizes and ρ_{23} values in terms of prediction. STACK showed significantly smaller MSE(P) values than model selection using single imputation.

Besides that, all three model selection methods (RR, M-STACK and STACK) using multiple imputation for imputing missing data perform better than model selection using single imputation in terms of model selection and prediction for logistic regression. M-STACK and STACK perform better than RR in terms of selecting true model M110 more frequently and also in terms of prediction. This shows that model selection using model selection criteria is better for model selection and prediction. The performance of the three methods can be arranged in the order STACK > M-STACK > RR in terms of model selection and prediction for logistic regression.

In the Logistic regression, STACK performs better than model averaging using multiple imputation for all sample size in terms of prediction. Model averaging using multiple imputation for imputing missing data performs better than model averaging using single imputation for small sample sizes, missing percentages and ρ_{23} in terms of prediction. There are no difference between model averaging using single imputation and multiple imputation for large sample size. MSE(P) was lowest when inclusive strategy was used

with model averaging and model selection using single and multiple imputation. Negative and positive correlations of the same magnitude have the same effect on prediction for model averaging and model selection (STACK) using all three model-building strategies. There are no significant effects of ρ_{23} in terms of prediction in logistic regression.

Finally, the RR method is a gold standard approach but it is more computationally intensive when repeated analyses are required. The proposed method, STACK is a sensible alternative to RR and M-STACK method. RR and STACK provides similar parameter estimates if there is no model selection is required. STACK incorporate suitable model selection process and parameter estimation. STACK is computationally easier compared to RR method when numerous covariates are included in model-building. Moreover, Wood et al. [2008] stated that their stacked dataset method using backward stepwise selection approach is an alternative method RR but it is not a substitute for RR method. However, the STACK method used in this research can be an alternative to RR since the STACK method described in this research used all subset regression.

Although there are no difference between the M-STACK and STACK method using model selection criteria for prediction, STACK provides better prediction and parameter estimation than M-STACK if there is no model selection is required. As stated in Appendix A by Wood et al. [2008], the parameter estimates of STACK method is approximately similar as M-STACK method. M-STACK method is computationally easier compared to STACK. Therefore, researchers can use M-STACK for analysing data with missing values and also if numerous covariates are available. This will allow researchers to obtain results faster compared to STACK method.

In conclusion, all three methods (RR, M-STACK and STACK) using multiple imputation perform better than model selection using a single imputation method (as discussed in Section 4.2) for both linear model and logistic regression. Since M-STACK and STACK perform better than RR in terms of model selection and prediction for both models, the researcher should use stacked imputed data using weighted regression for analysing data sets with missing values. Generally, STACK performs better than M-STACK in terms of model selection and prediction in most of the circumstances investigated here. Model averaging performs slightly better than STACK in terms of prediction for linear models. Therefore, researchers should use STACK for analysing data with missing values for model selection but use model averaging for prediction in linear models. Whereas researchers should use STACK for model selection and prediction for logistic regression. In addition, researchers should use an inclusive imputation strategy for prediction in linear models and logistic regression. In line with the discussion in Section 4.3, researchers should carry out analysis using STACK with AIC_c as a model selection criterion and model averaging using AIC_c based weights for both linear model and Logistic regression,

and also use highly correlated auxiliary variables where they are available in imputation models.

In this chapter, we were interested in comparing all three model selection methods and model averaging for multiply-imputed data sets. All three model-building strategies (non-overlapping variable sets, inclusive and restrictive strategies) were investigated for both best model selection method (STACK) and model averaging. In the next chapter, we will explore the STACK (model selection) and model averaging in a real life dataset to investigate the performance of the proposed model-building strategies and methods for model selection and prediction.

Chapter 6

Application of Model Selection and Model Averaging to the Gateshead Millennium Study

In this chapter, some of the methods discussed earlier in the thesis are applied to the analysis of data from the Gateshead Millennium Study (GMS), a longitudinal study of child growth which suffers from a moderate amount of missing data. The purpose of the modelling is to predict children's weight or weight standard deviation score (SDS) later in childhood from weights (or weight SDS) recorded in the first year of life. It was concluded in Chapter 4 and Chapter 5 that model averaging and model selection (STACK) perform best for prediction and also to determine the factors to be included when making predictions. Therefore, both these model-building approaches will be applied to combine results from multiply-imputed data sets using all three model-building strategies (non-overlapping variable sets, inclusive and restrictive). The dataset will be explored in the first section and formal modelling of children's weight at school entry and at eight years will be carried out in the following section.

6.1 Data Description of Gateshead Millennium Study

Various studies have found significant associations between rapid infancy weight gain and later overweight, leading to the suggestion that prevention and treatment of childhood obesity should begin as early as the first year of life. The Gateshead Millennium Study (GMS) is a prospective cohort study of feeding and growth in infancy. This study was set up primarily to explore the relationship between child development and feeding in the first year of life, but was later extended to continue to follow up the children throughout

childhood. Babies born between 1 June 1999 and 31 May 2000 in the Gateshead area of northeast England were recruited to the study shortly after birth. There is a total of 1029 babies of 1011 mothers, 524 boys and 505 girls, representing 83% of all births in the region that year. The children were studied prospectively using parent report questionnaire shortly after birth, at 6 weeks and at 4, 8 and 12 months. The cohort has since been re-traced at school entry, parent report questionnaires completed at 5-8 years, and a range of anthropometric and body composition measures collected at age 7-8 years [Wright et al., 2011].

Table 6.1: Description of Variables for GMS

Variables	Descriptions	Unit
X_1	Birth weight	kilograms (kg)
X_2	Weight at 6 weeks	kilograms (kg)
X_3	Weight at 4 months	kilograms (kg)
X_4	Weight at 8 months	kilograms (kg)
X_5	Weight at 12 months	kilograms (kg)
X_6	Gestational age	weeks
Y_1	Weight at school entry	kilograms (kg)
Y_2	Weight at 8 years	kilograms (kg)

Table 6.1 shows a description of the variables that are used in the analysis reported here. The dependent variables are the weight at school entry (Y_1) and weight at 8 years (Y_2). The independent variables are birth weight (X_1) , weight at 6 weeks (X_2), weight at 4 months (X_3), weight at 8 months (X_4), weight at 12 months (X_5) and gestational age (X_6). All these variables are quantitative and continuous except for gestational age which was rounded to the nearest whole number of weeks.

Table 6.2 shows the descriptive statistics for boys and girls separately. There are no missing data for baby's birth weights and gestational ages for boys or girls. The weight at school entry was missing for 29.6% of children and weight at eight years was missing for 42%. On average 17% of weights from the first year of life were missing.

Figure 6.1a shows the weights at school entry for male and female children, which are very similar on average. There are a number of exceptionally overweight children, especially female children. The female child whose weight at school entry was more than 50kg will be removed from the modelling, as it is an extreme outlier. Figure 6.1b shows the weight at eight years for male and female children. Again, these are very similar distributions. There are outliers in the weight at eight years for both male and female children who are exceptionally overweight.

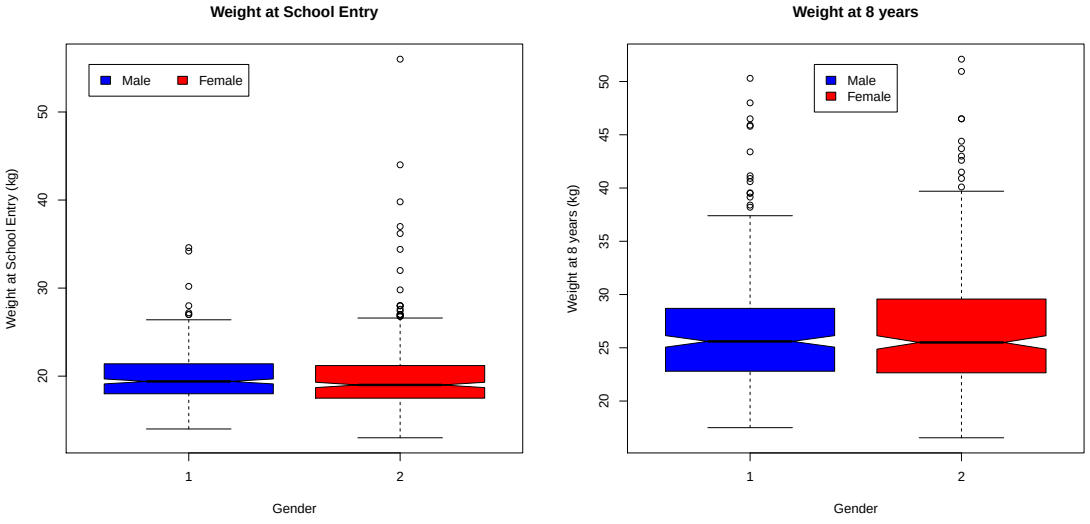
Table 6.2: Descriptive statistics

(a) Boys

Statistics	Variables							
	X_6	X_1	X_2	X_3	X_4	X_5	Y_1	Y_2
Mean	39.31	3.38	4.91	6.89	9.13	10.53	19.77	26.48
Standard deviation	1.88	0.58	0.65	0.86	1.00	1.19	2.85	5.45
Median	40.00	3.42	4.89	6.86	9.12	10.48	19.40	25.60
Minimum	29.00	1.36	3.18	4.42	6.62	7.54	14.00	17.50
Maximum	43.00	4.96	6.80	9.75	13.28	14.30	34.60	50.30
complete cases (n)	524	524	437	445	327	435	357	297
missing observation (n_{mis})	0	0	87	79	197	89	167	227
percentage of missing(m)	0	0	16.8%	15.1%	37.6%	17.0%	31.9%	43.3%

(b) Girls

Statistics	Variables							
	X_6	X_1	X_2	X_3	X_4	X_5	Y_1	Y_2
Mean	39.41	3.27	4.53	6.31	8.36	9.81	19.82	26.75
Standard deviation	1.80	0.59	0.58	0.82	0.98	1.21	4.10	5.88
Median	40.00	3.30	4.52	6.26	8.32	9.67	19.00	25.50
Minimum	27.00	0.84	2.76	4.09	5.27	6.24	13.00	16.55
Maximum	43.00	5.37	6.51	8.85	11.66	15.70	56.00	52.10
complete cases (n)	505	505	415	430	323	423	367	300
missing observation (n_{mis})	0	0	90	75	182	82	138	205
percentage of missing(m)	0	0	17.8%	14.9%	36.0%	16.2%	27.3%	40.6%

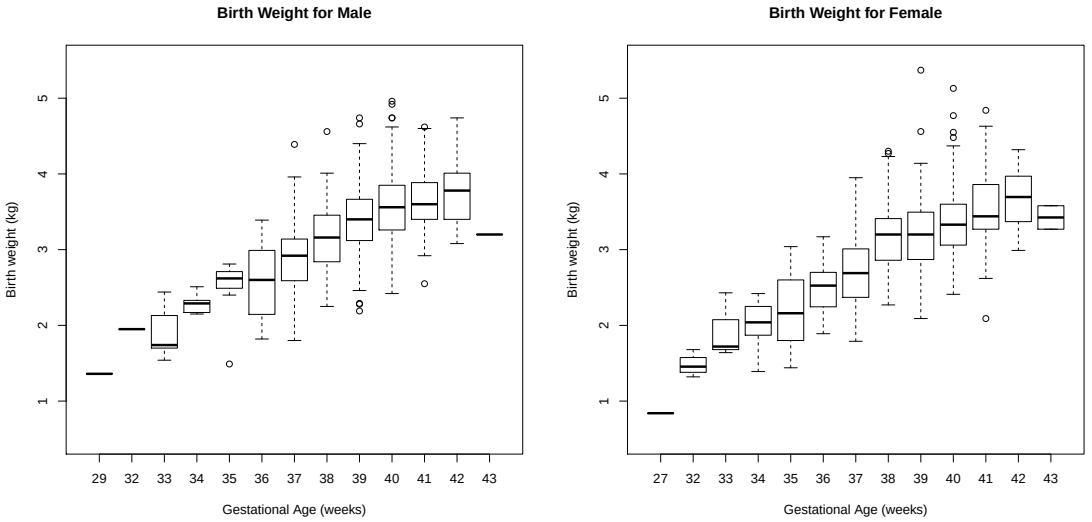


(a) Weight at school entry (Y_1)

(b) Weight at eight years (Y_2)

Figure 6.1: Weight at school entry and weight at eight years for boys and girls separately

Figure 6.2a and Figure 6.2b show the relationship between the birth weight and gestational age for male and female babies respectively. Gestational age gives an idea about the baby's growth and development during pregnancy and whether a baby can be expected to live outside the uterus. Generally, the median of birth weight increases as gestational age increases. The apparent decrease after 42 weeks gestational age is likely to be a result of very small numbers and mis-reporting of dates of conception. Premature babies (born before 37 weeks gestational age) have low birth weight compared to babies born after 37 weeks gestational age, so premature babies might have to be removed from the analysis.



(a) Male babies

(b) Female babies

Figure 6.2: The relationship between birth weight and gestational age for both male and female babies

Table 6.3: Correlations of Weights

Variables	gestational age	birth weight	weight at 6 weeks	weight at 4 months	weight at 8 months	weight at 12 months	weight at school entry	weight at 8 years
gestational age	1	0.4099	0.1604	0.1691	0.0873	0.0787	0.0509	-0.0652
birth weight	0.4099	1	**0.6642**	0.5065	0.5071	0.4557	0.1926	0.2540
weight at 6 weeks	0.1604	0.6642	1	**0.7847**	**0.6848**	**0.6223**	0.1731	0.4377
weight at 4 months	0.1691	0.5065	0.7847	1	**0.8609**	**0.7493**	0.2135	0.3966
weight at 8 months	0.0873	0.5071	0.6848	0.8609	1	**0.9046**	0.2307	0.4786
weight at 12 months	0.0787	0.4557	0.6223	0.7493	0.9046	1	0.2206	0.4976
weight at school entry	0.0509	0.1926	0.1731	0.2135	0.2307	0.2206	1	0.3254
weight at 8 years	-0.0652	0.2540	0.4377	0.3966	0.4786	0.4976	0.3254	1

Table 6.3 shows the Pearson correlations between all pairs of these variables. There is a moderate correlation ($\rho = 0.4099$) between gestational age and birth weight of a baby but the correlations between the gestational age and other weights are low and decreasing with age. There are stronger positive relationships between the weight at

8 years and first year baby weights compared to weight at school entry and first year baby weights. Moreover, the first year baby weights are highly correlated with their neighbouring weights. These relationships are more clearly shown in the scatter plots of Figure 6.3. Neighbouring weights appear to be good candidate variables for imputation purposes.

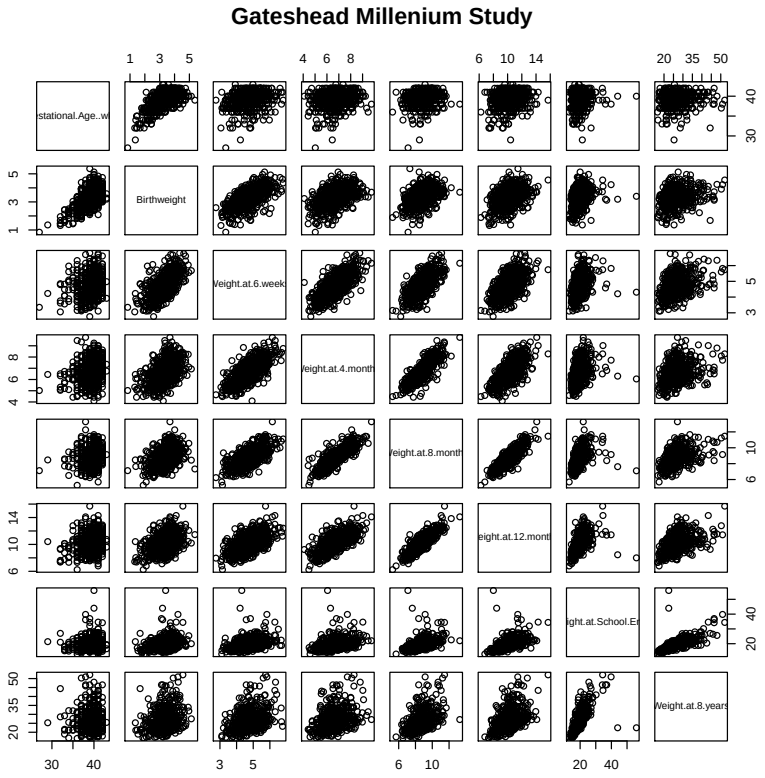


Figure 6.3: Relationship between the weight at school entry, weight at eight years and the first year baby weights

The raw weights (except those at school entry) were converted to Standard Deviation Scores (SDS) compared to the British 1990 growth reference [Freeman et al., 1995] using a Box-Cox transformation. The SDS or Z-scores represent the difference between the actual weight and the population mean weight in units of the standard deviation. Converting raw weights to standard deviation scores is intended to result in the transformed data at any given age, having an approximate standard Normal distribution with mean 0 and variance 1 in the reference population.

Table 6.4 shows the descriptive statistics based on weight Z-scores for boys and girls separately. Figure 6.4 shows the weight Z-score at eight years for both male and female children. The median weight Z-scores at eight years for male children (median=1.09) is higher than for female children (median=0.88), but both are much higher than the

reference value of 0. There are outliers in the weight Z-scores at eight years for both male and female children, male children have weight Z-scores that are as low as -2.60 and as high as 4.56 whereas some female children have weight Z-scores that are as low as -2.80 and as high as 5.22.

Table 6.4: Descriptive statistics - weight SDS

(a) Boys

Statistics	Variables					
	X_1	X_2	X_3	X_4	X_5	Y_2
Mean	-0.20	0.00	0.06	0.24	0.10	1.14
Standard deviation	1.07	1.00	0.98	1.02	1.03	1.26
Median	-0.14	-0.01	0.13	0.26	0.15	1.09
Minimum	-3.87	-3.13	-3.04	-2.62	-2.87	-2.60
Maximum	2.78	2.89	3.40	3.86	3.41	4.56
complete cases (n)	524	438	443	327	435	264
missing observation (n_{mis})	0	86	81	197	89	260
percentage of missing(m)	0	18.30%	17.20%	41.80%	18.90%	55.20%

(b) Girls

Statistics	Variables					
	X_1	X_2	X_3	X_4	X_5	Y_2
Mean	-0.19	-0.14	-0.09	0.06	0.02	1.00
Standard deviation	1.13	1.01	1.07	1.10	1.11	1.39
Median	-0.18	-0.13	-0.10	0.02	-0.06	0.88
Minimum	-3.89	-3.58	-3.66	-4.05	-4.05	-2.80
Maximum	3.96	2.72	2.92	3.21	3.52	5.22
complete cases (n)	505	420	430	323	423	249
missing observation (n_{mis})	0	85	75	182	81	255
percentage of missing(m)	0	18.70%	16.50%	40.10%	17.80%	56.20%

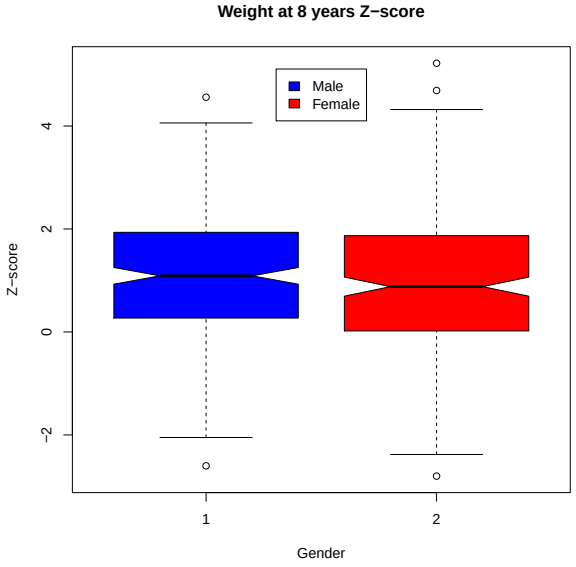


Figure 6.4: Weight Z-scores at eight years for both male and female children

Table 6.5: Correlations of weight Z-scores

Variables	gestational age	bwtz	wtz6w	wtz4m	wtz8m	wtz12m	wtz8y
gestational age	1	0.2212	0.2359	0.1786	0.0887	0.0681	-0.0594
bwtz	0.2212	1	**0.7394**	**0.5227**	0.4497	0.3966	0.2291
wtz6w	0.2359	0.7394	1	**0.8514**	**0.6977**	**0.6077**	0.2386
wtz4m	0.1786	0.5227	0.8514	1	**0.9008**	**0.7798**	0.1700
wtz8m	0.0887	0.4497	0.6977	0.9008	1	**0.9255**	0.2038
wtz12m	0.0681	0.3966	0.6077	0.7798	0.9255	1	0.1517
wtz8y	-0.0594	0.2291	0.2386	0.1700	0.2038	0.1517	1

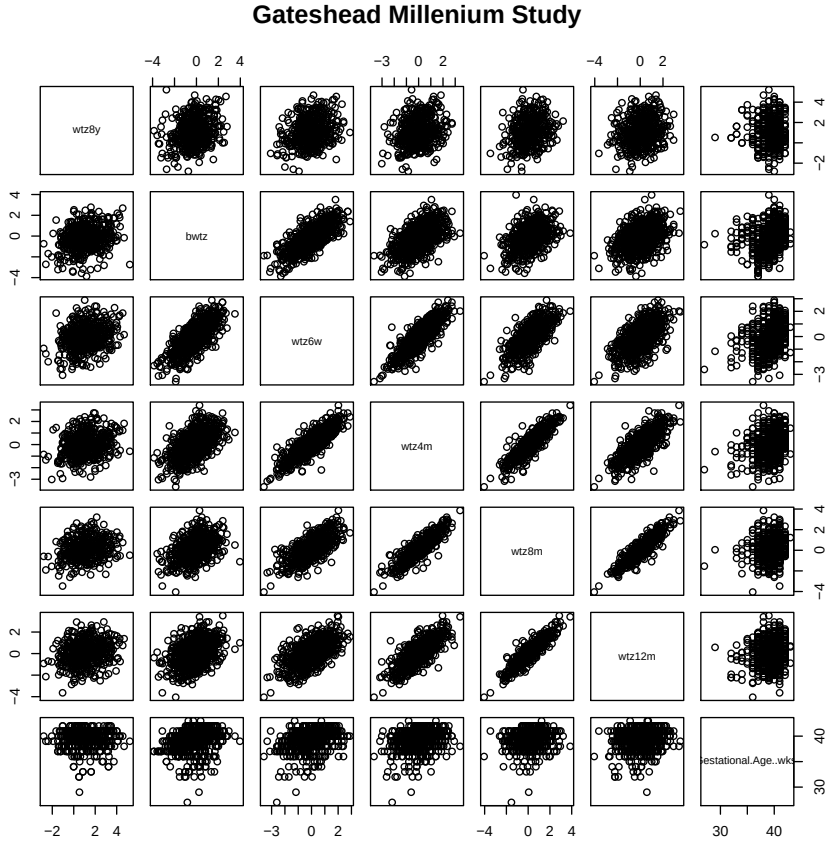


Figure 6.5: Relationship between the weight Z-scores at eight years and the first year baby weights

Table 6.5 shows the Pearson correlations between the weight Z-scores and gestational age and Figure 6.5 shows the corresponding scatter plots. The correlations between gestational age and weight Z-scores are generally low. There are weak positive relationships between the weight Z-scores at eight years and at earlier ages. Besides that, the first year weight Z-scores are highly correlated with their neighbouring weight Z-scores and

these correlations are higher than those for raw first year weights. Therefore, using the neighbouring weight Z-scores appears to be a good strategy for imputation purposes.

6.2 Model-building and Results

In this section, we will discuss the results on prediction of children's weight at school entry based on first year weights, prediction of children's weight at eight years based on first year weights and prediction of children's weight Z-scores at eight years based on first year weights Z-scores. The most commonly used technique for dealing with missing data is the method of complete case analysis, where the analysis is carried out using only babies that have no missing values for any of the variables used in the model.

The complete case analysis was carried out using model selection and model averaging as discussed in Chapter 4. For model selection in complete case analysis, model selection criterion AIC_c and BIC were allowed to choose a model based on any combinations of covariates. AIC_c and BIC based weights were used for model averaging. The incomplete data analysis was carried out using model averaging and STACK with all three model-building strategies (non-overlapping variable sets, the inclusive and restrictive strategies) as discussed in Chapter 5. The gestational age was used as an auxiliary variable for non-overlapping variable sets and the restrictive strategy whereas, for the inclusive strategy, the first year weights and the gestational age were used for both the imputation and prediction models. Note that there are strong correlation between the covariates but they are weakly correlated with the response variables. This favour the assumptions of STACK method and model averaging for prediction in the context of GMS analysis.

As discussed in Section 6.1, there are some outliers (premature babies and heavy weight children) which will affect the prediction and imputation. The complete cases analysis and incomplete case analysis were carried out initially with the outliers. The results showed that the MSE(P) values are much higher for analysis with outliers compared to without outliers. Therefore, the outliers, the premature baby (gestational age < 30 weeks) and heavy weight children (weight at school entry > 50kg) were removed. The complete cases for prediction of children's weight at school entry were 209 male babies and 238 female babies, whereas the complete cases for prediction of children's weight at 8 years were 207 male babies and 220 female babies. The complete cases for prediction of children's weight Z-scores at 8 years were 189 male babies and 194 female babies.

In addition, in the GMS data, there are missing data in the response variables as well as the covariates. This is different than the setting of simulation studies in Chapter 4 and Chapter 5 where there were no missing data in the response variables. All missing data

were imputed using the "norm" method in the R package MICE. The non-overlapping variable sets, inclusive and restrictive strategies were used for imputation and prediction models. For non-overlapping variable sets and restrictive strategy, the missing values were imputed using gestational age whereas for inclusive strategy, the missing values were imputed using first year baby weights and gestational age. The imputation model for non-overlapping and restrictive strategy is the same but the full prediction model for restrictive and inclusive strategies is the same.

The average MSE(P) for complete cases (MSE(P)-CC) was calculated based on estimates of multiply-imputed data. The cross validation was carried out to assess whether the predicted values from the chosen model are accurately predict responses. The cross validation test was carried out with 10% of complete case data and the estimation was carried out based 90% incomplete dataset. Here 10% of observations is omitted from the analysis and the response for that observation is predicted using the model derived from the remaining 90% of observations. The average MSE(P) for 10% of complete case data (MSE(P)-CV) was calculated based on estimates of 90% multiply-imputed data.

6.2.1 Complete case analysis

The complete case analysis was carried out using both non-overlapping variable sets and inclusive/restrictive strategies for prediction models since there are no imputations involved. Model selection criteria AIC_c and BIC were allowed to choose a model based on any combination of variables for non-overlapping variable sets (without gestational age in prediction model) and also for incusive/restrictive strategy (with gestational age in prediction model).

Table 6.6: Estimates and MSE(P) for prediction of weight at school entry for male children in complete case analysis

Approaches	**Model averaging**				**Model selection**			
Strategies	**Non-over**		**Restrictive/Inclusive**		**Non-over**		**Restrictive/Inclusive**	
Model selection Criterion	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**
Constant	1.9437	2.4843	5.8758	3.4876	1.5769	2.4092	8.4534	2.4092
Birth weight	0.5737	0.6068	0.8054	0.7141	0.5539	-	0.8902	-
Weight at 6 weeks	0.2488	0.3977	0.1945	0.3901	-	-	-	-
Weight at 4 months	-0.1975	0.0174	-0.1706	0.0177	-	-	-	-
Weight at 8 months	0.6572	0.6539	0.6413	0.6484	0.5899	0.6780	0.5815	0.6780
Weight at 12 months	1.1271	1.2776	1.1182	1.2753	1.0265	1.0519	1.0116	1.0519
Gestational Age	-	-	-0.1856	-0.1655	-	-	-0.1981	-
Average MSE(P)	**7.9714**	**45.1143**	**5.0988**	**5.6785**	**4.2418**	**4.3187**	**4.1658**	**4.3187**
Error variance (σ_ε^2)	4.2716	4.5616	4.1910	4.4476	4.2622	4.3395	4.1858	4.3395

Table 6.6 shows the estimates and mean squared error prediction (MSE(P)) for prediction of weight at school entry for complete case analysis of male children. The results showed that MSE(P) values for model selection using restrictive/inclusive strategy is the lowest. The factors that contribute to predict weight at school entry for male children are birth weight, weight at 8 months, weight at 12 months and gestational age. If weight at 12 months increase by 1 kg, the weight at school entry will increase by 1.0116kg.

Table 6.7: Estimates and MSE(P) for prediction of weight at school entry for female children in complete case analysis

Approaches	**Model averaging**				**Model selection**			
Strategies	**Non-over**		**Restrictive/Inclusive**		**Non-over**		**Restrictive/Inclusive**	
Model selection Criterion	**AIC$_c$**	**BIC**	**AIC$_c$**	**BIC**	**AIC$_c$**	**BIC**	**AIC$_c$**	**BIC**
Constant	-0.2717	-0.3371	8.0352	2.6164	-0.3780	-0.3780	9.7372	-0.3780
Birth weight	0.0911	0.0981	0.4995	0.3161	-	-	-	-
Weight at 6 weeks	0.2007	0.1556	0.3389	0.2414	-	-	-	-
Weight at 4 months	0.3720	0.2200	0.3945	0.2429	-	-	-	-
Weight at 8 months	-0.6542	-0.5336	-0.6523	-0.5249	-	-	-	-
Weight at 12 months	2.1906	2.0859	2.1643	2.0868	2.0537	2.0537	2.0852	2.0537
Gestational Age	-	-	-0.3075	-0.2761	-	-	-0.2626	-
Average MSE(P)	**10.1900**	**12.8232**	**15.6472**	**79.0300**	**10.1108**	**10.1108**	**9.9661**	**10.1108**
Error variance (σ^2_ε)	10.0519	10.0980	9.8507	9.8804	10.1533	10.1533	10.0080	10.1533

Table 6.7 shows the estimates and MSE(P) for prediction of weight at school entry for complete case analysis of female children. The results showed that MSE(P) values for model selection using restrictive/inclusive strategy is the lowest. The factors that contribute to predict weight at school entry for female children are weight at 12 months and gestational age. If weight at 12 months increase by 1 kg, the weight at school entry will increase by 2.0852kg. There is a negative relationship between weight at school entry and gestational age in predicting weight at school entry for both male and female children.

Table 6.8: Estimates and MSE(P) for prediction of weight at eight years for male children in complete case analysis

Approaches	**Model averaging**				**Model selection**			
Strategies	**Non-over**		**Restrictive/Inclusive**		**Non-over**		**Restrictive/Inclusive**	
Model selection Criterion	**AIC$_c$**	**BIC**	**AIC$_c$**	**BIC**	**AIC$_c$**	**BIC**	**AIC$_c$**	**BIC**
Constant	-1.1712	-0.6485	6.4679	1.5049	-1.2973	-1.2973	9.9483	-1.2973
Birth weight	0.3085	0.5589	0.8775	0.8300	-	-	-	-
Weight at 6 weeks	1.8432	1.9145	1.8958	1.9518	1.9763	1.9763	2.2149	1.9763
Weight at 4 months	0.2025	0.5933	0.2518	0.5897	-	-	-	-
Weight at 8 months	1.0008	1.6132	0.9911	1.6080	-	-	-	-
Weight at 12 months	1.5305	1.6760	1.5156	1.6715	1.7110	1.7110	1.6950	1.7110
Gestational Age	-	-	-0.3710	-0.3308	-	-	-0.3114	-
Average MSE(P)	**105.5651**	**455.2460**	**41.9434**	**142.9630**	**20.9590**	**20.9590**	**20.7242**	**20.9590**
Error variance (σ^2_ε)	21.7621	25.2349	21.5658	25.1836	21.0608	21.0608	20.8248	21.0608

Table 6.8 shows the estimates and MSE(P) for prediction of weight at eight years for complete case analysis of male children. The results showed that MSE(P) values for model selection using restrictive/inclusive strategy is the lowest. The factors that contribute to predict weight at eight years for male children are weight at 6 weeks, weight at 12 months and gestational age. If weight at 6 weeks increase by 1 kg, the weight at eight years will increase by 2.2149kg.

Table 6.9: Estimates and MSE(P) for prediction of weight at eight years for female children in complete case analysis

Approaches	**Model averaging**				**Model selection**			
Strategies	**Non-over**		**Restrictive/Inclusive**		**Non-over**		**Restrictive/Inclusive**	
Model selection Criterion	**AIC$_c$**	**BIC**	**AIC$_c$**	**BIC**	**AIC$_c$**	**BIC**	**AIC$_c$**	**BIC**
Constant	-1.5930	-1.6752	12.3727	2.0006	-1.7077	-1.7077	17.8742	-1.7077
Birth weight	0.5972	0.5685	1.4627	1.1120	-	-	1.4878	-
Weight at 6 weeks	0.3738	0.3440	0.2540	0.3778	-	-	-	-
Weight at 4 months	-0.1886	-0.2156	-0.2238	-0.2195	-	-	-	-
Weight at 8 months	-0.7920	-0.7212	-0.8252	-0.7291	-	-	-	-
Weight at 12 months	3.0444	2.9223	2.9982	2.9168	2.8817	2.8817	2.6809	2.8817
Gestational Age	-	-	-0.5120	-0.4096	-	-	-0.5705	-
Average MSE(P)	**26.9278**	**33.5009**	**74.4343**	**223.16000**	**21.0932**	**21.0932**	**20.4596**	**21.0932**
Error variance (σ^2_ε)	21.0664	21.1452	20.5372	20.6062	21.1895	21.1895	21.1895	20.5530

Table 6.9 shows the estimates and MSE(P) for prediction of weight at eight years for complete case analysis of female children. The results showed that MSE(P) values for model selection using restrictive/inclusive strategy is the lowest. The factors that contribute to predict weight at eight years for female children are birth weight, weight at 12 months and gestational age. If weight at 12 months increase by 1 kg, the weight at eight years will increase by 2.6809kg. There is a negative relationship between weight at eight years and gestational age in predicting weight at eight years for both male and female children.

Table 6.10: Estimates and MSE(P) for prediction of weight at eight years Z-scores for male children in complete case analysis

Approaches	**Model averaging**				**Model selection**			
Strategies	**Non-over**		**Restrictive/Inclusive**		**Non-over**		**Restrictive/Inclusive**	
Model selection Criterion	**AIC$_c$**	**BIC**	**AIC$_c$**	**BIC**	**AIC$_c$**	**BIC**	**AIC$_c$**	**BIC**
Constant	1.1671	1.1755	3.8446	2.3854	1.1838	1.1838	4.9337	1.1838
Birth weight Z-score	0.3242	0.3428	0.3508	0.3569	0.3579	0.3579	0.4050	0.3579
Weight at 6 weeks Z-score	0.0252	0.1051	0.0530	0.1116	-	-	-	-
Weight at 4 months Z-score	0.0030	0.0816	0.0312	0.0880	-	-	-	-
Weight at 8 months Z-score	0.2216	0.1562	0.2117	0.1516	-	-	-	-
Weight at 12 months Z-score	-0.0745	0.0213	-0.0787	0.0194	-	-	-	-
Gestational Age	-	-	-0.0971	-0.0960	-	-	-0.0953	-
Average MSE(P)	**1.3442**	**1.4084**	**2.6039**	**7.8115**	**1.3561**	**1.3561**	**1.3298**	**1.3561**
Error variance (σ^2_ε)	1.3513	1.4152	1.3266	1.3746	1.3633	1.3633	1.3369	1.3633

Table 6.10 shows the estimates and MSE(P) for prediction of weight at eight years Z-scores for complete case analysis of male children. The results showed that MSE(P) values for model selection using restrictive/inclusive strategy is the lowest. The factors that contribute to predict weight at eight years Z-scores for male children are birth weight Z-scores and gestational age. If birth weight Z-scores increase by 1, the weight at eight years Z-scores will increase by 0.4050. There is a negative relationship between weight at eight years Z-scores and gestational age in predicting weight at eight years Z-scores for both male children.

Table 6.11: Estimates and MSE(P) for prediction of weight at eight years Z-scores for female children in complete case analysis

Approaches	**Model averaging**				**Model selection**			
Strategies	**Non-over**		**Restrictive/Inclusive**		**Non-over**		**Restrictive/Inclusive**	
Model selection Criterion	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**
Constant	1.0312	1.0362	1.7230	1.1833	1.0306	1.0678	1.0306	1.0678
Birth weight Z-score	0.1603	0.2211	0.1634	0.2227	-	-	-	-
Weight at 6 weeks Z-score	0.4929	0.4336	0.5103	0.4379	0.5673	0.4609	0.5673	0.4609
Weight at 4 months Z-score	-0.4902	-0.3949	-0.4997	-0.3984	-0.4406	-	-0.4406	-
Weight at 8 months Z-score	0.3716	0.3458	0.3740	0.3383	-	-	-	-
Weight at 12 months Z-score	0.3275	0.3307	0.3254	0.3302	0.4107	-	0.4107	-
Gestational Age	-	-	-0.0531	-0.0431	-	-	-	-
Average MSE(P)	**1.7818**	**1.8257**	**3.7717**	**4.2667**	**1.7050**	**1.7801**	**1.7050**	**1.7801**
Error variance (σ_ε^2)	1.7910	1.8351	1.7772	1.8128	1.7138	1.7893	1.7138	1.7893

Table 6.11 shows the estimates and MSE(P) for prediction of weight at eight years Z-scores for complete case analysis of female children. The results showed that MSE(P) values for model selection using restrictive/inclusive strategy and non-overlapping variable sets are the lowest. The factors that contribute to predict weight at eight years Z-scores for female children are weight at 6 weeks Z-scores, weight at 4 months Z-scores and weight at 12 months Z-scores. If weight at 6 weeks increase by 1, the weight at eight years Z-scores will increase by 0.5673.

Generally, in all three predictions of weight at school entry, weight at eight years and weight Z-scores at eight years, BIC performs poorly for predictions using model averaging. The MSE(P) values for model averaging based on BIC weights is much higher compared to those based on AIC_c weights. This is due to the effects of BIC's penalty term (more strict than AIC_c), where smaller models are given more weight in model averaging based on BIC weights. The AIC_c performs better than BIC in selecting the best model for predicting weight at school entry, weight at eight years and weight at eight years Z-scores for both male and female children. Table 6.12 shows the comparison between parameter values used in the simulation studies and those for the GMS data analysis. The effects of these parameters will be observed in the prediction analysis of

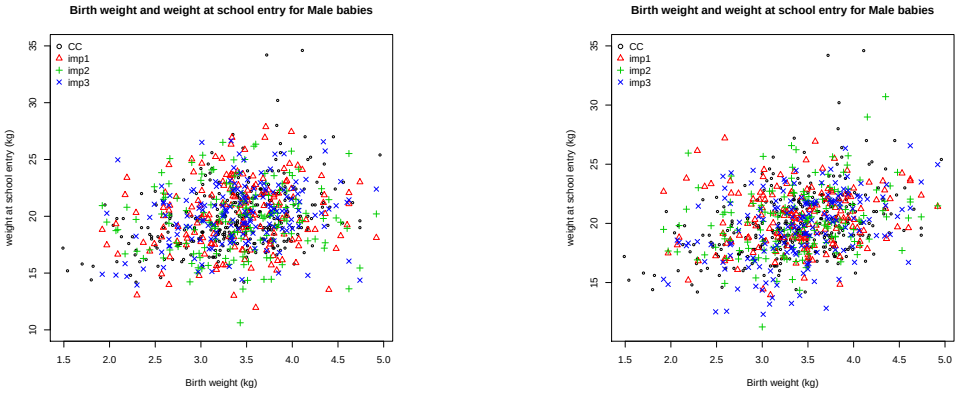
GMS data to compare the effects of these parameters in simulation setting and real-data application.

Table 6.12: Comparison between parameters used in simulation studies and GMS data

Parameters	**Simulation studies**	**GMS data**
m	0,25,50	$0 \leq m \leq 55$
n	50, 100, 200, 400	524 (boys) and 505 (girls)
σ_ε^2	$\frac{1}{16}, 1, 16$	1.5, 4, 10, 21
Number of parameters ($\beta's$)	up to 4	up to 7
correlations between auxiliary variable and covariates	$-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75$	-0.06 to 0.41
correlations between covariates	0	0 to 0.93
number of covariates	2	5
number of auxiliary variable	1	1
number of models for non-overlapping	4	32
number of models for restrictive/inclusive	8	64

6.2.2 Prediction of weight at school entry using multiple imputation

The incomplete data were imputed using all three model-building strategies (non-overlapping variable sets, restrictive and inclusive strategies) for both the STACK method and model averaging. Figure 6.6 shows the distribution of imputed values for weight at school entry for male children using non-overlapping variable sets, restrictive and inclusive strategies. The distribution of imputed values for weight at school entry for male children using inclusive strategy are closer to observed values compared to those for non-overlapping variable sets and restrictive strategy. Therefore, imputation using inclusive strategy are better than using non-overlapping variable sets and restrictive strategy for weight at school entry for male children.

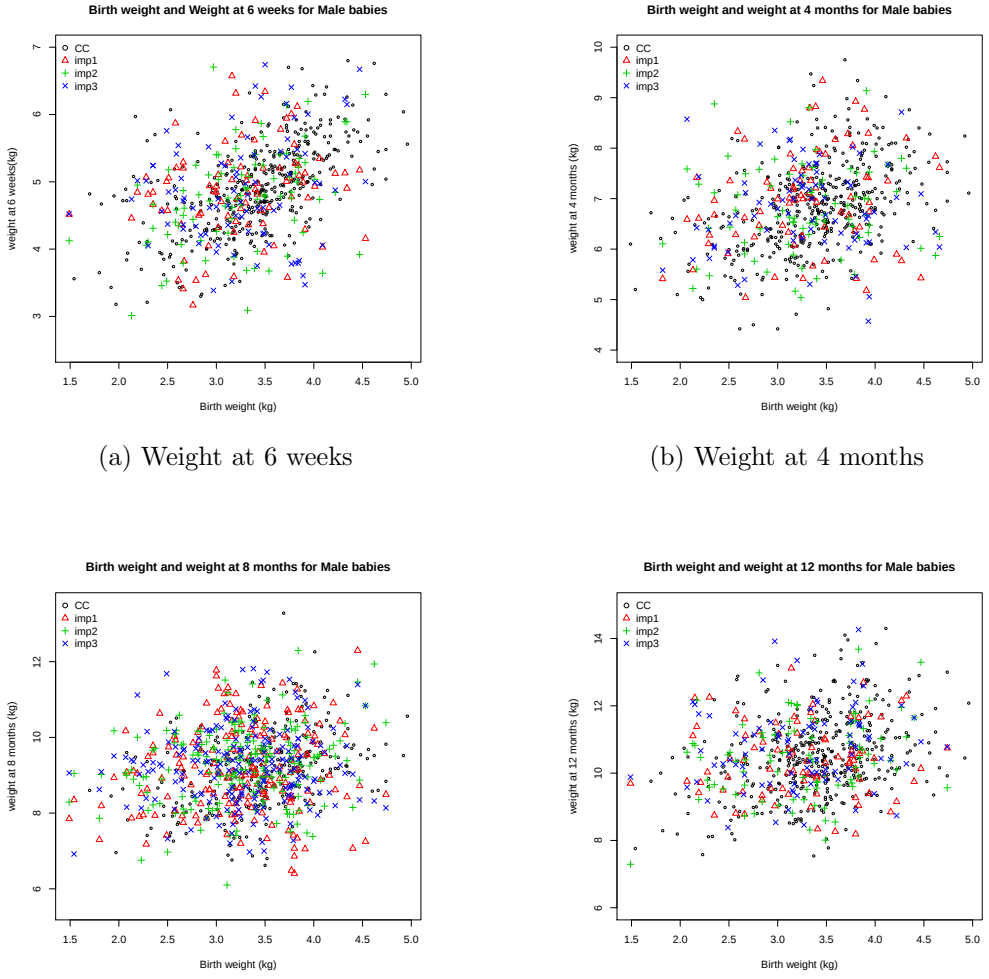


(a) Non-overlapping and restrictive strategy

(b) Inclusive strategy

Figure 6.6: Distribution of imputed values for weight at school entry for male children using non-overlapping, restrictive and inclusive strategies using multiple imputation

Figure 6.7 shows the distribution of imputed values for first year baby's weights for male babies using non-overlapping variable sets and restrictive strategy where both strategies use the same imputation model. The weight at 6 weeks, weight at 4 months, weight 8 months and weight 12 months are imputed using gestational age. The distribution of imputed values for weight at 6 weeks, weight at 4 months, weight 8 months and weight 12 months using using non-overlapping variable sets and restrictive strategy are closed to observed values.



(a) Weight at 6 weeks

(b) Weight at 4 months

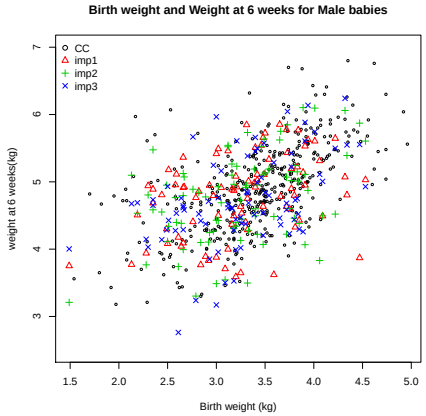
(c) Weight at 8 months

(d) Weight at 12 months

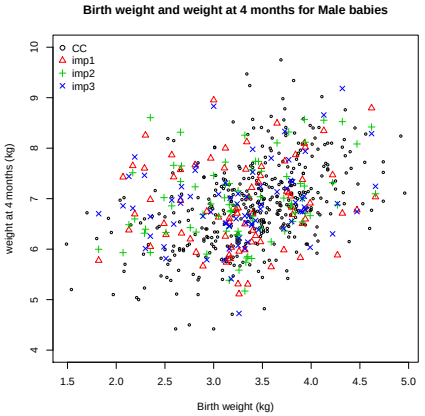
Figure 6.7: Distribution of imputed values for first year baby's weights for male babies using non-overlapping and restrictive strategy using multiple imputation

Figure 6.8 shows the distribution of imputed values for first year baby's weights for male babies using inclusive strategy. Each first year baby's weight is imputed using the rest of first year baby's weights and gestational age. The distribution of imputed values for weight at 6 weeks, weight at 4 months, weight 8 months and weight 12 months

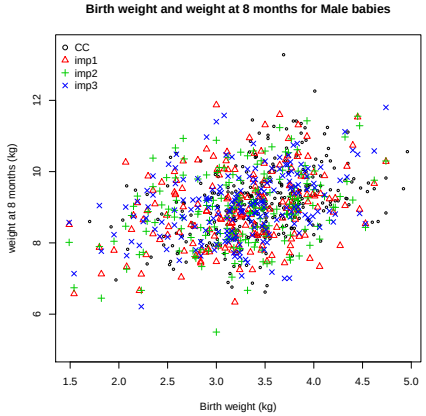
using inclusive strategy are closer to observed values. Although the imputed values of weight at 6 weeks, weight at 4 months, weight 8 months and weight 12 months using non-overlapping variable sets and restrictive strategy are closed to observed values, but the distribution of imputed values using inclusive strategy are better and more centered to observed values.



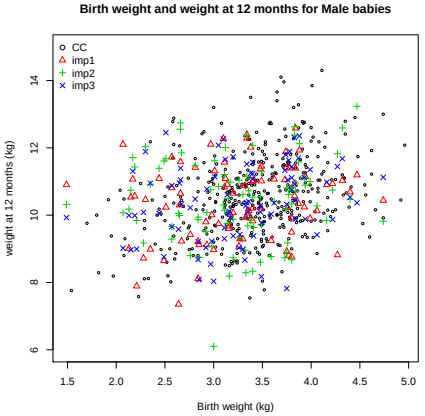
(a) Weight at 6 weeks



(b) Weight at 4 months



(c) Weight at 8 months



(d) Weight at 12 months

Figure 6.8: Distribution of imputed values for first year baby's weights for male babies using inclusive strategy using multiple imputation

Table 6.13 shows the parameter estimates and MSE(P) for prediction of weight at school entry for male children. The average MSE(P)-CC is the lowest for STACK using inclusive strategy for prediction model. Two different sets of 10% cross-validation test (CV1 and CV2) were used to calculate MSE(P). Since there are not many outliers for weight at school entry for male children, there is not much difference between MSE(P) for both cross validation sets. The results showed that MSE(P) values for STACK using

inclusive strategy for prediction model and model chosen by AIC_c is the lowest. The MSE(P) values for cross-validation test are higher than the MSE(P)-CC. The factors that contribute to predict weight at school entry for male children are birth weight, weight at 6 weeks, weight at 4 month, weight at 8 months and weight at 12 months. If birth weight increase by 1 kg, the weight at school entry will increase by 0.7239kg. There is a negative relationship of weight at 4 month in predicting weight at school entry for male children.

Table 6.13: Estimates and MSE(P) for prediction of weight at school entry for male children using multiple imputation

Approaches	**Model averaging**						**STACK**					
Strategies	**Non-over**		**Restrictive**		**Inclusive**		**Non-over**		**Restrictive**		**Inclusive**	
Model selection Criterion	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**
Constant	5.5490	5.8413	4.0514	4.7168	3.3628	3.2413	5.0202	4.8955	4.0173	5.2030	2.5159	2.5159
Birth weight	0.6356	0.7073	0.5304	0.6005	0.9377	0.9413	0.6477	0.6225	0.4496	0.5067	0.7239	0.7239
Weight at 6 weeks	0.3518	0.4256	0.2263	0.2699	0.1872	0.1725	0.3228	0.2690	0.3956	0.3446	0.3167	0.3167
Weight at 4 months	0.0155	0.0708	-0.1094	-0.0690	-0.5542	-0.4895	-0.1347	0	-0.1055	0	-0.6070	-0.6070
Weight at 8 months	0.5101	0.5305	0.6710	0.6851	0.7491	0.7009	0.6105	0.5860	0.5449	0.5291	0.9294	0.9294
Weight at 12 months	0.6267	0.6519	0.5645	0.5707	0.9984	1.0556	0.6107	0.5892	0.6397	0.6205	0.8736	0.8736
Gestational Age	-	-	0.0983	0.1263	-0.0106	-0.0079	-	-	0.0383	0	0	0
Average MSE(P)-CC	4.7656	6.9862	4.5459	9.9704	5.5908	8.8505	4.4322	4.4322	4.4583	4.4583	**4.2361**	**4.2361**
Average MSE(P)-CV1	7.9665	17.0231	14.5525	38.5906	6.6376	8.4025	5.5643	5.6220	5.8316	5.7821	**5.4056**	**5.4056**
Average MSE(P)-CV2	7.5230	14.1765	9.4636	23.3386	5.3800	6.6351	5.6279	5.6279	5.7187	5.6186	**5.3996**	**5.3996**

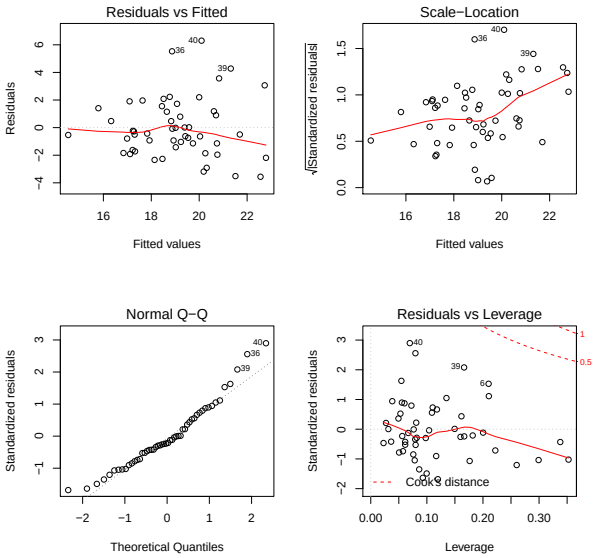


Figure 6.9: Residuals for male children using inclusive strategy and model selection criterion AIC_c using multiple imputation for prediction of weight at school entry

Figure 6.9 shows the residuals based on CV1 for male children using inclusive strategy and model selection criterion AIC_c. It indicates that the the spread of the residuals is increasing as the fitted values changes, which is called heteroskedasticity. Since the

deviation deviations from the straight line in normal Q-Q are minimal, this indicates that residuals are approximately normally distributed.

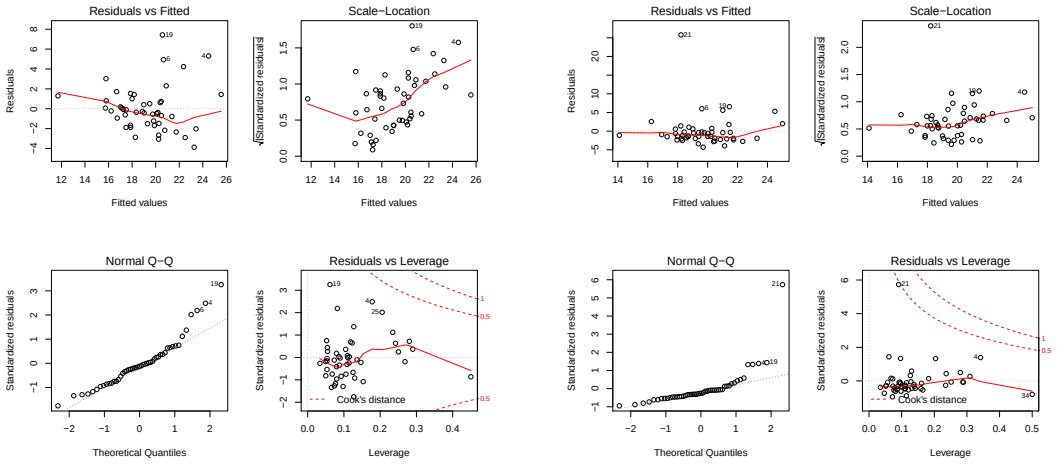
Table 6.14: Estimates and MSE(P) for prediction of weight at school entry for female children using multiple imputation

Approaches	**Model averaging**						**STACK**					
Strategies	**Non-over**		**Restrictive**		**Inclusive**		**Non-over**		**Restrictive**		**Inclusive**	
Model selection Criterion	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**
Constant	4.9262	5.6559	9.8042	7.0530	8.7603	5.6214	4.4595	4.5531	12.5232	12.5232	10.1894	10.1894
Birth weight	-0.1382	-0.0192	0.3554	0.3372	0.7818	0.6952	-0.2378	0	0.2924	0.2924	0.6978	0.6978
Weight at 6 weeks	0.5273	0.5704	0.6552	0.7119	-0.1895	-0.0286	0.4714	0.4281	0.3956	0.3956	0	0
Weight at 4 months	0.2604	0.3337	0.2178	0.3232	0.7278	0.7235	0.1539	0	0.2417	0.2417	0.5626	0.5626
Weight at 8 months	0.2826	0.3222	0.4215	0.4455	-1.0221	-1.0027	0.3656	0.3931	0.3836	0.3836	-1.0410	-1.0410
Weight at 12 months	1.1017	1.1812	1.0237	1.1074	2.3109	2.2893	1.0145	1.0211	0.9814	0.9814	2.3237	2.3237
Gestational Age	-	-	-0.2030	-0.1707	-0.2500	-0.2241	-	-	-0.2496	-0.2496	-0.2635	-0.2635
Average MSE(P)-CC	12.0432	33.7168	8.3389	12.4550	7.1680	10.8208	8.0525	8.0525	7.7610	7.7609	**6.8760**	**6.8760**
Average MSE(P)-CV1	13.9910	40.1831	9.8000	12.1061	**5.9506**	6.9151	6.4515	6.4321	6.2137	6.2137	6.4343	6.4343
Average MSE(P)-CV2	25.5326	45.6606	23.1714	27.2795	20.9613	21.8894	19.8003	19.8219	19.6686	**19.6550**	20.8172	20.7851

Table 6.14 shows the estimates and MSE(P) for prediction of weight at school entry for female children. Two different sets of 10% complete cases (CV1 and CV2) were used to calculate MSE(P). Since there is an outlier (heavy weight child) in CV2, the MSE(P) for dataset CV2 is much higher than CV1. This shows that heavy weight child (there is a overweight child in CV2 compared to to other child in that dataset) affect the prediction of weight at school entry for female children. For CC, the results showed that MSE(P) values for STACK using inclusive strategy for prediction model and the model chosen by AICc is the lowest. Whereas for CV1, the results showed that MSE(P) values for model averaging using inclusive strategy for prediction model is the lowest. For CV2, the MSE(P) values for STACK using restrictive strategy for prediction model is the lowest. The MSE(P) values for cross-validation test, CV2 is higher than the MSE(P) for CC. The factors that contribute to predict weight at school entry for male children are birth weight, weight at 6 weeks, weight at 4 month, weight at 8 months, weight at 12 months and gestational age. There is a strong relationship between weight at 12 months and weight at school entry for female children. If weight at 12 months increase by 1 kg, the weight at school entry will increase by 2.3109kg. There is a negative relationship of weight at 8 month and gestational age in predicting weight at school entry for female children.

Figure 6.10a shows the residuals based on CV1 for female children using inclusive strategy and model selection criterion AIC_c. It indicates that the the spread of the residuals is increasing as the fitted values changes, which is called as heteroskedasticity. Since the deviations from the straight line in normal Q-Q are minimal, this indicates that residuals for prediction on weight at school entry for female children are normally distributed. Figure 6.10b shows the residuals based on CV2 for female children using inclusive strategy

and model selection criterion AIC_c. It indicates that there is effect of heteroskedasticity for residuals based on CV2. The normal Q-Q shows that the residuals based on CV2 are heavy-tailed and not normal. This is due to the effects of outlier (heavy weight children) in the CV2 dataset.



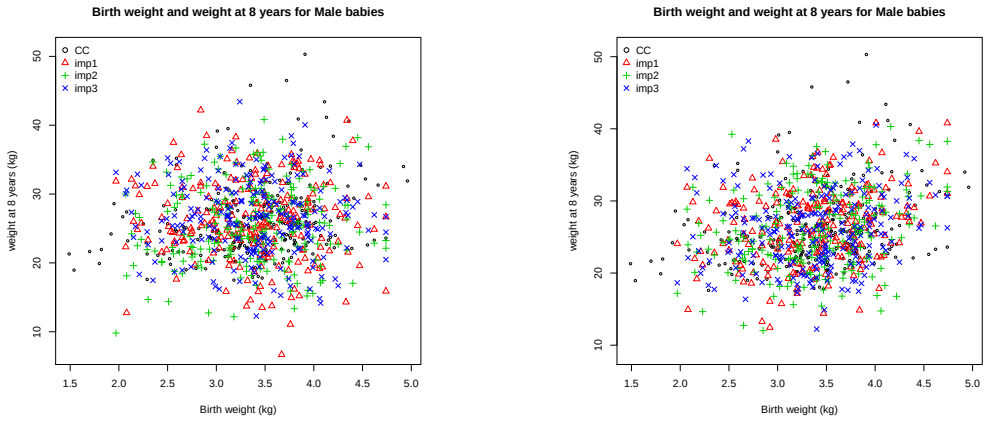
(a) residuals based on CV1

(b) residuals based on CV2

Figure 6.10: Residuals for female children using inclusive strategy and model selection criterion AIC_c using multiple imputation for prediction of weight at school entry

6.2.3 Prediction of weight at eight years using multiple imputation

Since the imputed values for first year baby's weights (6 weeks, weight at 4 months, weight 8 months and weight 12 months) for both male and female babies using non-overlapping variable sets, restrictive and inclusive strategies are similar as discussed in Section 6.2.2, the distribution of imputed values for first year baby's weights are not discussed in this section. Figure 6.11 shows the distribution of imputed values for weight at eight years for male children using non-overlapping variable sets, restrictive and inclusive strategies. It is clearly showed that the distribution of imputed values for weight at eight years using inclusive strategy are better than using non-overlapping variable sets and restrictive strategy. The distribution of imputed values for weight at eight years for male children using inclusive strategy are centered, overlapping and closer to observed values compared to non-overlapping variable sets and restrictive strategy.



(a) Non-overlapping and restrictive strategy

(b) Inclusive strategy

Figure 6.11: Distribution of imputed values for weight at eight years for male children using non-overlapping, restrictive and inclusive strategies using multiple imputation

Table 6.15: Estimates and MSE(P) for prediction of weight at eight years for male children using multiple imputation

Approaches	**Model averaging**						**STACK**					
Strategies	**Non-over**		**Restrictive**		**Inclusive**		**Non-over**		**Restrictive**		**Inclusive**	
Model selection Criterion	**AIC$_c$**	**BIC**	**AIC$_c$**	**BIC**	**AIC$_c$**	**BIC**	**AIC$_c$**	**BIC**	**AIC$_c$**	**BIC**	**AIC$_c$**	**BIC**
Constant	6.8881	7.7402	5.9286	6.7875	-1.2950	-1.0733	6.7906	6.7906	5.9562	5.9562	1.7115	-1.8052
Birth weight	0.4186	0.4999	0.4971	0.5829	0.6952	0.7340	0.5797	0.5797	0.3976	0.3976	0.8671	0.6456
Weight at 6 weeks	0.5879	0.6600	0.5570	0.6609	0.8028	0.8062	0.3424	0.3424	0.6539	0.6539	1.0448	1.0984
Weight at 4 months	-0.0997	0.0260	-0.0874	0.0264	-0.9263	-0.7977	0	0	0	0	-0.7376	-0.7327
Weight at 8 months	1.0100	1.1118	0.9866	1.0402	1.4481	1.3813	0.9702	0.9702	0.9284	0.9284	1.1052	1.0885
Weight at 12 months	0.7919	0.8361	0.8496	0.8999	1.4112	1.4876	0.6475	0.6475	0.6832	0.6832	1.4660	1.4754
Gestational Age	-	-	0.0199	0.0617	0.0220	0.0501	-	-	0	0	-0.1024	0
Average MSE(P)-CC	30.8166	116.2432	22.0900	45.2881	23.2989	25.2626	21.4420	21.4435	21.1440	21.1440	**20.7578**	**20.7578**
Average MSE(P)-CV	**27.2317**	40.2609	27.3428	56.9328	27.3811	29.6481	31.8134	31.8134	30.9464	30.9464	30.4445	30.7693

Table 6.15 shows the estimates and MSE(P) for prediction of weight at eight years for male children. The results showed that MSE(P) values for model averaging using non-overlapping variable sets for prediction model and model chosen by AIC$_c$ is the lowest. The factors that contribute to predict weight at eight years for male babies are birth weight, weight at 6 weeks, weight at 4 month, weight at 8 months and weight at 12 months. There are positive effects of birth weight, weight at 6 weeks, weight at 8 months and weight at 12 months on prediction of weight at eight years. If weight at 12 months increase by 1 kg, the weight at eight years will increase by 0.7919kg. There is a negative relationship of weight at 4 month and gestational age in predicting weight at eight years for male children. Figure 6.12 shows the residuals for male children using non-overlapping variable sets and model selection criterion, AIC$_c$. It indicates that the the spread of the residuals are symmetrically distributed and tending to cluster towards

the middle of the plot. The normal Q-Q shows that the residuals are heavy-tailed and not normal.

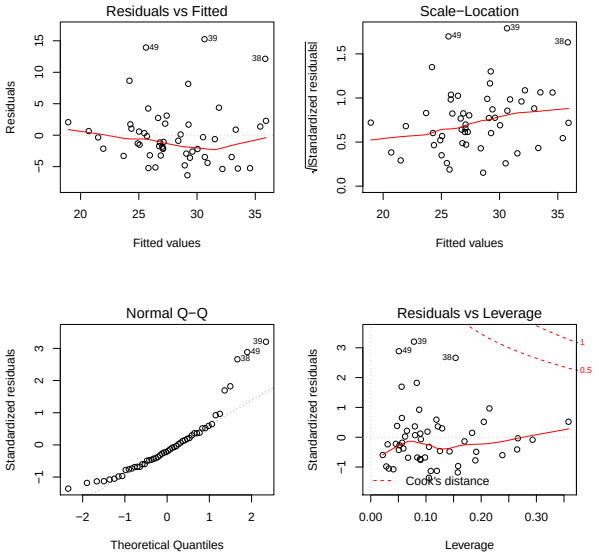
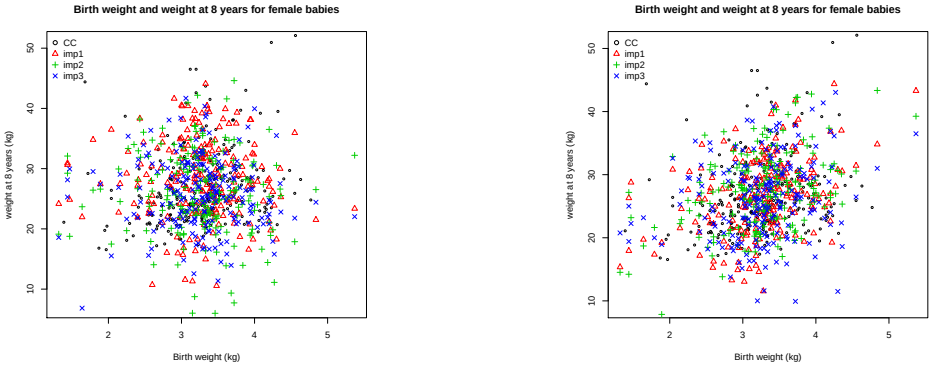


Figure 6.12: Residuals for male children using non-overlapping variable sets and model selection criterion, AIC_c using multiple imputation for prediction of weight at eight years



(a) Non-overlapping and restrictive strategy

(b) Inclusive strategy

Figure 6.13: Distribution of imputed values for weight at eight years for female children using non-overlapping, restrictive and inclusive strategy using multiple imputation

Figure 6.13 shows the distribution of imputed values for weight at eight years for female children using non-overlapping variable sets, restrictive and inclusive strategies. It is clearly showed that the distribution of imputed values for weight at eight years using inclusive strategy are better than using non-overlapping variable sets and restrictive strategy. The distribution of imputed values for weight at eight years for female

children using inclusive strategy are centered, overlapping and closer to observed values compared to non-overlapping variable sets and restrictive strategy. Therefore, the inclusive strategy (includes all four first year baby's weights and gestational age) is the best strategy for imputing missing values.

Table 6.16 shows the estimates and MSE(P) for prediction of weight at eight years for female children. The results showed that MSE(P) values for STACK using restrictive strategy for prediction model and model chosen by BIC is the lowest. The factors that contribute to predict weight at eight years for female children are birth weight, weight at 6 weeks, weight at 4 months, weight at 8 months, weight at 12 months and gestational age. There are positive relationship of birth weight, weight at 6 weeks, weight at 8 months and weight at 12 months on prediction of weight at eight years for female children. If weight at 12 months increase by 1 kg, the weight at eight years will increase by 1.2479kg. There is a negative relationship of weight at 4 month and gestational age in predicting weight at eight years for female children.

Table 6.16: Estimates and MSE(P) for prediction of weight at eight years for female children using multiple imputation

Approaches	**Model averaging**						**STACK**					
Strategies	**Non-over**		**Restrictive**		**Inclusive**		**Non-over**		**Restrictive**		**Inclusive**	
Model selection Criterion	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**
Constant	8.5121	9.6738	19.1958	17.5806	12.2690	9.6155	8.0925	8.0925	17.1399	16.9978	11.5053	11.5053
Birth weight	-0.2853	-0.1351	0.3096	0.1646	0.1315	0.1923	-0.6437	-0.6437	0	0	0	0
Weight at 6 weeks	0.7520	0.7882	0.6412	0.6573	1.0926	0.9761	1.0156	1.0156	0.8161	0.7133	1.5645	1.5645
Weight at 4 months	0.2091	0.3088	-0.3119	-0.1707	-0.2827	-0.1727	0	0	-0.2169	0	-0.4572	-0.4572
Weight at 8 months	0.6405	0.7021	0.6651	0.6896	-1.3961	-1.3904	0.5712	0.5712	0.6097	0.5743	-1.4758	-1.4758
Weight at 12 months	1.2326	1.3334	1.1826	1.2479	3.4942	3.4505	1.1504	1.1504	1.2447	1.2040	3.8219	3.8219
Gestational Age	-	-	-0.3110	-0.2648	-0.3485	-0.3247	-	-	-0.2543	-0.2559	-0.3628	-0.3628
Average MSE(P)-CC	28.1563	55.5994	22.4580	23.8397	21.9140	39.6080	22.9084	22.9084	22.4732	22.6426	**20.5759**	**20.5759**
Average MSE(P)-CV	23.7122	61.7798	17.3017	**16.0643**	23.5641	35.4340	16.6506	16.6506	16.2176	16.1085	17.2490	17.2490

Figure 6.14 shows the residuals for female children using restrictive strategy and model selection criterion, BIC. It indicates that the the spread of the residuals are symmetrically distributed and tending to cluster towards the middle of the plot. However, there are some outliers. Since the deviations from the straight line in normal Q-Q are minimal, this indicates that residuals are approximately normally distributed.

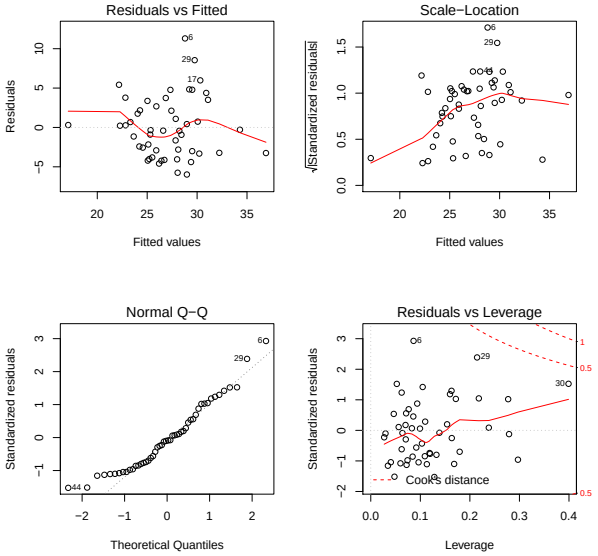
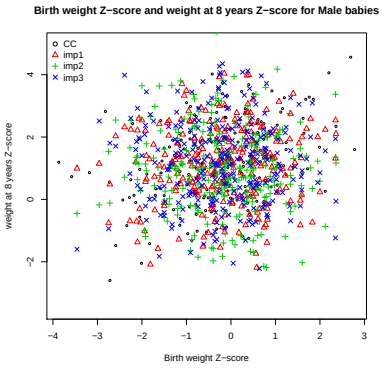


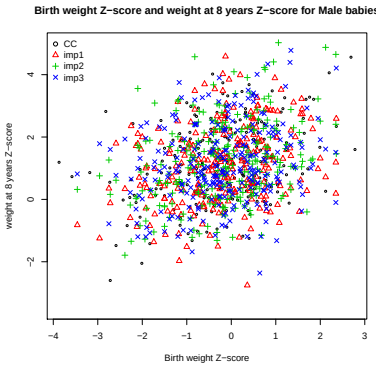
Figure 6.14: Residuals for female children using restrictive strategy and model selection criterion, BIC using multiple imputation for prediction of weight at eight years

6.2.4 Prediction of weight Z-scores at eight years using multiple imputation

A similar analysis on prediction of weight at eight years as discussed in Section 6.2.3 was carried out using Z-scores for all weights. Figure 6.15 shows the distribution of imputed values for weight Z-scores at eight years for male children using non-overlapping variable sets, restrictive and inclusive strategies. It is clearly showed that the distribution of imputed values for weight Z-scores at eight years using inclusive strategy are better than using non-overlapping variable sets and restrictive strategy. The distribution of imputed values for weight Z-scores at eight years using inclusive strategy are centered and closer to observed values compared to non-overlapping variable sets and restrictive strategy.

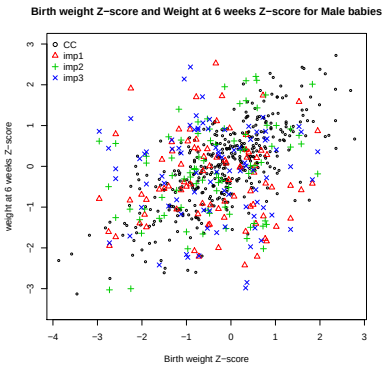


(a) Non-overlapping and restrictive strategy

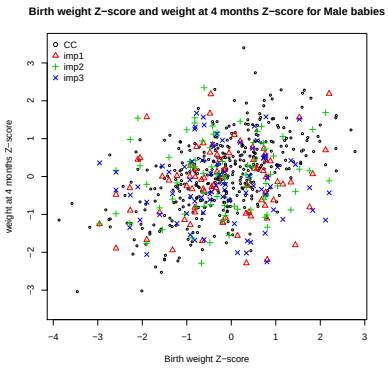


(b) Inclusive strategy

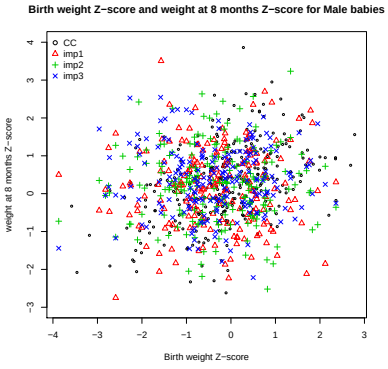
Figure 6.15: Distribution of imputed values for weight Z-scores at eight years for male children using non-overlapping, restrictive and inclusive strategies using multiple imputation



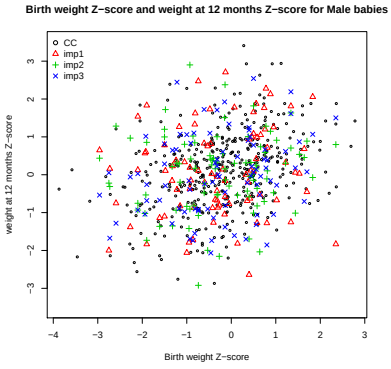
(a) Weight at 6 weeks Z-scores



(b) Weight at 4 months Z-scores



(c) Weight at 8 months Z-scores

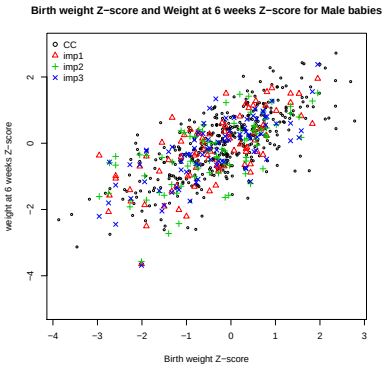


(d) Weight at 12 months Z-scores

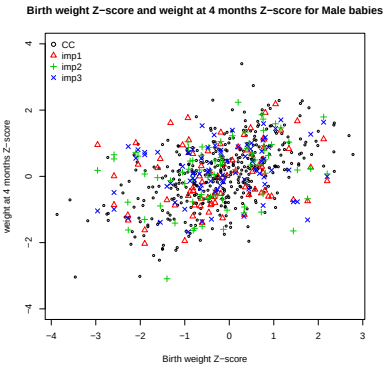
Figure 6.16: Distribution of imputed values for first year baby's weight Z-scores for male babies using non-overlapping and restrictive strategy using multiple imputation

Figure 6.16 shows the distribution of imputed values for first year baby's weight Z-scores for male children using non-overlapping variable sets and restrictive strategy. The distribution of imputed values for first year baby's weight Z-scores using using non-overlapping variable sets and restrictive strategy are closer to observed values but a bit scattered.

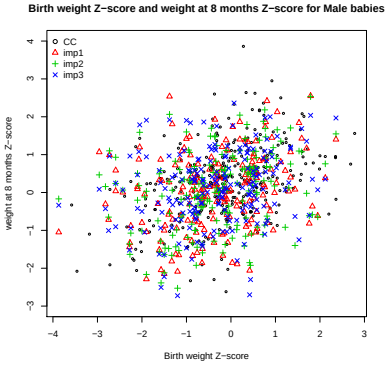
Figure 6.17 shows the distribution of imputed values for first year baby's weight Z-scores for male babies using inclusive strategy. The distribution of imputed values for first year baby's weight Z-scores using using inclusive strategy are closer to observed values and centered. The distribution of imputed values using inclusive strategy are better than the distribution of imputed values using non-overlapping variable sets and restrictive strategy.



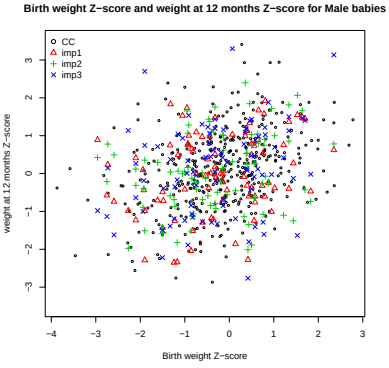
(a) Weight Z-scores at 6 weeks



(b) Weight Z-scores at 4 months



(c) Weight Z-scores at 8 months



(d) Weight Z-scores at 12 months

Figure 6.17: Distribution of imputed values for first year baby's weight Z-scores for male babies using inclusive strategy using multiple imputation

Table 6.17: Estimates and MSE(P) for prediction of weight Z-scores at eight years for male children using multiple imputation

Approaches	**Model averaging**						**STACK**					
Strategies	**Non-over**		**Restrictive**		**Inclusive**		**Non-over**		**Restrictive**		**Inclusive**	
Model selection Criterion	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**
Constant	1.1505	1.1554	1.8182	1.4598	2.0185	1.4554	1.1445	1.1503	2.0192	1.1503	2.9200	2.9200
Birth weight	0.1521	0.1724	0.1554	0.1647	0.3198	0.3462	0.1573	0.1570	0.1568	0.1478	0.2937	0.2937
Weight at 6 weeks	0.0602	0.1123	0.0321	0.0700	0.1580	0.1403	0	0	0	0	0.1898	0.1898
Weight at 4 months	0.0831	0.1218	0.0274	0.0574	-0.0958	-0.0046	0.0825	0.0809	0.0582	0.0554	-0.2012	-0.2012
Weight at 8 months	0.1405	0.1842	0.0943	0.0989	0.1598	0.1031	0.0379	0	0.0550	0.0538	0.2999	0.2999
Weight at 12 months	-0.0028	0.0257	-0.0136	0.0071	-0.0589	0.0115	-0.0373	0	0	0	-0.1445	-0.1445
Gestational Age	-	-	-0.0215	-0.0167	-0.0387	-0.0383	-	-	-0.0221	0	-0.0434	-0.0434
Average MSE(P)-CC	**1.3717**	1.4014	1.6703	2.6733	1.4297	2.6842	1.4124	1.4124	1.4211	1.4058	1.4814	1.5383
Average MSE(P)-CV	**1.7148**	1.7908	1.6737	1.7384	2.0937	3.0264	1.7656	1.7451	1.7358	1.7542	1.8441	1.8441

Table 6.17 shows the estimates and MSE(P) for prediction of weight Z-scores at eight years for male children. The results showed that MSE(P) values for model averaging using non-overlapping variable sets for prediction model and model chosen by AIC_c is the lowest. The factors that contribute to predict weight Z-scores at eight years for male children are birth weight Z-scores, weight Z-scores at 6 weeks, weight Z-scores at 4 months, weight Z-scores at 8 months and weight Z-scores at 12 months. If birth weight Z-scores increase by 1 unit, the weight Z-scores at eight years will increase by 0.1521 unit. There is a negative relationship between weight Z-scores at 12 months and weight Z-scores at eight years for male children.

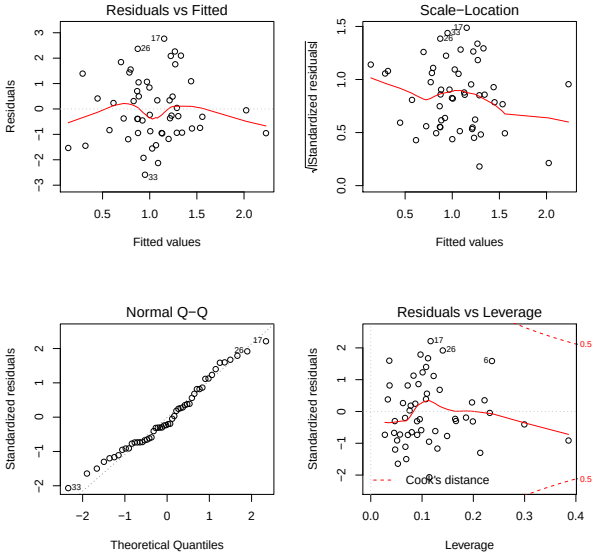


Figure 6.18: Residuals for male children using non-overlapping variables sets and model selection criterion, AIC_c using multiple imputation for prediction of weight Z-scores at eight years

Figure 6.18 shows the residuals for male children using non-overlapping variable sets and model selection criterion, AIC_c. It indicates that the the spread of the residuals are symmetrically distributed and tending to cluster towards the middle of the plot. Since the deviation deviations from the straight line in normal Q-Q are minimal, this indicates that residuals are approximately normally distributed.

Table 6.18: Estimates and MSE(P) for prediction of weight Z-scores at eight years for female children using multiple imputation

Approaches	**Model averaging**						**STACK**					
Strategies	**Non-over**		**Restrictive**		**Inclusive**		**Non-over**		**Restrictive**		**Inclusive**	
Model selection Criterion	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**	**AIC_c**	**BIC**
Constant	1.0186	1.0177	0.7262	0.8601	2.4371	1.5751	1.0540	1.0643	0.9545	0.9630	3.3093	3.3093
Birth weight	0.0348	0.0655	0.0575	0.0815	-0.0625	-0.0447	0.0963	0.0965	0	0	-0.0833	-0.0833
Weight at 6 weeks	0.1606	0.1770	0.1949	0.2160	0.7180	0.6100	0.1365	0.1265	0.2083	0.1856	0.8482	0.8482
Weight at 4 months	-0.0121	0.0431	0.0142	0.0643	-0.5694	-0.5184	-0.0571	0	-0.0755	0	-0.6462	-0.6462
Weight at 8 months	0.1271	0.1591	0.0738	0.0909	0.3252	0.3400	0.0853	0.0915	0.0832	0.0780	0.4048	0.4048
Weight at 12 months	0.0696	0.0972	0.0383	0.0581	0.1940	0.2337	0.0647	0	0.0514	0	0	0
Gestational Age	-	-	0.0140	0.0202	-0.0536	-0.0514	-	-	0	0	-0.0578	-0.0578
Average MSE(P)-CC	1.7609	**1.7522**	1.8097	2.3136	2.0447	2.5465	1.8205	1.8549	1.8122	1.8551	1.8992	1.8629
Average MSE(P)-CV	2.1742	2.2320	2.2083	2.5337	2.9279	4.7631	**2.1569**	2.1683	2.1799	2.1747	2.5328	2.5328

Table 6.18 shows the estimates and MSE(P) for prediction of weight Z-scores at eight years for female children. The results showed that MSE(P) values for STACK using non-overlapping variable sets for prediction model and model chosen by AIC_c is the lowest. The factors that contribute to predict weight Z-scores at eight years for female children are birth weight Z-scores, weight Z-scores at 6 weeks, weight Z-scores at 4 months, weight Z-scores at 8 months and weight Z-scores at 12 months. There are positive effects of birth weight Z-scores, weight Z-scores at 6 weeks, weight Z-scores at 8 months and weight Z-scores at 12 months on prediction of weight Z-scores at eight years for female children. There is a negative relationship between weight Z-scores at 4 months and weight Z-scores at eight years for female children. If weight Z-scores at 6 weeks increase by 1 unit, the weight Z-scores at eight years will increase by 0.1365 unit.

Figure 6.19 shows the residuals for female children using non-overlapping variable set and model selection criterion, AIC_c. It indicates that the the spread of the residuals are symmetrically distributed and tending to cluster towards the middle of the plot. However, there is a outlier. Since the deviation deviations from the straight line in normal Q-Q are minimal, this indicates that residuals are approximately normally distributed.

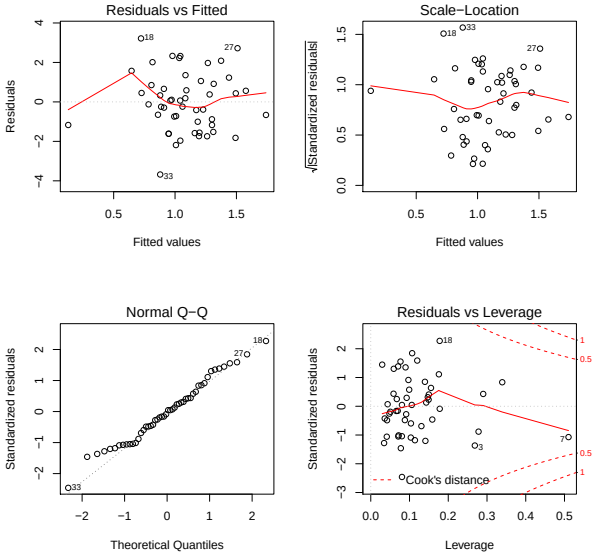


Figure 6.19: Residuals for female children using non-overlapping variable set and model selection criterion, AIC_c using multiple imputation for prediction of weight Z-scores at eight years

6.3 Gateshead Millennium Study Simulation Results

The results from the real-data analysis suggest that, in this study, the best approach was STACK with an inclusive model-building strategy. The conclusions from the simulation studies in Chapter 4 and Chapter 5 were different, indicating that model averaging with an inclusive strategy was best, or possibly STACK with non-overlapping variable sets. Hence, in this section, a simulation study was carried out to identify the reasons for these contradictory results. This simulation study was based on the simulation design discussed in Chapter 4. The aim was to mimic the conditions for predicting weight at school entry using the parameter values in Table 6.12. The analysis was carried out for a sample size $n = 500$, error variance $\sigma^2_\varepsilon = 16$ and various missing percentages ($m = 0, 10, 25, 26, 28, 30$ and 40). There were five covariates (X_1, X_2, X_3, X_4 and X_5) and one auxiliary variable (X_6). The covariance matrix of $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5, X_6)$ in the simulation study was

$$\Sigma = \begin{pmatrix} 1 & \rho_{12} & (\rho_{12})^2 & (\rho_{12})^3 & (\rho_{12})^4 & \rho_{16} \\ \rho_{12} & 1 & \rho_{12} & (\rho_{12})^2 & (\rho_{12})^3 & \rho_{26} \\ (\rho_{12})^2 & \rho_{12} & 1 & \rho_{12} & (\rho_{12})^2 & \rho_{36} \\ (\rho_{12})^3 & (\rho_{12})^2 & \rho_{12} & 1 & \rho_{12} & \rho_{46} \\ (\rho_{12})^4 & (\rho_{12})^3 & (\rho_{12})^2 & \rho_{12} & 1 & \rho_{56} \\ \rho_{16} & \rho_{26} & \rho_{36} & \rho_{46} & \rho_{56} & 1 \end{pmatrix} \tag{6.1}$$

where $\rho_{ij} = \rho_{ji}$ denotes the correlation between X_i and X_j. The correlations in simulation study 1 are $\rho_{12} = 0.75$ and $\rho_{i6} = 0.1$ for all $i = 1, 2, 3, 4, 5$. Whereas the correlations in simulation study 2 are $\rho_{12} = 0.75$ and $\rho_{i6} = 0.5$. All the model-building strategies discussed previously were investigated in this study. It was decided to use AIC_c as the criterion because AIC_c performs better than BIC in terms of model selection and prediction as concluded in Chapter 4 and Chapter 5.

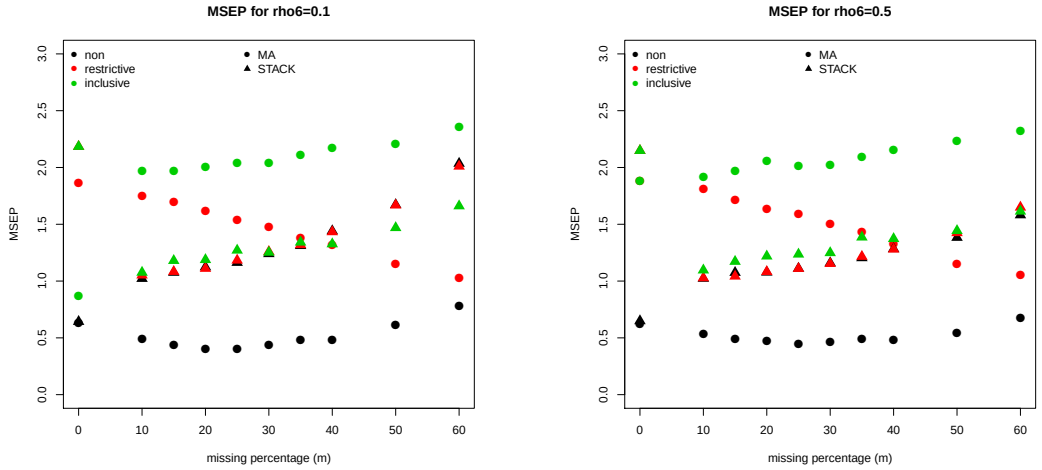
Table 6.19: MSE(P) for prediction of weight at school entry (GMS simulation) for $\sigma_\varepsilon = 4$ and $n = 500$ using multiple imputation

correlation	**Simulation Study 1 - $\rho_{i6} = 0.1$**						Simulation Study 2 - $\rho_{i6} = 0.5$					
Strategies	**Non-over**		**Restrictive**		**Inclusive**		**Non-over**		**Restrictive**		**Inclusive**	
Approaches	**MA**	**STACK**	**MA**	**STACK**	**MA**	**STACK**	**MA**	**STACK**	**MA**	**STACK**	**MA**	**STACK**
m=0	0.6315	0.6432	1.8686	2.1842	1.8686	2.1842	0.6235	0.6494	1.8819	2.1486	1.8819	2.1486
m=10	0.4882	1.0217	1.7513	1.0518	1.9726	1.0761	0.5362	1.0230	1.8090	1.0276	1.9198	1.0962
m=15	0.4397	1.0756	1.6964	1.0832	1.9702	1.1799	0.4928	1.0748	1.7159	1.0423	1.9725	1.1721
m=20	0.4044	1.1238	1.6145	1.1105	2.0058	1.1879	0.4730	1.0772	1.6314	1.0824	2.0581	1.2191
m=25	0.4061	1.1628	1.5419	1.1831	2.0397	1.2710	0.4504	1.1114	1.5895	1.1097	2.0138	1.2360
m=30	0.4381	1.2400	1.4778	1.2588	2.0374	1.2524	0.4688	1.1595	1.5077	1.1534	2.0207	1.2487
m=35	0.4789	1.3114	1.3834	1.3195	2.1147	1.3417	0.4927	1.2041	1.4348	1.2160	2.0945	1.3861
m=40	0.4839	1.4417	1.3206	1.4329	2.1742	1.3265	0.4853	1.2841	1.3264	1.2801	2.1523	1.3728
m=50	0.6106	1.6722	1.1494	1.6681	2.2064	1.4693	0.5433	1.3833	1.1505	1.1248	2.2347	1.4432
m=60	0.7852	2.0353	1.0315	2.0103	2.3601	1.6608	0.6766	1.5817	1.0585	1.6495	2.3218	1.6143

Table 6.19 shows the MSE(P) for prediction of weight at school entry (GMS simulation)for $\sigma_\varepsilon = 4$ and $n = 500$. Figure **??** and Figure **??** show the MSE(P) for model averaging and STACK via AIC_c using non-overlapping variable sets, restrictive and inclusive strategies for each ρ_{i6} ($\rho_{i6} = 0.1$ and $\rho_{i6} = 0.5$). The results suggest that model averaging with the non-overlapping strategy is the best method when an auxiliary variable is available. Model averaging using non-overlapping variable sets performs better than STACK using all three model-building strategies where the MSE(P) for model averaging using non-overlapping variable sets is the lowest. The MSE(P) for model averaging using restrictive strategy decreases as missing percentage increases, in contrast to the MSE(P) for model averaging and STACK using inclusive strategy which are increases as missing percentage increases.

In the simulation settings of Chapter 5, there is no correlation between the covariates but there is a correlation with the auxiliary variable. Whereas in the real-data analysis, there are moderate to high correlations between covariates and low correlations with the auxiliary variable. This appears to go some way towards explaining the contradictory results between Chapter 5 and real-data analysis (GMS). In real-data analysis, the best method is STACK using an inclusive strategy. This coincides reasonably well with the results of the simulation study discussed in this section given that the correlations with the auxiliary variable (gestational age) were low and the percentages of missing data

moderately high for some of the covariates in the real-life dataset. These results indicate that the correlation with covariates and missing percentages all play an important role in determining the best method and the best model-building strategy for prediction. There are no effects of correlation between auxiliary variable in terms of prediction as missing percentage increases and also between the three model-building strategies.



(a) MSEP for $\rho_{i6} = 0.1$

(b) MSEP for $\rho_{i6} = 0.5$

Figure 6.20: MSEP for model averaging and STACK via AIC_c using non-overlapping, restrictive and inclusive strategies for GMS simulation using multiple imputation

6.4 Discussion and Conclusions

The real-data analysis using GMS dataset was carried out to investigate the performance of the proposed methods and model-building strategies. Only the AIC_c criterion has been discussed in this chapter. As discussed and concluded in Chapter 4 and Chapter 5, AIC_c performs better than BIC in both model averaging and STACK. BIC performs very poorly for model averaging where the MSE(P) values for model averaging based on BIC weights are very much higher compared to those based on AIC_c weights. This is due to the effect of BIC's penalty term and also the effect of highly correlated covariates. In addition, BIC tends to give higher weights for smaller models to reduce the effects of highly correlated covariates. On the other hand, AIC_c tends to incorporate the effects of highly correlated covariates by choosing a larger model.

On the basis of the simulation results from Chapter 5, using highly correlated covariates in the imputation model can be expected to improve the imputation step and hence the overall analysis of a dataset with missing values. This coincides with the results

obtained in Chapter 6 by using an inclusive strategy with highly correlated covariates. However, the choice of model-building strategy in Chapter 4 and Chapter 5 was either non-overlapping variable sets or the inclusive strategy for prediction. This is because the correlation between the auxiliary variable and the variable to be imputed is high in Chapter 4 and Chapter 5, and there are no correlations between the covariates in the simulation settings there.

The inclusive strategy performs better in choosing the best model for prediction of weight at school entry and weight at eight years. This shows that both imputation and prediction models are interrelated. If the imputation model is misspecified, then the prediction analysis will be poorer. The inclusive strategy performs better for imputing the missing responses (weight at school entry and weight at eight years) and covariates (weight at 6 weeks, weight at 4 months, weight at 8 months and weight at 12 months). Since the covariates are highly correlated, the inclusion of all variables in the imputation model yields better imputed values. The non-overlapping variable sets and restrictive strategies perform poorer since the correlation between gestational age (an auxiliary variable) and the covariates are very low. This is in agreement with Collins et al. [2001] who stated that the inclusive strategy reduces the chance of inadvertently omitting an important cause of missingness and also brings the possibility of noticeable gains in terms of increased efficiency and reduced bias.

The real-data analysis (GMS dataset) suggests that, in this study, the best approach was STACK with an inclusive model-building strategy. The conclusion from the simulation study in Chapter 5 was different, indicating that model averaging with an inclusive strategy was the best method. A simulation study for predicting weight at school entry was carried out to identify the reasons for these contradictory results. It revealed that the MSE(P) for model averaging using restrictive strategy decreases as missing percentage increases, in contrast to the MSE(P) for model averaging and STACK using inclusive strategy which are increases as missing percentage increases. As a result, model averaging with non-overlapping strategy is the best method when an auxiliary variable is available. This helps to explain the contradictory results between Chapter 5 and real-data analysis, where it indicates the effects of highly correlated covariates and auxiliary variable. The use of highly correlated covariates and auxiliary variable as well as the percentages of missing values play an important role in determining the best method of prediction in the presence of missing data.

Moreover, the effects of outliers on the imputation and prediction steps are highlighted by the analysis of the real-life dataset. Although the observations with premature babies and extremely heavy weight child were removed from the prediction and imputation analysis, there are still deleterious effects of a heavy weight child for prediction of weight

at school entry for female children (since one of the children is heavier compared to others in that group). As discussed in Section 6.2.2, the MSE(P) value is higher for the cross-validation test with the heavy weight female child compared to those obtained without the heavy weight child. This is clearly shown by residuals where the residuals are not normally distributed for cross-validation test with the heavy weight female child. Another important issue is there are more heavy weight female children in this GMS dataset compared to male children, and the prediction of their weights at any time point should be considered separately in order to avoid any mis-interpretation of results in further research.

In conclusion, the proposed method, STACK with an inclusive strategy, performs better than other approaches in terms of prediction and variable selection when missing percentage is high and the correlations with the auxiliary variable is low. The inclusive strategy performs better in imputing missing values if the covariates are highly correlated. If an auxiliary variable is available, the researcher could use the non-overlapping strategy to improve the imputation. Researchers should use model averaging with non-overlapping variable sets for analysing data. Alternatively, researchers can use STACK with an inclusive strategy for prediction if auxiliary variable is not available.

Chapter 7

Conclusion

This chapter summarizes the main achievements of the work presented in this thesis, its contribution and novel aspects, as well as suggestions and recommendations for future work.

7.1 Review of Objectives and Guidelines

Model selection and model averaging in linear model and Logistic regression become complicated in the presence of missing data. The main aim of this research is to provide a comparison between model selection and model averaging so that guidelines can be drawn up for how to apply them in the presence of missing data. Five primary objectives were outlined in Section 1.2 and these will now be reviewed and discussed in the context of four guidelines for researchers who intend to use model selection or model averaging in the presence of missing data.

1. *Which imputation method is best for selecting and fitting additive linear model and Logistic regression? Single imputation or Multiple Imputation?*

   Model selection and model averaging using multiple imputation perform better than single imputation for selecting and fitting additive linear model and Logistic regression. Simulation studies showed that model selection and model averaging using multiple imputation is better than using single imputation for all missing percentages, sample sizes and ρ_{23} in terms of prediction for both linear model and Logistic regression. Therefore, multiple imputation is better for imputing missing data in the context of model-building.

2. *Which model-building strategy is better for imputation and prediction? Inclusive, Restrictive or Non-overlapping variable sets?*

   Imputing missing data using a correct imputation model is essential. When correlations among the covariates are low, one should generally use model selection with non-overlapping variable sets (use highly correlated auxiliary variables only in the imputation model) if the interest of the research in the presence of missing data is to identify which variables to be included when making predictions. The choice of auxiliary variables is usually based on personal judgements. There are no best guidelines for choosing the auxiliary variables. It is advisable to use non-overlapping variable sets if there is highly correlated auxiliary variable is available. Alternatively, researchers can use an inclusive strategy since the inclusive strategy performs better than the non-overlapping variable sets if the covariates are highly correlated and there is a higher missing percentage. The inclusive strategy also generally performs better than the restrictive strategy and non-overlapping variable sets in terms of prediction for extreme circumstances. Therefore, the researcher should use the inclusive strategy for imputation and prediction models. This approach has the added advantage of reducing the distinction between covariates and auxiliary variables, since all variables are available for use in both the imputation and prediction models.

3. *Which model selection criterion is better for model selection and model averaging? AIC, AIC_c or BIC?*

   Based on our simulation studies, model selection criterion, AIC_c performs better than AIC and BIC for larger error variance and in making predictions. AIC_c is known theoretically to be less biased than AIC for small sample size and this is proven through simulation studies. There is not much difference between the model chosen by AIC_c and BIC in terms of prediction for M-STACK method in the real data analysis. BIC performs very poorly for model averaging, where the MSE(P) values based on BIC weights were very much higher compared to those based on AIC_c weights. This is due to the effect of BIC's penalty term. BIC's penalty is more strict than AIC_c and it strongly discourages choosing a model with many parameters, so the smaller models are given more weight in model averaging using BIC compared to AIC_c. Therefore, researchers should carry out model selection and model averaging using model selection criterion, AIC_c.

4. *Which model-building approach is better for prediction? M-STACK or model averaging?*

   STACK performs better than the other model selection methods in terms of variable selection and prediction in most circumstances. In the restricted simulation studies, model averaging performs slightly better than STACK in terms of prediction. However, STACK performs better in the real data analysis on the GMS data. This is due to highly correlated covariates in the GMS study. There is a strong effect of highly correlated covariates and higher missing percentages in the poor performance of model averaging. On the other hand, the highly correlated covariates improve the performance of STACK in the GMS study. Model averaging using non-overlapping variable sets performs better only if an auxiliary variable is available. However, STACK using an inclusive strategy performs well in general for most circumstances. Therefore, researchers should use STACK using an inclusive strategy for model-building in the presence of missing data for making predictions and also for variable selection when there is no auxiliary variable is available.

7.2 Research Contributions

The major contributions to science of the work described in this thesis are listed below.

1. Compared model selection and model averaging in the presence of missing data, in terms of prediction.
2. Proposed a novel model selection method, a modified version of the STACK method (M-STACK), for model selection with multiply-imputed data sets.
3. Proposed STACK and M-STACK methods using all subset regression for model selection.
4. Proposed model averaging procedures for logistic regression based on averaging the estimated probability.
5. Compared inclusive and restrictive strategies for building appropriate imputation and prediction models for model averaging with multiply-imputed data sets, and compared the outcomes with model selection methods.
6. Diagnosed the effects of using highly correlated variables on building the imputation and prediction models.

7. Compared the performance of model selection criteria AIC_c and BIC as the weights in model averaging for linear model and Logistic regression in the presence of missing data.

8. Provided guidelines for model selection and model averaging in the presence of missing data using multiple imputation.

7.3 Limitations and Recommendations for Further Work

Although the research has achieved its original aims, there were some unavoidable limitations. First, due to time constraints, this research was focussed on simple simulation settings with three variables and no correlation between the covariates. Second, this research was focused on MCAR mechanism. There are no explorations of model averaging and model selection under MAR or MNAR mechanisms for either linear models or logistic regression. Finally, this research is restricted to continuous data. Model selection and model averaging of categorical data will introduce additional challenges. Therefore, a list of recommendations for future work are proposed. There are a number of areas that warrant further investigation.

1. Model averaging using the EM-based AIC developed by Ibrahim et al. [2008] needs to be explored in terms of prediction and parameter estimation. Ibrahim et al. [2008] and Claeskens and Consentino [2008] proposed an EM-based AIC for data with missing values, and claimed that model averaging using EM-based AIC can improve the predictions and can be a better choice of model-building approach since model selection method will introduce additional uncertainty into the model-building process.

2. Model averaging and STACK was solely tested on a real dataset when fitting a linear model. Model averaging for Logistic regression performs slightly better than STACK using multiply-imputed data sets in simulation studies, but the performance might be different in the real data analysis for fitting Logistic regression. Therefore, model averaging and STACK method need to be tested using real data on Logistic regression.

3. STACK using highly correlated covariates performs better than model averaging in the real data analysis when fitting a linear model. Highly correlated covariates might have strong effects on imputation and prediction in the real data analysis for fitting Logistic regression. Therefore, the effects of highly correlated covariates on model selection and model averaging for both linear model and Logistic regression

needs to be investigated using an extended Monte Carlo study along the lines of those discussed in Chapter 5 and Section 6.3.

4. The effects and use of binary covariates in both imputation and prediction models need to be further explored for both linear model and Logistic regression.

5. Compare model selection and model averaging under MAR or MNAR mechanisms for linear models and logistic regression.

Appendix A

R-script for Model averaging using Multiple Imputation for Linear Regression

```
n<-100
sigma<-1
rho23<-0
rho12<-0
rho13<-0
beta0<-1
beta1<-1
beta2<-1
beta3<-0
k0<-1          #k is number of parameters#
k1<-2
k2<-3
mu<-c(0,0,0)
sigmax<-matrix(c(1,rho12,rho13,rho12,1,rho23,rho13,rho23,1),3,3)#covariance matrix#
nsim<-1000

coef.AIC<-matrix(c(0,0,0),nrow=1, ncol=3)
coef.AICc<-matrix(c(0,0,0),nrow=1, ncol=3)
coef.BIC<-matrix(c(0,0,0),nrow=1, ncol=3)
LL.mat<-matrix(nrow=nsim, ncol=4)
AIC.best<-matrix(0, nrow=100, ncol=1)
AICc.best<-matrix(0, nrow=100, ncol=1)
```

```
BIC.best<-matrix(0, nrow=100, ncol=1)

#to create test values
prob.test<-seq(0.05,1, by=0.1)
z1.test <-qnorm(prob.test)
z2.test <-z1.test
x.test <- matrix(0,nrow=100, ncol=2)
for (ii in 1:10){
for (jj in 1:10){
x.test[(ii-1)*10+jj, ]<- c(z1.test[ii],z2.test[jj]) }   }

#To find y.test
beta0<-1
beta1<-1
beta2<-1
x1.test<-x.test[,1]
x2.test<-x.test[,2]
y.test<-beta0 + beta1*(x1.test)+ beta2*(x2.test)

#To find X3
x3<-matrix(0, nr=100, nc=1)
for (iii in 1:100){
x1<-matrix(c(x.test[iii,1],x.test[iii,2]),nr=2, nc=1)
mu1<-matrix(c(0,0),nr=2, nc=1)
sigmax<-matrix(c(1,rho12,rho13,rho12,1,rho23,rho13,rho23,1),3,3) #covariance matrix#
sigmax11<-matrix(c(1,rho12,rho12,1),nr=2, nc=2)
sigmax22<-matrix(c(1),nr=1, nc=1)
sigmax12<-matrix(c(rho13,rho23),nr=2, nc=1)
t.sigmax12<-t(sigmax12)                                    #to find transpose#
inv.sigmax11<-solve(sigmax11)                              #to find inverse#
mu2<-(t.sigmax12%*%inv.sigmax11)%*%(x1-mu1)                #to find mean x3#
sigmax2<-sigmax22-(t.sigmax12%*%inv.sigmax11%*%sigmax12)   #to find variance x3#
x3[iii]<-mu2
}
x3.test<-x3
test.values<-data.frame(y.test, x1.test, x2.test, x3.test)

for(i in 1:nsim){
x<-mvrnorm(n,mu,sigmax)
```

```
m<-25                          #m is percentage of missing#
nmiss<-n*(m/100)
nmiss<-round(nmiss)
e<-rnorm(n,0,sigma)                    #e is error term#
X1<-x[,1]
x2<-x[,2]
X3<-x[,3]
y<-beta0 + beta1*(X1)+ beta2*(x2)+ e
x2miss<-rep(NA, times=nmiss)
x2nmiss<-x2[seq(n-nmiss)]
X2<-cbind(c(x2nmiss,x2miss))
#dataset/model for imputation-non-overlapping variable sets#
dat.x<-data.frame(X2,y,X3)
#define number of multiple imputation, D=10#
imp<-mice(dat.x, method="norm", m=10)
mat<-complete(imp,"long")

imp1<-complete(imp,1)  #to retrieve the imputation data set 1#
imp2<-complete(imp,2)
imp3<-complete(imp,3)
imp4<-complete(imp,4)
imp5<-complete(imp,5)
imp6<-complete(imp,6)
imp7<-complete(imp,7)
imp8<-complete(imp,8)
imp9<-complete(imp,9)
imp10<-complete(imp,10)
comp.imp<-list(imp1, imp2, imp3, imp4, imp5, imp6, imp7, imp8, imp9, imp10)

for (k in 1:length(comp.imp)){
dat.xy<-data.frame(y, X1, comp.imp[k])  #non-overlapping variable sets#
M000<-lm(y~1, dat.xy)
M100<-lm(y~X1, dat.xy)
M010<-lm(y~X2, dat.xy)
M110<-lm(y~X1+X2, dat.xy)
model.list<-list(M000, M100, M010, M110)
M000LL<-logLik(M000) #to obtain log-likelihood value from the output for each model#
M100LL<-logLik(M100)
M010LL<-logLik(M010)
```

```
M110LL<-logLik(M110)
LL<-matrix(c(M000LL,M100LL,M010LL,M110LL),nrow=1,ncol=4)
a.LL<-(LL[1]+LL[2]+LL[3]+LL[4])/4

M000AIC<-(2*(LL[1]))-(2*(k0+1))          #calculating AIC, AICc and BIC #
M000BIC<-(2*(LL[1]))-((k0+1)*log(n))
M000AICc<-(2*(LL[1]))-(2*(k0+1)*(n/(n-k0-2)))
M100AIC<-(2*(LL[2]))-(2*(k1+1))
M100BIC<-(2*(LL[2]))-((k1+1)*log(n))
M100AICc<-(2*(LL[2]))-(2*(k1+1)*(n/(n-k1-2)))
M010AIC<-(2*(LL[3]))-(2*(k1+1))
M010BIC<-(2*(LL[3]))-((k1+1)*log(n))
M010AICc<-(2*(LL[3]))-(2*(k1+1)*(n/(n-k1-2)))
M110AIC<-(2*(LL[4]))-(2*(k2+1))
M110BIC<-(2*(LL[4]))-((k2+1)*log(n))
M110AICc<-(2*(LL[4]))-(2*(k2+1)*(n/(n-k2-2)))
M000LLW.AIC<-exp((M000AIC-a.LL)/2)
M100LLW.AIC<-exp((M100AIC-a.LL)/2)
M010LLW.AIC<-exp((M010AIC-a.LL)/2)
M110LLW.AIC<-exp((M110AIC-a.LL)/2)
M000LLW.AICc<-exp((M000AICc-a.LL)/2)
M100LLW.AICc<-exp((M100AICc-a.LL)/2)
M010LLW.AICc<-exp((M010AICc-a.LL)/2)
M110LLW.AICc<-exp((M110AICc-a.LL)/2)
M000LLW.BIC<-exp((M000BIC-a.LL)/2)
M100LLW.BIC<-exp((M100BIC-a.LL)/2)
M010LLW.BIC<-exp((M010BIC-a.LL)/2)
M110LLW.BIC<-exp((M110BIC-a.LL)/2)

#to obtain weights for model from AIC for each model#
W.M000LLAIC<-M000LLW.AIC / (M000LLW.AIC+M100LLW.AIC+M010LLW.AIC+M110LLW.AIC)
W.M100LLAIC<-M100LLW.AIC / (M000LLW.AIC+M100LLW.AIC+M010LLW.AIC+M110LLW.AIC)
W.M010LLAIC<-M010LLW.AIC / (M000LLW.AIC+M100LLW.AIC+M010LLW.AIC+M110LLW.AIC)
W.M110LLAIC<-M110LLW.AIC / (M000LLW.AIC+M100LLW.AIC+M010LLW.AIC+M110LLW.AIC)
#to obtain weights for model from AICc for each model#
W.M000LLAICc<-M000LLW.AICc / (M000LLW.AICc+M100LLW.AICc+M010LLW.AICc+M110LLW.AICc)
W.M100LLAICc<-M100LLW.AICc / (M000LLW.AICc+M100LLW.AICc+M010LLW.AICc+M110LLW.AICc)
W.M010LLAICc<-M010LLW.AICc / (M000LLW.AICc+M100LLW.AICc+M010LLW.AICc+M110LLW.AICc)
W.M110LLAICc<-M110LLW.AICc / (M000LLW.AICc+M100LLW.AICc+M010LLW.AICc+M110LLW.AICc)
```

```
#to obtain weights for model from BIC for each model#
W.M000LLBIC<-M000LLW.BIC / (M000LLW.BIC+M100LLW.BIC+M010LLW.BIC+M110LLW.BIC)
W.M100LLBIC<-M100LLW.BIC / (M000LLW.BIC+M100LLW.BIC+M010LLW.BIC+M110LLW.BIC)
W.M010LLBIC<-M010LLW.BIC / (M000LLW.BIC+M100LLW.BIC+M010LLW.BIC+M110LLW.BIC)
W.M110LLBIC<-M110LLW.BIC / (M000LLW.BIC+M100LLW.BIC+M010LLW.BIC+M110LLW.BIC)

SUM.W.AIC<-W.M000LLAIC+W.M100LLAIC+W.M010LLAIC+W.M110LLAIC
SUM.W.AICc<-W.M000LLAICc+W.M100LLAICc+W.M010LLAICc+W.M110LLAICc
SUM.W.BIC<-W.M000LLBIC+W.M100LLBIC+W.M010LLBIC+W.M110LLBIC

coef.M000.0<-coef(M000)[1]
coef.M000.0<-ifelse(is.na(coef.M000.0), 0, coef.M000.0)

coef.M100.0<-coef(M100)[1]
coef.M100.1<-coef(M100)[2]
coef.M100.0<-ifelse(is.na(coef.M100.0), 0, coef.M100.0)
coef.M100.1<-ifelse(is.na(coef.M100.1), 0, coef.M100.1)

coef.M010.0<-coef(M010)[1]
coef.M010.2<-coef(M010)[2]
coef.M010.0<-ifelse(is.na(coef.M010.0), 0, coef.M010.0)
coef.M010.2<-ifelse(is.na(coef.M010.2), 0, coef.M010.2)

coef.M110.0<-coef(M110)[1]
coef.M110.1<-coef(M110)[2]
coef.M110.2<-coef(M110)[3]
coef.M110.0<-ifelse(is.na(coef.M110.0), 0, coef.M110.0)
coef.M110.1<-ifelse(is.na(coef.M110.1), 0, coef.M110.1)
coef.M110.2<-ifelse(is.na(coef.M110.2), 0, coef.M110.2)

#to find averaged model using AIC weights
coef.AM.AIC.0<-(W.M000LLAIC*coef.M000.0)+(W.M100LLAIC*coef.M100.0)
                +(W.M010LLAIC*coef.M010.0)+(W.M110LLAIC*coef.M110.0)
coef.AM.AIC.1<-((W.M100LLAIC*coef.M100.1)+(W.M110LLAIC*coef.M110.1))
                /(W.M100LLAIC+W.M110LLAIC)
coef.AM.AIC.2<-((W.M010LLAIC*coef.M010.2)+(W.M110LLAIC*coef.M110.2))
                /(W.M010LLAIC+W.M110LLAIC)
```

```
#to find averaged model using AICc weights
coef.AM.AICc.0<-(W.M000LLAICc*coef.M000.0)+(W.M100LLAICc*coef.M100.0)
                  +(W.M010LLAICc*coef.M010.0)+(W.M110LLAICc*coef.M110.0)
coef.AM.AICc.1<-((W.M100LLAICc*coef.M100.1)+(W.M110LLAICc*coef.M110.1))
                  /(W.M100LLAICc+W.M110LLAICc)
coef.AM.AICc.2<-((W.M010LLAICc*coef.M010.2)+(W.M110LLAICc*coef.M110.2))
                  /(W.M010LLAICc+W.M110LLAICc)
#to find averaged model using BIC weights
coef.AM.BIC.0<-(W.M000LLBIC*coef.M000.0)+(W.M100LLBIC*coef.M100.0)
                +(W.M010LLBIC*coef.M010.0)+(W.M110LLBIC*coef.M110.0)
coef.AM.BIC.1<-((W.M100LLBIC*coef.M100.1)+(W.M110LLBIC*coef.M110.1))
                /(W.M100LLBIC+W.M110LLBIC)
coef.AM.BIC.2<-((W.M010LLBIC*coef.M010.2)+(W.M110LLBIC*coef.M110.2))
                /(W.M010LLBIC+W.M110LLBIC)

#to find averaged model coefficient using AIC weights after MI#
coef.AIC[1]<-coef.AIC[1] + coef.AM.AIC.0
coef.AIC[2]<-coef.AIC[2]+ coef.AM.AIC.1
coef.AIC[3]<-coef.AIC[3] + coef.AM.AIC.2
#to find averaged model coefficient using AICc weights after MI#
coef.AICc[1]<-coef.AICc[1] + coef.AM.AICc.0
coef.AICc[2]<-coef.AICc[2] + coef.AM.AICc.1
coef.AICc[3]<-coef.AICc[3] + coef.AM.AICc.2
#to find averaged model coefficient using BIC weights after MI#
coef.BIC[1]<-coef.BIC[1] + coef.AM.BIC.0
coef.BIC[2]<-coef.BIC[2] + coef.AM.BIC.1
coef.BIC[3]<-coef.BIC[3] + coef.AM.BIC.2
} #end of imputation loop#

#to find averaged model coefficient using  AIC weights after MI#
coef.AIC[1]<-coef.AIC[1]/length(comp.imp)
coef.AIC[2]<-coef.AIC[2]/length(comp.imp)
coef.AIC[3]<-coef.AIC[3]/length(comp.imp)
#to find averaged model coefficient using  AICc weights after MI#
coef.AICc[1]<-coef.AICc[1]/length(comp.imp)
coef.AICc[2]<-coef.AICc[2]/length(comp.imp)
coef.AICc[3]<-coef.AICc[3]/length(comp.imp)
```

```
#to find averaged model coefficient using  BIC weights after MI#
coef.BIC[1]<-coef.BIC[1]/length(comp.imp)
coef.BIC[2]<-coef.BIC[2]/length(comp.imp)
coef.BIC[3]<-coef.BIC[3]/length(comp.imp)

#to find y.test for model selected via AIC#
AIC.y.est<- coef.AIC[1] + coef.AIC[2]*(x1.test)
            + coef.AIC[3]*(x2.test)
#to find y.test for model selected via AICc#
AICc.y.est<- coef.AICc[1] + coef.AICc[2]*(x1.test)
            + coef.AICc[3]*(x2.test)
#to find y.test for model selected viaBIC#
BIC.y.est<- coef.BIC[1] + coef.BIC[2]*(x1.test)
            + coef.BIC[3]*(x2.test)

#to find MSE(p) using y.est for for model selected via AIC#
AIC.best<-AIC.best+(AIC.y.est-y.test)^2
AICc.best<-AICc.best+(AICc.y.est-y.test)^2
BIC.best<-BIC.best+(BIC.y.est-y.test)^2

} #end of simulation loop#

AIC.best<-AIC.best/nsim
AICc.best<-AICc.best/nsim
BIC.best<-BIC.best/nsim
```

Appendix B

R-script for Model Selection (RR) using Multiple Imputation for Linear Regression

```
n<-100
sigma<-1
rho23<-0
rho12<-0
rho13<-0
beta0<-1
beta1<-1
beta2<-1
beta3<-0
mu<-c(0,0,0)
sigmax<-matrix(c(1,rho12,rho13,rho12,1,rho23,rho13,rho23,1),3,3) #covariance matrix#
nsim<-1000
coef.est<-matrix(c(0,0,0),nrow=1, ncol=3)
MSEP.best<-matrix(0, nrow=100, ncol=1)
t.model.M000<-0
t.model.M100<-0
t.model.M010<-0
t.model.M110<-0
coef.M000.MI.0<-0
coef.M100.MI.0<-0
coef.M100.MI.1<-0
coef.M010.MI.0<-0
```

```
coef.M010.MI.2<-0
coef.M110.MI.0<-0
coef.M110.MI.1<-0
coef.M110.MI.2<-0
std.err.M000.MI.0<-0
std.err.M100.MI.0<-0
std.err.M100.MI.1<-0
std.err.M010.MI.0<-0
std.err.M010.MI.2<-0
std.err.M110.MI.0<-0
std.err.M110.MI.1<-0
std.err.M110.MI.2<-0

#to create test values
prob.test<-seq(0.05,1, by=0.1)
z1.test <-qnorm(prob.test)
z2.test <-z1.test
x.test <- matrix(0,nrow=100, ncol=2)
for (ii in 1:10){
for (jj in 1:10){
x.test[(ii-1)*10+jj, ]<- c(z1.test[ii],z2.test[jj]) }  }

#To find y.test#
beta0<-1
beta1<-1
beta2<-1
x1.test<-x.test[,1]
x2.test<-x.test[,2]
y.test<-beta0 + beta1*(x1.test)+ beta2*(x2.test)

#To find X3#
x3<-matrix(0, nr=100, nc=1)
for (iii in 1:100){
x1<-matrix(c(x.test[iii,1],x.test[iii,2]),nr=2, nc=1)
mu1<-matrix(c(0,0),nr=2, nc=1)
sigmax<-matrix(c(1,rho12,rho13,rho12,1,rho23,rho13,rho23,1),3,3) #covariance matrix#
sigmax11<-matrix(c(1,rho12,rho12,1),nr=2, nc=2)
sigmax22<-matrix(c(1),nr=1, nc=1)
sigmax12<-matrix(c(rho13,rho23),nr=2, nc=1)
```

```
t.sigmax12<-t(sigmax12)                                    #to find transpose#
inv.sigmax11<-solve(sigmax11)                              #to find inverse#
mu2<-(t.sigmax12%*%inv.sigmax11)%*%(x1-mu1)                #to find mean x3#
sigmax2<-sigmax22-(t.sigmax12%*%inv.sigmax11%*%sigmax12)   #to find variance x3#
x3[iii]<-mu2
}
x3.test<-x3
test.values<-data.frame(y.test, x1.test, x2.test, x3.test)

#to run 1000 simulations#
for(i in 1:nsim){   #simulation loop starts#
x<-mvrnorm(n,mu,sigmax)
m<-50                          #m is percentage of missing#
nmiss<-n*(m/100)
nmiss<-round(nmiss)
e<-rnorm(n,0,sigma)            #e is error term#
X1<-x[,1]
x2<-x[,2]
X3<-x[,3]
y<-beta0 + beta1*(X1)+ beta2*(x2)+ e
x2miss<-rep(NA, times=nmiss)
x2nmiss<-x2[seq(n-nmiss)]
X2<-cbind(c(x2nmiss,x2miss))
#dataset/model for imputation-non-overlapping variable sets#
dat.x<-data.frame(X2,y,X3)

#define number of MI, D=10#
imp<-mice(dat.x, method="norm", m=10)
mat<-complete(imp,"long")
imp1<-complete(imp,1)  #to retrieve the imputation data set 1#
imp2<-complete(imp,2)
imp3<-complete(imp,3)
imp4<-complete(imp,4)
imp5<-complete(imp,5)
imp6<-complete(imp,6)
imp7<-complete(imp,7)
imp8<-complete(imp,8)
imp9<-complete(imp,9)
imp10<-complete(imp,10)
```

```
comp.imp<-list(imp1, imp2, imp3, imp4, imp5, imp6, imp7, imp8, imp9, imp10)

{ #run M110#
for (k in 1:length(comp.imp)) #start imp loop M110{
dat.xy<-data.frame(y, X1, comp.imp[k])
M110<-lm(y~X1+X2, dat.xy)

#to retrieve covariance matrix#
covmat.M110<-vcov(M110)
#to retrieve coefficients#
coef.M110.0<-coef(M110)[1]
coef.M110.1<-coef(M110)[2]
coef.M110.2<-coef(M110)[3]
#to retrieve standard error#
std.err.M110.0<-coef(summary(M110))[, "Std. Error"][1]
std.err.M110.1<-coef(summary(M110))[, "Std. Error"][2]
std.err.M110.2<-coef(summary(M110))[, "Std. Error"][3]
coef.M110.0<-ifelse(is.na(coef.M110.0), 0, coef.M110.0)
coef.M110.1<-ifelse(is.na(coef.M110.1), 0, coef.M110.1)
coef.M110.2<-ifelse(is.na(coef.M110.2), 0, coef.M110.2)

#to find total coefficient after MI#
coef.M110.MI.0<-coef.M110.MI.0 + coef.M110.0
coef.M110.MI.1<-coef.M110.MI.1 + coef.M110.1
coef.M110.MI.2<-coef.M110.MI.2 + coef.M110.2
#to find total std.error after MI#
std.err.M110.MI.0<-std.err.M110.MI.0 + std.err.M110.0
std.err.M110.MI.1<-std.err.M110.MI.1 + std.err.M110.1
std.err.M110.MI.2<-std.err.M110.MI.2 + std.err.M110.2
}  #end of imputation loop M110#

#to find averaged coefficient after MI#
coef.M110.MI.0<-coef.M110.MI.0/length(comp.imp)
coef.M110.MI.1<-coef.M110.MI.1/length(comp.imp)
coef.M110.MI.2<-coef.M110.MI.2/length(comp.imp)
#to find averaged std.error after MI#
std.err.M110.MI.0<-std.err.M110.MI.0/length(comp.imp)
std.err.M110.MI.1<-std.err.M110.MI.1/length(comp.imp)
std.err.M110.MI.2<-std.err.M110.MI.2/length(comp.imp)
```

```
#To omit coefficients based on RR#
coef.M110.model.2<-ifelse((coef.M110.MI.2/std.err.M110.MI.2)<(-1.96)
     | (coef.M110.MI.2/std.err.M110.MI.2)>1.96, coef.M110.MI.2, 0)

if(coef.M110.model.2==0){  #test beta2 of M110#
{ #run M100#
for (k in 1:length(comp.imp)) {   # starts imputation loop for M100#
dat.xy<-data.frame(y, X1, comp.imp[k])
M100<-lm(y~X1, dat.xy)
covmat.M100<-vcov(M100)
coef.M100.0<-coef(M100)[1]
coef.M100.1<-coef(M100)[2]
std.err.M100.0<-coef(summary(M100))[, "Std. Error"][1]
std.err.M100.1<-coef(summary(M100))[, "Std. Error"][2]
coef.M100.0<-ifelse(is.na(coef.M100.0), 0, coef.M100.0)
coef.M100.1<-ifelse(is.na(coef.M100.1), 0, coef.M100.1)

#to find total coefficient after MI#
coef.M100.MI.0<-coef.M100.MI.0 + coef.M100.0
coef.M100.MI.1<-coef.M100.MI.1 + coef.M100.1
#to find total std.error after multiple imputation#
std.err.M100.MI.0<-std.err.M100.MI.0 + std.err.M100.0
std.err.M100.MI.1<-std.err.M100.MI.1 + std.err.M100.1
} #end imputation loop M100#

#to find averaged coefficient after MI#
coef.M100.MI.0<-coef.M100.MI.0/length(comp.imp)
coef.M100.MI.1<-coef.M100.MI.1/length(comp.imp)
#to find averaged std.error after MI#
std.err.M100.MI.0<-std.err.M110.MI.0/length(comp.imp)
std.err.M100.MI.1<-std.err.M110.MI.1/length(comp.imp)
coef.M100.MI.0<-ifelse(is.na(coef.M100.MI.0), 0, coef.M100.MI.0)
coef.M100.MI.1<-ifelse(is.na(coef.M100.MI.1), 0, coef.M100.MI.1)
#To omit coefficients based on RR#
coef.M100.model.1<-ifelse((coef.M100.MI.1/std.err.M100.MI.1)<(-1.96)
                  |(coef.M100.MI.1/std.err.M100.MI.1)>1.96, coef.M100.MI.1, 0)

if(coef.M100.model.1==0){  #test beta1 of M100#
{ #run M010#
```

```
for (k in 1:length(comp.imp))  { #starts imputation loop M010#
dat.xy<-data.frame(y, X1, comp.imp[k])
M010<-lm(y~X2, dat.xy)
covmat.M010<-vcov(M010)
coef.M010.0<-coef(M010)[1]
coef.M010.2<-coef(M010)[2]
std.err.M010.0<-coef(summary(M010))[, "Std. Error"][1]
std.err.M010.2<-coef(summary(M010))[, "Std. Error"][2]
coef.M010.0<-ifelse(is.na(coef.M010.0), 0, coef.M010.0)
coef.M010.2<-ifelse(is.na(coef.M010.2), 0, coef.M010.2)

#to find total coefficients after MI#
coef.M010.MI.0<-coef.M010.MI.0 + coef.M010.0
coef.M010.MI.2<-coef.M010.MI.2 + coef.M010.2
#to find total std.error after multiple imputation#
std.err.M010.MI.0<-std.err.M010.MI.0 + std.err.M010.0
std.err.M010.MI.2<-std.err.M010.MI.2 + std.err.M010.2
} #end imputation loop M010#

#to find averaged coefficient after MI#
coef.M010.MI.0<-coef.M010.MI.0/length(comp.imp)
coef.M010.MI.2<-coef.M010.MI.2/length(comp.imp)
#to find averaged std.error after MI#
std.err.M010.MI.0<-std.err.M010.MI.0/length(comp.imp)
std.err.M010.MI.2<-std.err.M010.MI.2/length(comp.imp)
coef.M010.MI.0<-ifelse(is.na(coef.M010.MI.0), 0, coef.M010.MI.0)
coef.M010.MI.2<-ifelse(is.na(coef.M010.MI.2), 0, coef.M010.MI.2)

#To omit coefficients based on RR#
coef.M010.model.2<-ifelse((coef.M010.MI.2/std.err.M010.MI.2)<(-1.96)
     | (coef.M010.MI.2/std.err.M010.MI.2)>1.96, coef.M010.MI.2, 0)

if(coef.M010.model.2==0){ #beta2 of M010#
{ #run M000#
for (k in 1:length(comp.imp)) { #starts imputation loop M000#
dat.xy<-data.frame(y, X1, comp.imp[k])
M000<-lm(y~1, dat.xy)
covmat.M000<-vcov(M000)
coef.M000.0<-coef(M000)[1]
```

```
std.err.M000.0<-coef(summary(M000))[, "Std. Error"][1]
coef.M000.0<-ifelse(is.na(coef.M000.0), 0, coef.M000.0)

#to find total coefficient after MI#
coef.M000.MI.0<-coef.M000.MI.0 + coef.M000.0
#to find total std.error after multiple imputation#
std.err.M000.MI.0<-std.err.M000.MI.0 + std.err.M000.0
} #end imputation loop M000#

#to find averaged coefficient after MI#
coef.M000.MI.0<-coef.M000.MI.0/length(comp.imp)
#to find averaged std.error after MI#
std.err.M000.MI.0<-std.err.M000.MI.0/length(comp.imp)
coef.M000.MI.0<-ifelse(is.na(coef.M000.MI.0), 0, coef.M000.MI.0)

model.M000<-1
y.est.M000<- coef.M000.MI.0                    #to find y.test for model M000#
MSEP.best<-MSEP.best+(y.est.M000-y.test)^2     #to find mse.p#
coef.est[1]<-coef.est[1] + coef.M000.MI.0
t.model.M000<-t.model.M000 + model.M000         #to find total number of M000#
} #end M000#

#for model M010# } else {
model.M010<-1
#to find y.test for model M010#
y.est.M010<- coef.M010.MI.0+ coef.M010.MI.2*(x2.test)
MSEP.best<-MSEP.best+(y.est.M010-y.test)^2   #to find mse.p#
coef.est[1]<-coef.est[1] + coef.M010.MI.0
coef.est[3]<-coef.est[3] + coef.M010.MI.2
t.model.M010<-t.model.M010 + model.M010    #to find total number of M010#
} #end else loop M010# } #end M010#

#for model M100# } else {
model.M100<-1
#to find y.test for model M100#
y.est.M100<- coef.M100.MI.0+ coef.M100.MI.1*(x1.test)
MSEP.best<-MSEP.best+(y.est.M100-y.test)^2      #to find mse.p#
coef.est[1]<-coef.est[1] + coef.M100.MI.0
coef.est[2]<-coef.est[2] + coef.M100.MI.1
```

```
t.model.M100<-t.model.M100 + model.M100     #to find total number of M100#
} #end else loop M100# } #end M100#

#for model M110# } else {
model.M110<-1
#to find y.test for model M110#
y.est.M110<- coef.M110.MI.0+ coef.M110.MI.1*(x1.test)+ coef.M110.MI.2*(x2.test)
MSEP.best<-MSEP.best+(y.est.M110-y.test)^2         #to find mse.p#
coef.est[1]<-coef.est[1] + coef.M110.MI.0
coef.est[2]<-coef.est[2] + coef.M110.MI.1
coef.est[3]<-coef.est[3] + coef.M110.MI.2
t.model.M110<-t.model.M110 + model.M110     #to find total number of M110#
} #end else loop M110# } #end M110#
} #end of simulation loop#

coef.est<-coef.est/nsim
MSEP.best<-MSEP.best/nsim
t.no.model<-matrix(c(t.model.M000, t.model.M100, t.model.M010,
                     t.model.M110),nrow=1,ncol=4)
```

Appendix C

R-script for Model Selection (M-STACK) using Multiple Imputation for Linear Regression

```
n<-100
sigma<-1
rho23<-0
rho12<-0
rho13<-0
beta0<-1
beta1<-1
beta2<-1
beta3<-0
k0<-1        #k is number of parameters#
k1<-2
k2<-3
mu<-c(0,0,0)
sigmax<-matrix(c(1,rho12,rho13,rho12,1,rho23,rho13,rho23,1),3,3) #covariance matrix#
nsim<-1000
coef.AIC<-matrix(c(0,0,0), nrow=1, ncol=3)
coef.AICc<-matrix(c(0,0,0), nrow=1, ncol=3)
coef.BIC<-matrix(c(0,0,0), nrow=1, ncol=3)
LL.mat<-matrix(nrow=nsim, ncol=4)
AIC.mat<-matrix(nrow=nsim, ncol=4)
AICc.mat<-matrix(nrow=nsim, ncol=4)
BIC.mat<-matrix(nrow=nsim, ncol=4)
```

```
AIC.best<-matrix(0, nrow=100, ncol=1)
AICc.best<-matrix(0, nrow=100, ncol=1)
BIC.best<-matrix(0, nrow=100, ncol=1)
n.AIC.model<-matrix(c(0,0,0,0), nrow=1, ncol=4)
n.AICc.model<-matrix(c(0,0,0,0), nrow=1, ncol=4)
n.BIC.model<-matrix(c(0,0,0,0), nrow=1, ncol=4)

#to create test values
prob.test<-seq(0.05,1, by=0.1)
z1.test <-qnorm(prob.test)
z2.test <-z1.test
x.test <- matrix(0,nrow=100, ncol=2)
for (ii in 1:10) {
  for (jj in 1:10) {
     x.test[(ii-1)*10+jj, ]<- c(z1.test[ii],z2.test[jj]) }  }

#To find y.test#
beta0<-1
beta1<-1
beta2<-1
x1.test<-x.test[,1]
x2.test<-x.test[,2]
y.test<-beta0 + beta1*(x1.test)+ beta2*(x2.test)

#To find X3#
x3<-matrix(0, nr=100, nc=1)
for (iii in 1:100) {
x1<-matrix(c(x.test[iii,1],x.test[iii,2]),nr=2, nc=1)
mu1<-matrix(c(0,0),nr=2, nc=1)
sigmax<-matrix(c(1,rho12,rho13,rho12,1,rho23,rho13,rho23,1),3,3) #covariance matrix#
sigmax11<-matrix(c(1,rho12,rho12,1),nr=2, nc=2)
sigmax22<-matrix(c(1),nr=1, nc=1)
sigmax12<-matrix(c(rho13,rho23),nr=2, nc=1)
t.sigmax12<-t(sigmax12)                                  #to find transpose#
inv.sigmax11<-solve(sigmax11)                            #to find inverse#
mu2<-(t.sigmax12%*%inv.sigmax11)%*%(x1-mu1)              #to find mean x3#
sigmax2<-sigmax22-(t.sigmax12%*%inv.sigmax11%*%sigmax12) #to find variance x3#
x3[iii]<-mu2
}
```

```
x3.test<-x3
test.values<-data.frame(y.test, x1.test, x2.test, x3.test)

#to run 1000 simulations#
for(i in 1:nsim) {  #simulation loop starts#
x<-mvrnorm(n,mu,sigmax)
m<-50
nmiss<-n*(m/100)
nmiss<-round(nmiss)             #m is percentage of missing#
w1<-1/10                        #weight for X1#
w2<-(1-(nmiss/n))/10            #weight for X2#
W1big<-rep(w1, times=(n*10))
W2big<-rep(w2, times=(n*10))
e<-rnorm(n,0,sigma)             #e is error term#
X1<-x[,1]
x2<-x[,2]
X3<-x[,3]
y<-beta0 + beta1*(X1)+ beta2*(x2)+ e
x2miss<-rep(NA, times=nmiss)
x2nmiss<-x2[seq(n-nmiss)]
X2<-cbind(c(x2nmiss,x2miss))
#dataset/model for imputation-non-overlapping variable sets#
dat.x<-data.frame(X2,y,X3)
#define number of multiple imputation, D=10#
imp<-mice(dat.x, method="norm", m=10)
imp1<-complete(imp,1)   #to retrieve the imputation data set 1#
mat<-complete(imp,"long")

#to obtain get stacked data after imputation#
ybig<-rep(y, times=10)
x1big<-rep(X1, times=10)
data.xy<-data.frame(ybig, x1big, mat)
M000<-lm(ybig~1, data.xy)
M100<-lm(ybig~x1big, data.xy, weights=(W1big))
M010<-lm(ybig~X2, data.xy, weights=(W2big))
M110<-lm(ybig~x1big+X2, data.xy, weights=(W2big))
model.list<-list(M000, M100, M010, M110)
```

```
#to obtain log-likelihood value from the output for each model#
M000LL<-logLik(M000)
M100LL<-logLik(M100)
M010LL<-logLik(M010)
M110LL<-logLik(M110)

LL<-matrix(c(M000LL,M100LL,M010LL,M110LL),nrow=1,ncol=4)
M000AIC<-(2*(LL[1]))-(2*(k0+1))           #calculating AIC, AICc and BIC#
M000BIC<-(2*(LL[1]))-((k0+1)*log(n))
M000AICc<-(2*(LL[1]))-(2*(k0+1)*(n/(n-k0-2)))
M100AIC<-(2*(LL[2]))-(2*(k1+1))
M100BIC<-(2*(LL[2]))-((k1+1)*log(n))
M100AICc<-(2*(LL[2]))-(2*(k1+1)*(n/(n-k1-2)))
M010AIC<-(2*(LL[3]))-(2*(k1+1))
M010BIC<-(2*(LL[3]))-((k1+1)*log(n))
M010AICc<-(2*(LL[3]))-(2*(k1+1)*(n/(n-k1-2)))
M110AIC<-(2*(LL[4]))-(2*(k2+1))
M110BIC<-(2*(LL[4]))-((k2+1)*log(n))
M110AICc<-(2*(LL[4]))-(2*(k2+1)*(n/(n-k2-2)))

#to form a matrix of LL for all the models#
LL<-matrix(c(M000LL,M100LL,M010LL,M110LL),nrow=1,ncol=4)
#to form a matrix of AIC for all the models#
AIC<-matrix(c(M000AIC,M100AIC,M010AIC,M110AIC),nrow=1,ncol=4)
#to form a matrix of AICc for all the models#
AICc<-matrix(c(M000AIC,M100AIC,M010AIC,M110AIC),nrow=1,ncol=4)
#to form a matrix of BIC for all the models#
BIC<-matrix(c(M000BIC,M100BIC,M010BIC,M110BIC),nrow=1,ncol=4)

AIC.mat[i,]<-AIC
AICc.mat[i,]<-AICc
BIC.mat[i,]<-BIC
LL.mat[i,]<-LL

max.AIC<- max.col(AIC)            #find the maximum column in AIC matrix#
max.AICc<- max.col(AICc)          #find the maximum column in AICc matrix#
max.BIC<- max.col(BIC)            #find the maximum column in BIC matrix#
```

```
#find the frequency of selected models#
model<-list("M000", "M100", "M010", "M110")
AIC.model<-model[[max.AIC]]
n.AIC.M000<-ifelse(AIC.model=="M000", 1, 0)
n.AIC.M100<-ifelse(AIC.model=="M100", 1, 0)
n.AIC.M010<-ifelse(AIC.model=="M010", 1, 0)
n.AIC.M110<-ifelse(AIC.model=="M110", 1, 0)

AICc.model<-model[[max.AICc]]
n.AICc.M000<-ifelse(AICc.model=="M000", 1, 0)
n.AICc.M100<-ifelse(AICc.model=="M100", 1, 0)
n.AICc.M010<-ifelse(AICc.model=="M010", 1, 0)
n.AICc.M110<-ifelse(AICc.model=="M110", 1, 0)

BIC.model<-model[[max.BIC]]
n.BIC.M000<-ifelse(BIC.model=="M000", 1, 0)
n.BIC.M100<-ifelse(BIC.model=="M100", 1, 0)
n.BIC.M010<-ifelse(BIC.model=="M010", 1, 0)
n.BIC.M110<-ifelse(BIC.model=="M110", 1, 0)

AIC.model<-model.list[[max.AIC]]
AICc.model<-model.list[[max.AICc]]
BIC.model<-model.list[[max.BIC]]

coef.AIC.model.0<-coef(AIC.model)[1]
coef.AIC.model.1<-coef(AIC.model)[2]
coef.AIC.model.2<-coef(AIC.model)[3]
coef.AICc.model.0<-coef(AICc.model)[1]
coef.AICc.model.1<-coef(AICc.model)[2]
coef.AICc.model.2<-coef(AICc.model)[3]
coef.BIC.model.0<-coef(BIC.model)[1]
coef.BIC.model.1<-coef(BIC.model)[2]
coef.BIC.model.2<-coef(BIC.model)[3]

coef.AIC.model.0<-ifelse(is.na(coef.AIC.model.0), 0, coef.AIC.model.0)
coef.AIC.model.1<-ifelse(is.na(coef.AIC.model.1), 0, coef.AIC.model.1)
coef.AIC.model.2<-ifelse(is.na(coef.AIC.model.2), 0, coef.AIC.model.2)
coef.AICc.model.0<-ifelse(is.na(coef.AICc.model.0), 0, coef.AICc.model.0)
coef.AICc.model.1<-ifelse(is.na(coef.AICc.model.1), 0, coef.AICc.model.1)
```

```
coef.AICc.model.2<-ifelse(is.na(coef.AICc.model.2), 0, coef.AICc.model.2)
coef.BIC.model.0<-ifelse(is.na(coef.BIC.model.0), 0, coef.BIC.model.0)
coef.BIC.model.1<-ifelse(is.na(coef.BIC.model.1), 0, coef.BIC.model.1)
coef.BIC.model.2<-ifelse(is.na(coef.BIC.model.2), 0, coef.BIC.model.2)

#to find y.test using model selected via AIC#
AIC.y.est<- coef.AIC.model.0 + coef.AIC.model.1*(x1.test)
            + coef.AIC.model.2*(x2.test)
#to find y.test using model selected via AICc#
AICc.y.est<- coef.AICc.model.0 + coef.AICc.model.1*(x1.test)
            + coef.AICc.model.2*(x2.test)
#to find y.test using model selected via BIC#
BIC.y.est<- coef.BIC.model.0 + coef.BIC.model.1*(x1.test)
            + coef.BIC.model.2*(x2.test)

#to find MSE(p) using y.est for model selected via AIC#
AIC.best<-AIC.best+(AIC.y.est-y.test)^2
AICc.best<-AICc.best+(AICc.y.est-y.test)^2
BIC.best<-BIC.best+(BIC.y.est-y.test)^2

#to find total coefficient over simulation loop#
coef.AIC[1]<-coef.AIC[1] + coef.AIC.model.0
coef.AIC[2]<-coef.AIC[2] + coef.AIC.model.1
coef.AIC[3]<-coef.AIC[3] + coef.AIC.model.2
coef.AICc[1]<-coef.AICc[1] + coef.AICc.model.0
coef.AICc[2]<-coef.AICc[2] + coef.AICc.model.1
coef.AICc[3]<-coef.AICc[3] + coef.AICc.model.2
coef.BIC[1]<-coef.BIC[1] + coef.BIC.model.0
coef.BIC[2]<-coef.BIC[2] + coef.BIC.model.1
coef.BIC[3]<-coef.BIC[3] + coef.BIC.model.2

#to find total frequency of each model selected over simulation loop#
n.AIC.model[1]<-n.AIC.model[1] + n.AIC.M000
n.AIC.model[2]<-n.AIC.model[2] + n.AIC.M100
n.AIC.model[3]<-n.AIC.model[3] + n.AIC.M010
n.AIC.model[4]<-n.AIC.model[4] + n.AIC.M110
n.AICc.model[1]<-n.AICc.model[1] + n.AICc.M000
n.AICc.model[2]<-n.AICc.model[2] + n.AICc.M100
n.AICc.model[3]<-n.AICc.model[3] + n.AICc.M010
```

```
n.AICc.model[4]<-n.AICc.model[4] + n.AICc.M110
n.BIC.model[1]<-n.BIC.model[1] + n.BIC.M000
n.BIC.model[2]<-n.BIC.model[2] + n.BIC.M100
n.BIC.model[3]<-n.BIC.model[3] + n.BIC.M010
n.BIC.model[4]<-n.BIC.model[4] + n.BIC.M110
} #end of simulation loop#

#To find averaged MSE(P)#
AIC.best<-AIC.best/nsim
AICc.best<-AICc.best/nsim
BIC.best<-BIC.best/nsim
#To find averaged coefficients#
coef.AIC<-coef.AIC/nsim
coef.AICc<-coef.AICc/nsim
coef.BIC<-coef.BIC/nsim
```

Appendix D

R-script for Model Selection (STACK) using Multiple Imputation for Linear Regression

```
n<-100
sigma<-1
rho23<-0
rho12<-0
rho13<-0
beta0<-1
beta1<-1
beta2<-1
beta3<-0
k0<-1        #k is number of parameters#
k1<-2
k2<-3
mu<-c(0,0,0)
sigmax<-matrix(c(1,rho12,rho13,rho12,1,rho23,rho13,rho23,1),3,3) #covariance matrix#
nsim<-1000
coef.AIC<-matrix(c(0,0,0), nrow=1, ncol=3)
coef.AICc<-matrix(c(0,0,0), nrow=1, ncol=3)
coef.BIC<-matrix(c(0,0,0), nrow=1, ncol=3)
LL.mat<-matrix(nrow=nsim, ncol=4)
AIC.mat<-matrix(nrow=nsim, ncol=4)
AICc.mat<-matrix(nrow=nsim, ncol=4)
BIC.mat<-matrix(nrow=nsim, ncol=4)
```

```
n.AIC.model<-matrix(c(0,0,0,0), nrow=1, ncol=4)
n.AICc.model<-matrix(c(0,0,0,0), nrow=1, ncol=4)
n.BIC.model<-matrix(c(0,0,0,0), nrow=1, ncol=4)
AIC.best<-matrix(0, nrow=100, ncol=1)
AICc.best<-matrix(0, nrow=100, ncol=1)
BIC.best<-matrix(0, nrow=100, ncol=1)

coef.AIC.MI.0<-0
coef.AIC.MI.1<-0
coef.AIC.MI.2<-0
coef.AICc.MI.0<-0
coef.AICc.MI.1<-0
coef.AICc.MI.2<-0
coef.BIC.MI.0<-0
coef.BIC.MI.1<-0
coef.BIC.MI.2<-0

#to create test values#
prob.test<-seq(0.05,1, by=0.1)
z1.test <-qnorm(prob.test)
z2.test <-z1.test
x.test <- matrix(0,nrow=100, ncol=2)
for (ii in 1:10) {
  for (jj in 1:10) {
     x.test[(ii-1)*10+jj, ]<- c(z1.test[ii],z2.test[jj]) } }

#To find y.test#
beta0<-1
beta1<-1
beta2<-1
x1.test<-x.test[,1]
x2.test<-x.test[,2]
y.test<-beta0 + beta1*(x1.test)+ beta2*(x2.test)

#To find X3#
x3<-matrix(0, nr=100, nc=1)
for (iii in 1:100) {
x1<-matrix(c(x.test[iii,1],x.test[iii,2]),nr=2, nc=1)
mu1<-matrix(c(0,0),nr=2, nc=1)
```

```
sigmax<-matrix(c(1,rho12,rho13,rho12,1,rho23,rho13,rho23,1),3,3) #covariance matrix#
sigmax11<-matrix(c(1,rho12,rho12,1),nr=2, nc=2)
sigmax22<-matrix(c(1),nr=1, nc=1)
sigmax12<-matrix(c(rho13,rho23),nr=2, nc=1)
t.sigmax12<-t(sigmax12)                                    #to find transpose#
inv.sigmax11<-solve(sigmax11)                              #to find inverse#
mu2<-(t.sigmax12%*%inv.sigmax11)%*%(x1-mu1)                #to find mean x3#
sigmax2<-sigmax22-(t.sigmax12%*%inv.sigmax11%*%sigmax12)   #to find variance x3#
x3[iii]<-mu2
}
x3.test<-x3
test.values<-data.frame(y.test, x1.test, x2.test, x3.test)

#to run 1000 simulations#
for(i in 1:nsim)   { #simulation loop starts#
x<-mvrnorm(n,mu,sigmax)
m<-50
nmiss<-n*(m/100)
nmiss<-round(nmiss)              #m is percentage of missing#
w1<-1/10                         #weight for X1
w2<-(1-(nmiss/n))/10             #weight for X2
W1big<-rep(w1, times=(n*10))
W2big<-rep(w2, times=(n*10))
e<-rnorm(n,0,sigma)              #e is error term#
X1<-x[,1]
x2<-x[,2]
X3<-x[,3]
y<-beta0 + beta1*(X1)+ beta2*(x2)+ e
x2miss<-rep(NA, times=nmiss)
x2nmiss<-x2[seq(n-nmiss)]
X2<-cbind(c(x2nmiss,x2miss))
#dataset/model for imputation-non-overlapping variable sets#
dat.x<-data.frame(X2,y,X3)
imp<-mice(dat.x, method="norm", m=10)    #define number of MI, D=10#
imp1<-complete(imp,1)             #to retrieve the imputation data set 1#
mat<-complete(imp,"long")


imp1<-complete(imp,1)  #to retrieve the imputation data set 1#
```

```
imp2<-complete(imp,2)
imp3<-complete(imp,3)
imp4<-complete(imp,4)
imp5<-complete(imp,5)
imp6<-complete(imp,6)
imp7<-complete(imp,7)
imp8<-complete(imp,8)
imp9<-complete(imp,9)
imp10<-complete(imp,10)
comp.imp<-list(imp1, imp2, imp3, imp4, imp5, imp6, imp7, imp8, imp9, imp10)

#to obtain get stacked data after imputation#
ybig<-rep(y, times=10)
x1big<-rep(X1, times=10)
data.xy<-data.frame(ybig, x1big, mat)
M000<-lm(ybig~1, data.xy)
M100<-lm(ybig~x1big, data.xy, weights=(W1big))
M010<-lm(ybig~X2, data.xy, weights=(W2big))
M110<-lm(ybig~x1big+X2, data.xy, weights=(W2big))
model.list<-list(M000, M100, M010, M110)

#to obtain log-likelihood value from the output for each model#
M000LL<-logLik(M000)
M100LL<-logLik(M100)
M010LL<-logLik(M010)
M110LL<-logLik(M110)
LL<-matrix(c(M000LL,M100LL,M010LL,M110LL),nrow=1,ncol=4)

M000AIC<-(2*(LL[1]))-(2*(k0+1))        #calculating AIC, AICc and BIC#
M000BIC<-(2*(LL[1]))-((k0+1)*log(n))
M000AICc<-(2*(LL[1]))-(2*(k0+1)*(n/(n-k0-2)))
M100AIC<-(2*(LL[2]))-(2*(k1+1))
M100BIC<-(2*(LL[2]))-((k1+1)*log(n))
M100AICc<-(2*(LL[2]))-(2*(k1+1)*(n/(n-k1-2)))
M010AIC<-(2*(LL[3]))-(2*(k1+1))
M010BIC<-(2*(LL[3]))-((k1+1)*log(n))
M010AICc<-(2*(LL[3]))-(2*(k1+1)*(n/(n-k1-2)))
M110AIC<-(2*(LL[4]))-(2*(k2+1))
M110BIC<-(2*(LL[4]))-((k2+1)*log(n))
```

```
M110AICc<-(2*(LL[4]))-(2*(k2+1)*(n/(n-k2-2)))

#to form a matrix of LL for all the models#
LL<-matrix(c(M000LL,M100LL,M010LL,M110LL),nrow=1,ncol=4)
#to form a matrix of AIC for all the models#
AIC<-matrix(c(M000AIC,M100AIC,M010AIC,M110AIC),nrow=1,ncol=4)
#to form a matrix of AICc for all the models#
AICc<-matrix(c(M000AIC,M100AIC,M010AIC,M110AIC),nrow=1,ncol=4)
#to form a matrix of BIC for all the models#
BIC<-matrix(c(M000BIC,M100BIC,M010BIC,M110BIC),nrow=1,ncol=4)

AIC.mat[i,]<-AIC
AICc.mat[i,]<-AICc
BIC.mat[i,]<-BIC
LL.mat[i,]<-LL

max.AIC<- max.col(AIC)              #find the maximum column in AIC matrix#
max.AICc<- max.col(AICc)            #find the maximum column in AICc matrix#
max.BIC<- max.col(BIC)              #find the maximum column in BIC matrix#

#find the frequency of selected model#
model<-list("M000", "M100", "M010", "M110")
AIC.model<-model[[max.AIC]]
n.AIC.M000<-ifelse(AIC.model=="M000", 1, 0)
n.AIC.M100<-ifelse(AIC.model=="M100", 1, 0)
n.AIC.M010<-ifelse(AIC.model=="M010", 1, 0)
n.AIC.M110<-ifelse(AIC.model=="M110", 1, 0)
AICc.model<-model[[max.AICc]]
n.AICc.M000<-ifelse(AICc.model=="M000", 1, 0)
n.AICc.M100<-ifelse(AICc.model=="M100", 1, 0)
n.AICc.M010<-ifelse(AICc.model=="M010", 1, 0)
n.AICc.M110<-ifelse(AICc.model=="M110", 1, 0)
BIC.model<-model[[max.BIC]]
n.BIC.M000<-ifelse(BIC.model=="M000", 1, 0)
n.BIC.M100<-ifelse(BIC.model=="M100", 1, 0)
n.BIC.M010<-ifelse(BIC.model=="M010", 1, 0)
n.BIC.M110<-ifelse(BIC.model=="M110", 1, 0)

for (k in 1:length(comp.imp))  {
```

```
dat.xy<-data.frame(y, X1, comp.imp[k])
MM000<-lm(y~1, dat.xy)
MM100<-lm(y~X1, dat.xy)
MM010<-lm(y~X2, dat.xy)
MM110<-lm(y~X1+X2, dat.xy)
model.list2<-list(MM000, MM100, MM010, MM110)
best.model.AIC<-model.list2[[max.AIC]]
best.model.AICc<-model.list2[[max.AICc]]
best.model.BIC<-model.list2[[max.BIC]]

coef.AIC.0<-coef(best.model.AIC)[1]
coef.AIC.1<-coef(best.model.AIC)[2]
coef.AIC.2<-coef(best.model.AIC)[3]
coef.AICc.0<-coef(best.model.AICc)[1]
coef.AICc.1<-coef(best.model.AICc)[2]
coef.AICc.2<-coef(best.model.AICc)[3]
coef.BIC.0<-coef(best.model.BIC)[1]
coef.BIC.1<-coef(best.model.BIC)[2]
coef.BIC.2<-coef(best.model.BIC)[3]

coef.AIC.0<-ifelse(is.na(coef.AIC.0), 0, coef.AIC.0)
coef.AIC.1<-ifelse(is.na(coef.AIC.1), 0, coef.AIC.1)
coef.AIC.2<-ifelse(is.na(coef.AIC.2), 0, coef.AIC.2)
coef.AICc.0<-ifelse(is.na(coef.AICc.0), 0, coef.AICc.0)
coef.AICc.1<-ifelse(is.na(coef.AICc.1), 0, coef.AICc.1)
coef.AICc.2<-ifelse(is.na(coef.AICc.2), 0, coef.AICc.2)
coef.BIC.0<-ifelse(is.na(coef.BIC.0), 0, coef.BIC.0)
coef.BIC.1<-ifelse(is.na(coef.BIC.1), 0, coef.BIC.1)
coef.BIC.2<-ifelse(is.na(coef.BIC.2), 0, coef.BIC.2)

#to find total model coefficient for model selected via AIC after MI#
coef.AIC.MI.0<-coef.AIC.MI.0 + coef.AIC.0
coef.AIC.MI.1<-coef.AIC.MI.1 + coef.AIC.1
coef.AIC.MI.2<-coef.AIC.MI.2 + coef.AIC.2
#to find total model coefficient for model selected via AICc after MI#
coef.AICc.MI.0<-coef.AICc.MI.0 + coef.AICc.0
coef.AICc.MI.1<-coef.AICc.MI.1 + coef.AICc.1
coef.AICc.MI.2<-coef.AICc.MI.2 + coef.AICc.2
#to find total model coefficient for model selected via BIC after MI#
```

```
coef.BIC.MI.0<-coef.BIC.MI.0 + coef.BIC.0
coef.BIC.MI.1<-coef.BIC.MI.1 + coef.BIC.1
coef.BIC.MI.2<-coef.BIC.MI.2 + coef.BIC.2
} #end of imputation loop#

#to find averaged model coefficient for model selected via AIC after MI#
coef.AIC.MI.0<-coef.AIC.MI.0/length(comp.imp)
coef.AIC.MI.1<-coef.AIC.MI.1/length(comp.imp)
coef.AIC.MI.2<-coef.AIC.MI.2/length(comp.imp)
#to find averaged model coefficient for model selected via AICc after MI#
coef.AICc.MI.0<-coef.AICc.MI.0/length(comp.imp)
coef.AICc.MI.1<-coef.AICc.MI.1/length(comp.imp)
coef.AICc.MI.2<-coef.AICc.MI.2/length(comp.imp)
#to find averaged model coefficient for model selected via BIC after MI#
coef.BIC.MI.0<-coef.BIC.MI.0/length(comp.imp)
coef.BIC.MI.1<-coef.BIC.MI.1/length(comp.imp)
coef.BIC.MI.2<-coef.BIC.MI.2/length(comp.imp)

#to find y.test for model selected via AIC#
AIC.y.est<- coef.AIC.MI.0 + coef.AIC.MI.1*(x1.test)
             + coef.AIC.MI.2*(x2.test)
#to find y.test for model selected via AICc#
AICc.y.est<- coef.AICc.MI.0 + coef.AICc.MI.1*(x1.test)
             + coef.AICc.MI.2*(x2.test)
#to find y.test for model selected via BIC#
BIC.y.est<- coef.BIC.MI.0 + coef.BIC.MI.1*(x1.test)
             + coef.BIC.MI.2*(x2.test)

#to find MSE(p) using y.est for model selected via AIC#
AIC.best<-AIC.best+(AIC.y.est-y.test)^2
AICc.best<-AICc.best+(AICc.y.est-y.test)^2
BIC.best<-BIC.best+(BIC.y.est-y.test)^2

#to find total coefficient over simulation loop#
coef.AIC[1]<-coef.AIC[1] + coef.AIC.MI.0
coef.AIC[2]<-coef.AIC[2] + coef.AIC.MI.1
coef.AIC[3]<-coef.AIC[3] + coef.AIC.MI.2
coef.AICc[1]<-coef.AICc[1] + coef.AICc.MI.0
coef.AICc[2]<-coef.AICc[2] + coef.AICc.MI.1
```

```
coef.AICc[3]<-coef.AICc[3] + coef.AICc.MI.2
coef.BIC[1]<-coef.BIC[1] + coef.BIC.MI.0
coef.BIC[2]<-coef.BIC[2] + coef.BIC.MI.1
coef.BIC[3]<-coef.BIC[3] + coef.BIC.MI.2

#to find total frequency of each model selected over simulation loop#
n.AIC.model[1]<-n.AIC.model[1] + n.AIC.M000
n.AIC.model[2]<-n.AIC.model[2] + n.AIC.M100
n.AIC.model[3]<-n.AIC.model[3] + n.AIC.M010
n.AIC.model[4]<-n.AIC.model[4] + n.AIC.M110
n.AICc.model[1]<-n.AICc.model[1] + n.AICc.M000
n.AICc.model[2]<-n.AICc.model[2] + n.AICc.M100
n.AICc.model[3]<-n.AICc.model[3] + n.AICc.M010
n.AICc.model[4]<-n.AICc.model[4] + n.AICc.M110
n.BIC.model[1]<-n.BIC.model[1] + n.BIC.M000
n.BIC.model[2]<-n.BIC.model[2] + n.BIC.M100
n.BIC.model[3]<-n.BIC.model[3] + n.BIC.M010
n.BIC.model[4]<-n.BIC.model[4] + n.BIC.M110
} #end of simulation loop#

AIC.best<-AIC.best/nsim      #To find averaged MSE(P)#
AICc.best<-AICc.best/nsim
BIC.best<-BIC.best/nsim
coef.AIC<-coef.AIC/nsim      #To find averaged coefficients#
coef.AICc<-coef.AICc/nsim
coef.BIC<-coef.BIC/nsim
```

Appendix E

R-script for Model Selection (STACK) using Multiple Imputation for Logistic Regression

```
n<-100
rho23<-0
rho12<-0
rho13<-0
beta0<-1
beta1<-1
beta2<-1
beta3<-0
k0<-1              #k is number of parameters#
k1<-2
k2<-3
mu<-c(0,0,0)
sigmax<-matrix(c(1,rho12,rho13,rho12,1,rho23,rho13,rho23,1),3,3) #covariance matrix#
nsim<-1000
coef.AIC<-matrix(c(0,0,0), nrow=1, ncol=3)
coef.AICc<-matrix(c(0,0,0), nrow=1, ncol=3)
coef.BIC<-matrix(c(0,0,0), nrow=1, ncol=3)
LL.mat<-matrix(nrow=nsim, ncol=4)
AIC.mat<-matrix(nrow=nsim, ncol=4)
```

```
AICc.mat<-matrix(nrow=nsim, ncol=4)
BIC.mat<-matrix(nrow=nsim, ncol=4)
n.AIC.model<-matrix(c(0,0,0,0), nrow=1, ncol=4)
n.AICc.model<-matrix(c(0,0,0,0), nrow=1, ncol=4)
n.BIC.model<-matrix(c(0,0,0,0), nrow=1, ncol=4)
AIC.best<-matrix(0, nrow=100, ncol=1)
AICc.best<-matrix(0, nrow=100, ncol=1)
BIC.best<-matrix(0, nrow=100, ncol=1)

coef.AIC.MI.0<-0
coef.AIC.MI.1<-0
coef.AIC.MI.2<-0
coef.AICc.MI.0<-0
coef.AICc.MI.1<-0
coef.AICc.MI.2<-0
coef.BIC.MI.0<-0
coef.BIC.MI.1<-0
coef.BIC.MI.2<-0

###to create test values
prob.test<-seq(0.05,1, by=0.1)
z1.test <-qnorm(prob.test)
z2.test <-z1.test
x.test <- matrix(0,nrow=100, ncol=2)
for (ii in 1:10) {
  for (jj in 1:10) {
     x.test[(ii-1)*10+jj, ]<- c(z1.test[ii],z2.test[jj]) }  }

#To find y.test#
beta0<-1
beta1<-1
beta2<-1
x1.test<-x.test[,1]
x2.test<-x.test[,2]
LP.test<-beta0 + beta1*(x1.test)+ beta2*(x2.test)
p.test<-exp(LP.test)/(1+exp(LP.test))
y.test<-rbinom(100, 1, p.test)
```

```
#To find X3#
x3<-matrix(0, nr=100, nc=1)
for (iii in 1:100) {
x1<-matrix(c(x.test[iii,1],x.test[iii,2]),nr=2, nc=1)
mu1<-matrix(c(0,0),nr=2, nc=1)
sigmax<-matrix(c(1,rho12,rho13,rho12,1,rho23,rho13,rho23,1),3,3) #covariance matrix#
sigmax11<-matrix(c(1,rho12,rho12,1),nr=2, nc=2)
sigmax22<-matrix(c(1),nr=1, nc=1)
sigmax12<-matrix(c(rho13,rho23),nr=2, nc=1)
t.sigmax12<-t(sigmax12)                                    #to find transpose#
inv.sigmax11<-solve(sigmax11)                              #to find inverse#
mu2<-(t.sigmax12%*%inv.sigmax11)%*%(x1-mu1)                #to find mean x3#
sigmax2<-sigmax22-(t.sigmax12%*%inv.sigmax11%*%sigmax12)   #to find variance x3#
x3[iii]<-mu2
}
x3.test<-x3
test.values<-data.frame(y.test, x1.test, x2.test, x3.test)

###to run 1000 simulations
for(i in 1:nsim)  { #simulation loop starts#
x<-mvrnorm(n,mu,sigmax)
m<-50
nmiss<-n*(m/100)
nmiss<-round(nmiss)           #m is percentage of missing#
w1<-1/10                      # weight for X1
w2<-(1-(nmiss/n))/10          # weight for X2
W1big<-rep(w1, times=(n*10))
W2big<-rep(w2, times=(n*10))
X1<-x[,1]
x2<-x[,2]
X3<-x[,3]
LP<-beta0 + beta1*(X1)+ beta2*(x2)
p<-exp(LP)/(1+exp(LP))
y<-rbinom(n, 1, p)
x2miss<-rep(NA, times=nmiss)
x2nmiss<-x2[seq(n-nmiss)]
X2<-cbind(c(x2nmiss,x2miss))
#dataset/model for imputation-non-overlapping variable sets#
dat.x<-data.frame(X2,y,X3)
```

```
#define number of multiple imputation, D=10#
imp<-mice(dat.x, method="norm", m=10)
imp1<-complete(imp,1)             # to retrieve the imputation data set 1#
mat<-complete(imp,"long")
imp1<-complete(imp,1)  #to retrieve the imputation data set 1#
imp2<-complete(imp,2)
imp3<-complete(imp,3)
imp4<-complete(imp,4)
imp5<-complete(imp,5)
imp6<-complete(imp,6)
imp7<-complete(imp,7)
imp8<-complete(imp,8)
imp9<-complete(imp,9)
imp10<-complete(imp,10)
comp.imp<-list(imp1, imp2, imp3, imp4, imp5, imp6, imp7, imp8, imp9, imp10)

#to obtain get stacked data after imputation#
ybig<-rep(y, times=10)
x1big<-rep(X1, times=10)
data.xy<-data.frame(ybig, x1big, mat)
M000<-lrm(ybig~1, data.xy)
M100<-lrm(ybig~x1big, data.xy, weights=(W1big))
M010<-lrm(ybig~X2, data.xy, weights=(W2big))
M110<-lrm(ybig~x1big+X2, data.xy, weights=(W2big))
model.list<-list(M000, M100, M010, M110)

#to obtain log-likelihood value from the output for each model#
M000LL<-logLik(M000)
M100LL<-logLik(M100)
M010LL<-logLik(M010)
M110LL<-logLik(M110)
LL<-matrix(c(M000LL,M100LL,M010LL,M110LL),nrow=1,ncol=4)

M000AIC<-(2*(LL[1]))-(2*(k0+1))           #calculating AIC, AICc and BIC#
M000BIC<-(2*(LL[1]))-((k0+1)*log(n))
M000AICc<-(2*(LL[1]))-(2*(k0+1)*(n/(n-k0-2)))
M100AIC<-(2*(LL[2]))-(2*(k1+1))
M100BIC<-(2*(LL[2]))-((k1+1)*log(n))
M100AICc<-(2*(LL[2]))-(2*(k1+1)*(n/(n-k1-2)))
```

```
M010AIC<-(2*(LL[3]))-(2*(k1+1))
M010BIC<-(2*(LL[3]))-((k1+1)*log(n))
M010AICc<-(2*(LL[3]))-(2*(k1+1)*(n/(n-k1-2)))
M110AIC<-(2*(LL[4]))-(2*(k2+1))
M110BIC<-(2*(LL[4]))-((k2+1)*log(n))
M110AICc<-(2*(LL[4]))-(2*(k2+1)*(n/(n-k2-2)))

#to form a matrix of LL for all the models#
LL<-matrix(c(M000LL,M100LL,M010LL,M110LL),nrow=1,ncol=4)
#to form a matrix of AIC for all the models#
AIC<-matrix(c(M000AIC,M100AIC,M010AIC,M110AIC),nrow=1,ncol=4)
#to form a matrix of AICc for all the models#
AICc<-matrix(c(M000AIC,M100AIC,M010AIC,M110AIC),nrow=1,ncol=4)
#to form a matrix of BIC for all the models#
BIC<-matrix(c(M000BIC,M100BIC,M010BIC,M110BIC),nrow=1,ncol=4)

AIC.mat[i,]<-AIC
AICc.mat[i,]<-AICc
BIC.mat[i,]<-BIC
LL.mat[i,]<-LL

max.AIC<- max.col(AIC)           #find the maximum column in AIC matrix#
max.AICc<- max.col(AICc)         #find the maximum column in AICc matrix#
max.BIC<- max.col(BIC)           #find the maximum column in BIC matrix#

model<-list("M000", "M100", "M010", "M110") #find the frequency of selected model#
AIC.model<-model[[max.AIC]]
n.AIC.M000<-ifelse(AIC.model=="M000", 1, 0)
n.AIC.M100<-ifelse(AIC.model=="M100", 1, 0)
n.AIC.M010<-ifelse(AIC.model=="M010", 1, 0)
n.AIC.M110<-ifelse(AIC.model=="M110", 1, 0)
AICc.model<-model[[max.AICc]]
n.AICc.M000<-ifelse(AICc.model=="M000", 1, 0)
n.AICc.M100<-ifelse(AICc.model=="M100", 1, 0)
n.AICc.M010<-ifelse(AICc.model=="M010", 1, 0)
n.AICc.M110<-ifelse(AICc.model=="M110", 1, 0)
BIC.model<-model[[max.BIC]]
n.BIC.M000<-ifelse(BIC.model=="M000", 1, 0)
n.BIC.M100<-ifelse(BIC.model=="M100", 1, 0)
```

```
n.BIC.M010<-ifelse(BIC.model=="M010", 1, 0)
n.BIC.M110<-ifelse(BIC.model=="M110", 1, 0)

for (k in 1:length(comp.imp))  {
dat.xy<-data.frame(y, X1, comp.imp[k])

MM000<-lrm(y~1, dat.xy)
MM100<-lrm(y~X1, dat.xy)
MM010<-lrm(y~X2, dat.xy)
MM110<-lrm(y~X1+X2, dat.xy)
model.list2<-list(MM000, MM100, MM010, MM110)
best.model.AIC<-model.list2[[max.AIC]]
best.model.AICc<-model.list2[[max.AICc]]
best.model.BIC<-model.list2[[max.BIC]]

coef.AIC.0<-coef(best.model.AIC)[1]
coef.AIC.1<-coef(best.model.AIC)[2]
coef.AIC.2<-coef(best.model.AIC)[3]
coef.AICc.0<-coef(best.model.AICc)[1]
coef.AICc.1<-coef(best.model.AICc)[2]
coef.AICc.2<-coef(best.model.AICc)[3]
coef.BIC.0<-coef(best.model.BIC)[1]
coef.BIC.1<-coef(best.model.BIC)[2]
coef.BIC.2<-coef(best.model.BIC)[3]

coef.AIC.0<-ifelse(is.na(coef.AIC.0), 0, coef.AIC.0)
coef.AIC.1<-ifelse(is.na(coef.AIC.1), 0, coef.AIC.1)
coef.AIC.2<-ifelse(is.na(coef.AIC.2), 0, coef.AIC.2)
coef.AICc.0<-ifelse(is.na(coef.AICc.0), 0, coef.AICc.0)
coef.AICc.1<-ifelse(is.na(coef.AICc.1), 0, coef.AICc.1)
coef.AICc.2<-ifelse(is.na(coef.AICc.2), 0, coef.AICc.2)
coef.BIC.0<-ifelse(is.na(coef.BIC.0), 0, coef.BIC.0)
coef.BIC.1<-ifelse(is.na(coef.BIC.1), 0, coef.BIC.1)
coef.BIC.2<-ifelse(is.na(coef.BIC.2), 0, coef.BIC.2)

#to find total model coefficient for model selected via AIC after MI#
coef.AIC.MI.0<-coef.AIC.MI.0 + coef.AIC.0
coef.AIC.MI.1<-coef.AIC.MI.1 + coef.AIC.1
coef.AIC.MI.2<-coef.AIC.MI.2 + coef.AIC.2
```

```
#to find total model coefficient for model selected via AICc after MI#
coef.AICc.MI.0<-coef.AICc.MI.0 + coef.AICc.0
coef.AICc.MI.1<-coef.AICc.MI.1 + coef.AICc.1
coef.AICc.MI.2<-coef.AICc.MI.2 + coef.AICc.2
#to find total model coefficient for model selected via BIC after MI#
coef.BIC.MI.0<-coef.BIC.MI.0 + coef.BIC.0
coef.BIC.MI.1<-coef.BIC.MI.1 + coef.BIC.1
coef.BIC.MI.2<-coef.BIC.MI.2 + coef.BIC.2
} #end of imputation loop#

#to find averaged model coefficient for model selected via AIC after MI#
coef.AIC.MI.0<-coef.AIC.MI.0/length(comp.imp)
coef.AIC.MI.1<-coef.AIC.MI.1/length(comp.imp)
coef.AIC.MI.2<-coef.AIC.MI.2/length(comp.imp)
#to find averaged model coefficient for model selected via AICc after MI#
coef.AICc.MI.0<-coef.AICc.MI.0/length(comp.imp)
coef.AICc.MI.1<-coef.AICc.MI.1/length(comp.imp)
coef.AICc.MI.2<-coef.AICc.MI.2/length(comp.imp)
##to find averaged model coefficient for model selected via BIC after MI#
coef.BIC.MI.0<-coef.BIC.MI.0/length(comp.imp)
coef.BIC.MI.1<-coef.BIC.MI.1/length(comp.imp)
coef.BIC.MI.2<-coef.BIC.MI.2/length(comp.imp)

#to find y.test for model selected via AIC#
LP.AIC.y.est<- coef.AIC.MI.0 + coef.AIC.MI.1*(x1.test)
               + coef.AIC.MI.2*(x2.test)
#to find y.test for model selected via AICc#
LP.AICc.y.est<- coef.AICc.MI.0 + coef.AICc.MI.1*(x1.test)
               + coef.AICc.MI.2*(x2.test)
#to find y.test for model selected via BIC#
LP.BIC.y.est<- coef.BIC.MI.0 + coef.BIC.MI.1*(x1.test)
               + coef.BIC.MI.2*(x2.test)

p.AIC.y.est<-exp(LP.AIC.y.est)/(1+exp(LP.AIC.y.est))
AIC.y.est<-rbinom(100, 1, p.AIC.y.est)
p.AICc.y.est<-exp(LP.AICc.y.est)/(1+exp(LP.AICc.y.est))
AICc.y.est<-rbinom(100, 1, p.AICc.y.est)
p.BIC.y.est<-exp(LP.BIC.y.est)/(1+exp(LP.BIC.y.est))
BIC.y.est<-rbinom(100, 1, p.BIC.y.est)
```

```
#to find MSE(p) using y.est for model selected via AIC#
AIC.best<-AIC.best+(p.AIC.y.est-p.test)^2
AICc.best<-AICc.best+(p.AICc.y.est-p.test)^2
BIC.best<-BIC.best+(p.BIC.y.est-p.test)^2

###to find total coefficient over simulation loop
coef.AIC[1]<-coef.AIC[1] + coef.AIC.MI.0
coef.AIC[2]<-coef.AIC[2] + coef.AIC.MI.1
coef.AIC[3]<-coef.AIC[3] + coef.AIC.MI.2
coef.AICc[1]<-coef.AICc[1] + coef.AICc.MI.0
coef.AICc[2]<-coef.AICc[2] + coef.AICc.MI.1
coef.AICc[3]<-coef.AICc[3] + coef.AICc.MI.2
coef.BIC[1]<-coef.BIC[1] + coef.BIC.MI.0
coef.BIC[2]<-coef.BIC[2] + coef.BIC.MI.1
coef.BIC[3]<-coef.BIC[3] + coef.BIC.MI.2

#to find total frequency of each model selected over simulation loop#
n.AIC.model[1]<-n.AIC.model[1] + n.AIC.M000
n.AIC.model[2]<-n.AIC.model[2] + n.AIC.M100
n.AIC.model[3]<-n.AIC.model[3] + n.AIC.M010
n.AIC.model[4]<-n.AIC.model[4] + n.AIC.M110
n.AICc.model[1]<-n.AICc.model[1] + n.AICc.M000
n.AICc.model[2]<-n.AICc.model[2] + n.AICc.M100
n.AICc.model[3]<-n.AICc.model[3] + n.AICc.M010
n.AICc.model[4]<-n.AICc.model[4] + n.AICc.M110
n.BIC.model[1]<-n.BIC.model[1] + n.BIC.M000
n.BIC.model[2]<-n.BIC.model[2] + n.BIC.M100
n.BIC.model[3]<-n.BIC.model[3] + n.BIC.M010
n.BIC.model[4]<-n.BIC.model[4] + n.BIC.M110
} #end of simulation loop#

AIC.best<-AIC.best/nsim     #To find averaged MSE(P)#
AICc.best<-AICc.best/nsim
BIC.best<-BIC.best/nsim
coef.AIC<-coef.AIC/nsim     #To find averaged coefficients#
coef.AICc<-coef.AICc/nsim
coef.BIC<-coef.BIC/nsim
```

Bibliography

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

Andridge, R. R. and Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1):40–64.

Brand, J. P. L. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets.* PhD thesis, Erasmus University, Rotterdam.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, 53(2):603–618.

Burham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A practical Information-Theoretic Approach.* Springer-verlag New York, Inc., New York.

Chatterjee, S. and Simonoff, J. S. (2013). *Handbook Regression Analysis.* John Wiley & Sons, Inc, New Jersey.

Chaurasia, A. and Harel, O. (2012). Using AIC in multiple linear regression framework with multiply imputed data. *Health Services and Outcomes research Methodology*, 12:219–233.

Chen, Q. and Wang, S. (2013). Variable selection for multiply imputed data with application to dioxin exposure study. *Statistics in Medicine*, 32:3646–3659.

Claeskens, G. and Consentino, F. (2008). Variable selection with incomplete covariate data. *Biometrics*, 64:1062–1069.

Claeskens, G. and Hjort, N. L. (2008). *Model selection and model averaging.* Cambridge University Press, Cambridge.

Clark, T. G. and Altman, D. G. (2003). Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *Journal of Clinical Epidemiology*, 56:28–37.

Collins, L. M., Schafer, J. L., and Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4):330–351.

Consentino, F. and Claeskens, G. (2011). Missing covariates in logistic regression, estimation and distribution selection. *Statistical Modelling*, 11(2):159–183.

Davison, A. C. (2003). *Statistical Models*. Cambridge University Press, New York.

Donner, A. (1982). The relative effectiveness of procedures commonly used in multiple regression analysis dealing with missing values. *The American Statistician*, 36(4):378–381.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carre, G., Marquez, J. R. G., Gruber, B., Lafourcade, B., Leitao, P. J., Munkemuller, T., McClean, C., Osborne, P. E., Reineking, B., Schroder, B., Skidmore, A. K., Zurell, D., and Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecograpyhy*, 36:027–046.

Everitt, B. S. (2006). *The Cambridge Dictionary of Statistics*. Cambridge University Press, Cambridge, third edition.

Freeman, J., Cole, T., Chinn, S., Jones, P., White, E., and Preece, M. (1995). Cross sectional stature and weight reference curves for the UK, 1990. *Archives of Disease in Childhood*, 73(1):17–24.

Garcia, R. I., Ibrahim, J. G., and Zhu, H. (2010). Variable selection for regression models with missing data. *Statistica Sinica*, 20(1):149–165.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, 75:1175–1189.

Hardt, J., Herke, M., and Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing X: A warning against including too many in small sample research. *BMC Medical Research Methodology*, 12:184.

Harrell, F. E. (2001). *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis*. Springer, Inc., New York.

Hens, N., Aerts, M., and Molenberghs, G. (2006). Model selection for incomplete and design-based samples. *Statistics in Medicine*, 25:2502–2520.

Heymans, M. W., van Buuren, S., Knol, D. L., van Mechelen, W., and Vet, H. C. W. (2007). Variable selection under multiple imputation using the bootstrap in prognostic study. *BMC Medical Research Methodology*, 7:33.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98:879–945.

Hocking, R. R. (2003). *Methods and Applications of Linear Models.* John Wiley & Sons, New Jersey, second edition.

Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76:297–307.

Hurvich, C. M. and Tsai, C. L. (1990). The impact of model selection on inference in linear regression. *American Statistician*, 44:214–217.

Ibrahim, J. G., Zhu, H., and Tang, N. (2008). Model selection criteria for missing-data problems using EM algorithm. *Journal of the American Statistical Association*, 103(484):1648–1658.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data.* John Wiley and Sons, Inc., New Jersey, second edition.

Lu, F. and Petkova, E. (2014). A comparative study of variable selection methdos in the context of developing psychiatric screening instuments. *Statistics in Medicine*, 33:401–421.

Maghsoudi, S. H., Haghdoost, A. A., and Baneshi, M. R. (2014). Selection of variables that inference drug injection in prison: Comparison of methods with multiple imputed data sets. *Addiction and Health*, 6(1-2):36–44.

Mela, C. F. and Kopalle, P. K. (2002). The impact of collinearity on regression analysis: The asymmetric effect of negative and positive correlations. *Applied Economics*, 34:667–677.

Mevik, B. H. and Cederkvist, H. R. (2004). Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). *Journal of Chemometrics*, 18(9):422–429.

Moons, K. G. M., Donders, R. A. R. T., Stijnen, T., and Harrell, F. E. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*, 59:1092–1101.

Nagakawa, S. and Freckleton, R. P. (2011). Model averaging, missing data and multiple imputation: a case study for behavioural ecology. *Behavioural Ecology and Sociobiology*, 65:103–116.

Nakagawa, S. and Freckleton, R. P. (2008). Missing inaction: the dangers of ignoring missing data. *Trends in Ecology and Evaluation*, 23:592–596.

Osborne, J. W. (2013). *Best Practise in Data Cleaning*. Sage Publication, Inc., California.

Patrician, P. A. (2002). Focus on research methods: Multiple imputation for missing data. *Research in Nursing and Health*, 25:76–84.

Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation*, 7(4):353–383.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, 47(3):537–560.

Royston, P. and White, I. R. (2011). Multiple imputation by chained equations (mice): Implementation in stata. *Journal of Statistical Software*, 45(4):1–20.

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

Rust, R. T., Simester, D., Brodie, R. J., and Nilikant, V. (1995). Model selection criteria: An investigation of relative accuracy, posterior probabilities and combination of criteria. *Management Science*, 41(2):322–333.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman and Hall, London.

Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, 33(4):545–571.

Schomaker, M. and Heumann, C. (2011). Model averaging in factor analysis: an analysis of Olympic decathlon data. *Journal of Quantitative Analysis in Sports*, 7(1):1–15.

Schomaker, M. and Heumann, C. (2014). Model selection and model averaging after multiple imputation. *Computational Statistics and Data Analysis*, 71:758–770.

Schomaker, M., Wan, A. T. K., and Heumann, C. (2010). Frequentist model averaging with missing observations. *Computational Statistics and Data Analysis*, 54:336–3347.

Sinharay, S., Stern, H. S., and Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, 6(4):317–329.

Studenmund, A. H. (2006). *Using Economics: A Practical Guide*. Addison-Wesley, New York, fifth edition.

Symonds, M. R. E. and Moussalli, A. (2011). A brief guide to model selection, multimodal inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioural Ecology and Sociobiology*, 65:13–21.

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. CRC Press, Boca Raton.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–31.

Verbeke, G., Molenberghs, G., and Beunckens, C. (2008). Formal and informal model selection with incomplete data. *Statistical Science*, 23(2):201–218.

Vergouw, D., Heymans, M. W., Peat, G. M., Kuijpers, T., Croft, P. R., de Vet, H. C. W., van der Horst, H. E., and van der Windt, D. A. W. M. (2010). The search for stable prognostic models in multiple imputed data sets. *BMC Medical Research Methodology*, 10:81.

Vergouwe, Y., Royston, P., Moons, K. G. M., and Altman, D. G. (2010). Development and validation of a prediction model with missing predictor data: a pratical approach. *Journal of Clinical Epidermiology*, 63:205–214.

Wallach, D. and Goffinet, B. (1989). Mean squared error of prediction as a criterion for evaluating and comparing system models. *Ecological Modelling*, 44:299–306.

Wan, Y., Datta, S., Conklin, D. J., and Kong, M. (2015). Variable selection models based on multiple impuatation with an application for predicting median effective dose and maximum effect. *Journal of Statistical Computation and Simulation*, 85(9):1902–1916.

White, I. R. and Royston, P. (2009). Imputing missing covariate values for the cox model. *Statistics in Medicine*, 28:1982–1998.

White, I. R., Royston, P., and Wood, A. M. (2011). Muliple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399.

Wood, A. M., White, I. R., and Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, 27:3227–3246.

Wright, M. W., Cox, K. M., Sherriff, A., Franco-Villoria, M., Pearce, M. S., Adamson, A. J., and core team, G. M. S. (2011). To what extent do weight gain and eating avidity during infancy predict later adposity? *Public Health Nutrition*, 15(4):656–662.

Yang, X., Belin, T. R., and Boscardin, W. J. (2005). Imputation and variable selection in linear regression models with missing covariates. *Biometrics*, 61:498–506.

Yu, S. S., Gelman, A., Hill, J., and Yajima, M. (2011). Multiple imputation with diagnostics (mi) in r: Opening windows into the black box. *Journal of Statsitical Software*, 45(2):1–31.