Population as a Cohort: Increasing statistical power in a cohort analysis using register data in addition to population survey datasets

A Thesis

submitted to
Indian Institute of Science Education and Research Pune
in partial fulfillment of the requirements for the
BS-MS Dual Degree Programme

by

Arya P V

Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA.

December, 2020

Supervisor: Dr. Tommi Härkänen
Co-Supervisor: Professor Sangita Kulathinal

Certificate

This is to certify that this dissertation entitled Population as a Cohort: Increasing statistical power in a cohort analysis using register data in addition to population survey datasetstowards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Arya P V, BS-MS student at Indian Institute of Science Education and Research under the supervision of Dr. Tommi Härkänen, Research Manager and Senior statistician at Finnish Institute for Health and Welfare (THL), Helsinki, Finland, Adjunct Professor at Department of Mathematics and Statistics, University of Helsinki, Finland and Professor Sangita Kulathinal, Professor of statistics at Department of Mathematics and Statistics, University of Helsinki, Finland , during the academic year 2015-2020.

Tommi Härkänen

Dr. Tommi Härkänen

Committee:

Dr. Tommi Härkänen

Professor Sangita Kulathinal

Dr. Anindya Goswami

This thesis is dedicated to my mother, Pushpa for her endless love and support

Declaration

I hereby declare that the matter embodied in the report entitled Population as a Cohort: Increasing statistical power in a cohort analysis using register data in addition to population survey datasets are the results of the work carried out by me at the Finnish Institute for Health and Welfare (THL), Helsinki, Finland and at Department of Mathematics and Statistics, University of Helsinki, Finland, under the supervision of Dr. Tommi Härkänen, Professor Sangita Kulathinal and the same has not been submitted elsewhere for any other degree.

Arya P V

Acknowledgments

I would like to express my sincere gratitude to Professor Sangita Kulathinal and Dr Tommi Härkänen for providing me with an excellent opportunity to work on such an important and exciting project, which not only is of great significance but also has wide-reaching applications in statistics and data science. I am also grateful to Professor Sangita Kulathinal for helping me through the courses which were required at the early stages of the project and assisted in bringing the project to a good pace right after it's commencement. I would also like to show my appreciation towards Professor Uttara Naik Nimbalkar for her guidance thoughout my past two years. I am indebted to the National Institute for Health and Welfare, Helsinki, Finland and the University of Helsinki for providing me financial support and data access during of the course of this project. I am also thankful to Tossavainen Sanna for helping me with the paperwork related to the work and finding a good accommodation. I am incredibly grateful to my parents and siblings for always being there for me with their relentless support and encouragement. Thanks to my friend Akhil Mithran whom I could always count on.

Abstract

Individual-level records from health and social services are routinely being generated, collected and maintained centrally in nation-wide registers. These records, when combined with a cohort study/survey, may increase the statistical power of association between the outcome and risk factors. The biggest challenge in combining data from the population and the survey is missing risk factor data. Methods to handle missing data within the survey are well developed and widely used. Multiple Imputation (MI) is one such widely used methods of handling missing data.

MI is popular because it avoids the potential bias and efficiency loss resulting from a complete-case analysis (CCA). This thesis studies how MI handles missing data in comparison with CCA for different types of covariates such as continuous and categorical covariates, for time-to-event and binary outcomes data. It also discusses the ways to include the time-to-event data in the presence of right censoring and delayed entry in the imputation model. Furthermore, an empirical study has conducted on the population-level ischemic heart disease event data provided by the Finnish Institute for Health and Welfare (THL), that contains missing data in the selected risk factors. The overall results show that the MI method, with a sufficient number of imputation and iterations, is preferred in most scenarios.

Keywords; Missing data, Multiple imputation, time to event data, delayed entry, right censored data, imputation model, complete case analysis, illness death model.

Contents

Introduction

Missing data or incomplete data inevitably occur in survey-based research when the response of the respondents is not recorded. It may be due to the intentional refusal to answer some particular survey questions because of reasons pertaining to privacy and lack of awareness or even unwillingness to answer, thinking it as a waste of time.

Although the source of missing values may differ from case to case, the troubles resulting from them are similar. Missing data are a prevailing problem in clinical-based studies and observational studies, especially in longitudinal ones, and often the data deficiency occurs in covariates. This is mainly due to the fact that many of the associated covariates or risk factors are often measured or recorded using surveys that in turn, will have an abundant amount of missing data. However, survival outcomes such as death and disease times are usually recorded in hospitals or some other health registers since it can be measured easily and accurately. For example, in surveys that collect the data of risk factors such as smoking status or alcohol consumption measurements, participants might be reluctant to give their daily usage values. Considering the expense of collecting the data in surveys, starting over again to minimise or to completely eliminate missing data is practically not feasible.

Rubin (1987) [4] proposed multiple imputation (MI) method to provide statistical inference, which has then become a popular method for handling missing data. Further studies by Rubin (1996) [12] described that, for the ultimate users who in general have access only to complete-data, the MI by the database constructor is the method of choice. The basic notion behind MI is substituting each of the missing values by multiple plausible values obtained from the distribution of the observed data. As a result, multiple complete data sets with the same non-missing part but different missing part are generated. Performing MI on survival data brings up extra efforts than usual because of the inclusion of time to event (TTE) data as a predictor for the imputation of a missing covariate. Several research papers proposed

that including the event indicator D and the log of the observed event or censoring time T as predictors in the imputation model will results in reliable imputations of the missing values. However, White and Royston(2009) [13] demonstrated using simulation studies that the usage of Nelson-Aalen estimate of the cumulative hazard along with the event indicator in the imputation model shows the best results.

MI has the ability to incorporate all sources of variability and uncertainty. Hence, most of the time, the MI method is capable of making valid inference from incomplete data. In prognostic research areas, Multivariate Imputation with Chained Equations (MICE) is currently considered the golden standard. MICE is a special MI technique which imputes multivariate missing data in a variable by variable basis in a smart and flexible way. MICE is an iterative approach, in which each of the incomplete covariates is imputed one at a time based on a unique regression model specified by the user and using the other covariates as predictors. In cases where the incomplete data are present in covariates of the analysis model, it is necessary to include the outcome as a predictor of the imputation model as well. Complete case analysis is another easy way of handling missing data. It is done by selecting those rows which do not have any missing values in it, i.e., including only the participants for which we have no missing data on the covariates of interest.

In this thesis, we assess the efficiency of the MI method on handling the missing data present in the categorical and continuous covariates for the analysis model that have a binary response and then for the TTE response through simulation studies. With the simulation studies, we are also comparing the accuracy of the estimates obtained by handling the missing data using MI to CCA. From the results obtained from the simulation studies, we are implementing both the MI and CCA method to handle the missing data in the Ischemic heart disease(IHD) data set of the Finnish population and modelling it using a progressive illness death model.

Chapter 1

Missing Data

1.1 Missing data or incomplete data

The problem of having missing data and its corresponding difficulties frequently occur in statistical analysis as well as in many fields of research. However, disregarding the process and cause behind the missing data often bring along some statistical issues such as bias and loss of efficiency. Missing data mechanism by Rubin(1976) [2] helps to give a clear idea about the process behind missing data mechanism. Let X be a $n \times p$ matrix that contains partially observed values in the data set and let R be the indicator vector indicating the missingness in X, i.e. $R = 0$ indicates missing values. The general expression denoting the missing data model is $P(R|X^{obs}, X^{miss}, \theta)$ where θ denotes the parameter vecor associated with missing data model, X^{miss} denotes the missing part and X^{obs} denotes the non missing part of X i.e. $X = (X^{miss}, X^{obs})$. The missing data model describes the level of dependence of the distribution of R on X

1.1.1 Missing completely at random (MCAR)

This is the case where the missingness of the data is independent of the observed and unobserved data. Here, the likelihood of having missing values in the data only depends on

the net probability of being missing, θ:

$$P(R = 0|X^{obs}, X^{miss}, \theta) = P(R = 0|\theta) \tag{1.1}$$

In other words, there does not exist any systematic differences between participants that have missing data and those with complete data. However, since the MCAR assumptions restrictive, it is the most ideal to hold in the real life scenarios An instance of MCAR is the inability of the participants to attend the survey due to severe rain.

1.1.2 Missing at random (MAR)

This happens when the missingness of data is systematically related to the observed but not to the unobserved data. These types of missingness occur when the reason for the missingness is correlated to an observed variable or some of the observed variables. Here, the missingness probability differs between groups however probability of having missing values in the group is same. Hence, for the case at hand, the likelihood of having missing values depends on both the parameters θ and the observed values,

$$P(R = 0|X^{miss}, X^{obs}, \theta) = P(R = 0|X^{obs}, \theta) \tag{1.2}$$

For instance, in a survey, the female participants will be more reluctant to respond to a question related to their daily alcohol consumption measurements.

1.1.3 Missing not at random (MNAR)

This happens when the missingness of the data is systematically related to the unobserved data, i.e., the missingness is related to events or factors which are not measured by the researcher. Under the MNAR condition, the likelihood of having missing data differs between groups or even between the individual points in the same group due to reasons which depends on missing information. Hence, the general missing data model cannot be simplified in this case.

$$P(R = 0|X^{miss}, X^{obs}, \theta) \tag{1.3}$$

An example is that the participants with high alcohol consumption are most likely to refuse to answer the questionnaire related to alcohol consumption measurements. Hence, the probability of having non-response for such cases depends on the missingness of the high alcohol consumption values.

These categorization of missing data gives a better insight about how to tackle this problem and how to proceed further with the analyses.

1.2 Methods for Handling missing data

Several researchers have proposed different methods for dealing with missing data problems such as mean substitution, regression substitution, MI, listwise deletion/CCA, pairwise deletion and expectation-maximization technique etc. Among these methods, we are particularly focusing on MI and CCA, which is found to be popular and widely used methods to tackle the missing data problems.

1.2.1 Complete case analysis (CCA)

CCA is used to describe a statistical method that only includes participants for which we have no missing data on the covariates of interest in the analysis. This means that the participants with any missing data are excluded. CCA is preferred because it is the easiest and more computationally efficient among all the methods available for handling missing data. In cases of MCAR, CCA is accepted as the best method for handling it, since the parameter estimation remains unbiased even by the absence of data that is unobserved. When the exposure or confounders in the main analysis are MNAR, CCA is a valid approach since it gives a less biased result even when compared to other sophisticated techniques such as MI. It may give biased results under the MAR assumption since the chance of being a complete case depends on the observed values of the outcome. This method works efficiently when only a small proportion of data are missing. However, in most other cases discarding cases may lead to a reduction of statistical power.

1.2.2 Multiple Imputation (MI)

Rubin (1978 and 1987) [3, 4] suggested MI method which is now one of the most commonly used approaches in handling the missing data problem under the assumption of MAR. It is based on the Bayesian paradigm that involves drawing missing values from a posterior predictive distribution conditioned on the observed data. The central concept behind it is to replace each missing values by a set of "m" imputed values which in turn generates m different data sets with the same non-missing part but different missing parts. Based on Schafer (1999) [5], a small number of imputation, m would be enough, which generally ranges from 5 to 10 times. But recently, White et al. (2011) [14] suggested in their study that m should be at least equal to the percentage of incomplete cases in the data. This is now considered as the standard rule for choosing the appropriate m. MI technique consists of three main steps:

1. Generating multiply imputed datasets: Consider the simplest case where there is only one incomplete variable X while the rest of the variables Z and the response Y are complete. The first step is the construction of the imputation model $f(X|Z, Y; \beta)$ from the observed X, and drawing estimated $\hat{\beta}$ values with their variance-covariance matrix S_β from this model. Then from $N(\hat{\beta}, S_\beta)$, β's are drawn and thereafter from the posterior predictive distribution, $f(X|Z, Y; \beta')$ the missing values are generated. Now, the whole process is repeated m times to generate (m) multiple imputed data sets.

2. Analysing the multiply imputed datasets: After the imputation, m complete data sets are generated and analyses are done separately to the m imputed data sets in order to obtain the parameter estimate of interest.

3. Pooling the estimates: The estimates obtained from each of the m imputed data sets are combined or pooled to obtain a single parameter. The method of pooling is done using Rubin's rules, i.e. let $\alpha^{(k)}$ be the point estimate of the kth imputed data set $(k = 1, \ldots, m)$, $W^{(k)}$ be the estimated variance of $\alpha^{(k)}$ and B be the variance between the imputations of the same covariate. Then the pooled estimate is given by:

$$\hat{\alpha} = \frac{1}{m} \sum_{k=1}^{m} \alpha^{(k)} \tag{1.4}$$

and the overall variance is:

$$Var(\hat{\alpha}) = W + (1 + \frac{1}{m})B \tag{1.5}$$

where W and B are the corresponding "within" and "between" imputation variance given by;

$$W = \frac{1}{m}\sum_{k=1}^{m} W^{(k)} \tag{1.6}$$

$$B = \frac{1}{m-1}\sum_{k=1}^{m} (\alpha^{(k)} - \hat{\alpha})^2 \tag{1.7}$$

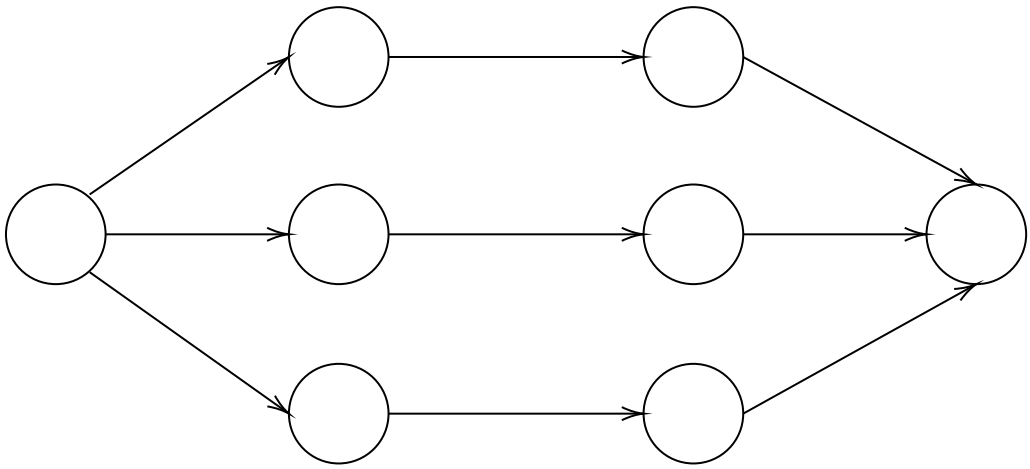


Figure 1.1: The schematic diagram representing the 3 stages of MI method

Rubin (1987)[4] mentioned that the advantage of imputation especially with the MI method is it's ability to incorporate standard complete data methods of analysis on the imputed data sets and it's ability to incorporate the data collector's knowledge. Another visible advantage is it's increased efficiency to make valid inferences as compared to other single imputation methods. Furthermore, when combining the imputed complete data inference according to Rubin's rule, it incorporate all sources of variability and uncertainty, in the form of within imputation and between-imputation variance.

Chapter 2

Multiple Imputation in multivariate missing data

Missing data may be present in different kinds of data, which includes multivariate data as well. Let $X = (X_1, .., X_p)$ be the vector of covariates of interest. In this multivariate setting, the imputation of the variable X_j that contains missing data will requires all or some of the predictors in X_{-j}; where X_{-j} represents all the covariates in X other than X_j and $j \in (1, .., p)$.

There are several practical problems associated with the imputation of missing values in these types of data. This includes a circular dependence, when the missing data in two incomplete covariates are dependent due to the high correlation between those covariates and a possible occurrence of collinearity, when p is large and the size of the data n is small. Various approaches are used in MI to overcome these difficulties, among which the most commonly used ones are the Joint modeling (JM) and the Fully Conditional Specific (FCS).

2.1 Joint Modeling (JM)

Joint modeling (JM) works under the assumption that a multivariate distribution can fully describe the data. Imputations for the missing values are created as draws from the fitted

multivariate distribution conditioned on the observed data, which is used as the imputation model. The imputation model can be based on any multivariate distribution. However, the multivariate normal distribution is the one most widely applied.

JM primarily involves specifying a multivariate distribution for the missing data and drawing imputation from their conditional distributions by Markov Chain Monte Carlo (MCMC) technique. Hence the basic idea behind this approach is that, in a general missing data pattern, missing data can be in anywhere in X. As a result, the distribution from which imputations are to be drawn varies from row to row. For instance, let the missingness pattern in the i^{th} row be $r_{(i)} = (0, 0, 1, 1)$. Here, the imputations are drawn from a bivariate distribution $P_i(X_1^{miss}, X_2^{miss}|X_3, X_4, \theta_{1,2})$ whereas when $r(i') = (0, 1, 1, 1)$, it needs to be drawn from the univariate distribution $P_i(X_1^{miss}|X_2, X_3, X_4, \theta_1)$. The JM method is not preferred much due to its ideal assumption of the having a multivariate model for imputation, i.e it is not robust to misspecification of model which sometimes hard to cope with.

2.2 Fully Conditional Specification (FCS)

FCS method impute missing data present in the multivariate case in a variable-by-variable manner (Van Buuren et al. 2006 [15], Van Buuren 2007 [16]). In this method, the specification of an imputation model is required for each variable and subsequently, imputations are created in an iterative manner.

Contrary to the joint modeling, FCS assumes the existence of a multivariate joint distribution which may or may not actually exist, from the individual univariate imputation model specified for each of the missing variables, i.e. let X be the set of incomplete variables, R be the missingness indicating vector, Z be the rest of the variables and Y be the response that is observed completely. In theory, FCS describes the multivariate distribution $P(X, Z, Y, R|\phi)$ from individual conditional distribution $P(X_j|X_{-j}, Z, Y, R, \theta)$ specified separately for each of the missing variables X_j. This conditional distribution $P(X_j|X_{-j}, Z, Y, R, \theta)$ is the one used for the imputation of the missing values in variable X_j conditioned on the rest of the missing variable X_{-j}, Z, R and Y.

The algorithm starts by substituting the missing values in each X_j by some simple single imputation strategy. Then the algorithm proceeds by repeatedly imputing the missing val-

ues in each variable in an iterative fashion, at each stage conditioning on the most recent imputations of the other missing variables. Let $x_j = (x_j^{obs}, x_j^{miss})$, y and z be the elements in the variables X, Y and Z respectively. Then the t^{th} iteration is given by

$$\begin{aligned}
\theta_1^t &\sim f(\theta_1) f(x_1^{obs} | x_{-1}^{(t)}, z, y, \theta_1) \\
x_1^{miss(t)} &\sim f(x_1^{miss} | x_{-1}^{(t)}, z, y, \theta_1^{(t)}) \\
\theta_2^t &\sim f(\theta_2) f(x_2^{obs} | x_{-2}^{(t)}, z, y, \theta_2) \\
x_2^{miss(t)} &\sim f(x_2^{miss} | x_{-2}^{(t)}, z, y, \theta_2^{(t)}) \\
&\vdots \\
\theta_p^t &\sim f(\theta_p) f(x_p^{obs} | x_{-p}^{(t)}, z, y, \theta_p) \\
x_p^{miss(t)} &\sim f(x_p^{miss} | x_{-p}^{(t)}, z, y, \theta_p^{(t)})
\end{aligned}$$

where $x_j^{(t)} = (x_j^{Obs}, x_j^{(t)})$ denote the vector of observed and imputed values at the t^{th} iteration. After the cycle reaches convergence, the current draws are taken as the first set of imputed values. The cycle is then repeated until the desired number of imputations has been achieved.

Similarly, multiple number (m) of imputed data sets are created and the statistical quantity of interest is estimated by following the three steps of Rubin's rule. The `mice` package designed by Van Buuren(2012) [1] performs MI in the FCS perspective, which initiates the iteration by replacing the missing values by randomly selected observed values from the same variable.

2.2.1 Convergence

The flexibility of mice algorithm defines different imputation models for different variables. This is beneficial if there are particular properties of the data that requires to be preserved. Even so, this flexibility can also cause problems in the estimation. Defining different imputation model can cause slow convergence or non-convergence of the imputation model.
So assessing convergence of the imputation model is one of the crucial steps, which should be

done for each imputed variables, but specifically for those variables with a high proportion of missingness. The convergence of the imputation model implies that the data augmentation algorithm has reached an appropriate stationary posterior distribution.

Convergence is mostly checked visually from the trace plots which is plotted for all or selected parameters against the iteration number. Long term trends in trace plots are the indicates a slow convergence to stationary. A stationary process has its mean and variance unchanging over time. On convergence, the different streams of curves denoting each imputation should be freely intermingled together, without showing any definite trends

Similar algorithms with the FCS have been used by the researchers under different names: stochastic relaxation, variable-by-variable imputation, switching regruch asessions, sequential regressions, ordered pseudo-Gibbs sampler, partially incompatible MCMC, iterated univariate imputation, chained equations etc.

Chapter 3

Imputation models

Consider the simplest case when we have only one missing variable X, a vector of completely observed outcome Y and rest of the complete covariate vectors Z. Here, the imputation model can be assumed as $P(X|Y, Z, \theta)$. Then, the MI formally involves drawing the missing values in X^{miss}, where $X = (X^{miss}, X^{miss})$, from a predictive probability distribution

$$P(X^{miss}|X^{obs}, Y, Z) = \int P(X^{miss}|X^{obs}, Y, Z; \theta)P(\theta|X^{obs}, Y, Z)d\theta \tag{3.1}$$

Which in practice, obtained by fitting the model $P(X|Y, Z, \theta)$, called the imputation model based on the observed cases X and yields an estimate $\hat{\theta}$ with a variance covariance matrix S_θ. Thereafter the values of θ, θ^*'s are drawn from its posterior which can be approximated as $\mathcal{N}(\hat{\theta}, S_\theta)$. Then at the final step the imputations of X^{miss} are drawn from the distributions $P(X|Y, Z; \theta^*)$.

Method	type
Predictive mean matching	numeric
Bayesian linear regression	numeric
Linear regression, non-Bayesian	numeric
Unconditional mean imputation	numeric
Two-level linear model	numeric
Logistic regression	factor,
Multinomial logit model	factor, >2 levels
Ordered logit model	ordered, >2 levels
Linear discriminant analysis	factor
Classification and regression trees (CART)	any
Random sample from the observed data	any

Table 3.1: Imputation models for the different types of covariates

However, these whole processes gets complicates when we have missing data in the covariates of a survival outcome / time to event outcome model.

3.1 Survival analysis a general overview

Survival analysis is a branch of statistics that deals with the study of data on times of events in individual life histories. The major component in survival analysis are,

Definition 3.1.1. *The time-to-event, T is a random variable which measures the time duration between the starting time and observation time of the event or the censoring time. In general $T \geq 0$ and $T = 0$ at the start event.*

3.1.1 Censoring of a data

Censoring in a data happens when we are not able to collect the complete set of data for a subject due to various reasons such as time limitation due end of the study period. In most of the cases, it is impossible to avoid data censoring. Hence, it is important to understand

and incorporate it in the model as well. When the observation time of the event of interest is independent of censoring then it's called independent censoring.

Definition 3.1.2. *Right censoring: Censoring that occurs before the observation of event i.e. the event is observed after the end of the study. Let C is the variable indicating censoring time. Assume T, the time at which the event observed and C are independent. Defined as,*

$$U := \min(T, C) \tag{3.2}$$

and the censoring indicator is given by,

$$D = \begin{cases} 1 & \text{if data is uncensored } (t \leq C) \\ 0 & \text{if data is censored } (t > C) \end{cases} \tag{3.3}$$

Remark 3.1.1. *If censoring is informative i.e it gives information about the time of event, then the model must include it as a random event C, which makes the analysis complicated. Hence, it is enough to show that whether the censoring is stochastically independent of T, then we can say C is uninformative.*

3.1.2 Fundamental functions in survival analysis

Given the time to event random variable T,

Definition 3.1.3. *1. Survival Function defines the probability of observing the event after the time t*

$$S(t) = P(T \geq t) \tag{3.4}$$

2. Hazard Function denotes the instantaneous rate of occurrence of the event, defined as

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt}. \tag{3.5}$$

3. Cumulative hazard $H(t)$ is the summation of all hazard rates until time point t

$$\begin{aligned} \text{For discrete } T : H(t) &= \sum_{t_i \in T, t_i \leq t} h(t) \\ \text{For continuous } T : H(t) &= \int_0^t h(u) du \end{aligned} \tag{3.6}$$

The Nelson-Aalen estimator is a non-parametric method which is used to estimate the cumulative hazard rate function from censored survival data.

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i} \tag{3.7}$$

where d_i denotes the number of individuals who event of interest happened at t and n_i is the number of individuals at risk of having the event just prior to time t_i

3.1.3 Proportional hazard model

Modeling a survival data is the next crucial step in survival analysis. In most cases a proportional hazard model (PH) is used for regression due to its simplicity.

Definition 3.1.4. *Proportional Hazard model: It is a semi-parametric method for estimating the hazard function. The core assumption of a proportional hazard model is that all individuals have the same hazard function with a unique scaling factor. Hence the shape of the hazard function curve is the same for all individuals, and only a scalar multiple changes per individual. Let* $\mathbf{x_i} = (x_{1i}, x_{2i}, .., x_{pi})$ *are the covariates if an individual* i *and* $h_0(t)$ *is the baseline hazard at* t*,*

$$h(t, x_i) = h_0(t) \times \exp(\sum_{j=1}^{p} \beta_j x_{ij}) \tag{3.8}$$

i.e. hazard at t *for given* x_i = (*baseline hazard at* t) $\times$ (*Risk factor*; $\exp(\beta \mathbf{x_i})$)*. Further more, the ratio of this with another subject, by keeping all the covariates except* k*th covariate* ($k \leq p$) *constant for both and* $x_{ik} - x_{jk} = 1$*, is called the hazard ratio and will looks:*

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t) \exp(\sum_{l=1}^{p} \beta x_{il})}{h_0(t) \exp(\sum_{l=1}^{p} \beta x_{jl})} = \frac{\exp(\beta x_{ik})}{\exp(\beta x_{jk})} = \exp(\beta) \tag{3.9}$$

i.e. in case of this multiplicative hazards model, $\exp\{\beta\}$ *is called the hazard ratio.*

3.1.4 Multistate illness death model

Multi-state models are the models often used for describing the development of longitudinal failure time data. A multi-state model is defined as the model for the stochastic process,

which at any time point occupies in one of a set of discrete states. The change of state is called a transition, or also called as an event. However, it is essential to distinguish between an event, e.g. death and a state such as dead. The state structure defines the states and which transitions from a state to, to a state are possible. It is possible to make a figure of the state structure as in the Figure 3.1. The full statistical model specifies the state structure and the form of the hazard function for each possible transition.

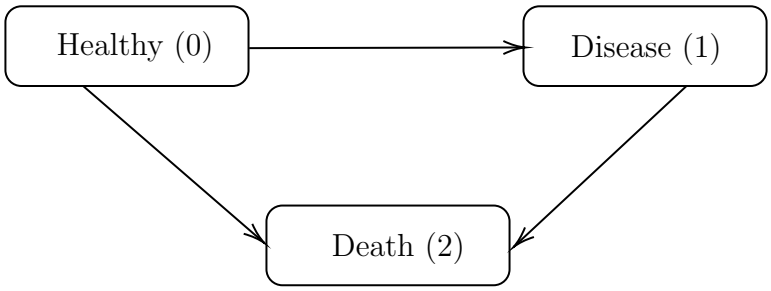


Figure 3.1: The pictorial representation of the multistate illness death model

The progressive illness death model also called the disability model, is a multistate model with three states. The illness death model of Figure 3.1 can be constructed by splitting the pathways to the dead state into two, based on whether the death happened from the healthy or the diseased state. The interpretation of this figure is that the current state contains the information on how many and which states have been visited previously, and the order they have been visited in, but not the times of transitions. This is the most commonly used model in the medical research areas.

3.2 Missing data in covariates of a survival outcome data

The main aim behind the MI technique is to impute the missing value so that the uncertainty of the imputed value is also accounted for. So, the imputed values are the plausible values generated from the predictive distribution rather than the actual value if it is not missing. Therefore, the variance of the predictive distribution is incorporated in the further analysis. MI method uses a selected model such as the regression model in the prediction of missing values based on observed data. In order to incorporate the uncertainty, instead of picking

one value, many values are chosen stochastically for each of the missing values, and the uncertainty is described in the variance-covariance matrix of the estimates that are used to predict missing values. The crucial step for carrying out appropriate imputation lies in the selection of right imputation model.

The imputation of missing data in covariates of a survival model is usually found to be challenging. For handling survival data sets that have missing values in the covariates by the MI method is done by including the time to event outcome in the imputation model. However, the inclusion of time to event data should be done carefully since the outcome is the pair of observation, the length of time during which no event was observed T (or the follow-up time t_i if the event is not observed throughout the study) and an event indicator of whether the end of that time period corresponds to an event or just the end of the observation D. Several research papers proposed different ways to include time to event outcome into the imputation model. A prominent research paper by Vann Buure et al(1999) [9] used the D, T and $log(T)$ as the predictors in the imputation model. Some researchers used D and $log(T)$, or D and T in their imputation models; however, the rationale behind all these usages are still not clear. White and Royston(2009) [13] demonstrated that inclusion of Nelson-Aalen estimate of cumulative hazard and the event indicator in the imputation model is the best method. They have also demonstrated that the suitable imputation model for a normal or binary missing covariate X is a linear or logistic regression on the Nelson-Aalen estimate $\hat{H}_0(t_i)$, the event indicator D and the remaining covariates Z. The imputation model of a normal variable X with missing data will be as follows,

$$X|T, D, Z \approx \mathcal{N}(\gamma_0 + \gamma_1 D + \gamma_2 H_0(T) + \gamma_3 Z, \sigma^2) \tag{3.10}$$

for a binary variable X with missing data, we get the imputation model as,

$$\operatorname{logit} p(X = 1|T, D, Z) \approx \gamma_0 + \gamma_1 D + \gamma_2 H_0(T) + \gamma_3 Z \tag{3.11}$$

Chapter 4

Bias Estimation

The data sets after dealing with the missing data are fitted to an analysis model. After that, the bias of the coefficient estimates with the true parameter is then calculated which is done separately for both MI and CCA, solely for comparison purposes, using the methods given below:

4.1 Mean absolute error (MAE)

Definition 4.1.1. *Mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon, defined by*

$$MAE = \frac{\sum_{i=1}^{n} |\beta_{true} - \beta_{estimate}|}{n} \tag{4.1}$$

Expressed in words, the MAE is the average over the verification sample of the absolute values of differences between the actual and estimated values. The MAE is a linear score which means that all the individual differences are weighted equally in the average

4.2 Root mean square error (RMSE)

Definition 4.2.1. *RMSE is a quadratic scoring rule which also measures the mean magnitude of the error in the obtained parameter estimate, i.e. it's the square root of the average of squared differences between estimated and actual parameter value.*

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\beta_{true} - \beta_{estimate})^2}{n}} \tag{4.2}$$

It describes the degree of coincidence among true and estimated values. Since squaring of the errors comes before averaging, the RMSE gives a relatively high weightage to large errors. This implies that the RMSE is most useful when large errors are particularly undesirable.

4.3 Comparison with the ratio of closeness with the true value

This method simply counts the number of times the estimate obtained from the data set handled by MI method outperforms the estimates obtained from the CCA method. It is defined as the average number of times when the estimate based on the MI method is closer to the true parameter than the estimate based on the CCA method,

$$\begin{aligned} d &= \frac{1}{K}\sum_{k=1}^{K} 1\{|\hat{\beta}_k^{\mathrm{MI}} - \beta| < |\hat{\beta}_k^{\mathrm{CCA}} - \beta|\} \\ r &= d/(1-d) \end{aligned} \tag{4.3}$$

where, the K denotes the total number of simulated data sets, β indicates the true value of the estimate when the data is complete without any missing values. $\hat{\beta}^{\mathrm{MI}}$ denotes the estimates obtained by handling the missing data using MI technique and the $\hat{\beta}^{\mathrm{CCA}}$ denotes the estimate obtained from CCA.

Chapter 5

Simulation Study

In this section, we aim to verify whether the MI method is the appropriate one over CCA when the amount of missing data is very high. We also want to see how efficient the MI method for a large population sample with a high amount of missing data . Here, the accuracy of the estimates is quantified using the underlying bias between the coefficient estimates with the true parameter. The accuracies obtained from both the methods are in turn used for carrying out the comparison. A quick look at the necessary variables that are required to be included in the imputation model for the better result are also discussed. This simulation study has been carried out in two parts, the first one which has a logistic regression analysis model with missing values in the covariates and the second one which has a survival model that has missing values in its covariates.

5.1 Study Design

In the construction of the covariates, we've considered the simplest case where there are two normally distributed variables without any missing values in it and only one incomplete normally distributed variable. In further studies, we proceeded by adding missing variables one at a time until a total of three missing covariates were present at the same time, of which two of them were normally distributed and one with categorical values. It is important to note that the assessment of the performance of CCA and MI method for handling the missing data is done by evaluating the accuracy of the estimated coefficients obtained using RMSE,

MAE and by comparing the ratio mentioned in section 4.3 from the previous chapter. Also, these two methods are examined with respect to the variation in the relative size of missing data present in the sample. By comparing the results obtained from the two methods, we will be able to come to the conclusion whether the MI method is the better choice for handling the missing data as proposed in most of the recent research papers.

Covariates

The detailed description of the simulation settings and variable construction are as follows:

- A data set of size 10^5 is generated with covariates x_1, x_2, x_3, x_4 and x_5. The variables x_1, x_2, x_3 and x_4 are generated as continuous and the final variable x_5 as a categorical variable with 4 categories.
- Among the variables, x_1 and x_2 are constructed from a random normal distribution independent of all the remaining variables and the construction is such that, $x_1 \sim \mathcal{N}(2,\, 1^2)$, and $x_2 \sim \mathcal{N}(-2,\, 1^2)$.
- From the third variable onwards, missing values are introduced. Hence, in order to incorporate the correlation to facilitate the MI, it is constructed as $x_3 \sim \mathcal{N}(0.5\ x_1 + 0.5\ x_2,\, 2^2)$
- The variable x_4 is generated based on the previous three variables, the complete variables x_1, x_2 and the partially observed variable x_3, in order to incorporate the correlation for imputation and the construction goes like, $x_4 \sim \mathcal{N}(0.33\ x_1 + 0.33\ x_2 + 0.33\ x_3,\, 2^2)$.
- The final categorical covariate x_5 with four categories is generated by conditioning the probability of each category with respect to the four covariates defined previously,

$$\begin{aligned} lp_1 &= (-2) + 0.5x_1 + 0.2x_2 + 0.1x_3 + 0.2x_4 \\ lp_2 &= (-2) + 0.5x_1 + 0.2x_2 + 0.1x_3 + 0.2x_4 \\ lp_3 &= 0.5x_1 + 0.2x_2 + 0.1x_3 + 0.2x_4 \\ lp_4 &= 0 \end{aligned} \tag{5.1}$$

Covariates	mean/number of observations for each category	standard deviation/ percentage of observation for each category
x_1	2	1
x_2	-2	1
x_3	$0.5x_1 + 0.5x_2$	2
x_4	$0.3x_1 + 0.3x_2 + 0.3x_3$	2
x_5	$x_5^{(1)} \approx 33000$ $x_5^{(2)} \approx 7000$ $x_5^{(3)} \approx 53000$ $x_5^{(4)} \approx 7000$	$x_5^{(1)} \approx 33\%$ $x_5^{(2)} \approx 7\%$ $x_5^{(3)} \approx 53\%$ $x_5^{(4)} \approx 7\%$

Table 5.1: Descriptive statistics of the simulate covariates

$$P_1 = \exp(lp_1)/\exp(\sum_{i=1}^{4} lp_i) \qquad P_3 = \exp(lp_3)/\exp(\sum_{i=1}^{4} lp_i) \tag{5.2}$$
$$P_2 = \exp(lp_2)/\exp(\sum_{i=1}^{4} lp_i) \qquad P_4 = 1/\exp(\sum_{i=1}^{4} lp_i)$$

The probability of the response variable is constructed in such a way that there are two less frequent categories (categories 2 and 4) while the other two are more frequent (1 and 3).

As mentioned previously, the missing values are placed in the variables x_3, x_4 and x_5 with the percentage of missingness as 99%, 95%, 90% and 80% of the total data size. The rest of the variables x_1, x_2 and the response are generated without any missing values in it. Similarly, about 240 data sets are simulated which is then subjected to both CCA and MI method.

5.1.1 Outcome

Binary Outcome

A binary outcome y, without any missing values, is generated for the first set of simulation studies. The construction is done as follows:

$$lp = -1 + 0.05\ x_1 + 0.2\ x_2 + 0.1\ x_3 + 0.02x_4 + \log(1.5)(x_5 = 1) + \log(5)(x_5 = 2) + \log(2)(x_5 = 3) \quad (5.3)$$

$$y_i = \begin{cases} 0 & \text{if } u_i \geq \texttt{plogis}(lp_i) \\ 1 & \text{if } u_i < \texttt{plogis}(lp_i) \end{cases} \quad (5.4)$$

for some random number $u_i \in \mathcal{U}(0, 1)$ where $i = 1, ..., 10^5$

Survival outcome / Time to event outcome

For the final simulation, we've used a TTE outcome and the time scale used here is age. Here, our objective was to replicate the real data situation. For this purpose, we've used the mortality statistics, provided by Statistics Finland [6], to determine shape and scale parameters of a TTE response which in turn follows a Weibull distribution. The construction of the whole set up is carried out as follows:

Here we are considering the subjects with delayed entries, i.e. the subjects are not observed from time(here our time variable is age) 0 but only from a later entry time, x_t, that is, the subject is only observed conditionally on having survived until x_t Let the delayed time denoted by $x_t \sim \mathcal{U}(0, 50)$ is chosen as the starting age when the individual or subject enters the cohort study. The failure time of each of the individual is defined as

$$l_p = 0.05x_1 + 0.2x_2 + 0.1x_3 + 0.02x_4 + \log 5(x_5 = 2) + \log 2(x_5 = 3) + \log 1.5(x_5 = 4) \quad (5.5)$$

$$T = (a/b)(t/b)^{a-1} exp(l_p) \quad (5.6)$$

where a is the shape parameter, b being the scale parameter and $l_p = \sum \beta_i x_i$, denotes the linear progression of the covariates.
The individual is censored after an age of 100. So, the observed response variable "T_i" and

the event indicator δ_i for an i^{th} individual or observation :

$$T_i = min(t_i, 100) \tag{5.7}$$

$$\delta_i = \begin{cases} 0 & \text{if the response was censored } (t_i > 100) \\ 1 & \text{if the event was observed } (t_i \leq 100). \end{cases} \tag{5.8}$$

5.2 Result of the simulation study

5.2.1 Missing covariates with a binary outcome

Results of the simulation setup with the binary outcome are presented in Table 5.2, Figure 5.2 and Figure 5.1. This section explains the results only from the simulation study, which contains all the partially observed covariates. The results are arranged in terms of the percentage of missing data present in the 240 simulated data sets.

The Table 5.2 denotes the ratio of the number of times MI methods outperformed the CCA method. From the values, we can see that most of its entries are greater than or close to 1, which implies that the MI technique works very well for a model with a binary outcome and missing data in the covariates. However, for the second category in the categorical variable x_5, MI doesn't perform well as, compared to CCA.

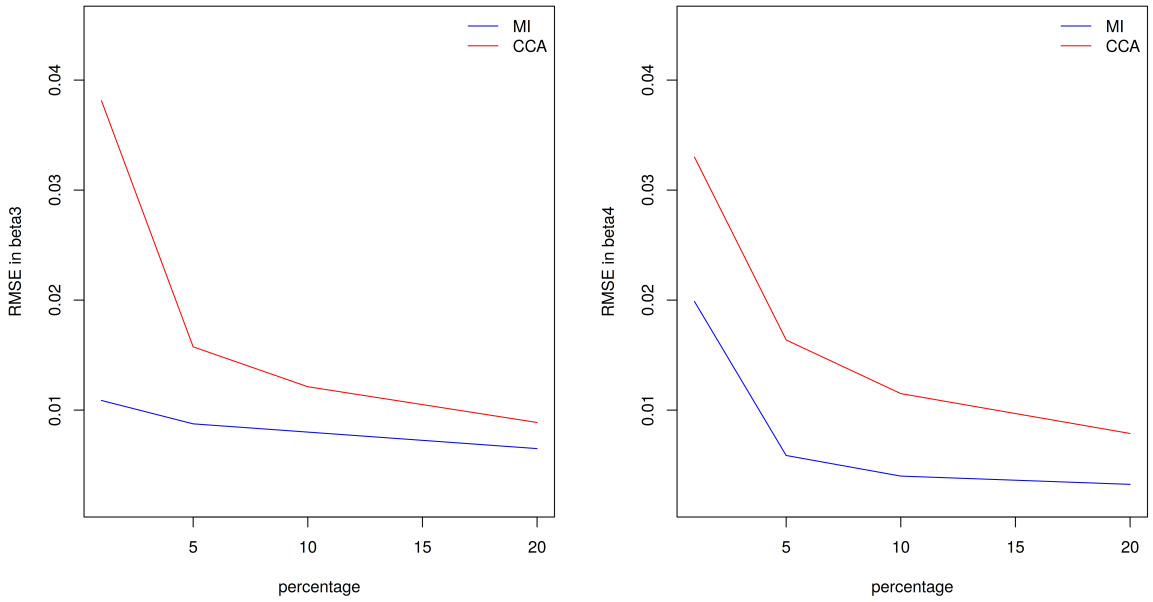
Table 5.2: The table depicts the ratio comparison of the accuracy of estimates obtained by the imputation methods with that of CCA, the ratio (r) given in the equation (4.3).

	1 percent	5 percent	10 percent	20 percent
x3	0.388	0.367	0.312	0.217
x4	0.446	0.438	0.383	0.342
x5_class2	0.021	0	0	0
x5_class3	0.338	0.3	0.275	0.271
x5_class4	0.479	0.5	0.458	0.342

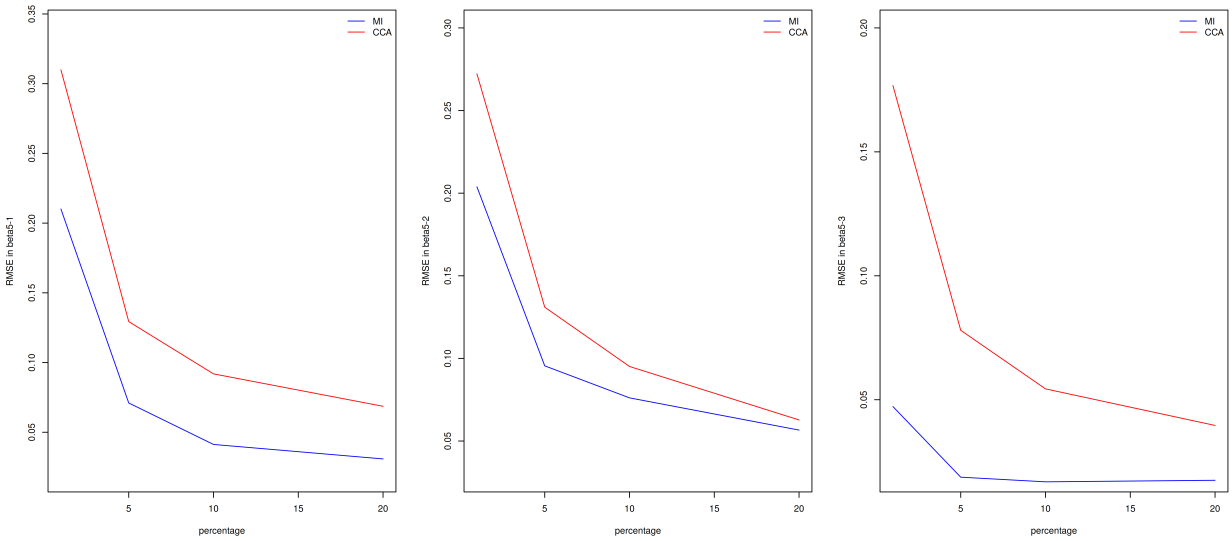
This is probably in view of the fact that this category doesn't provide sufficient details for its imputation by the MI method, due to its low frequency in x_5 as depicted in the Table 5.1.

This table doesn't mean that one method, say MI, is far better than the other, since it only counts the number of times the parameter estimate obtained by MI is closer, than CCA method, to the actual value and not how much accurate the estimate is. This enhances the importance of the use of other evaluation techniques, such as RMSE and MAE.

MAE and RMSE are plotted against the percentage of observed values or the non missing values. The plot of RMSE (Figure 5.1) and MAE (Figure 5.2) shows almost same trend and in all the plots, it is evident that MI performs better than CCA for a high percentage of missing values. Even in the less frequent category, category 2 of the variable x_5, MI performs better than CCA when the non missing part is only 1% and 5% (or when 99% and 95% of the total size is missing) of the total size and then coincide in the later part of the plots.

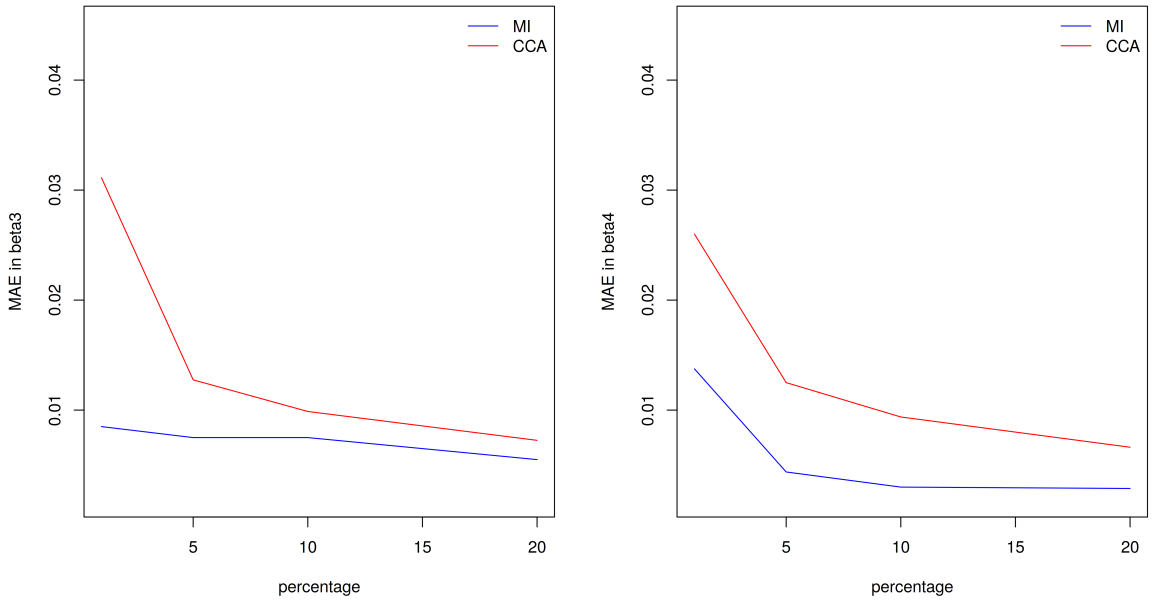


(a) RMSE in coefficient estimate of x_3 is plotted in the left panel and x_4 is in the right panel of the figure.

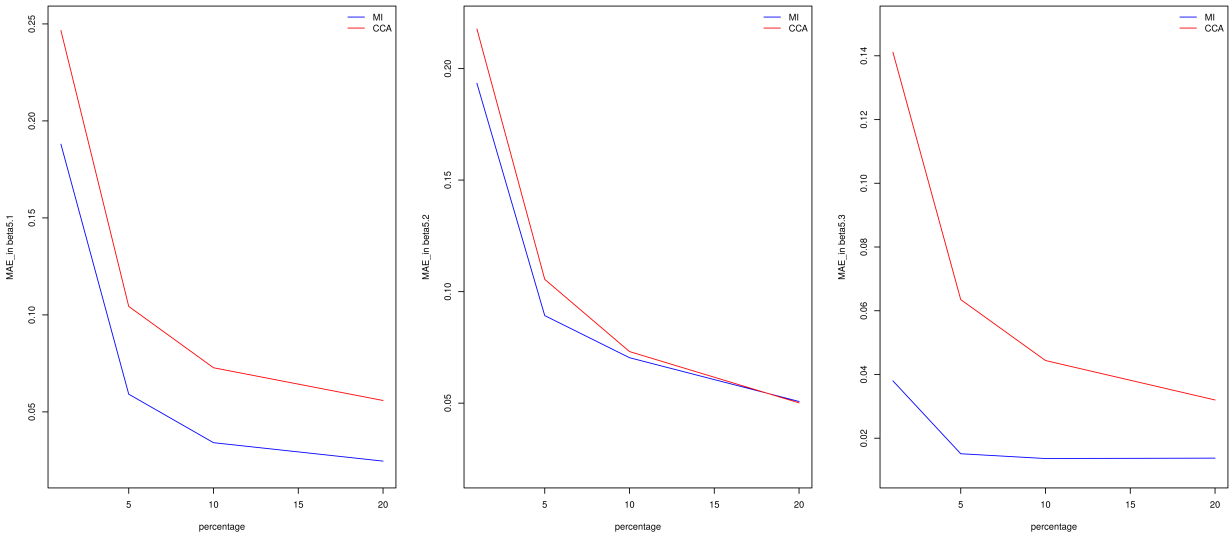


(b) RMSE in the coefficient estimate of categories of x_5 is plotted in this figure. The leftmost plot represent the 1st category, the centre being the second and the rightmost the 3rd category of the variable x_5.

Figure 5.1: Root mean square error in the case of a binary outcome. The blue curve indicates the RMSE of the estimate obtained by MI method while the red curve indicates the same obtained from CCA.



(a) MAE in coefficient estimate of x_3 is plotted in the left panel and x_4 is in the right panel of the figure.



(b) MAE in the coefficient estimate of categories of x_5 is plotted in this figure. The leftmost plot represent the 1st category, the centre being the second and the rightmost the 3rd category of the variable x_5.

Figure 5.2: Mean absolute error in case of a binary outcome. The blue curve indicates the MAE of the estimate obtained by MI method while the red curve indicates the same obtained from CCA.

5.2.2 Missing covariates in a Survival outcome data

The simulation study with the survival outcome model having missing covariates addresses our empirical study in the next section. As in the previous simulation study, here also the results are arranged in decreasing amount of missing values. The results are described in the Table 5.3, Figure 5.3 and in Figure 5.4. Here, we tried different ways of imputation such as, using logistic regression model as the imputation model for the categorical variables and linear regression model for continuous variables as in White and Royston(2009). And using classification and regression trees (CART) method for the imputation of both the categorical and continuous variables.

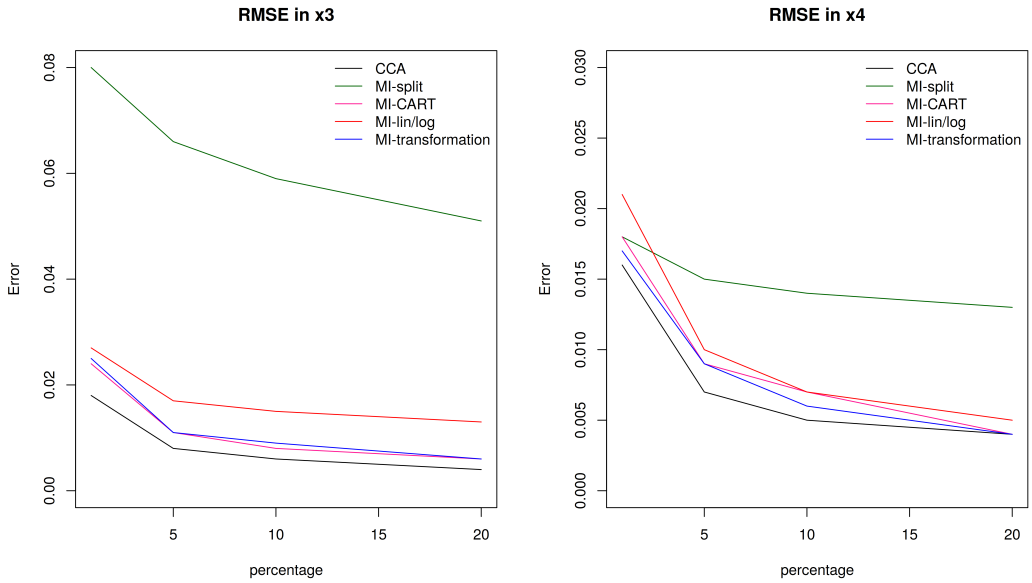
Table 5.3: The table depicts the ratio comparison of the accuracy of estimates obtained by the imputation methods with that of CCA, the ratio (r) given in the equation (4.3)

	x3	x4	x5 Cat2	x5 Cat3	x5 Cat4
Imputation using split					
10	0.009	0.591	0	0	0.274
20	0	0.123	0	0	0.145
30	0	0.022	0	0	0.102
40	0	0.009	0	0	0.03
Imputation with CART model					
10	0.539	0.927	0.113	0.943	0.911
20	0.463	0.539	0.087	0.852	0.756
30	0.463	0.612	0.058	0.896	0.705
40	0.386	0.634	0.026	0.669	0.549
Imputation with GLM model					
10	0.519	0.681	0.601	1.174	1.097
20	0.145	0.436	0.463	0.756	0.881
30	0.134	0.559	0.339	0.657	0.927
40	0.082	0.58	0.191	0.529	0.866
Imputation using T and $log(T)$					
10	0.612	0.896	0.118	1.043	0.959
20	0.463	0.646	0.068	0.911	0.646
30	0.463	0.623	0.063	0.866	0.669
40	0.317	0.669	0.026	0.837	0.570

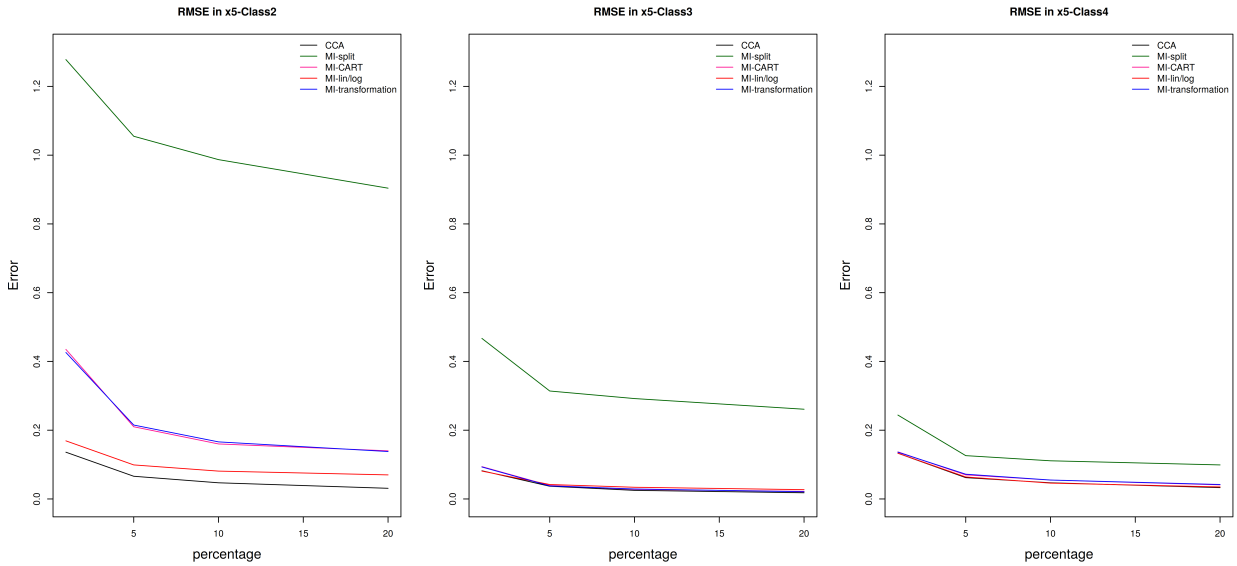
As mentioned earlier, it is a common practice to use the event indicator D, observed event or censoring time T and $log(T)$ in the imputation model, which is also assessed in this study. Along with all three setting, we tried to experiment another way of defining the imputation model by including D and a time group variable T_m which is obtained by splitting the survival time into time group element.

From the Table 5.3 it can be seen that most of the values are less than 1 which indicates that almost all the times the coefficient estimated from the CCA is closer to the actual estimate value than all the other imputation. However, inference can be drawn only after quantifying the closeness using other methods like MAE and RMSE

The Figure 5.3 represents the RMSE and the Figure 5.4 represents the MAE values of the simulated data sets. According to the plots of MAE and RMSE it is clear that the imputation model which includes the time group as covariate is the least preferred one. While the rest of the three MI method and CCA method performs almost equally well for the coefficient estimate.

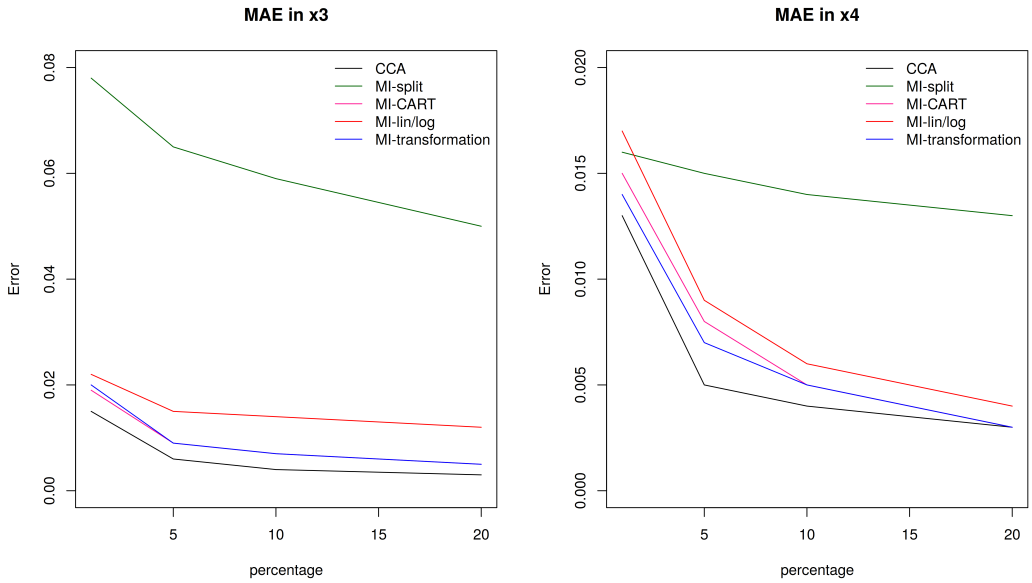


(a) RMSE in coefficient estimate of x_3 is plotted in the left panel and x_4 is in the right panel of the figure. The RMSE values are evaluated for different imputation model for the comparison purpose

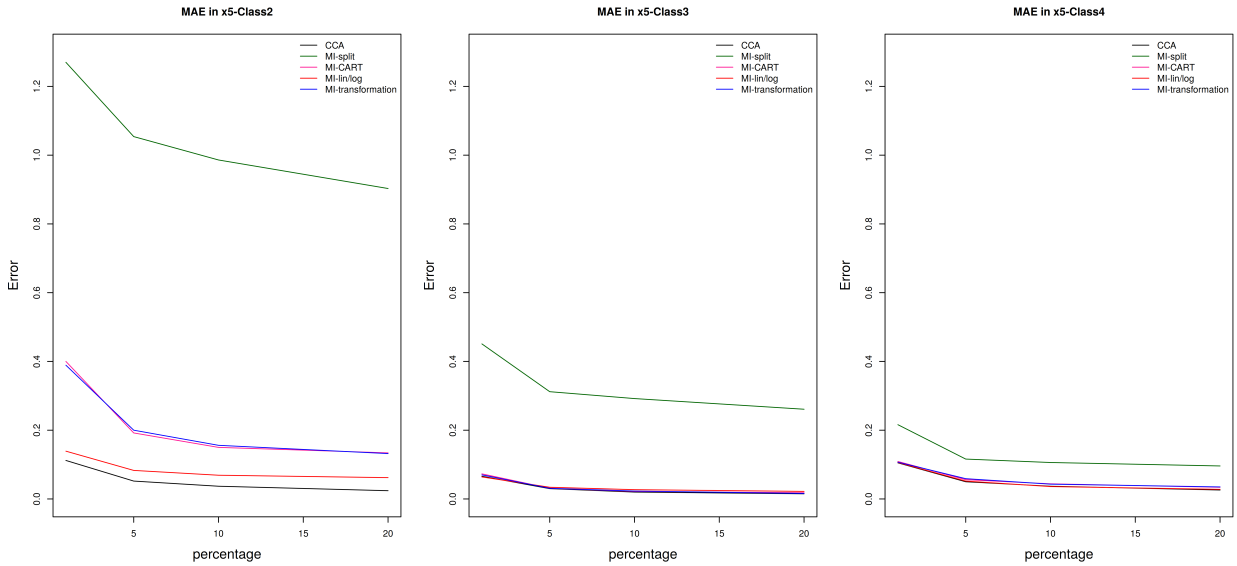


(b) RMSE in the coefficient estimate of categories of x_5 is plotted in this figure. The leftmost plot represent the 1st category, the centre being the second and the rightmost the 3rd category of the variable x_5. The RMSE values are evaluated for different imputation model for the comparison purpose

Figure 5.3: Root mean square error: The black line denotes the CCA, and the rest of the lines indicate MI method applied in various manner. The dark green line indicate the imputation model with splitted survival time, the pink line indicates the imputation method which use logistic regression model for missing categorical covariates and linear model for the continuous covariate. The blue line indicates the common practice which use T, and $log(T)$ in the imputation model



(a) MAE in coefficient estimate of x_3 is plotted in the left panel and x_4 is in the right panel of the figure. The MAE values are evaluated for different imputation model for the comparison purpose



(b) MAE in the coefficient estimate of categories of x_5 is plotted in this figure. The leftmost plot represent the 1st category, the centre being the second and the rightmost the 3rd category of the variable x_5. The MAE values are evaluated for different imputation model for the comparison purpose

Figure 5.4: Mean absolute error: The black line denotes the CCA, and the rest of the lines indicate MI method applied in various manner. The dark green line indicate the imputation model with splitted survival time, the pink line indicates the imputation method which use logistic regression model for missing categorical covariates and linear model for the continuous covariate. The blue line indicates the common practice which use T, and $log(T)$ in the imputation model

Chapter 6

Empirical study on Ischemic Heart Disease follow up data

Individual-level records from health and social services are routinely being generated, collected and maintained centrally in nation-wide registers in Finland by the Finnish Institute for Health and Welfare (THL). In this section, data from population register and Care Register for Health Care (Hilmo) are used to carry out an empirical study of handling the missing data problem in the ischemic heart disease (IHD) data. The Finnish population and health care registers are presumed to have a well covered and recorded health details of the population, and the register-based analyses can provide very accurate estimates of population health. However, the registers lack essential information on risk factors such as smoking status and alcohol consumption, which are considered as relevant risk factors for IHD. Hence, we are combining the Finnish population register data with the survey data that was collected in 2000 health surveys, with a follow-up data until 2018, which comprised of these relevant risk factors and an enormous amount of missing data. After handling the missing data with the proposed methods, the population data which contain both the register data and the survey data is then modelled by a multi-state prognostic illness death model. The multi-state illness death models allow subjects to move among a finite number of states, as in the Figure 3.1, during a follow-up period.

The data set is restricted to those people who are in the age of 30 years or older and the ones who are in the healthy phase at the beginning of the study, which is at the 1st July 2000.

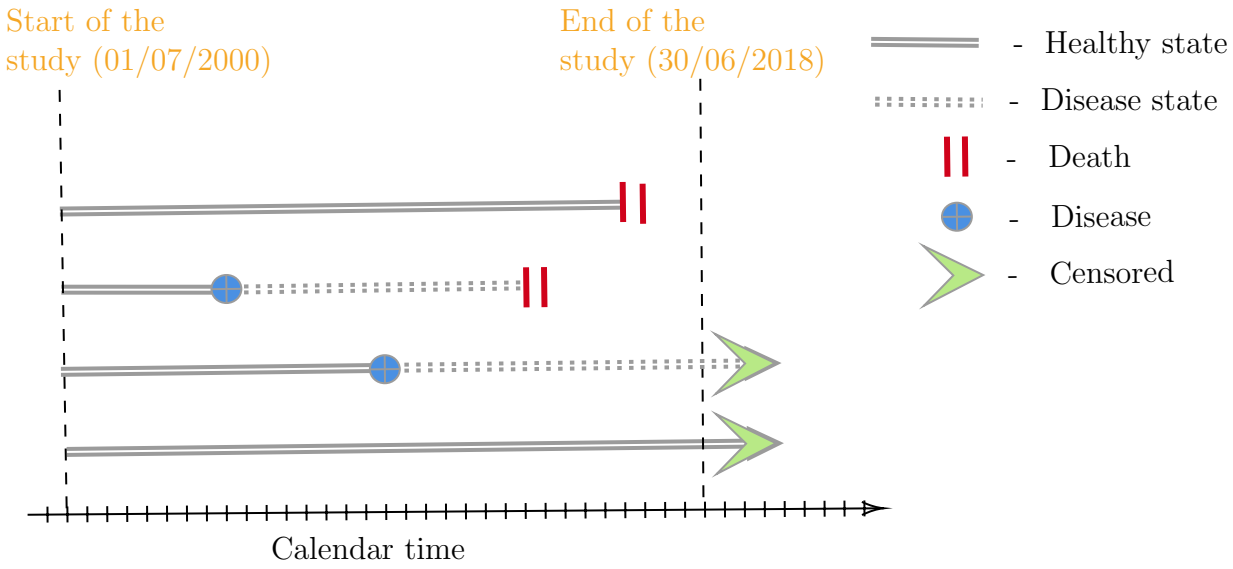


Figure 6.1: A pictorial timeline depicting the possible pathways of the study cohort

The constrains made according to our educated guess, one of which is that the individuals below 30 years are less likely to be diagnosed with IHD in this 18 years period. The other constrain, which only allows those people who are not yet diagnosed with the IHD, is due to the fact that those people who had already diagnosed with IHD will be probably in control of their risk factor values with the help of medication or care. This will affect the imputation model by adding noise to those risk factors. This is the reason why, in this study, we are only including the individuals starting in the same state, namely the healthy. Moreover, the diseased individuals have a generally higher risk to die because of the disease, hence incorporating the diseased individuals will require a different imputation model. A schematic representation of this setting is provided in the Figure 6.1.

The population data set consists of 496,709 individuals that includes 12,229 observations from the survey data. The associated risk factors, which numbers up to 16 in total, are composed of 6 categorical variable and 10 continuous variable as described in the Table 6.2 . From the Table 6.2 we can also see that there are about 97% of missing values present in all the risk factors.

Prior to applying MI to deal with the missing data, CCA is done. The complete cases which do not have any missing values are only 6987 observations which are slightly more than 50% which is then analysed by a proportional hazard Weibull regression model. In order to give an elaborate comparison between CCA and the full population data, the number of transitions

Covariates	Missing value number (%)	mean (SD)	min value	1st quartile	median	3rd quartile	max value
X_3	486328 (97.91)	6.0 (1.11)	1.9	5.2	5.9	6.6	11.7
X_4	486328 (97.91)	1.3 (0.376)	0.2	1	1.2	1.5	3.4
X_5	[486287 (97.90)	139.6 (22.41)	68	124	138	154	245
X_6	486287 (97.90)	81.8 (12.212)	0	74	82	90	134
X_7	486316 (97.91)	137.6 (22.081)	66	122	136	150	236
X_8	486320 (97.91)	80.8 (11.532)	0	74	80	88	131
X_9	485525 (97.75)	77.0 (15.865)	29.1	65.9	75.9	86.8	169.2
X_{10}	485541 (97.75)	167.9 (9.986)	135.5	160	168	175	198
X_{13}	489432 (98.53)	7.5 (5.995)	1	2	7	12	50
X_{16}	486451 (97.93)	3833.8 (8939.833)	0	0	633.5	3708.4	229394.1

(a) The descriptive statistics for the continuous covariates

Covariates	Number of missing values(%)	number in each categories	% in each category
X_1	0	1 - 238559	1 - 48.03
		2 - 258150	2 - 51.97
X_2	485378 (97.72)	0 - 9064	0 - 80
		1 - 2267	1- 20.01
X_{11}	486479 (97.94)	1 - 3317	1 - 32.42,
		2 - 5396	2 - 52.75,
		3 - 1412	3 - 13.8,
		4 - 105	4 - 1.03
X_{12}	486719 (97.99)	1 - 2140	1 - 21.42
		2 - 684	2 - 6.85
		3 - 7166	3 - 71.73
X_{14}	486255 (97.90)	0 - 7642	0 - 73.1
		1 - 2812	1 - 26.9
X_{15}	485790 (97.80)	0 - 2579	0 - 23.62
		1 - 3714	1 - 34.01
		2 - 4626	2 - 42.37

(b) The descriptive statistics for the categorical covariates

Table 6.1: Descripive Statistics

	Healthy	Disease	Death
Healthy	5349	5574	1306
Disease	0	2653	2921
Death	0	0	4227

(a) Number of transition in the population data

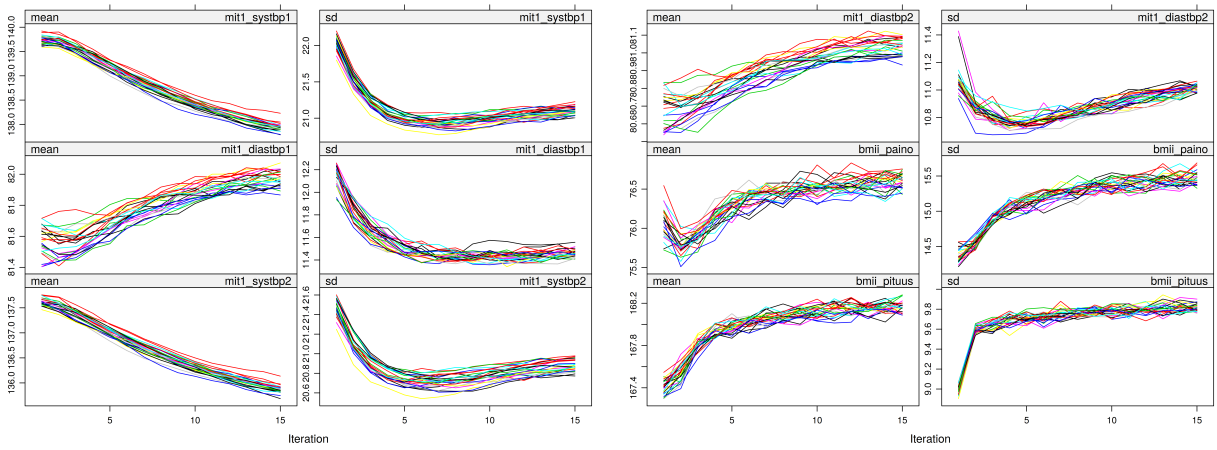
	Healthy	Disease	Death
Healthy	3761	2770	456
Disease	0	1660	1110
Death	0	0	1566

(b) Number of transition for the complete cases setting

Table 6.2: Transition numbers

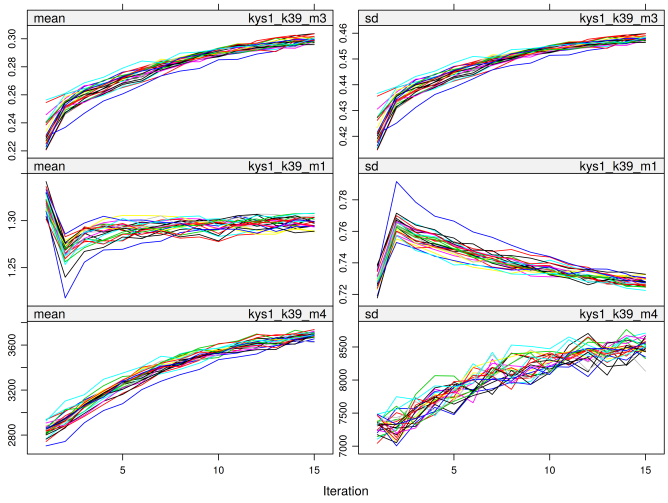
are calculated and described in the Table 6.2b. After analysing with the CCA method, the data is subjected to MI method. The imputation method we've used here is CART, and the covariates selected in each of the imputation model is based on it's Kendall rank correlation value with the response. Along with the covariates in the Table 6.2 the Nelson Aalen estimate of cumulative hazard and event indicators for the death and the disease are also used in the imputation models. After the imputation process, the convergence of each of the imputed covariates are assessed. The Figure 6.2 represents trace plot of the imputed covariates with slow on no convergence while the rest of the covariates are had their streams well intermingled and are free from any kind of trends which in turn indicates the convergence.
The imputed data sets are then analysed using the PH Weibull regression model. The estimates obtained from both the methods are provided in the Table 6.3 .

From the MAE and RMSE values obtained in the simulation study, it is evident that both the MI and CCA works almost equally efficient in handling the missing values present in the covariates of a survival model. As a result, we expected the value of the estimates obtained from both the methods to be close to each other. The empirical study using the IHD data set clearly demonstrated the same. Hence with some improvement in the MI method such as modifying the number of iteration and number of imputation, it will provide to a better result than the CCA. The estimates obtained from the CCA and MI is demonstrated in the Table 6.3.



(a)

(b)



(c)

Figure 6.2: Trace plots of the covariates that shows slow or no convergence

Covariates	Transition from state 1 to 2				Transition from state 2 to 3				Transition from state 1 to 3			
	Estimates		Std.error		Estimates		Std.error		Estimates		Std.error	
	MI	CCA	MI	CCA	MI	CCA	MI	CCA	MI	CCA	MI	CCA
$X_1^{(2)}$	-0.069 (-0.132)	-0.739	0.01	0.065	-0.188 (-0.231)	-0.385	0.014	0.147	-0.032	-0.496	0.012	0.104
$X_2^{(2)}$	0.211	0.182	0.008	0.051	0.032	0.682	0.012	0.113	0.249 (0.291)	0.916	0.010	0.087
X_3	-0.064 (0.035)	-0.025	0.003	0.019	-0.030	-0.072	0.004	0.046	-0.086	-0.284	0.004	0.036
X_4	-0.505 (-0.242)	-0.771	0.01	0.068	0.178 (0.081)	0.12	0.014	0.146	-0.326 (-0.084)	0.183	0.011	0.108
X_5	0.001	-0.008	0.001	0.004	-0.002	-0.016	0.001	0.009	0.002	0.014	0.001	0.006
X_6	0.002	-0.002	0.001	0.004	-0.002	0.025	0.001	0.011	-0.001	0.013	0.001	0.005
X_7	-0.004	0.011	0.001	0.004	0.002	0.017	0.001	0.009	-0.005	-0.013	0.001	0.006
X_8	0.009	-0.007	0.001	0.004	-0.006	-0.019	0.001	0.012	0.004	-0.018	0.001	0.007
X_9	-0.005	0.009	0	0.002	0.000	-0.009	0.000	0.004	-0.006	-0.005	0.000	0.003
X_{10}	0.013	-0.019	0.001	0.004	0.006	-0.011	0.001	0.008	0.017	-0.007	0.001	0.006
$X_{11}^{(2)}$	-0.019	-0.178	0.007	0.043	-0.135 (-0.186)	-0.325	0.010	0.107	-0.413 (-0.25)	-0.16	0.008	0.078
$X_{11}^{(3)}$	-0.068 (-0.044)	-0.357	0.011	0.067	-0.017 (-0.195)	-0.257	0.015	0.155	-0.304 (-0.343)	-0.223	0.012	0.113
$X_{11}^{(4)}$	-0.093 (0.007)	-0.38	0.034	0.193	0.046 (-0.466)	-0.893	0.051	0.585	-0.440 (0.414)	0.229	0.043	0.331
$X_{12}^{(2)}$	0.028	-0.13	0.016	0.178	0.092	0.423	0.021	0.318	0.075 (0.158)	1.543	0.019	0.258
$X_{12}^{(3)}$	0.002	-0.039	0.014	0.102	0.022	-0.608	0.022	0.205	0.032	0.626	0.016	0.168
X_{13}	0.031	-0.017	0.001	0.005	0.000	0.035	0.001	0.011	0.041	0.006	0.001	0.009
$X_{14}^{(2)}$	0.039	0	0.009	0.059	-0.033	0.082	0.013	0.134	-0.022 (0.055)	-0.05	0.011	0.109
$X_{15}^{(2)}$	-0.168 (-0.092)		0.016		0.110		0.024		0.077 (0.103)		0.018	
$X_{15}^{(3)}$	-0.009 (0.123)	0.247	0.014	0.052	0.050 (0.001)	-0.057	0.022	0.124	0.099 (0.126)	-0.3	0.016	0.083
X_{16}	0	0.006	0	0.001	0.000	0.002	0.000	0.001	0.000	-0.005	0.000	0.001

Table 6.3: Ischemic heart disease data: results for the parameter estimate in the analysis model, where the imputation model is chosen based on the Kendall rank correlation method(the Pearson correlation[1]).

[1]The values in brackets are the coefficient estimates of the analysis model when variables selection in the imputation model is based on the Pearson correlation that has high deviation from the ones obtained using Kendall rank correlation

Chapter 7

Conclusion

In this thesis, we have discussed the two of the widely used methods for handling missing data, namely MI and CCA. For MI itself, we've discussed several ways to model the imputation model in order to achieve the best results. Generally, for the majority of regression models, CCA gives unbiased and accurate results when the missingness is assumed as MCAR and sometime even for MNAR. Accordingly, there are situations in which CCA analyses are more efficient than MI analyses, under the MAR assumption (Little, 1992 [17]). Nevertheless, unlike CCA, MI is valid for all MAR cases and can use information included in the incomplete cases and auxiliary variables to improve the accuracy of the estimates. Our aim in this study was to assess the MI method when the involved missing data amount was very high, about 80 – 99% of the total size of the sample.

From the simulation studies, we observed that both the MI and CCA based estimates work almost equally well in the case of survival outcome. At the same time, MI estimate is more accurate than that obtained from CCA in case of a binary outcome. However, this is not always the case. In order to have a fair comparison between the MI and CCA, we should use the number of imputation to be approximately equal to the percentage of missing value as demonstrated by White et al. (2011) [14]. With the current computation power, carrying out the imputation with $m \approx 99$ will be difficult. Furthermore, an increase in the number of iteration until attaining the convergence for all the variable or at least to the variables with a high fraction of missing values would also improve the estimation of our statistical quantity of interest, at the cost of more computational resources.

From the empirical study, we can see a major drawback in the CCA method, which is, the 1^{st} category in the variable "X_{15}" gets eliminated in the complete case analysis. This happens because the rows corresponding to the 1^{st} category of the variable "X_{15}" always had at least one missing value in some other covariates in the same row. This can always happen to the categorical covariates with a rare category in a data having an enormous amount of missing values when subjected to CCA analysis.

In conclusion, we still propose the MI method as a good method for handling the data sets with an enormous amount of missing data after incorporating the improvements in the number of imputation, number of iteration and the selection of covariates both in the imputation and analysis model.

Chapter 8

Appendix

Appendix 1: In order to incoporate the delayed entry situation we've modified the R function, `nelsonaalen`, in the `mice` package which only allows the entry at time $t_0 = 0$, for calculating the Nelson Aalen estimate for cumulative hazrad. The R code for the modified function is given as,

```
##############################################################################
                            Nelson-Aalen estimate
##############################################################################
nelson_aalen <- function(data, timevar, statusvar, starttime) {
  mice:::install.on.demand("survival")
  if (!is.data.frame(data))
    stop("Data must be a data frame")
  timevar <- as.character(substitute(timevar))
  statusvar <- as.character(substitute(statusvar))
  time <- data[, timevar]
  status <- data[, statusvar]
  hazard1 <- survival::
    basehaz(survival::coxph(survival::Surv(time, status) ~ 1))
  idx1 <- match(time, hazard1[, "time"])
  haz1 <- hazard1[idx1, "hazard"]
  if(missing(starttime)){
```

```
    return(haz1) }

  else {
    starttime <- as.character(substitute(starttime))
    ini_time <- data[, starttime]
    X <- rep(1, nrow(data))
    cox_out <- survival::coxph(survival::Surv(ini_time, time, status) ~ X)
    cox_out$coefficients["X"] <- 0
    haz <- predict(cox_out, newdata = data.frame(ini_time, time,  0, 1),
                   type = "expected")
    return(haz)
  }
}
```

Appendix 2: The correlation matrix measured using the Kendall's Tau method for the IHD data variables along with the disease indicator(stat1), death indicator(stat2) and the Nelson Aalen estimate for healthy to disease(na12) and healthy to death(na13) transitions, is given in the following Table 8.1

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	na12	na13	stat1	stat2
X_1	1	-0.14	0.00	0.26	-0.05	-0.13	-0.05	-0.14	-0.39	-0.59	-0.04	-0.23	-0.35	-0.23	-0.04	-0.30	0.14	0.17	-0.16	0.061
X_2	-0.14	1	0.03	-0.06	-0.1	0.01	-0.10	0.02	0.03	0.14	-0.054	0.16	0.24	0.19	0.08	0.18	-0.16	-0.17	-0.03	0.02
X_3	0.00	0.03	1	0.06	0.12	0.15	0.13	0.15	0.03	-0.03	-0.02	-0.02	0.02	0.06	-0.03	0.02	0.13	0.10	0.09	0.02
X_4	0.26	-0.06	0.06	1	-0.06	-0.06	-0.05	-0.05	-0.28	-0.14	0.09	0.05	-0.07	0.07	0.01	0.04	0.06	0.07	-0.13	0.01
X_5	-0.05	-0.1	0.12	-0.06	1	0.38	0.86	0.36	0.09	-0.09	-0.08	-0.14	-0.08	-0.05	-0.07	-0.08	0.20	0.19	0.10	0.08
X_6	-0.13	0.01	0.15	-0.06	0.38	1	0.38	0.83	0.22	0.10	0.01	0.09	0.08	0.09	0.06	0.10	0.05	0.01	0.09	0.01
X_7	-0.05	-0.10	0.13	-0.05	0.86	0.38	1	0.37	0.09	-0.08	-0.08	-0.13	-0.07	-0.05	-0.07	-0.06	0.20	0.19	0.12	0.08
X_8	-0.14	0.02	0.15	-0.05	0.36	0.83	0.37	1	0.23	0.12	0.02	0.1	0.09	0.09	0.07	0.11	0.04	0.00	0.09	0.01
X_9	-0.38	0.03	0.03	-0.28	0.09	0.22	0.09	0.23	1	0.41	-0.02	0.15	0.21	0.18	0.03	0.20	-0.05	-0.08	0.15	-0.07
X_{10}	-0.59	0.14	-0.03	-0.14	-0.09	0.10	-0.08	0.12	0.41	1	0.08	0.30	0.31	0.24	0.08	0.33	-0.19	-0.21	0.1	-0.09
X_{11}	-0.04	-0.05	-0.02	0.09	-0.08	0.008	-0.085	0.02	-0.02	0.08	1	0.12	0.02	0.02	0.11	0.07	0.00	-0.03	-0.03	-0.08
X_{12}	-0.23	0.16	-0.02	0.05	-0.14	0.09	-0.13	0.11	0.15	0.30	0.12	1	0.19	0.37	0.56	0.6	-0.16	-0.19	0.02	-0.13
X_{13}	-0.35	0.24	0.02	-0.07	-0.08	0.08	-0.07	0.09	0.21	0.31	0.02	0.19	1	0.45	-0.36	0.52	-0.24	-0.24	0.02	-0.02
X_{14}	-0.23	0.19	0.06	0.07	-0.05	0.09	-0.05	0.09	0.18	0.24	0.02	0.37	0.45	1	-0.03	0.58	-0.13	-0.14	0.08	-0.04
X_{15}	-0.04	0.08	-0.03	0.01	-0.07	0.06	-0.07	0.07	0.03	0.08	0.11	0.56	-0.36	-0.03	1	0.19	-0.08	-0.10	0.00	-0.1
X_{16}	-0.30	0.18	0.02	0.04	-0.08	0.10	-0.06	0.11	0.20	0.33	0.07	0.6	0.52	0.58	0.19	1	-0.15	-0.18	0.08	-0.09
na12	0.14	-0.16	0.13	0.06	0.20	0.05	0.20	0.04	-0.05	-0.19	0.00	-0.16	-0.24	-0.13	-0.08	-0.15	1	0.87	-0.1	0.03
na13	0.17	-0.17	0.10	0.07	0.19	0.01	0.19	0.00	-0.08	-0.21	-0.03	-0.19	-0.24	-0.14	-0.10	-0.18	0.87	1	-0.15	0.05
stat1	-0.16	-0.03	0.09	-0.13	0.10	0.09	0.12	0.09	0.15	0.1	-0.03	0.02	0.02	0.08	0.002	0.08	-0.1	-0.15	1	-0.18
stat2	0.061	0.02	0.02	0.01	0.08	0.01	0.08	0.01	-0.07	-0.09	-0.08	-0.13	-0.02	-0.04	-0.	-0.09	0.03	0.05	-0.18	1

Table 8.1: Pair wise correlation matrix between the covariates associated to IHD data, using Kendall's Tau method

Bibliography

[1] van Buuren, S . Flexible imputation of missing data, Boca Raton, FL: Chapman & Hall/CRC, 2012.

[2] Rubin, Donald B. "Inference and Missing Data." Biometrika, vol. 63, no. 3, 1976, pp. 581–592. JSTOR, www.jstor.org/stable/2335739. Accessed 19 Oct. 2020.

[3] Rubin, Donald B. Bayesian Inference for Causal Effects: The Role of Randomization. Ann. Statist. 6 (1978), no. 1, 34–58. doi:10.1214/aos/1176344064. https://projecteuclid.org/euclid.aos/1176344064.

[4] Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons Inc., New York. http://dx.doi.org/10.1002/9780470316696

[5] Schafer JL. Multiple imputation: a primer. Stat Methods Med Res. 1999 Mar;8(1):3-15. doi: 10.1177/096228029900800102. PMID: 10347857.

[6] Reference data.

http://pxnet2.stat.fi/PXWeb/pxweb/en/StatFin/StatFin_vrm_kuol/statfin_kuol_pxt_12ap.px

[7] Andersen PK, Keiding N. Multi-state models for event history analysis. Stat Methods Med Res. 2002 Apr;11(2):91-115. doi: 10.1191/0962280202SM276ra. PMID: 12040698.

[8] Von Hippel, P.T. (2007), Regression with missing YS: An improved strategy for analysing multiply imputed data. Sociological Methodology, 37: 83-117. doi:10.1111/j.1467-9531.2007.00180.x

[9] van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999 Mar 30;18(6):681-94. doi: 10.1002/(sici)1097-0258(19990330)18:6¡681::aid-sim71¿3.0.co;2-r. PMID: 10204197.

[10] Huo, Z. (2015). A Comparsion of Multiple Imputation Methods for Missing Covariate Values in Recurrent Event Data (Dissertation). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-256602

[11] Kruijk, Monique van der. Multiple imputation with chained equations and survival outcomes A simulation study. (2015).

[12] Rubin, Donald B. "Multiple Imputation After 18 Years." Journal of the American Statistical Association, vol. 91, no. 434, 1996, pp. 473–489. JSTOR, www.jstor.org/stable/2291635.

[13] ¨ Royston P, Carlin JB, White IR. Multiple Imputation of Missing Values: New Features for Mim. The Stata Journal. 2009;9(2):252-264. doi:10.1177/1536867X0900900205

[14] White, I. R., Royston, P., Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. Statistics in medicine, 30(4), 377–399. https://doi.org/10.1002/sim.4067

[15] S. Van Buuren, J. P.L. Brand, C. G.M. Groothuis-Oudshoorn D. B. Rubin (2006) Fully conditional specification in multivariate imputation, Journal of Statistical Computation and Simulation, 76:12, 1049-1064, DOI: 10.1080/10629360600810434

[16] van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research. 2007;16(3):219-242. doi:10.1177/0962280206074463

[17] Roderick J. A. Little (1992) Regression with Missing X's: A Review, Journal of the American Statistical Association, 87:420, 1227-1237, DOI: 10.1080/01621459.1992.10476282

[18] Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009 Jun 29;338:b2393. doi: 10.1136/bmj.b2393. PMID: 19564179; PMCID: PMC2714692.

[19] Ofer Harel, Emily M Mitchell, Neil J Perkins, Stephen R Cole, Eric J Tchetgen Tchetgen, BaoLuo Sun, Enrique F Schisterman, Multiple Imputation for Incomplete Data in Epidemiologic Studies, American Journal of Epidemiology, Volume 187, Issue 3, March 2018, Pages 576–584, https://doi.org/10.1093/aje/kwx349