EVALUATE COGNITIVE IMPAIRMENT ON TIME TO DEMENTIA WITH HRS DATA THROUGH MULTIPLE IMPUTATION AND VARIABLE SELECTION

by

YILU JIN

(Under the Direction of Xiao Song)

ABSTRACT

The Health and Retirement Study (HRS) is designed to have a thorough investigation of retirement policy, health and well-being of the US elderly. The main focus of this thesis is to predict the time to dementia by demographic characteristics and self-respond questions measured in HRS. The Cox proportional hazard (PH) model was adopted in the analysis. However, there is a certain percentage of missingness in the HRS data. In order to ensure the accuracy of the study results, multiple imputation was used to deal with the missing data. Since not all variables are related to the response, we could reduce the number of covariates contained in the final model to improve the precision of model prediction. Stepwise and LASSO selection methods were then conducted on multiply-imputed datasets to select significant variables related to dementia. We evaluate the performance of the methods on selecting important variables by simulation studies.

INDEX WORDS: Cox PH model, Multiple imputation, Variable selection, Data simulation, Cognitive impairment, Dementia

EVALUATE COGNITIVE IMPAIRMENT ON TIME TO DEMENTIA WITH HRS DATA THROUGH MULTIPLE IMPUTATION AND VARIABLE SELECTION

by

Yilu Jin

B.E., Nanjing Agricultural University, 2017

A Thesis Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Master of Science

Athens, Georgia

2019

EVALUATE COGNITIVE IMPAIRMENT ON TIME TO DEMENTIA WITH HRS DATA THROUGH MULTIPLE IMPUTATION AND VARIABLE SELECTION

by

Yilu Jin

Approved:

Major Professor: Xiao Song

Committee: Ming Zhang
Hanwen Huang

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2019

Acknowledgments

I would like to express my deep gratitude to Dr.Song, my research advisor, for her patient guidance, passionate encouragement and useful critiques of this research work. Your expertise was priceless in the devising of the research topic and methodology. Your advice and assistance have kept my progress on schedule. I would also like to thank Dr.Zhang and Dr.Huang, my committee members, for generously offering their time and their assistance with resolving my questions as well as the collection of my data. Thank you very much for your helpful comments and suggestions. My grateful thanks are also extended to you all for your instruction on my two-years study. Finally, I wish to thank my parents for supporting and encouraging my pursuit throughout my study and career.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Cognitive impairment (CI) is a major health problem among the elderly, which has received increasing attention due to its importance to human health. CI can lead to a decline in the quality of life, reduce life expectancy[1], and has a high mortality rate among the elderly[2]. Dementia, a terminal disease of CI, is defined as a decline in memory and cognitive functions that causes a loss of independent ability. This is a common phenomenon that has a great influence on individuals, families and government programs[3].

In the past 25 years, great progress has been made on assessing the effects of demographic, medical and lifestyle factors on dementia[4, 5]. As is known to all, CI and dementia disease are not evenly distributed in the population since they are highly correlated with age. Moreover, education level is another factor which has a connection with CI. Those with higher education level are less likely to develop cognitive impairment. On the other hand, low levels of education are thought to be linked to cognitive impairment because people with low levels of education have less cognitive reserves, poorer physical health in old age, and a greater risk of cognitive decline. In this study, we have evaluated more factors related to CI as a function of time to dementia.

The HRS was launched in 1992 by The University of Michigan's Institute, primarily sponsored by the National Institute on Aging (NIA), with additional funding from the US Social Security Administration (SSA)[6]. Its main objective is to meet the national demand for reliable data that enable research and analysis to support policies on retirement, health, and well-being of people as they age. The survey gathered information on population, income,

assets, health, cognition, family structure and relationships, health care utilization and costs, housing, work status and history, expectations and insurance. The HRS is administered by the University of Michigan Institute for Social Research (ISR). It consists of six cohorts: the original HRS, first interviewed in 1992; the AHEAD cohort, first interviewed in 1993; depressed, first interviewed in 1998; war babies (WB) cohort, first interviewed in 1998; the early baby boomers, first interviewed in 2004; and the mid-baby boomers born, first interviewed in 2010[7]. In this study, we adopted all of these six cohorts. However, since the measurements of memory-related or dementia disease started from 1998, only the measurement-results at 1998, which is my baseline, and measurements from 1998 forward were counted in my study. For cohort 5 and 6, the baseline are the measurement-results at 2004 and 2010, respectively.

1.2 Data Overview

The HRS is a large, nationally representative biennial interview longitudinal survey of people over the age of 50 and their spouses. This survey is carried out both in person and on the telephone. Core interview is conducted once every two years and each interview last for about two hours, including face-to-face interview at participant's home and follow-up interview at home or on phone. The study now has 12 waves from 1992 to 2014, each with approximately 18-23,000 participants. In order to make data more accessible to researchers, the RAND center for the Study of Aging created the RAND HRS data products. The RAND HRS Longitudinal File is a user-friendly version of a subset of the HRS[7].

The variables in the HRS cover a very large range of measures (demographics, the health, health insurance, income, Social Security, wealth, family structure, retirement etc.), including the extensive measures that can be used to judge the decline of cognition and the starting point of cognitive impairment. In the HRS, the cognitive function can be evaluated by word-recall and mental status measurements which include: the self-reported memory; the immediate and delay word recall; backwards counting; date naming; president/vice-president

naming etc. The higher the score, the lower the likelihood of cognitive impairment. The variables of word-recall, mental status tests and other factors related to health at baseline were used as covariates to predict the future risks of suffering from memory-related disease in elderly. Other than self-respondent cognitive variables, we also used demographic variables like age, gender, and education years etc. All of these covariates were taken from the RAND HRS longitudinal dataset (RAND).

Nearly half of the variables used in my study are binary, and remaining half are either continuous or multinomial covariates. Ten out of thirty-seven variables are continuous, i.e., education years, age, BMI, immediate word recall (IMRC) and delay word recall (DLRC) etc. The rest of the variables are either nomial or ordinal variables, such as, cohort indicator (COHORT), race, religion, and backwards counting test (BWC20) etc are nominal variables; self-report of the health (SHLT), backwards counting test (BWC20), and self-reported memory (SLFMEM) are ordinal variables.

Since the primary focus of this study is the self-respondent's cognitive measurements, a subset including 44 variables and 22,197 observations was extracted from the RAND data. Among the 44 variables, 9 variables (R4MEMRY∼R9MEMRY, R10DEMEN∼R12DEMEN) indicating the status of doctor diagnosed memory-related disease were used to construct the survival time, censored time and event status for survival data. The event here means the observation was diagnosed with a memory-related or dementia disease[1]. The time was measured in years-scale with 1998 as the starting point. The variables related to cognition in section A and B in the RAND file were included in this study, except the variables that has collinearity property with other variables, such as total word recall summary score. Most variables contain missing values except gender and cohort indicator. The missing percentages of most variables are around 11% to 18%.

[1]From 2010 forward, the outcome of doctor diagnosed health problem changed from memory-related disease to dementia. Since the consistence between these two outcomes in results, dementia was used to represent these two for convenience.

1.3 Purpose of Study

The motivation for this study is to analyze the cognition-related measurements in the RAND data. Subsequently our main interest became whether the self-report cognitive evaluations at the baseline can be used to effectively predict the time when a subject will be diagnosed as dementia by a doctor. As the RAND contains the biennial investigation of the dementia status of the elderly people, survival analysis[8] is a feasible way to achieve this purpose. Moreover, this data meets the two requirements of survival analysis: 1) subjects are usually followed for a certain period; 2) the time of the occurrence of interested-event can be observed. We adopt the proportional hazards model[9] to evaluate the effect of cognition-related measurements on time to dementia. This could be used in the future to predict the future risk of cognitive development. The objective of this thesis is to present thorough evaluation of relevant factors at baseline and identify important covariates that affect the dementia status of the observations. However, there is a large amount of missing values in the dataset. Variable selection is generally carried out on the complete dataset, otherwise, significant missing will cause unreliable analysis results. Thus, it is necessary to handle the missing values with common multiple imputation methods.

1.4 Literature Review

1.4.1 Cognitive Impairment and Dementia

Cognition impairment is a common disease among elderly Americans, and dementia is the terminal illness. The definition of severe cognition impairment is that subject cannot successfully answer any cognitive screen questions in the survey interview. Senile dementia includes different causes and clinical manifestations: Alzheimer's, non-Alzheimer's, stroke-related, vascular etc[4]. Nowadays, a lot of studies on the effects of demographic and lifestyle factors on CI and dementia diseases have been presented in literatures.

Kenneth et al.[10] have studied the influence of medical, demographic, and social trends on the cognition health in old adults. The comparison of prevalence of CI and two-year mortality rates between the 1993-1995 and 2002-2004 was measured. The CI was evaluated by 35-point cognition scale for self-respondents aged 70 or older, and the evaluation of memory and judgment by proxy respondent were also considered in cognitive assessment. The results showed that there was a higher rate of CI among people aged 70 or older in 1993 than in 2002. Moreover, those with moderate or severe CI have higher mortality in both years. Although education can prevent the development of CI, higher education was accompanied by a higher 2-year mortality among those who have CI.

Eileen et al.[11] also conducted a research on assessment of cognition using HRS data by performing a new measurement using subsample. A detailed neuropsychiatric assessment (Aging, Demographics, and Memory Study [ADAMS]) was adopted in the subsample to verify the effectiveness of HRS survey in predicting cognition. ADAMS has identified three degrees of cognition outcomes: Demented, cognitively impaired without dementia (CIND) and normal cognitive function. The author applied the multinomial logistic regression model to predict the diagnosis using the measurements related to cognition or noncognitive in HRS and ADAMS. However, this approach does not taken account of the censoring observations in the HRS data.

1.4.2 Survival Analysis and Cox PH Model

Survival analysis[8] is a group of methods used to analyze data where the outcome variable is survival time that subject to censoring. Dependent variables in survival analysis include time to event and event status. Time to event is the time until the occurrence of an event of interest, and event status indicates if the event happens or not. There are three types of methods in survival analysis that can be used to analyze the relationship between predictor variables and survival time: non-parametric, semi-parametric, and parametric methods. Kaplan Meier[12] is a widely used non-parametric function for estimating

and graphing survival probability as a function of time. Parametric approach assumes that the distribution of survival follows a specific probability distribution, such as, exponential, Weibull, or lognormal distribution. The most popular and applicable semi-parametric model is the Cox proportional hazard model[13] which does not require the specification of baseline hazard function. The Cox regression model is a simple and convenient analysis method, which can describe the relationship between hazard function and predictors.

Censoring is a common problem that needs to be taken into account in survival analysis, which refers to a subject did not undergo the event of interest during follow-up period. The most common situation is right censoring, for example, a person was not observed to experience the event before the end of the study or he dropped out in the middle of the study. Cox PH model was adopted in my thesis to deal with the right-censored time to dementia in HRS data.

1.4.3 Techniques For Missing Data

The common solutions on missing data are listwise deletion and pairwise deletion provided with the default statistical package. Unless data is missing completely at random (MCAR), which means missingness is not dependent on the data at all, these two methods would induce bias and reduce the power of analysis. Another way of handling missing data is single imputation, which means imputing each missing value under the specified model. This imputation method can provide a complete data to facilitate subsequent statistical analysis that requires complete data. However, there are some problems with this method. Since it fits in the missing value as if they were known, limitation of missing-value variability would occur in statistical analysis. Other ad-hoc imputation methods, such as mean imputation, would cause bias and inference errors.

Multiple Imputation (MI), proposed by Rubin[14], is a widely used method to solve problems of missing data, which will reduce bias compared with methods above. This method repeatedly fills in the missing cells in incomplete data and generates corresponding two or

more complete data sets. MI performs well with data under an ignorable missing mechanism, including missing completely at random (MCAR) and missing at random (MAR)[15] while it may produce biased results under a non-ignorable missing mechanism-missing not at random (MNAR). MAR occurs when the pattern of missingness only depends on the observed data, not the unobserved data. MNAR occurs when the probability of missingness on a variable is related to the value of that variable itself or other unobserved variables given the observed variables.

1.4.4 Variable Selection

Variable selection is a critical step in model building. Several variable selection methods for Cox model were mentioned in literatures. Purposeful selection[16], proposed by Hosmer and Lemeshow, can avoid overfitting and too strict in initial screen. Starting with fitting individual univariate Cox models, non-significant variables were deleted at $20\% \sim 25\%$ level. Then, multivariate Cox model would be fitted with all the remaining covariates.

Forward selection starts with a model without covariate, then adds one covariate at a time until no more can be added to the model given a specific criterion. Backward selection is an inverse process opposite to forward selection. It starts from the model containing all the covariate, then deletes a covariate at one time until no more can be removed based on a specific threshold. Stepwise selection, first proposed by Efroymson[17], is a combination of forward selection and backward selection. Best subset selection examines all subsets of variables and select the best model among them. However, one common drawback of these approaches is that the variations induced in the variable selection process is not accounted in inference based on the final selected model. When the number of observations is less than the candidate covariates, namely high-dimensional data, penalized regression is a better way to achieve variable selection.

The least absolute shrinkage and selection operator (LASSO), introduced by Tibshirani[18], is a general variable selection method which maximizes a penalized version of the likeli-

hood. For Cox model, the likelihood is replaced with partial likelihood[19]. In some cases, however, it is desirable to select predictors in groups. Yuan and Lin[20] proposed the group LASSO method, which penalizes the grouped coefficients in a similar way to traditional lasso. However, Fan and Li[21] have shown that the traditional lasso estimator may not be completely effective and its selection result may be inconsistent. The major reason is that the traditional lasso applies same tuning parameter on each regression coefficient. In order to solve this problem, Zou[22] proposed adaptive LASSO which generally adds adaptive weights used for penalizing different coefficients in the L_1 penalty. Similar method has been developed for Cox proportional hazard model[23].

After multiple imputation, if the variable selection methods are applied to each multiply-imputed dataset separately, the selection results of each dataset may be inconsistent, making it difficult to obtain the final selected model. Therefore, many methods have been mentioned in the literatures for variable selection with missing data. Wood et al.[24] provided a method using backward variable selection on a stacked multiply-imputed dataset. They also proposed a stepwise selection approach for multiply-imputed data via repeated application of Rubin's rule. Chen and Wang[25] used group LASSO to select the grouped coefficients of each variable in multiple imputed datasets. Schomaker et al.[26] proposed using the stepwise and LASSO selection methods on each imputed dataset separately.

Chapter 2

Methodology

2.1 Cox Proportional Hazard Model

The Cox proportional hazard regression model was first introduced by Cox[13] in a seminal paper (1972). Let T denote the survival time, $X_1, ..., X_k$ be the covariates. The proportional hazards model assumes that the hazard of the failure

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{\Pr\left[(t \leq T < t + \Delta t) \mid T \geq t\right]}{\Delta t} \\ &= h_0(t) \exp\left(\beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}\right). \end{aligned}$$

Here $h_0(t)$ is a unspecified baseline function. It is assumed that the hazard ratio does not vary over time. The `survival` package in R can be used for fitting the Cox model.

The standard inference for the Cox regression approach is based on the partial likelihood. Let X_i and β denote as $X_i = (x_{i1}, ..., x_{ik})$ and $\beta = (\beta_1, \beta_2, ..., \beta_k)^T$ respectively. When there is no tied event times, the partial likelihood is given by

$$L(\beta) = \prod\nolimits_{r \in E} \frac{\exp\left(\beta^T X_{k_r}\right)}{\sum_{k \in R_r} \exp\left(\beta^T X_k\right)}, \tag{2.1}$$

where E is a set of indicators of the failure times, R_r is the set of indicators of the individuals at risk at time T_r. Given $\mathcal{L}(\beta) = \log L(\beta)$, the parameter β is estimated by maximizing the log partial likelihood

$$\hat{\beta} = \arg\max \mathcal{L}(\beta).$$

When there exists tied events, the permutations in the calculation of true partial likelihood function can make the calculation very time-consuming. In this case, Breslow[27], Efron[28],

and Kalbfleisch and Prentice et al.[29] have proposed several approximated partial likelihood functions. Hertz-picciotto and Rockhill[30] showed that, in most cases, Efron approximation is the best method. Consequently, the default method of tie handling is Efron in `survival` package.

2.2 Multiple Imputation

2.2.1 Software Packages for Multiple Imputation

For this study, we used two packages: Amelia II[31] and MICE[32] for multiple imputation (MI). Amelia II imputes the missing data based on multivariate normal assumption of the variables. However, real data often has complex data structures and is usually composed of many variables, which are usually subject to different distributions, such as Poisson data. It is difficult to assign the joint distribution of all variables. Thus, under certain circumstances, it is not reasonable to assume missing data is multivariate normal (MVN). MICE (multivariate imputation by chained equations), which is based on "fully conditional specifications" (FCS). Starting with selecting a variable with the lowest missing rate, this variable will be imputed through an appropriate regression model given other variables and observed values. The next variable will be imputed in the same way with the imputed variable. Each different model is assigned to different variables until all the variables are imputed. The whole procedure will be iterated until convergence. MICE can deal with different types of variable: numeric, binary, multinomial, ordinary counts and mixed variables.

Both packages assume that missing data is MAR. The assumption of MAR will be more reliable if more variables from the dataset are included in MI process, not only those that would show up in the final analysis model. In general, Amelia II will have a better performance in imputing multivariate normal data, while MICE does a better job in imputing missing data with lots of different types of variables.

2.2.2 Rubin's Rule

Multiple imputation (MI) from either method creates M imputed datasets, and then regression analysis is conducted on each imputed dataset, which will result in M estimates of the parameters and their covariance estimates. Then we combine these results according to the following rules. The estimate of parameters after MI can be obtained by

$$\hat{\beta}_{MI} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}^{(m)}, \tag{2.2}$$

where $\hat{\beta}^{(m)}$ is the estimate of β in the m^{th} imputed data set $d^{(m)}$. The covariance of β after MI is

$$\widehat{Cov}\left(\hat{\beta}_{MI}\right) = \widehat{W} + \frac{M+1}{M}\hat{B}, \tag{2.3}$$

with the average within imputation covariance

$$\widehat{W} = \frac{1}{M}\sum_{m=1}^{M}\widehat{Cov}\left(\hat{\beta}^{(m)}\right),$$

and the between imputation covariance

$$\hat{B} = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}^{(m)} - \hat{\beta}_{MI}\right)\left(\hat{\beta}^{(m)} - \hat{\beta}_{MI}\right)'.$$

The confidence interval then can be calculated by a t_v distribution

$$\hat{\beta}_{MI} \pm t_v\sqrt{\widehat{Cov}\left(\hat{\beta}_{MI}\right)}, \tag{2.4}$$

where the degrees of freedom v is given by $v = \left[1 + \left(\frac{M}{M+1}\right)\frac{\widehat{W}}{\hat{B}}\right]^2 (M-1)$. More details can be found in Rubin[14].

2.3 Variable Selection

We consider two approaches for variable selection. One is the approach proposed in the R package MAMI[33], the other is the group LASSO approach[20, 34].

2.3.1 Variable Selection in MAMI

For each imputed data set, the variables are selected using a variable selection approach, such as AIC and LASSO. If a variable is selected in at least one imputed data set, it will be officially selected, but its frequency of being selected determines its overall impact. After multiple imputation, the estimates of β can be obtained by

$$\hat{\beta}_{MI} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}^{(m)},$$

where $\hat{\beta}^{(m)}$ is the estimates of the parameters for the m^{th} imputed data set, with the estimates being zero for unselected variables. According to (2.3), the overall variance of the estimator after multiple imputation is:

$$\widehat{Var}\left(\hat{\beta}_{j,MI}\right) = \frac{1}{M}\sum_{m=1}^{M}\widehat{Var}\left(\hat{\beta}_{j}^{(m)}\right) + \frac{M+1}{M(M-1)}\sum_{m=1}^{M}\left(\hat{\beta}_{j}^{(m)} - \hat{\beta}_{j,MI}\right)^{2}.$$

Therefore, the confidence interval can also be calculated using (2.4). However, this method still may not solve the problem that the parameters in regression models may still be biased[26].

For Cox model, two variable selection approaches are implemented in MAMI, one is based on AIC (mAIC) and the other is based on LASSO (mLASSO). The AIC for the Cox model is

$$AIC = -2\mathcal{L}\left(\hat{\beta}\right) + 2p,$$

where p is the dimension of parameters β. Tibshirani[19] applied his own LASSO method to Cox model. the LASSO parameter estimate of β is

$$\hat{\beta}_{LE}(\lambda) = \arg\min_{\beta}\left\{ -\mathcal{L}(\beta) + \lambda\sum_{j=1}^{p}|\beta_j|\right\},$$

where λ is a set of candidate tuning parameters $\lambda = \{\lambda_1, ..., \lambda_L\}$. The tuning parameter λ is selected by cross-validation (CV), and the purpose is to minimize the error of the difference between training- and test- likelihood (the partial likelihood deviance). In `MAMI`, for the m^{th} imputed dataset, the variance of the LASSO estimator is estimated by bootstrapping[33].

2.3.2 Group LASSO

Yuan and Lin[20] proposed a group LASSO (gLASSO) approach for variable selection in linear regression model after multiple imputation. Here we extend this approach to the Cox model. The coefficients of the same variable in the M imputed dataset are constrained to a group. In regular LASSO, each coefficient has different constraints. However, the group LASSO removes a set of variables from the model by shrinking the parameter of this set to zero, and retains the important set of variables on which the hazard function depends.

The covariates are divided into K groups and $\beta_{(j)}$ is the regression coefficient of the j^{th} covariate group, where $j = 1, ..., K$. The group LASSO estimate is obtained by minimizing

$$Q_n(\beta, \lambda) = -\mathcal{L}_n(\beta) + \lambda \sum_{j=1}^{K} ||\beta_{(j)}||,$$

$$\hat{\beta}_n(\lambda) = \arg\min_\beta Q_n(\beta, \lambda),$$

where $\mathcal{L}_n(\beta)$ is the partial log likelihood for the Cox model. $||\cdot||$ represents the L_2 norm.

Applying group LASSO to the multiply-imputed datasets is implemented as follows. Treating the M imputed datasets as one big dataset and letting x_j^m be jth covarites in the m^{th} imputed dataset with coefficient β_j^m, define $\beta_j = (\beta_j^1, ..., \beta_j^m)$. If there is no significant correlation between X_j and the survival time, $\beta_j^1, ..., \beta_j^m$ are all zero. If X_j is significant, then none of $\beta_j^1, ..., \beta_j^m$ should be zero. The group LASSO method is implemented with the R package `SGL`[34] by setting alpha = 0. The tuning parameter is chosen by cross validation via the function `cvSGL`.

Chapter 3

Simulation Study

To investigate the performance of different variable selection methods, we use simulated data. In this way, we can control which variables are important to the response and know if the method has selected the right variables. All methods are evaluated based on 200 simulated data sets, each with 100 observations and 20 variables.

3.1 Generating Survival Data

In this simulation, there were n = 100 subjects and p =20 variables, which were generated from a multivariate normal distribution with mean $\mu = 0$ and the correlation structure of compound symmetric with the correlation coefficient $\rho = 0.3$ between any two variables. Among 20 variables, the variables $(X_1, X_5, X_{10}, X_{11}, X_{15}, X_{20})$ are important. The true Cox model is

$$h\left(t|\mathbf{x}\right) = 0.5\exp(0.5X_1 - 1.2X_5 + 1.8X_{10} + 0.5X_{11} - 1.2X_{15} + 1.8X_{20}).$$

The coefficients values from negative to positive and from small to large were taken into account in this study. Then we tested if the variable selection methods were successful in finding them.

The censoring time was generated from an exponential distribution with mean c and truncated at d, which corresponds to the maximum follow up time. By changing the parameter c and d we can vary the overall censoring rate. The censoring rate was controlled to be approximately 30%.

3.2 Generating Missing Data

The missing data was generated for the 20 candidate covariates. The first ten variables, $X_1, ..., X_{10}$, were completely observed, and the remaining ten variables, $X_{11}, ..., X_{20}$, all had missing values. The ignorable missing mechanism, MAR, was considered in this simulation. Two different nonresponse rates were also taken into account, resulting in 60% and 35% of complete cases, respectively.

Under MAR, the indicator R_{ij} is generated to specify whether X_{ij} is missing, $R_{ij} = 1$ if X_{ij} is missing and $R_{ij} = 0$ if X_{ij} is not missing. Suppose the missing-probability of variable X_{ij} depends on $X_{i(j-10)}$ such that

$$\text{logit}\{Pr(R_{ij} = 1|X_{i(j-10)})\} = \alpha_0 + 0.5X_{i(j-10)},$$

where $j = 11, \ldots, 20$, α_0 is controlled to get 60% and 35% complete cases, separately.

3.3 Measurements of Performance

To compare the performance of variable selection among the three methods, we used the following three criteria:

1. Sensitivity:
$$\text{Sen} = \frac{\text{\# of selected important variables}}{\text{\# of true important variables}};$$
2. Specificity:
$$\text{Spe} = \frac{\text{\# of unselected unimportant variables}}{\text{\# of true unimportant variables}};$$
3. Geometric mean of sensitivity and specificity:
$$\text{G} = \sqrt{\text{Sensitivity} \times \text{Specificity}}.$$

Sensitivity assess the proportion of correctly choosing important variables, while specificity measures the rate at which unimportant variables are correctly eliminated. In addition, the

geometric mean of sensitivity and specificity was used in this study to measure the degree of the correctness after the combination of sensitivity and specificity. These measures all range from 0 to 1, with the property that the larger the value, the better the selection performance.

3.4 Results

For each setup, 200 replicates of simulations were conducted. Although the more times of imputation the better the results, the processing time would be too long which is time-consuming. In order to obtain a balance, within each replicate, five imputed-datasets were generated in each incomplete data using R package Amelia II. We compared the variable selection, estimation and coverage probabilities of the four methods.

3.4.1 Variable Selection

In my missing data simulation, first half of variables $X_{10} \sim X_{20}$ are fully observed while missing values are only generated in second half of variables $X_{10} \sim X_{20}$. These two sets of covariates, complete set and incomplete set, are used to assess the effect of missing level. Moreover, two missing scenarios were designed to evaluate the variable selection methods: 60% CC under MAR; 35% CC under MAR. In order to intuitively illustrate the selection performance, Figure 3.1 is presented to show the selected frequency of all variables among different methods under the missing mechanism MAR. It represents comparison among four different methods: 1) the gLASSO on the multiply-imputed datasets; 2) the mLASSO on the multiply-imputed datasets; 3) the mAIC on the multiply-imputed datasets; 4) the LASSO on the complete-cases (CC). It is worth mentioning here, for CC, the LASSO is implemented in R using the function SGL by setting alpha = 1 and the tuning parameter is chosen by cvSGL.

The absolute value of coefficient has a considerable impact on selected percentage of important variables, i.e., X_{10} and X_{20} have higher percentages of selection than X_1 and X_{11}. Both mAIC and mLASSO tend to over-select more unimportant variables. Furthermore,

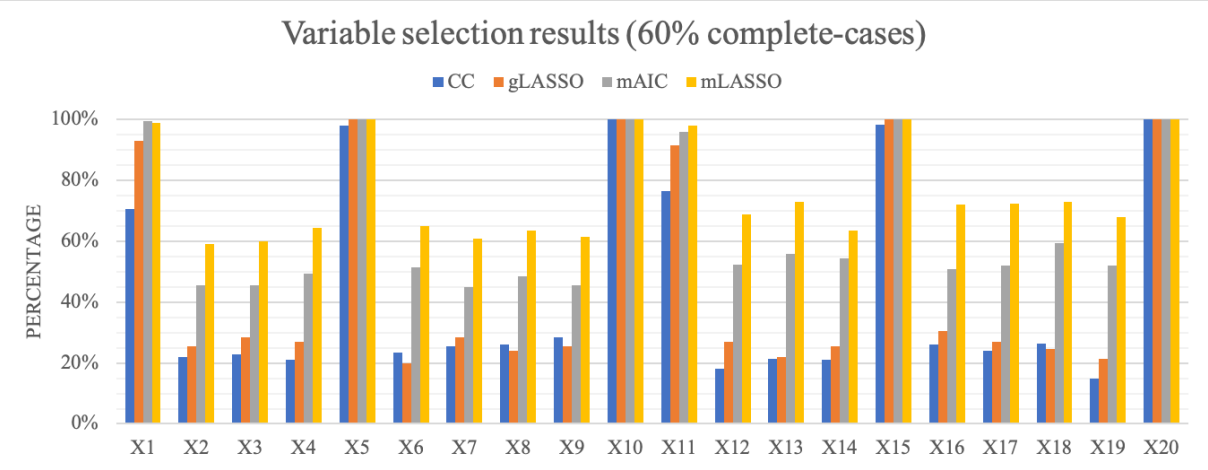

Figure 3.1: Percentage of each variable selected in the model using the four methods among 200 replicates with 60% complete cases. The important variables are $X_1, X_5, X_{10}, X_{11}, X_{15}, X_{20}$ with censoring rate 30%. Variables $X_1 \sim X_{10}$ are complete, while variables $X_{11} \sim X_{20}$ have missing values resulting in 60% complete cases. The missing mechanism is MAR. Each missing replicate is multiply imputed with Amelia II.

no significant difference is observed on the selection percentage between complete observed covariates and incomplete observed covariates. Specifically, for two important variables X_1 and X_{11} with the same coefficient values, we can see that X_1 has an approximate same selection percentage as X_{11}.

Comparing Figure 3.2 against Figure 3.1, the performance of CC on 35% complete-cases is worse than on 60% complete-cases. It is reasonable to conjecture that the higher the percentage of missing, the worse the performance of CC. For gLASSO, the performance is only slighly worsened when the complete cases decreased from 60% to 35%. However, mAIC and mLASSO tend to select more unimportant variables when the missing rate increases, with no obvious difference in selecting important variables.

When comparing gLASSO against CC, the selected percentages of the important variables by gLASSO are higher than by CC. In contrast, CC has comparatively lower probabilities of selecting unimportant variables than gLASSO. To some extent, CC has a better performance

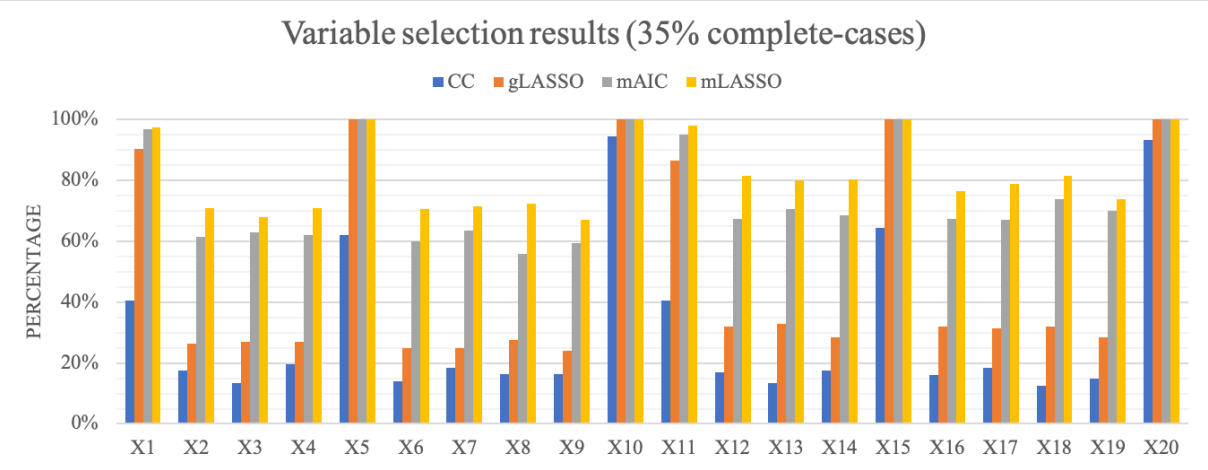

Figure 3.2: Percentage of each variable selected in the model using the four methods among 200 replicates with 35% complete cases. The important variables are $X_1, X_5, X_{10}, X_{11}, X_{15}, X_{20}$ with censoring rate 30%. Variables $X_1 \sim X_{10}$ are complete, while variables $X_{11} \sim X_{20}$ have missing values resulting in 35% complete cases. The missing mechanism is MAR. Each missing replicate is multiply imputed with Amelia II.

on removing unimportant variables. For mAIC and mLASSO, although their probabilities of selecting important variables are high, they both tend to over-select unimportant variables. The potential reason is that the variable selected in any imputed data set by mAIC and mLASSO will be formally presented in the final model.

The mean sensitivities, mean specificities and corresponding geometric means are reported in Table 3.1 for all the four methods. gLASSO, mAIC, and mLASSO have relatively higher sensitivities of selecting important variables than CC on both 60% and 35% complete-cases. CC and gLASSO have relatively higher specificity of removing unimportant variables than mAIC and mLASSO. After combing the performance of selecting important variables and eliminating unimportant variables, gLASSO outperforms other three methods since it has the highest geometric mean.

Table 3.1: Mean sensitivity (SEN), specificity (SPE) and their geometric mean (G) for the four methods among 200 replicates. Variables $X_1 \sim X_{10}$ are complete, while variables $X_{11} \sim X_{20}$ have missing values resulting in 60% or 35% complete cases. The missing mechanism is MAR. Each missing replicate is multiply imputed with Amelia II.

	60% complete cases		
Amelia II	**SEN**	**SPE**	**G**
CC	0.906	0.770	0.835
gLASSO	0.974	0.745	0.852
mAIC	0.993	0.494	0.700
mLASSO	0.995	0.339	0.581
	35% complete cases		
Amelia II	**SEN**	**SPE**	**G**
CC	0.659	0.839	0.743
gLASSO	0.962	0.715	0.829
mAIC	0.987	0.350	0.587
mLASSO	0.993	0.254	0.502

3.4.2 Estimation

We also compared the estimates from the four approaches. The coefficient estimator and confidence interval of each covariate under mAIC and mLASSO come with the variable selection results. In order to obtain the estimates and confidence intervals for CC, the model is refitted with the selected covariates. For gLASSO, the selected variables are used to refit the model for each imputed dataset separately, then Rubin's rule is used to combine the inference from each dataset.

The coefficient estimates of six important variables and fourteen unimportant variables for 60% complete-cases were presented in the Box plots in Figure 3.3. Four horizontal lines at (-1.2, 0, 0.5, 1.8) were drawn as reference. For important variables, one can see that gLASSO has the best performance among the four scenarios, since the medians of important variables fall at the reference line and the range of 1st to 3rd quantile for each variable is quite narrow.

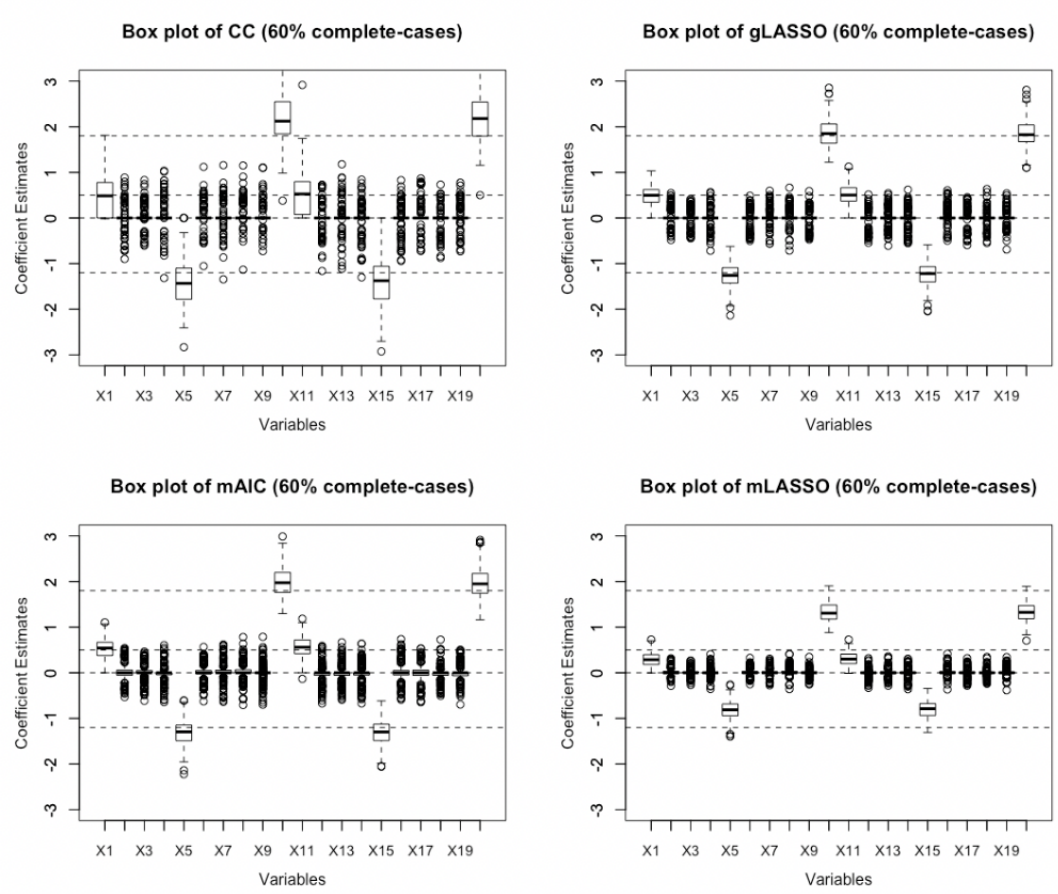

Figure 3.3: Estimates of each variable using the four methods among 200 replicates with 60% complete cases. The true model is $h\left(t|\mathbf{x}\right) = 0.5\exp(0.5X_1 - 1.2X_5 + 1.8X_{10} + 0.5X_{11} - 1.2X_{15} + 1.8X_{20})$ with censoring rate 30%. Variables $X_1 \sim X_{10}$ are complete, while variables $X_{11} \sim X_{20}$ have missing values resulting in 60% complete cases. The missing mechanism is MAR. Five (m=5) imputed datasets are generated in each replicate. Each missing replicate is multiply imputed with Amelia II.

While for unimportant variables, mLASSO have the least biased estimates as its ranges of outliners are pretty narrow. CC has the worst performance since it produces much biases on coefficient estimations.

3.4.3 Confidence Intervals

Empirical coverage probabilities of the 95% confidence intervals were calculated for each variable under different scenarios. For each covariate $X_i (i = 1, ...20)$, the empirical coverage probability can be obtained by

$$\text{Coverage Probability } (X_i) = \frac{\#\text{ of CI}(X_i) \text{ covering true parameter}}{200},$$

where the numerator is the number of confidence intervals covering the true parameter of X_i among 200 replicates.

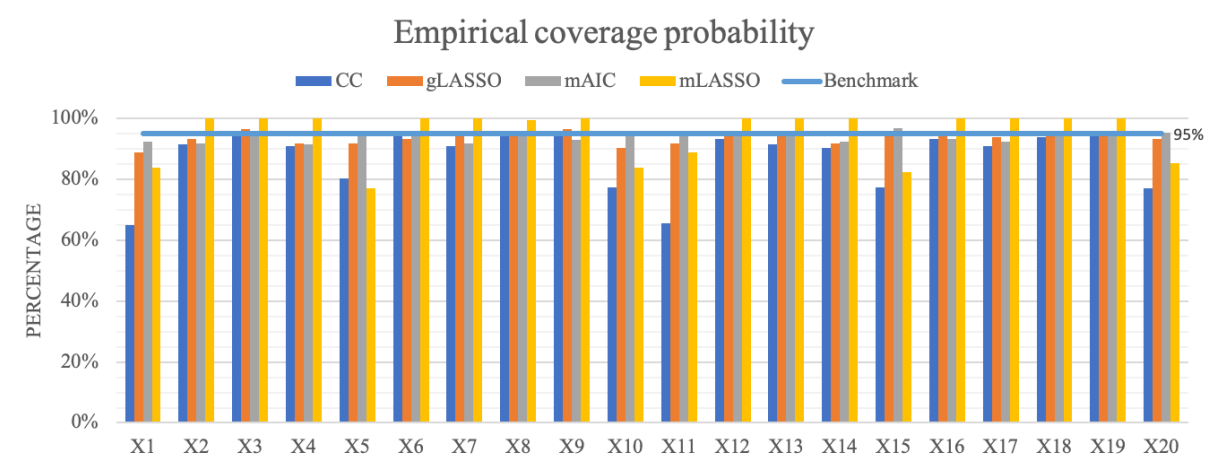

Figure 3.4: Percentage of each variable whose true parameter is covered by the 95% confidence interval given by the four methods among 200 replicates. The important variables are $X_1, X_5, X_{10}, X_{11}, X_{15}, X_{20}$ with censoring rate 30%. Variables $X_1 \sim X_{10}$ are complete, while variables $X_{11} \sim X_{20}$ have missing values resulting in 60% complete cases. The missing mechanism is MAR. Five (m=5) imputed datasets are generated in each replicate. Each missing replicate is multiply imputed with Amelia II.

The empirical coverage probability of each variable on 60% complete-cases is presented in Figure 3.4 for the four methods. The benchmark of 95% confidence interval was drawn

as reference. The shorter the distance between the probability and benchmark, the better the performance. CC has a terrible performance on estimating confidence interval of important variables. gLASSO and mAIC give the coverage probabilities of both important and unimportant variables slightly lower than 95%. The coverage probabilities of most unimportant variables given by mLASSO are higher than the benchmark. The majority of coverage probabilities from four methods are mildly under 95%. Thus, these four methods generally underestimate the true 95% confidence interval.

Table 3.2: Average coverage-probabilities of 95% confidence intervals. $(X_1, X_5, X_{10}, X_{11}, X_{15}, X_{20})$ are important and $(X_2 \sim X_4, X_6 \sim X_9, X_{12} \sim X_{14}, X_{16} \sim X_{19})$ are unimportant variables. Variables $X_1 \sim X_{10}$ are complete, while variables $X_{11} \sim X_{20}$ have missing values resulting in 60% complete cases. The missing mechanisms is MAR. Each missing replicate is multiply imputed with Amelia II.

	60% complete case			
Covariates	**CC**	**gLASSO**	**mAIC**	**mLASSO**
Important	74%	92%	96%	84%
Unimportant	93%	95%	94%	100%
All	87%	94%	94%	95%

The average coverage-probabilities of important, unimportant and all variables on 60% complete-cases are reported in Table 3.2. For important variables, the gLASSO and mAIC methods have relatively better performances than mLASSO and CC, where the average coverage probability of the confidence intervals of mAIC is 96% which is higher than 95% and gLASSO gives the average coverage probability of 92% which is close to 95%. On the other hand, for unimportant variables, all of the four methods have coverage probabilities close to the nominal level. Overall, gLASSO and mAIC perform better than the other methods.

Chapter 4

Application

The RAND data is longitudinal data containing a very large range of measures of elderly in the United States. The outcome of interest is time to diagnosis of dementia. The subjects were followed up every two years. The 35 variables used as candidate covariates were listed in Table 4.3 which were selected from sections A and B from the RAND file[7]. More details about the variables can be found in the RAND file. Most variables contain missing values except gender and cohort indicator. The missing percentages of most variables are around 11% to 18%. The missing data were multiply imputed with MICE by specifying each variable's type, assuming the missing mechanism is MAR. After 5 multiply-imputed datasets were generated, the three methods (gLASSO, mLASSO and mAIC) were applied to them to select important cognitive factors related to dementia. CC was applied to complete-cases dataset directly. Cross-validation was used to select the tuning parameters in CC, gLASSO and mLASSO.

The variable selection results from these four methods are shown in Table 4.1 and 4.2. The variables with category 1 are all binary, and the reference groups for them are 0. The reference group for BWC20 is also 0. The Table 4.1 and the top part of Table 4.2 show the 19 variables which were selected by all four methods (except HOMCAR, NRSHOM, and MEDUC), and the bottom part of Table 4.2 lists the variables which were selected by three methods or less. Among the 19 variables, the HR estimators of the variables AGE, EDUCATION, BMI, HOSP, IMRC, DLRC, DY, DW, PRES, and VP were pretty similar among all four methods. According to the 95% CI, these 10 variables are all significant at the 0.05 level. A larger value of AGE, EDUCATION, or HOSP is accompanied with a higher hazard of dementia. Conversely, for BMI, IMRC, DLRC, DY, DW, PRES, and VP, a higher

value is associated with lower hazard to develop dementia. For the remaining 9 variables, they are only found significant by three or less methods. Moreover, for the nominal and ordinal covariates (COHORT, SLFMEM, SHLT, and BWC20), they are partially significant in their sub-categories. In addition to the 19 common variables, other 16 variables were also selected by three or less methods which are not significantly associated with the outcome at the 0.05 significant level.

SER7 is merely found insignificantly by CC, while NRSHOM is merely found significantly by CC. According to my simulation results, CC gives untrustworthy confidence intervals comparing to other three methods. Hence, SER7 is likely to be significant, while NRSHOM is likely to be insignificant. Moreover, mAIC and mLASSO pick RACE (group 2) as a significant variable. Also, BWC20 (group2) is only found significantly by gLASSO. However, the nearest boundaries of these confidence intervals are very close to 1. It is inconclusive whether these variables are related to hazard of dementia.

Among the thirteen cognitive related measures BWC20, PRES, VP, SCIS, CACT, DY, MO, YR, DW, IMRC, DLRC, SER7, and VOCAB, only 6 variables PRES, VP, DY, DW, IMRC, DLRC are found significant by all the four methods. Generally, the larger the scores of these 6 variables, the less the hazard to develop dementia.

Table 4.1: Variable selection results of Complete-Cases (CC), gLASSO, mLASSO and mAIC for 35 variables on HRS data.

Variable	**Category**	**CC** HR (95%CI)	**gLASSO** HR (95%CI)	**mLASSO** HR (95%CI)	**mAIC** HR (95%CI)
COHORT	1	0.708 (0.376, 1.334)	0.865 (0.456, 1.642)	1.436 (0.971, 2.122)	0.892 (0.469, 1.697)
	2	0.522 (0.274, 0.994)	0.650 (0.339, 1.233)	1.094 (0.790, 1.514)	0.672 (0.351, 1.285)
	3	0.414 (0.217, 0.790)	0.451 (0.237, 0.860)	0.790 (0.574, 1.088)	0.478 (0.250, 0.912)
	4	0.250 (0.118, 0.530)	0.353 (0.173, 0.720)	0.721 (0.505, 1.031)	0.402 (0.195, 0.827)
	5	0.299 (0.140, 0.640)	0.323 (0.159, 0.660)	0.692 (0.490, 0.976)	0.385 (0.186, 0.796)
	6	0.323 (0.141, 0.737)	0.490 (0.230, 1.042)	–	0.592 (0.275, 1.273)
	0	reference	reference	reference	reference
AGE		1.047 (1.033, 1.060)	1.052 (1.040, 1.064)	1.052 (1.041, 1.063)	1.051 (1.039, 1.064)
EDUCATION		1.045 (1.024, 1.066)	1.040 (1.022, 1.058)	1.027 (1.009, 1.045)	1.037 (1.018, 1.057)
BMI		0.974 (0.961, 0.987)	0.979 (0.968, 0.990)	0.982 (0.971, 0.994)	0.980 (0.969, 0.992)
HOSP	1	1.256 (1.100, 1.434)	1.208 (1.078, 1.353)	1.193 (1.068, 1.333)	1.217 (1.094, 1.354)
DRUGS	1	1.128 (0.972, 1.310)	1.138 (0.990, 1.309)	1.129 (0.975, 1.307)	1.112 (0.888, 1.393)
IMRC		0.891 (0.849, 0.935)	0.900 (0.859, 0.942)	0.905 (0.864, 0.949)	0.899 (0.859, 0.942)
DLRC		0.897 (0.863, 0.935)	0.897 (0.862, 0.934)	0.899 (0.863, 0.936)	0.894 (0.857, 0.933)
SER7		0.964 (0.926, 1.002)	0.953 (0.923, 0.985)	0.949 (0.919, 0.981)	0.942 (0.912, 0.974)
DY	1	0.830 (0.718, 0.960)	0.833 (0.734, 0.944)	0.857 (0.757, 0.969)	0.837 (0.734, 0.955)
YR	1	0.747 (0.557, 1.002)	0.842 (0.671, 1.058)	0.874 (0.681, 1.122)	0.849 (0.597, 1.207)
DW	1	0.648 (0.480, 0.876)	0.623 (0.486, 0.798)	0.663 (0.506, 0.868)	0.620 (0.478, 0.805)
CACT	1	0.876 (0.721, 1.065)	0.898 (0.759, 1.063)	0.852 (0.714, 1.016)	0.819 (0.677, 0.992)
PRES	1	0.616 (0.481, 0.789)	0.636 (0.516, 0.784)	0.668 (0.538, 0.829)	0.638 (0.506, 0.805)
VP	1	0.679 (0.579, 0.796)	0.750 (0.655, 0.859)	0.761 (0.660, 0.877)	0.739 (0.643, 0.848)
SLFMEM	2	1.130 (0.852, 1.498)	1.210 (0.957, 1.531)	1.003 (0.838, 1.200)	1.197 (0.947, 1.516)
	3	1.300 (0.992, 1.703)	1.375 (1.099, 1.719)	1.144 (0.942, 1.389)	1.359 (1.082, 1.705)
	4	1.867 (1.411, 2.471)	2.020 (1.603, 2.546)	1.665 (1.338, 2.071)	1.994 (1.573, 2.527)
	5	2.697 (1.897, 3.835)	2.428 (1.829, 3.224)	1.957 (1.473, 2.598)	2.380 (1.769, 3.201)
	1	reference	reference	reference	reference
HOMCAR	1	1.017 (0.817, 1.265)	1.037 (0.862, 1.248)	1.032 (0.881, 1.209)	–
NRSHOM	1	1.596 (1.052, 2.421)	1.102 (0.799, 1.519)	1.093 (0.806, 1.482)	–
MEDUC		–	1.023 (1.001, 1.046)	1.015 (0.990, 1.040)	1.021 (0.994, 1.048)

Table 4.2: (Continuing) Variable selection results of CC, gLASSO, mLASSO and mAIC for 35 variables on HRS data.

Variable	**Category**	**Complete-Case** HR(95% CI)	**gLASSO** HR(95% CI)	**mLASSO** HR(95% CI)	**mAIC** HR(95% CI)
SHLT	2	1.113 (0.887, 1.395)	1.085 (0.894, 1.316)	1.070 (0.882, 1.299)	1.093 (0.901, 1.326)
	3	1.237 (0.990, 1.547)	1.149 (0.951, 1.389)	1.055 (0.915, 1.215)	1.164 (0.963, 1.407)
	4	1.261 (0.989, 1.607)	1.169 (0.955, 1.432)	1.078 (0.916, 1.269)	1.198 (0.979, 1.466)
	5	1.610 (1.208, 2.146)	1.590 (1.250, 2.024)	1.465 (1.175, 1.827)	1.643 (1.274, 2.118)
	1	reference	reference	reference	reference
GENDER	2	1.087 (0.915, 1.292)	1.085 (0.978, 1.203)	1.051 (0.938, 1.178)	1.087 (0.980, 1.207)
	1	reference	reference	reference	reference
BWC20	1	1.740 (0.756, 4.002)	1.086 (0.590, 2.001)	1.066 (0.644, 1.764)	1.114 (0.649, 1.913)
	2	0.971 (0.762, 1.237)	0.823 (0.680, 0.996)	0.827 (0.676, 1.012)	0.826 (0.569, 1.197)
MOMLIV	1	0.842 (0.661, 1.073)	–	0.866 (0.689, 1.089)	0.852 (0.611, 1.188)
MO	1	0.832 (0.601, 1.152)	–	0.977 (0.768, 1.243)	–
RELIGION	2	0.912 (0.796, 1.045)	–	0.921 (0.883, 1.032)	0.945 (0.788, 1.134)
	3	0.792 (0.561, 1.116)	–	0.889 (0.675, 1.171)	0.886 (0.586, 1.338)
	4	0.743 (0.543, 1.018)	–	0.869 (0.704, 1.073)	0.874 (0.575, 1.328)
	5	0.727 (0.410, 1.288)	–	0.794 (0.539, 1.192)	0.796 (0.376, 1.682)
	1	reference	–	reference	reference
SCIS	1	1.374 (0.732, 2.578)	–	1.007 (0.705, 1.439)	–
VETRN	1	0.919 (0.757, 1.116)	–	1.002 (0.892, 1.127)	–
VOCAB		–	–	1.016 (0.990, 1.042)	1.015 (0.969, 1.064)
DADLIV	1	–	–	0.898 (0.664, 1.215)	0.945 (0.626, 1.426)
RACE	2	–	–	0.845 (0.715, 0.999)	0.795 (0.677, 0.933)
	3	–	–	0.849 (0.669, 1.078)	0.799 (0.607, 1.053)
	1	–	–	reference	reference
DOCTOR	1	–	–	0.990 (0.799, 1.226)	0.969 (0.767, 1.225)
MRCT		–	–	1.064 (0.989, 1.145)	1.083 (1.003, 1.168)
FEDUC		–	–	1.000 (0.990, 1.015)	–
MPART	1	–	–	1.037 (0.752, 1.430)	–
OUTPT	1	–	–	0.990 (0.891, 1.099)	–

Table 4.3: Thirty-five candidate variables considered in variable selection and their labels on HRS data.

Variable	**Label** & **Category**
COHORT	Sample cohort: 0.Hrs/Ahead ovrlap; 1.Ahead; 2.Coda; 3.Hrs; 4.WarBabies; 5.Early BabyBoomers; 6.Mid BabyBoomers
BWC20	Whether the Respondent was able to successfully count backwards for 10 continuous numbers from 20: 0.Incorrect; 1.Correct, 2nd try; 2.Correct, 1st try
PRES / VP	Whether Respondent was able to name the president / vice-president: 0.Incorrect; 1.Correct
SCIS / CACT	Whether the Respondent was able to correctly name scissors /cactus: 0.Incorrect; 1.Correct
DY / MO / YR / DW	Whether the Respondent was able to report today's date correctly, including the day of month, month, year, and day of week, respectively: 0.Incorrect; 1.Correct
DOCTOR	Whether the Respondent reports any doctor visit in the reference period: 0.No; 1.Yes
HOMCAR	Whether the Respondent reports any home health care in the reference period: 0.No; 1.Yes
HOSP	Whether the Respondent reports any overnight hospital stay in the reference period: 0.No; 1.Yes
NRSHOM	Whether the Respondent reports any overnight nursing home stay in the reference period: 0.No; 1.Yes
DRUGS / OUTPT	Whether the Respondent reports regular use of prescription drugs / outpatient surgery: 0.No; 1.Yes
SHLT	Respondent's self-reported general health status: 1.Excellent; 2.Very good; 3.Good; 4.Fair; 5.Poor
SLFMEM	Self-reported general rating of memory: 1.Excellent; 2.Very good; 3.Good; 4.Fair; 5.Poor
IMRC / DLRC	Measures for immediate word recall / delayed word recall
SER7	The number of correct subtractions in the serial 7s test
VOCAB	The sum of scores over five words
AGE	Age (years) at interview end month
EDUCATION	Years of education
BMI	Body Mass Index=kg/m2
GENDER	1.Male; 2.Female
RACE	1.White/Caucasian; 2.Black/African American; 3.Other
RELIGION	1.Protestant; 2.Catholic; 3.Jewish; 4.None/no pref; 5.Other
MPART	Whether living with a partner who is not the Respondent's spouse: 0.No; 1.Yes
VETRN	Veteran status: 0.No; 1.Yes
DADLIV / MOMLIV	Whether a Respondent's father / mother is still alive: 0.No; 1.Yes
FEDUC / MEDUC	Father's / mother's years education
MRCT	# of marriages

Chapter 5

Discussion and Conclusion

We have compared CC and three different methods for variable selection for the Cox PH model after multiple imputation. The mAIC method uses AIC criteria to select important variables on each multiply-imputed dataset. The mLASSO method, a method similar to mAIC, combines the results of lasso variable selection on each multiply-imputed dataset. The third method is gLASSO, which deals with variable selection after multiple imputation. The basic idea is to constrain the coefficients of the same variable in each imputed data set to a group, then lasso will be applied to the grouped and constrained variables to obtain a consistent result from variable selection. The last method is CC, which simply applies lasso variable selection on complete-cases dataset.

Simulation study was used to compare the methods. We considered different values for the β, from negative to positive and from small to large, in order to reflect if the coefficients would affect the performance of variable selection. We also compared the approaches under different degree of missingness. Overall, gLASSO has better performance than the other three methods.

According to the results of application on HRS data, the demographic and health characteristics of AGE, EDUCATION, and HOSP are significantly associated with hazard of dementia. A higher value is related with a higher hazard to develop dementia. The health and cognitive related factors of BMI, IMRC, DLRC, DY, DW, PRES, and VP are negatively correlated with the development of dementia at 0.05 significant level. Specifically, the higher the scores of these variables, the lower the hazards of developing dementia.

Bibliography

[1] Gallo, Joseph J., Robert Schoen, and Richard Jones. "Cognitive impairment and syndromal depression in estimates of active life expectancy: the 13-year follow-up of the Baltimore Epidemiologic Catchment Area sample." *Acta Psychiatrica Scandinavica* 101.4 (2000): 265-273.

[2] Agüero-Torres, Hedda, et al. "Mortality from dementia in advanced age: a 5-year follow-up study of incident dementia cases." *Journal of clinical epidemiology* 52.8 (1999): 737-743.

[3] SHAPSE, STEVEN N. "THE DIAGNOSTIC AND STATISTICAL MANUAL OF MENTAL DISORDERS i." (2008).

[4] Manton, K. G., X. L. Gu, and S. V. Ukraintseva. "Declining prevalence of dementia in the US elderly population." *Adv Gerontol* 16 (2005).

[5] Kempen, Gertrudis IJM, et al. "Morbidity and quality of life and the moderating effects of level of education in the elderly." *Social science & medicine* 49.1 (1999): 143-149.

[6] Michael J. Behm, Mark J. Bernstein, Shauna Ryder Diggs, et al. "The Health and Retirement Study Data Book". [January 2017] Available at: http://hrsonline.isr.umich.edu/sitedocs/databook/inc/pdf/HRS-Aging-in-the-21St-Century.pdf

[7] Bugliari, Delia, et al. "RAND HRS Longitudinal File 2014 (V2) Documentation." (2018).

[8] Kleinbaum, David G., and Mitchel Klein. *Survival analysis*. Vol. 3. New York: Springer, 2010.

[9] Fox, John. "Cox proportional-hazards regression for survival data." *An R and S-PLUS companion to applied regression* 2002 (2002).

[10] Langa, Kenneth M., et al. "Trends in the prevalence and mortality of cognitive impairment in the United States: is there evidence of a compression of cognitive morbidity?." *Alzheimer's & Dementia* 4.2 (2008): 134-144.

[11] Crimmins, Eileen M., et al. "Assessment of cognition using surveys and neuropsychological assessment: the Health and Retirement Study and the Aging, Demographics, and Memory Study." *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 66.suppl_1(2011) :i162-i171.

[12] Efron, Bradley. "Logistic regression, survival analysis, and the Kaplan-Meier curve." *Journal of the American statistical Association* 83.402 (1988): 414-425.

[13] Cox, David R. "Regression models and life-tables." *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972): 187-202.

[14] Rubin, Donald B. "Multiple imputation after 18+ years." *Journal of the American statistical Association* 91.434 (1996): 473-489.

[15] Rubin, Donald B. "Inference and missing data." *Biometrika* 63.3 (1976): 581-592.

[16] Bursac, Zoran, et al. "Purposeful selection of variables in logistic regression." *Source code for biology and medicine* 3.1 (2008): 17.

[17] Efroymson, M. A. "Multiple regression analysis." *Mathematical methods for digital computers* (1960): 191-203.

[18] Tibshirani, Robert. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996): 267-288.

[19] Tibshirani, Robert. "The lasso method for variable selection in the Cox model." *Statistics in medicine* 16.4 (1997): 385-395.

[20] Yuan, Ming, and Yi Lin. "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1 (2006): 49-67.

[21] Fan, Jianqing, and Runze Li. "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American statistical Association* 96.456 (2001): 1348-1360.

[22] Zou, Hui. "The adaptive lasso and its oracle properties." *Journal of the American statistical association* 101.476 (2006): 1418-1429.

[23] Zhang, Hao Helen, and Wenbin Lu. "Adaptive Lasso for Cox's proportional hazards model." *Biometrika* 94.3 (2007): 691-703.

[24] Wood, Angela M., Ian R. White, and Patrick Royston. "How should variable selection be performed with multiply imputed data?." *Statistics in medicine* 27.17 (2008): 3227-3246.

[25] Chen, Qixuan, and Sijian Wang. "Variable selection for multiply-imputed data with application to dioxin exposure study." *Statistics in medicine* 32.21 (2013): 3646-3659.

[26] Schomaker, M. and C. Heumann (2014). Model selection and model averaging after multiple imputation. *Computational Statistics and Data Analysis* 71, 758–770.

[27] Breslow, Norman. "Covariance analysis of censored survival data." *Biometrics* (1974): 89-99.

[28] Efron, Bradley. "The efficiency of Cox's likelihood function for censored data." *Journal of the American statistical Association* 72.359 (1977): 557-565.

[29] Kalbfleisch, John D., and Ross L. Prentice. "Marginal likelihoods based on Cox's regression and life model." *Biometrika* 60.2 (1973): 267-278.

[30] Hertz-Picciotto, Irva, and Beverly Rockhill. "Validity and efficiency of approximation methods for tied survival times in Cox regression." *Biometrics* (1997): 1151-1156.

[31] Honaker, James, Gary King, and Matthew Blackwell. "Amelia II: A program for missing data." *Journal of statistical software* 45.7 (2011): 1-47.

[32] Buuren, S. van, and Karin Groothuis-Oudshoorn. "mice: Multivariate imputation by chained equations in R." *Journal of statistical software* (2010): 1-68.

[33] Schomaker, M. and C. Heumann (2017). *Model averaging and model selection after multiple imputation using the R-package MAMI*; http://mami.r-forge.r-project.org/

[34] Simon, Noah, et al. "A sparse-group lasso." *Journal of Computational and Graphical Statistics* 22.2 (2013): 231-245.