ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

Dealing with missing values on variables using multiple imputation methods to Cox regression analysis

Νικόλαος Παπαδημητρίου

ΕΡΓΑΣΙΑ

Που υποβλήθηκε στο Τμήμα Στατιστικής του Οικονομικού Πανεπιστημίου Αθηνών ως μέρος των απαιτήσεων για την απόκτηση Διπλώματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

December 2020

Dedication

To my family

Acknowledgements

There are a lot of people without whom this thesis might not have been written and to whom I am greatly indebted.

First of all, I would like to thank my supervisor Vasileios Vasdekis from the Department of Statistics who gave me the opportunity to gain this valuable experience working with him. At many stages in the course of this thesis I benefited from his advice and always his support and encouragement were genuine. His positive outlook and confidence in my research inspired me and gave me confidence. His careful editing contributed enormously to the production of this thesis.

Secondly, I would like to thank my parents, who always have been a source of encouragement and inspiration to me throughout my life. A very special thank you for the countless ways you have actively supported me in my determination to find and realise my potential.

Finally, I would like to thank the Professors Dimitrios Varsos and Evangelos Raptis from the Department of Mathematics of the National and Kapodistrian University of Athens whose passion for teaching set a new standard for anyone involved in education and development or any other endeavour in which a human being seeks to support.

Memoir

Nikolaos Papadimitriou was born and grew up in Athens. He obtained his Master's degree in Statistics of the Athens University of Economics and Business. His educational background includes a Bachelor's degree in Mathematics of the National and Kapodistrian University of Athens.

Abstract

Nikolaos Papadimitriou

Dealing with missing values on variables using multiple imputation methods to Cox regression analysis

December 2020

In the field of Survival Analysis, where the complete case analysis is the common method, we exclude cases with missing values. In order to take advantage of the whole dataset, we propose multiple imputation methods to cope with missing data. To implement these methods, a fully observed variable is necessary to exist in the dataset. This fully observed variable offers closest to the real values estimations of the other variable with missing values. More specifically, the proposed multiple imputation methods in this thesis are the following: the semi-parametric predictive mean matching and the non-parametric nearest neighbor multiple imputation. In order to evaluate the performance of the aforementioned methods, Cox regression analysis is employed. In the end, the methods are compared in terms of efficiency, robustness and consistency.

Περίληψη

Νικόλαος Παπαδημητρίου

Αντιμετωπίζοντας ελλειπούσες τιμές σε μεταβλητές χρησιμοποιώντας μεθόδους πολλαπλής απόδοσης τιμών στην παλινδρόμηση Cox

Δεκέμβριος 2020

Στο πεδίο της Ανάλυσης Επιβίωσης, όπου η complete case analysis είναι η κοινή μέθοδος, αποκλείουμε παρατηρήσεις με ελλειπούσες τιμές. Προκειμένου να εκμεταλλευτούμε ολόκληρο το σύνολο δεδομένων, προτείνουμε πολλαπλές μεθόδους απόδοσης τιμών για την αντιμετώπιση των δεδομένων που λείπουν. Για την εφαρμογή αυτών των μεθόδων, απαιτείται μια πλήρως παρατηρούμενη μεταβλητή στο σύνολο δεδομένων. Αυτή η πλήρως παρατηρούμενη μεταβλητή προσφέρει πλησιέστερες εκτιμήσεις των πραγματικών τιμών της άλλης μεταβλητής με τις ελλειπούσες τιμές. Πιο συγκεκριμένα, οι προτεινόμενες μέθοδοι πολλαπλού καταλογισμού σε αυτή τη διπλωματική εργασία είναι οι εξής: η ημι-παραμετρική predictive mean matching και η μη παραμετρική nearest neighbor multiple imputation. Προκειμένου να αξιολογηθεί η απόδοση των προαναφερθεισών μεθόδων, χρησιμοποιείται ανάλυση παλινδρόμησης Cox. Εν τέλει, οι μέθοδοι συγκρίνονται ως προς την αποτελεσματικότητα, την ευρωστία και τη συνέπεια.

Contents


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

List of Tables


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Chapter 1

Introduction

One of the most popular ways to specify the relationship between survival time and variables is Cox regression analysis. It estimates the coefficients of the model using the partial likelihood function without the need of specifying the baseline hazard function. The estimators are normally distributed, consistent and semi-parametrically efficient. In case of time-independent variables, Cox regression has the proportional hazards property. There are often situations though, in which variables are not fully observed. If complete case analysis is implemented, excluding the cases which the variable has missing data, it has been shown that Cox regression loses efficiency and also leads to biased regression coefficient estimates. When missingness depends on the survival outcome and some fully observed variables, missing mechanism is considered as missing at random (MAR). When missingness does not depend on failure time (failure ignorable MAR), complete case analysis can produce some valid results. But, when missingness does not depend on censoring time but may depend on failure time, complete case analysis fails to produce valid results.

To deal with missing data many methods have been developed in order to produce valid results when performing Cox regression analysis. The methods developed are categorized as parametric, semi-parametric and non-parametric. We will use two multiple imputation methods. Predictive mean matching, which is a semi-parametric method based on multiple imputation by chain equations and nearest neighbor multiple imputation, which is a non-parametric method where two working models are used, one for the missing probability and one for the missing value, in order to create an imputation set with possible donors for the missing data of the variable. In this study, we will not implement a misspecified situation for the predictive mean matching method. We will investigate the performance of nearest neighbor multiple imputation method, as well as the performance of nearest neighbor multiple imputation when the model of the missing probability is misspecified. The purpose of this study is to compare the performance of Complete Case analysis, Predictive Mean Matching and Nearest Neighbor multiple imputation methods on different situations and compare their results.

The structure of this study will be the following: In chapter 2, there is a review of survival analysis. What is survival analysis, definitions and theory of what we will use in this study. In chapter 3, we give brief definitions of the missing mechanisms. In chapter 4, we review the methods we will implement in this study.First, we present the parametric complete case analysis. Subsections 4.2.1, 4.2.2 introduce the semi-parametric and non-parametric methods respectively, the predictive mean matching and the nearest neighbor multiple imputation. In chapter 5, we apply all the previous methods to real data, which have to do with a case of a heart transplant and a simulated study is conducted to evaluate

the proposed methods. Lastly in chapter 6, there is a brief discussion about the results of the simulation as well as some thoughts for future researching.

Chapter 2

Survival Analysis

Survival analysis is a collection of data from different independent individuals for which the outcome variable of interest is time until an event occurs. Time is a continuous variable and takes values depending on the experiment or the study we want to conduct. Time could be hours, days, months or years from the beginning of follow-up of an individual until an event occurs. Event is an incidence or experience of interest that may happen to an individual. By event we mean death, recovery, disease incidence, disease recession etc. In a survival analysis, we refer to the time variable as survival time, because it shows us the time of an individual during some follow-up period. We also refer to the event as a failure, because an individual failed to "survive" and the event of interest happened.

What makes survival analysis unique from other forms of analyses are the so called censored data. In short, censoring happens to our collection of data when we have some information about individual survival time, but we don't know the exact time when the event occurred during the follow-up period. There are three types of censored data. Right-censored, left-censored and interval-censored. Most survival data is right-censored. Right-censoring happens when for some individuals, the failure times have not been observed even though the follow-up period has ended. They may reach the event after the follow-up period. As a result, these observed data are censored. Left-censoring is the exact opposite of right-censoring. The event of interest of some individuals has occurred before the follow-up period starts. Lastly, interval-censoring happens when during the follow-up period, the event occurs within a time interval (t_1, t_2). There are three assumptions about censoring for survival data: independent censoring, random censoring and non-informative censoring. Independent censoring means that within any subgroup of interest, the subjects who are censored at time t should be representative of all the subjects in that subgroup who remained at risk at time t with respect to their survival experience. The assumption of independence is the most useful of the three types for drawing correct inferences that compare the survival experience of two or more groups. Random censoring means that subjects who are censored at time t should be representative of all the study subjects who remained at risk at time t with respect to their survival experience. Random censoring is a stronger assumption and more restrictive than independent censoring. Lastly, we consider the assumption of non-informative censoring. Whether censoring is non-informative or informative depends on two distributions: 1) the distribution of the time-to-event random variable and 2) the distribution of time-to-censorship random variable. Non-informative censoring occurs if the distribution of survival times T provides no information about the distribution of censoring times C and vice versa. Otherwise, the censoring is informative. The assumption of non-informative censoring is justifiable when censoring is independent

and/or random. Nevertheless, these three assumptions are not equivalent.

We introduce basic mathematical terminology and notation for survival analysis. We denote by a capital T the random variable for an individual's survival time. T can any number equal to or greater than zero. By a small letter t we denote any specific value of interest for the random variable T. Finally, we denote by δ_t the indicator function to define if we have failure or censorship. Meaning, $\delta_t = 1$ for failure or $\delta_t = 0$ for censorship. We next introduce two quantitative functions widely used in survival analysis. These are the survival function, denoted by $S(t)$ and the hazard function, denoted by $h(t)$. The survival function $S(t)$ gives us the probability that a person survives longer than some specified time t; $S(t) = P(T > t)$. The survival function is fundamental, because obtaining survival probabilities for different values of t is crucial and important for survival analysis. In most cases, we are more interested in how long the individuals in a study live, than how quickly they die. We continue with the hazard function, denoted by $h(t)$, which is given by the following formula:

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

The interpretation of the hazard function in practical terms is not an easy task. In essence, the hazard function $h(t)$ gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t. For instance, you are driving your car and the speedometer shows 100kmh. It means that if you continue driving this way, then in one hour you cover 100km. This is not absolute though, because if you slow down or speed up or even stop at some point you may or may not cover that distance in one hour. That is what potential really means. Note here, that the survival function's main focus is not failing in contrast to the hazard function which is the event occurring. Thus, the hazard function gives us the opposite side of information of the survival function. The relationship between the two can be described from the following formulas:

$$S(t) = exp\Big[- \int_0^t h(u)du \Big] \quad \& \quad h(t) = -\Big[\frac{dS(t)/dt}{S(t)} \Big].$$

The modeling for the survival data though, is rather challenging and difficult. The main problems are the nature of data, meaning that there is an inherent aging process when subjects are followed over time and the presence of censoring data. The hazard function described above captures the essence of the aging process. Thus, a regression like model is built around the hazard function. The hazard function is a rate, so it must be strictly positive. However, for a statistical model we need a property to be parameterized in a way that the allowable range of parameter values is infinite. This also helps for the estimation of the parameters. To fix this problem we parameterize the hazard function as

$$h(t) = e^{\beta_0},$$

where $\beta_0 = ln(\theta_0)$ and is thus unconstrained. Given the above form, we include variables by being additive on the log scale as follows:

$$ln[h(t, x)] = \beta_0 + \beta_1 x$$

and the hazard function is

$$h(t, x) = e^{\beta_0 + \beta_1 x}.$$

A fully parametric model accomplishes two goals simultaneously. It describes the basic underlying distribution of survival time (which is called the error component) and it characterizes how that distribution changes as a function of the variables (which is called the systematic component). It is favorable to have a model which accomplishes both goals but in our case we are interested in the systematic component. This categorizes the above model as a semi parametric regression model. The baseline hazard function $h_0(t)$ makes the the model semiparametric. the baseline hazard function is a generalization of the intercept or constant term found in parametric regression models. Consequently, the hazard function is the product of two functions:

$$h(t, x, \beta) = h_0(t)r(x, \beta).$$

The function $h_0(t)$ characterizes how the hazard function changes as a function of survival time. The other function $r(x, \beta)$, characterizes how the hazard function changes as a function of subject variables. The above model is called Cox regression model because it was proposed by Cox in 1972. The semiparametric property is what makes the Cox model popular. Even though the baseline hazard function is unspecified, the Cox model produces good regression coefficient estimates, hazard ratios of interest and survival curves for a wide variety of data situations. In other words, the Cox model is a robust model which the results will approximate the results of the correct parametric model. The measure of effect, which is the hazard ratio is calculated without having to estimate the baseline hazard function. The hazard ratio for two subjects with variable values denoted x_1 and x_2 is:

$$HR(t, x_1, X_0) = \frac{h(t, x_1, \beta)}{h(t, x_0, \beta)} HR(t, x_1, x_0) = \frac{h_0(t)r(x_1, \beta)}{h_0(t)r(x_0, \beta)} = \frac{r(x_1, \beta)}{r(x_0, \beta)}.$$

Cox model is also referred to as Cox proportional hazards model. The term proportional hazards refers to the fact that the hazard functions are multiplicatively related, meaning their ratio is constant over survival time. This is a very important assumption and there are methods assessing its validity and existence in the model. Other parameterizations exist like the additive relative hazard model whose function is

$$h(t, x, \beta) = h_0(t)(1 + x\beta)$$

but will not be included and referred to further in this study.

We stated above that the distribution of survival time can be specified through the hazard function. Now that we have specified the hazard function we can use it to specify the survival function, so we have

$$S(t, x, \beta) = e^{-H(t,x,\beta)},$$

where $H(t, x, \beta)$ is the cumulative hazard function at time t for a subject with variable x. We assume that the survival time is continuous so

$$H(t, x, \beta) = \int_0^t h(u, x, \beta)du = r(x, \beta) \int_0^t h_0(u)du = r(x, \beta)H_0(t).$$

Combining the two equations we have the following:

$$S(t, x, \beta) = e^{-r(x,\beta)H_0(t)} = [e^{-H_0(t)}]^{r(x,\beta)} = [S_0(t)]^{r(x,\beta)},$$

where $S_0(t) = e^{-H_0(t)}$ is the baseline survival function.

We proceed with the estimation method of the model parameters. As with logistic regression, the maximum likelihood (ML) estimates of the Cox model parameters are derived by maximizing a likelihood function, usually denoted as L. The likelihhood function is a mathematical expression which describes the joint probability of obtaining the data observed on the subjects in the study as a function of the unknown parameters in the model being considered. The likelihood method proposed by Cox for estimating the model parameters is a little different. The method proposed by Cox takes into consideration the probabilities for each subject who fails and does not consider probabilities for subjects who are censored. That is the reason for the method's name, which is called "partial" likelihood instead of (complete) likelihood. The full likelihood, under the assumption of independent observations, is obtained by multiplying the respective contributions of the observed (t, β, x), a value of $f(t, \beta, x)$ for a non censored observation (c=1) and a value of $S(t, \beta, x)$ for censored observations (c=0). The expression is the following:

$$[f(t, \beta, x)]^c \times [S(t, \beta, x)]^{1-c},$$

where $c = 0$ or $c = 1$. Thus, the likelihood function, because of the independence assumption, is the following:

$$L(\beta) = \prod_{i=1}^{n} \{[f(t_i, \beta, x_i)]^{c_i} \times [S(t_i, \beta, x_i)]^{1-c_i}\}.$$

To obtain the maximized likelihood with respect to the parameter of interest β, we maximize the log-likelihood function:

$$ln(L(\beta)) = \sum_{i=1}^{n} \{c_i ln[f(t_i, \beta, x_i)] + (1 - c_i) ln[S(t_i, \beta, x_i)]\}.$$

The log function is monotone, so the maximum value of β is the same. However, computing the maximum with the log function is simpler. The partial likelihood proposed by Cox is given from the following expression:

$$L_p(\beta) = \prod_{i=1}^{n} \left[\frac{e^{x_i \beta}}{\sum_{j \in R(t_i)} e^{x_j \beta}} \right]^{c_i},$$

where the sum in the denominator is over all subjects in the risk set at time t_i, denoted by $R(t_i)$. For the same reasons as mentioned above, we use the log function to calculate the maximum from the following expression:

$$ln(L_p(\beta)) = \sum_{i=1}^{m} \left\{ x_i \beta - ln \left[\sum_{j \in R(t_i)} e^{x_j \beta} \right] \right\}.$$

We obtain the maximum partial likelihood estimator by differentiating the log partial likelihood function with respect to β, setting the derivative equal to zero and solving for the unknown parameter. The derivative with respect to β is

$$\frac{\partial L_p(\beta)}{\partial \beta} = \sum_{i=1}^{m} \left\{ x_i - \frac{\sum_{j \in R(t_i)} x_j e^{x_j \beta}}{\sum_{j \in R(t_i)} e^{x_j \beta}} \right\} = \sum_{i=1}^{m} \left\{ x_i - \sum_{j \in R(t_i)} w_{ij}(\beta) x_j \right\} = \sum_{i=1}^{m} \{x_i - \overline{x}_{w_i}\},$$

where

$$w_{ij}(\beta) = \frac{e^{x_j \beta}}{\sum_{j \in R(t_i)} e^{x_j \beta}}$$

and

$$\overline{x}_{w_i} = \sum_{j \in R(t_i)} w_{ij}(\beta) x_j,$$

which we denote as $\hat{\beta}$.

The estimator of the variance of the estimator of the coefficient is obtained from the inverse of the negative of the second derivative of the log partial likelihood as shown in the following expression:

$$\frac{\partial^2 L_p(\beta)}{\partial \beta^2} = -\sum_{i=1}^{m} \left\{ \frac{\left[\sum_{j \in R(t_i)} e^{x_j \beta}\right]\left[\sum_{j \in R(t_i)} x_j^2 e^{x_j \beta}\right] - \left[\sum_{j \in R(t_i)} x_j e^{x_j \beta}\right]^2}{\left[\sum_{j \in R(t_i)} e^{x_j \beta}\right]^2} \right\},$$

which can be simplified by using the definition of $w_{ij}(\beta)$ and becomes:

$$\frac{\partial^2 L_p(\beta)}{\partial \beta^2} = -\sum_{i=1}^{m} \sum_{j \in R(t_i)} w_{ij}(\beta)(x_j - \overline{x}_{w_i})^2.$$

The negative of the second derivative of the log partial likelihood is called the observed information and we will denote it as

$$\mathbf{I}(\beta) = -\frac{\partial^2 L_p(\beta)}{\partial \beta^2}.$$

If the model contains more than one variable then the result is called observed information matrix. The variance of the estimated coefficient $\hat{\beta}$ is

$$\hat{Var}(\hat{\beta}) = \mathbf{I}(\hat{\beta})^{-1}.$$


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Chapter 3

Missing Mechanisms

Missing data or missing values occur when no data value is stored for the variable in an observation. Rubin classified missing data into three categories. In his theory every datum has some likelihood of being missing. Missingness can then be categorised as missing at random (MAR), missing completely at random (MCAR) and missing not at random (MNAR). Missing data can be handled similarly as censored data.

The terms missing at random and missing completely at random are used to describe assumptions about missing data that are needed for standard implementations of multiple imputation, but the meanings of these terms are often confused. When we have the missing completely at random mechanism (MCAR), means that the missing observations are a random subset of all observations. Thus, the missing and observed values will have similar distributions. There is nothing systematic going on that makes some data more likely to be missing than others. Missing at random (MAR) means there might be systematic differences between the missing and observed values, but these can be explained by other observed variables which are fully observed. On the other hand, the missing not at random (MNAR) has a systematic relationship between the missing values and the missing data which must be considered.

Missing completely at random (MCAR) and missing at random (MAR) are considered ignorable mechanisms, because we don't have to include any information about the missing data when we deal with the missing data. Missing not at random (MNAR) mechanism is considered non-ignorable, because the missing data have to be modeled and you have to figure out the reasons why the data are missing and predict their possible values. In this study the missing at random (MAR) mechanism will be used.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Chapter 4

Review of Methods

4.1 Complete Case Analysis

The complete case (CC) analysis uses only the observations which have all the variables observed and is based on the partial likelihood to estimate $\mathbf{B} = (\beta_x, \beta_z)$. Let $r_i(\mathbf{B}, t) = e^{\beta_x X_i + \beta_z Z_i} \equiv r_i^{(0)}(\mathbf{B}, t)$ and $r_i^{(1)} = (X_i Z_i)^t r_i(\mathbf{B}, t)$. The estimators are the solution from the following equations

$$U_{cc} = \sum_{i=1}^{n} \left[\delta_{t_i} \delta_{x_i} \left\{ X_i Z_i - \frac{S_{cc}^{(1)}(\mathbf{B}, T_i)}{S_{cc}^{(0)}(\mathbf{B}, T_i)} \right\} \right] = 0$$

where $\delta_t = I[T \leq C]$ is the censoring indicator, δ_x is the missing indicator ($\delta_x = 1$ if X is observed; otherwise 0), $S_{cc}^{(m)}(\mathbf{B}, T_i) = n^{-1} \sum_{j=1}^{n} \delta_{x_i} I(T_j \geq T_i) r_j^{(m)}(\mathbf{B}, T_i)$ for $m = 0, 1$. The CC analysis is simple to implement and is widely used. It does perform fairly well when the missingness depends on Z and the missing rate is not greater than 25%. Its inconsistency may also be caused when missingness depends on failure time T or censoring indicator δ_t.

4.2 Multiple Imputation Methods

4.2.1 Predictive Mean Matching

Predictive mean matching (PMM) is a semi-parametric imputation method. It is very close to the regression method in a sense except that for each missing value it fills in a value randomly among the observed donor values from an observation whose regression-predicted values are closest to the regression-predicted value for the missing value from the simulated regression model[2].

Predictive mean matching is a quite easy and versatile method to use. It performs fairly well even when the target variable is transformed, meaning that $log(Y)$ often yields to similar results as $exp(Y)$. The method can be also used for categorical data. The fact that there is no need to define an explicit model for the distribution of the missing values makes the method less vulnerable to model misspecification to some extent.

For univariate X with missing values we denote as Z_{obs} the subset of n_1 rows of predictor variable Z for which X is observed and as Z_{mis} the complementing subset of n_0 rows of Z for which X is missing. The vector containing the n_1 observed data in X is denoted as X_{obs} and the vector of n_0 imputed values in X is indicated as $\dot{X}$. The bootstrap multiple imputation model $\dot{X} = \dot{\beta}_0 + X_{mis}\dot{\beta}_1 + \dot{\epsilon}$, where $\dot{\epsilon} \sim N(0, \dot{\sigma}^2)$ and $\dot{\beta}_0, \dot{\beta}_1, \dot{\sigma}$ are the least squares estimates calculated from a bootstrap sample taken from the observed data, is estimated from the following steps:

1. Draw a bootstrap sample $(\dot{X}_{obs}, \dot{Z}_{obs})$ of size n_1 from (X_{obs}, Z_{obs}).
2. Calculate the cross-product matrix $\dot{S} = \dot{Z}_{obs}^T \dot{Z}_{obs}$.
3. Calculate $\dot{V} = (\dot{S} + diag(\dot{S})\kappa)^{-1}$, with some small κ.
4. Calculate regression weights $\dot{\beta} = \dot{V}\dot{Z}_{obs}^T \dot{X}_{obs}$.
5. Calculate $\dot{\sigma}^2 = (\dot{X}_{obs} - \dot{Z}_{obs}\dot{\beta})^T(\dot{X}_{obs} - \dot{Z}_{obs}\dot{\beta})/(n_1 - q - 1)$.
6. Draw n_0 independent $N(0, 1)$ variates in vector $\dot{c}_2$.
7. Calculate the n_0 values $X_{imp} = Z_{mis}\dot{\beta} + \dot{c}_2\dot{\sigma}$.

According to Andridge and Little[3] there are four distinguished methods to select a donor once the metric has been defined. Although various metrics exist to define the distance between cases, Rubin[4] and Little[5] proposed the predictive mean matching metric.

Let $\hat{y}_i$ denote the predicted value of the rows with an observed y_i where $i = 1, ..., n_1$. Likewise, let $\hat{y}_j$ denote the predicted value of the rows with missing y_j where $j = 1, ...n_0$.

1. Choose a threshold η, and take all i for which $|\hat{X}_i - \hat{X}_j| < \eta$ as candidate donors for imputing j. Randomly sample one donor from the candidates and take its y_i as replacement value.
2. Take the closest candidate, i.e the case i for which $|\hat{X}_i - \hat{X}_j|$ is minimal as the donor. This is known as "nearest neighbor hot deck" or "closest predictor".
3. Find the d candidates for which $|\hat{X}_i - \hat{X}_j|$ is minimal and sample on of them. Usual values for d are 3, 5 and 10. There is also an adaptive method to specify the number of donors[6].

4. Sample one donor with a probability that depends on $|\hat{X}_i - \hat{X}_j|$[7].

Additionally, we distinguish four types of matching:

- $Type\ 0 : \hat{X} = Z_{obs}\hat{\beta}$ is matched to $\hat{X}_j = Z_{mis}\hat{\beta}$;
- $Type\ 1 : \hat{X} = Z_{obs}\hat{\beta}$ is matched to $\dot{X}_j = Z_{mis}\dot{\beta}$;
- $Type\ 2 : \dot{X} = Z_{obs}\dot{\beta}$ is matched to $\dot{X}_j = Z_{mis}\dot{\beta}$;
- $Type\ 3 : \dot{X} = Z_{obs}\dot{\beta}$ is matched to $\ddot{X}_j = Z_{mis}\ddot{\beta}$.

The estimate of β is denoted as $\hat{\beta}$ and $\dot{\beta}$ is a value randomly drawn from the posterior distribution of β. Sampling variability is ignored with Type 0 matching, which leads to improper imputations. Type 2 solves this, but it is insensitive to the process of taking random draws of β if there are only a few variables. In the extreme case, with a single Z, the set of candidate donors based on $|\dot{X}_i - \dot{X}_j|$ remains unchanged under different values of $\dot{\beta}$, so the same donor(s) get selected too often. A small adaptation of the matching distance that seems to alleviate the problem is type 1. The difference between Type 0, Type 2 and Type 1 is that in Type 1 matching only $Z_{mis}\dot{\beta}$ varies stochastically and does not cancel out any more. Type 3 creates two draws for β, one for the donor set and one for the recipient set.

4.2.2 Nearest Neighbor Multiple Imputation

The multiple imputation that Rubin[8] proposes is a parametric method using the posterior distribution of the variable with missing data to create the imputing set. Here, we present a non parametric multiple imputation method using two generalized linear models one for the missing probability and one for the missing values of the missing variable and a metric in order to create the imputing sets. The two models will be fitted on a non-parametric bootstrap sample of the original dataset in order to embody the uncertainty of parameter estimates from the working models. This results in proper multiple imputation[8][11].

1. **Estimation of the predictive scores on a bootstrap sample.**

   Instead of using a parametric distribution we use all the variables we have available Y, δ_t, Z including the variable with the missing data X to create a bootstrap sample[9]. We continue by estimating the baseline hazard function $H_0(t)$ using the Nelson-Aalen estimator[10] on the bootstrap sample. The Nelson-Aalen estimator is a non-parametric estimator of the cumulative hazard rate function in case of censored or incomplete data and is qiven by $\hat{H_0}(t) = \sum_{t_i \leq t} \frac{a_i}{n_i}$, with a_i the number of events at t_i and n_i the total individuals at risk at t_i.

   Then we fit a logistic regression model with variables Y, δ_t and Z as the variables to the missing indicator δ_x to derive a predictive score for missingness. This score shows the relationship between the missingness and Y, δ_t and Z. We standardize the fitted values by subtracting the mean and by dividing with the standard deviation and denote the standardized score by $S_{\delta_x}^{c(B)}$.

   We continue by fitting generalized linear model with $H_0(t), \delta_t$ and Z as the variables to the variables with missing data X. This score shows the relationship between X and $H_0(t), \delta_t, Z$. We also standardize the fitted values by subtracting the mean and by dividing with the standard deviation and denote the standardized score by $S_x^{c(B)}$.

2. **Using the Euclidean metric to define the imputing set.**

   For each missing subject in the original dataset, two predictive standardized scores are derived from the two regression models obtained from the bootstrap sample. Implementing the Euclidean metric, the distance between subject j in the original dataset and subject k in the bootstrap sample is then defined as:

   $$d(j,k) = \sqrt{w_1[S_x^c(j) - S_x^{c(B)}(k)]^2 + w_2[S_{\delta_x}^c(j) - S_{\delta_x}^{c(B)}(k)]^2},$$

   where w_1 and w_2 are non negative weights that sum to one. The set is consisted of by subjects who have their X observed and have a small distance from subject j in terms of the metric d considering the two predictive scores.

3. **Random draw from the imputing set.**

   When the imputing set is created, a value for the variable with missing data X is randomly drawn from the imputing set. Meaning that NNMI(NN, w_1, w_2) method imputes values to X only from the subjects with X observed.

4. **Repeat Steps 1 to 3 independently M times.**

   The imputation will be complete after repeating the above steps M times. Each time an imputed is derived from a different bootstrap sample. Once the M imputed

datasets are obtained, the MI analysis established by Rubin[8] is implemented. On our study, a Cox regression analysis with X and Z as the variables is conducted on the M imputed datasets to estimate the Cox regression coefficients β_x and β_z. For both β_x and β_z the estimate is the average of the M corresponding regression coefficients denoted as $\hat{\beta}_x$, $\hat{\beta}_z$ and the final variance denoted as $var(\hat{\beta})$ is the sum of the sample variances of the M regression coefficient estimates and the average denoted as U_β of the M variance estimates of $\hat{\beta}$. The quantity $[\hat{\beta} - \beta]/\sqrt{var(\hat{\beta})}$ approximately follows a t-distribution with degrees of freedom $\nu = (M-1)\Big[1 + \Big\{\frac{U_\beta M}{M+1}\Big\}/B_\beta\Big]^2$. We use a value of 10 or higher for M.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Chapter 5

Use of Methods

5.1 Real Data

We demonstrate the above methods on dataset that contained 184 heart transplant cases in 1980 with 27 cases missing (15% missing rate). The dataset is extracted from the Stanford Heart Transplant data. The survival time is measured from the date of transplant in days with the censoring status. The additional variables collected are the age, which is the patient age at first transplant measured in years and a mismatch score variable which is subject to missing.

Table 5.1: Description of the Heart Transplant Data

Variable	Mean	Standard Deviation	Missing
Age	41.092	11.035	0
Mismatch Score	1.116	0.577	27

We will make use the previous methods on the Heart Transplant dataset and compare the results for the CC, PMM and NNMI methods. We examine how close the results are for each method. The missing rate is 15% so we expect all three methods to perform fairly well.

All three methods have a non significant p-value for the mismatch score variable and a highly significant p-value for the age variable. All three methods produce very similar results, but PMM and NNMI offer tighter confidence intervals for the variables compared always to complete case analysis.

Table 5.2: Cox regression estimation

Variable	Estimation	95% Confidence Interval	p-value
	Complete Case Analysis		
Age	0.170	(-0.188, 0.529)	0.352
Mismatch Score	0.029	(0.007, 0.051)	0.009
	Predictive Mean Matching		
Age	0.169	(-0.188, 0.527)	0.223
Mismatch Score	0.029	(0.008, 0.049)	0.006
	Nearest Neighbor Multiple Imputation		
Age	0.169	(-0.188, 0.526)	0.225
Mismatch Score	0.028	(0.008, 0.049)	0.006

5.2 Simulated Data

In this section we will perform simulation studies to compare CC, PMM and NNMI methods when performing Cox regression with two independent variables which one is subject to missing and the other is a fully observed variable predictive to the missing one. We investigate the performance of each method on different situations such as sample size, mis-specification of one of the working models in NNMI and increasing missing rate from 10% up to 65% under a situation of dependent censoring. We used R to write the simulation program implementing the following R libraries: The libraries survminer, survival for the Cox regression, the library mice for the predictive mean matching method (PMM) using the functions **mice** and **complete** and the library NNMIS for the nearest neighbor multiple imputation method (NNMI) using the functions **NNMIS** and **coxph.pool** . For the PMM method, the function **mice** will detect which variables in the dataset have missing values. We then choose the multiple imputation by chained equations method we want and the number of imputations. Then, using the function **complete** our missing cases are filled with imputed values presenting us a complete dataset. For the NNMI method, the **NNMIS** function performs the algorithm we mentioned in section 6 and the function **coxph.pool** estimates Cox regression model, taking into account the additional uncertainty that arises due to a finite number of imputations of the missing data.

The variables for each of the 1000 independent datasets are generated from the following distributions:Z, which is the predictive variable to the missing variable, is generated from a $U(0,1)$ distribution. The variable X subject to missing is generated from a $Bernoulli[p(Z)]$ distribution, where $p(Z)$ is based on a logit link $p(Z) = \frac{1}{1+e^{\alpha_0+\alpha_z Z}}$. The failure times T, C are generated either from an exponential distribution with hazard rate $e^{\beta_x X+\beta_z Z}$ or a Weibull distribution with a hazard rate of $(e^{\beta_x X+\beta_z Z})\tau t^{\tau-1}$. We define Y to be the minimum between the two failure times, $Y = min(T, C)$ and censoring indicator to be $\delta_t = I(T \leq C)$. The missing indicator $\delta_x(\delta_x = 1$ if X is observed is generated from a $Bernoulli[p(Z, Y)]$ distribution , where $p(Z, Y)$ is based on a logit link $p(Z, Y) = \frac{1}{1+e^{\eta_0+\eta_z Z+\eta_y Y}}$. The coefficients are selected to give the desired censoring rate and missing rate.

The fully-observed analysis (FO) will be used before any missingness is applied and we obtain the Cox regression coefficients for each simulated dataset . The FO analysis is implemented for us to have it as a comparison measure for the other methods. In CC we obtain the Cox regression coefficients after missingness has been applied on the simulated datasets.

In PMM method we estimate $\dot{\beta}$ and $\hat{\beta}$ by a bootstrap sample under the normal linear model, then we calculate $\dot{\eta}(i, j) = |Z_i^{obs}\hat{\beta} - Z_j^{mis}\dot{\beta}|$ with $i = 1, ..., n_1$ and $j = 1, ..., n_0$. We construct n_0 sets A_j, each containing $d = 5$ candidate donors, from X_{obs} such that $\sum_d \dot{\eta}(i, j)$ is minimum for all $j = 1, ..., n_0$ (ties are broken randomly). We draw one donor i_j from A_j randomly for $j = 1, ..., n_0$. Afterwards, we calculate the imputations $\dot{X}_j = X_i$, for $j = 1, ..., n_0$. We use type 1 stochastic matching distance for the imputations.

For the NNMI method, a logistic regression model will be fitted to X, with $Z, \delta_t, \hat{H}_0(t)$ as the variables, to derive the conditional distribution of X given the observed data and the predictive score of X. Another logistic regression model will be fitted to the missing indicator δ_x, with Z, Y as the variables, to derive the missing probability and the predictive score of δ_x. For NNMI, we will investigate how the method performs when the logistic regression model for the missing probability is misspecified. When there is no misspecification the method is denoted as $NNMI_1$. When the variable Y is missing from

the model, the method is denoted as $NNMI_2$ and is denoted as $NNMI_3$ when the variable Z is missing from the model. We set the number of imputations $M = 10$, the nearest neighbor imputation sets to contain five imputed values $NN = 5$ and the weights w_1, w_2 to be either (0.8,0.2) or (0.2,0.8).

The measures we examine in this simulation are:root mean square mean error (RMSE), coverage rate (CR), standard errors (SD,SE) and the estimates (EST) from Cox regression model. Our coefficient estimates are the average of 1000 point estimates. The (SD) is the empirical standard error produced by computing the standard error of the 1000 previous estimates and the (SE) is the average of the estimated standard errors produced from the Cox regression models. The RMSE is defined as the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results. In our case $RMSE = \sqrt{bias^2 + SD^2}$. The CR is the proportion of the 1000 samples for which the known population parameter is contained in the confidence interval. That proportion is an estimate for the empirical coverage probability for the CI.

All the tables of the simulation study can be found in the appendices from 3 to 10. The FO analysis without missing values has the lowest root mean square error (RMSE) and produces coverage rates near the nominal level 95%. The CC analysis as expected produces biased cox regression coefficient estimates with large RMSE for the coefficient β_z and the coverage rate is lower most of the times than NNMI and PMM. PMM outperforms NNMI on the situations when the sample size is small, but NNMI performs excellently on all situations when the sample is large, producing accurate estimates with small RMSE and coverage rates close to nominal level. NNMI produces more accurate estimates of the coefficients when the weights are $w_1 = 0.2, w_2 = 0.8$. As the missing rate increases, the coefficients from CC analysis produce more bias as expected (Table 6). The PMM, as the missing rate increases, does not produce coverage rates close to the nominal level especially for the coefficient β_x. NNMI on the other hand produces coverage rates close to nominal level and is more consistent in all missing rates except for the situation with missing rate at 45% (Table 5), where the produced standard errors when $N = 100$ are too large. When the working logistic regression model is misspecified, without the variable Z, it produces better estimates for both coefficients compared to the other form of misspecification.

When the failure and censoring times are generated from Weibull distributions (Tables 7-10), all methods produce similar results to those generated with exponential failure and censoring times. This is reasonable, because PMM and NNMI do not need to specify the underlying distribution of failure and censoring times while performing estimation.

PMM and NNMI manage variables with missing data with high missing rate better than standard CC analysis. They produce more accurate coefficient estimates with reasonable standard errors. Also, NNMI proved to be quite robust to misspecification when the working logistic regression model for missing probability was misspecified either with Z or Y missing.

Chapter 6

Conclusion

In this study we investigated the performance of complete case analysis and two multiple imputation methods on various missing rates. The simulation results showed that CC analysis breaks down at about 30% missing rate, PMM breaks down at about 45% and NNMI seems to be the most consistent method in all missing rates. On lower missing rates and when having small sample situations, PMM estimates were better than those of NNMI. NNMI estimates were precise enough even when the working logistic regression model for missing probability was misspecified. Furthermore, both multiple imputation methods had no problem estimating values either from Exponential or from Weibull failure and censoring times, performing the same.

The sample depends on how high the missing rate will be. The sample affects the number of donors for the multiple imputation methods for each missing variable observation. Increasing missing rate means an increase also on the size of the sample. Nevertheless, we need to see how the methods perform even in small sample sizes in order for us to evaluate the performance of PMM and NNMI.

The missingness mechanism assumed in this study depends on the observed data (MAR mechanism). In some cases the missingness may very well depend on some unobserved data (MNAR mechanism). A possible way to figure the impact of the unobserved data is by performing sensitivity analysis. The results will indicate which mechanism should be followed in the study. Violation of the MAR assumption in this study may not affect the results because of the nature of the multiple imputation methods being semi-parametric (PMM) and non-parametric (NNMI).

Although, can be quite challenging and tough, a study can be conducted comparing complete case analysis (CC), predictive mean matching PMM and nearest neighbor multiple imputation (NNMI) to show if the same results stand when more than one variable is subject to missing under the MAR mechanism.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Appendix A

Matrices with Exponential Times

Table A.1: $Z \sim Uniform[0,1], T \sim Exponential[e^{ln(2)X-1.2Z}], C \sim Exponential[e^{-3X+0.1Z}], X \sim Bernoulli[p(Z) = \frac{1}{1+e^{0.25-0.5Z}}], \delta_x = Bernoulli[p(Z,Y) = \frac{1}{1+e^{ln(5.3)+3.5Z-0.8Y}}]$

Method	EST[a]	SD[b]	SE[c]	RMSE[d]	CR[e]	EST	SD	SE	RMSE	CR
N=100										
FO	0.701	0.309	0.294	0.309	0.939	-1.256	0.492	0.471	0.496	0.944
CC	0.735	0.324	0.310	0.327	0.952	-1.340	0.526	0.504	0.544	0.943
$NNMI_1(0.8, 0.2)$	0.815	0.322	0.316	0.344	0.900	-1.182	0.497	0.485	0.497	0.896
$NNMI_2(0.8, 0.2)$	0.814	0.315	0.314	0.338	0.917	-1.184	0.498	0.485	0.499	0.895
$NNMI_3(0.8, 0.2)$	0.803	0.326	0.311	0.344	0.915	-1.168	0.505	0.479	0.506	0.877
$NNMI_1(0.2, 0.8)$	0.817	0.322	0.316	0.345	0.933	-1.182	0.502	0.485	0.502	0.896
$NNMI_2(0.2, 0.8)$	0.820	0.300	0.311	0.326	0.932	-1.189	0.505	0.482	0.505	0.895
$NNMI_3(0.2, 0.8)$	0.793	0.325	0.314	0.341	0.933	-1.158	0.503	0.480	0.505	0.879
PMM	0.730	0.322	0.295	0.324	0.925	-1.206	0.473	0.470	0.473	0.948
N=500										
FO	0.694	0.133	0.128	0.133	0.945	-1.198	0.197	0.202	0.197	0.957
CC	0.715	0.139	0.134	0.141	0.940	-1.271	0.211	0.215	0.223	0.941
$NNMI_1(0.8, 0.2)$	0.686	0.142	0.135	0.142	0.942	-1.193	0.184	0.206	0.184	0.974
$NNMI_2(0.8, 0.2)$	0.686	0.141	0.134	0.141	0.941	-1.196	0.183	0.206	0.183	0.985
$NNMI_3(0.8, 0.2)$	0.674	0.140	0.134	0.141	0.941	-1.181	0.183	0.204	0.184	0.972
$NNMI_1(0.2, 0.8)$	0.689	0.141	0.138	0.141	0.962	-1.187	0.184	0.206	0.185	0.973
$NNMI_2(0.2, 0.8)$	0.692	0.138	0.135	0.139	0.942	-1.196	0.183	0.205	0.183	0.979
$NNMI_3(0.2, 0.8)$	0.660	0.136	0.134	0.140	0.949	-1.169	0.182	0.203	0.185	0.978
PMM	0.695	0.145	0.128	0.145	0.907	-1.185	0.207	0.202	0.208	0.947

Note:Censoring rate:0.35; Missing rate:0.10.
[a] Average of 1000 point estimates.
[b] Empirical standard deviation.
[c] Average estimated standard error.
[d] Root mean square error:square root of bias$+SD^2$.
[e] Coverage rate of 1000 95% confidence intervals.

Table A.2: $Z \sim Uniform[0,1], T \sim Exponential[e^{ln(2)X-1.2Z}], C \sim Exponential[e^{-3X+0.1Z}], X \sim Bernoulli[p(Z) = \frac{1}{1+e^{0.25-0.5Z}}], \delta_x = Bernoulli[p(Z,Y) = \frac{1}{1+e^{ln(5.3)+2Z-2Y}}]$

Method	EST[a]	SD[b]	SE[c]	RMSE[d]	CR[e]	EST	SD	SE	RMSE	CR
N=100										
FO	0.701	0.309	0.294	0.309	0.939	-1.256	0.492	0.471	0.496	0.944
CC	0.836	0.362	0.348	0.389	0.941	-1.155	0.613	0.571	0.615	0.940
$NNMI_1(0.8, 0.2)$	0.737	0.398	0.389	0.401	0.936	-1.155	0.491	0.503	0.493	0.946
$NNMI_2(0.8, 0.2)$	0.736	0.380	0.376	0.383	0.967	-1.172	0.503	0.507	0.503	0.947
$NNMI_3(0.8, 0.2)$	0.722	0.399	0.384	0.400	0.944	-1.151	0.495	0.501	0.498	0.946
$NNMI_1(0.2, 0.8)$	0.711	0.402	0.408	0.402	0.940	-1.132	0.484	0.501	0.489	0.947
$NNMI_2(0.2, 0.8)$	0.743	0.351	0.362	0.354	0.980	-1.173	0.495	0.505	0.496	0.951
$NNMI_3(0.2, 0.8)$	0.687	0.416	0.397	0.416	0.946	-1.116	0.489	0.499	0.496	0.946
PMM	0.777	0.441	0.298	0.449	0.820	-1.151	0.501	0.469	0.504	0.931
N=500										
FO	0.694	0.133	0.128	0.133	0.945	-1.198	0.197	0.202	0.197	0.957
CC	0.824	0.149	0.148	0.198	0.862	-1.124	0.244	0.242	0.256	0.938
$NNMI_1(0.8, 0.2)$	0.708	0.173	0.174	0.174	0.959	-1.151	0.210	0.216	0.216	0.937
$NNMI_2(0.8, 0.2)$	0.710	0.154	0.159	0.155	0.955	-1.176	0.214	0.218	0.215	0.944
$NNMI_3(0.8, 0.2)$	0.689	0.162	0.167	0.162	0.964	-1.142	0.211	0.213	0.218	0.928
$NNMI_1(0.2, 0.8)$	0.685	0.192	0.192	0.192	0.954	-1.121	0.205	0.212	0.220	0.921
$NNMI_2(0.2, 0.8)$	0.723	0.148	0.154	0.151	0.953	-1.178	0.211	0.216	0.212	0.943
$NNMI_3(0.2, 0.8)$	0.651	0.179	0.179	0.183	0.939	-1.115	0.206	0.209	0.222	0.919
PMM	0.697	0.172	0.127	0.172	0.866	-1.148	0.215	0.201	0.221	0.925

Note:Censoring rate:0.35; Missing rate:0.30.
[a] Average of 1000 point estimates.
[b] Empirical standard deviation.
[c] Average estimated standard error.
[d] Root mean square error:square root of bias$+SD^2$.
[e] Coverage rate of 1000 95% confidence intervals.

Table A.3: $Z \sim Uniform[0,1], T \sim Exponential[e^{ln(2)X-1.2Z}], C \sim Exponential[e^{-3X+0.1Z}], X \sim Bernoulli[p(Z) = \frac{1}{1+e^{0.25-0.5Z}}], \delta_x = Bernoulli[p(Z,Y) = \frac{1}{1+e^{1.3+0.5Z-2Y}}]$

Method	EST[a]	SD[b]	SE[c]	RMSE[d]	CR[e]	EST	SD	SE	RMSE	CR
N=100										
FO	0.701	0.309	0.294	0.309	0.939	-1.256	0.492	0.471	0.496	0.944
CC	0.862	0.425	0.402	0.458	0.936	-0.873	0.716	0.664	0.787	0.901
$NNMI_1(0.8, 0.2)$	0.825	0.495	0.458	0.512	0.957	-1.223	0.515	0.534	0.515	0.976
$NNMI_2(0.8, 0.2)$	0.828	0.498	0.461	0.516	0.957	-1.239	0.534	0.544	0.536	0.975
$NNMI_3(0.8, 0.2)$	0.820	0.506	0.457	0.521	0.956	-1.221	0.514	0.537	0.515	0.995
$NNMI_1(0.2, 0.8)$	0.761	0.490	0.472	0.495	0.998	-1.170	0.495	0.525	0.495	0.975
$NNMI_2(0.2, 0.8)$	0.809	0.431	0.420	0.446	0.979	-1.230	0.500	0.545	0.500	0.997
$NNMI_3(0.2, 0.8)$	0.726	0.500	0.477	0.501	0.977	-1.155	0.490	0.521	0.492	0.975
PMM	0.829	0.541	0.300	0.558	0.707	-1.134	0.518	0.470	0.523	0.927
N=500										
FO	0.694	0.133	0.128	0.133	0.945	-1.198	0.197	0.202	0.197	0.957
CC	0.846	0.168	0.169	0.228	0.866	-0.844	0.289	0.276	0.458	0.719
$NNMI_1(0.8, 0.2)$	0.750	0.203	0.204	0.210	0.955	-1.132	0.218	0.224	0.228	0.943
$NNMI_2(0.8, 0.2)$	0.755	0.184	0.182	0.194	0.941	-1.161	0.220	0.229	0.223	0.957
$NNMI_3(0.8, 0.2)$	0.742	0.198	0.200	0.204	0.959	-1.128	0.214	0.223	0.226	0.948
$NNMI_1(0.2, 0.8)$	0.702	0.223	0.225	0.223	0.980	-1.091	0.207	0.217	0.233	0.924
$NNMI_2(0.2, 0.8)$	0.760	0.172	0.175	0.185	0.950	-1.152	0.211	0.223	0.216	0.961
$NNMI_3(0.2, 0.8)$	0.686	0.216	0.217	0.216	0.962	-1.091	0.207	0.216	0.234	0.915
PMM	0.731	0.205	0.127	0.209	0.788	-1.129	0.223	0.201	0.234	0.901

Note:Censoring rate:0.35; Missing rate:0.45.
[a] Average of 1000 point estimates.
[b] Empirical standard deviation.
[c] Average estimated standard error.
[d] Root mean square error:square root of bias+SD^2.
[e] Coverage rate of 1000 95% confidence intervals.

Table A.4: $Z \sim Uniform[0,1], T \sim Exponential[e^{ln(2)X-1.2Z}], C \sim Exponential[e^{-3X+0.1Z}], X \sim Bernoulli[p(Z) = \frac{1}{1+e^{0.25-0.5Z}}], \delta_x = Bernoulli[p(Z,Y) = \frac{1}{1+e^{1.5+1.5Z-2Y}}]$

Method	EST[a]	SD[b]	SE[c]	RMSE[d]	CR[e]	EST	SD	SE	RMSE	CR
N=100										
FO	0.701	0.309	0.294	0.309	0.939	-1.256	0.492	0.471	0.496	0.944
CC	0.748	1.357	1.354	1.358	0.942	-2.065	0.973	0.912	1.302	0.874
$NNMI_1(0.8, 0.2)$	1.075	2.044	4.105	2.079	0.959	-1.207	0.537	0.499	0.537	0.938
$NNMI_2(0.8, 0.2)$	1.120	2.026	4.101	2.070	0.960	-1.238	0.531	0.499	0.532	0.940
$NNMI_3(0.8, 0.2)$	1.045	2.043	4.106	2.073	0.982	-1.195	0.536	0.499	0.536	0.938
$NNMI_1(0.2, 0.8)$	0.764	1.062	2.694	1.065	0.982	-1.179	0.517	0.491	0.517	0.959
$NNMI_2(0.2, 0.8)$	0.960	1.660	3.802	1.681	0.963	-1.237	0.522	0.497	0.523	0.949
$NNMI_3(0.2, 0.8)$	0.793	1.544	3.521	1.547	0.959	-1.159	0.522	0.491	0.523	0.948
PMM	0.965	1.471	2.396	1.496	0.712	-1.226	0.514	0.473	0.515	0.926
N=500										
FO	0.694	0.133	0.128	0.133	0.945	-1.198	0.197	0.202	0.197	0.957
CC	0.639	0.259	0.232	0.264	0.915	-1.819	0.364	0.357	0.718	0.592
$NNMI_1(0.8, 0.2)$	0.708	0.262	0.249	0.263	0.937	-1.191	0.203	0.213	0.203	0.965
$NNMI_2(0.8, 0.2)$	0.728	0.231	0.223	0.234	0.956	-1.202	0.200	0.211	0.200	0.976
$NNMI_3(0.8, 0.2)$	0.707	0.255	0.244	0.256	0.931	-1.190	0.201	0.212	0.202	0.976
$NNMI_1(0.2, 0.8)$	0.694	0.272	0.255	0.272	0.935	-1.178	0.200	0.212	0.201	0.975
$NNMI_2(0.2, 0.8)$	0.766	0.217	0.209	0.229	0.933	-1.214	0.199	0.211	0.200	0.972
$NNMI_3(0.2, 0.8)$	0.685	0.260	0.244	0.260	0.948	-1.177	0.197	0.211	0.198	0.977
PMM	0.752	0.274	0.131	0.280	0.662	-1.214	0.218	0.203	0.218	0.931

Note:Censoring rate:0.35; Missing rate:0.65.
[a] Average of 1000 point estimates.
[b] Empirical standard deviation.
[c] Average estimated standard error.
[d] Root mean square error:square root of bias$+SD^2$.
[e] Coverage rate of 1000 95% confidence intervals.

Appendix B

Matrices with Weibull Times

Table B.1: $Z \sim Uniform[0,1], T \sim Weibull[e^{ln(2)X-1.2Z}, 1.5], C \sim Weibull[e^{-3X+0.1Z}, 1.4], X \sim Bernoulli[p(Z) = \frac{1}{1+e^{0.25-0.5Z}}], \delta_x = Bernoulli[p(Z,Y) = \frac{1}{1+e^{2.8+3Z-2Y}}]$

Method	EST[a]	SD[b]	SE[c]	RMSE[d]	CR[e]	EST	SD	SE	RMSE	CR
N=100										
FO	0.702	0.304	0.293	0.304	0.941	-1.227	0.483	0.463	0.484	0.939
CC	0.849	0.312	0.306	0.349	0.927	-1.237	0.527	0.502	0.528	0.946
$NNMI_1(0.8, 0.2)$	0.766	0.327	0.323	0.336	0.959	-1.222	0.486	0.480	0.486	0.951
$NNMI_2(0.8, 0.2)$	0.768	0.311	0.315	0.320	0.960	-1.236	0.484	0.482	0.485	0.957
$NNMI_3(0.8, 0.2)$	0.739	0.322	0.316	0.326	0.955	-1.205	0.483	0.474	0.483	0.955
$NNMI_1(0.2, 0.8)$	0.756	0.332	0.336	0.338	0.972	-1.196	0.479	0.483	0.479	0.955
$NNMI_2(0.2, 0.8)$	0.778	0.299	0.316	0.311	0.968	-1.232	0.480	0.483	0.481	0.957
$NNMI_3(0.2, 0.8)$	0.726	0.337	0.326	0.339	0.949	-1.178	0.479	0.476	0.479	0.959
PMM	0.748	0.338	0.293	0.342	0.921	-1.207	0.489	0.465	0.490	0.943
N=500										
FO	0.692	0.133	0.128	0.133	0.944	-1.208	0.201	0.200	0.201	0.953
CC	0.828	0.142	0.133	0.196	0.816	-1.217	0.226	0.216	0.226	0.940
$NNMI_1(0.8, 0.2)$	0.699	0.136	0.138	0.136	0.954	-1.169	0.206	0.203	0.208	0.930
$NNMI_2(0.8, 0.2)$	0.711	0.135	0.135	0.136	0.945	-1.190	0.206	0.215	0.206	0.945
$NNMI_3(0.8, 0.2)$	0.686	0.131	0.135	0.131	0.952	-1.161	0.202	0.202	0.209	0.919
$NNMI_1(0.2, 0.8)$	0.690	0.142	0.163	0.142	0.950	-1.149	0.204	0.203	0.210	0.922
$NNMI_2(0.2, 0.8)$	0.726	0.132	0.134	0.136	0.947	-1.197	0.206	0.205	0.206	0.944
$NNMI_3(0.2, 0.8)$	0.672	0.133	0.137	0.134	0.954	-1.149	0.204	0.201	0.210	0.920
PMM	0.706	0.137	0.127	0.137	0.923	-1.200	0.211	0.200	0.211	0.945

Note:Censoring rate:0.35; Missing rate:0.10.
[a]Average of 1000 point estimates.
[b]Empirical standard deviation.
[c]Average estimated standard error.
[d]Root mean square error:square root of bias+SD^2.
[e]Coverage rate of 1000 95% confidence intervals.

Table B.2: $Z \sim Uniform[0,1], T \sim Weibull[e^{ln(2)X-1.2Z}, 1.5], C \sim Weibull[e^{-3X+0.1Z}, 1.4], X \sim Bernoulli[p(Z) = \frac{1}{1+e^{0.25-0.5Z}}], \delta_x = Bernoulli[p(Z,Y) = \frac{1}{1+e^{2+Z-2Y}}]$

Method	EST[a]	SD[b]	SE[c]	RMSE[d]	CR[e]	EST	SD	SE	RMSE	CR
N=100										
FO	0.702	0.304	0.293	0.304	0.941	-1.227	0.483	0.463	0.484	0.939
CC	0.911	0.350	0.340	0.413	0.920	-0.992	0.596	0.563	0.631	0.923
$NNMI_1(0.8, 0.2)$	0.877	0.194	0.365	0.267	0.994	-1.295	0.488	0.482	0.497	0.995
$NNMI_2(0.8, 0.2)$	0.869	0.177	0.327	0.250	0.994	-1.321	0.504	0.485	0.519	0.995
$NNMI_3(0.8, 0.2)$	0.867	0.210	0.328	0.273	0.993	-1.286	0.484	0.475	0.491	0.997
$NNMI_1(0.2, 0.8)$	0.895	0.229	0.373	0.305	0.996	-1.283	0.455	0.484	0.462	0.995
$NNMI_2(0.2, 0.8)$	0.853	0.153	0.321	0.222	0.996	-1.317	0.507	0.485	0.521	0.993
$NNMI_3(0.2, 0.8)$	0.799	0.270	0.336	0.290	0.991	-1.208	0.465	0.470	0.465	0.995
PMM	0.821	0.402	0.293	0.422	0.839	-1.187	0.499	0.462	0.499	0.943
N=500										
FO	0.692	0.133	0.128	0.133	0.944	-1.208	0.201	0.200	0.201	0.953
CC	0.893	0.151	0.146	0.251	0.722	-0.957	0.251	0.239	0.349	0.799
$NNMI_1(0.8, 0.2)$	0.727	0.137	0.155	0.141	0.962	-1.162	0.169	0.207	0.173	0.985
$NNMI_2(0.8, 0.2)$	0.748	0.128	0.145	0.139	0.951	-1.199	0.166	0.212	0.166	0.985
$NNMI_3(0.8, 0.2)$	0.720	0.134	0.151	0.137	0.962	-1.162	0.169	0.206	0.0.173	0.986
$NNMI_1(0.2, 0.8)$	0.704	0.141	0.165	0.141	0.969	-1.136	0.170	0.206	0.181	0.964
$NNMI_2(0.2, 0.8)$	0.770	0.120	0.143	0.143	0.961	-1.197	0.171	0.212	0.171	0.985
$NNMI_3(0.2, 0.8)$	0.703	0.140	0.164	0.141	0.984	-1.142	0.171	0.206	0.180	0.965
PMM	0.731	0.161	0.126	0.166	0.871	-1.192	0.202	0.200	0.202	0.942

Note:Censoring rate:0.35; Missing rate:0.30.
[a] Average of 1000 point estimates.
[b] Empirical standard deviation.
[c] Average estimated standard error.
[d] Root mean square error:square root of bias+SD^2.
[e] Coverage rate of 1000 95% confidence intervals.

Table B.3: $Z \sim Uniform[0,1], T \sim Weibull[e^{ln(2)X-1.2Z}, 1.5], C \sim Weibull[e^{-3X+0.1Z}, 1.4], X \sim Bernoulli[p(Z) = \frac{1}{1+e^{0.25-0.5Z}}], \delta_x = Bernoulli[p(Z,Y) = \frac{1}{1+e^{1.4+0.5Z-2Y}}]$

Method	EST[a]	SD[b]	SE[c]	RMSE[d]	CR[e]	EST	SD	SE	RMSE	CR
N=100										
FO	0.702	0.304	0.293	0.304	0.941	-1.227	0.483	0.463	0.484	0.939
CC	0.922	0.403	0.381	0.464	0.914	-0.893	0.691	0.638	0.756	0.906
$NNMI_1$(0.8, 0.2)	0.862	0.413	0.401	0.446	0.939	-1.183	0.414	0.504	0.414	0.966
$NNMI_2$(0.8, 0.2)	0.891	0.376	0.386	0.425	0.941	-1.233	0.437	0.520	0.438	0.976
$NNMI_3$(0.8, 0.2)	0.850	0.408	0.399	0.438	0.940	-1.176	0.417	0.505	0.418	0.966
$NNMI_1$(0.2, 0.8)	0.853	0.436	0.428	0.464	0.933	-1.140	0.405	0.500	0.409	0.975
$NNMI_2$(0.2, 0.8)	0.859	0.343	0.375	0.381	0.981	-1.204	0.420	0.522	0.420	0.977
$NNMI_3$(0.2, 0.8)	0.849	0.434	0.428	0.461	0.953	-1.131	0.408	0.499	0.413	0.974
PMM	0.884	0.478	0.296	0.515	0.778	-1.195	0.520	0.464	0.520	0.923
N=500										
FO	0.692	0.133	0.128	0.133	0.944	-1.208	0.201	0.200	0.201	0.953
CC	0.916	0.169	0.162	0.280	0.717	-0.846	0.269	0.267	0.444	0.715
$NNMI_1$(0.8, 0.2)	0.758	0.160	0.179	0.172	0.990	-1.177	0.211	0.213	0.212	0.970
$NNMI_2$(0.8, 0.2)	0.776	0.141	0.154	0.163	0.950	-1.213	0.206	0.220	0.206	0.970
$NNMI_3$(0.8, 0.2)	0.748	0.159	0.172	0.168	0.970	-1.177	0.212	0.212	0.0.213	0.951
$NNMI_1$(0.2, 0.8)	0.734	0.175	0.198	0.180	0.998	-1.139	0.204	0.210	0.212	0.949
$NNMI_2$(0.2, 0.8)	0.790	0.127	0.152	0.160	0.942	-1.209	0.208	0.218	0.209	0.970
$NNMI_3$(0.2, 0.8)	0.726	0.173	0.191	0.177	0.988	-1.133	0.206	0.210	0.214	0.971
PMM	0.753	0.118	0.126	0.188	0.816	-1.176	0.213	0.200	0.215	0.922

Note:Censoring rate:0.35; Missing rate:0.45.
[a] Average of 1000 point estimates.
[b] Empirical standard deviation.
[c] Average estimated standard error.
[d] Root mean square error:square root of bias+SD^2.
[e] Coverage rate of 1000 95% confidence intervals.

Table B.4: $Z \sim Uniform[0,1], T \sim Weibull[e^{ln(2)X-1.2Z}, 1.5], C \sim Weibull[e^{-3X+0.1Z}, 1.4], X \sim Bernoulli[p(Z) = \frac{1}{1+e^{0.25-0.5Z}}], \delta_x = Bernoulli[p(Z,Y) = \frac{1}{1+e^{2+0.5Z-2Y}}]$

Method	EST[a]	SD[b]	SE[c]	RMSE[d]	CR[e]	EST	SD	SE	RMSE	CR
N=100										
FO	0.702	0.304	0.293	0.304	0.941	-1.227	0.483	0.463	0.484	0.939
CC	0.800	2.107	9.868	2.110	0.932	-1.714	0.992	0.897	1.118	0.923
$NNMI_1(0.8, 0.2)$	0.691	0.516	0.689	0.516	0.991	-1.235	0.438	0.486	0.440	0.997
$NNMI_2(0.8, 0.2)$	0.794	0.513	0.585	0.523	0.946	-1.250	0.439	0.479	0.442	0.998
$NNMI_3(0.8, 0.2)$	0.695	0.500	0.680	0.500	0.991	-1.234	0.436	0.490	0.437	0.997
$NNMI_1(0.2, 0.8)$	0.578	0.49	0.655	0.505	0.946	-1.209	0.443	0.487	0.443	0.997
$NNMI_2(0.2, 0.8)$	0.700	0.429	0.524	0.429	0.952	-1.251	0.423	0.583	0.426	0.998
$NNMI_3(0.2, 0.8)$	0.588	0.490	0.656	0.501	0.994	-1.206	0.443	0.481	0.443	0.996
PMM	0.846	0.420	0.293	0.447	0.821	-1.166	0.483	0.463	0.484	0.938
N=500										
FO	0.692	0.133	0.128	0.133	0.944	-1.208	0.201	0.200	0.201	0.953
CC	0.544	0.263	0.251	0.302	0.885	-1.577	0.373	0.358	0.530	0.817
$NNMI_1(0.8, 0.2)$	0.706	0.262	0.268	0.262	0.946	-1.194	0.199	0.210	0.199	0.967
$NNMI_2(0.8, 0.2)$	0.740	0.226	0.239	0.231	0.967	-1.208	0.202	0.210	0.202	0.962
$NNMI_3(0.8, 0.2)$	0.704	0.256	0.266	0.256	0.945	-1.193	0.199	0.209	0.0.199	0.969
$NNMI_1(0.2, 0.8)$	0.701	0.270	0.272	0.270	0.949	-1.182	0.197	0.208	0.198	0.962
$NNMI_2(0.2, 0.8)$	0.799	0.218	0.229	0.242	0.946	-1.222	0.203	0.209	0.205	0.962
$NNMI_3(0.2, 0.8)$	0.696	0.261	0.271	0.261	0.948	-1.181	0.195	0.208	0.196	0.968
PMM	0.743	0.162	0.126	0.169	0.851	-1.170	0.214	0.199	0.216	0.932

Note:Censoring rate:0.35; Missing rate:0.65.
[a] Average of 1000 point estimates.
[b] Empirical standard deviation.
[c] Average estimated standard error.
[d] Root mean square error:square root of bias$+SD^2$.
[e] Coverage rate of 1000 95% confidence intervals.

References

LA Escobar and WQ Meeker Jr (1992), Assessing influence in regression analysis with censored data. Biometrics 48, 507–528. Page 519.
Heitjan and Little 1991; Schenker and Taylor 1996. "Partially parametric techniques for multiple imputation" Computational Statistics & Data Analysis
Andridge, R. R., and R. J. A. Little. 2010. "A Review of Hot Deck Imputation for Survey Non-Response." International Statistical Review 78 (1): 40–64.
1986. "Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations." Journal of Business Economics and Statistics 4 (1): 87–94.
Little, R. J. A. 1988. "Missing-Data Adjustments in Large Surveys (with Discussion)." Journal of Business Economics and Statistics 6 (3): 287–301.
Schenker, N., and J. M. G. Taylor. 1996. "Partially Parametric Techniques for Multiple Imputation." Computational Statistics and Data Analysis 22 (4): 425–46.
Siddique, J., and T. R. Belin. 2008. "Multiple Imputation Using an Iterative Hot-Deck with Distance-Based Donor Selection." Statistics in Medicine 27 (1): 83–102.
Donald B. Rubin. "Multiple Imputation for Nonresponse in Surveys" JOHN WILEY & SONS
B. Efron. "Bootstrap Methods: Another Look at the Jackknife" The Annals of Statistics Vol. 7, No. 1 (Jan., 1979), pp. 1-26
Ian R. White, Patrick Royston. "Imputing missing covariate values for the Cox model" Statististics in Medicine. 2009; 28:1982–1998
Søren Feodor Nielsen. "Proper and improper multiple imputation" Department of Statistics and Operations Research
Efron, B., and R. J. Tibshirani. 1993. An Introduction to the Bootstrap. London: Chapman & Hall.
Erler, N. S., D. Rizopoulos, J. Van Rosmalen, V. W. V. Jaddoe, O. H. Franco, and E. M. Lesaffre. 2016. "Dealing with Missing Covariates in Epidemiologic Studies: A Comparison Between Multiple Imputation and a Full Bayesian Approach." Statistics in Medicine 35 (17): 2955–74.
Chiu-Hsieh Hsu, Mandi Yu "Cox regression analysis with missing covariates via nonparametric multiple imputation". Statistical Methods in Medical Research 2019, Vol. 28(6) 1676–1688
Ofer Harel, Xiao-Hua Zhou "Multiple imputation: Review of theory, implementation and software" STATISTICS IN MEDICINE Statist. Med. 2007; 26:3057–3077 Published online 29 January 2007 in Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/sim.2787
Lihong Qi, Ying-FangWangb and Yulei Hec "A comparison of multiple imputation and fully augmented weighted estimators for Cox regression with missing covariates". Statistics in Medicine (wileyonlinelibrary.com) DOI: 10.1002/sim.4016
Chiu-Hsieh Hsu, Qi Long, Yisheng Li and Elizabeth Jacobs "A Nonparametric Mul-

tiple Imputation Approach for Data with Missing Covariate Values with Application to Colorectal Adenoma Data", Journal of Biopharmaceutical Statistics, 24:3, 634-648, DOI: 10.1080/10543406.2014.888444
Donald B. Rubin (1996) Multiple Imputation after 18+ Years, Journal of the American Statistical Association, 91:434, 473-489
Efron, B. "Bootstrap Methods: Another Look at the Jackknife." The Annals of Statistics, vol. 7, no. 1, 1979, pp. 1–26. JSTOR, www.jstor.org/stable/2958830. Accessed 10 Dec. 2020
David W. Hosmer Jr., Stanley Lemeshow, Susanne May "Applied Survival Analysis: Regression Modeling of Time-to-Event Data, 2nd Edition" ‚JOHN WILEY & SONS, ISBN: 978-0-471-75499-2
Ralf Bender, Thomas Augustin, Maria Blettner "Generating survival times to simulate Cox proportional hazards models", Statistics in Medicine Volume 24, Issue 11 1713-1723
James M. Robins, Andrea Rotnitzky and Lue Ping Zhao "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed", Journal of the American Statistical Association Vol. 89, No. 427 (Sep., 1994), pp. 846-866 (21 pages)
R library package "NNMIS" https://cran.r-project.org/web/packages/NNMIS/NNMIS.pdf
R library package "mice" https://cran.r-project.org/web/packages/mice/mice.pdf


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ