QATAR UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

PROPORTIONAL HAZARD REGRESSION MODEL UNDER PARTLY INTERVAL-CENSORING ASSUMPTION WITH APPLICATION TO PRISON DATA

BY

SHAIKHA AHMED ABDULLA

A Thesis Submitted to

the College of Arts and Sciences

in Partial Fulfillment of the Requirements for the Degree of

Masters of Science in Applied Statistics

June 2020

COMMITTEE PAGE

The members of the Committee approve the Thesis of Shaikha Ahmed Abdulla defended on 17/05/2020.

Dr. Faiz Ahmed Elfaki
Thesis/Dissertation Supervisor

Dr. Mohamed Chaouch
Co- Supervisor

Approved:

Ibrahim AlKaabi, Dean, College of Arts and Sciences

ABSTRACT

ABDULLA, SHAIKHA AHMED, Masters : June : 2020, Applied Statistics

Title: Proportional Hazard Regression Proportional Hazard Regression Model Under Partly Interval-Censoring Assumption with Application to Prison Data

Supervisor of Thesis : Faiz Ahmed Mohamed Elfaki

In this thesis the analysis of well-known model in survival study that is Cox proportional hazard regression model via prison Partly Interval Censored (PIC) data is used. The maximum likelihood estimate was considered to obtain the estimated of the model parameter and the survival function and then the results were compared. In this model several imputation techniques are used that is; left point, mean and median. In contrast, the data needed to be modified to PIC data for the proposed of the researcher's needs. Likewise, simulation data was generated where the failure rates were taken based on prison PIC data was also used to further compare these three imputation methods of estimation.

From the prison data set and simulation study for this particular case, we can conclude that the Cox model proved to be feasible and works well in terms of estimation the survival function, likelihood ratio test and their P-value. In additional to that, based on imputation techniques, the mean and median showed better results with respect to estimate of the survival function.

DEDICATION

I dedicate my thesis to my family. A special feeling of gratitude to my loving parents, who have supported me throughout the process.

It is also dedicated to my doctor, who taught me that even the largest task can be accomplished if it is done one step at a time.

ACKNOWLEDGMENTS

First and foremost, I must acknowledge my limitless thanks to Allah, the Ever-Magnificent; the Ever-Thankful, for His help and bless. I am grateful to some people, who worked hard with me from the beginning till the completion of the present thesis particularly my supervisor Dr. Faiz, who has been always generous during all phases of the thesis. Also to my co-supervisor, Dr. Mohamed Chaouch for helping me to improve the work to a higher level. My special thanks are extended to the Management of penal and correctional institutions in Qatar for their help in collecting the data. I also would like to express my wholehearted thanks to my family for their generous support they provided me throughout my entire life and particularly through the process of pursuing the master's degree. Last but not least, deepest thanks go to all people who took part in making this thesis real.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

This chapter introduces the general framework of the survival data analysis, and discusses different censoring schemes. The statistical problem as well as the objectives of this thesis are also introduced in this chapter.

1.1 Background

Statistical problems emerge when examining the occurrence of events and the occurrence of time in a population. An event in this context involves the qualitative transformation of the observed person occurring at a specific period **(**Emmert-Streib & Dehmer, 2019). In the health care settings, the event can be the time until death or cure of a person, computed from a specific treatment or disease onset. Statistical analysis in these contexts entails survival analysis that is used when examiners are interested in the time until an event occurs (Emmert-Streib & Dehmer, 2019). Survival analysis includes different analysis techniques for examining data with time as the outcome variable. Time in these instances corresponds to the period until a specific event occurs. Examples of events include heart attack, death, product wear out, or parole violation, etc.

Based on these different examples, it is apparent that different fields, such as behavioral sciences, social sciences, marketing, engineering, medicine, and biology use survival analysis (Zhang et al., 2013). For example, the sample size in a clinical trial can be diagnosed by survival analysis when the test requires the comparison of the mean or a specific percentile concerning the survival distribution (Emmert-Streib & Dehmer, 2019). The basis of the approach to this test is the accelerated failure time model, which can be applied directly to design reliability studies when comparing the reliability of differentially manufactured products. Survival analysis can also be used to examine the failure of mechanical devices and among couples treated for fertility issues to model

the time to pregnancy. Another instance of the application of survival analysis is in engineering in which it can be used to test the durability of electrical or mechanical components where the researcher uses the method to track items and the life span of material to predict the reliability of the product (Fauzi et al., 2015). These examples show that survival analysis explores and simulates the changes in survival probability at the time of the event. Estimation is based on data of participants offering information about the event time. The exact starting point and ending point are not required since observations do not always begin at zero as a participant can enter into the study at any time. Time is relative where all participants are placed to a common initial point in which the time is zero and all participants have survival probabilities equal to one (Emmert-Streib & Dehmer, 2019).

The uniqueness of survival data concerns that not all participant's experience the event (like a heart attack) towards the end of the observation time, which means that for some patients their real survival times will be unknown. In turn, this creates the censoring phenomenon, which must be considered during the analysis to ensure valid inferences. Censoring is a factor that complicates the estimation of survival analysis as it causes incomplete information. Censorship, nevertheless, allows the examiner to compute lifetimes for participants who have not been subjected to an experiment. Notably, the participants who did not experience the targeted event must be part of the investigation because eliminating biases influence the outcomes of every participant experiencing the targeted event. They must be included in which they can be separated from those who experienced the targeted event through a variable indicating censorship.

1.2 Censoring Mechanisms

Survival analysis uses different censoring techniques. It is crucial to note that censoring is independent of the future importance of the threat for a specific participant (Schober & Vetter, 2018).

Right censoring occurs when the participant enters at the beginning of the examination and terminates before the targeted event happens. The participants may not experience the event by staying longer than the examination period or may not have been part of the examination in which they leave early without experiencing the event (Rabe-Hesketh & Skrondal, 2012).

Left censoring occurs when the analyst fails to observe the birth event. It is vital to mention the idea of length-biased sampling that happens when the study objective is to analyze the participants who experienced the event already to examine whether they will undergo the event again. Interval censoring is experienced when the period between observations or the follow-up time is discontinuous, which can be quarterly, monthly, or weekly. Left truncation or late entry happens when participants may have experienced the targeted event before being examined (Selvin, 2004).

Partly interval-censored data entail interval-censored observations and exact observations, which mostly occurs in health studies and clinical trials requiring periodic follow-ups with patients (Guure, et al., 2006). Here, the failure period is determined precisely for estimated participants simultaneously with the rest at the fixed period **(**Kim 2003; Alharpy and Ibrahim 2013; and Zyoud et al., 2016).

It is also crucial to be aware that most survival times are skewed, which limits the effectiveness of analysis techniques that are based on normal data distribution. In turn, this emphasizes the importance of examining statistical techniques for analyzing time-to-event information. Examples of these techniques include parametric methods

such as; Weibull, exponential, log-logistic, and log-normal, Gompertz given by George et al., (2014) and non-parametric such as; Kaplan Meier, Nelson-Alan, life table, and semi-parametric such as; Cox proportional hazard (Abbas et al., 2019). These models impose varying distributional propositions on the hazard. The final decision, nevertheless, regarding their application is based on the specific research question, how the model fits the actual data, and other practical matters such as challenges when approximating with the available interpretability and software. Parametric models, for example, assume that a survival function is based on a parametric distribution such as a Weibull distribution or an exponential distribution. The benefit of parametric models is to make the survival functions smooth. It is easy to suggest the behavior of these models rather than using a technique to make the functions smooth after initially estimating the function. Covariates can also be integrated easily in a parametric technique and inference method (Abbas et al., 2019). The only drawback is that parametric models must describe the data effectively, which may be true or untrue since methods such as visualization techniques or hypothesis testing may be required for testing the model (Abbas et al., 2019). Non-parametric models entail non-parametric density assessments in the availability of censoring. The benefit of the model lies in its flexibility and the ability of its complexity to develop with the observation numbers. The main drawback of the model is concerning the difficulty in integrating covariates, which makes it challenging to explain how the survival functions of people differ. Another disadvantage **is** that survival functions are not smooth. The semi-parametric model deals with the integration of covariates issues. The model breaks the instantaneous risk or hazard into a non-parametric baseline that all participants share and a relative risk that explains how each covariate influences risk (Abbas et al., 2019). In turn, this leads to a time-varying baseline risk and enables patients to possess various

survival functions in the same fitted model. The drawback of the model is that the survival function is not smooth. Besides, for correct inferences and good predictions, two propositions including proportional hazards and linearity between log-hazard and covariates must be satisfied.

This thesis aims to apply Cox Proportional Hazard Regression Model (PHRM) with partly interval censoring (PIC) data in the social field rather than the medical or engineering fields.

1.3 Cox Proportional Hazard Regression Model (PHRM)

The PHRM is a survival model used to analyse failure time data. Cox (1972) proposed a PHRM for the analysis of censored survival data that allows the inclusion of covariates. The hazard rate is,

$$h(t,z) = h_0(t)\exp[f(z_i)], \tag{1.1}$$

where $h_0(t)$ is the underlying hazard function, $f(z_i) = \beta_1 z_{i1} + \beta_2 z_{i2} + \ldots + \beta_p z_{ip} = \beta^T z_i$, z_i represent covariates (the covariates such as gender, age, type of treatments, etc.), and $\beta_i (i = 1,2,\ldots,p)$ is the unknown regression coefficients. Cox (1972, 1975) obtained estimates of β and asymptotic covariance matrix using a partial likelihood argument. Breslow (1974) proposed an estimate for the underlying hazards rate assuming that the hazard rate was constant between death times.

1.4 Problem Statement

PHRM has received considerable attention in the statistical literature as many studies involve assessing of covariates in the presence of right-censored, left-censored and interval-censored. For example, in the medical field, in a study of patients with lung cancer, an individual could die from lung cancer (the event of interest) and some others

individual still alive when study end. From industrial reliability, the failure is attributed to the malfunctioning of one of three components (motherboard, disc driver, or power supply) in a study of computer component systems.

Several approaches have been proposed when the failure time is known. A challenging twist to the problem arises when the failure time is unknown exactly but can be narrowed down within interval. In this case, there is a need to tackle such problems first in order to establish good inferences, or our inferences will not be reliable.

Significant effort has been done to consider Cox model, based on right censored failure time data oppositely to the effort based on interval censored data. However, to the best knowledge of this researcher, no one has considered the case of prison PIC data. This means that the research for PIC data is still ongoing. In this research it is proposed to study and develop a Cox PHRM based on prison PIC data to assess the effect of covariates on the model via imputation techniques.

1.5 Research Objectives

The essential aim of this research is to develop a flexible method for prison partly interval-censored. By using the time in the prison and release time for each prisoner (i.e. response variable), and some other incorporates explanatory variables information as in social applications, where most of the methodological work has focused on estimation of survival function without covariate adjustment. Hence, the major objectives of this research are:

1. To simplify Cox proportional hazard model under maximum likelihood estimation framework on the basis of prison partly interval-censored data via imputation techniques.

2. Evaluating the effect of explanatory variables (covariates) on the models and the survival function.
3. Compare the proposed model with existing model.
4. Evaluating the performance of the proposed approaches using simulation data.
5. Applying the proposed techniques to real prison partly interval-censored data.

1.6 Scope of the Thesis

The researcher has officially received the approval of Dean of the faculty for the purposes of the study and support for the data collection. The research supervisor Dr. Faiz Elfaki has accompanied the researcher Miss Shaikha Ahmedi in a field trip to Management of penal and correctional institutions in Qatar to collect data for this study. After collecting all the required data, it was documented an excel sheet then transform into R-software. The collected data contained some information for criminal nationality, age, social status, education level gender, the crime and the time for instance in the prison and release time which is the response time.

This thesis is limited to use the Cox PHRM to fit a flexible model for prison partly interval-censored data, using several socio-demographic variables. The literature review summarized in chapter 2 shins the lights on several studies linked with partly interval censoring methods, Cox PHRM, and application studies related to prisons. Chapter 3 present the Cox PHRM and introduces the maximum likelihood estimation approach which will be used to estimate the parameters. In chapter 4, real prison data as well as a simulation study were performed to evaluate the accuracy of the used methodology. Finally, chapter 5 is devoted to summarize the output of this thesis and gives several perspectives for a future work.

CHAPTER 2: LITERATURE REVIEW

This chapter will present a background about the partly interval-censored, Cox Model and its applications in addition to the literature review that provides some previous studies that have been done relevant to our study.

2.1 Partly Interval-Censored (PIC)

Peto and Peto (1972) mentioned and analyzed data comprising of either exact or interval-censored observations when deriving asymptotically efficient rank invariant test techniques to detect differences between two sets of independent observations. It is possible to correctly observe the failure times for a part of participants using this data while for the other participants the failure times happen in a specific time assessment (Nesi et al., 2015). PIC data, thus, comprises of both exactly observed and interval-censored which mean that some targeted events are exactly observed while the remaining events stay within intervals (Fauzi et al., 2015; Wu et al., 2019). PIC data occurs mostly in situations that entail periodic assessment (Nesi et al., 2015). The Weibull distribution model is used to develop partly interval-censored data (Fauzi et al., 2015). In a study to compare treatment survival functions using the imputation procedure for PIC data based on Weibull distribution assessment and established that the accuracy of the estimated sample size affects the power of the sample (Nesi et al., 2015).

Partly interval-censored data models also require the identification of a sample and a control sample to allow for observations over a vector parameter to be made for both samples' times. The proportional hazards model is mostly used in these cases for which established a asymptotic properties of generalized log-rank class tests based on data with PIC (Lane et al., 1986; Zhao et al., 2008). The model then produces a survivor function that estimates the probability of the survival periods in the future (Lane et al.,

1986). The term with observed events dominates the likelihood function for PIC data (Wu et al., 2019). Disregarding the interval-censored observations from the entire data set, nevertheless, leads to enlarged standard error and estimation bias (Wu et al., 2019). It is challenging to fit the correct model to PIC data because of factors such as violations of independence and linearity of the data, ignoring vital covariates in studies, which in turn affect variances, biases, and variances' estimate of the parameter estimates. These factors compel researchers to approximate the model being fitted. Besides, fitting a model may raise both analytic and descriptive value, which emphasizes the importance of avoiding violating the assumptions to ensure that the truth value of the model is high (Binder, 1992). Imputation methods can also estimate PIC with non-parametric approximations (Zyoud et al., 2016). In their research the found that the best approaches in this regard entail mean and median imputation and random imputation as they produce better outcomes compared with other techniques. Other techniques that can be used to examine PIC such as maximum likelihood, Expectation Maximization (EM) algorithm, and multiple imputation method and etc. Another proposed technique to estimate functions for partly interval-censored data is the semi-parametric Cox proportional hazards regression models and weighting technique model and the censoring complete model (Elfaki et al., 2013). In their studied also highlight the importance of the generalized missing data principle in the context of semiparametric models and the application of the generalized profile data for non-identically distributed samples.

When examining a failure time distribution, it is vital to ensure that the sample comprises of both items with known failure time and items with only a lower bound of the failure time. The later items have a censored survival time. Observations that have not failed by the end of the examination or those that are eliminated from the study for

other reasons besides failure use censoring (Lane et al., 1986). When designing the required sample, it is crucial to consider the availability of ties between the survival times observed, which allows for the selection of a fitting model to the data in the study, and the time dependence of variables because the independent variable value stays constant over the study time interval (Lane et al., 1986).

In this thesis, we will use PIC based on prison and simulation data sets that applied to survival model such as Cox model via imputation methods.

2.2 Cox Proportional Hazard Regression Model (PHRM)

Cox (1972) developed the PHRM to manage continuous time survival data. The PHRM refers to a technique for examining the effect of different variable on the period a specific event takes to occur (Liu, 2017). The assumption behind this model is that the core hazard rate, not the survival time, represents the covariates and independent variables' function. The model as described in (1.1) and was presented by (Liu, 2017) as;

$$log\left[\frac{h(t_i)}{h_0(t_i)}\right] = \beta_1 z_{1i} + \beta_2 z_{2i} + \beta_3 z_{3i} + \cdots \beta_k z_{ki}\,, \tag{2.1}$$

where $h(t_i)$ is the hazard function which is the probability a target event occurs at time t_i assuming the participant survived at and beyond t_i. The baseline hazard is represented by $h_0(t_i)$ refers to the hazard of the respective participant given all independent variables are equal to zero. In equation (2.1), $z_1, z_2 \ldots, z_k$ represent a set of k covariates while $\beta_1, \beta_2, \ldots, \beta_k$ represent the corresponding regression coefficients. To interpret the PHRM model, Hazard Ratio (HR) is used. HR refers to the projected hazard function based on two separate values of a predictor variable (George et al., 2014). For instance, an event can possibly occur if the HR is greater than 1 and less

probable to occur if the HR is less than 1. The covariates vector is linked to the model in which β represents the unknown parameter and defines the covariates' effects (Kumar & Klefsjö, 1994). The assumption concerning the multiplicative covariates' effect and the hazard baseline rate means that the share of the rates of hazard for two process variables experienced at time *t* concerning the set of covariates x_1 and x_2 correspondingly stays constant regarding time and proportional to each other, which demonstrates why the model is referred to as the Cox PHRM (Kumar & Klefsjö, 1994).

Cox (1972) proposed the PHRM as mentioned in chapter equation (1.1) with considered being partly because the function for the partial likelihood used for inferences was considered a function of β (the vector of regression parameters). However, several researchers are not compelled to handle the baseline hazard function, which leads to efficiency as the resulting β estimator is equivalent asymptomatically to the β estimator offered by the complete likelihood function. The model also lacks underlying assumptions and can estimate the possible failure time, which makes it beneficial in predicting failures (Lane et al., 1986). The explanatory variables affect the model by multiplying the hazard $h_0(t)$ by the function $exp(\beta_i z_i)$ of the explanatory variables' deviations from their mean values, which is the underlying assumption of the model (Lane et al., 1986). The exponential function $exp(\beta_i z_i)$ also simplifies the estimation of the vector of regression parameters (Lane et al., 1986). The model, nevertheless, has *several assumption* such as the true model differing from the traditional model through missing covariates, dependent observations, nonlinear exponential argument, hazard functions not being proportional, and the assumption that the process producing the censoring in right censored data is separate from the remaining lifetime (Binder, 1992). The proportional hazards model is not a truly *non-parametric* model because of its reliance on the vector regression parameter. Its

baseline hazard function$h_0(t)$, nevertheless, is considered to be random without the need for distributional assumptions to estimate it or β. The model is *semi-parametric* where $exp(\beta_i z_i)$ the parametric part is and $h_0(t)$ is the semi-parametric part, and $exp(\beta_i z_i)$ is the independent variable function. The main assumption based on this model is that the independent variable does not change based on the time interval being used as it remains constant over time (Liu, 2017). The model has other assumptions. For example, the proportional hazard assumption states that in a regression-based environment, the hazard functions representing survival curves for two or more strata (identified through specific value selections for the interested study) must be proportional over time (constant relative hazard). Based on this assumption, the baseline hazard function is common to all participants in a study, which means that all participants have the same baseline risk (Liu, 2017).

Compared to other discriminant analysis procedures, the PHM offers extra data about the possible time to an event offered by the model (Lane et al., 1986). The extra data is contained in the estimated survivor function for a given item with z as an independent variable vector.

The PHM should be based on the design approach because violating the assumptions of the model will lead to estimates of β. The design-based approach model produces consistent estimate of the exact underlying parameters with few efficiency losses compared to a pure model if the model is universally true for all participants. The model-based approach, however, may lead to misleading outcomes if it fails in certain respects (Binder, 1992). If β is model free, which means that the PHM assumptions do not limit it. When the underlying study participants follow the PHM the usefulness of β is enhanced, this in turn does not exempt the researcher from fitting and classifying the applicable explanatory variables (Binder, 1992). Due to the use of a hazard function

the PHRM does not require the analyst to assume a specific survival distribution for the data (George et al., 2014). Studies find the baseline hazard to possess beneficial data because the baseline hazard rate is the reference in a survival model that shifts as a function of time (Royston & Lambert, 2011). The absolute effect of an exposure relies on the time since the origin and the size of the essential hazard rate even when the proportional hazard assumption is rational (Royston & Lambert, 2011).

Bender et al. (2005) used the exponential $exp(\beta_i z_i)$ distribution extensively to generate survival times in most simulation analysis, which leads to the underutilization of other distributions. The main reason for this is the lack of obviousness when generating survival times based on pre-specified PHRM measurements. In additional to that Bender et al. (2005) developed the general association linking the hazard and the survival time of the PHRM.

Several researches used the model in their studied such as; Royston & Lambert (2011) developed relation that can generate survival times using any compatible distribution with the proportional hazards such as Gompertz, Weibull, and exponential distribution. Bender et al., (2005) used several practical situations that are flexible distributions than the exponential distribution when examining the features of the PHRM.

George et al., (2014) used a stratified PHRM that can accommodate a different baseline hazard from stratum to stratum or fitting a model that entail time-varying covariates. A crucial issue about the PHRM concerns the understanding of the true coefficients in which the impact of the covariates must be translated from the hazards to the survival times. The reason for this is that rather than the hazard function, individual survival time data are required by the software packages regarding the Cox model. It is easy to translate the coefficients from hazard to survival time in the presence

of a constant baseline hazard function, which is why the exponential distribution is widely used.

Vaida and Xu (2000) presented a PHRM with random effects in the log relative risk in which the effects affect the design matric subjectively. Their model is useful in examining clustered survival data. Also, Vaida and Xu (2000) extended the stratified models immediately in which the parameter numbers do not increase due to various baseline hazards based on the assumption that ties do not exist as the baseline hazards non-parametric maximum likelihood estimate (NPMLE) has masses at the observed periods only.

Examining the underlying assumptions of the PHRM for all predictors examined in the model is vital to ensure accuracy. For example, studies recommend plotting the Schoenfeld residual versus time to evaluate the PHRM for a continuous predictor. If random scatters around zero appear in the Schoenfeld residual, then the assumption of the model is valid (Schober & Vetter, 2018). For categorical predictors, the Kaplan-Meier survival curves' log-log transformation for various categories can be compared. The curves under the PHRM will be nearly parallel without intersecting after separating (George et al., 2014).

Another important idea regarding the PHRM is that the covariates values can change with time, particularly in follow up situations. There are, therefore, two types of covariate, time-dependent and fixed. Fixed covariates occur if their values do not change with time, for instance race or sex. Time dependent covariates occur if the difference between their values for two separate participants changes with time, for instance cholesterol in serum. Practically, some observations may occur simultaneously, which the classical PHRM cannot handle. In such case, alternative models can be used. The PHRM also faces the issue of collinearity. Fan and Li (2002)

developed a smoothly clipped absolute deviation (SCAD) penalty in the PHRM to solve such issues. However, in this thesis the use of PIC data based on PHRM via imputations techniques.

2.3 Applications

Tripodi et al., (2010) examined whether being employed is related to criminal behavior for people freed from prison, particularly concerning the duration between the prison release and re-imprisonment , furthermore, it examined the relationship between employment and recidivism for parolees freed from prisons in Texas, whether being employed after being freed from prison is related with reduced potential for re-incarceration, and whether being employed is related with more time to re-incarceration. In their study**,** they analyzed administrative data from a random sample of 250 male parolees released from prisons in Texas among 2001 and 2005. They obtained pre-prison and in-prison information from the statewide data of the Executive Service of the Texas Department of Criminal Justice (TDCJ). They also received post-prison information from the Parole Services Department of (TDCJ) using the case files of the parolees. In their study later they analyzed the combined data for the selected participants. PHRM is used to analyze the impact of employment on re-incarceration over time. The model was sufficient for the study since recidivism did not occur for a part of the participants before data collection ended, which censored the data. The study found that while being employed is not related to a substantial reduction in the potential for re-incarceration, being employed is related to a substantial amount of time to re-incarceration Re-incarcerated parolees who are employed spent more time away from crime in the community before going back to prison.

The study by Benda et al., (2002) was about environmental factors that predict the survival of inmates in the community without experiencing recidivism. The purpose

of the study was to detect the factors that predict the length of time graduates of boot camps remain in the community without being re-incarcerated. Specifically, the study sought to determine the dynamic and static factors that predict re-incarceration or recidivism among boot camp graduates in the Department of Correction. In their study they selected 480 male participants in a boot camp in one southern state through questionnaires administered by a psychologist, and used the questionnaires to collect additional data for the study such as race, marital status, committed offences, return offenses, incarceration time, and age. Besides, that they obtained the ratio measurement level of recidivism concerning the survived days in the community. The study used the PHRM to assess the relative recidivism (hazard function) rate throughout the follow-up interval of three years based on the predictors. The PHRM was used because of its flexibility concerning the reliance of the re-incarceration hazard on time and the ability to allow them to examine the impact of predictors on recidivism. The study found that factors such as the perceptions of inmates regarding boot camps as just an proper place to early release, resilience, future success expectations, peer association and influence, past criminal history, socio-demographic features, personal attributes, personality, and age at first arrest strongly predicted recidivism.

The study by Benda (2003) was about recidivism among boot camp graduates involving male non-violent offenders. The study also sought to determine whether adults in adult boot camps early starters and late starters experienced a different criminal rate of recidivism, besides, to examine the crucial care giving factors to understand the extent to which they predict criminal recidivism and explore the differences in impacts on criminal behaviors.

Their research involved 601 male participants graduates in the study from the only boot camp in one southern state and obtained various features of the participants

such as age, number of children, legal annual income, education, race, marital status, employment status, family structure, gun carriage, drug selling, and recidivism through questionnaires. The research also determined the ratio of the measurement level of recidivism for the survived days in the community. The study used the PHRM to examine the recidivism hazard rate (parole or arrest violation) of different aspects of developmental and general models. The analysis was based on the age at which the participants began engaging in illegal acts. The study found caregiver factors to be inversely related to the recidivism hazard while carrying weapons, drug sales and use, gang membership, peer relationship with criminals, social skills' deficits, and low self-esteem were found to be positively related to the recidivism hazard. The results were observed irrespective of the age at which participants began engaging in criminal activities.

Another study by Benda et al., (2005) was about the life-course theory factors that predict recidivism, the gender differences in the predictors, and issues about the impact of boot camp. The study aimed to investigate, determine and explore potential gender differences in the components of the life-course theory that predict recidivism; and recidivism abuse that happens at various life span stages; in addition to open discussions regarding the potential detrimental impacts of boot camp. In their study they selected 601 male and 120 female graduates from the boot camp in one of the southern states and used two questionnaires to obtain information such as age, education, age for the first arrest, race, childhood physical and sexual abuse, existing living status, job status, gang membership, weapon carriage, and drug selling. The PHRM was used in their study to analyze the gender differences while the non-parametric examination of survival curves was used to explore the time until the first parole violation or felony arrest of participants using standard life table techniques. The

study found that specific positive views about the boot camp program were related to low recidivism hazard rates. Present sexual assaults, adolescent physical and sexual maltreatment, and sexual abuse during childhood were also associated with high recidivism hazard rates. Ameliorating experiences such as full-time jobs and the presence of a conventional partner substantially reduced the hazard rates of many examined predictors.

Cloyes et al., (2010) studied the rates of recidivism among offenders suffering from a mental illness and who are returning to prison. Also, in their study they engaged in the study to explore further issues regarding whether specific prisoners with serious mental illness exist at the State prison in Utah, the criteria to be used in identifying this population, and the way to compare with other prisoners. The objective of their study was to determine, measure, and explain the part of the prison population in Utah State Prison between 1998 and 2002 that met the Severe Mental Illness (SMI) criteria and to compute time from prisoner release to re-incarceration for SMI offenders compared to non-SMI offenders. The researchers involved all individuals released from the Utah State Prison from January 1, 1998, to December 31, 2002, together with all release events that included 14, 621 real meaningful release events and 9,245 unique cases, and also conducted a systematic review of records of all identified cases related to SMI and gathered data concerning mental health intervention in prisons, prison resource use and management, and demographics. The study used the Kaplan-Meier techniques to perform the survival analysis in which time from prison release to re-incarceration for the SMI group was compared to that involving the non-SMI group. The study found substantial differences between the non-SMI and SMI group were due to factors associated with resource use and clinical symptoms, not demographics, release conditions, or offense features. The study also found that SMI offenders had a higher

rate of recidivism.

Hill et al. (2008) engaged in a study to identify criminal risk factors by examining forensic psychiatric reports about sexual homicide perpetrators in Germany. The study sought to collect data about the risk factors that predict future sexual homicide; to explore the legal outcomes of the sexual homicide, assess the factors that affect release from prison or a forensic hospital, evaluate the rates of criminal recidivism, and determine the risk factors for violent nonsexual and sexual reoffending. The researchers assessed court reports on 166 men who had been involved in a sexual homicide for the period between 1945 and 1991 to identify clinical, criminal, and socio-demographic factors. The study also examined the German federal criminal records for follow-up information regarding the incarceration duration in a forensic hospital or prison following the last sexual homicide and regarding reconvictions and further detentions for 139 offenders. The study used the Kaplan-Meier technique for survival analysis to evaluate the influence of risk factors on the potential to be released and to measure rates of recidivism following release as a function of time at risk. The main findings of the study were that high sexual recidivism was associated with young age at the sexual homicide period while past nonsexual and sexual delinquency, high scores in risk evaluation tools, and psychopathic symptoms led to increased non-sexual violent recidivism. The study also found that high recidivism rate with violent re-offenses was related to age-based factors such as young age during the first sexual offense, at homicide, and during release and detention duration.

The study by Jung et al., (2010) was about the rates of recidivism and survival time among male ex-inmates freed from the Allegheny County Jail in 2003. The study objective was to examine recidivism based on racial disparity among ex-inmates and to explore the relationship between recidivism and race with ex-inmates. The study also,

compared recidivism rates across race by first generating inmate historical information concerning their entry and release date documentations. A sample of 12,545 participants was included in which 46.9 per cent were black while 53.1 per cent were white. The study used the Kaplan-Meier involving log-rank tests and the PHRM to explore whether black ex-inmates recidivated within a shorter period than white ex-inmates. The Kaplan-Meier technique compared the survival curves across race while the log-rank test identified the statistical significance of the compared differences. The PHRM investigated racial differences in the risk of recidivism. The study found that the rate of recidivism for three years stood at 55.9 per cent. Black men were also found to experience recidivism at a higher rate compared to white men. The survival analysis also demonstrated the existence of racial disparity in recidivism and the recidivism rate of black male to be within a shorter period than that of the white men. The study also found the covariates and interaction impacts of a race to be substantial.

Mackie et al. (2001) studied post-transplantation alcohol consumption and the risk factors related to recidivism. The objective of the study was to compare survival rates for participants who experienced transplantation for ALD with participants who experienced transplantation for other kinds of chronic liver illnesses. The study also sought to evaluate post-transplantation consumption of alcohol, assess the existing screening procedure, and evaluate the potential risk factors that can be used to identify patients at a higher risk of recidivism. In their study, also they used a self-report questionnaire to evaluate pre-and post-transplantation alcohol consumption and patient notes to examine recidivism risk factors. The study sample comprised of 49 participants who experienced transplantation for Alcoholic Liver Disease (ALD) between May 1996 and November 1999 and 49 participants who experienced transplantation due to non-alcohol induced chronic liver illness for comparison objectives. The study used the

Kaplan-Meier technique to determine survival rates for 1- and 2 years while the log-rank test compared the rates. The study found high rates of recidivism, even though most participants did not drink heavily at a damaging level. The study also found that participants in the ALD group who consumed alcohol took a long time to do so in comparison to participants outside the ALD group, even though participants who returned to heavy drinking in both groups did so rapidly. Women were also found to experience low recidivism rates than men while age and socioeconomic status had no significant effect. Divorce was the only social risk factor that significantly influenced recidivism rates.

The study by Ostermann (2015) was about the post-release life of all former inmates using the existing information for those freed from prison in 2006, in New Jersey. The study sought to examine the performance of former inmate in their transition back into the community. In additional to that the study also, used three recidivism indicators including technical parole violations, a conviction for new crimes, and arrest for new crimes, and grouped participants into sets based on the release mechanism experienced such as unconditional, mandatory parole, and discretionary release. The study used the PHRM to separate the impact of parole supervision while controlling for identified post-release recidivism predictors. The study found that inmates freed to supervision after a three-year follow-up engaged less in new offenses compared to those freed unconditionally. A high percentage of paroled inmate's recidivated immediately after being released.

Rainforth et al., (2003) examined recidivism rates among former inmates who learned about the Transcendental Meditation (TM) technique in a prison in California. The study sought to explore participants from the Bleick and Abrams study who incarcerated at Folsom Prison by tracking their re-offending rate for 15 years following

their release. 120 inmates at Folsom Prison learned the TM technique between 1975 and 1982. The inmates had been paroled by October 1982. Moreover, the study also, selected 128 non-meditating participants as the control group, and obtained extra background and demographic data for both participants including rule violations before entering the study, period served during the considered term, past commitment record, age at parole, age at first commitment, age at first arrest, drug abuse history, military discharge and service, employment history, educational achievement, IQ, marital status, and ethnicity. PHRM was used to estimate the relative decline in recidivism risk as a result of treatment to measure the size of the treatment impact. The study also used a split population technique based on the Weibull distribution to describe the data for both groups in the study. The study found that TM led to permanent rehabilitation instead of just postponing the commencement of re-offending. The TM group also experienced less severe re-offending compared to the control. TM combined with group therapy significantly reduced recidivism compared to TM alone and group therapy alone.

However, from the above applications and to the best of our knowledge, no one has considered the case of prison PIC data. This means that the research for PIC data is still ongoing. In this study, PHRM will be used based on prison PIC data to assess the effect of covariates (such as age, gender, social status, and nationality) on the model via imputation techniques.

CHAPTER 3: METHODOLOGY

3.1 Introduction

Cox Proportional Hazards Regression Model (PHRM) is one of the most popular models being used extensively in survival analysis. This model envisages the assessment of the significance of various covariates in the survival times of subjects or individuals through the hazard function.

A well-recognized technique for analyzing survival data is Cox model, is based on a modelling approach and aims at exploring the effects of several variables on survival simultaneously. The Cox Model analyses the survival of patients in the clinical trials and the model facilitates to isolate the effects of treatment from the effects of other variables. By theoretical deduction, the model can also be used, if the other variables, which cannot be easily controlled in a clinical trial but affect the patient survival apart from the treatment, are also known.

Let Tbe a non-negative random variable representing the failure time of an individual in the population. Generally, the values of Thave a probability distribution that is Probability Density function (PDF) $f(t)$, however, the cumulative distribution function (CDF) is:

$$F(T) = P(T \leq t) = \int_0^t f(u)du\,, \qquad (3.1)$$

which gives the probability that the event has duration t. The survival function $S(t)$ is defined as the complement of the CDF of T. That is;

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(u)du\ . \qquad (3.2)$$

The survival function gives the probability of being alive at duration t. $S(t) = 1$ at $t = 0$and $S(t) = 0$ at $t = \infty$, which indicate that the survival function begins at $S(t) = 1$ and as t increases to ∞, decreases to 0.

Likewise, the hazard function is an important concept in survival analysis which we can say is a kind of density function $f(t)$. For which it is conditional while $f(t)$ is an unconditional probability.

According to Lee et al. (2003) and Brostrom (2019), the hazard function also known as instantaneous failure rate which is defined as the probability that an event lies in an interval $(t, t + \Delta t)$ given that it has not happened prior to t as follow:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t / T \geq t)}{\Delta t}, \quad t > 0 \quad . \qquad (3.3)$$

The function (3.3) can be seen as the probability a person dies in a short interval $(t, t + \Delta t)$ where the individual has already survived at the time t.

By integrating the hazard function, we obtain the cumulative hazard function which is comparatively easier to estimate non-parametric models than hazard and density functions.

Brostom (2012) and Lee et al., (2003) introduced the cumulative hazard function as:

$$H(t) = \int_0^t h(u)du. \qquad (3.4)$$

The conditional probability in the numerator of equation (3.3) may be written as the ratio of the joint probability that $T$is in the interval $(t, t + \Delta t)$ and $T > t$, to the probability of the condition $T > t$. The former may be written as $f(t)dt$for small dt,

while the latter is $S(t)$ by definition. Dividing by dt and taking the limit, we have;

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{d}{dt}F(t)}{S(t)} = \frac{\frac{d}{dt}(1-S(t))}{S(t)} = \frac{-\frac{d}{dt}S(t)}{S(t)} = \frac{-S'(t)}{S(t)}. \quad (3.5)$$

Equation (3.5) suggest that;

$$h(t) = -\frac{d}{dt}\ln\big(S(t)\big). \quad (3.6)$$

Then;

$$log\, S(t) = -\int_0^t h(u)du + cons\, tan\, t. \quad (3.7)$$

The constant in (3.7) will be equal zero provided that $S(0) = 1$, then (3.7) became;

$$S(t) = exp\left[-\int_0^t h(u)du\right]. \quad (3.8)$$

By substitute (3.8) into (3.5) we have;

$$f(t) = h(t)S(t) = h(t)\, exp\left[-\int_0^t h(u)du\right]. \quad (3.9)$$

As shown, the hazard and the survival functions are mathematically associated. Because of convenience and practicality, the hazard function is used in the regression model. In the next paragraph, we will introduce our model that related to the mentioned above functions.

Cox (1972) interpreted the association between the hazard rate and covariates via the

following model:

$$ln[\,h(t)] = ln[\,h_0(t)] + \sum_{i=1}^{p} z_i\beta_i. \qquad (3.10)$$

Equation (3.10) can be written as:

$$h_z(t) = h_0(t)e^{\sum_{i=1}^{p} z_i\beta_i}\,, \qquad (3.11)$$

where $h_0(t)$, $z_i$and β_i are defined in equation (1.1). Therefore, the regression model as a linear form based on equation (3.11) is given as:

$$\frac{h(t)}{h_0(t)} = e^{\sum_{i=1}^{p} z_i\beta_i} = e^{z_1\beta_1} \times e^{z_2\beta_2} \times \ldots\ldots \times e^{z_p\beta_p}. \qquad (3.12)$$

Equation (3.12) represents the prime assumption of PHRM that is the proportional hazards that is defined by $h(t)$ and $h_0(t)$ from the two independent distribution and $e^{\sum_{i=1}^{p} z_i\beta_i}$ is positive proportional constant that does not depend on t. However, the proportional hazards would not be used for all the cases. This assumption must always be carefully examined and this could be done by using some methods such as Schoenfeld residuals (Allison, 2014).

The predictive form of PHRM can be written in terms of the survival function as;

$$S(t/z_i,\beta) = S_0(t)^{exp[\beta^T z_i]},$$

where $S_0(t)$ is the baseline survival function at time t, which corresponds to the

baseline hazard $h_0(t)$ as $S_0(t) = exp[-\int_0^t h(u)du]$, $S(t/z_i, \beta)$ is the probability of surviving beyond time t given predictors and β & z_i is defined in equation (1.1).

An interesting characteristic of proportional hazards model is that to estimate the regression coefficients, only the ranks of the failure times are needed. The real failure times are only used to generate the ranks. Therefore, regardless of whether the time values are in days, months, or years, the same regression coefficient estimates will be achieved.

The PHRM is a regression model for time to event data assuming that the covariates (age, gender, treatment, etc) will affect the survival times. It enables to test the difference between survival times of different groups of patients allowing other factors (covariate) to be taken into account. The two term $h_0(t)$ and β, PHRM is called a semi parametric model as $h_0(t)$ is non parametric and $\beta$is parametric part. Moreover, the parametric part in equation (3.12) need to be estimated. In PHRM the unknown parameters β_i $(i = 1,2,..p)$ can be estimated by partial likelihood (Brostrom, 2012).

3.2 Maximum Likelihood Estimation (MLE)

MLE is known as the likelihood function of the sample data based on a mathematical expression. Also, the MLE is a numerical technique used for estimating the unknown parameters of a given model (distribution), using some observed data by maximize the probability of observing the data from the joint probability distribution given a specific probability distribution and its parameters in MLE is that their variances may be approximated routinely by the inversion of the observed information matrix.

Cox and Oakes (1984) introduced the likelihood function as

$$L(\beta) = \prod_{i=1}^{n}[f(t_i, \beta)^{\delta_i}(1 - F(t_i, \beta))^{1-\delta_i}], \qquad (3.13)$$

where $t_i = min(T_i, c)$,

$$\delta_i = \begin{cases} 1 & , T_i \leq c \\ 0 & , T_i > c \end{cases}$$

and $T_i (i = 1, \ldots, n)$ is a sample from a random variable T having a pdf given by $f(t, \beta)$, and a CDF given by $F(t, \beta)$; where β is the parameter vector and c is the censoring constant.

The estimator β is $\hat{\beta}$ which is the point in the parameter space that maximizes the likelihood function. The MLE is invariant under parameter transformation. Let a vector of parameters be $\beta = (\beta_1, \ldots, \beta_p)$; suppose that $\zeta(\beta)$ is a function of β, not necessarily one to one or differentiable. Then, the MLE of $\zeta(\beta)$ is $\hat{\zeta}(\beta)$, is given by $\zeta(\hat{\beta})$, where $\hat{\beta}$ is the MLE of β.

The MLE is asymptotically sufficient. This can be seen by expanding the log-likelihood function in a Taylor series. The resulting expression for the density function can be shown to have the asymptotic factorization given by

$$f(t, \beta) = f(t, \hat{\beta}) \, exp\left\{-\frac{1}{2}(\hat{\beta} - \beta)' I_0(\beta)(\hat{\beta} - \beta) + o_p(1)\right\}, \qquad (3.14)$$

where $\hat{\beta}$ is the maximum likelihood estimator of β, $o_p(1)$ a term that approaches 0 and $I_0(\beta)$ is the Fisher information matrix given by;

$$I_{0,ij}(\hat{\beta}) = E\left(\frac{-\partial^2 \, log\,(L(\beta))}{\partial \beta_i \partial \beta_j}\right) i, j = 1, \ldots, p. \qquad (3.15)$$

The factorization given by equation (3.14) establishes the asymptotic sufficiency of the maximum likelihood estimator (Cox and Hinkley, 1974). This means that, asymptotically, the estimator $\hat{\beta}$ contains all the information in the sample about β which explains the good asymptotic properties of the MLE.

3.3 Computation of Maximum Likelihood Estimator

Taken the log of equation (3.13), we have;

$$l(\beta) = log\, L\,(\beta). \tag{3.16}$$

If the first partial derivatives of equation (3.16) exist and the MLE does not occur on the boundary of the parameter space, then, the estimator is the solution of the system of simultaneous equations given by

$$\frac{\partial}{\partial \beta} l(\beta) = 0, \tag{3.17}$$

also, the first derivative of $l(\beta)$ is also known as the score vector defined by$U(\beta)$. The maximum likelihood is the solution of

$$U(\beta) = 0. \tag{3.18}$$

Generally, using the EM algorithm or the Newton-Raphson methods can solve equation (3.18)

3.4 Estimation in PHRM

Assume that $t_i = t_1, t_2, \ldots, t_p$ be the failure times (release time in our case) with one failure at each time and let $R(t_i)$ be the set of subjects at risk at time t_i, who are stay in jail and under observation just before t. We indicate with i the label of the subject who fails at t_i so that its vector of covariates is z. Then the full likelihood is;

$$L(\beta) = \prod_{i=1}^{p} L_i(\beta) = \frac{h_i(t_i,z)}{\sum_{i \in R(t_i)} h_i(t,z)}. \tag{3.19}$$

From equation (1.1) we have;

$$L_i(\beta) = \frac{h_0(t_i)\, exp(\beta^T z_i)}{\sum_{i \in R(t_i)} h_0(t_i)\, exp(\beta^T z_i)}.$$

The baseline hazards we cancel out, then we have the final form of partial likelihood;

$$L(\beta) = \prod_{i=1}^{p} L_i(\beta) = \prod_{i=1}^{p} \frac{exp(\beta^T z_i)}{\sum_{i \in R(t_i)} exp(\beta^T z_i)} \ . \qquad (3.20)$$

The log partial likelihood is given by;

$$l(\beta) = log\, L\,(\beta) = log[\prod_{i=1}^{p} \frac{exp(\beta^T z_i)}{\sum_{i \in R(t_i)} exp(\beta^T z_i)}] \,, \qquad (3.21)$$

where $l(\beta)$ indicates that a function depends on the unknown parameters β, the values of z being known. The large sample properties of the maximum likelihood estimators of β based on equation (3.21) have been shown to be the same as those of any estimator from complete likelihood (Cox, 1975; Tsiatis, 1981; Andersen and Gill, 1982). It is worth mentioning that equation (3.21) was given the name "partial likelihood" by (Cox 1975), as he derived the full likelihood $L(\beta)$ based on equation (1.1), and showed that inference on β could be made using $l(\beta)$, which coincides with that found in equation (3.21), and depends on β only. In large samples, the normal distribution with the score vector is the distribution used for approximated the value of β, which is estimated by maximizing the likelihood from the first derivative, and a variance-covariance matrix, which is estimated from the likelihood function based on second derivative.

$$l(\beta) = log\, L\,(\beta) = \sum_{i=1}^{p} \sum \sum_{i \in R(t_i)} exp(\beta^T z_i) [\beta^T z_i - log. \qquad (3.22)$$

The score function is defined as the first derivative of equation (3.22) shown as;

$$\frac{\partial}{\partial \beta_i}[l(\beta)] = \sum_{i \in R_i} \left[z - \frac{\sum_{i \in R_i} z\, exp(\beta^T z)}{\sum_{i \in R_i} exp(\beta^T z)} \right]. \qquad (3.23)$$

Summing equation (3.16) over all failure times, we have the pth component of the score $U(\beta)$

$$U(\beta) = \sum_{i=1}^{p} \frac{\partial l(\beta)}{\partial \beta_i}. \qquad (3.24)$$

By taking the second derivative of equation (3.22), an expression is obtained which has the form of a variance. For example, the derivative of (3.23) with respect to β_p is;

$$\frac{\partial^2 l(\beta)}{\partial \beta_p^2} = -\sum_{i \in R_i} \left[\frac{\sum_{i \in R_i} z^2\ exp(\beta^T z)}{\sum_{i \in R_i} exp(\beta^T z)} - \left(\frac{\sum_{i \in R_i} z^2\ exp(\beta^T z)}{\sum_{i \in R_i} exp(\beta^T z)} \right)^2 \right]. \qquad (3.25)$$

The negative value of equation (3.25) is the partial likelihood observed information matrix $I(\beta)$. The inverse of $I(\beta)$ which is evaluated at $\hat{\beta}$, that is $I^{-1}(\beta)$, is the estimated covariance matrix of $\hat{\beta}$. Equation (3.25) is also known as minus the Hessian Matrix is used to produce the standard errors for the regression coefficients. Based on evaluated $\hat{\beta}$ the maximum partial likelihood estimator, then asymptotically $\hat{\beta} \sim N(\beta_0, I^{-1}(\hat{\beta}))$, where is the inverse of information matrix at $\beta = \hat{\beta}$ at β_0 is a true value.

3.5 Likelihood Ratio Test (LRT)

As a goodness-of-fit test, the LRT will utilized to compare between two models. In other words, a complex model will be compared to a simpler model in order to find out that it fits the dataset better or not. For a large sample size, the a chi-square is approximated distribution of the LRT. The degrees of freedom of this distribution is equal to the difference in the number of coefficients in two models. This test is defined as:

$$LR = -2\left[L_{subset} - L_{full}\right] = -2\ln\left(\frac{l_{subset}}{l_{full}}\right).$$

The -2 in LR equation adjusts the test in a way that the chi-square distribution can be used to approximate the distribution of the test.

CHAPTER 4: RESULTS AND ANALYSIS

In this chapter, we will illustrate the implementation of the methods discussed in earlier chapters using two data sets. The first one is prison data, the second one is from generated data. All calculations were computed using R software.

4.1 Prison Data

We applied the proposed method to a modified real prison data. The data consist of 1730 criminals where variables such as age, gender, social status and nationality are measured. To apply this data set to survival study, we consider the release from prison as event of interest. Then those who are not released, are treated as censored subjects. However, for PIC we consider those released from the prison as exact value and we consider 20 months as interval censored so that we can achieve the PIC assumption (we do try less and more than 20 months but the result is better when it was 20). This study is used to apply the survival data analysis to social studied as well to implemented to compare the variables in the data sets which have more effective in a crime. Moreover, in this study, we implemented to compare the effects in crime for; 30 years or older and below 30-year-old based on age variable, for marriage and non-marriage prisoners, based on a social status variable, for males and females prisoners based on gender, and finally for prisoners from Gulf Cooperation Council (GCC) countries and others based on nationality variables.

To set up the data as the PIC data, for each age group the subjects are divided as follows; for 30 years or older subjects, 655 were right-censored, 550 were interval censored and 525 were exact data. Likewise, for below 30 years old subjects, 332 were right-censored, 890 were interval censored and 508 were exact data. The same scenario was followed for categorizing other variables.

Table 4.1 shows the result based on our model mentioned above the data set.

The results show that the variables age, gender and social status are highly significant compared to the nationality as per the Likelihood Ration Test (LRT) and their p-value criterion. Occupancy was found to be higher between the ages of 30 and above relative to younger prisoners, which is shown in Figure 4.3. The study also indicates that males commit more crimes compared to females, as shown in Figure 4.4.

Figure 4.2 shows that there is significant different between married and single for the social status in interval of more than 10 month to 250 months. In additional to that single have longer crime compare to married as shown in the Figure 4.2. For the nationality variable, there is no significant different between Gulf nation and others nation as shown in Figure 4.1. Note that the significance level we consider in this analysis is 0.05.

These results indicate the age, gender and social status are strong factors effect to commit crimes. Figure 4.3 shown the younger prison (less than 30 years) commit slightly more crimes compared to prisoners have age more than 30, males commit more crimes compared to females, as shown in Figure 4.4 and the single prisons have longer crime compare to married as shown in the Figure 4.2.

Figure 4.5 showed the log(-log) of survival function based on age group for which the two lines of age 30 years or older and younger than 30 years are parallel, for which highlighted validity or one of the assumptions of Cox PHRM. The results confirm we reject the null hypothesis for age because the p-value <0.05 where μ_1: the mean for younger prisoners (less than 30 years) and μ_2: the mean for prisoners who is age more than 30 ($H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$) as well as for social status reject the null hypothesis because the p-value <0.05, where μ_3: the mean for single prisoners and μ_4:the mean for married prisoners ($H_0: \mu_3 = \mu_4$ vs $H_1: \mu_3 \neq \mu_4$).

Table 4.1: Result from Prison data set based on Cox Model.

Variable	Coefficient	Exp(Coef)	SE	LRT* (P-value)
Nationality	-0.17853	0.83560	0.07285	6.00 (0.01431)
Social Status	-0.26010	0.77097	0.07417	12.29 (0.00046)
Age	-0.13351	0.87502	0.02613	24.33 (6.81e-4)
Gender	-1.28100	0.27800	0.012010	30.5 (3.27e-08)

LRT*: Likelihood Ratio Test

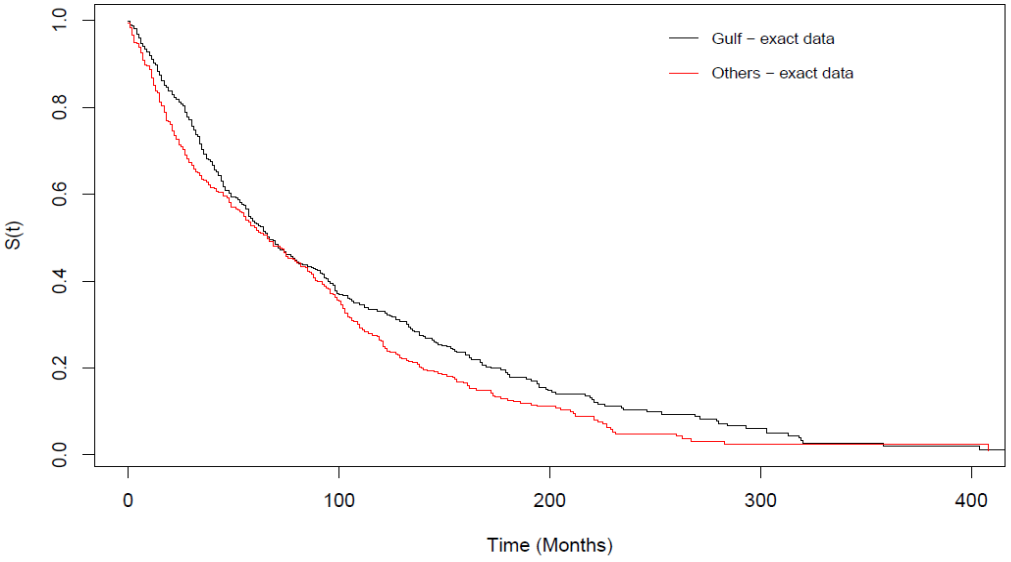

Figure 4.1: The survival function for Nationality (Gulf and others)

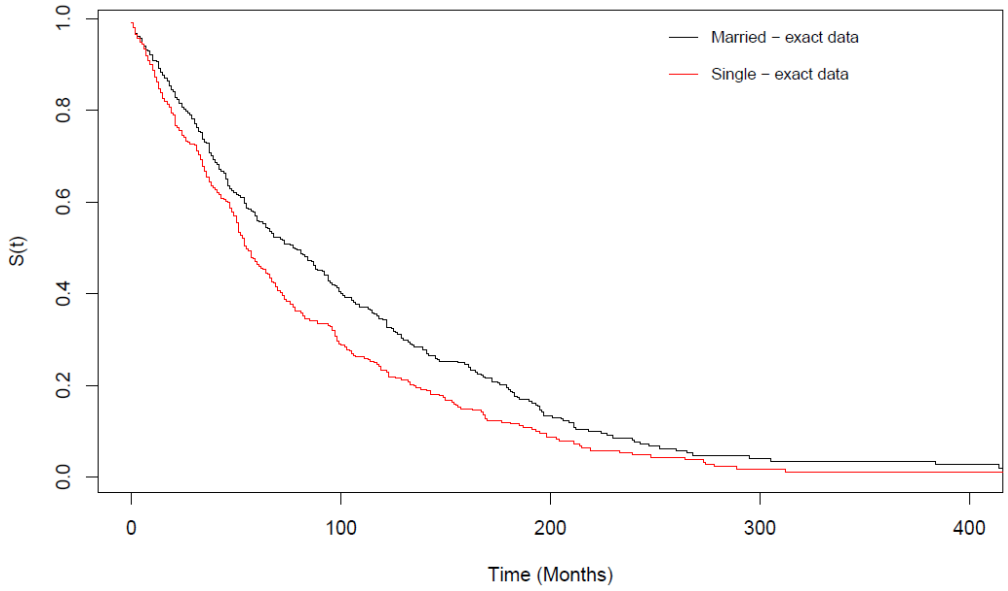

Figure 4.2: The survival function for Social Status (Married and Single)

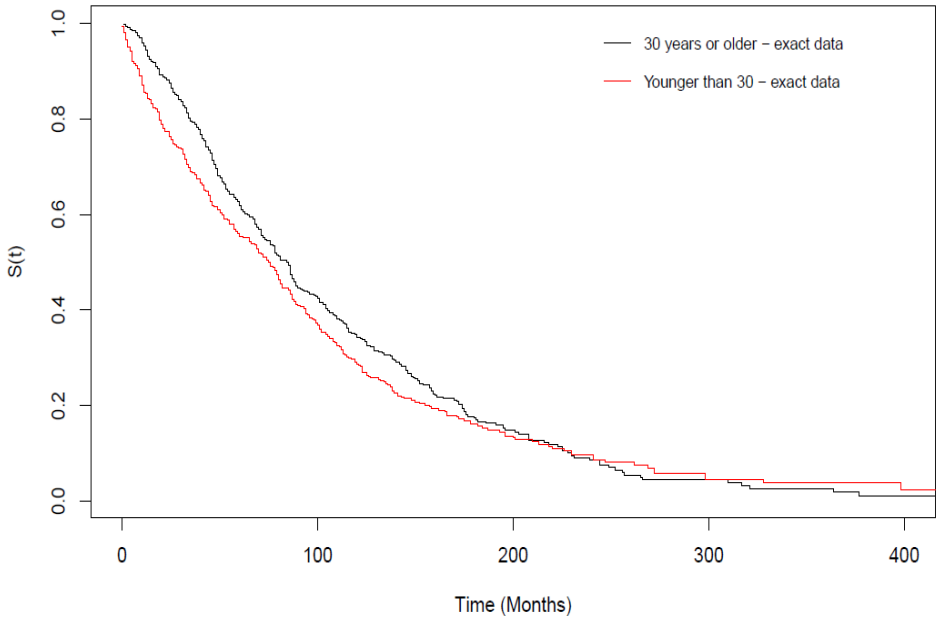

Figure 4.3: The survival function for the two failure rates of Age variable

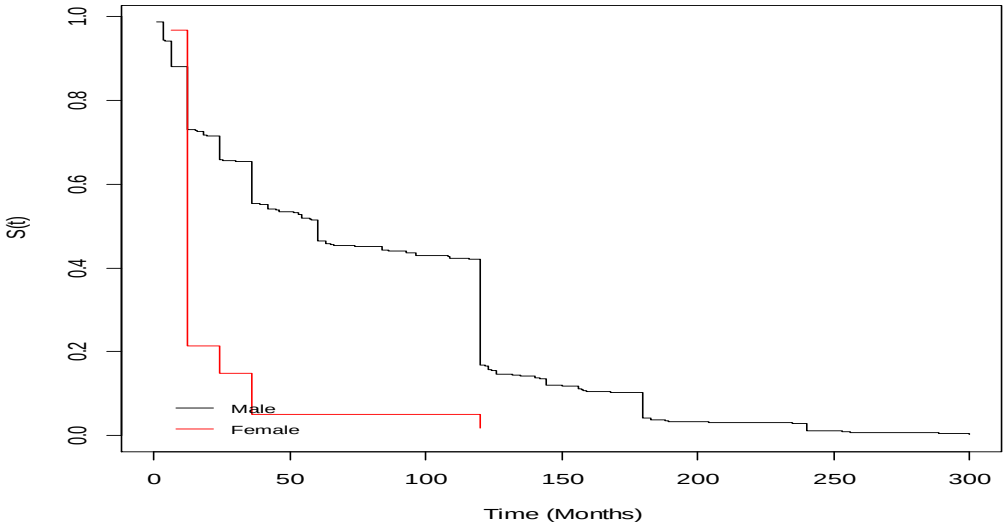

Figure 4.4: The survival function for Gender (Male and Female)

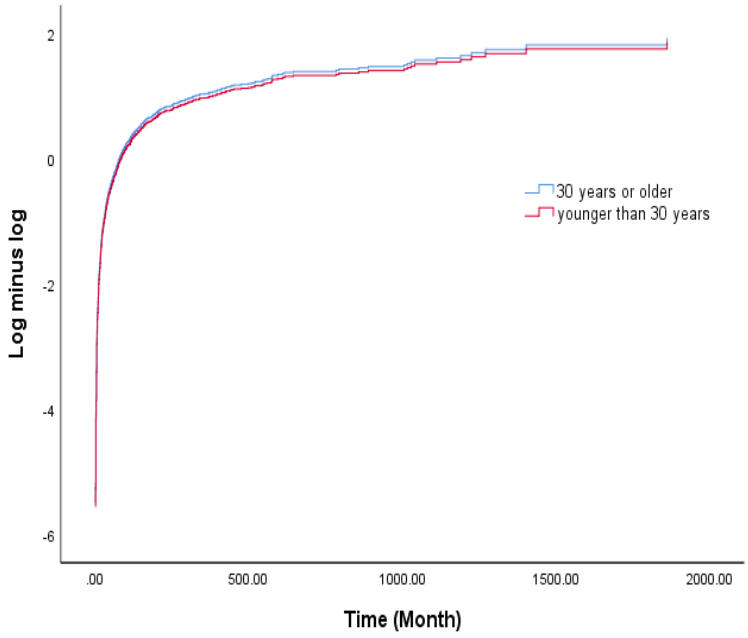

Figure 4.5: The Log minus log of survival function for Age group.

4.2 Simulated Data

A simulation study can be defined as a technique for performing computer experiments involving certain types of mathematical and logical models explaining the behavior of a particular system (Rubinstein, 1981). Simulation has been used most widely in statistics to analyze and research the conduct of statistical procedures, in particular when the problem cannot be solved analytically or when an analytical solution is not easy to work with.

The technique consists of setting up a large number of samples. The samples are then individually reckoned in terms of statistics of interest, and the overall statistics of interest is used to study distribution properties. The simulations can also be used to generate estimates of the mean, variance, coverage probability of confidence intervals. In this simulation study, the objective is to compare the survival function that obtained from imputation techniques based on nonparametric model.

In this section, we present our published manuscript about the analysis of our model using simulation data sets but will be here in more details. Note that the materials of this section have been reproduced from our article by Ahmedi et al., (2020).

In order to examine the influence of the Cox model on this prison data set and to compare the variables in the data sets, a simulation analysis was carried out on the basis of a prison real data set. (that have been mentioned in section 4.1 in this thesis). We generate data based on the distribution of Weibull since we find that the distribution of Weibull is suitable for real data (Real data histogram graphics were found to be similar to Weibull distribution curves relative to other distributions such as Lognormal and normal as shown in Figures 4.6, 4.7 & 4.8. In additional to that the AIC based on Weibull is 4311.244, for Lognormal 4361.268 and for normal is 4557.028 which confirm the mentioned results in the Figures. For each variable, the sample was taken 1000 times. (age, gender, social status, and nationality).

Firstly, to generate the data for the variable age (30 years or older and below 30 years) we used the mean and standard deviation as -0.13351 and 0.02613 based on different percentage of exact observation (0%, 25%, 50%, and 75%) in the PIC data. We used the left point, mean and median imputations against the exact observation based on our Cox model to obtained estimated survival function for the two groups of age variable that is 30 years or older and younger than 30 years. The estimates approach in Figure 4.11 and Figure 4.12 which is shown similar results compared with the one obtained by left point. However, the group younger than 30 years develops more crime earlier than those in older than 30 years, suggesting that our left point approach provides an acceptable approximation to the estimate, when we have more exact in the data compare when we have less exact as shown in Figure 4.9 and 4.10.

The mean imputation is used to obtain the estimated survival function for the

two groups of age variable that is 30 years or older and younger than 30 years against that exact data with different percentages based on our model mentioned in chapter 3, as shown in Figures 4.13, 4.14, 4.15 and 4.16. The Figures shows that the two groups of age variables have similar results as compared with the one obtained by exact data with different percentages, which indicate that our mean point are better. Similarly, the result obtained by using median imputation it looks similar to the one obtained by mean point as shown Figure 4.17, 4.18, 4.19 and 4.20.

In summary, the result obtained by our methods that is; left, median and mean imputations for exact observation more than 0% percentages are significant with respect to P-value that shown in Table 4.2. Moreover, the result obtained by mean and median imputations are better than the one obtained by left imputation especially when we have more exact observation in PIC data (Table 4.2 and Figures 4.13 to 4.20).

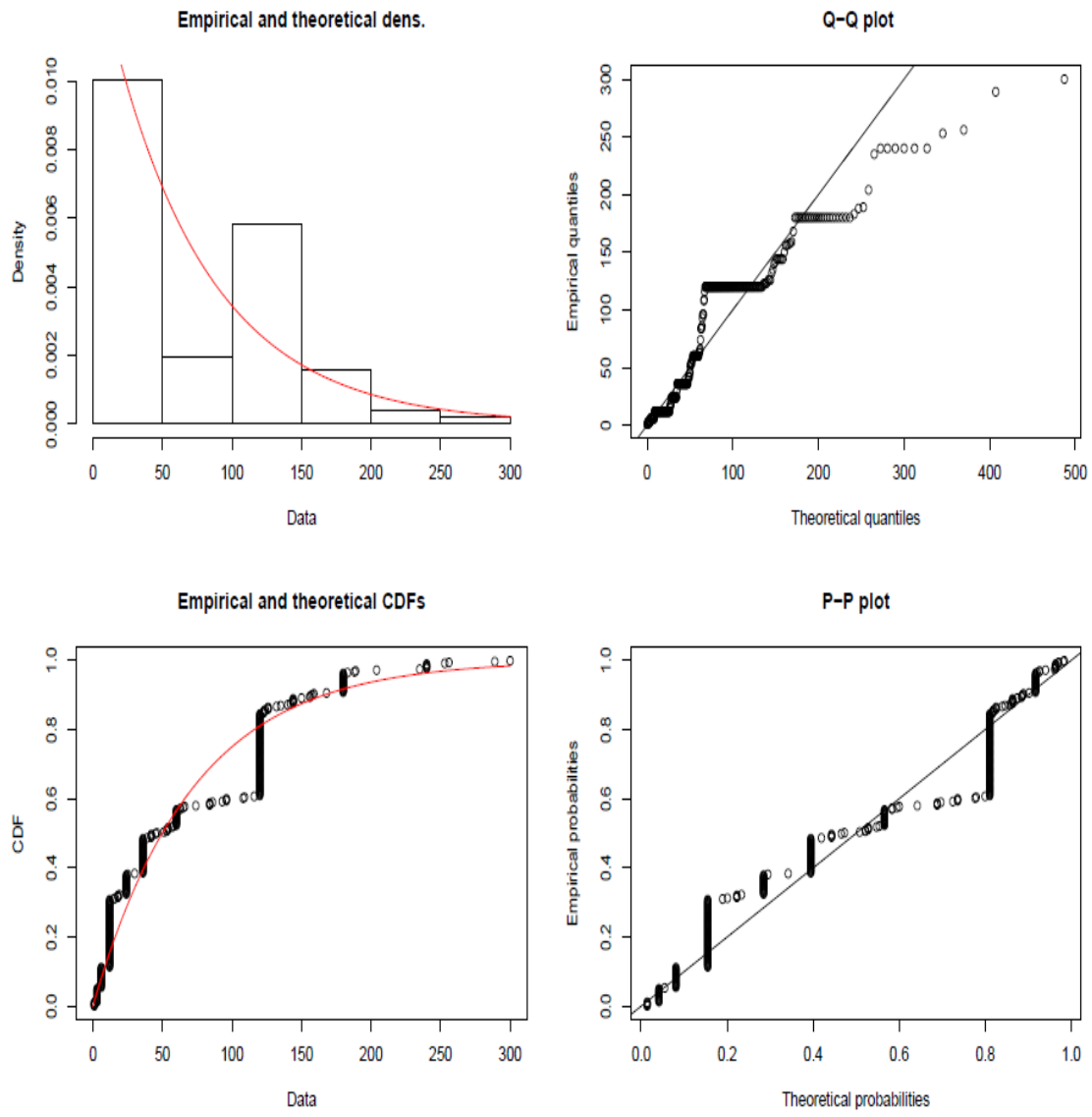

Figure **4.6**: Estimated of density function, empirical quantiles, cumulative density function and Empirical probabilities based on Weibull Distribution.

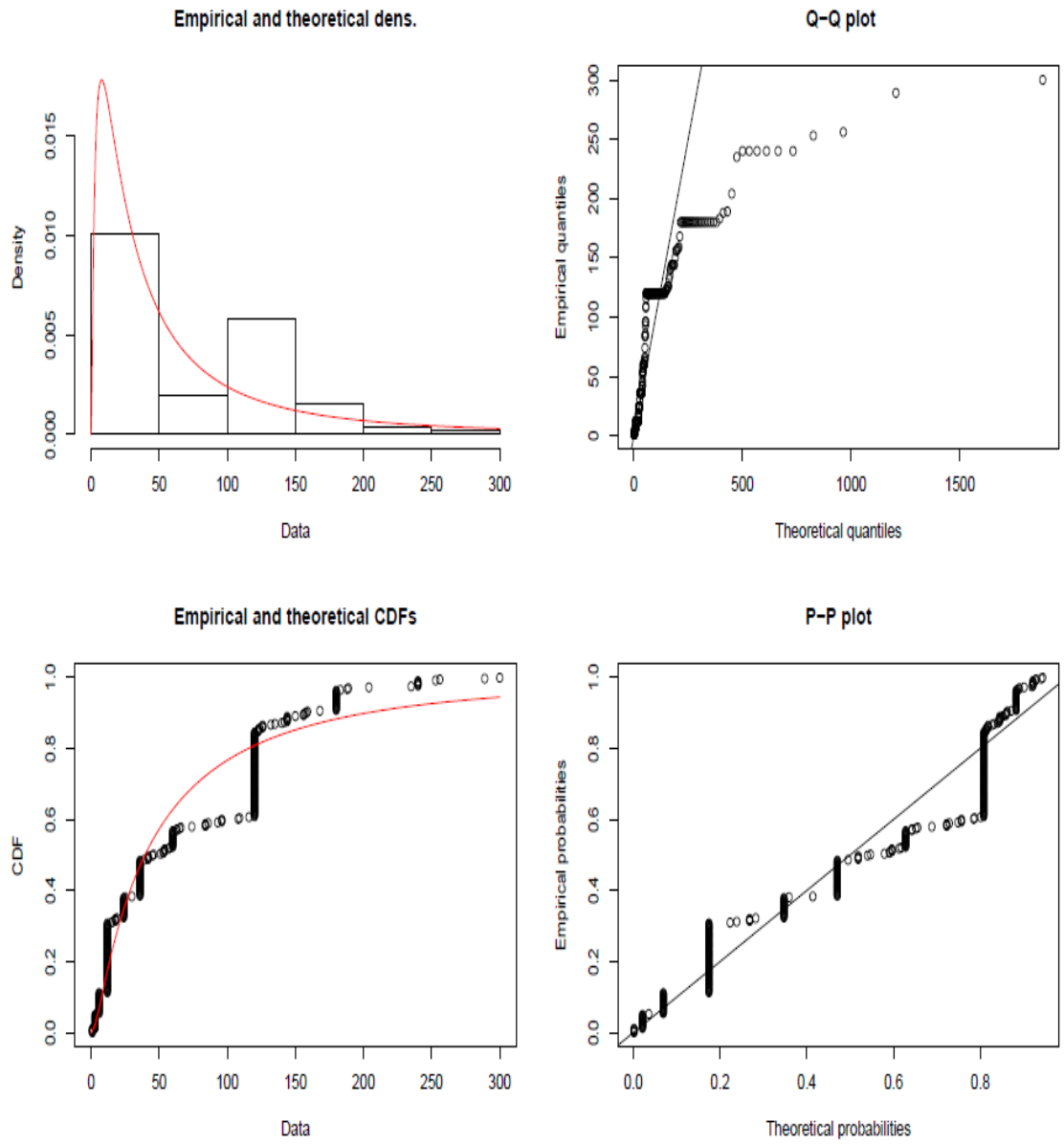

Figure **4.7**: Estimated of density function, empirical quantiles, cumulative density function and Empirical probabilities based on Lognormal Distribution.

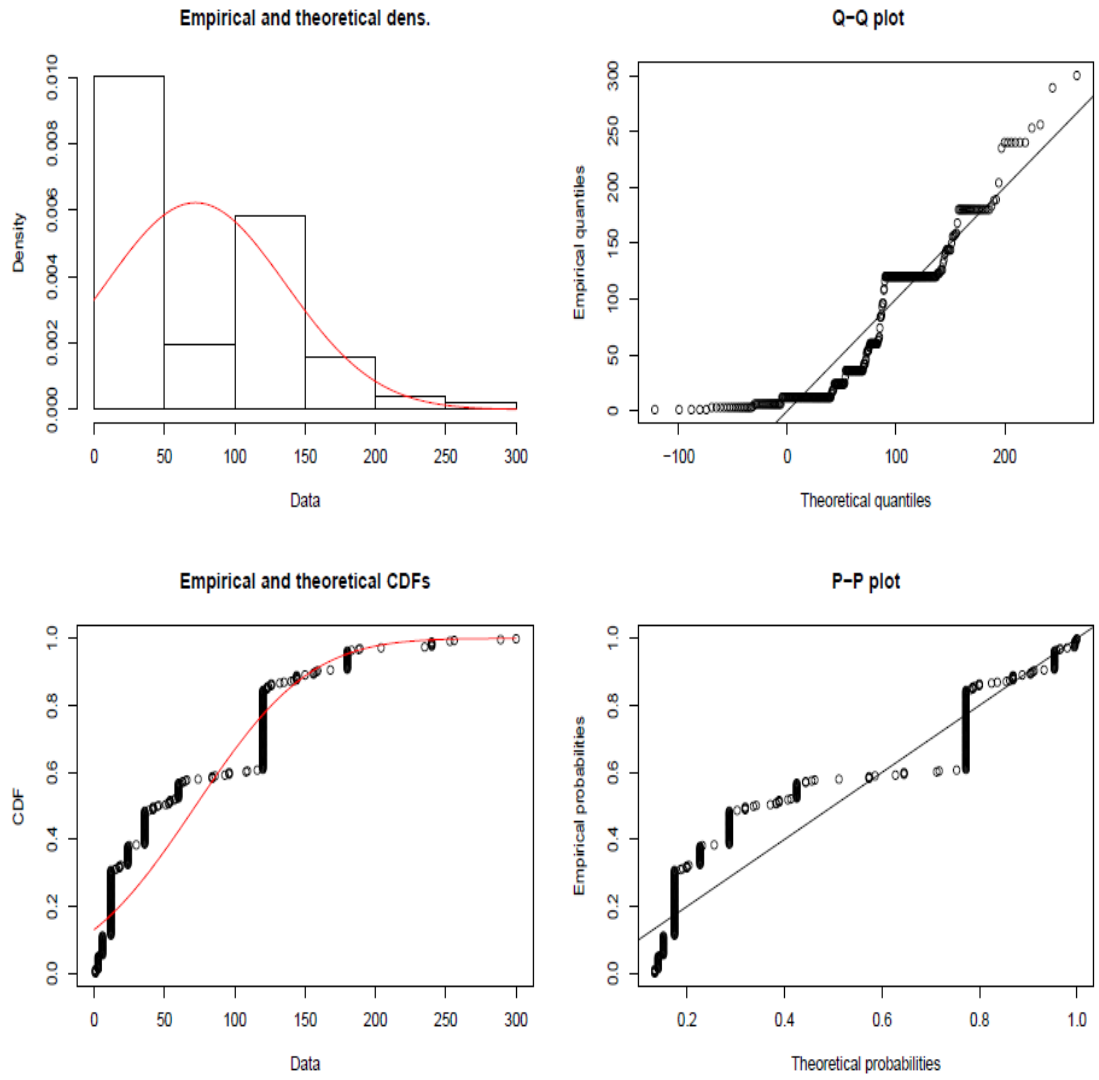

Figure **4.8**: Estimated of density function, empirical quantiles, cumulative density function and Empirical probabilities based on Normal Distribution.

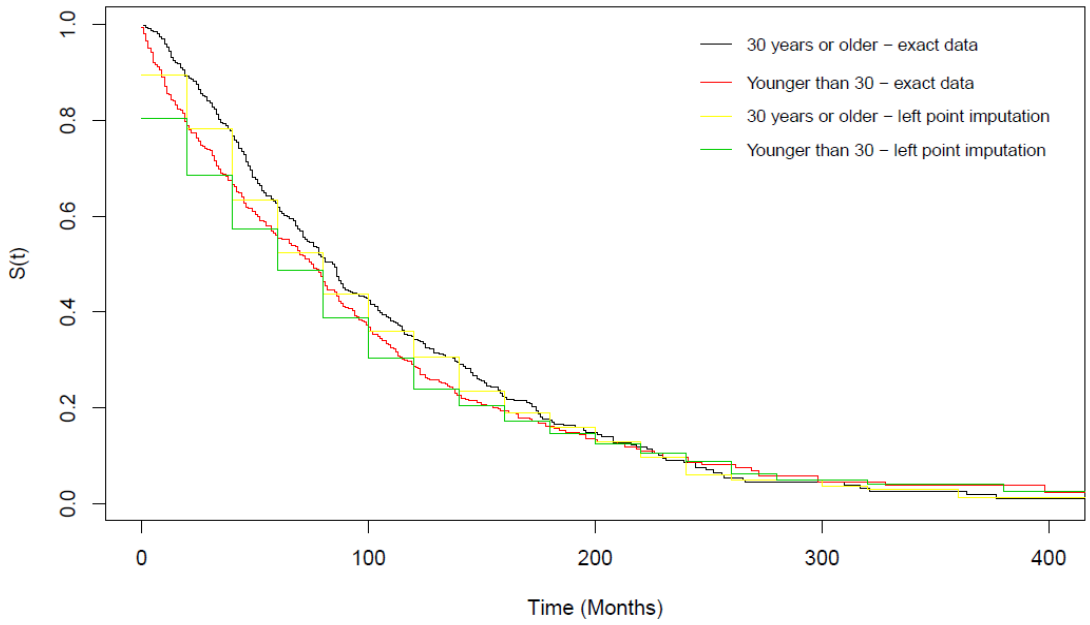

Figure 4.9: The survival function obtained by left point with 0% exact data for the two failure rates of age variable.

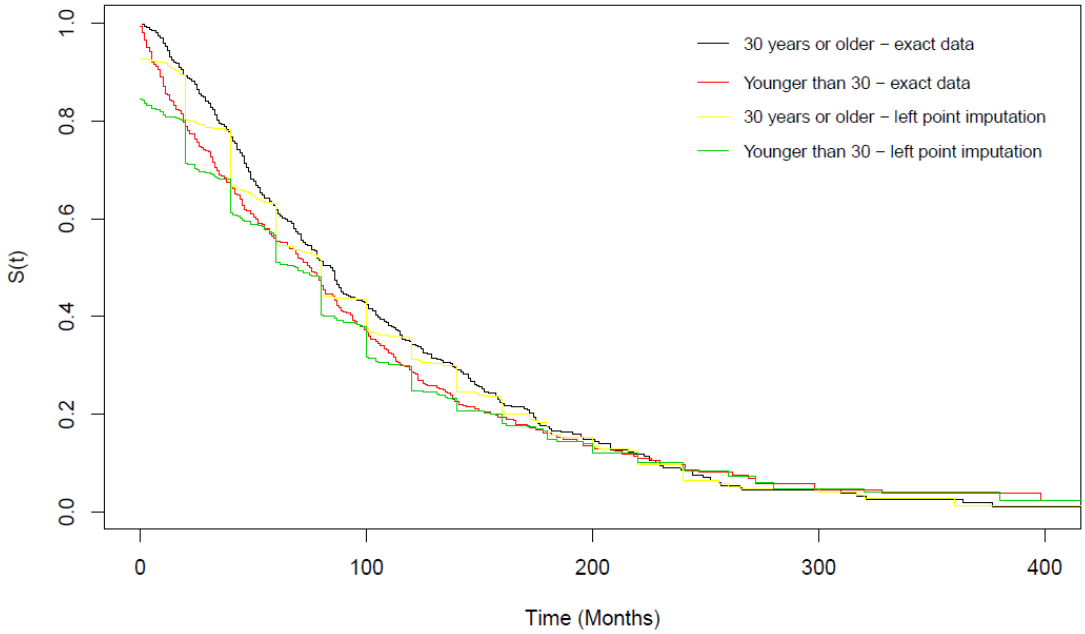

Figure 4.10: The survival function obtained by left point with 25% exact data for two failure rates of age variable.

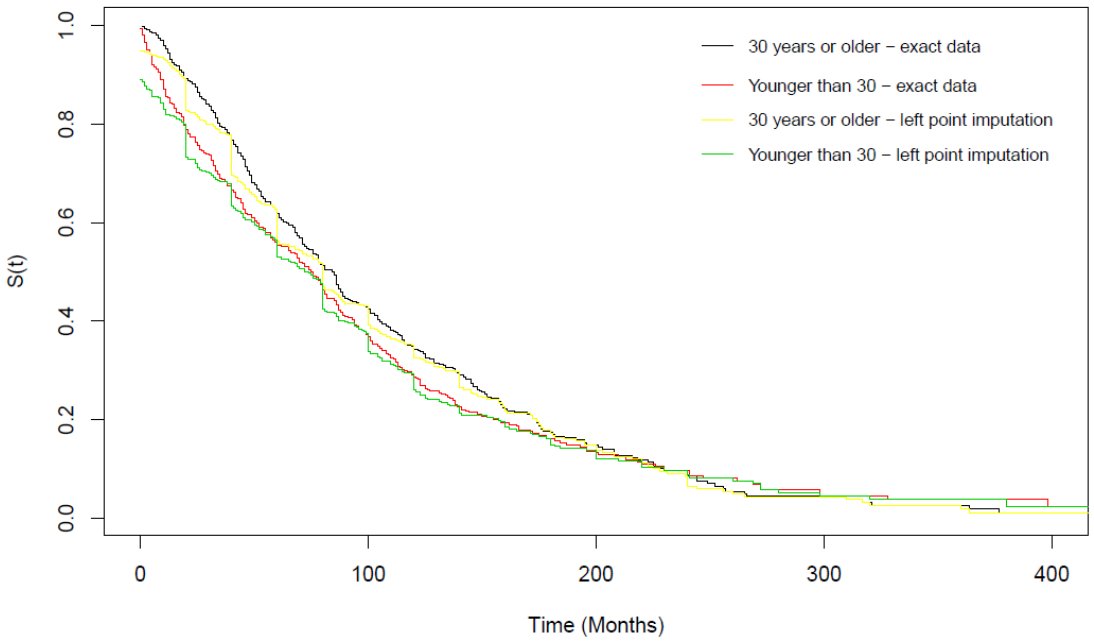

Figure 4.11: The survival function obtained by left point with 50% exact data for two failure rates of age variable.

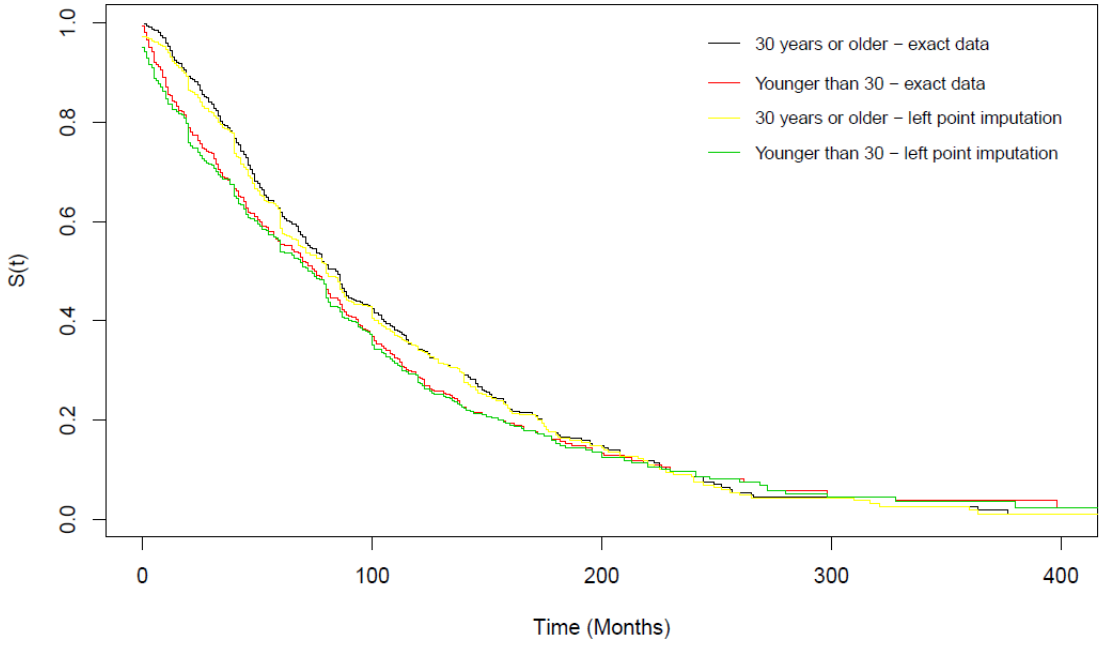

Figure 4.12: The survival function obtained by left point with 75% exact data for two failure rates of age variable.

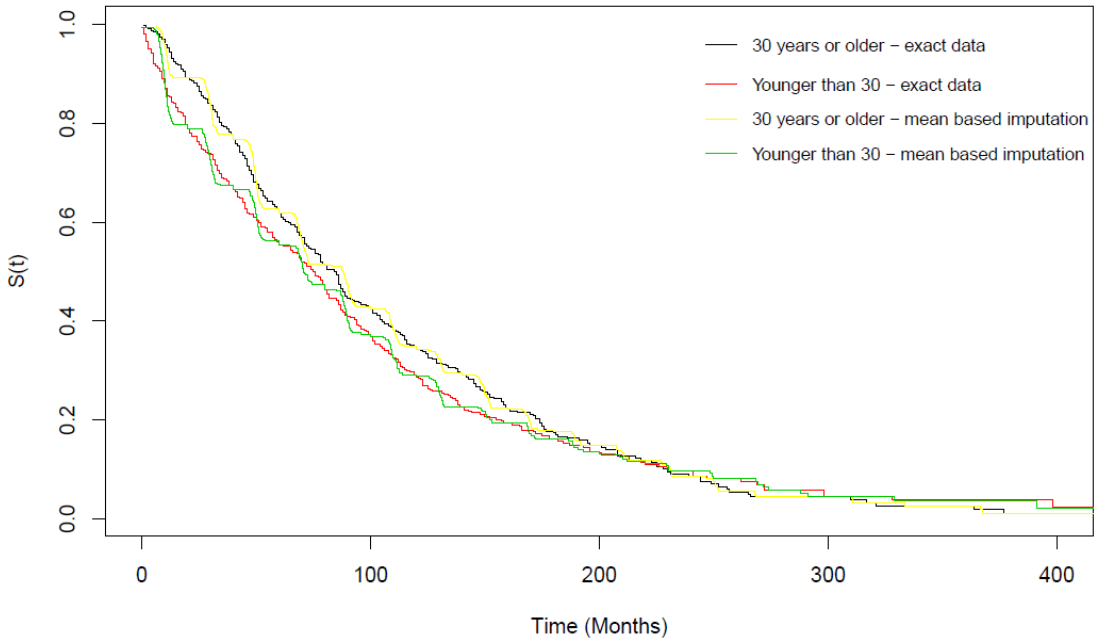

Figure 4.13: The survival function obtained by mean point with 0% exact data for two failure rates of age variable.

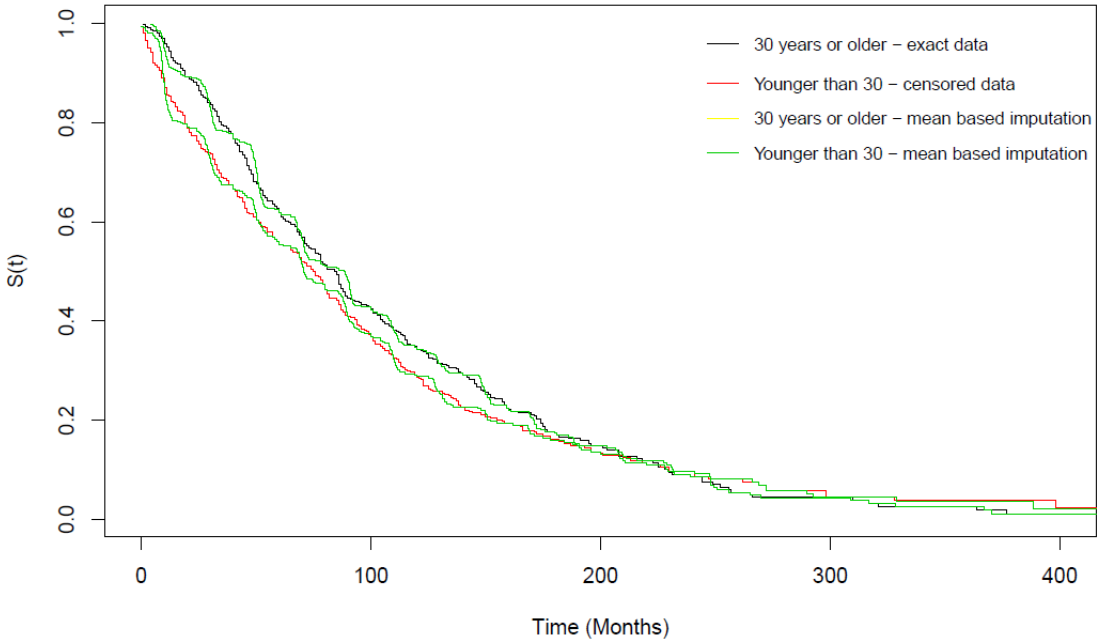

Figure 4.14: The survival function obtained by mean point with 25% exact data for two failure rates of age variable.

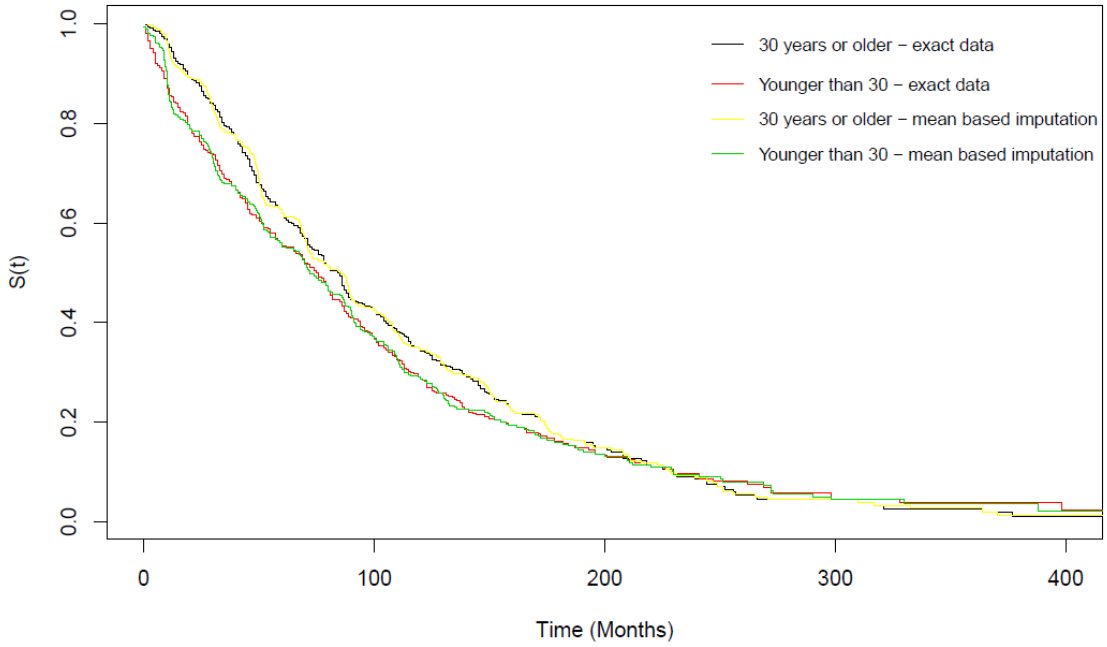

Figure 4.15: The survival function obtained by mean point with 50% exact data for two failure rates of age variable.

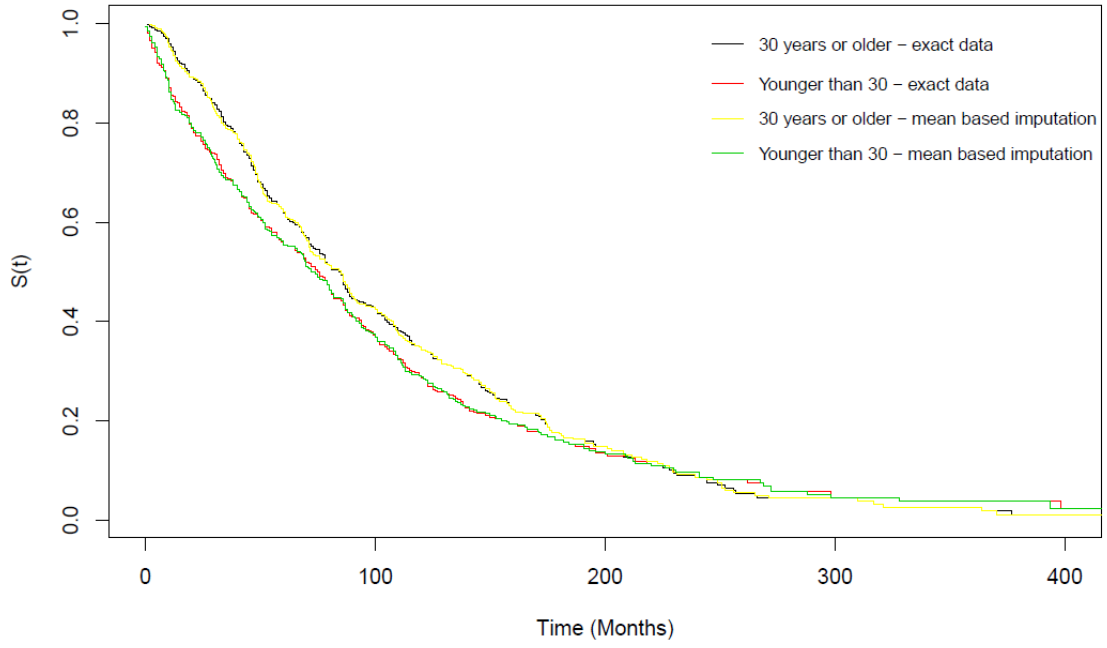

Figure 4.16: The survival function obtained by mean point with 75% exact data for two failure rates of age variable.

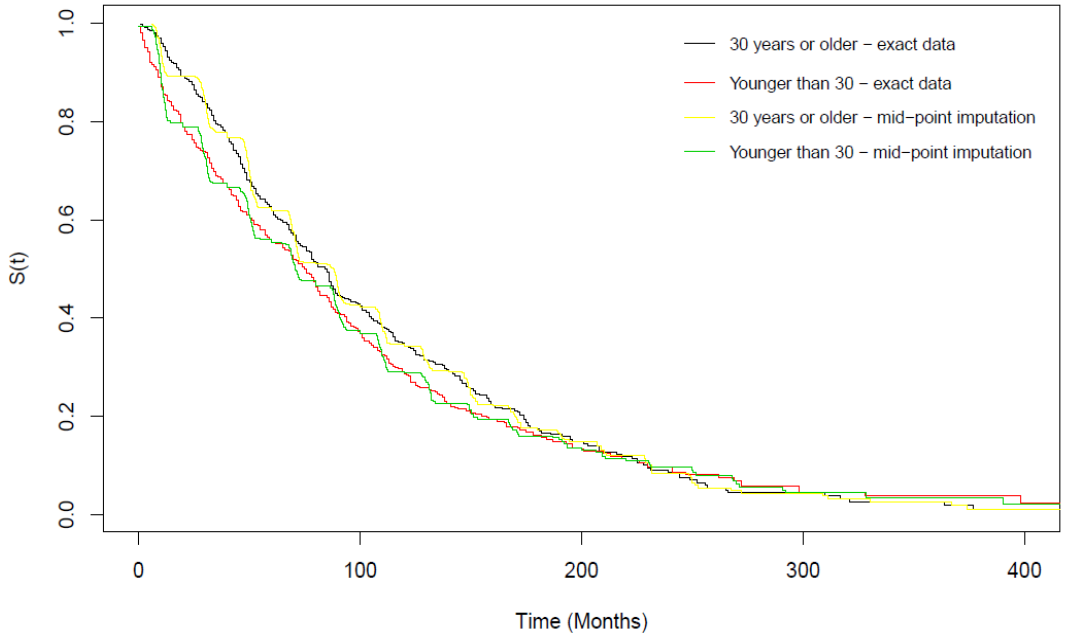

Figure 4.17: The survival function obtained by median point with 0% exact data for two failure rates of age variable.

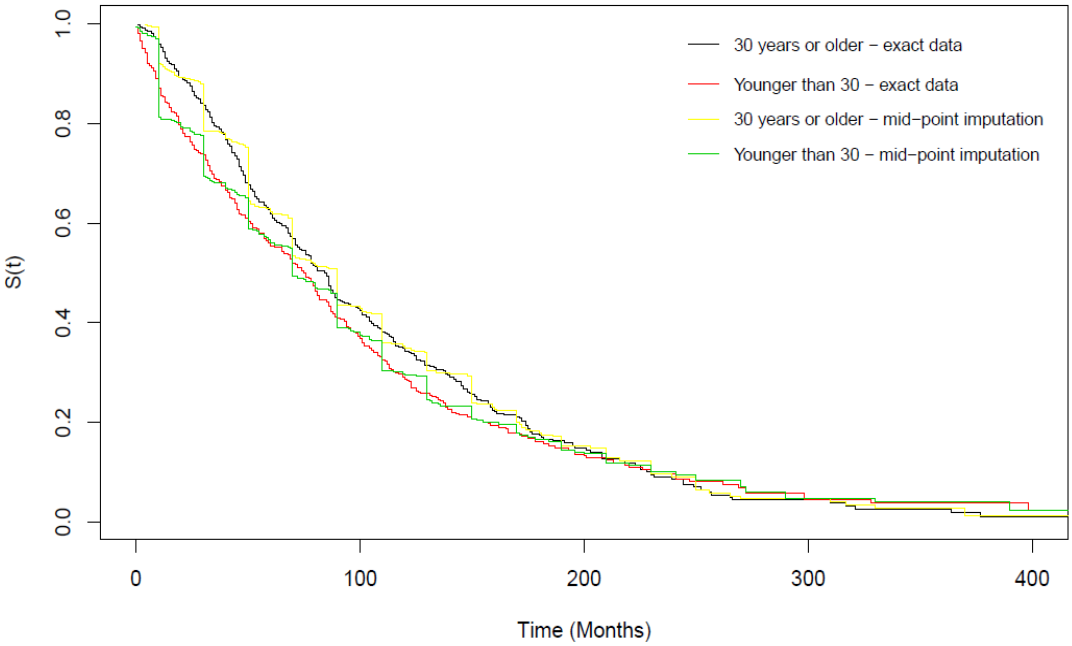

Figure 4.18: The survival function obtained by median point with 25% exact data for two failure rates of age variable.

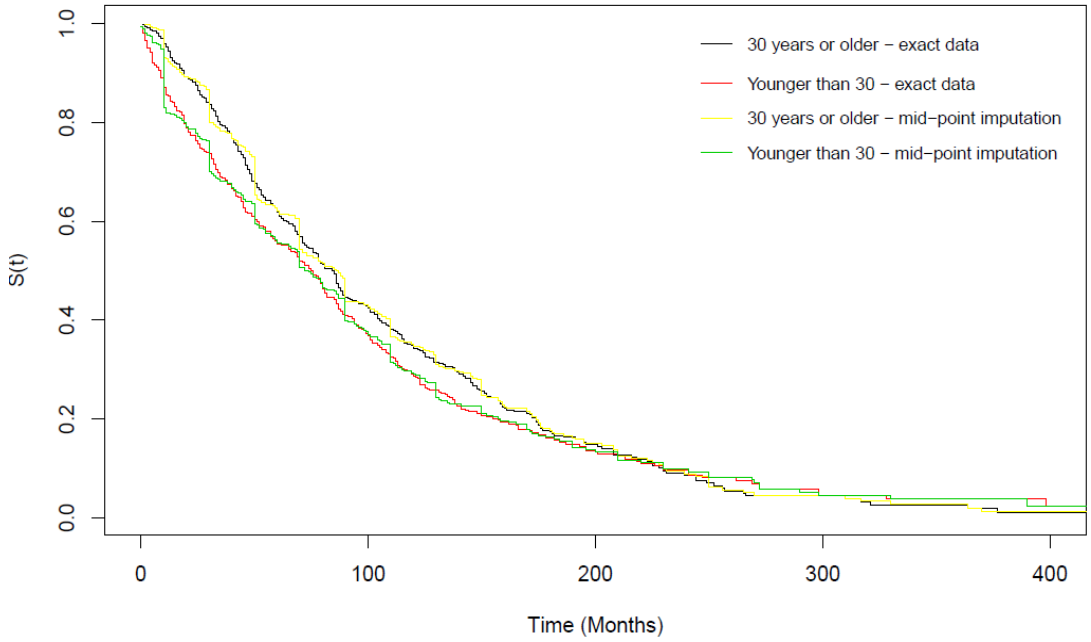

Figure 4.19: The survival function obtained by median point with 50% exact data for two failure rates of age variable.

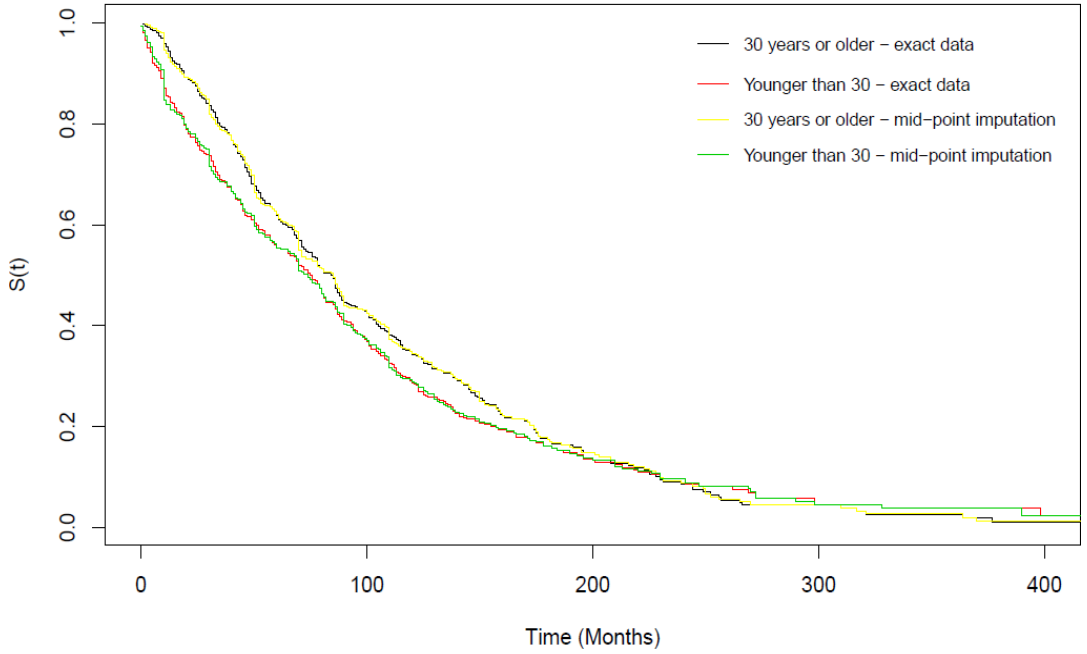

Figure 4.20: The survival function obtained by median point with 75% exact data for two failure rates of age variable.

Table 4.2: Result from simulation data for Age variable based on Cox Model.

	Imputation	Coefficient	Exp(Coef)	SE	P-value	LRT*
0%Exact	Left-point	-0.11864	0.88813	0.07314	0.105	2.63(0.105)
	Median	-0.12267	0.88455	0.07314	0.0935	2.81(0.0937)
	Mean	-0.12994	0.87814	0.07313	0.0756	3.15(0.0758)
25%Exact	Left-point	-0.11940	0.8846	0.07314	0.103	2.66(0.0103)
	Median	-0.12502	0.88248	0.07312	0.0873	2.92(0.0876)
	Mean	-0.13581	0.87301	0.07312	0.0633	3.44(0.0635)
50%Exact	Left-point	-0.12315	0.88413	0.07314	0.0922	2.83(0.0925)
	Median	-0.12974	0.87833	0.07313	0.076	3.14(0.0763)
	Mean	-0.13331	0.87520	0.07314	0.0683	3.32(0.0686)
75%Exact	Left-point	-0.12869	0.87924	0.07314	0.0785	3.09(0.0787)
	Median	-0.13254	0.87587	0.07312	0.0699	3.28(0.0701)
	Mean	-0.13351	0.87502	0.07312	0.0679	3.33(0.0681)

Table 4.3: Result from simulation data for social status variable based on Cox Model.

	Imputation	Coefficient	Exp(Coef)	SE	P-value	LRT*
0%Exact	Left-point	-0.2602	0.7709	0.0742	0.000453	12.29(0.00046)
	Median	-0.26309	0.7709	0.0742	0.000453	12.75(0.00039)
	Mean	-0.27289	0.76118	0.07418	0.000453	13.52(0.00024)
25%Exact	Left-point	-0.26225	0.76932	0.07422	0.00041	12.47(0.00041)
	Median	-0.26077	0.77046	0.07418	0.000439	12.35(0.00044)
	Mean	-0.26685	0.76579	0.07418	0.00321	12.93(0.00032)
50%Exact	Left-point	-0.2602	0.7709	0.00742	0.000453	12.29(0.00046)
	Median	-0.25682	0.77350	0.07417	0.000535	11.98(0.00054)
	Mean	-0.26027	0.77084	0.07416	0.000449	12.31(0.00045)
75%Exact	Left-point	-0.2612	0.7701	0.0742	0.00043	12.39(0.00048)
	Median	-0.25895	0.77186	0.07417	0.000481	12.18(0.00048)
	Mean	-0.26071	0.77050	0.07418	0.00044	12.34(0.00044)

The results also confirm that we reject the null hypothesis for variable social status, which mean that there is different between single prison and married prisoners based on Table 4.3. The mean imputation showed slightly better result compare to left and median imputation with respect to the value of p-value in Table 4.3.

Secondly, to generate the data for the variable social status (marriage and single) we used the mean and standard deviation as -0.2601 and 0.07417 based on different percentage of exact observation (0%, 25%, 50%, and 75%) in the PIC data.

Figures 4.21, 4.22, 4.23 and 4.24 show the results obtained based on our model by left point with different percentages of exact observations. These results are almost similar as in one obtained by left point for variable age. However, the left point imputation technique shows better results when we have more exact in case of 25%, 50% and 75% compared to 0% especially after 100 months. The single group have loner survival compare to marriage group which indicate that the single group are more crime than marriage group.

Figures 4.25, 4.26, 4.27, 4.28, 4.29, 4.30, 4.31 and 4.32 show the results obtained based on our model by mean and median imputations, respectively. The figures showed that almost similar to the one obtained by exact data except the one obtained by left point with exact 0% and 25%. Also, as we found in the left point, the single group showed that have more crime compare to marriage group. However, the mean, median and left imputations showed significant results with respect to likelihood ratio test and their P-value as shown in Table 4.3.

This result indicate that the Cox model can be easy implemented to PIC social data sets via simple imputations techniques.

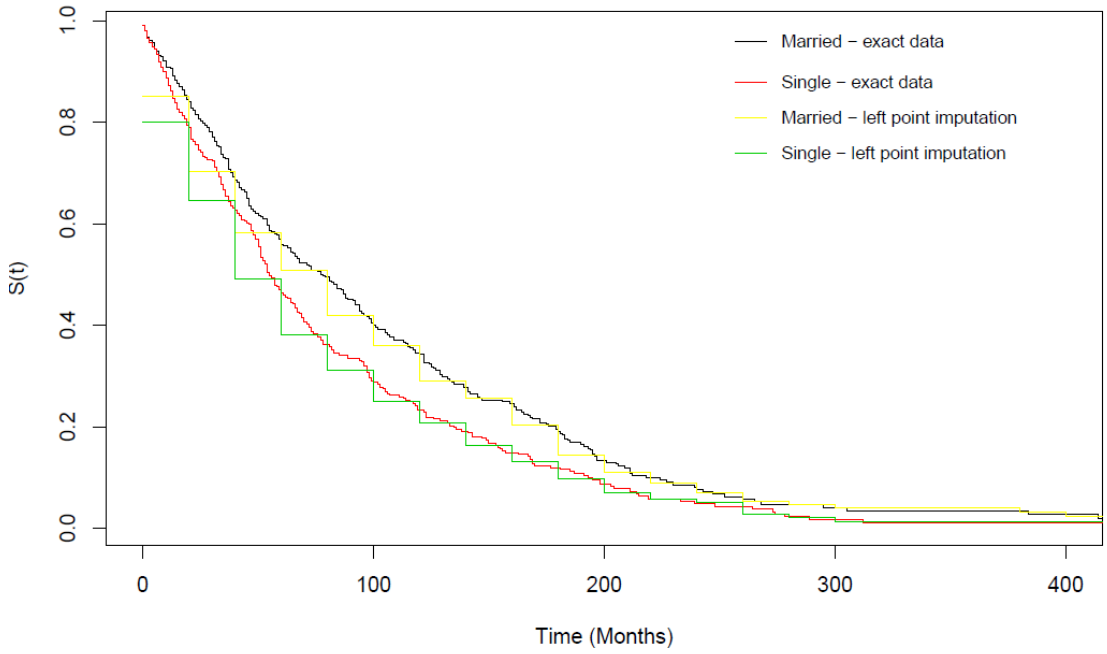

Figure 4. 21: The survival function obtained by left point with 0% exact data for social status variable (married and single)

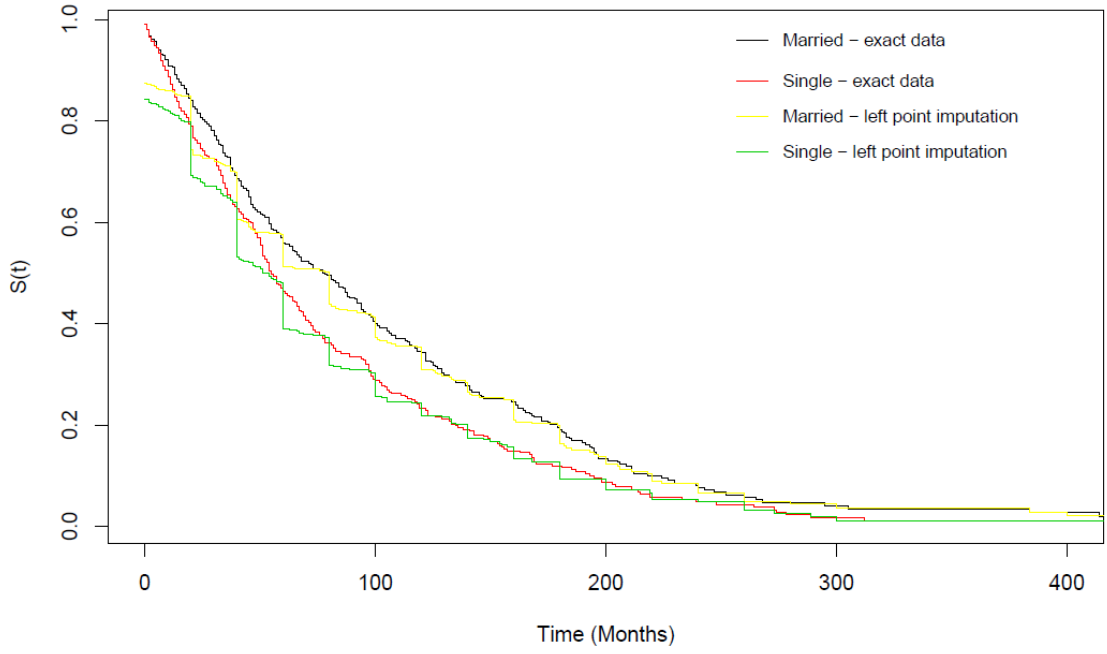

Figure 4.22: The survival function obtained by left point with 25% exact data for social status variable (married and single)

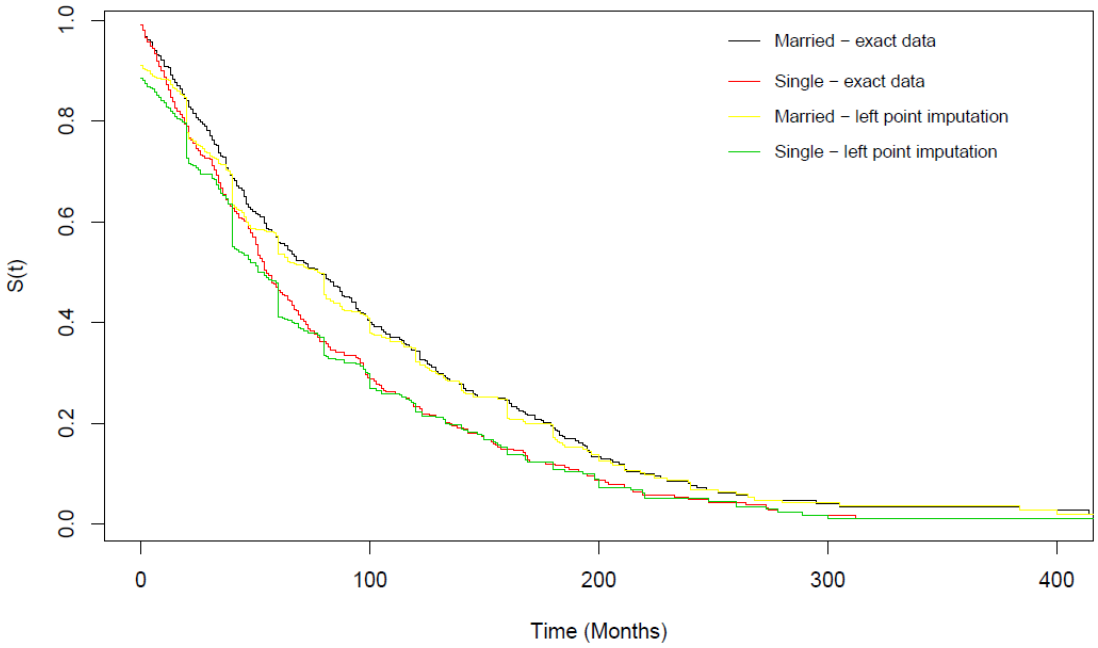

Figure 4.23: The survival function obtained by left point with 50% exact data for social status variable (married and single)

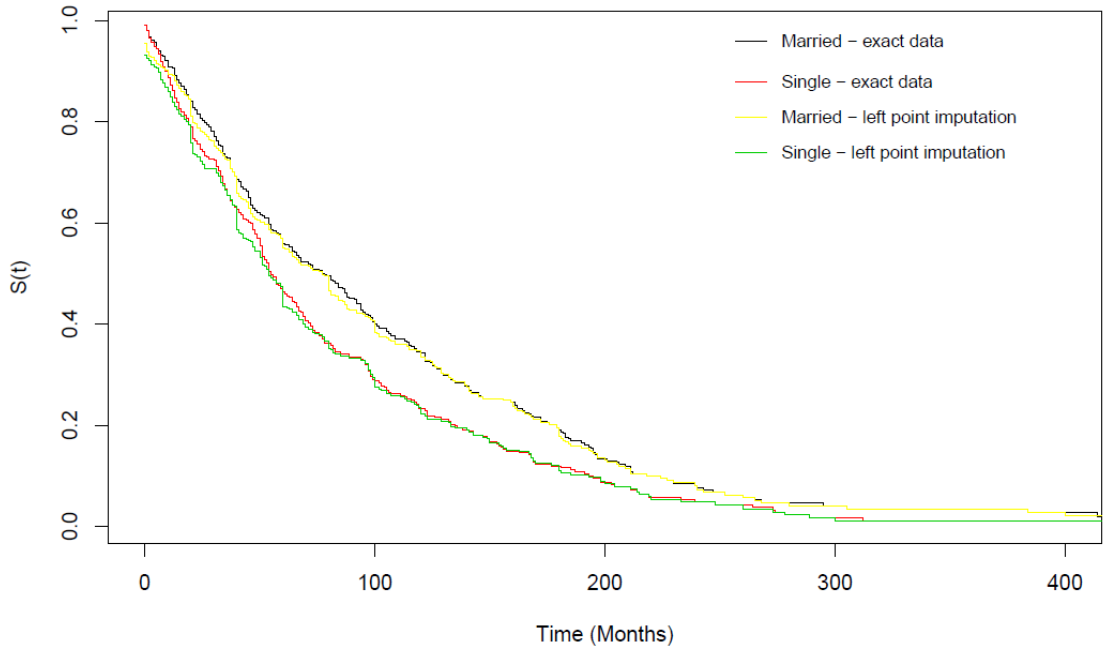

Figure 4.24: The survival function obtained by left point with 75% exact data for social status variable (married and single)

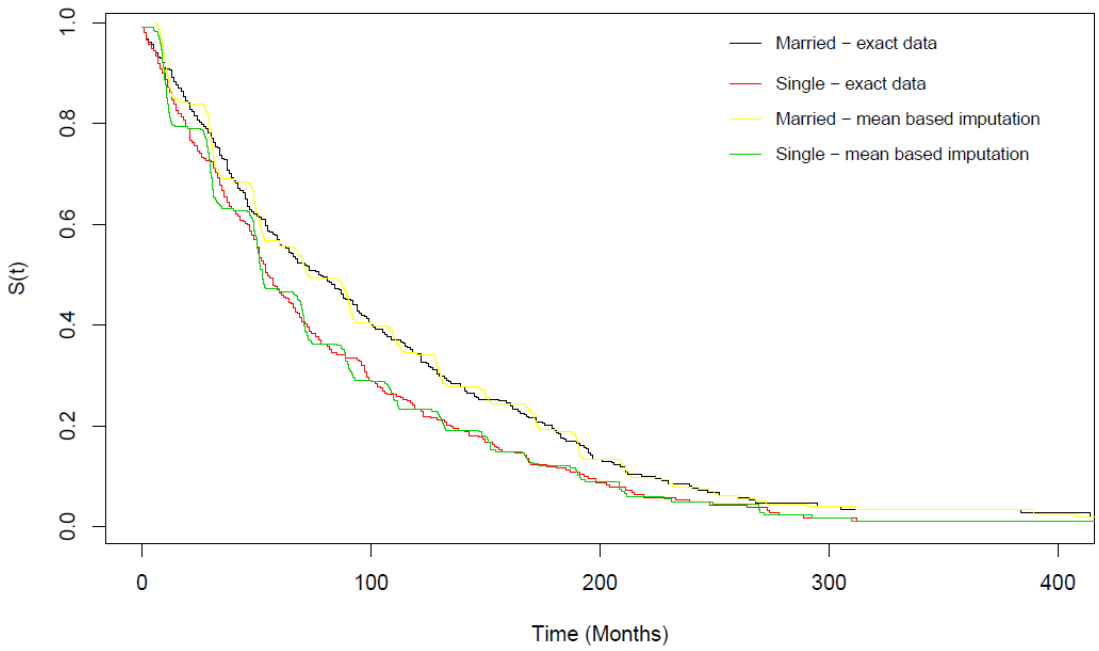

Figure 4.25: The survival function obtained by mean point with 0% exact data for social status variable (married and single)

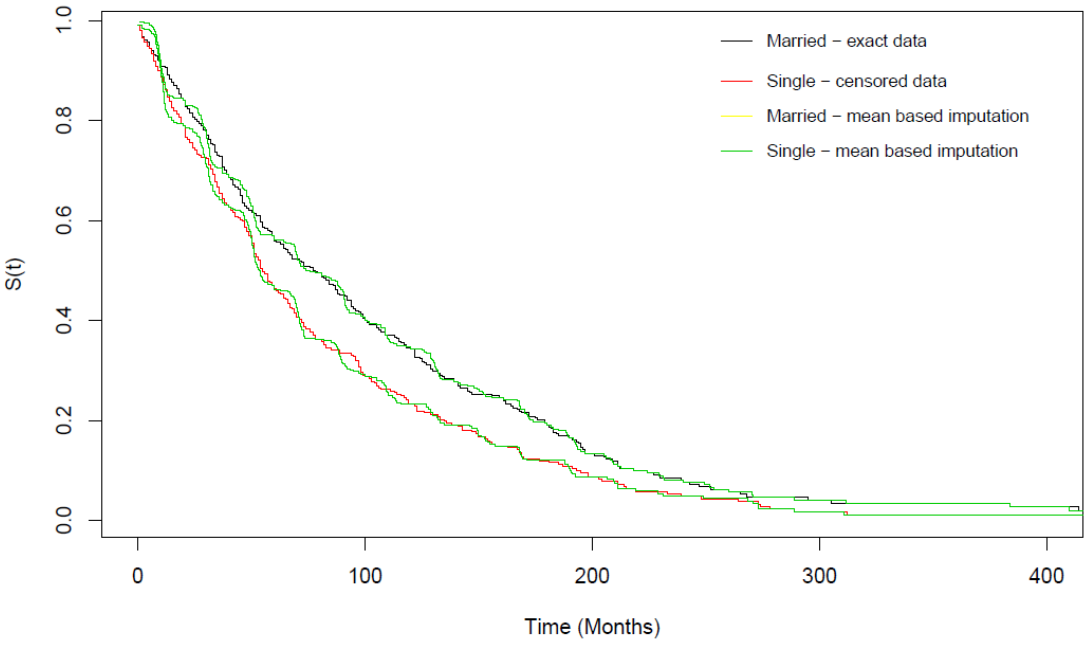

Figure 4.26: The survival function obtained by mean point with 25% exact data for social status variable (married and single)

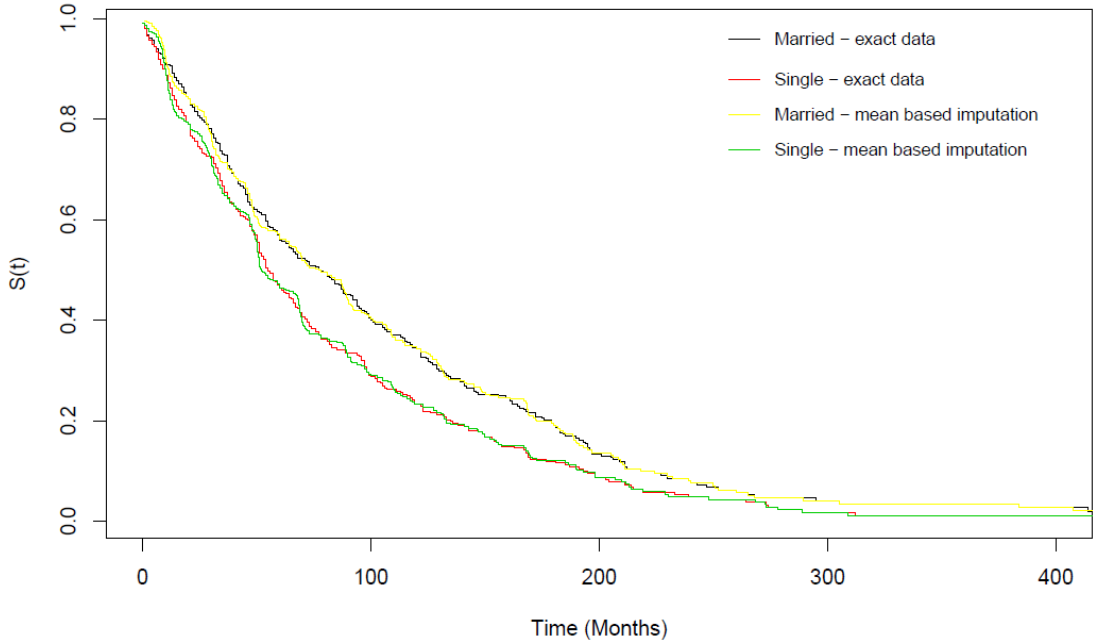

Figure 4.27: The survival function obtained by mean point with 50% exact data for social status variable (married and single)

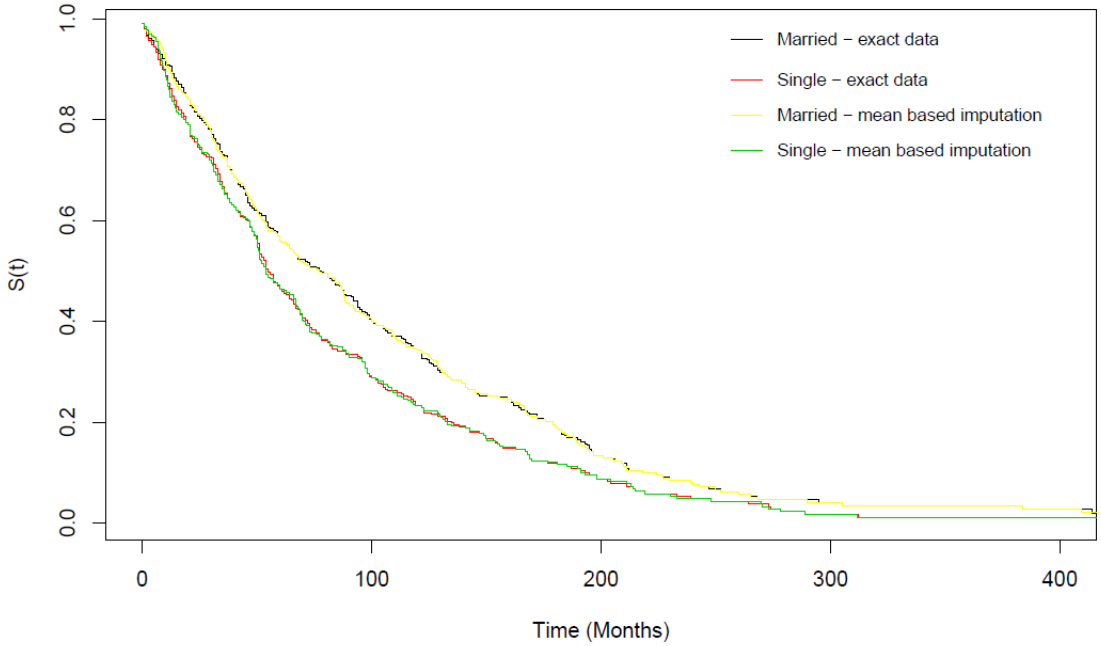

Figure 4.28: The survival function obtained by mean point with 75% exact data for social status variable (married and single)

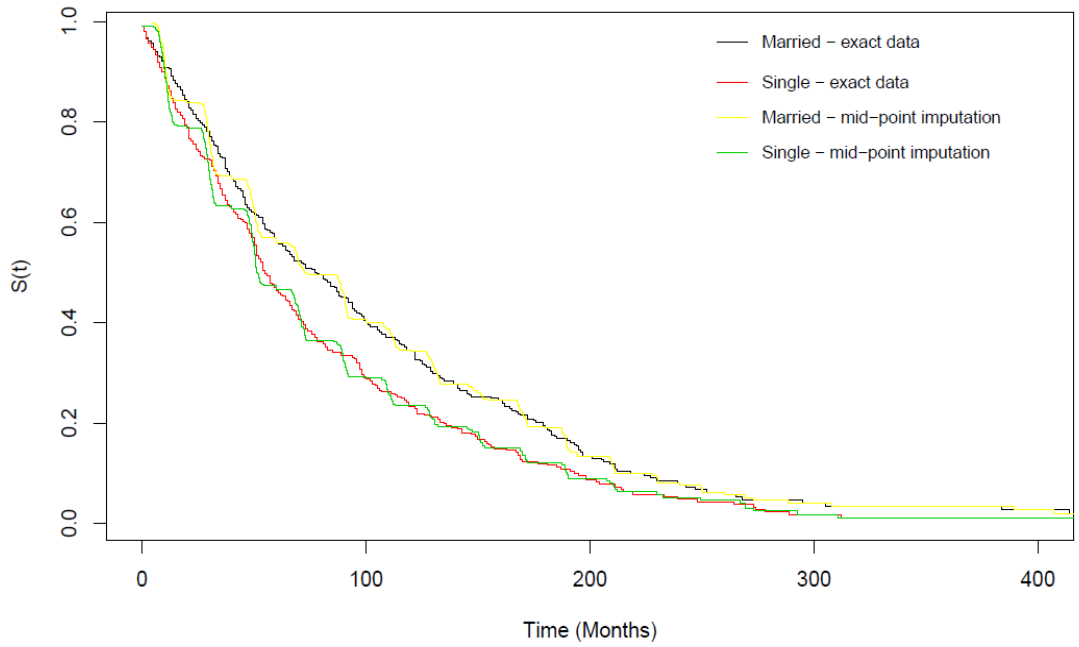

Figure 4.29: The survival function obtained by median point with 0% exact data for social status variable (married and single)

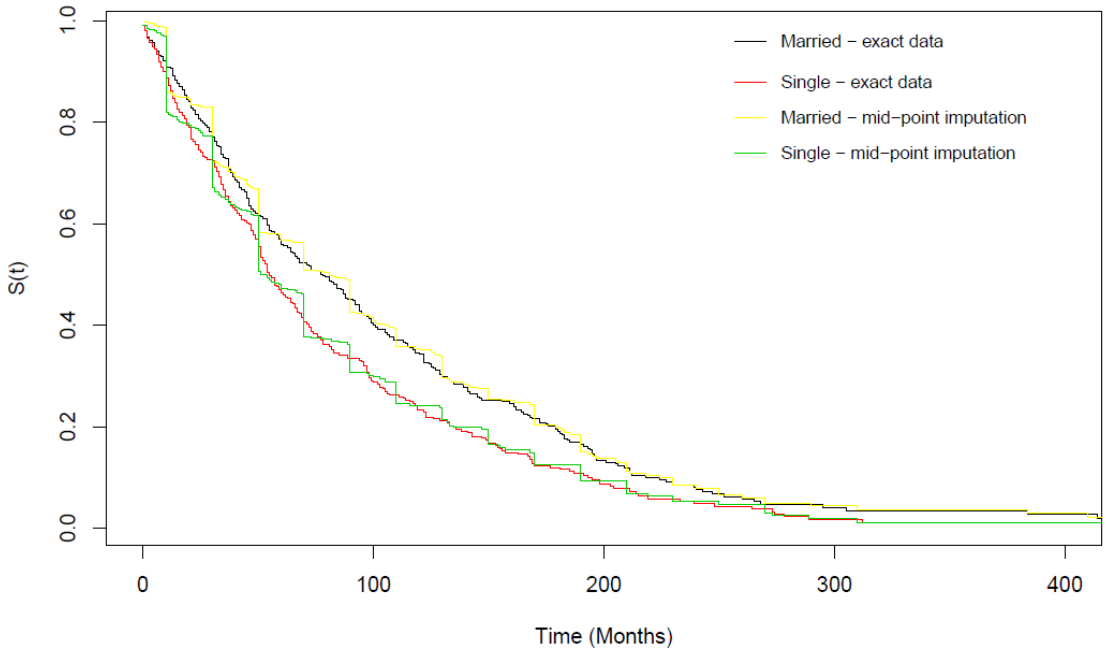

Figure 4.30: The survival function obtained by median point with 25% exact data for social status variable (married and single)

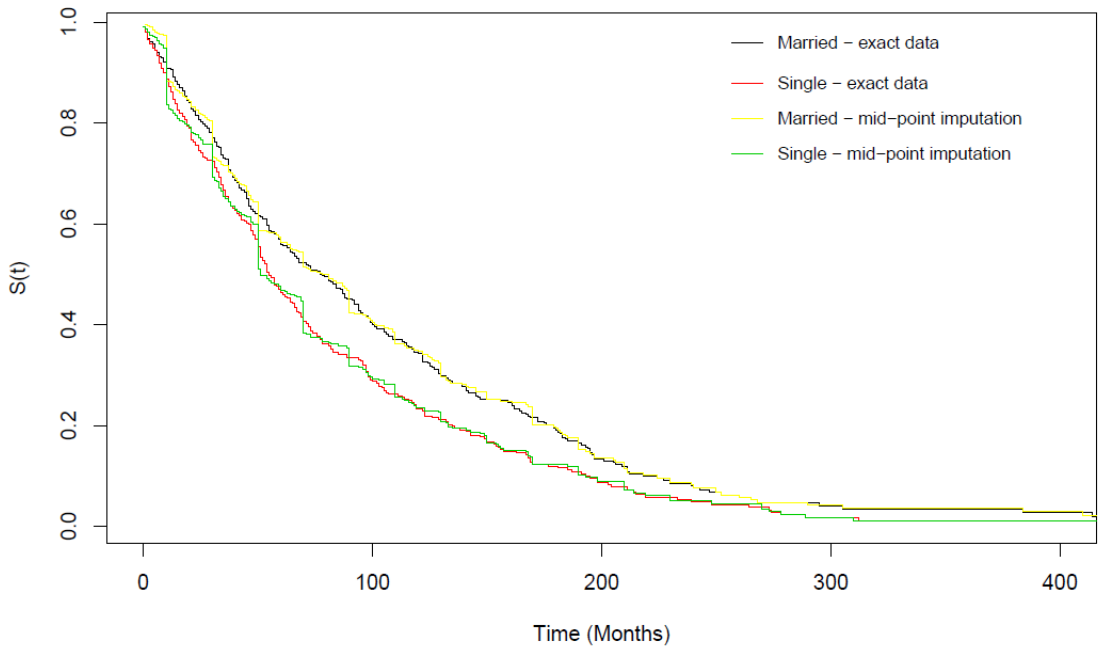

Figure 4.31: The survival function obtained by median point with 50% exact data for social status variable (married and single)

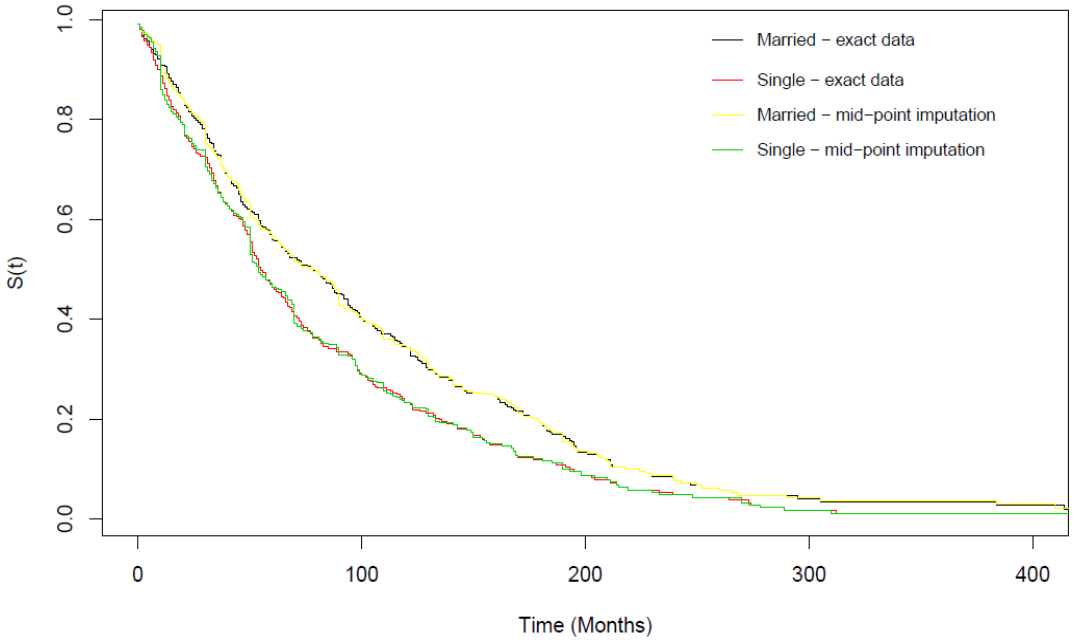

Figure 4.32: The survival function obtained by median point with 75% exact data for social status variable (married and single)

Table 4.4: Result from simulation data for nationality variable based on Cox Model

	Imputation	Coefficient	Exp(Coef)	SE	P-value	LRT*
0%Exact	Left-point	-0.16212	0.85034	0.07282	0.0260	4.95(0.02607)
	Median	-0.16175	0.85065	0.07282	0.0263	4.93(0.02641)
	Mean	-0.17063	0.84313	0.07296	0.0194	5.46(0.0194)
25%Exact	Left-point	-0.16733	0.84592	0.07283	0.0216	5.27(0.02165)
	Median	-0.17175	0.84219	0.07284	0.0184	5.55(0.0184)
	Mean	-0.1773	0.8375	0.0729	0.015	5.91(0.01507)
50%Exact	Left-point	-0.16853	0.84491	0.07282	0.0206	5.35(0.01533)
	Median	-0.17493	0.83951	0.07284	0.0163	5.76(0.01638)
	Mean	-0.17679	0.83795	0.07288	0.0153	5.88(0.01533)
75%Exact	Left-point	-0.17536	0.83916	0.07284	0.0161	5.79(0.01613)
	Median	-0.18028	0.83504	0.07286	0.0134	6.11(0.01341)
	Mean	-0.17922	0.83592	0.07288	0.0139	6.04(0.01398)

Table 4.4 showed that the results from simulation study for nationality variable (Gulf Cooperation Council and non-Gulf Cooperation Council). These results confirm the result obtained earlier, which we reject the null hypothesis for nationality based on PIC when the exact value more than 25% there is a different between the two groups compare to when the exact value below than 25% there is slightly different.

Thirdly, for the covariate nationality (Gulf and others) we generate the data based on the mean and standard deviation as -0.1785 and 0.07285 via the exact observation with 0%, 25%, 50%, and 75% in the PIC data.

Figures 4.33, 4.34, 4.35, 4.36, 4.37, 4.38, 4.39, 4.40, 4.41, 4.42, 4.43 and 4.44 showed the result of the estimation of survival function obtained by Cox proportional hazard model (exact observation-Cox) compared and imputation techniques that is; left point, mean, and median. The Figures look almost similar in case of the one obtained by mean and median, but little difference compared with one obtained by left point.

Based on Figures above mentioned, there is little difference in survival function between the two types of failures (Gulf and others). However, the mean and median are better estimate of the survival function based on the similarity in the graph and likelihood ratio with their P-value (Table 4.4). Based on the ratio between Gulf and

others the left imputation showed to be better in term of the P-value.

The findings above-mentioned correspond to exactly observation with the findings obtained by Cox's model on the same data set. Their findings showed the estimates of the survival function to be very similar, with the survival function that obtained mean and median imputation techniques. On other hand, the left point showed different results compare with exact data for all different percentages from the exact data based on PIC.

Based on the two types of failures the Gulf and others, their results look similar but there is slightly different in the begging from 10 to 90 months and also from 95 to 325 months, but lately the Gulf's look have longer survival compare to others, which indicate that the Gulfs may stayed longer in prison compared to others.

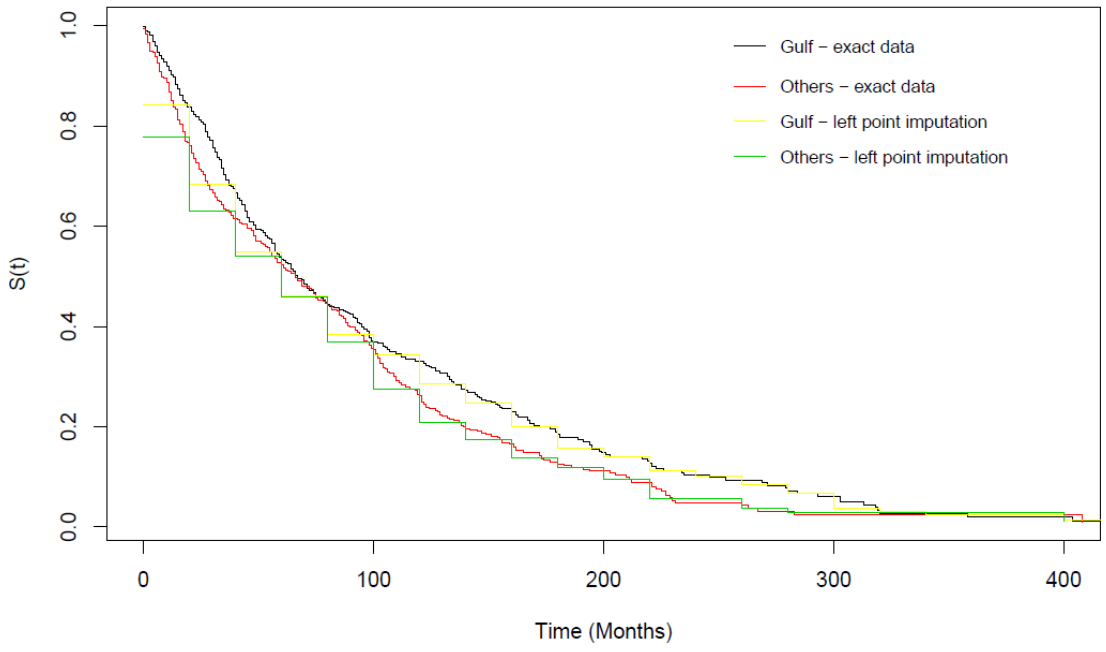

Figure 4.33: The survival function obtained by left point with 0% exact data for nationality covariate (Gulf and others)

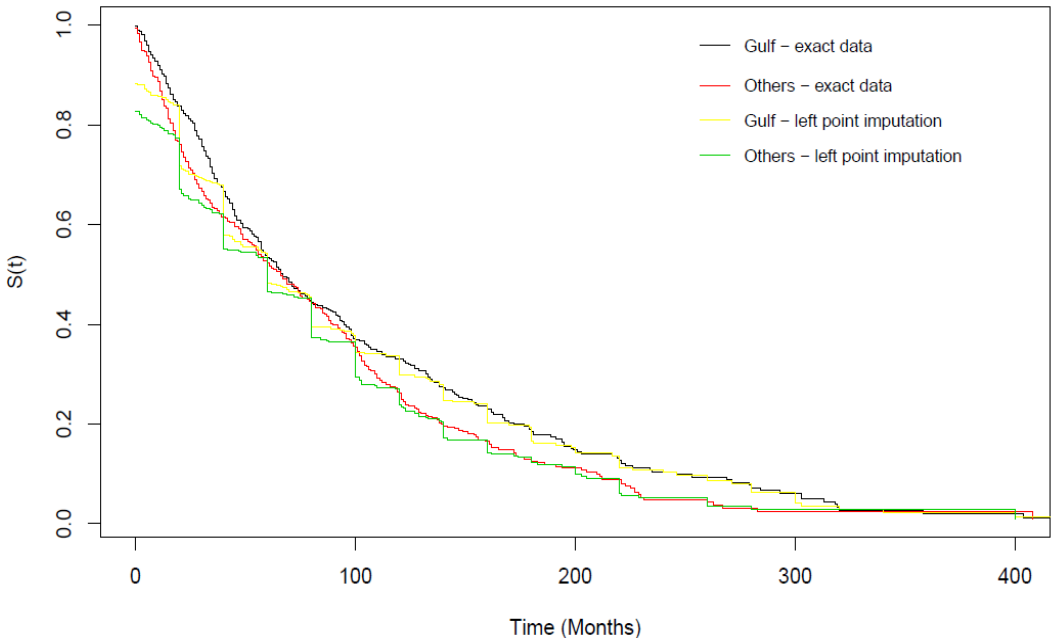

Figure 4.34: The survival function obtained by left point with 25% exact data for nationality covariate (Gulf and others)

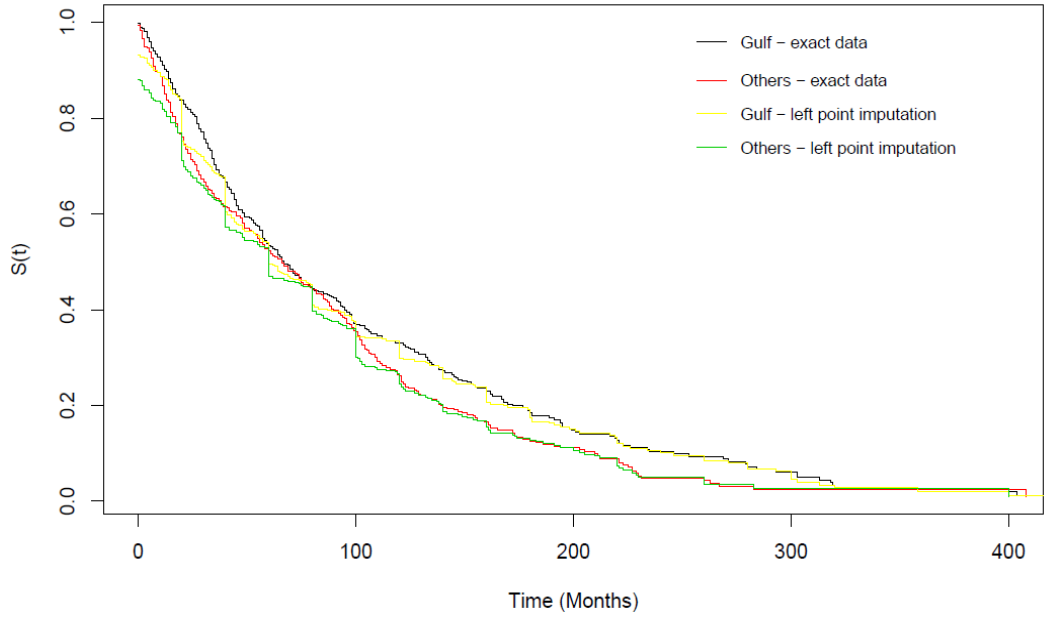

Figure 4.35: The survival function obtained by left point with 50% exact data for nationality covariate (Gulf and others)

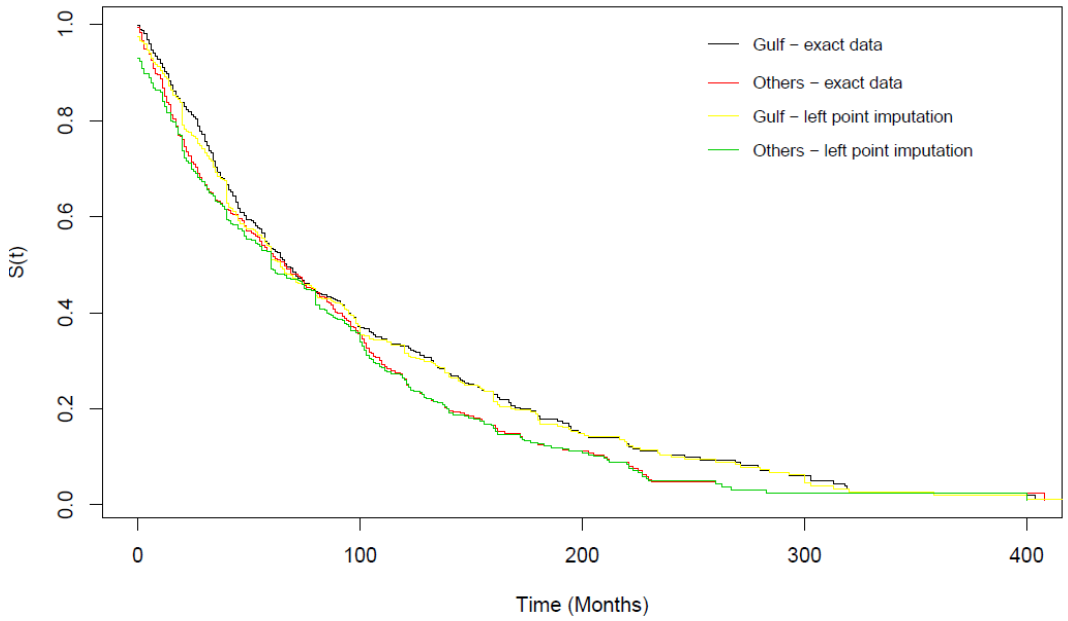

Figure 4.36: The survival function obtained by left point with 75% exact data for nationality covariate (Gulf and others)

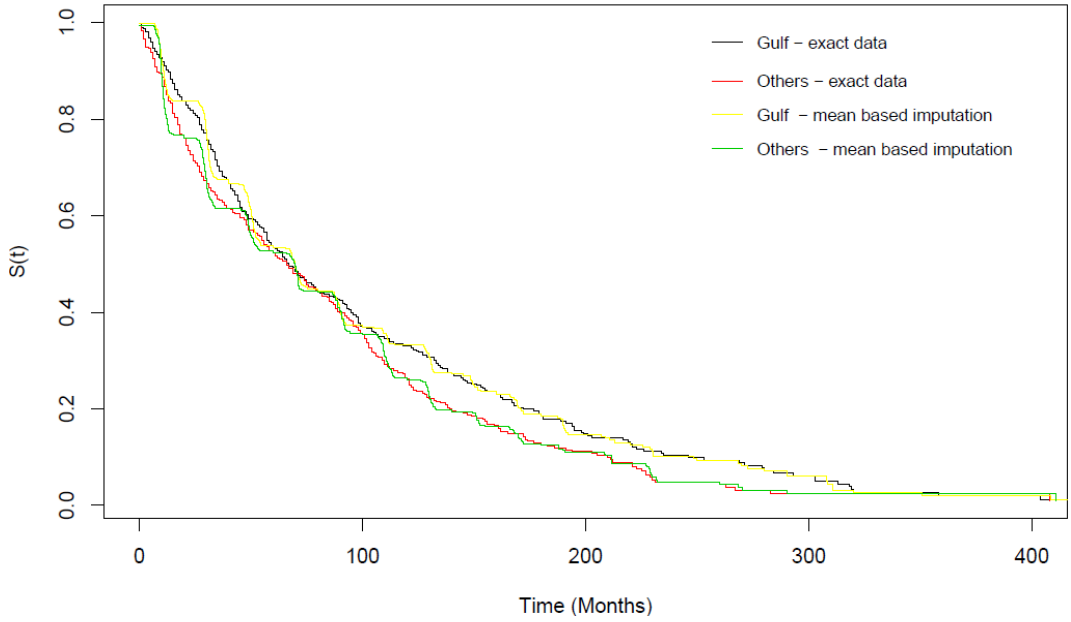

Figure 4.37: The survival function obtained by mean point with 0% exact data for nationality covariate (Gulf and others)

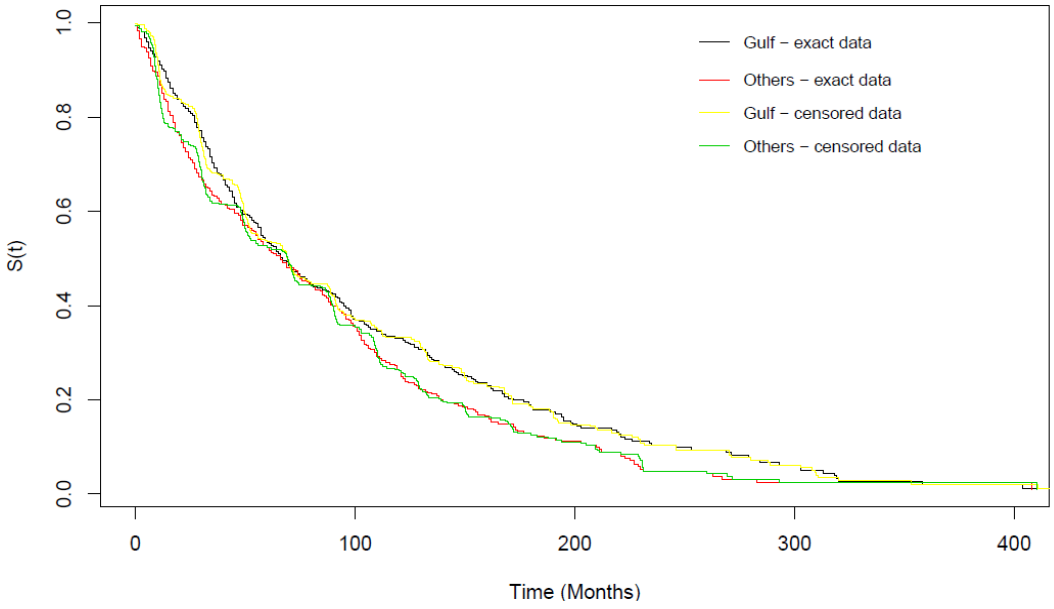

Figure 4.38: The survival function obtained by mean point with 25% exact data for nationality covariate (Gulf and others)

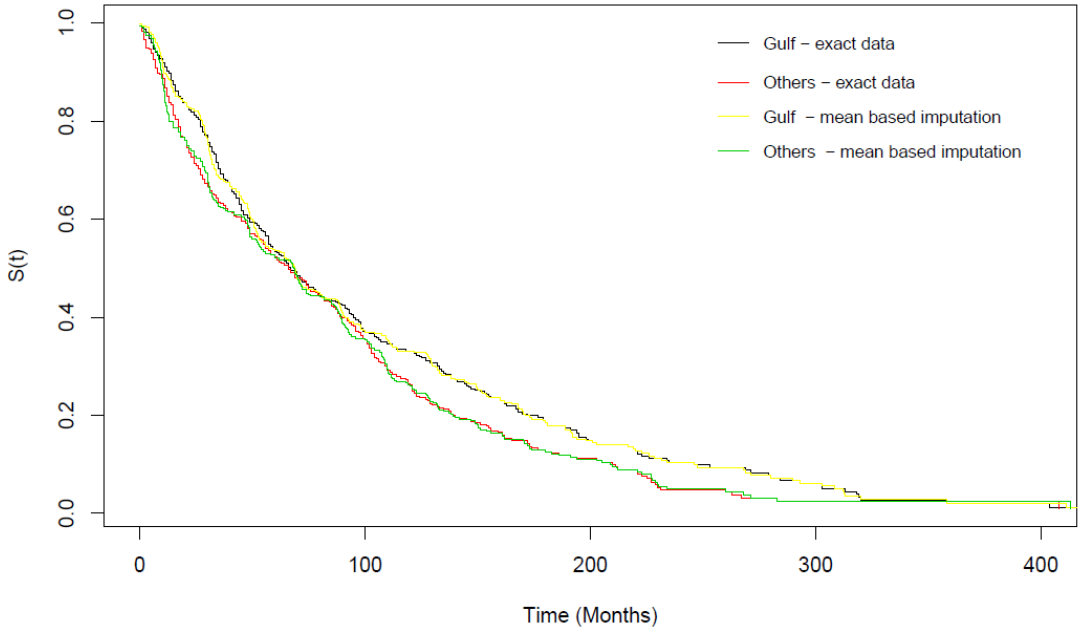

Figure 4.39: The survival function obtained by mean point with 50% exact data for nationality covariate (Gulf and others)

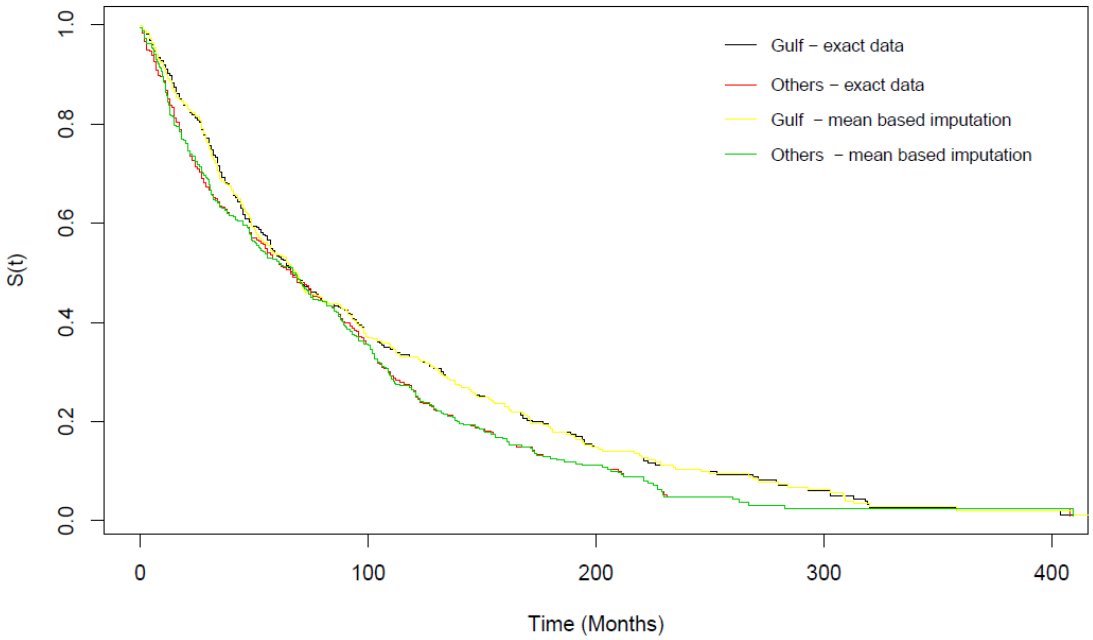

Figure 4.40: The survival function obtained by mean point with 75% exact data for nationality covariate (Gulf and others)

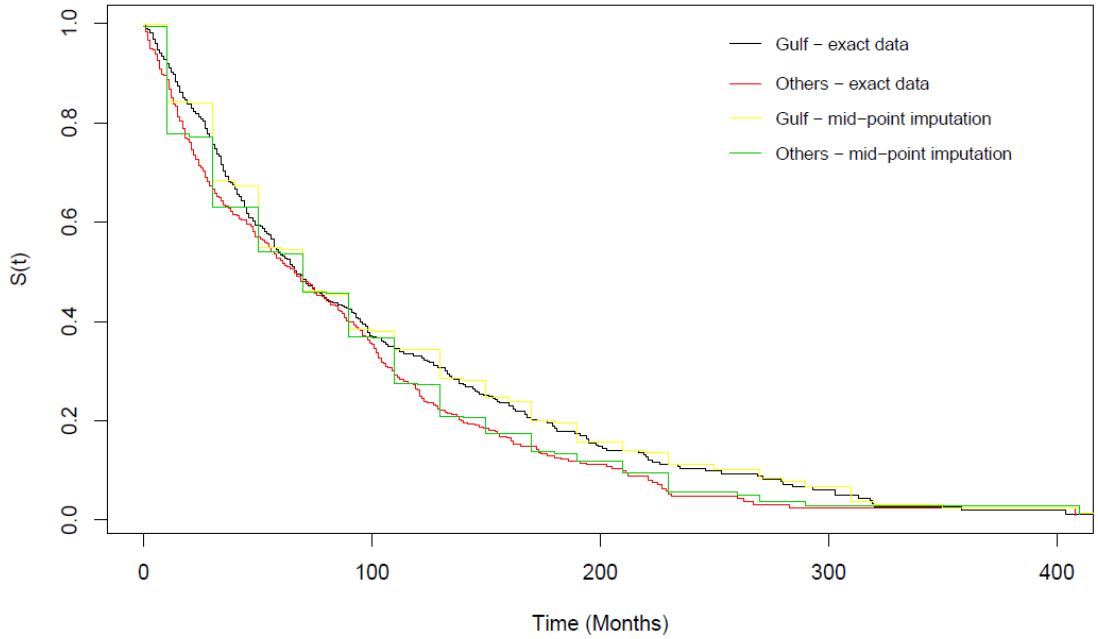

Figure 4.41: The survival function obtained by median point with 0% exact data for nationality covariate (Gulf and others)

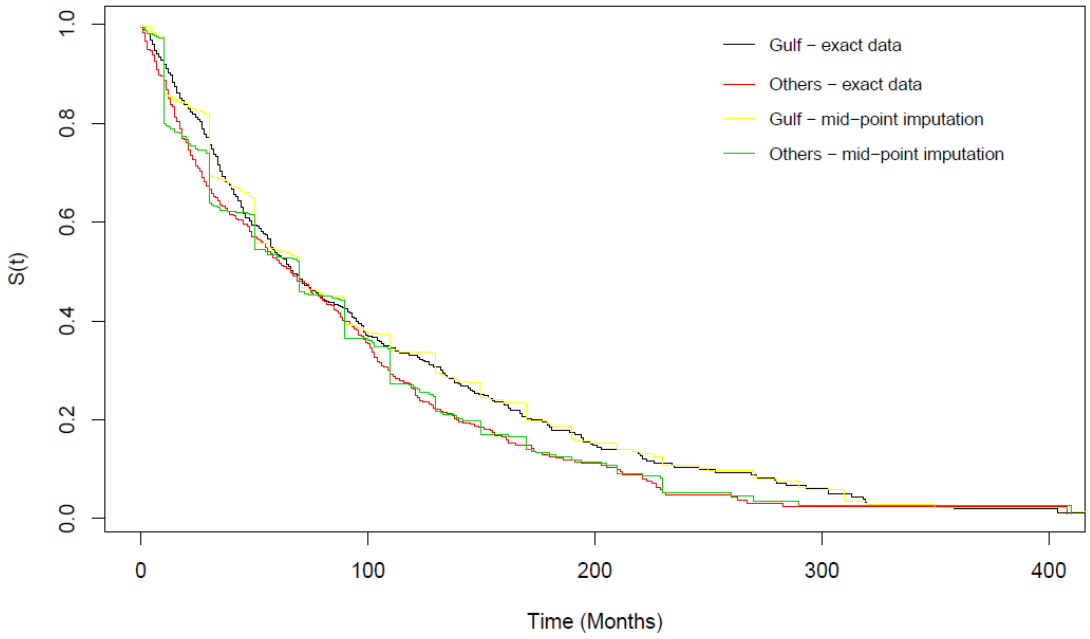

Figure 4.42: The survival function obtained by median point with 25% exact data for nationality covariate (Gulf and others)

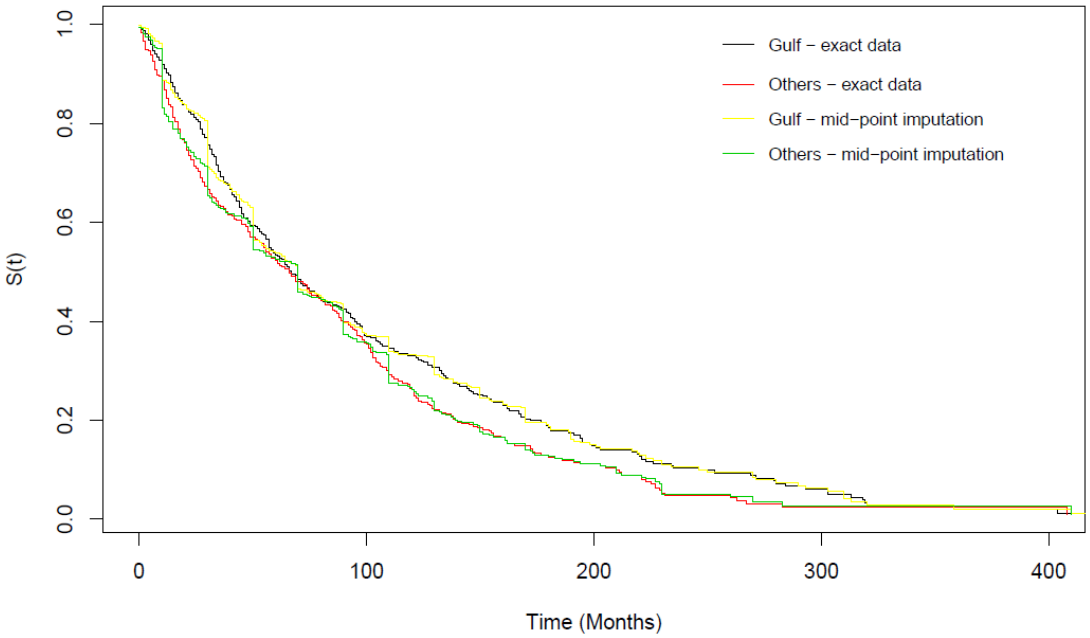

Figure 4.43: The survival function obtained by median point with 50% exact data for nationality covariate (Gulf and others)

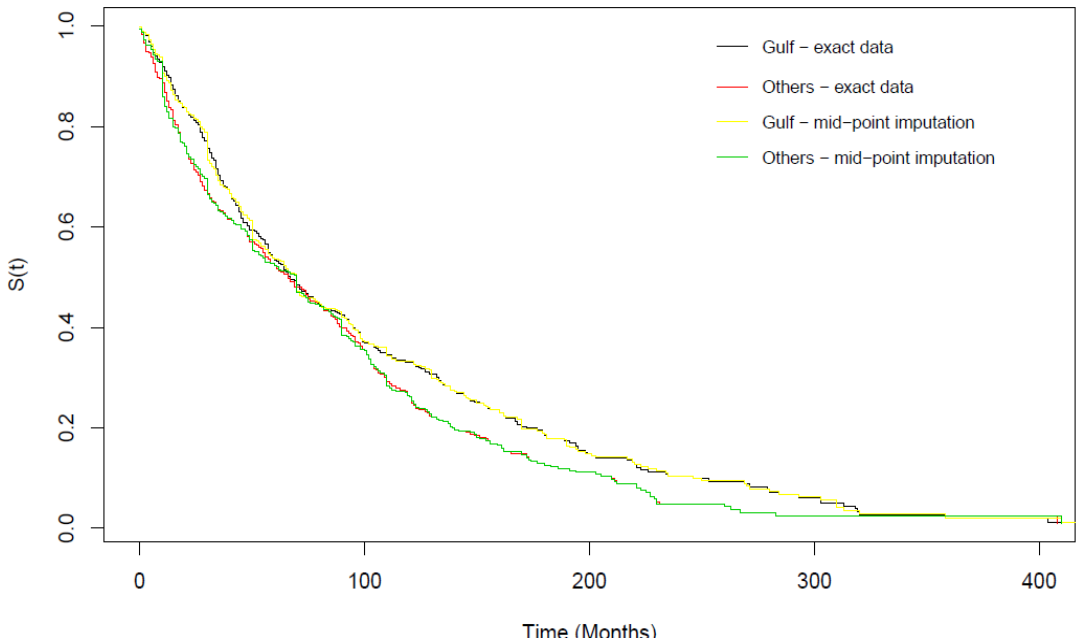

Figure 4.44: The survival function obtained by median point with 75% exact data for nationality covariate (Gulf and others)

In summary, we fit the simulation data based on the Cox model for the two failure rates for different covariates that is age, social status, and nationality (we don't used gender as one of covariates due to the only 5% of the original data set from female). The Figures 4.1 to 4.44 showed the failure time against the survival function for the two types of failures rates obtained by exact observation Cox compare to the one obtained by imputation techniques that is; left point, mean point and median. In all the figures the survival function curve fell between the two confident intervals, and also, these Figures substantiate the non-significant effect of the two failures of rates. Although there are small differences between the two failures rates, especially to the one obtained by left imputations. Clearly, the rest of the Figures showed very similar predicted of survival function patterns. However, the likelihood ratio test shows significant results in our imputation techniques which indicates that these techniques can easy to be use for social data set as well as simulation data.

CHAPTER 5: CONCLUSION AND SUGGESTIONS FOR FURTHER RESEARCH

This chapter present the conclusion for which summarizes the results obtained in the previous chapters, and suggestions for future studies are presented later in the second section.

5.1 Conclusion

The primary purpose of the study in this thesis is to look into the study of the Cox proportional hazard regression model based on imputation techniques for prison PIC data. This method will be compared for the different imputation techniques with different percentages of exact data based on simulation study as well as the covariates in the model. In additional to that, left, mean, and median imputations based on the Cox's PHR approach utilizing the estimate of the survival function.

In this thesis, the maximum likelihood estimation based on Newton-Raphson method was used to obtain the survival function estimates, and comparisons were made with existing one under the assumption of Cox's PHRM (chapter three).

The partly interval-censored for prison data and simulation data was found preferable compared to interval censored data (0% exact), because the likelihood for PIC data for Cox's model has a much simpler form than the likelihood corresponding to the Cox regression hazard model with censored data. Moreover, the maximum likelihood estimates with Newton-Raphson does not always require the inversion of large matrices of large values. Furthermore, the other methods (such as EM algorithm) can become overwhelming when the number of subjects is large, and get worse when there are multiple random coefficients for each subject, this result it similar to the one founded by Zyoud et al., (2016).

To analyze survival data, based on imputation techniques with partly interval

censored data, at least one failure rate must be present. An example of this is failure rates for the marriage & single (social status), Gulf & others (nationality), male & female (gender), 30 years or older & younger than 30 years (age) data, which was used in this thesis. In this data sets two failure rates were identified that is failure rate marriage and single for example when age is used as covariate. For simulation data, the failure times were generated via the prison failure data set. According to the survival study, we should have one of the failures to be at least longer survival compare to other failure, so in a case of social status for example, the study found that the single and marriage are similar at the begging of the survival curve. However, later we conform that the single failure has loner survival which indicate that single is more active in crime and stay loner at prison compare to marriage. Moreover, the survival method used in this thesis was found to be acceptable and easy to implement for PIC data based on application of social data set.

Based on chapter Four, the left point, mean point and median point imputations based on Cox PHRM are discussed. Data used to these methods need be modified (depending on data characteristics and the researcher's needs) as PIC and interval data. In comparison between the imputation techniques, the median point & mean point found to be reasonable in term of survival function estimation, P-value and likelihood ratio test (LRT).

Survival function and LRT with their P-value for the two type of failure rates were calculated from the prison data set and simulation data using full iteration of Newton-Raphson. It was discovered that the censored observations from simulation have influence on the model and should be studied and taken account of for further research. Nonparametric model shows better result and can easy implement base on partly interval censored data via the imputation techniques compared with interval

censored.

From the practical applications of this research finding significant in age and social status, because the younger prisoners (less than 30 years) commit more crimes compared to prisoners with age more than 30 and the single prisoners have more crime compare to married prisoners. However, from this finding we recommend to the social institutions to focus on these groups of members of society to make various programs and awareness campaigns aimed at defining these groups to awareness and the risks of crime and spreading the culture of planning a good life.

Finally, R software were used as procedures to obtain the results. As explained in earlier chapters in this thesis, R software is capable of doing calculation involving large matrix sizes. However, the program codes were built using R software for our model via the imputation techniques as showed in Appendix A.

5.2 Suggestions for Further Research

The results obtained from simulation data and the real data set in this thesis showed that simple imputation methods via proportional hazard regression model is easy to implement and preferable. Likewise, the results showed that the mean and median is better than left imputation in the computation of the estimate of the survival function measures. However, more work needs to be done for left point imputation and others imputation methods such as random, midpoint and multiple imputation as well as different lengths of interval. One of the most obvious is to try the procedure of a Markov Chain Monte Carlo EM algorithm so as to achieve more precise and unbiased estimates.

The data used in this thesis contain only four variables as age, gender, social status and nationality. More variables are required such as education level, family status, psychology status, and previous crime.

REFERENCES

Abbas, S. A., Subramanian, S., Ravi, P., Ramamoorthy, S., & Munikrishnan, V. (2019). An Introduction to Survival Analytics, Types, and Its Applications. *Biomechanics*, 33.

Alharpy, A. M., & Ibrahim, N. A. (2013). Parametric Tests for Partly Interval-censored Failure Time Data under Weibull Distribution via Multiple Imputation. *Journal of Applied Sciences, 13(4), 621–626.*

Ahmedi, S., Elfaki, F. A. M., Lukman, I., & Kabbashi, N. A. (2020). Cox's Model for Prison Partly Interval Censored Data. *Journal of Physics: Conference Series, 1489*, 012032. doi: 10.1088/1742-6596/1489/1/012032.

Allison, P. D. (2014). *Event history and survival analysis*. Los Angeles: SAGE.

Anderson, P. K. and R. D. Gill. (1982). "Cox's Regression Model for Counting Processes: A Larg Sample Study." *The Annals of Statistics 10(4):1100–1120*

Benda, B. B. (2003). Survival analysis of criminal recidivism of boot camp graduates using elements from general and developmental explanatory models. *International Journal of Offender Therapy and Comparative Criminology*, *47*(1), 89-110.

Benda, B. B., Harm, N. J., & Toombs, N. J. (2005). Survival analysis of recidivism of male and female boot camp graduates using life-course theory. *Journal of Offender Rehabilitation*, *40*(3-4), 87-113.

Benda, B. B., Toombs, N. J., & Peacock, M. (2002). Ecological factors in recidivism: A survival analysis of boot camp graduates after three years. *Journal of*

Offender Rehabilitation, 35(1), 63-85.

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine, 24*(11), 1713-1723.

Binder, D. A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika, 79*(1), 139-147.

Breslow, N. (1974). Covariance Analysis of Censored Survival Data. *Biometrics, 30(1), 89-99. doi:10.2307/2529620*

Brostrom, G (2012). *Event History Analysis with R. CRC Press.*

Cloyes, K. G., Wong, B., Latimer, S., & Abarca, J. (2010). Time to prison return for offenders with serious mental illness released from prison: A survival analysis. *Criminal Justice and Behavior, 37*(2), 175-187.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological), 34*(2), 187-202.

Cox, D. R. (1975). "Partial likelihood". Biometrika. 62: 269-276.

Cox, D.R. and Oakes D.(1984). *Analysis Of Survival Data* Chapman and Hall, London.

Elfaki, F. A. M., Abobakar, A., Azram, M., & Usman, M. (2013). Survival Model for Partly Interval-Censored Data with Application to Anti D in Rhesus D Negative Studies. *World Academy of Science, Engineering and Technology, International Journal of Biological, Biomolecular, Agricultural, Food and Biotechnological Engineering*, *7*(5), 347-350.

Emmert-Streib, F., & Dehmer, M. (2019). Introduction to Survival Analysis in

Practice. *Machine Learning and Knowledge Extraction*, *1*(3), 1013-1038.

Fauzi, N. A. M., Elfaki, F. A. M., & Ali, Y. (2015). Some Method On Survival Analysis Via Weibull Model In the Present of Partly Interval Censored: A Short Review. *International Journal of Computer Science and Network Security (IJCSNS)*, *15*(4), 48.

Fan, J., & Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, *30*(1), 74-99.

Fox, J. (2002). Cox proportional-hazards regression for survival data. *An R and S-PLUS companion to applied regression*, *2002*.

George, B., Seals, S., & Aban, I. (2014). Survival analysis and regression models. *Journal of Nuclear Cardiology*, *21*(4), 686-694.

Guure, C. B., Ibrahim, N. A., & Adam, M. B. (2006). On partly censored data with the Weibull distribution.

Hill, A., Habermann, N., Klusmann, D., Berner, W., & Briken, P. (2008). Criminal recidivism in sexual homicide perpetrators. *International Journal of Offender Therapy and Comparative Criminology*, *52*(1), 5-20.

Jung, H., Spjeldnes, S., & Yamatani, H. (2010). Recidivism and survival time: Racial disparity among jail ex-inmates. *Social Work Research*, *34*(3), 181-189.

Kim, J. S. (2003) "Maximum likelihood estimation for the proportional hazards model with party interval-censored data," *J R. Statist. Soc., Series B65, pp. 489-502.*

Kumar, D., & Klefsjö, B. (1994). Proportional hazards model: a review. *Reliability*

Engineering & System Safety, *44*(2), 177-188.

Lane, W. R., Looney, S. W., & Wansley, J. W. (1986). An application of the Cox proportional hazards model to bank failure. *Journal of Banking & Finance*, *10*(4), 511-531.

Lee, Elisa T., and John Wang. (2003) Statistical methods for survival data analysis. *Vol. 476. John Wiley & Sons.*

Liu, L. (2017). *Heart Failure: Epidemiology and research methods*. Elsevier Health Sciences.

Mackie, J., Groves, K., Hoyle, A., Garcia, C., Garcia, R., Gunson, B., & Neuberger, J. (2001). Orthotopic liver transplantation for alcoholic liver disease: a retrospective analysis of survival, recidivism, and risk factors predisposing to recidivism. *Liver Transplantation*, *7*(5), 418-427.

Nesi, C. N., Shimakura, S. E., Ribeiro Junior, P. J., & Mio, L. L. M. D. (2015). Survival analysis: a tool in the study of post-harvest diseases in peaches. *Revista Ceres*, *62*(1), 52-61.

Ostermann, M. (2015). How do former inmates perform in the community? A survival analysis of rearrests, reconvictions, and technical parole violations. *Crime & Delinquency*, *61*(2), 163-187.

Peto, R., & Peto, J. (1972). Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society: Series A (General)*, *135*(2), 185-198.

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata*. College Station, TX: Stata Press Publication.

Rainforth, M. V., Alexander, C. N., & Cavanaugh, K. L. (2003). Effects of the transcendental meditation program on recidivism among former inmates of Folsom Prison: Survival analysis of 15-year follow-up data. *Journal of Offender Rehabilitation*, *36*(1-4), 181-203.

Royston, P., & Lambert, P. C. (2011). Flexible parametric survival analysis using Stata: beyond the Cox model.

Rubinstein, R. (1981) S*imulation and the Mote Carlo Method*. New York: Wiley.

Schober, P., & Vetter, T. R. (2018). Survival analysis and interpretation of time-to-event data: The tortoise and the hare. *Anesthesia and analgesia*, *127*(3), 792.

Selvin, S. (2004). *Statistical analysis of epidemiologic data*. Estados Unidos: Oxford University Press.

Teachman, J. D. (1983). Analyzing social processes: Life tables and proportional hazards models. *Social Science Research, 12(3), 263–301. doi: 10.1016/0049-089x(83)90015-7*

Tripodi, S. J., Kim, J. S., & Bender, K. (2010). Is employment associated with reduced recidivism? The complex relationship between employment and crime. *International Journal of Offender Therapy and Comparative Criminology*, *54*(5), 706-720.

Tsiatis, A. (1981). "*A Large Sample Study of Cox's Regression Model.*" The

Annals of Statistic. 9: 93-108.

Vaida, F., & Xu, R. (2000). Proportional hazards model with random effects. *Statistics in medicine*, *19*(24), 3309-3324.

Wu, Y., Chambers, C. D., & Xu, R. (2019). Semiparametric sieve maximum likelihood estimation under cure model with partly interval censored and left truncated data for application to spontaneous abortion. *Lifetime data analysis*, *25*(3), 507-528.

Zhang, W., Ota, T., Shridhar, V., Chien, J., Wu, B., & Kuang, R. (2013). Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. *PLoS computational biology*, *9*(3), e1002975.

Zhao, X. Zhao, and Q.J. Sun (2008). "Generalized log-rank test for partly interval-censored failure time data," *Biometrical Journal, Vol. 3, pp. 375-385.*

Zyoud, A., Elfaki, F. A., & Hrairi, M. (2016). Nonparametric estimate based in imputations technique for interval and partly interval censored data. *Science International (Lahore)*, *28*(2), 879-884.

APPENDIX A

SAMPLES OF PROGRAM CODE IN R

Left Imputations for Age with 0% exact.

```
require(survival)
dat <- read.table('ASimW100.txt',header=T)
dat1 <- dat[dat$Age==1,]
dat2 <- dat[dat$Age==0,]
cxt1=coxph(Surv(dat1$eve,dat1$cens)~Age,data=dat1)
ek1 <-survfit(cxt1)
pdf(file="D:/code/Age/left-00.pdf", width = 9, height =6)
plot(ek1$time,ek1$surv,type="s",col=1,lty=1,
xlab="Time (Months)",ylab="S(t)",xlim=range(c(0,400)))
cxt2=coxph(Surv(dat2$eve,dat2$cens)~Age,data=dat2)
ek2<-survfit(cxt2)
lines(ek2$time,ek2$surv,type="s",col=2,lty=1)
legend(240,1,lty=1,col=1, "30 years or older - exact data", bty="n",cex=0.8)
legend(240,0.92,lty=1,col=2, "Younger than 30 - exact data", bty="n",cex=0.8)
data <- read.table('ASimWint100.txt',header=T)
dat3 <- data[data$Age==1,]
dat4 <- data[data$Age==0,]
pm <- dat3$left
cxt3=coxph(Surv(pm,dat3$cens)~Age,data=dat3)
ek3 <-survfit(cxt3)
lines(ek3$time,ek3$surv,type="s",col=6,lty=3)
```

```
#plot(ek1$time,ek1$surv,type="s",col=1,lty=1,
xlab='Time  (Months)",ylab="S(t)",xlim=range(c(0,530)))
pm2 <- dat4$left
cxt4=coxph(Surv(pm2,dat4$cens)~Age,data=dat4)
ek4<-survfit(cxt4)
lines(ek4$time,ek4$surv,type="s",col=4,lty=3)
legend(240,0.85,lty=3,col=6, "30 years or older - left point imputation",
bty="n",cex=0.8)
legend(240,0.78,lty=3,col=4, "Younger than 30 - left point imputation",
bty="n",cex=0.8)
dev.off()
```

Median Imputations for Age with 0% exact.

```
require(survival)
dat <- read.table('ASimW100.txt',header=T)
dat1 <- dat[dat$Age==1,]
dat2 <- dat[dat$Age==0,]
cxt1=coxph(Surv(dat1$eve,dat1$cens)~Age,data=dat1)
ek1 <-survfit(cxt1)
pdf(file="D:/code/Age/mid-00.pdf",  width = 9, height =6)
plot(ek1$time,ek1$surv,type="s",col=1,lty=1,
xlab='Time  (Months)",ylab="S(t)",xlim=range(c(0,400)))
cxt2=coxph(Surv(dat2$eve,dat2$cens)~Age,data=dat2)
ek2<-survfit(cxt2)
lines(ek2$time,ek2$surv,type="s",col=2,lty=1)
```

```
legend(240,1,lty=1,col=1, "30 years or older - exact data", bty="n",cex=0.8)
legend(240,0.92,lty=1,col=2, "Younger than 30 - exact data", bty="n",cex=0.8)
data <- read.table('ASimWint100.txt',header=T)
dat3 <- data[data$Age==1,]
dat4 <- data[data$Age==0,]
p <-1:nrow(dat3)
for (i in 1:nrow(dat3)){p[i] <- mean(runif(10,dat3$left[i],dat3$right[i]))}
pm <-ifelse(is.finite(p),p,dat3$left)
cxt3=coxph(Surv(pm,dat3$cens)~Age,data=dat3)
ek3 <-survfit(cxt3)
lines(ek3$time,ek3$surv,type="s",col=6,lty=3)
#plot(ek1$time,ek1$surv,type="s",col=1,lty=1,
xlab="Time (Months)",ylab="S(t)",xlim=range(c(0,530)))
p <-1:nrow(dat4)
for (i in 1:nrow(dat4)){p[i] <- mean(runif(10,dat4$left[i],dat4$right[i]))}
pm2 <-ifelse(is.finite(p),p,dat4$left)
cxt4=coxph(Surv(pm2,dat4$cens)~Age,data=dat4)
ek4<-survfit(cxt4)
lines(ek4$time,ek4$surv,type="s",col=4,lty=3)
legend(240,0.85,lty=3,col=6, "30 years or older - mid-point imputation",
bty="n",cex=0.8)
legend(240,0.78,lty=3,col=4, "Younger than 30 - mid-point imputation",
bty="n",cex=0.8)
dev.off()
```

Mean Imputations for social status with 50% exact.

```
require(survival)
dat <- read.table('SSimW100.txt',header=T)
dat1 <- dat[dat$Status==1,]
dat2 <- dat[dat$Status==0,]
cxt1=coxph(Surv(dat1$eve,dat1$cens)~Status,data=dat1)
ek1 <-survfit(cxt1)
pdf(file="D:/code/marital status/mean-50.pdf", width = 9, height =6)
plot(ek1$time,ek1$surv,type="s",col=1,lty=1,
xlab="Time (Months)",ylab="S(t)",xlim=range(c(0,400)))
cxt2=coxph(Surv(dat2$eve,dat2$cens)~Status,data=dat2)
ek2<-survfit(cxt2)
lines(ek2$time,ek2$surv,type="s",col=2,lty=1)
legend(250,1,lty=1,col=1, "Married - exact data", bty="n",cex=0.8)
legend(250,0.92,lty=1,col=2, "Single - exact data", bty="n",cex=0.8)
data <- read.table('SSimWint50.txt',header=T)
dat3 <- data[data$Status==1,]
dat4 <- data[data$Status==0,]
p <-1:nrow(dat3)
for (i in 1:nrow(dat3)){p[i] <- mean(runif(10,dat3$left[i],dat3$right[i]))}
pm <-ifelse(is.finite(p),p,dat3$left)
cxt3=coxph(Surv(pm,dat3$cens)~Status,data=dat3)
ek3 <-survfit(cxt3)
lines(ek3$time,ek3$surv,type="s",col=6,lty=3)
#plot(ek1$time,ek1$surv,type="s",col=1,lty=1,
```

```
xlab="Time  (Months)",ylab="S(t)",xlim=range(c(0,530)))
p <-1:nrow(dat4)
for (i in 1:nrow(dat4)){p[i]  <- mean(runif(10,dat4$left[i],dat4$right[i]))}
pm2 <-ifelse(is.finite(p),p,dat4$left)
cxt4=coxph(Surv(pm2,dat4$cens)~Status,data=dat4)
ek4<-survfit(cxt4)
lines(ek4$time,ek4$surv,type="s",col=4,lty=3)
legend(250,0.85,lty=3,col=6,  "Married  - mean based imputation",  bty="n",cex=0.8)
legend(250,0.78,lty=3,col=4,  "Single  - mean based imputation",  bty="n",cex=0.8)
dev.off()
```

Mean Imputations for Nationality with 75% exact.

```
require(survival)
dat <- read.table('NSimW100.txt',header=T)
dat1 <- dat[dat$Nationality==1,]
dat2 <- dat[dat$Nationality==0,]
cxt1=coxph(Surv(dat1$eve,dat1$cens)~Nationality,data=dat1)
ek1 <-survfit(cxt1)
pdf(file="D:/code/Nationality/mean-75.pdf",  width = 9, height  =6)
plot(ek1$time,ek1$surv,type="s",col=1,lty=1,
xlab="Time  (Months)",ylab="S(t)",xlim=range(c(0,400)))
cxt2=coxph(Surv(dat2$eve,dat2$cens)~Nationality,data=dat2)
ek2<-survfit(cxt2)
lines(ek2$time,ek2$surv,type="s",col=2,lty=1)
legend(250,1,lty=1,col=1,  "Gulf  - exact data", bty="n",cex=0.8)
```

```
legend(250,0.92,lty=1,col=2, "Others - exact data", bty="n",cex=0.8)
data <- read.table('NSimWint75.txt',header=T)
dat3 <- data[data$Nationality==1,]
dat4 <- data[data$Nationality==0,]
p <-1:nrow(dat3)
for (i in 1:nrow(dat3)){p[i] <- mean(runif(10,dat3$left[i],dat3$right[i]))}
pm <-ifelse(is.finite(p),p,dat3$left)
cxt3=coxph(Surv(pm,dat3$cens)~Nationality,data=dat3)
ek3 <-survfit(cxt3)
lines(ek3$time,ek3$surv,type="s",col=6,lty=3)
#plot(ek1$time,ek1$surv,type="s",col=1,lty=1,
xlab="Time (Months)",ylab="S(t)",xlim=range(c(0,530)))
p <-1:nrow(dat4)
for (i in 1:nrow(dat4)){p[i] <- mean(runif(10,dat4$left[i],dat4$right[i]))}
pm2 <-ifelse(is.finite(p),p,dat4$left)
cxt4=coxph(Surv(pm2,dat4$cens)~Nationality,data=dat4)
ek4<-survfit(cxt4)
lines(ek4$time,ek4$surv,type="s",col=4,lty=3)
legend(250,0.85,lty=3,col=6, "Gulf - mean based imputation", bty="n",cex=0.8)
legend(250,0.78,lty=3,col=4, "Others - mean based imputation", bty="n",cex=0.8)
 dev.off()
```