Quantile Regression for Sequentially Observed Bivariate Survival Data

by

© **Leila Torabi**

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Science.

Department of Mathematics and Statistics
Memorial University

May 2024

St. John's, Newfoundland and Labrador, Canada

Abstract

Quantile regression is an extension to the traditional linear regression. It offers a flexible way to assess the effects of covariates on the quantiles of the conditional distribution of a random variable for a given set of covariates. Since the effects of covariates can be assessed at any quantile of the conditional distribution of the response variable, it provides a better understanding of the effects of covariates comparing with traditional regression models. In this study, we consider a parametric conditional quantile regression model for survival data with time-fixed covariates. We propose a multi-stage estimation procedure to estimate the effects of covariates on the quantiles of marginal distributions of sequentially observed bivariate survival times. We model the dependency between survival times with a Clayton copula. Our estimation method is based on the martingale estimating equations. We study the bias and precision of the parameter estimates obtained with the proposed method, as well as investigate their large sample properties with simulation studies. Finally, the method is illustrated by analyzing a colon cancer data set.

To God
who gave me an amazing family
and
to the light of my eyes, my parents
and my siblings,
who were always there for me.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisors, Dr. Candemir Cigsar and Dr. Zhaozhi Fan, for their continuous support, enlightening discussions, and critical feedback during my master's studies. I deeply appreciate their immense knowledge, guidance, patience, and kindness throughout this time. I could not have imagined having better supervisors and mentors for my master's studies.

My sincere gratitude also goes to all faculty and staff in the Department of Mathematics and Statistics. In particular, I would like to thank Dr. Yildiz Yilmaz for giving me the best possible introduction to Survival Analysis, as well as all my other graduate professors who graciously shared their knowledge and wisdom to help me discover the beauty of Statistics.

I am also indebted to my parents and my siblings for their unconditional support and love throughout the years.

Statement of contributions

Dr. Candemir Cigsar and Dr. Zhaozhi Fan proposed the research question investigated throughout this thesis. The overall study was designed by Dr. Candemir Cigsar, Dr. Zhaozhi Fan and Leila Torabi. The simulation study was conducted and the manuscript was drafted by Leila Torabi. Dr. Candemir Cigsar and Dr. Zhaozhi Fan supervised the study and contributed to the final manuscript.

Table of contents

References 108

List of tables

List of figures

Chapter 1

Introduction

In this chapter, we introduce the goal of this thesis and some concepts frequently used throughout the thesis. The layout of this chapter is as follows. In Section 1.1, we introduce the general concept of multistate models. Our focus is on the illness-death model. Furthermore, we provide a general discussion of data types and introduce an illustrative example. In Section 1.2, we examine dependence modeling using copulas. Section 1.3 includes a summary of quantile regression. A comprehensive literature review is presented in Section 1.4. Finally, we outline an overview of the thesis in the concluding section of this chapter.

1.1 Multistate Models

Many processes longitudinally evolve over time in a stochastic way. Such stochastic processes are of interest in many studies. For example, people born and die, which is a process of interest in demography and insurance studies (Daley and Vere-Jones, 2003). Repairable systems may subject to repeated failures over their lifetime, which is an important subject in the field of industrial engineering (Rigdon and Basu, 2000). The process of marriage and divorce, which may occur multiple times during the lifetime of an individual, may be of interest in social sciences (Aalen et al., 2008). In medicine, patients may experience cancer relapse after the removal of a tumor and then die (Lawless, 2003). A common feature of these processes is that subjects may experience a well-defined event or multiple events at least once during their lifetimes at random

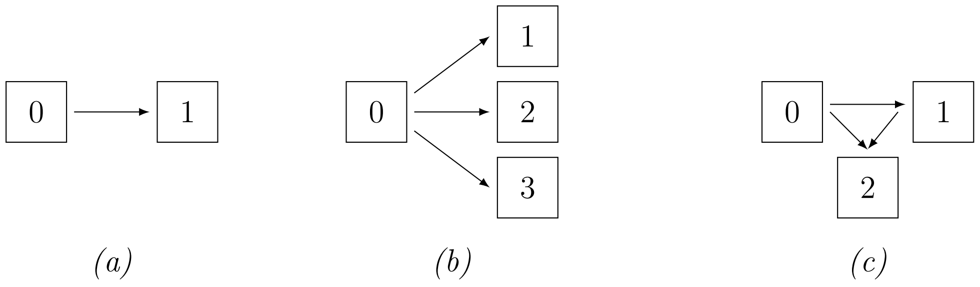


Figure 1.1: Multistate diagrams of (a) survival model, (b) competing risks model, (c) the illness-death model.

time instants. If the observations are only made at equally spaced time instants, the time is of discrete type. Otherwise, it is continuous. The event occurrence times, or shortly the event times, define the *state* of a process, which means that processes are assumed to be in a certain, well-defined state at any given time point. The set of states is called the *state space*, which usually includes finite number of distinct elements. Event history is used as a generic term to define such processes.

Multistate models provide a canonical framework for the statistical analysis of event history data. In a multistate model, subjects can be placed in different states during their follow-up. Moving from one state to another state is called a transition, which is only possible at the event times. The transitions are governed by intensity functions, which specify probabilistic characteristics of the event history processes for making transitions in multistate models. Some well-known examples of multistate models are depicted in Figure 1.1. For example, the multistate diagram of the classical survival model is given in (a), where the state "0" denotes *alive* and the state "1" denotes *death.* Individuals start at the state "0" and make a transition to the state "1" at the time of death. Diagram (b) in Figure 1.1 shows the competing risks model with three causes, in which the state "0" is *alive* and the states "1" and "2" are *death from cause 1* and *death from cause 2*, respectively, and the state "3" is *death from other causes.*

The illness-death model, also known as the disability model, is a widely utilized multistate model in the medical literature. The multistate diagram (c) in Figure 1.1 depicts the illness-death model. In this model, individuals begin at the state "0" representing a *healthy* state. Then, they move to the state "1", usually called the *ill* state or the *intermediate* state, at a given time instant when they experience a well-defined event. The state "2" is an absorbing state, called the *death* state. Transition from states "0" or "1" to state "2" are possible. If we let the notation "$a \to b$" denotes a transition from the state a to the state b, whole possible paths for subjects in the illness-death model are $0 \to 1 \to 2$ or $0 \to 2$, with states "0" and "1" being transient and state "2" being an absorbing state. In this thesis, we consider the path $0 \to 1 \to 2$ of the illness-death model. It should be noted that in some settings it is also possible for individuals to make $1 \to 0$ transitions, for example, when recovering from a sickness and regaining health. Such a model is called bidirectional illness-death model or illness-death model with recovery (Andersen et al., 2002). Many multistate models and statistical methods for their analysis have been discussed by Cook and Lawless (2018).

An important issue about the statistical analysis of event history data is the censoring. Typically, the entire path of all individuals included in a study is not always observed by the end of a study because of limited follow-up times of individuals. For instance, in certain clinical studies, within the context of the illness-death model, it may not be possible to observe exact event times for $0 \to 1$, $0 \to 2$, or $1 \to 2$ transitions for patients who withdraw from the study before its completion. In such cases, the follow-up of an individual is called right censored. More details about different censoring mechanisms can be found in Lawless (2003, Chapter 2).

In the sequentially observed bivariate survival (lifetime) data, the time and related information from the origin (study entry) to the occurrence of illness, as well as the time from illness onset to death are examined. Specifically, the duration spent in state "0" (prior to illness) and the subsequent time spent in state "1" are considered. There are important challenges in the analysis of sequentially observed bivariate survival data. As aforementioned, one of the significant challenges is dealing with censoring. Right censoring occurs when subjects do not experience the event of interest during the study period, resulting in incomplete information regarding their survival times. Particularly, in the context of sequentially observed two survival times, if the first survival time is censored, the second survival time becomes unobservable which may result in non-identifiability issue. An example of this issue in the context of sequentially observed bivariate survival times is given by Lawless and Yilmaz (2011). As a result, estimating the marginal distribution of the second gap time becomes challenging without information on the first gap time. Addressing these issues is essential for accurate estimation and interpretation of the illness-death model with bivariate survival data. Another challenge in the analysis of sequential survival time data is the dependency between two survival times. Since sequentially observed bivariate survival times are observed from the same individual, their independence does not hold. This dependency causes the second survival time to be subject to induced dependent censoring, meaning a dependent variable censors the second gap time. The presence of induced dependent censoring of the second survival time introduces additional complexities in the analysis. Specifically, due to the longer first survival times, there is an increased probability that the second survival time becomes censored. Consequently, the observed second survival times will predominantly consist of shorter times, leading to a disproportionality in the data. Failing to account for this induced dependent censoring when estimating the survival time distribution using standard methods may result in biased outcomes (Visser, 1996 ; Wang and Wells, 1998 ; Lin et al., 1999; Lawless and Yilmaz, 2011).

In addition to the challenges mentioned above, another crucial aspect of analyzing bivariate sequential lifetime data is understanding the effects of covariates on each survival time. Many researchers are therefore interested in applying regression models to make inferences on how covariates influence the timing of events and transitions within the illness-death model. While a class of parametric survival models, called accelerated failure time (AFT) models, are widely applied to understand how risk factors or treatments influence event times, estimation of the parameters in the AFT models with likelihood based methods usually provide limited information about the conditional distribution of survival times given the values of covariates. In many cases, the effects of covariates on a survival time are heterogeneous, meaning that risk factors may have varying impacts at different stages of the study period. For example, covariates tend to have a significant impact on the likelihood of survival at the beginning of the study period but often diminish or even become insignificant as time progresses. The Cox proportional hazards (Cox PH) model is another widely used regression technique in survival analysis. This model assumes that the hazard ratio remains constant over time and models the effect of covariates on the hazard rate. To address the potential violation of the constant hazard ratio assumption in the Cox PH model and to obtain a thorough understanding about the effects of covariates on the distribution of survival times, an alternative approach is the use of quantile regression (QR) models. QR model provides a dynamic perspective, offering a relationship between covariates and survival time based on different quantiles of the conditional distribution of the survival times given a set of covariates. This method is especially valuable for assessing the influence of covariates at various quantiles of survival times. QR models are robust with respect to outliers in the estimation of the effects of covariates, making them particularly useful for exploring heterogeneity effects (Koenker, 2003). We provide a more detailed explanation of QR in Section 1.3 and Section 2.3 of this thesis. We would also like to note that we use the terms survival time, gap time of lifetime interchangeably throughout this thesis. All mean the elapsed time spend in "0" or "1" state before making a transition in the illness-death model.

1.1.1 Data Types

Data obtained from event history processes usually include event occurrence times, an indicator about the type of the state occupied at that time and other relevant information about the characteristics of processes included in the study cohort, called explanatory variables or covariates. In this context, the illness-death model provides a comprehensive framework for capturing the progression of a disease, enabling a thorough statistical analysis of disease processes and providing insights into the potential influence of intermediate events on the probability of mortality. Data sets employed in this model typically contain information on the times at which a specific incident or related events occurred. Specifically, the response measurement of interest within the illness-death model often involves measuring the elapsed time from a well-defined origin to the occurrence of an event of interest. These time intervals are commonly referred to as survival or gap times. Unless explicitly mentioned, we assume these response measurements are continuous and represent observations derived from a random sample of study participants. The time origin for each subject should be precisely defined, ensuring all subjects are as comparable as possible at this point. The time origin may vary depending on the study. For instance, it can be the birth of an individual, entry into a study, the onset of a disease, or the commencement of a treatment. Analyzing data sets often involves considering either the global time scale, which encompasses calendar time or the time elapsed since study entry, or the local time scale, which entails the elapsed time between occurrences of well-defined events, commonly referred to as gap time or sojourn time. Techniques for analyzing data sets within the illness-death model are based on either the calendar time scale or periods of time between subsequent events. Most of the data types in illness-death model and other multistate models are illustrated with examples by Cook and Lawless (2018).

An important aspect of the illness-death model is to study the disease progression, particularly through sequentially observed bivariate survival data. This type of data captures the time and related information from the origin to the occurrence of illness and the time from illness onset to death. The time from the origin to death without experiencing the intermediate event in the illness-death model is disregarded in this context. This setup is often employed when the primary interest lies in understanding the relationship between the elapsed time for an individual to visit the ill state (State 1) after entering the study and the subsequent time until death (State 2). As a result, the data observed from an individual potentially include the elapsed time from study entry to entering State 1 and the time from State 1 to entering State 2, which in essence constitutes a sequentially observed bivariate gap times for individuals in the study cohort.

In the illness-death model and sequentially observed bivariate gap times, data types extend beyond event times and state indicators to include explanatory variables such as demographic and clinical factors. In the context of disease progression, regression analysis allows us to quantitatively assess the impact of the covariates on the timing of events, providing a more comprehensive understanding of the relationships between explanatory variables and gap times. In this thesis, our primary focus is quantile regression analysis of sequentially observed bivariate gap times which are subject to right censoring.

1.1.2 Data Example

Colon cancer is a significant public health concern, ranking forth in common cancer types and third in cancer-related deaths in Canada (Fitzgerald et al., 2022). Treatment options involve chemotherapy, radiation, and surgery, with notable drugs being fluorouracil (5FU) or 5FU plus levamisole. Although surgical intervention during early stages can remove an affected tissue, residual microscopic cancer may lead to recurrence of disease and death of an individual within 5 years. A clinical trial of colon cancer patients was executed to assess the efficacy of a drug therapy (Moertel et al., 1990; Lin et al., 1999). In this study, levamisole with or without 5FU was compared to a placebo concerning the recurrence and survival rates of colon cancer patients. A total of 929 participants were included in the randomized clinical trail.

Out of 929 patients, 315 patients were allocated to the placebo group, 310 patients were included in the levamisole therapy group and 304 patients were included in the 5FU plus levamisole therapy group. Patients in the study were monitored for a minimum of 18 months, with the maximum follow-up extending to approximately 9 years. The average follow-up duration for participants was 6.5 years. The central objective of this study was to assess the potential effects of levamisole therapy and 5FU plus levamisole therapy on two critical time intervals, the interval from the surgical removal of colon cancer to the onset of cancer recurrence, and subsequently, from the recurrence of colon cancer to the occurrence of death.

The data set includes crucial information regarding several timeframes. This encompasses the duration from the initial event, which is the study registration with the removal of affected tissue, to the intermediate event, which is the recurrence of cancer, or censoring. Furthermore, it tracks the time from recurrence to either death or censoring. Additionally, the data set includes the censoring status of patients associated with both the first and second type of events. The covariates in this data set include the type of treatment, gender, age at the registration time, and the time from surgery to registration, which categorizes the duration into short and long. The other covariates are tumor-related characteristics, including presence or absence of colon obstruction due to a tumor, presence or absence of colon perforation, a binary variable indicating adherence of the tumor to nearby organs, number of cancer-affected lymph nodes, categorical variable indicating tumor differentiation grade, ordinal categorical variable indicating the extent of local spread of the cancer, and presence or absence of more than 4 positive lymph nodes. A comprehensive analysis of the data set and the utilization of quantile regression are provided in Chapter 4.

1.2 Dependence Modeling with Copulas

In the analysis of event history data, the assumption of independent gap times is a strong one, and often not valid in applications even after conditioning on the values of the available covariates (Cook and Lawless, 2018, Chapter 6). This implies that the occurrence time of one event can significantly impact the occurrence time of another event, leading to a dependency between the gap times. To illustrate this concept, consider the study of patients diagnosed with colon cancer as explained in the previous section. The time interval from the removal of the affected tissues to the onset of colon cancer recurrence, and the subsequent duration from recurrence to death are likely to exhibit a dependent relationship. Modeling the dependence between these two gap times may provide valuable insights about the effectiveness of treatments, and better understanding of the progression of cancer can be obtained, as well as correct decisions regarding treatment planning and patient care can be made.

A review of methods to deal with dependency in sequentially observed gap times in the context of recurrent event process is given by Cook and Lawless (2007, Section 4.2). These methods are based on conditional models, random effects models and copulas. Conditional models are used to model the conditional distribution of the second gap time, given the value of the first gap time. Such a conditional approach generally do not provide simple forms for the marginal distributions of the gap times. As a result, effects of covariates on the marginal distribution of the second gap time may not be easily interpreted. Random effects models suffer from a similar problem as well. Additionally, they require the specification of a distribution for random effects, which is untestable. Also, they may modify the marginal distribution of the second gap time in an unrealistic way (see Aalen et al., 2008, Section 6.7). Copulas offer a powerful alternative to address these challenges. Copulas are a statistical tool that separates the joint distribution of variables from their individual marginal distributions and act as functions that link the individual marginal distributions to their joint multivariate distribution. Copula models describe the dependence structure between variables independently of their individual distributions, allowing for greater flexibility in modeling various forms of dependence. We mathematically define the copula function in Section 2.2. Many properties and technical details of copulas can be found in Joe (1997) and Nelsen (2007).

In comparison to other techniques for evaluating dependence structures, copulas have a number of advantages. First, they make it possible to model non-linear dependencies among variables, which is very helpful in studies where complex relationships and interactions between the covariates and outcome variables are not linear. Second, they are very flexible to model dependencies among variables with different marginal distributions. This is an important feature because marginal distributions can be specified by the modeling needs. For example, copulas can be used to simulate the joint distribution of two gap times in a way that the first gap time follows a specified distribution and the second gap time follows another specified distribution. The dependence structure between the two gap times is then modeled by a copula function. The use of copulas allows for studying the dependence structure and the marginal effects separately since the dependence parameters are not part of the marginal distributions. It should also be noted that copulas remain unchanged under transformations of the marginal distributions, which facilitates inference procedures.

There are several different types of copulas, including Gaussian, Archimedean, and Student's t copulas. Most of the copula families are rigorously studied by Joe (1997). In this thesis, single parameter bivariate copula models are taken into account, which is specifically chosen to explore and model the dependence structure between two sequentially observed, positive-valued variables of interest. More specifically, we use Clayton family (Clayton, 1978) of copulas, which are members of the one-parameter Archimedean copula family. The use of Clayton family for modeling the dependency of bivariate survival data has been suggested by Oakes (1982). This approach can help uncover complex forms of dependence, whether linear, non-linear, or asymmetric, and provide a comprehensive understanding of how these variables interact with each other. It should be noted that the methods developed in this thesis can be applied under other copula functions as well. Furthermore, the adequacy of the assumed copula model can be tested (Genest et al., 2009; Lawless and Yilmaz, 2011).

1.3 Quantile Regression

Quantile regression (QR), introduced by Koenker and Bassett Jr. (1978), was developed to investigate the effects of explanatory variables on the entire conditional distribution of a response variable, in relation to a given set of covariates. While the classical least squares method estimates the conditional mean function of the response variable across covariate values, QR estimates the conditional quantile function of the response variable as a linear form of the covariates. This approach not only provides a more comprehensive description of functional changes, encompassing both the tails and the center of the distribution, but also offers a robust and flexible alternative to traditional linear regression. Thus, it enables a thorough analysis of outcomes across the entire range without assuming specific distribution patterns, which is particularly useful when dealing with variable error variance or outliers. This approach provides valuable insights into conditional response beyond mean regression and can serve as a supplementary or alternative method to least squares analysis, especially in cases where underlying assumptions are problematic (Koenker, 2003).

In various clinical and medical research studies, making inferences based on the conditional quantiles of the outcome is often more desirable than relying solely on analysis based on the AFT model. For instance, specific risk factors may have a higher impact on low birth weight, which is associated with adverse outcomes such as infant mortality and chronic diseases, compared to their effects on the mean birth weight (Yang et al., 2019). Ignoring changing patterns of risk factor effects in the AFT regression analysis may lead to an oversight in capturing the strong association between risk factors and mortality among survivors of low birth weight infants. In fact, in the survival analysis, QR provides more flexible modelling of survival data, enabling a more thorough exploration of how different factors impact survival times across various quantiles of the distribution. Unlike the traditional Cox proportional hazards (PH) and AFT models, QR does not restrict the variation of the coefficients for different quantiles of the conditional distribution of the response variable, given a set of covariates. This approach yields robust results when assessing factors affecting time-to-event situations (Wei, 2022). For a comprehensive understanding of the QR model in standard settings, a detailed overview is provided by Koenker (2003).

In this thesis, our goal is to explore the impact of various covariates on specific time intervals between the first and second events. We employ QR model to investigate how alterations in one covariate, while keeping others constant, influence the duration of life for a certain proportion of individuals. For example, in a data set related to colon cancer testing the effectiveness of a drug, we inquire about how this intervention affects the survival duration of a predetermined percentile of patients. Essentially, our study delves into the relationships between covariates and dependent gap times across different quantiles for a more comprehensive understanding.

1.4 Literature Review

Multistate models and illness-death models have been extensively studied for several decades. A comprehensive review of these models, along with examples and additional applications, can be found in the book written by Hougaard (2000) and Cook and

Lawless (2018). These sources provide valuable insights into the complexities and practical applications of these models with many examples, particularly in the field of biomedicine. Ma et al. (2008) conducted a thorough examination of multistate models from both biomedical and engineering reliability perspectives.

Analyzing multistate models are usually based on two different time scales. These are the time since entering the study and the elapsed time since the last event, referred to as gap or sojourn times. In many applications, the gap times are the primary time scale. However, the stochastic ordering structure of events poses challenges for statistical analysis when the focus is on gap times. These challenges include dependent censoring, which can lead to biased parameter estimation (Lawless and Yilmaz, 2011). In recent years, various statistical methods have been proposed to address the study of gap times within the framework of multistate models. Some studies have focused on nonparametric estimation of the gap time distribution, while others have investigated different parametric and semiparametric regression models to analyze the effects of covariates on gap times. Huang and Wang (2005) discussed bivariate recurrence times, where two distinct event types alternate over time and are subject to right censoring. They proposed nonparametric estimators for the joint distribution of bivariate recurrence times and the marginal distribution of the first recurrence time. In their approach, it is assumed that the correlation between the bivariate recurrence times is characterized by a latent variable. Similarly, Huang et al. (2016) examined nonparametric estimators for the conditional bivariate cumulative incidence distribution of the bivariate gap time. They also proposed a modified Kendall's tau measure to assess the association between two successive gap times in the presence of competing risks. They concluded that nonparametric estimators and association measures can be viewed as inverse probability censoring weighted (IPCW) estimators. Huang and Louis (1998), Wang and Wells (1998), Lin et al. (1999), and Zeng and Lin (2007) provided a comprehensive explanation of nonparametric and semiparametric estimators in the analysis of gap times, most of them focus on a single gap time.

The association between survival times has been extensively studied with the development of global dependence measures such as Spearman's rho and Kendall's tau (Betensky and Finkelstein, 1999). These measures capture the overall association pattern between dependent pairs across the entire study area and offer simplicity of interpretation. However, they may not capture the local association structure, which may vary over time. To assess local association in the context of correlated survival data, various approaches exist, including frailty models, marginal methods, and copulas. Copulas, which originated in probabilistic metric spaces, have gained popularity in analyzing the association between dependent survival times, offering a means to capture local association structures that may vary over time. This function has been used to model joint distributions as a function of each marginal distribution and a dependence parameter in various contexts, including bivariate survival data and multistate modeling. The literature on copulas has rapidly expanded in the recent years, with studies exploring their statistical properties and applications. Many general references on this subject can be found in Cook and Lawless (2018, Section 6.5). Some studies have demonstrated the utility of copula-based methods in capturing complex relationships between variables and accounting for dependencies among survival times. Frees and Valdez (1998) demonstrated the statistical properties and the use of copulas by analyzing information from insurance companies on losses and expenses. They illustrated how copulas can be fitted to the data and discussed their utility in calculating reinsurance costs and forecasting expenditures for specified losses. For a more in-depth understanding of copula models, additional references include Joe (1997), Kurowicka and Cooke (2006), and Nelsen (2007). Further insights into the application of copulas in the context of multistate modeling for lifetime data are given by Cook and Lawless (2018, Section 6.5).

Several studies have applied copulas to analyze illness-death model and assess the association between consecutive gap times or sequentially observed survival times. Lakhal-Chaieb et al. (2010) proposed nonparametric estimation of association using inverse probability censoring weighting (IPCW), and estimated conditional gap time distributions using Kendall's tau and Clayton copula. Lawless and Yilmaz (2011), using a pseudolikelihood function, employed copula-based parametric and semiparametric estimation methods to account for dependency between gap times. Rotolo et al. (2013) presented a simulation method based on copulas to generate clustered multistate survival data. Diao and Cook (2014) applied copulas to characterize composites for the joint analysis of multiple multistate processes. Their study focused on utilizing copula-based methods to model the dependence structure among various processes simultaneously. In a similar vein, Meyer and Romeo (2015) presented a copula-based Bayesian semiparametric analysis for recurrent failure time data. Their research emphasized the use of copulas within a Bayesian framework to analyze the recurrent nature of failure events over time. Barthel et al. (2019) employed D-vine copulas to capture the dependency between waiting times in recurrent events subjected to right censoring. Bedair et al. (2021) conducted an analysis of correlated recurrent event data utilizing copula-frailty models. They utilized the Monte Carlo expectation-maximization (MCEM) algorithm to estimate the model parameters, considering the joint influence of frailty and copulas in capturing the correlation among recurrent events.

Regression analysis is commonly used for modeling the relationship between predictor variables and a response variable in lifetime data settings. In the context of gap times, a significant focus lies on modeling the hazard functions associated with gap times. Various techniques have been developed to address these analyses and explore the effects of covariates on complex data structures. Schaubel and Cai (2004) proposed generalized estimating equations to fit proportional hazards regression models for gap times observed sequentially. Their approach did not require specifying the functional relationship between the gap times. However, it is worth noting that the proportional hazards model they use might not fit data well, and their estimation focuses on the conditional survival function of the gap times. Huang and Chen (2017) introduced a regression model to analyze the effects of covariates on bivariate gap time data with missing first gap time information. They developed a methodology to account for the missing data and explore the covariate relationships in this setting. Lee et al. (2018) developed a semiparametric regression model to estimate the effects of covariates in settings where individuals may experience different types of recurrent events. They treated the dependency between two alternating events and among different bivariate gap time pairs within each subject with random effects. They showed that their proposed estimation method has advantages over the rank-based estimation method. When considering gap time distributions in illness-death models, Huang (2019) extended semiparametric mixture models for competing risks data and adapted them to consider the subsequent event, incorporating a copula function to model the dependent structure between successive events.

QR has recently emerged as a popular approach for directly modeling censored survival times. It has been extended to handle correlated failure times. Peng and Huang (2008) developed a QR model for survival data using martingale-based estimating equations. They utilized this approach to estimate QR coefficients for survival times, with minimization of L_1-type convex functions. They used grid-based estimation procedure to estimate the parameters of QR model for the first gap time, which is a nonparametric method. We discuss this method in Section 3.2 in some detail. Peng and Fine (2009) extended QR to handle competing risk data. Luo et al. (2013) investigated QR for recurrent gap time data. They expanded the martingale-based estimating equation method originally developed for univariate survival data to study the gap times between successive recurrent events. They discarded the last censored gap time for those subjects who have at least one complete gap time. Hsieh et al. (2013) employed QR to fit semi-competing risk data with dependent censoring. To address dependent censoring, they assumed a parametric copula model for the joint distribution of the two event times, while leaving the marginal distributions unspecified. Sun et al. (2016) generalized QR to counting processes and demonstrated its application to recurrent events. Li et al. (2017) proposed nonparametric and semiparametric estimators for quantile association in bivariate survival data. Their approach captured the dynamic association between two gap times without relying on assumptions about the marginal distributions.

Inference for QR in survival data becomes challenging when parametric assumptions about the marginal distributions are not feasible. To address this issue, several researchers have proposed the use of copula functions to approximate the marginal distributions and model QR. Wang et al. (2019) introduced a copula-based quantile regression model for longitudinal data. Building upon this framework, Wang and Shan (2021) developed composite QR, extending copula-based methodologies to account for intra-subject dependence in longitudinal data. They proposed constructing the correlation matrix through copulas, enabling the incorporation of complex dependence patterns between response variables within subjects. Ghasemzadeh et al. (2022) employed the Gaussian copula to develop a QR model for correlated mixed bivariate discrete and continuous data. They obtained maximum likelihood estimates of the model parameters using an EM algorithm and utilized a Monte Carlo technique to derive confidence intervals for the estimated parameters.

1.5 Overview of the Thesis

In this thesis, we develop an estimation method for quantile regression models of bivariate survival data, where “healthy” individuals undergo sequential events of “illness” followed by “death”, in generic terms. The time to these events are subject to right censoring. As a result, a more detailed investigation of the covariates effects in disease progression and insights on the timing of events based on cohort characteristics can be obtained. Traditional regression models for survival data, such as AFT and Cox PH models, have limitations in capturing the changing impacts of covariates at different quantiles of the conditional distribution of the survival times, given a set of covariates. On the other hand, QR models are used to estimate such effects on the entire distribution of the survival times.

Most of the methods in the QR pertaining to the analysis of survival time focus on the analysis of settings in which individuals can only experience a single event, such as, death. In this study, we employ QR model to examine the direct effect of covariates on various quantiles of the distributions of two gap times, which are sequentially observed. There are however important challenges in analyzing such data. These challenges include issues related to the identifiability of marginal survival distributions for the second gap time, as well as induced dependent censoring. Ignoring dependent censoring in the occurrence of the second event could lead to biased results. To overcome these challenges, we employ QR approach, incorporating copula functions to model the dependence structure between successive events. Copulas provide flexible modeling of bivariate survival times and are well-recognized in the literature.

Our method is based on the martingale estimating equations, and can be used to estimate covariate effects on quantiles of marginal distributions of the first and second gap times. The estimates are obtained after an application of the Newton-Raphson algorithm developed to solve the martingale estimating equations. We applied a multi-stage estimation procedure to obtain the estimates of the model parameters related to the second gap time. First, parameters related to the first gap time and copula are estimated, and then the estimated parameters are plugged-in the martingale estimating equations developed for the condition distribution of the second gap time given the first gap time and covariates. The variances of the parameters are estimated using the sandwich method. Our extensive simulations demonstrate the effectiveness of the proposed method, as compared with the Peng-Huang method, which provides unbiased results only for the first gap time and under independent assumption for the second gap time. Moreover, the practical applicability of the developed method is showcased through the analysis of a real-world medical data set, emphasizing its potential in clinical studies.

The remaining part of the thesis is organized as follows. In Chapter 2, the notations used in this thesis and some background information on technical terms are introduced. A comprehensive analysis of QR model for the sequentially observed bivariate survival data, emphasizing the modeling of dependency between sequential gap times using the Clayton copula are provided in Chapter 3. Furthermore, a real-world data set from the medical field is analyzed in Chapter 4, demonstrating the practical applicability of the developed method in a clinical study. A summary of thesis and future extensions to our work are presented in Chapter 5.

Chapter 2

Notation and Background

In this chapter, we introduce frequently used notation in this thesis and some background information on technical terms. Section 2.1 includes important concepts in the context of sequentially observed bivariate survival data. Copulas to model the dependency among random variables are discussed in Section 2.2. Finally, the concept of quantile regression is discussed in Section 2.3.

2.1 Notation and Fundamental Concepts

In this thesis, we use stochastic processes to model life history data. In particular, we focus on sequentially observed bivariate survival (gap) times. Our goal here is to introduce our notation and fundamental concepts frequently used in the remaining parts of this thesis. A more detailed study of stochastic processes pertaining to point processes can be found in point-process textbooks; e.g., Andersen et al. (1993) and Daley and Vere-Jones (2003).

A stochastic process, denoted by $\{X(t), t \in \Gamma\}$, is a set of random variables indexed by the element t in the index set Γ. Stochastic processes can be categorized as discrete-time stochastic processes or continuous-time stochastic processes based on their index parameter t, $t \in \Gamma$. In particular, a stochastic process, $\{X(t), t = 0, 1, 2, ...\}$ is a discrete-time stochastic process including a countable collection of random variables with a non-negative index parameter t. A continuous-time stochastic process, denoted by $\{X(t), t \geq 0\}$, includes an uncountable set of random variables indexed by the non-negative real numbers. In general, for any index set Γ, a stochastic process $\{X(t), t \in \Gamma\}$ means that $X(t)$ is a random variable for each $t \in \Gamma$ and denotes the state occupied at time t. The set of all possible states is called the state space. In this thesis, we consider state spaces with a finite number of elements. For example, in the illness-death model, as depicted in Figure 1.1, the state space includes the elements 0, 1 and 2. In this case, for example, we can define $X(t) = 0$, which means that the process $\{X(t), t \geq 0\}$ is in the state 0 (healthy state) at time t.

Many event history models, including classical survival and illness-death models, can be considered as a point process model. A broad survey of these models can be found in Cox and Isham (1980) or Andersen et al. (1993). In these models, times of event occurrences constitute point processes, which can be described by counting the number of events as they happen during the follow-up of processes, which leads to the term *counting process*. In a simple definition, counting processes are the number of well-defined events occurred throughout the duration of the follow-up period of processes included in a study. In this section, we discuss fundamentals of counting processes and introduce our basic notation. More rigorous treatment of counting processes in the context of event history analysis can be found in Andersen et al. (1993).

In this thesis, we consider continuous-time counting processes. Let the random variable $N(t)$ represent the number of events that have occurred up to, and including, t, where $t \geq 0$. Also, let $N(s,t)$ denote the number of event occurrences over the interval $(s,t]$ so that $N(s,t) = N(t) - N(s)$ for all $0 \leq s < t < \infty$, where $N(0) = 0$ and expectation of $N(t)$, denoted by $E[N(t)]$, is finite for all $t > 0$. A counting process, denoted by $\{N(t), t \geq 0\}$, is a stochastic process with a collection of random variables that indicate the number of a well-defined event over a specified period of time. The random variable, $N(t)$, as a function of t, is a right-continuous integer-valued step function with jump of size one only.

Survival analysis or time-to-event analysis deals with non-negative random variables, representing the random occurrence time of a well-defined event, such as the death of an individual or the recurrence of a disease. Suppose that the non-negative random variable T represents the survival time in a study. In this thesis, all functions are defined over the interval $[0,\infty)$, unless otherwise stated. The cumulative distribution function (c.d.f.) and probability density function (p.d.f.) of the random variable

T are respectively defined as $F(t) = \Pr(T \leq t)$ and $f(t) = (d/dt)F(t)$, $t > 0$. The survival function $S(t)$ is then given by

$$S(t) = 1 - F(t) = \Pr(T > t) = \int_t^\infty f(x)dx, \qquad t \geq 0. \tag{2.1}$$

Another important concept in the survival analysis is the hazard function of the random variable T, denoted by $h(t)$. Given that the person has not experienced the event of interest up until time t, the hazard function is defined as the instantaneous conditional rate of failure (event) or death at time t (Lawless, 2003, p. 9). It is mathematically defined as

$$h(t) = \lim_{dt \to 0} \frac{\Pr(t \leq T < t + dt | T \geq t)}{dt}, \qquad t \geq 0. \tag{2.2}$$

It can be shown that $S(t) = \exp\Big(-\int_0^t h(s)ds\Big)$ and $f(t) = -(d/dt)S(t)$.

Counting processes are usually described by their intensity functions. Let $H(t) = \{N(s), 0 \leq s < t\}$ denotes the history of the counting process $\{N(t), t \geq 0\}$ at time t. Note that the history, $H(t)$, includes all information about the counting process $\{N(t), t \geq 0\}$ over the interval $[0, t)$. The intensity function of the counting process $\{N(t), t \geq 0\}$, denoted by $\lambda(t|H(t))$, gives the instantaneous conditional probability that an event occurs over a small time interval $[t, t+\Delta t)$ in the limit as Δt approaches 0, given the history of the process $H(t)$. It is mathematically defined as follows.

$$\lambda(t|H(t)) = \lim_{\Delta t \to 0} \frac{\Pr(N(t, t + \Delta t) = 1 | H(t))}{\Delta t}, \qquad t \geq 0, \tag{2.3}$$

where $N(t, t + \Delta t)$ gives the number of events in the interval $(t, t + \Delta t]$. We assume that at most a single event may occur at any given time instant. In the continuous time scale, the intensity function completely defines a counting process (Cook and Lawless, 2007).

The martingale theory is an important concept to model counting processes. Within this concept, the Doob-Meyer decomposition theorem (Andersen et al., 1993, pp. 66-67) states that any counting process $\{N(t), t \geq 0\}$ can be decomposed into the sum of a martingale, denoted by $M(t)$, and a predictable increasing process given by $\int_0^t \lambda(s|H(s))ds$; that is,

$$N(t) = M(t) + \int_0^t \lambda(s|H(s))ds, \qquad t \geq 0, \tag{2.4}$$

where a predictable process is a process that its value can be anticipated at any given time based on information available up to that time (Aalen et al., 2008). This result can be further represented as martingale increments, given by

$$dN(t) = dM(t) + \lambda(t|H(t))dt, \tag{2.5}$$

where $dN(t)$ denotes the number of jumps made by the counting process $\{N(t), t \geq 0\}$ in the infinitesimal time interval $[t, t+dt)$. Using (2.3), for infinitesimal time intervals $[t, t+dt)$, we have

$$\lambda(t|H(t))dt = \Pr(dN(t) = 1|H(t)), \qquad t \geq 0. \tag{2.6}$$

In the continuous time setting, $dN(t)$ is a $0 - 1$ valued binary variable. So, the probability of a jump in the interval $[t, t+dt)$ is given by the expected number of jumps in the interval; that is,

$$\lambda(t|H(t))dt = E(dN(t)|H(t)), \qquad t \geq 0. \tag{2.7}$$

If we let $dM(t) = dN(t) - \lambda(t|H(t))dt$, from (2.5) and (2.7) we have

$$E(dM(t)|H(t)) = 0, \qquad t \geq 0, \tag{2.8}$$

which is the definition of a martingale with respect to the history $H(t)$ (Aalen et al., 2008, Section 2.2.1). Since $M(t) = \int_0^t dM(u)$ and $N(t) = \int_0^t dN(u)$, it is easy to see that

$$M(t) = N(t) - \int_0^t \lambda(s|H(s))ds, \qquad t \geq 0, \tag{2.9}$$

is a martingale with respect to the history $H(t)$. From this development, the result (2.9) can be interpreted as, at any time t, the counting process $\{N(t), t \geq 0\}$ can be decomposed into summation of two parts, a zero mean random white-noise, given by $M(t)$ and a cumulative intensity process given by $\int_0^t \lambda(s|H(s))ds$, which is a predictable process.

We now explain how to express the survival model in terms of counting processes. We first discuss the case in which there is no censoring present. Suppose the survival time, represented by T, is not subject to the right censoring. In this case, we define the counting process $\{N(t), t \geq 0\}$, where $N(t) = I\{T \leq t\}$ and $I(\cdot)$ is a typical $0-1$ valued indicator function. Suppose that the individual experiences the event of interest at time T. Then, the value of the random variable $N(t)$ is 0 if $t < T$ and 1 if $T \leq t$. Note that, from (2.3), we have

$$\lambda(t)dt = \Pr(dN(t) = 1|H(t)). \tag{2.10}$$

The result in (2.10) can be written as

$$\Pr(dN(t) = 1|H(t)) = \begin{cases} h(t)dt, & \text{for } T \geq t \\ 0, & \text{for } T < t, \end{cases} \tag{2.11}$$

where $h(t)$ is the hazard function of T as defined in (2.2). As discussed by Aalen et al. (2008, Section 1.4.2), the censoring is unavoidable in survival settings. Now suppose the survival time T, is subject to the right censoring denoted by C. In this case, we define the counting process $\{N(t), t \geq 0\}$ where

$$N(t) = I\{\min(T, C) \leq t, \delta = 1\}, \tag{2.12}$$

where $\delta = I(T \leq C)$, called the censoring indicator, and $I(\cdot)$ is a typical $0-1$ valued indicator function.

To deal with censoring, we next introduce at-risk function, $Y(t)$, as an indicator function taking the value of 1 if the process is under observation and at-risk of observing an event at time t; otherwise, it is 0. Under the observation scheme, in which the process is continuously observed from the time origin and subject to right censoring, the at-risk function is given by

$$\begin{aligned} Y(t) &= I(T \geq t, C \geq t) \\ &= \begin{cases} 1, & \text{if individual is under observation just before or at time } t, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{2.13}$$

Note that, under the assumption of the independent censoring (Aalen et al., 2008, p. 30), the intensity process takes the form

$$\lambda(t) = Y(t)h(t), \ \ t \geq 0. \tag{2.14}$$

With at-risk function notation, the martingale associated with the counting process $N(t)$ can be written as

$$\begin{aligned} M(t) &= N(t) - \int_0^t Y(s)h(s)ds, \\ &= N(t) - \Lambda_T\big(\min(t, T, C)|H(t)\big), \end{aligned} \tag{2.15}$$

where $\Lambda_T(\cdot|H(t))$ is the cumulative hazard function of T conditional on the past. From (2.15), we obtain that

$$E(M(t)|H(t)) = 0, \ \ t \geq 0. \tag{2.16}$$

In many event history studies, there are more than one event of interest. As discussed in the previous chapter, the main data type considered in this thesis is the sequentially observed bivariate survival data as a part of the illness-death model. We next extend our notation to this setting. Suppose that T_1 and T_2 denotes survival (gap) times, representing the elapsed time from the initial "healthy" state to the "disease" state and from the "disease" state to the "death" state, respectively. Note that T_1 and T_2 are sequentially observed and may not be independent. We therefore need to address the dependency between T_1 and T_2. For $t_1 \geq 0$ and $t_2 \geq 0$, the joint distribution function of T_1 and T_2 is defined as

$$F\left(t_1, t_2\right) = \Pr(T_1 \leq t_1, T_2 \leq t_2), \tag{2.17}$$

and the joint survivor function is given by

$$S\left(t_1, t_2\right) = \Pr(T_1 \geq t_1, T_2 \geq t_2). \tag{2.18}$$

The marginal distribution functions of T_1 and T_2 are given by $F_1(t_1) = F(t_1, \infty)$ and $F_2(t_2) = F(\infty, t_2)$, and the marginal survivor functions are $S_1(t_1) = S(t_1, 0)$ and $S_2(t_2) = S(0, t_2)$, respectively. The hazard rate of the conditional distribution of T_2

given $T_1 = t_1$ is

$$h_{2|1}(t|t_1) = \lim_{dt\to 0} \frac{\Pr(T_2 < t + dt | T_2 \geq t, T_1 = t_1)}{dt}, \qquad t > 0, \tag{2.19}$$

and the conditional intensity function of T_2 given $T_1 = t_1$ is

$$\lambda_{2|1}(t|t_1, H(t)) = \lim_{dt\to 0} \frac{\Pr(N(t, t + dt) = 1 | T_1 = t_1, H(t))}{dt}, \qquad t > 0. \tag{2.20}$$

The gap time variables (T_1, T_2) are subject to a potential right censoring time C. We define that $(t_1, t_2) = (\min(T_1, C), \min(T_2, C - t_1))$ and $(\delta_1, \delta_2) = (I[T_1 = t_1], I[T_2 = t_2])$. Note that t_1 and t_2 in (t_1, t_2) are either observed gap times T_1 and T_2 or the censoring time, and δ_1 and δ_2 in (δ_1, δ_2) are event indicators of the first and second types of events, respectively. Suppose $\{N_1(t), t \geq 0\}$, where $N_1(t) = I(\min(T_1, C) \leq t, \delta = 1)$, is the counting process of the first event. As explained before, for any $t \geq 0$, the martingale process associated with the counting process of the first event $N_1(t)$ is defined as

$$\begin{aligned} M_1(t) &= N_1(t) - \int_0^t Y_1(s) h_{T_1}(s) ds, \\ &= N_1(t) - \Lambda_{T_1}\big(\min(t, T_1, C) | H(t)\big), \end{aligned} \tag{2.21}$$

which gives

$$E(M_1(t)) = 0, \qquad \text{for } t \geq 0. \tag{2.22}$$

The function $Y_1(t) = I(T_1 \geq t, C \geq t)$ takes the value of one if the process is at-risk of experiencing the first event just before time t, otherwise it is zero. The function $h_{T_1}(t)$ is the hazard function of T_1 and $\Lambda_{T_1}\big(\,\cdot\,|H(t)\big)$ is the conditional cumulative hazard function of T_1, given $H(t)$.

We next define the counting variable $N_2(t)$ to denote the number of transition from state "illness" to state "death" during the time interval $(0, t]$. The process $\{N_2(t), t \geq 0\}$ forms the counting process associated with the second event, where

$$N_2(t) = I(\min(T_2, C - t_1) \leq t, \delta_1 = 1, \delta_2 = 1). \tag{2.23}$$

As explained by Yip and Lam (1997), the martingale process associated with the counting process $\{N_2(t), t \geq 0\}$ is defined as

$$\begin{aligned} M_2(t) &= N_2(t) - \int_0^t Y_2(s) h_{2|1}(s|t_1) ds, \\ &= N_2(t) - \Lambda_{2|1}\big(\min(t, T_2, C - t_1) | T_1 = t_1, H(t)\big), \end{aligned} \tag{2.24}$$

which is a zero mean martingale given the history $H(t)$. In Equation (2.24), the random variable

$$Y_2(t) = I(T_2 \geq t, C - t_1 \geq t, T_1 = t_1, \delta_1 = 1), \;\; t \geq 0, \tag{2.25}$$

which takes the value of one if an individual is at-risk of having the second event after experiencing the first event, otherwise it is zero. The function $h_{2|1}(\cdot|t_1)$ is the conditional hazard rate of T_2, given $T_1 = t_1$, and the cumulative hazard function of the second gap time given the all information up to time t and $T_1 = t_1$ is given by $\Lambda_{2|1}\big(\,\cdot\,|T_1 = t_1, H(t)\big)$.

Now suppose that we have a random sample of n processes in a survival study. The first and second gap times of the ith individual, $i = 1, 2, ..., n$, are denoted by T_{1i} and T_{2i}, respectively, which are subject to a right-censoring time denoted by C_i. Additionally, event indicators for the first and second events are denoted by δ_{1i} and δ_{2i}, $i = 1, 2, ..., n$. The maximum likelihood estimation can be applied to estimate the parameters in a given model with the likelihood function for the sequentially observed bivariate survival data. Let $\{(t_{1i}, t_{2i}, \delta_{1i}, \delta_{2i}), i = 1, 2, ..., n\}$ denote the observed data. Then, the likelihood function is given by

$$L = \prod_{i=1}^{n} \Big[\frac{\partial^2 F(t_{1i}, t_{2i})}{\partial t_{1i} \partial t_{2i}}\Big]^{\delta_{1i}\delta_{2i}} \Big[\frac{\partial F_1(t_{1i})}{\partial t_{1i}} - \frac{\partial F(t_{1i}, t_{2i})}{\partial t_{1i}}\Big]^{\delta_{1i}(1-\delta_{2i})} \Big[1 - F_1(t_{1i})\Big]^{(1-\delta_{1i})}. \tag{2.26}$$

The model parameters in the c.d.f. F_1 of T_1 and the joint c.d.f. F of T_1 and T_2 can be estimated by maximizing the log of the likelihood function (2.26). In the next section, we will briefly discuss how to do this by using copulas, as well as how to obtain the estimates of the parameters in the marginal c.d.f. F_2 of T_2.

2.2 Modeling Dependence with Copulas

Copula is a statistical tool used to model the dependence between different variables in a multivariate distribution. This function combines known individual distributions and a correlation structure to create a joint distribution function, representing a way of combining these variables into a multivariate uniform distribution. The idea behind the copula approach is to break down the joint distribution into two parts, the individual distributions of each variable and a special function called the copula, which determines how the variables are related to each other. The copula function is all about describing the dependence between variables, while the individual distributions only describe the marginal characteristics of each variable separately. In the remaining part of this section, we briefly introduce copula modeling and related concepts. For a more in-depth and rigorous exploration of the subject, additional resources can be found in Nelsen (2007) and Joe (1997).

Suppose the p-dimensional columnwise random vector $(X_1, X_2, ..., X_p)^T$, where the notation T stands for the transpose of a vector or matrix, represents the outcomes that we wish to analyze, with marginal cumulative distribution functions $F_1, F_2, ..., F_p$, respectively. Then there exists a p-dimensional copula function $C : [0,1]^p \to [0,1]$ such that for the vector $(X_1 = x_1, X_2 = x_2, ..., X_p = x_p)^T$ in R^p, the copula function C is defined by

$$C\left(F_1(x_1), F_2(x_2), ..., F_p(x_p)\right) = F(x_1, x_2, ..., x_p). \tag{2.27}$$

In this thesis, we focus on bivariate random variables. So, for convenience from now on, we take $p = 2$, and consider a pair of random variables, for example X_1 and X_2, with marginal cumulative distribution functions, $F_1(x_1) = P(X_1 \le x_1)$ and $F_2(x_2) = \Pr(X_2 \le x_2)$, respectively, and a joint c.d.f., $F(x_1, x_2) = \Pr(X_1 \le x_1, X_2 \le x_2)$. According to Sklar (1959), any joint c.d.f. $F(x_1, x_2)$ has a unique copula function $C(u_1, u_2)$, where $0 \le u_1, u_2 \le 1$, such that

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2)), \tag{2.28}$$

for all $(x_1, x_2) \in R^2$. Note that any multivariate distribution function F can be expressed in this manner, and a copula representation exists for every multivariate distribution function. The joint c.d.f. of X_1 and X_2 on the unit square $C(u_1, u_2)$ is a copula, with marginal distributions of X_1 and X_2 following a uniform distribution on $(0, 1)$.

There are some desirable properties of copulas. We listed important ones here. More discussion can be found in Nelsen (2007) and Joe (1997). One notable benefit is the flexibility to utilize various families as marginal distributions for the random variables. The copula construction does not constrain the choice of marginal distributions. This result is especially useful when analyzing sequentially observed bivariate survival gap times that are anticipated to exhibit unique characteristics. Furthermore, copulas allow for the study of the dependence structure between random variables, separated from the marginal distributions. This aspect provides researchers valuable insights and a deeper understanding of the interrelationships among the variables under consideration. Some subject areas that have used copulas to understand relationships among multivariate observations include epidemiological and actuarial studies, biological, medical, epidemiological studies, industry, etc. (Frees and Valdez, 1998).

In the copula modeling, various methods exist to construct copulas, with one of the most commonly employed approaches being the Archimedean approach (Nelsen, 2007). Archimedean copulas take on a specific form, which can be expressed as

$$C(u_1, u_2, ..., u_p) = \varphi^{-1}(\varphi(u_1) + \varphi(u_2) + ... + \varphi(u_p)), \tag{2.29}$$

where $\varphi(\cdot)$ represents the generator of the copula. Different copula families emerge based on the choice of $\varphi(\cdot)$. Each of these families possesses distinct characteristics, making them suitable for different types of dependence patterns observed in the data. Frequently used Archimedean families include the Clayton family, Gumbel-Hougaard family and Frank family. In this thesis, our focus is on the Clayton copula, which is widely used for modeling positive lower tail dependence. However, proposed method can be applied under different copulas as well.

The Clayton copula can be expressed in the following form.

$$C_\phi(u, v) = \left(u^{-\phi} + v^{-\phi} - 1\right)^{-\frac{1}{\phi}}, \tag{2.30}$$

where $\phi > 0$ is the Clayton copula parameter, and u and v are real numbers in $(0, 1)$. The generator function of the Clayton copula is represented as $\varphi_\phi(t) = t^{-\phi} - 1$

(Nelsen, 2007). This generator function plays a crucial role in defining the Clayton copula and characterizes its unique dependence structure. It is essential to investigate the connection between the Clayton copula parameter, denoted as ϕ, and Kendall's tau, represented by τ_ϕ in this thesis. Kendall's tau is a widely used measure of rank correlation, providing a quantification of the association between variables in copula modeling. In the Clayton copula, this relationship is mathematically defined as

$$\tau_\phi = \frac{\phi}{\phi+2}, \quad \phi > 0. \tag{2.31}$$

We use Clayton copula to model the dependency between the first and the second gap times, T_1 and T_2, in the sequentially observed bivariate survival data. The choice of the Clayton copula is recommended for its suitable structure in modeling the dependency between two survival times subject to right censoring (Oakes, 1982). So the likelihood function (2.26) can be expressed in terms of the copula function using the relationship between the joint c.d.f. $F(t_{1i}, t_{2i})$ and the copula function, $C(F_1(t_{1i}), F_2(t_{2i}))$, given by

$$\begin{aligned} F(t_{1i}, t_{2i}) &= C(F_1(t_{1i}), F_2(t_{2i})), \\ &= \left(F_1(t_{1i})^{-\phi} + F_2(t_{2i})^{-\phi} - 1\right)^{-\frac{1}{\phi}}. \end{aligned} \tag{2.32}$$

Substituting this relationship, the likelihood function (2.26) becomes

$$\begin{aligned} L = &\prod_{i=1}^{n} \Big[\frac{\partial^2 C(F_1(t_{1i}), F_2(t_{2i}))}{\partial t_{1i} \partial t_{2i}}\Big]^{\delta_{1i}\delta_{2i}} \\ &\times \Big[\frac{\partial F_1(t_{1i})}{\partial t_{1i}} - \frac{\partial C(F_1(t_{1i}), F_2(t_{2i}))}{\partial t_{1i}}\Big]^{\delta_{1i}(1-\delta_{2i})} \\ &\times \Big[1 - F_1(t_{1i})\Big]^{(1-\delta_{1i})} \\ = &\prod_{i=1}^{n} \Big[\frac{\partial^2 \left(F_1(t_{1i})^{-\phi} + F_2(t_{2i})^{-\phi} - 1\right)^{-\frac{1}{\phi}}}{\partial t_{1i} \partial t_{2i}}\Big]^{\delta_{1i}\delta_{2i}} \\ &\times \Big[\frac{\partial F_1(t_{1i})}{\partial t_{1i}} - \frac{\partial \left(F_1(t_{1i})^{-\phi} + F_2(t_{2i})^{-\phi} - 1\right)^{-\frac{1}{\phi}}}{\partial t_{1i}}\Big]^{\delta_{1i}(1-\delta_{2i})} \\ &\times \Big[1 - F_1(t_{1i})\Big]^{(1-\delta_{1i})} \\ = &\prod_{i=1}^{n} \Big[(\phi+1)\left(F_1(t_{1i})^{-\phi} + F_2(t_{2i})^{-\phi} - 1\right)^{-\frac{1}{\phi}-2} F_1(t_{1i})^{-\phi-1} f_1(t_{1i}) F_2(t_{2i})^{-\phi-1} f_2(t_{2i})\Big]^{\delta_{1i}\delta_{2i}} \end{aligned}$$

$$
\times \Big[f_1(t_{1i}) + \frac{1}{\phi}\left(F_1(t_{1i})^{-\phi} + F_2(t_{2i})^{-\phi} - 1\right)^{-\frac{1}{\phi}-1} F_1(t_{1i})^{-\phi-1} f_1(t_{1i})\Big]^{\delta_{1i}(1-\delta_{2i})}
\times \Big[1 - F_1(t_{1i})\Big]^{(1-\delta_{1i})}. \tag{2.33}
$$

In this thesis, we use the likelihood method to estimate the parameters related to the first gap time T_1 and the copula parameter in the analysis of the second gap time. This approach follows the methodology suggested by Lawless and Yilmaz (2011), ensuring that our parameter estimates remain consistent and reliable throughout the analytical process. To achieve this, we utilize the “optim” function in R software to find the maximum likelihood estimation of these parameters.

2.3 Quantile Regression

As mentioned in Chapter 1, quantile regression (QR) assesses the effects of covariates across the entire response distribution. Quantiles, also known as percentiles, play a crucial role in data summarization. Among these, the median, which represents the $50th$ percentile, divides the ordered observations into two equal parts based on their magnitude. In notation, we denote the τth quantile as q_τ, where τ is a numeric value within the range of 0 and 1 ($0 < \tau < 1$).

Let Y be a random variable with the c.d.f. $F(y)$. For any given value $0 < \tau < 1$, the τth quantile of Y, denoted by $F^{-1}(\tau)$, is defined as

$$
Q_Y(\tau) = F^{-1}(\tau) = \inf\{y : F(y) > \tau\}, \tag{2.34}
$$

for any $0 < \tau < 1$ (Koenker, 2003). Replacing the c.d.f. F with its empirical distribution function, given by

$$
F_n(y) = n^{-1}\sum_{i=1}^{n} I(Y_i < y), \tag{2.35}
$$

leads to a sample quantile. The τth sample quantile is then given by

$$
Q_Y(\tau) = F_n^{-1}(\tau) = \inf\{y : F_n(y) > \tau\}. \tag{2.36}
$$

Based on the conditional mean of Y given $\mathbf{x}$, indicated as $E(Y|\mathbf{x})$, linear regression describes the average relationship between a set of regressors ($\mathbf{x}$) and an outcome variable (Y). Least squares estimation method, which minimizes the sums of squares of the residuals, is a standard method to estimate the linear regression parameters. However, as noted by Koenker (2003), the whole conditional distribution of Y given $\mathbf{x}$ cannot be described by this conditional mean $E(Y|\mathbf{x})$. When data are skewed or heteroscedastic, for instance, classical least squares regression cannot be utilized to effectively describe the connection between response and covariates. However, the QR allows us to create a distinct QR line for each quantile value.

QR models are modeling conditional quantiles rather than conditional means, extending mean regression to examine the entire conditional distribution of the response variable. As a result, the location, scale, and shape of the distribution can all be thoroughly studied to give a complete picture of how the covariates affect the entire response distribution (Koenker, 2003). QR is a statistical technique used to estimate conditional quantiles of a response variable Y given a set of explanatory variables represented by a design matrix $\mathbf{x}$, which is a p-dimensional vector. For a given quantile level τ, the τth linear conditional quantile, denoted as $Q_Y(\tau|\mathbf{x})$, represents the conditional quantile of Y given $\mathbf{x}$. In the QR framework, it is expressed as

$$Q_Y(\tau|\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}(\tau), \quad \tau \in (0,1), \tag{2.37}$$

where $Q_Y(\tau|\mathbf{x}) = {F_Y}^{-1}(\tau|\mathbf{x})$ and the vector of parameters denoted by $\boldsymbol{\beta}(\tau)$ represents the effects of $\mathbf{x}$ on the τth conditional quantile of the response variable Y. Additionally, this model can also be written as a standard linear model with the formula $Y = \mathbf{x}^T\boldsymbol{\beta}(\tau) + \varepsilon$, where ε stands for a random error term such that $Q_\varepsilon(\tau) = 0$.

Estimating $\boldsymbol{\beta}(\tau)$ in QR involves minimizing a specific objective function, which depends on the choice of τ. The loss function aims to find the values of $\boldsymbol{\beta}(\tau)$ that best capture the relationship between conditional quantile of Y and $\mathbf{x}$. As explained by Koenker (2003), estimate of $\boldsymbol{\beta}(\tau)$, denoted by $\hat{\boldsymbol{\beta}}(\tau)$, can be obtained by solving the following minimization problem.

$$\begin{aligned}\hat{\boldsymbol{\beta}}(\tau) &= \underset{\boldsymbol{\beta}(\tau)\in R^p}{\arg\min}\left\{\sum_{i\in\{i:y_i\geq\mathbf{x}_i^T\boldsymbol{\beta}(\tau)\}}\tau\left|y_i-\mathbf{x}_i^T\boldsymbol{\beta}(\tau)\right| + \sum_{i\in\{i:y_i<\mathbf{x}_i^T\boldsymbol{\beta}(\tau)\}}(1-\tau)\left|y_i-\mathbf{x}_i^T\boldsymbol{\beta}(\tau)\right|\right\}\\ &= \underset{\boldsymbol{\beta}(\tau)\in R^p}{\arg\min}\sum_{i=1}^{n}\rho_\tau(y_i-\mathbf{x}_i^T\boldsymbol{\beta}(\tau)) \qquad (2.38)\end{aligned}$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$, as a piecewise linear function, is the check loss function in the QR model and $I(\cdot)$ is the indicator function.

One special case is the median regression, where the quantile level τ is equal to 0.5. In this case, the goal is to find the value of the parameter vector $\boldsymbol{\beta}(\tau)$ that minimizes the sum of the absolute values of the differences between the observed response values (y_i) and the corresponding predicted values $(\mathbf{x}_i^T \boldsymbol{\beta}(\tau))$ for all data points in the data set. Note that

$$\hat{\boldsymbol{\beta}}(0.5) = \underset{\boldsymbol{\beta}(\tau) \in R^p}{\arg\min} \sum_{i=1}^{n} \left| y_i - \mathbf{x}_i^T \boldsymbol{\beta}(0.5) \right|, \tag{2.39}$$

represents the estimated parameter vector for the conditional median.

QR model for survival data is complicated by censoring. In this regard, Peng and Huang (2008) proposed a quantile regression model for survival data using martingale-based estimating equations. Their method is explained in Section 3.2 in more detail. This becomes even more challenging when handling gap times, especially, the second gap time, which is dependent on the first gap time. We will propose a new method to address this issue in Section 3.3.

Chapter 3

Estimation for Quantile Regression of Sequential Lifetime Data

In this chapter, we discuss analyzing sequentially observed bivariate survival times. Correlated gap times pose challenges in the analysis, including induced dependent censoring and non-identifiability. These challenges are addressed with the estimation method considered in this chapter. This chapter is organized as follows. Parametric estimation of sequential gap times with copula is discussed in Section 3.1. Parametric estimation of quantile regression of sequential gap times based on the Peng-Huang method and the proposed method are explained in Section 3.2 and Section 3.3, respectively. To validate the martingale property associated with the counting process of the gap times, both with and without covariates, we conduct Monte Carlo simulations. The results of this study are presented in Section 3.4. Finally, Section 3.5 includes a summary of a set of simulation studies under various settings to demonstrate the performance of the proposed method.

3.1 Parametric Estimation of Sequentially Observed Bivariate Gap Times: A Copula Approach

As discussed in Section 1.1 and by Lawless and Yilmaz (2011), sequentially observed gap times may subject to dependent censoring. This issue is an important restriction on the use of some analytical methods developed for the analysis of the marginal distribution of the second and subsequent gap times. Identifiability issue is another challenge for the analysis of such data, which arises for the second and subsequent gap times when the first gap time is censored. A parametric model for the joint distribution of the first and second gap times can be adopted to address these issues with sequentially observed bivariate gap times. In this section, we discuss the dependence modeling of two sequentially observed gap times by using copula functions. Our discussion is based on the method developed by Lawless and Yilmaz (2011). Our goal is to explain how copulas can be utilized to estimate the dependence parameter in settings with sequentially observed bivariate gap times. This parameter exists in the joint cumulative distribution and survival functions of gap times, and connects the joint distribution function to its marginal distributions with a dependence parameter.

In many studies, there is an interest in assessing the effects of covariates on the marginal distribution of the gap times between successive events. Many methods require the independence of within-subject gap times, conditional on the given values of covariates. In order to do this, either the gap times must be independent or the covariate vector must accurately represent the relevant information. In general, the independence assumption of the successive gap times is very strong, and does not hold in many studies even after conditioning on the available covariates, but if they are, then modeling, analysis and interpretation of the effects of the covariates are straightforward. Also, majority of the available methods for analyzing gap times concentrate on modelling the conditional hazard function of the second gap times given the values of the first gap times and available covariates. These methods may not provide a direct interpretation of the covariate effects on the marginal distribution of the second and subsequent gap times. In this section, we consider a linear regression setup for gap times on the basis of the accelerated failure time (AFT) model, which can provide a direct assessment of covariate effects on the marginal distribution of the gap times. We focus on the sequentially observed bivariate gap times, in which the dependency is modeled through copulas. As discussed by Lawless and Yilmaz (2011), the method can be extended to settings with more than two sequentially observed gap times.

Suppose that n independent randomly chosen individuals are included in a study. All individuals in the study cohort starts at a well-defined initial state at the beginning of their follow-ups, and may sequentially experience two events. For the ith individual, $i = 1, 2, ..., n$, we let T_{1i} and T_{2i} represent the elapsed time from the initial state to the first event and the elapsed time between the first and second events, respectively. Since the events can only occur sequentially, the gap time T_{2i} cannot be observed unless the gap time T_{1i} has already been observed for the ith individual. We next let $(t_{1i}, t_{2i}) = (\min(T_{1i}, C_i), \min(T_{2i}, C_i - t_{1i}))$ and $(\delta_{1i}, \delta_{2i}) = (I[T_{1i} = t_{1i}], I[T_{2i} = t_{2i}])$, where C_i is the random right censoring time for the ith individual, $i = 1, 2, ..., n$. Note that t_{1i} and t_{2i} are the observed quantities for the ith individual at the end of the study and δ_{1i} and δ_{2i} are the event indicators associated with the first and second gap times, respectively. If $\delta_{1i} = 0$, the second gap time T_{2i} is unobservable. Suppose that $\tilde{\mathbf{x}}_i$ is a $(p-1)$ dimensional vector including the values of $(p-1)$ covariates for the ith individual, $i = 1, 2, ..., n$. Let $\mathbf{x}_i = (1, \tilde{\mathbf{x}}_i^T)^T$ be a p dimensional vector. Note that, for notational convenience, we consider time-fixed covariates in the vector $\mathbf{x}_i$. However, the discussion can be extended to the settings in which covariates may depend on the time variable as well.

We respectively denote the joint distribution and survival functions of T_{1i} and T_{2i}, given $\mathbf{x}_i$, by

$$F(t_1, t_2|\mathbf{x}_i) = \Pr(T_{1i} \leq t_1, T_{2i} \leq t_2|\mathbf{x}_i), \tag{3.1}$$

and

$$S(t_1, t_2|\mathbf{x}_i) = \Pr(T_{1i} > t_1, T_{2i} > t_2|\mathbf{x}_i), \tag{3.2}$$

where $0 < t_1 < \infty$ and $0 < t_2 < \infty$ for $i = 1, 2, ..., n$. For a given vector of covariates $\mathbf{x}_i$, the marginal distribution of T_{1i}, denoted by $F_1(t_1|\mathbf{x}_i)$, $i = 1, 2, ..., n$, can be estimated by standard methods in the theory of survival analysis (see, e.g., Lawless, 2003). The formulation of the second gap time T_{2i}, given the vector of covariates $\mathbf{x}_i$, $i = 1, 2, ..., n$, is usually considered through the conditional cumulative distribution function (c.d.f.) of T_{2i} given t_{1i} and $\mathbf{x}_i$, and the conditional hazard function of T_{2i} given t_{1i} and $\mathbf{x}_i$ which are given by

$$F_{2|1}(t_2|\mathbf{x}_i, t_{1i}) = \Pr(T_{2i} \leq t_2|T_{1i} = t_{1i}, \mathbf{x}_i), \tag{3.3}$$

and

$$h_{2|1}(t_2|\mathbf{x}_i, t_{1i}) = \lim_{dt \to 0} \frac{\Pr(T_{2i} < t_2 + dt|T_{2i} \geq t_2, T_{1i} = t_{1i}, \mathbf{x}_i)}{dt}, \tag{3.4}$$

respectively (Lawless, 2003, Section 11.3.1). Since T_{2i} can only be observed if T_{1i} is observed; that is, $T_{1i} < \infty$ for $i = 1, 2, ..., n$, the marginal distribution function of T_{2i} is interpreted as

$$F_2(t_2) = \Pr(T_{2i} \leq t_2 | T_{1i} < \infty, \mathbf{x}_i), \tag{3.5}$$

where $0 < t_2 < \infty$.

In this study, our goal is to estimate the effects of covariates on the marginal distribution of T_{2i}, for a given vector of fixed covariates $\mathbf{x}_i$ under settings where T_{1i} and T_{2i} may depend on each other. Copula models provide a strong framework for formulating the dependence between gap times, and can be utilized to estimate the marginal distribution of the second gap time. We therefore consider a parametric model for the joint distribution of T_{1i} and T_{2i} using a copula formulation for addressing the dependency between them. In this case, all accelerated failure time (AFT) regression model for $T_{ki}, k = 1, 2$, can be written in the following form.

$$T_{ki} = e^{\mathbf{x}_i^T \boldsymbol{\beta}_k + \epsilon_{ki}}, \tag{3.6}$$

where $p \times 1$ dimensional vector $\boldsymbol{\beta}_k$ includes the regression coefficients and ϵ_{ki} is the random error term for the kth gap time of the ith individual. The parameters $\boldsymbol{\beta}_k$ in the regression Model (3.6) can be estimated by using a maximum likelihood estimation procedure. The likelihood function pertaining to the analysis of only the first gap time is given by

$$L(\boldsymbol{\beta}_1) = \prod_{i=1}^{n} f_1(t_{1i}|\mathbf{x}_i; \boldsymbol{\beta}_1)^{\delta_{1i}} (1 - F_1(t_{1i}|\mathbf{x}_i; \boldsymbol{\beta}_1))^{1-\delta_{1i}}, \tag{3.7}$$

which uses the data $\{(t_{1i}, \delta_{1i}); \mathbf{x}_i;\ i = 1, 2, ..., n\}$. In this chapter, we only consider parametric specification of models for T_1 and T_2. Therefore, if there is no ambiguity, we drop the $\boldsymbol{\beta}_k$ notation from the likelihood functions and other related functions. The estimation of parameters in Model (3.6) can be obtained by maximizing the likelihood function (3.7) or its log likelihood function. The asymptotic properties of the estimator based on this method have been well-established (e.g. see Lawless, 2003, Chapter 6). These properties include the consistency of the maximum likelihood estimators of $\boldsymbol{\beta}_1$, as well as their regular asymptotic standard normal distribution results as the sample size n increases. We use the log likelihood function $\ell = \log L$, where L is given in (3.7) to obtain the estimate of parameters in the distribution of T_1. To do this, we use the "optim" function in R software, and obtain the estimates of parameters.

The likelihood function considering the data $\{(t_{1i}, t_{2i}, \delta_{1i}, \delta_{2i}); \mathbf{x}_i; i = 1, 2, ..., n\}$ can be expressed as follows.

$$L = \prod_{i=1}^{n} \Big[\frac{\partial^2 C(F_1(t_{1i}|\mathbf{x}_i), F_2(t_{2i}|T_{1i} < \infty, \mathbf{x}_i))}{\partial t_{1i} \partial t_{2i}}\Big]^{\delta_{1i}\delta_{2i}} \\ \times \Big[\frac{\partial F_1(t_{1i}|\mathbf{x}_i)}{\partial t_{1i}} - \frac{\partial C(F_1(t_{1i}|\mathbf{x}_i), F_2(t_{2i}|T_{1i} < \infty, \mathbf{x}_i))}{\partial t_{1i}}\Big]^{\delta_{1i}(1-\delta_{2i})} \Big[1 - F_1(t_{1i}|\mathbf{x}_i)\Big]^{(1-\delta_{1i})}, \tag{3.8}$$

where

$$C(u, v) = F(F_1^{-1}(u), F_2^{-1}(v)). \tag{3.9}$$

With the Clayton copula specification of (3.9); that is, $C_\phi(u, v) = \big(u^{-\phi} + v^{-\phi} - 1\big)^{-\frac{1}{\phi}}$, the likelihood function (3.8) can be written as follows.

$$L = \prod_{i=1}^{n} \Big[(\phi + 1) \big(F_1(t_{1i}|\mathbf{x}_i)^{-\phi} + F_2(t_{2i}|T_{1i} < \infty, \mathbf{x}_i)^{-\phi} - 1\big)^{-\frac{1}{\phi}-2} F_1(t_{1i}|\mathbf{x}_i)^{-\phi-1} f_1(t_{1i}|\mathbf{x}_i) \\ F_2(t_{2i}|T_{1i} < \infty, \mathbf{x}_i)^{-\phi-1} f_2(t_{2i}|T_{1i} < \infty, \mathbf{x}_i)\Big]^{\delta_{1i}\delta_{2i}} \\ \times \Big[f_1(t_{1i}|\mathbf{x}_i) + \frac{1}{\phi} \big(F_1(t_{1i}|\mathbf{x}_i)^{-\phi} + F_2(t_{2i}|T_{1i} < \infty, \mathbf{x}_i)^{-\phi} - 1\big)^{-\frac{1}{\phi}-1} F_1(t_{1i}|\mathbf{x}_i)^{-\phi-1} \\ f_1(t_{1i}|\mathbf{x}_i)\Big]^{\delta_{1i}(1-\delta_{2i})} \\ \times \Big[1 - F_1(t_{1i}|\mathbf{x}_i)\Big]^{(1-\delta_{1i})}. \tag{3.10}$$

The derivation of the likelihood function (3.10) can be found in Lawless and Yilmaz (2011). The parameters $\boldsymbol{\beta}_2$ in the marginal distribution of the second gap time and the copula parameter ϕ are obtained with a two stage estimation procedure. In the first stage, the parameters $\boldsymbol{\beta}_1$ in the marginal distribution of the first gap time T_1 are obtained by maximizing the log of the likelihood function (3.7). In the second stage, the estimates of the model parameters of the first gap time are plugged in the functions F_1 and f_1 given in (3.10). Then, the log of the resulting likelihood function was maximized to obtain the estimates of the regression model parameters $\boldsymbol{\beta}_2$ in the marginal distribution of the second gap time T_2 and the copula parameter ϕ in the Clayton copula. The consistency and asymptotic distribution of the estimators obtained with this estimation method have been discussed by Lawless and Yilmaz (2011). Based on an extensive simulation study, they showed that the estimators of $\boldsymbol{\beta}_2$ and ϕ are consistent and their limiting distributions are normal as n increases. In Section 3.3, we use this method to estimate the copula parameter in our models. To do this, we use the "optim" function in R software to maximize the log of the likelihood function (3.10), and obtain the estimate of the copula parameter ϕ.

3.2 Peng-Huang Approach for Parameter Estimation in Quantile Regression

As discussed in Chapter 1, classical regression models may not provide a complete understanding of the effects of covariates on the distribution of gap times. In contrast, quantile regression (QR) models offer a more flexible and robust way to examine covariate effects at various conditional quantiles of the gap times given the values of covariates. In this section, we introduce the method by Peng and Huang (2008), which we refer to as the Peng-Huang method, which was proposed to estimate parameters in QR models for survival data. For technical details, we will refer to the paper by Peng and Huang (2008), where the method was proposed. To explain this method, we use a similar notation introduced in the previous section. We would like to note that the Peng-Huang method is proposed for the analysis of the survival times and may be biased if it is used for the analysis of the second gap time, where there is dependence between the first and second gap times. In Section 3.5, we naively apply the Peng-Huang method for the analysis of the second gap time in a simulation study to discuss the bias in the estimates of model parameters.

As defined in Section 2.3, the conditional quantile of a random variable Y, given $\mathbf{x}$ and τ, is defined as

$$Q_Y(\tau|\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}(\tau), \qquad \tau \in (0,1), \tag{3.11}$$

where $\mathbf{x} = (1, \tilde{\mathbf{x}}^T)^T$ is a $p \times 1$ vector including the $(p-1) \times 1$ vector of covariates $\tilde{\mathbf{x}}$ and $\boldsymbol{\beta}(\tau) = (\beta_0(\tau), \beta_1(\tau), ..., \beta_{p-1}(\tau))^T$ is a $p \times 1$ vector of regression parameters representing the effects of covariates on the τth quantile of the response variable Y, allowing for potential changes with varying values of τ. To proceed, we define $z_1 = \min(T_1, C)$, where C is the censoring time. The observed data set consists of n independent and identically distributed (i.i.d.) replicates of $\{z_1, \delta_1, \mathbf{x}\}$, denoted by $\{(z_{1i}, \delta_{1i}); \mathbf{x}_i; i = 1, 2, ..., n\}$, where $z_{1i} = \min(T_{1i}, C)$. It is assumed that the censoring time C is independent of T_1 conditional on $\mathbf{x}$. Note that if we define the response variable Y as $Y = \log T_1$, where $T_1 = \exp\big(\mathbf{x}^T \boldsymbol{\beta}_{\mathbf{1}}(\tau) + \epsilon_1\big)$, the quantile regression model given in (3.11) can be equivalently expressed as

$$Q_{T_1}(\tau|\mathbf{x}) = \exp\big(\mathbf{x}^T \boldsymbol{\beta}_{\mathbf{1}}(\tau)\big), \qquad \tau \in (0, 1). \tag{3.12}$$

Model (3.12) can also be written as a standard linear model with the formula

$$\log T_1 = \mathbf{x}^T \boldsymbol{\beta}_{\mathbf{1}}(\tau) + \epsilon_1, \tag{3.13}$$

where ϵ_1 stands for a random error term such that $Q_{\epsilon_1}(\tau) = 0$.

The application of QR model to the survival data poses challenges due to the presence of censoring. However, one approach to estimate the parameters $\boldsymbol{\beta}_{\mathbf{1}}(\tau)$ in the quantile regression model given in (3.13) with censored data is to utilize the martingale theory developed for the analysis of counting process. A good source on this theorem from the survival analysis perspective is given by Aalen et al. (2008). Let $h_{T_1}(t_1|\mathbf{x})$, $t_1 > 0$, denotes the hazard function of the first gap time T_1 conditional on $\mathbf{x}$, where

$$\begin{aligned} \int_0^{t_1} h_{T_1}(u|\mathbf{x}) du &= -\log S(t_1|\mathbf{x}), \\ &= -\log(1 - \Pr(T_1 \leq t_1|\mathbf{x})), \\ &= -\log(\Pr(T_1 > t_1|\mathbf{x})). \end{aligned} \tag{3.14}$$

Following the discussion given in Section 2.1, we next define the counting process $\{N_1(t), t \geq 0\}$, where $N_1(t) = I(\min(T_1, C) \leq t, \delta_1 = 1)$ and the at-risk indicator function $Y_1(t) = I(T_1 \geq t, C \geq t)$. We assume that the value of $Y_1(t)$ is known at time t. The martingale process associated with the counting process $\{N_1(t), t \geq 0\}$

is denoted by $\{M_1(t), t \geq 0\}$, where $M_1(t) = N_1(t) - \Lambda_{T_1}(t|\mathbf{x})$ and $d\Lambda_{T_1}(t|\mathbf{x}) = Y_1(t)h_{T_1}(t|\mathbf{x})dt$.

Suppose that there are n independent such processes included in a study, each denoted by $\{N_{1i}(t), t \geq 0\}$, $i = 1, 2, ..., n$, with the associated intensity functions $\Lambda_{T_{1i}}(t|\mathbf{x}_i)$, where $\mathbf{x}_i = (1, x_{i1}, x_{i2}, ..., x_{i,p-1})^T$. Note that, in this case, $\Lambda_{T_{1i}}(t|\mathbf{x}_i)$ can be written as $\Lambda_{T_{1i}}(\min(t, z_{1i})|\mathbf{x}_i)$. Consequently, $M_{1i}(t)$ is a martingale associated with the counting process $\{N_{1i}(t), t \geq 0\}$, $i = 1, 2, ..., n$. It can be shown that $E[M_{1i}(t)|\mathbf{x}_i] = 0$ for $t \geq 0$ (Aalen et al., 2008). We then obtain

$$E\left\{ n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{x}_i M_{1i}(t) \middle| \mathbf{x}_i \right\} = \mathbf{0}, \quad t \geq 0. \tag{3.15}$$

Since $\exp\big(\mathbf{x}_i^T \boldsymbol{\beta}_\mathbf{1}^*(\tau)\big) > 0$, where $\boldsymbol{\beta}_\mathbf{1}^*(\tau)$ denotes the true value of $\boldsymbol{\beta}_\mathbf{1}(\cdot)$ in Model (3.12), we have

$$E\left\{ n^{-\frac{1}{2}} \sum_{i=1}^{n} \left[\mathbf{x}_i M_{1i}\big(e^{\mathbf{x}_i^T \boldsymbol{\beta}_\mathbf{1}^*(\tau)}\big) \middle| \mathbf{x}_i \right] \right\} = \mathbf{0}, \quad \tau \in (0, 1). \tag{3.16}$$

Therefore, for a given τ, $\tau \in (0, 1)$, we obtain a $p \times 1$ system of unbiased estimating equations given by $n^{\frac{1}{2}} \mathbf{S_1}(\tau, \boldsymbol{\beta}_\mathbf{1}) = \mathbf{0}$, where $\mathbf{0}$ is a $p \times 1$ vector of zeros and

$$\mathbf{S_1}(\tau, \boldsymbol{\beta}_\mathbf{1}^*) = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \left[N_{1i}\big(e^{\mathbf{x}_i^T \boldsymbol{\beta}_\mathbf{1}^*(\tau)}\big) - \Lambda_{T_1}(\min\left\{e^{\mathbf{x}_i^T \boldsymbol{\beta}_\mathbf{1}^*(\tau)}, z_{1i}\right\} \middle| \mathbf{x}_i) \right]. \tag{3.17}$$

Consequently, the following unbiased estimating equations, forming a $p \times 1$ system of equations, can be simultaneously solved to obtain the estimate of $\boldsymbol{\beta}_\mathbf{1}^*(\tau)$.

$$\begin{aligned}
n^{-\frac{1}{2}} \sum_{i=1}^{n} \left[N_{1i}\big(e^{\mathbf{x}_i^T \boldsymbol{\beta}_\mathbf{1}^*(\tau)}\big) - \Lambda_{T_1}(\min\left\{e^{\mathbf{x}_i^T \boldsymbol{\beta}_\mathbf{1}^*(\tau)}, z_{1i}\right\} \middle| \mathbf{x}_i) \right] &= 0, \\
n^{-\frac{1}{2}} \sum_{i=1}^{n} x_{i1} \left[N_{1i}\big(e^{\mathbf{x}_i^T \boldsymbol{\beta}_\mathbf{1}^*(\tau)}\big) - \Lambda_{T_1}(\min\left\{e^{\mathbf{x}_i^T \boldsymbol{\beta}_\mathbf{1}^*(\tau)}, z_{1i}\right\} \middle| \mathbf{x}_i) \right] &= 0, \\
n^{-\frac{1}{2}} \sum_{i=1}^{n} x_{i2} \left[N_{1i}\big(e^{\mathbf{x}_i^T \boldsymbol{\beta}_\mathbf{1}^*(\tau)}\big) - \Lambda_{T_1}(\min\left\{e^{\mathbf{x}_i^T \boldsymbol{\beta}_\mathbf{1}^*(\tau)}, z_{1i}\right\} \middle| \mathbf{x}_i) \right] &= 0, \\
&\vdots \\
n^{-\frac{1}{2}} \sum_{i=1}^{n} x_{i,p-1} \left[N_{1i}\big(e^{\mathbf{x}_i^T \boldsymbol{\beta}_\mathbf{1}^*(\tau)}\big) - \Lambda_{T_1}(\min\left\{e^{\mathbf{x}_i^T \boldsymbol{\beta}_\mathbf{1}^*(\tau)}, z_{1i}\right\} \middle| \mathbf{x}_i) \right] &= 0.
\end{aligned} \tag{3.18}$$

Note that using the definition of the cumulative hazard function of T_1,

$$\int_0^t h_{T_1}(u|\mathbf{x})du = -\log(1 - F_1(t|\mathbf{x})), \tag{3.19}$$

we can write the $p \times 1$ vector of unbiased estimating functions given in (3.17) in the following form.

$$\mathbf{S_1}(\tau, \boldsymbol{\beta}_\mathbf{1}^*) = n^{-1}\sum_{i=1}^n \mathbf{x}_i\Big[N_{1i}\big(e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)}\big) + \log\Big(1 - F_1(\min\big\{e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)}, z_{1i}\big\}\Big|\mathbf{x}_i)\Big)\Big]. \tag{3.20}$$

Peng and Huang (2008) proposed to use the functions defined in (3.20) to estimate the parameters $\boldsymbol{\beta}_\mathbf{1}^*(\tau)$ of the QR model given in (3.12). They developed an algorithm based on the martingale estimating equations to minimize an L_1-type convex function. We next explain this method in more detail.

From the definition of QR, we have

$$F_1\big(e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)}\big|\mathbf{x}_i\big) = \tau, \tag{3.21}$$

where $\tau \in (0, 1)$. From the relation given in (3.14), we have

$$\Lambda_{T_1}\Big(\min\big\{e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)}, z_{1i}\big\}\Big|\mathbf{x}_i\Big) = -\log\Bigg\{1 - \Pr\Big(T_1 \le \min\big\{e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)}, z_{1i}\big\}\Big|\mathbf{x}_i\Big)\Bigg\}. \tag{3.22}$$

Note that, in the probability given on the right hand side of (3.22), if $e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)} \le z_{1i}$, we have

$$\Pr\Big(T_1 \le e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)}\Big) \le \Pr(T_1 \le z_{1i}), \tag{3.23}$$

which gives

$$-\log\Big(1 - \Pr\Big(T_1 \le e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)}\big|\mathbf{x}_i\Big)\Big) \le -\log\Big(1 - \Pr\big(T_1 \le z_{1i}\big|\mathbf{x}_i\big)\Big). \tag{3.24}$$

On the other hand, when $z_{1i} \le e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)}$, we have

$$\Pr(T_1 \le z_{1i}) \le \Pr\Big(T_1 \le e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)}\Big), \tag{3.25}$$

which gives

$$-\log\Big(1-\Pr\big(T_1 \le z_{1i}\big|\mathbf{x}_i\big)\Big) \le -\log\Big(1-\Pr\Big(T_1 \le e^{\mathbf{x}_i^T\boldsymbol{\beta}_{\mathbf{1}}^{*}(\tau)}\big|\mathbf{x}_i\Big)\Big). \tag{3.26}$$

By incorporating (3.24) and (3.26) into (3.22), we obtain the following result.

$$\begin{aligned}
\Lambda_{T_1}\Big(\min\big\{e^{\mathbf{x}_i^T\boldsymbol{\beta}_{\mathbf{1}}^{*}(\tau)}, z_{1i}\big\}\Big|\mathbf{x}_i\Big) &= \min\Big\{-\log\Big(1-\Pr\Big(T_1 \le e^{\mathbf{x}_i^T\boldsymbol{\beta}_{\mathbf{1}}^{*}(\tau)}\big|\mathbf{x}_i\Big)\Big),\\
&\quad -\log\Big(1-\Pr\big(T_1 \le z_{1i}\big|\mathbf{x}_i\big)\Big)\Big\},\\
&= \min\Big\{-\log\big(1-\tau\big), -\log\Big(1-F_{T_1}(z_{1i}\big|\mathbf{x}_i)\Big)\Big\},\\
&= \min\Big\{G(\tau), G(F_{T_1}(z_{1i}\big|\mathbf{x}_i))\Big\},\\
&= \int_0^{\min\{\tau, F_1(z_{1i}\big|\mathbf{x}_i)\}} dG(u),\\
&= \int_0^{\tau} I[u \le F_1(z_{1i}\big|\mathbf{x}_i)]dG(u),\\
&= \int_0^{\tau} I[z_{1i} \ge F_1{}^{-1}(u\big|\mathbf{x}_i)]dG(u),\\
&= \int_0^{\tau} I[z_{1i} \ge Q_{T_1}(u\big|\mathbf{x}_i)]dG(u),\\
&= \int_0^{\tau} I[z_{1i} \ge e^{\mathbf{x}_i^T\boldsymbol{\beta}_{\mathbf{1}}^{*}(u)}]dG(u),
\end{aligned} \tag{3.27}$$

where $G(x) = -\log(1-x)$ for $0 \le x < 1$. Note that $G(x)$ is a strictly increasing function of x. Substituting (3.27) into (3.17), the $p \times 1$ vector of unbiased estimating functions becomes

$$\mathbf{S_1}(\boldsymbol{\beta}_{\mathbf{1}}^{*}, \tau) = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\Big[N_{1i}\big(e^{\mathbf{x}_i^T\boldsymbol{\beta}_{\mathbf{1}}^{*}(\tau)}\big) - \int_0^{\tau} I[z_{1i} \ge e^{\mathbf{x}_i^T\boldsymbol{\beta}_{\mathbf{1}}^{*}(u)}]dG(u)\Big], \tag{3.28}$$

which can be used to estimate $\boldsymbol{\beta}_{\mathbf{1}}^{*}(\tau)$.

To deal with censored data, the Peng-Huang method uses the grid-based estimation procedure to estimate $\boldsymbol{\beta}_{\mathbf{1}}(\tau)$. This procedure involves discretizing the covariate space into a grid of values, and estimating the regression quantiles at these grid points. In this method, the attention is to the estimation of $\{\boldsymbol{\beta}_{\mathbf{1}}(\tau); \tau \in (0, \tau_u)\}$, where $\tau_u \in (0,1)$, instead of attempting to estimate $\boldsymbol{\beta}_{\mathbf{1}}(\tau)$ for all value of $\tau \in (0,1)$.

The quantity τ_u, $\tau_u \in (0,1)$, is a deterministic constant subject to certain identifiability constraints due to censoring. In other words, with the Peng and Huang method, the parameters in Model (3.12) are identifiable up to a certain quantile point τ_u. The value of τ_u can only be determined from a given data set. It should be noted that the *crq*() function in the package *quantreg* in R software provides the estimation of $\hat{\boldsymbol{\beta}}_{\mathbf{1}}(\tau)$ only up to the quantile τ_u. Under this constraint, the Peng-Huang method utilizes a grid-based estimation method to estimate $\boldsymbol{\beta}_{\mathbf{1}}(\tau)$ at discrete points, specifically $\{\tau_j; j = 0, 1, ..., L_{(n)}\}$, on the grid $||S_{L_{(n)}}|| = \{0 = \tau_0 < \tau_1 < \tau_2 < ... < \tau_{L_{(n)}}\}$. Their approach initiates by solving for $\boldsymbol{\beta}_{\mathbf{1}}(0)$ using the equation $\exp\big(\mathbf{x}_i^T\boldsymbol{\beta}_{\mathbf{1}}(0)\big) = 0$. Subsequently, for each τ_j within $\{\tau_j; j = 0, 1, ..., L_{(n)}\}$, a set of unbiased estimating equations is constructed as follows.

$$
\begin{aligned}
n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{x}_i \Big[N_{1i}(e^{\mathbf{x}_i^T \boldsymbol{\beta}^*_{\mathbf{1}}(\tau_1)}) - \int_0^{\tau_1} I[z_{1i} \geq e^{\mathbf{x}_i^T \boldsymbol{\beta}^*_{\mathbf{1}}(u)}] dG(u) \Big] &= 0, \\
n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{x}_i \Big[N_{1i}(e^{\mathbf{x}_i^T \boldsymbol{\beta}^*_{\mathbf{1}}(\tau_2)}) - \int_0^{\tau_2} I[z_{1i} \geq e^{\mathbf{x}_i^T \boldsymbol{\beta}^*_{\mathbf{1}}(u)}] dG(u) \Big] &= 0, \\
&\vdots \\
n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{x}_i \Big[N_{1i}(e^{\mathbf{x}_i^T \boldsymbol{\beta}^*_{\mathbf{1}}(\tau_{L_{(n)}})}) - \int_0^{\tau_{L_{(n)}}} I[z_{1i} \geq e^{\mathbf{x}_i^T \boldsymbol{\beta}^*_{\mathbf{1}}(u)}] dG(u) \Big] &= 0. \qquad (3.29)
\end{aligned}
$$

Consequently, the unbiased estimating functions for $j = 1, 2, ..., L_{(n)}$ are formulated as

$$
\mathbf{S}_{\mathbf{1}}(\boldsymbol{\beta}_{\mathbf{1}}, \tau_j) = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \Big[N_{1i}(e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{1}}(\tau_j)}) - \int_0^{\tau_j} I[z_{1i} \geq e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{1}}(u)}] dG(u) \Big]. \qquad (3.30)
$$

Note that the integral in (3.30) can be written as

$$
\begin{aligned}
\int_0^{\tau_j} I[z_{1i} \geq e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{1}}(u)}] dG(u) &= \sum_{k=0}^{j-1} \int_{\tau_k}^{\tau_{k+1}} I[z_{1i} \geq e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{1}}(u)}] dG(u), \\
&\approx \sum_{k=0}^{j-1} \int_{\tau_k}^{\tau_{k+1}} I[z_{1i} \geq e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{1}}(\tau_k)}] dG(u), \\
&= \sum_{k=0}^{j-1} I[z_{1i} \geq e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{1}}(\tau_k)}] \big(G(\tau_{k+1}) - G(\tau_k) \big). \qquad (3.31)
\end{aligned}
$$

Thus, a sequence of unbiased estimating equations can be established by setting

$$n^{\frac{1}{2}}\mathbf{S_1}(\boldsymbol{\beta_1},\tau_j)=\mathbf{0},\quad j=1,2,...,L_{(n)}, \tag{3.32}$$

which implies that the equations in

$$n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\Big\{N_{1i}(e^{\mathbf{x}_i^T\boldsymbol{\beta_1}(\tau_j)})-\sum_{k=0}^{j-1}I[z_{1i}\geq e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta_1}}(\tau_k)}]\big(G(\tau_{k+1})-G(\tau_{k+1})\big)\Big\}=\mathbf{0}, \tag{3.33}$$

are solved in a sequential manner for $\boldsymbol{\beta_1}(\tau_j)$. As discussed by Peng and Huang (2008), due to the non-continuity of the equations in (3.33), an exact root of (3.33) might not exist. Thus, $\hat{\boldsymbol{\beta_1}}(\tau_j)$, $j=1,2,...,L_{(n)}$, is defined as a generalized solution of the system of equations with components given in (3.29). Furthermore, all functions in $\mathbf{S_1}(\boldsymbol{\beta},\tau_j)$ are monotonically non-decreasing, and thus, they represent the gradient of a convex function. Consequently, the root of the vector $\mathbf{S_1}(\boldsymbol{\beta_1},\tau_j)$ corresponds to the minimizer of the vector of convex functions (Peng and Huang, 2008).

Convex optimization is obtained as follows. Since the absolute function $f(x)=|x|$ is convex, we have

$$\begin{aligned}N_{1i}(e^{\mathbf{x}_i^T\boldsymbol{\beta_1}(\tau)})&=I(z_{1i}\leq e^{\mathbf{x}_i^T\boldsymbol{\beta_1}(\tau)},\delta_{1i}=1),\\&=I(\log z_{1i}\leq \mathbf{x}_i^T\boldsymbol{\beta_1}(\tau),\delta_{1i}=1),\\&=I(\log z_{1i}-\mathbf{x_i}^T\boldsymbol{\beta_1}(\tau)\leq 0,\delta_{1i}=1),\end{aligned} \tag{3.34}$$

which allows us to express the convex function for $N_{1i}(e^{\mathbf{x}_i^T\boldsymbol{\beta_1}(\tau)})$ by considering $\boldsymbol{\beta_1}(\tau)=\mathbf{h}$ as

$$|\delta_{1i}\log z_{1i}-\delta_{1i}\mathbf{x}_i^T\mathbf{h}|. \tag{3.35}$$

Taking this result into account and assuming a very large constant R^*, the convex function for $\mathbf{S_1}(\boldsymbol{\beta_1},\tau)$ becomes

$$\begin{aligned}L_j(\mathbf{h})=&\sum_{i=1}^{n}|\delta_{1i}\log z_{1i}-\delta_{1i}\mathbf{x}_i^T\mathbf{h}|+|R^*+\sum_{i=1}^{n}\delta_{1i}\mathbf{x}_i^T\mathbf{h}|\\&+|R^*-2\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{h}\sum_{k=0}^{j-1}I(z_{1i}\geq e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_1(\tau_k)})[G(\tau_{k+1})-G(\tau_k)]|.\end{aligned} \tag{3.36}$$

In the case where R^* is a large constant, $L_j(\mathbf{h})$ given in (3.36) is simplified to

$$\begin{aligned} L_j(\mathbf{h}) =& \sum_{i=1}^{n} \big\{ \delta_{1i} | \log z_{1i} - \mathbf{x}_i^T \mathbf{h} | + \delta_{1i} \mathbf{x}_i^T \mathbf{h} \\ & - 2\mathbf{x}_i^T \mathbf{h} \sum_{k=0}^{j-1} I(z_{1i} \geq e^{\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_1(\tau_k)})[G(\tau_{k+1}) - G(\tau_k)] \big\}, \\ =& \sum_{i=1}^{n} \big\{ \delta_{1i} | \log z_{1i} - \sum_{r=1}^{p} x_{ir} h_r | + \delta_{1i} \sum_{r=1}^{p} x_{ir} h_r \\ & - 2 \sum_{r=1}^{p} x_{ir} h_r \sum_{k=0}^{j-1} I(z_{1i} \geq e^{\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_1(\tau_k)})[G(\tau_{k+1}) - G(\tau_k)] \big\}. \end{aligned} \tag{3.37}$$

Next, for $r = 1, 2, ..., p$, we can differentiate (3.35) as follows.

$$\begin{aligned} \frac{\partial}{\partial h_r} | \log z_{1i} - \mathbf{x}_i^T \mathbf{h} | &= \frac{\partial}{\partial h_r} \begin{cases} \log z_{1i} - \mathbf{x}_i^T \mathbf{h}, & \text{if } z_{1i} \geq e^{\mathbf{x}_i^T \mathbf{h}}, \\ -\log z_{1i} + \mathbf{x}_i^T \mathbf{h}, & \text{if } z_{1i} < e^{\mathbf{x}_i^T \mathbf{h}}, \end{cases} \\ &= \begin{cases} -x_{ir}, & \text{if } z_{1i} \geq e^{\mathbf{x}_i^T \mathbf{h}}, \\ x_{ir}, & \text{if } z_{1i} < e^{\mathbf{x}_i^T \mathbf{h}}, \end{cases} \\ &= -x_{ir} I(z_{1i} \geq e^{\mathbf{x}_i^T \mathbf{h}}) + x_{ir} I(z_{1i} < e^{\mathbf{x}_i^T \mathbf{h}}), \end{aligned} \tag{3.38}$$

and

$$\frac{\partial}{\partial h_r} \mathbf{x}_i^T \mathbf{h} = x_{ir}. \tag{3.39}$$

From the results given in (3.38) and (3.39) for $r = 1, 2, .., p$, we find

$$\begin{aligned} \frac{\partial}{\partial h_r} \big[\delta_{1i} | \log z_{1i} - \mathbf{x}_i^T \mathbf{h} | + \delta_{1i} \mathbf{x_i}^T \mathbf{h} \big] &= -\delta_{1i} x_{ir} I(z_{1i} \geq e^{\mathbf{x}_i^T \mathbf{h}}) + \delta_{1i} x_{ir} I(z_{1i} < e^{\mathbf{x}_i^T \mathbf{h}}) + \delta_{1i} x_{ir}, \\ &= -\delta_{1i} x_{ir} I(z_{1i} \geq e^{\mathbf{x}_i^T \mathbf{h}}) + \delta_{1i} x_{ir} I(z_{1i} < e^{\mathbf{x}_i^T \mathbf{h}}) \\ &\quad + \delta_{1i} x_{ir} I(z_{1i} \geq e^{\mathbf{x}_i^T \mathbf{h}}) + \delta_{1i} x_{ir} I(z_{1i} < e^{\mathbf{x}_i^T \mathbf{h}}), \\ &= 2\delta_{1i} x_{ir} I(z_{1i} < e^{\mathbf{x}_i^T \mathbf{h}}), \\ &= 2 x_{ir} N_{1i}(e^{\mathbf{x}_i^T \mathbf{h}}). \end{aligned} \tag{3.40}$$

So, when we replace $\mathbf{h}$ with $\boldsymbol{\beta_1}(\tau)$ in (3.40), we obtain

$$\frac{\partial}{\partial h_r}\big[\delta_{1i}|\log z_{1i} - \mathbf{x}_i^T\mathbf{h}| + \delta_{1i}\mathbf{x}_i^T\mathbf{h}\big]\Big|_{\mathbf{h}=\boldsymbol{\beta_1}(\tau)} = 2x_{ir}N_{1i}(e^{\mathbf{x}_i^T\boldsymbol{\beta_1}(\tau)}). \tag{3.41}$$

As a result, we obtain that

$$\begin{aligned}\frac{\partial}{\partial h_r}L_j(\mathbf{h})\Big|_{\mathbf{h}=\boldsymbol{\beta_1}(\tau)} =& \sum_{i=1}^{n}\Big\{2x_{ir}N_{1i}(e^{\mathbf{x}_i^T\boldsymbol{\beta_1}(\tau)}) \\ &\qquad - 2x_{ir}\sum_{k=0}^{j-1} I(z_{1i} \geq e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_1(\tau_k)})[G(\tau_{k+1}) - G(\tau_k)]\Big\}, \\ =& \, 2\sum_{i=1}^{n} x_{ir}\Big\{N_{1i}(e^{\mathbf{x}_i^T\boldsymbol{\beta_1}(\tau)}) \\ &\qquad - \sum_{k=0}^{j-1} I(z_{1i} \geq e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_1(\tau_k)})[G(\tau_{k+1}) - G(\tau_k)]\Big\}. \end{aligned} \tag{3.42}$$

Consequently, the root of the derivative of the convex function $L_j(\mathbf{h})$ with respect to h_r corresponds to the root of $\mathbf{S_1}(\boldsymbol{\beta_1}, \tau_j)$. This result implies that the minimizer of the L_1-type convex objective function $L_j(h)$ serves as the root of $\mathbf{S_1}(\boldsymbol{\beta_1}, \tau_j)$, which is given by $\hat{\boldsymbol{\beta}}_1(\tau_j)$.

To estimate the variance of the estimators, Peng and Huang (2008) apply the resampling approach based on the technique developed by Jin et al. (2001). This resampling approach allows them to estimate the asymptotic variance of the estimators, providing insights into the precision and uncertainty associated with the estimators in the context of regression quantile processes. The consistency and asymptotic distribution of the estimators obtained with the Peng-Huang method have been also discussed by Peng and Huang (2008). They showed the uniform consistency and weak convergence of the estimated regression quantile process by providing four regularity conditions. It is worth noting that estimates of model parameters $\boldsymbol{\beta}_1(\tau)$, variance estimates of them, and coverage probabilities can be obtained with the *crq*() function in the package *quantreg* in R software.

In section 3.5.2, we naively apply the Peng-Huang method for the analysis of second gap time. In the last part of this section, we extend our notation to deal with this setting. To initiate our analysis, we define z_2 as the minimum of T_2 and $C - T_1$; that is, $z_2 = \min(T_2, C - T_1)$, where C represents the censoring time. In this context, the observed data set includes n i.i.d. replicates denoted by $\{(t_{1i}, z_{2i}, \delta_{1i}, \delta_{2i}); \mathbf{x}_i;\ i = 1, 2, ..., n\}$. The QR model for the second gap time T_2 is given by

$$Q_{T_2}(\tau|\mathbf{x}) = \exp\bigl(\mathbf{x}^T \boldsymbol{\beta}_{\mathbf{2}}(\tau)\bigr),\ \text{for } \tau \in (0, 1), \tag{3.43}$$

where $\boldsymbol{\beta}_{\mathbf{2}}(\tau) = (\beta_{02}(\tau), \beta_{12}(\tau), ..., \beta_{p-1,2}(\tau))$ represents the effect of covariates on $\log T_2$. Specifically

$$\log T_2 = \beta_{02}(\tau) + \beta_{12}(\tau)x_1 + \beta_{22}(\tau)x_2 + ... + \beta_{p-1,2}(\tau)x_{p-1} + \epsilon_2, \tag{3.44}$$

where ϵ_2 is the error term such that $Q_{\epsilon_2}(\tau|\mathbf{x}) = 0$.

Next we define the counting process for the second event as $\{N_2(t), t \geq 0\}$, where $N_2(t) = I(z_2 \leq t, \delta_1 = 1, \delta_2 = 1)$, which counts the second type of events. The martingale process associated with the counting process of the second event is denoted as $\{M_2(t), t \geq 0\}$, where

$$M_2(t) = N_2(t) - \Lambda_{2|1}(\min(t, z_2)|T_1 = t_1, \mathbf{x}). \tag{3.45}$$

We discuss this martingale structure in the next section in more detail. Similar to the first gap time, we derive functions for estimating $\boldsymbol{\beta}_{\mathbf{2}}(\tau)$. The $p \times 1$ system of unbiased estimating functions is given by

$$\mathbf{S_2}(\tau, \boldsymbol{\beta}_2^*) = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \Bigl[N_{2i}(e^{\mathbf{x}_i^T \boldsymbol{\beta}_2^*(\tau)}) - \Lambda_{2|1}(\min\bigl\{e^{\mathbf{x}_i^T \boldsymbol{\beta}_2^*(\tau)}, z_{2i}\bigr\}\Big| t_{1i}, \mathbf{x}_i)\Bigr], \tag{3.46}$$

where $\boldsymbol{\beta}_2^*(\tau)$ is the true value of $\boldsymbol{\beta}_{\mathbf{2}}(\tau)$ and $E(n^{\frac{1}{2}}\mathbf{S_2}(\tau, \boldsymbol{\beta}_2^*)) = \mathbf{0}$. Similar to the Peng-Huang method explained for T_1, the system of estimating functions for solving $\boldsymbol{\beta}_2^*(\tau)$ can be expressed as

$$\mathbf{S_2}(\tau_j, \hat{\boldsymbol{\beta}}_2) = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \Bigl\{N_{2i}(e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{2}}(\tau_j)}) - \sum_{k=0}^{j-1} I\bigl(z_{2i} \geq e^{\mathbf{x_i^T} \hat{\boldsymbol{\beta}}_2(\tau_k)}\bigr)[G_{T_2}(\tau_{k+1}) - G_{T_2}(\tau_k)]\Bigr\}, \tag{3.47}$$

for $j = 1, 2, ..., L$. Once again the task of finding solutions for $\mathbf{S_2}(\tau_j, \hat{\boldsymbol{\beta}}_2)$ is equivalent to locating the minimizer of the following L_1-type convex objective function given by

$$\begin{aligned} L_j(\mathbf{h}) = & \sum_{i=1}^{n} |\delta_{2i} \log z_{2i} - \delta_{2i}\mathbf{x}_i^T\mathbf{h}| + |M^* + \sum_{i=1}^{n} \delta_{2i}\mathbf{x}_i^T\mathbf{h}| \\ & + |M^* - 2\sum_{i=1}^{n} \mathbf{x}_i^T\mathbf{h} \sum_{k=0}^{j-1} I(z_{2i} \geq e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_2(\tau_k)})[G_{T_2}(\tau_{k+1}) - G_{T_2|}(\tau_k)]|, \end{aligned} \tag{3.48}$$

where M^* represents a significantly large value. The solution to this minimization problem can be easily obtained using the *crq*() function in R package *quantreg*.

We would like to note that solving the system of unbiased equations given in (3.46) with the Peng-Huang method gives the estimates of parameters in the conditional cumulative intensity function

$$\Lambda_{2|1}(t|T_1 = t_1, \mathbf{x}), \qquad t > 0. \tag{3.49}$$

In this case, the interpretation of the estimates is not straightforward as it is based on the value of a given $T_1 = t_1$. However, if the goal is to estimate the effect of covariates $\mathbf{x}$ on the quantile of the marginal distribution of the second gap time T_2, the Peng-Huang method only provides unbiased results of

$$\Lambda_{2|1}(t|T_1 = t_1, \mathbf{x}) = \Lambda_2(t|\mathbf{x}), \tag{3.50}$$

for all $t > 0$. In essence, the result (3.50) holds true if T_1 and T_2 are independent. In the next section, we discuss a method that can address this issue even when T_1 and T_2 are not independent. In Section 3.5.2, we naively apply the Peng-Huang method to estimate the effects of covariates on the conditional quantile of the marginal distribution of T_2, given the value of the covariates $\mathbf{x}$, when T_1 and T_2 are not independent.

3.3 Parametric Estimation for Quantile Regression of Sequential Gap Times with Copulas

The theory of martingales associated with counting processes provides powerful tools to estimate model parameters in QR. For example, as discussed in the previous section, parameters related to the distribution of the first gap time in sequentially observed bivariate gap times can be estimated with the Peng-Huang method. However, this method may result in biased estimates of the parameters in the marginal distribution of the second gap time when the two gap times are not independent. In this section, we introduce an approach to estimate the parameters in the marginal distribution of the first and second gap times for the quantile regression method. This method is based on the theory of generalized estimating equations, which is similar to the Peng-Huang method in a way that it provides a system of estimating equations. Different from the Peng-Huang method, we utilize the Newton-Raphson algorithm to solve the system of equations to obtain the estimates of model parameters and their variance estimates. The Newton-Raphson algorithm is a widely-used iterative numerical technique for approximating solutions to equations (Lange et al., 2010, Chapter 14). It should be noted that the application of the Newton-Raphson algorithm for estimation purposes is a major change comparing with the Peng-Huang method, which is based on a grid type estimation procedure. We also apply the Newton-Raphson algorithm to the marginal distribution of the second gap time in our estimation procedure.

We first explain our parameter estimation method for the marginal distribution of the first gap time. More specifically, our goal here is to estimate the parameters $\boldsymbol{\beta_1}(\tau)$ in the conditional QR model for the marginal distribution of the first gap time, given a set of covariate values; that is,

$$Q_{T_1}(\tau|\mathbf{x}) = \exp\bigl(\mathbf{x}^T\boldsymbol{\beta_1}(\tau)\bigr), \qquad \tau \in (0,1), \tag{3.51}$$

where $\mathbf{x} = (1, \tilde{\mathbf{x}}^T)^T$ is a $p \times 1$ vector including the $(p-1) \times 1$ vector of covariates $\tilde{\mathbf{x}}$ and $\boldsymbol{\beta_1}(\tau) = (\beta_{01}(\tau), \beta_{11}(\tau), ..., \beta_{p-1,1}(\tau))^T$ is a $p \times 1$ vector of regression parameters representing the effects of covariates on the τth quantile of the log of the first gap time T_1, allowing for potential changes with varying values of τ.

We assume n independent individuals subject to at most two sequentially observed events. We let $\Lambda_{T_{1i}}(t|\mathbf{x}_i)$ be the cumulative intensity function of the counting process $\{N_{1i}(t), t \geq 0\}$, $i = 1, 2, ..., n$, in which $N_{1i}(t)$ takes the value of 1 if the first type of the event occurs in $(0, t]$. Otherwise, it is equal to 0. Note that, as in the Peng-Huang method, the martingale structure is preserved for the quantities $M_{1i}(t) = N_{1i}(t) - \Lambda_{T_{1i}}(\min(t, z_{1i})|\mathbf{x}_i)$, where $z_{1i} = \min(T_{1i}, C_i)$ and $i = 1, 2, ..., n$. It follows from this result that

$$E\Bigg\{n^{-\frac{1}{2}}\sum_{i=1}^{n}\Big[\mathbf{x}_i M_{1i}(e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)})\Big|\mathbf{x}_i\Big]\Bigg\}=\mathbf{0}, \tag{3.52}$$

where $\boldsymbol{\beta}_\mathbf{1}^*(\tau)$ denotes the true value of $\boldsymbol{\beta}_\mathbf{1}(\cdot)$ in Model (3.51), $\tau\in(0,1)$ and $\mathbf{0}$ is a $p\times 1$ vector of zeros. From the above result and the martingales $M_{1i}(t)$, we obtain the $p\times 1$ vector of unbiased estimating functions

$$\mathbf{S_1}(\tau,\boldsymbol{\beta}_\mathbf{1}^*)=n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\Big[N_{1i}\big(e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)}\big)-\Lambda_{T_1}(\min\big\{e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)},z_{1i}\big\}\Big|\mathbf{x}_i)\Big], \tag{3.53}$$

which provides a $p\times 1$ system of unbiased estimating equations with the components

$$\begin{aligned}
&n^{-\frac{1}{2}}\sum_{i=1}^{n}\Big[N_{1i}\big(e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)}\big)-\Lambda_{T_1}(\min\big\{e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)},z_{1i}\big\}\Big|\mathbf{x}_i)\Big]=0,\\
&n^{-\frac{1}{2}}\sum_{i=1}^{n}x_{i1}\Big[N_{1i}\big(e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)}\big)-\Lambda_{T_1}(\min\big\{e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)},z_{1i}\big\}\Big|\mathbf{x}_i)\Big]=0,\\
&n^{-\frac{1}{2}}\sum_{i=1}^{n}x_{i2}\Big[N_{1i}\big(e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)}\big)-\Lambda_{T_1}(\min\big\{e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)},z_{1i}\big\}\Big|\mathbf{x}_i)\Big]=0,\\
&\qquad\qquad\vdots\\
&n^{-\frac{1}{2}}\sum_{i=1}^{n}x_{i,p-1}\Big[N_{1i}\big(e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)}\big)-\Lambda_{T_1}(\min\big\{e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)},z_{1i}\big\}\Big|\mathbf{x}_i)\Big]=0.
\end{aligned} \tag{3.54}$$

Since

$$\Lambda_{T_1}(t|\mathbf{x})=\int_0^t h_{T_1}(u|\mathbf{x})du=-\log(1-F_1(t|\mathbf{x})),\quad t>0, \tag{3.55}$$

we can write the $p\times 1$ vector of estimating functions given in (3.53) as follows.

$$\begin{aligned}
\mathbf{S_1}(\tau,\boldsymbol{\beta}_\mathbf{1}^*)=n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\Big[&N_{1i}\big(e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)}\big)\\
&+\log\Big(1-F_1(\min\big\{e^{\mathbf{x}_i^T\boldsymbol{\beta}_\mathbf{1}^*(\tau)},z_{1i}\big\}\Big|\mathbf{x}_i)\Big)\Big].
\end{aligned} \tag{3.56}$$

Therefore, for a p-dimensional problem, the system of unbiased estimating equations $\mathbf{S_1}(\tau,\boldsymbol{\beta}_\mathbf{1}^*)=\mathbf{0}$ includes p functions, where $\mathbf{S_1}(\tau,\boldsymbol{\beta}_\mathbf{1}^*)=[S_{11}(\tau,\boldsymbol{\beta}_\mathbf{1}^*),S_{12}(\tau,\boldsymbol{\beta}_\mathbf{1}^*),...,$

$S_{1p}(\tau, \boldsymbol{\beta}_{\mathbf{1}}^{*})]^T$ and each component $S_{1j}(\tau, \boldsymbol{\beta}_{\mathbf{1}}^{*})$, for $j = 0, 1, .., p-1$, is defined as

$$S_{1j}(\tau, \boldsymbol{\beta}_{\mathbf{1}}^{*}) = n^{-1} \sum_{i=1}^{n} x_{ij} \Big[N_{1i}\big(e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{1}}^{*}(\tau)}\big) \\ + \log\Big(1 - F_1(\min\big\{e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{1}}^{*}(\tau)}, z_{1i}\big\}\Big|\mathbf{x}_i)\Big)\Big], \tag{3.57}$$

where $x_{i0} = 1$.

The goal is to simultaneously solve the equations in $\mathbf{S_1}(\tau, \boldsymbol{\beta}_{\mathbf{1}}^{*}) = \mathbf{0}$ for the elements in the p-dimensional vector $\boldsymbol{\beta}_1(\tau) = (\beta_{01}(\tau), \beta_{11}(\tau), ..., \beta_{p-1,1}(\tau))^T$. To achieve this, we apply the Newton-Raphson algorithm, which iteratively updates the estimation of $\boldsymbol{\beta}_{\mathbf{1}}(\tau)$ as follows. Let $\boldsymbol{\beta}_{\mathbf{1}}(\tau)_0$ be an initial value of $\boldsymbol{\beta}_{\mathbf{1}}(\tau)$. The iterative scheme includes updating $\boldsymbol{\beta}_{\mathbf{1}}(\tau)_k$, for $k = 0, 1, 2, ...$, using the equation

$$\boldsymbol{\beta}_{\mathbf{1}}(\tau)_{k+1} = \boldsymbol{\beta}_{\mathbf{1}}(\tau)_k - [\mathbf{J}(\boldsymbol{\beta}_{\mathbf{1}}(\tau)_k)]^{-1}\mathbf{S_1}(\tau, \boldsymbol{\beta}_{\mathbf{1}}(\tau)_k), \tag{3.58}$$

where $\boldsymbol{\beta}_{\mathbf{1}}(\tau)_k$ and $\boldsymbol{\beta}_{\mathbf{1}}(\tau)_{k+1}$ are the parameter estimates in the kth and $(k+1)$st iterations, respectively. The $p \times p$ matrix $\mathbf{J}(\boldsymbol{\beta}_{\mathbf{1}}(\tau))$ is the derivative of the vector $\mathbf{S_1}(\tau, \boldsymbol{\beta}_{\mathbf{1}})$, which is given by

$$\mathbf{J}(\boldsymbol{\beta}_{\mathbf{1}}(\tau)) = \frac{\partial \mathbf{S_1}(\tau, \boldsymbol{\beta}_{\mathbf{1}})}{\partial {\boldsymbol{\beta}_{\mathbf{1}}}^T(\tau)}, \tag{3.59}$$

where

$$\frac{\partial \mathbf{S_1}(\tau, \boldsymbol{\beta}_{\mathbf{1}})}{\partial {\boldsymbol{\beta}_{\mathbf{1}}}^T(\tau)} = -n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \Big(\frac{f_1\big(e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{1}}(\tau)}\big|\mathbf{x}_i\big)\mathbf{x}_i^T e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{1}}(\tau)}}{1 - F_1\big(e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{1}}(\tau)}\big|\mathbf{x}_i\big)} \Big) I\{e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{1}}(\tau)} < z_{1i}\}. \tag{3.60}$$

This iterative technique is based on the differentiation of the estimating functions and requires the derivative of the vector of estimating functions given in (3.56) with respect to the $p \times 1$ vector $\boldsymbol{\beta}_{\mathbf{1}}(\tau)$. The stopping criterion for the Newton-Raphson algorithm in the proposed method employed in this thesis is based on the convergence criterion $|\boldsymbol{\beta}_{\mathbf{1}}(\tau)_{k+1} - \boldsymbol{\beta}_{\mathbf{1}}(\tau)_k| < \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} > \mathbf{0}$ is a prespecified vector of error terms used to stop the algorithm. If the convergence criterion is satisfied in the $(k+1)$st iteration, the values in $\boldsymbol{\beta}_{\mathbf{1}}(\tau)_{k+1}$ are used for the corresponding values of the estimators in $\hat{\boldsymbol{\beta}}_1(\tau)$.

To calculate the model-based variance estimates of the estimated parameters in the vector $\hat{\boldsymbol{\beta}}_1$, we employ the sandwich estimator (Lawless, 2003, p. 553). Let $Cov(\hat{\boldsymbol{\beta}}_1)$

be the $p \times p$ variance-covariance matrix of $\hat{\boldsymbol{\beta}}_1$ with the elements $Cov(\hat{\boldsymbol{\beta}}_{1i}, \hat{\boldsymbol{\beta}}_{1j})$, $i, j = 0, 1, 2, ..., p-1$. The estimator of the $p \times p$ covariance matrix of $\hat{\boldsymbol{\beta}}_1$ is given by

$$\hat{Cov}(\hat{\boldsymbol{\beta}}_1) = \frac{1}{n}[\boldsymbol{A_n}(\hat{\boldsymbol{\beta}}_1)]^{-1}[\boldsymbol{B_n}(\hat{\boldsymbol{\beta}}_1)]\Big[[\boldsymbol{A_n}(\hat{\boldsymbol{\beta}}_1)]^{-1}\Big]^T, \tag{3.61}$$

where the $p \times p$ matrix $\boldsymbol{A_n}$ is

$$\boldsymbol{A_n}(\boldsymbol{\beta_1}) = -\frac{1}{n}\frac{\partial \mathbf{S_1}(\tau, \boldsymbol{\beta_1})}{\partial \boldsymbol{\beta_1}^T(\tau)}, \tag{3.62}$$

and the $p \times p$ matrix $\boldsymbol{B_n}$ is

$$\begin{aligned}\boldsymbol{B_n}(\boldsymbol{\beta_1}) = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\mathbf{x}_i\Big[N_{1i}\big(e^{\mathbf{x}_i^T\boldsymbol{\beta_1}(\tau)}\big) + \log\Big(1 - F_1(\min\{e^{\mathbf{x}_i^T\boldsymbol{\beta_1}(\tau)}, z_{1i}\}|\mathbf{x}_i)\Big)\Big]\Bigg] \\ \times \Bigg[\mathbf{x}_i\Big[N_{1i}\big(e^{\mathbf{x}_i^T\boldsymbol{\beta_1}(\tau)}\big) + \log\Big(1 - F_1(\min\{e^{\mathbf{x}_i^T\boldsymbol{\beta_1}(\tau)}, z_{1i}\}|\mathbf{x}_i)\Big)\Big]\Bigg]^T.\end{aligned} \tag{3.63}$$

In Section 3.5, we used this method to estimate $\boldsymbol{\beta_1}(\tau)$ in a simulation study and compare our results to those obtained with the Peng-Huang method.

We next discuss the estimation of the parameters in the conditional QR model, given a vector of covariates, for the second gap time in sequentially observed bivariate gap times setting. The marginal distributions of the first and second gap times can be of different types. In this case, our main goal is to estimate the parameters $\boldsymbol{\beta_2}(\tau)$ in the conditional QR model for the marginal distribution of the second gap time; that is, to estimate $\boldsymbol{\beta_2}(\tau)$ in the model

$$Q_{T_2}(\tau|\mathbf{x}) = \exp\big(\mathbf{x}^T\boldsymbol{\beta_2}(\tau)\big), \qquad \tau \in (0, 1), \tag{3.64}$$

where $\mathbf{x} = (1, \tilde{\mathbf{x}}^T)^T$ is a $p \times 1$ vector including the $(p-1) \times 1$ vector of covariates $\tilde{\mathbf{x}}$ and $\boldsymbol{\beta_2}(\tau) = (\beta_{02}(\tau), \beta_{12}(\tau), ..., \beta_{p-1,2}(\tau))^T$ is a $p \times 1$ vector of regression parameters representing the effects of covariates on the τth quantile of the log of the second gap time T_2. An important aspect of our estimation method is that we do not assume the independence of the first and second gap times, T_1 and T_2, respectively. We model the dependency between the first and second gap times using a Clayton copula, which defines the joint distribution of the two gap times, accounting for their interdependence.

For n independent individuals, suppose that the counting process $\{N_{2i}(t),\ t \geq 0\}$, $i = 1, 2, ..., n$, has the associated cumulative intensity function $\Lambda_{2i}(t|H_i(t))$, where $H_i(t) = \{N_{1i}(s),\ N_{2i}(s),\ \mathbf{x}_i;\ 0 \leq s < t\}$. The counting random variable $N_{2i}(t)$ takes the value of 1 if the ith individual experiences the second type of event in $(0, t]$. Otherwise, its value is zero. Note that, since the first and second types of events are sequentially observed, it is necessary that $N_{1i}(t) = 1$ if $N_{2i}(t) = 1$. As discussed by Yip and Lam (1997), the martingale structure is preserved for the quantities $M_{2i}(t) = N_{2i}(t) - \Lambda_{2|1}(\min(t, z_{2i})|T_{1i} = t_{1i}, \mathbf{x}_i)$, for $i = 1, 2, ..., n$, where $z_{2i} = \min(T_{2i}, C_i - T_{1i})$. We investigate this structure in an empirical study in Section 3.4 through a Monte Carlo simulation study.

Similar to the previous discussion, since the $M_{2i}(t)$ are martingales, we have

$$E\left\{ n^{-\frac{1}{2}} \sum_{i=1}^{n} \Big[\mathbf{x}_i M_{2i}\big(e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{2}}^{*}(\tau)}\big) \Big| \mathbf{x}_i \Big] \right\} = \mathbf{0}, \tag{3.65}$$

where $\boldsymbol{\beta}_{\mathbf{2}}^{*}(\tau)$ denotes the true value of $\boldsymbol{\beta}_{\mathbf{2}}(\tau)$ in Model (3.64), $\tau \in (0, 1)$ and $\mathbf{0}$ is a $p \times 1$ vector of zeros. As a result, for a given τ, we obtain $p \times 1$ vector of unbiased estimation functions $n^{\frac{1}{2}} \mathbf{S}_{\mathbf{2}}(\tau, \boldsymbol{\beta}_{\mathbf{2}}^{*})$, where

$$\mathbf{S}_{\mathbf{2}}(\tau, \boldsymbol{\beta}_{\mathbf{2}}^{*}) = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \Big[N_{2i}\big(e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{2}}^{*}(\tau)}\big) - \Lambda_{2|1}(\min\big\{e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{2}}^{*}(\tau)}, z_{2i}\big\} \Big| t_{1i}, \mathbf{x}_i) \Big]. \tag{3.66}$$

Therefore, $n^{\frac{1}{2}} \mathbf{S}_{\mathbf{2}}(\tau, \boldsymbol{\beta}_{\mathbf{2}}^{*}) = \mathbf{0}$ provides a $p \times 1$ system of unbiased estimating equations with the components

$$\begin{aligned}
&n^{-\frac{1}{2}} \sum_{i=1}^{n} \Big[N_{2i}\big(e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{2}}^{*}(\tau)}\big) - \Lambda_{2|1}(\min\big\{e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{2}}^{*}(\tau)}, z_{2i}\big\} \Big| t_{1i}, \mathbf{x}_i) \Big] = 0, \\
&n^{-\frac{1}{2}} \sum_{i=1}^{n} x_{i1} \Big[N_{2i}\big(e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{2}}^{*}(\tau)}\big) - \Lambda_{2|1}(\min\big\{e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{2}}^{*}(\tau)}, z_{2i}\big\} \Big| t_{1i}, \mathbf{x}_i) \Big] = 0, \\
&n^{-\frac{1}{2}} \sum_{i=1}^{n} x_{i2} \Big[N_{2i}\big(e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{2}}^{*}(\tau)}\big) - \Lambda_{2|1}(\min\big\{e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{2}}^{*}(\tau)}, z_{2i}\big\} \Big| t_{1i}, \mathbf{x}_i) \Big] = 0, \\
&\qquad\qquad\qquad\qquad\vdots \\
&n^{-\frac{1}{2}} \sum_{i=1}^{n} x_{i,p-1} \Big[N_{2i}\big(e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{2}}^{*}(\tau)}\big) - \Lambda_{2|1}(\min\big\{e^{\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{2}}^{*}(\tau)}, z_{2i}\big\} \Big| t_{1i}, \mathbf{x}_i) \Big] = 0.
\end{aligned} \tag{3.67}$$

We use these unbiased estimating equations to estimate the parameters in the vector $\boldsymbol{\beta}_{\mathbf{2}}^{\mathbf{*}}(\tau)$. Note that from the definition of the cumulative hazard function, we have

$$\Lambda_{2|1}(t|t_1, \mathbf{x})) = \int_0^t h_{2|1}(u|t_1, \mathbf{x}))du = -\log\big(1 - F_{2|1}(t|t_1, \mathbf{x})\big). \tag{3.68}$$

Also, as discussed in Section 2.2, the Clayton copula is given by

$$C_\phi(u, v) = \big(u^{-\phi} + v^{-\phi} - 1\big)^{-\frac{1}{\phi}}, \tag{3.69}$$

where $\phi > 0$ is the Clayton copula parameter, and u and v are standard uniformly distributed random variables. From (3.68) and (3.69), we can show that

$$\Lambda_{2|1}(\min\left\{e^{\mathbf{x}_i^T\boldsymbol{\beta}_2^*(\tau)}, z_{2i}\right\}\Big|t_{1i}, \mathbf{x}_i) =$$

$$\begin{cases} 0, & \text{if } \delta_1 = 0, \\ -\log\Big(1 - F_1(t_1)^{-\phi-1}[F_1(t_1)^{-\phi} + F_2(e^{\mathbf{x}_i^T\boldsymbol{\beta}_2^*(\tau)})^{-\phi} - 1]^{-\frac{1}{\phi}-1}\Big), & \text{if } t = e^{\mathbf{x}_i^T\boldsymbol{\beta}_2^*(\tau)} < z_{2i}, \\ -\log\Big(1 - F_1(t_1)^{-\phi-1}[F_1(t_1)^{-\phi} + F_2(z_{2i})^{-\phi} - 1]^{-\frac{1}{\phi}-1}\Big), & \text{if } z_{2i} < t = e^{\mathbf{x}_i^T\boldsymbol{\beta}_2^*(\tau)}, \end{cases} \tag{3.70}$$

where F_1 and F_2 are the c.d.f. of the first and the c.d.f. of the second gap times, respectively, and $\delta_1 = I(T_1 < C)$, which is the event indicator for the first event. For brevity, we do not explicitly show the dependency of the functions F_1 and F_2 to the parameters in (3.70). However, note that $F_1(t) = F_1(t; \boldsymbol{\beta}_1)$, where $\boldsymbol{\beta}_1$ are the $p \times 1$ vector of parameters in the model $\log T_1 = \tilde{\mathbf{x}}^T\boldsymbol{\beta}_1 + \epsilon_1$, and ϵ_1 is the random error term. We explain the computation of $\Lambda_{2|1}(\min\left\{e^{\mathbf{x}_i^T\boldsymbol{\beta}_2^*(\tau)}, z_{2i}\right\}\Big|t_{1i}, \mathbf{x}_i)$ for a given data set in more detail in Section 3.4.3.

It should be noted that the parameters $\boldsymbol{\beta}_1$ in the marginal distribution of the first gap time T_1 and the copula parameter ϕ appear in the conditional cumulative intensity function (3.70). We apply a two-stage estimation procedure for the estimation of them. In the first stage, the estimates of the parameters $\boldsymbol{\beta}_1$ are obtained by maximizing the log of likelihood function $\ell_1(\boldsymbol{\beta}_1) = \log L(\boldsymbol{\beta}_1)$, where $L(\boldsymbol{\beta}_1)$ is given in (3.7). Let $\hat{\boldsymbol{\beta}}_1$ denote the vector of these estimates, which are the maximum likelihood estimates of the parameters in $\boldsymbol{\beta}_1$. In the second stage, the estimates in $\hat{\boldsymbol{\beta}}_1$ are plugged in the functions F_1 and f_1 in the log likelihood function $L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \phi)$ given in (3.10); that is, $\ell_2(\hat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}_2, \phi) = \log L(\hat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}_2, \phi)$, which is a profile log likelihood function for $\boldsymbol{\beta}_2$ and ϕ. Then, the function $\ell_2(\hat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}_2, \phi)$ is maximized to obtain the estimate of the Clayton copula parameter ϕ given in (3.70). Let $\hat{\phi}$ denotes the value of this estimate. We then replace $\boldsymbol{\beta}_1$ and ϕ with $\hat{\boldsymbol{\beta}}_1$ and $\hat{\phi}$ in (3.70). It should be noted that both estimators $\hat{\boldsymbol{\beta}}_1$ and $\hat{\phi}$ obtained in this approach are consistent under some regularity conditions. The regularity conditions and asymptotic properties of the maximum likelihood estimates $\hat{\boldsymbol{\beta}}_1$ are well-established. The regularity conditions for $\hat{\phi}$ can be found in Shih and Louis (1995) and Andersen (2005). The maximum likelihood estimates of $\boldsymbol{\beta}_1$ and ϕ can also be obtained by directly maximizing $\ell_2(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \phi)$. However, the two-stage estimation method is computationally more efficient than the direct maximization method. Our simulation results show that the estimates are very close in both methods for sufficiently large sample sizes.

Another point worth noting is that the estimating equations given in (3.67) are not *bona fide* unbiased estimating equations after replacing $\boldsymbol{\beta}_1$ and ϕ with $\hat{\boldsymbol{\beta}}_1$ and $\hat{\phi}$, respectively. However, as the sample size n approaches infinity, the expectation given in (3.65) converges to $\mathbf{0}$. Therefore, the estimating functions given in (3.67) are unbiased in the limit as $n \to \infty$, after plugging-in $\hat{\boldsymbol{\beta}}_1$ and $\hat{\phi}$ in place of $\boldsymbol{\beta}_1$ and ϕ, respectively. We investigated this result in empirical settings with a Monte Carlo simulation study in Section 3.4.3.

We first replace $\boldsymbol{\beta}_1$ and ϕ in (3.70) with $\hat{\boldsymbol{\beta}}_1$ and $\hat{\phi}$, respectively, and then incorporating (3.70) into (3.66). We then employ the Newton-Raphson method to the p estimating equations given in (3.67), to estimate $\boldsymbol{\beta_2}(\tau)$ in Model (3.64). As explained before, the proposed method using the Newton-Raphson algorithm takes the form

$$\boldsymbol{\beta}_2(\tau)_{k+1} = \boldsymbol{\beta}_2(\tau)_k - \Big(\mathbf{J}(\boldsymbol{\beta_2}(\tau))\Big)^{-1} \mathbf{S_2}(\tau, \boldsymbol{\beta}_2(\tau)_k), \tag{3.71}$$

where $\boldsymbol{\beta_2}(\tau)_k$ and $\boldsymbol{\beta_2}(\tau)_{k+1}$ are the parameter estimates in the kth and $(k+1)$st iterations, respectively. With the Clayton copula (3.69), the conditional p.d.f. and c.d.f. of T_{2i}, given $T_{1i} = t_1$ and $\mathbf{x}_i = \mathbf{x}_i$, are respectively given by

$$f_{2|1}\big(t\big|t_1, \mathbf{x}_i\big) = \frac{\frac{\partial^2}{\partial t_1 \partial t} C_\phi(F_1(t_1), F_2(t))}{f_1(t_1)}$$

$$= \frac{\frac{\partial^2}{\partial t_1 \partial t}(F_1(t_1)^{-\phi} + F_2(t)^{-\phi} - 1)^{-\frac{1}{\phi}}}{f_1(t_1)}$$

$$= \frac{\frac{\partial}{\partial t}(-\frac{1}{\phi})(F_1(t_1)^{-\phi} + F_2(t)^{-\phi} - 1)^{-\frac{1}{\phi}-1}(-\phi) f_1(t_1) F_1(t_1)^{-\phi-1}}{f_1(t_1)}$$

$$= (\phi + 1) F_1(t_1)^{-\phi-1}\big[F_1(t_1)^{-\phi} + F_2(t)^{-\phi} - 1\big]^{-\frac{1}{\phi}-2} F_2(t)^{-\phi-1} f_2(t), \quad (3.72)$$

and

$$F_{2|1}\big(t \big| t_1, \mathbf{x}_i\big) = F_1(t_1)^{-\phi-1}\big[F_1(t_1)^{-\phi} + F_2(t)^{-\phi} - 1\big]^{-\frac{1}{\phi}-1}. \quad (3.73)$$

For $z_{2i} = \min(T_{2i}, C_i - T_{1i})$ and with replacing $t = \exp\big(\mathbf{x}_i^T \boldsymbol{\beta_2}(\tau)\big)$ in (3.72) and (3.73), the $p \times p$ matrix $\mathbf{J}(\boldsymbol{\beta_2}(\tau))$ is defined by

$$\mathbf{J}(\boldsymbol{\beta_2}(\tau)) = \frac{\partial \mathbf{S_2}(\tau, \boldsymbol{\beta_2})}{\partial \boldsymbol{\beta_2}^T(\tau)}, \qquad \tau \in (0, 1), \quad (3.74)$$

where

$$\frac{\partial \mathbf{S_2}(\tau, \boldsymbol{\beta}_2)}{\partial \boldsymbol{\beta_2}^T(\tau)} = -n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{x}_i \Big(\frac{f_{2|1}\big(e^{\mathbf{x}_i^T \boldsymbol{\beta_2}(\tau)} \big| t_1, \mathbf{x}_i\big) \mathbf{x}_i^T e^{\mathbf{x}_i^T \boldsymbol{\beta_2}(\tau)}}{1 - F_{2|1}\big(e^{\mathbf{x}_i^T \boldsymbol{\beta_2}(\tau)} \big| t_1, \mathbf{x}_i\big)} \Big) I\{e^{\mathbf{x}_i^T \boldsymbol{\beta_2}(\tau)} < z_{2i}\}. \quad (3.75)$$

Convergence in the algorithm with (3.71) is assessed based on the absolute difference between the calculated value of $\boldsymbol{\beta}_2(\tau)$ parameters at the kth and $(k+1)$st steps. Specifically, the algorithm is stopped when $|\boldsymbol{\beta}_2(\tau)_{k+1} - \boldsymbol{\beta}_2(\tau)_k| < \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} > \mathbf{0}$ is a prespecified number. If the convergence criterion is met, the estimated parameters in the $(k+1)$st iteration are returned as the estimated values of the parameters in $\boldsymbol{\beta}_2(\tau)$. This iterative process effectively estimates the parameters in the vector $\boldsymbol{\beta_2}(\tau)$.

Once again, we employ the sandwich estimator to calculate the model-based variance estimates of the estimated parameter $\hat{\boldsymbol{\beta}}_2$. In this case,

$$C\hat{o}v(\hat{\boldsymbol{\beta}}_2) = \frac{1}{n}[\boldsymbol{A_n}(\hat{\boldsymbol{\beta}}_2)]^{-1}[\boldsymbol{B_n}(\hat{\boldsymbol{\beta}}_2)]\Big[[\boldsymbol{A_n}(\hat{\boldsymbol{\beta}}_2)]^{-1}\Big]^T, \quad (3.76)$$

where the $p \times p$ matrix $\boldsymbol{A_n}$ is given by

$$\boldsymbol{A_n}(\boldsymbol{\beta_2}) = -\frac{1}{n}\frac{\partial \mathbf{S_2}(\tau, \boldsymbol{\beta_2})}{\partial \boldsymbol{\beta_2}^T(\tau)}, \quad (3.77)$$

and the $p \times p$ matrix $\boldsymbol{B_n}$ is

$$\begin{aligned}\boldsymbol{B_n}(\boldsymbol{\beta_2}) =& \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{x}_i\Big[N_{2i}\big(e^{\mathbf{x}_i^T\boldsymbol{\beta_2}(\tau)}\big) + \log\Big(1 - F_{2|1}(\min\{e^{\mathbf{x}_i^T\boldsymbol{\beta_2}(\tau)}, z_{1i}\}|\mathbf{x}_i, t_{1i})\Big)\Big]\right] \\ & \times \left[\mathbf{x}_i\Big[N_{2i}\big(e^{\mathbf{x}_i^T\boldsymbol{\beta_2}(\tau)}\big) + \log\Big(1 - F_{2|1}(\min\{e^{\mathbf{x}_i^T\boldsymbol{\beta_2}(\tau)}, z_{1i}\}|\mathbf{x}_i, t_{1i})\Big)\Big]\right]^T.\end{aligned} \tag{3.78}$$

It should be noted that the estimate of $Cov(\hat{\boldsymbol{\beta}}_2)$ given in (3.76) does not take into account the estimation of $\boldsymbol{\beta}_1$ and ϕ in the previous stages. This approach works fine for large sample sizes like the ones considered in our simulation studies presented in Section 3.5 and the data analysis given in Section 4.2. However, it may provide conservative values in some cases with small sample sizes. In such cases, a bootstrap procedure can be used to calculate the standard errors of the estimates of parameters in $\boldsymbol{\beta}_2$ (Casella and Berger, 2002, Section 10.1.4). We further discuss this issue in Section 5.2.

The asymptotic properties of the estimators obtained with this method are not analytically investigated in this thesis. However, we conducted a Monte Carlo simulation study to discuss the accuracy of the standard normal approximations for the quantity $\sqrt{n}(\hat{\beta}_{2k}(\tau)-\beta^*_{2k}(\tau))/\sqrt{V\hat{a}r(\hat{\beta}_{2k}(\tau))}$, $k = 0, 1, 2, ..., p-1$, with normal quantile-quantile (Q-Q) plots. The results of this study are presented in Section 3.5 and suggest that the standard normal approximations for $\sqrt{n}(\hat{\beta}_{2k}(\tau) - \beta^*_{2k}(\tau))/\sqrt{V\hat{a}r(\hat{\beta}_{2k}(\tau))}$ are adequate for sufficiently large sample sizes. Furthermore, the results of the simulation studies given in Section 3.5 show that the estimators $\hat{\beta}_{2k}(\tau)$ are consistent. Details of this simulation study can be found in Section 3.5.

3.4 Martingale Structure for Sequentially Observed Two Events: A Simulation Study

In this section, we first introduce how to simulate sequentially observed bivariate gap times with copulas. Then, we presented the results of two Monte Carlo simulation studies. The primary objective of the simulation studies is to investigate whether the counting processes associated with events constitute a zero-mean martingale structure in settings with sequentially observed gap times. This result is important for the development of the estimation methods discussed in Sections 3.2 and 3.3. Specifically, we aimed to investigate the expectation of the martingales associated with the counting processes of the first and the second event times in some empirical settings. We first explain how to generate sequentially observed bivariate gap times in the next section. Then, we present our simulation study conducted to investigate this for the first event in Section 3.4.2. Then, in Section 3.4.3, we discuss the simulation study conducted to incorporate the second event while considering the dependency between the two gap times using the Clayton copula structure.

3.4.1 Simulation of Sequentially Observed Bivariate Gap Times

In this section, we explain how to generate sequentially observed bivariate survival (gap) times from n independent subjects. Let T_{1i} and T_{2i} be survival times of two sequentially observed events, where T_{1i} is the time to the first event and T_{2i} is the time between the first and second events for the ith subject, $i = 1, 2, ..., n$. In the illness-death model framework, T_{1i} and T_{2i} represent the elapsed times for the ith individual staying in the *healthy* state before moving to the *ill* state and staying in the *ill* state before moving to the *death* state, respectively.

The algorithm to generate sequentially observed bivariate survival data in this thesis is given as follows:

1. Generate $U_{1i} = u_{1i}$ from a standard uniform distribution for $i = 1, 2, ..., n$.
2. Set $T_{1i} = t_{1i}$, where $t_{1i} = F_1^{-1}(u_{1i})$ and F_1 is the marginal c.d.f. of the first gap time, T_{1i}, for $i = 1, 2, ..., n$.
3. For $i = 1, 2, ..., n$, generate censoring time $C_i = c_i$ for the ith individual from the Uniform(ψ_1, ψ_2) distribution, where the values of ψ_1 and ψ_2 are selected so that the desired censoring proportion in the data is achieved.
4. If $t_{1i} \leq c_i$, let $t^*_{1i} = t_{1i}$ and $\delta_{1i} = 1$, for $i = 1, 2, ..., n$. Otherwise, set $t^*_{1i} = c_i$, $\delta_{1i} = 0$, $t_{2i} = 0$, and $\delta_{2i} = 0$, for $i = 1, 2, ..., n$.
5. Generate $U_{2i} = u_{2i}$ from a standard uniform distribution for $i = 1, 2, ..., n$.

6. For $i = 1, 2, ..., n$, if $\delta_{1i} = 1$, calculate $T_{2i} = t_{2i}$ as the solution of

$$u_{2i} = F_{2|1}(t_{2i}|t_{1i}), \tag{3.79}$$

where $F_{2|1}(t_{2i}|t_{1i})$ is the conditional c.d.f. of T_{2i} given $T_{1i} = t_{1i}$.

7. For $i = 1, 2, ..., n$, if $\delta_{1i} = 1$ and $t_{2i} \leq c_i - t_{1i}$, let $t^*_{2i} = t_{2i}$ and $\delta_{2i} = 1$. If $\delta_{1i} = 1$ and $t_{2i} > c_i - t_{1i}$, set $t^*_{2i} = c_i - t_{1i}$ and $\delta_{2i} = 0$.

The above algorithm generates the data $\{(t^*_{1i}, t^*_{2i}, \delta_{1i}, \delta_{2i});\ i = 1, 2, ..., n\}$ for n independent individuals. It should be noted that if the first gap time T_{1i} is censored then the second gap time T_{2i} is unobservable. In this generated data set such a situation corresponds to the cases in which $\delta_{1i} = 0$, which gives $(t^*_{1i}, 0, 0, 0)$ as the generated datum. In the remaining sections of this chapter, we applied similar steps given above to generate data in Monte Carlo simulations with R software. It should also be noted that the conditional c.d.f. of $F_{2|1}$ given in the right-hand side of (3.79) can be written as

$$F_{2|1}(t|t_{1i}) = \frac{\frac{\partial}{\partial t_{1i}} F(t_{1i}, t)}{\frac{\partial}{\partial t_{1i}} F_1(t_{1i})}, \qquad t \geq 0, \tag{3.80}$$

where F is the joint c.d.f. of T_{1i} and T_{2i} and F_1 is the marginal c.d.f. of T_{1i}, $i = 1, 2, ..., n$. As discussed in Section 2.2, the joint c.d.f. F can be represented with the copula $C(F_1(t_{1i}), F_2(t_{2i}))$, where F_2 is the marginal c.d.f. of T_{2i}. With the bivariate Archimedean copulas (Genest and Rivest, 1993), the conditional c.d.f. given in (3.80) therefore takes the form of

$$\begin{aligned} F_{2|1}(t|T_{1i} = t_{1i}) &= \frac{\frac{\partial}{\partial t_{1i}} C(F_1(t_{1i}), F_2(t))}{\frac{\partial}{\partial t_{1i}} F_1(t_{1i})}, \\ &= \frac{\frac{\partial}{\partial t_{1i}} \varphi^{-1}(\varphi(F_1(t_{1i})) + \varphi(F_2(t)))}{\frac{\partial}{\partial t_{1i}} \varphi^{-1}(\varphi(F_1(t_{1i})))}, \\ &= \frac{\varphi^{-1(1)}(\varphi(F_1(t_{1i})) + \varphi(F_2(t)))}{\varphi^{-1(1)}(\varphi(F_1(t_{1i})))}, \end{aligned} \tag{3.81}$$

where φ is the generator of the Archimedean copula and $\varphi^{-1(1)} = (\partial/\partial t_{1i})\varphi^{-1}$. In this thesis, we use the Clayton copula to model the dependency between two gap times.

As mentioned in Section 2.2, the generator function of the Clayton copula is

$$\varphi(t) = t^{-\phi} - 1, \ \ \phi > 0, \tag{3.82}$$

which generates

$$C_\phi(u,v) = \left(u^{-\phi} + v^{-\phi} - 1\right)^{-\frac{1}{\phi}}, \ \ 0 \le u, v \le 1, \tag{3.83}$$

where ϕ is the copula (dependence) parameter. The inverse of the generator function of the Clayton copula is

$$\varphi^{-1}(z) = (z+1)^{-\frac{1}{\phi}}, \ \ \phi > 0, \tag{3.84}$$

and

$$\varphi^{-1(1)}(z) = \frac{\partial \varphi^{-1}(z)}{\partial z} = -\frac{1}{\phi}(z+1)^{-\frac{1}{\phi}-1}, \ \ \phi > 0. \tag{3.85}$$

In (3.79), we have $u_{2i} = F_{2|1}(t_{2i}|t_{1i})$ and from (3.81), we can write

$$u_{2i} = \frac{\varphi^{-1(1)}\{\varphi[F_1(t_{1i})] + \varphi[F_2(t_{2i})]\}}{\varphi^{-1(1)}\{\varphi[F_1(t_{1i})]\}}. \tag{3.86}$$

From the results given in (3.82), (3.84) and (3.85), we can show that

$$F_2(t_{2i}) = \Big[{u_{2i}}^{-\frac{\phi}{\phi+1}} {u_{1i}}^{-\phi} - {u_{1i}}^{-\phi} + 1\Big]^{-\frac{1}{\phi}}, \ \ \phi > 0, \tag{3.87}$$

which gives

$$t_{2i} = {F_2}^{-1}\Big(\big[1 - {u_{1i}}^{-\phi}(1 - {u_{2i}}^{-\frac{\phi}{\phi+1}})\big]^{-\frac{1}{\phi}}\Big), \tag{3.88}$$

where ${F_2}^{-1}$ is the inverse of marginal c.d.f. of T_{2i}, for $i = 1, 2, ..., n$. Note that the value of t_{2i} in (3.88) depends on the value of t_{1i} trough the Clayton copula structure. We use the result (3.88) to generate the dependent gap times t_{1i} and t_{2i} for $i = 1, 2, ..., n$ in the simulations.

3.4.2 Martingale Associated with the Counting Process of the First Event

Formation of the zero-mean martingale associated with the counting process of the first event is well established (see, for example, Aalen et al., 2008, Andersen et al., 1993; Daley and Vere-Jones, 2003). We conducted a simulation study to investigate this property in an empirical setting. We present the results of this study in this section. Our goal is to get some insight about the amount of variability around the zero-mean martingale in an empirical setting. This investigation is also useful to understand the martingale structure associated with the second event, which is discussed in Section 3.4.3. In particular, we are interested in the investigation of whether or not $E\{M_1(t|\mathbf{x})\} = 0$, for $t \geq 0$, where $M_1(t) = N_1(t) - \Lambda_{T_1}(\min(t, T_1, C)|H(t))$, $\{N_1(t), t \geq 0\}$ is a counting process of the first event with the associated cumulative intensity function $\Lambda_{T_1}(t|H(t))$ defined in Equation (2.21) given in Section 2.1, $H(t) = \{N_1(s), \mathbf{x}; 0 \leq s < t\}$, T_1 is the event time of the first event and C is the censoring time.

In the simulation study, we considered a sample size of 200 individuals (i.e. $n = 200$) and approximately 20% right censoring. For each individual, we generated two covariates:

a. Covariate 1 (X_1): A continuous covariate drawn from a standard normal distribution.

b. Covariate 2 (X_2): A discrete covariate drawn from a Bernoulli distribution with the probability of success set at 0.5.

In each simulation run r, $r = 1, 2, ..., R$, we obtained the survival data $D_r = \{(t^*_{1i}, \delta_{1i});\ i = 1, 2, ..., n\}$ as follows:

a. We set the simulation index $r = 1$.

b. For $i = 1, 2, ..., n$, we generated the error term $\epsilon_{1i} = e_{1i}$ from a standard extreme value distribution, and $X_{1i} = x_{1i}$ and $X_{2i} = x_{2i}$ as explained above.

c. Then, we calculated

$$y_{1i} = \beta_{01} + \beta_{11}x_{1i} + \beta_{21}x_{2i} + e_{1i}\ , \quad \text{for} \ \ i = 1, 2, ..., n. \tag{3.89}$$

d. The observed values of the survival times of individuals were then $t_{1i} = \exp(y_{1i})$, $i = 1, 2, ..., n$.

e. For $i = 1, 2, ..., n$, we then generated the right-censoring time $C_i = c_i$ from Uniform$(0, \psi)$ distribution, where the parameter ψ was selected so that approximately 20% of the processes were right censored.

f. We then set $t^*_{1i} = \min(t_{1i}, c_i)$ and $\delta_{1i} = 1$ if $t_{1i} \leq c_i$; otherwise $\delta_{1i} = 0$, for $i = 1, 2, ..., n$. The data set in the rth simulation run was then given by $D_r = \{(t^*_{1i}, \delta_{1i}), x_{1i}, x_{2i};\ i = 1, 2, ..., n\}$.

g. For the process i, $i = 1, 2, ..., n$, we set $N_{1i}(t) = 1$ if $t^*_{1i} \leq t$ and $\delta_{1i} = 1$; otherwise $N_{1i}(t) = 0$ at various time points t.

h. Then, we calculated the cumulative hazard function of the ith process, $i = 1, 2, ..., n$, at each time point t as follows.

$$\Lambda_{T_{1i}}(t|x_{1i}, x_{2i}) = \int_0^t Y_{1i}(s)\ h_{T_{1i}}(s|x_{1i}, x_{2i})ds, \qquad (3.90)$$

where $h_{T_{1i}}(s)$ is the hazard function of the ith individual and $Y_{1i}(s)$ is the at-risk function of the ith individual as defined in Section 2.1.

i. For $i = 1, 2, ..., n$, we then calculated the martingale $M_{1i}(t) = N_{1i}(t)\ - \Lambda_{T_{1i}}(\min(t, t_{1i}, c_i)|H_i(t))$ at each time point t. Subsequently, we computed $\bar{M}_{1r}(t) = \frac{1}{n}\sum_{i=1}^{n} M_{1i}(t)$ at each time point t.

j. We then increased r by 1 and repeated the previous steps $R - 1$ times.

We conducted $R = 1000$ iterations of the simulation process to estimate the martingale expectation at various time points t. At each time point t, we computed the mean of the martingale $\bar{M}_{1r}(t)$ based on 1000 runs, and obtained the empirical means of martingales at each time point t. That is, we calculated $\bar{M}_1(t) = \frac{1}{R}\sum_{r=1}^{R} \bar{M}_{1r}(t)$ at various t points.

The results are presented in Figure 3.1 and Figure 3.2 as pointwise dot plots of $\bar{M}_1(t)$ at various time points t. The dots are connected with straight lines for an easy visual interpretation. The plot in Figure 3.1 shows the results of $\bar{M}_1(t)$ when the values of the parameters β_{01}, β_{11} and β_{21} in the model were all set to

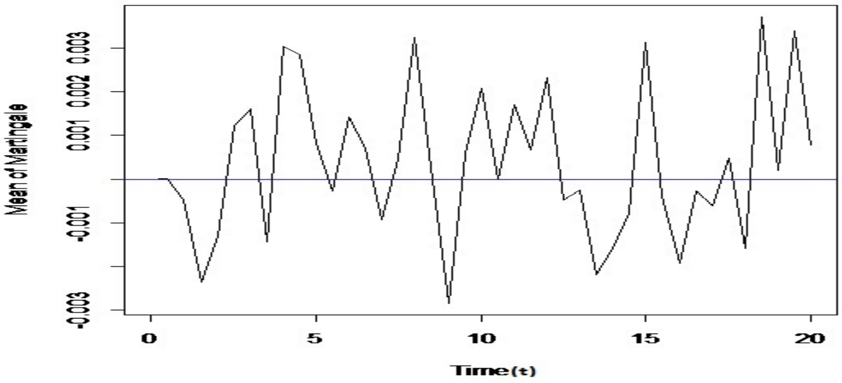
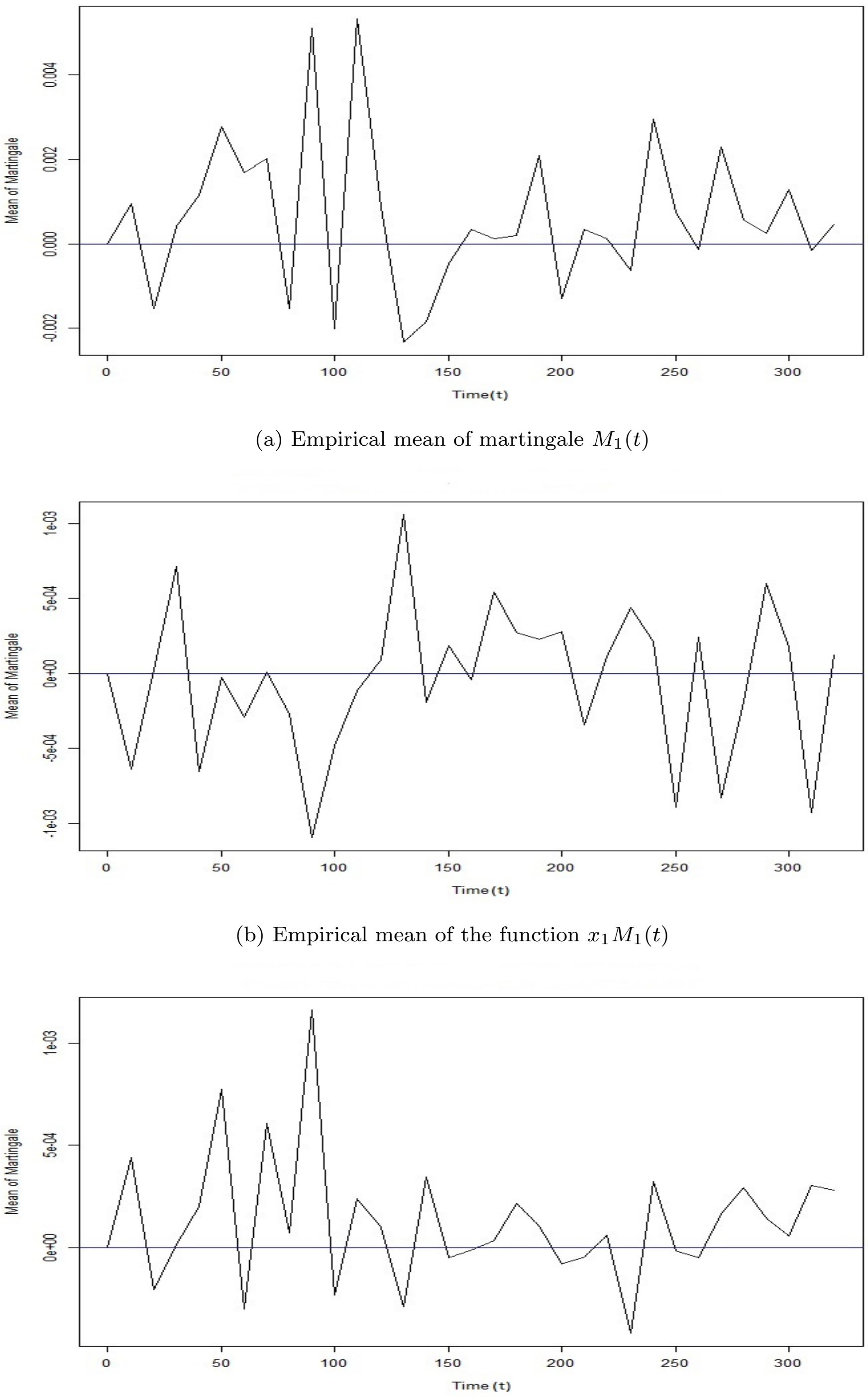


Figure 3.1: Empirical mean of the martingale $M_1(t)$ associated with the counting process of the first event without covariates (i.e. $\beta_{01} = \beta_{11} = \beta_{21} = 0$ in (3.89)) at various time points t.

zero (i.e. $\beta_{01} = \beta_{11} = \beta_{21} = 0$ in Model (3.89)). It is clear from the plot that the empirical setting supports that the mean-zero martingale structure is present for the case considered in our simulation setting. The results are close to the zero line with a random occurrence pattern above and below the zero line, which is, a typical white-noise pattern of a martingale difference sequence.

The plots given in Figure 3.2 include the values of $\bar{M}_1(t)$ when the data were generated with Model (3.89) where $\beta_{01} = 0$, $\beta_{11} = 3.5$ and $\beta_{21} = 2$. The pattern in plots (a), (b) and (c) suggest that the martingale structure is preserved for $M_{1i}(t)$, $x_{1i}M_{1i}(t)$ and $x_{2i}M_{1i}(t)$, respectively, for $i = 1, 2, ..., n$. This empirical evidence suggests that it is reasonable to assume that the estimating functions in $\mathbf{x}_i M_{1i}(t)$, where $\mathbf{x}_i = (1, x_{1i}, x_{2i})^T$, are unbiased; that is, $E(\mathbf{x}_i M_{1i}(t) \Big| \mathbf{x}_i) = \mathbf{0}$, where $\mathbf{0} = (0, 0, 0)^T$, for $i = 1, 2, ..., n$. This result suggests that the expectation of the martingales arising from the counting processes for the first event are mean-zero valued. As a result, the estimating equations based on them can be used for the estimation of the coefficients $\beta_{01}(\tau)$, $\beta_{11}(\tau)$, and $\beta_{21}(\tau)$ at different values of quantile τ in our survival model. To achieve this, as discussed in Section 3.3, we employ the Newton-Raphson algorithm in the proposed method, an efficient iterative numerical technique used to find the roots of the system of equations with respect to $\beta_{01}(\tau)$, $\beta_{11}(\tau)$, and $\beta_{21}(\tau)$ at different values of a given quantile τ, $\tau \in (0, 1)$.



(a) Empirical mean of martingale $M_1(t)$

(b) Empirical mean of the function $x_1 M_1(t)$

(c) Empirical mean of the function $x_2 M_2(t)$

Figure 3.2: Empirical martingale structures $\mathbf{x}M_1(t)$, where $\mathbf{x} = (1, x_1, x_2)^T$, associated with the counting process of the first event at various time points t. The data were generated where $\beta_{01} = 0$, $\beta_{11} = 3.5$ and $\beta_{21} = 2$ in Model (3.89).

3.4.3 Martingale Associated with the Counting Process of the Second Event

In this section, we consider the martingale structure related to the counting process of the second event observed after the occurrence of the first event. We would like to note that the first and second events can be of different types. Our aim is to investigate whether the expectation of the martingale associated with the counting process of the second event time is a mean-zero structure in an empirical setting. To do this, we modeled the dependency between the first and second gap times using a Clayton copula, which defines the joint distribution of the two gap times, accounting for their interdependence.

To investigate the martingale structure in an empirical setting, we conducted a simulation study. The generation of the gap times of the first event was similar to that of explained in the previous section. To do this, we generated n i.i.d. error terms ϵ_{1i} for the first gap times from the standard extreme value distribution. We then proceeded the algorithm explained in the previous section to obtain the data $D_{1r} = \{(t^*_{1i}, \delta_{1i}), x_{1i}, x_{2i}; i = 1, 2, ..., n\}$ in the rth simulation run, where $r = 1, 2, ..., R$. The regression model to generate this data was given in the step c of the previous data generation algorithm; that is, for $i = 1, 2, ..., n$, $y_{1i} = \beta_{01} + \beta_{11}x_{1i} + \beta_{21}x_{2i} + e_{1i}$, where $y_{1i} = \log(t_{1i})$, t_{1i} is the observed value of T_{1i} and e_{1i} is the generated value of ϵ_{1i}.

To generate the data for the second event, we first obtained the data D_{1r} in the rth simulation run. We then applied the algorithm given below. Note that, to establish the dependency between T_{1i} and T_{2i}, we generated error terms ϵ_{2i} using a Clayton copula, taking into account the relationship with the previously generated error terms ϵ_{1i}.

a. We set $r = 1$, and obtained e_{1i} and c_i, for $i = 1, 2, ..., n$, as well as $D_{1r} = \{(t^*_{1i}, \delta_{1i}), x_{1i}, x_{2i}; i = 1, 2, ..., n\}$.

b. We defined $u_{1i} = F_1(e_{1i})$, $i = 1, 2, ..., n$, where F_1 represents the c.d.f. of the distribution of ϵ_{1i}.

c. We generated $U_{2i} = u_{2i}$, $i = 1, 2, ..., n$, from a standard uniform distribution, and calculated $\epsilon_{2i} = e_{2i}$, where

$$e_{2i} = F_2^{-1}\left(\left[1 - u_{1i}^{-\phi}(1 - u_{2i}^{-\frac{\phi}{\phi+1}})\right]^{\frac{-1}{\phi}}\right), \tag{3.91}$$

where F_2^{-1} represents the inverse of the c.d.f. of ϵ_{2i}, for $i = 1, 2, ..., n$, which was the standard extreme value distribution; that is $F_2^{-1}(x) = -\log(-\log(x))$, where $0 < x < 1$.

d. We then calculated

$$y_{2i} = \beta_{02} + \beta_{12}x_{1i} + \beta_{22}x_{2i} + e_{2i}, \tag{3.92}$$

and $t_{2i} = \exp(y_{2i})$, $i = 1, 2, ..., n$.

e. For $i = 1, 2, ..., n$,

- if $\delta_{1i} = 0$, we let $t^*_{2i} = 0$ and $\delta_{2i} = 0$,
- if $\delta_{1i} = 1$ and $t_{1i} + t_{2i} \leq c_i$, we let $t^*_{2i} = t_{2i}$ and $\delta_{2i} = 1$, and
- if $\delta_{1i} = 1$ and $t_{1i} + t_{2i} > c_i$, we let $t^*_{2i} = c_i - t_{1i}$ and $\delta_{2i} = 0$.

Then, we obtained the data set $D_{2r} = \{(t^*_{1i}, t^*_{2i}, \delta_{1i}, \delta_{2i}), x_{1i}, x_{2i};\ i = 1, 2, ..., n\}$ in the rth simulation run.

The above steps of the algorithm generates the data $D_{2r} = \{(t^*_{1i}, t^*_{2i},\ \delta_{1i},\ \delta_{2i}), x_{1i}, x_{2i}; i = 1, 2, ..., n\}$ for n independent individuals. This procedure was also used to obtain the results of the simulation study presented in Section 3.5. To obtain the results of the simulation study of this section, we further applied the following steps from f to i. These steps were used to calculate the martingale values associated with the second event at various time points t.

f. For a given value of t, we set the value of $Y_{2i}(t)$ and $N_{2i}(t)$, $i = 1, 2, ..., n$, using the following conditions. For $i = 1, 2, ..., n$,

- if $\delta_{1i} = 0$ or $t < t^*_{1i}$, we set $Y_{2i}(t) = 0$ and $N_{2i}(t) = 0$,
- if $\delta_{1i} = 1$, $\delta_{2i} = 0$ and $t < t^*_{1i} + t^*_{2i}$, we set $Y_{2i}(t) = 1$ and $N_{2i}(t) = 0$,
- if $\delta_{1i} = 1$, $\delta_{2i} = 1$ and $t < t^*_{1i} + t^*_{2i}$, we set $Y_{2i}(t) = 1$ and $N_{2i}(t) = 0$,

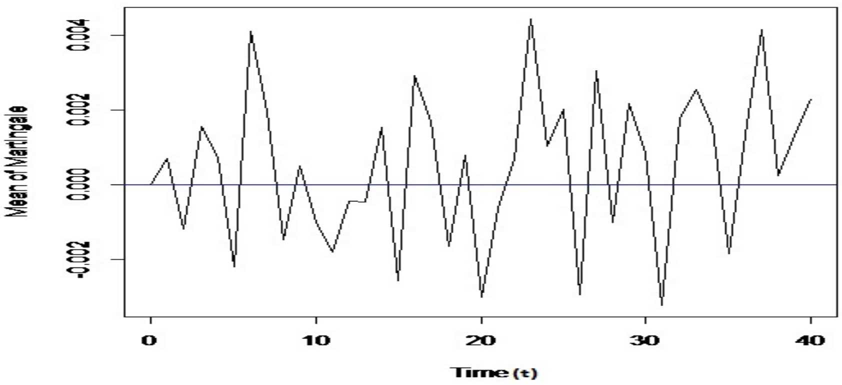


Figure 3.3: Mean of the martingale $M_2(t)$ associated with the counting process of the second event without covariates (i.e. $\beta_{01} = \beta_{11} = \beta_{21} = 0$ in Model (3.89) and $\beta_{02} = \beta_{12} = \beta_{22} = 0$ in Model (3.92)) at various time points t.

- if $\delta_{1i} = 1$, $\delta_{2i} = 0$ and $t \geq t_{1i}^* + t_{2i}^*$, we set $Y_{2i}(t) = 0$ and $N_{2i}(t) = 0$, and
- if $\delta_{1i} = 1$, $\delta_{2i} = 1$ and $t \geq t_{1i}^* + t_{2i}^*$, we set $Y_{2i}(t) = 0$ and $N_{2i}(t) = 1$.

g. Then, we calculated the cumulative hazard function of the ith process, $i = 1, 2, ..., n$, at each time point t as follows.

$$\Lambda_{2|1}(t|t_{1i}^*, x_{1i}, x_{2i}) = \int_0^t Y_{2i}(s)\ h_{2|1}(s|t_{1i}^*, x_{1i}, x_{2i})ds, \tag{3.93}$$

where $h_{2|1}$ is the conditional hazard function, given t_{1i}^*, x_{1i}, and, x_{2i}, and $Y_{2i}(s)$ is the at-risk function of the second event for the ith individual as defined in Step f.

h. We then calculated the martingale $M_{2i}(t)$ for $i = 1, 2, ..., n$ at each time point t, where $M_{2i}(t) = N_{2i}(t) - \Lambda_{2|1}(t|t_{1i}^*, x_{1i}, x_{2i})$. Subsequently, we computed $\bar{M}_{2r}(t) = \frac{1}{\delta_{1.}} \sum_{i=1}^{n} M_{2i}(t)$, where $\delta_{1.} = \sum_{i=1}^{n} \delta_{1i}$, for each time point t.

i. We then increased r by 1 and repeated the previous steps $R - 1$ times.

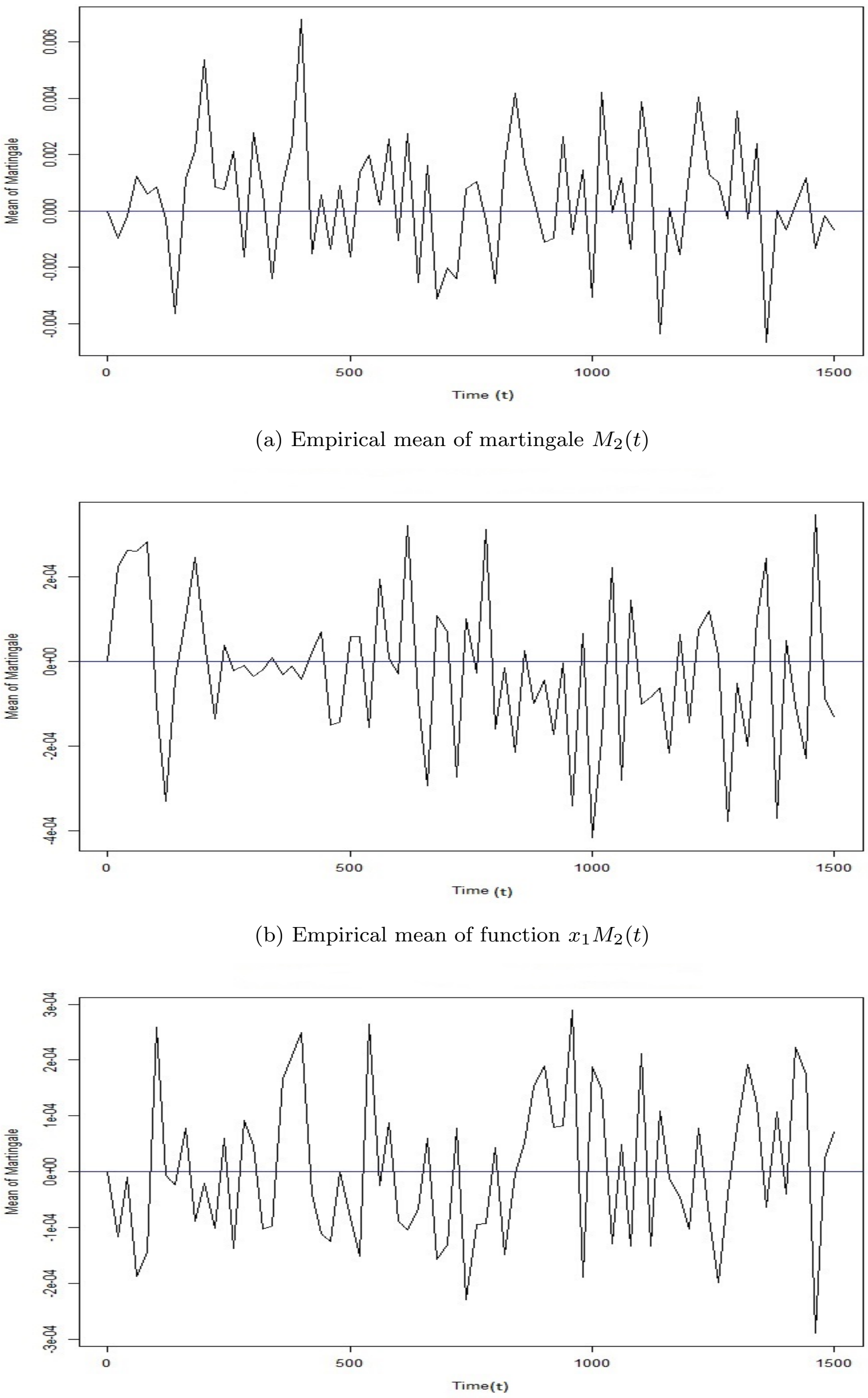
We fitted the simulation runs R at 1000, and estimated the martingale expectation at various time points t. At each time point t, we computed the mean of the martingale for 1000 iterations and obtained the empirical means of martingales at each time point t. That is, we calculated $\bar{M}_2(t) = \frac{1}{R} \sum_{r=1}^{R} \bar{M}_{2r}(t)$ for various t points. It should be noted that

$$h_{2|1}(t|t^*_{1i}, x_{1i}, x_{2i}) = -\log\Big(1 - F_1(t^*_{1i})^{-\phi-1}\big[F_1(t^*_{1i})^{-\phi} + F_2(t)^{-\phi} - 1\big]^{-\frac{1}{\phi}-1}\Big), \quad (3.94)$$

where $F_1(t^*_{1i}) = F_1(t^*_{1i}, \beta_{01}, \beta_{11}, \beta_{21})$ is the c.d.f. of T_{1i} evaluated at t^*_{1i} and $F_2(t) = F_2(t, \beta_{02}, \beta_{12}, \beta_{22})$ is the c.d.f. of T_{2i}

The results are presented in Figure 3.3 and Figure 3.4 as pointwise dot plots of $\bar{M}_2(t)$ for various values of t. Once again, the dots are connected with straight lines for an easy visual interpretation. The plot in Figure 3.3 shows the results of $\bar{M}_2(t)$ when the values of the parameters β_{01}, β_{11} and β_{21} in Model (3.89) and β_{02}, β_{12} and β_{22} in Model (3.92) were all set at zero (i.e. $\beta_{01} = \beta_{11} = \beta_{21} = \beta_{02} = \beta_{12} = \beta_{22} = 0$). Our conclusion is similar to that given in the previous section. It is clear from the plot that the empirical setting supports the mean-zero martingale structure for the case considered in our simulation study. The results are close to the zero line with a random occurrence pattern above and below the zero line, which is similar to a white-noise pattern.

The plots given in Figure 3.4 include the results of $\bar{M}_2(t)$ at various time points t when the data were generated with $\beta_{01} = 0$, $\beta_{11} = 2$, and $\beta_{21} = 1$ in Model (3.89) and with $\beta_{02} = 0$, $\beta_{12} = 3.5$ and $\beta_{22} = 2$ in Model (3.92). The patterns in these plots reveal that the martingale structure associated with the second event is preserved for $M_{2i}(t)$, $x_{1i}M_{2i}(t)$ and $x_{2i}M_{2i}(t)$ as well. This empirical evidence suggests that the estimating functions in $\mathbf{x}_i M_{2i}(t)$, where $\mathbf{x}_i = (1, x_{1i}, x_{2i})^T$, are reasonably unbiased; that is, for $i = 1, 2, ..., n$, $E(\mathbf{x}_i M_{2i}(t)) = \mathbf{0}$. In other words, the empirical results showed that it is reasonable to assume that the expectation of the martingales arising from the counting processes for the second event are mean-zero valued, and the estimating equations based on them can be used for the estimation of the coefficients $\beta_{02}(\tau)$, $\beta_{12}(\tau)$, and $\beta_{22}(\tau)$ at different values of quantile τ. Since our proposed estimation method of $\boldsymbol{\beta_2}(\tau)$ is based on the expectation given in (3.65), the result of this simulation study gives some insight about the validity of this martingale structure in an empirical setting.



(a) Empirical mean of martingale $M_2(t)$

(b) Empirical mean of function $x_1 M_2(t)$

(c) Empirical mean of function $x_2 M_2(t)$

Figure 3.4: Empirical martingale structures $\mathbf{x}M_2(t)$, where $\mathbf{x} = (1, x_1, x_2)^T$, associated with the counting process of the second event at various time points t. The data were generated where $\beta_{01} = 0$, $\beta_{11} = 2$ and $\beta_{21} = 1$ in Model (3.89) and $\beta_{02} = 0$, $\beta_{12} = 3.5$ and $\beta_{22} = 2$ in Model (3.92).

3.5 Monte Carlo Simulations

In this section, we present the results of Monte Carlo simulations conducted to assess and compare the performance of the Peng-Huang and proposed methods for the estimation of the model parameters in the marginal distributions of the first and second gap times. We also show the result of a simulation study conducted to investigate the asymptotic normality of the estimators based on the proposed method for the estimation of parameters related to the second gap time.

3.5.1 Comparative Analysis: Proposed and Peng-Huang Methods for Parameter Estimation in Quantile Regression Model for the First Gap Time

The goal of our first simulation study in this section is to estimate the model parameters related to the first gap time in sequentially observed bivariate gap times settings. This case corresponds to the regular survival analysis with quantile regression. The Peng-Huang method discussed in Section 3.2 provides estimates of the model parameters, which possesses the uniform consistency and standard weak convergence to a normal distribution properties as the sample size increases. Details of these results are given in Peng and Huang (2008). Since we do not analytically investigate these properties for the proposed method, it is important to compare its performance with that of the Peng-Huang method. To do this, we conducted a Monte Carlo simulation study based on $R = 1000$ runs, and examined the bias and precision of estimators obtained from the Peng-Huang method and proposed method.

For $i = 1, 2, ..., n$, we generated the first gap time t_{1i} from the model

$$\log t_{1i} = \beta_{01} + \beta_{11} x_{1i} + \beta_{21} x_{2i} + e_i, \tag{3.95}$$

where e_i is the value of the error terms ϵ_i independently generated from the standard extreme value distribution, and x_{1i} and x_{2i} are the values of covariates X_{1i} and X_{2i}. The former covariate was generated from the standard normal distribution $N(0, 1)$ and the latter covariate was generated from a Bernoulli distribution with the probability of success set at 0.5; i.e. $Ber(0.5)$. We took $\beta_{01} = 0$, $\beta_{11} = 3.5$ and $\beta_{21} = 2$. The details of the data generation process are given in Section 3.4.2.

We generated the censoring times of individuals independently from the Uniform(0, ψ) distribution, where ψ was selected to obtain the desired approximate censoring proportion. For this simulation study, we set the value of ψ so that we obtained approximately 20% censoring out of $n = 200$ subjects. The data set generated in the rth simulation run then consisted of $D_{1r} = \{(t_{1i}^*, \delta_{1i}), x_{1i}, x_{2i}; i = 1, 2, ..., n\}$ where $t_{1i}^* = \min(t_{1i}, c_i)$, $\delta_{1i} = I(t_{1i} \leq c_i)$ and $r = 1, 2, ..., R$.

We fitted the model

$$Q_{T_1}(\tau|x_1, x_2) = \exp\{\beta_{01}(\tau) + \beta_{11}(\tau)x_1 + \beta_{21}(\tau)x_2\}, \tag{3.96}$$

where $\tau \in (0, 1)$, and $\boldsymbol{\beta}_1(\tau) = (\beta_{01}(\tau), \beta_{11}(\tau), \beta_{21}(\tau))^T$ are the parameters estimated with the generated data D_{1r}, for $r = 1, 2, ..., R$. We used the Peng-Huang method of Section 3.2 and the proposed method of Section 3.3 to estimate the parameters in Model (3.96) using the same generated data D_{1r}, $r = 1, 2, ..., R$. The estimates are obtained at $\tau = 0.1$, 0.3, 0.5 and 0.7 quantile points. For $k = 0, 1$ and 2, we let $\tilde{\beta}_{k1,r}(\tau)$ and $\hat{\beta}_{k1,r}(\tau)$ be the estimates of the parameter $\beta_{k1}(\tau)$ based on the Peng-Huang and proposed methods, respectively, obtained at the rth run of the simulation. At each τ point, we calculated the empirical bias (*EmpBias*) in the estimation of the parameter $\beta_{k1}(\tau)$, $k = 0, 1$ and 2, using the formula

$$EmpBias_{k1}(PH) = \frac{1}{R}\sum_{r=1}^{R}\big(\tilde{\beta}_{k1,r}(\tau) - \beta_{k1,r}(\tau)\big), \tag{3.97}$$

where $EmpBias_{k1}(PH)$ is the calculated bias using the Peng-Huang method. Similarly, we calculated

$$EmpBias_{k1}(PM) = \frac{1}{R}\sum_{r=1}^{R}\big(\hat{\beta}_{k1,r}(\tau) - \beta_{k1,r}(\tau)\big), \tag{3.98}$$

which gives the calculated bias based on the proposed method. Note that the parameter $\beta_{01}(\tau)$ in Model (3.96) denotes the τth quantile of the error term ϵ, which follows a standard extreme value distribution, and $\beta_{11}(\tau)$ and $\beta_{21}(\tau)$ are the values of β_{11} and β_{21} given in Model (3.95).

The calculation of the empirical standard deviations of the estimates $\tilde{\beta}_{k1}(\tau)$, where

$\bar{\tilde{\beta}}_{k1}(\tau) = \frac{1}{R}\sum_{r=1}^{R}\tilde{\beta}_{k1,r}(\tau)$, for $k = 0, 1$ and 2, was based on the formula

$$EmpSD(\tilde{\beta}_{k1}(\tau)) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\tilde{\beta}_{k1,r}(\tau) - \bar{\tilde{\beta}}_{k1}(\tau)\right)^2}. \tag{3.99}$$

A similar formula was used for the calculation of $EmpSD(\hat{\beta}_{k1}(\tau))$, for $k = 0, 1$ and 2, where $\bar{\hat{\beta}}_{k1}(\tau) = \frac{1}{R}\sum_{r=1}^{R}\hat{\beta}_{k1,r}(\tau)$. We also calculated an estimate of the variance of $\tilde{\beta}_{k1}(\tau)$ and $\hat{\beta}_{k1}(\tau)$, for $k = 0, 1$ and 2 as follows. In each simulation run r, $r = 1, 2, ..., R$, we obtained $\hat{Var}(\tilde{\beta}_{k1,r}(\tau))$ using the $crq()$ function in the *quantreg* package of the R software. For $k = 0, 1$ and 2, we then obtained

$$AveSE(\tilde{\beta}_{k1}(\tau)) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\hat{Var}(\tilde{\beta}_{k1,r}(\tau))}. \tag{3.100}$$

To obtain $\hat{Var}(\hat{\beta}_{k1,r}(\tau))$, we used the 3×3 sandwich estimator of the variance-covariance matrix given in (3.61), and took its corresponding diagonal element to find $\hat{Var}(\hat{\beta}_{k1,r}(\tau))$ for $k = 0, 1$ and 2. We then reported

$$AveSE(\hat{\beta}_{k1}(\tau)) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\hat{Var}(\hat{\beta}_{k1,r}(\tau))}. \tag{3.101}$$

Finally, we obtained the proportion of the estimated 95% standard normal distribution based confidence intervals for $\beta_{k1}(\tau)$, $k = 0, 1$ and 2, using the values of $\tilde{\beta}_{k1,r}(\tau)$ and $\hat{Var}(\tilde{\beta}_{k1,r}(\tau))$, as well as the values of $\hat{\beta}_{k1,r}(\tau)$ and $\hat{Var}(\hat{\beta}_{k1,r}(\tau))$. Since our simulation study was based on $R = 1000$ simulation runs, an empirical coverage rate below 0.9365 or above 0.9635 would indicate that the coverage rate is significantly different than the normal value of 0.95 $(0.95 \pm 1.96\sqrt{(0.05 \times 0.95)/100})$.

We presented the results of our simulation study in Table 3.1. The results based on the Peng-Huang method show that, for all τ values, the magnitude of biases in the estimates of $\beta_{01}(\tau)$, $\beta_{11}(\tau)$ and $\beta_{21}(\tau)$ are small. For example, the absolute values of EmpBias in all scenarios are less than 0.06. We also observe a similar pattern with the proposed method. Overall, the values of the EmpBias obtianed with two methods are close to each other under the same scenarios. In all scenarios, the values of AveSE and EmpSD are similar in the Peng-Huang method. We also observe a similar result for

Table 3.1: Simulation results for the first gap time when approximately 20% of the first gap time is censored.

		Proposed Method				Peng-Huang Method			
τ		EmpBias	AveSE	EmpSD	Cov95	EmpBias	AveSE	EmpSD	Cov95
0.1	$\hat{\beta}_0$	-0.004	0.224	0.233	0.943	-0.001	0.211	0.225	0.947
	$\hat{\beta}_1$	0.012	0.315	0.317	0.940	0.011	0.316	0.333	0.916
	$\hat{\beta}_2$	-0.009	0.179	0.189	0.937	-0.008	0.196	0.179	0.966
0.3	$\hat{\beta}_0$	0.005	0.298	0.279	0.961	-0.004	0.209	0.217	0.928
	$\hat{\beta}_1$	0.054	0.473	0.459	0.976	0.054	0.346	0.323	0.952
	$\hat{\beta}_2$	-0.006	0.163	0.181	0.932	-0.007	0.192	0.173	0.956
0.5	$\hat{\beta}_0$	0.033	0.259	0.239	0.955	0.033	0.253	0.239	0.905
	$\hat{\beta}_1$	-0.003	0.403	0.371	0.959	-0.003	0.401	0.372	0.917
	$\hat{\beta}_2$	0.003	0.238	0.216	0.967	0.003	0.229	0.217	0.907
0.7	$\hat{\beta}_0$	0.029	0.288	0.243	0.970	0.036	0.386	0.343	0.965
	$\hat{\beta}_1$	-0.002	0.446	0.485	0.928	-0.003	0.549	0.555	0.948
	$\hat{\beta}_2$	0.031	0.201	0.259	0.940	0.032	0.308	0.330	0.936

the proposed method. For the cross comparison of the values of AveSE and EmpSD obtained from two methods, we observe that both methods provided similar results when $\tau = 0.1$ and $\tau = 0.5$ for all $\beta_{k1}(\tau)$, $k = 0, 1, 2$, parameters. When $\tau = 0.3$, the values of AveSE and EmpSD are similar with the Peng-Huang method comparing with those values obtained with the proposed method. We observe an opposite pattern when $\tau = 0.7$. As explained in Section 3.2, this result is probably caused by the grid-based method used by the Peng-Huang method. The proposed method produced more scenarios where the empirical coverage rates (i.e., Cov95 values in Table 3.1) of $\beta_{k1}(\tau)$ within the threshold values (7 out of 12) comparing with those obtained with the Peng-Huang method (4 out of 12). However, most of the Cov95 values are very close to the interval $(0.9365, 0.9635)$.

To sum up, the simulation study considered in this section reveals that the proposed method provides similar estimates of the model parameter $\beta_{k1}(\tau)$ and their variance estimates, to those obtained with the Peng-Huang method in all scenarios considered in Table 3.1. Finally, it should be noted that our simulation results presented in Table 3.1 for the Peng-Huang method are consistent with the simulation results given by Peng and Huang (2008).

3.5.2 Comparative Analysis: Proposed and Peng-Huang Methods for Parameter Estimation in Quantile Regression Model for the Second Gap Time

In this section, we focus on the quantile regression model for the second gap time after the occurrence of the first event. We do not assume independence of the first and second gap times. As noted before, the first and second events can be of different types. We conducted a Monte Carlo simulation study to assess the magnitude of the bias and precision of estimators with QR model based on the proposed method and the naive Peng-Huang method.

We conducted $R = 1000$ Monte Carlo simulation runs. As explained in the previous section, to generate gap time t_{1i}, we considered the following model. For $i = 1, 2, ..., n$,

$$\log t_{1i} = \beta_{01} + \beta_{11}x_{1i} + \beta_{21}x_{2i} + e_{1i}, \tag{3.102}$$

where x_{1i} and x_{2i} are the values of the covariates generated from $N(0, 1)$ and $Ber(0.5)$ distributions, respectively. We used the following model to generate the second gap time t_{2i}. For $i = 1, 2, ..., n$,

$$\log t_{2i} = \beta_{02} + \beta_{12}x_{1i} + \beta_{22}x_{2i} + e_{2i}, \tag{3.103}$$

where the same values of x_{1i} and x_{2i} were used. In Models (3.102) and (3.103), we specified $\beta_{01} = 0$, $\beta_{11} = 2$, $\beta_{21} = 1$, $\beta_{02} = 0$, $\beta_{12} = 3.5$ and $\beta_{22} = 2$, and generated the values of the error terms $\epsilon_{1i} = e_{1i}$ and $\epsilon_{2i} = e_{2i}$ from the standard extreme value distribution. To create the dependency between ϵ_{1i} and ϵ_{2i} values, we used the Clayton copula model with the dependence parameter ϕ so that the dependence between the gap times T_1 and T_2 was established. Note that the Kendall's tau parameter τ_ϕ and the Clayton copula parameter ϕ have the relation $\tau_\phi = \frac{\phi}{\phi+2}$. We also generated a censoring value for each individual from a Uniform$(0, \psi)$ distribution, where the value of ψ was chosen so that an approximate proportion of censoring of the first gap times was obtained. Then in each simulation run, $r = 1, 2, ..., R$, we obtained the data set $D_{2r} = \{(t^*_{1i}, t^*_{2i}, \delta_{1i}, \delta_{2i}), x_{1i}, x_{2i};\ i = 1, 2, ..., n\}$, where $t^*_{1i} = \min(t_{1i}, c_i)$, $\delta_{1i} = I\{t_{1i} \leq c_i\}$, $t^*_{2i} = \min(t_{2i}, c_i - t_{1i})$, $\delta_{2i} = I\{\delta_{1i} = 1 \text{ and } t_{2i} \leq c_i - t_{1i}\}$, for $i = 1, 2, ..., n$. The details of data generation is explained in Section 3.4.3.

The factors of this simulation study included the sample size n (n = 200, 500), the approximate censoring proportion of the first gap time ($c\%$ = 20%, 40%) and the Kendall's tau coefficient τ_ϕ ($\tau_\phi = 0.2, 0.4, 0.6$). We applied a full factorial simulation design with these factors, and presented the results in 12(= 2 × 2 × 3) tables each defined by a combination of (n, $c\%$, τ_ϕ) as a simulation scenario.

At each simulation run r, $r = 1, 2, ..., R$, we fitted the conditional QR model

$$Q_{T_2}(\tau|x_1, x_2) = \exp\{\beta_{02}(\tau) + \beta_{12}(\tau)x_1 + \beta_{22}(\tau)x_2\}, \tag{3.104}$$

using the generated data set D_{2r} and obtained the estimates of model parameters $\beta_{02}(\tau)$, $\beta_{12}(\tau)$ and $\beta_{22}(\tau)$, and their variance estimates using the Peng-Huang and proposed methods at four different quantile values of the distribution of ϵ_2 (τ = 0.1, 0.3, 0.5 and 0.7). It should be noted that the estimated parameters obtained from the naive Peng-Huang method are considered as the initial values for estimating the parameters $\beta_{k2}(\tau)$, $k = 0, 1, 2$, in the Newton-Raphson algorithm in the proposed method given in (3.71). For each k, where $k = 0, 1$, and 2, $\tilde{\beta}_{k2,r}(\tau)$ and $\hat{\beta}_{k2,r}(\tau)$ denote the estimates of the parameter $\beta_{k2}(\tau)$ obtained from the Peng-Huang and proposed methods in the rth simulation run, respectively. At each τ point, we computed the empirical bias (*EmpBias*) in estimating the parameter $\beta_{k2}(\tau)$, $k = 0, 1$ and 2, using the formula

$$EmpBias_{k2}(PH) = \frac{1}{R}\sum_{r=1}^{R}\big(\tilde{\beta}_{k2,r}(\tau) - \beta_{k2,r}(\tau)\big), \tag{3.105}$$

where $EmpBias_{k2}(PH)$ represents the bias calculated using the Peng-Huang method. Similarly, we calculated

$$EmpBias_{k2}(PM) = \frac{1}{R}\sum_{r=1}^{R}\big(\hat{\beta}_{k2,r}(\tau) - \beta_{k2,r}(\tau)\big), \tag{3.106}$$

which gives the empirical bias calculated using the proposed method. Note that the parameter $\beta_{02}(\tau)$ in Model (3.104) denotes the τth quantile of the error term ϵ_2, following a standard extreme value distribution, and $\beta_{12}(\tau)$ and $\beta_{22}(\tau)$ are the values of β_{12} and β_{22} given in Model (3.103).

We computed the empirical standard deviations of the estimates $\tilde{\beta}_{k2}(\tau)$, where

$\bar{\tilde{\beta}}_{k2}(\tau) = \frac{1}{R}\sum_{r=1}^{R}\tilde{\beta}_{k2,r}(\tau)$, for $k = 0, 1$ and 2, using the formula

$$EmpSD(\tilde{\beta}_{k2}(\tau)) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\tilde{\beta}_{k2,r}(\tau) - \bar{\tilde{\beta}}_{k2}(\tau)\right)^2}. \tag{3.107}$$

We applied a similar formula to compute $EmpSD(\hat{\beta}_{k2}(\tau))$, for $k = 0, 1$ and 2, where $\bar{\hat{\beta}}_{k2}(\tau) = \frac{1}{R}\sum_{r=1}^{R}\hat{\beta}_{k2,r}(\tau)$. Additionally, we calculated the estimates of the variance for $\tilde{\beta}_{k2}(\tau)$ and $\hat{\beta}_{k2}(\tau)$, for $k = 0, 1$ and 2 as follows. In each simulation run r, $r = 1, 2, ..., R$, we obtained $\hat{Var}(\tilde{\beta}_{k2,r}(\tau))$ utilizing the *crq*() function in the *quantreg* package in R software. Subsequently, for $k = 0, 1$, and 2, we then calculated

$$AveSE(\tilde{\beta}_{k2}(\tau)) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\hat{Var}(\tilde{\beta}_{k2,r}(\tau))}. \tag{3.108}$$

This computation provided the average of the estimated standard deviation for $\tilde{\beta}_{k2}(\tau)$ across the simulation runs. As for the proposed method, we calculated $\hat{Var}(\hat{\beta}_{k2,r}(\tau))$ by utilizing the 3×3 sandwich estimator of the variance-covariance matrix as explained in (3.76). Extracting the corresponding diagonal element provided the value of $\hat{Var}(\hat{\beta}_{k2,r}(\tau))$ for $k = 0, 1$ and 2. Subsequently, we reported

$$AveSE(\hat{\beta}_{k2}(\tau)) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\hat{Var}(\hat{\beta}_{k2,r}(\tau))}. \tag{3.109}$$

Finally, we determined the proportion of the estimated 95% confidence intervals for $\beta_{k2}(\tau)$, $k = 0, 1$ and 2. These intervals were based on the regular standard normal approximations and computed using the values of $\tilde{\beta}_{k2,r}(\tau)$ and $\hat{Var}(\tilde{\beta}_{k2,r}(\tau))$, as well as the values of $\hat{\beta}_{k2,r}(\tau)$ and $\hat{Var}(\hat{\beta}_{k2,r}(\tau))$. Given that our simulation study was based on $R = 1000$ simulation runs, an empirical coverage rate below 0.9365 or above 0.9635 would indicate a significant deviation from the expected normal value of 0.95 for the coverage rate.

The results for scenario ($n = 200$, $c\% = 0.2$, $\tau_\phi = 0.2$) is given in Table 3.2 for the values of $\tau = 0.1, 0.3, 0.5$ and 0.7. Upon comparing the magnitude of the empirical bias of the two methods, the results given in Table 3.2 indicate that the estimates of the model parameters based on the proposed method typically exhibit lower bias when compared to those obtained with the Peng-Huang method. Furthermore, the proposed method consistently outperforms the naive Peng-Huang method, yielding more accurate coverage probabilities across various quantiles and parameters. These results, along with the calculated coverage probabilities, indicate that the estimation of the model parameters with the proposed method is more precise and reliable, effectively capturing the true parameter values within the specified 95% standard normal approximation based confidence intervals, comparing with the naive Peng-Huang method.

It is noteworthy that, with both methods, as censoring increases, the empirical bias also increases, accompanied by an increase in the average of the estimated standard error (AveSE) and empirical standard deviation (EmpSD). Additionally as empirical coverage rates (Cov95) indicates, the confidence level becomes less accurate as censoring increases. Conversely, with an increase in the sample size n, the empirical bias decreases. This reduction in bias is accompanied by a decrease in the average estimated standard error (AveSE) and empirical standard deviation (EmpSD). Importantly, as the sample size n increases, the confidence level becomes more accurate, contributing to a more reliable estimation of the parameters.

The results in all tables indicates that the naive Peng-Huang method produced the values of the average estimated standard errors (AveSE) and empirical standard deviations (EmpSD) closely aligned. However, the absolute values of the empirical bias computed using this method yielded higher values, and less accurate values of the confidence level, compared with the proposed method. On the other hand, the values of the average estimated standard errors (AveSE) and empirical standard deviations (EmpSD) may not always be close to each other in the proposed method. Nevertheless, the empirical bias computed using this method is lower, and the empirical coverage rates (Cov95) are closer to the nominal 0.95 level in most of the scenarios, compared with the naive Peng-Huang method. This suggests that the effect of variability in estimation of the model parameters is not as significant when utilizing the proposed method, emphasizing its robustness in providing accurate parameter estimates.

Table 3.2: Simulation results for the second gap time when approximately 20% of the first gap time is censored, Kendall's Tau equals 0.2 and $n = 200$.

		Proposed Method				Peng-Huang Method			
τ		EmpBias	AveSE	EmpSD	Cov95	EmpBias	AveSE	EmpSD	Cov95
0.1	$\hat{\beta}_0$	0.055	0.175	0.166	0.973	-0.062	0.163	0.182	0.906
	$\hat{\beta}_1$	-0.012	0.140	0.125	0.969	0.013	0.122	0.135	0.842
	$\hat{\beta}_2$	-0.029	0.263	0.249	0.969	-0.048	0.239	0.265	0.920
0.3	$\hat{\beta}_0$	0.075	0.178	0.157	0.962	-0.083	0.155	0.160	0.845
	$\hat{\beta}_1$	-0.001	0.145	0.121	0.959	0.002	0.117	0.124	0.872
	$\hat{\beta}_2$	-0.044	0.278	0.249	0.961	-0.073	0.245	0.246	0.904
0.5	$\hat{\beta}_0$	-0.088	0.193	0.216	0.938	-0.129	0.165	0.182	0.819
	$\hat{\beta}_1$	-0.006	0.126	0.135	0.942	-0.011	0.124	0.141	0.908
	$\hat{\beta}_2$	-0.010	0.254	0.227	0.944	-0.066	0.251	0.274	0.916
0.7	$\hat{\beta}_0$	-0.105	0.206	0.215	0.933	-0.191	0.205	0.210	0.753
	$\hat{\beta}_1$	-0.015	0.147	0.157	0.938	-0.055	0.146	0.156	0.887
	$\hat{\beta}_2$	-0.070	0.289	0.301	0.943	-0.119	0.283	0.311	0.889

Table 3.3: Simulation results for the second gap time when approximately 20% of the first gap time is censored, Kendall's Tau equals 0.4 and $n = 200$.

		Proposed Method				Peng-Huang Method			
τ		EmpBias	AveSE	EmpSD	Cov95	EmpBias	AveSE	EmpSD	Cov95
0.1	$\hat{\beta}_0$	0.028	0.156	0.146	0.957	-0.108	0.143	0.163	0.799
	$\hat{\beta}_1$	-0.003	0.134	0.117	0.963	0.003	0.115	0.117	0.838
	$\hat{\beta}_2$	-0.004	0.215	0.164	0.968	-0.069	0.212	0.244	0.894
0.3	$\hat{\beta}_0$	0.037	0.211	0.194	0.961	-0.167	0.147	0.158	0.748
	$\hat{\beta}_1$	-0.004	0.100	0.115	0.948	0.013	0.109	0.116	0.885
	$\hat{\beta}_2$	0.016	0.221	0.212	0.955	-0.067	0.211	0.229	0.920
0.5	$\hat{\beta}_0$	-0.172	0.190	0.196	0.949	-0.201	0.162	0.174	0.720
	$\hat{\beta}_1$	-0.007	0.135	0.124	0.955	0.0082	0.123	0.135	0.910
	$\hat{\beta}_2$	-0.017	0.230	0.231	0.946	-0.121	0.229	0.258	0.890
0.7	$\hat{\beta}_0$	-0.183	0.157	0.166	0.938	-0.310	0.186	0.206	0.598
	$\hat{\beta}_1$	-0.008	0.122	0.108	0.968	-0.009	0.140	0.157	0.878
	$\hat{\beta}_2$	-0.018	0.195	0.172	0.963	-0.169	0.274	0.309	0.886

Table 3.4: Simulation results for the second gap time when approximately 20% of the first gap time is censored, Kendall's Tau equals 0.6 and $n = 200$.

		Proposed Method				Peng-Huang Method			
τ		EmpBias	AveSE	EmpSD	Cov95	EmpBias	AveSE	EmpSD	Cov95
0.1	$\hat{\beta}_0$	0.083	0.138	0.149	0.936	-0.156	0.138	0.162	0.716
	$\hat{\beta}_1$	0.007	0.091	0.103	0.945	0.017	0.097	0.110	0.809
	$\hat{\beta}_2$	0.005	0.206	0.223	0.938	-0.063	0.207	0.231	0.893
0.3	$\hat{\beta}_0$	0.200	0.104	0.112	0.941	-0.224	0.134	0.138	0.569
	$\hat{\beta}_1$	-0.002	0.101	0.095	0.960	0.006	0.097	0.105	0.899
	$\hat{\beta}_2$	-0.003	0.125	0.151	0.930	-0.104	0.185	0.206	0.906
0.5	$\hat{\beta}_0$	-0.303	0.091	0.106	0.948	-0.355	0.148	0.158	0.392
	$\hat{\beta}_1$	0.002	0.113	0.116	0.939	0.004	0.110	0.120	0.884
	$\hat{\beta}_2$	0.002	0.175	0.120	0.960	-0.128	0.215	0.234	0.876
0.7	$\hat{\beta}_0$	-0.430	0.144	0.152	0.946	-0.480	0.183	0.194	0.312
	$\hat{\beta}_1$	-0.001	0.110	0.138	0.948	0.004	0.137	0.147	0.881
	$\hat{\beta}_2$	-0.002	0.182	0.172	0.961	-0.196	0.249	0.280	0.865

Table 3.5: Simulation results for the second gap time when approximately 40% of the first gap time is censored, Kendall's Tau equals 0.2 and $n = 200$.

		Proposed Method				Peng-Huang Method			
τ		EmpBias	AveSE	EmpSD	Cov95	EmpBias	AveSE	EmpSD	Cov95
0.1	$\hat{\beta}_0$	0.081	0.227	0.261	0.937	-0.301	0.237	0.263	0.667
	$\hat{\beta}_1$	-0.016	0.246	0.219	0.970	0.022	0.212	0.242	0.848
	$\hat{\beta}_2$	-0.037	0.418	0.394	0.974	-0.126	0.402	0.434	0.857
0.3	$\hat{\beta}_0$	0.165	0.223	0.243	0.963	-0.395	0.227	0.231	0.534
	$\hat{\beta}_1$	-0.027	0.169	0.207	0.926	-0.036	0.194	0.208	0.904
	$\hat{\beta}_2$	-0.045	0.403	0.349	0.968	-0.174	0.364	0.385	0.878
0.5	$\hat{\beta}_0$	-0.369	0.196	0.227	0.929	-0.501	0.206	0.308	0.326
	$\hat{\beta}_1$	-0.036	0.171	0.197	0.933	-0.060	0.174	0.178	0.875
	$\hat{\beta}_2$	-0.056	0.309	0.346	0.925	-0.282	0.311	0.345	0.843
0.7	$\hat{\beta}_0$	-0.414	0.234	0.275	0.931	-0.744	0.222	0.327	0.154
	$\hat{\beta}_1$	-0.076	0.171	0.162	0.968	-0.145	0.184	0.189	0.820
	$\hat{\beta}_2$	-0.108	0.294	0.309	0.941	-0.352	0.312	0.346	0.784

Table 3.6: Simulation results for the second gap time when approximately 40% of the first gap time is censored, Kendall's Tau equals 0.4 and $n = 200$.

		Proposed Method				Peng-Huang Method			
τ		EmpBias	AveSE	EmpSD	Cov95	EmpBias	AveSE	EmpSD	Cov95
0.1	$\hat{\beta}_0$	0.084	0.178	0.196	0.935	-0.393	0.178	0.230	0.513
	$\hat{\beta}_1$	-0.003	0.181	0.159	0.980	0.008	0.159	0.214	0.885
	$\hat{\beta}_2$	-0.005	0.351	0.331	0.975	-0.143	0.303	0.364	0.894
0.3	$\hat{\beta}_0$	0.099	0.222	0.197	0.971	-0.585	0.195	0.244	0.207
	$\hat{\beta}_1$	-0.005	0.197	0.161	0.965	0.011	0.161	0.169	0.895
	$\hat{\beta}_2$	-0.019	0.294	0.314	0.922	-0.185	0.295	0.321	0.877
0.5	$\hat{\beta}_0$	-0.356	0.220	0.198	0.966	-0.722	0.191	0.203	0.137
	$\hat{\beta}_1$	-0.010	0.194	0.165	0.971	-0.917	0.165	0.174	0.892
	$\hat{\beta}_2$	0.021	0.292	0.314	0.924	-0.257	0.292	0.321	0.846
0.7	$\hat{\beta}_0$	-0.402	0.216	0.236	0.936	-0.926	0.210	0.214	0.055
	$\hat{\beta}_1$	-0.020	0.173	0.166	0.974	-0.057	0.167	0.180	0.874
	$\hat{\beta}_2$	0.027	0.314	0.301	0.969	-0.345	0.302	0.324	0.766

Table 3.7: Simulation results for the second gap time when approximately 40% of the first gap time is censored, Kendall's Tau equals 0.6 and $n = 200$.

		Proposed Method				Peng-Huang Method			
τ		EmpBias	AveSE	EmpSD	Cov95	EmpBias	AveSE	EmpSD	Cov95
0.1	$\hat{\beta}_0$	0.085	0.169	0.152	0.970	-0.407	0.172	0.199	0.354
	$\hat{\beta}_1$	0.008	0.143	0.109	0.963	0.028	0.149	0.155	0.790
	$\hat{\beta}_2$	-0.007	0.231	0.271	0.932	-0.149	0.297	0.306	0.819
0.3	$\hat{\beta}_0$	0.209	0.131	0.120	0.963	-0.634	0.153	0.182	0.108
	$\hat{\beta}_1$	-0.008	0.134	0.105	0.971	-0.017	0.114	0.135	0.834
	$\hat{\beta}_2$	-0.009	0.171	0.190	0.926	-0.200	0.228	0.271	0.872
0.5	$\hat{\beta}_0$	-0.335	0.148	0.126	0.964	-0.876	0.168	0.270	0.013
	$\hat{\beta}_1$	-0.009	0.135	0.120	0.965	0.011	0.126	0.148	0.819
	$\hat{\beta}_2$	-0.011	0.206	0.218	0.930	-0.238	0.227	0.268	0.876
0.7	$\hat{\beta}_0$	-0.452	0.162	0.153	0.957	-1.119	0.171	0.282	0.021
	$\hat{\beta}_1$	-0.010	0.130	0.153	0.937	-0.016	0.139	0.163	0.896
	$\hat{\beta}_2$	-0.015	0.218	0.234	0.928	-0.302	0.230	0.267	0.735

Table 3.8: Simulation results for the second gap time when approximately 20% of the first gap time is censored, Kendall's Tau equals 0.2 and $n = 500$.

		Proposed Method				Peng-Huang Method			
τ		EmpBias	AveSE	EmpSD	Cov95	EmpBias	AveSE	EmpSD	Cov95
0.1	$\hat{\beta}_0$	0.020	0.086	0.097	0.939	-0.056	0.101	0.106	0.851
	$\hat{\beta}_1$	-0.005	0.073	0.065	0.961	0.011	0.078	0.084	0.865
	$\hat{\beta}_2$	-0.002	0.141	0.123	0.960	-0.048	0.152	0.162	0.902
0.3	$\hat{\beta}_0$	0.066	0.083	0.089	0.937	-0.096	0.100	0.104	0.767
	$\hat{\beta}_1$	-0.009	0.040	0.037	0.958	0.010	0.075	0.080	0.908
	$\hat{\beta}_2$	-0.008	0.106	0.118	0.944	-0.058	0.149	0.156	0.896
0.5	$\hat{\beta}_0$	-0.074	0.072	0.088	0.939	-0.113	0.106	0.110	0.754
	$\hat{\beta}_1$	-0.010	0.048	0.043	0.953	-0.014	0.081	0.084	0.884
	$\hat{\beta}_2$	-0.011	0.103	0.094	0.957	-0.091	0.159	0.167	0.884
0.7	$\hat{\beta}_0$	-0.105	0.059	0.092	0.935	-0.172	0.128	0.133	0.695
	$\hat{\beta}_1$	-0.013	0.061	0.043	0.960	-0.049	0.090	0.099	0.875
	$\hat{\beta}_2$	-0.014	0.059	0.099	0.945	-0.128	0.180	0.195	0.871

Table 3.9: Simulation results for the second gap time when approximately 20% of the first gap time is censored, Kendall's Tau equals 0.4 and $n = 500$.

		Proposed Method				Peng-Huang Method			
τ		EmpBias	AveSE	EmpSD	Cov95	EmpBias	AveSE	EmpSD	Cov95
0.1	$\hat{\beta}_0$	0.018	0.057	0.036	0.952	-0.125	0.096	0.130	0.680
	$\hat{\beta}_1$	-0.002	0.043	0.052	0.939	0.056	0.067	0.074	0.894
	$\hat{\beta}_2$	-0.002	0.174	0.162	0.960	-0.059	0.138	0.152	0.898
0.3	$\hat{\beta}_0$	0.037	0.058	0.053	0.958	-0.192	0.089	0.094	0.440
	$\hat{\beta}_1$	-0.005	0.044	0.059	0.947	0.013	0.070	0.073	0.887
	$\hat{\beta}_2$	-0.003	0.121	0.126	0.944	-0.091	0.133	0.143	0.872
0.5	$\hat{\beta}_0$	-0.085	0.084	0.093	0.949	-0.249	0.102	0.148	0.361
	$\hat{\beta}_1$	-0.007	0.088	0.052	0.951	0.011	0.075	0.084	0.900
	$\hat{\beta}_2$	0.010	0.133	0.117	0.951	-0.131	0.152	0.161	0.843
0.7	$\hat{\beta}_0$	-0.103	0.132	0.117	0.959	-0.314	0.129	0.162	0.361
	$\hat{\beta}_1$	-0.007	0.046	0.041	0.957	-0.021	0.088	0.099	0.902
	$\hat{\beta}_2$	-0.013	0.166	0.144	0.959	-0.189	0.181	0.189	0.803

Table 3.10: Simulation results for the second gap time when approximately 20% of the first gap time is censored, Kendall's Tau equals 0.6 and $n = 500$.

		Proposed Method				Peng-Huang Method			
τ		EmpBias	AveSE	EmpSD	Cov95	EmpBias	AveSE	EmpSD	Cov95
0.1	$\hat{\beta}_0$	0.033	0.085	0.048	0.964	-0.144	0.086	0.094	0.612
	$\hat{\beta}_1$	-0.001	0.061	0.043	0.951	0.036	0.065	0.070	0.879
	$\hat{\beta}_2$	0.004	0.120	0.092	0.958	-0.060	0.131	0.139	0.896
0.3	$\hat{\beta}_0$	0.096	0.101	0.110	0.947	-0.254	0.083	0.086	0.203
	$\hat{\beta}_1$	-0.002	0.099	0.079	0.957	0.010	0.063	0.064	0.893
	$\hat{\beta}_2$	-0.002	0.124	0.119	0.958	-0.102	0.119	0.128	0.853
0.5	$\hat{\beta}_0$	-0.136	0.087	0.070	0.953	-0.341	0.089	0.095	0.106
	$\hat{\beta}_1$	-0.002	0.093	0.101	0.940	0.091	0.067	0.071	0.905
	$\hat{\beta}_2$	0.001	0.117	0.092	0.957	-0.143	0.126	0.141	0.803
0.7	$\hat{\beta}_0$	-0.143	0.142	0.151	0.948	-0.493	0.111	0.116	0.064
	$\hat{\beta}_1$	-0.001	0.097	0.074	0.953	0.007	0.081	0.088	0.906
	$\hat{\beta}_2$	0.003	0.157	0.131	0.960	-0.204	0.160	0.172	0.749

Table 3.11: Simulation results for the second gap time when approximately 40% of the first gap time is censored, Kendall's Tau equals 0.2 and $n = 500$.

		Proposed Method				Peng-Huang Method			
τ		EmpBias	AveSE	EmpSD	Cov95	EmpBias	AveSE	EmpSD	Cov95
0.1	$\hat{\beta}_0$	0.079	0.161	0.196	0.939	-0.246	0.148	0.155	0.559
	$\hat{\beta}_1$	-0.009	0.137	0.142	0.937	0.014	0.127	0.129	0.848
	$\hat{\beta}_2$	-0.004	0.152	0.135	0.961	-0.110	0.241	0.267	0.872
0.3	$\hat{\beta}_0$	0.157	0.135	0.099	0.968	-0.327	0.128	0.137	0.346
	$\hat{\beta}_1$	-0.010	0.119	0.126	0.935	-0.017	0.106	0.114	0.894
	$\hat{\beta}_2$	-0.016	0.291	0.273	0.965	-0.183	0.202	0.225	0.850
0.5	$\hat{\beta}_0$	-0.303	0.084	0.077	0.964	-0.434	0.134	0.138	0.190
	$\hat{\beta}_1$	-0.031	0.170	0.151	0.963	-0.057	0.109	0.115	0.867
	$\hat{\beta}_2$	-0.015	0.111	0.128	0.940	-0.240	0.193	0.219	0.787
0.7	$\hat{\beta}_0$	-0.406	0.106	0.144	0.934	-0.667	0.131	0.136	0.027
	$\hat{\beta}_1$	-0.055	0.171	0.152	0.963	-0.126	0.109	0.115	0.733
	$\hat{\beta}_2$	-0.019	0.158	0.173	0.942	-0.329	0.187	0.208	0.618

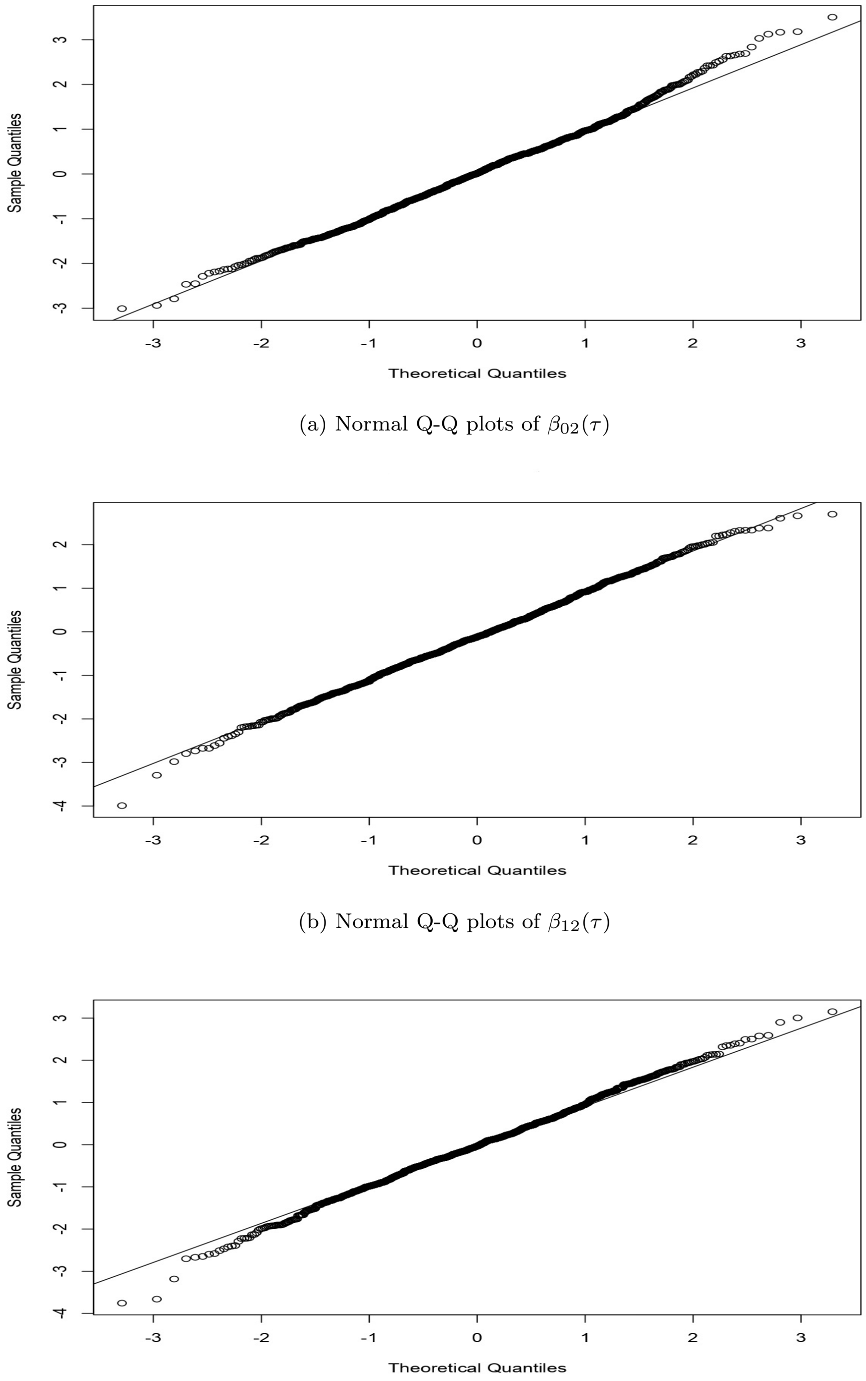
Table 3.12: Simulation results for the second gap time when approximately 40% of the first gap time is censored, Kendall's Tau equals 0.4 and $n = 500$.

		Proposed Method				Peng-Huang Method			
τ		EmpBias	AveSE	EmpSD	Cov95	EmpBias	AveSE	EmpSD	Cov95
0.1	$\hat{\beta}_0$	0.082	0.139	0.183	0.938	-0.347	0.117	0.128	0.278
	$\hat{\beta}_1$	-0.007	0.073	0.058	0.963	0.065	0.100	0.105	0.863
	$\hat{\beta}_2$	0.005	0.182	0.176	0.960	-0.130	0.178	0.211	0.855
0.3	$\hat{\beta}_0$	0.179	0.186	0.190	0.937	-0.510	0.111	0.119	0.146
	$\hat{\beta}_1$	-0.007	0.049	0.067	0.941	0.070	0.090	0.098	0.882
	$\hat{\beta}_2$	0.008	0.127	0.138	0.931	-0.192	0.174	0.188	0.773
0.5	$\hat{\beta}_0$	-0.311	0.115	0.098	0.963	-0.691	0.117	0.123	0.115
	$\hat{\beta}_1$	-0.008	0.149	0.116	0.964	-0.031	0.099	0.103	0.879
	$\hat{\beta}_2$	0.008	0.173	0.130	0.961	-0.245	0.180	0.198	0.727
0.7	$\hat{\beta}_0$	-0.393	0.152	0.172	0.941	-0.883	0.130	0.140	0.111
	$\hat{\beta}_1$	-0.009	0.087	0.115	0.938	-0.043	0.105	0.113	0.878
	$\hat{\beta}_2$	0.018	0.175	0.198	0.960	-0.323	0.187	0.211	0.626

3.5.3 Normal Quantile-Quantile (Q-Q) Plots of the Estimated Parameters

In this section, we present the results of our last simulation study conducted to investigate the adequacy of the asymptotic normality of the estimators based the proposed method. To do so, we generated the data set D_{2r}, $r = 1, 2, ..., R = 1000$, as explained before and obtained the standard normal quantile-quantile (Q-Q) plots of the estimates of parameters in Model (3.103) using the proposed method. When $\tau = 0.5$ and approximately 20% of the first gap times are censored, the Q-Q plots of $(\hat{\beta}_{k2}(\tau) - \beta_{k2}(\tau))/\sqrt{\hat{Var}(\hat{\beta}_{k2}(\tau))}$, $k = 0, 1$ and 2, with a sample size of 150, 250 and 500 are given in Figure 3.5, Figure 3.6 and Figure 3.7, respectively. The Q-Q plots in these figures indicate the plausibility of the normal distribution approximates of the estimates of $\beta_{02}(\tau)$, $\beta_{12}(\tau)$, and $\beta_{22}(\tau)$ with the proposed method. Consequently, we can use the standard normal approximation to make decisions regarding the significance of covariates for sufficiently large sample sizes.

When $\tau = 0.5$ and approximate 40% of the first gap times are censored, the Q-Q plots for the investigation of the estimates of $\beta_{02}(\tau)$, $\beta_{12}(\tau)$, and $\beta_{22}(\tau)$ are presented in Figure 3.8, Figure 3.9, and Figure 3.10 for sample sizes of 150, 250, and 500, respectively. These standard normal probability (Q-Q) plots indicate that, under heavy right-censoring (40% censoring), the normal approximations are off in the extreme upper and lower tails of the distribution when $n = 150$ and $\tau = 0.5$ for the distribution of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$. However, the normal distribution is more plausible when $n = 250$ and 500.

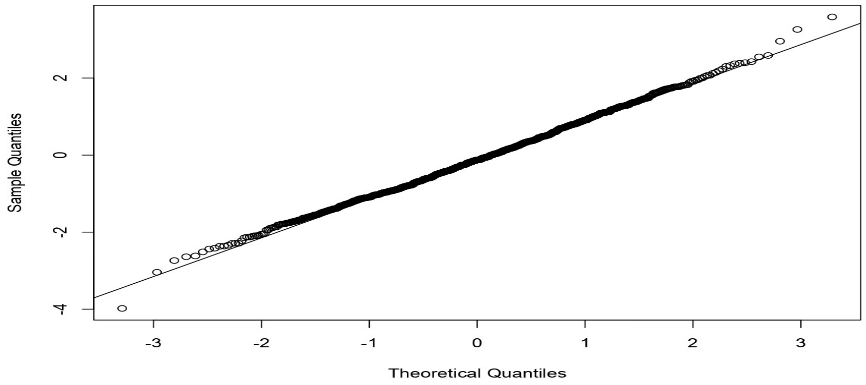


(a) Normal Q-Q plots of $\beta_{02}(\tau)$

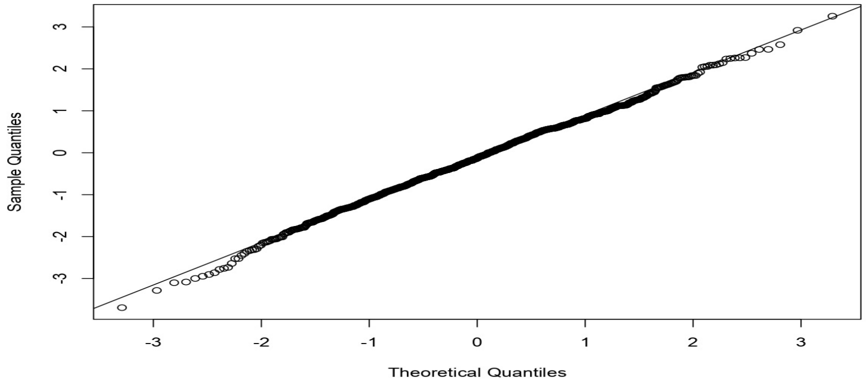
(b) Normal Q-Q plots of $\beta_{12}(\tau)$

(c) Normal Q-Q plots of $\beta_{22}(\tau)$

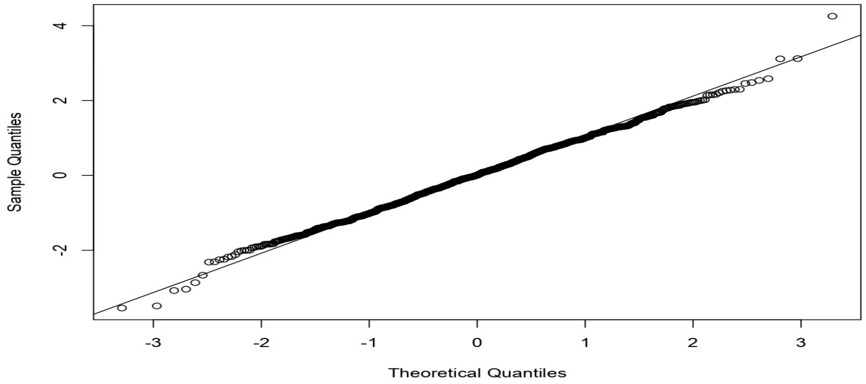
Figure 3.5: Normal Q-Q plots of $\beta_{02}(\tau)$, $\beta_{12}(\tau)$ and $\beta_{22}(\tau)$ with a sample size of 150 and 20% censored data when $\tau = 0.5$.



(a) Normal Q-Q plots of $\beta_{02}(\tau)$

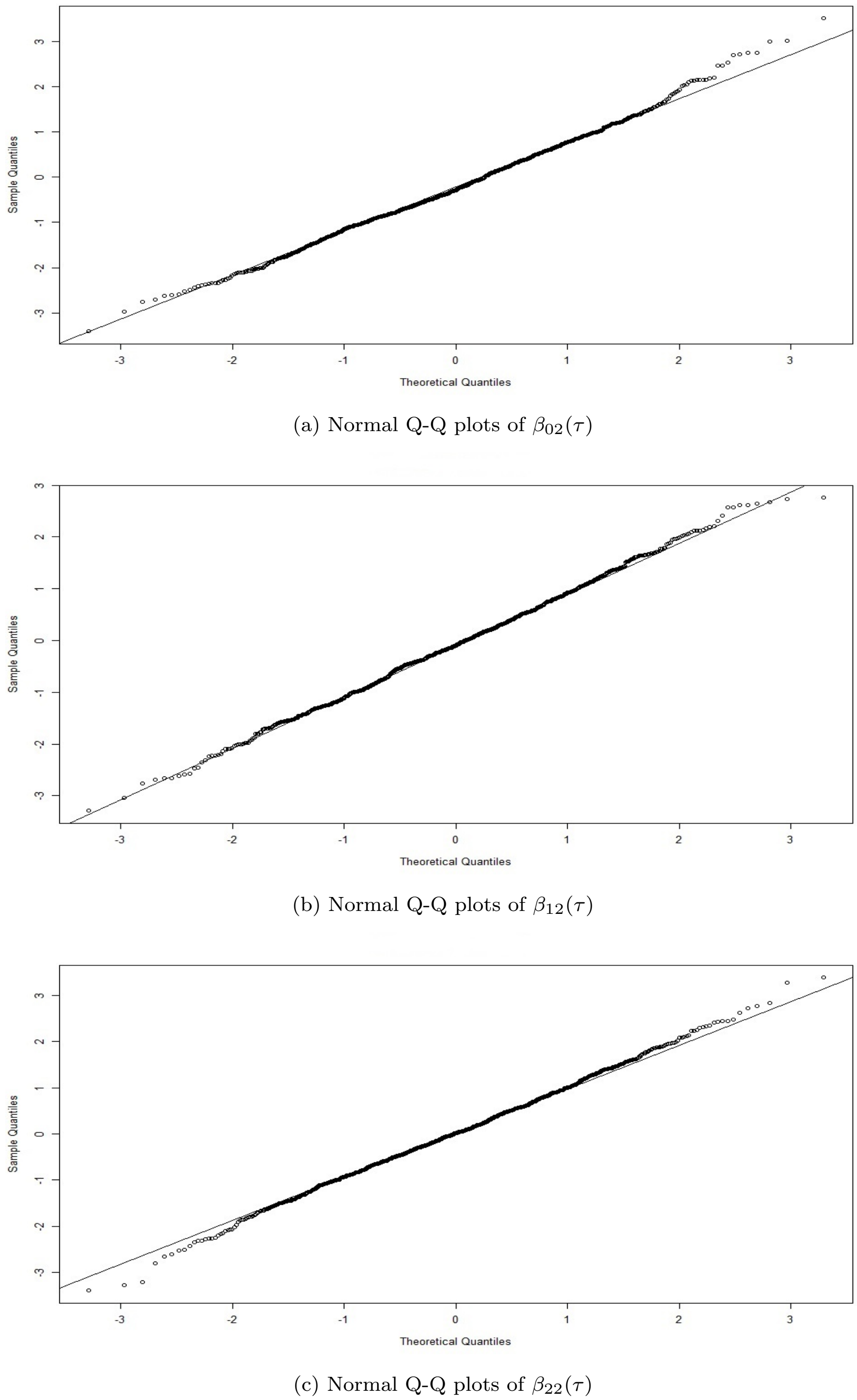


(b) Normal Q-Q plots of $\beta_{12}(\tau)$



(c) Normal Q-Q plots of $\beta_{22}(\tau)$

Figure 3.6: Normal Q-Q plots of $\beta_{02}(\tau)$, $\beta_{12}(\tau)$ and $\beta_{22}(\tau)$ with a sample size of 250 and 20% censored data when $\tau = 0.5$.

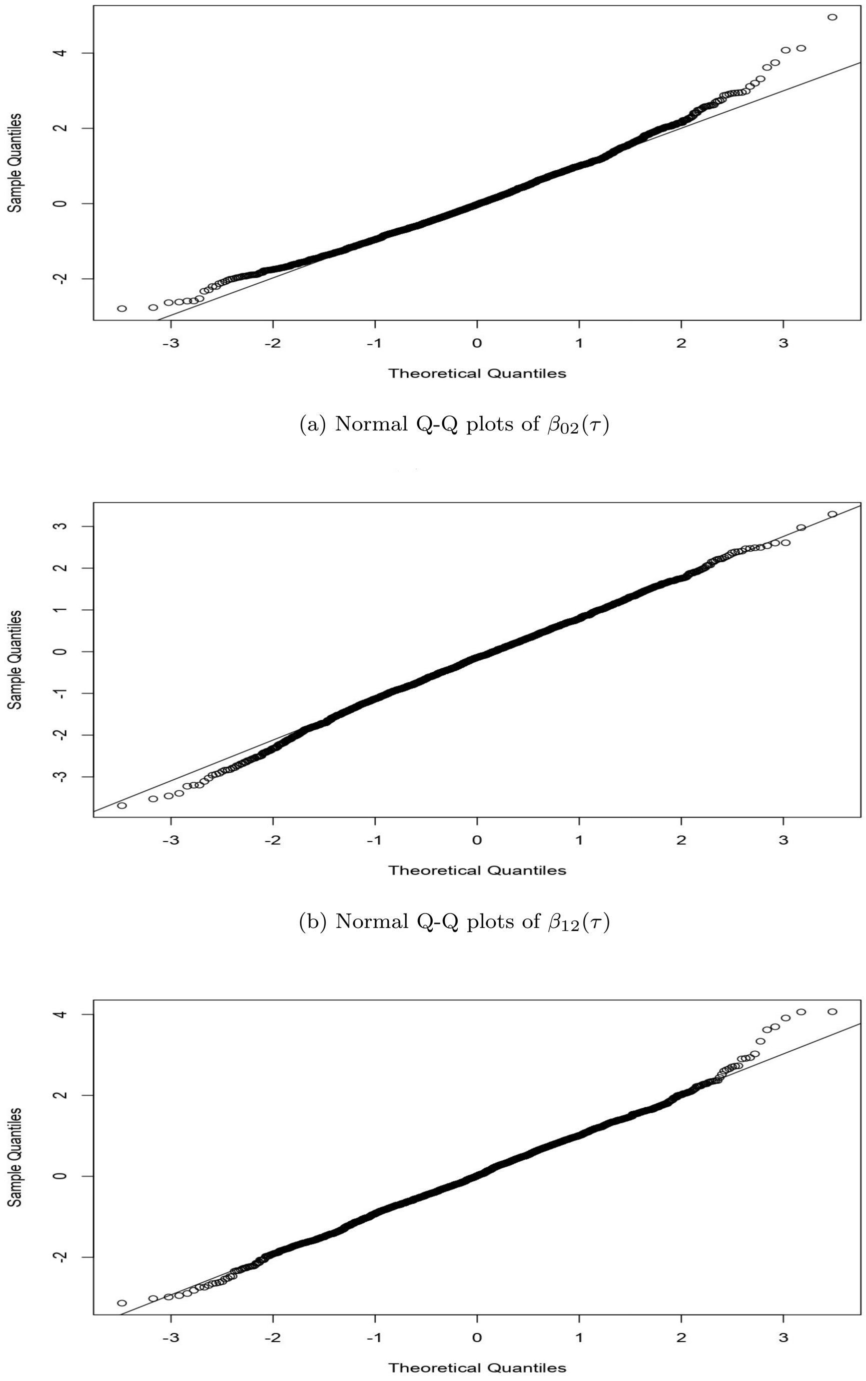


(a) Normal Q-Q plots of $\beta_{02}(\tau)$

(b) Normal Q-Q plots of $\beta_{12}(\tau)$

(c) Normal Q-Q plots of $\beta_{22}(\tau)$

Figure 3.7: Normal Q-Q plots of $\beta_{02}(\tau)$, $\beta_{12}(\tau)$ and $\beta_{22}(\tau)$ with a sample size of 500 and 20% censored data when $\tau = 0.5$.

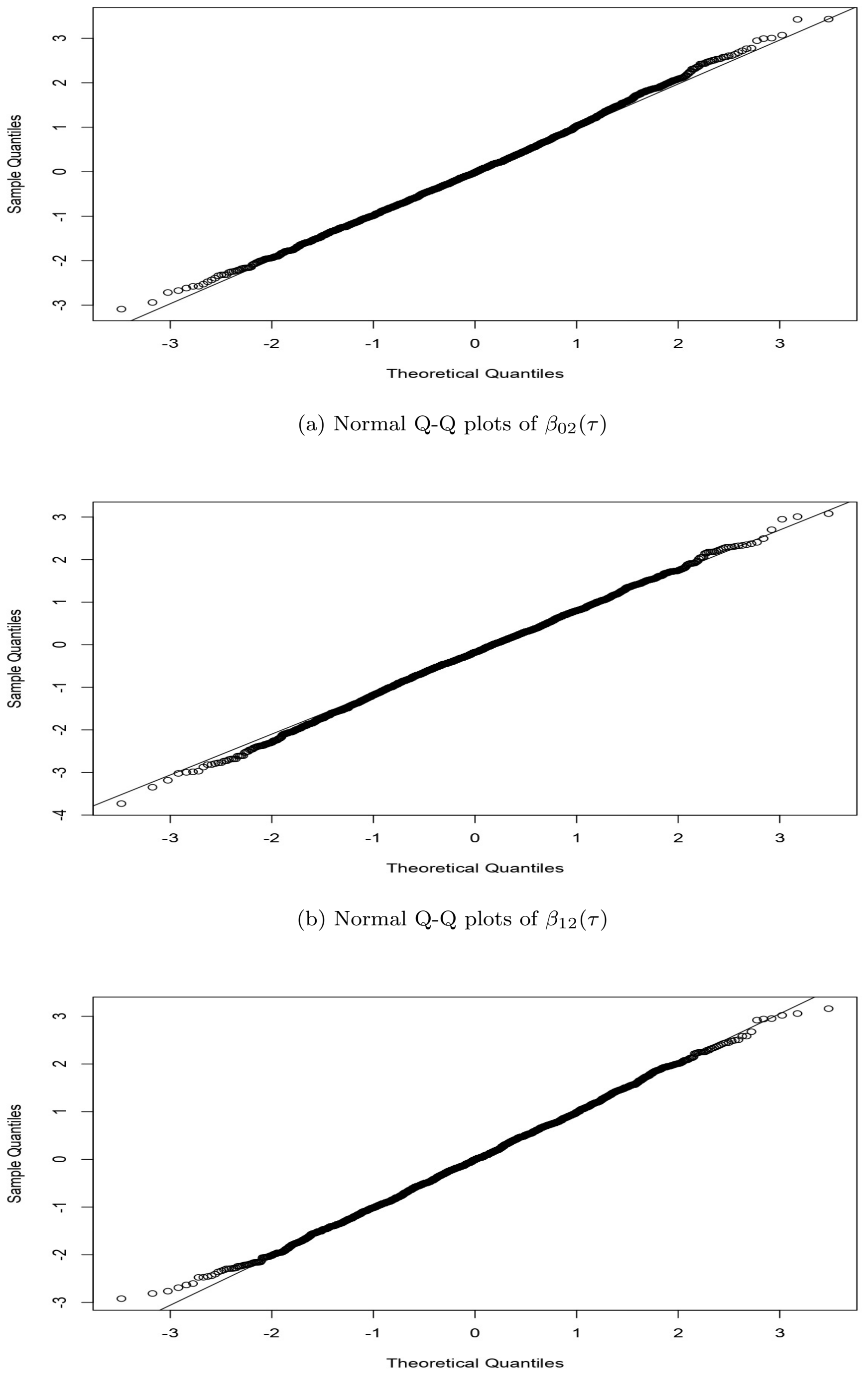


(a) Normal Q-Q plots of $\beta_{02}(\tau)$

(b) Normal Q-Q plots of $\beta_{12}(\tau)$

(c) Normal Q-Q plots of $\beta_{22}(\tau)$

Figure 3.8: Normal Q-Q plots of $\beta_{02}(\tau)$, $\beta_{12}(\tau)$ and $\beta_{22}(\tau)$ with a sample size of 150 and 40% censored data when $\tau = 0.5$.

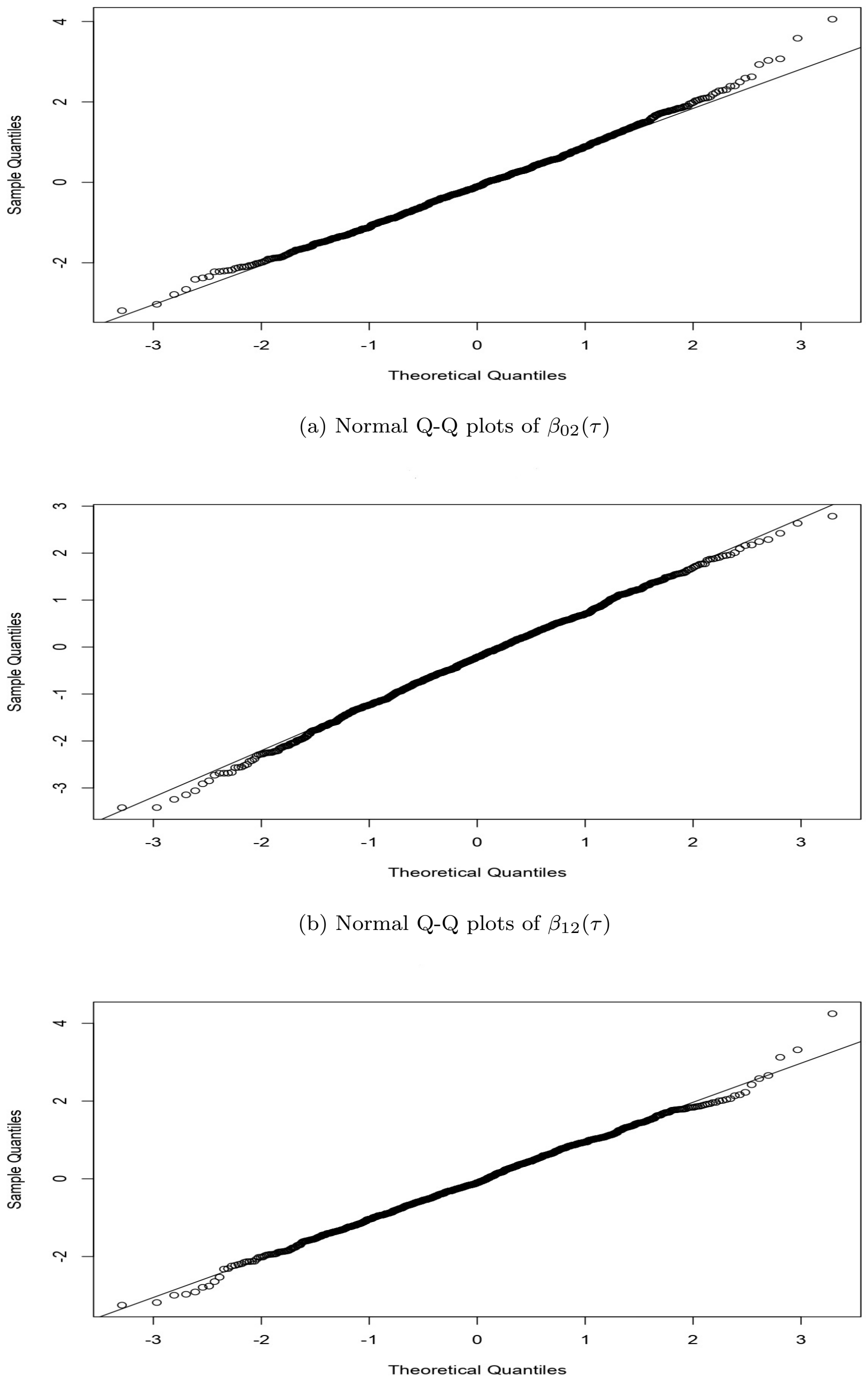


(a) Normal Q-Q plots of $\beta_{02}(\tau)$

(b) Normal Q-Q plots of $\beta_{12}(\tau)$

(c) Normal Q-Q plots of $\beta_{22}(\tau)$

Figure 3.9: Normal Q-Q plots of $\beta_{02}(\tau)$, $\beta_{12}(\tau)$ and $\beta_{22}(\tau)$ with a sample size of 250 and 40% censored data when $\tau = 0.5$.



(a) Normal Q-Q plots of $\beta_{02}(\tau)$

(b) Normal Q-Q plots of $\beta_{12}(\tau)$

(c) Normal Q-Q plots of $\beta_{22}(\tau)$

Figure 3.10: Normal Q-Q plots of $\beta_{02}(\tau)$, $\beta_{12}(\tau)$ and $\beta_{22}(\tau)$ with a sample size of 500 and 40% censored data when $\tau = 0.5$.

Chapter 4

Analysis of Colon Cancer Data Set

In this chapter, we apply our proposed method to analyze a colon cancer data set, which consists of survival times and several covariates obtained from patients diagnosed with Duke's Stage C colon cancer. The data set can be found in `survival` package in R software, encompassing individuals distributed among the Observation (placebo control), Levamisole (treatment), and Levamisole plus 5FU (treatment) groups.

The rest of this chapter is organized as follows. In Section 4.1, we provide a descriptive analysis of the data set. Subsequently, in Section 4.2, we apply quantile regression models to the data set and employ the proposed method and the Peng-Huang method to make inference on the parameters of these models.

4.1 Descriptive Analysis of Colon Cancer Data Set

The Colon Cancer data set given in the `survival` package in R, comprises data on 929 patients who received surgery of removal of tumors. We introduce that data set in Section 1.1.2. It consists of 14 variables, which are listed below.

1. **rx:** Treatment type (Obs = Observation(placebo), Lev = Levamisole, Lev+5FU= Levamisole+Fluorouracil)
2. **sex:** Gender (0=Female, 1=Male)
3. **age:** Age in years at the registration time

4. **obstruct:** Presence or absence of colon obstruction due to a tumor (0=Absence, 1=Presence)

5. **perfor:** Presence or absence of colon perforation (0=Absence, 1=Presence)

6. **adhere:** Adherence of the tumor to nearby organs (0=Not adherence, 1=Adherence)

7. **nodes:** Number of cancer-affected lymph nodes

8. **time:** Days until an event or censoring

9. **status:** Censoring status (0=Censored, 1=Not censored)

10. **differ:** Tumor differentiation grade (1=Well, 2=Moderate, 3=Poor)

11. **extent:** Extent of local spread of the cancer (1=Submucosa, 2=Muscle, 3=Serosa, 4=Contiguous structures)

12. **surg:** Time from surgery to registration (0=Short, 1=Long)

13. **node4:** Presence of more than four positive lymph nodes (0=Absence, 1=Presence)

14. **etype:** Event type (1=Recurrence, 2=Death)

The patients have an average age of around 60 years, ranging from 18 to 85 years. Out of these patients, 43 individuals died without facing a recurrence of colon cancer, resulting in their exclusion from our analysis. Consequently, we focus on analyzing the data of 886 patients who either experienced the first event or were censored in terms of their first gap time. In this cohort of patients, a total of 463 individuals experienced a recurrence of colon cancer, whereas 423 patients remained cancer-free and alive throughout the duration of the study. Additionally, 409 patients experienced a recurrence of colon cancer and ultimately died. Furthermore, there were 54 patients who, despite experiencing a recurrence of colon cancer, were still alive at the end of the study. The second gap times of those patients were censored. Therefore, a total of 477 patients have censored gap times.

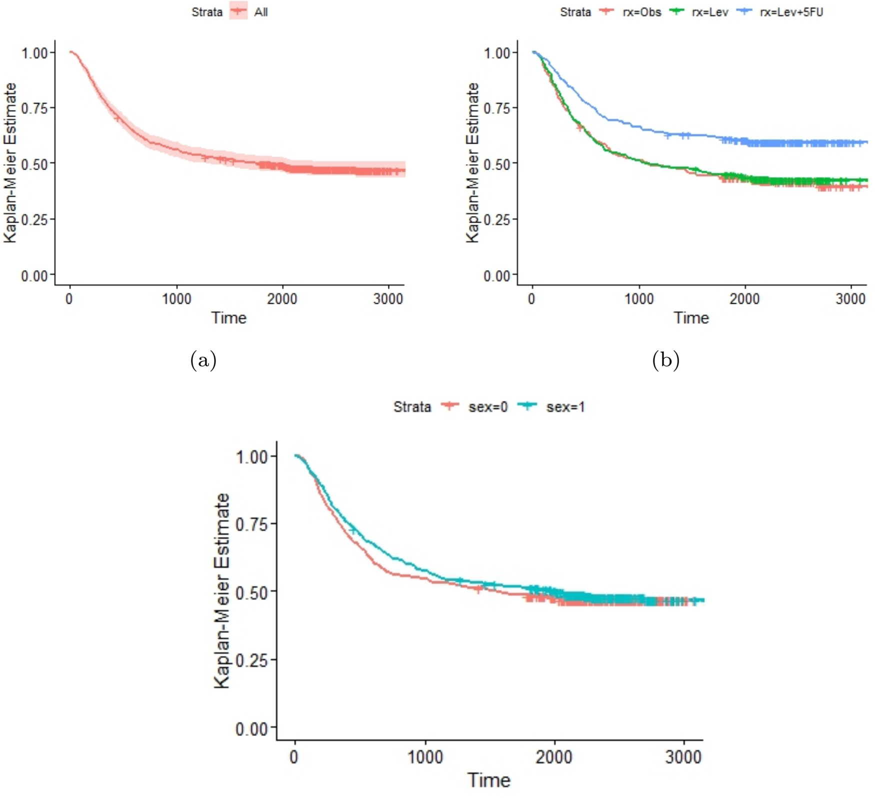
The mean of the first gap times, denoting the time until recurrence of colon cancer, was approximately 1418 days. In terms of treatments, patients received either placebo

Table 4.1: Descriptive Analysis of Colon Cancer Data.

	Obs **(N=300)**	**Lev** **(N=300)**	**Lev+5FU** **(N=286)**	**Overall** **(N=886)**
T_1				
Mean(SD)	1285 (984)	1328 (1014)	1653 (981)	1418 (1005)
Median[Min, Max]	1056 [20, 3192]	1040 [19, 3329]	2060 [8, 3309]	1625 [8, 3329]
status 1				
Censored	125 (41.67%)	128 (42.7%)	170 (59.4%)	423 (47.7%)
Not Censored	175 (58.33%)	172 (57.3%)	116 (40.6%)	463 (52.3%)
T_2				
Mean(SD)	335 (479)	309 (476)	185 (371)	278 (450)
Median[Min, Max]	130 [0, 2725]	110 [0, 2515]	0 [0, 2184]	34 [0, 2725]
status 2				
Censored	147 (49%)	149 (49.7%)	181 (63.3%)	477 (53.8%)
Not Censored	153 (51%)	151 (50.3%)	105 (36.7%)	409 (46.2%)
Age				
Mean(SD)	59 (12)	60 (12)	60 (12)	60 (12)
Median[Min, Max]	60 [18,85]	61 [27, 83]	62 [26, 81]	61 [18, 85]

(300 patients), Lev (300 patients), or Lev+5FU (286 patients). Comparing the treatments, patients who received Lev+5FU exhibited the longest average first gap time (1653 days), followed by those who received Lev (1328 days) and then placebo (1285 days). However, it is important to note that more patients had censored data for the first gap time in the Lev+5FU group (170 patients) compared to the placebo (125 patients) and Lev (128 patients) groups. When considering the recurrence of colon cancer, the placebo group had the highest number of patients experiencing the first event (175 patients), followed by the Lev group (172 patients) and then the Lev+5FU group (116 patients). However, the Lev+5FU group included more patients who did not observe either the first or second event (181 patients) compared to the placebo (147 patients) and Lev (179 patients) groups at the end of the study. Furthermore, a comparison of deaths among patients who experienced a recurrence of colon cancer revealed that the Lev+5FU group had the lowest number of deaths (105 patients), followed by the Lev group (151 patients), and then the placebo group (153 patients). Also, the patients in the placebo group lived longer on the average after the recurrence of colon cancer compared to the treatment groups. Table 4.1 shows the summary statistics for the control (Obs) and treatment groups (Lev and Lev+5FU).

We next present the results of our analysis conducted to investigate the statistical significance of the effects of covariates within the context of the analysis concerning the recurrence of colon cancer and the mortality of patients experienced colon cancer recurrence. This analysis encompasses separate investigations of the first and second gap times. Subsequently, we present a comprehensive summary of findings derived from both the log-rank test and Kaplan-Meier analysis in the following paragraphs.



(a)

(b)

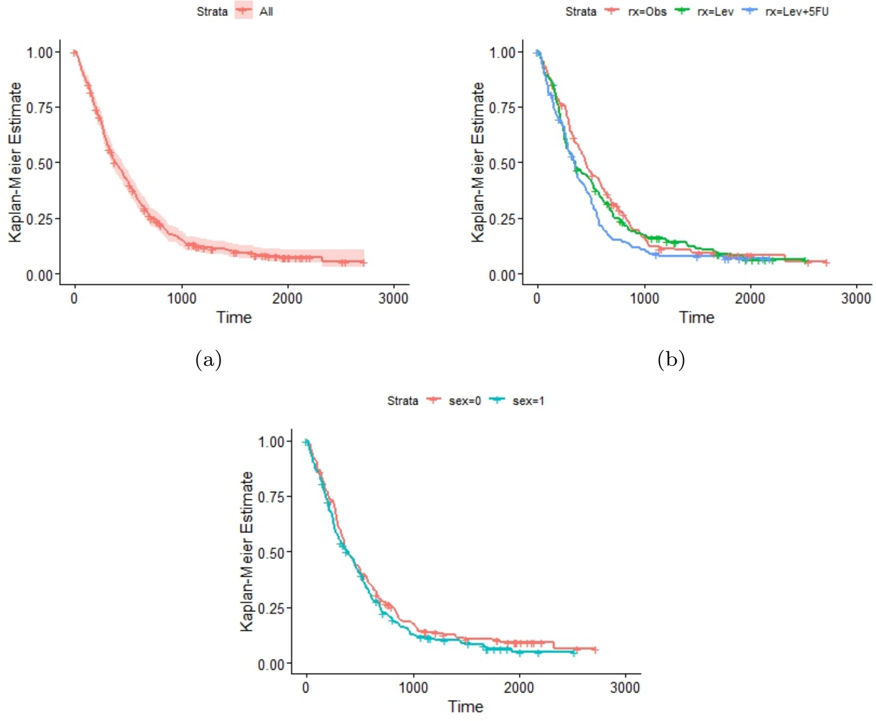
(c)

Figure 4.1: Plots of the Kaplan-Meier survival function estimates for the duration time (in days) from surgery to recurrence of colon cancer; (a) all patients, (b) breakdown with respect to treatment groups, (c) breakdown with respect to the sex covariate.

The plot (a) given in Figure 4.1 depicts the Kaplan-Meier estimates of the survivor function for the first gap time of all the patients included in the study. At time 365 days (after one year), 667 patients were at risk and 220 patients experienced the recurrence of colon cancer. The Kaplan-Meier estimate was 0.752 with an approximate 95% CI given by (0.724,0.781), indicating that around 75% of the patients had not yet experienced the recurrence of colon cancer by the end of the first year. After two years, 531 patients were at risk and 355 patients experienced the recurrence event.

The estimated survival probability was 0.599 (approximate 95% CI: 0.568-0.632), indicating that around 60% of the patients had not yet experienced the recurrence of colon cancer by the end of the second year. At the end of the study, more than half of the patients in the cohort have not experienced the recurrence of colon cancer. We also present the plots of the Kaplan-Meier estimates of the survival probability for each treatment group (Obs, Lev, and Lev+5FU), which are given in Figure 4.1(b). Patients treated with Lev+5FU showed the highest survival rates at all time points, indicating that this treatment was more effective in prolonging the time to recurrence of colon cancer compared to Lev alone and Obs (i.e. placebo group). The log-rank test indicates that there is a significant difference (p-value < 0.0001) between the survival probabilities of time to colon cancer recurrence for patients in the Lev+5FU group compared with those in Lev and Obs groups, whereas there is no significant difference for the comparison of Lev and Obs groups (p-value$= 0.8$). The last plot in Figure 4.1 shows the Kaplan-Meier estimates of time to cancer recurrence for male and female patients, separately. The plot indicates that there is not significant difference between male and female patients (p-value$= 0.4$).

We also analyzed the time from recurrence of colon cancer to death (i.e. the second gap time) by using the Kaplan-Meier estimates of the survival probabilities and log-rank test. It should be noted that, even though these tools produce unbiased results in the analysis of the first gap times, they may induce significant bias for the analysis of the second and subsequent gap times when the gap times are not independent (Lin et al., 1999; Lawless and Yilmaz, 2011). We therefore refer to the Kaplan-Meier estimate as the naive Kaplan-Meier estimate in this case. The graphical representations of survival probabilities of the time from colon cancer recurrence to death over time are depicted as plots of the naive Kaplan-Meier estimates in Figure 4.2; for all patients in plot (a), separately for patients in specific treatment groups in plot (b), and separately for patients with respect to the sex covariate in plot (c). From the plot (a), we observe that the median survival time to death after cancer recurrence is 374 days with an approximate 95% CI $(341, 448)$ and the naive Kaplan-Meier estimate of the survival probability is less than 0.25 at the 1000 day (approximate 95% CI: $0.121 - 0.19$). A comparison of the plots given in Figure 4.2(b) reveals that the lowest survival rates are in the Lev+5FU group in most of the time points. Notably, patients in the Obs (placebo) group exhibited the highest survival rates for the second gap time, followed by the Lev treatment group, indicating an adverse effect of the Lev and Lev+5FU



(a)

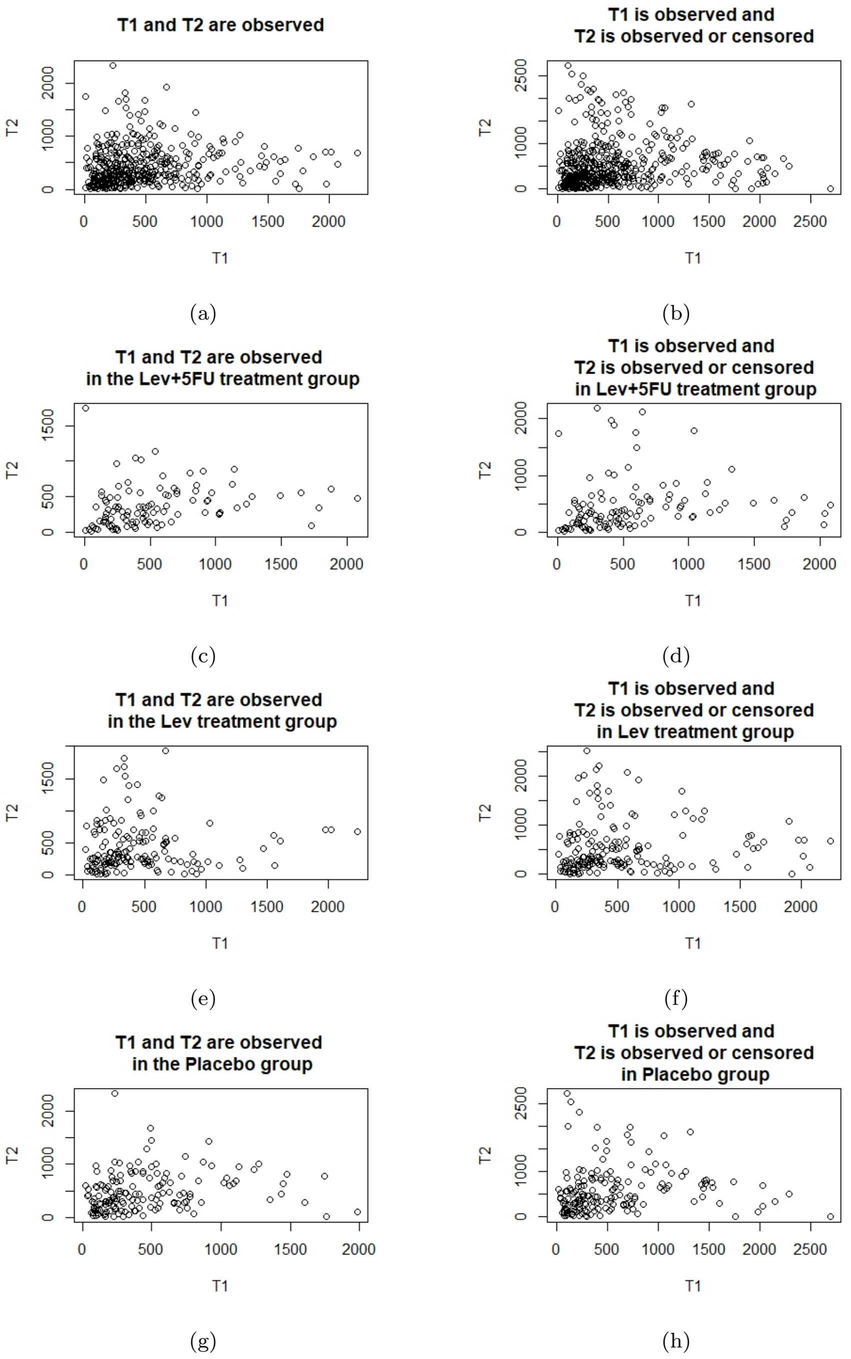
(b)

(c)

Figure 4.2: Plots of the naive Kaplan-Meier survival function estimates for the duration time (in days) from recurrence of colon cancer to death; (a) all patients, (b) breakdown with respect to treatment groups, (c) breakdown with respect to the sex covariate.

treatments on the patients. The plot of the naive Kaplan-Meier estimates given in Figure 4.2(c) denotes that the estimated survival probabilities of female patients are higher than those of male patients in most of the time.

We present the scatter plots of the first gap times T_1 and the second gap times T_2 in Figure 4.3 under different settings. The plots (a), (c), (e) and (g) are based on patients with complete T_1 and T_2 observations. Patients with censored observations are discarded. The other plots in Figure 4.3 are based on patients with complete T_1 observations and complete or censored T_2 observations. In all settings, the scatter plots reveal a positive but relatively weak association between the two gap times (the estimated Kendall's tau coefficients are 0.168 and 0.172, respectively), an important aspect to consider in our analysis. Due to the exclusion of patients with censored first gap times, there is a substantial proportion of patients (47.7%) who are not represented in the plots given in Figure 4.3. This exclusion may lead to a misleading



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 4.3: Scatter plots of the first gap time against the second gap time under different settings.

Table 4.2: The results of the log-rank tests for the first and second gap times in colon cancer patients.

	Patients (n)	**p-value** (T_1)	**p-value** (T_2)
Sex			
Male	459		
Female	427	0.4	0.2
Treatment			
Observation	300		
Lev	300	0.8	0.4
Lev+5FU	286	8e-06*	0.02*
Age			
$\leq 60y$	438		
$> 60y$	448	0.9	0.01*
Obstruction			
Yes	171		
No	715	0.03*	0.1
Perforation			
Yes	27		
No	859	0.2	0.2
Adherence to nearby organs			
Yes	129		
No	757	0.005*	0.5
Involved nodes			
$1 - 4$	644		
> 4	242	$< 2e - 16$*	9e-07*
Differentiation			
Well	90		
Moderate	631	0.6	0.5
Poor	143	0.01*	0.1
Extent			
Submucosa	20		
Muscle	101	0.5	0.1
Serosa	726	0.02*	0.06*
Contiguous structure	39	4e-04*	0.1
Time from surgery to registration			
Short	654		
Long	232	0.01*	0.6

*Significant at 0.05 level.

interpretation of dependency. To address this issue, we investigate the dependency through dependence modeling with copulas in the next section. Also, the plots in Figure 4.3 indicate that the Clayton copula might be an appropriate model for the dependency between T_1 and T_2.

Table 4.2 provides a comprehensive summary of the results of the log-rank tests applied for testing the statistical significance of various covariates related to the first and second gap times in colon cancer patients. The goal of the analysis is to assess the significance of these factors on the recurrence of colon cancer and survival of patients after the cancer recurrence. In Table 4.2, the significant covariates (p-value$<$ 0.05) are denoted by "*" notation. In the analysis of the first gap time T_1, the covariates sex, age and perforation are not significant (p-value$>$ 0.05). However, the treatment Lev+5FU is strongly significant comparing with the treatment group (Lev) and the placebo group (Obs) in the analysis of the first gap times (both p-values$<$ 0.0001). Furthermore, obstruction (p-value$=$ 0.03), adherence to nearby organs (p-value$=$ 0.005), nodes (p-value$<$ 0.0001), differentiation (p-value$=$ 0.002), extent (p-value$<$ 0.0001), and time from surgery to registration (p-value$=$ 0.01) are significant covariates in the analysis of the first gap time. In the examination of the second gap time, age (p-value$=$ 0.01) and nodes (p-value$<$ 0.0001) are significant factors in the duration from the cancer recurrence to death. As aforementioned, the log-rank test may lead to wrong results in the analysis of the second gap times, when there is dependency between the first and second gap times. Because of this reason, all covariates are included in the analysis given in the next section for a comprehensive assessment.

4.2 Quantile Regression Model for Colon Cancer Data Set

In this section, we employ quantile regression (QR) to analyze the colon cancer data set, aiming to understand the effects of the covariates on the time from the study entry to cancer recurrence as well as time from cancer recurrence to death. We first focus on the time until the recurrence of colon cancer (i.e. the first gap time T_1). The conditional QR model for log of the first gap time T_1, given the values of covariate $\mathbf{x}$, is given by

$$Q_{\log(T_1)}(\tau|\mathbf{x}) = (1, \mathbf{x}^T)^T \boldsymbol{\beta}(\tau), \qquad \tau \in (0,1), \tag{4.1}$$

where $\boldsymbol{\beta}$ denotes a vector of regression parameters and $\mathbf{x}$ is a vector of observed covariates including sex, age, obstruction, perforation, adherence to nearby organs, the number of involved nodes, differentiation, extent, time from surgery to registration, presence of more than 4 positive lymph nodes and treatment. Thus, the regression model of $\log(T_1)$ for the analysis of the first gap time is given by

$$\begin{aligned}\log(T_1) = \beta_{01} &+ \beta_{11}\text{sex} + \beta_{21}\text{age} + \beta_{31}\text{obstruct} + \beta_{41}\text{perfor} + \beta_{51}\text{adhere} \\ &+ \beta_{61}\text{nodes} + \beta_{71}\text{differ} + \beta_{81}\text{extent} + \beta_{91}\text{surg} + \beta_{10,1}\text{node4} \\ &+ \beta_{11,1}\text{rx\_Lev} + \beta_{12,1}\text{rx\_Lev5FU} + \epsilon_1,\end{aligned} \tag{4.2}$$

where ϵ_1 follows the logistic distribution. The covariate rx_Lev takes the value of 1 if a patient is under Lev treatment, otherwise it is 0. Similarly, rx_Lev5FU is 1 if the patient is under Lev+5FU treatment, otherwise it is 0. The values of other covariates are given in the previous section.

We fit Model (4.1), where $\log(T_1)$ is given in (4.2) with the proposed method and the Peng-Huang (PH) method for the quantile value τ, where $\tau = 0.1$, 0.3 and 0.5. The estimates of the parameters and their estimated standard deviations, as well as p-values for testing $H_0 : \beta_{k1} = 0$ against $H_1 : \beta_{k1} \neq 0$, $k = 1, 2, ..., 12$, are presented in Table 4.3. Lawless (2003) demonstrated the adequacy of the log-logistic distribution for both the first and second gap times in analyzing this colon cancer data set. Consequently, we used the log-logistic models with the c.d.f.

$$F(t) = 1 - \left[1 + (\frac{t}{\alpha})^\gamma\right]^{-1}, \quad t > 0,\ \alpha > 0,\ \gamma > 0, \tag{4.3}$$

for both first and second gap times. The parameters α and γ were estimated using the maximum likelihood estimation method.

The parameter estimates and their estimated standard deviations are given in Table 4.3. The results show that the estimated standard deviations derived from the proposed method closely aligns with those obtained from the Peng-Huang method. In the majority of cases, the estimates of the standard deviations acquired through the proposed method are lower than those obtained with the Peng-Huang method. It

Table 4.3: Quantile regression (based on the *Proposed* method and the Peng-Huang *PH* method) estimate for the time until the recurrence of colon cancer.

τ	Method		Intercept	sex	age	obstruct	perfor	adhere	nodes	differ	extent	surg	node4	rx-Lev	rx-Lev5FU
			$\hat{\beta}_{01}$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\beta}_{31}$	$\hat{\beta}_{41}$	$\hat{\beta}_{51}$	$\hat{\beta}_{61}$	$\hat{\beta}_{71}$	$\hat{\beta}_{81}$	$\hat{\beta}_{91}$	$\hat{\beta}_{10,1}$	$\hat{\beta}_{11,1}$	$\hat{\beta}_{12,1}$
0.1	Proposed	Estimate	7.726	0.089	-0.009	-0.638	0.288	0.023	-0.052	-0.228	-0.410	-0.190	-0.620	0.027	0.485
		SD	0.375	0.090	0.004	0.144	0.277	0.138	0.033	0.098	0.079	0.107	0.213	0.120	0.108
		p-value	< 0.001	0.324	0.027*	< 0.001*	0.299	0.865	0.108	0.020*	< 0.001*	0.076	0.004*	0.821	< 0.001*
	PH	Estimate	7.734	0.087	-0.008	-0.619	0.278	0.023	-0.050	-0.222	-0.401	-0.184	-0.602	0.026	0.473
		SD	0.780	0.125	0.006	0.182	0.267	0.152	0.029	0.152	0.175	0.171	0.205	0.140	0.151
		p-value	< 0.001	0.487	0.197	< 0.001*	0.298	0.880	0.080	0.144	0.022*	0.281	0.003*	0.851	0.002*
0.3	Proposed	Estimate	9.443	0.318	0.001	-0.380	0.126	-0.279	-0.053	-0.203	-0.868	-0.442	-0.697	0.031	0.711
		SD	0.730	0.120	0.006	0.166	0.356	0.190	0.035	0.133	0.186	0.141	0.244	0.152	0.148
		p-value	< 0.001	0.008*	0.854	0.022*	0.724	0.141	0.123	0.127	< 0.001*	0.002*	0.004*	0.839	< 0.001*
	PH	Estimate	9.539	0.313	0.0008	-0.376	0.122	-0.273	-0.052	-0.200	-0.881	-0.439	-0.690	0.029	0.706
		SD	0.775	0.165	0.006	0.226	0.248	0.178	0.029	0.152	0.210	0.158	0.219	0.197	0.179
		p-value	< 0.001	0.057	0.892	0.096	0.621	0.125	0.073	0.189	< 0.001*	0.005*	0.002*	0.881	< 0.001*
0.5	Proposed	Estimate	11.884	-0.061	0.011	-0.392	-0.375	-0.460	-0.096	-0.303	-1.240	-0.600	-1.222	0.080	1.258
		SD	2.038	0.189	0.008	0.226	0.443	0.254	0.054	0.207	0.590	0.254	0.357	0.191	0.371
		p-value	< 0.001	0.746	0.188	0.082	0.398	0.071	0.072	0.143	0.036*	0.018*	< 0.001*	0.676	< 0.001*
	PH	Estimate	12.562	-0.071	0.012	-0.414	-0.410	-0.467	-0.095	-0.319	-1.429	-0.658	-1.270	0.059	1.367
		SD	1.613	0.263	0.010	0.274	0.377	0.296	0.042	0.304	0.360	0.247	0.404	0.227	0.508
		p-value	< 0.001	0.787	0.262	0.131	0.276	0.115	0.026*	0.294	< 0.001*	0.008*	0.002*	0.796	0.007*

should be noted that, for this data set, the Peng-Huang method provided reasonable estimates of the model parameters when the quantile τ is less than 0.5 (i.e. median of the distribution of $\log T_1$). The inability to estimate the regression coefficients for higher quantiles in the Peng-Huang method may be attributed to the substantial number of censored observations in the tails of the survival function. We discussed this issue in Section 3.2. For all quantiles $\tau = 0.1$, 0.3 and 0.5, the covariates extent, node4 and the treatment Lev+5FU are found to be significant with both methods. When $\tau = 0.1$, in addition to the mentioned covariates, the covariates age and obstruct are also significant. When $\tau = 0.5$, alongside the mentioned covariates, the covariate sex and obstruct are identified as other significant factors.

In the analysis of the time from the recurrence of colon cancer to the death, we employed QR to gain deeper insights into the relationship between various covariates and the duration until the death occurrence. The regression model for the second gap time is formulated as follows.

$$\begin{aligned}\log(T_2) = \beta_{02} &+ \beta_{12}\text{sex} + \beta_{22}\text{age} + \beta_{32}\text{obstruct} + \beta_{42}\text{perfor} + \beta_{52}\text{adhere}\\ &+ \beta_{62}\text{nodes} + \beta_{72}\text{differ} + \beta_{82}\text{extent} + \beta_{92}\text{surg} + \beta_{10,2}\text{node4}\\ &+ \beta_{11,2}\text{rx\_Lev} + \beta_{12,2}\text{rx\_Lev5FU} + \epsilon_2,\end{aligned} \tag{4.4}$$

where T_2 represents the time from the recurrence of colon cancer to the death and ϵ_2 is the error term from the logistic distribution with the c.d.f. given in (4.3). In the proposed method, we employed the log-logistic models for both gap times and used the Clayton copula for modeling the dependency between the gap times. Because of its one-to-one relation with the gamma frailty model (Goethals et al., 2008) and its flexibility in modeling bivariate survival data (Oakes, 1989), the Clayton copula model has been used to model the dependency in other studies. Since we analyze this data set to illustrate the method developed in this thesis, we did not check the adequacy of the Clayton copula model. However, an approach given in Lawless and Yilmaz (2011) can be used for this purpose. We estimated the parameters of the log-logistic models with the two-stage maximum likelihood estimation method as explained in Section 3.3. This approach involved optimizing the log-likelihood function for the parameters in the distribution of error terms and the copula model. The dependency between two gap times is considered in the analysis of the second gap time using the proposed method, but ignored in the naive Peng-Huang method. It should be noted that maximum likelihood estimate of the Kendall's Tau coefficient for this data set is 0.182, indicating a lower level of dependency between the gap times. Table 4.4 provides parameter estimates across various quantiles of the time from recurrence of colon cancer to death, with the proposed method and naive Peng-Huang method. Notably, the table reveals that standard deviation of estimated parameters derived from the proposed method is lower compared to the Peng-Huang method, suggesting enhanced precision and stability with the proposed method.

The results presented in Table 4.4 indicate the significance of Lev+5FU treatment for the quantiles $\tau = 0.3$, 0.5 and 0.7 of the marginal distribution of the second gap time; that is, the time from recurrence of colon cancer to death, in both methods. For example, the proposed method yields $\hat{\beta}_{12,2} = -0.399$ when $\tau = 0.5$. The result can be interpreted as follows. The median survival time for the individuals in the Lev+5FU treatment group is $\exp(-0.399) = 0.677$ times shorter than those in the placebo comparison group, while keeping other covariates the same. Comparing this result with the result given in Table 4.3, we observe that the effect of the Lev+5FU treatment changes its direction. That is, while it was significantly beneficiary in terms of extending the time from the tumor removal to the recurrence of the colon cancer, it significantly decreases the time from cancer recurrence to death. Such result may be interpreted as an adverse effect of the Lev+5FU treatment for the distribution of the second gap time, suggesting that across various patient percentiles, Lev+5FU treatment is associated with a reduction in the time from colon cancer recurrence to death compared to the placebo group (Obs), while controlling for other covariates. Note that the Lev treatment is not significant at all τ values, compared with the placebo group.

Moreover, the results presented in Table 4.4 show that, in addition to treatment Lev+5FU, the age covariate is significant at quantiles $\tau = 0.1$ and 0.3, in both methods. Additionally, at $\tau = 0.5$, the covariate node4 is also identified as significant. Further analysis reveals that at $\tau = 0.7$, covariates obstruct and adhere are also identified as significant factors, alongside Lev+5FU treatment. Additionally, the estimation of the coefficients of QR model for the second gap time is close for both the proposed method and the Peng-Huang method in some cases, as shown in Table 4.4, which is due to the small dependency between the two gap times in this data set.

Table 4.4: Quantile regression estimate (based on the *Proposed* method and the Peng-Huang *PH* method) for the time from the recurrence of colon cancer to the death.

τ	Method		Intercept	sex	age	obstruct	perfor	adhere	nodes	differ	extent	surg	node4	rx-Lev	rx-Lev5FU
			$\hat{\beta}_{02}$	$\hat{\beta}_{12}$	$\hat{\beta}_{22}$	$\hat{\beta}_{32}$	$\hat{\beta}_{42}$	$\hat{\beta}_{52}$	$\hat{\beta}_{62}$	$\hat{\beta}_{72}$	$\hat{\beta}_{82}$	$\hat{\beta}_{92}$	$\hat{\beta}_{10,2}$	$\hat{\beta}_{11,2}$	$\hat{\beta}_{12,2}$
0.1	Proposed	Estimate	8.294	-0.184	-0.029	-0.276	-0.204	0.244	-0.012	-0.178	-0.236	0.223	-0.345	-0.097	-0.164
		SD	0.842	0.210	0.009	0.243	0.679	0.216	0.036	0.199	0.197	0.246	0.364	0.309	0.247
		p-value	< 0.001	0.245	< 0.001*	0.092	0.701	0.211	0.524	0.219	0.188	0.292	0.267	0.687	0.471
	PH	Estimate	8.343	-0.244	-0.004	-0.409	-0.260	0.271	-0.023	-0.245	-0.259	0.259	-0.404	-0.125	-0.178
		SD	1.203	0.271	0.011	0.337	0.728	0.344	0.052	0.279	0.280	0.278	0.414	0.313	0.350
		p-value	< 0.001	0.498	0.008*	0.413	0.779	0.477	0.824	0.523	0.399	0.424	0.405	0.757	0.639
0.3	Proposed	Estimate	8.323	-0.238	-0.022	-0.362	0.478	-0.053	-0.030	-0.242	-0.148	0.073	-0.353	-0.071	-0.505
		SD	0.522	0.138	0.005	0.160	0.293	0.143	0.025	0.145	0.116	0.140	0.208	0.146	0.183
		p-value	< 0.001	0.083	< 0.001 *	0.023	0.100	0.702	0.222	0.091	0.205	0.611	0.092	0.628	0.006 *
	PH	Estimate	8.263	-0.239	-0.022	-0.364	0.482	-0.055	-0.031	-0.245	-0.147	0.071	-0.350	-0.071	-0.503
		SD	0.717	0.155	0.006	0.226	0.265	0.210	0.054	0.184	0.163	0.160	0.314	0.171	0.237
		p-value	< 0.001	0.123	< 0.001*	0.110	0.071	0.802	0.574	0.188	0.364	0.650	0.261	0.678	0.033*
0.5	Proposed	Estimate	8.453	-0.075	-0.010	-0.359	0.312	-0.143	-0.031	-0.198	-0.304	-0.053	-0.332	-0.163	-0.390
		SD	0.653	0.138	0.005	0.154	0.326	0.132	0.020	0.115	0.196	0.134	0.148	0.142	0.158
		p-value	< 0.001	0.680	0.049*	0.032*	0.395	0.279	0.152	0.076	0.179	0.740	0.013*	0.508	0.035*
	PH	Estimate	8.437	-0.075	-0.010	-0.358	0.311	-0.141	-0.031	-0.197	-0.303	-0.052	-0.330	-0.163	-0.389
		SD	0.853	0.182	0.008	0.253	0.389	0.234	0.044	0.221	0.169	0.192	0.302	0.205	0.248
		p-value	< 0.001	0.587	0.065	0.020*	0.339	0.283	0.123	0.087	0.123	0.696	0.026*	0.253	0.013*
0.7	Proposed	Estimate	8.324	-0.063	-0.010	-0.317	0.547	-0.251	-0.043	-0.061	-0.181	-0.080	-0.109	-0.245	-0.461
		SD	0.605	0.111	0.005	0.123	0.347	0.164	0.019	0.115	0.097	0.120	0.168	0.146	0.158
		p-value	< 0.001	0.597	0.052	0.012*	0.121	0.132	0.022*	0.586	0.069	0.515	0.528	0.100	0.004*
	PH	Estimate	8.274	-0.059	-0.010	-0.309	0.538	-0.248	-0.043	-0.063	-0.177	-0.078	-0.106	-0.239	-0.452
		SD	1.314	0.271	0.012	0.353	0.640	0.390	0.052	0.370	0.270	0.310	0.407	0.308	0.357
		p-value	< 0.001	0.817	0.385	0.037*	0.392	0.519	0.041*	0.047	0.503	0.796	0.788	0.427	0.019*

Chapter 5

Summary and Future Work

This chapter includes a summary and conclusion of the thesis in Section 5.1 and future work in Section 5.2.

5.1 Summary and Conclusion

The research conducted in this thesis focuses on the analysis of sequentially observed bivariate survival data, where individuals may experience two same or different types of events sequentially. An application of such a setting is given in the illness-death model. An important issue in the illness-death model is to understand the effects of treatments, interventions and some certain covariates on the marginal distributions of time-to-events associated with the *ill* state as well as the *death* state observed after the *ill* state. Quantile regression (QR) provides a more comprehensive understanding of the effects of covariates on the response variable compared with the classical regression models. The effects of covariates can be measured on various quantiles of the distribution of the response variable for a given set of covariates. Because of this reason, QR has received a considerable attention recently. The applications of QR in the context of classical survival analysis, in which the occurrence time of a single event is of interest, has been well investigated. However, inference about the effects of covariates on the marginal distribution of the second and subsequent gap (survival) times has not been discussed thoroughly, when the gap times are dependent. Statistical methods that cannot address this dependency may lead to biased results for the estimation of covariate effects for the marginal distribution of the second gap time.

In this thesis, we considered an estimation method which can be applied to estimate the effects of covariates on the marginal distribution of sequentially observed two gap times. We address the dependency between two gap times with copulas. Major challenges include issues related to the non-identifiability of marginal survival distributions for the second gap time and the complexities introduced by the issue of induced dependent censoring. If the first survival time is censored, the second survival time becomes unobservable, which may result in non-identifiability issue. As a result, estimating the marginal distribution of the second gap time becomes challenging without information on the first gap time. Additionally, dependency between two gap times causes the second gap time to be subject to induced dependent censoring, meaning a dependent variable censors the second gap time.

Our estimation method is based on the martingale estimating equations. We introduce this method in Section 3.3, first for QR of the first gap time and then for that of the second gap time. In the case of the first gap time, our method is similar to the Peng-Huang method, which is discussed in Section 3.2. However, instead of a grid-based estimation tool, on which the Peng-Huang method is based, we utilize the Newton-Raphson algorithm to solve the system of unbiased estimating equations to obtain the parameter estimations in QR model for the distribution of the first gap time. Our simulation results presented in Section 3.5.1 indicate that the proposed method is competitive with the Peng-Huang method for the simulation scenarios considered in this case.

In the second part of Section 3.3, we extend our discussion to the estimation of the parameters in QR model of the marginal distribution of the second gap time with the proposed method. In this case, martingale estimating equations include parameters related to the first gap time and dependence parameters in the copulas. The estimation of these parameters is carried out with a two-stage procedure proposed by Lawless and Yilmaz (2011). In the first stage, we maximize the log likelihood function for the parameters related to the first gap time using an accelerated failure time (AFT) survival regression model. Then, we plug in these estimates to the log likelihood function of the parameters related to the copula and an AFT model for the second gap time. In particular, we focus on the Clayton copula model, which belongs to one-parameter Archimedean copula family. For the marginal distribution of the gap times, we consider AFT type regression models where the error terms follow the standard extreme value distribution. It should be noted that the other models for copula and error terms can be used in a similar way in the proposed method.

In Section 3.5.2, we apply both the proposed method and the naive Peng-Huang method to estimate QR for the second gap time. Subsequently, we present the results of a Monte Carlo simulation study and compare these two methods. Our findings reveal that parameters estimated using the naive Peng-Huang method are affected by dependency, resulting in high bias and lower precision compared with the estimators obtained with the proposed method.

Additionally, we also investigate the martingale structure related to the second gap time in an empirical setting in Section 3.4.3. Also, we present the results of another Monte Carlo simulation study conducted to discuss the accuracy of the standard normal approximations for the estimators of model parameters using the proposed method. To do this, we utilize the normal quantile-quantile (Q-Q) plots. The results are given in Section 3.5 and show that the standard normal approximations for the estimators of parameters using the proposed method are adequate for sufficiently large sample sizes. Finally, we illustrate the proposed method by analyzing the colon cancer data set in Chapter 4.

5.2 Future Work

In this final section of the thesis, we discuss some future extensions to our work. The proposed method given in Section 3.3 can be extended to address some of the limitations of this study. For example, the estimation of $Cov(\hat{\boldsymbol{\beta}}_2)$ is based on the sandwich type estimator given in (3.76). As discussed in Section 3.3, this estimator ignores the variability caused by the estimation of $\boldsymbol{\beta}_1$ and ϕ in the previous stages and may provide conservative values in settings with small sample sizes. Other than applying a nonparametric bootstrap procedure, an estimator of $Cov(\hat{\boldsymbol{\beta}}_2)$ which takes into account the estimation of $\boldsymbol{\beta}_1$ and ϕ in the previous stages can be developed. This approach requires the extension of the estimator (3.76) to include likelihood equations for the parameters in the regression model of the first gap time and copula model. As a result, the method may become computationally less efficient. We list this extension as a future work and investigate it in more detail by comparing this approach with the one given in Section 3.3 as well as bootstrapping method. Other future works are listed below.

First, throughout the thesis we considered only time-fixed covariates. As a future work, we intend to take into account the case in which covariates change over time. Because time-varying covariates are of interest in some epidemiological and public health research with bivariate sequences of gap times, such an extension of the proposed method to such settings will be investigated as a future work. It should be noted that the settings on which the values of time-varying covariates are fixed over each gap time can be handled in a straightforward way with the proposed method. For example, models $\log T_1 = \beta_{01} + \beta_{11}x_1 + \beta_{21}x_2 + \epsilon_1$ and $\log T_2 = \beta_{02} + \beta_{12}y_1 + \beta_{22}y_2 + \epsilon_2$ can be specified for the first and second gap times, respectively. Then, the proposed estimation method given in Section 3.3 can be directly applied. The settings in which the gap times may vary over each gap time are more complicated. Therefore, it is important to carefully consider the role of covariates in the analysis and to fit models that appropriately account for the relationship between the gap times and the covariates. The use of advanced modeling techniques, such as time-varying covariate models, can help to better capture the relationship between the gap times and the covariates, and will be investigated as a future work.

Second, in this thesis, it is supposed that the first gap time and all covariate values are observable and not missing. It could be interesting to extend the proposed method to deal with QR model of bivariate sequential gap time to include the situations where the first gap times and/or values of covariates are missing. The situation in which the first gap times are missing in classical survival analysis settings has been discussed by Huang and Chen (2017). They considered the analysis of HIV-infected subjects, and defined the first gap time as the time from the initial contraction of HIV to diagnosis of HIV and the second gap time as the time between HIV and AIDS diagnoses. In practice, the initial HIV contracting time is usually not available; hence, the first gap time is missing. We will investigate such setting as a future work.

Third, unidirectional sequential gap times are considered in this thesis. It can be useful to evaluate bidirectional sequential gap times. In the regression model, forward and backward gap times and covariates can be considered to fit the linear or QR models to predict the mean or quantile of the distribution of the next gap time. For example, in medical studies, alternative sequential gap times could be used to analyze the time between hospital readmissions and the covariates that influence them. Therefore, as the third future work, we will investigate the illness-death model with recovery, and extend the proposed method to analyze data from this model.

References

Aalen, O., Borgan, O., & Gjessing, H. (2008). *Survival and Event History Analysis: A Process Point of View*. Springer Science & Business Media.

Andersen, E. W. (2005). Two-stage estimation in copula models used in family studies. *Lifetime Data Analysis*, *11*, 333–350.

Andersen, P. K., Abildstrom, S. Z., & Rosthøj, S. (2002). Competing risks as a multi-state model. *Statistical Methods in Medical Research*, *11*(2), 203–215.

Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer Science & Business Media.

Andersen, P. K., & Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, *11*(2), 91–115.

Barthel, N., Geerdens, C., Czado, C., & Janssen, P. (2019). Dependence modeling for recurrent event times subject to right-censoring with d-vine copulas. *Biometrics*, *75*(2), 439–451.

Bedair, K. F., Hong, Y., & Al-Khalidi, H. R. (2021). Copula-frailty models for recurrent event data based on monte carlo em algorithm. *Journal of Statistical Computation and Simulation*, *91*(17), 3530–3548.

Betensky, R. A., & Finkelstein, D. M. (1999). An extension of kendall's coefficient of concordance to bivariate interval censored data. *Statistics in Medicine*, *18*(22), 3101–3109.

Casella, G., & Berger, R. (2002). *Statistical Inference (2rd ed.)*. Duxbury Press; Pacific Grove, CA.

Chang, S.-H. (2004). Estimating marginal effects in accelerated failure time models for serial sojourn times among repeated events. *Lifetime Data Analysis*, *10*(2), 175–190.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence.

Biometrika, *65*(1), 141–151.

Cook, R. J., & Lawless, J. F. (2007). *The Statistical Analysis of Recurrent Events*. Springer.

Cook, R. J., & Lawless, J. F. (2018). *Multistate Models for the Analysis of Life History Data*. Chapman and Hall/CRC.

Cox, D. R., & Isham, V. (1980). *Point Processes* (Vol. 12). CRC Press.

Daley, D. J., Vere-Jones, D., et al. (2003). *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer.

Diao, L., & Cook, R. J. (2014). Composite likelihood for joint analysis of multiple multistate processes via copulas. *Biostatistics*, *15*(4), 690–705.

Fitzgerald, N., CHE, L. S. P., & Holmes, E. (2022). Projected estimates of cancer in canada in 2022. *Canadian Medical Association. Journal*, *194*(17), E601–E607.

Frees, E. W., & Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, *2*(1), 1–25.

Fu, T.-C., Su, D.-H., & Chang, S.-H. (2016). Serial association analyses of recurrent gap time data via kendall's tau. *Biostatistics*, *17*(1), 188–202.

Genest, C., Rémillard, B., & Beaudoin, D. (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, *44*(2), 199–213.

Genest, C., & Rivest, L.-P. (1993). Statistical inference procedures for bivariate archimedean copulas. *Journal of the American Statistical Association*, *88*(423), 1034–1043.

Ghasemzadeh, S., Ganjali, M., & Baghfalaki, T. (2022). Quantile regression via the em algorithm for joint modeling of mixed discrete and continuous data based on gaussian copula. *Statistical Methods and Applications*, *31*(5), 1181–1202.

Giovannucci, E. (2002). Modifiable risk factors for colon cancer. *Gastroenterology Clinics*, *31*(4), 925–943.

Goethals, K., Janssen, P., & Duchateau, L. (2008). Frailty models and copulas: similarities and differences. *Journal of Applied Statistics*, *35*(9), 1071–1079.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data* (Vol. 564). Springer.

Hsieh, J.-J., Ding, A. A., Wang, W., & Chi, Y.-L. (2013). Quantile regression based on semi-competing risks data.

Huang, C.-H. (2019). Mixture regression models for the gap time distributions and illness–death processes. *Lifetime Data Analysis*, *25*(1), 168–188.

Huang, C.-H., & Chen, Y.-H. (2017). Regression analysis for bivariate gap time with missing first gap time data. *Lifetime Data Analysis*, *23*(1), 83–101.

Huang, C.-Y., Wang, C., & Wang, M.-C. (2016). Nonparametric analysis of bivariate gap time with competing risks. *Biometrics*, *72*(3), 780–790.

Huang, C.-Y., & Wang, M.-C. (2005). Nonparametric estimation of the bivariate recurrence time distribution. *Biometrics*, *61*(2), 392–402.

Huang, Y., & Louis, T. A. (1998). Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika*, *85*(4), 785–798.

Jin, Z., Ying, Z., & Wei, L. (2001). A simple resampling method by perturbing the minimand. *Biometrika*, *88*(2), 381–390.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts.* CRC press.

Kessing, L., Hansen, M., Andersen, P., & Angst, J. (2004). The predictive effect of episodes on the risk of recurrence in depressive and bipolar disorders–a life-long perspective. *Acta Psychiatrica Scandinavica*, *109*(5), 339–344.

Koenker, R. (2003). *Quantile Regression.* Cambridge University Press.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 33–50.

Kurowicka, D., & Cooke, R. M. (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling.* John Wiley & Sons.

Lakhal-Chaieb, L., Cook, R. J., & Lin, X. (2010). Inverse probability of censoring weighted estimates of kendall's τ for gap time analyses. *Biometrics*, *66*(4), 1145–1152.

Lange, K., Chambers, J., & Eddy, W. (2010). *Numerical Analysis for Statisticians* (Vol. 1). Springer.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data* (2nd ed.). Wiley, New York.

Lawless, J. F., & Yilmaz, Y. E. (2011). Semiparametric estimation in copula models for bivariate sequential survival times. *Biometrical Journal*, *53*(5), 779–796.

Lee, C. H., Huang, C.-Y., Xu, G., & Luo, X. (2018). Semiparametric regression analysis for alternating recurrent event data. *Statistics in Medicine*, *37*(6), 996–1008.

Li, R., Cheng, Y., Chen, Q., & Fine, J. (2017). Quantile association for bivariate survival data. *Biometrics*, *73*(2), 506–516.

Lin, D., Sun, W., & Ying, Z. (1999). Nonparametric estimation of the gap time distribution for serial events with censored data. *Biometrika*, *86*(1), 59–70.

Lin, D. Y., Wei, L.-J., Yang, I., & Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(4), 711–730.

Luo, X., & Huang, C.-Y. (2011). Analysis of recurrent gap time data using the weighted risk-set method and the modified within-cluster resampling method. *Statistics in Medicine*, *30*(4), 301–311.

Luo, X., Huang, C.-Y., & Wang, L. (2013). Quantile regression for recurrent gap time data. *Biometrics*, *69*(2), 375–385.

Ma, Z., Krings, A. W., & Hiromoto, R. E. (2008). Multivariate survival analysis (ii): An overview of multi-state models in biomedicine and engineering reliability. In *2008 international conference on biomedical engineering and informatics* (Vol. 1, pp. 536–541).

Meyer, R., & Romeo, J. S. (2015). Bayesian semiparametric analysis of recurrent failure time data using copulas. *Biometrical Journal*, *57*(6), 982–1001.

Moertel, C. G., Fleming, T. R., Macdonald, J. S., Haller, D. G., Laurie, J. A., Goodman, P. J., … others (1990). Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *New England Journal of Medicine*, *322*(6), 352–358.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to Linear Regression Analysis*. John Wiley & Sons.

Nelsen, R. B. (2007). *An Introduction to Copulas*. Springer Science & Business Media.

Oakes, D. (1982). A concordance test for independence in the presence of censoring. *Biometrics*, 451–455.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, *84*(406), 487–493.

Olesen, A. V., & Mortensen, P. B. (2002). Readmission risk in schizophrenia: selection explains previous findings of a progressive course of disorder. *Psychological Medicine*, *32*(7), 1301–1307.

Peng, L., & Fine, J. P. (2009). Competing risks quantile regression. *Journal of the American Statistical Association*, *104*(488), 1440–1453.

Peng, L., & Huang, Y. (2008). Survival analysis with quantile regression models. *Journal of the American Statistical Association*, *103*(482), 637–649.

Putter, H., Fiocco, M., & Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, *26*(11), 2389–2430.

Rigdon, S. E., & Basu, A. P. (2000). *Statistical Methods for the Reliability of Repairable Systems*. New York, NY: Wiley.

Rotolo, F., Legrand, C., & Van Keilegom, I. (2013). A simulation procedure based on copulas to generate clustered multi-state survival data. *Computer Methods and Programs in Biomedicine*, *109*(3), 305–312.

Schaubel, D. E., & Cai, J. (2004). Regression methods for gap time hazard functions of sequentially ordered multivariate failure time data. *Biometrika*, *91*(2), 291–303.

Schober, P., & Vetter, T. R. (2018). Repeated measures designs and analysis of longitudinal data: If at first you do not succeed—try, try again. *Anesthesia and Analgesia*, *127*(2), 569.

Shih, J. H., & Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 1384–1399.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges.

Soutinho, G., & Meira-Machado, L. (2022). Nonparametric estimation of the distribution of gap times for recurrent events. *Statistical Methods and Applications*, 1–26.

Sun, X., Peng, L., Huang, Y., & Lai, H. J. (2016). Generalizing quantile regression for counting processes with applications to recurrent events. *Journal of the American Statistical Association*, *111*(513), 145–156.

Veeramachaneni, S., Mudunuru, V. R., et al. (2023). Survival analysis of colon cancer data using quantile regression. *Research Journal of Pharmacy and Technology*, *16*(3), 1401–1408.

Visser, M. (1996). Nonparametric estimation of the bivariate survival function with an application to vertically transmitted aids. *Biometrika*, *83*(3), 507–518.

Wang, H. J., Feng, X., & Dong, C. (2019). Copula-based quantile regression for longitudinal data. *Statistica Sinica*, *29*(1), 245–264.

Wang, K., & Shan, W. (2021). Copula and composite quantile regression-based estimating equations for longitudinal data. *Annals of the Institute of Statistical Mathematics*, *73*(3), 441–455.

Wang, M.-C., & Chang, S.-H. (1999). Nonparametric estimation of a recurrent survival function. *Journal of the American Statistical Association*, *94*(445), 146–153.

Wang, W., & Wells, M. T. (1998). Nonparametric estimation of successive duration times under dependent censoring. *Biometrika*, *85*(3), 561–572.

Wei, B. (2022). Quantile regression for censored data in haematopoietic cell transplant research. *Bone Marrow Transplantation*, *57*(6), 853–856.

Yang, M., Luo, S., & DeSantis, S. (2019). Bayesian quantile regression joint models: inference and dynamic predictions. *Statistical Methods in Medical Research*, *28*(8), 2524–2537.

Yip, P., & Lam, K. (1997). Anonparametric inference procedure for an illness-death model. *Stochastic Analysis and Applications*, *15*(1), 125–135.

Zeng, D., & Lin, D. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *69*(4), 507–564.