Multiple Imputation in Censored Quantile Regression

by

©**Ummay Nayeema Islam**

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Masters in Statistics

Department of Mathematics and Statistics
Memorial University

August 2024

St John's , Newfoundland and Labrador, Canada

Abstract

Quantile regression is a natural extension to the traditional linear regression. Instead of modeling the conditional mean of the response variable, quantile regression models the conditional quantiles of the response. With properly selected quantiles, the quantile regression model provides a better understanding of the relationship between the response and the covariates comparing with the traditional regression models Recently this method is introduced to the area of survival analysis, where censoring is a natural characteristic of the data. To address the challenges posed by censored data, especially the specification issue at high quantiles, we propose a novel approach that employs multiple imputations of censored observations using the Buckley-James method, originally developed in the framework of classical quantile regression analysis. Our method not only ensures consistent estimators of the model parameters, but also achieves asymptotically normality when the sample size approaches infinity. Notably, it overcomes the limitations of traditional censored quantile regression, particularly in estimating extreme quantiles. Extensive simulation studies demonstrate the efficacy of our approach. Additionally, we apply our method to a Health Maintenance Organization (HMO) dataset as an illustration.

Acknowledgment

With the grace of Almighty Allah, I am happy to express my sincerest appreciation to my supervisor, Dr. Zhaozhi Fan for his mentorship throughout the entire journey of my MSc thesis. This thesis would not have been done without the extensive support from my supervisor, both professionally and personally. I would like to take this opportunity to extend my gratitude to my professors, specially Dr. Alwell Oyet, Dr. Wasimul Bari, and Dr. Jafar Ahmed Khan. They made me interested in statistics through their outstanding teaching and research. My sincere thanks goes also to all my friends and colleagues in Bangladesh and here in Canada. I am specially grateful to my beloved parents, my family members for their support and encouragement during my entire study period. Finally, I owe thanks to a very special person, my husband, Md. Abu Sufian for his continued and unfailing love, support and understanding during my pursuit of my future career. With due respect, I would like to dedicate this thesis to my beloved family members.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Survival analysis comprises a collection of statistical theories and techniques for analyzing data where the main response variable is the time until an event of interest occurs. (Lawless, 2011). Time variable is also known as survival time, failure time or lifetime. Typical event of interest include death, disease, relapse from remission, recovery etc. The life time or survival time may be measured in days, months, years or any designated experience of interest that may happen to an individual.

The main aspect which differentiates survival analysis from other areas in statistics is that survival data are usually subject to censoring. Observations are called censored when information about their survival times is incomplete. A censored observation is defined as an observation with incomplete information about the **"time-to-event"**. Censoring can be classified into three main categories, of which are right censoring, left censoring and interval censoring. The most usually seen form of censoring is right censoring. The lifetime of an individual is said to be right censored, when his/her lifetime becomes incomplete at the right side of the starting point of follow-up period. In left censoring, the event of interest occurs before a particular time point but the exact time point of occurring the event of interest is unknown (Kalbfleisch & Prentice, 2011a). In case of interval censoring, individual's event time is known to fall between two specific time points.

Let T be a non-negative random variable that represents the failure time of an individual from a homogeneous population. The probability distribution of T can be described in several useful ways in survival analysis. Three particularly useful functions are the survivor function, the probability density function and the hazard function, denoted by $S(t)$, $f(t)$ and $\lambda(t)$, respectively. In the context of survival analysis, for a continuous domain $[0, \infty)$ with probability density function $f(t)$ and distribution function $F(t)$, the survivor function, $S(t)$ is defined as the probability that an event does not occur at or before time t. Mathematically, the survivor function $S(t)$ can be expressed as,

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(u)du, \quad t > 0$$

The hazard function is defined as the instantaneous rate at which failures occur for subjects that are surviving at time t (Lawless, 2011). The hazard function, $\lambda(t)$ is defined as follows.

$$\begin{aligned}
\lambda(t) &= \lim_{\Delta t \to 0} \frac{P[t < T \leq t + \Delta t | T \geq t]}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{P[t < T \leq t + \Delta t, T \geq t]}{\Delta t} \frac{1}{P(T \geq t)} \\
&= \frac{f_T(t)}{S(t)} \\
&= -\frac{d}{dt} log S(t).
\end{aligned}$$

Many parametric, semiparametric and nonparametric approaches have been proposed for estimating survival and hazard functions. In continuous-time framework, parametric models assume continuous parametric distributions, such as exponential, Weibull, log normal, log logistic distributions. Survival analysis can also be performed without considering any distributional assumption. Kaplan-Meier estimator (Kaplan & Meier, 1958) also known as product limit estimator is the most common nonparametric technique for estimating the survival function $S(t)$. Let n_i be the number of individuals at risk at t_i and d_i be the number of events at t_i. The Kaplan-Meier estimator has

been defined as,

$$\hat{S}(t) = \prod_{i|t_i \leq t} \frac{n_i - d_i}{n_i}.$$

Another important concept in survival analysis is the cumulative hazard function $\Lambda(t)$, which is given by $\Lambda(t) = \int_0^t \lambda(u)du$, $t > 0$. The cumulative hazard function $\Lambda(t)$ is most naturally estimated using the Nelson-Aalen estimator (Nelson, 1972),

$$\hat{\Lambda}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i} = \sum_{t_i \leq t} \hat{\lambda}_i.$$

This estimator is a right-continuous step function, where the magnitude of increments is calculated by the empirical hazard estimates. Often, its a matter of interest to determine whether two or more samples originate from the same survivor function. One commonly used method for this comparison is the log-rank test (Lawless, 2011). The log-rank test is a non-parametric test that compares the survival distributions of two or more groups. It is especially useful in clinical trials and medical research to compare the efficacy of different treatments.

One of the primary goals of survival analysis is to examine relationship between failure time and explanatory variables under censoring. Since survival time takes non-negative values, the classical linear regression models are not appropriate for modeling survival time unless the restriction is removed by transforming these times in such a way which takes all possible values from the real line. Two most popular survival regression models are proportional hazards (PH) model and accelerated failure time (AFT) model (Klein, Moeschberger, et al., 2003). The proportional hazards model (Cox, 1972) is extensively utilized in survival analysis to assess the impact of explanatory variables on survival time by modeling the hazard function. Let $h(t|X)$ be the hazard function at time t conditional on the vector of covariates X. The PH regression model assumes the form

$$h(t|X) = h_0(t)e^{X'\beta},$$

where $h_0(t)$ is an arbitrary baseline hazard rate and $\beta = (\beta_1, \beta_2, \cdots, \beta_p)'$ is a $(p \times 1)$ parameter vector. The PH model is a semi-parametric model because the baseline

hazard rate is treated non-parametrically, while the multiplicative part $e^{X'\beta}$ is parametric.

The natural logarithm of survival time is modeled under the AFT survival regression model, which is widely used due to its ability to provide explicit interpretations of the regression parameters. Let T_i $(i = 1, 2, \ldots, n)$ be the logarithm of failure time of the i^{th} subject and X_i be the $(p \times 1)$ vector of covariates. The AFT model with the $(p \times 1)$ vector of regression parameters β is,

$$T_i = X_i'\beta + \varepsilon_i,$$

where the ε_i are independent and identically distributed (iid) random error variables from a distribution function, F, such as normal distribution, extreme value distribution or log logistic distribution. The AFT model is called semi-parametric when the distribution of errors is not specified.

Inference procedures for the AFT model have been extensively developed by various researchers, including Prentice (1978), Buckley and James (1979), Tsiatis (1990), Ritov (1990), and Wei, Ying, and Lin (1990), among others. These methodologies are derived without specifying the distribution of F. However, they require the independent error terms to be homogeneous. For further details, we refer to the works of Cox and Oakes (2018), Kalbfleisch and Prentice (2011b), and Klein and Moeschberger (2003). Both the AFT model and the Cox PH model assume that covariates influence the location of the distribution of transformed survival times, not the shape. This assumption can restrict the nature of the impact that covariates can have on the survival times. In practice, there are numerous situations where the focus is on the tail of the survival distribution. The paper entitled "Reappraising Medfly Longevity: A Quantile Regression Survival Analysis" by Koenker and Geling (2001) monitored that mortality rates of Mediterranean fruit flies declined at the oldest observed ages. This medfly experiments contradict the traditional assumption of Gompertz form in survival distribution that hazard is log-linear (Koenker & Geling, 2001). Censored quantile regression (CQR) offers a flexible and robust semi-parametric approach which

provides a comprehensive view by allowing the analysis of the effects of covariates across the entire distribution of survival times, rather than at a single point.

Another approach to handling censored data is to treat them as a type of missing observation and apply missing data mechanisms to them. There is an extensive body of literature on methodologies for handling missing data outside the realm of survival analysis. It is only in recent years that these techniques have begun to be implemented to address missing event time information in censored observations in the context of survival analysis. Multiple imputation (MI) (Rubin, 1978) is a widely adopted method to analyze data subject to various missing data mechanisms. The theoretical underpinnings of multiple imputation are rooted in Bayesian statistics. A study conducted by Moghaddam et al. (2022) proposed a Bayesian imputation of censored survival data, demonstrating an improved visualization and analysis of survival data by treating censored data as incomplete observations and imputing them for a more comprehensive analysis. Another research considered nonparametric multiple imputation scheme in handling missing failure times in censored data (Taylor, Murray, & Hsu, 2002).

In this thesis, we propose a non-Bayesian or model-based multiple imputation scheme for analyzing survival data and construct quantile regression with the multiply imputed data sets. The purpose of this thesis is to provide some theoretical basis and foundation for the use of quantile regression to multiply imputed data sets in survival analysis. The scenarios we examine are relatively straightforward; nonetheless, our results offer positive evidence. Specifically, for higher quantiles, CQR often fails to yield results, whereas quantile regression using multiply imputed data successfully provides outcomes.

The rest of this thesis proceeds as follows. In Chapter 2 we give an introduction to quantile regression and CQR. Chapter 3 develops the proposed model-based multiple imputation method in CQR. This chapter also includes consistency and asymptotic properties of parameter estimators. In Chapter 4, we present the results of extensive simulation studies and apply the method to Health Maintenance Organization

(HMO) data to illustrate the performance of the proposed method. Finally, Chapter 5 portrays the conclusion of this thesis.

Chapter 2

Introduction to Quantile Regression

The classical linear regression model (CLRM) is focused on the conditional mean; that is, the CLRM summarizes the relationship between the response and the predictors by using conditional mean of the target for each fixed value of the predictors. Undoubtedly, CLRMs have some appealing properties. Under ideal circumstances, they provide a detailed picture of the relationship between the response and the covariates. A natural extension of linear regression model is quantile regression (QR). It is commonly used when the assumptions of the CLRM fail to meet, in particular, linearity, homoscedasticity, independence, or normality). The QR approach was introduced by Koenker and Bassett (1978) in their seminal paper titled "Regression Quantiles", which extends the traditional linear regression model to the conditional quantiles of the response variable, in contrast to focusing solely on its conditional mean. QR does not concentrate only on the conditional mean, instead, it can be used to estimate the entire conditional distribution of a response variable for a given set of covariates. Moreover, it offers valuable information hidden in tails. An interesting example of QR is self-thinning of tropical plants in Chihuahuan desert of the south western US (Cade & Guo, 2000), where the effects of increasing germination densities of seedlings

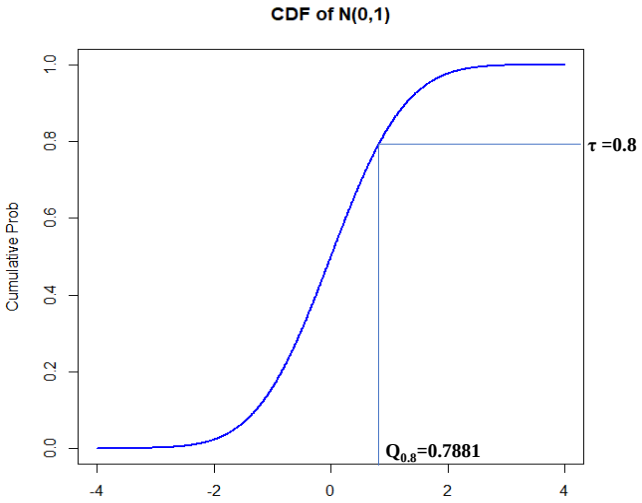

Figure 2.1: A CDF for the standard normal distribution

on the decline in densities of final mature plants were best captured at the higher plant densities associated with upper quantiles. Addressing issues like self-thinning problem where extremes are important, QR plays a significant role in such situations. These attractive features of the QR approach has led to its growing popularity in various fields, such as econometrics (Machado & Mata, 2005), biostatistics (Y. Wei, Pere, Koenker, & He, 2006), micro-array data analysis (Huang et al., 2008) and more.

2.1 Quantile and Quantile Function

Let a real-valued random variable X follow the standard normal distribution, characterized by its (right-continuous) cumulative distribution function (CDF) $F(x)$, as shown in Figure 2.1. Whereas for any τ, where $0 < \tau < 1$, the τ^{th} quantile of X is defined as its inverse; that is,

$$Q(\tau) = F^{-1}(\tau) = inf\{x : F(x) \geq \tau\}. \tag{2.1}$$

The inverse function $F^{-1}(\tau)$ is the minimum value of x for which $F(x) = \tau$. The plot in Figure in 2.1 shows that $F(0.7881) = 0.8$ and $F^{-1}(0.8) = Q(0.8) = 0.7881$. It implies that the τ^{th} quantile of the distribution is the value of the inverse of CDF at

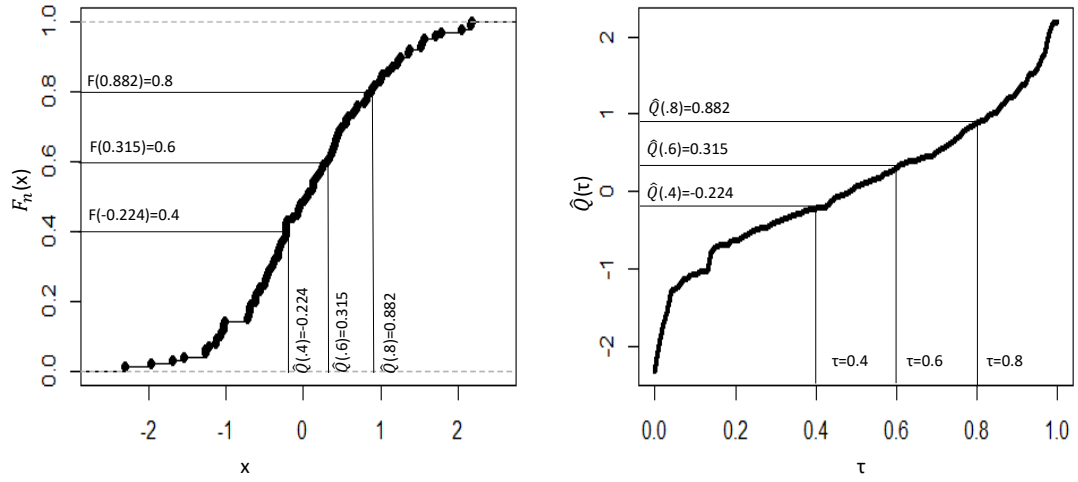

Figure 2.2: Empirical CDF and QF of standard normal distribution

τ.

For a sample size n, the empirical distribution function of X, which is the distribution function associated with the empirical measure of a sample, can be defined as

$$F_n(x) = n^{-1} \sum_{i=1}^{n} I(X_i < x). \tag{2.2}$$

Thus, for a given n, the τ^{th} sample quantile can be represented by,

$$\hat{Q}(\tau) = F_n^{-1}(\tau) = inf\{x : F(x) \geq \tau\}. \tag{2.3}$$

For illustration purposes, we independently generated $n = 100$ observation from the standard normal distribution and plotted the empirical quantile function in Figure 2.2. It is observed that both sample CDF and the sample quantile function are monotonic non-decreasing functions.

It is widely known that the mean of a random variable X, defined as the center μ, can be obtained by the following minimization problem

$$\mu = \arg\min_{\alpha} E(X - \alpha)^2, \tag{2.4}$$

that is the minimization of the squared sum of deviations. The median, instead, minimizes the mean absolute deviations about any point, say α. Therefore, the expression becomes,

$$Me = \arg\min_{\alpha} E|X - \alpha|. \tag{2.5}$$

Presenting quantiles as specific centers of the distribution of X, the τ^{th} quantile can be obtained by minimizing the following absolute sum of deviations (Hao & Naiman, 2007), that is,

$$Q(\tau) = \arg\min_{\alpha} E[\rho_\tau(X - \alpha)], \tag{2.6}$$

where $\rho_\tau(\cdot)$, called the check loss function, satisfies

$$\begin{aligned} \rho_\tau(u) &= (\tau - I(u < 0))u, \\ &= [\tau I(u > 0) + (1 - \tau)I(u < 0)]|u|, \end{aligned} \tag{2.7}$$

for some $\tau \in (0, 1)$. In 2.7, positive and negative deviations are weighted by τ and $(\tau - 1)$, correspondingly. The loss function is illustrated in Figure 2.3. If $\tau = 0.5$, the check loss function $\rho_\tau(u)$ becomes the absolute value of u. That is,

$$\begin{aligned} \rho_{0.5}(u) &= (0.5 - I(u < 0))u, \\ &= 0.5 \begin{cases} u \quad if \quad u > 0, \\ -u \quad if \quad u < 0, \end{cases} \\ &= 0.5|u|. \end{aligned}$$

Under the loss function $\rho_\tau(\cdot)$, we try to minimize

$$\begin{aligned} E[\rho_\tau(X - \alpha)] &= \int_{x \in \mathbb{R}} \rho_\tau(t - \alpha) dF(t), \\ &= (\tau - 1) \int_{-\infty}^{\alpha} (t - \alpha) dF(t) + \tau \int_{\alpha}^{\infty} (t - \alpha) dF(t). \end{aligned} \tag{2.8}$$

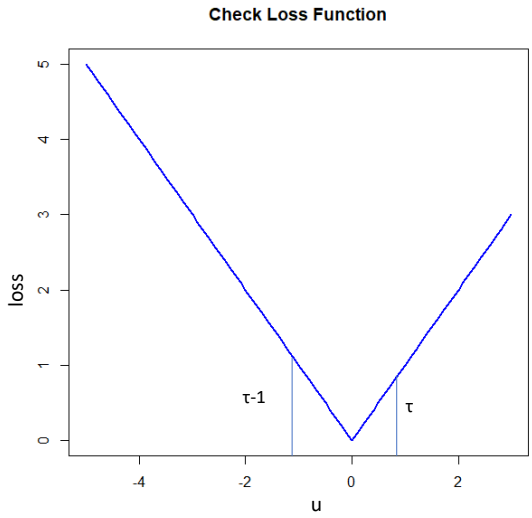

Figure 2.3: The check loss function $\rho_\tau(u)$ for a certain τ

Differentiating the expected loss in Equation 2.8 with respect to α and setting the partial derivative equal to zero lead to solution of the minimization problem.

$$\begin{aligned}
\frac{\partial}{\partial\alpha}E[\rho_\tau(X-\alpha)] &= \frac{\partial}{\partial\alpha}(\tau-1)\int_{-\infty}^{\alpha}(t-\alpha)dF(t) + \frac{\partial}{\partial\alpha}\tau\int_{\alpha}^{\infty}(t-\alpha)dF(t), \\
&= (1-\tau)\int_{-\infty}^{\alpha}dF(t) - \tau\int_{\alpha}^{\infty}dF(t), \\
&= \int_{-\infty}^{\alpha}dF(t) - \tau[\int_{-\infty}^{\alpha}dF(t) + \int_{\alpha}^{\infty}dF(t)], \\
&= F(\alpha) - \tau\int_{-\infty}^{\infty}dF(t), \\
&= F(\alpha) - \tau, \\
&= 0.
\end{aligned}$$

Because of the monotonicity, minimization can occur at any element of the set $\{x : F(x) = \tau\}$. If the solution is distinct, then $\alpha = F^{-1}(\tau)$. Otherwise, the smallest element will be chosen from a set of τ quantiles. It is noticed that choosing $\tau = 0.5$ yields the median quantile.

There is a close relation between sample quantiles and order statistics. Given a

random sample of size n, $x_1, x_2, ..., x_n$, the data values can be arranged in ascending order of magnitude. Let $x_{(1)}, x_{(2)}, ..., x_{(i)}, ..., x_{(n)}$ denote the ordered observations, where $x_{(1)} \leq x_{(2)}, ..., \leq x_{(i)}, ..., \leq x_{(n)}$. For $i = 1, 2, ..., n$ the notation $x_{(i)}$ is referred to as the i^{th} order statistic of the sample. In terms of sample quantile, the $(k/n)^{th}$ sample quantile corresponds to the k^{th} order statistic.

Let us consider a large sample, $x_1, x_2, ..., x_n$ drawn from a population with probability density function $f(x)$ and quantile function $Q(\cdot)$, the distribution of sample quantile follows approximately normal with mean $Q(\tau)$ and finite variance $\dfrac{\tau(1-\tau)}{n} \cdot \dfrac{1}{f(Q_\tau)^2}$ (Walker, 1968).

2.2 Quantile Regression Model

In order to set up a quantile regression model (QRM), it is reasonable to start with classical linear regression model and make a comparison between them.

Given a $(p \times 1)$ vector of explanatory vectors, $x_i^T = (x_{i1}, ..., x_{ip})$, the CLRM takes the form,

$$Y_i = X_i^T \beta + \varepsilon_i, \qquad i = 1, 2, ..., n, \tag{2.9}$$

where, $\beta^T = (\beta_1, ..., \beta_p)$ is a p-dimensional vector of unknown parameters and $\varepsilon_i, \quad i = 1, 2, ..., n$ is an error term defined as follows.

$$\varepsilon_i \overset{iid}{\sim} F_\varepsilon \quad \text{with} \;\; E(\varepsilon_i) = 0 \;\; \text{and} \;\; Var(\varepsilon_i) = \sigma_\varepsilon^2.$$

The CLRM focuses on the expectation of the target conditional distribution, that is, $E(Y|X = x) = X\beta$. The Model 2.9 can be generalized to the matrix form as follows. Let

$$Y = X\beta + \varepsilon,$$

be the CLRM, where $Y^T = (y_1, y_2, ..., y_n)$ and $\varepsilon^T = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$ be the $(n \times 1)$ random vectors of response variables and corresponding residuals, respectively. The $(n \times p)$ matrix X is called the design matrix where x_i^T is the covariate vector in the

i^{th} row; that is,

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Different that from CLRM, the QRM gives a global perspective on the relationships between variables by enabling the analysis of the conditional distribution of Y on X at various quantiles so that we can examine the effects of covariates on the response variable at a certain quantile rather than focusing only on the mean effect. The conditional quantile of a random variable Y can be defined as,

$$Q_Y(\tau|X = \mathbf{x}) = inf\{y : P(Y \leq y|X = x) \geq \tau\},\ 0 < \tau < 1.$$

Analogous to the CLRM, the linear quantile regression at τ^{th} $(0 < \tau < 1)$ quantile can be modeled as

$$Y = X\beta(\tau) + \varepsilon(\tau), \tag{2.10}$$

where, $\varepsilon^T(\tau) = (\varepsilon_1(\tau), ..., \varepsilon_n(\tau))$ is a $n \times 1$ vector of quantile errors. Under the assumption that $Q_\varepsilon(\tau|X) = 0$, Model 2.10 becomes,

$$Q_Y(\tau|X = x) = X\beta(\tau),$$

where $\beta^T(\tau) = (\beta_1(\tau), ..., \beta_p(\tau))$ is a $(p \times 1)$ column vector of unknown quantile coefficients at τ^{th} quantile and may change with different values of τ.

2.3 Estimation of Parameters in Quantile Regression

In CLRM, estimation of regression coefficient β is evaluated by the minimization of the function,

$$\arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^{'}\beta)^2.$$

In QRM, regression coefficient at τ^{th} quantile, denoted by $\beta(\tau)$, can be estimated by minimizing

$$\arg\min_{\beta(\tau)} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^{'}\beta(\tau)),$$

where ρ_τ is defined in 2.7. Thus, the objective function, $R(\beta(\tau)) = \sum_{i=1}^{n} \rho_\tau(y_i - Q_Y(\tau|X = \mathbf{x}))$ is a weighted sum of absolute deviations. As shown by Koenker and Bassett (1978), the following can be optimized to yield the parameter estimation of the QRM:

$$\hat{\beta}(\tau) = \arg\min_{\beta(\tau)} l(\tau) = \arg\min_{\beta(\tau)} \left(\sum_{y_i \geq x_i^{'}\beta} \tau(y_i - x_i^{'}\beta(\tau)) - \sum_{y_i < x_i^{'}\beta} (1-\tau)(y_i - x_i^{'}\beta(\tau)) \right) \tag{2.11}$$

When τ in 2.11 is 0.5,

$$\hat{\beta}(\tau) = \arg\min_{\beta(\tau)} \sum_{i=1}^{n} |y_i - x_i^{'}\beta(\tau)|, \tag{2.12}$$

which is known as the median regression (Koenker & Bassett Jr, 1978).

2.4 Quantile Based Approach for Censored Data

Censored quantile regression (CQR) is a useful addition to survival analysis, which has been increasingly popular in econometric analysis, industrial life testing, and the health sciences. CQR is not only robust to outliers but also to the misspecification of error distribution, called the heteroscedasticity, as well as to scale transformations of the response variables.

To assess the relationship between a survival outcome and a set of explanatory variables (or covariates), the accelerated failure time (AFT) model serves as a substitute for linear regression in survival analysis. Let, T_i be the log of the survival time. For a p-dimensional vector of covariates X, the AFT model is represented by the following equation.

$$T_i = X_i^{'}\beta + \epsilon_i, \qquad i = 1, 2, ..., n,$$

where β is an unknown $(p \times 1)$ vector regression coefficient and the ϵ_i $(i = 1, ..., n)$ denote the error terms which are iid with the distribution function F_ε.

To introduce CQR, Powell (1984, 1986) proposed least absolute deviations (LAD) for the estimation of regression parameters β. In this approach, Powell considered censored Tobit model with left censoring at zero. Let, $C_i, i = 1, 2, ..., n$, be the natural logarithm of the left censored values which are observed and therefore fixed. For the observed survival time, $Y_i = \max(T_i, C_i)$, the CQR is defined by

$$Q_{T_i}(\tau|X_i) = X_i^{'}\beta(\tau) + F^{-1}(\tau), \qquad 0 < \tau < 1, \tag{2.13}$$

and the quantile regression coefficients at τ can be estimated as,

$$\hat{\beta}(\tau) = \arg\min_{\beta(\tau)} \sum_{i=1}^{n} \rho_\tau(y_i - \max(C_i, x_i^{'}\beta(\tau))). \tag{2.14}$$

Here, $\rho_\tau(\cdot)$ is the check loss function defined in Equation 2.7. In Powell's approach, the censored least absolute deviation (CLAD), denoted by $\hat{\beta}_n$, is defined as,

$$\hat{\beta}_n = \arg\min n^{-1} \sum_{i=1}^{n} |y_i - max(0, x_i^{'}\beta)|$$

Honore, Khan, and Powell (2002) proposed a method for extending the CQR estimator under fixed censoring to models with random censoring using the Kaplan-Meier estimator. Later, this methodology was applied to Powell (1984, 1986) estimators.

In this thesis, we consider only the right censoring. The CQR with right censored data can be similarly modeled by replacing $\max(C_i, X_i^{'}\beta)$ with $\min(C_i, X_i^{'}\beta)$ in Equation (2.14). Let T_i and C_i be the corresponding log of failure time and log of censoring time for the i^{th} individual, $i = 1, 2, ..., n$, the random variable Y_i= $\min(T_i, C_i)$,is the observed time. Let δ_i be the censoring indicator; that is,

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

Thus, the CQR model for the $\tau^{th}(0 < \tau < 1)$ quantile is given by

$$Q_{T_i}(\tau|X_i) = X_i^{'}\beta(\tau), \qquad 0 < \tau < 1. \tag{2.15}$$

Then, for a given value of τ, an estimate of the CQR parameter, $\beta(\tau)$ can be obtained as,

$$\hat{\beta}(\tau) = \arg\min_{\beta(\tau)} \sum_{i=1}^{n} \rho_\tau(y_i - \min(C_i, x_i'\beta(\tau))). \tag{2.16}$$

Several methods have been proposed on CQR to estimate the model parameters. Portnoy (2003) introduced a recursively reweighted estimator, which is a direct generalization of the Kaplan–Meier estimator, to estimate regression quantile parameters and established $\sqrt{n}$ convergences of the proposed estimators. Peng and Huang (2008) used counting processes and martingale theory to develop a CQR model based on the Nelson-Aalen estimator. These three methods to estimate CQR were described by Koenker (2008). The "quantreg" package in the statistical software R implements these methods with options of choosing one of these three methods for the inference. A simple three-step estimation procedure was suggested by Chernozhukov and Hong (2002) for CQR models. In this approach, the authors illustrated this procedure with an extramarital affair example. Wang and Wang (2009) experimented a locally weighted CQR approach, which relaxes the stringent assumptions of existing literature, such as, unconditional independence of survival time and censored time, global linearity etc. Further, the authors studied the proposed method via simulations and illustrated it with an acute myocardial infarction dataset. For survival data with random censoring, Yin, Zeng and Li (2014) proposed a varying-coefficient quantile regression model. Chernozhukov, Fernandez-Val, and Kowalski (2015) created a censored quantile instrumental variable (CQIV) estimator that integrates both the model and endogenous variables. Wu and Yin (2013) proposed a multiple imputation method to cure rate quantile regression for censored data with a survival fraction.

To handle higher dimensional variables, a semiparametric copula-based estimator for conditional quantiles was investigated for both complete or right-censored data (De Backer, El Ghouch, & Van Keilegom, 2017). For the estimation of linear quantile regression with right censored responses, Backer et al (2019) avoided classical

approaches on "check loss" function instead, they investigated a novel approach to estimate quantile coefficients by minimizing an alternative measure of distance.

Chapter 3

Non-Bayesian Multiple Imputation in Censored Quantile Regression

3.1 Introduction

Allan and Wishart (1930) first developed a statistical method to replace a missing value. However, their work became well-known following the publication of Little and Rubin's book in (1978), which introduced the term "imputation" to the world of statistical literature. The earlier imputation methods impute a unique value for each missing observation. This single imputation approach disregards the sampling variability thus provides underestimated standard errors and wrong confidence intervals. Multiple imputation (MI) was first proposed by Donald B. Rubin which creates multiple (say, m) values for each missing value and reflects the uncertainty about prediction of unknown missing values.

Censored survival data may be treated as a kind of missing data. Under the circumstance of censored observations, the term "multiple imputation" refers to the process of replacing a censored datum with several imputed values. Wei and Tanner (1991) proposed two different data augmentation algorithms of multiple imputation to the analysis of censored regression data. Taylor, Murray and Hsu (2002) proposed

non-parametric schemes in multiple imputation to impute censored survival times. Royston (2001) presented a simple method of imputation by substituting a censored survival time with a randomly imputed value sampled from log-normal distribution. MI is generally based on Bayesian framework. Thus, the methodology for MI involves executing posterior distribution of censored data given the observed data. In this thesis, a non-Bayesian approach was proposed to impute censored survival times. The key idea is to replace each censored time with an imputed time obtained by fitting a suitable regression model, and then execute a quantile regression model.

3.2 Proposed Method

In this section, we introduce the proposed method developed to adjust censoring in quantile regression model. We first review the accelerated failure time (AFT) model and then the MI method in this section.

3.2.1 Accelerated Failure Time Model

Accelerated failure time (AFT) model has been a widely used method to investigate the direct effects of covariates on mean survival times. Let T_i and C_i be the corresponding logarithm of failure time and right censoring time for the $i^{th}(i = 1, 2, ..., n)$ subject. Define the observed time $Y_i = T_i \wedge C_i$ and $\delta_i = I(T_i \leq C_i)$, where $\wedge$ is the minimum operator and $I(\cdot)$ is the function used to indicate whether the event time is right censored ($\delta_i = 0$) or not ($\delta_i = 1$). The observed data is a triplet (Y_i, δ_i, X_i), where X_i is a $(p \times 1)$ vector of covariates and p is the number of covariates. Under the assumption that T_i is independent of C_i conditional on X_i, the AFT model can be represented as

$$T_i = \beta_0 + X_i'\beta + \varepsilon_i, \qquad i = 1, 2, ..., n,$$

where β_0 is the intercept, β is a $(p \times 1)$ vector of coefficients corresponding to $(1 \times p)$ vector of covariates X_i, and ε_i is the independent and identically distributed (i.i.d)

random error with mean zero and finite variance. We consider the semi-parametric AFT model where the distribution of ε_i is not specified.

3.2.2 Buckley-James Estimation for AFT Model

Buckley-James (BJ) estimation approach, proposed by Buckley and James (1979), simply relies on an iterative solution to the standard least squares normal equations that have been modified to account for the censoring.

In the absence of censoring ($\delta_i = 1$ for $i = 1, 2, ..., n$), the survival regression model can be considered as classical linear regression model where least square estimation plays a significant role in data analysis. The ordinary least square (OLS) estimators of β_0 and β can be obtained by minimizing the objective function,

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \beta_0 - X_i'\beta\right)^2, \tag{3.1}$$

with respect to intercept and regression parameters. The minimization of (2.1) yields the following estimating equation for regression parameter.

$$\begin{aligned}\sum_{i=1}^{n} X_i(Y_i - X_i'\hat{\beta}) &= 0\\ \sum_{i=1}^{n} X_i Y_i &= \sum_{i=1}^{n} X_i X_i'\hat{\beta}\end{aligned} \tag{3.2}$$

The intercept β_0 can be estimated by, $\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{\beta})$ where $\varepsilon_i(\hat{\beta}) = Y_i - X_i'\hat{\beta}$.

In the presence of right censoring, Buckley and James (1979) proposed to replace censored observation (C_i) with the conditional expectation $E(Y_i|Y_i > C_i)$. Thus, the BJ weighted response variable has the form,

$$Y_i^* = \delta_i Y_i + (1-\delta_i)E(Y_i|Y_i > C_i, X_i, \delta_i). \tag{3.3}$$

Lemma 3.2.2.1. *If* $Y_i^* = \delta_i Y_i + (1-\delta_i)E(Y_i|Y_i > C_i, X_i, \delta_i)$, *then* $E(Y_i^*) = \beta_0 + X_i'\beta = E(Y_i)$.

Proof:

$$\begin{aligned}
E(Y_i^*) &= E(E(Y_i^*|\delta_i)) \\
&= E(Y_i^*|\delta_i = 1).P(\delta_i = 1) + E(Y_i^*|\delta_i = 0)P(\delta_i = 0) \\
&= E(Y_i|\delta_i = 1).P(\delta_i = 1) + E(E(Y_i|Y_i > C_i)|\delta_i = 0).P(\delta_i = 0) \\
&= E(Y_i|\delta_i = 1).P(\delta_i = 1) + E(Y_i|Y_i > C_i).P(\delta_i = 0) \\
&= E(Y_i|\delta_i = 1).P(\delta_i = 1) + E(Y_i|\delta_i = 0).P(\delta_i = 0) \\
&= E(Y_i) = \beta_0 + X_i'\beta
\end{aligned}$$

This completes the proof of Lemma 3.2.2.1.

Let, $\acute{Y_i} = E(Y_i|Y_i > C_i, X_i, \delta_i)$. Note that when $\delta_i = 0$,

$$\begin{aligned}
\acute{Y_i} &= E(Y_i|Y_i > C_i, X_i, \delta_i) \\
&= \beta_0 + X_i'\beta + E(\varepsilon_i|\varepsilon_i > Y_i - (\beta_0 + X_i'\beta)) \\
&= \beta_0 + X_i'\beta + \frac{\int_{Y_i-(\beta_0+X_i'\beta)}^{\infty} u dF_\beta(u)}{1 - F_\beta\{Y_i - (\beta_0 + X_i'\beta)\}} \\
&= \beta_0 + X_i'\beta + \frac{\int_{\varepsilon_i}^{\infty} u dF_\beta(u)}{S_\beta(\epsilon_i)} \qquad (3.4)
\end{aligned}$$

where $\varepsilon_i = Y_i - \beta_0 - X_i'\beta, i = 1, 2, ..., n$, are the error terms. $F_\beta(\varepsilon)$ and $S_\beta(\varepsilon)$ stand for the distribution function and survival function of the error terms, respectively. Using Kaplan-Meier estimator, the cdf $\hat{F}_\beta(\varepsilon)$ can be obtained in a non-parametric way. This approach gives the Kaplan-Meier estimator $\hat{S}_\beta(\varepsilon)$ as follows.

$$\begin{aligned}
\hat{S}_\beta(\varepsilon) &= 1 - \hat{F}_\beta(\varepsilon) \\
&= 1 - \left[1 - \prod_{\varepsilon_i \le \varepsilon} \left(\frac{n-i}{n-i+1}\right)^{\delta_i}\right] \\
&= \prod_{\varepsilon_i \le \varepsilon} \left(\frac{n-i}{n-i+1}\right)^{\delta_i}.
\end{aligned}$$

In the above formula, $\hat{F}_\beta(\varepsilon)$ will not tend to 1 if the largest residual is censored. The convention in BJ method is to adopt the largest ε_i, denoted by $\varepsilon_{(n)}$, be uncensored and thus $\hat{F}_\beta(\varepsilon_{(n)}) = 1$. Now, the Buckley-James estimator $\hat{\beta}$ can be obtained after

the modification in Equation (3.2) which satisfies,

$$\hat{\beta}^{BJ} = \sum_{i=1}^{n} \left(X_i X_i'\right)^{-1} \sum_{i=1}^{n} \left(X_i Y_i^*\right) \tag{3.5}$$

Computation of BJ estimators follow an iterative procedures which can be described below

Step 1. Obtain a starting value of β, say, $\hat{\beta}^1$

Step 2. Compute $Y_i^*(\hat{\beta}^j)$ in Equation (3.4) at j^{th} iteration.

Step 3. Obtain $\hat{\beta}^{j+1}$ using the result given in (3.5).

Step 4. Go back to Step 2 until convergence criterion $|\hat{\beta}^{j+1} - \hat{\beta}_j| < 0.01$ is met.

3.2.3 The Buckley-James Multiple Imputation Method in Quantile Regression

Without loss of generality, let us assume that the observed failure times are y_j for $j = 1, ..., n_1$, and the censored ones are for $j = n_1 + 1, ..., n_1 + n_0$ where $n_1 + n_0 = n$. Based on the available data set $\{(y_j, x_j, \delta_j), j = 1, ..., n\}$, we can obtain the Buckley-James estimators of regression coefficients $\hat{\beta}_0^{BJ}, \hat{\beta}_1^{BJ}, ..., \hat{\beta}_p^{BJ}$. The predicted values of the j^{th} observation will be $X_j'\hat{\beta}^{BJ}$, $j = 1, ..., n$. Correspondingly, the residuals for the n_1 observed failure times can be calculated as, $\hat{r}_j = y_j - X_j'\hat{\beta}^{BJ}$, $j = 1, ..., n$.
In this thesis, we propose to predict the censored failure times by

$$\tilde{y}_j = x_j'\hat{\beta}^{BJ} + \hat{r}_j - \hat{r}(\tau), \quad j = 1, ...n,$$

where $\hat{r}_j$ is a randomly selected (with replacement) residual from those available observations, and $\hat{r}(\tau)$ is the τ^{th} sample quantile of those residuals. The $\tilde{y}_j$ are imputed values of censored ones. Now let,

$$\mathcal{Y}_{\mathcal{J}} = \begin{cases} y_j, \text{if} \ \ , j = 1, ..., n_1 \\ \tilde{y}_j, j = n_1 + 1, ..., n_1 + n_0 \end{cases}$$

Now, $\{\mathcal{Y}_{\mathcal{J}}, x_j\}$, $j = 1, ..., n$ becomes a data without censoring. Hence a quantile regression model can be applied to fit this data.

Note 1: One other option for this imputation is to utilize the value $\hat{E}(Y_j|Y_j > C_j, X_j = x_j)$ at convergence. Our proposed method is an approximation of those quantities, plus the introduced residual terms. It is more straight forward to apply our method. Furthermore, the added residual term could remedy possible overfitting, caused by using $\hat{E}(Y_j|Y_j > C_j, X_j = x_j)$.

Note 2: Using $x_j'\hat{\beta}^{BJ} + \hat{r}_j - \hat{r}(\tau)$ as an imputation leads to a simplified justification for the model performance. This result will be seen when we prove the consistency and asymptotic normality.

Note 3: Introducing randomly selected modified residuals enables us to obtain more estimates of the quantile regression coefficients when applying different subsamples of the residuals. By taking an average of these estimates, we could efficiently control extra variability introduced by adding randomly selected residuals. To be specific, if we select m subsamples of residuals (with replacement), and for the k^{th} subsample, we define,

$$\tilde{y}_j^{(k)} = x_j'\hat{\beta}^{BJ} + \hat{r}_j^{(k)} - \hat{r}(\tau), \quad k = 1, ...m,$$

and use it to replace the censoring. The k^{th} augmented sample leads to an estimate of quantile regression coefficient at the τ^{th} level, $\hat{\beta}^{(k)}(\tau)$, $k = 1, ..., m$. Our multiple imputation estimate then takes the form,

$$\hat{\beta}^*(\tau) = \frac{1}{m}\sum_{k=1}^{m}\hat{\beta}^{(k)}(\tau).$$

3.3 Consistency and Asymptotic Normality

In this section, we establish the consistency and asymptotic normality of the $k^{th}(k = 1, 2, ..., m)$ estimator $\hat{\beta}^{(k)}(\tau)$ which can be obtained by minimizing

$$R_n(\beta(\tau)) = \sum_{i=1}^{n_1} \rho_\tau(y_i - x_i'\beta(\tau)) + \sum_{j=n_1+1}^{n} \rho_\tau(\tilde{y_j} - x_j'\beta(\tau)).$$

We can rewrite $R_n(\beta(\tau))$ as,

$$\begin{aligned} R_n(\beta(\tau)) =& \sum_{i=1}^{n_1} \rho_\tau(y_i - x_i'\beta(\tau)) + \sum_{j=n_1+1}^{n} \rho_\tau(y_j - x_j'\beta(\tau)) + \\ & \sum_{j=n_1+1}^{n} \rho_\tau(\tilde{y_j} - x_j'\beta(\tau)) - \sum_{j=n_1+1}^{n} \rho_\tau(y_j - x_j'\beta(\tau)) \\ =& \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta(\tau)) + \left(\sum_{j=n_1+1}^{n} \rho_\tau(\tilde{y_j} - x_j'\beta(\tau)) - \sum_{j=n_1+1}^{n} \rho_\tau(y_j - x_j'\beta(\tau)) \right) \end{aligned}$$

We next redefine $\tilde{y}_j$ to recover y_j as follows,

$$\begin{aligned} \tilde{y}_j &= x_j'\hat{\beta}^{BJ} + \hat{r}_j - \hat{r}(\tau), \\ &= x_j'\beta + \varepsilon_j + x_j'(\hat{\beta}^{BJ} - \beta) + (\hat{r}_j - \hat{r}(\tau) - \varepsilon_j), \\ &= y_j + x_j'(\hat{\beta}^{BJ} - \beta) + (\hat{r}_j - \hat{r}(\tau) - \varepsilon_j). \end{aligned}$$

From Buckley and James (1979), we get that $(\hat{\beta}^{BJ} - \beta) \xrightarrow{p} 0$ as $n \to \infty$. Utilizing the identity given by Knight (1998), which states,

$$\rho_\tau(u - v) - \rho_\tau(u) = -v\psi_\tau(u) + \int_0^v (I(u \le s) - I(u \le 0))ds$$

where $\rho_\tau(\cdot)$ is the quantile regression check function defined before and $\psi_\tau(u) = \tau - I(u < 0)$ is the quantile influence function.

$$\begin{aligned} \rho_\tau(y_j - x_j'\beta(\tau) - (\hat{r}(\tau) + \varepsilon_j - \hat{r}_j)) - \rho_\tau(y_j - x_j'\beta(\tau)) &= -(\hat{r}(\tau) + \varepsilon_j - \hat{r_j})\psi_\tau(y_j - x_j'\beta(\tau)) \\ &+ \int_0^{\hat{r}(\tau)+\varepsilon_j-\hat{r_j}} \big(I(y_j - x_j'\beta(\tau) \le s) - I(y_j - x_j'\beta(\tau) \le 0)\big) \end{aligned}$$

which does not depend neither β nor $\beta(\tau)$ since $y_j - x_j'\beta(\tau) = \varepsilon_j(\tau)$. Thus, the second term of the objective function,

$$\sum_{j=n_1+1}^{n} \left(\rho_\tau(y_j - x_j'\beta(\tau) - (\hat{r}(\tau) + \varepsilon_j - \hat{r_j})) - \rho_\tau(y_j - x_j'\beta(\tau)) \right)$$

can be considered as a remainder $R(c)$, which is a constant as a function of $\beta(\tau)$. Minimizing $R_n(\beta(\tau))$ is equivalent to minimize

$$R_n(\beta(\tau)) \approx \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta(\tau)) + n_0 R(c). \tag{3.6}$$

To prove consistency and asymptotic normality, the following conditions are required.

A1. Let Y_i be an independent random variables with continuous CDF F_i and continuous density $f_i(\varepsilon)$ uniformly bounded away from 0 and ∞ at points $\varepsilon_i(\tau)$, $i = 1, 2, ..., n$.

A2. There exists positive definite matrices C_0 and C_1 such that

1. $lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i x_i' = C_0$
2. $lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f_i(\varepsilon_i(\tau)) x_i x_i' = C_1$
3. $\frac{sup_i ||x_i||}{\sqrt{n}} \to 0$

Theorem 3.3.1. *Under regularity conditions $A1 - A2$ listed above,*

$$\hat{\beta}^{(k)}(\tau) \xrightarrow{p} \beta_0(\tau).$$

Proof:

The first step in the proof is to calculate the probability limit of the minimand. It is convenient to normalize the minimand $R_n(\beta(\tau))$ by subtracting off its value at the true parameter $\beta_0(\tau)$, which clearly does not affect the minimizing value $\hat{\beta}^{(k)}(\tau)$. That is,

$$\begin{aligned}
\hat{\beta}^{(k)}(\tau) &= \arg\min_{\beta(\tau)} \frac{1}{n} \Big[R_n(\beta(\tau)) - R_n(\beta_0(\tau)) \Big] \\
&= \arg\min_{\beta(\tau)} \frac{1}{n} \sum_{i=1}^{n} \Big[\rho_\tau(y_i - x_i'\beta(\tau)) - \rho_\tau(y_i - x_i'\beta_0(\tau)) \Big] \\
&= \arg\min_{\beta(\tau)} \frac{1}{n} \sum_{i=1}^{n} \Big[\rho_\tau(y_i - x_i'\beta_0(\tau) + x_i'\beta_0(\tau) - x_i'\beta(\tau)) - \rho_\tau(y_i - x_i'\beta_0(\tau)) \Big] \\
&= \arg\min_{\beta(\tau)} \frac{1}{n} \sum_{i=1}^{n} \Big[\rho_\tau(\varepsilon_i(\tau) - x_i'\delta(\tau)) - \rho_\tau(\varepsilon_i(\tau)) \Big],
\end{aligned}$$

where, $\delta(\tau) = \beta(\tau) - \beta_0(\tau)$.

But since

$$-||x_i|| \cdot ||\delta(\tau)|| \leq |\varepsilon_i(\tau) - x_i'\delta(\tau)| - |\varepsilon_i(\tau)| \leq ||x_i|| \cdot ||\delta(\tau)||$$

by the triangle and Cauchy-Schwarz inequalities (Cauchy–Schwarz inequality, 2001), the normalized minimand is a sample average of i.i.d. random variables with finite first and second moments. So, according to Khintchine's law of large number (Law of large numbers, 2002),

$$\begin{aligned} R_n(\beta(\tau)) - R_n(\beta(\tau_0)) \rightarrow^p \bar{R}(\delta(\tau)) & \\ =& E\big[R_n(\beta(\tau_0)) - R_n(\beta(\tau_0))\big] \\ =& E\big[\rho_\tau(\varepsilon_i(\tau) - x_i'\delta(\tau)) - \rho_\tau(\varepsilon_i(\tau))\big] \\ =& E\big[(\varepsilon_i(\tau) - x_i'\delta(\tau))sgn_\tau(\varepsilon_i(\tau) - x_i'\delta(\tau)) - \varepsilon_i(\tau)sgn_\tau(\varepsilon_i(\tau))\big] \\ =& E\big[(\varepsilon_i(\tau) - x_i'\delta(\tau))\{sgn_\tau(\varepsilon_i(\tau) - x_i'\delta(\tau)) - sgn_\tau(\varepsilon_i(\tau))\}\big] \\ =& E\left[\int_{x_i'\delta(\tau)}^{0} (\nu - x_i'\delta(\tau))f(\nu|x_i)d\nu\right] \end{aligned} \tag{3.7}$$

Here, $\rho_\tau(u)$ has been expressed in terms of τ weighted *sign* function (Fitzenberger, 1997). That is,

$$sgn_\tau(u) \equiv \tau I(u > 0) - (1 - \tau)I(u < 0).$$

The last four equalities use the fact that

$$E[(x_i'\delta(\tau))sgn_\tau\{\varepsilon_{\tau_i}\}] = E[E[(x_i'\delta(\tau))sgn_\tau\{\varepsilon_{\tau_i}\}|x_i]] = 0.$$

The integral in Equation 3.7 is well-defined for both positive and negative values of $x_i'\beta(\tau)$, under the standard convention $\int_a^b dF = -\int_b^a dF$. By inspection, $\bar{R}(\delta(\tau))$ equals zero at $\delta(\tau) = \beta(\tau) - \beta_0(\tau) = 0$, and is non-negative otherwise since the sign of the integrand is same as the sign of the lower limit $x_i'\delta(\tau)$. Furthermore, $R_n(\beta_\tau) - R_n(\beta_0(\tau))$ is convex for all n (Hjort & Pollard, 1993), so is its probability limit $\bar{R}(\beta(\tau) - \beta_0(\tau))$. Thus, if $\beta(\tau) = \beta_0(\tau)$ is a unique local minimizer, it is also a

global minimizer, implying the consistency of $\hat{\beta}^{(k)}(\tau)$. But, by the Leibnitz' rule

$$\begin{aligned}\frac{\partial \bar{R}(\delta(\tau))}{\partial \delta(\tau)} =& - E[x_i \cdot \int_{x_i'\delta(\tau)}^{0} f(\nu|x_i)d\nu] \\ \frac{\partial \bar{R}(0)}{\partial \delta(\tau)} =& 0\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2 \bar{R}(\delta(\tau))}{\partial \delta(\tau)\partial \delta(\tau)'} =& E[x_i x_i' \cdot f(x_i'\delta(\tau)|x_i)] \\ \frac{\partial^2 \bar{R}(\delta(\tau))}{\partial \delta(\tau)\partial \delta(\tau)'} =& E[x_i x_i' \cdot f(0|x_i)] \equiv C\end{aligned}$$

which is positive definite. So $\delta(\tau) = 0 = \beta(\tau) - \beta_0(\tau)$ is indeed a unique local (and global) minimizer of $\bar{R}(\delta(\tau)) = \bar{R}(\beta(\tau) - \beta_0(\tau))$, and thus

$$\hat{\beta}^{(k)}(\tau) \xrightarrow{p} \beta_0(\tau)$$

We next study the theorems about the asymptotic notmality of the estimator proposed in the previous section.

Theorem 3.3.2. *From Koenker (2008), under regularity conditions $A1 - A2$ listed above,*

$$\sqrt{n}(\hat{\beta}(\tau) - \beta_0(\tau)) \sim N\big(0, \tau(1-\tau)C_1^{-1}C_0C_1^{-1}\big).$$

Theorem 3.3.3. *For fixed $m \to \infty$ and $n \to \infty$, the estimator $\hat{\beta}^*(\tau)$ is defined as the average of m such τ^{th} quantile estimators. By the properties of the normal distribution and the Central Limit Theorem, we can show that*

$$\sqrt{n}(\hat{\beta}^*(\tau) - \beta_0(\tau)) \xrightarrow{D} N\bigg(0, \tau(1-\tau)C_1^{-1}C_0C_1^{-1}\bigg).$$

Note that, when n and m both are fixed, the averaged estimator $\hat{\beta}^*(\tau)$ is more stable than using just one imputation.

Remarks:

- The sampling variance was estimated by,

$$\begin{aligned}\hat{V}ar(\hat{\beta}^*(\tau)) =& \frac{1}{m}\sum_{k=1}^{m}\hat{V}ar(\hat{\beta}^{(k)}(\tau)) + \\ & \frac{1}{m-1}\sum_{k=1}^{m}(\hat{\beta}^{(k)}(\tau) - \hat{\beta}^*(\tau))^2 \\ =& I + II\end{aligned}$$

- if $n \to \infty$ and $m \to \infty$,

$$\hat{V}ar(\hat{\beta}^*(\tau)) - \hat{V}ar(\hat{\beta}^{(k)}(\tau)) \xrightarrow{p} 0$$

  Hence, term (I) dominates the variance estimation.

Chapter 4

Numerical Studies

To evaluate the performance of the proposed method, especially for small sample sizes, we conducted extensive simulation studies. This section compares and displays the simulation results for the censored quantile method (CQR) proposed by Portnoy (2003) and quantile regression estimation method with BJ multiple imputation given in the previous Chapter.

4.1 Simulation setup

We consider two cases in our simulation studies.

Case 1

Let T_i and C_i be the logarithm of failure time and censoring time. A set of data $(x_i, T_i, C_i, \delta_i)$ was generated from the following model.

$$T_i = b_0 + b_1 x_{1i} + e_i, \qquad i = 1, 2, ..., n,$$

where $b_0 = 2$, $b_1 = -0.2$, $x_{1i} \sim$ U(0,5) and the e_i are iid N(0,1). To introduce right censoring, the censoring variable C_i^1 and C_i^2 were generated from the uniform distribution on the interval (0,10) and (0,6) resulting in 15% and 26% censoring rates,

respectively. The observed response variable is Y_i =min(T_i,C_i). First, we created a data frame of 10000 observations including (T_i, C_i^1, C_i^2, x_i), i $= 1, ...10000$. We conducted τ^{th} quantile regression using the quantreg package in R. These estimates are considered as true parameters of a specific value of τ.

For each censoring rate, we randomly selected 1000 random samples with n=100,200 and 500. The current research conducted Monte Carlo simulation in R programming to compare the performance of the multiply imputed quantile regression (MIQR) with that of Portnoy's CQR. The R package quantreg (version 4.24), were used to get the results of CQR. The existing *crq* function in quantreg package provided confidence intervals with coverage probability (CP) below $1 - \alpha < 0.95$. To avert the situation we executed crq bootstrapping. The mean of the bias was reported based on 1000 simulation runs and 250 bootstrap samples were taken for estimating the standard error (SE) of the estimates and the calculation of the CP of a 95% confidence interval of the model parameters. Table 4.1 summarizes simulation results at different quantiles 0.1,0.25, 0.5,0.75 and 0.9.

It can be seen that irrespective of sample size and censoring level, at τ =0.1,0.25 and 0.5 MIQR and CQR have quite similar performance in terms of bias and standard error (SE). Moreover, the bootstrap confidence intervals of both methods have coverage probabilities close to the nominal level 1-α= 0.95. It is also observed that for small sample size, censored quantile regression (CQR) fails to provide estimates at higher quantiles (τ=0.9) while multiply imputed QR provides quantile estimates with small bias. Moreover, at higher censoring rate (0.26) CQR method produces NA.

Table 4.1: Bias, SE(standard error) and CP(coverage probability) to the estimators of β_0 and β_1 using methods (CQR, MIQR) at different quantiles (τ=0.1,0.25,0.5,0.75,0.9) for censoring rates (CR) 15% and 26%

				β_0			β_1		
τ	CR	n	Method	Bias	SE	CP	Bias	SE	CP
0.1	15%	100	CQR	-0.0150	0.3549	0.946	0.0026	0.1296	0.962
			MIQR	-0.0242	0.3872	0.953	0.0027	0.1413	0.968
		200	CQR	-0.0143	0.2511	0.949	0.0028	0.0884	0.963
			MIQR	-0.0283	0.2659	0.951	0.0030	0.0932	0.969
		500	CQR	-0.0149	0.1551	0.961	0.0017	0.0543	0.960
			MIQR	-0.0314	0.1650	0.965	0.0019	0.0573	0.964
	26%	100	CQR	-0.0240	0.3521	0.957	0.0105	0.1298	0.958
			MIQR	-0.0372	0.3875	0.965	0.0095	0.1422	0.963
		200	CQR	-0.0090	0.2550	0.966	0.0025	0.0894	0.966
			MIQR	-0.0358	0.2726	0.967	0.0031	0.0947	0.961
		500	CQR	-0.0075	0.1575	0.952	-0.0021	0.0546	0.950
			MIQR	-0.0355	0.1687	0.957	-0.0018	0.0578	0.953
0.25	15%	100	CQR	0.0121	0.2935	0.945	-0.0074	0.1083	0.960
			MIQR	-0.0408	0.3189	0.952	-0.0075	0.1165	0.958
		200	CQR	0.0002	0.2106	0.952	-0.0038	0.0736	0.959
			MIQR	-0.0501	0.2217	0.941	-0.0042	0.0769	0.954
		500	CQR	-0.0017	0.1273	0.963	-0.0044	0.0440	0.955
			MIQR	-0.0487	0.1364	0.941	-0.0051	0.0469	0.950
	26%	100	CQR	-0.0123	0.3015	0.962	0.0046	0.1099	0.955
			MIQR	-0.0970	0.3292	0.941	0.0039	0.1188	0.960
		200	CQR	0.0042	0.2115	0.951	-0.0038	0.0731	0.950
			MIQR	-0.0810	0.2277	0.943	-0.0038	0.0779	0.953

τ	CR	n	Method	Bias	SE	CP	Bias	SE	CP
		500	CQR	0.0061	0.1315	0.948	-0.0062	0.0451	0.953
			MIQR	-0.0752	0.1433	0.907	-0.0070	0.0487	0.944
0.5	15%	100	CQR	0.0338	0.2821	0.956	-0.0049	0.1032	0.952
			MIQR	-0.0828	0.3068	0.936	-0.0025	0.1118	0.955
		200	CQR	0.0264	0.1944	0.948	-0.0030	0.0681	0.952
			MIQR	-0.0921	0.2132	0.908	-0.0001	0.0739	0.955
		500	CQR	0.0214	0.1218	0.948	-0.0042	0.0420	0.953
			MIQR	-0.0955	0.1316	0.864	-0.0003	0.045	0.942
	26%	100	CQR	0.0199	0.2996	0.957	0.0016	0.1086	0.958
			MIQR	-0.1944	0.3289	0.890	0.0098	0.1180	0.950
		200	CQR	0.0286	0.2060	0.958	-0.0051	0.0715	0.962
			MIQR	-0.1819	0.2270	0.868	0.0031	0.0780	0.951
		500	CQR	0.0205	0.1278	0.954	-0.0024	0.0439	0.958
			MIQR	-0.1881	0.1402	0.711	0.0049	0.0478	0.948
0.75	15%	100	CQR	0.0256	0.3204	0.950	-0.0103	0.1175	0.948
			MIQR	-0.1162	0.3499	0.918	-0.0066	0.1270	0.946
		200	CQR	0.0238	0.2267	0.945	-0.0078	0.0785	0.958
			MIQR	-0.1279	0.2421	0.909	-0.0017	0.0833	0.962
		500	CQR	0.0098	0.1387	0.952	-0.0053	0.0479	0.963
			MIQR	-0.1356	0.1474	0.825	-0.0002	0.0505	0.955
	26 %	100	CQR	NA	NA	NA	NA	NA	NA
			MIQR	-0.2774	0.3909	0.858	0.0122	0.1417	0.955
		200	CQR	0.0222	0.2488	0.963	-0.0065	0.0856	0.958
			MIQR	-0.2698	0.2660	0.801	0.0081	0.0910	0.962
		500	CQR	0.0100	0.1513	0.951	-0.0047	0.0515	0.955
			MIQR	-0.2810	0.1627	0.570	0.0110	0.0550	0.955

τ	CR	n	Method	Bias	SE	CP	Bias	SE	CP
0.9	15%	100	CQR	NA	NA	NA	NA	NA	NA
			MIQR	-0.0704	0.6908	0.925	0.0010	0.2485	0.971
		200	CQR	NA	NA	NA	NA	NA	NA
			MIQR	-0.1310	0.3204	0.899	0.0099	0.1117	0.954
		500	CQR	0.0121	0.1784	0.956	0.0060	0.0618	0.962
			MIQR	-0.1422	0.1883	0.851	0.0109	0.0646	0.969
	26%	100	CQR	NA	NA	NA	NA	NA	NA
			MIQR	-0.2044	0.8784	0.862	0.0224	0.3094	0.978
		200	CQR	NA	NA	NA	NA	NA	NA
			MIQR	-0.2840	0.3892	0.809	0.0201	0.1311	0.974
		500	CQR	NA	NA	NA	NA	NA	NA
			MIQR	-0.3115	0.2110	0.623	0.0252	0.0713	0.947

Case 2

To gain more insight, we further execute another simulation. In the second set of simulations, the data were generated from similar setting as in the previous case, but the number of covariates was increased. The failure times were generated from the log-linear model,

$$T_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + e_i \qquad i = 1, 2, ..., n$$

where, $\boldsymbol{x_i} = (1, x_{i1}, x_{i2})^T$, $\boldsymbol{x_1}$ was simulated from Unif(0,5) and $\boldsymbol{x_2}$ from Bernoulli(0.5). ϵ_i was generated from $N(0, 1)$. The initial quantile regression coefficients are (b_0, b_1, b_2) =(2,0.2,1.2). The censoring times were generated as $C_i \sim Uniform(0, \theta)$. We take θ=(20,12) to produce 15% and 26% of censoring. Similar to previous setup, at different quantiles, with n=100,200 and 500, the results of CQR and MIQR were summarized in Table 4.2. At higher quantile level (τ=0.75,0.9), the performance of the multiply imputed quantile method is satisfactory with two covariates including one discrete

and one continuous variates. From the simulation results, it is observed that for τ=0.75, CQR produces NA for n=100 at 15% and 26% of censoring rates. It is also depicted that for extreme quantile, such as $\tau = 0.9$, Portnoy's method fails to obtain estimates even for small sample size ($n = 100$) at lower censoring rate (15%). Even if we increased sample size (n=500) at the censoring rate (26%), CQR did not always create estimates of the model parameters. But MIQR succeeded to obtain reasonable estimates.

Table 4.2: Bias, SE(standard error) and CP(coverage probability) to the estimators of β_0, β_1 and β_2 using methods (CQR, MIQR) at different quantiles (τ=0.1,0.25,0.5,0.75,0.9) for censoring rates (CR) 15% and 26%

				β_0			β_1			β_2		
τ	CR	n	Method	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP
0.1	15%	100	CQR	0.0499	0.3981	0.953	-0.0065	0.1340	0.959	0.0014	0.3603	0.975
			MIQR	0.0618	0.4249	0.955	-0.0009	0.1424	0.971	0.0355	0.3843	0.975
		200	CQR	0.0588	0.2965	0.948	-0.0115	0.0928	0.971	-0.0077	0.2591	0.962
			MIQR	0.0677	0.3095	0.959	-0.0058	0.0962	0.972	0.0258	0.2690	0.972
		500	CQR	0.0647	0.1736	0.932	-0.0134	0.0564	0.950	-0.0169	0.1600	0.959
			MIQR	0.0717	0.1805	0.933	-0.0078	0.0583	0.957	0.0192	0.1658	0.959
	26%	100	CQR	0.0935	0.4019	0.946	-0.0110	0.1390	0.957	-0.0224	0.3678	0.964
			MIQR	0.1082	0.4312	0.951	-0.0010	0.1475	0.969	0.0380	0.3927	0.970
		200	CQR	0.0684	0.3075	0.940	-0.0121	0.0969	0.964	-0.0100	0.2675	0.962
			MIQR	0.0768	0.3207	0.945	-0.0016	0.1001	0.967	0.0567	0.2778	0.962
		500	CQR	0.0530	0.1775	0.930	-0.0100	0.0583	0.948	-0.0131	0.1654	0.963
			MIQR	0.0601	0.1859	0.937	0.0006	0.0604	0.957	0.0512	0.1716	0.954
0.25	15%	100	CQR	0.0063	0.3335	0.959	0.0069	0.1141	0.971	-0.0329	0.3020	0.956
			MIQR	0.0089	0.3377	0.961	0.0125	0.1137	0.975	-0.0009	0.3039	0.955
		200	CQR	0.0111	0.2481	0.958	0.0052	0.0771	0.976	-0.0331	0.2152	0.959
			MIQR	0.0124	0.2459	0.960	0.0108	0.0755	0.970	-0.0020	0.2123	0.963
		500	CQR	0.0164	0.1419	0.950	0.0027	0.0457	0.965	-0.0390	0.1296	0.935
			MIQR	0.0184	0.1429	0.954	0.0082	0.0456	0.957	-0.0070	0.1297	0.948
	26%	100	CQR	0.0247	0.3395	0.969	0.0078	0.1187	0.960	-0.0487	0.3102	0.966
			MIQR	0.0295	0.3367	0.959	0.0164	0.1152	0.959	0.0082	0.3054	0.963
		200	CQR	0.0217	0.2554	0.960	0.0032	0.0797	0.967	-0.0319	0.2208	0.963
			MIQR	0.0228	0.2485	0.950	0.0125	0.0763	0.958	0.0235	0.2133	0.954
		500	CQR	0.0055	0.1461	0.926	0.0051	0.0477	0.940	-0.0291	0.1364	0.949
			MIQR	0.0077	0.1452	0.930	0.0141	0.0467	0.942	0.0269	0.1329	0.935
0.5	15%	100	CQR	0.0116	0.3126	0.963	-0.0003	0.1074	0.970	-0.0213	0.2848	0.961
			MIQR	-0.0360	0.2960	0.948	-0.0009	0.0995	0.960	-0.0221	0.2669	0.948

τ	CR	n	Method	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP
		200	CQR	0.0226	0.2269	0.951	-0.0010	0.0705	0.967	-0.0206	0.1954	0.963
			MIQR	-0.0266	0.2166	0.953	-0.0011	0.0664	0.962	-0.0219	0.1863	0.959
		500	CQR	0.0269	0.1336	0.952	-0.0038	0.0432	0.976	-0.0239	0.1228	0.956
			MIQR	-0.0226	0.1280	0.966	-0.0040	0.0409	0.967	-0.0237	0.1162	0.950
	26%	100	CQR	0.0276	0.3327	0.966	-0.0004	0.1152	0.967	-0.0250	0.3039	0.965
			MIQR	-0.0535	0.2963	0.966	-0.0007	0.0995	0.946	-0.0302	0.2679	0.949
		200	CQR	0.0318	0.2366	0.960	-0.0048	0.0749	0.965	-0.0201	0.2070	0.964
			MIQR	-0.0530	0.2149	0.943	-0.0043	0.0662	0.949	-0.0196	0.1853	0.938
		500	CQR	0.0147	0.1407	0.943	-0.0008	0.0459	0.956	-0.0143	0.1302	0.952
			MIQR	-0.0690	0.1271	0.914	-0.0008	0.0405	0.942	-0.0135	0.1150	0.933
0.75	15%	100	CQR	NA	NA	NA	NA	NA	NA	NA	NA	NA
			MIQR	-0.1221	0.3362	0.947	-0.0040	0.1135	0.964	-0.0491	0.3026	0.962
		200	CQR	-0.0002	0.2526	0.961	-0.0002	0.0779	0.967	-0.0177	0.2183	0.960
			MIQR	-0.1058	0.2411	0.937	-0.0064	0.0736	0.968	-0.0507	0.2071	0.946
		500	CQR	0.0081	0.1490	0.961	-0.0035	0.0483	0.965	-0.0060	0.1364	0.960
			MIQR	-0.0992	0.1424	0.897	-0.0089	0.0457	0.958	-0.0418	0.1295	0.935
	26%	100	CQR	NA	NA	NA	NA	NA	NA	NA	NA	NA
			MIQR	-0.1752	0.3321	0.923	-0.0145	0.1118	0.967	-0.0734	0.2984	0.953
		200	CQR	0.0114	0.2724	0.963	-0.0041	0.0864	0.969	-0.0053	0.2395	0.970
			MIQR	-0.1784	0.2415	0.879	-0.0137	0.0743	0.954	-0.0723	0.2079	0.948
		500	CQR	0.0012	0.1596	0.948	-0.0010	0.0523	0.955	0.0003	0.1494	0.963
			MIQR	-0.1844	0.1433	0.732	-0.0114	0.0458	0.941	-0.0643	0.1295	0.940
0.9	15%	100	CQR	NA	NA	NA	NA	NA	NA	NA	NA	NA
			MIQR	-0.0891	0.4272	0.945	-0.0128	0.1418	0.969	-0.1207	0.3860	0.958
		200	CQR	NA	NA	NA	NA	NA	NA	NA	NA	NA
			MIQR	-0.0758	0.3121	0.950	-0.0126	0.0958	0.965	-0.1183	0.2692	0.960
		500	CQR	0.0481	0.1897	0.955	-0.0088	0.0609	0.966	-0.0540	0.1733	0.944
			MIQR	-0.0708	0.1815	0.937	-0.0142	0.0581	0.958	-0.1005	0.1654	0.907
	26%	100	CQR	NA	NA	NA	NA	NA	NA	NA	NA	NA
			MIQR	-0.1715	0.4296	0.942	-0.0222	0.1438	0.966	-0.1588	0.3903	0.958
		200	CQR	NA	NA	NA	NA	NA	NA	NA	NA	NA
			MIQR	-0.1611	0.3148	0.917	-0.0236	0.0977	0.954	-0.1465	0.2711	0.935
		500	CQR	NA	NA	NA	NA	NA	NA	NA	NA	NA
			MIQR	-0.1605	0.1873	0.864	-0.0195	0.0599	0.950	-0.1459	0.1708	0.891

4.2 An application: HMO-HIV data

As an illustration, we applied the proposed method to HMO-HIV data from Hosmer and Lemeshow (Hosmer, Lemeshow, & May, 1999). A total of 100 patients were chosen and followed till death due to AIDS or AIDS related complications, until the end of the study or until the subject was lost to follow-up. The outcome variable of interest is survival time (month) to death after a confirmed diagnosis of HIV. Two covariates were considered: age (in years) at the start of follow-up and status of prior IV drug use (1=yes, 0=no).

Table 4.3: Estimated parameters (EP), their standard errors (SE) and corresponding 95% confidence intervals (CI) from fitting both the proposed quantile regression model (MIQR) and censored quantile regression (CQR) at three quartiles, $\tau = 0.25, 0.5, \text{and } 0.75$

		MIQR			CQR		
τ	Covariates	EP	SE	CI	EP	SE	CI
0.25	Intercept	4.369	0.919	(2.569, 6.170)	4.601	1.039	(2.564,6.637)
	Age	-0.077	0.025	(-0.126, -0.028)	-0.081	0.028	(-0.136,-0.027)
	Drug	-0.717	0.441	(-1.581, 0.147)	-0.780	0.457	(-1.675,0.115)
0.5	Intercept	5.071	0.641	(3.816, 6.327)	5.668	0.741	(4.215,7.120)
	Age	-0.079	0.017	(-0.113, -0.046)	-0.090	0.019	(-0.127,-0.053)
	Drug	-0.770	0.212	(-1.185, -0.354)	-0.864	0.253	(-1.360,-0.368)
0.75	Intercept	5.566	0.508	(4.570, 6.562)	NA	NA	NA
	Age	-0.075	0.014	(-0.102, -0.047)	NA	NA	NA
	Drug	-1.143	0.250	(-1.632, -0.654)	NA	NA	NA
0.9	Intercept	5.430	0.643	(4.169,6.691)	NA	NA	NA
	Age	-0.056	0.020	(-0.095,-0.017)	NA	NA	NA
	Drug	-1.43	0.262	(-1.949,-0.919)	NA	NA	NA

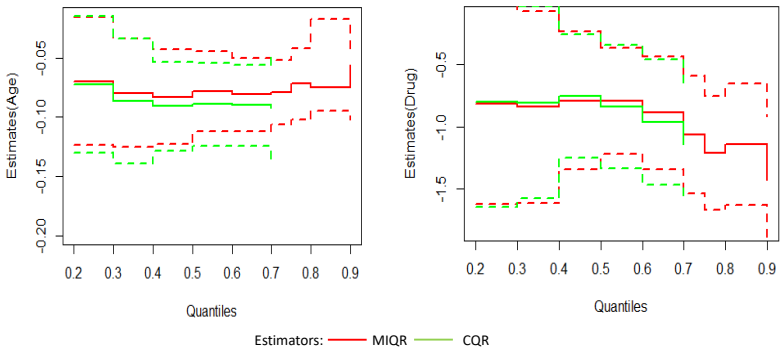

Figure 4.1: Estimators for the covariate effects in HMO data and dashed lines represnt the 95% CI for corresponding estimates

Of the subjects who were alive at the end of study or lost to follow-up, we only got partial or incomplete observation of survival time. In survival analysis, these incomplete data are referred as censored and HMO-HIV data included 20% of these type. A total of 250 bootstrap samples were performed for both methods (Portnoy's CQR and proposed MIQR). Table 4.3 displays the estimated quantile regression coefficients under Portnoy's CQR method and the proposed MIQR method, including the variance estimates and the 95% pointwise confidence intervals. Figure 4.1 provides an overall summary of the coefficient movements along with the 95% pointwise confidence intervals across the quantile levels. The findings indicate that both age and prior use of drugs are negatively related predictors, implying that increases in either variable are associated with shorter survival times. In this application of our proposed method, we can observe from Table 4.3 that Portnoy's method fails to get estimates of the covariates for some bootstrap samples for higher quantiles, such as $\tau = 0.75$ and $\tau = 0.9$. Our proposed method MIQR fulfilled the purpose to obtain estimates with standard errors and 95% confidence intervals of covariates age and drug. The two methods worked equally well for lower quantiles.

Chapter 5

Conclusion

In the thesis, we introduced a novel approach that utilizes model-based multiple imputations of censored observations through the Buckley-James estimation approach. Our method ensures consistent estimators of model parameters and achieves asymptotic normality when $n \to \infty$. This advancement addresses and overcomes the limitations of traditional censored quantile regression methods, particularly in estimating upper quantiles.

Through extensive simulation studies, we demonstrated the efficacy of our approach. The simulation studies reveals that our proposed method worked equally well with the Portnoy's CQR for cases with lower censoring rate. When the censoring rates are high, $\tau = 0.75$ or 0.9, our method outperforms the method of Portnoy. Additionally, the application of our method to a Health Maintenance Organization (HMO) dataset provided a practical illustration of its utility, further validating the theoretical findings. In the future, we plan to extend the proposed method to multivariate failure times by employing the copula method.

References

Allan, F., & Wishart, J. (1930). A method of estimating the yield of a missing plot in field experimental work. *The Journal of Agricultural Science*, *20*(3), 399–406.

Buckley, J., & James, I. (1979). Linear regression with censored data. *Biometrika*, *66*(3), 429–436.

Cade, B. S., & Guo, Q. (2000). Estimating effects of constraints on plant performance with regression quantiles. *Oikos*, *91*(2), 245–254.

Cauchy–Schwarz inequality. (2001). *Cauchy–schwarz inequality — Wikipedia, the free encyclopedia.* Retrieved from `https://en.wikipedia.org/wiki/Cauchy%E2%80%93Schwarz_inequality`

Chernozhukov, V., Fernández-Val, I., & Kowalski, A. E. (2015). Quantile regression with censoring and endogeneity. *Journal of Econometrics*, *186*(1), 201–221.

Chernozhukov, V., & Hong, H. (2002). Three-step censored quantile regression and extramarital affairs. *Journal of the American statistical Association*, *97*(459), 872–882.

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

Cox, D. R. (2018). *Analysis of survival data*. Chapman and Hall/CRC.

De Backer, M., El Ghouch, A., & Van Keilegom, I. (2017). Semiparametric copula quantile regression for complete or censored data.

De Backer, M., Ghouch, A. E., & Van Keilegom, I. (2019). An adapted loss function for censored quantile regression. *Journal of the American Statistical Association*,

114(527), 1126–1137.

Fitzenberger, B. (1997). Computational aspects of censored quantile regression. *Lecture Notes-Monograph Series*, 171–186.

Hao, L., & Naiman, D. Q. (2007). *Quantile regression* (No. 149). Sage.

Hjort, N., & Pollard, D. (1993). Asymptotics for minimisers of convex processes technical report. *Yale University*.

Honore, B., Khan, S., & Powell, J. L. (2002). Quantile regression under random censoring. *Journal of Econometrics*, *109*(1), 67–105.

Hosmer, D. W., Lemeshow, S., & May, S. (1999). Regression modeling of time to event data. *New York*.

Huang, L., Zhu, W., Saunders, C. P., MacLeod, J. N., Zhou, M., Stromberg, A. J., & Bathke, A. C. (2008). A novel application of quantile regression for identification of biomarkers exemplified by equine cartilage microarray data. *BMC bioinformatics*, *9*, 1–8.

Kalbfleisch, J. D., & Prentice, R. L. (2011a). *The statistical analysis of failure time data*. John Wiley & Sons.

Kalbfleisch, J. D., & Prentice, R. L. (2011b). *The statistical analysis of failure time data*. John Wiley & Sons.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, *53*(282), 457–481.

Klein, J. P., Moeschberger, M. L., et al. (2003). *Survival analysis: techniques for censored and truncated data* (Vol. 1230). Springer.

Knight, K. (1998). Limiting distributions for l1 regression estimators under general conditions. *Annals of statistics*, 755–770.

Koenker, R. (2008). Censored quantile regression redux. *Journal of Statistical Software*, *27*, 1–25.

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

Koenker, R., & Geling, O. (2001). Reappraising medfly longevity: a quantile regres-

sion survival analysis. *Journal of the American Statistical Association*, *96*(454), 458–468.

Law of large numbers. (2002). *Law of large numbers — Wikipedia, the free encyclopedia.* Retrieved from `https://en.wikipedia.org/wiki/Law_of_large_numbers`

Lawless, J. F. (2011). *Statistical models and methods for lifetime data.* John Wiley & Sons.

Machado, J. A., & Mata, J. (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of applied Econometrics*, *20*(4), 445–465.

Moghaddam, S., Newell, J., & Hinde, J. (2022). A bayesian approach for imputation of censored survival data. *Stats*, *5*(1), 89–107.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, *14*(4), 945–966.

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

Portnoy, S. (2003). Censored regression quantiles. *Journal of the American Statistical Association*, *98*(464), 1001–1012.

Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of econometrics*, *25*(3), 303–325.

Powell, J. L. (1986). Censored regression quantiles. *Journal of econometrics*, *32*(1), 143–155.

Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika*, *65*(1), 167–179.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *The Annals of Statistics*, 303–328.

Royston, P. (2001). The lognormal distribution as a model for survival time in cancer, with an emphasis on prognostic factors. *Statistica Neerlandica*, *55*(1), 89–104.

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological

bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the american statistical association* (Vol. 1, pp. 20–34).

Taylor, J. M., Murray, S., & Hsu, C.-H. (2002). Survival estimation and testing via multiple imputation. *Statistics & probability letters*, *58*(3), 221–232.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*, 354–372.

Walker, A. (1968). A note on the asymptotic distribution of sample quantiles. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *30*(3), 570–575.

Wang, H. J., & Wang, L. (2009). Locally weighted censored quantile regression. *Journal of the American Statistical Association*, *104*(487), 1117–1128.

Wei, G. C., & Tanner, M. A. (1991). Applications of multiple imputation to the analysis of censored regression data. *Biometrics*, 1297–1309.

Wei, L.-J., Ying, Z., & Lin, D. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika*, *77*(4), 845–851.

Wei, Y., Pere, A., Koenker, R., & He, X. (2006). Quantile regression methods for reference growth charts. *Statistics in medicine*, *25*(8), 1369–1382.

Wu, Y., & Yin, G. (2013). Cure rate quantile regression for censored data with a survival fraction. *Journal of the American Statistical Association*, *108*(504), 1517–1531.

Yin, G., Zeng, D., & Li, H. (2014). Censored quantile regression with varying coefficients. *Statistica Sinica*, 855–870.

Appendix

Table A.1: Bias, SE(standard error) and CP(coverage probability) to the estimators of β_0 and β_1 using $e_i \sim$ Gumbel$(0, 1)$ at different quantiles (τ=0.1,0.25,0.5,0.75,0.9) for censoring rates (CR) 16% and 28%

				β_0			β_1		
τ	CR	n	Method	Bias	SE	CP	Bias	SE	CP
0.1	16%	100	CQR	-0.0930	0.6423	0.944	0.0156	0.2385	0.945
			MIQR	-0.0801	0.6833	0.945	0.0174	0.2544	0.960
		200	CQR	-0.0481	0.4593	0.949	-0.0001	0.1637	0.963
			MIQR	-0.0456	0.4702	0.954	0.0013	0.1681	0.965
		500	CQR	-0.0546	0.2867	0.945	0.0058	0.1005	0.955
			MIQR	-0.0551	0.2944	0.960	0.0071	0.1032	0.960
	28%	100	CQR	-0.0743	0.6493	0.950	0.0018	0.2434	0.960
			MIQR	-0.0673	0.6868	0.958	0.0036	0.2593	0.966
		200	CQR	-0.0799	0.4685	0.951	0.0132	0.1648	0.959
			MIQR	-0.0828	0.4778	0.962	0.0153	0.1685	0.960
		500	CQR	-0.0631	0.2860	0.958	0.01165	0.0992	0.967
			MIQR	-0.0636	0.2913	0.967	0.0126	0.1016	0.965

				β_0			β_1		
τ	CR	n	Method	Bias	SE	CP	Bias	SE	CP
0.25	16%	100	CQR	0.0237	0.4245	0.954	-0.0101	0.1570	0.964
			MIQR	-0.0755	0.4270	0.958	-0.0012	0.1572	0.960
		200	CQR	0.0435	0.2981	0.947	-0.0186	0.1061	0.956
			MIQR	-0.0531	0.2972	0.929	-0.0116	0.1054	0.940
		500	CQR	0.0455	0.1843	0.930	-0.0171	0.0642	0.935
			MIQR	-0.0537	0.1874	0.939	-0.0091	0.0647	0.938
	28%	100	CQR	0.03951	0.4301	0.944	-0.0153	0.1589	0.967
			MIQR	-0.1184	0.4196	0.937	-0.0023	0.1550	0.948
		200	CQR	0.0563	0.3042	0.947	-0.0206	0.1063	0.957
			MIQR	-0.1116	0.2941	0.928	-0.0061	0.1022	0.943
		500	CQR	0.0450	0.1864	0.942	-0.0145	0.0649	0.961
			MIQR	-0.1157	0.1823	0.907	-0.0023	0.0630	0.952
0.5	16%	100	CQR	0.0392	0.3262	0.954	-0.0144	0.1189	0.952
			MIQR	-0.2012	0.3640	0.901	0.0039	0.1312	0.941
		200	CQR	0.0407	0.2213	0.941	-0.0164	0.0772	0.939
			MIQR	-0.1885	0.2509	0.868	-0.0009	0.0866	0.937
		500	CQR	0.0470	0.1389	0.922	-0.0168	0.0480	0.934
			MIQR	-0.1795	0.1560	0.784	-0.0021	0.0532	0.935
	28%	100	CQR	0.0343	0.3437	0.964	-0.0112	0.1248	0.965
			MIQR	-0.3882	0.3656	0.792	0.0179	0.1325	0.953
		200	CQR	0.0626	0.2379	0.955	-0.0205	0.0824	0.963
			MIQR	-0.3701	0.2608	0.674	0.0108	0.0893	0.931
		500	CQR	0.0512	0.1475	0.935	-0.0170	0.0499	0.950
			MIQR	-0.3858	0.1626	0.356	0.0177	0.0548	0.934

				β_0			β_1		
τ	CR	n	Method	Bias	SE	CP	Bias	SE	CP
0.75	16%	100	CQR	NA	NA	NA	NA	NA	NA
			MIQR	-0.2083	0.3424	0.898	0.0103	0.1232	0.954
		200	CQR	0.0237	0.2102	0.960	-0.0082	0.0731	0.962
			MIQR	-0.1995	0.2384	0.857	0.0067	0.0821	0.959
		500	CQR	0.0271	0.1300	0.934	-0.0085	0.0447	0.938
			MIQR	-0.1898	0.1462	0.728	0.0053	0.0494	0.943
	28%	100	CQR	NA	NA	NA	NA	NA	NA
			MIQR	-0.4980	0.3980	0.728	0.03990	0.1408	0.949
		200	CQR	NA	NA	NA	NA	NA	NA
			MIQR	-0.4686	0.2768	0.560	0.0341	0.0931	0.937
		500	CQR	0.0298	0.1483	0.949	-0.0088	0.0499	0.963
			MIQR	-0.4718	0.1683	0.201	0.0347	0.0557	0.906
0.9	16%	100	CQR	NA	NA	NA	NA		NA
			MIQR	-0.1651	0.4488	0.894	-0.0012	0.1606	0.974
		200	CQR	NA	NA	NA	NA	NA	NA
			MIQR	-0.1762	0.2534	0.851	-0.0001	0.0874	0.960
		500	CQR	0.01207	0.1423	0.937	-0.0104	0.0488	0.946
			MIQR	-0.1674	0.1527	0.788	-0.00020	0.0520	0.949
	28%	100	CQR	NA	NA	NA	NA	NA	NA
			MIQR	-0.4318	0.5568	0.780	0.0295	0.1961	0.973
		200	CQR	NA	NA	NA	NA	NA	NA
			MIQR	-0.4379	0.3101	0.629	0.0279	0.1041	0.952
		500	CQR	NA	NA	NA	NA	NA	NA
			MIQR	-0.4470	0.1812	0.308	0.0296	0.0604	0.932