On Rank-Based Inference for Quantile Regression

by

Yuan Sun

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2020

Doctoral Committee:

Professor Xuming He, Chair
Professor Moulinath Banerjee
Assistant Professor Yang Chen
Professor Peter Xuekun Song

Yuan Sun

yuansun@umich.edu

ORCID ID: 0000-0002-9989-2197

For all the people

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Quantile regression is a useful tool for testing the possible effect of covariates, especially when the effect is heterogeneous. Classical methods designed to test the effect at one quantile level can be sensitive to the quantile level choice. In this dissertation, we propose a regional quantile regression rank test as a generalization of the rank test at an individual quantile level. The proposed test statistic allows us to detect the treatment effect for a prespecified quantile interval by integrating the regression rank scores over the quantile region of interest. A new model-based bootstrap method is constructed to estimate the null distribution of the test statistic. A simulation study is conducted to demonstrate the validity and usefulness of the proposed test. We also illustrate the power of the proposed test using sub-samples from the 2016 US birth weight data.

We then generalize the regional quantile regression rank test to censored quantile regression settings. We propose a censored version of the regression rank score using the redistribution of the probability mass for each censored observation. The model-based bootstrap algorithm is also generalized to implement the test. We illustrate the advantage of the proposed method through simulation and apply our method to study how the early environment condition influences the survival time of the bighorn sheep.

In a related study, we consider the genome-wide association study where the goal is to select genes that are associated with an outcome of interest. One major challenge for the genome-wide association study is how to handle the possible interactions between the genes and the environment. We propose a two-stage model, including one that relies on the conditional quantile levels of the outcome variables, to allow the genes to have comprehensive interactions with the environment. We use the two-stage model to study a lung cancer data set to identify new genes that can potentially influence lung cancer patients' survival time.

CHAPTER I

Introduction

It has been a classical question in statistics to study the relationship between the predictor X and the outcome Y. The least squares regression, which assumes the conditional mean of Y can be expressed as a function of X, is probably the most common tool to answer this question. Besides focusing on the conditional mean of Y, it is also useful to study how X affects the conditional quantile of Y because of two reasons. Firstly, the conditional quantiles are less influenced by the outliers than the conditional mean; secondly, when the errors are heterogeneous, the effects of X at various conditional quantiles of Y vary and are also different from its effect on the conditional mean.

Quantile regression, which was first studied in *Koenker and Bassett* (1978), can be used to study the effect of X on the conditional quantiles of Y. A linear quantile regression model can be represented as

$$y_i = x_i^T \beta(\tau) + e_{i,\tau}, \quad i = 1, 2, ...n, \tag{1.1}$$

where $x_i = (x_{i1}, x_{i2}, ...x_{ip}) \in \mathbb{R}^p$ with $x_{i1} = 1$, $\beta(\tau) = (\beta_1(\tau), ..., \beta_p(\tau)) \in \mathbb{R}^p$ and $e_{i,\tau}$ are independent errors. For identifiability, we require that at any quantile level $\tau \in (0, 1)$, the conditional τth quantile of $e_{i,\tau}$ given x_i is 0. One can assume that Model (1.1) holds locally at a specific τ or globally at any $\tau \in (0, 1)$. To ensure model validity, we require $x_i^T \beta(\tau)$ to be a monotone increasing function of τ given any x_i if Model (1.1) is assumed to hold globally.

The quantile regression estimates of $\beta(\tau)$ are obtained by

$$\hat{\beta}(\tau) = \underset{t \in \mathbb{R}^p}{argmin} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T t), \tag{1.2}$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$ is the check loss function proposed in *Koenker and Bassett* (1978). This linear optimization problem can be easily solved for all τ in (0,1) as discussed in *Koenker* (2005). *Koenker and Machado* (1999) introduced the likelihood ratio test, the Wald test and the rank test for inference in the quantile regression settings. Interested readers may refer to *Koenker* (2005) for a comprehensive introduction to quantile regression. A review of the more recent developments in quantile regression can be found in *Koenker et al.* (2017).

In this thesis, we develop new methods based on the global quantile regression model. Traditionally, to use a quantile regression, we first choose a quantile level τ and then carry out the estimation and inference at the specified τ. However, it can be challenging to choose a proper quantile level in practice. Moreover, tests may have reduced power since information at a single τ is limited. Therefore in Chapter 2, we propose a regional quantile regression rank test that allows us to detect the effect for

the covariates over a pre-specified quantile region. It can be shown that the proposed test statistic converges to a mixed chi-square distribution, and we construct a new model-based bootstrap method to estimate this distribution. The main idea for the bootstrap algorithm is to sample from the conditional quantile functions estimated under the null hypothesis but over a slightly wider region then the target quantile region. Our model-based bootstrap approximates the data generative procedure consistently and uniformly over a given region of τ, and it can also be used to approximate the null distribution for other test statistics.

In biomedical studies, it happens quite often that the responses are not fully observed for some individuals due to censoring. Censored quantile regression can be used instead of quantile regression to study this type of data. Multiples estimation procedures have been proposed in the literature for censored quantile regression, but inference methods are relatively limited. In Chapter 3, we generalize the regional quantile regression rank test to the censored setting. One major challenge in the generalization is that the regression rank score, which is used to construct the test statistics, is undefined for the censored quantile regression. To conquer this difficulty, we utilize the redistribution of mass idea and define the regression rank score for the censored version. We also generalize the model-based bootstrap method to the censored quantile regression setting to implement the test.

In Chapter 4, we consider an applied problem that is different from the previous chapters. In genome-wide association studies, the genes may have comprehensive interactions with the environment. Classical methods usually model the interactions as the product between genes and the environment, which only represents a special

form of interaction. To model the interactions more flexibly, we propose a two-stage model. In the first stage, we calculate the conditional percentile for each individual using quantile regression with all the environmental covariates. In the second stage, we select important genes by regressing the conditional quantile levels on the gene factors. The two-stage model can select genes that only have the marginal effects as well as genes that have comprehensive interactions with the environment.

CHAPTER II

Model-based Bootstrap for Detection of Regional Quantile Treatment Effects

2.1 Introduction

The detection of treatment effects is an important problem in a wide variety of applications and has been studied by many researchers under different settings. In this chapter, we focus on testing the hypothesis of no treatment effect against the alternative that the effect is significant for the upper or lower tail of the outcome distribution. There are at least two reasons why this particular class of alternatives is worth considering. Firstly, in some applications the evaluation of the treatment effect at one tail is of direct concern. For example, when financial institutions compare the risks among different portfolios, they need to focus on the lower tail of the return distribution so that they can be better prepared for the worst case scenarios. Secondly, there are cases where the treatment effect is minimal except at low or high quantile levels. In those cases any tests designed to detect mean or median

differences may have poor power. For example, it is shown later in the paper as we analyze the 2016 US birth data that maternal hypertension is a risk factor for low birth weight, and the hypertension effect on birth weight is much more obvious at the lower tail of the birth weight distribution. In such cases, a statistical test aimed at detecting the effect in the lower tail is more useful than the conventional tests on the mean treatment effects.

Quantile regression is the basis of a natural solution for the above-stated problems. A common approach is to choose a quantile level (say 0.9 quantile) and test whether the quantile regression coefficient for the treatment is significant. However, the test results may be sensitive to the choice of the individual quantile level and the test may lose power when the data are sparse around that quantile level of choice.

An improvement to individual quantile regression analysis is to consider the treatment effect over a quantile region. *He et al.* (2010) proposed a covariate-adjusted expected shortfall test (COVES), which uses quantile regression to select the observations that lie in the upper or lower quantiles and compare the covariate-adjusted means of the selected observations. COVES has been shown to be quite powerful but the test is designed for randomized trials. *Koenker* (2010) suggested an alternative test using regression rank scores over a quantile region, following the quantile rank scores proposed in *Gutenbrunner and Jurečková* (1992) and *Gutenbrunner et al.* (1993). The distribution of the test statistic under the null hypothesis is approximated by a chi-square distribution, but the chi-square approximation is only valid for i.i.d errors.

In this chapter, we consider the regional quantile regression rank test in the more

realistic case with the heterogeneous models. In this case the proposed test converges to a mixed chi-square distribution under the null hypothesis, but the mixture coefficients depend on the unknown conditional densities of the regression errors over a quantile region, whose estimates tend to be numerically unstable. An alternative way to carry out the inference is to use the bootstrap. However, commonly used bootstrap methods in regression are not directly applicable to this setting. We propose a new model-based bootstrap algorithm which aims to mimic the data generative procedure. This bootstrap algorithm enables us to generate the quantile regression model under the null hypothesis globally and to consistently estimate the null distribution of the proposed test statistic.

Applicable beyond the proposed test, our model-based bootstrap is a general bootstrap algorithm for global quantile regression analysis and is useful for a variety of settings. For example, the proposed bootstrap can be used to build the confidence band of the quantile coefficients over certain region. It can also be used in other hypothesis testing problems because the model-based structure in our bootstrap provides the flexibility to generate the desired model under the null hypothesis.

2.2 Review of quantile regression rank score

In this section we provide a brief review of the regression rank score.

Recall the linear quantile regression Model (1.1)

$$y_i = x_i^T \beta(\tau) + e_{i,\tau}, \quad i = 1, 2, ...n.$$

Letting $Q_y(\tau|x)$ be the τth quantile of y given x, we can write (1.1) equivalently as $Q_{y_i}(\tau|x_i) = x_i^T\beta(\tau)$. Therefore at the population level, for given x_i, we can express y_i as

$$y_i = x_i^T\beta(u_i), \quad u_i \sim \text{Uniform}(0,1). \tag{2.1}$$

In other words, we can view y_i as being generated from the quantile process $x_i^T\beta(u_i)$. This is an important observation for the development of our bootstrap method later in the chapter.

The quantile regression estimates of $\beta(\tau)$ are obtained by

$$\hat{\beta}(\tau) = \underset{t\in\mathbb{R}^p}{argmin} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T t), \tag{2.2}$$

which can be transformed into a dual problem

$$\hat{a}(\tau) = \underset{a\in[0,1]^n}{argmax}\{a^T y \,|\, X^T a = (1-\tau)X^T\mathbf{1}_n\}, \tag{2.3}$$

where $\hat{a}(\tau) = (\hat{a}_1(\tau), ..., \hat{a}_n(\tau))$ is an n-dimensional vector. By the duality between (2.2) and (2.3), we have

$$\hat{a}_i(\tau) = \begin{cases} 1 & y_i > x_i^T\hat{\beta}(\tau) \\ \in (0,1) & y_i = x_i^T\hat{\beta}(\tau) \\ 0 & y_i < x_i^T\hat{\beta}(\tau), \end{cases} \tag{2.4}$$

Thus $\hat{a}_i(\tau)$ is essentially an indicator whether the ith observation is above the fitted τ-quantile. Let $\hat{\tau}_i = \inf\{\tau : \hat{a}_i(\tau) < 1\}$, the ith observations should lie roughly at the

$\hat{\tau}_i$-quantile. Namely knowing $\hat{a}_i(\tau)$ for any $\tau \in (0,1)$ is equivalent to knowing the relative position of the ith observation after the covariate is adjusted for. *Gutenbrunner and Jurečková* (1992) named $\hat{a}_i(\tau)$ as the regression rank score, because $\hat{a}_i(\tau)$ can be interpreted as a generalization of ranks in the regression setting. Notice that $\hat{a}_i(\tau) - (1-\tau)$ is also an approximation of the score function of quantile regression $\Psi_\tau(u) = \tau - I(u < 0)$ evaluated at $x_i^T \hat{\beta}(\tau)$. The regression rank scores $\hat{a}_i(\tau)$ have been used to construct rank-based test in *Koenker and Machado* (1999) and *Wang* (2009) among others for the local quantile models.

In this section we are interested in detecting the treatment effect over a quantile region, and we integrate the regression rank score $\hat{a}_i(\tau)$ against an non-decreasing score function $\varphi(\cdot)$. Namely, define $\hat{b} = (\hat{b}_1, ..., \hat{b}_n)^T$ where

$$\hat{b}_i = \int_{\tau_a}^{\tau_b} \hat{a}_i(\tau) d\varphi(\tau). \tag{2.5}$$

on an interval $[\tau_a, \tau_b]$ that is specified by users. If a observation is above most quantiles over $[\tau_a, \tau_b]$ after the covariate adjustment, it is expected to have a relatively large $\hat{b}_i$.

The score function $\varphi(\cdot)$ provides flexibility in assigning different weights at different quantile levels. Two typical choices of $\varphi(\cdot)$ are:

- Wilcoxon score: $\varphi(t) = t$, which assigns weights evenly.
- Normal score: $\varphi(t) = \Phi^{-1}(t)$, which assigns more weights at upper and lower tails.

We use $\hat{b}_i$ to construct the regional quantile regression rank test statistic in the next section.

2.3 Proposed method and main results

2.3.1 Test statistic

In this section, we consider the following model

$$y_i = x_{i1}^T \beta_1(\tau) + x_{i2}^T \beta_2(\tau) + e_{i,\tau}, \quad i = 1, 2, ...n, \tag{2.6}$$

where x_{i1} is a p-dimensional vector, x_{i2} is a q-dimensional vector. The error $e_{i,\tau}$ are assumed to be independent but not necessarily identically distributed with the natural constraint that $Q_{e_{i,\tau}}(\tau|x_{i1,}, x_{i2}) = 0$. We assume the model holds globally at any $\tau \in (0, 1)$ since our goal is to detect the treatment effect over a region of τ.

We are interested in testing the hypothesis

$$H_0\text{: } \beta_2(\tau) = 0 \ \forall \tau \in (0,1) \ \text{ vs } \ H_1\text{: } \beta_2(\tau) \neq 0 \text{ for } \tau \in [\tau_a, \tau_b],$$

where $[\tau_a, \tau_b]$ is the user-specified subset of (0,1) and should be chosen to target the region of interest.

For convenience, write the design matrix of (2.6) as $X = [X_1, X_2]$. Let $\hat{X}_2 = X_1(X_1^T X_1)^{-1} X_1^T X_2$, which is the projection of X_2 into the space spanned by the columns of X_1. If we fit the quantile regression with only X_1 as the explanatory variable, $\hat{b}$ calculated under this null model represents the ranks after adjusting for X_1. If the null hypothesis is true, $X_2 - \hat{X}_2$ is expected to be orthogonal to $\hat{b}$

asymptotically, since no variations in $\hat{b}$ can be further explained by $X_2 - \hat{X}_2$. To help understand this orthogonality, we recall that the residuals are orthogonal to the design matrix in the least squares regression. For the quantile regression $\hat{a}_i(\cdot)$ plays similar roles as the residuals and can be shown to be orthogonal to design variables used in the quantile regression. A rigorous argument follows from Lemma 2 and Equation (2.15) of in Section 2.6.

Our test statistic will be constructed based on the above observation. But instead of using the integral version of $\hat{b}$ defined in (2.5), we will employ a grid of points in τ and replace $\hat{b}$ with a weighted sum. More precisely, consider a set of $M+1$ ordered and evenly spaced grid points

$$S = (\tau_0, \tau_1, ..., \tau_M), \tag{2.7}$$

where $[\tau_a, \tau_b]$ is a proper subset of $[\tau_0, \tau_M]$. With S and a differentiable score function $\varphi(\cdot)$ specified, we define

$$\tilde{b}_i = \sum_{\tau_m \in S \cap [\tau_a, \tau_b]} \hat{a}_i(\tau_m)\varphi'(\tau_m)(\tau_m - \tau_{m-1}), \tag{2.8}$$

where $\hat{a}_i$ is given in (2.3) and calculated under the null model.

The employment of these grid points in calculating $\tilde{b}$ is mainly to facilitate the bootstrap used later. Since $\hat{a}_i(\tau)$ is a piecewise linear function with $O(n \log n)$ break points (*Portnoy* (1991)), $\hat{b}_i$ defined in (2.5) can be written as a sum of $O(n \log n)$ terms, and $\tilde{b}_i$ is an approximation of $\hat{b}_i$ with a sum of roughly $M+1$ terms.

It is worth pointing out that only $\hat{a}_i$ evaluated at grid points within $[\tau_a, \tau_b]$ are

used in calculating (2.8) to focus on our region of interest $[\tau_a, \tau_b]$. But the grid points need to be defined on $[\tau_0, \tau_M]$, which is strictly larger than $[\tau_a, \tau_b]$. To get reliable estimation of $\hat{a}(\cdot)$ at the end points using the bootstrap, $\beta(\cdot)$ should be estimated accurately over a slightly larger quantile region.

Now we define our proposed test statistic as

$$T_n = S_n^T Q_n^{-1} S_n, \tag{2.9}$$

where

$$S_n = n^{-1/2}(X_2 - \hat{X}_2)^T \tilde{b},$$
$$Q_n = n^{-1}(X_2 - \hat{X}_2)^T (X_2 - \hat{X}_2).$$

A larger value of T_n will be in favor of the alternative hypothesis. We shall show in Section 2.6 that under some regularity assumptions, S_n converges to a zero mean normal distribution with variance Σ taking the form

$$\Sigma = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sum_{\tau_m \in S} c_{\tau_m} (x_{2i} - \hat{x}_{2i} - K_n^{\tau_m} x_{1i})(x_{2i} - \hat{x}_{2i} - K_n^{\tau_m} x_{1i})^T, \tag{2.10}$$

where c_{τ_m} is a constant depending on $\varphi(\cdot)$ and $K_n^{\tau_m}$ is a matrix involves the conditional densities of y_i given x_i evaluated at τ_m-quantile. In principle, we could estimate the densities using kernel or spline methods. However, the results are often numerically unstable. Thus instead of estimating this covariance matrix to standardize the test statistic, we will use the bootstrap as our preferred approach.

The matrix Q_n can be viewed as an approximate standardization because it can

be shown that Σ is equal to Q_n times a constant when the model is homogeneous. With the usage of Q_n, T_n will behave closer to a standard chi-square distribution asymptotically and the resulting test may have better power when the model is close to homogeneous. In theory, many choices of Q_n would work, but the specific choice used here is consistent with the common choice for the quantile regression rank tests.

In the proposed test statistics T_n, the quantities M, $\varphi(\cdot)$ and $[\tau_a, \tau_b]$ need to be specified by the users. Therefore a discussion of how to choose them are in order.

1. Choice of M: The number of grid points M should be between the order of $n^{1/4}$ and of $n^{1/2}$ for our theory to work. But in practice, the choice of M does not have notable influence on the result as long as M is not too extreme. For example, we find that 50 or 100 can be a suitable choice for M for a wide range of problems.

2. Choice of $\varphi(\cdot)$: The score function $\varphi(\cdot)$ may influence the power of the test. *Koenker* (2010) showed how the optimal score function can be selected under the simpler model with i.i.d errors, if the error density is known. Since the density is unknown in practise and moreover we allow heterogeneity, it is unrealistic to aim for an optimal score function. We compared the power of our test with the most commonly used Wilcoxon score and Normal score under a variety of settings by simulation and the differences are not major. We therefore recommend using the Wilcoxon score for simplicity.

3. Choice of $[\tau_a, \tau_b]$: The quantile region $[\tau_a, \tau_b]$ should be used to target the region of interest, such as the lower tail of birth-weight or the upper tail of the loss

from an investment portfolio. In the typical quantile regression settings, we usually choose a value τ, whether a specific value of τ is better than another nearby value of τ is difficult to answer. The choice of one interval over another has the same question around it. But from the numerical results in Section 2.4 and 2.5, we note that the power of our test is shown to be stable over a range of reasonable choices of $[\tau_a, \tau_b]$. In other words, choosing a specific value of τ in the analysis is associated with less robust analysis results than choosing an interval $[\tau_a, \tau_b]$.

2.3.2 Model-based bootstrap

In this subsection, we propose a model-based bootstrap method to approximate the distribution of T_n under the null hypothesis.

There are quite a few established bootstrap methods under the quantile regression setting. The paired bootstrap, the generalized bootstrap (*Chatterjee and Bose* (2005)) and the wild bootstrap (*Feng et al.* (2011)) are examples of those methods that have been implemented in the R package *quantreg*. However these methods cannot be directly applied here.

The paired bootstrap does not generate bootstrap samples under H_0 when the data are not from the null model. The same goes with the generalized bootstrap. One possible solution is to keep x_{i2} unchanged and sample (y_i^*, x_{i1}^*) with replacement from (y_i, x_{i1}). The resulting bootstrap data set would be $(y_i^*, x_{i1}^*, x_{i2})$. But the correlation between x_{i1} and x_{i2} can not be preserved under such a subsampling scheme.

The wild bootstrap uses the coefficients $\hat{\beta}_1(\tau)$ and residuals $\hat{e}_{i,\tau}$ obtained from the

τth quantile regression fitted under H_0. The bootstrap data set will be (y_i^*, x_{i1}, x_{i2}) where $y_i^* = x_{i1}^T\hat{\beta}(\tau) + w_i|\hat{e}_{i,\tau}|$, and w_i is generated independently from a specially designed distribution to make sure the bootstrap is consistent at the τ-quantile. The wild bootstrap is useful for inference at a single quantile level. Since our test statistic consists of estimation from multiple quantiles, no weight distribution would work in this framework.

We propose a new bootstrap scheme that generates data globally under H_0. The key idea is that as shown in (2.1), we can write our linear quantile regression model equivalently as $y_i = x_i^T\beta(u_i)$, where $u_i \sim \text{Uniform}(0,1)$. We keep x_{i1} and x_{i2} fixed and generate bootstrap samples y_i^* from $x_{i1}^T\hat{\beta}_1(u_i)$. Namely, we view $x_{i1}^T\hat{\beta}_1(\cdot)$ as a quantile process for the bootstrap, where $\hat{\beta}_1(\cdot)$ is estimated under the null model.

Although the quantile function $x_i^T\beta(\cdot)$ is monotonously increasing at any x_i, the estimate $x^T\hat{\beta}(\cdot)$ is only guaranteed to be monotone at $x = \bar{x}$. Thus $x_{i1}^T\hat{\beta}_1(\cdot)$ may not be a valid quantile process. This is the reason why we introduce the set of grid points S defined in (2.7). Let $\tilde{\beta}_1(\tau)$ be the linear interpolation of $\{\hat{\beta}_1(\tau_m), m \in S\}$. *Neocleous and Portnoy* (2008) showed that when M increases in the order between $n^{1/4}$ and $n^{1/2}$, the probability that $x_{i1}^T\tilde{\beta}_1(\cdot)$ is monotonously increasing will converge to 1. At the same time, $\tilde{\beta}_1(\cdot)$ is a good enough approximation to $\hat{\beta}_1(\cdot)$. Thus we propose to generate y_i^* from an asymptotically valid quantile process $x_{i1}^T\tilde{\beta}_1(u_i)$. The detailed algorithm of this model-based bootstrap method is given as follows:

Step 1: Fit the linear quantile regression under H_0 and obtain the estimator $\hat{\beta}_1(\tau)$ for $\tau \in S \cap [\tau_0, \tau_M]$. Calculate T_n using (2.9).

Step 2: Let $\tilde{\beta}_1(\tau)$ be the linear interpolation of $\{\hat{\beta}_1(\tau_m), m \in S\}$.

Namely $\tilde{\beta}_1(\tau) = \frac{\tau_{m+1}-\tau}{\tau_{m+1}-\tau_m}\hat{\beta}_1(\tau_m) + \frac{\tau-\tau_m}{\tau_{m+1}-\tau_m}\hat{\beta}_1(\tau_{m+1})$ when $\tau_m < \tau < \tau_{m+1}$, $m = 0, ..., M-1$. Let $\tilde{\beta}_1(\tau) = \hat{\beta}_1(\tau_0)$ for $\tau < \tau_0$, and $\tilde{\beta}_1(\tau) = \hat{\beta}_1(\tau_M)$ for $\tau > \tau_M$.

Step 3: For $i = 1, ..., n$, generate $u_i \sim$ Uniform$(0, 1)$ independently, and then construct a bootstrap sample (y_i^*, x_{i1}, x_{i2}), where $y_i^* = x_{i1}^T\tilde{\beta}_1(u_i)$.

Step 4: Calculate T_n^* from (2.9) with the bootstrap sample.

Step 5: Repeat Steps 3 and 4 for B times to get $\{T_{n1}^*, T_{n2}^*, ..., T_{nB}^*\}$, where B is a prespecified integer. The resulting p-value is calculated by $B^{-1}\sum_b \mathbb{I}(T_n > T_{nb}^*)$.

The model-based bootstrap can be used for other forms of test statistics. For example, the same bootstrap method can be used to approximate the distribution of $\sup_{\tau\in S\cap[\tau_a,\tau_b]}|\tilde{\beta}_2(\tau)|$ under H_0, which may also be used as a test statistic for regional treatment effect detection. We will discuss this supremum-based test in more detail in Section 2.4.

2.3.3 Asymptotic properties

Let f_i be the density of y_i given x_i. To study the asymptotic properties of the proposed test, we impose the following regularity conditions:

(A1) $\max_i \| x_i \| \leq L$, where L is a positive constant and $\|\cdot\|$ denotes the L^2 norm.

(A2) The densities f_i are bounded away from 0 and infinity at $x_i^T\beta(\tau)$ uniformly for i and $\tau \in [\tau_0, \tau_M]$, where $0 < \tau_0 < \tau_a$ and $\tau_b < \tau_M < 1$. Furthermore, $|f_i(c_1) - f_i(c_2)| = O(|c_1 - c_2|)$ uniformly in i as $|c_1 - c_2| \to 0$.

(A3) The limits $Q := \lim_{n\to\infty} \frac{1}{n} \sum x_i x_i^T$ and $D_x^\tau := \lim_{n\to\infty} \frac{1}{n} \sum f_i(x_i^T \beta(\tau)) x_i x_i^T$ exist, and are positive definite at any $\tau \in [\tau_0, \tau_M]$.

(A4) $\varphi(\cdot)$ is a nondecreasing differentiable function with bounded variation.

(A5) $S = (\tau_0, \tau_1, ..., \tau_M)$ is a set of ordered and evenly spaced grid points where $n^{1/4} \ll M \ll n^{1/2}$.

The regularity conditions are stated under fixed designs. When x_i is a random variable, all the calculations can be carried out conditioning on x_i. Replacing (A1) and (A3) with corresponding moment conditions, our results still hold for random designs as well.

Condition (A1) assumes that the covariate space lies within a compact set. This assumption is necessitated by heterogeneity because if the quantile regression model is linear over an unbounded set of x at multiple τ values, the quantile functions $x^T\beta(\tau_1)$ and $x^T\beta(\tau_2)$ may cross unless they are vertical shifts. (A2) and (A3) are common sufficient conditions used to establish the uniform Bahadur representation for the quantile regression estimates. We restrict our attention to $[\tau_0, \tau_M]$ instead of the whole interval $(0, 1)$. To study the asymptotic behavior of $\hat{\beta}(\tau)$ as τ approaches 0 or 1 requires much stronger assumptions on f_i. And for our study, we need to work on a set slightly larger than our region of interest $[\tau_a, \tau_b]$, which can be chosen to be a compact subset of $(0, 1)$.

Theorem 2.1: With regularity conditions (A1)-(A4), we have under H_0,

(i) $T_n \Rightarrow \bar{\chi}^2$, a mixed chi-square distribution as a weighted sum of q chi-square variables of one degree of freedom.

Further assume (A5) holds, then

(ii) The bootstrap estimator $\hat{\beta}_1^*(\tau)$ is a consistent estimator of $\beta_1(\tau)$ uniformly for $\tau \in S \cap [\tau_a, \tau_b]$.

(iii) Given the data, the conditional distribution of T_n^* will converge to the same mixed chi-square distribution $\bar{\chi}^2$.

Theorem 2.1(i) shows that our test statistic will converge to a mixed chi-square distribution under H_0 while Theorem 2.1(ii) and 2.1(iii) show that the conditional bootstrap distribution approximates to the same mixed chi-square distribution. Hence our model-based bootstrap is consistent for inference. The proof of these results relies on the empirical process theory and is given in Section 2.6.

2.4 Simulation

In this section, we present some empirical results of our proposed test by Monte Carlo simulations.

2.4.1 Settings

The number of replications in each simulation and the bootstrap replication size are both set to 1000 throughout this section. We first generated our data from the following model that was considered in *He et al.* (2010),

$$y_i = 5 + x_{i1} + x_{i2} + \big(1 + \gamma \mathbb{I}(e_i > 0)\mathbb{I}(d_i = 0)\big)e_i, \quad i = 1, ..., m + n, \tag{2.11}$$

where the treatment indicator $d_i = 1$ for $i = 1, ..., m$ and $d_i = 0$ for $i = m+1, ..., m+n$. Let $\gamma = 0$ under H_0 and $\gamma = 1.35$ under H_1. We consider testing whether the coefficient of the treatment indicator γ is zero. By design, the treatment effect only exists in the upper tail under the alternative. We considered the following three different settings based on model (2.11):

(i) $x_{i1} \sim$ Uniform$(5, 12)$, $x_{i2} \sim N(8, 8)$ and $e_i \sim N(0, 5)$, and they are mutually independent. This represents a randomized trial with $i.i.d$ errors.

(ii) $x_{i1} \sim$ Uniform$(5, 12)$ when $d_i = 1$, but $x_{i1} \sim$ Uniform$(5, 20)$ when $d_i = 0$. In addition, $x_{i2} \sim N(8, 8)$ and $e_i \sim N(0, x_{i1})$ are independently generated. This represents a non-randomized trial with heterogeneous errors.

(iii) $x_{i1} \sim$ Uniform$(5, 12)$ when $d_i = 1$; otherwise x_{i1} is generated from the t distribution truncated to [0, 250] with 2 degrees of freedom and non-centrality parameter equal to 15. The variables $\{x_{i2}\}$ and $\{e_i\}$ are generated from the same distributions as (ii). Compared to (ii), x_{i1} is generated from a distribution with heavier tails.

Under these settings, $\{x_{i2}\}$ is generated from a normal distribution, which violates (A1) that $\{x_i\}$ should lie in a compact set. However, since the coefficient of x_{i2} is a constant of τ in these settings, we still have valid quantile functions even when the range of x_{i2} extends to the whole line, so our theory applies to the model with trivial modifications.

In addition, we also evaluated the performance of the proposed method when the effect of multiple covariates are simultaneously tested in the following model:

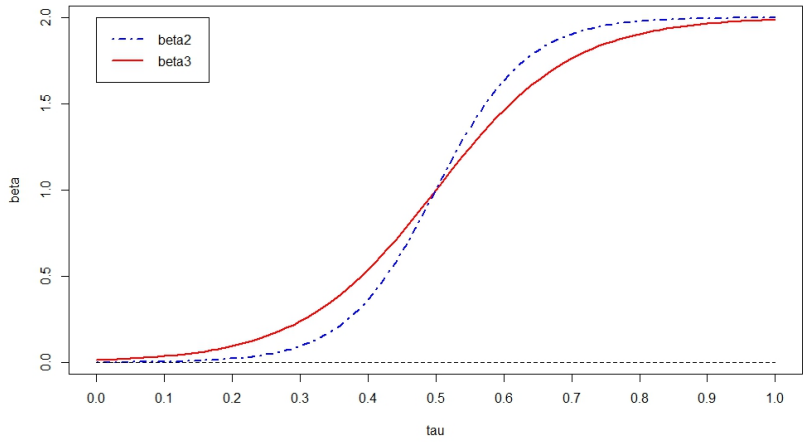

Figure 2.1: Curves of quantile coefficients of model (iv) in simulation under the alternative.

(iv)

$$y_i = \beta_0(u_i) + x_{i1}\beta_1(u_i) + x_{i2}\beta_2(u_i) + x_{i3}\beta_3(u_i), \quad i = 1, ..., n, \tag{2.12}$$

where $u_i \sim \text{Uniform}(0,1)$, $x_{i2} \sim \text{Uniform}(0,2)$, $x_{i3} \sim \text{Uniform}(0,2)$, and $x_{i1} \sim \text{Uniform}(1,3)$ when $x_{i2} < 1$ but $x_{i1} \sim \text{Uniform}(0,2)$ when $x_{i2} \geq 1$. Furthermore, let $\beta_0(\tau) = \Phi^{-1}(\tau)$, $\beta_1(\tau) = \tau^2$. Under H_0, we use $\beta_2(\tau) = \beta_3(\tau) = 0$. Under H_1, we use $\beta_2(\tau) = \frac{\exp(15(\tau-0.5))}{1+\exp(15(\tau-0.5))}$ and $\beta_3(\tau) = \frac{\exp(10(\tau-0.5))}{1+\exp(10(\tau-0.5))}$. As shown in Figure 2.1, the effect of x_{i1} and x_{i2} are larger at the upper tail under the alternative by design.

We consider the problem of testing the null hypothesis H_0: $\beta_2(\tau) = \beta_3(\tau) = 0, \forall \tau \in (0,1)$, but the test will focus on upper quantiles.

We first compared the proposed regional quantile regression rank (RQRR) test

with the quantile regression rank (QRR) test that focuses on one fixed quantile proposed by *Koenker and Machado* (1999) to see if we can benefit from considering a quantile region. To show the necessity of the proposed bootstrap method, we also considered the proposed QRRQ test statistic with the critical value approximated by the chi-square distribution based on the working assumption of $i.i.d$ errors.

When $q = 1$, we further compared the performance of our test to other three methods that focus on the overall treatment effect: the COVES test proposed by *He et al.* (2010); the test based on simultaneous confidence band; the supremum-based test. The latter two methods are described as follows.

To build simultaneous confidence bands, we use a method similar to what is considered in *Chernozhukov and Fernández-Val* (2004). A level $1 - \alpha$ confidence band of $\beta_2(\cdot)$ over $[\tau_a, \tau_b]$ can be built based on the statistic

$$T_n^{sup} = \sup_{\tau \in [\tau_a, \tau_b]} |\sqrt{n}\tilde{\beta}_2(\tau)|,$$

where $\tilde{\beta}_2(\tau)$ is the linear interpolation of the coefficient estimate $\hat{\beta}_2(\tau)$. The distribution of $\tilde{\beta}_2(\tau)$ is approximated by the m out of n bootstrap, where $m = 20 + n^{1/2}$. The null hypothesis is rejected if 0 is contained nowhere in the confidence band.

To carry out the supremum-based test, we use the model-based bootstrap scheme introduced in Section 2.3.2 to generate the bootstrap sample (y_i^*, x_{1i}, x_{2i}) where $y_i^* = x_{i1}^T \tilde{\beta}_1(u_i)$, with $\tilde{\beta}_1(\cdot)$ estimated under the restricted model. The bootstrap test statistics $T_n^{sup,*} = \sup_{\tau \in [\tau_a, \tau_b]} |\sqrt{n}\tilde{\beta}_2^*(\tau)|$ can then be obtained from the bootstrap sample. The null distribution of T_n^{sup} is approximated by the empirical distribution

of $T_n^{sup,*}$.

Both the test based on simultaneous confidence bands and the supremum-based test utilize the test statistics T_n^{sup}. The difference is that the bootstrap is conducted under the full model to build the confidence band while the bootstrap is conducted under the null hypothesis for the supremum-based test. Also notice that the difference between the supremum-based test and the RQRR test lies in the test statistics. Thus comparing these two tests is basically comparing the performance of a supremum-based statistic versus a rank-based statistic under our settings.

2.4.2 Results

We first set the quantile region to be $[0.7, 0.99]$ and $[0.85, 0.99]$ to compare the performance of the proposed RQRR test with the QRR test at one quantile level, the RQRR test with the chi-square approximation, and the COVES test. The results are summarized in Table 2.1.

For the randomized trail we considered in model (i), all the tests have reasonable type I error rates. For model (ii), both the COVES and the RQRR with chi-square approximations are not valid theoretically. According to our simulation results, COVES fails to control the type I errors, and the RQRR test with chi-square approximations is acceptable. This is actually consistent with our knowledge that the RQRR with chi-square approximations is reasonably robust under heterogeneity (*Kocherginsky et al.* (2005)). For the more extreme example where x_{i1} has a heavy right tail in model (iii), however, it is obvious that both COVES and the RQRR with chi-square approximations are not valid anymore, while our proposed RQRR has empirical

	Model (i)				**Model (ii)**			
	$m = n = 50$		$m = n = 100$		$m = n = 50$		$m = n = 100$	
	α level	power	α level	power	α level	power	α level	power
QRR(0.70)	0.050	0.327	0.047	0.586	0.041	0.273	0.052	0.524
QRR(0.80)	0.050	0.567	0.046	0.889	0.047	0.493	0.037	0.797
QRR(0.85)	0.041	0.682	0.046	0.953	0.041	0.576	0.049	0.889
QRR(0.90)	0.035	0.723	0.042	0.983	0.038	0.622	0.043	0.926
QRR(0.95)	0.036	0.431	0.025	0.974	0.024	0.488	0.034	0.891
COVES	0.066	0.909	0.052	0.998	0.070	0.917	0.088	0.999
RQRR(0.70,0.99)	0.043	0.712	0.048	0.960	0.045	0.628	0.046	0.917
RQRR_{b}(0.70,0.99)	0.047	0.707	0.049	0.956	0.047	0.627	0.046	0.908
RQRR(0.85,0.99)	0.041	0.816	0.044	0.995	0.041	0.702	0.046	0.970
RQRR_{b}(0.85,0.99)	0.045	0.834	0.046	0.994	0.048	0.725	0.051	0.972
	Model (iii)				**Model (iv)**			
	$m = n = 100$		$m = n = 300$		$n = 100$		$n = 200$	
	α level	power	α level	power	α level	power	α level	power
QRR(0.70)	0.064	0.162	0.048	0.338	0.045	0.356	0.048	0.657
QRR(0.80)	0.058	0.290	0.065	0.571	0.042	0.494	0.040	0.844
QRR(0.85)	0.062	0.358	0.064	0.675	0.042	0.532	0.043	0.892
QRR(0.90)	0.053	0.428	0.074	0.752	0.041	0.537	0.040	0.897
QRR(0.95)	0.046	0.438	0.057	0.789	0.034	0.384	0.037	0.784
COVES	0.522	0.943	0.915	1.000	NA	NA	NA	NA
RQRR(0.70,0.99)	0.115	0.557	0.148	0.868	0.042	0.619	0.047	0.930
RQRR_{b}(0.70,0.99)	0.056	0.381	0.058	0.724	0.048	0.629	0.047	0.930
RQRR(0.85,0.99)	0.102	0.678	0.128	0.930	0.049	0.625	0.040	0.944
RQRR_{b}(0.85,0.99)	0.061	0.551	0.063	0.853	0.052	0.657	0.040	0.947

Table 2.1: Comparison of the empirical type I error rate and the power out of 1000 simulation samples. In the table, QRR(τ) stands for the quantile regression rank test conducted at the τth quantile proposed in *Koenker and Machado* (1999); RQRR(τ_a,τ_b) stands for the regional quantile regression rank test with chi-square approximations at the quantile region $[\tau_a,\tau_b]$, while $\text{RQRR}_{\text{b}}(\tau_a,\tau_b)$ stands for the proposed regional quantile regression rank test with the model-based bootstrap. For COVES, the cutoff quantile level is set to be 0.75 as in *He et al.* (2010).

type I errors close to the nominal level. The results from model (iv) show that our proposed test also has satisfactory performance when testing the effect of multiple continuous covariates simultaneously. Overall our proposed test works more broadly than the COVES and the RQRR with the chi-square approximation. In particular, the proposed test remains valid under heterogeneous cases where the other two tests may fail.

Table 2.1 also illustrates the advantages of the proposed RQRR over the QRR test at one quantile level in terms of power stability. Firstly, we observe that the empirical power of the QRR test heavily depends on the choice of τ. Its power tends to increase as τ increases to some value because the magnitude of treatment effect also increases. But the power will decrease if we further increase τ due to the inflation in variance. On the other hand, it is quite obvious from our results that the proposed RQRR test is less sensitive to the choice of the quantile region. Secondly, the proposed RQRR test with a reasonably-chosen quantile interval is more powerful than the QRR test at many individual quantile levels. For example, the power for the proposed RQRR test with quantile region $[0.85, 0.99]$ is higher than the QRR test with $\tau = 0.90$ for all the settings we considered. Therefore we can benefit from utilizing the extra information provided over a quantile region to achieve more stable statistical power.

We then set the quantile region to be $[0.7, 0.95]$, $[0.85, 0.95]$, $[0.7, 0.99]$ and $[0.85, 0.99]$ to compare the performance of the proposed RQRR test with the methods based on simultaneous confidence bands and the supreme-based test. For this comparison, we only present the results under model (ii) in Table 2.2. The results under model (i)

	Model (ii)			
	$m = n = 50$		$m = n = 100$	
	α	power	α	power
RQRR_b(0.70,0.95)	0.047	0.571	0.042	0.895
RQRR_b(0.85,0.95)	0.043	0.701	0.051	0.955
RQRR_b(0.70,0.99)	0.047	0.627	0.046	0.908
RQRR_b(0.85,0.99)	0.048	0.725	0.051	0.972
CF(0.70,0.95)	0.015	0.399	0.018	0.767
CF(0.85,0.95)	0.021	0.448	0.021	0.785
CF(0.70,0.99)	0.071	0.516	0.083	0.810
CF(0.85,0.99)	0.074	0.530	0.084	0.811
Max(0.70,0.95)	0.052	0.605	0.045	0.924
Max(0.85,0.95)	0.057	0.607	0.049	0.923
Max(0.70,0.99)	0.090	0.546	0.147	0.871
Max(0.85,0.99)	0.092	0.547	0.150	0.871

Table 2.2: Comparison of the empirical type I error rate and power out of 1000 simulation samples for the tests based on simultaneous confidence bands (CF), the supremum-based test (Max), and the RQRR test over $[\tau_a, \tau_b]$.

and (iii) tell a similar story.

From Table 2.2, we notice that the methods based on simultaneous confidence bands and the supremum-based test do not control the type I error well when we set the upper quantile level to be 0.99. This is because the estimation of the coefficients are unreliable when τ is close to one for data with moderate sample sizes. The RQRR test can be roughly seen as analyzing the average treatment effect over the quantile region, so it is able to handle relatively extreme tails better.

2.5 Data analysis

2.5.1 The birth weight data

In this subsection, we illustrate the power of the proposed RQRR test with the 2016 US birth weight data. Because the size of the full data is large, we were able to conduct the proposed RQRR test and the QRR test at one quantile level with sub-samples of the full data set and compare their number of rejections.

The 2016 US birth weight data set is produced by the National Center for Health Statistics and is available to the public online[1]. The data set contains the infant and maternal health characteristics along with paternal demographic information of the births occurred in the US during 2016. In particular, we restricted our focus to 32,169 white mothers whose ages are between 36 and 40 and we aimed to study the relationship between birth weight and the maternal history of hypertension. Besides the indicator for maternal hypertension, mothers' education level, mothers' weight before delivery and indicator for smoking during pregnancy were included as confounding variables. Notice that these variables were also considered in the birthweight data collected at Baystate Medical Center, Springfield, Mass during 1986 (*Hosmer and Lemeshow* (2010)).

We first fitted linear quantile regressions with the full data set at different τ. From Figure 2.2, we can see that the coefficient of hypertension is significantly less than zero at all the quantile levels and the hypertension effect decreases in magnitude as the quantile level increases. Namely hypertension has a negative effect on birth weight and its effect is more severe at the lower tail. Given the size of the data, we

[1]The data set is available for download at *www.cdc.gov/nchs/data_access/vitalstatsonline.htm*

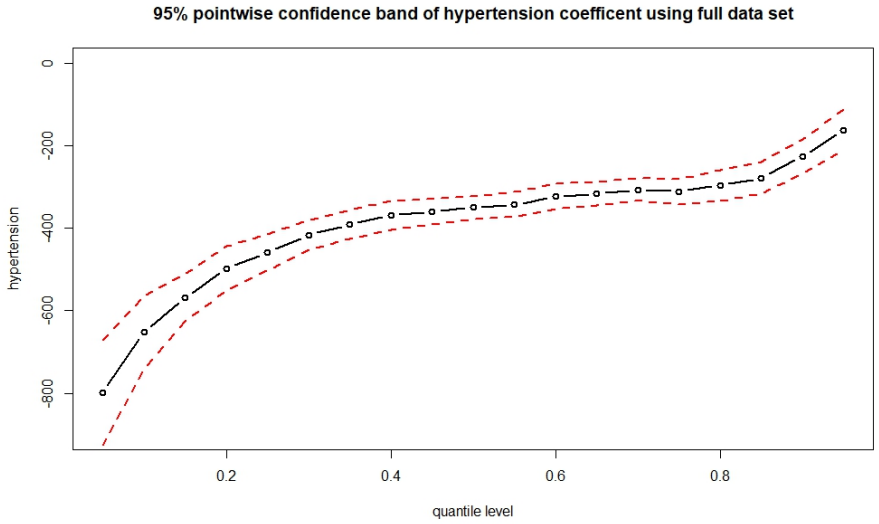

Figure 2.2: 95% pointwise confidence band of the hypertension effect.

view the full data estimates as good proxies to the true parameters.

In the next step, we subtracted the median hypertension effect estimated with the full data from the birth weight to check whether the quantile coefficient varies with τ. To compare the performance of the proposed RQRR test with the QRR test at on quantile level when we have limited sample sizes, we sub-sampled ($n = 200 - 800$) from the full data and compared the number of rejections out of 500 sub-sampled data sets.

The results are summarized in Table 2.3, which is consistent to what we observed in the simulations. The power of both the proposed RQRR test and the QRR test depends on the choice of quantile interval/level. Though the treatment of hypertension becomes more significant at lower tails, the QRR test will suffer from low power if we choose τ to be as small as 0.01 due to higher variances in those quantile esti-

test	counts of rejection at 0.05 level		
	$n = 200$	$n = 400$	$n = 800$
$\text{RQRR}_{\text{b}}(0.01,0.10)$	103	150	223
$\text{RQRR}_{\text{b}}(0.01,0.15)$	94	141	216
$\text{RQRR}_{\text{b}}(0.05,0.25)$	72	92	171
QRR(0.01)	44	54	76
QRR(0.05)	81	120	194
QRR(0.10)	69	108	185
QRR(0.20)	40	65	104
Least squares	50	63	69

Table 2.3: Number of rejections out of 500 sub-sampled birthweight data sets.

mates. For the RQRR test, the intervals $[0.01, 0.1]$ and $[0.01, 0.15]$ are better choices compared to the interval $[0.05, 0.25]$. Comparing the RQRR test with the QRR test, the former is less sensitive to the choice of the quantile interval/level. Even when the quantile interval is chosen to be [0.05,0.25], the power of the RQRR test is not much worse. When both the quantile interval and the quantile level are reasonably chosen, the RQRR test tends to perform better than the QRR test in general. The least squares regression is included in the comparison, and its power is clearly lower than the RQRR test, because it aims to detect the difference in the mean, which is less obvious than in the lower tail of the birthweight distribution.

2.5.2 S&P 500 index

In this subsection, we looked at the S&P 500 index data to test if there exist any differences in the risk of investing in different sectors. We collected the S&P 500 index of the financial, energy and information technology sectors from January 2, 2015 to January 26, 2018, which has a total of 773 data points; see Figures 2.3 for

the time series plots. Let x_t be the index at time t, the return r_t is calculated by $100 \log \frac{x_t}{x_{t-1}}$.

We first compared the financial sector with the energy sector using the following AR-like model

$$r_t = \beta_0(\tau) + \beta_1(\tau) r_{t-1} + \beta_2(\tau) \mathbb{I}_{energy} + \beta_3(\tau) r_{t-1} \mathbb{I}_{energy} + g(r_{t-1}) e_{t,\tau},$$

where $\mathbb{I}_{energy}$ is the indicator for the energy sector. We assume that $\{e_{t,\tau}\}$ is independent over t so that the proposed RQRR is still valid for this time series data. Writing the error term in the form of $g(r_{t-1})e_{t,\tau}$ allows heterogeneity. We are interested in testing

$$H_0\text{: } \beta_2(\tau) = \beta_3(\tau) = 0 \ \forall \tau \in (0,1) \text{ vs}$$
$$H_1\text{: } \beta_2(\tau) \neq 0 \text{ or } \beta_3(\tau) \neq 0 \text{ for some } \tau.$$

We focused on the lower tail under the alternative since the occurrences of large negative returns are the risk we are concerned with. The proposed RQRR and the QRR at multiple quantile intervals/levels are conducted.

	test statistics	p-value
$\text{RQRR}_\text{b}(0.01,0.05)$	0.0006	0.001
$\text{RQRR}_\text{b}(0.01,0.10)$	0.0035	0.005
$\text{RQRR}_\text{b}(0.05,0.25)$	0.0578	0.000
QRR(0.01)	3.332	0.036
QRR(0.05)	4.445	0.012
QRR(0.10)	3.014	0.049
QRR(0.20)	7.111	0.001

Table 2.4: Comparing risks (low returns) between the financial and the energy sectors.

According to the results summarized in Table 2.4, we are able to reject the null hypothesis and claim that the risk level of the financial and the energy sectors are different with all the tests at 0.05 level, but not always at the 0.01 level. Also notice that the p-values of the RQRR are consistently smaller than the QRR, which may indicate that the RQRR is more powerful.

We also compared the financial sector with the information technology sector using a similar model

$$r_t = \beta_0(\tau) + \beta_1(\tau) r_{t-1} + \beta_2(\tau)\mathbb{I}_{info} + \beta_3(\tau) r_{t-1}\mathbb{I}_{info} + g(r_{t-1})e_{t,\tau}.$$

From the results in Table 2.5, the tests fail to reject the null hypothesis at 0.05 level. Thus the risk between the financial and the information technology sectors can be

quite similar.

Note that in this analysis, the risk refers to the potential losses in the daily returns of each sector given the previous day's return, in the spirit of *Engle and Manganelli* (2004). The quantification of risk conditional on the recent past supplements the common risk measures on the marginal return distributions such as the Value-at-Risk measures. In fact, the one day 5% Value-at-Risk for the financial, information and energy sectors are 1.62, 1.61 and 2.04, respectively.

	test statistics	p-value
$\text{RQRR}_{\text{b}}(0.01,0.05)$	10^{-7}	0.999
$\text{RQRR}_{\text{b}}(0.01,0.10)$	0.0004	0.586
$\text{RQRR}_{\text{b}}(0.05,0.25)$	0.0143	0.194
QRR(0.01)	0.4576	0.633
QRR(0.05)	0.0333	0.967
QRR(0.10)	2.3979	0.091
QRR(0.20)	1.9106	0.148

Table 2.5: Testing risk between finance and information technology sectors.

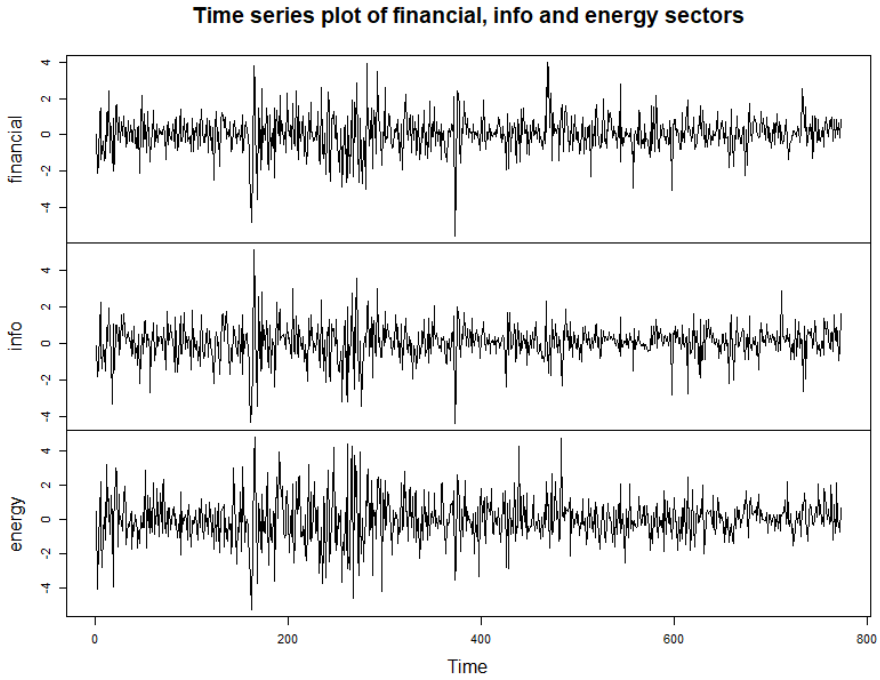

Figure 2.3: Time series plot of the financial,info and the energy sectors.

2.6 Proof

In this section, we present the proof of Theorem 2.1.

We first study the limiting distribution of the test statistic T_n as shown in Theorem 1(i). Let $Z_n = \sqrt{n}(\hat{\beta}(\tau) - \beta(\tau))$. Let d_i be a d-dimensional vector that is uniformly bounded and write $D_{nd}^{\tau} = \frac{1}{n}\sum f_i(x_i^T\beta(\tau))d_i x_i^T$. We assume that the limit of D_{nd}^{τ} exists and is positive definite. We plug in $d_i = x_i$ and $d_i = \hat{x}_{i2} - x_{i2}$ in latter part of the proof. Notice that by conditions (A1) and (A3), the assumptions we made on d_i will be satisfied when $d_i = x_i$ and $d_i = \hat{x}_{i2} - x_{i2}$.

Let $\hat{G}_n^d(t) = n^{-1/2}\sum d_i \mathbb{I}(y_i \le x_i^T t)$ and $G_n^d(t) = n^{-1/2}\sum d_i F_i(x_i^T t)$. Lemma 1 shows that $G_n^d(t)$ is a good approximation of $\hat{G}_n^d(t)$ using results from empirical process theory. Furthermore define $\hat{W}_n^d = n^{-1/2}\sum d_i(\hat{a}_i(\tau) - (1-\tau))$ and $W_n^d = n^{-1/2}\sum d_i(\tilde{a}_i(\tau) - (1-\tau))$ where $\tilde{a}_i(\tau) = \mathbb{I}(y_i \ge x_i^T\beta(\tau))$. Recall $\hat{a}_i(\tau) \approx \mathbb{I}(y_i \ge x_i^T\hat{\beta}(\tau))$ is the main component of our test statistics and $\tilde{a}_i(\tau)$ follows i.i.d binomial distributions which is easy to analysis. Lemma 2 establishes the relationship between $\hat{W}_n^d$ and W_n^d. Theorem 1(i) then follows combining the results of Lemma 1 and Lemma 2. With out loss of generality, we write $\tau_0 = \epsilon$ and $\tau_M = 1-\epsilon$ though out the proof.

Lemma 1: $\sup_{\epsilon\le\tau\le 1-\epsilon} \|\hat{G}_n^d(\hat{\beta}(\tau)) - \hat{G}_n^d(\beta(\tau)) - G_n^d(\hat{\beta}(\tau)) + G_n^d(\beta(\tau))\| = o_p(1)$.

Proof. For any d-dimensional vector v, define the class of function $\mathcal{G}$ over a compact set $\mathcal{T} \in \mathbb{R}^{p+q}$ as

$$\mathcal{G} = \big\{v^T d_i \mathbb{I}(y_i \le x_i^T t), \quad t \in \mathcal{T}\big\}.$$

It is obvious that $\mathcal{G}$ is a VC subgraph class and $E(g^2)$ is bounded for any $g \in \mathcal{G}$.

Thus $v^T(\hat{G}_n^d(t) - G_n^d(t))$ is stochastically equicontinuous over $\mathcal{T}$ with semi-metric

$$\rho(t_1,t_2) = \big\{E\big(v^T d_i \mathbb{I}(y_i \le x_i^T t_1) - v^T d_i \mathbb{I}(y_i \le x_i^T t_2)\big)^2\big\}^{1/2}.$$

Since

$$\begin{aligned}\rho(t_1,t_2)^2 &\le (v^T d_i)^2 E\big(\mathbb{I}(x_i^T t_2 \le y_i \le x_i^T t_1) + \mathbb{I}(x_i^T t_1 \le y_i \le x_i^T t_2)\big)\\ &= 2(v^T d_i)^2 O(\|t_1 - t_2\|)\\ &= O(\|t_1 - t_2\|),\end{aligned}$$

and $\hat{\beta}(\tau)$ is a consistent estimator of $\beta(\tau)$ uniformly for $\tau \in [\epsilon, 1-\epsilon]$ (This result can be proved using similar and easier argument as the proof of theorem 1(ii).), we have

$$\sup_{\epsilon \le \tau \le 1-\epsilon} |v^T(\hat{G}_n^d(\hat{\beta}(\tau)) - G_n^d(\hat{\beta}(\tau))) - v^T(\hat{G}_n^d(\beta(\tau)) + G_n^d(\beta(\tau)))| = o_p(1)$$

by the definition of equicontinuity. The lemma is hence proved since v is arbitrary.

□

Lemma 2: $\|\hat{W}_n^d - W_n^d + D_{nd}^\tau Z_n\| = O(\sqrt{n}\|\hat{\beta}(\tau) - \beta(\tau)\|^2) + o_p(1)$ uniformly over $\tau \in [\epsilon, 1-\epsilon]$.

Proof. By simple manipulation, we can write

$$\hat{W}_n^d = W_n^d - D_{nd}^\tau Z_n + R_1 - R_2 - R_3,$$

where

$$R_1 = n^{-1/2} \sum d_i \mathbb{I}(y_i = x_i^T \hat{\beta}(\tau)) \hat{a}_i(\tau),$$

$$R_2 = \hat{G}_n^d(\hat{\beta}(\tau)) - \hat{G}_n^d(\beta(\tau)) - G_n^d(\hat{\beta}(\tau)) + G_n^d(\beta(\tau)),$$

$$R_3 = G_n^d(\hat{\beta}(\tau)) - G_n^d(\beta(\tau)) - D_{nd}^\tau Z_n.$$

When y_i is continuous, $\sum \mathbb{I}(y_i = x_i^T \hat{\beta}(\tau)) = p + q$ almost surely for any τ. Since $|\hat{a}_i(\tau)| \le 1$ and d_i bounded, $R_1 = O(n^{-1/2})$ uniformly.

By Lemma 1, R_2 is uniformly $o_p(1)$.

Now consider R_3. By Taylor expansion,

$$
\begin{aligned}
&\| G_n^d(\beta(\tau)+n^{-1/2}\Delta) - G_n^d(\beta(\tau)) - D_{nd}^{\tau}\Delta \| \\
=&\| \frac{1}{\sqrt{n}} \sum d_i n^{-1/2}(x_i^T\Delta) \int_0^1 \Big(f_i\big(x_i^T\beta(\tau)+n^{-1/2}(x_i^T\Delta)s\big) - f_i(x_i^T\beta(\tau)) \Big) ds \| \\
=&\| \frac{1}{\sqrt{n}} \sum d_i n^{-1/2}(x_i^T\Delta) \int_0^1 O\big(n^{-1/2}(x_i^T\Delta)s\big) ds \| \\
=&\| \frac{1}{\sqrt{n}} \sum d_i O\big(n^{-1}(x_i^T\Delta)^2\big) \| \\
=&O(n^{-1/2}\|\Delta\|^2).
\end{aligned}
$$

Let $\Delta = \sqrt{n}(\hat{\beta}(\tau)-\beta(\tau))$,

$$R_3 = O(\sqrt{n}\|\hat{\beta}(\tau)-\beta(\tau)\|^2).$$

□

Proof of Theorem 1(i): Set $d_i = x_i$. By the constraints in (2.3), $\hat{W}_n^x = n^{-1/2}\sum x_i(\hat{a}_i(\tau) - (1-\tau)) = 0$. Thus from Lemma 2, we have

$$D_{nx}^{\tau} Z_n = W_n^x + O(\sqrt{n}\|\hat{\beta}(\tau)-\beta(\tau)\|^2) + o_p(1).$$

Namely,

$$Z_n(1+o_p(1)) = (D_{nx}^{\tau})^{-1} W_n^x + o_p(1).$$

By similar argument as in Lemma 1, $\mathcal{W} = \big\{ x_i(\tilde{a}_i(\tau) - (1-\tau)), \quad \tau \in [\epsilon, 1-\epsilon] \big\}$

is a VC subgraph class with bounded envelope. Thus $\mathcal{W}$ is Donsker. Then we have $Z_n = O_p(1)$ since the limit of D_{nx}^{τ} is positive definite by (A3). Therefore we have the uniform Bahadur representation for quantile regression

$$Z_n = (D_{nx}^{\tau})^{-1} W_n^x + o_p(1). \tag{2.13}$$

By Lemma 2 and (2.13),

$$\hat{W}_n^d = W_n^d - D_{nd}^{\tau}(D_{nx}^{\tau})^{-1} W_n^x + o_p(1). \tag{2.14}$$

Notice that the above derivation holds for linear quantile regression model generally. Now we consider the model under H_0 where only x_{i1} is included. Set $d_i = x_{i2} - \hat{x}_{i2}$, from (2.14) we get

$$n^{-1/2}\sum_{i=1}^{n}(x_{2i}-\hat{x}_{2i})(\hat{a}_i(\tau)-(1-\tau)) = n^{-1/2}\sum_{i=1}^{n}(x_{2i}-\hat{x}_{2i}-K_n^{\tau}x_{1i})(\tilde{a}_i(\tau)-(1-\tau))+o_p(1), \tag{2.15}$$

where $K_n^{\tau} = (X_2 - \hat{X}_2)^T \Gamma_n^{\tau} X_1 (X_1^T \Gamma_n^{\tau} X_1)^{-1}$ and Γ_n^{τ}=diag$(f_i(x_{i1}^T \beta_1(\tau))$.

Since (2.15) holds uniformly for $\tau \in [\epsilon, 1-\epsilon]$,

$$\begin{aligned} &n^{-1/2}\sum_{i=1}^{n}(x_{2i} - \hat{x}_{2i})\tilde{b}_i = \\ &n^{-1/2}\sum_{i=1}^{n}\sum_{\tau_m \in S}(x_{2i} - \hat{x}_{2i} - K_n^{\tau_m} x_{1i})(\tilde{a}_i(\tau_m) - (1-\tau_m))\varphi'(\tau_m)(\tau_m - \tau_{m-1}) + o_p(1). \end{aligned} \tag{2.16}$$

By Lindeberg-Feller CLT, (2.16) converge to a Normal distribution of mean 0 and

variance

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{\tau_m\in S}\varphi'^2(\tau_m)(\tau_m-\tau_{m-1})^2(1-\tau_m)\tau_m(x_{2i}-\hat{x}_{2i}-K_n^{\tau_m}x_{1i})(x_{2i}-\hat{x}_{2i}-K_n^{\tau_m}x_{1i})^T. \tag{2.17}$$

Thus $T_n = S_n^T Q_n^{-1} S_n$ converges to a mixed chi-square distribution. □

We now want to study the consistency of our model-based bootstrap. Parallel to the notations in the original space, we have the following notations in the bootstrap space:

$\hat{W}_n^{d*} = n^{-1/2}\sum d_i(\hat{a}_i^*(\tau) - (1-\tau))$ where $\hat{a}_i^*(\tau)$ is the regression rank score under H_0 for the bootstrap sample.

$W_n^{d*} = n^{-1/2}\sum d_i(\tilde{a}_i^*(\tau) - (1-\tau))$ where $\tilde{a}_i^*(\tau) = \mathbb{I}(y_i^* \geq x_{i1}^T\tilde{\beta}_1(\tau))$.

$\hat{G}_n^{d*}(t) = n^{-1/2}\sum d_i\mathbb{I}(y_i^* \leq x_i^T t)$.

$G_n^{d*}(t) = n^{-1/2}\sum d_i E^*\mathbb{I}(y_i^* \leq x_i^T t)$.

$Z_n^* = \sqrt{n}(\hat{\beta}_1^*(\tau) - \hat{\beta}_1(\tau))$.

We first show that $\hat{\beta}_1^*(\tau)$ is a consistent estimator of $\beta_1(\tau)$. The relationship corresponding to Lemma 2 under the bootstrap space is given the Lemma 3. Combining the above results, we can finally establish the consistency of our bootstrap algorithm in Theorem 1(iii).

Proof of Theorem 1(ii): Write $\tilde{y}_i = x_{i1}^T\bar{\beta}_1(u_i)$, where $\bar{\beta}_1(\cdot)$ is the linear interpolation

of $\{\beta_1(\tau_m), m \in S\}$. By LLN,

$$\begin{aligned} \| \frac{1}{n} E^*[\sum \rho_\tau(\tilde{y}_i - x_{i1}^T\beta_1) - \sum \rho_\tau(\tilde{y}_i - x_{i1}^T\bar{\beta}_1(\tau))] \\ - \frac{1}{n}[\sum \rho_\tau(\tilde{y}_i - x_{i1}^T\beta_1) - \sum \rho_\tau(\tilde{y}_i - x_{i1}^T\bar{\beta}_1(\tau))] \| = o_{p^*}(1). \end{aligned} \tag{2.18}$$

Note that the expectation above is taken with respect to u_i. Since

$$\begin{aligned} &\| \frac{1}{n}[\sum \rho_\tau(\tilde{y}_i - x_{i1}^T\beta_1) - \sum \rho_\tau(\tilde{y}_i - x_{i1}^T\bar{\beta}_1(\tau))] \\ &\qquad - \frac{1}{n}[\sum \rho_{\tau'}(\tilde{y}_i - x_{i1}^T\beta_1') - \sum \rho_{\tau'}(\tilde{y}_i - x_{i1}^T\bar{\beta}_1(\tau'))] \| \\ \leq & c_1|\tau - \tau'| + c_2 \| \beta_1 - \beta_1' \|, \end{aligned}$$

$\frac{1}{n}[\sum \rho_\tau(\tilde{y}_i - x_{i1}^T\beta_1) - \sum \rho_\tau(\tilde{y}_i - x_{i1}^T\bar{\beta}_1(\tau))]$ is stochastically equicontinuous. Thus the convergence in (2.18) is uniform over $\tau \in [\tau_a, \tau_b]$ and β_1 in a compact set $\mathcal{B}$. We know that

$$\hat{\beta}_1^*(\tau) = \underset{\beta_1}{argmin} \sum \rho_\tau(y_i^* - x_{i1}^T\beta_1) - \rho_\tau(\tilde{y}_i - x_{i1}^T\bar{\beta}_1(\tau)),$$

and

$$\bar{\beta}_1(\tau) = \underset{\beta_1}{argmin}\, E^*[\sum \rho_\tau(\tilde{y}_i - x_{i1}^T\beta_1) - \rho_\tau(\tilde{y}_i - x_{i1}^T\bar{\beta}_1(\tau))].$$

The minimizer $\bar{\beta}_1(\tau)$ is also unique for $\tau \in [\tau_a, \tau_b]$. Notice that

$$\begin{aligned} &\| \frac{1}{n}\sum \rho_\tau(\tilde{y}_i - x_{i1}^T\beta_1) - \frac{1}{n}\sum \rho_\tau(y_i^* - x_{i1}^T\beta_1) \| \\ =& O(\frac{1}{n}\sum |y_i^* - \tilde{y}_i|) \\ =& O(\| \tilde{\beta}_1(\tau) - \bar{\beta}_1(\tau) \|). \end{aligned} \tag{2.19}$$

Since $\sup_{\tau_a \le \tau \le \tau_b} \parallel \hat{\beta}_1(\tau) - \beta_1(\tau) \parallel = o_p(1)$, $\sup_{\tau_a \le \tau \le \tau_b} \parallel \tilde{\beta}_1(\tau) - \bar{\beta}_1(\tau) \parallel = o_p(1)$. Thus by (2.18) and (2.19),

$$sup_{\tau \in [\tau_a, \tau_b], \beta_1 \in \mathcal{B}} \parallel \frac{1}{n} E^*[\sum \rho_\tau(\tilde{y}_i - x_{i1}^T \beta_1) - \sum \rho_\tau(\tilde{y}_i - x_{i1}^T \bar{\beta}_1(\tau))] - \frac{1}{n}[\sum \rho_\tau(y_i^* - x_{i1}^T \beta_1) - \sum \rho_\tau(\tilde{y}_i - x_{i1}^T \bar{\beta}_1(\tau))] \parallel = o_p^*(1) + o_p(1).$$

Let $\mathcal{B}_\delta(\bar{\beta}_1(\tau))$ be a ball of radius δ centered at $\bar{\beta}_1(\tau)$ with L^∞ norm. For any $b(\tau)$ in the boundary of $\mathcal{B}_\delta(\beta_1(\tau))$,

$$\begin{aligned} & \frac{1}{n} \sum \rho_\tau(y_i^* - x_{i1}^T b(\tau)) - \frac{1}{n} \sum \rho_\tau(y_i^* - x_{i1}^T \bar{\beta}_1(\tau)) \\ \ge & \frac{1}{n} E^* \sum \rho_\tau(y_i^* - x_{i1}^T b(\tau)) - \frac{1}{n} E^* \sum \rho_\tau(y_i^* - x_{i1}^T \bar{\beta}_1(\tau)) - o_p^*(1) - o_p(1) \\ \ge & \epsilon(\tau) - o_p^*(1) - o_p(1), \end{aligned}$$

where $\epsilon(\tau) \ge 0$ and the inequality is strict for some $\tau \in [\epsilon, 1-\epsilon]$. Namely,

$$P^*(\inf_{sup|\bar{\beta}_1(\tau) - b(\tau)| = \delta} \sup_{\tau \in [\tau_a, \tau_b]} \sum \rho_\tau(y_i^* - x_{i1}^T b(\tau)) - \sum \rho_\tau(y_i^* - x_{i1}^T \bar{\beta}_1(\tau)) \le 0]) \to 0$$

in P. By the convexity of ρ_τ,

$$P^*(\inf_{sup|\bar{\beta}_1(\tau) - b(\tau)| \ge \delta} \sup_{\tau \in [\tau_a, \tau_b]} \sum \rho_\tau(y_i^* - x_{i1}^T b(\tau)) - \sum \rho_\tau(y_i^* - x_{i1}^T \bar{\beta}_1(\tau)) \le 0]) \to 0$$

in P. Also notice that $\bar{\beta}_1(\tau) = \beta_1(\tau)$ for $\tau \in S$. Thus we have the desired result. □

Lemma 3: $G_n^{d*}(\hat{\beta}_1^*(\tau)) = G_n^{d*}(\hat{\beta}_1(\tau)) - D_{nd}^\tau Z_n^* + O_p(\sqrt{n} \parallel \hat{\beta}_1^*(\tau) - \hat{\beta}_1(\tau) \parallel^2) + o_{p^*}(1) + o_p(1)$ uniformly for $\tau \in S \cap [\tau_a, \tau_b]$.

Proof. Write $G_n^{d*}(\hat{\beta}_1(\tau)+\delta) - G_n^{d*}(\hat{\beta}_1(\tau)) = A_1 + A_2$ where

$$\begin{aligned} A_1 = n^{-1/2}\sum d_i E^*[\mathbb{I}(u_i \le \epsilon)\big(\mathbb{I}\big(x_{i1}^T\hat{\beta}_1(\epsilon) \le x_{i1}^T(\hat{\beta}_1(\tau)+\delta)\big) \\ - \mathbb{I}\big(x_{i1}^T\hat{\beta}_1(\epsilon) \le x_{i1}^T\hat{\beta}_1(\tau)\big)\big)] \\ + n^{-1/2}\sum d_i E^*[\mathbb{I}(u_i \ge 1-\epsilon)\big(\mathbb{I}\big(x_{i1}^T\hat{\beta}_1(1-\epsilon) \le x_{i1}^T(\hat{\beta}_1(\tau)+\delta)\big) \\ - \mathbb{I}\big(x_{i1}^T\hat{\beta}_1(1-\epsilon) \le x_{i1}^T\hat{\beta}_1(\tau)\big)\big)], \end{aligned}$$

$$\begin{aligned} A_2 = n^{-1/2}\sum d_i E^*[\mathbb{I}(\epsilon < u_i < 1-\epsilon)\big(\mathbb{I}\big(x_{i1}^T\tilde{\beta}_1(u_i) < x_{i1}^T(\hat{\beta}_1(\tau)+\delta)\big) \\ - \mathbb{I}\big(x_{i1}^T\tilde{\beta}_1(u_i) < x_{i1}^T\hat{\beta}_1(\tau)\big)\big)]. \end{aligned}$$

From Theorem 1 of *Neocleous and Portnoy* (2008), $x_{i1}^T\tilde{\beta}_1(\tau)$ is strictly monotone uniformly on $[\epsilon, 1-\epsilon]$ with probability tending to 1. Therefore A_1 is $o_p(1)$ for any $\delta \to 0$.

Let $\Delta = \tilde{\beta}_1(u_i) - \beta_1(u_i)$, write A_2 as

$$\begin{aligned} & n^{-1/2}\sum d_i E^*[\mathbb{I}(\epsilon < u_i < 1-\epsilon)\big(\mathbb{I}(x_{i1}^T\beta_1(u_i) + x_{i1}^T\Delta) < x_{i1}^T\hat{\beta}_1(\tau) + x_{i1}^T\delta\big) \\ & \qquad - \mathbb{I}\big(x_{i1}^T\beta_1(u_i) + x_{i1}^T\Delta) < x_{i1}^T\hat{\beta}_1(\tau)\big)] \\ =& n^{-1/2}\sum d_i E^*[\mathbb{I}(\epsilon < u_i < 1-\epsilon)\mathbb{I}\big(x_{i1}^T\hat{\beta}_1(\tau) - x_{i1}^T\Delta < x_{i1}^T\beta_1(u_i) \\ & \qquad < x_{i1}^T\hat{\beta}_1(\tau) - x_{i1}^T\Delta - x_{i1}^T\delta\big)\mathbb{I}(x_{i1}^T\delta < 0)] \\ & + n^{-1/2}\sum d_i E^*[\mathbb{I}(\epsilon < u_i < 1-\epsilon)\mathbb{I}\big(x_{i1}^T\hat{\beta}_1(\tau) - x_{i1}^T\Delta - x_{i1}^T\delta < x_{i1}^T\beta_1(u_i) \\ & \qquad < x_{i1}^T\hat{\beta}_1(\tau) - x_{i1}^T\Delta\big)\mathbb{I}(x_{i1}^T\delta \ge 0)]. \end{aligned}$$

We only need to consider the case when $x_{i1}^T\delta < 0$, since the situation when $x_{i1}^T\delta \ge 0$ is symmetric.

When $x_{i1}^T\delta < 0$,

$$
\begin{aligned}
& n^{-1/2}\sum d_i E^*[\mathbb{I}(\epsilon < u_i < 1-\epsilon)\mathbb{I}\big(x_{i1}^T\hat{\beta}_1(\tau) - x_{i1}^T\Delta < x_{i1}^T\beta_1(u_i) < x_{i1}^T\hat{\beta}_1(\tau) - x_{i1}^T\Delta - x_{i1}^T\delta\big)] \\
=& n^{-1/2}\sum d_i \int_{max\{x_{i1}^T\hat{\beta}_1(\tau)-x_{i1}^T\Delta,\, x_{i1}^T\beta_1(\epsilon)\}}^{min\{x_{i1}^T\hat{\beta}_1(\tau)-x_{i1}^T\Delta-x_{i1}^T\delta,\, x_{i1}^T\beta_1(1-\epsilon)\}} f_i(c)dc \\
=& n^{-1/2}\sum d_i \int_{max\{x_{i1}^T\hat{\beta}_1(\tau)-x_{i1}^T\Delta,\, x_{i1}^T\beta_1(\epsilon)\}}^{min\{x_{i1}^T\hat{\beta}_1(\tau)-x_{i1}^T\Delta-x_{i1}^T\delta,\, x_{i1}^T\beta_1(1-\epsilon)\}} f_i(x_{i1}^T\beta_1(\tau)) + O(|c - x_{i1}^T\beta_1(\tau)|)dc \\
=& n^{-1/2}\sum \Big(d_i f_i(x_{i1}^T\beta_1(\tau))\big(- max\{x_{i1}^T\hat{\beta}_1(\tau) - x_{i1}^T\Delta,\, x_{i1}^T\beta_1(\epsilon)\} \\
& + min\{x_{i1}^T\hat{\beta}_1(\tau) - x_{i1}^T\Delta - x_{i1}^T\delta,\, x_{i1}^T\beta_1(1-\epsilon)\}\big) \\
& + O(\parallel \Delta \parallel\parallel \delta \parallel) + O(\parallel \delta \parallel^2) + O(\parallel \hat{\beta}_1(\tau) - \beta_1(\tau) \parallel\parallel \delta \parallel)\Big) \\
=& n^{-1/2}\sum d_i f_i(x_{i1}^T\beta_1(\tau))(-x_{i1}^T\delta) + O(\sqrt{n} \parallel \Delta \parallel\parallel \delta \parallel) + O(\sqrt{n} \parallel \delta \parallel^2) \\
& + O(\sqrt{n} \parallel \hat{\beta}_1(\tau) - \beta_1(\tau) \parallel\parallel \delta \parallel) + R_1,
\end{aligned}
$$

where

$$
\begin{aligned}
R_1 =& O\Big(n^{-1/2}\sum d_i f_i(x_{i1}^T\beta_1(\tau))\big(\mathbb{I}(x_{i1}^T\hat{\beta}_1(\tau) - x_{i1}^T\Delta < x_{i1}^T\beta_1(\epsilon)) \\
& + \mathbb{I}(x_{i1}^T\hat{\beta}_1(\tau) - x_{i1}^T\Delta - x_{i1}^T\delta > x_{i1}^T\beta_1(1-\epsilon))\big)\Big).
\end{aligned}
$$

For $\tau \in [\tau_a, \tau_b]$, R_1 converges to zero in probability if Δ and δ are $o(1)$.

Recall $\Delta = \tilde{\beta}_1(u_i) - \beta_1(u_i)$, which is $O_p(n^{-1/2})$ uniformly over $u_i \in [\epsilon, 1-\epsilon]$ by Theorem 1 of *Neocleous and Portnoy* (2008). Let $\delta = \hat{\beta}_1^*(\tau) - \hat{\beta}_1(\tau) = o_{p^*}(1) + o_p(1)$ for $\tau \in S \cap [\tau_a, \tau_b]$. Thus

$$
A_2 = D_{nd}^{\tau} Z_n^* + O_p(\sqrt{n} \parallel \hat{\beta}^*(\tau) - \hat{\beta}(\tau) \parallel^2) + o_{p^*}(1) + o_p(1), \tag{2.20}
$$

we have the desired result. □

Proof of Theorem 1(iii): Similar to Lemma 1,

$$\sup_{\tau\in S\cap[\tau_a,\tau_b]} \|\hat{G}_n^{d*}(\hat{\beta}_1^*(\tau)) - \hat{G}_n^{d*}(\hat{\beta}_1(\tau)) - G_n^{d*}(\hat{\beta}_1^*(\tau)) + G_n^{d*}(\hat{\beta}_1(\tau))\| = o_{p^*}(1) + o_p(1).$$

This is because

$$\mathcal{G}^* = \left\{v^T d_i \mathbb{I}(y_i^* \le x_i^T t), \quad t \in \mathcal{T}\right\}$$

is a VC subgraph class and $\hat{\beta}_1^*(\tau)$ is consistent for $\hat{\beta}_1(\tau)$ uniformly over $\tau \in S\cap[\tau_a, \tau_b]$. Thus we have

$$\hat{W}_n^{d*} = W_n^{d*} - D_{nd}^{\tau} Z_n^* + R_1^* - R_2^* - R_3^*,$$

where

$$R_1^* = n^{-1/2} \sum d_i \mathbb{I}(y_i^* = x_{i1}^T \hat{\beta}_1^*(\tau)) \hat{a}_i^*(\tau),$$

$$R_2^* = \hat{G}_n^{d*}(\hat{\beta}_1^*(\tau)) - \hat{G}_n^{d*}(\hat{\beta}_1(\tau)) - G_n^{d*}(\hat{\beta}_1^*(\tau)) + G_n^{d*}(\hat{\beta}_1(\tau)),$$

$$R_3^* = G_n^{d*}(\hat{\beta}_1^*(\tau)) - G_n^{d}(\hat{\beta}_1(\tau)) - D_{nd}^{\tau} Z_n^*.$$

Since R_1^* and R_2^* are $o_{p^*}(1) + o_p(1)$, by Lemma 4, we have

$$\hat{W}_n^{d*} = W_n^{d*} - D_{nd}^{\tau} Z_n^* + O_p(\sqrt{n} \parallel \hat{\beta}_1^*(\tau) - \hat{\beta}_1(\tau) \parallel^2) + o_{p^*}(1) + o_p(1).$$

Set $d_i = x_{i1}$,

$$D_{nx}^{\tau} Z_n^* = W_n^{x*} + O_p(\sqrt{n} \parallel \hat{\beta}^*(\tau) - \hat{\beta}(\tau) \parallel^2)) + o_{p^*}(1) + o_p(1).$$

Thus $Z_n^* = (D_{nx}^\tau)^{-1} W_n^{x*} + o_{p^*}(1) + o_p(1)$ and

$$\hat{W}_n^{d*} = W_n^{d*} - D_{nd}^\tau (D_{nx}^\tau)^{-1} W_n^{x*} + o_{p^*}(1) + o_p(1).$$

Set $d_i = \hat{x}_{i2} - x_{i2}$,

$$n^{-1/2} \sum_i (x_{i2} - \hat{x}_{i2})(\hat{a}_i^*(\tau) - (1-\tau)) =$$

$$n^{-1/2} \sum_i (x_{i2} - \hat{x}_{i2} - K_n^\tau x_{1i})(\tilde{a}_i^*(\tau) - (1-\tau)) + o_{p^*}(1) + o_p(1).$$

Therefore

$$S_n^* = n^{-1/2} \sum_{i=1}^n \sum_{\tau_m \in S} (x_{i2} - \hat{x}_{i2} - K_n^{\tau_m} x_{i1})(\tilde{a}_i^*(\tau_m) - (1-\tau_m))\varphi'(\tau_m)(\tau_m - \tau_{m-1}) + o_{p^*}(1) + o_p(1). \tag{2.21}$$

Comparing equation (2.15) with (2.21), their right hand sides are exactly the same except that we have $\tilde{a}_i^*$ instead of $\tilde{a}_i$ for the bootstrapped test statistics. Recall $\tilde{a}_i(\tau) = \mathbb{I}(y_i \geq x_{i1}^T \beta(\tau))$ and $\tilde{a}_i^*(\tau) = \mathbb{I}(y_i^* \geq x_{i1}^T \tilde{\beta}_1(\tau))$. Consider a set $\mathcal{D}$ where $x_{i1}^T \tilde{\beta}_1(\tau)$ is strictly monotone for $\tau \in [\epsilon, 1-\epsilon]$. On $\mathcal{D}$, $\tilde{a}_i^*(\tau)$ given data independently follows the same binary distribution as $\tilde{a}_i(\tau)$. Therefore the conditional distribution of T_n^* given data will convergence to the same limiting distribution as T_n on $\mathcal{D}$. We then have the desired results since $P(\mathcal{D}) \to 1$ as $n \to \infty$ by Theorem 1 of *Neocleous and Portnoy* (2008). □

CHAPTER III

Rank-based Inference for Censored Quantile Regression

3.1 Introduction

In the previous chapter, we have already seen that the quantile regression is particularly useful when the effect of the covariates to the response varies in τ. This phenomenon can often be observed in biomedical studies, where the effect of certain treatment is expected to depend on certain unobserved aspects of the patients.

In biomedical studies, the responses are usually censored from the right because patients may drop out of the study and a clinical trial will terminate after certain period of time. The accelerated failure time model and the Cox proportional hazard model are popular regression models to study censored outcomes. However, these two models do not capture the heterogeneity of the treatment effect. Therefore it is useful to develop estimation and inference schemes for quantile regression with censored outcomes.

Censored quantile regression was first studied in *Powell* (1984, 1986), where the censored time is assumed to be fixed. *Wang and Fygenson* (2009) developed methods for longitudinal data with fixed censoring. *Ying et al.* (1995), *Zhou* (2006) and *Bang and Tsiatis* (2002) among others proposed different estimating equations assuming the censored time is independent of the survival time;

A less stringent and more common assumption, which is called standard right censoring, assumes the censored time is conditionally independent of the survival time given the covariates. Multiple methodologies have been proposed under this standard right censoring assumptions, and they can be classified into two groups by whether the linear quantile regression model is assumed to held locally at one τ or globally at any τ.

Under the local linear quantile regression model, *Wang and Wang* (2009) proposed a method using redistribution of mass idea; *Leng and Tong* (2013) proposed an alternative method by inversing censoring probability; *Backer et al.* (2019) constructed an adapted loss function for censored quantile regression. Though these methods are different, they share the same feature that the conditional distribution of either the survival time or the censored time needs to be estimated nonparametrically to carry out the estimation.

Under the global linear quantile regression model, two popular methods were proposed by *Portnoy* (2003) and *Peng and Huang* (2008). Portnoy proposed an iterative self-consistency algorithm based on the idea of redistribution of mass. Peng and Huang's method constructed their estimation equation by clever usage of the martingale feature of censored data. No estimation of conditional distributions is re-

quired for Portnoy's method or Peng and Huang's method. Because when the global quantile model is assumed, the conditional distribution of the survival is defined by the coefficients function of τ. But the global linear assumption is stronger than the local one.

Asymptotic normality has been established for all the above-mentioned methods. The covariance of the estimated coefficients takes complicated form involving the conditional densities of the survival and censored times. Therefore inference for censored quantile regression is usually carried out by building a confidence interval of the interested coefficient using the bootstrap. Though building the confidence interval is sufficient for some testing purposes, there are scenarios where a more flexible testing procedure is preferred. For example, suppose the goal is to test the significance of a coefficient over a quantile region. Shown by the simulation results in Chapter 2, for quantile regression without censoring, the rank-based test outperformed the method based on a confidence band of the coefficient over the selected region. Another example is when comparing two nested models, one needs to test whether several coefficients simultaneously equal to zero. In this case, building a confidence interval individually for each coefficient will lead to the multiple testing problem, which will need to be adjusted with a possible loss of power.

In this chapter, we propose a rank-based test under the global linear quantile regression model with random right censoring. The rank-based test allows the users to study the effect of one or more coefficients over any pre-specified quantile region. There are two major challenges to conduct the rank-based test. Firstly, for quantile regression without censoring, the rank-based test is constructed with the

regression rank score, which is the solution of the dual problem of optimizing the quantile loss function. However, the regression rank score is not naturally defined for censored settings. We propose a regression rank score for censored quantile regression with a similar redistribution of mass idea in *Portnoy* (2003) and *Wang and Wang* (2009). Secondly, the bootstrap is required to implement our test since the exact analytic form of the limiting distribution of our test statistics is complicated. In a hypothesis testing framework, the bootstrap sample should be generated from the null hypothesis to ensure bootstrap consistency. Sampling schemes like paired bootstrap or perturbing the minimand that generate the bootstrap samples from the full model can not be used in our context. Therefore we propose a new bootstrap algorithm that mimics the true data generating procedure under the null hypothesis. This bootstrap algorithm is an extension of the model-based bootstrap proposed in Chapter 2 to censored quantile regression.

In conclusion, recent research on censored quantile regression has focused more on the estimation and inference methods are relatively limited. In this paper, we focus on the inference part and propose a rank-based test that complements what is available in the literature. We also propose a model-based bootstrap that can be used for the general hypothesis testing framework for global censored quantile regression.

3.2 Main results

3.2.1 Test statistics

By convention, let T_i be the survival time which may not be fully observed. Let C_i denotes the censoring time and $Y_i := \min(T_i, C_i)$ is the observed outcome. Let $\Delta_i = \mathbb{I}(T_i \leq C_i)$ be the event indicator. Further, we assume that given covariates x_i, the survival time T_i and censoring time C_i are independent. This standard assumption is commonly assumed in the survival analysis literature.

We consider a random sample of size n that follows the linear quantile model

$$T_i = x_{i1}^T \beta_1(\tau) + x_{i2}^T \beta_2(\tau) + e_{i,\tau}, \quad \forall \tau \in (0, \tau_U], \quad i = 1, 2, ...n, \tag{3.1}$$

where $x_{i1} \in \mathbb{R}^p$, $x_{i2} \in \mathbb{R}^q$ and the conditional τth quantile of $e_{i,\tau}$ given x_{i1} and x_{i2} is 0. Notice that we assume the above linear relationship holds up to quantile level τ_U, which denotes the largest quantile level where the coefficient is identifiable. Two popular methods has been proposed for the estimation of Model (3.1) in *Portnoy* (2003) and *Peng and Huang* (2008) respectively. Both methods estimate $\beta(\tau)$ sequentially at a set of $M + 1$ grid points $\mathcal{S} = (t_0, t_1, ..., t_M)$, where $t_M \leq \tau_U$. In this chapter, we utilize these available methods for the estimation of $\beta(\tau)$.

We are interested in testing

$$H_0\text{: } \beta_2(\tau) = 0, \ \forall \tau \in (0, \tau_U]$$

vs

$$H_1\text{: } \beta_2(\tau) \neq 0 \text{ for } \tau \in [\tau_a, \tau_b],$$

where $[\tau_a, \tau_b]$ is a user-specific strict subset of $[t_1, t_M]$.

Similar to the previous chapter, we would like to use the regression rank score $\hat{a}_i(\tau)$ to construct our test statistics. But we can not use $\hat{a}_i(\tau)$ directly in the current censored setting because of two reasons. Firstly, unlike the uncensored case where $\hat{a}_i(\tau)$ is the solution of the dual problem (2.3), $\hat{a}_i(\tau)$ is undefined for the censored case. Secondly, the original $\hat{a}_i(\tau)$ does not take the effect of censoring into consideration. Therefore, a regression rank score for the censored case needs to be developed.

To overcome the difficulties, observe that as shown in (2.4), $\hat{a}_i(\tau) = \mathbb{I}(T_i > x_i^T \hat{\beta}(\tau))$ unless the outcomes are exactly on the fitted line. However for both censored and uncensored cases, the number of points lying on fitted line is bounded by a constant independent of n uniformly in τ. Therefore the difference between $\hat{a}_i(\tau)$ and $\mathbb{I}(T_i > x_i^T \hat{\beta}(\tau))$ is of smaller order and most asymptotic properties will not be influenced if $\hat{a}_i(\tau)$ is replaced with $\mathbb{I}(T_i > x_i^T \hat{\beta}(\tau))$.

We use the redistribution of mass idea motivated by *Portnoy* (2003) and *Wang and Wang* (2009) to account for the censoring. Suppose we already obtained the $\hat{\beta}(\tau)$ on the set of grid points $\mathcal{S} = (t_0, t_1, ..., t_M)$. Let $\tilde{\beta}(\tau)$ be the linear interpolation between these grid points. For censored observations , define

$$\hat{\tau}_i = \inf_{t_0 \le \tau \le t_M} \{x_i^T \tilde{\beta}(\tau) \ge C_i\}. \tag{3.2}$$

Set $\hat{\tau}_i = t_M$ if $C_i > x_i^T \tilde{\beta}(\tau)$.

We then define for each observation a weight $w_i(\tau)$ as

$$\hat{w}_i(\tau) = \begin{cases} \frac{\tau - \hat{\tau}_i}{1 - \hat{\tau}_i} & \Delta_i = 0, \tau \geq \hat{\tau}_i \\ 1 & \Delta_i = 0, \tau < \hat{\tau}_i \\ 1 & \Delta_i = 1 \end{cases} \tag{3.3}$$

and the regression rank score for censored version as

$$\hat{a}_i^c(\tau) = 1 - \hat{w}_i(\tau)\mathbb{I}(Y_i - x_i^T\tilde{\beta}(\tau) < 0). \tag{3.4}$$

The intuition is that when defining $\hat{a}_i^c(\tau)$, what matters is the sign of $T_i - x_i^T\tilde{\beta}(\tau)$ but not its exact value. When $\Delta_i = 0$ and $\tau < \hat{\tau}_i$, which means C_i lies above $x_i^T\tilde{\beta}(\tau)$, T_i is also above $x_i^T\tilde{\beta}(\tau)$ since T_i is no smaller then C_i. Thus in this case, we can assign $\hat{w}_i(\tau)$ to be 1, which is equivalent to replacing the unobserved T_i with the observed C_i. When $\Delta_i = 0$ and $\tau \geq \hat{\tau}_i$, C_i is below the fitted line and T_i can either lie below or above $x_i^T\tilde{\beta}(\tau)$. We assign $\hat{w}_i(\tau) = (\tau - \hat{\tau}_i)/(1 - \hat{\tau}_i)$, which is the probability T_i is below $x_i^T\tilde{\beta}(\tau)$ given $T_i > C_i$. Notice that when there is no censoring, $\hat{a}_i^c(\tau) = \mathbb{I}(T_i > x_i^T\hat{\beta}(\tau))$ is asymptotically equivalent to the regression rank score for the uncensored cases.

Remark: Although $x_i^T\beta(\tau)$ is monotone in τ for any x_i, $x_i^T\hat{\beta}(\tau)$ may not be monotone. So strictly speaking when $\Delta_i = 0$ and $\tau \geq \hat{\tau}_i$, C_i may still lie above the fitted line. By *Portnoy and Lin* (2010), $x_i^T\hat{\beta}(\tau)$ is monotone with probability going to 1. Therefore the above statement is true asymptotically.

Write the design matrix of (3.1) as $X = [X_1, X_2]$, Let $\hat{X_2} = X_1(X_1^TX_1)^{-1}X_1^TX_2$

be the projection of X_2 into the spaces spanned by columns of X_1. Let $Q_n = n^{-1}(X_2 - \hat{X}_2)^T(X_2 - \hat{X}_2)$. Define

$$S(\tau) = n^{-1/2} \sum_i (x_{i2} - \hat{x}_{i2}) \hat{a}_i^c(\tau). \tag{3.5}$$

where $\hat{a}_i^c(\tau)$ is defined in (3.4) and calculated under the restricted model that only includes X_1. $S(\tau)$ is the main component of our test statistics. Intuitively, $\hat{a}_i^c(\tau)$ represents the relative position of observation i at τth level after adjusting for X_1. If the null hypothesis is true, no more variation in $\hat{a}_i^c(\tau)$ can be further explained by $X_2 - \hat{X}_2$. Thus we expect the norm of $S(\tau)$ to be close the 0 if the null hypothesis is true.

Based on $S(\tau)$, we construct the following two test statistics,

$$\mathcal{T}_1 = \Big(\sum_{t_m \in \mathcal{S} \cap [\tau_a, \tau_b]} S(t_m)(t_m - t_{m-1}) \Big)^T Q_n^{-1} \Big(\sum_{t_m \in \mathcal{S} \cap [\tau_a, \tau_b]} S(t_m)(t_m - t_{m-1}) \Big), \tag{3.6}$$

$$\mathcal{T}_2 = \sum_{t_m \in \mathcal{S} \cap [\tau_a, \tau_b]} \big(S(t_m)^T Q_n^{-1} S(t_m) \big)(t_m - t_{m-1}). \tag{3.7}$$

For $\mathcal{T}_1$, we first take a weighted sum of $S(\tau)$ over all the grid points in $[\tau_a, \tau_b]$. This way, our test statistics can detect the effect of X_2 over the $[\tau_a, \tau_b]$ region instead of a single quantile level. $\mathcal{T}_1$ is probably more natural and is equivalent to the test statistics proposed in the previous chapter for the uncensored case when a Wilcoxon score function is used.

One possible defect of $\mathcal{T}_1$ is that if the quantile region of interest $[\tau_a, \tau_b]$ is relatively

large, it is possible that the effect is negative at the lower quantile level but positive at the upper quantile level. In this case, the power of $\mathcal{T}_1$ may suffer because it can be roughly seen as the average effect over this region. Therefore we propose another test statistic $\mathcal{T}_2$ where the weighted sum is taken over the square of $S(\tau)$. Thus $\mathcal{T}_2$ is expected to have better power in the aforementioned scenario. The performance of $\mathcal{T}_1$ and $\mathcal{T}_2$ is compared numerically in Section 3.3.

3.2.2 Bootstrap algorithm

The asymptotic properties of $\mathcal{T}_1$ and $\mathcal{T}_2$ will be studied in the next subsection, where the limiting distributions of $\mathcal{T}_1$ and $\mathcal{T}_2$ are shown to take relative complicated forms. Therefore we use the bootstrap to approximate the distribution of $\mathcal{T}_1$ and $\mathcal{T}_2$ under H_0, which is common in the censored quantile regression literature.

However, common resampling schemes like the paired bootstrap or resampling by perturbing the minimand can not be used for our purpose because these methods generate the bootstrap samples under the full model instead of the restricted model. In this subsection, we propose a new bootstrap algorithm which generalizes the model-based bootstrap introduced in the previous chapter to the censored case.

Recall we assume that given x_i, T_i and C_i are independent. This enables us to generate T_i^* and C_i^* independently while keeping x_i fixed. To generate T_i^*, notice that under H_0, we have $Q_\tau(T_i|x_{i1}) = x_{i1}^T\beta_1(\tau)$. Therefore it is nature to set $T_i^* = x_{i1}^T\tilde{\beta}_1(u_i)$ where $u_i \sim U(0,1)$.

To resample C_i^*, we estimate $G(\cdot|x_{i1}, x_{i2})$ using a local Kaplan–Meier (KM) esti-

mator. Specifically, let

$$\hat{G}(y|x) = 1 - \prod_{i=1}^{n} \left(1 - \frac{B_{ni(x)}}{\sum_j \mathbb{I}(y_i < y_j) B_{nj}(x)}\right)^{\mathbb{I}(Y_i<y,\Delta_i=0)}, \tag{3.8}$$

where $B_{nj}(x) = \frac{K((x-x_j)/h_n)}{\sum_k K((x-x_k)/h_n)}$, K is a selected kernel density function and h_n is a sequence of bandwidth that tends to 0. Then set $C_i^* = \hat{G}^{-1}(v_i|x_{i1}, x_{i2})$ where $v_i \sim U(0,1)$ independent of u_i.

Remark: It is difficult to get an accurate estimate of $G(\cdot|x_{i1}, x_{i2})$ using the local KM estimator unless $p+q$ is small. Alternatively, since the role of T_i and C_i are symmetric, we could fit a censored quantile regression $Q_{C_i}(\tau|x_{i1}, x_{i2}) = x_{i1}^T\gamma_1(\tau) + x_{i2}^T\gamma_2(\tau)$ and let $C_i^* = x_{i1}^T\hat{\gamma}_1(v_i) + x_{i2}^T\hat{\gamma}_2(v_i)$. Again, $v_i \sim U(0,1)$ is independent of u_i. This approach requires the additional assumption that a linear quantile model also holds for C_i.

When the above algorithm is implemented, however, one may encounter non-identifiability issues in multiple steps. Notice that we choose $[\tau_a, \tau_b]$ to be inside $(0, \tau_U]$ to avoid the non-identifiability issue of $\beta(\tau)$ over our region of interest. But the non-identifiability is unavoidable when generating bootstrap samples.

When generating T_i^*, u_i ranges from 0 up to 1 but $\hat{\beta}_1(\tau)$ is only attainable until τ_U. Fortunately, since our test only focus on $[\tau_a, \tau_b]$, the exact value of $\hat{\beta}_1(\tau)$ when $\tau > \tau_U$ has no influence on the results. Therefore we can let $\tilde{\beta}_1(\tau) = \hat{\beta}_1(\tau_U)$ for $\tau > \tau_U$.

A similar non-identifiability issue also occurs when generating C_i^*. The problem is slightly trickier in this case but we can assign a very large value for C_i^* when

$\hat{G}^{-1}(\cdot|x_{i1}, x_{i2})$ is unidentifiable at the generated v_i level. This is because if we look locally at any x_i in the domain, C_i is unidentifiable in the population when the largest attainable value of T_i is smaller than the largest attainable value of C_i. Because C_i will always be censored by T_i for $C_i > \sup T_i$, we have no information about the distribution of C_i when $C_i > \sup T_i$. But since survival time T_i is what we are really interested in, the exact value of C_i is not important, as long as we know T_i can be observed in this case.

We are now ready to summarize the detailed algorithm of the proposed bootstrap. The following algorithm uses $\mathcal{T}_1$ as the test statistics and the local KM to re-sample C_i^*; the algorithm using $\mathcal{T}_2$ as the test statistics or censored quantile regression to resample C_i^* is similar.

Step 1: Fit the censored quantile regression under H_0 using Portnoy's or Peng and Huang's method. Calculate $\mathcal{T}_1$ using (3.6).

Step 2: For $i = 1, ..., n$, generate $u_i \sim U(0, 1)$. Let $T_i^* = x_{i1}\tilde{\beta}(u_i)$, where $\tilde{\beta}(\tau)$ is the linear interpolation of $\{\hat{\beta}_1(\tau_m), m \in \mathcal{S}\}$ calculated under the restricted model. Set $\tilde{\beta}_1(\tau) = \hat{\beta}_1(\tau_U)$ for $\tau > \tau_U$.

Step 3: For $i = 1, ..., n$, generate $v_i \sim U(0, 1)$ independent of u_i. Let $C_i^* = \hat{G}^{-1}(v_i|x_{i1}, x_{i2})$, where $\hat{G}(\cdot|x_{i1}, x_{i2})$ is estimated using the local KM estimator described in (3.8). Set C_i^* to be a very large number if $\hat{G}^{-1}(\cdot|x_{i1}, x_{i2})$ is undefined at v_i.

Step 4: Construct a bootstrap sample $(Y_i^*, \Delta_i^*, x_{i1}, x_{i2})$. Calculate $\mathcal{T}_1^*$ at this bootstrap sample.

Step 5: Repeat steps 2 to 4 for B times to get $\{\mathcal{T}_{11}^*, \mathcal{T}_{12}^*, ..., \mathcal{T}_{1B}^*\}$. The resulting p-value is calculated by $B^{-1}\sum_b \mathbb{I}(\mathcal{T}_1 > \mathcal{T}_{1b}^*)$.

3.2.3 Asymptotic properties

In this subsection, we study the asymptotic properties of $\mathcal{T}_1$ and $\mathcal{T}_2$ and show that the proposed bootstrap inference is consistent. For simplicity, the results of this subsection will be proved assuming that the Portnoy's method is used for the estimation of $\hat{\beta}(\tau)$. The $\hat{\beta}(\tau)$ estimated using Portnoy's method and Peng and Huang's method are similar numerically (*Koenker* (2008)) and the equivalence of these two methods is discussed in *Peng* (2012).

The following regularity conditions are assumed:

(B1) Let $\epsilon = t_0 < 2\epsilon = t_1 < ... < t_M \leq \min(1-\epsilon, \tau_U)$ be a set of grid points where $n^{-1/2} \ll t_j - t_{j-1} \ll n^{-1/4}$, $j = 2, ..., M$. Assume (3.1) is identifiable over $(0, \tau_U]$.

(B2) there is no censoring below 2ϵ. Namely for any $\tau < 2\epsilon$, $x_i^T\beta(\tau) < C_i$.

(B3) $\|x_i\|$ is bounded uniformly in i.

(B4) Given x, the conditional density $f(t|x)$ and $g(t|x)$ have uniformly bounded and strictly positive derivatives with respect to t, for any $t \in x^T\beta(\tau)$, $\tau \in [2\epsilon, \tau_U]$.

(B5) $F(t|x)$ and $G(t|x)$ have bounded second partial derivatives (uniformly in t) with respect to x.

(B6) The matrix $\mathbb{E}(xx^T)$ and $D(t) = \mathbb{E}\big(xx^T f(x^T\beta(t)|x)(1 - G(x^T\beta(t)|x))\big)$ are positive definite for $t \in [\epsilon, \tau_U]$.

(B7) The kernel density function K is positive, with compact support, and Lipschitz-continuous of order 1. Furthermore, $\int K(u)du = 1$, $\int uK(u)du = 0$, $\int K^2(u)du < \infty$ and $\int |u|^2 K(u)du < \infty$.

(B8) The bandwidth satisfies $h_n = c_n n^{-1/2+\gamma_0}$, with $c_n \to c$, where c is a constant, $0 < \gamma_0 < 1/4$.

(B1) controls the distance of adjacent grid points to be of order between $n^{-1/2}$ and $n^{-1/4}$. The same order is required in *Portnoy and Lin* (2010) to establish the asymptotic normality of $\hat{\beta}(\tau)$ estimated with Portnoy's method. This order is also required in Chapter 2 to show the consistency of the model-based bootstrap without censoring. (B2) is required by the Portnoy's method to ensure that it is valid to use quantile regression without censoring to estimate $\beta(\cdot)$ at t_0th quantile level. In (B3) we assume that the covariates are bounded, which is seemingly restrictive. However, since we assume the linear quantile model globally, the quantile function $x^T\beta(\tau_1)$ and $x^T\beta(\tau_1)$ will cross eventually if x is allowed to go to infinity, unless $x^T\beta(\tau_1)$ and $x^T\beta(\tau_1)$ are parallel, which precludes heterogeneity. (B4) and (B6) are common assumptions assumed when studying the asymptotic properties of censored quantile regression. Notice that assuming $D(t)$ to be positive definite until τ_Uth level implies (3.1) is identifiable up to τ_U. (B5), (B7) and (B8) are required in Theorem 2.1 of *Gonzalez-Manteiga and Cadarso-Suarez* (1994) where the asymptotic behavior of $\hat{G}(t|x)$ is studied.

Theorem 3.1: Under regularity conditions (B1)-(B8), we have under H_0,

(i) $S(t)$ converges to a zero mean Gaussian process for $t \in [\tau_a, \tau_b]$. Therefore $\mathcal{T}_1 \Rightarrow \bar{\chi}^2$, where $\bar{\chi}^2$ is a mixed chi-square distribution as a weighted sum of q chi-square variables of one degree of freedom; The limiting distribution of $\mathcal{T}_2$ is a time integral of a squared Gaussian process.

(ii) Given the data, the conditional distribution of $\mathcal{T}_1^*$ will converge to the same limiting distribution as $\mathcal{T}_1$; the same can be said for $\mathcal{T}_2^*$.

Theorem 3.1 shows the consistency of our bootstrap method. Theorem 3.1(i) follows from the proof in *Portnoy and Lin* (2010). To establish 3.1(ii), the key is to show that the conditional distribution of $S^*(\tau) = n^{-1/2}\sum_i (x_{i2} - \hat{x}_{i2})\hat{a}_i^*(\tau)$ given the original data has the same limiting distribution as $S(\tau)$, which follows if the conditional distribution of $\sqrt{n}(\hat{\beta}^*(t) - \hat{\beta}(t))$ given data converges to the same limit as $\sqrt{n}(\hat{\beta}(t) - \beta(t))$. By using results from the product-integration theory (*Gill and Johansen* (1990)), we could expand $\sqrt{n}(\hat{\beta}^*(t) - \hat{\beta}(t))$ and $\sqrt{n}(\hat{\beta}(t) - \beta(t))$ as a Bahadur representation for censored quantile regression. Then we would have the desired results by studying the two expansions term by term. A detailed proof is provided in Section 3.5.

3.3 Simulations

In this section, we evaluate the performance of our method in finite samples using Monte Carlo simulations. More specifically we compare the performance of $\mathcal{T}_1$, $\mathcal{T}_2$

and the following test statistic focuses on one quantile level τ,

$$\mathcal{T}_3 = S(\tau)^T Q_n^{-1} S(\tau). \tag{3.9}$$

Notice that $\mathcal{T}_3$ is a special case for $\mathcal{T}_2$ and $\mathcal{T}_2$ when we set $\tau_a = \tau_b$. We also show that the results of our methods are similar whether we use local KM or censored quantile regression as the model to bootstrap C_i^*.

In the simulation we consider the following model

$$\log(T_i) = \beta_0(u_i) + z_{i1}\beta_1(u_i) + z_{i2}\beta_2(u_i) + z_{i3}\beta_3(u_i), \quad i = 1, ..., n, \tag{3.10}$$

where $u_i \sim U(0,1)$. Generate $z_{i1} \sim U(1,3)$ when $z_{i2} < 1$ and $z_{i1} \sim U(0,2)$ when $z_{i2} \geq 1$; z_{i2} and z_{i3} are generated from $U(0,2)$ independently. Let $\beta_0(\tau) = \Phi^{-1}(\tau)$, where $\Phi(\tau)$ is the cdf for the standard normal distribution; $\beta_1(\tau) = \tau^2$. Under the null model, set $\beta_2(\tau) = \beta_3(\tau) = 0$.

We consider 2 cases for $\beta_2(\tau)$ and $\beta_3(\tau)$ under the alternative.

In case (i), let $\beta_2(\tau) = \frac{2\exp(15(\tau-0.5))}{1+\exp(15(\tau-0.5))}$ and $\beta_3(\tau) = \frac{2\exp(10(\tau-0.5))}{1+\exp(10(\tau-0.5))}$. Set $\log(C_i) \sim U(-0.5z_{i1}, 5-0.5z_{i1})$ under H_0 and $\log(C_i) \sim U(2-0.5z_{i1}, 7-0.5z_{i1})$ under H_1. In this case, the effect of z_{i1} and z_{i2} is always positive and is more significant at the upper tail. Case (i) is designed to capture the scenarios when by previous knowledge the effect of z_{i1} and z_{i2} is suspected to be minimal except at upper quantile level and $[\tau_a, \tau_b]$ is chosen to focus on the upper tail (see *He et al.* (2010) for a real example).

For case (ii), let $\beta_2(\tau) = -2\mathbb{I}(\tau < 0.4) + 20(\tau - 0.4)\mathbb{I}(0.4 < \tau < 0.6) + 2\mathbb{I}(\tau > 0.6)$, and $\beta_3(\tau) = -3\mathbb{I}(\tau < 0.4) + 30(\tau - 0.4)\mathbb{I}(0.4 < \tau < 0.6) + 3\mathbb{I}(\tau > 0.6)$. Set $\log(C_i) \sim$

$U(-z_{i1}, 5 - z_{i1})$ under H_0 and $\log(c_i) \sim U(2 - z_{i1}, 7 - z_{i1})$ under H_1. In this case, the effect of z_{i2} and z_{i3} changes from negative to positive as τ increases. Case (ii) is designed to capture the scenarios when the goal is to detect a overall effect of z_{i2} and z_{i3} and $[\tau_a, \tau_b]$ is chosen to cover a relatively large quantile region.

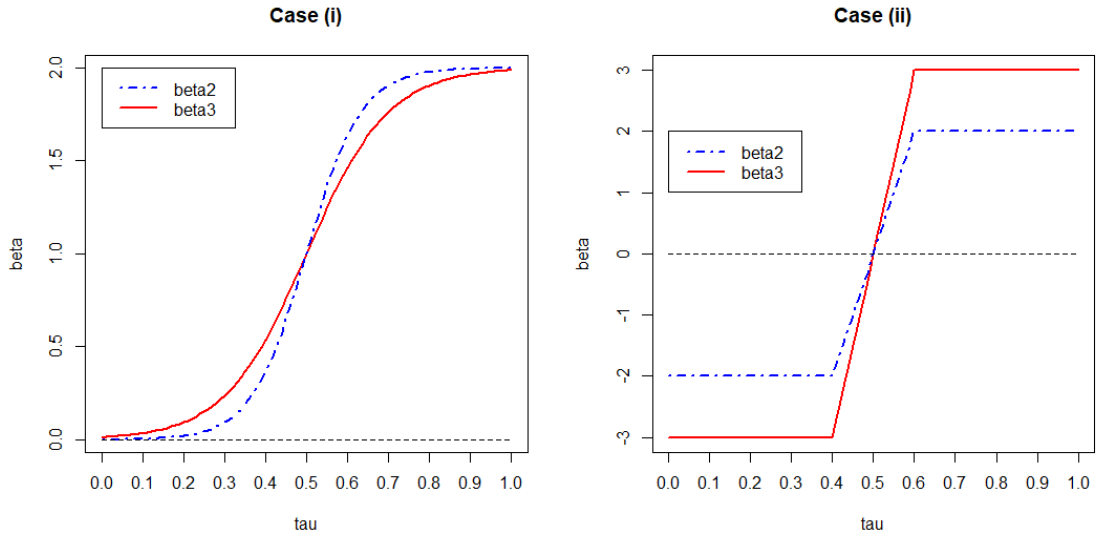

Figure 3.1: Curves of quantile coefficients of case (i) and (ii) under the alternative.

The simulation size is 1000 and the bootstrap sample size is 500 throughout this simulation. The results of our simulation under case (i) and (ii) are summarized in Table 3.1 and 3.2 respectively.

	$n = 100$		$n = 200$		$n = 500$	
	Type I error	Power	Type I error	Power	Type I error	Power
$\mathcal{T}_1^{km}(0.50, 0.85)$	0.046	0.365	0.047	0.651	0.057	0.979
$\mathcal{T}_1^{km}(0.75, 0.85)$	0.056	0.438	0.029	0.798	0.057	0.998
$\mathcal{T}_2^{km}(0.50, 0.85)$	0.049	0.352	0.048	0.647	0.054	0.979
$\mathcal{T}_2^{km}(0.75, 0.85)$	0.056	0.433	0.029	0.797	0.055	0.998
$\mathcal{T}_1^{qr}(0.50, 0.85)$	0.057	0.372	0.048	0.660	0.059	0.981
$\mathcal{T}_1^{qr}(0.75, 0.85)$	0.061	0.440	0.037	0.809	0.056	0.998
$\mathcal{T}_2^{qr}(0.50, 0.85)$	0.061	0.361	0.047	0.653	0.056	0.982
$\mathcal{T}_2^{qr}(0.75, 0.85)$	0.062	0.435	0.036	0.805	0.057	0.998
$\mathcal{T}_3^{qr}(0.50)$	0.044	0.165	0.055	0.280	0.054	0.549
$\mathcal{T}_3^{qr}(0.75)$	0.059	0.431	0.041	0.772	0.057	0.998
$\mathcal{T}_3^{qr}(0.85)$	0.064	0.390	0.046	0.784	0.058	0.998

Table 3.1: Comparison of the empirical type I error rate and power under case (i) out of 1000 simulation samples. $\mathcal{T}_1^{km}(\tau_a, \tau_b)$ stands for the test statistic $\mathcal{T}_1$ over τ in $[\tau_a, \tau_b]$ with C_i^* sampled from the the local KM estimator. Similarly, $\mathcal{T}_2^{qr}[\tau_a, \tau_b]$ stands for the test statistic $\mathcal{T}_2$ over τ in $[\tau_a, \tau_b]$ with C_i^* sampled from the censored quantile regression model. $\mathcal{T}_3^{qr}[\tau]$ stands for the test statistic $\mathcal{T}_3$ at τ.

When the nominal type I error is 0.05, the standard derivation of the empirical type I error is 0.007. From Table 3.1, all the tests we considered seem to be reasonable because the empirical type I errors fall into two standard derivation of 0.05 except two entries. To compare the power, for reference, the largest possible standard derivation for empirical power is 0.016, which is achieved when the true power is 0.5. Whether we use the local KM or the censored quantile regression to sample C_i^* provides similar

results. The performance of $\mathcal{T}_1$ and $\mathcal{T}_2$ are also similar in this setting. Comparing $\mathcal{T}_1/\mathcal{T}_2$ to $\mathcal{T}_3$, we notice that the power of $\mathcal{T}_3$ heavily relies on the chosen quantile level τ and it can be difficult to choose a good quantile level in practice. Though the power of $\mathcal{T}_1/\mathcal{T}_2$ also depends on $[\tau_a, \tau_b]$, it is less sensitive. And the power of $\mathcal{T}_1/\mathcal{T}_2$ targeting the region $[0.75, 0.85]$ is higher than $\mathcal{T}_3$ at 0.75 or 0.85 level. This illustrates the advantage of considering a quantile region instead of a single τ in global quantile regression.

From Table 3.2, the major distinction from case (ii) to case (i) is that in case (ii), $\mathcal{T}_2$ has higher power than $\mathcal{T}_1$ under the same $[\tau_a, \tau_b]$. It is because by design, the signs of $\beta_2(\tau)$ and $\beta_3(\tau)$ changes from negative to positive as τ increases. Thus the negative effect near τ_a and the positive effect near τ_b is averaged out to some degree when $\mathcal{T}_1$ is used. This problem is avoided when $\mathcal{T}_2$ is used instead. Therefore according to our simulation results, $\mathcal{T}_2$ is preferable to the more natural $\mathcal{T}_1$ overall.

	$n = 100$		$n = 200$		$n = 500$	
	Type I error	Power	Type I error	Power	Type I error	Power
$\mathcal{T}_1^{km}(0.40, 0.60)$	0.043	0.079	0.058	0.097	0.042	0.098
$\mathcal{T}_1^{km}(0.10, 0.70)$	0.046	0.350	0.057	0.621	0.049	0.935
$\mathcal{T}_2^{km}(0.40, 0.60)$	0.042	0.104	0.057	0.164	0.057	0.340
$\mathcal{T}_2^{km}(0.10, 0.70)$	0.044	0.763	0.063	1.000	0.051	1.000
$\mathcal{T}_1^{qr}(0.40, 0.60)$	0.044	0.076	0.059	0.096	0.046	0.097
$\mathcal{T}_1^{qr}(0.10, 0.70)$	0.045	0.351	0.063	0.625	0.051	0.936
$\mathcal{T}_2^{qr}(0.40, 0.60)$	0.047	0.102	0.061	0.160	0.044	0.343
$\mathcal{T}_2^{qr}(0.10, 0.70)$	0.046	0.766	0.064	1.000	0.053	1.000
$\mathcal{T}_3^{qr}(0.30)$	0.040	0.770	0.060	0.969	0.056	0.999
$\mathcal{T}_3^{qr}(0.50)$	0.043	0.077	0.054	0.076	0.049	0.060
$\mathcal{T}_3^{qr}(0.70)$	0.039	0.116	0.045	0.240	0.064	0.623

Table 3.2: Comparison of the empirical type I error rate under case (ii) out of 1000 simulation samples.

3.4 Natural mortality in bighorn sheep

In this section, we apply our method to study the effect of early environment conditions on the natural mortality of adult bighorn sheep using the data analyzed in *Douhard et al.* (2019)[1]. The data set contains the survival time of 351 bighorn sheep born at Ram Mountain in Alberta, Canada, from 1973 to 2010. Other covariates

[1]The data set is available for download from the Dryad Digital Repository: *https://doi.org/10.5061/dryad.6bm4228*

included in the data set are sex, adult environment condition and the indicator of whether cougar predation exists. The environment condition is measured as the 3-year average of the average mass of the 15-month-old yearlings. Because we are interested in the natural mortality rate, the lifetimes of sheep that were shot by hunters are considered as censored. In the data set, 19 out of 191 female sheep are censored and 53 out of 160 male sheep are censored.

We use the log of survival time as the response and sex, cougar, early environment condition, adult environment condition and the interaction between sex and the early environment condition as predictors. Results from the Cox proportional hazard model used in *Douhard et al.* (2019) show that female sheep with a better early environment tend to live longer (p-value = 0.0042). But this phenomenon is not observed for male (p-value = 0.1747), though the interaction between sex and early environment is also not significant (p-value = 0.4341). This seemingly contradicting result may imply that the test does not have enough power to detect the early environment effect on male or the interaction between sex and early environment.

The Cox proportional hazard model assumes that the effect of a covariate on the hazard ratio is a constant, which precludes many forms of heterogeneity. Alternatively, we fit the model with censored quantile regression with the same covariates and the results are shown in Figure 3.2. We first notice that the estimated coefficients for adult environment condition and cougar are non-constant over τ. This implies that heterogeneity exists and the Cox proportional hazard model may not be adequate. According to the figure, the effect of early environment condition for female is significant for a large range of τ. But the effect of early environment con-

dition for male is only significant at τ around 0.2. It is very difficult to detect the early environment effect on male if one only looks at single τ because it is hard to know which τ to look at beforehand and multiplicity adjustment would be needed if one conduct test at several quantile levels individually.

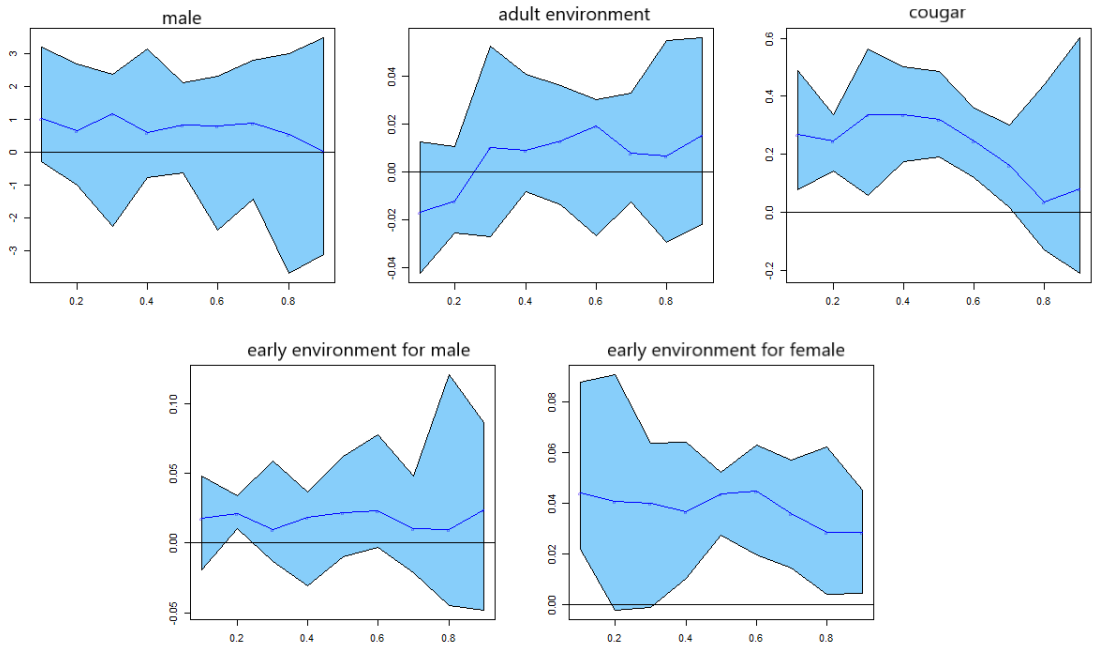

Figure 3.2: Pointwise confidence band for the censored quantile regression model coefficients.

In the next step, we conduct the proposed rank-based test. Because we aim to test the overall effect of early environment conditions on male/female, we should choose $[\tau_a, \tau_b]$ to cover a large quantile region. With $\mathcal{T}_2$ and $[\tau_a, \tau_b] = [0.01, 0.8]$, we detect that the interaction between early environment and sex is significant (p-value = 0.016). Furthermore, with $\mathcal{T}_2$ and $[\tau_a, \tau_b] = [0.01, 0.8]$, we detect that male with

better early environment tends to live longer (p-value = 0.020), and the same holds for female with p-value = 0.000. According to our analysis, good early environment condition has a positive effect on the survival time for both male and female sheep, and the effect on female sheep is greater than the male sheep. We are able to arrive at the same conclusion if $\mathcal{T}_1$ is used or we set $[\tau_a, \tau_b]$ to be other quantile regions like [0.1,0.8], [0.1,0.7], etc., indicating the robustness of the proposed test over the choices of the quantile regions.

3.5 Proof

In this section we present the proof of Theorem 3.1. when Portnoy's method is used for the estimation.

Notice that the weight $\hat{w}_i$ defined in (3.3) for our test is slightly differently from the weight defined in *Portnoy and Lin* (2010) for the estimation. Suppose C_i is first crossed between $x_i^T\hat{\beta}(t_j)$ and $x_i^T\hat{\beta}(t_{j+1})$, this observation is actually considered as uncensored at t_{j+1}th quantile (namely $\hat{w}_i(t_{j+1}) = 1$) by *Portnoy and Lin* (2010). Because by their estimation algorithm only $\hat{\beta}(t_1)$,...,$\hat{\beta}(t_j)$ have been obtained at this point and C_i has not been crossed by t_jth quantile. For our inference $\hat{\beta}(\cdot)$ has already been estimated at all the grid points and $\hat{\tau}_i$ and $\hat{w}_i$ are calculated by (3.2) and (3.3).

Throughout the proof, $\hat{w}_i$ will present the weight describe in *Portnoy and Lin* (2010) unless otherwise distinguished as $\hat{w}_i^{PL}$ (weight defined in *Portnoy and Lin* (2010)) or $\hat{w}_i^{SH}$ (weight defined in (3.3)).

The proof is established in four steps.

3.5.1 Step 1: establish the distribution of the test statistics $\mathcal{T}_1$ and $\mathcal{T}_2$

We fit the censored quantile regression model model with only x_{i1} as the covariate and assume H_0 is true.

For censored observation i, define τ_i such that $x_{i1}^T\beta(\tau_i) = C_i$. Let w_i be the true weight where $\hat{r}_i$ in (3.3) is replaced with τ_i. Let d_i be a random vector with bounded support. Write

$$\Psi_{k+1}(w(\underline{\beta}_k, t_{k+1}), b) = \sum_i d_i \Big(\mathbb{I}(\Delta_i = 1)\psi(Y_i - x_{i1}^T b, t_k) \\ + \mathbb{I}(\Delta_i = 0)\big(w(\underline{\beta}_k, t_{k+1})\psi(C_i - x_{i1}^T b, t_k) - (1 - w(\underline{\beta}_k, t_{k+1}))\psi(Y_i^* - x_{i1}^T b, t_k)\big)\Big), \tag{3.11}$$

where $\underline{\beta}_k$ denotes $\beta(\cdot)$ evaluated at grid points $t_0, ..., t_k$ and $\psi(u, t) = t - \mathbb{I}(u < 0)$. By equation (13) of *Portnoy and Lin* (2010), for $\|\theta - \beta(t_{l+1})\| = O(n^{-1/2})$ and any $l < M$,

$$\Psi_{l+1}(w(\underline{\beta}_l, t_{l+1}), \theta)) - \Psi_{l+1}(w(\underline{\beta}_l, t_{l+1}), \beta(t_{l+1})) - D^T V X(\theta - \beta(t_{l+1})) = O_p(n^{1/4}\log n), \tag{3.12}$$

where V is a diagonal matrix with $V_{ii} = f_i(x_{i1}^T\beta(t_{l+1}))[1 - G_i(x_{i1}^T\beta(t_{l+1})]$ and D is a $n \times p$ matrix with d_i^T as the ith row.

By same argument as equation (14) of *Portnoy and Lin* (2010), we have

$$\Psi_{l+1}(w(\hat{\underline{\beta}}_l, t_{l+1}), \theta)) = \sum_{i \in CI_l} d_i(\hat{w}_i - w_i)\mathbb{I}(C_i < x_{i1}^T\hat{\theta}) + \Psi_{l+1}(w(\underline{\beta}_l, t_{l+1}), \beta(t_{l+1})) \\ - D^T V X(\theta - \beta(t_{l+1})) + O_p(n^{1/4}\log n). \tag{3.13}$$

Set $\theta = \hat{\beta}_{t_{l+1}}$. If the weight w is defined as our paper, we have $\Psi_{l+1}(w^{SH}(\underline{\hat{\beta}}_l, t_{l+1}), \theta)) = \sum_i (\hat{a}_i^c(t_{l+1}) + 1 - t)$. For a censored observation i, $\hat{w}_i^{PL}(t_{l+1})$ is different from $\hat{w}_i^{SH}(t_{l+1})$ if the observation is first crossed between t_l and t_{l+1} level, which is of order δ_n. And in this case, $1 - \hat{w}_i^{PL}(\underline{\hat{\beta}}_l, t_{l+1})\mathbb{I}(Y_i - x_{i1}^T\hat{\beta}(t_{l+1}) < 0) = 0$ and $\hat{a}_i^c(t_{l+1}) = O(t_{l+1} - \hat{\tau}_i^{SH}) = O(\delta_n)$. Therefore $\Psi_{l+1}(w^{PL}(\underline{\hat{\beta}}_l, t_{l+1}), \theta)) = \sum_i (\hat{a}_i^c(t_{l+1}) + 1 - t) + O(n\delta_n)$.

Let $a_i^c(t) := 1 - w_i(t)\mathbb{I}(Y_i - x_i^T\beta(t) < 0)$, by direct calculation we have at each $t = t_k$,

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_i d_i(1 - t - \hat{a}_i^c(t)) =& \frac{1}{\sqrt{n}}\sum_i d_i(1 - t - a_i^c(t)) \\ &+ \big(\frac{1}{n}\sum_i d_i x_{i1}^T f_i(x_{i1}^T\beta(t))[1 - G_i(x_{i1}^T\beta(t)]\big)\sqrt{n}(\hat{\beta}(t) - \beta(t)\big) \\ &+ \frac{1}{n}\sum_i d_i \frac{\sqrt{n}(\hat{\tau}_i - \tau_i)}{(1-\tau_i)^2}\mathbb{I}(Y_i > C_i)\mathbb{I}(x_{i1}^T\hat{\beta}(t) \geq C_i) + o_p(1).\end{aligned} \tag{3.14}$$

Notice that terms similar to the first two terms also appears in the derivation of quantile regression without censoring while the third term appears because the true weight w_i is estimated by $\hat{w}_i$. It is easy to see that the first term $W_{n,d}(t) := \frac{1}{\sqrt{n}}\sum_i d_i(1 - t - a_i^c(t))$ converges to a zero mean Gaussian process $W(t)$. It is shown in Portnoy and Lin (2010) that the third term $DT_{n,d}(t) := \frac{1}{n}\sum_i d_i \frac{\sqrt{n}(\hat{\tau}_i - \tau_i)}{(1-\tau_i)^2}\mathbb{I}(Y_i > C_i)\mathbb{I}(x_{i1}^T\hat{\beta}(t) \geq C_i)$ converges to

$$DT_d(t) = \int_0^t B_n(u)\Gamma_d(u)du + o_p(1), \tag{3.15}$$

where

$$\Gamma_d(t) = \frac{g_i(x_{i1}^T\beta(t))}{(1-t)(1 - G_i(x_{i1}^T\beta(t)))}\mathbb{E}(d_i x_i^T), \tag{3.16}$$

and

$$B_n(t) = \sqrt{n}(\hat{\beta}(t) - \beta(t)). \tag{3.17}$$

Let $d_i = x_{i1}$ in (3.14), we have

$$D(t)B_n(t) = \int_0^t B_n(u)\Gamma_{x_1}(u)du + W_{n,x_1}(t) + o_p(1), \tag{3.18}$$

where $D(t) = \lim \frac{1}{n}\sum_i x_{i1}x_{i1}^T f_i(x_{i1}^T\beta(t))[1 - G_i(x_{i1}^T\beta(t)]$.

Since $W_{n,x_{i1}}(t)$ converges to a zero mean Gaussian process $W_{x_{i1}}(t)$, we have under the null hypothesis $B_n(t)$ converges weakly to a Gaussian process $B(t)$ satisfying

$$D(t)B(t) = \int_0^t B(u)\Gamma_{x_1}(u)du + W_{x_1}(t). \tag{3.19}$$

Let $d_i = x_{i2} - \hat{x}_{i2}$, equation (3.14) becomes

$$\frac{1}{\sqrt{n}}\sum_i (x_{i2} - \hat{x}_{i2})\hat{a}_i^c(t) = W_{n,x_{i2}-\hat{x}_{i2}}(t) + D_{x_{i2}-\hat{x}_{i2}}(t)B_n(t) + DT_{n,x_{i2}-\hat{x}_{i2}}(t) + o_p(1). \tag{3.20}$$

Thus under the null hypothesis, $\frac{1}{\sqrt{n}}\sum_i (x_{i2} - \hat{x}_{i2})\hat{a}_i^c(t)$ converges to a zero mean Gaussian process. Therefore $\mathcal{T}_1$ will converge to a mixed chi-square distribution as a weighted sum of q chi-square variables of $df = 1$. The limiting distribution of $\mathcal{T}_2$ is more complicated, which the time intergral of a squared Gaussian process. For our concern, we do not need to know the exact distribution of $\mathcal{T}_1$ and $\mathcal{T}_2$ since we will approximate their distribution under the null hypothesis using the bootstrap.

3.5.2 Step 2: establish the consistency of $\hat{\beta}^*$ and the bootstrap version of equation (3.14).

Let $\hat{\tau}_i^*$ and τ_i^* satisfy $x_{i1}^T\tilde{\beta}^*(\hat{\tau}_i) = C_i^*$ and $x_{i1}^T\tilde{\beta}(\tau_i) = C_i^*$ respectively.

For the subsequent derivations, we restrict our analysis on the set where $x_i^T\tilde{\beta}(\tau)$ is monotone in τ, which is true with probability tending to 1 as shown in *Portnoy and Lin* (2010). Within this set, T_i^* is generated from a valid quantile process $x_i^T\tilde{\beta}(\tau)$ and many arguments in Portnoy and Lin (2010) still hold in the bootstrap space. Following *Portnoy and Lin* (2010), We shall show by induction that for $k = 1, ..., M$

$$\sum_{i \in CI_k} |\hat{\tau}_i^* - \tau_i^*| \leq d_{k,n} \tag{3.21}$$

$$\|\hat{\beta}^*(t_k) - \hat{\beta}(t_k)\| \leq 2r_1 n^{-1} d_{k,n} \tag{3.22}$$

where $d_{k,n} = R_n\sqrt{n}(1 + 2r_1r_2E_n^*\delta_n)^{k-1}$, $R_n = n^{-1/2}\|\Psi_{l+1}^*(w(\underline{\hat{\beta}}_l, t_{l+1}), \hat{\beta}(t_{l+1}))\|$, $E_n^* = O_{p^*}(1)$ is a random bound, r_1 and r_2 are constant given in the derivation below.

First consider $k = 1$, by our bootstrap design where is no censoring for $\tau \leq t_1$, thus $\sum_{i \in CI_1} |\hat{\tau}_i^* - \tau_i^*| = 0$. Since there is no censoring at t_1 level, $\|\hat{\beta}^*(t_1) - \hat{\beta}(t_1)\| \leq 2r_1n^{-1}d_{1,n}$ is given by Theorem 2.1 where the root-n consistency of $\hat{\beta}^*$ for the model-based bootstrap without censoring is proved. Assume (3.21) and (3.22) are satisfied when $k = l$. At t_{l+1} level, let CI_l be the set of censored observations that have been

crossed at t_lth level,

$$\begin{aligned}\sum_{i=1}^{n}|w_i^*(\hat{\underline{\beta}}_l^*,t_{l+1})-w_i^*(\hat{\underline{\beta}}_l,t_{l+1})| &= \sum_{i\in CI_l}|w_i^*(\hat{\underline{\beta}}_l^*,t_{l+1})-w_i^*(\hat{\underline{\beta}}_l,t_{l+1})| \\ &\quad + \sum_{x_{i1}^T\hat{\beta}^*(t_l)<C_i^*<x_{i1}^T\hat{\beta}(t_l)}|w_i^*(\hat{\underline{\beta}}_l^*,t_{l+1})-w_i^*(\hat{\underline{\beta}}_l,t_{l+1})| \\ &= \sum_{i\in CI_l}\frac{(1-t_{l+1}|\hat{\tau}_i^*-\tau_i^*|)}{(1-\hat{\tau}_i^*)}+\sqrt{n}E_n\delta_n \\ &\le \sum_{i\in CI_l}\frac{1-\epsilon}{\epsilon^2}|\hat{\tau}_i^*-\tau_i^*|+\sqrt{n}E_n\delta_n \\ &\le \frac{1-\epsilon}{\epsilon^2}d_{l,n}(1+\tilde{E}_n\delta_n),\end{aligned} \tag{3.23}$$

where E_n and $\tilde{E}_n$ are two random bounds. By Lemma 4.1 of *He and Shao* (1996), we have on $\{\theta:\|\theta-\hat{\beta}(t_{l+1})\|\le Kn^{-1/2}\}$,

$$\begin{aligned}&\Psi_{l+1}^*\big(w(\hat{\underline{\beta}}_l,t_{l+1}),\theta\big)-\Psi_{l+1}^*\big(w(\hat{\underline{\beta}}_l,t_{l+1}),\hat{\beta}(t_{l+1})\big) \\ &\qquad -\mathbb{E}\bigg(\Psi_{l+1}^*\big(w(\hat{\underline{\beta}}_l,t_{l+1}),\theta\big)-\Psi_{l+1}^*\big(w(\hat{\underline{\beta}}_l,t_{l+1}),\hat{\beta}(t_{l+1})\big)\bigg)=O_p^*(n^{1/4}\log n).\end{aligned} \tag{3.24}$$

Studying the expectation term in the above equation,

$$
\begin{aligned}
&\mathbb{E}\Big(\Psi^*_{l+1}\big(w(\underline{\hat{\beta}}_l, t_{l+1}), \theta\big) - \Psi^*_{l+1}(w(\underline{\hat{\beta}}_l, t_{l+1}), \hat{\beta}(t_{l+1}))\Big)\\
&= \sum_i d_i \mathbb{E}\Big((\mathbb{I}(T_i^* \le x_i^T\theta) - \mathbb{I}(T_i^* \le x_i^T\hat{\beta}(t_{l+1})))\mathbb{I}(T_i^* \le C_i^*)\\
&\quad + w_i^*(t_{l+1})(\mathbb{I}(C_i^* \le x_i^T\theta) - \mathbb{I}(C_i^* \le x_i^T\hat{\beta}(t_{l+1})))\mathbb{I}(T_i^* > C_i^*)\Big)\\
&= \sum_i d_i \mathbb{E}\Big((\mathbb{I}(T_i^* \le x_i^T\theta) - \mathbb{I}(T_i^* \le x_i^T\hat{\beta}(t_{l+1})))\mathbb{I}(x_i^T\hat{\beta}(t_{l+1}) \le C_i^*)\\
&\quad + w_i^*(t_{l+1})(\mathbb{I}(C_i^* \le x_i^T\theta) - \mathbb{I}(C_i^* \le x_i^T\hat{\beta}(t_{l+1})))\mathbb{I}(T_i^* > x_i^T\hat{\beta}(t_{l+1}))\Big) + O_{p^*}(1)\\
&= \sum_i d_i \mathbb{E}\big((\mathbb{I}(T_i^* \le x_i^T\theta) - \mathbb{I}(T_i^* \le x_i^T\hat{\beta}(t_{l+1})))\mathbb{I}(x_i^T\hat{\beta}(t_{l+1}) \le C_i^*)\big) + O_{p^*}(1)\\
&= \sum_i d_i \underbrace{\mathbb{E}\big(\mathbb{I}(T_i^* \le x_i^T\theta) - \mathbb{I}(T_i^* \le x_i^T\hat{\beta}(t_{l+1}))\big)}_{I}\underbrace{\mathbb{E}(\mathbb{I}(x_i^T\hat{\beta}(t_k) \le C_i^*))}_{II} + O_{p^*}(1).
\end{aligned}
\tag{3.25}
$$

The second equality follows because the probability that both T_i^* and C_i^* are between $x_i^T\theta$ and $x_i^T\hat{\beta}$ is of order n^{-1} for $\|\theta - \hat{\beta}(t_{l+1})\| = O(n^{-1/2})$. The third equation follows because the probability that C_i^* is between $x_i^T\theta$ and $x_i^T\hat{\beta}$ is of order $n^{-1/2}$. And w_i^* is $O_{p^*}(n^{-1/2})$ for such terms.

Now we want to calculate the expectation of I and II. By our bootstrap $T_i^* = x_i^T\tilde{\beta}(u_i)$ for $2\epsilon < u_i < \min(1-\epsilon, \tau_U)$. When $u_i < 2\epsilon$ or $u_i > 1-\epsilon$, it is impossible for T_i^* to lie between $x_i^T\theta$ and $x_i^T\hat{\beta}$ with probability tending to 1 by the asymptotic monotonicity of $\tilde{\beta}(\cdot)$.

Let $\Delta = \tilde{\beta}(u_i) - \beta(u_i)$, we have for $2\epsilon < u_i < \min(1-\epsilon, \tau_U)$

$$\begin{aligned} I &= \mathbb{E}\big(\mathbb{I}(x_i^T\theta - x_i^T\Delta < x_i^T\beta(u_i) < x_i^T\hat{\beta}(t_{l+1}) - x_i^T\Delta) \\ &= \int_{x_i^T\theta - x_i^T\Delta}^{x_i^T\hat{\beta}(t_{l+1}) - x_i^T\Delta} f_i(c)dc \\ &= \int_{x_i^T\theta - x_i^T\Delta}^{x_i^T\hat{\beta}(t_{l+1}) - x_i^T\Delta} f_i(x_i^T\beta(t_{l+1})) + O(c - x_i^T\beta(t_{l+1}))dc \\ &= f_i(x_i^T\beta(t_{l+1}))(x_i^T\hat{\beta}(t_{l+1}) - x_i^T\theta) + O(n^{-1}). \end{aligned} \tag{3.26}$$

Now consider II. By our bootstrap design $C_i^* = \hat{G}^{-1}(v_i|x_{i1}, x_{i2})$ for $v_i < \tau_{Vi}$, where τ_{Vi} is the largest value $G_i^{-1}(\cdot)$ is identifiable. Notice that $\tau_{Vi} > G_i(x_i^T\beta(t_{l+1}))$ because both censored and uncensored outcome can be observed at t_{l+1} level. When $v_i > \tau_{Vi}$, $x_i^T\hat{\beta}(t_{l+1}) \leq C_i^*$ since we impute a very large value for C_i^*. Thus

$$\mathbb{P}(C_i^* < x_i^T\hat{\beta}(t_{l+1})) = \mathbb{P}(v_i < \hat{G}_i(x_i^T\hat{\beta}(t_{l+1}))) = \hat{G}_i(x_i^T\hat{\beta}(t_{l+1})). \tag{3.27}$$

By Theorem 2.1 of *Gonzalez-Manteiga and Cadarso-Suarez* (1994),

$$\sup_t \sup_x |\hat{G}(t|x) - G(t|x)| = O_p((\log n)^{1/2} n^{-1/4-\gamma_0/2}), \tag{3.28}$$

where $0 < \gamma_0 < 1/4$. Thus

$$II = 1 - \hat{G}(x_i^T\hat{\beta}(t_{l+1})) = 1 - G(x_i^T\beta(t_{l+1})) + O_p(n^{-1/4}\log n), \tag{3.29}$$

and

$$\begin{aligned}&\mathbb{E}\bigg(\Psi^*_{l+1}\big(w(\underline{\hat{\beta}}_l, t_{l+1}), \theta\big) - \Psi^*_{l+1}(w(\underline{\hat{\beta}}_l, t_{l+1}), \hat{\beta}(t_{l+1}))\bigg)\\ &= \sum_i d_i f_i(x_i^T\beta(t_{l+1})(1 - G(x_i^T\beta(t_{l+1})(x_i^T\hat{\beta}(t_{l+1}) - x_i^T\theta) + O_p(n^{1/4}\log n).\end{aligned} \tag{3.30}$$

Then we have

$$\begin{aligned}\Psi^*_{l+1}(w(\underline{\hat{\beta}}^*_l, t_{l+1}), \theta)) = &\sum_{i\in CI^*_l} d_i(\hat{w^*_i} - w^*_i)\mathbb{I}(C^*_i < x_i^T\theta) + \Psi^*_{l+1}(w(\underline{\hat{\beta}}_l, t_{l+1}), \hat{\beta}(t_{l+1}))\\ &- D^TVX(\theta - \hat{\beta}(t_{l+1})) + O_p(n^{1/4}\log n) + o^*_p(1).\end{aligned} \tag{3.31}$$

Set $\theta = \hat{\beta}^*(t_{l+1})$ and $d_i = x_i$. (This is possible because if $\|\hat{\beta}^*(t_{l+1}) - \hat{\beta}(t_{l+1})\| \geq Cn^{-1/2}$ for C large enough, the gradient condition can not be satisfied by (3.31).)

$$\begin{aligned}\|\hat{\beta}^*(t_{l+1}) - \hat{\beta}(t_{l+1})\| = \|(X^TVX)^{-1}\bigg(&\sum_{i\in CI^*_l} x_i(\hat{w^*_i} - w^*_i)\mathbb{I}(C^*_i < x_i^T\hat{\beta}^*(t_{l+1}))\\ &- \Psi^*_{l+1}(w(\underline{\hat{\beta}}^*_l, t_{l+1}), \hat{\beta}^*(t_{l+1})) + \Psi^*_{l+1}(w(\underline{\hat{\beta}}_l, t_{l+1}), \hat{\beta}(t_{l+1})) + O_p(n^{1/4}\log n) + o^*_p(1)\bigg).\end{aligned} \tag{3.32}$$

By (B4) and (B6), there exist a $a > 0$ such that the biggest eigen value of $(X^TVX)^{-1} \leq an^{-1}$. Let

$$r_1 = an^{-1}\frac{1-\epsilon}{\epsilon^2}, \tag{3.33}$$

we have

$$\begin{aligned}\|\hat{\beta}^*(t_{l+1}) - \hat{\beta}(t_{l+1})\| \leq & an^{-1}\Bigg(\sum_{i\in CI_l^*} (\hat{w_i^*} - w_i^*) + \|\Psi_{l+1}^*(w(\hat{\underline{\beta}}_l^*, t_{l+1}), \hat{\beta}^*(t_{l+1}))\| \\ & + \|\Psi_{l+1}^*(w(\hat{\underline{\beta}}_l, t_{l+1}), \hat{\beta}(t_{l+1}))\| + O_p(n^{1/4}\log n)\Bigg) \\ \leq & an^{-1}\left(\frac{1-\epsilon}{\epsilon^2} d_{l,n}(1+\tilde{E}_n\delta_n) + n^{1/2}R_n\right) \\ \leq & r_1 n^{-1} d_{l,n}(1+\tilde{E}\delta_n) + r_1 n^{-1} d_{l,n} \\ \leq & 2r_1 n^{-1} d_{l+1,n}.\end{aligned} \tag{3.34}$$

for $E_n^* \geq \tilde{E}_n/2r_1r_2$. This shows that (3.22) holds. To show equation (3.21) is correct, consider

$$\sum_{i\in CI_{l+1}} |\hat{\tau}_i^* - \tau_i^*| \leq \sum_{i\in CI_l} |\hat{\tau}_i^* - \tau_i^*| + \sum_i |\hat{\tau}_i^* - \tau_i^*| \mathbb{I}(x_i^T\hat{\beta}^*(t_l) < C_i^* < x_i^T\hat{\beta}^*(t_{l+1})). \tag{3.35}$$

We aim to bound the last term in the above equation.

Let $j = j(i)$ such that $t_j \leq \hat{\tau}_i^* \leq t_{j+1}$. Since both $x_i^T\tilde{\beta}^*(\hat{\tau}_i^*)$ and $x_i^T\tilde{\beta}(\tau_i^*)$ equal to C_i^*, we have

$$0 = x_i^T\big(\tilde{\beta}^*(\hat{\tau}_i^*) - \tilde{\beta}(\hat{\tau}_i^*)\big) + x_i^T\big(\tilde{\beta}(\hat{\tau}_i^*) - \tilde{\beta}(\tau_i^*)\big). \tag{3.36}$$

Define $\hat{\alpha}_i^*$ such that $\tilde{\beta}^*(\hat{\tau}_i^*) = \hat{\beta}^*(t_j) + \hat{\alpha}_i^*(\hat{\beta}^*(t_{j+1}) - \hat{\beta}^*(t_j))$. Expand the first term in (3.36) as

$$x_i^T\big(\tilde{\beta}^*(\hat{\tau}_i^*) - \tilde{\beta}(\hat{\tau}_i^*)\big) = \hat{\alpha}_i^* x_i^T(\hat{\beta}^*(t_j) - \hat{\beta}(t_j)) + (1-\hat{\alpha}_i^*)x_i^T(\hat{\beta}^*(t_{j+1}) - \hat{\beta}(t_{j+1})). \tag{3.37}$$

Let $h_i(\cdot)$ be the right derivative of $x_i^T\tilde{\beta}(\cdot)$, by Taylor expansion, with probability 1,

$$x_i^T\big(\tilde{\beta}(\hat{\tau}_i^*) - \tilde{\beta}(\tau_i^*)\big) = (\hat{\tau}_i^* - \hat{\tau}_i)h_i(t_j) + O(\delta_n^2). \tag{3.38}$$

Thus, we have

$$\sqrt{n}(\hat{\tau}_i^* - \tau_i^*) = h_i(t_j)x_i^T B_{i.j}^* + O(\delta^2), \tag{3.39}$$

where

$$B_{i,j}^* = \sqrt{n}\bigg(\hat{\alpha}_i^*(\hat{\beta}^*(t_j) - \hat{\beta}(t_j)) + (1-\hat{\alpha}_i^*)(\hat{\beta}^*(t_{j+1}) - \hat{\beta}(t_{j+1}))\bigg). \tag{3.40}$$

Therefore

$$\begin{aligned}
\sum_{i\in CI_{l+1}} |\hat{\tau}_i^* - \tau_i^*| \leq & d_{l,n} + \sum_i \big(n^{-1/2}h_i(t_j)x_i^T B_{i,j}^* \mathbb{I}(x_i^T\hat{\beta}^*(t_l) < C_i^* < x_i^T\hat{\beta}^*(t_{l+1}))\big) + O_{p^*}(n\delta_n^2) \\
\leq & d_{l,n} + \sum_i \big(r_2 n^{-1/2} x_i^T B_{i,j}^* \delta_n\big) \\
\leq & d_{l,n} + 2r_1 r_2 d_{l,n}(1+\tilde{E}_n\delta_n)\delta_n \\
\leq & d_{l+1,n}.
\end{aligned} \tag{3.41}$$

In the second line we replace $\mathbb{I}(x_i^T\hat{\beta}^*(t_l) < C_i^* < x_i^T\hat{\beta}^*(t_{l+1}))$ with its expectation

which is of order δ_n. The error incurred by this replacement is dominated by $d_{l,n}$

$$\begin{aligned}
&\mathbb{E}\bigg(\sum_i \big(\mathbb{I}(x_i^T\hat{\beta}^*(t_l) < C_i^* < x_i^T\hat{\beta}^*(t_{l+1})) - \mathbb{E}\big(\mathbb{I}(x_i^T\hat{\beta}^*(t_l) < C_i^* < x_i^T\hat{\beta}^*(t_{l+1}))\big)\big)\bigg)^2 \\
&= \sum_i \mathbb{E}\big(\mathbb{I}(x_i^T\hat{\beta}^*(t_l) < C_i^* < x_i^T\hat{\beta}^*(t_{l+1})) - \mathbb{E}\big(\mathbb{I}(x_i^T\hat{\beta}^*(t_l) < C_i^* < x_i^T\hat{\beta}^*(t_{l+1}))\big)\big)^2 \\
&= O(n\delta_n).
\end{aligned} \tag{3.42}$$

3.5.3 Step 3: study the asymptotic behavior of $DT^*_{n,d}$

From (3.31), we have at each $t = t_k$

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_i d_i(1-t-\hat{a}_i^{c*}(t)) = &\frac{1}{\sqrt{n}}\sum_i d_i(1-t-a_i^{c*}(t)) \\
&+ \big(\frac{1}{n}\sum_i d_i x_i^T f_i(x_i^T\beta(t))[1-G_i(x_i^T\beta(t)]\big)\sqrt{n}(\hat{\beta}^*(t)-\hat{\beta}(t)\big) \\
&+ \frac{1}{n}\sum_i d_i \frac{\sqrt{n}(\hat{\tau}_i^*-\tau_i^*)}{(1-\tau_i^*)^2}\mathbb{I}(Y_i^* > C_i^*)\mathbb{I}(x_i^T\hat{\beta}^*(t) \geq C_i^*) + o_p^*(1),
\end{aligned} \tag{3.43}$$

where $a_i^*(t) = 1 - w_i^*(t)\mathbb{I}(Y_i^* - x_i^T\hat{\beta}(t) < 0)$. Write $W^*_{n,d_i}(t) := \frac{1}{\sqrt{n}}\sum d_i(t - a_i^{c*}(t))$ and $DT^*_{n,d}(t) := \frac{1}{n}\sum_i d_i \frac{\sqrt{n}(\hat{\tau}_i^*-\tau_i^*)}{(1-\tau_i^*)^2}\mathbb{I}(Y_i^* > C_i^*)\mathbb{I}(x_i^T\hat{\beta}^*(t) \geq C_i^*)$, we will study the asymptotic of $DT^*_{n,d}$. The arguments to study $DT_{n,d}$ in *Portnoy and Lin* (2010) can also be adjusted to the bootstrap case.

Let $j = j(i)$ such that $t_j \le \hat{\tau}_i^* \le t_{j+1}$.

$$\begin{aligned} DT_{n,d}^*(t) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} d_i \frac{\sqrt{n}(\hat{\tau}_i^* - \tau_i^*)}{(1-\tau_i^*)^2} \mathbb{I}\big(x_i^T\hat{\beta}^*(t_j) \le C_i^* \le x_i^T\hat{\beta}^*(t_{j+1})\big)\mathbb{I}\big(T_i^* \ge C_i^*\big) + o_p^*(1) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} \frac{d_i x_i^T B_{i,j}^*}{(1-t_j)^2 h_i(t_j)} \mathbb{I}\big(x_i^T\hat{\beta}^*(t_j) \le C_i^* \le x_i^T\hat{\beta}^*(t_{j+1})\big)\mathbb{I}\big(T_i^* \ge C_i^*\big) + o_p^*(1) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} \frac{d_i x_i^T B_{i,j}^*}{(1-t_j)^2 h_i(t_j)} \mathbb{I}\big(x_i^T\hat{\beta}^*(t_j) \le C_i^* \le x_i^T\hat{\beta}^*(t_{j+1})\big)\mathbb{I}\big(T_i^* \ge x_i^T\hat{\beta}^*(t_{j+1})\big) + o_p^*(1). \end{aligned} \tag{3.44}$$

The second equality follows from (3.39) and notice that

$$\frac{1}{1-\tau_i^*} = \frac{1}{1-t_j}\left(1 + \frac{\tau_i^* - \hat{\tau}_i^* + \hat{\tau}_i^* - t_j}{1-\tau_i^*}\right) = \frac{1}{1-t_j}(1 + O(\delta_n)). \tag{3.45}$$

In the third equality we replace $\mathbb{I}\big(T_i^* \ge C_i^*\big)$ with $\mathbb{I}\big(T_i^* \ge x_i^T\hat{\beta}^*(t_{j+1})$. The second line and the third line only differ if T_i^* is between $x_i^T\hat{\beta}^*(t_j)$ and $x_i^T\hat{\beta}^*(t_{j+1})$, which is of order δ_n.

Define the event

$$D_{ij} = \{\hat{\beta}^*(t_l)\}_{l=1}^{j+1} \bigcap \mathbb{I}\big(x_i^T\hat{\beta}^*(t_j) \le C_i^* \le x_i^T\hat{\beta}^*(t_{j+1})\big) \bigcap \mathbb{I}\big(T_i^* \ge x_i^T\hat{\beta}^*(t_{j+1})\big). \tag{3.46}$$

Notice that C_i^* is not used in calculating $\hat{\beta}^*(t_j)$ and $\hat{\beta}^*(t_{j+1})$. Thus given $D_{i,j}$, C_i^* are i.i.d with distribution

$$\begin{aligned} \frac{\hat{G}_i(c) - \hat{G}_i(x_i^T\hat{\beta}^*(t_j))}{\hat{G}_i(x_i^T\hat{\beta}^*(t_{j+1})) - \hat{G}_i(x_i^T\hat{\beta}^*(t_j))} &= \frac{G_i(c) - G_i(x_i^T\hat{\beta}^*(t_j))}{G_i(x_i^T\hat{\beta}^*(t_{j+1})) - G_i(x_i^T\hat{\beta}^*(t_j))} + o_p(1) \\ &= \frac{c - x_i^T\hat{\beta}^*(t_j)}{x_i^T\hat{\beta}^*(t_{j+1}) - x_i^T\hat{\beta}^*(t_j)} + o_p(1). \end{aligned} \tag{3.47}$$

Since $C_i^*|D_{ij}$ is approximately uniform on on $[x_i^T\hat{\beta}^*(t_j), x_i^T\hat{\beta}^*(t_{j+1})]$, $\hat{\alpha}_i^*$ in $B_{i,j}^*$ is also approximately uniform. Therefore we want to replace B_{ij}^* with

$$\bar{B}_j^* = \mathbb{E}(B_{ij}^*|D_{ij}) = \frac{\sqrt{n}}{2}((\hat{\beta}^*(t_j) - \hat{\beta}(t_j)) + (\hat{\beta}^*(t_{j+1}) - \hat{\beta}(t_{j+1}))). \tag{3.48}$$

Let d_{ij} be in difference of the ij term of $DT_{n,d}^*(t)$ when B_{ij}^* is replaced with $\bar{B}_j^*$. We have $\mathbb{E}(d_{ij}|D_{ij}) = 0$ and

$$\mathbb{E}(\frac{1}{n}\sum_i\sum_j d_{ij}|D_{ij})^2 =\leq \frac{1}{n^2}\sum_i\sum_{j_1,j_2}\mathbb{E}|d_{i,j_1}d_{i,j_2}| = O(\frac{M^2}{n}). \tag{3.49}$$

Therefore we have

$$\begin{aligned}
DT_{n,d}^*(t) &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^k \frac{d_i x_i^T B_j^*}{(1-t_j)^2 h_i(t_j)}\mathbb{I}\big(x_i^T\hat{\beta}(t_j) \le C_i^* \le x_i^T\hat{\beta}(t_{j+1})\big)\mathbb{I}\big(T_i^* \ge x_i^T\hat{\beta}(t_{j+1})\big) + o_p^*(1) \\
&= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^k \frac{d_i x_i^T B_j^*}{(1-t_j)h_i(t_j)}\frac{\hat{G}(x_i^T\hat{\beta}^*(t_{j+1})) - \hat{G}(x_i^T\hat{\beta}^*(t_j))}{1-\hat{G}_i(x_i^T\hat{\beta}^*(t_j))} + o_p^*(1) \\
&= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^k \frac{d_i x_i^T B_j^*}{(1-t_j)h_i(t_j)}\frac{G(x_i^T\beta(t_{j+1})) - G(x_i^T\beta(t_j))}{1-G_i(x_i^T\beta(t_j))} + o_p^*(1) \\
&= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^k \frac{d_i x_i^T B_j^*}{(1-t_j)h_i(t_j)}\frac{g_i(x_i^T\beta(t_j))h_i(t_j)\delta_n}{1-G_i(x_i^T\beta(t_j))} + o_p^*(1) \\
&= \sum_{j=1}^k B_j^*\delta_n\frac{1}{n}\sum_{i=1}^n \frac{d_i x_i^T g_i(x_i^T\beta(t_j))}{(1-t_j)(1-G(x_i^T\beta(t_j)))} + o_p^*(1).
\end{aligned} \tag{3.50}$$

By LLN, the inner sum converges to $\Gamma_d(t_j)$ in probability and the outer sum is the Riemann sum of integrating $B_n^*(t)\Gamma_d(t)$ from 2ϵ to t, which is equivalent to integrating from 2ϵ to t because $g_i(x_i^T\beta(t)) = 0$ for $t < 2\epsilon$ since there is no censoring below 2ϵ

level. Therefore for fixed $t \in [t_1, t_M]$, $DT^*_{n,d}(t)$ converges to

$$DT^*_d(t) := \int_0^t B^*_n(u)\Gamma_d(u)du + o^*_p(1). \tag{3.51}$$

The above convergence is uniform by tightness argument as Step 7 of*Portnoy and Lin* (2010).

3.5.4 Step 4: establish the conditional distribution of $\mathcal{T}^*_1$ and $\mathcal{T}^*_2$

In equation (3.43), set $d_i = x_{i1}$, we have

$$D(t)B^*_n(t) = \int_0^t B^*_n(u)\Gamma_{x_{i1}}(u)du + W^*_{n,x_{i1}}(t) + o^*_p(1). \tag{3.52}$$

Let $d_i = x_{i2} - \hat{x}_{i2}$, equation (3.43) becomes

$$\frac{1}{\sqrt{n}}\sum_i (x_{i2} - \hat{x}_{i2})\hat{a}^{c*}_i(t) = W^*_{n,x_{i2}-\hat{x}_{i2}}(t) + D_{x_{i2}-\hat{x}_{i2}}(t)B^*_n(t) + DT^*_{n,x_{i2}-\hat{x}_{i2}}(t) + o^*_p(1). \tag{3.53}$$

If we can show that given the data, $\frac{1}{\sqrt{n}}\sum_i (x_{i2} - \hat{x}_{i2})\hat{a}^{c*}_i(t)$ converges to the same process as $\frac{1}{\sqrt{n}}\sum_i (x_{i2} - \hat{x}_{i2})\hat{a}^{c}_i(t)$, then it follows immediately that the conditional distribution of $\mathcal{T}^*_1/\mathcal{T}^*_2$ will converge to the same limiting distribution as $\mathcal{T}_1/\mathcal{T}_2$.

Solving $B_n(t)$ in (3.18) by Theorem 10 in *Gill and Johansen* (1990), we have

$$B_n(t) = D^{-1}(t)W_{n,x_{i1}}(t) + \int_0^t \mathcal{I}(s,t)W_{n,x_{i1}}(s)D^{-1}(s)\Gamma_{x_{i1}}(s)ds + o_p(1), \tag{3.54}$$

where $\mathcal{I}(s,t) = \mathbf{\Pi}_{u\in(s,t]}(I_p + D^{-1}(u))\Gamma_{x_{i1}}(u)du$. Solving $B_n^*(t)$ in (3.52),

$$B_n^*(t) = D^{-1}(t)W_{n,x_{i1}}^*(t) + \int_0^t \mathcal{I}(s,t)W_{n,x_{i1}}^*(s)D^{-1}(s)\Gamma_{x_{i1}}(s)ds + o_p^*(1). \tag{3.55}$$

Thus we only need to look at the limiting distribution of $W_{n,x_{i1}}(t)$ and $W_{n,x_{i1}}^*(t)$, which is relative easy to study since $a_i(t)$ and $a_i^*(t)$ take simpler forms.

By simple calculation, $\mathbb{E}(a_i^c(t)) = 1 - \mathbb{P}(T_i < x_i^T\beta(t)) = 1 - t$. For probability tending to 1, $x_i^T\tilde{\beta}(t)$ is monotone and $\mathbb{E}^{c*}(a_i^*(t)) = 1 - \mathbb{P}(T_i^* < x_i^T\beta(t)) = 1 - t$.

Now consider $\mathbb{E}(1 - a_i^c(t))^2$,

$$\mathbb{E}(1-a_i^c(t))^2 = \tau - \mathbb{P}(T_i > x_i^T\beta(t)|T_i > C_i)\mathbb{P}(T_i < x_i^T\beta(t)|T_i > C_i)\mathbb{P}(T_i > C_i). \tag{3.56}$$

Let u_i and v_i be independent standard uniform distribution

$$\mathbb{P}(T_i > C_i) = \mathbb{P}(v_i < G_i(x_i^T\beta(u_i)) = (1-\tau_U) + \tau_U \int_{2\epsilon}^{\tau_U} G_i(x_i^T\beta(u))du. \tag{3.57}$$

Notice that T_i will always be smaller than C_i when $u_i < 2\epsilon$ since we are assuming no censoring below 2ϵ. And T_i will always be greater than C_i when $u_i > \tau_U$ since τ_U is the highest quantile level where T_i is identifiable.

Calculating $\mathbb{P}(T_i > x_i^T\beta(t)|T_i > C_i)$ and $\mathbb{P}(T_i < x_i^T\beta(t)|T_i > C_i)$, we have

$$\mathbb{E}(1-a_i^c(t))^2 = \tau - \frac{\left((1-\tau_U) + \tau_U \int_t^{\tau_U} G_i(x_i^T\beta(u))du\right)\left(\tau_U \int_{2\epsilon}^t G_i(x_i^T\beta(u))du\right)}{(1-\tau_U) + \tau_U \int_{2\epsilon}^{\tau_U} G_i(x_i^T\beta(u))du}. \tag{3.58}$$

Repeat the same calculation for the bootstrap space,

$$\begin{aligned}\mathbb{E}^*(1-a_i^{c*}(t))^2 &= \tau - \frac{\big((1-\tau_U)+\tau_U\int_t^{\tau_U}\hat{G}_i(x_i^T\tilde{\beta}(u))du\big)\big(\tau_U\int_{2\epsilon}^{t}\hat{G}_i(x_i^T\tilde{\beta}(u))du\big)}{(1-\tau_U)+\tau_U\int_{2\epsilon}^{\tau_U}\hat{G}_i(x_i^T\tilde{\beta}(u))du} \\ &= \mathbb{E}(1-a_i^c(t))^2+o_p(1).\end{aligned} \tag{3.59}$$

Thus $W_{n,x_{i1}}(t)$ and $W^*_{n,x_{i1}}(t)$ converges to same Gaussion process. Therefore $B_n(t)$ and $B_n^*(t)$ converges to same Gaussion process by (3.54) and (3.55), $\frac{1}{\sqrt{n}}\sum_i(x_{i2}-\hat{x}_{i2})\hat{a}_i^{c*}(t)$ and $\frac{1}{\sqrt{n}}\sum_i(x_{i2}-\hat{x}_{i2})\hat{a}_i^{c}(t)$ converges to same Gaussion process by (3.20) and (3.53). Then we have the desired result.

CHAPTER IV

A Two-Stage Model for Genome-Wide Association Study

4.1 Introduction

A major goal of genome-wide association study (GWAS) is to identify the gene markers that are related to a response variable through an exhaustive search among all gene variants available. Besides the genetic covariates, clinical or environmental covariates may also be present in the study. By convention, let G represent the genetic covariates and E represents all the non-genetic covariates in this chapter. It is well-known that many diseases are influenced by the marginal effect of G and E as well as their interactions (*Hunter* (2005)). Therefore it is important to include the G and E interactions into the modeling.

The most common way to deal with the interaction is to add the G $\times$ E terms into the model. However, when the dimension of G is large, the inclusion of the G $\times$ E terms makes the model more complicated to work with. Furthermore, G $\times$ E

only captures one special type of interactions and the true interaction can be more general.

In this chapter, we propose a two-stage model as a solution to the aforementioned problems. In the first stage, we calculate the conditional percentile for each individual adjusting for all the E factors with a global quantile regression model. In the second stage, we select G factors that are associated with the conditional percentile with the least squares regression. We believe that our method is simple to implement and can identify important gene markers where the G and E interactions are taken into account automatically.

In this chapter, we introduce the two-stage model and apply this method to detect genes that are associated with the survival time of lung cancer patients. Future work includes studying the theory and comparing the two-stage model with alternative methods systematically.

4.2 Two-stage model

4.2.1 Model set-up

Imagine two individuals who are exposed to the same environment but differ a lot in their observed outcomes. This difference may be due to chance, but it may also be caused by other factors (e.g., genes) that can not be observed by bare eyes. Notice that the genes are determined when people are born while the environments are factors that people are exposed to later throughout their lifetime. This motivates us to consider the impact of genes and the environment separately in two stages.

To set up the model, consider n i.i.d observations (y_i, x_i, z_i), where $y_i \in \mathbb{R}$ is the outcome of interest, $x_i = (1, x_{i1}, ..., x_{ip})$ is the E factors and $z_i = (1, z_{i1}, ..., z_{iq})$ is the G factors. For now, assume both p and q are finite and less than n. Imagine that for each individual, there exists an unobserved intrinsic score $r_i \sim U(0, 1)$ determined by z_i through the following model:

$$\text{logit}(r_i) = z_i^T \gamma + \epsilon_i, \tag{4.1}$$

where $\gamma = (\gamma_0, \gamma_1, ..., \gamma_q)$ and ϵ_i are i.i.d errors with mean 0. This r_i measures one's susceptibility to a larger y_i determined by the gene factors. We then assume the observed outcome y_i is determined by r_i and x_i combined through the model:

$$y_i = x_i^T \beta(r_i), \tag{4.2}$$

where $\beta(\tau) = (\beta_0(\tau), \beta_1(\tau), ..., \beta_p(\tau))$.

To understand this model better, consider an imaginary scenario where y_i is the yield of a specific type of corns, while x_i is the assignment to a rich land or barren land. In this case, r_i can be interpreted as the yielding ability determined by genes z_i. If a corn has large r_i but is planted in the barren land, its yield should be lower than the yield if it was planted the rich land. But its yield might still be higher than most corns that are also planted in the barren land.

According to our model, for certain $j > 0$, if $\gamma_j = 0$, z_{ij} has no effect on y_i. if $\gamma_j \neq 0$, z_{ij} is associated with y_i, but whether z_{ij} has an interaction with x_i requires further investigation. More specifically, for $\gamma_j \neq 0$, if $\beta_k(\tau)$ is a constant for any

$k = 1, ..., p$, z_{ij} only has a mean effect on y_i. In this context, the mean effect of z_{ij} refers to the effect of z_{ij} on y_i through $\beta_0(\tau)$. On the other hand, if $\beta_k(\tau)$ is not a constant for certain $k > 0$, z_{ij} has an interaction effect with x_{ik} besides the mean effect as long as $\beta_0(\tau)$ is not a constant. If $\beta_k(\tau)$ is not a constant but $\beta_0(\tau)$ is a constant, z_{ij} has an interaction effect with x_{ik} but no mean effect. Since $\beta_k(\tau)$ belongs to wide range of functions, the form of the interaction between x_{ij} and z_{ik} is allowed to be quite flexible when an interaction exists.

In this chapter, we aim to identify genes associated with the outcome instead of studying the interactions between the genes and the environment. Therefore our goal is to select j such that $\gamma_j \neq 0$ for $j = 1, ..., q$. According to the above discussion, when $\gamma_j \neq 0$, z_{ij} has either a mean effect, or an interaction effect with x_i, or both. Therefore compared to the more classical models that does not consider the interaction or only includes $x_{ik}z_{ij}$-types of interactions, our model allows genes that interact with the environments in more flexible ways to be identified.

4.2.2 Model fitting

In the first stage, we work with (4.2) and solve for r_i. This can be achieved by fitting the global quantile regression model

$$Q_{y_i}(\tau|x_i) = x_i^T \beta(\tau) \tag{4.3}$$

to get the estimate $\hat{\beta}(\tau)$ for all τ. Then define

$$\hat{r}_i = \inf\{\tau : x_i^T \hat{\beta}(\tau) \geq y_i\}. \tag{4.4}$$

Truncate $\hat{r}_i$ to $[\epsilon, 1-\epsilon]$ for both computing and theoretical convenience.

In the second stage, we replace r_i in (4.1) with its estimate $\hat{r}_i$ and fit a least squares regression with the model

$$\text{logit}(\hat{r}_i) = z_i^T \gamma + \epsilon_i. \tag{4.5}$$

Though r_i are independent, $\hat{r}_i$ are weakly correlated since they are all estimated with the same data set. With $\text{logit}(\hat{r}_i)$ as responses, $\sqrt{n}(\hat{\gamma} - \gamma)$ still converge to a normal distribution of mean 0. But the variance could be inflated due to the correlation among $\hat{r}_i$ and we estimate the variance by paired bootstrap. We can then conduct the Wald test and claim that z_j is associated with y_i if $\hat{\gamma}_j$ is significantly non-zero.

4.2.3 Extension

4.2.3.1 Model misspecification

In Model (4.1), we assume that $\text{logit}(r_i)$ is linear in z_i. The choice of the logit link function here is quite arbitrary and can possibly be misspecified. *Li and Duan* (1989) studies the behaviours of the regressions when the link function might be misspecified. Suppose the true model takes the general form

$$r_i = g(z_i^T \gamma^*, \epsilon_i), \tag{4.6}$$

where $g(\cdot)$ is an unknown link function. *Li and Duan* (1989) shows that under certain assumptions, $\gamma_j = c\gamma_j^*$ for $j = 1, ..., q$, for some scalar c. Namely the slopes for the misspecified model is proportional to the slopes of the true model. Thus although

the magnitude of γ_j is not interpretable, we can still conduct the hypothesis test $H_0 : \gamma_j = 0,\ j = 1, ..., q$. *Li and Duan* (1989) shows that the Wald test for the above hypothesis has the correct asymptotic distribution under the null with proper scaling, if r_i is observed. Our scenario is more complicated since r_i is replaced by $\hat{r_i}$. We believe that inference is still valid with the paired bootstrap, but more work is required to confirm our conjecture.

4.2.3.2 Censoring in the outcomes

In biomedical studies, it is quite often that the outcome of interest is censored. It is useful to modify our two-stage model to accommodate this scenario.

Recall for the censoring case, T_i denotes the survival time which is censored from the right by C_i, and $Y_i := \min(T_i, C_i)$. In the first stage, we fit a global censored quantile regression model as discussed in Chapter 3 to get the estimate $\hat{\beta}(\tau)$ for any τ. Similarly, we want to define

$$\hat{r}_i = \inf\{\tau : x_i^T \hat{\beta}(\tau) \geq Y_i\}. \tag{4.7}$$

For censored quantile regression, $\beta(\tau)$ may be unidentifiable for $\tau > \tau_U$. Thus it is possible that $x_i^T \hat{\beta}(\tau) \leq Y_i$ for any $\tau \leq \tau_U$ where τ_U is the largest identifiable quantile level. Set $\hat{r}_i = \tau_U$ in this case.

Notice that $\hat{r}_i$ is censored if either Y_i is censored or $\hat{r}_i$ is set to be τ_U. In the second stage, we fit the Cox proportional hazard model

$$\lambda(t) = \lambda_0(t) \exp(\gamma_1 z_{i1} + ... + \gamma_q z_{iq}), \tag{4.8}$$

where $\lambda(t)$ is the hazard ratio for $\hat{r}_i$. We then select $\hat{\gamma}_j$ that is significantly non-zero.

4.2.3.3 High dimension in G

In GWAS studies, it is reasonable to assume that the dimension of the environmental/clinical covariates are finite and small compared to the sample size. But the dimensions of the genetic covariates are usually high. In this scenario, the first stage of the model fitting is unaffected since z_i is not involved.

In the second stage, when the dimension of z_i is high, there are two options. The first option is to fit the least squares regression (4.5) with one z_{ij} as covariates at a time and control the family-wise type I error rate or the false discovery rate. The second option is to use shrinkage methods by working with the loss function

$$\sum \left(\text{logit}(r_i) - z_i^T \gamma\right)^2 + P_\lambda(\gamma), \tag{4.9}$$

where $P_\lambda(\cdot)$ is a penalty like LASSO, adaptive LASSO or SCAD.

4.3 Application to the lung cancer data

Lung cancer is the most common type of cancer worldwide and there has been plenty of research studying the genetic factors associated with the development of lung cancer (*Bossé and Amos* (2018)). In this section, we use the proposed two-stage model to identify the genes that are associated with the survival time of cancer patients with the data set studied in *Shedden et al.* (2008). This data set contains 442 lung cancer subjects with lung adenocarcinomas from 6 contribution hospitals in the

US. The data set includes clinical information such as patients' age, gender, smoking history, the grade of cancer and whether adjuvant chemotherapy is used. The gene expression of 12402 genes is measured at 22283 probes for each subject. For some of the genes, the data set contains its expression level measured at multiple probes. There exist multiple methods in literature to calculate the expression index for each gene from the probe intensity matrix and the singular value decomposition (SVD) is one of them (*Hu et al.* (2006)). Therefore, we use the first principle component of the SVD to represent the gene expression. The outcome of interest is the survival time, which has a 47% censoring rate.

In the first stage, we fit the following censored quantile regression model

$$\log(T_i) \sim \text{Gender} + \text{Race} + \text{Chemotherapy} + \text{Smoke} + \text{Grade} + bs(\text{Age}), \quad (4.10)$$

where Chemotherapy is the indicator of whether adjuvant chemotherapy is used, Smoke is a categorical variable recording the smoking history and Grade is a categorical variable recording the grade of cancer. The only continuous variable is Age and we model it here with a B-spline basis with one knot located at the median.

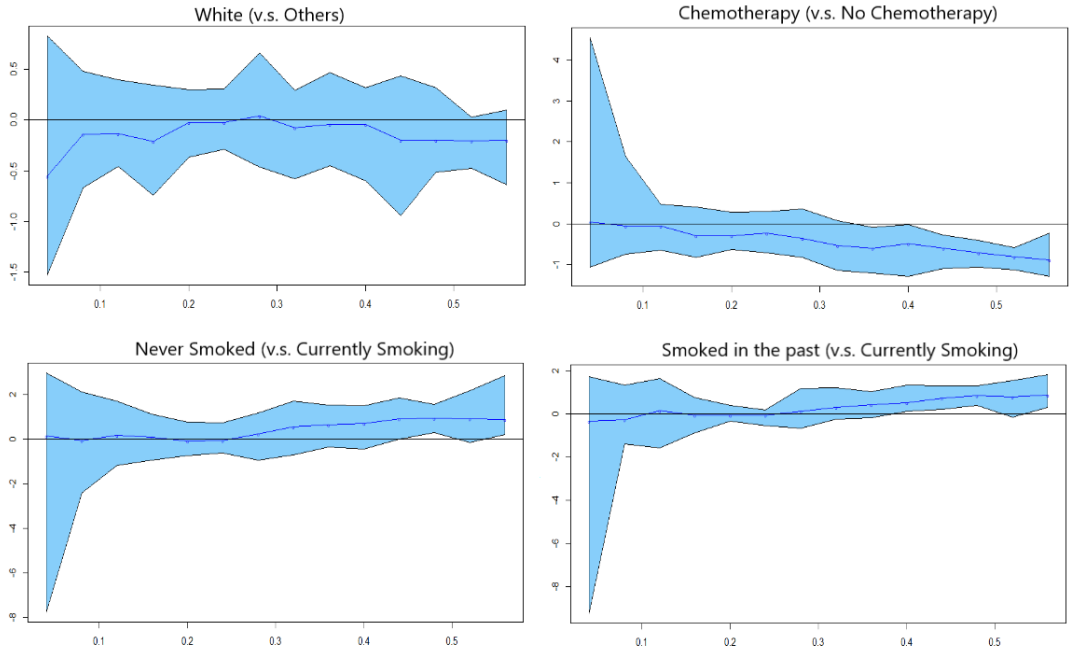

Figure 4.1: Pointwise confidence band for the coefficients of Race, Chemotherapy and Smoke.

Figure 4.1 shows fitted coefficients of some of the variables. The effect of smoking history and adjuvant chemotherapy vary among quantile levels. The effect of race seems insignificant at all quantile levels. We do not have to remove the insignificant covariates because our goal in the first stage is to estimate r_i. Similar to prediction, it is not necessary to find a parsimonious model to get an accurate estimate of r_i as long as we have enough sample size.

In the second stage, since z_i is in high dimension, one can either work with one z_{ij} at a time or use the shrinkage methods. Since gene expressions can be highly correlated, the performance of the shrinkage methods is usually unsatisfactory.

Therefore, for each j, we fit the Cox proportional hazard model with the jth gene as the only covariate. We control the false discovery rate with Benjamini–Hochberg procedure (*Benjamini and Hochberg* (1995)). Let $p(1), ..., p(q)$ be the ordered p-value of the q test. The Benjamini–Hochberg (BH) procedure finds the largest k such that $p_{(k)} \leq \frac{k}{q}\alpha$ and reject $H_{(j)} : j = 1, ..., k$.

Setting $\alpha = 0.05$, we are able to identify 175 significant genes. Here we list 10 genes with the smallest p-value: SCGB1D2, ARNTL2, ZNF185, ZC2HC1A, PLEK2, KLK6, RPL39L, GOLT1B, VEGFC, CHEK1. Among them, there already exists literature confirming that lung cancer progression can be influenced by ARNTL2 (*Brady et al.* (2016)), ZNF185 (*Wang et al.* (2016)), PLEK2 (*Wu et al.* (2020)), KLK6 (*Nathalie et al.* (2009)), VEGFC (*Jiang et al.* (2013)) and CHEK1 (*Sen et al.* (2017)). For other genes, we are unable to find results about their association with lung cancer.

We also analyze the data using the classical model. We fit the Cox proportional harzard model with covariates

$$\sim \text{Gender} + \text{Race} + \text{Chemotherapy} + \text{Smoke} + \text{Grade} + bs(\text{Age}) + \text{Gene}_j, \quad (4.11)$$

and control the false discovery rate with the BH procedure. We are able to identify 228 significant genes, more than the 175 genes identified using the two-stage model. But we notice that there exist 27 genes that are identified by the two-stage model but not by the classical model. In Figure 4.2, we randomly select 100 genes whose p-value either calculated with the two-stage model or the classical model is less than 0.001 and plot their p-values calculated with two methods. We observe that for some of

the genes, the p-values calculated with the two methods are similar, but there exist genes that the p-values differ greatly. The result is reasonable because if a gene only has a mean effect on the survival time, the classical method is usually more powerful than our two-stage model. But the two-stage model has complementary power detecting genes that have interactions effect with the environment. From the clinical perspective, it is certainly as important, if not more important, to identify those genes.

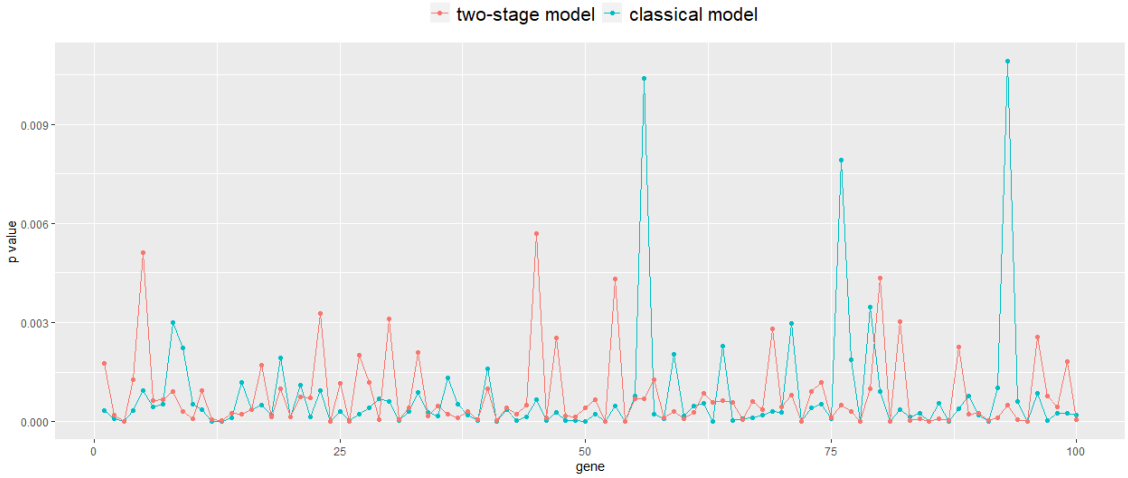

Figure 4.2: Compare the p-value of the two-stage model and classical model.

In conclusion, we identity 175 genes that are possibly associated with the survival of lung cancer patients. Some of the genes have already been studied and reported in the literature, but some of our findings are new. It will be worthwhile for scientists to conduct further research to study the mechanisms of how these new identified genes impact the progression of lung cancer.

4.4 Future work

In this chapter, we introduce the two-stage model and apply this model to find genes that are associated with the progression of lung cancer. Our work in this chapter is to provide a new approach that has the potential to accommodate general forms of gene-environment interactions. Additional work is needed to fully investigate the potential of the method. Our future work will focus on the following two aspects.

First, we want to study the asymptotic properties of the proposed method. Specifically, as discussed in Section 4.2.3.1, we want to show that the paired bootstrap is consistent when Model (4.5) is allowed to be misspecified and the responses $\text{logit}(\hat{r}_i)$ are weakly correlated.

Second, we want to compare the two-stage model to the varying index coefficient model (VICM) proposed in *Ma and Song* (2015). *Ma and Song* (2015) considered the model

$$y_i = \sum_{l=1}^{p} m_l(z_i^T \beta_l) x_{il} + \epsilon_i, \tag{4.12}$$

where $\beta_l = (\beta_{l1}, ..., \beta_{lq})$ and $m_l(\cdot)$ is some unknown smooth function. The two-stage model and the VICM look somewhat similar and both models allow the interactions between z_i and x_i to be non-linear. But the interpretation of these two models and how the coefficients are estimated are quite different. The two-stage model we proposed features a latent variable r_i, affording model heterogeneity in a transparent way. It would be interesting to study the possible connection between these two methods and compare their performance under different settings.

BIBLIOGRAPHY

BIBLIOGRAPHY

Backer, M. D., A. E. Ghouch, and I. V. Keilegom (2019), An adapted loss function for censored quantile regression, *Journal of the American Statistical Association*, *114*(527), 1126–1137.

Bang, H., and A. A. Tsiatis (2002), Median regression with censored cost data, *Biometrics*, *58*(3), 643–649.

Benjamini, Y., and Y. Hochberg (1995), Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society*, *57*(1), 289–300.

Bossé, Y., and C. I. Amos (2018), A decade of GWAS results in lung cancer, *Cancer Epidemiol Biomarkers Prev*, *27*(4), 363–379.

Brady, J. J., et al. (2016), An Arntl2-driven secretome enables lung adenocarcinoma metastatic self-sufficiency, *Cancer Cell*, *29*(5), 697–710.

Cameron, R. H., and W. T. Martin (1945), Evaluation of various wiener integrals by use of certain sturm-liouville differential equations, *Bulletin of the American Mathematical Society*, *51*(2), 73–90.

Chatterjee, S., and A. Bose (2005), Generalized bootstrap for estimating equations, *The Annuals of Statistics*, *33*(1), 414–436.

Chernozhukov, V., and I. Fernández-Val (2004), Subsampling inference on quantile regression processes, *The Indian Journal of Statistics*, *67*(2), 253–276.

Douhard, M., M. Festa-Bianchet, J. Landes, and F. Pelletier (2019), Trophy hunting mediates sex-specific associations between early-life environmental conditions and adult mortality in bighorn sheep, *Journal of Animal Ecology*, *88*(5), 734–745.

Engle, R., and S. Manganelli (2004), CAViaR: conditional autoregressive value at risk by regression quantiles, *Journal of Business and Economic Statistics*, *22*(4), 367–381.

Feng, X., X. He, and J. Hu (2011), Wild bootstrap for quantile regression, *Biometrika*, *98*(4), 995–999.

Feng, Y., Y. Chen, and X. He (2015), Bayesian quantile regression with approximate likelihood, *Bernoulli*, *21*(2), 832–850.

Gill, R. D., and S. Johansen (1990), A survey of product-integration with a view toward application in survival analysis, *The Annals of Statistics*, *18*(4), 1501–1555.

Gonzalez-Manteiga, W., and C. Cadarso-Suarez (1994), Asymptotic properties of a generalized kaplan-meier estimator with some applications, *Journal of Nonparametric Statistics*, *4*(1), 65–78.

Gutenbrunner, C., and J. Jurečková (1992), Regression rank scores and regression quantiles, *The Annals of Statistics*, *20*(1), 305–330.

Gutenbrunner, C., J. Jurečková, R. Koenker, and S. Portnoy (1993), Tests of linear hypotheses based on regression rank scores, *Journal of Nonparametric Statistics*, *2*(4), 307–331.

He, X., and Q.-M. Shao (1996), A general bahadur representation of m-estimators and its application to linear regression with nonstochastic designs, *The Annals of Statistics*, *24*(6), 2608–2630.

He, X., Y.-H. Hsu, and M. Hu (2010), Detection of treatment effects by covariate-adjusted expected shortfall, *Bulletin of the American Mathematical Society*, *4*(4), 2114–2125.

Hosmer, D., and S. Lemeshow (2010), *Applied Logistic Regression, 2nd edition*, Wiley-Interscience Publication.

Hu, J., F. A. Wright, and F. Zou (2006), Estimation of expression indexes for oligonucleotide arrays using the singular value decomposition, *Journal of the American Statistical Association*, *101*(473), 41–50.

Hunter, D. J. (2005), Gene–environment interactions in human diseases, *Nature Reviews Genetics*, *6*, 287–298.

Jiang, H., W. Shao, and W. Zhao (2013), VEGF-C in non-small cell lung cancer: Meta-analysis, *Clinica Chimica Acta*, *427*, 94–99.

Kocherginsky, M., X. He, and Y. Mu (2005), Practical confidence intervals for regression quantiles, *Journal of Computational and Graphical Statistics*, *14*(1), 41–55.

Koenker, R. (2005), *Quantile regression*, Cambridge University Press.

Koenker, R. (2008), Censored quantile regression redux, *Journal of Statistical Software*, *27*(6).

Koenker, R. (2010), Rank tests for heterogeneous treatment effects with covariates, *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jureckova,*, *7*, 134–142.

Koenker, R., and G. Bassett (1978), Regression quantiles, *Econometrica*, *46*(1), 33–50.

Koenker, R., and J. A. F. Machado (1999), Goodness of fit and related inference processes for quantile regression, *Journal of the American Statistical Association*, *94*(448), 1296–1310.

Koenker, R., V. Chernozhukov, X. He, and L. Peng (2017), *Handbook of Quantile Regression*, Chapman and Hall/CRC.

Leng, C., and X. Tong (2013), A quantile regression estimator for censored data, *Bernoulli*, *19*(1), 344–361.

Li, K.-C., and N. Duan (1989), Regression analysis under link violation, *The Annals of Statistics*, *17*(3), 1009–1052.

Ma, S., and P. X.-K. Song (2015), Varying index coefficient models, *Journal of the American Statistical Association*, *110*(509), 341–356.

Nathalie, H.-V., et al. (2009), High kallikrein-related peptidase 6 in non-small cell lung cancer cells: An indicator of tumour proliferation and poor prognosis, *Journal of Cellular and Molecular Medicine*, *13*(9B), 4014–4022.

Neocleous, T., and S. Portnoy (2008), On monotonicity of regression quantile functions, *Statistics and Probability Letters*, *78*(10), 1226–1229.

Peng, L. (2012), Self-consistent estimation of censored quantile regression, *Journal of Multivariate Analysis*, *105*(1), 368–379.

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

Portnoy, S. (1991), Asymptotic behavior of the number of regression quantile breakpoints, *SIAM Journal on Scientific and Statistical Computing*, *12*(4), 867–883.

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

Portnoy, S., and G. Lin (2010), Asymptotics for censored regression quantiles, *Journal of Nonparametric Statistics*, *22*(1), 115–130.

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.

Rodriguesa, T., J.-L. Dortet-Bernadetc, and Y. Fanc (2019), Pyramid quantile regression, *Journal of Computational and Graphical Statistics*, *28*(3), 732–746.

Sen, T., et al. (2017), CHK1 inhibition in small-cell lung cancer produces single-agent activity in biomarker-defined disease subsets and combination activity with cisplatin or olaparib, *Cancer Research*, *77*(14), 3870–84.

Shedden, K., et al. (2008), Gene expression-based survival prediction in lung adenocarcinoma: A multi-site, blinded validation study, *Nature medicine*, *14*(8), 822–827.

Veena, V., K. Rajan, V. Saritha, S. Preethi, K. Chandramohan, K. Jayasree, S. T. S, and K. Sujathan (2017), DNA replication licensing proteins for early detection of lung cancer, *Asian Pacific journal of cancer prevention*, *18*(11), 3041–3047.

Wang, H. (2009), Inference on quantile regression for heteroscedastic mixed models, *Statistica Sinica*, *19*(3), 1247–1261.

Wang, H. J., and M. Fygenson (2009), Inference for censored quantile regression models in longitudinal studies, *The Annals of Statistics*, *37*(2), 756–781.

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

Wang, J., H. H. Huang, and F. B. Liu (2016), ZNF185 inhibits growth and invasion of lung adenocarcinoma cells through inhibition of the akt/gsk3β pathway, *Journal of Biological Regulators & Homeostatic Agents*, *30*(3), 683–691.

Wu, D.-M., S.-H. Deng, J. Zhou, R. Han, T. Liu, T. Zhang, J. Li, J.-P. Chen, and Y. Xu (2020), PLEK2 mediates metastasis and vascular invasion via the ubiquitin-dependent degradation of SHIP2 in non-small cell lung cancer, *International Journal of Cancer*, *146*(9), 2563–2575.

Yang, X., N. N. Narisetty, and X. He (2018), A new approach to censored quantile regression estimation, *Journal of Computational and Graphical Statistics*, *27*(2), 417–425.

Ying, Z., S. H. Jung, and L. J. Wei (1995), Survival analysis with median regression models, *Journal of the American Statistical Association*, *90*(429), 178–184.

Zhou, L. (2006), A simple censored median regression estimator, *Statistica Sinica*, *16*, 1043–1058.