On Testing the Change-point in the Longitudinal Bent Line Quantile Regression Model

Nanshi Sha

Submitted in partial fulfillment of the
Requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2011

Abstract

On Testing the Change-point in the Longitudinal Bent Line Quantile Regression Model

Nanshi Sha

The problem of detecting changes has been receiving considerable attention in various fields. In general, the change-point problem is to identify the location(s) in an ordered sequence that divides this sequence into groups, which follow different models. This dissertation considers the change-point problem in quantile regression for observational or clinical studies involving correlated data (e.g. longitudinal studies) . Our research is motivated by the lack of ideal inference procedures for such models.

Our contributions are two-fold. First, we extend the previously reported work on the bent line quantile regression model [Li et al. (2011)] to a longitudinal framework. Second, we propose a score-type test for hypothesis testing of the change-point problem using rank-based inference. The proposed test in this thesis has several advantages over the existing inferential approaches. Most importantly, it circumvents the difficulties of estimating nuisance parameters (e.g. density function of unspecified error) as required for the Wald test in previous works and thus is more reliable in finite sample performance. Furthermore, we demonstrate, through a series of simulations, that the proposed methods also outperform the extensively used bootstrap methods by providing more accurate and computationally efficient confidence intervals. To illustrate the usage of our methods, we apply them to two datasets from real studies:

the Finnish Longitudinal Growth Study and an AIDS clinical trial. In each case, the proposed approach sheds light on the response pattern by providing an estimated *location of abrupt change* along with its 95% confidence interval at any quantile of interest – a key parameter with clinical implications. The proposed methods allow for different change-points at different quantile levels of the outcome. In this way, they offer a more comprehensive picture of the covariate effects on the response variable than is provided by other change-point models targeted exclusively on the conditional mean. We conclude that our framework and proposed methodology are valuable for studying the change-point problem involving longitudinal data.

KEY WORDS: Change-point; Piecewise linear; Rank score test; Longitudinal data; Adiposity rebound; Plateau; HIV treatment.

Contents

List of Tables

List of Figures

Acknowledgments

I would like to express my sincere gratitude to many people who supported me in completing this thesis. First and foremost, I would like to thank Professor Ying Wei, my thesis advisor at Columbia University, Department of Biostatistics. I first got to know Ying in 2007 when I took her graduate course "Design of Medical Experiments", one of my favorite statistics classes, from which I had my first lecture on principles in the design and analysis of controlled experiments. Later on I started my dissertation project under Ying's guide. Her incredible breadth and depth knowledge in statistics, timely help and thoughtful suggestions have always been a source of confidence. Through numerous discussions with her, whether in her office or online, I have not only accumulated knowledge in biostatistics, but also learned how to conduct scientific research. This thesis would not have been possible without her generous assistance and supervision. I thank her for her sage guidance, insightful criticisms, and patient encouragement.

In addition to my interactions with Ying, I have been fortunate in my internships and collaborations during my graduate studies. I would like to thank Drs. Naihua Duan, Yuanjia Wang and Huiping Jiang, my supervisers during my first summer internship at the New York State Psychiatric Institute. Dr. Duan's rigorous scholarship has had an profound

effect on my later research while Drs. Wang and Jiang have broadened my horizons on interdisciplinary research from an applied perspective. At the Merck research laboratories, I had many thought-provoking discussions with Drs. Peggy Wong, Sabrina Wan and Gary Chen, who enabled me to obtain some hands-on experience in biometric studies and encouraged me to develop novel methods for testing gene pathway activation. I would also like to thank Drs. Hui Quan, Lynn Wu and Huiling Li at Sanofi-Aventis, who have provided me valuable experiences in clinical trials and help me to understand the underlying motivations of pharmaceutical biostatistics research. I am also thankful to Drs. Guohua Li, Lena Sun, Julia Sobol, Hannah Wunsch, Caleb Ing, Teeda Pinyavat, Radhika Dinavahi and Teresa Mulaikal at the Department of Anesthesiology. Due to Dr. Li's guidance and coordination, my statistical input in epidemiology and anesthesiology projects were widely accepted and well appreciated. My collaborations with them are very enjoyable and fruitful. I must also thank Dr. Mitchell Berman and Mr. Brian Thumm for providing extremely useful training on the management of databases of electronic medical records, making me more careful and precise.

I am very thankful to all the faculty, staff and fellow colleagues in the Biostatistics Department at Columbia. The comprehensive coursework and seminars have been key to my understanding of statistical research. I am grateful to Professors Ian McKeague, Bin Cheng, Mary Beth Terry, Wei-Yann Tsai, Zhezhen Jin, and Ken Cheung for serving on my oral examination committee and/or doctoral thesis committee. I am particularly indebted to Professor Bin Cheng, who was my mentor during my first year here. I would also highlight my biostatistics fellow, Jimmy Duong, who has made great efforts to edit my thesis.

Last but not least, I owe special thanks to my parents Yanbu Sha and Jinqiu Yu for their unconditional support for my pursuit of scientific truth. Since I left my hometown and started college at the University of Science and Technology of China in 2002, the time I have spent with them has declined. However, they always stand by my decisions regardless of distance. Finally, this endeavor would have been much more challenging without the unwavering support and constant love from Nan Lin, my darling girlfriend.

To my family

Chapter 1

Introduction

Change-point problems frequently arise in public health, medical, and behavioral, as well as in biological, agricultural and geographical sciences. For effective communication, it is crucial to report key statistical findings in a way that non-statisticians can appreciate [Brownson and Remington (2002)]. Change-point models are desirable for this purpose since they usually condense key functional information into a few parameters with clear interpretations. Depending on context, a change-point parameter may also be referred to as a break-point or cusp-point.

The general form of a change-point problem is estimating an unknown time point t, where an ordered sequence of observations $y_1, \ldots, y_n$ naturally fall into two groups, $\{y_i\}_{i=1,\ldots,t}$ and $\{y_i\}_{i=t+1,\ldots,n}$. Each group obeys a model with a distinct analytic form. The index usually represents time, but in general it can be any associated variable, e.g. dosage of new drug under investigation. A simple example in the univariate setting is the *level-change model*: the random variables $\{y_i\}_{i=1,\ldots,t}$ follow an independent identically distributed (iid) $N(\mu_1, \sigma^2)$ while $\{y_i\}_{i=t+1,\ldots,n}$ are iid $N(\mu_2, \sigma^2)$ with $\mu_1 \neq \mu_2$. Another common example in the regression setting is the *two-phase regression model* [Gallant and Fuller (1973)]: the response variables $\{y_i\}_{i=1,\ldots,t}$ obey a linear model a_1+b_1x, based on some explanatory variable x while $\{y_i\}_{i=t+1,\ldots,n}$ follow another model $a_2 + b_2x$, which has a similar form but with a different set of parameters.

A wide variety of theoretical challenges have arisen in modeling change-point patterns. In the mid 20th century, Page (1954) proposed an inspection scheme for

detecting an abrupt change of the parameter in one direction, known as *continuous inspection schemes*. This is probably when the change-point problem was first formulated. Later on, many authors [Box and Tiao (1965); Hinkley (1971); Fu and Curnow (1990); Liang et al. (1990); Banerjee and McKeague (2007); Li et al. (2011)] have considered extensions in various settings and directions including binary outcomes, multivariate outcomes, time series data, stochastic processes, hazard rate regressions, quality control problems, sequential applications and multiple change-points problems. The general techniques employed include maximum likelihood approach, nonlinear least squares [Srivastava and Worsley (1986)], nonparametric methods [Csörgo and Horvath (1988)], as well as Bayesian methods [Barry and Hartigan (1993)]. All these techniques have been developed to solve the change-point problems in various applications.

It is conceivable that a change-point problem may simply be viewed as a special case of nonlinear regression. From a theoretical perspective, however, the problem is nonregular. For instance, let us take the simplest case of the level-change model above. Here the log-likelihood function is not differentiable with respect to the change-point parameter t; consequently, the standard likelihood approach does not apply. In this situation, more assumptions are thus generally required. One such assumption is the existence of one single change-point. Based on this assumption, theoretical derivations of the asymptotic properties follow: consistency and asymptotic normality of the change-point estimate [Feder (1975); Krisnaiah and Miao (1988)], as well as the asymptotic distribution of the test statistics [Worsley (1979)], which are derived to test whether there exists a change-point and, where/when it is. Recent results include, but are not limited to, the estimation of the number and the locations of multiple change-points [Fearnhead (2006)].

From an application perspective, the method of nonlinear least squares is straightforward for generating the estimate and can be readily implemented with existing

statistical softwares. However, there is no a simple general result on the asymptotic theory of the change-point estimate [Krisnaiah and Miao (1988)]. Consequently, researchers frequently have to resort to the computationally intensive methods such as the resampling methods to construct confidence intervals [Efron and Tibshirani (1986); Hinkley and Schechtman (1987)].

The classical least squares regression models examine the covariate effect on the response by focusing on the conditional mean. However, they do not address another equally important question of whether the covariate effect, if there is any, is homogeneous across different quantiles (or percentiles) of the response variable. Questions in the latter form, more often than not, are important topics of modern epidemiology research such as studies of obesity studies and low birthweight investigations. Quantile regression emerges as one of the indispensable took-kits for addressing such research questions by providing a more comprehensive picture of the covariate effects on the response variable distribution [Koenker and Hallock (2001); Hao and Naiman (2007)]. Change-point problems naturally arise in these studies, as will be explained in greater detail in the following section. Quantile regression by allowing the change-point to vary across different quantiles, offers a flexible setup for studying these problems. As a result, the development of proper inferential tools motivates our research.

Extra challenges arise from studies with correlated data, such as longitudinal data and clustered data. The former frequently arise in large cohort studies where repeated measures are taken from one single subject and are thus correlated. Time series data from surveillance studies can also be considered as longitudinal data. Clustered data, on the other hand, are commonly seen in multi-center studies, where each participating center is usually treated as a cluster. The subjects within one cluster are thus correlated. In either case, the complex correlation structure needs to be appropriately taken into account for valid inference. However, statistical methods for detecting change-points for longitudinal data are still limited. Wu and Yang (2008)

proposed a method based on a transition function for longitudinal binary data, and Rosenfield et al. (2010) described a method based on statistical control theory for longitudinal physiological time series data.

1.1 Motivating examples

There are various types of observational studies and experimental studies which involve change-point phenomena [Pawitan (2005)]. In this section we describe two real examples which motivate our research. One comes from the Finnish Longitudinal Growth Study [Sorva et al. (1990)]. The other one is from an AIDS clinical study [Park and Wu (2006)].

1.1.1 The Finnish Longitudinal Growth Study

Our first motivating example is a pediatric study with longitudinal measures. In pediatrics and nutritional epidemiology, there is a time period in childhood, usually between 5 to 7 years of age, termed adiposity rebound (AR) [Diez (1994)]. AR is a critical period for the regulation of energy balance and adult obesity risk and thus generates extensive research interest [Rolland-Cachera et al. (1984); Siervogel et al. (1991); Prokopec and Bellisle (1993); Reilly et al. (2005)]. Furthermore, an early AR (younger age at the point of AR) is associated with not only higher BMI in adolescence [Rolland-Cachera et al. (1984); Prokopec and Bellisle (1993)] but also an increased risk of adult obesity, even after adjusting for the BMI at AR, maternal BMI, and paternal BMI – three other known risk factors for adult obesity. [Whitaker et al. (1998)].

The dataset is from the Finnish Longitudinal Growth Study [Sorva et al. (1990); Pere (2000)]. The observations consist of longitudinal measurements on height and

weight for 2514 Finnish children. Weight (kg) and height (m) were measured for each participant using standard techniques, and BMI (kg/m^2) was calculated by BMI = weight/height^2. As described in more detail in Pere (2000), this dataset has been cleaned to remove a small proportion of children with low or missing birth weight or other suspicious measurements. After cleaning, there are 1140 boys and 1162 girls with ages ranging from 1 to 18 years old.

In Figure 1.1, we plot the BMI against age in years for boys and girls, respectively. One can observe a clear pattern for both male and female subjects: The BMI decreases

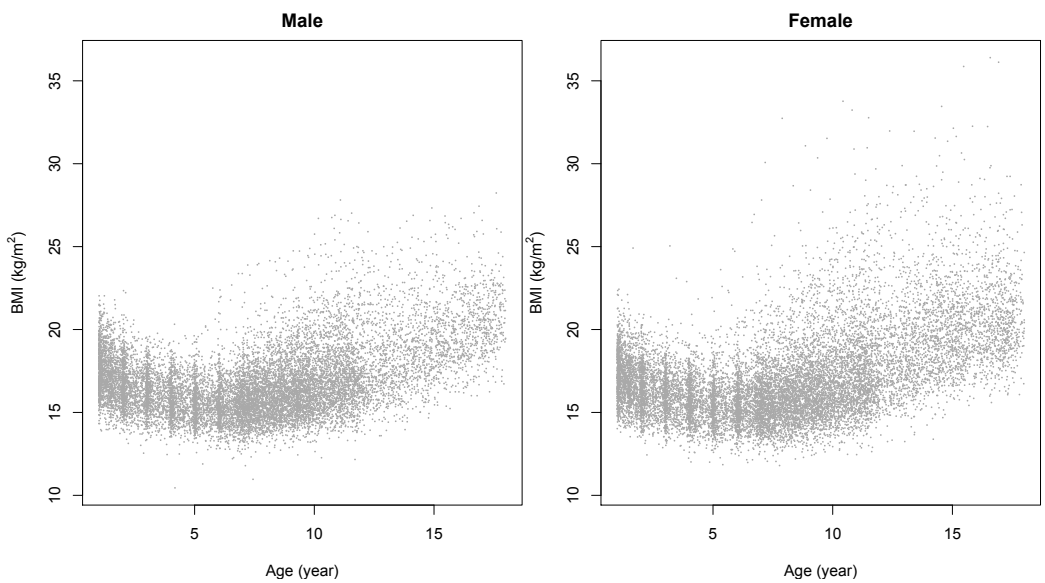

Figure 1.1: Body Mass Index (BMI) plotted against Age in years for 1140 boys and 1162 girls.

from 1 year of age until somewhere between 5 and 7 years of age, and then starts to increase. In fact, pediatric considerations support the idea of a change-point. As described in Whitaker et al. (1998), "body fatness normally declines to a minimum, a point called adiposity rebound (AR), before increasing again into adulthood." It is of clinical interest to estimate the mean age at AR for the general population. A more informative analysis is to characterize the ages at AR for heavy, medium and

lean subjects. This is particularly helpful for the clinicians to study the relationship between the onset of childhood obesity and more severe adult disease.

To accommodate data arising from such applications, it is natural to model the BMI using segmented regression models that allow for different slopes in different time domains, e.g. before and after AR. It is particularly of clinical interest to determine the AR, i.e. when the BMI begins to increase again after infancy. Moreover, compared to the question of estimating the time of AR on average, it is more meaningful to determine the time of AR for lean, medium and heavy subjects typified by, say the 10th, 50th and 90th percentiles of BMI. Such information is crucial for further investigation of the relationships between childhood obesity, early adiposity rebound and risks of subsequent adulthood diseases, and would also contribute to the development of protocols for future studies.

1.1.2 An AIDS clinical study

In many medical studies, investigators assess the efficacy of new treatments longitudinally and measure the response to the treatment over time. One commonly observed response pattern consists of two stages, a steep improvement at the beginning of the treatment, followed by a period where the condition has stabilized [Deeks et al. (2004); Hunt et al. (2003)].

Another motivating example is the AIDS (acquired immune deficiency syndrome) clinical study developed by the AIDS Clinical Trials Group (ACTG). There are 171 patients enrolled in one of the three treatment arts and received the antiretrovial therapies (ART). They were followed every 4 weeks in the first 2 months, and every 8 weeks thereafter. Some patients might not exactly follow the designed schedule of clinical visits. Our main purpose is to characterize the response patterns of the CD4 cell measurements, an important marker for assessing immunologic response, in the

ART treatment. Figure 1.2 shows the CD4 cell count response from 171 patients during the 120 weeks of ART treatment. Each solid line in Figure 1.2 connects the CD4 cell counts from each patient's scheduled visit. In general, a steep increase in the CD4 cell counts is observed during the first few months regardless of the baseline severely. Then the CD4 cell counts reach a plateau indicating stabilized conditions. Such treatment response patterns of antiretroviral therapies (ART) have drawn much attention in the medical community –[Staszewski et al. (1999); Renaud et al. (1999); Tarwater et al. (2001)], just to name a few articles. In fact, pharmacological considerations support the idea of a change-point: the function of ART is the suppression of plasma HIV RNA, allowing a significant increase in the CD4 cell counts until the ART, virologic and other factors reach a balance, or a *plateau* as termed by Tarwater et al. (2001).

To accommodate data arising from such applications, it is natural to model the treatment efficacy using segmented regressions that allow for different slopes in different time regions. And it is particularly of clinical interest to determine the change-point indicating when the patients' conditions could be stabilized. Moreover, compared to the question of when the patients stabilize in average, perhaps a more meaningful question is what CD4 cell count level could be achieved by certain percentages of the patients and when does the stabilization occur. An example of a percentage of interest is the 0.10 quantile (10% percentile) of CD4 cell counts. Here, the severely infected patients are represented by those whose observations fall below this quantile function. Such information helps determine the appropriate length of the therapy depending on severity of infection, which in turn contributes to formulation of clinical protocols.

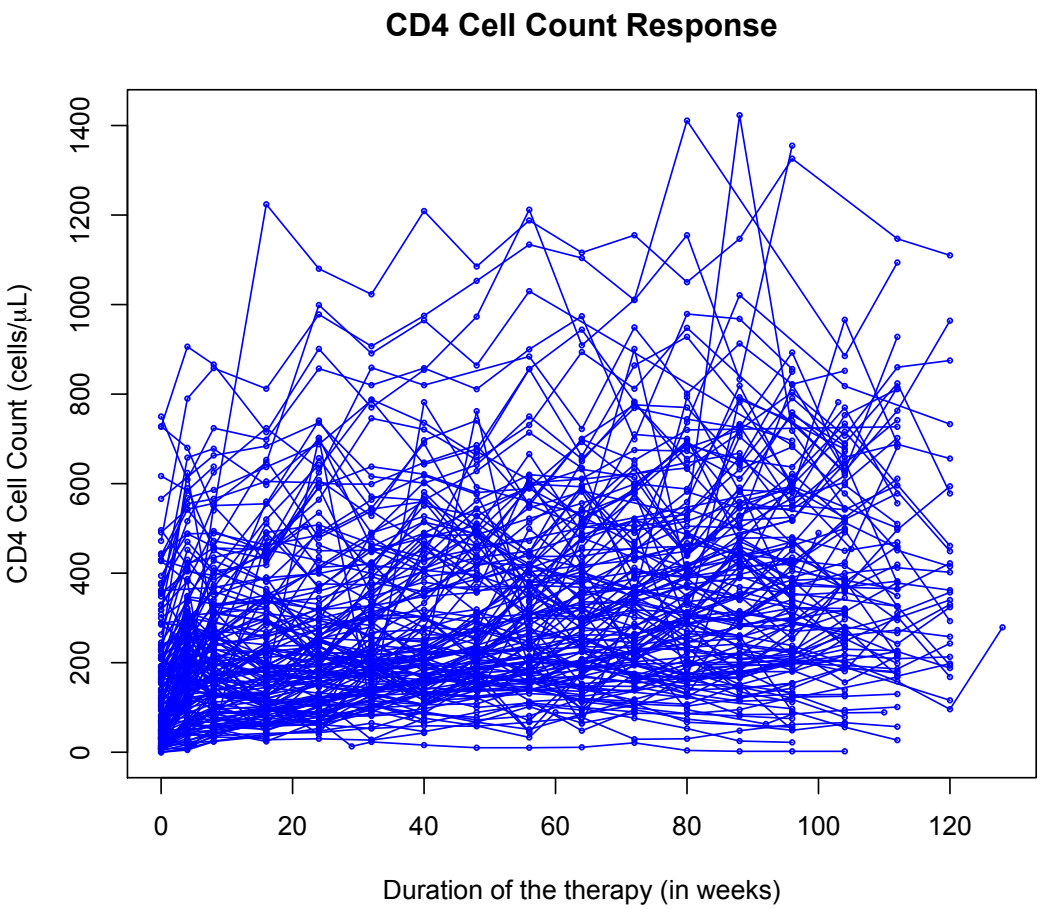

Figure 1.2: Observed CD4 Cell Counts from 171 HIV positive patients under a 120-week-long highly active antiretroviral therapy.

1.2 Our contribution

To address the aforementioned questions, we propose to conduct quantile analysis on the change-points. Although an extensive literature can be found on segmented least squares regressions models [Quandt (1958, 1960); Robison (1964); Feder (1975)], analogous work done for segmented *quantile regression* models is limited. For cross-sectional data, Li et al. (2011) developed a bent line quantile model.

Our contributions are two-fold. First, we generalize previous work [Li et al. (2011)] to a longitudinal framework. The asymptotic properties of the longitudinal estimate are established. However estimating the limiting variance-covariance matrix is known to be somewhat arduous, mostly due to the involvement of the unspecified unknown density function. To circumvent the difficulties of estimating the limiting variance-covariance matrix of the estimators, we propose a score-type test on the change-point derived from rank-based inference. The asymptotic distribution of the test statistic is also derived. The proposed test could also be utilized to construct confidence intervals. Through a series of simulations designed to compare the performance of the proposed test and the extensively used bootstrap method, we come to the conclusion that our test has a couple of attractive characteristics including better coverage accuracy and computational efficiency.

1.3 Structure of this dissertation

The rest of the thesis is organized as follows. In Chapter 2, we first review the change-point models developed in various areas. Then we summarize the fundamentals of linear quantile regression. In Chapter 3, we extend the original bent line quantile model to longitudinal settings. An estimation algorithm for the model parameters is described. The asymptotic properties of the parameter estimates are

studied. In Chapter 4, a rank score test statistic on the change-point is proposed and its asymptotic distribution is established. In Chapter 5, a series of simulation studies are conducted to evaluate the performance of the proposed test in finite sample sizes. Furthermore, an important application of the rank score test, construction of confidence intervals, is described, and its performance is compared against the popular bootstrap method. In Chapter 6, we applied these methods to the Finnish Longitudinal Growth Study and the AIDS clinical study mentioned earlier. In Chapter 7, we summarize the important findings in previous chapters and discuss some future directions. Chapter 8 provides complete proofs for all the theorems.

Chapter 2
Review of linear quantile regression

2.1 Overview

Before turning to the longitudinal aspects of our model, this chapter provides a brief overview on the existing methods on change-point models in various areas, followed by a complete review on fundamentals of linear quantile regression. In the first section, we review two versions of the bent-line regression model developed based on mean regression and quantile regression. In the following sections, the estimation of the linear quantile regression is outlined, followed by some discussions of the computational aspects and asymptotic properties. Fundamental inference methods for quantile regression are also compared and summarized.

2.2 Bent line regressions

Most frequently in regression problems, one functional form of the outcome is assumed throughout the entire domain of interest. However, in many applications it is more appropriate to separate the domain where different parametric forms are assumed respectively, resulting in segmented regression models. Feder (1975) studies a very broad class of segmented additive models, and derives the asymptotic results via Taylor expansions. He found the asymptotic consistency property of the regression

estimate depends on the "unsmoothness" at the change-point, i.e. the lowest order of the derivative that differs on two sides of the change-point. The main conclusions include: (1) the estimates have a rate of convergence of $n^{-1/2}(\log\log n)^{1/2}$ under suitable identifiability conditions; (2) the rate of convergence $n^{-1/2}$ is achieved if the first order derivatives disagree on the two sides of the change-point. Among the segmented regression models, one important special case is that of bent line regression, or broken line regression [Chappell (1989)]. The model takes a continuous piecewise form with single change-point

$$y_i = \begin{cases} a_1 + b_1 t_i + e_i, \text{if } t_i \leq t; \\ a_2 + b_2 t_i + e_i, \text{if } t_i > t, \end{cases} \tag{2.2.1}$$

where $i = 1, \ldots, n$ indexes the outcome y_i, whose slope in covariate t_i changes abruptly at t, $\{e_i\}$ are normally and independently distributed error terms with mean zero and variance σ^2. Due to the requirement of continuity at t, a_2 must satisfy

$$a_2 = a_1 + (b_1 - b_2)t, \tag{2.2.2}$$

based on which, equation (2.2.1) may be re-expressed

$$y_i = \begin{cases} a_1 + b_1 t_i + e_i, & \text{if } t_i \leq t; \\ a_1 + b_1 t + b_2(t_i - t) + e_i, & \text{if } t_i > t. \end{cases} \tag{2.2.3}$$

Upon letting $u_i = \max(t_i, t)$ and $v_i = \min(0, t_i - t)$, the compact formulation

$$y_i = a_1 + b_1 u_i + b_2 v_i + e_i, i = 1, \ldots, n \tag{2.2.4}$$

is equivalent to equation (2.2.3). Consequently, the conditional log-likelihood given fixed t is

$$\begin{aligned} l(a_1, b_1, b_2, \sigma|t) = &- \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) \\ &- \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a_1 - b_1 u_i b_2 v_i)^2, i = 1, \ldots, n \end{aligned} \tag{2.2.5}$$

By substituting the least squares estimates into equation (2.2.5), we have

$$l(\widehat{a}_1, \widehat{b}_1, \widehat{b}_2, \widehat{\sigma}|t) = -\frac{n}{2}\left(1 + \log(2\pi\widehat{\sigma}^2)\right). \tag{2.2.6}$$

Therefore one can obtain the unconditional maximum likelihood estimate of a_1, b_1, b_2 and σ by further minimizing $\widehat{\sigma}$ over t. The value of t where minimum is realized is $\widehat{t}$, the maximum likelihood estimate of t. $\widehat{a}_2$ can be calculated by substituting these estimates into the continuity constraint (2.2.2).

In general, the change-point estimates of segmented regression are not asymptotically normally distributed. To be more precise, the rate of convergence is $n^{-1/2}(\log\log n)^{1/2}$. However the bent line regression is one exception. By Feder (1975)'s result, Chappell (1989) shows the rate of convergence is $n^{-1/2}$ rather than $n^{-1/2}(\log\log n)^{1/2}$ based on the fact that the lowest order of derivative that disagree at t is 1. As n increases, $\sqrt{n}(\widehat{t}-t)$ converges to a normal distribution with mean zero and variance $D(t)GD^\top(t)$. Similarly $\sqrt{n}(\widehat{a}_1 - a_1, \widehat{b}_1 - b_1, \widehat{a}_2 - a_2, \widehat{b}_2 - b_2)$ converges to a multivariate normal distribution with mean $(0, 0, 0, 0)$ and variance G, with $D(t) = (1, t, 1, t)/(b_1 - b_2)$ and

$$G = \sigma^2 \begin{pmatrix} \frac{T_1}{v_1} & \frac{-S_1}{v_1} & 0 & 0 \\ \frac{-S_1}{v_1} & \frac{n_1}{v_1} & 0 & 0 \\ 0 & 0 & \frac{T_2}{v_2} & \frac{-S_2}{v_2} \\ 0 & 0 & \frac{-S_2}{v_2} & \frac{n_2}{v_2} \end{pmatrix} \tag{2.2.7}$$

where

$$
\begin{aligned}
S_1 &= \sum_{t_i \le t} t_i, S_2 = \sum_{t_i > t} t_i, \\
T_1 &= \sum_{t_i \le t} t_i^2, T_2 = \sum_{t_i > t} t_i^2, \\
v_1 &= \sum_{t_i \le t} t_i^2 - \frac{1}{n_1} \left(\sum_{t_i \le t} t_i \right)^2, \\
v_2 &= \sum_{t_i > t} t_i^2 - \frac{1}{n_2} \left(\sum_{t_i > t} t_i \right)^2.
\end{aligned}
\tag{2.2.8}
$$

The problem of formally examining the equality of b_1 and b_2 in model (2.2.1) has received much attention. The hypothesis of equality may be of interest depending on context. More importantly, the asymptotic results are established based on the identifiability assumption $b_1 \neq b_2$. If the change-point is known, a classic procedure for testing the equality of slopes, with

$$H_0 : b_1 = b_2 \text{ vs. } H_1 : b_1 \neq b_2, \tag{2.2.9}$$

is an F test developed from likelihood ratio statistic, also known as the "Chow test" (1960).

$$F_c = (n-3)(SS_s - SS_t)/SS_t \tag{2.2.10}$$

follows a F distribution with 1 and n-3 degrees of freedom, where SS_s and SS_t refer to the residual sum of squares from the single regression and the regression segmented at t, respectively. In most cases, however, the change-point is unknown. The Chow test is not appropriate. Testing for a quadratic term is suggested where the data are approximately evenly spaced. Null hypothesis which presumes the existence of t may be tested via likelihood ratio statistic. For example, the problem

$$H_0 : t = t_0 \text{ vs. } H_1 : t \neq t_0, \tag{2.2.11}$$

may be tested by the statistic $(n-4)(SS_{\widehat{t}}-SS_{t_0})/SS_{\widehat{t}}$, which follows an F distribution with 1 and n-4 degrees of freedom.

In some applications, the upper or lower quantile or all quantiles are of interest. This time quantile regression provides natural tools for modeling. The idea of bent line model has been extended from least squares regression to quantile regression. The analogue of (least squares) bent line regression in quantile regression is the *bent line quantile regression* [Li et al. (2011)]. Bent line quantile regression takes similar form as model (2.2.3),

$$y_i = \begin{cases} a_{1,\tau} + b_{1,\tau}t_i + x_i^\top \gamma_\tau + e_i(\tau), \text{if } t_i \leq t_\tau; \\ a_{1,\tau} + b_{1,\tau}t_\tau + b_{2,\tau}(t_i - t_\tau) + x_i^\top \gamma_\tau + e_i(\tau), \text{if } t_i > t_\tau. \end{cases} \tag{2.2.12}$$

where $\tau \in (0,1)$ is the quantile level of interest, all the parameters are $(a_{1,\tau}, b_{1,\tau}, b_{2,\tau}, t_\tau)$ are τ-specific, x_i is some covariate vector with constant slope vector γ_τ. As commonly assumed on the error terms in quantile regression, the conditional quantile $Q_{e_i}(\tau|t_i)$ is zero. As n tends to ∞, $\widehat{\boldsymbol{\theta}}_{n,\tau}$ converges to multivariate normal with mean zero vector and covariance matrix Σ_τ. One major difference from the least squares estimate is that the limiting covariance matrix involves unknown density function of the error term evaluated at the τth quantile, which is generally difficult to estimate well by conventional methods. In this thesis, we further extend the bent line quantile regression model into a longitudinal setting. To overcome the difficulty of estimating unknown density function, we propose an alternative test on the change-point parameter. Before we introduce the longitudinal bent line quantile regression, we need some preparation on linear quantile regression model, which are summarized in the following sections.

2.3 Estimation in linear quantile regression

In this section, we review the linear quantile regression model

$$y_i = x_i^\top \beta + e_i(\tau), \tag{2.3.13}$$

where $i = 1, \ldots, n$, $x_i = (1, x_{i,1}, \ldots, x_{i,p-1})^\top$ consists of an intercept and (p-1) covariates associated with the ith observation, $\{e_i(\tau)\}$ are *iid* random errors whose τth quantile is assumed to be zero for identifiability. We denote the cumulative distribution function (CDF) as $F(.)$ and probability density (PDF) as $f(.)$. We assume $f(.)$ is strictly positive at $F^{-1}(\tau)$, the τth quantile of e_i.

The quantile regression estimate $\widehat{\beta}$ can be obtained by solving the minimization problem taking the form,

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau \left(y_i - x_i^\top \beta \right). \tag{2.3.14}$$

Note that since we are discussing quantile regression, the estimator $\widehat{\beta}$ is τ-specific. Thus $\widehat{\beta}$ is notational shorthand for $\widehat{\beta}_\tau$. Here $\rho_\tau(\cdot)$ is a simple piecewise linear function illustrated in Figure 2.1,

$$\rho_\tau(u) = \begin{cases} \tau \cdot |u|, & \text{if } u \geq 0, \\ (1-\tau) \cdot |u|, & \text{if } u < 0; \end{cases} \tag{2.3.15}$$

which can also be written compactly using indicator function $I\{\cdot\}$ as

$$\rho_\tau(u) = (\tau - I\{u < 0\}) \cdot u.$$

When we apply $\rho_\tau(\cdot)$ to the error $y_i - x_i^\top \beta$, the error magnitude $|y_i - x_i^\top \beta|$ is weighted by τ or $1 - \tau$ based on error sign. Due to this effect, the piecewise linear

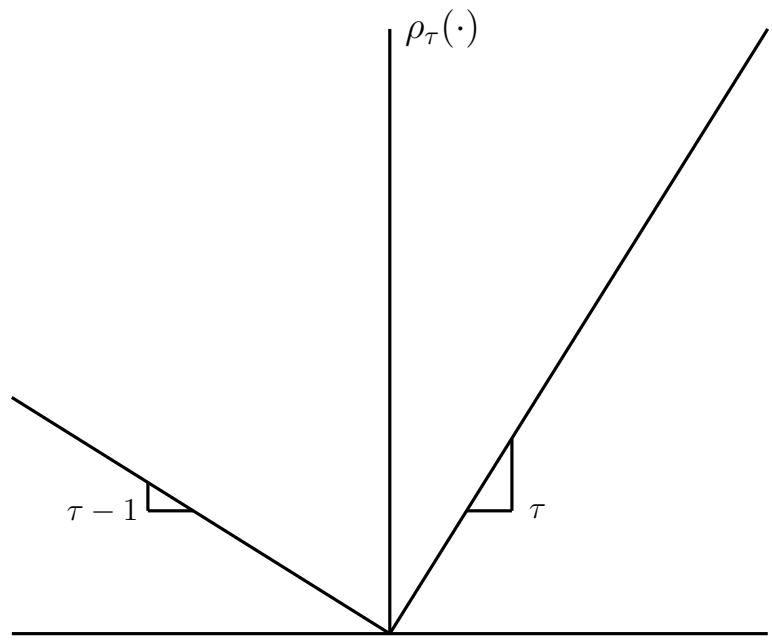

Figure 2.1: Quantile regression ρ_τ function.

form of the objective function achieves a balance between the proportions of observations falling above and below the fitted line, or plane in higher dimensions, $x^\top \widehat{\beta}$. In the simple case of $x_i = 1$, minimizing this objective function requires that: (1) the proportion of observations below the fitted line $\widehat{\beta}$ is at most τ; (2) the proportion of observations above the fitted line $\widehat{\beta}$ is at most $1 - \tau$. This is equivalent to the condition that $\widehat{\beta}$ is a τth sample quantile of $\{y_i\}_{i=1,\ldots,n}$.

2.4 Computational aspects

Koenker and Bassett (1978) proposed to extend this optimization interpretation of ordinary sample quantile to the estimation of linear parametric models for conditional quantile functions. Recall the fact that minimizing the sum of squared errors $\sum_{i=1}^n (y_i - \xi)^2$ over $\xi \in \mathbb{R}$ yields the sample mean $\widehat{\xi} = \bar{y}$ and minimization of $\sum_{i=1}^n (y_i - x_i^\top \beta)^2$ leads to the conditional mean function, $E[y|x] = x^\top \beta$. Similarly, minimizing the weighted sum of absolute errors $\sum_{i=1}^n \rho_\tau(y_i - \xi)$ gives the unconditional τth sample quantile, and minimizing

$$\sum_{i=1}^n \rho_\tau(y_i - x_i^\top \beta) \tag{2.4.16}$$

with respect to the $p-$dimensional parameter β leads to an estimate of the τth *conditional quantile function* of y given the covariate vector, x.

Further examination upon the optimization problem (2.4.16)

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \rho_\tau(y_i - x_i^\top \beta) \tag{2.4.17}$$

reveals that the problem may be reformulatd as minimizing a linear function

$$\left(\tau 1_n^\top, (1-\tau)1_n^\top, 0\right)\begin{pmatrix} u \\ v \\ \beta \end{pmatrix}$$

subject to some linear constrains

$$\left(I_n, -I_n, X\right)\begin{pmatrix} u \\ v \\ \beta \end{pmatrix} = y \tag{2.4.18}$$

by using $2n$ artificial variables $\{u_i, v_i : 1, \ldots, n\}$ to represent the positive and negative parts of the error vector $y - X\beta$, where X now denotes the usual $n \times p$ regression design matrix; $y = (y_1, \ldots, y_n)^\top$ denotes the $n-$response vector; 1_n denotes an $n-$vector of 1 and I_n an $n \times n$ identity matrix.

Such problems as (2.4.18) which aim to optimize a linear function subject to linear constraints are *linear programs* (LP). Hence problem (2.4.18) may be reformulated as

$$\min_{(\beta,u,v)\in\mathbb{R}^p\times\mathbb{R}_+^{2n}} \{\tau 1_n^\top u + (1-\tau)1_n^\top v \mid X\beta + u - v = y\} \tag{2.4.19}$$

where $\mathbb{R}_+^{2n} = \{a = (a_1, \ldots, a_{2n}) \in \mathbb{R}^{2n} : a_i \geq 0\}$ and 1_n denotes an $n-$vector of 1's. This idea was elaborated in Koenker and Bassett (1978).

Koenker and D'Orey (1987, 1993) proposed one algorithm which solves the general quantile regression problem in an efficient way. Their algorithm is an extension of the medium regression algorithm of Barrodale and Roberts (1974), which typifies the class of so called exterior point algorithms for solving linear programming problems. Koenker (2005) described the modified algorithm of Barrodale and Roberts in the following vivid terms: "we travel from vertex along the edges of the polyhedral

constraint set, choosing at each vertex the path of steepest descent, until we arrive at the optimum."

For practical problems of moderate size, the exterior point methods are competitive with least squares in terms of computational expenses. However, for quantile regression problems with p fixed and $n \to \infty$, the modified algorithm of Barrodale and Roberts exhibits rapid $O(n^2)$ growth in CPU time. In this sense it is not effective for large scale problems. The work of Karmarker (1984) initiated a dramatic reappraisal of computational methods for linear programming. Instead of traversing the outer surface, one takes Newton steps from the interior of a deformed version of the constrained set toward the boundary. This approach produced extremely effective interior point algorithms which dramatically improve the computational efficiency. Therefore these methods are particularly effective for large scale quantile regressions. For such problems, Portnoy and Koenker (1997) have shown that a combination of interior point methods with some effective problem preprocessing render large scale quantile regression computations competitive even with least squares computations.

2.5 Asymptotic properties

If quantile regression is expected to "work", the minimal requirement is asymptotic consistency, $\|\widehat{\beta}_n - \beta\| \to 0$ in probability as $n \to \infty$. For the linear quantile regression model with independent error

$$y_i = x_i^\top \beta + e_i, \tag{2.5.20}$$

where $i = 1, \ldots, n$, Bantli and Hallin (1999) demonstrated that the following conditions are sufficient,

Condition R1.

$\sqrt{n}(a_n(\epsilon) - \tau) \to \infty$ and $\sqrt{n}(\tau - b_n(\epsilon)) \to \infty$ with

$$\begin{aligned} a_n(\epsilon) &= n^{-1} \sum_i F_i(x_i^\top \beta(\tau) - \epsilon), \\ b_n(\epsilon) &= n^{-1} \sum_i F_i(x_i^\top \beta(\tau) + \epsilon) \end{aligned} \tag{2.5.21}$$

for $\epsilon > 0$, where F_i denotes the CDF of e_i.

Condition R2. There exists $d_1 > 0$ such that

$$\liminf_{n\to\infty} \inf_{\|u\|=1} n^{-1} \sum_i I\{|x_i^\top u| < d_1\} = 0. \tag{2.5.22}$$

Condition R3. There exists $d_2 > 0$ such that

$$\limsup_{n\to\infty} \sup_{\|u\|=1} n^{-1} \sum_i (x_i^\top u)^2 \le d_2. \tag{2.5.23}$$

Condition R2 ensures that the $\{x_i\}$'s are not concentrated in any linear subspace of $\mathbb{R}^p$ and is necessary for identifiability. Condition R3 controls the rate of growth of the $\{x_i\}$'s and holds under the classical condition that $n^{-1}\sum_i x_i x_i^\top$ converges to a positive definite matrix. Now we move forward to the following regularity conditions which are required for the asymptotic properties of the estimator.

Condition R4. The distribution functions F_i of e_i are absolutely continuous, with continuous densities f_i uniformly bounded away from 0 and ∞ at $F^{-1}(\tau)$.

Condition R5. There exist positive definite matrices D_0 and $D_1(\tau)$ such that

$$\begin{aligned} &(i) \lim_{n\to\infty} n^{-1} x_i x_i^\top = D_0, \\ &(ii) \lim_{n\to\infty} n^{-1} f_i(F^{-1}(\tau)) x_i x_i^\top = D_1(\tau), \\ &(iii) \max_{i=1,\ldots,n} \|x_i\|/\sqrt{n} \to 0. \end{aligned} \tag{2.5.24}$$

We emphasize that the behavior of the conditional response density in a neighbor-

hood of the conditional quantile model is crucial to the asymptotic behavior of $\widehat{\beta}_n$. One may find Conditions R5(i) and R5(iii) familiar throughout the literature on M-estimators for regression models; Condition R5(ii) facilitates notational convenience.

Theorem 2.5.1 Koenker and Bassett (1978) *Under Conditions R1 and R2,*

$$\sqrt{n}(\widehat{\beta}_n - \beta) \xrightarrow{D} N(0, \tau(1-\tau)D_1^{-1}D_0D_1^{-1}). \tag{2.5.25}$$

There is an extensive literature developing some form of linear representation of the quantile regression estimator. One could mention Jurečková and Sen (1984), Portnoy (1984), Koenker and Portnoy (1987), Gutenbrunner and Jurečková (1992), He and Shi (1996), all of whom provided some variant of the linear representation for $\widehat{\beta}_n$:

$$\sqrt{n}(\widehat{\beta}_n - \beta) = n^{-1}D_1^{-1}\sum_{i=1}^{n} x_i\psi_\tau(\widehat{e}_i) + o(1). \tag{2.5.26}$$

where $\widehat{e}_i$ denotes the residual from the model. The beauty of such representations lies in the fact that they express a rather complicated nonlinear estimator as a normalized sum of iid random variables, based on which the limiting distribution readily follows.

2.6 Wald test in quantile regression

In this section we consider the hypothesis testing problem

$$H_0 : \beta = 0 \text{ vs. } H_1 : \beta \neq 0. \tag{2.6.27}$$

One can take the Wald approach and examine the asymptotic normality of the $\widehat{\beta}_n(\tau)$. However, as shown in (2.5.25) and (2.5.24), the asymptotic variance-covariance matrix of the estimator involves the density function of the unspecified error. The nuisance

quantity $s(t) = \frac{1}{f(F^{-1}(t))}$ has to be estimated. For the case of iid errors, Siddiqui (1960) estimate $s(t)$ by using a simple difference quotient of the empirical quantile function,

$$\widehat{s}_n(t) = [\widehat{F}_n^{-1}(t + h_n) - \widehat{F}_n^{-1}(t)]/2h_n, \tag{2.6.28}$$

where $\widehat{F}_n^{-1}$ is an estimate of F^{-1} and h_n is a bandwidth that tends to zero as $n \to \infty$. Different choices of bandwidth are discussed [Siddiqui (1960); Bofinger (1975); Hall and Sheather (1988)] and have effect on the performance of the Wald test. Non-iid error settings are even more challenging [Hendricks and Koenker (1992); Powell (1991)]. As we see, the presence of the sparsity function hampers the application of the Wald-type test. Moreover, it has been shown that, in a quantile regression setup, a Wald-type test is generally unstable at small sample sizes or at extreme quantiles [Chen and Wei (2005); Kocherginsky et al. (2005)].

2.7 Rank score test in quantile regression

To overcome the difficulties of estimating the unspecified error density function, which is an infinite dimensional nuisance parameter, we turn to an alternative rank-based approach. In this section, we provide a brief review on the rank-based inference framework summarized in Koenker and Machado (1999).

We first describe the rank-based inference framework for the one-sample problem of estimating the τth quantile ξ of the outcome y_i. We need to minimize $\sum \rho_\tau(y_i - \xi)$ by formulating it as the following linear programming problem

$$\min_{(\xi,u,v)\in\mathbb{R}\times\mathbb{R}_+^{2n}} \{\tau 1_n^\top u + (1-\tau) 1_n^\top v \mid 1_n\xi + u - v = y\}. \tag{2.7.29}$$

In general, the original formulation of any linear programming problem is called the basic problem, which has a corresponding dual problem. Here the basic problem is

to generate the order statistic or quantile in the one-sample setting and the dual problem is to generate the ranks.

Rank-based inference can be generalized to the regression setting using the regression rank score process introduced by Gutenbrunner and Jurečková (1992) ("GJ"). The regression rank score process arises from the dual problem of linear programming as formulated in (2.7.31). It provides a natural generalization to the previously described statistical duality of the order statistics and ranks in the one-sample problem. As such, it can be viewed as providing a fundamental link between quantile regression and the classical theory of rank statistics [Hájek and Šidák (1967)]. The rank score process may also be interpreted as an implementation of the Rao score for quantile regression inference.

The regression rank score process for the restricted form of the linear model

$$y_i = x_i\beta + z_i\gamma + e_i \tag{2.7.30}$$

is given by

$$\widehat{\mathbf{a}}_n(\tau) = \arg\max\{Y^\top \mathbf{a} \mid X^\top \mathbf{a} = (1-\tau)X^\top 1_n, \mathbf{a} \in [0,1]^n\} \tag{2.7.31}$$

where $Y = (y_1, \ldots, y_n)^\top$. The problem posed in (2.7.31) is the formal dual problem corresponding to the (basic) quantile regression linear program under the restriction imposed by $H_0 : \beta = 0$. Based on Theorem a of Gutenbrunner and Jurečková (1992), Theorem 5.1 of Gutenbrunner et al. (1993) ("GJKP") may be extended to conform to the conditions of the location shift model (2.7.30). The rank score test of the hypothesis H_0 is based on the statistic

$$T_n = S_n^\top M_n^{-1} S_n / \tau(1-\tau), \tag{2.7.32}$$

where

$$
\begin{aligned}
S_n &= n^{-1/2}(X - \widehat{X})^\top \widehat{\beta}_n, \\
\widehat{X} &= Z(Z^\top Z)^{-1} Z^\top X, \\
M_n &= (X - \widehat{X})^\top (X - \widehat{X})/n, \\
\widehat{\beta}_n &= \left(- \int \psi_\tau(t) d\widehat{a}_{n,i}(t) \right)_{i=1}^n ,
\end{aligned}
\tag{2.7.33}
$$

and $\psi_\tau(u) = \tau - I\{u < 0\}$ is the τth-quantile score function, the piecewise first derivative of $\rho_\tau(u)$.

Several regularity conditions are listed.

G1. $X_{i1} \equiv 1$, for $i = 1, \ldots, n$, i.e. the design matrix contains an intercept.

G2. $D_n = n^{-1} X_n^\top X_n \to D_0$, a positive definite matrix. [same as R2(i)]

G3. $n^{-1} \sum \|X_i^4\| = O(1)$,

G4. $\max \|X_i\| = O(n^{1/4}/\log n)$,

Theorem 2.7.2 GJKP (1993) *Under conditions G1-G4,*

$$
T_n \xrightarrow{D} \chi_q^2 \tag{2.7.34}
$$

where q is the dimension of β.

A crucial feature of the rank score test T_n is that under H_0, it does not involve any estimation of the nuisance parameter $s(\tau)$. This is a substantial advantage over Wald type approaches to testing in quantile regression.

2.8 Resampling methods and the bootstrap

There are also extensive literature on developing bootstrap methods in quantile regression to circumvent the problem of estimating nuisance parameters. We review

a few popular methods based on resampling approaches.

One of the most popular methods is the (x, y)pair method or paired bootstrap. (x_i^*, y_i^*) is drawn with replacement from the n pairs $\{(x_i, y_i) : i = 1, \ldots, n\}$ of the original sample, each with equal probability of $1/n$. This form of bootstrap has been widely employed in applications of quantile regression. Given the bootstrap realizations $\widehat{\beta}_b^*(\tau) : b = 1, \ldots, B$, there are several options which can be considered for constructing tests and confidence intervals. Most straightforwardly, one can compute the empirical covariance matrix of the realizations and construct tests and confidence intervals directly from it. Buchinsky (1995) concludes that the (x, y)-pair method performs well based on an extensive Monte Carlo experiment comparing several variants of the bootstrap.

Alternatively, one can form percentile intervals as in Efron and Tibshirani (1993). Many important practical aspects of the implementation of the bootstrap, including the important question of how to choose the number of replications, R, are treated by Andrews and Buchinsky (2001).

There are also several proposals on some refinements of the (x, y)-pair bootstrap based on smoothing [De Angelis et al. (1993), Horowitz (1998)]. Another related approach developed to refine the bootstrap inference is provided by saddle point methods introduced by Daniels (1954). Spady (1991) has explored the saddle-point approach for a median regression estimator for bimodal responses. More recent work by de Jongh et al. (1994) reconfirms the advantage of this method. Parzen et al. (1994) proposed a method based on resampling the subgradient, $S_n(\beta) = -n^{-1/2} \sum_{i=1}^n x_i \psi_\tau(y_i - x_i^\top \beta)$. For problems involving high parametric dimensions, He and Hu (2002) developed an approach which aims to alleviate the computational burden by considering using a Markov chain resampler based on the solutions to the marginal coefficient-wise version of the estimating equations for general M-estimation problems. This approach has been implemented in quantile regression by Kochergin-

sky et al. (2005) and turns out attractive in large problems involving high parametric dimensions.

2.9 Monte Carlo comparison of methods

In this section, we conclude this chapter with a report on a small-scale Monte Carlo simulation designed to compare the performance of the methods reviewed earlier. This is a confirmation study similar to the ones in Koenker (2005). Three classes of methods are included: two of the inverted rank score methods, three of the computationally less demanding Wald methods, and three of the resampling methods. The study also focuses exclusively on the problem of confidence intervals for median regression parameters. Three models are considered: a pure location-shift model and two more complicated location-scale-shift models. All eight methods are implemented as options in the function `summary.rq` of the Quantreg package in `R`[R Development Core Team (2009)]. They include the following eight methods:

`Riiid` - rank score test inversion assuming iid errors

`Rnid` - rank score test inversion assuming independent, not identically distributed (nid) errors

`Wiid` - Wald confidence interval assuming iid errors, with scalar sparsity estimate

`Wker` - Wald confidence interval assuming nid errors, with Powell's estimate

`Wnid` - Wald confidence interval assuming nid errors, with Siddiqui's estimate

`Bxy` - Bootstrap confidence interval using (x, y)-pair method

`Bpwy` - Bootstrap confidence interval by "Parzen-Wei-Ying" method [Parzen et al. (1994)]

`Bmcmb` - Bootstrap confidence interval by Markov chain marginal bootstrap method [He and Hu (2002)]

For each method, we calculate the confidence intervals that are intended to have coverage of 0.95, the median length of the confidence intervals and the average CPU time for each parameter. These results are based on $1,000$ Monte Carlo replications. Our objective is to assess whether the intervals have the desired coverage. If not, is the coverage too high or too low? It could also be interesting to compare the coverage of the intervals for different coefficients among different methods.

As the first model, we consider a classical linear regression model in which the covariates exert a pure location shift effect taking the form:

$$y_i = a + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + e_i. \tag{2.9.35}$$

The e_i's are taken to be iid Student t on three degrees of freedom, and each of the covariates is also independently generated from the $t(3)$ distribution. The true coefficients are set $a = b_1 = b_2 = b_3 = 0$; so the conditional median of the response given the covariates is zero.

Regarding coverage of the intervals, the two Wald methods designed to adapt to non-iid error models exhibit some size distortion, i.e. they fail to maintain the nominal type I error rate. The remaining methods perform reasonably well, having quite accurate coverage and very similar median lengths. The Wald methods have an advantage in terms of computational efficiency. Mean lengths between methods are quite similar except in the case of the *Rnid* method, which has a small number of realizations with infinite length.

In model 2, we consider a case in which a single covariate has an effect on both the location and the scale of the conditional distribution of the response variable.

$$y_i = a + b_1 x_i + (1 + x_i)^2 e_i. \tag{2.9.36}$$

In this model, we take the e_i as iid standard normal, and the x_i's are generated from $\chi_3^2/3$. Again we take the coefficients $a = b_1 = 0$, and so the conditional median of the response is zero.

Table 2.2 reports the results from model 2. Not surprisingly, the methods based on the iid error assumption exhibit more or less size distortion, with `Riid` showing minor deviation and `Wiid` suffering serious distortion. On the other hand, the rank test inversion assuming nid errors performs well, as do the bootstrap procedures. The Wald interval with Powell's sandwich estimate `Wker` is overly optimistic. In contrast, the Wald interval with Siddiqui's sandwich estimate `Wnid` is somewhat too conservative, and has interval lengths which are consequently larger than all others.

In model 3, we consider a variant of the previous model 2:

$$y_i = a + b_1 x_i + b_2 x_i^2 + (1 + x_i)^2 e_i \tag{2.9.37}$$

where the quantile regression model is specified such that the median function of y_i is a quadratic function in x_i. Again we take the coefficients $a = b_1 = b_2 = 0$, so the conditional median of the response is zero.

Table 2.3 reports the results from model 3, treating the estimated model as quadratic in x_i. Now it is of interest to distinguish the three separate coefficients. Again the rank-based intervals assuming nid errors perform quite well, as do the bootstrap procedures. The Wald iid and nid intervals are somewhat unstable. The computational cost of the rank-based intervals are roughly comparable for the Wald type methods.

In summary, the two rank inversion methods show reliable performances in different settings, as well as the three bootstrap methods. The Wald methods, however, exhibit some instability. These observations are in line with the reports in Koenker (2005), Chen and Wei (2005), and Kocherginsky et al. (2005).

Table 2.1: Confidence interval performance: model 1

	Riid	Rnid	Wiid	Wker	Wnid	Bxy	Bpwy	Bmcmb
Coverage								
n=100 a	0.926	0.935	0.929	0.952	0.997	0.954	0.956	0.945
n=100 b_1	0.967	0.971	0.929	0.902	0.995	0.973	0.975	0.967
n=500 a	0.946	0.949	0.936	0.955	0.981	0.947	0.950	0.942
n=500 b_1	0.952	0.954	0.951	0.906	0.984	0.961	0.967	0.961
Length								
n=100 a	0.538	0.554	0.543	0.604	0.863	0.596	0.610	0.560
n=100 b_1	0.378	0.390	0.339	0.340	0.535	0.402	0.421	0.420
n=500 a	0.237	0.239	0.238	0.249	0.302	0.249	0.250	0.238
n=500 b_1	0.149	0.150	0.114	0.143	0.182	0.155	0.157	0.153
CPU time								
n=100 a	0.004	0.015	0.006	0.005	0.005	0.011	0.012	0.013
n=100 b_1	0.004	0.014	0.006	0.005	0.005	0.011	0.012	0.013
n=500 a	0.015	0.029	0.007	0.009	0.006	0.043	0.051	0.042
n=500 b_1	0.015	0.029	0.007	0.009	0.007	0.043	0.050	0.042

`Riid` - rank score test inversion assuming iid errors

`Rnid` - rank score test inversion assuming independent, not identically distributed (nid) errors

`Wiid` - Wald confidence interval assuming iid errors, with scalar sparsity estimate

`Wker` - Wald confidence interval assuming nid errors, with Powerll's estimate

`Wnid` - Wald confidence interval assuming nid errors, with Siddiqui's estimate

`Bxy` - Bootstrap confidence interval using (x, y) pair method

`Bpwy` - Bootstrap confidence interval using "Parzen-Wei-Ying" method

`Bmcmb` - Bootstrap confidence interval using Markov chain marginal bootstrap method

Table 2.2: Confidence interval performance: model 2

	Riid	Rnid	Wiid	Wker	Wnid	Bxy	Bpwy	Bmcmb
Coverage								
n=100 a	0.886	0.958	0.909	0.816	0.997	0.965	0.962	0.981
n=500 b_1	0.924	0.956	0.537	0.918	0.966	0.946	0.952	0.948
n=100 a	0.893	0.952	0.924	0.780	0.996	0.955	0.950	0.984
n=500 b_1	0.913	0.957	0.533	0.899	0.960	0.940	0.944	0.943
Length								
n=100 a	1.946	2.508	2.184	1.804	4.296	2.574	2.604	3.090
n=500 b_1	4.153	4.982	1.709	4.113	5.663	4.784	4.869	4.907
n=100 a	0.846	1.037	0.968	0.776	1.466	1.074	1.081	1.292
n=500 b_1	1.789	2.125	0.752	1.780	2.281	2.066	2.075	2.063
CPU time								
n=100 a	0.003	0.009	0.006	0.005	0.005	0.007	0.008	0.010
n=500 b_1	0.003	0.009	0.005	0.005	0.005	0.007	0.008	0.010
n=100 a	0.006	0.014	0.006	0.008	0.005	0.025	0.031	0.027
n=500 b_1	0.006	0.014	0.006	0.008	0.005	0.025	0.031	0.026

`Riid` - rank score test inversion assuming iid errors

`Rnid` - rank score test inversion assuming independent, not identically distributed (nid) errors

`Wiid` - Wald confidence interval assuming iid errors, with scalar sparsity estimate

`Wker` - Wald confidence interval assuming nid errors, with Powerll's estimate

`Wnid` - Wald confidence interval assuming nid errors, with Siddiqui's estimate

`Bxy` - Bootstrap confidence interval using (x, y) pair method

`Bpwy` - Bootstrap confidence interval using "Parzen-Wei-Ying" method

`Bmcmb` - Bootstrap confidence interval using Markov chain marginal bootstrap method

Table 2.3: Confidence interval performance: model 3

	Riid	Rnid	Wiid	Wker	Wnid	Bxy	Bpwy	Bmcmb
Coverage								
n=100 a	0.819	0.929	0.921	0.846	0.994	0.952	0.964	0.979
n=100 b_1	0.819	0.945	0.609	0.880	0.920	0.927	0.949	0.947
n=100 b_2	0.866	0.944	0.390	0.838	0.755	0.913	0.942	0.936
n=500 a	0.833	0.939	0.920	0.905	0.985	0.957	0.960	0.970
n=500 b_1	0.825	0.938	0.532	0.920	0.904	0.932	0.945	0.905
n=500 b_2	0.877	0.947	0.310	0.903	0.782	0.930	0.943	0.909
Length								
n=100 a	2.339	3.384	3.086	2.770	5.217	3.305	3.504	4.337
n=100 b_1	7.628	11.326	4.932	9.703	10.955	10.821	11.675	12.985
n=100 b_2	4.318	5.898	1.454	5.238	3.841	5.660	6.236	7.281
n=500 a	1.021	1.448	1.307	1.408	1.837	1.478	1.517	1.637
n=500 b_1	3.484	4.919	1.924	4.848	4.819	5.085	5.141	4.609
n=500 b_2	2.023	2.622	0.542	2.571	2.094	2.687	2.732	2.527
CPU time								
n=100 a	0.004	0.012	0.006	0.005	0.005	0.009	0.010	0.011
n=100 b_1	0.004	0.012	0.007	0.005	0.005	0.009	0.010	0.011
n=100 b_2	0.004	0.012	0.007	0.005	0.005	0.009	0.010	0.011
n=500 a	0.012	0.025	0.007	0.009	0.006	0.037	0.043	0.036
n=500 b_1	0.011	0.024	0.007	0.009	0.006	0.036	0.043	0.035
n=500 b_2	0.012	0.025	0.007	0.009	0.006	0.036	0.044	0.035

`Riid` - rank score test inversion assuming iid errors
`Rnid` - rank score test inversion assuming nid errors
`Wiid` - Wald confidence interval assuming iid errors, with scalar sparsity estimate
`Wker` - Wald confidence interval assuming nid errors, with Powerll's estimate
`Wnid` - Wald confidence interval assuming nid errors, with Siddiqui's estimate
`Bxy` - Bootstrap confidence interval using (x, y) pair method
`Bpwy` - Bootstrap confidence interval using "Parzen-Wei-Ying" method
`Bmcmb` - Bootstrap interval using Markov chain marginal bootstrap method

Chapter 3

Bent line quantile regression for longitudinal data

3.1 Overview

This chapter describes the model framework we have employed to answer the questions described in Chapter 1, characterization of the change-point problem in quantile regression for correlated data. In the first section, the notation and the model that we are going to use will be introduced, based on which, the estimates of the change-point and the regression coefficients are derived. The asymptotic behavior of these estimates are established in the second section. In the third section, we discuss some difficulties with existing inferential approaches.

3.2 Model specification and Notation

In this section, we extend the original bent line quantile regression model for independent data developed by Li et al. (2011) into a longitudinal setting. To establish the notation, suppose we have m subjects, with each measured n_i times, resulting in a total of $n = \sum_i^m n_i$ observations. Let $y_{i1}, y_{i2}, \ldots, y_{i,n_i}$ be the outcome values measured from the ith subject at different time points $t_{i1}, t_{i2}, \ldots, t_{i,n_i}$, which may or may not

be evenly spaced. For a given quantile level $\tau \in (0,1)$, we model y_{ij} by

$$y_{ij} = a_\tau + b_{1,\tau}(t_{ij} - t_\tau)_- + b_{2,\tau}(t_{ij} - t_\tau)_+ + \mathbf{x}_{ij}^\top \boldsymbol{\gamma}_\tau + e_{ij}(\tau), \tag{3.2.1}$$

where $i = 1, \ldots, m,\ j = 1, \ldots, n_i,\ (u)_+ = \max(u, 0),\ (u)_- = \min(u, 0),\ \mathbf{x}_{ij}$ is a p-dimensional vector of linear covariates with constant slopes, and $e_{ij}(\tau)$ denotes the error term, whose τth quantile, given $(t_{ij}, \mathbf{x}_{ij}^\top)$, is zero. $(a_\tau,\ b_{1,\tau},\ b_{2,\tau},\ \boldsymbol{\gamma}_\tau^\top,\ t_\tau)$ are the parameters. Here t_τ represents the time when the slope coefficient with respect to t_{ij} changes abruptly from $b_{1,\tau}$ to $b_{2,\tau}$, is thus referred to as a change-point or cusp-point [Li et al. (2011), Chu and Zhao (2004), Park and Kim (2004)],

We have intentionally written the error term in model (3.2.1) as $e_{ij}(\tau)$, which is τ-specific. This may help distinguish quantile regression from most other regression models that appear in the same form. Note that specification of the error $e_{ij}(\tau)$ in Model (3.2.1) is very general and flexible, even allowing for the error term to depend on the covariates as well. If the error terms are independent of the covariates, the regression coefficients are invariant to the quantile levels. In that case, the model is comparable to those well studied segmented mean regression models with i.i.d. normal errors [Hinkley (1971), Johnstone and Siegmund (1989)] or correlated errors [Lee (1993), Piepho and Ogutu (2003)].

To simplify the presentation, we denote the conditional τth quantile function in (3.2.1)

$$g(\tilde{x}_{ij}; \boldsymbol{\theta}_\tau) = a_\tau + b_{1,\tau}(t_{ij} - t_\tau)_- + b_{2,\tau}(t_{ij} - t_\tau)_+ + \mathbf{x}_{ij}^\top \boldsymbol{\gamma}_\tau, \tag{3.2.2}$$

where $\tilde{x}_{ij} = (t_{ij}, \mathbf{x}_{ij}^\top)^\top$, and $\boldsymbol{\theta}_\tau = (a_\tau, b_{1,\tau}, b_{2,\tau}, \boldsymbol{\gamma}_\tau^\top, t_\tau)^\top$ is the $(p+4)$-dimensional unknown parameter vector. An estimator of $\boldsymbol{\theta}_\tau$ can be expressed by

$$\widehat{\boldsymbol{\theta}}_{n,\tau} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^{p+4}} \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \rho_\tau \left(y_{ij} - g(\tilde{x}_{ij}; \boldsymbol{\theta}) \right), \tag{3.2.3}$$

where $\boldsymbol{\theta} = (a, b_1, b_2, \boldsymbol{\gamma}^\top, t)^\top$, and $\rho_\tau(u) = u(\tau - I\{u < 0\})$ is the *quantile regression loss function*. Due to the non-convexity of the objective function in (3.2.3), $\widehat{\boldsymbol{\theta}}_{n,\tau}$ is obtained via the profile estimate procedure discussed in Li et al. (2011) with a slight modification.

Specifically, let $\boldsymbol{\eta}_\tau = (a_\tau, b_{1,\tau}, b_{2,\tau}, \boldsymbol{\gamma}_\tau^\top)^\top$ denote the true parameters excluding t_τ. Our objective function in (3.2.3) can be expressed in the partitioned parameter space $\{(\boldsymbol{\eta}, t) \in \mathbb{R}^{p+3} \times \mathbb{R}\}$

$$Q_{n,\tau}(\boldsymbol{\eta}, t) = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \rho(y_{ij} - g(\tilde{x}_{ij}; \boldsymbol{\eta}, t)). \tag{3.2.4}$$

A profile estimate of $\boldsymbol{\eta}_\tau$ at a fixed t is given by

$$\widehat{\boldsymbol{\eta}}_{n,\tau}(t) = \arg\min_{\boldsymbol{\eta}} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \rho_\tau \left(y_{ij} - g(\tilde{x}_{ij}; \boldsymbol{\eta}, t)\right), \tag{3.2.5}$$

where $\boldsymbol{\eta} = (a, b_1, b_2, \boldsymbol{\gamma}^\top)^\top$. Then an estimate of t_τ is given by

$$\widehat{t}_{n,\tau} = \arg\min_{t} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \rho_\tau \left(y_{ij} - g(\tilde{x}_{ij}; \widehat{\boldsymbol{\eta}}_{n,\tau}(t), t)\right). \tag{3.2.6}$$

Finally $\widehat{\boldsymbol{\theta}}_{n,\tau}$ is obtained from

$$\widehat{\boldsymbol{\eta}}_{n,\tau}(\widehat{t}_{n,\tau}). \tag{3.2.7}$$

In some applications, investigators might wish to incorporate prior knowledge on the parameters. For example $b_{2,\tau} < b_{1,\tau}$, which means the improvement in the second stage will be slower than that in the first stage. Another example is $b_{2,\tau} = 0$, meaning that the patient's condition will no longer improve in the second stage, but stabilizes. Such information can also be easily implemented by adding the corresponding constraints into the estimation procedures (3.2.5) and (3.2.6).

3.3 Asymptotic behavior of $\widehat{\boldsymbol{\theta}}_{n,\tau}$

Before we present the asymptotic properties of $\widehat{\boldsymbol{\theta}}_{n,\tau}$, we introduce the following notation.

$$
\begin{aligned}
h(\tilde{x}_{ij};\boldsymbol{\theta}) &= (I\{t_{ij} \le t\}, t_{ij}I\{t_{ij} \le t\}, I\{t_{ij} > t\}, t_{ij}I\{t_{ij} > t\}, \mathbf{x}_{ij}^{\top})^{\top}, \\
g_1(\tilde{x}_{ij};\boldsymbol{\theta}) &= a + b_{1,\tau}(t_{ij} - t) + \mathbf{x}_{ij}^{\top}\boldsymbol{\gamma}, \\
g_2(\tilde{x}_{ij};\boldsymbol{\theta}) &= a + b_{2,\tau}(t_{ij} - t) + \mathbf{x}_{ij}^{\top}\boldsymbol{\gamma}, \\
C_{n,\tau} &= n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i} E\{h(\tilde{x}_{ij};\boldsymbol{\theta}_\tau)h(\tilde{x}_{ij};\boldsymbol{\theta}_\tau)^{\top}\}, \\
D_{n,\tau} &= n^{-1}\frac{\partial E\sum_{i=1}^{m}\sum_{j=1}^{n_i}\psi_\tau(y_{ij} - g(\tilde{x}_{ij};\boldsymbol{\theta}))h(\tilde{x}_{ij};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\tau} \\
C_\tau &= \lim_{n\to\infty} C_{n,\tau}, \text{ and } D_\tau = \lim_{n\to\infty} D_{n,\tau}.
\end{aligned}
\tag{3.3.8}
$$

where $\psi_\tau(u) = \tau - I\{u < 0\}$, called *$\tau$-quantile score function*, is the piecewise first derivative of $\rho_\tau(u)$. The regularity conditions are listed in Chapter 8 with brief discussions. Next we are going to establish the asymptotic behavior of $\widehat{\boldsymbol{\theta}}_{n,\tau}$.

Theorem 3.3.3 *Under the Conditions A0-A5 (See Chapter 8) and $b_{1,\tau} \neq b_{2,\tau}$, $\widehat{\boldsymbol{\theta}}_{n,\tau}$ has the following Bahadur representation:*

$$
\widehat{\boldsymbol{\theta}}_{n,\tau} - \boldsymbol{\theta}_\tau = -n^{-1}D_{n,\tau}^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\psi_\tau(e_{ij})h(\tilde{x}_{ij},\boldsymbol{\theta}_\tau) + o_p(n^{-1/2}), \tag{3.3.9}
$$

Representation (3.3.9) implies that

$$
n^{1/2}(\widehat{\boldsymbol{\theta}}_{n,\tau} - \boldsymbol{\theta}_\tau) \xrightarrow{D} N(0, \Sigma_\tau), \tag{3.3.10}
$$

where $\Sigma_\tau = \tau(1-\tau)D_\tau^{-1}C_\tau D_\tau^{-1}$.

Remark 1. When $b_{1,\tau} = b_{2\tau}$, i.e. the change-point does not exist, the estimation is ill-conditioned.

Remark 2. Under the condition $b_{1,\tau} \neq b_{2,\tau}$, the optimization problem in (3.2.3) is equivalent to solving the following estimation equation for $\boldsymbol{\theta}$,

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} \psi_\tau(y_{ij} - g(\tilde{x}_{ij};\boldsymbol{\theta}))h(\tilde{x}_{ij};\boldsymbol{\theta}) = 0. \tag{3.3.11}$$

These estimation equations have a nice interpretation: the quantile subgradient condition holds on both sides of the change-point t_τ at $\widehat{\boldsymbol{\theta}}_{n,\tau}$.

Remark 3. The estimation of the limiting variance-covariance matrix Σ_τ involves nuisance parameters, that is, the density function of e_{ij} evaluated at the τth quantile.

Theorem 1 in Li et al. (2011) is a special case of the above theorem with $n_i \equiv 1$, that is, each subject has only one outcome measured. The proof of Theorem 3.3.3 follows similar arguments in Li et al. (2011) with modifications tailored for longitudinal settings. Technical details are included in Chapter 8.

3.4 Existing inferential approaches and their limitations

Thus far in this chapter we have focused on the asymptotic behavior of the estimate $\widehat{\boldsymbol{\theta}}_{n,\tau}$. In order to develop inference tools, one can explore the asymptotic normality of the quantile estimators $\widehat{\boldsymbol{\theta}}_{n,\tau}$ of the parameters $\boldsymbol{\theta}_\tau$ in model 3.2.1. However, following the proof of Theorem 3.3.3, one can find that the asymptotic variance-covariance matrix of these estimators involve a function of the unspecified density of the error terms. This is difficult to estimate reliably, seriously hampering the use of a Wald-type test. Moreover, it has been shown that, in a quantile regression set up,

a Wald-type test is generally unstable at small sample sizes [Chen and Wei (2005), Kocherginsky et al. (2005)]. On the other hand, the use of likelihood ratio-type tests, e.g. ρ-test for linear models [Koenker and Machado (1999)], is also difficult, again due to the challenges in estimating the density of the errors.

It is a somewhat unhappy fact, as Koenker put it in his book *Quantile Regression*, that the asymptotic precision of quantile regression estimates depends on the reciprocal of a density function evaluated at the quantile of interest – a quantity Tukey (1965) termed the "sparsity function". This quantity characterizes the sample information in the neighborhood of the τth quantile. There is extensive literature on the estimation of the sparsity function [Koenker and Bassett (1982), Welsh (1987)]. However, dissatisfaction existed with these methods, motivating development of methods based on the bootstrap, e.g. [Efron and Tibshirani (1986), Parzen et al. (1994), Buchinsky (1994), Horowitz (1998)]. Unfortunately, in longitudinal bent line regression, bootstrap based methods suffer from a more serious problem of expensive computational costs, due to the repeated profile estimation involved, which is known to be computationally demanding. To avoid all these challenging issues, we turn to the ideas of rank based inference. Specifically, we extend the quantile rank score test proposed in Gutenbrunner et al. (1993) for the linear model with iid error terms. To our best knowledge, our attempt is the first to extend the rank score test to non-linear models.

Chapter 4

Hypothesis testing on the change-point

4.1 Overview

As discussed in Chapter 2, there are major difficulties, e.g. estimation of nuisance parameter or high computational cost, with existing approaches, summarized in sections (2.6) and (2.8) for inference on the change-point in model (3.2.1). To overcome these difficulties, we apply the rank-based approach introduced in section (2.7). The rank score test given by Gutenbrunner et al. (1993) provided an attractive alternative way for hypothesis testing problems in linear quantile regression by avoiding direct estimation of the error densities. In this chapter we extend the original rank score test (2.7.32) for linear models to the non-linear setting (3.2.1). In the first section, the rank score test statistic T_n for the change-point is derived. The asymptotic properties of T_n is established in the second section. In the third section, we discuss some possible variations to incorporate different study designs.

4.2 Rank score test on the change-point

Defining the outcome vector $Y = (y_{11}, \ldots, y_{m,n_m})^\top$, the error vector $\epsilon = (e_{11}, \ldots, e_{m,n_m})^\top$, the design matrix $\mathcal{W}(t_\tau) = (1, (t_{ij} - t_\tau)_-, (t_{ij} - t_\tau)_+, x_{ij}^\top)_{n\times(p+3)}$, and the regression

coefficient vector $\boldsymbol{\eta}_\tau = (a_\tau, b_{1,\tau}, b_{2,\tau}, \boldsymbol{\gamma}_\tau)$, the model (3.2.1) can then be rewritten in the following matrix form

$$Y = \mathcal{W}(t_\tau)\boldsymbol{\eta}_\tau + \epsilon. \tag{4.2.1}$$

Suppose we wish to test the null hypothesis that the change-point t_τ is at a pre-specified location t_0, i.e. $H_0 : t_\tau = t_0$ versus the alternative hypothesis $H_1 : t_\tau \neq t_0$. When H_0 is true, $\mathcal{W}(t_0)$ is the design matrix. As we aim to derive a rank score test statistic under H_0, the notation of the design matrix can be simplified and rewritten as $\mathcal{W}(t_0) = \mathcal{W}_0$ for easier presentation.

We further define

$$z(t_{ij}; \boldsymbol{\eta}) = b_1 I\{t_{ij} \leq t_0\} + b_2 I\{t_{ij} > t_0\} \tag{4.2.2}$$

as the first derivative of $g(\tilde{x}_{ij}; \boldsymbol{\eta})$ with respect to t. Then we further denote

$$\begin{aligned} \mathcal{Z}(\widehat{\boldsymbol{\eta}}) &= (z(t_{1,1}; \widehat{\boldsymbol{\eta}}), \ldots, z(t_{m,n_m}; \widehat{\boldsymbol{\eta}}))^\top, \\ \mathcal{P}_0 &= \mathcal{W}_0(\mathcal{W}_0^\top \mathcal{W}_0)^{-1}\mathcal{W}_0^\top, \quad \mathcal{Z}(\widehat{\boldsymbol{\eta}})^* = (I - \mathcal{P}_0)\mathcal{Z}(\widehat{\boldsymbol{\eta}}), \end{aligned} \tag{4.2.3}$$

where $\widehat{\boldsymbol{\eta}}$ minimizes $\sum_{ij} \rho_\tau (y_{ij} - g(\tilde{x}_{ij}; \boldsymbol{\eta}))$. $z^*(t_{ij}; \widehat{\boldsymbol{\eta}})$ is the element of $Z^*(\widehat{\boldsymbol{\eta}})$. The orthogonal transformation in the definition of $\mathcal{Z}(\widehat{\boldsymbol{\eta}})^*$ in equation (4.2.3) ensures the asymptotic independence between $z^*(t_{ij}; \widehat{\boldsymbol{\eta}})$ and $\psi_\tau(\widehat{e}_{ij})$, where $\psi_\tau(\cdot)$ is the τth quantile score funtion mentioned earlier.

The rank score test on the change-point parameter is based on

$$S_n = n^{-1/2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \psi_\tau(\widehat{e}_{ij}) z^*(t_{ij}; \widehat{\boldsymbol{\eta}}) \tag{4.2.4}$$

where $\widehat{e}_{ij} = y_{ij} - \mathbf{w}_{ij}\widehat{\boldsymbol{\eta}}$. Letting

$$V_n = n^{-1} \sum_{i=1}^{m} z^*(t_{i\cdot}; \widehat{\boldsymbol{\eta}})^\top \widehat{A}_i z^*(t_{i\cdot}; \widehat{\boldsymbol{\eta}}), \tag{4.2.5}$$

where $z^*(t_{i\cdot}; \cdot) = (z^*(t_{i,1}; \cdot), \ldots, z^*(t_{i,n_i}; \cdot))^\top$, and $\widehat{A}_i$ is a $n_i \times n_i$ matrix whose (j, j')th element is $\psi_\tau(\widehat{e}_{ij})\psi_\tau(\widehat{e}_{ij'})$. Here j and j' index the rows and columns of $\widehat{A}_i$, respectively.

Now we define the rank score test statistics in quadratic form as

$$T_n = S_n^2 / V_n. \tag{4.2.6}$$

4.3 Asymptotic distribution of T_n

We now present the main result of the asymptotic property of the proposed rank score test in the following theorem but defer the proof to Chapter 8.

Theorem 4.3.4 *: Under H_0 and the assumptions (A0), (A1*), and (A3)-(A7) in Chapter 8, the distribution of T_n converges to χ_1^2 as $n \to \infty$.*

4.4 Variants of T_n incorporating dependence structures

The rank score we derived in the previous section does not require any specification of the dependence structure among the repeated measures $y_{11}, \ldots, y_{m,n_m}$ or equivalently, $e_{11}, \ldots, e_{m,n_m}$. In some applications where the correlation structure is

already known from the study design, the original unspecified version of T_n can be easily tailored to incorporate the specific variance-covariance correlation matrix for efficiency considerations. For example, in microarray studies, a compound symmetry (C.S.) structure is commonly assumed to take into account the probe effects [Wang and He (2007)]. The rank score test can be tailored to incorporate different specified structures, based on the following facts,

$$\begin{aligned}\mathrm{Var}(S_n) = n^{-1}\sum_i (z^*(t_{i,i};\widehat{\boldsymbol{\eta}}))^2\tau(1-\tau) \\ + \sum_i \sum_{j\neq j'} z^*(t_{i,j};\widehat{\boldsymbol{\eta}})z^*(t_{i,j'};\widehat{\boldsymbol{\eta}})\mathrm{Cov}(\psi(e_{i,j}),\psi(e_{i,j'}))\end{aligned} \tag{4.4.7}$$

and

$$\begin{aligned}\mathrm{Cov}(\psi(e_{i,j}),\psi(e_{i,j'})) &= \mathrm{Cov}(I(e_{i,j}<0), I(e_{i,j'}<0)) \\ &= P(e_{i,j}<0, e_{i,j'}<0) - P(e_{i,j}<0)P(e_{i,j'}<0) \\ &= P(e_{i,j}<0, e_{i,j'}<0) - \tau^2, \text{ for all } j\neq j',\end{aligned} \tag{4.4.8}$$

where the last equation holds under the null hypothesis. In the following subsections, we discuss some possible variations for common study designs by specifying and estimating the joint probability $P(e_{ij}<0, e_{ij'}<0)$ in (4.4.8) accordingly.

4.4.1 Compound symmetry

Compound symmetry (C.S.), also referred to as the Exchangeable structure, is perhaps the most commonly seen correlation structure. It arises from a random effects model with a common random subject level, e.g. [Liang and Zeger (1986)]. The rank score test T_n can easily incorporate this *a priori* knowledge via a corresponding adaptation only on the V_n part, by modifying $P(e_{i,j}<0, e_{i,j'}<0)=\delta$ for all $j\neq j'$

and some $\delta \geq 0$. Specifically, let

$$V_n^{\text{CS}}(\delta) = n^{-1} \sum_i (z^*(t_{i,i}; \widehat{\boldsymbol{\eta}}))^2 \tau(1-\tau) + \sum_i \sum_{j \neq j'} z^*(t_{i,j}; \widehat{\boldsymbol{\eta}}) z^*(t_{i,j'}; \widehat{\boldsymbol{\eta}})(\delta - \tau^2) \quad (4.4.9)$$

with

$$\delta = P(e_{11} < 0, e_{12} < 0). \quad (4.4.10)$$

We can define the rank score test statistic incorporating C.S. structure as

$$T_n^{\text{CS}} = S_n^2 / V_n^{\text{CS}}(\widehat{\delta}) \quad (4.4.11)$$

where

$$\widehat{\delta} = (\sum_i L_i)^{-1} \sum_i \sum_{j \neq j'} I(\widehat{e}_{i,j} < 0, \widehat{e}_{i,j'} < 0) \quad (4.4.12)$$

with $L_i = n_i(n_i - 1)/2$.

4.4.2 Unspecified correlation structure in study designs following a fixed time schedule

As opposed to studies where the measurement time t_{ij}'s are taken randomly on a finite interval, another commonly seen design is where the t_{ij}'s follow a fixed time schedule, which we designate as FD. For example, the patients enrolled in a clinical study will have their measurement taken every 4 weeks in the first two months and every 8 weeks thereafter. In this case, we modify the V_n and let

$$V_n^{FD} = n^{-1} \sum_i (z^*(t_{i,i}; \widehat{\boldsymbol{\eta}}))^2 \tau(1-\tau) + n^{-1} \sum_i \sum_{j \neq j'} z^*(t_{i,j}; \widehat{\boldsymbol{\eta}}) z^*(t_{i,j'}; \widehat{\boldsymbol{\eta}}) v_{j,j'}. \quad (4.4.13)$$

Here we define $v_{j,j'} = \frac{1}{\#\{i:n_i \geq j, n_i \geq j'\}-1} \sum_i (\psi(\widehat{e}_{i,j}) - \bar{\psi}_j)(\psi(\widehat{e}_{i,j}) - \bar{\psi}_{j'}) I(n_i \geq j, n_i \geq j')$. In the previous formula, $\bar{\psi}_j = \frac{1}{\#\{i:n_i \geq j\}} \sum_i \psi(\widehat{e}_{i,j}) I(n_i \geq j)$, where, abusing notation, j applies to both j and j'. $v_{j,j'}$ is the sample covariance estimate, i.e. $\widehat{\mathrm{Cov}}(\psi(\widehat{e}_{i,j}), \psi(\widehat{e}_{i,j'}))$. The corresponding rank score test for this design is defined

$$T_n^{FD} = S_n^2 / V_n^{FD}. \tag{4.4.14}$$

4.4.3 Time spacing-dependent structure

A third possibility for adapting the known correlation structure is to allow $P(e_{i,j} < 0, e_{i,j'} < 0)$, denoted by $p_{jj'}$, to be a function of the measurement time spacing $d_{jj'} = |x_j - x_{j'}|$ where $x_j, x_{j'}$ are the measurement times associated with $e_{ij}, e_{ij'}$. Naturally, we model the probability $p_{jj'}$ through the following logit link,

$$\log(\frac{p_{jj'}}{1 - p_{jj'}}) = f_0(d_{jj'}), \tag{4.4.15}$$

where the $f_0(.)$ is a nonparametric function we approximate by a regression spline. Splines are piecewise polynomials that satisfy certain smoothness conditions between pieces. The space of the splines is determined by the order of the polynomials and the location of knots. Since we are estimating the function f_0 on the interval $[0, M]$, let $0 = s_0 < s_1 < \ldots < s_k = M$ be a partition of the interval. Using the s_i's as knots, we have $K = k + l$ normalized B-spline basis functions of order $l + 1$ that form a basis for the linear spline space. We write these basis functions into a vector $\pi(d) = (B_1(d), \ldots, B_K(d))^\top$. As our simulations reveal that the performance of the final statistic is not sensitive to the knot selection (not elaborated), we therefore use knots that are quantiles of the observed x_{ij}'s with uniform percentile ranks. We also use cubic splines with $l = 3$, but linear or quadratic splines can be considered if we think that f_0 is less smooth. Readers are referred to He et al. (2002) for issues on the

determination of order and knot selection. For more details about the construction of those basis functions, the readers are referred to He and Shi (1994). Let $f_0(d)$ be approximated by $\pi(d)^\top \nu$, where $\nu \in R^K$ is the spline coefficient vector. This parameterizes our logistic regression model (4.4.15) so that our regression problem becomes

$$\log(\frac{p_{jj'}}{1-p_{jj'}}) \doteq \pi(d_{jj})^\top \nu. \tag{4.4.16}$$

where $\doteq$ means both sides are of the same order when $k \to \infty$. To estimate ν, we construct observed binary outcomes as $I\{\widehat{e}_{ij} < 0, \widehat{e}_{ij'} < 0\}$ from each pair of residuals $(\widehat{e}_{ij}, \widehat{e}_{ij'})$, the associated log-likelihood function can be written as

$$\widehat{M}_n(\nu) = n^{-1} \sum_{i=1}^{n} \sum_{j<j'} -\log(1+\exp(\pi(t_{ij}-t_{ij'})^\top \nu)) + I_{\{\widehat{e}_{ij}<0, \widehat{e}_{ij'}<0\}} \pi(t_{ij}-t_{ij'})^\top \nu \tag{4.4.17}$$

The maximum likelihood estimate can be expressed by $\widehat{\nu} = \arg\max_\alpha \widehat{M}_n(\nu)$. Consequently, we plug in fitted $p_{jj'}$ defined by $\widehat{p}_{jj'} = \frac{1}{1+e^{-\pi_{jj'}^\top \widehat{\nu}}}$ into the equation (4.4.8) and obtain

$$\begin{aligned} V_n^{SP} = V_{2,n} = n^{-1} \sum_{ij} (g_t^*(\mathbf{w}_{ij}; \widehat{\boldsymbol{\eta}}, t_0))^2 \tau(1-\tau) \\ + n^{-1} \sum_{i} \sum_{j \neq j'} (g_t^*(\mathbf{w}_{ij}; \widehat{\boldsymbol{\eta}}, t_0))^2 (\widehat{p}_{jj'} - \tau^2). \end{aligned} \tag{4.4.18}$$

The corresponding rank score test for this design is defined as

$$T_n^{SP} = S_n^2 / V_n^{SP}. \tag{4.4.19}$$

4.4.4 Summary

We have discussed some variations of the T_n. There are still many other options which can be explored depending on the study design, e.g. AR-1, m-dependence, just to name a few. More can be found in Liang and Zeger (1986). In fact, the performances between these variations are similar as will be shown in simulations. As expected, under the correct specifications, the asymptotic distribution of the rank score test statistics all converge to χ_1^2 as the T_n does.

Chapter 5
Simulation studies

5.1 Overview

In this chapter, we will conduct a series of Monte Carlo simulations for the proposed rank score test. These simulations studies assess the Type I errors of the test statistic and compare the performance to the extensively used bootstrap method. All our computer simulations presented in this chapter have been implemented by `R` [R Development Core Team (2009)] version 2.10.0.

5.2 Model description

This section describes the simulated models we utilize for investigations of finite-sample performance of the rank score test. The generated data from the models mimic those we encountered in applications. To incorporate the longitudinal nature of the data, the error terms employ similar within-subject dependent structures as described in He et al. (2002) and Moyeed and Diggle (1994). We first draw data of 50 subjects with 10 measurements each using the following model,

$$y_{ij} = a + b_1(t_{ij} - t_\tau)_- + b_2(t_{ij} - t_\tau)_+ + b_3 x_i + \mathcal{G}(t_{ij}) + e_{ij}, i = 1, \ldots, m,\ j = 1, \ldots, n_i. \tag{5.2.1}$$

Here the t_{ij}s are independent random samples from uniform distribution on $(0, 1)$, the true change-point t_τ is set at 0.6 for all quantile levels $\tau \in (0, 1)$, x_i are random samples drawn from $Binomial(1, 0.5)$, and the other parameters are set at $(a, b_1, b_2, b_3) = (1, 2, -3, 1)$. In addition, we have two sources of error. e_{ij} is a random noise. And we also have a *subject-level error component* $\mathcal{G}(t_{ij})$, where $\mathcal{G}(\cdot)$ is a stationary Gaussian process with zero mean and autocovariance function $\gamma(u) = 0.4\exp(-r|u|)$ for some value of $r \in [0, 1]$. For the ith subject, $(\mathcal{G}(t_{i1}), \ldots, \mathcal{G}(t_{i,n_i}))$ simply follows a n_i-dimensional multivariate normal distribution with zero mean and variance-covariance matrix with (j, j')th element $\gamma(t_{ij}-t_{ij'}) = 0.4\exp(-r|t_{ij}-t_{ij'}|)$. This Gaussian process induces correlation between y_{ij} and $y_{ij'}$, which depends on the measurement spacing $|t_{ij} - t_{ij'}|$. For example, $\mathrm{Cov}(y_{i1}, y_{i2}) = \gamma(t_{i1} - t_{i2}) = 0.4\exp(-r|t_{i1} - t_{i2}|)$. Note that the value of r controls the strength of correlation between the measurements from the same subject. The extreme cases of $e^{-r} \approx 0$ or 1 correspond to no correlation at all or perfect correlation, respectively. This formulation of the error structure, used in [He et al. (2002)], is more flexible than traditional error correlation structures, e.g. Compound symmetry, AR-1, and m-dependent, and thus is useful for simulating longitudinal data where the within subject correlation is not constant.

The following three different cases are considered in this study.

Case 1 (Normal error model): The composite error term takes the form $\epsilon_{ij} = \mathcal{G}(t_{ij}) + e_{ij}$, where the subject-level error component is $\mathcal{G}(t_{ij})(i = 1, \ldots, n)$ just described, and the random noise e_{ij}'s are generated from $N(0, 0.1)$.

Case 2 (Normal mixture error model): The composite error term takes the form $\epsilon_{ij} = \mathcal{G}(t_{ij}) + e_{ij}$, where $\mathcal{G}(t_{ij})$ is the same as in Case 1, but the random noise e_{ij}'s are generated from the mixture distribution $.95 * N(0, 0.1) + .05 * N(0, 12.5)$. This is a small deviation from Case 1 because the e_{ij}s are generated from a mixture of two normal distributions so that 95% of the time the error is sampled from $N(0, 0.1)$ but the other 5% of the time it comes from $N(0, 12.5)$, which can

be interpreted as an "outlier distribution". The variance of y_{ij} is the sum of the variance of $\mathcal{G}(t_{ij})$ and the variance of e_{ij}. Therefore, without outliers, the variance of y_{ij} is $0.4 + 0.1 = 0.5$, and from the outlier distribution, the variance of y_{ij} is $0.4 + 12.5 = 12.9$. Even though the ratio of the two variances from e_{ij}, 12.5 versus 0.1, would appear to suggest a much larger relative magnitude, in fact, the standard deviation of the outlier distribution is only about 5 times that of the other distribution, which is a reasonable range for outliers.

Case 3 (Heteroscedastic error model): In this case, we consider a somewhat more realistic situation. The composite error term now takes the heteroscedastic form $\epsilon_{ij} = \mathcal{G}(t_{ij}) + (1/2 + t_{ij}/15)e_{ij}$, where $\mathcal{G}(t_{ij})$ and e_{ij} are the same as in Case 1. What differentiates this model from Case 1 is that the variance of outcome y_{ij} increases over time t_{ij}. This situation also exhibits more deviation than the homogeneous error model (5.2.1) considered in Case 1. For this model, the slope coefficients vary across the quantile levels, and more specifically, the slopes, left and right of the change-point t_τ, are $b_1 + Q_\tau(e_{ij})/15$ and $b_2 + Q_\tau(e_{ij})/15$ respectively. By not assuming homogeneous errors, this heteroscedastic model allows for more flexible modeling of real world scenarios. In deriving the limiting distribution of our test statistic T_n, one of the assumption we make is homogeneous error density at the τth quantile. This model violates the assumption of homogeneous error density, and thus is useful to evaluate whether the performance of our proposed test is sensitive to such deviations.

In Cases 1 and 2, we have $\mathrm{Var}(Y_{ij}) = 0.5$ and $\mathrm{Cov}(Y_{ij}, Y_{il}) = 0.4\exp(-r|t_{ij} - t_{il}|)$. In practice, exact information about the covariance matrix V of the repeated measures Y_{ij} within the ith subject is usually unknown. It is a substantial advantage for the rank score test that it does not require this as part of the input.

5.3 Type I errors

In this section, we investigate the Type I error rate of the rank score test using the models (5.2.1) previously mentioned. For these models, the τth conditional quantile function $a + Q_\tau(\epsilon_{ij}) + b_1(t_{ij} - t_\tau)_- + b_2(t_{ij} - t_\tau)_+ + b_3 x_i$ is linear in t_{ij} on both sides of the change-point t_τ, where $Q_\tau(\epsilon_{ij})$ is the τth quantile of the composite error $\epsilon_{ij} = \mathcal{G}(t_{ij}) + e_{ij}$. We test the null hypothesis $H_0 : t_\tau = 0.6$ versus $H_1 : t_\tau \neq 0.6$. The number of Monte Carlo samples used for estimating the Type I error rate is $10,000$ in each scenario, so that the standard error of the estimate is around 0.2%. The results are summarized in Table 5.1.

Table 5.1 summarizes the results in terms of Type I error rates averaged over $10,000$ simulations for the 3 cases, normal error, normal mixture error, and heteroscedastic error. For each case, we report the results for 5 different values of within-subject correlations e^{-r} ranging from 0.1 to 0.9. Our simulation results indicate that without any knowledge of the dependence structure in the data, our test exhibit decent finite-sample performance in terms of preserving the nominal level of type I error rate. This is helpful because, in practice, the true correlation structure of longitudinal data may vary from subject to subject so that it usually cannot be estimated in any reliable way.

5.4 Comparisons of confidence intervals

In the previous section, we have evaluated the Type I error rate of the rank score test under different settings. In this section, we introduce an important application of the proposed test, constructing of confidence intervals for the change-point parameter in model (3.2.1). In what follows, we describe the procedure of constructing confidence

intervals for the change-point t_τ of model (3.2.1), based on inverting the rank score test. Its performance is then compared to the extensively used resampling method.

5.4.1 Rank score test inversion

The confidence intervals are constructed based on inversion of the proposed rank score test, finding a set of null values that do not lead to rejection at the pre-specified level. This approach circumvents the somewhat complicated problem of estimating limiting nuisance parameters, i.e. the conditional error densities, and thus offers a reliable method of forming confidence intervals in non-iid error settings [Koenker (1996)].

The algorithm we implemented for rank score test inversion is described as follows.

Rank score test inversion:

(1): Estimate $\widehat{t}_\tau$ using the profile estimation (3.2.6);

(2): Define a fine grid in a neighborhood around $\widehat{t}_\tau$, $\mathcal{T} = \{\widehat{t}_\tau \pm \delta k : k = 1, 2, \ldots, K\}$ for some positive K and δ;

(3): Test $H_0 : t_\tau = \xi$ for every $\xi \in \mathcal{T}$ using the proposed test at significance level 0.05;

(4): Define $t_\tau^L = \min\{\xi \in \mathcal{T} : \xi \text{ not rejected by } T\}$, $t_\tau^U = \max\{\xi \in \mathcal{T} : \xi \text{ not rejected by } T\}$ as the lower and upper limit of a 95% confidence interval for t_τ.

5.4.2 Resampling methods

There is extensive literature on the resampling methods in quantile regression including the paired bootstrap method [Buchinsky (1994)], the subgradient bootstrap method (also known as "Parzen-Wei-Ying" method) [Parzen et al. (1994)]. More recent development includes the method of perturbing the minimand (also known as "Jin-Ying-Wei" method) [Jin et al. (2001)], and the Markov chain marginal bootstrap

method [He and Hu (2002)], just to name a few. More comprehensive review is provided in [Koenker (2005)]. The main advantage of the bootstrap methods over earlier methods, e.g. Koenker and Bassett (1982), and Welsh (1987), is that they circumvent the somewhat difficult issue of estimating nuisance parameters, that is the sparsity function.

In this section, we will compare our methods based on the rank score test to two of the extensively used bootstrap methods. We describe the two methods one by one. First is the Subject bootstrap. It is the direct extension of Paired bootstrap for longitudinal data and works as follows.

Subject bootstrap Subject level triples $\{(y_{ij}, t_{ij}, x_{ij}) : j = 1, \cdots, n_i\}$ are drawn at random from the original observations *with replacement*. Whenever a subject is drawn the associated measures are all included. For each resampled sample the estimator $\widehat{t}^*_\tau$ is computed. Then this procedure is repeated B times which yields bootstrap realizations whose percentiles lead to the the end-points of a confidence interval. The algorithm for generating percentile bootstrap confidence interval is outlined. (1): Obtain bootstrap samples of size B, by resampling subjects with replacement;
(2): Estimate the change-point $\widehat{t}^*_{\tau,b}$ from each bootstrap sample, $\widehat{t}^*_{\tau,1}, \widehat{t}^*_{\tau,2}, \ldots, \widehat{t}^*_{\tau,B}$;
(3): Construct a 95% confidence interval of t_τ by $\left[2\widehat{t}_\tau - Q_{0.975}(\widehat{t}^*_{\tau,b}), 2\widehat{t}_\tau - Q_{0.025}(\widehat{t}^*_{\tau,b})\right]$, where $Q_p(\widehat{t}^*_{\tau,b})$ indicates the pth quantile of the B bootstrap estimates $\widehat{t}^*_{\tau,b}$.

"Jin-Ying-Wei" (JYW) Bootstrap method – Perturbing the minimand Next, we describe another popularly used bootstrap, that is "Jin-Ying-Wei" bootstrap. This bootstrap method involves perturbing or weighting the terms in the objective function by a positive random variable. It has been shown to perform well even in small sample size. We outline the algorithm as follows.

(1): Based on the original objective function

$$Q_n(\boldsymbol{\theta}_\tau) = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\rho_\tau\left(y - g(\tilde{x}_{ij};\boldsymbol{\theta}_\tau)\right), \tag{5.4.2}$$

a perturbed version is constructed by reweighting the terms in equation (5.4.2) associated with each subject using $Gamma(1,1)$ random variable v_i,

$$Q_n^*(\boldsymbol{\theta}_\tau) = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}v_i\rho_\tau\left(y - g(\tilde{x}_{ij};\boldsymbol{\theta}_\tau)\right), \tag{5.4.3}$$

(2): By minimizing $Q_n^*(\boldsymbol{\theta})$, we obtain bootstrap estimate $\widehat{\boldsymbol{\theta}}^*_{n,\tau}$.
(3): Calculate sample variance of the bootstrap estimate $\{\widehat{\boldsymbol{\theta}}^*_{n,\tau}\}$, and then construct a 95% confidence interval using normal approximation.

5.4.3 Numerical comparison of methods

Now we focus on the aforementioned two available approaches to inference for the change-point in longitudinal bent line quantile regression. Comparisons are conducted showing the estimated confidence accuracy (or coverage probability), interval length, and user CPU time required for computing each confidence interval. Comparisons have been evaluated under normal, normal mixture and heteroscedastic error models, respectively.

We consider generating 200 Monte Carlo samples and calculating the 95% C.I. by inverting the rank score test and the bootstrap. For the bootstrap method, the bootstrap sample size B is also set to 200. The standard error for the estimated coverage is 1.5%. Tables 5.2-5.4 summarizes the estimated coverage of C.I., median interval length, and CPU-time consumed for each C.I. based on 200 Monte Carlo samples for each case, at different quantile levels as well as within-subject correlation strengths.

In general, both the rank score inversion and JWY bootstrap preserves nominal 95% confidence level while the subject bootstrap consistently under-performed. The estimated coverage of JYW bootstrap CI is the highest but the computational expense is also the highest among the three methods. Under this relative small sample size setting, 50 subjects with correlated measures, our QQ-plots (not shown in this thesis) indicate the JWY bootstrap perform well by better approximating the distribution of $\widehat{t}_{n,\tau} - t_\tau$ using $\widehat{t}^*_{n,\tau} - \widehat{t}_{n,\tau}$ while the approximate provided by subject bootstrap exhibits considerable discrepancy. In terms of the computational expense, the rank score test inversion obviously enjoys a major advantage since its user CPU time consumed for generating each confidence interval is less than a fraction of those by the bootstrap methods, e.g. approximately 2 vs. > 38 seconds (for a relative small size of 200 bootstrap replicates). As discussed previously, the main reason is that the bootstrap involves repeated profile estimation (3.2.3) which is especially time-consuming.

Table 5.4 summarizes the performance using heteroscedastic error models in Case 3. The estimated confidence levels are not seriously affected compared to Tables 5.2 and 5.3. This again confirms that the rank score test can be robust to the deviation of the homogeneous error assumption (A1*). In conclusion, generating confidence interval for the change-point t_τ by inverting rank score tests works well and is also computationally efficient.

5.5 Rank score test statistic T_n vs. correlation structure specific T_n^{CS}

In the previous simulations, we have seen that the rank score test statistic T_n works well without any specification of the within-subject dependence structure. As discussed in section (4.4), the modified version of the rank score test statistic T_n^{cs}

is tailored to account for prior knowledge of the C.S. correlation structure. In this section, we summarize the simulation studies conducted to compare the performance of the two rank score test statistics on data with true C.S. correlation structure.

5.5.1 Additional model description

The simulation study is based on the following models which generate C.S. correlation structure. We keep a similar set-up as in (5.2) but change the correlation structure of the error term, modifying $\mathcal{G}_i(t_{ij}) + e_{ij}$. The model is

$$y_{ij} = a + b_1(t_{ij} - t_\tau)_- + b_2(t_{ij} - t_\tau)_+ + b_3 x_i + \epsilon_{ij}, \tag{5.5.4}$$

where $i = 1, \ldots, n, j = 1, \ldots, m_i$, and the composite error term $\alpha_i + e_{ij}$ is defined in two parts, a random subject-level intercept α_i and a random noise e_{ij}. The following three different cases are considered in this study.

Case 4 (Normal error model, C.S.): The composite error term takes a C.S. form from $\epsilon_{ij} = \alpha_i + e_{ij}$, where the subject-level intercept α_i's are generated randomly from $N(0, 0.4)$, and the random noise e_{ij}'s are generated from $N(0, 0.1)$. This is the typical model where the C.S. or Exchangeable correlation structure arises. $\text{Cov}(\epsilon_{ij}, \epsilon_{ij'}) = 0.4$ for all $j \neq j'$.

Case 5 (Normal mixture error model, C.S.): The composite error term takes the same form as in Case 4, where the subject-level intercept α_i's are generated randomly from $N(0, 0.4)$, and the random noise e_{ij}'s are generated from the mixture distribution $0.95 * N(0, 0.1) + 0.05 * N(0, 12.5)$. This is different from Case 1 because it allows a potential outlier distribution.

Case 6 (Heteroscedastic error model, C.S. analogue): The composite error term takes a heteroscedastic form $\epsilon_{ij} = \alpha_i + (1/2 + t_{ij}/15)e_{ij}$, where the subject-level intercept

α_i's are generated randomly from $N(0, 0.4)$, and the random noise e_{ij}'s are generated from $N(0, 0.1)$.

In each case, we generate 200 datasets, each consisting of $m = 50$ subjects with $n_i = 10$ measurements. For these rank score tests, we evaluate the performance at $\tau = 0.5$ and $\tau = 0.9$.

5.5.2 Comparisons

Table 5.5 summarizes the results in terms of estimated coverage, median interval length, and CPU time based on 200 Monte Carlo samples. The rank score C.I. based on the rank score test statistic T_n better preserves the type I error rate than the one based on the T_n^{CS} does, which is exclusively tailored for data with C.S. correlation structure. Let "Rankscore" denotes the rank score test inversion procedure based on the rank score test statistic T_n defined in Chapter 4; let "Rankscore (CS)" denotes the rank score inversion procedure based on the correlation specific statistic T_n^{CS} defined in 4.4.1. When the true correlation is correctly specified (Cases 4 and 5), the specifically tailored version T_n^{CS} generates slightly shorter confidence intervals. Nevertheless the difference in coverage between the two methods is minimal. On the other hand, when the true correlation is mis-specified (Case 6), the rank score test statistic T_n, which makes no assumption about the correlation structure, outperforms its C.S. specific counterpart. This shows that the performance of the specifically tailored statistic is sensitive to correct specification of the true correlation structure. Hence caution should be taken when knowledge of the correlation structure is unclear. In conclusion, the rank score test statistic T_n is recommended for general application because misspecification of the correlation structure would undermine the performance.

Table 5.1: The average Type I error rate of T_n with $m = 50$ under normal, normal mixture and heteroscedastic errors in $10,000$ simulations

corr.†	normal error		normal mixture error		heteroscedastic error	
e^{-r}	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.5$	$\tau = 0.9$
0.1	0.0538	0.0485	0.0512	0.0503	0.0532	0.0491
0.3	0.0532	0.0458	0.0504	0.0442	0.0541	0.0466
0.5	0.0524	0.0457	0.0491	0.0460	0.0489	0.0435
0.7	0.0548	0.0457	0.0518	0.0449	0.0527	0.0422
0.9	0.0519	0.0439	0.0488	0.0451	0.0530	0.0436

†: e^{-r} controls the strength of the inter-subject correlation.
The nominal Type I error rate is 0.05.
See (5.2) for model specifications.

Table 5.2: Performance of 95% confidence intervals for t_τ under *normal error*

		Coverage			Interval Length			User CPU Time (sec.)		
	e^{-r}†	Rankscore	Bjyw	Bsub	Rankscore	Bjyw	Bsub	Rankscore	Bjyw	Bsub
$\tau = 0.5$	0.1	0.960	0.985	0.915	0.211	0.211	0.198	2.83	89.56	50.18
	0.3	0.960	0.985	0.870	0.188	0.189	0.173	2.83	89.34	50.41
	0.5	0.970	0.990	0.860	0.191	0.209	0.173	2.76	89.39	50.31
	0.7	0.935	0.985	0.845	0.215	0.204	0.181	2.78	89.35	50.05
	0.9	0.945	0.960	0.845	0.187	0.200	0.175	2.49	89.06	46.51
$\tau = 0.9$	0.1	0.950	0.960	0.820	0.313	0.320	0.264	2.22	71.43	38.12
	0.3	0.930	0.980	0.815	0.284	0.294	0.243	2.22	71.50	38.14
	0.5	0.970	0.975	0.815	0.270	0.278	0.226	2.22	71.65	38.26
	0.7	0.975	0.960	0.820	0.296	0.282	0.240	2.23	71.73	38.21
	0.9	0.950	0.985	0.810	0.272	0.282	0.215	2.21	71.66	38.21

Normal error model : $y_{ij} = a + b_1(t_{ij} - t_\tau) + b_2(t_{ij} - t_\tau) + b_3 x_i + \mathcal{G}(t_{ij}) + e_{ij}, e_{ij} \sim N(0, 0.1)$

†: e^{-r} controls the strength of the inter-subject correlation.

Rankscore: Rank score test inversion.

Bjyw: "Jin-Ying-Wei" bootstrap method.

Bsub: Subject bootstrap.

Table 5.3: Performance of 95% confidence intervals for t_τ under *normal mixture error*

		Coverage			Interval Length			User CPU Time (sec.)		
	e^{-r}†	Rankscore	Bjyw	Bsub	Rankscore	Bjyw	Bsub	Rankscore	Bjyw	Bsub
$\tau = 0.5$	0.1	0.920	0.985	0.855	0.210	0.220	0.201	1.76	65.11	34.75
	0.3	0.945	0.975	0.865	0.207	0.217	0.194	1.79	64.84	34.61
	0.5	0.960	0.980	0.885	0.210	0.211	0.179	1.72	64.87	34.60
	0.7	0.970	0.990	0.835	0.213	0.214	0.176	1.77	64.75	34.57
	0.9	0.955	0.990	0.850	0.203	0.197	0.180	1.76	64.72	34.58
$\tau = 0.9$	0.1	0.930	0.980	0.810	0.329	0.329	0.291	3.09	88.77	49.38
	0.3	0.945	0.950	0.765	0.360	0.360	0.309	3.11	88.68	49.51
	0.5	0.935	0.970	0.810	0.342	0.342	0.290	3.02	88.78	49.24
	0.7	0.970	0.980	0.815	0.372	0.309	0.273	3.00	88.68	49.39
	0.9	0.960	0.975	0.765	0.355	0.333	0.287	2.76	82.90	46.03

Normal mixture error model : $y_{ij} = a + b_1(t_{ij} - t_\tau) + b_2(t_{ij} - t_\tau) + b_3 x_i + \mathcal{G}(t_{ij}) + e_{ij}, e_{ij} \sim 0.95 * N(0, 0.1) + 0.05 * N(0, 12.5)$

†: e^{-r} controls the strength of inter-subject correlation.

Rankscore: Rank score test inversion.

Bjyw: "Jin-Ying-Wei" bootstrap method.

Bsub: Subject bootstrap.

Table 5.4: Performance of 95% confidence intervals for t_τ under *heteroscedastic error*

		Coverage			Interval Length			User CPU Time (sec.)		
	e^{-r}†	Rankscore	Bjyw	Bsub	Rankscore	Bjyw	Bsub	Rankscore	Bjyw	Bsub
$\tau = 0.5$	0.1	0.950	0.970	0.905	0.201	0.203	0.177	2.79	89.52	50.63
	0.3	0.935	0.980	0.885	0.188	0.188	0.166	2.77	89.92	50.57
	0.5	0.955	0.985	0.890	0.174	0.174	0.160	2.78	89.68	50.64
	0.7	0.960	0.975	0.870	0.200	0.200	0.164	2.73	89.46	50.19
	0.9	0.920	0.970	0.895	0.170	0.170	0.152	2.38	82.99	46.54
$\tau = 0.9$	0.1	0.940	0.965	0.785	0.306	0.302	0.254	2.91	89.22	49.72
	0.3	0.955	0.965	0.810	0.289	0.276	0.236	2.85	89.12	49.80
	0.5	0.955	0.985	0.825	0.260	0.269	0.214	2.77	89.39	49.29
	0.7	0.960	0.970	0.840	0.247	0.253	0.217	2.82	89.21	49.61
	0.9	0.935	0.965	0.825	0.243	0.242	0.192	2.55	83.25	45.89

Heteroscedastic error model : $y_{ij} = a + b_1(t_{ij} - t_\tau) + b_2(t_{ij} - t_\tau) + b_3 x_i + \mathcal{G}(t_{ij}) + (1/2 + t_{ij}/15)e_{ij}, e_{ij} \sim N(0, 0.1)$

†: e^{-r} controls the strength of the inter-subject correlation.

Rankscore: Rank score test inversion.

Bjyw: "Jin-Ying-Wei" bootstrap method.

Bsub: Subject bootstrap.

Table 5.5: Performance of 95% confidence intervals for t_τ under the *compound-symmetry* error structures

		Coverage		Interval Length		User CPU Time (sec.)	
	Models	Rankscore	Rankscore (CS)	Rankscore	Rankscore (CS)	Rankscore	Rankscore (CS)
$\tau = 0.5$	Case 4*	0.955	0.945	0.183	0.175	2.08	2.34
	Case 5†	0.960	0.950	0.229	0.180	2.17	2.39
	Case 6#	0.950	0.935	0.175	0.156	1.97	2.24
$\tau = 0.9$	Case 4	0.950	0.935	0.266	0.332	2.17	2.69
	Case 5	0.965	0.940	0.366	0.351	2.16	2.47
	Case 6	0.945	0.930	0.241	0.254	1.99	2.40

$*$: Case 4: Normal error model with C.S. structure;
†: Case 5: Normal mixture error model with C.S. structure;
#: Case 6: Heteroscedastic error model with C.S. analogue structure;
Rankscore: Rank score test inversion based on T_n;
Rankscore (CS): Rank score test inversion based on C.S. structured specified version T_n^{CS};
The confidence level is 0.95.

Chapter 6
Applications

6.1 Overview

In this chapter, we apply the longitudinal bent line model with the proposed rank score test to conduct statistical inference on the change-point parameter of interest in two different contexts, the Finnish Longitudinal Growth Study and the AIDS clinical study. In the first section, we characterize the growth pattern in Body Mass Index (BMI) for the children enrolled in the Finnish Longitudinal Growth Study. The adiposity rebound (AR), a crucial time associated with risk of higher BMI in adolescence and risk of adult obesity, is estimated for the lean, moderate, heavy and heavist subgroups, represented by the 0.1, 0.5, 0.9 and 0.95 quantile functions. From this longitudinal bent line model, we calculate the fitted BMI at AR, which is another known risk factor for adult obesity. This value is represented by the estimate of the regression parameter a_τ. In the second section, we apply the methods to an AIDS clinical study. We characterize the treatment response in CD4 cell counts, an important marker of the immunologic response, among HIV-infected population treated with the antiretroviral therapy (ART). After that, the time when CD4 cell count reaches a plateau as well as the associated stabilized CD4 cell count levels are obtained for the mildly ill and severely ill patients, represented by the 0.5 and 0.1 quantile functions. These examples demonstrate the value of the quantile inference approach for the change-point problems, especially for longitudinal data from

observational studies as well as from clinical trials.

6.2 Application to the Finnish Longitudinal Growth Study

In pediatrics and nutritional epidemiology, a time period termed adiposity rebound (AR) in childhood is a critical period for the regulation of energy balance and adult obesity risk and thus generates extensive research interest [Rolland-Cachera et al. (1984); Siervogel et al. (1991); Prokopec and Bellisle (1993); Reilly et al. (2005)]. AR usually occurs between 5 to 7 years of age, at which point body fatness has normally declined to a minimum, before increasing again into adulthood. An early AR (younger age at the point of AR) is associated with not only higher BMI in adolescence [Rolland-Cachera et al. (1984); Prokopec and Bellisle (1993)] but also an increased risk of adult obesity. After adjusting for BMI at AR, maternal BMI, paternal BMI - three other known risk factors for adulty obesity, early AR is still asssociated with risk for adult obesity [Whitaker et al. (1998)].

To illustrate the usage of the proposed methods, we apply them to the dataset from the Finnish Longitudinal Grow Study [Sorva et al. (1990)]. The data was collected retrospectively from health centers and schools. The observations consist of longitudinal measurements on weight and height for 2514 Finnish children. Weight (kg) and height (m) were measured for each participant using standardized techniques, and BMI (kg/m^2) was calculated from weight and height. As described in more detail in Pere (2000), the dataset has been cleaned to remove a small proportion of children with low or missing birth weight or other suspicious measurements. For our purposes of characterizing AR, we further edit the dataset to include measurements taken from 1 to 18 years of age. The resulting working dataset consists of 1140 males and 1162

females. This datset does not contain any missing observations in response variable or covaraites. In this analysis, we focus on making statistical inference on the time of AR at different conditional quantile levels of BMI, respectively.

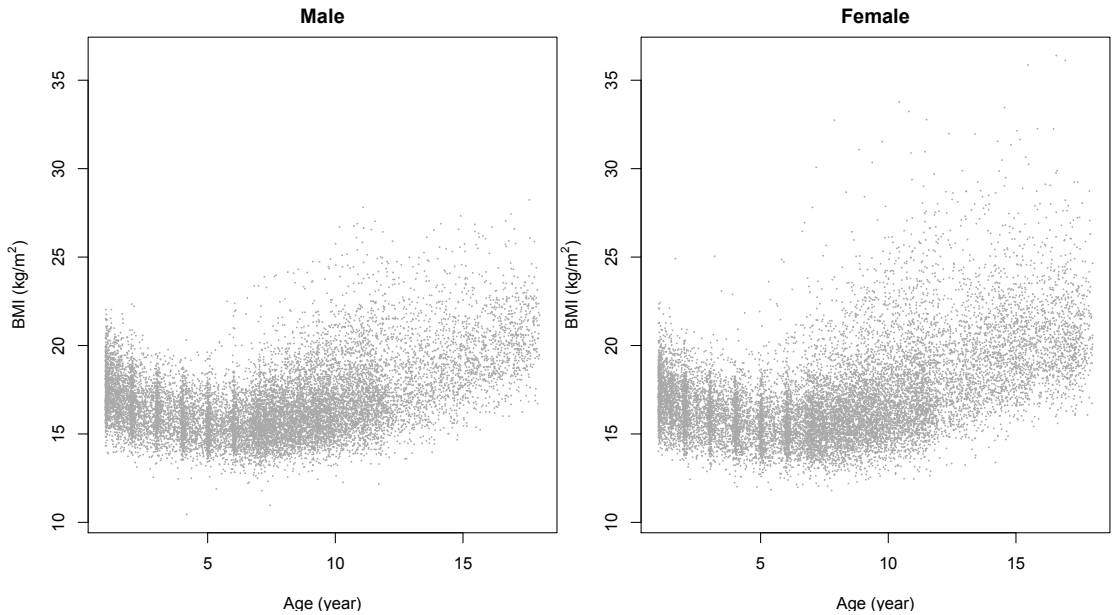

Figure 6.1: Body Mass Index (BMI) plotted against Age in years for 1140 males and 1162 females.

For any fixed quantile $\tau \in (0, 1)$, we model the Body Mass Index (BMI) for the male and female subgroups respectively by

$$\text{BMI}_{ij} = a_\tau + b_{1,\tau}(t_{ij} - t_\tau)_- + b_{2,\tau}(t_{ij} - t_\tau)_+ + e_{ij}, \tag{6.2.1}$$

where $i = 1, \ldots, n, j = 1, \ldots, m_i$, BMI_{ij} is the Body Mass Index from the ith subject at the jth measurement calculated by BMI = Weight/Height^2 and t_{ij} is the measurement time. All the parameters have clear interpretations: t_τ is the time of AR; a_τ is the BMI at AR; $b_{1,\tau}$ and $b_{2,\tau}$ are the slopes of change in BMI before and after AR.

For the purposes of this analysis, we have specified four quantiles of interest ($\tau = 0.1, 0.5, 0.9, 0.95$), which we define as representing the lean, moderate, heavy, and

heaviest subgroups. Before we turn to the model fitting, one may question whether the dataset support the claim that a time of AR exists in these quantile functions. This question can be formulated as testing the adequacy of a simple linear quantile regression model, i.e. one without a change-point, versus the bent line quantile regression model. As discussed in Li et al. (2011), a general test for nonlinearity (bent line in our case) for iid error model is assessing the significance of the quadratic term of the covariate associated with the change-point, using a Wald test. The reason to test for a quadratic term is not that one believes the quadratic model fits the data but that it is an easy test against a wide range of nonlinear alternatives. Due to the longitudinal nature of our datset, the Wald test statistic is not readily available. To overcome this issue, we calculate a bootstrap confidence interval for the coefficient associated with the quadratic term meantioned earlier. After that a decision of the existence of the change-point can be made. To be specific, we first calculate a 95% confidence interval of the coefficient for t_{ij}^2 by resampling the subjects with replacement. Then claim the existence of the change-point t_τ in the τth quantile function only if the bootstrap interval does not include zero. The resulting 95% confidence intervals are summarized in Table 6.1. Since none of them include zero, we claim the change-points exists in the four quantile functions. This result reconfirms the findings of AR in the pediatric literature [Rolland-Cachera et al. (1984); Siervogel et al. (1991); Prokopec and Bellisle (1993)].

Table 6.1: The 95% bootstrap confidence intervals (C.I.) of the coefficient associated with t_{ij}^2, the quadratic term of Age, in a linear quantile regression model

Quantile levels	95% C.I. for the coefficient associated with ${t_{ij}}^{2}$ †	
τ	Male	Female
0.1	(0.039, 0.043)	(0.037, 0.041)
0.5	(0.044, 0.048)	(0.043, 0.048)
0.9	(0.048, 0.062)	(0.039, 0.055)
0.95	(0.050, 0.070)	(0.032, 0.052)

†: We test the existence of the change-point by assessing the adequacy of a simple linear quantile regression model. A general test is testing the significance of Age t_{ij}^2 in a quadratic model. We claim the change-point exists in the τth quantile function if the confidence interval does not include zero.

After confirming the existence of the AR in the four quantile functions, we apply model (6.2.1) to the male and female subgroups respectively. Table 6.2 summarizes the estimated slopes in BMI change, the time of AR, and BMI at AR for the 4 quantiles. For example, the lean males and females (0.1 quantile) experienced relatively late AR; their estimated AR are 6.58 and 7.45 with 95% confidence intervals $(6.31, 6.93)$ and $(6.68, 7.82)$, respectively. For the males and females with moderate BMI (0.5 quantile), their estimated AR are 6.32 and 6.41 with 95% confidence intervals $(6.12, 6.47)$ and $(6.16, 6.56)$ respectively. The AR for the heavy males and females (0.9 quantile) are obtained as 5.47 and 5.11 with 95% confidence intervals $(5.16, 5.57)$ and $(4.78, 5.50)$ respectively. For the heaviest males and heaviest females (0.95 quantile) the estimated time of AR are 5.33 and 4.47 with 95% confidence intervals $(4.67, 5.57)$ and $(4.03, 4.83)$, respectively. The later confidence interval shed light

on the timing of adiposity rebound. From the lower limit, e.g. 4.03, one implication is that the public health institute needs to monitor the onset of adiposity rebound as early as 4 years of age for female subjects. The upper limit, e.g. 4.83, is consistent to Whitaker et al. (1998)'s definition of early adiposity rebound, that is earlier than 4.8 years of age. Therefore we identify a region in extreme higher quantile ($\tau \geq 0.95$) associated with early adiposity rebound for the female cohort. Therefore it is worthwhile to separate the subjects in the upper quantile for further investigation if more information becomes available. An extreme measurement is likely to be a sign of some unusual underlying physical condition.

Table 6.2: Fitted parameters and 95% confidence intervals (C.I.) for adiposity rebound t_τ

Cohort	Quantile†	$b_{1,\tau}$	$b_{2,\tau}$	t_τ	95% CI of t_τ	BMI at AR††
Male n=1140	$\tau = 0.1$	-0.33	0.34	6.58	(6.31, 6.93)	13.74
	$\tau = 0.5$	-0.40	0.44	6.32	(6.12, 6.47)	15.06
	$\tau = 0.9$	-0.56	0.57	5.47	(5.16, 5.57)	16.82
	$\tau = 0.95$	-0.58	0.65	5.33	(4.67, 5.57)	17.37
Female n=1162	$\tau = 0.1$	-0.23	0.45	7.45	(6.68, 7.82)	13.60
	$\tau = 0.5$	-0.36	0.51	6.41	(6.16, 6.56)	15.00
	$\tau = 0.9$	-0.49	0.65	5.11	(4.78, 5.50)	16.99
	$\tau = 0.95$	-0.60	0.74	**4.47**	(4.03, 4.83)	17.67

†: 0.1 quantile = 10th percentile, etc.;

$b_{1,\tau}$: The slope of BMI change before adiposity rebound;

$b_{2,\tau}$: The slope of BMI change after adiposity rebound;

t_τ: The time of adiposity rebound (AR); Early AR (< 4.8 years of age) is shown in bold.

††: a_τ indicate the BMI at AR.

For purposes of comparison, we fit the data using a longitudinal version of the east squares bent line regression [Chappell (1989)].

$$\text{BMI}_{ij} = a + b_1(t_{ij} - t)_- + b_2(t_{ij} - t)_+ + u_{ij}, \tag{6.2.2}$$

where $i = 1, \ldots, n, j = 1, \ldots, m_i$, BMI_{ij}, and t_{ij} are the same as in model (6.2.1).Error term denoted by u_{ij} has zero mean. Parameters b_1 and b_2 are the slopes of BMI change before and after AR. By minimizing the likelihood via profile argument, the estimated parameters are summarized in Table (6.3).

Table 6.3: Fitted parameters and 95% confidence intervals (C.I.) for adiposity rebound t_τ using least squire bent line regression

Cohort	b_1	b_2	t	95% CI of t	BMI at AR††
Male	-0.47	0.44	5.82	(5.50, 6.15)	15.14
Female	-0.39	0.52	5.88	(5.52, 6.23)	15.16

b_1: The slope of BMI change before adiposity rebound;

b_2: The slope of BMI change after adiposity rebound;

t: The time of adiposity rebound (AR);

a: The BMI at AR.

The estimated timings of AR from Model (6.2.2) for male and female cohorts are similar, i.e. 5.82 for male and 5.88 for female. Due to the longitudinal nature of the dataset, 95% confidence intervals of t are obtained via subject bootstrap procedure outlined in Chapter 5.

Unlike quantile regression, the parameters obtained from least squares regression are interpreted based on average, and thus do not differentiate based on conditional quantiles of BMI, e.g. between the relative lean subjects and relatively heavy. since

Model 6.2.2 assumes a common change-point t. Therefore Model 6.2.2 could not capture any early AR.

This example illustrates that the longitudinal bent line quantile regression sheds light on the BMI growth patterns by examining the AR at various quantiles. The resulting important yet straightforward parameters convey key information with clinical meaning at quantiles of interest, which least squares based regressions do not provide. In summary, the proposed method is helpful to characterize the BMI growth patterns, especially the time of AR. Hence it is a valuable tool for pediatricians in understanding children's growth patterns associated with adolescent and adult obesity.

6.3 Application to an AIDS clinical study

In this section, we illustrate the use of the longitudinal bent line model using the proposed Rank score test by applying it to a HIV (human immunodeficiency virus) dataset collected by the AIDS clinical trials group (ACTG) mentioned earlier. In this study, 517 HIV-positive patients were enrolled and randomly assigned to three treatments for 120 weeks. The CD4 cell count, an important marker for assessing immunologic response, were measured at the 4th week, 8th week, and every 8 weeks thereafter, which constitute a typical longitudinal dataset. Greater detail about this dataset can be found in Park and Wu (2006). This dataset does not contain any missing observations in response variable or covariates. The mechanism of how CD4 cell count change has not been fully explained but CD4 cell count has important implications for understanding the incidence of HIV-related opportunistic infections, especially tuberculosis [Williams et al. (2006)]. We illustrate the proposed method by applying the methodologies developed in previous chapters to model CD4 cell response in one of the three treatment arms. In this arm, 171 patients were treated

with an potent combination antiretroviral therapy (ART) which has been shown to dramatically extend the time to development of AIDS [Detels et al. (1998); Sterne et al. (2005)].

One important fact about the ART is that regardless of baseline CD4 cell count at ART initiation, a steeper increase occurs in CD4 cell count during the first 2 years after ART initiation, followed by a subsequent *plateau* indicating the stabilization period thereafter [Tarwater et al. (2004)]. This suggests the existence of a change-point in treatment efficacy at a certain time, when CD4 cell count reach the plateau. Our objective for the analysis is to characterize the population response in terms of CD4 cell gain during the ART treatment and conduct statistical inference on the change-point, the time CD4 reaches the plateau. For computational stability, we stabilize the large variance in CD4 cell counts by applying $\log 10$ transformation as in Park and Wu (2006). Specifically, the outcome is $\log \mathrm{CD4} = \log_{10}(\mathrm{CD4}/100)$ for model fitting. For any fixed quantile $\tau \in (0, 1)$, we model the transformed CD4 response by

$$\log \mathrm{CD4}_{ij} = a_\tau + b_{1,\tau}(t_{ij} - t_\tau)_- + b_{2,\tau}(t_{ij} - t_\tau)_+ + x_i\gamma_\tau + e_{ij}, \tag{6.3.3}$$

where $i = 1, \ldots, n, j = 1, \ldots, m_i$, $\log \mathrm{CD4}_{ij}$ is the $\log 10$ transformed CD4 cell response from the ith subject at the jth measurement, x_i is the baseline CD4 cell count dichotomized at its median, 96 cells/μL. The parameters have clear clinical interpretations: a_τ represents the stabilized response at plateau; $b_{1,\tau}$ and $b_{2,\tau}$ represent the slopes in response before and after it reaches the plateau.

For the purpose of characterizing the pattern of CD4 cell count in this analysis, we specify three quantiles, 0.1, 0.5, and 0.9, which we define as representing the severely ill, moderately ill, and mildly ill subgroups respectively. Before we turn to model fitting, we assess the existence of change-points for these quantiles by the methods described and used in the previous section. The 95% confidence intervals

of the coefficient associated with t_{ij}^2, for 0.1, 0.5, and 0.9 quantiles are calculated as $(-1.8\times10^{-4}, -4.7\times10^{-5})$, $(-8.9\times10^{-5}, -3.2\times10^{-5})$, and $(-7.6\times10^{-5}, 7.5\times10^{-6})$. As one can see, the third interval includes zero, indicating the existence of change-points only in the 0.1, and 0.5 quantile functions. Therefore we include the two quantiles for model fitting.

Table 6.4: The 95% bootstrap confidence intervals (C.I.) for $b_{1,\tau}$ and $b_{2,\tau}$ from the initial fitting

Quantile levels	95% C.I.† for $b_{1,\tau}$ and $b_{2,\tau}$	
τ	$b_{1,\tau}$	$b_{2,\tau}$
0.1	(0.009, 0.032)	(-0.002, 0.005)
0.5	(0.008, 0.014)	(-0.015, 0.006)

†: The number of bootstrap replications is set as 200.

After we apply model (6.3.3) to the dataset, the initial fitting of the longitudinal bent line model reveals that the estimated $b_{2,\tau}$'s, the slopes after the stabilization, are close to zero. Furthermore, we generate 95% bootstrap confidence intervals for $b_{1,\tau}$ and $b_{2,\tau}$, the slopes before and after stabilization for the 0.1 and 0.5 quantile functions. It turns out that only the $b_{1,\tau}$'s are significantly non-zero (See Table 6.4). For considerations of efficiency and easier interpretation, we fit the model again under the constraint that $b_{2,\tau} = 0$, for $\tau = 0.1$ and 0.5.

Figure 6.2 illustrates the fitted 0.1 and 0.5 quantile functions for only those patients whose baseline CD4 cell count are less than 96 cells/μL. The vertical dashed blue and red lines indicate the locations of the estimated change-points for the corresponding quantile functions. One can see an increasing trend in CD4 response after treatment initiation followed by a stabilization for each quantile function.

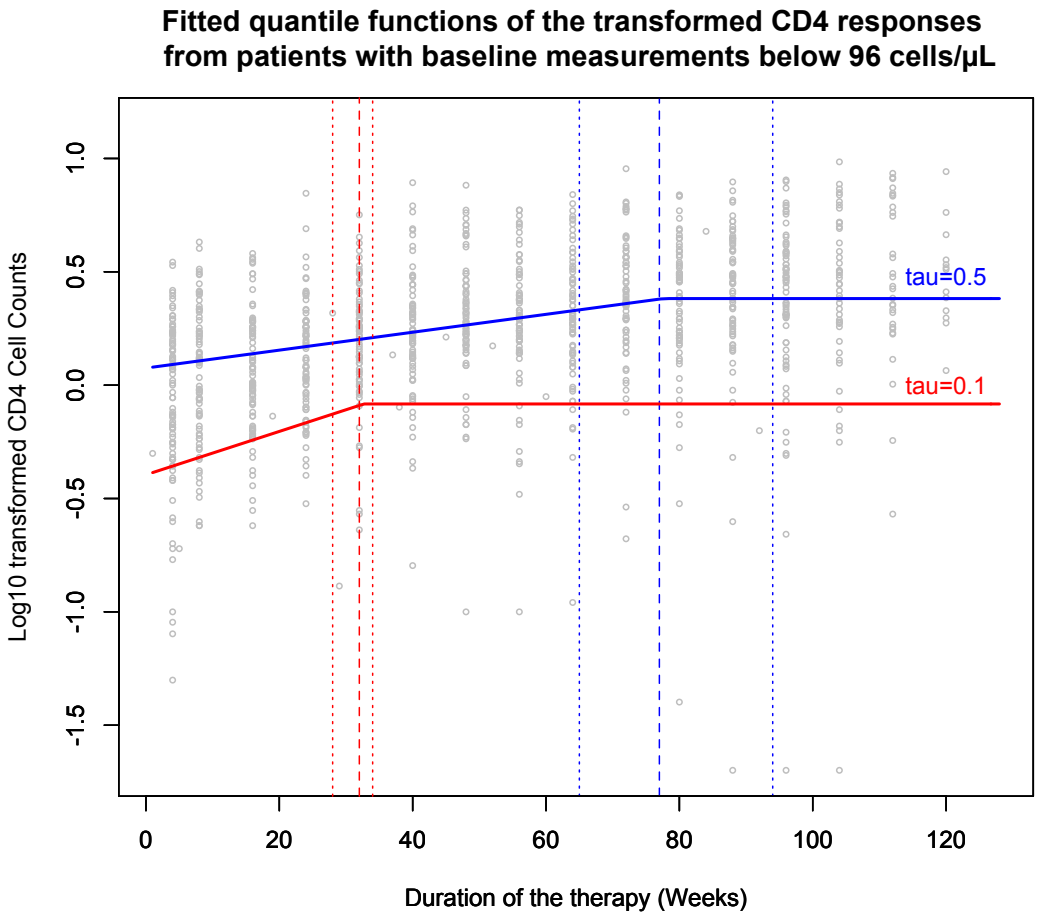

Figure 6.2: Solid bent lines depict estimated quantiles (e.g. 0.1 = 10% percentile) of the transformed CD4 response as a function of the therapy duration. The vertical dashed lines indicate the locations of estimated change-points at the 0.5 and 0.9 quantiles with 95% confidence intervals shown as vertical dotted lines respectively. Covariates included in the model are duration of the therapy and baseline CD4 cell count dichotomized at 96 cells/μL (median). To illustrate the different locations of the change-points, only those patients with baseline CD4 cell count less than 96 cells/μL and their estimated CD4 response quantiles are plotted.

Table 6.5: Fitted parameters and 95% confidence intervals for change-point t_τ

Quantile †	a_τ	$b_{1,\tau}$	γ_τ	t_τ	95% CI of t_τ	Plateau level#
$\tau = 0.1$	0.307	0.010	-0.390	32	(28,33)	203
$\tau = 0.5$	0.700	0.004	-0.314	77	(65,94)	501

†: 0.1 quantile = 10th percentile, etc.;

a_τ: The stabilized CD4 cell count on log10 scale;

$b_{1,\tau}$: The slope of transformed CD4 change before adiposity rebound;

$b_{2,\tau}$: The slope of transformed CD4 change after adiposity rebound;

γ_τ: The difference in log CD4 cell count, comparing those with baseline CD4 cell count less 96 cells/μL to those greater than 96;

t_τ: The week when CD4 cell count reaches the plateau; rounded to the nearest integer.

#: The stabilized CD4 cell count on the original scale, calculated by $100 \times 10^{a_\tau}$.

Furthermore the proposed rank score test enables us to calculate 95% confidence intervals of the change-points for the 0.1 and 0.5 quantile functions using the rank inversion procedure described in the previous chapter. Table 6.5 summarizes the fitted parameters: a_τ is the stabilized CD4 cell count on the $\log 10$ transformed scale; $b_{1,\tau}$ and $b_{2,\tau}$ are the slopes before and after the plateau occurs; γ_τ represents the difference in $\log 10$ transformed CD4 cell count between the patients whose baseline CD4 cell counts are below and above 96 cellsμL; t_τ is the time when the plateau occurs. The 95% confidence intervals for the change-points are $[28, 33]$ and $[65, 94]$ for the 0.1 and 0.5 quantile functions respectively. The fitted 0.1 quantile function reveals that the lower quantile ($\tau = 0.1$) of the transformed CD4 cell count reaches the plateau

of $a_{0.1} + \gamma_{0.1} = -0.083$ on the log scale, which is $100 \times 10^{-0.083} = 83$ cells/μL on the original scale, after the 32nd week. This indicates that approximately 90% of those patients, whose baseline CD4 cell count were less than 96, have their CD4 cell counts stabilized above 83 cells/μL after the 32nd week from the initiation of therapy. The 95% confidence interval for this change-point is $[28, 33]$. Similarly, the fitted 0.5 quantile function shows that the median ($\tau = 0.5$) transformed CD4 cell count reaches a plateau of $a_{0.5} + \gamma_{0.5} = 0.386$ on the log scale, which is $100 \times 10^{0.386} = 243$ cells/μL on the original scale, after the 78th week. This indicates that approximately 50% of those patients, whose baseline CD4 cell count were less than 96, have their CD4 cell count stabilized above 243 after the 78th week. The 95% confidence interval for this change-point is $[65, 94]$. The results for the patients whose baseline CD4 cell count are greater than 96 cells/μL are interpreted similarly: the times they reach plateau are the same as those with baseline CD4 cell count less than 96, by model formulation; However, the stabilized CD4 cell counts for those with baseline levels greather than 96 is $100 \times 10^{a_\tau}$, which is higher by a factor of $10^{-\gamma_\tau}$ compared to those less than 96.

Chapter 7
Conclusions and future work

In this chapter, we discuss and summarize the advantages as well as some limitations of the proposed method. In the first section, some important conclusions are listed. Next we discuss some issues on missing data. Finally we conclude this chapter by point out a few possible future directions on this topic in the third section.

7.1 Conclusions

This thesis achieves two goals. One is to extend the (iid) bent line quantile regression model to the longitudinal settings. The resulting longitudinal bent line quantile regression, as a special case of nonlinear quantile regression models, enjoy most of the virtue of quantile regression such as robustness to response outliers, and clear interpretations of conditional quantile functions. This extension is motivated by many longitudinal studies where the relationship between the repeatedly measured response and the covariates is not constantly linear but piecewise linear with an abrupt change in slope at certain unknown time point. Such data can be commonly found in epidemiology, medicine, nutrition, etc. [Pawitan (2005)]. In those scenarios, the longitudinal bent line quantile regression emerged as one of the indispensable tool-kits to characterize such piecewise linear pattern and estimate the change-point location especially when extreme quantile or all quantiles are of interest. The other

goal is to develop a reliable method for conducting statistical inference of the change-point, which frequently is of central interest. In order to overcome the difficulties of existing inferential approaches, e.g. estimating nuisance parameters in the Wald test and expensive computational cost of the bootstrap methods. We propose a score type test via the rank-based approach. Throughout a series of Monte Carlo simulations, we demonstrate the proposed rank score test is accurate, robust, and also computationally efficient. To illustrate the value of the proposed methods, we apply them to two real datasets. Important conclusions include: (1) for the Finnish Longitudinal Growth Study, a female subgroup in upper quantile ($\tau \geq 0.95$) of BMI experienced early adiposity rebound, one of the risk factors of adult obesity; (2) for the AIDS clinical study, the patients undergoing ART treatment exhibit different times when the treatment plateau occurs. Based on the aforementioned reasons, we recommend the proposed methods for practical use. Finally, we discuss some possible generalizations to the model (3.2.1) and the test (4.2.6).

7.2 Future work

Rank score test for slope parameters. In this thesis, we have focused on developing the rank score test statistic T_n for the change-point parameter t_τ, which is the parameter of central interest in most of the applications. As a next step, we will aim to derive rank score tests for other parameters, e.g. $b_{1,\tau}$ and $b_{2,\tau}$ in model (3.2.1), based on which, problems such as $H_0 : b_{2,\tau} = 0$ in the AIDS clinical study can be tested in an efficient manner. Furthermore, the rank score tests might be extended for general linear hypotheses. Important applications include, for example, assessing the existence of the change-point t_τ by formulating it as a hypothesis testing problem $H_0 : b_{1,\tau} = b_{2,\tau}$. Comparisons to the currently used computational intensive methods based on resampling will also be conducted.

Segmented linear quantile regression. The longitudinal bent line quantile regression model implicitly assumes a continuous response on the time domain. In some applications, however, not only the slope but the response itself may change abruptly. For example, the blood pressure monitored on patients undergoing vascular surgeries may show jumps when the artery is clamped on and off. For such models, the estimation procedure still works with minor modification. We conjecture that the rank score test on the change-point also holds due to the Rao score type interpretations.

Multiphase quantile regression. A third direction of extending the current longitudinal bent line model is to relax the implicit assumption of the existence of one single change-point. A natural extension is to consider multiple change-points. In the mean regression framework, such multiphase (also called segmented) regression models have been widely studied [see Hudson (1966), Hawkins (1976)]. In this family of models, one need to determine the number of change-points, sometimes even without *a prior* scientific knowledge. Jones and Dey (1995) determine the number of change-points that best fits the data using a modified version of Akaike's Information Criterion (AIC) to avoid overfitting. Similar ideas can be used for quantile regression with modified AIC or BIC [Schwarz (1978)]. Once the number of the change-points is determined, the profile estimation can be naturally extended to obtain the estimates of the regression coefficients and the change-point locations. Research needs to be carried out in this direction. In the framework of quantile regression, a possible solution will be based on similar criteria.

Handling missing data. In this thesis, both applications contain missing data. In the Finnish Longitudinal Growth Study, the data has been edited to discard low birth weight (3%) or unknown birth weight (2%). This missing mechanism belongs to Missing complete at random (MCAR) since the distribution of missing does not depend on covariates or outcomes and thus ignorable. Hence, the estimate from the complete data is unbiased and conclusion is valid. The missing data may exert higher impact

on the estimation for the AIDS clinical trial data, in which a considerable amount of patients dropped out after week 100. Our analysis made assumption that whether a patient will drop out from the study only depends on his/her baseline condition, but not his/her actual outcome. Under this Missing at random (MAR) assumption, the complete-set based analysis is valid, though not optimally efficient. Multiple imputations [Rubin (1987)] could be considered to further improve the estimation efficiency. In case that the MAR assumption does not hold, the complete set estimator could be biased. One may employ inverse probability weighting [Little and Rubin (2002)] or reconstructing weighted estimating equations as in [Robins et al. (1995)] to correct the bias. We refer to future research for handling the missing data in longitudinal bent line models.

Chapter 8

Proofs

8.1 Overview

In this chapter, we give complete proofs for previous results in the preceding chapters along with the regularity conditions required, followed by some explanation. An inequality similar to Hoeffding's inequality is given in Proposition 8.3.1. Then two lemmas are provided to support the proof of Theorems 3.3.3 and 4.3.4. In Lemma 8.4.1, we demonstrate an extremely useful procedure called the *chaining argument*, a general way to extend uniform convergence onto a compact set. The argument is invoked in Lemma 8.5.2 and also Theorems 3.3.3 and 4.3.4. Based on the two lemmas, the proofs of Theorems 3.3.3 and 4.3.4 follow at the end of the chapter. First, we introduce notation for the well-defined limiting quantile regression objective function and the associated marginal profile estimator,

$$\begin{aligned} Q_\tau(\boldsymbol{\eta}, t) &= \lim_{n\to\infty} Q_{n,\tau}(\boldsymbol{\eta}, t) \text{ in probability, and} \\ \boldsymbol{\eta}_\tau(t) &= \arg\min_{\boldsymbol{\eta}} Q_\tau(\boldsymbol{\eta}, t), \end{aligned} \tag{8.1.1}$$

and state the following regularity conditions.

8.2 Regularity conditions

(A0) Given $\tilde{x}_{ij} = (t_{ij}, \mathbf{x}_{ij}^\top)^\top$, the expected objective function $E[\rho_\tau(Y_{ij} - g(\tilde{x}_{ij}; \eta, t))]$ achieves its unique global minimum at true parameters $(\boldsymbol{\eta}_\tau, t_\tau) = (\alpha_\tau, b_{1,\tau}, b_{2,\tau}, \boldsymbol{\gamma}_\tau, t_\tau)$.

(A1) The distribution function of e_{ij} is absolutely continuous, with continuous densities f uniformly bounded away from 0 and ∞ at $F^{-1}(\tau)$.

(A2) t_{ij} has continuous density function $p(t)$ on a bounded support $[0, M]$, where $M > 0$.

(A3) $\max_{1 \le i \le n, 1 \le j \le m_i} \|\tilde{x}_{ij}\| = O(n^{1/4})$ and $\sum_{ij} \|\tilde{x}_{ij}\|^3 = O(n)$ as $n \to \infty$, where $\tilde{x}_{ij} = (t_{ij}, \mathbf{x}_{ij}^\top)^\top$.

(A4) Given $b_{1,\tau} \neq b_{2,\tau}$, there exists a nonnegative definite matrix C_τ, such that, $C_{n,\tau} \to C_\tau$.

(A5) Given $b_{1,\tau} \neq b_{2,\tau}$, there exists a full rank matrix D_τ, such that, $D_{n,\tau} \to D_\tau$.

(A6) $A_i = E[\psi(e_i)\psi^\top(e_i)] > 0$ for each i and $\sup_i \|A_i\| < \infty$.

(A7) The distribution function of e_{ij} has a Lebesgue density with a bounded first-order derivative f.

(A1*) There exists a constant b such that $f(F^{-1}(\tau)) = b$ for all i and j.

Condition A0 ensures the identifiability of the model (3.2.1). Conditions A1 and A3 are the standard regularity conditions in quantile regression. Conditions A0 to A3 together suffice for consistency of the estimates. Condition A6 ensures the existence of a consistent variance estimate of S_n. Condition A7 is required for Theorem 4.3.4. Note that Condition A1* requires a homogeneous error density, which is necessary for theoretical derivations. In practice, however, the rank score test statistic has been

shown to be quite robust against some deviations from this assumption via simulation studies in Chapter 5 .

We will proceed to state a proposition and two lemmas, based on which the complete proofs for Theorems 3.3.3 and 4.3.4 follow towards the end of this chapter. Since our results are stated for a given τ, from prior notation, $\psi_\tau(\cdot), \widehat{\boldsymbol{\theta}}_{n,\tau}, \boldsymbol{\theta}_\tau$, and $e_{ij}(\tau)$, we drop the dependence on τ, rewriting them as $\psi(\cdot)$, $\widehat{\boldsymbol{\theta}}_n$, $\boldsymbol{\theta}_0$ and e_{ij}. (Note that $\boldsymbol{\theta}_0$ does not denote the value of the parameter under some null, but rather it denotes the true parameter value.) For the sake of simpler presentation throughout the proofs of Proposition 8.3.1, Lemma 8.4.1 and Lemma 8.5.2, we choose the coordinate system such that $\boldsymbol{\theta}_0 = \mathbf{0}$.

Recall $h(\tilde{x}_{ij}; \boldsymbol{\theta}) = (I\{x_{ij} \le t\}, x_{ij} I\{x_{ij} \le t\}, I\{x_{ij} > t\}, x_{ij} I\{x_{ij} > t\}, \mathbf{x}_{ij}^\top\})^\top$. For notational convenience, we define

$$
\begin{aligned}
h_{ij} &= h(\tilde{x}_{ij}; \mathbf{0}), \\
\phi_i(\boldsymbol{\delta}) &= \sum_{j=1}^{n_i} h_{ij}[\psi(Y_{ij} - g(\tilde{x}_{ij}; \boldsymbol{\delta})) - \psi(Y_{ij} - g(\tilde{x}_{ij}; \mathbf{0}))] \\
\tilde{\phi}_i(\boldsymbol{\delta}) &= \phi_i(\boldsymbol{\delta}) - E[\phi_i(\boldsymbol{\delta})], \\
\tilde{\phi}(\boldsymbol{\delta}) &= \sum_{i=1}^{m} \tilde{\phi}_i(\boldsymbol{\delta}).
\end{aligned}
\tag{8.2.2}
$$

Now we state a proposition which facilitates the following proofs. The inequality we establish in the proposition, whose form is similar to the well-know Hoeffding's inequality [Hoeffding (1963)], provides an upper bound of the tail probability for the deviation of $\tilde{\phi}(\boldsymbol{\delta})$ from its mean. It is a modified version of concentration inequalities, specifically tailored for $\phi_i(\boldsymbol{\delta})$ here. As a result, our inequality relaxes the boundedness condition on the summands required for Hoeffding's inequality.

8.3 Proposition 8.3.1

Proposition 8.3.1 *Define*

$$\Delta = \left\{ \boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}\| \leq K\sqrt{\log n/n} \right\}, \textit{for some } K > 0. \tag{8.3.3}$$

For any $\lambda > 0$, $\boldsymbol{\delta} \in \Delta$, under Conditions A1 and A3,

$$P\{|\tilde{\phi}(\boldsymbol{\delta}) \geq \lambda n^{1/4} \log n|\} \leq 2\exp\{-\lambda \log n(1 + o(1))\} \tag{8.3.4}$$

where $\tilde{\phi}(\boldsymbol{\delta})$ is defined in (8.2.2).

Proof For the sake of clear presentation and without loss of generality, we assume $n_i = 2$ in model (3.2.1). The proof for other n_i can follow in a similar way with minor modifications.

First of all, by the Markov inequality, for $\delta > 0$ and any $\lambda_n > 0$, we have

$$P\{|\tilde{\phi}(\boldsymbol{\delta}) \geq \lambda_n|\} \leq e^{-t\lambda_n}[M(t) + M(-t)], \tag{8.3.5}$$

where $M(t)$ is the moment generating function of $\tilde{\phi}(\boldsymbol{\delta})$. Due to the between subject independence assumed by model (3.2.1), $M(t)$ can also be expressed as

$$M(t) = \prod_{i=1}^{m} M_i(t), \tag{8.3.6}$$

where

$$M_i(t) = E\exp\{t[\phi_i(\boldsymbol{\delta}) - E\phi_i(\boldsymbol{\delta})]\}. \tag{8.3.7}$$

Recall $y_{ij} = g(\tilde{x}_{ij}; \mathbf{0}) + e_{ij}$. By (8.2.2)(ii), we have

$$\phi_i(\boldsymbol{\delta}) = \sum_{j=1}^{2} -(I\{e_{ij} \leq F^{-1}(\tau) + g(\tilde{x}_{ij}; \boldsymbol{\delta}) - g(\tilde{x}_{ij}; 0)\} - I\{e_{ij} < F^{-1}(\tau)\})h_{ij}. \tag{8.3.8}$$

Under Condition A7, we can further denote

$$\begin{aligned} p_{i1} =& P\{e_{i1} \text{ is between 0 and } g(\tilde{x}_{i1};\boldsymbol{\delta}) - g(\tilde{x}_{i1};\mathbf{0})\}, \\ p_{i2} =& P\{e_{i2} \text{ is between 0 and } g(\tilde{x}_{i2};\boldsymbol{\delta}) - g(\tilde{x}_{i2};\mathbf{0})\}, \\ p_{i+} =& P\{ \text{ both } e_{i1} \text{ and } e_{i2} \text{ are between 0 and } g(\tilde{x}_{ij};\delta) - g(\tilde{x}_{ij};\mathbf{0})\}. \end{aligned} \tag{8.3.9}$$

Based on (8.3.7) and (8.3.9), direct calculation yields

$$\begin{aligned} M_i(t) =& (p_{i1} - p_{i+}) \exp\{-t(1-p_{i1})h_{i1}\mathrm{sgn}_{i1}(\boldsymbol{\delta}) + tp_{i2}h_{i2}\mathrm{sgn}_{i2}(\boldsymbol{\delta})\} \\ &+ (p_{i2} - p_{i+}) \exp\{tp_{i1}h_{i1}\mathrm{sgn}_{i1}(\boldsymbol{\delta}) - t(1-p_{i2})h_{i2}\mathrm{sgn}_{i2}(\boldsymbol{\delta})\} \\ &+ p_{i+} \exp\{-t(1-p_{i1})h_{i1}\mathrm{sgn}_{i1}(\boldsymbol{\delta}) - t(1-p_{i2})h_{i2}\mathrm{sgn}_{i2}(\boldsymbol{\delta})\} \\ &+ (1 - p_{i1} - p_{i2} + p_{i+}) \exp\{tp_{i1}h_{i1}\mathrm{sgn}_{i1}(\boldsymbol{\delta}) + tp_{i2}h_{i2}\mathrm{sgn}_{i2}(\boldsymbol{\delta})\}, \end{aligned} \tag{8.3.10}$$

where $\mathrm{sgn}_{ij}(\boldsymbol{\delta}) = \mathrm{sgn}(g(\tilde{x}_{ij};\boldsymbol{\delta}) - g(\tilde{x}_{ij};\mathbf{0}))$ for $j = 1, 2$, with $\mathrm{sgn}(\cdot)$ denoting the signum function of a real number s defined as follows:

$$\mathrm{sgn}(s) = \begin{cases} -1 & , \text{ if } s < 0 \ , \\ 0 & , \text{ if } s = 0 \ , \\ 1 & , \text{ if } s > 0 \ . \end{cases} \tag{8.3.11}$$

If $t = O(n^{-1/4})$, $|g(\tilde{x}_{ij};\boldsymbol{\delta}) - g(\tilde{x}_{ij};\mathbf{0})|$ is bounded by A3, and hence for $\boldsymbol{\delta} \in \Delta$,

$$\begin{aligned} p_{i1} &= f(F^{-1}(\tau))|g(\tilde{x}_{i1};\boldsymbol{\delta}) - g(\tilde{x}_{i1};\mathbf{0})| \le c^*\|\tilde{x}_{i1}\|\|\boldsymbol{\delta}\|, \\ p_{i2} &= f(F^{-1}(\tau))|g(\tilde{x}_{i2};\boldsymbol{\delta}) - g(\tilde{x}_{i2};\mathbf{0})| \le c^*\|\tilde{x}_{i2}\|\|\boldsymbol{\delta}\|, \\ p_{i+} &\le \max\{p_{i1}, p_{i2}\} \le c^* \max\{\|\tilde{x}_{i1}\|, \|\tilde{x}_{i2}\|\}\|\boldsymbol{\delta}\|, \end{aligned} \tag{8.3.12}$$

for some $c^* > 0$ since f is bounded. Moreover, in view of (8.3.12), taking the Taylor

series expansion of 8.3.10 reveals

$$\log M_i(t) \leq \log\left(1 + \sum_{k=1}^{K} c_k t^k + o(t^K)\right) \tag{8.3.13}$$

where

$$\begin{aligned} c_k =& \frac{p_{i1} - p_{i+}}{k!}\left[-\begin{pmatrix} h_{i1}\mathrm{sgn}_{i1}(\boldsymbol{\delta}) & h_{i2}\mathrm{sgn}_{i2}(\boldsymbol{\delta})\end{pmatrix}\begin{pmatrix} 1-p_{i1} \\ -p_{i2}\end{pmatrix}\right]^k + \\ & \frac{p_{i2} - p_{i+}}{k!}\left[-\begin{pmatrix} h_{i1}\mathrm{sgn}_{i1}(\boldsymbol{\delta}) & h_{i2}\mathrm{sgn}_{i2}(\boldsymbol{\delta})\end{pmatrix}\begin{pmatrix} -p_{i1} \\ 1-p_{i2}\end{pmatrix}\right]^k + \\ & \frac{p_{i+}}{k!}\left[-\begin{pmatrix} h_{i1}\mathrm{sgn}_{i1}(\boldsymbol{\delta}) & h_{i2}\mathrm{sgn}_{i2}(\boldsymbol{\delta})\end{pmatrix}\begin{pmatrix} 1-p_{i1} \\ 1-p_{i2}\end{pmatrix}\right]^k + \\ & \frac{1 - p_{i1} - p_{i2} + p_{i+}}{k!}\left[\begin{pmatrix} h_{i1}\mathrm{sgn}_{i1}(\boldsymbol{\delta}) & h_{i2}sgn_{i2}(\boldsymbol{\delta})\end{pmatrix}\begin{pmatrix} p_{i1} \\ p_{i2}\end{pmatrix}\right]^k \end{aligned} \tag{8.3.14}$$

is the coefficient for the kth power. It is not difficult to verify that $c_1 = 0$. Hence for sufficiently large n, there exists some positive constants c', c'' such that

$$\begin{aligned} \log M_i(t) &\leq \log\left(1 + c't^2\|\mathbf{p}_i^\top \mathbf{h}_i\|\right) \\ &\leq c't^2\|\mathbf{p}_i^\top \mathbf{h}_i\| \\ &\leq c''t^2\|\tilde{x}_i\boldsymbol{\delta}\|\|\mathbf{h}_i\| \end{aligned} \tag{8.3.15}$$

where

$$\mathbf{p}_i = (p_{i1}, \min(p_{1,1}, p_{i2}), p_{i2})^\top, \mathbf{h}_i = (h_{i1}^2, 2|h_{i1}h_{i2}|, h_{i2}^2)^\top, \mathbf{1} = (1,1,1)^\top. \tag{8.3.16}$$

Therefore, by Condition A3, for $t > 0$, and $t = O(n^{-1/4})$,

$$\begin{aligned} \log M(t) &= \sum_{i=1}^{m} M_i(t) \\ &\leq b' t^2 \|\boldsymbol{\delta}\| \|\tilde{\mathbf{x}}_i\|^3 \\ &\leq b'' t^2 (n \log n)^{1/2}. \end{aligned} \tag{8.3.17}$$

for some positive constants b', b''. Finally, by (8.3.5), with $t = n^{-1/4}$,

$$P\{|\tilde{\phi}(\boldsymbol{\delta})| \geq \lambda n^{1/4} \log n\} \leq 2 \exp\{-\lambda \log n + b'' \sqrt{\log n}\}. \tag{8.3.18}$$

.

Hence (8.3.4) follows immediately. The proof of Proposition (8.3.1) is complete.

8.4 Lemma 8.4.1

The proofs of Theorems 3.3.3 and 4.3.4 rely on the following two lemmas. The first one is crucial. It extends the inequality (8.3.4) uniformly in Δ (8.3.3) via a *chaining argument*.

Lemma 8.4.1 *Under Conditions A1 and A3, and letting $\tilde{\phi}(\boldsymbol{\delta})$ be as in (8.2.2),*

$$\sup_{\boldsymbol{\delta} \in \Delta} |\tilde{\phi}(\boldsymbol{\delta})| = o_p(n^{1/4} \log n), \tag{8.4.19}$$

with Δ as defined in (8.3.3).

Proof For the sake of clear presentation and without loss of generality, take $\|\boldsymbol{\delta}\|$ to be the sup-norm, $\|\boldsymbol{\delta}\| = \max(\delta_1, \ldots, \delta_p)$ for $\boldsymbol{\delta} \in \mathbb{R}^p$.

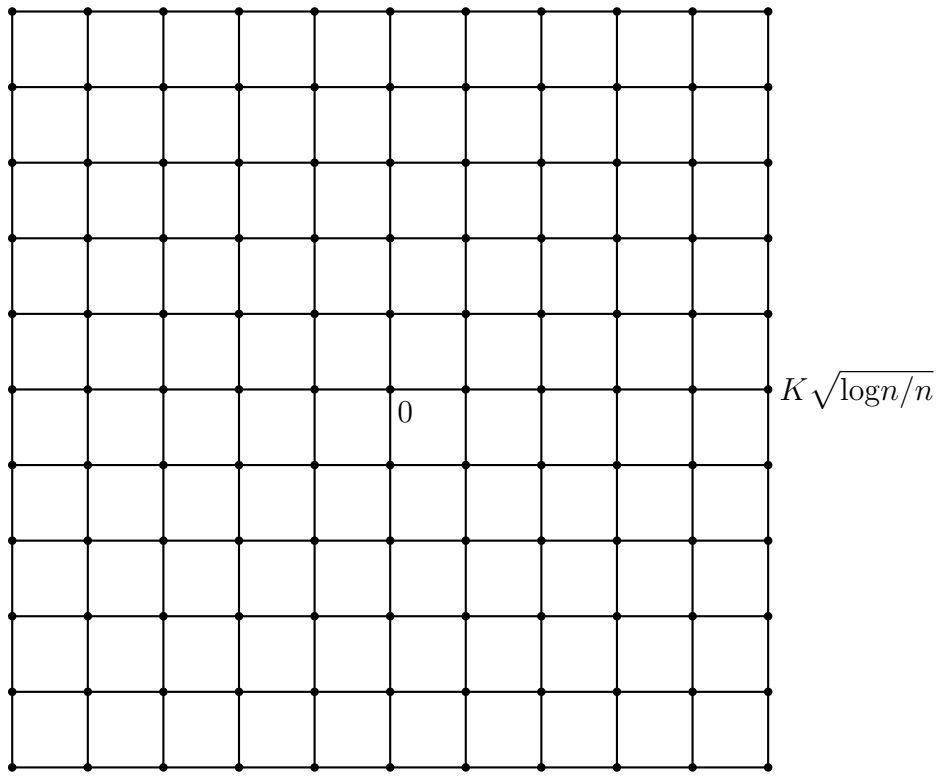

Figure 8.1: An illustration of the partition of $\Delta = \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}\| \leq K\sqrt{\log n/n}\}$ used in the chaining argument

The idea of the proof is to partition the cube (or sphere equivalent if Euclidean norm is employed) Δ into a number of smaller sub-cubes. Then we must show: (1) Inequality (8.3.4) holds on all the centers of the sub-cubes, and (2) for $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2$ sufficiently close, that is within one sub-cube, the difference $|\tilde{\phi}(\boldsymbol{\delta}_1) - \tilde{\phi}(\boldsymbol{\delta}_2)|$ is also $o_p(n^{1/4} \log n)$. Then the uniform convergence (8.4.19) can be implied by the triangle inequality.

Actually, we shall let $\boldsymbol{\delta}_i \in \Delta$ be the centers of sub-cubes with edges of length $2n^{-3}$, covering Δ (See Figure 8.1). Let $B = \{\boldsymbol{\delta} \in \Delta : \boldsymbol{\delta} = \boldsymbol{\delta}_k, \text{ for } 1 \leq k \leq n^{3p}\}$ denote the set containing the centers of the sub-cubes. Then $\#B \leq an^{3p}$ where $\#$ denotes the number of elements.

Hence, by Proposition (8.3.1), for any $\epsilon > 0$,

$$\begin{aligned} P\{\sup_{\boldsymbol{\delta}\in B}|\tilde{\phi}(\boldsymbol{\delta})| \geq (3p+2)n^{1/4}\log n\} &\leq an^{3p}P\{|\tilde{\phi}(\boldsymbol{\delta})| \geq (3p+2)n^{1/4}\log n\} \\ &\leq 2an^{-1} \to 0, \text{ as } n \to \infty. \end{aligned} \tag{8.4.20}$$

Consequently,

$$\sup_{\boldsymbol{\delta}\in B}|\tilde{\phi}(\boldsymbol{\delta})| = o_p(n^{1/4}\log n). \tag{8.4.21}$$

Next, for $\{\boldsymbol{\delta}_1, \boldsymbol{\delta}_2\} \subset \Delta$, with $\|\boldsymbol{\delta}_1 - \boldsymbol{\delta}_2\| \leq n^{-3}$, we consider

$$\begin{aligned} &|\phi(\boldsymbol{\delta}_1) - \phi(\boldsymbol{\delta}_2)| \\ \leq &\left|\sum_{i,j} h_{ij}[I\{e_{ij} < g(\tilde{x}_{ij};\boldsymbol{\delta}_1) - g(\tilde{x}_{ij};\mathbf{0})\} - I\{e_{ij} < g(\tilde{x}_{ij};\boldsymbol{\delta}_2) - g(\tilde{x}_{ij};\mathbf{0})\}]\right| \\ \leq &\sum_{i,j}|h_{ij}||I\{e_{ij} < g(\tilde{x}_{ij};\boldsymbol{\delta}_1) - g(\tilde{x}_{ij};\mathbf{0})\} - I\{e_{ij} < g(\tilde{x}_{ij};\boldsymbol{\delta}_2) - g(\tilde{x}_{ij};\mathbf{0})\}]|. \end{aligned} \tag{8.4.22}$$

Then by Condition A3,

$$|g(\tilde{x}_{ij};\boldsymbol{\delta}_1) - g(\tilde{x}_{ij};\boldsymbol{\delta}_2)| \leq c_3 n^{1/4} n^{-3} < c_3 n^{-2.5}. \tag{8.4.23}$$

Now for $(i_1, j_1) \neq (i_2, j_2)$,

$$\begin{aligned} &P\{|e_{i_1,j_1} - e_{i_2,j_2}| \leq c_3 n^{-2.5}\} \\ =&P\{e_{i_1,j_1} \in [e_{i_2,j_2} - c_3 n^{-2.5}, e_{i_2,j_2} + c_3 n^{-2.5}]\} \leq c_4 n^{-2.5}, \end{aligned} \tag{8.4.24}$$

since f is bounded. Hence

$$P\left\{\min_{(i_1,j_1)\neq(i_2,j_2)}|e_{i_1,j_1} - e_{i_2,j_2}| \leq c_3 n^{-2.5}\right\} \leq n(n-1)c_4 n^{-2.5} \to 0. \tag{8.4.25}$$

It follows that with probability tending to 1, at most two of the terms in (8.4.22)

(where each term corresponds to one (i, j)-pair) is nonzero, that is,

$$
\begin{aligned}
&P\left(\{|\phi(\boldsymbol{\delta}_1) - \phi(\boldsymbol{\delta}_2)| > \epsilon n^{1/4}\} \bigcap \{\min_{(i_1,j_1)\neq(j_2,j_2)} |e_{i_1,j_1} - e_{i_2,j_2}| > c_3 n^{-2.5}\}\right) \\
\leq & P(\sum_{(i,j)=(i_1^*,j_1^*) \text{ or } (i_2^*,j_2^*)} |h_{ij}| |I\{e_{ij} < g(\tilde{x}_{ij};\boldsymbol{\delta}_1) - g(\tilde{x}_{ij};\mathbf{0})\} \\
& \qquad - I\{e_{ij} < g(\tilde{x}_{ij};\boldsymbol{\delta}_2) - g(\tilde{x}_{ij};\mathbf{0})\}| > \epsilon n^{1/4}) \\
& \to 0, \text{ as } n \to \infty.
\end{aligned}
\tag{8.4.26}
$$

Hence, combining (8.4.25) and Condition A3, we obtain

$$
|\phi(\boldsymbol{\delta}_1) - \phi(\boldsymbol{\delta}_2)| \leq 2 \max |h_{ij}| = O_p(n^{1/4}). \tag{8.4.27}
$$

In addition, from (8.4.22),

$$
\begin{aligned}
& |E[\phi(\boldsymbol{\delta}_1) - \phi(\boldsymbol{\delta}_2)]| \\
\leq & \sum_{ij} |h_{ij}| |F(g(\tilde{x}_{ij};\boldsymbol{\delta}_1) - g(\tilde{x}_{ij};\mathbf{0})) - F(g(\tilde{x}_{ij};\boldsymbol{\delta}_2) - g(\tilde{x}_{ij};\mathbf{0}))|,
\end{aligned}
\tag{8.4.28}
$$

since f is bounded. Then using (8.4.23) and Condition A3, along with the fact that $\|\boldsymbol{\delta}_1 - \boldsymbol{\delta}_2\| \leq n^{-3}$, (8.4.28) is bounded. It follows that,

$$
|\tilde{\phi}(\boldsymbol{\delta}_1) - \tilde{\phi}(\boldsymbol{\delta}_2)| = O_p(n^{1/4}) \tag{8.4.29}
$$

uniformly for $(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) \subset \Delta$ with $\|\boldsymbol{\delta}_1 - \boldsymbol{\delta}_2\| \leq n^{-3}$. Thus combined with (8.4.21), we

obtain

$$
\begin{aligned}
&P\{\sup_{\boldsymbol{\delta}\in\Delta}|\tilde{\phi}(\boldsymbol{\delta})| > 2\epsilon n^{1/4}\log n\} \\
\leq &P\{\sup_{\boldsymbol{\delta}\in B}|\tilde{\phi}(\boldsymbol{\delta})| > \epsilon n^{1/4}\log n\} + P\{\sup_{\|\boldsymbol{\delta}_1-\boldsymbol{\delta}_2\|\leq n^{-3}}|\tilde{\phi}(\boldsymbol{\delta}_1)-\tilde{\phi}(\boldsymbol{\delta}_2)| > \epsilon n^{1/4}\log n\} \\
\to &0, \text{ as } n\to\infty.
\end{aligned}
\tag{8.4.30}
$$

Therefore, (8.4.19) holds immediately. The proof of Lemma 2 is hence complete.

8.5 Lemma 8.5.2

The following Lemma establishes the consistency of $\widehat{\boldsymbol{\theta}}_n$ for a given quantile level $\tau \in (0,1)$.

Lemma 8.5.2 (Consistency) *Under Conditions (A0) to (A3),* $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \to 0$ *in probability.*

Proof The proof of this lemma is essentially the same as that of Lemma 1 in Li et al. (2011) for the iid error case where a brief proof is outlined. The same arguments readily apply after identifying our objective function $\sum_{i=1}^{m}\sum_{j=1}^{n_i}\rho_\tau(Y_{ij} - g(\tilde{x}_{ij};\boldsymbol{\eta},t))$ in place of $S_{n,\tau}(\boldsymbol{\eta},t)$ in Li et al. (2011). Now we rephrase the proof with more detail.

With fixed t, the longitudinal bent line model (3.2.1) is linear in form. By the consistency of linear quantile regression provided by Koenker (2005), and conditions A1-A3, it follows that

$$
\|\widehat{\boldsymbol{\eta}}_{n,\tau}(t) - \boldsymbol{\eta}_\tau(t)\| = o_p(1), \tag{8.5.31}
$$

for any t in $[-M, M]$ with M a positive constant as in Condition A2.

Upon letting $t_k = -M + k/n^3, k = 0, 1, \ldots, [2Mn^3]$, we partition the interval $[-M, M]$ into $[2Mn^3]$ sub-intervals each centered at t_k and of length n^{-3}, covering

$[-M, M]$. By the *chaining argument* demonstrated in Lemma 8.4.1, we can extend (8.5.31) uniformly in $t : |t| \leq M$ and get

$$\sup_{|t|\leq M} \|\widehat{\boldsymbol{\eta}}_{n,\tau}(t) - \boldsymbol{\eta}_\tau(t)\| = o_p(1). \tag{8.5.32}$$

On the other hand, we know $|\widehat{t}_{n,\tau}| \leq M$ as implied by Condition A2, and $\boldsymbol{\eta}_\tau(t)$ is continuous in t. Based on (8.5.31) and the fact

$$\|\widehat{\boldsymbol{\eta}}_{n,\tau}(\widehat{t}_{n,\tau}) - \boldsymbol{\eta}_\tau(t_\tau)\| \leq \|\widehat{\boldsymbol{\eta}}_{n,\tau}(\widehat{t}_{n,\tau}) - \boldsymbol{\eta}_\tau(\widehat{t}_{n,\tau})\| + \|\boldsymbol{\eta}_\tau(\widehat{t}_{n,\tau}) - \boldsymbol{\eta}_\tau(t_\tau)\|, \tag{8.5.33}$$

it is clear that all we need to show is $|\widehat{t}_{n,\tau} - t_\tau| = o_p(1)$ for the consistency of $\widehat{\boldsymbol{\theta}}_n(\tau)$.

Based on Conditions A0-A2, the limiting objective function $Q_\tau(\boldsymbol{\eta}_\tau(t), t)$ is continuous in t, and uniquely minimized at t_τ. Since $\widehat{t}_{n,\tau}$ is bounded by M, the following condition suffices for the consistency of $\widehat{t}_{n,\tau}$,

$$\sup_{|t|\leq M} |Q_{n,\tau}(\widehat{\boldsymbol{\eta}}_{n,\tau}(t), t) - Q_\tau(\boldsymbol{\eta}_\tau(t), t)| = o_p(1). \tag{8.5.34}$$

Furthermore, we follow the similar idea used in Lemma 8.4.1 and partition the cube $\|\boldsymbol{\eta}\| \leq M'$ into n^{3+q} sub-cubes, each with edge length $2n^{-3}$ covering $\|\boldsymbol{\eta}\| \leq M'$. By the *chaining argument* used in the proof of Lemma 8.4.1, we obtain

$$\sup_{|t|\leq M; \|\boldsymbol{\eta}\|\leq M'} |Q_{n,\tau}(\boldsymbol{\eta}, t) - Q_\tau(\boldsymbol{\eta}, t)| = o_p(1). \tag{8.5.35}$$

Hence by (8.5.35) it follows that

$$\sup_{|t|\leq M} |Q_{n,\tau}(\widehat{\boldsymbol{\eta}}_{n,\tau}(t), t) - Q_\tau(\widehat{\boldsymbol{\eta}}_{n,\tau}(t), t)| = o_p(1). \tag{8.5.36}$$

In addition, since $Q_\tau(\cdot, t)$ is continuous in $\boldsymbol{\eta}$, the uniform convergence of $\widehat{\boldsymbol{\eta}}_{n,\tau}(t)$

(8.5.31) also implies

$$\sup_{|t|\leq M} |Q_\tau(\widehat{\boldsymbol{\eta}}_{n,\tau}(t), t) - Q_\tau(\boldsymbol{\eta}_\tau(t), t)| = o_p(1). \tag{8.5.37}$$

Finally, notice that the left hand side of (8.5.34) is bounded by

$$\sup_{|t|\leq M} |Q_{n,\tau}(\widehat{\boldsymbol{\eta}}_{n,\tau}(t), t) - Q_\tau(\widehat{\boldsymbol{\eta}}_{n,\tau}(t), t)| + \sup_{|t|\leq M} |Q_\tau(\widehat{\boldsymbol{\eta}}_{n,\tau}(t), t) - Q_\tau(\boldsymbol{\eta}_\tau(t), t)|. \tag{8.5.38}$$

Now putting (8.5.36) and (8.5.37) together, we obtain (8.5.34). Thus the proof of Lemma 8.5.2 is now complete.

8.6 Lemma 8.6.3

Next we state a lemma similar to Lemma (8.4.1), based on which the consistency of the variance estimate of S_n follows easily.

Lemma 8.6.3 *Under Condition A1-A3,*

$$\begin{aligned}\sup_{\boldsymbol{\delta}\in\Delta} &|\psi(e_{i,j_1} + g(\tilde{x}_{i,j_1}; \mathbf{0}) - g(\tilde{x}_{i,j_1}; \boldsymbol{\delta}))\psi(e_{i,j_2} + g(\tilde{x}_{i,j_2}; \mathbf{0}) - g(\tilde{x}_{i,j_2}; \boldsymbol{\delta})) - \psi(e_{i,j_1})\psi(e_{i,j_2})| \\ &= o_p(n^{1/4}\log n).\end{aligned} \tag{8.6.39}$$

where $\Delta = \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}\| \leq K(\log n/n)^{1/2}\}$ *for some constant* $K > 0$ *as defined in (8.3.3).*

Proof Recall that the score function $\psi(u) = \tau - I\{u < 0\}$. Based on which,

direct calculation shows

$$
\begin{aligned}
&\psi(e_{i,j_1}+g(\tilde{x}_{i,j_1};\mathbf{0})-g(\tilde{x}_{i,j_1};\boldsymbol{\delta}))\psi(e_{i,j_2}+g(\tilde{x}_{i,j_2};\mathbf{0})-g(\tilde{x}_{i,j_2};\boldsymbol{\delta}))-\psi(e_{i,j_1})\psi(e_{i,j_2})\\
=&-\tau(I\{e_{i,j_1}<g(\tilde{x}_{i,j_1};\boldsymbol{\delta})-g(\tilde{x}_{i,j_1};\mathbf{0})\}-I\{e_{i,j_1}<0\})\\
&-\tau(I\{e_{i,j_2}<g(\tilde{x}_{i,j_2};\boldsymbol{\delta})-g(\tilde{x}_{i,j_2};\mathbf{0})\}-I\{e_{i,j_2}<0\})\\
&+I\{e_{i,j_1}<g(\tilde{x}_{i,j_1};\boldsymbol{\delta})-g(\tilde{x}_{i,j_1};\mathbf{0})\}I\{e_{i,j_2}<g(\tilde{x}_{i,j_2};\boldsymbol{\delta})-g(\tilde{x}_{i,j_2};\mathbf{0})\}\\
&-I\{e_{i,j_1}<0\}I\{e_{i,j_2}<0\}\\
=&(I\{e_{i,j_1}<g(\tilde{x}_{i,j_1};\boldsymbol{\delta})-g(\tilde{x}_{i,j_1};\mathbf{0})\}-\tau)(I\{e_{i,j_2}<g(\tilde{x}_{i,j_2};\boldsymbol{\delta})-g(\tilde{x}_{i,j_2};\mathbf{0})\}-I\{e_{i,j_2}<0\})\\
&+(I\{e_{i,j_2}<0\}-\tau)(I\{e_{i,j_1}<g(\tilde{x}_{i,j_1};\boldsymbol{\delta})-g(\tilde{x}_{i,j_1};\mathbf{0})\}-I\{e_{i,j_1}<0\}).
\end{aligned}
\tag{8.6.40}
$$

Note that $|I\{e_{i,j_k}<g(\tilde{x}_{i,j_k};\boldsymbol{\delta})-g(\tilde{x}_{i,j_k};\mathbf{0})\}-\tau|\le 2$ for $k=1,2$; and hence, by Lemma 8.4.1, it follows that (8.6.39) holds immediately. The proof of Lemma 8.6.3 is complete.

8.7 Proof of Theorem 3.3.3

In this section, we give a complete proof of Theorem 3.3.3. The ideas of the proof are outlined first and the technical details follow.

With the preparation of the previous lemmas, it turns out that the key to the proof of Theorem 3.3.3 is to show that the following equation holds,

$$
\begin{aligned}
n^{-1/2}\|\sum_{i=1}^{m}\sum_{j=1}^{n_i}&[\psi(Y_{ij}-g(\tilde{x}_{ij};\widehat{\boldsymbol{\theta}}_n))h(\tilde{x}_{ij};\widehat{\boldsymbol{\theta}}_n)-\psi(e_{ij})h(\tilde{x}_{ij};\boldsymbol{\theta}_0)]\\
&-E[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\psi(Y_{ij}-g(\tilde{x}_{ij};\boldsymbol{\theta}))h(\mathbf{w}_{ij};\boldsymbol{\theta})]|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n}\|=o_p(1),
\end{aligned}
\tag{8.7.41}
$$

where the expectations are taken with respect to the underlying distribution of Y_{ij}.

To this end, we recall

$$Y_{ij} = g(\tilde{x}_{ij}; \boldsymbol{\theta}_0) + e_{ij} \text{ as in 3.2.1}, \psi(u) = u - I\{u < 0\}, \tag{8.7.42}$$

and define the following notation for easier reference,

$$\begin{aligned}
\boldsymbol{U}_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) &= \sum_{i=1}^{m} \sum_{j=1}^{n_i} [\psi(Y_{ij} - g(\tilde{x}_{ij}; \boldsymbol{\theta}))h(\tilde{x}_{ij}; \boldsymbol{\theta}) - \psi(Y_{ij} - g(\tilde{x}_{ij}; \boldsymbol{\theta}_0))h(\tilde{x}_{ij}; \boldsymbol{\theta}_0)] \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n_i} [u_{n1}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + u_{n2}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)], \\
\boldsymbol{U}_{n1}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) &= \sum_{i=1}^{m} \sum_{j=1}^{n_i} [-I\{e_{ij} < g(\tilde{x}_{ij}; \boldsymbol{\theta}) - g(\tilde{x}_{ij}; \boldsymbol{\theta}_0) < 0\} + I\{e_{ij} < 0\}]h(\tilde{x}_{ij}; \boldsymbol{\theta}_0), \\
\boldsymbol{U}_{n2}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) &= \sum_{i=1}^{m} \sum_{j=1}^{n_i} [h(\tilde{x}_{ij}; \boldsymbol{\theta}) - h(\tilde{x}_{ij}; \boldsymbol{\theta}_0)][\tau - I\{e_{ij} < g(\tilde{x}_{ij}; \boldsymbol{\theta}) - g(\tilde{x}_{ij}; \boldsymbol{\theta}_0)\}].
\end{aligned} \tag{8.7.43}$$

Compared to (8.7.42), it is not difficult to verify

$$U_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = U_{n1}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + U_{n2}(\boldsymbol{\theta}, \boldsymbol{\theta}_0). \tag{8.7.44}$$

Proof By applying Lemma 8.4.1 to $U_{n1}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - E[U_{n1}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)]$, we obtain

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \le K(\log n/n)^{1/2}} \| \sum_{i=1}^{m} \sum_{j=1}^{n_i} \boldsymbol{U}_{n1}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - E[\boldsymbol{U}_{n1}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)]\| = o_p(\sqrt{n}); \tag{8.7.45}$$

In addition, by Chebyshev's inequality, it follows

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \le K(\log n/n)^{1/2}} \| \sum_{i=1}^{m} \sum_{j=1}^{n_i} \boldsymbol{U}_{n2}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - E[\boldsymbol{U}_{n2}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)]\| = o_p(\sqrt{n}). \tag{8.7.46}$$

Combining (8.7.45), (8.7.46), and the consistency of $\widehat{\boldsymbol{\theta}}_n$ as established in Lemma 8.5.2

yields

$$n^{-1/2}\|\sum_{i=1}^{m}\sum_{j=1}^{n_i}[\psi(Y_{ij}-g(\tilde{x}_{ij};\widehat{\boldsymbol{\theta}}_n))h(\tilde{x}_{ij};\widehat{\boldsymbol{\theta}}_n)-\psi(e_{ij})h(\tilde{x}_{ij};\boldsymbol{\theta}_0)] - E\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\psi(Y_{ij}-g(\tilde{x}_{ij};\boldsymbol{\theta}))h(\tilde{x}_{ij};\boldsymbol{\theta})\right]\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n}\| = o_p(1). \tag{8.7.47}$$

The Taylor expansion of $E[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\psi(Y_{ij}-g(\tilde{x}_{ij};\boldsymbol{\theta}))h(\tilde{x}_{ij};\boldsymbol{\theta})]|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n}$ around $\boldsymbol{\theta}_0$, together with the fact that $E[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\psi(Y_{ij}-g(\tilde{x}_{ij};\boldsymbol{\theta}_0))h(\tilde{x}_{ij};\boldsymbol{\theta}_0)]=0$, leads to the following equations,

$$\begin{aligned}
&E\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\psi(Y_{ij}-g(\tilde{x}_{ij};\boldsymbol{\theta}))h(\tilde{x}_{ij};\boldsymbol{\theta})\right]\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n}\\
=&\frac{\partial E\sum_{i=1}^{m}\sum_{j=1}^{n_i}\psi(Y_{ij}-g(\tilde{x}_{ij};\boldsymbol{\theta}))h(\tilde{x}_{ij};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)+R_n\\
=&nD_{n,\tau}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)+R_n,
\end{aligned} \tag{8.7.48}$$

where $R_n = o_p(n^{1/2})$.

Since both $g(\tilde{x}_{ij};\boldsymbol{\theta})$ and $h(\tilde{x}_{ij};\boldsymbol{\theta})$ are differentiable with respect to $(a, b_1, b_2, \boldsymbol{\gamma}^\top)^\top$, together with Condition A1, the expectation $E[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\psi(Y_{ij}-g(\tilde{x}_{ij};\boldsymbol{\theta})h(\tilde{x}_{ij},\boldsymbol{\theta}))]$ is consequently differentiable with respect to those parameters. By same argument of Theorem 1 in Li et al. (2011), we can show that, under Conditions A1 and A2, the expectation $E[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\psi(Y_{ij}-g(\tilde{x}_{ij};\boldsymbol{\theta})h(\tilde{x}_{ij},\boldsymbol{\theta}))]$ is also differentiable with respect to t. Hence the matrix $D_{n,\tau}$ is well-defined.

In view of (8.7.47), (8.7.48) and (3.3.11), we obtain

$$n^{1/2}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0) = -n^{-1/2}D_{n,\tau}^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\psi(e_{ij})h(\tilde{x}_{ij};\boldsymbol{\theta}_0)+o_p(1). \tag{8.7.49}$$

Based on the central limit theorem, $-n^{-1/2}D_{n,\tau}^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\psi(e_{ij})h(\tilde{x}_{ij};\boldsymbol{\theta}_0)$ is asymptotically normal with limiting variance covariance matrix $\tau(1-\tau)D_\tau^{-1}C_\tau D_\tau^{-\top}$, where $^{-\top}$ denotes the transpose of the matrix inverse. Hence, the rest of Theorem 3.3.3 holds immediately.

Corollary 8.7.1 *Under conditions (A0)-(A3) and H_0, we have $\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0 = O_p(n^{-1/2})$.*

8.8 Proof of Theorem 4.3.4

Proof Recall that $S_n = n^{-1/2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\psi(\widehat{e}_{ij})z^*(t_{ij};\widehat{\boldsymbol{\eta}})$. Since the $z^*(t_{ij};\widehat{\boldsymbol{\eta}})$'s are the residuals from the least squares projection of the $z(t_{ij};\widehat{\boldsymbol{\eta}})$ into the space spanned by $\mathcal{W}(t_0)$, they differ from $z^{(0)}(t_{ij};\widehat{\boldsymbol{\eta}}) = z^*(t_{ij};\widehat{\boldsymbol{\eta}}) - E\{z^*(t_{ij};\widehat{\boldsymbol{\eta}})|\mathcal{W}(t_0)_{ij}\}$ by $O_p(n^{-1/2})$. It is easy to show that the limiting distribution of $T_n = S_n^2/V_n$ will not be affected if the $z^*(t_{ij};\widehat{\boldsymbol{\eta}})$'s are replaced by the $z^{(0)}(t_{ij};\widehat{\boldsymbol{\eta}})$'s; the latter enjoy between subject independence and are often easier to handle mathematically. For sake of simplicity, we will simply prove the results in this section by assuming the $z^*(t_{ij};\widehat{\boldsymbol{\eta}})$'s are independent between subjects. Similar arguments are used in Wei and He (2006).

Let $S_n^* = n^{-1/2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\psi(e_{ij})z^*(t_{ij};\boldsymbol{\eta}_0)$ where $z^*(t_{ij};\boldsymbol{\eta}_0)$ is defined in a similar way as $z^*(t_{ij};\widehat{\boldsymbol{\eta}})$ from (4.2.3) except that $\widehat{\boldsymbol{\eta}}$ is replaced by $\boldsymbol{\eta}_0$. Then the summands $\sum_{j=1}^{n_i}\psi(e_{ij})z^*(t_{ij};\boldsymbol{\eta}_0)$ are independent of each other and have mean zero. Due to the between subject independence,

$$\mathrm{Var}(S_n^*) = n^{-1}\sum_{i=1}^{m}\mathrm{Var}\left(\sum_{j=1}^{n_i}\psi(e_{ij})z^*(t_{ij};\boldsymbol{\eta}_0)\right) = n^{-1}\sum_{i=1}^{m}\mathbf{z}^*(t_i;\boldsymbol{\eta}_0)^\top A_i\mathbf{z}^*(t_i;\boldsymbol{\eta}_0), \tag{8.8.50}$$

where A_i are $n_i\times n_i$ matrices with j-j' entries $\psi(e_{ij})\psi(e_{ij'})$ defined in a similar fashion to $\widehat{A}_i$ in the Section 2. Just like $z^*(t_{ij};\widehat{\boldsymbol{\eta}})$ and $\mathbf{z}^*(t_i;\widehat{\boldsymbol{\eta}})$, $z^*(t_{ij};\boldsymbol{\eta}_0)$ and $\mathbf{z}^*(t_i;\boldsymbol{\eta}_0)$ are

also defined in a similar fashion. Let $V_n^* = n^{-1}\sum_{i=1}^m \mathbf{z}^*(t_i;\boldsymbol{\eta}_0)^\top A_i \mathbf{z}^*(t_i;\boldsymbol{\eta}_0)$; it follows from the CLT that

$$S_n^* \text{ is asymptotic} N(0, V_n^*). \tag{8.8.51}$$

By Lemma (8.6.3) and Corollary (8.7.1) along with the continuous mapping theorem, it is not difficult to show

$$\|V_n - V_n^*\| = o_p(1). \tag{8.8.52}$$

Hence, by combining (8.8.51), (8.8.52) and Slutsky's theorem, it is clear that all we need to show to prove Theorem 4.3.4 is

$$|S_n - S_n^*| = o_p(1). \tag{8.8.53}$$

Due to non-differentiability of the function S_n with respect to $\boldsymbol{\theta}$, standard theory does not apply for (8.8.53). Instead, we take an intermediate step. Consider any $\boldsymbol{\delta}$ such that $\|\delta\| \leq K(\log n/n)^{1/2}$ for some constant K. Let $r_n(\boldsymbol{\delta}) = \sum_{ij} \psi(Y_{ij} - g(\tilde{x}_{ij};\boldsymbol{\eta}_0 + \boldsymbol{\delta}, t_0))z^*(t_{ij};\boldsymbol{\eta}_0 + \boldsymbol{\delta}) - \psi(Y_{ij} - g(\tilde{x}_{ij},\boldsymbol{\eta}_0, t_0))z^*(t_{ij};\boldsymbol{\eta}_0) = \sum_{ij} R_{ij}(\boldsymbol{\delta})$,

Lemma 8.4.1 is invoked and we obtain,

$$\sup_{\|\boldsymbol{\delta}|\leq K(\log n/n)^{1/2}} |r_n(\boldsymbol{\delta}) - Er_n(\boldsymbol{\delta})| = o_p(n^{1/4}\log n). \tag{8.8.54}$$

Furthermore, by the partition on $r_n(\boldsymbol{\delta})$ similar to (8.7.43), and the fact that $E[\psi(y_{ij} - g(\tilde{x}_{ij};\boldsymbol{\eta}_0, t_0))] = 0$, it follows that

$$\begin{aligned} &E[r_n(\boldsymbol{\delta})] \\ &= \sum_{i,j} z^*(t_{ij};\boldsymbol{\eta}_0 + \boldsymbol{\delta})[F(F^{-1}(\tau) + g(\tilde{x}_{ij};\boldsymbol{\eta}_0 + \boldsymbol{\delta}, t_0) - g(\tilde{x}_{ij};\boldsymbol{\eta}_0, t_0)) - F(F^{-1}(\tau))]. \end{aligned} \tag{8.8.55}$$

Recall that $g(\tilde{x}_{ij};\boldsymbol{\eta}, t_0) = \mathbf{w}_{ij}^\top(\boldsymbol{\eta} - \boldsymbol{\eta}_0)$ is the conditional quantile function under H_0

defined in Chapter 3. A Taylor series expansion on F at 0 reveals,

$$\begin{aligned}
&E[r_n(\boldsymbol{\delta})] \\
=&\sum_{i,j} z^*(t_{ij};\boldsymbol{\eta}_0+\boldsymbol{\delta})\left(f(F^{-1}(\tau))\mathbf{w}_{ij}^\top\boldsymbol{\delta}+\frac{1}{2}f'(F^{-1}(\tau))(\mathbf{w}_{ij}^\top\boldsymbol{\delta})^2+o(\|\boldsymbol{\delta}\|^2)\right) \\
=&\sum_{i,j} z^*(t_{ij};\boldsymbol{\eta}_0+\boldsymbol{\delta})\left(\frac{1}{2}f'(F^{-1}(\tau))(\mathbf{w}_{ij}^\top\boldsymbol{\delta})^2+o(\|\boldsymbol{\delta}\|^2)\right),
\end{aligned} \tag{8.8.56}$$

where the orthogonalization projection $\sum_{i,j} z^*(t_{ij};\widehat{\boldsymbol{\eta}})\mathbf{w}_{ij} = \mathbf{0}$ as in (4.2.3)(ii), and Condition A7 are used in the last step of (8.8.56).

Combining (8.8.56) and Conditions (A2)-(A3), we obtain

$$\begin{aligned}
&\sup_{\|\boldsymbol{\delta}\|\le K(\log n/n)^{1/2}} |E[r_n(\boldsymbol{\delta})]| \\
=&\sup_{\|\boldsymbol{\delta}\|\le K(\log n/n)^{1/2}} E\left|\sum_{i,j}\frac{1}{2}z^*(t_{ij};\boldsymbol{\eta}_0+\boldsymbol{\delta})f'(F^{-1}(\tau))(\mathbf{w}_{ij}^\top\boldsymbol{\delta})^2\right| \\
=&O(\log n).
\end{aligned} \tag{8.8.57}$$

Combining (8.8.54) and (8.8.57), we have

$$\sup_{\|\boldsymbol{\delta}\|\le K(\log n/n)^{1/2}} |r_n(\boldsymbol{\delta})| = o_p(\sqrt{n}), \tag{8.8.58}$$

which together with Corollary (8.7.1), (8.8.53) follows immediately. Hence the proof of Theorem 4.3.4 is complete.

8.9 Proof of Remark 2

To prove (3.3.11), we need to show the two optimizations are equivalent. Due to the uniqueness of $\boldsymbol{\theta}_0$ by Condition A0, it suffices to verify the following two sets of

subgradient conditions are equivalent:

$$
\text{(I)}: \begin{cases} \sum_{i,j} \psi(y_{ij} - g(\tilde{x}_{ij}; \boldsymbol{\theta})) & = o_p(1); \\ \sum_{i,j} \psi(y_{ij} - g(\tilde{x}_{ij}; \boldsymbol{\theta}))(t_{ij} - t) I\{t_{ij} \leq t\} & = o_p(1); \\ \sum_{i,j} \psi(y_{ij} - g(\tilde{x}_{ij}; \boldsymbol{\theta}))(t_{ij} - t) I\{t_{ij} > t\} & = o_p(1); \\ \sum_{i,j} \psi(y_{ij} - g(\tilde{x}_{ij}; \boldsymbol{\theta}))(b_1 I\{t_{ij} \leq t\} + b_2 I\{t_{ij} > t\}) & = o_p(1); \end{cases}
$$

and (8.9.59)

$$
\text{(II)}: \begin{cases} \sum_{i,j} \psi(y_{ij} - g(\tilde{x}_{ij}; \boldsymbol{\theta})) I\{t_{ij} \leq t\} & = o_p(1); \\ \sum_{i,j} \psi(y_{ij} - g(\tilde{x}_{ij}; \boldsymbol{\theta})) t_{ij} I\{t_{ij} \leq t\} & = o_p(1); \\ \sum_{i,j} \psi(y_{ij} - g(\tilde{x}_{ij}; \boldsymbol{\theta})) I\{t_{ij} > t\} & = o_p(1); \\ \sum_{i,j} \psi(y_{ij} - g(\tilde{x}_{ij}; \boldsymbol{\theta})) t_{ij} I\{t_{ij} > t\} & = o_p(1). \end{cases}
$$

First of all, $\widehat{\boldsymbol{\theta}}_n$ obtained from (3.2.3) satisfies the subgradient conditions (I). Notice that under the assumption $b_1 \neq b_2$, direct calculation shows that (I)(i) and (I)(iv) implies (II)(iii) and then (II)(i). Next the rest of the verification in this direction, i.e. (II) $\Rightarrow$ (I), follows immediately. On the other hand, the reverse direction of the verification, i.e. (I) $\Rightarrow$ (II), is trivial. Hence the proof of Remark 2 is complete.

Bibliography

Andrews, D. W. K. and Buchinsky, M. (2001). Evaluation of a three-step method for choosing the number of bootstrap repetitions, *Journal of Econometrics* URL.

Banerjee, M. and McKeague, I. W. (2007). Estimating optimal step-function approximations to instantaneous hazard rates, *Bernoulli* **13**, 279–299, URL.

Bantli, F. E. and Hallin, M. (1999). L1-estimation in linear models with heterogeneous white noise, *Statistics & Probability Letters* **45**, 305–315, URL.

Barrodale, I. and Roberts, F. D. K. (1974). Solution of an overdetermined system of equations in the l1 norm [F4], *Communications of The ACM* **17**, 319–320.

Barry, D. and Hartigan, J. (1993). A Bayesian analysis for change point problems, *Journal of the American Statitical Association* **88**, 309–319.

Bofinger, E. (1975). Estimation of a Density Function Using Order Statistics, *Australian Journal of Statistics* **17**, 1–7.

Box, G. and Tiao, G. C. (1965). A Change in Level of a Non-Stationary Time Series, *Biometrika* **52**, 181–192, URL.

Brownson, R. and Remington, P. (2002). *Communicating Public Health Information Effectively: A Guide for Practitioners*, American Public Health Association, Washington DC.

Buchinsky, M. (1994). Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression, *Econometrica* **62**, 405–58, URL.

Buchinsky, M. (1995). Estimating the asymptotic covariance matrix for quantile regression models a Monte Carlo study, *Journal of Econometrics* **68**, 303–338, URL.

Chappell, R. (1989). Fitting bent lines to data, with application to allometry, *Journal of Theoretical Biology* **138**, 235–256.

Chen, C. and Wei, Y. (2005). Computational issues for quantile regression, *Sankhya* **67**, 399–417, URL.

Chu, P.-S. and Zhao, X. (2004). Bayesian Change-Point Analysis of Tropical Cyclone Activity: The Central North Pacific Case*, *Journal of Climate* **17**, 4893–4901, URL.

Csörgo, M. and Horvath, L. (1988). Nonparametric methods for changepoint problem, *Handbook of Statistics* **7**, 403–425.

Daniels, H. (1954). Saddlepoint Approximations in Statistics, *Annals of Mathematical Statistics* **25**, 631–650, URL.

De Angelis, D., Hall, P. and Young, G. A. (1993). Analytical and Bootstrap Approximations to Estimator Distributions in L1 Regression, *Journal of the American Statistical Association* **88**, 1310–1316, URL.

de Jongh, P., de Wet, T. and van Deventer, P. (1994). Saddlepoint Approximations for the Distributions of Regression Quantiles, *South African Statistical Journal* **38**.

Deeks, S. G., Kitchen, C. M., Liu, L., Guo, H., Gascon, R., Narváez, A. B., Hunt, P., Martin, J. N., Kahn, J. O., Levy, J., McGrath, M. S. and Hecht, F. M. (2004). Immune activation set point during early HIV infection predicts subsequent CD4+ T-cell changes independent of viral load, *Blood* **104**, 942–947, URL.

Detels, R., Muñoz, A., McFarlane, G., Kingsley, L., Bargolick, J., Giorgi, J., Schrager, L., Phair, J. and the Multicenter AIDS Cohort Study inestiagators (1998). Effectiveness of Potent Antiretroviral Therapy on Time to AIDS and Death in Mean With Known HIV Infection Duration, *JAMA* **280**.

Diez, W. (1994). Critical periods in childhood for the development of obesity, *Am J Clin Nutr.* **59**.

Efron, B. and Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy, *Statistical Science* **1**, 54–75, URL.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems, *Statistics and Computing* **16**.

Feder, P. I. (1975). On Asymptotic Distribution Theory in Segmented Regression Problems– Identified Case, *The Annals of Statistics* **3**, 49–83, URL.

Fu, Y.-X. and Curnow, R. (1990). Maximum Likelihood Estimation of Multiple Change Points, *Biometrika* **77**, 563–573.

Gallant, A. and Fuller, W. A. (1973). Fitting Segmented Polynomial Regression Models Whose Joint Points Have to be Estimated, *Journal of the American Statistical Association* **68**, 144–147, URL.

Gutenbrunner, C. and Jurečková, J. (1992). Regression Rank Scores and Regression Quantiles, *The Annals of Statistics* **20**, 305–330, URL.

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

Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests*, Academic Press, New York.

Hall, P. and Sheather, S. J. (1988). On the Distribution of a Studentized Quantile, *Journal of the Royal Statistical Society. Series B (Methodological)* **50**, 381–391, URL.

Hao, L. and Naiman, D. (2007). *Quantile Regression*, Sage Publications, Thousand Oaks.

Hawkins, D. M. (1976). Point Estimation of the Parameters of Piecewise Regression Models, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **25**, 51–57, URL.

He, X. and Hu, F. (2002). Markov Chain Marginal Bootstrap, *Journal of the American Statistical Association* **97**, 783–795.

He, X. and Shi, P. (1994). Convergence rate of b-spline estimators of nonparametric conditional quantile functions, *Journal of Nonparametric Statistics* **3**, 299–308, URL.

He, X. and Shi, P. (1996). Bivariate Tensor-Product B-Splines in a Partly Linear Model, *Journal of Multivariate Analysis* **58**, 162 – 181, URL.

He, X., Zhu, Z.-Y. and Fung, W.-K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure, *Biometrika* **89**, 579–590, URL.

Hendricks, W. and Koenker, R. (1992). Hierarchical Spline Models for Conditional Quantiles and the Demand for Electricity, *Journal of the American Statistical Association* **87**, 58–68, URL.

Hinkley, D. and Schechtman, E. (1987). Conditional bootstrap methods in meanshift model, *Biometrika* **74**, 85–93.

Hinkley, D. V. (1971). Inference in Two-Phase Regression, *Journal of the American Statistical Association* **66**, 736–743, URL.

Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables, *Journal of the American Statistical Association* **58**, 13–30, URL.

Horowitz, J. L. (1998). Bootstrap Methods for Median Regression Models, *Econometrica* **66**, 1327–1351, URL.

Hudson, D. J. (1966). Fitting Segmented Curves Whose Join Points Have to be Estimated, *J. Am. Statist. Ass.* **61**, 1097–129.

Hunt, P. W., Deeks, S. G., Rodriguez, B., Valdez, H., Shade, S. B., Abrams, D. I., Kitahata, M. M., Krone, M., Neilands, T. B., Brand, R. J., Lederman, M. M. and Martin, J. N. (2003). Continued CD4 cell count increases in HIV-infected adults experiencing 4 years of viral suppression on antiretroviral therapy, *AIDS* **17**, 1907–1915, URL.

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

Johnstone, I. and Siegmund, D. (1989). On Hotelling's Formula for the Volume of Tubes and Naiman's Inequality, *The Annals of Statistics* **17**, 184–194, URL.

Jones, R. and Dey, I. (1995). Determining one or more change points, *Chemistry and Physics of Lipids* **76**, 1–6.

Jurečková, J. and Sen, P. K. (1984). On Adaptive Scale-Equivariant M-Estimators in Linear Models, *Statistics & Decisions Supplement Issue* **1**, 31–46.

Karmarker, N. (1984). A new polynomial time algorithm for linear programming, *Combinatorica* **4**, 373–395.

Kocherginsky, M., He, X. and Mu, Y. (2005). Practical confidence intervals for regression quantiles, *J. Comput. Graph. Statis.* **14**, 41–55, URL.

Koenker, R. (1996). Rank Tests for Linear Models, in *Handbook of Statistics*, (ed. by C. Rao and G. Maddala), New York: North-Holland.

Koenker, R. (2005). *Quantile Regression*, Cambridge University Press, New York.

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

Koenker, R. and Bassett, G. W. (1982). Robust Tests for Heteroscedasticity Based on Regression Quantiles, *Econometrica* **50**, 43–61, URL.

Koenker, R. and D'Orey, V. (1993). A Remark on Computing Regression Quantiles, *Applied Statistics* **36**, 383–393.

Koenker, R. and Hallock, K. F. (2001). Quantile Regression, *Journal of Economic Perspectives* **15**, 143–156, URL.

Koenker, R. and Machado, J. A. F. (1999). Goodness of Fit and Related Inference Processes for Quantile Regression, *Journal of the American Statistical Association* **94**, 1296–1310, URL.

Koenker, R. and Portnoy, S. (1987). L-Estimation for Linear Models, *Journal of the American Statistical Association* **82**, 851–857, URL.

Koenker, R. W. and D'Orey, V. (1987). Algorithm AS 229: Computing Regression Quantiles, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **36**, 383–393, URL.

Krisnaiah, P. and Miao, B. (1988). Review about estimation of change points, *Handbook of Statistics* **7**, 375–402.

Lee, T.-S. (1993). Estimating coefficients of two-phase linear regression model with autocorrelated errors, *Statistics & Probability Letters* **18**, 113 – 120, URL.

Li, C., Wei, Y., Chappell, R. and He, X. (2011). Bent Line Quantile Regression with Application to an Allometric Study of Land Mammals' Speed and Mass, *Biometrics* **67**, 242–249, URL.

Liang, K.-Y., Self, S. G. and Xinhua, L. (1990). The Cox Proportional Hazards Model with Change Point: An Epidemiologic Application, *Biometrics* **46**, URL.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22, URL.

Little, R. and Rubin, D. (2002). *Statistical analysis with missing data*, Wiley, Hoboken, NJ.

Moyeed, R. and Diggle, P. (1994). RATES OF CONVERGENCE IN SEMI-PARAMETRIC MODELLING OF LONGITUDINAL DATA, *Australian Journal of Statistics* **36**, 75–93, URL.

Page, E. S. (1954). Continuous Inspection Schemes, *Biometrika* **41**, 100–115, URL.

Park, C.-W. and Kim, W.-C. (2004). Estimation of a regression function with a sharp change point using boundary wavelets, *Statistics Probability Letters* **66**, 435–448, URL.

Park, J.-G. and Wu, H. (2006). Backfitting and local likelihood methods for nonparametric mixed-effects models with longitudinal data, *Journal of Statistical Planning and Inference* **136**, 3760–3782, URL.

Parzen, M. I., Wei, L. J. and Ying, Z. (1994). A resampling method based on pivotal estimating functions, *Biometrika* **81**, 341–350, URL.

Pawitan, Y. (2005). Change-Point problem. Published online: July 15, 2005, in *Encyclopedia of Biostatistics*, 2nd edition. P. Armitage and T. Colton (eds), John Wiley & Sons.

Pere, A. (2000). Comparison of two methods of transforming height and weight to Normality., *The Annals of Human Biology* **27**, 35–45, URL.

Piepho, H. and Ogutu, J. (2003). Inference for the Break Point in Segmented Regression with Application to Longitudinal Data, *Biometrical Journal* **45**, 591–601, URL.

Portnoy, S. (1984). Asymptotic Behavior of M-Estimators of p Regression Parameters when p2/n is Large. I. Consistency, *The Annals of Statistics* **12**, 1298–1309, URL.

Portnoy, S. and Koenker, R. (1997). The Gaussian Hare and the Laplacian Tortoise: Computability of Squared- Error versus Absolute-Error Estimators, *Statistical Science* **12**, 279–296, URL.

Powell, J. L. (1991). Estimation of Monotonic Regression Models under Quantile Restrictions, in *Nonparametric and Semiparametric Methods in Econometrics*, edited by Barnett, W., Powell, J. and Tauchen, G., Cambridge University Press, Cambridge.

Prokopec, M. and Bellisle, F. (1993). Adiposity in Czech children followed from 1

month of age to adulthood: analysis of individual BMI patterns, *The Annals of Human Biology* **20**, 517–525, URL.

Quandt, R. E. (1958). The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes, *Journal of the American Statistical Association* **53**, 873–880, URL.

Quandt, R. E. (1960). Tests of the Hypothesis that a Linear Regression System Obeys Two Separate Regimes, *Journal of the American Statistical Association* **55**, 324–330, URL.

R Development Core Team, . (2009). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL, ISBN 3-900051-07-0.

Reilly, J., Armstrong, J., Dorosty, A., Emmett, P., Ness, A., Rogers, I., Steer, C. and Sherriff, A. (2005). Early life riskfactors for obesity in childhood: cohort study, *BMJ* **330**.

Renaud, M., Katlama, C., Mallet, A., Calvez, V., Carcelain, G., Tubiana, R., Jouan, M., Caumes, E., Agut, H., Bricaire, F., Debré, P. and Autran, B. (1999). Determinants of paradoxical CD4 cell reconstitution after protease inhibitor-containing antiretroviral regimen, *AIDS* **13**.

Robins, J., Rotnitzky, A. and Zhao, P. (1995). Analysis of Semiparametric Regression Models for Repeated Outcomes in Presence of Missing Data, *Journal of American Statistical Association* **90**.

Robison, D. E. (1964). Estimates for the Points of Intersection of Two Polynomial Regressions, *Journal of the American Statistical Association* **59**, 214–224, URL.

Rolland-Cachera, M., Deheeger, M., Bellisle, F., Sempe, M., Guilloud-Bataille, M. and Patois, E. (1984). Adiposity rebound in children: a simple indicator for predicting obesity, *The American Journal of Clinical Nutrition* **39**, 129–135, URL.

Rosenfield, D., Zhou, E., Wilhelm, F. H., Conrad, A., Roth, W. T. and Meuret, A. E. (2010). Change point analysis for longitudinal physiological data: Detection of cardio-respiratory changes preceding panic attacks, *Biological Psychology* **84**, 112 – 120, URL, psychobiology of Respiration and the Airways.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*, Wiley, New York.

Schwarz, G. (1978). Estimating the Dimension of a Model, *The Annals of Statistics* **6**, 461–464, URL.

Siddiqui, M. (1960). Distribution of Quantiles from a Bivariate Population, *Journal of Researh of the Natioinal Bureau of Standards* **64**, 145–150.

Siervogel, R. M., Roche, A. F., Guo, S., Mukherjee, D. and Chumlea, W. C. (1991). Patterns of change in weight/stature2 from 2 to 18 years: findings from long-term serial data for children in the Fels longitudinal growth study., *Int J Obes.* **15**, 479–485, URL.

Sorva, R., Lankinen, S., Tolppaen, E.-M. and Perheentupa, J. (1990). Variation of growth in height and weight of children. II. After Infancy., *Acta Paediatrica Scandinavica* **3**, 498–506.

Spady, R. H. (1991). Saddlepoint Approximations for Regression Models, *Biometrika* **78**, 879–889, URL.

Srivastava, M. and Worsley, K. (1986). Likelihood ratio tests for a change in the nultivariate normal mean, *Journal of the American Statistical Association* **81**, 199–205.

Staszewski, S., Morales-Ramirez, J., Tashima, K., Rachlis, A., Skiest, D., Stanford, J., Stryker, R., Johnson, P., Labriola, D., Farina, D. and Manion, N., D.J. Ruiz (1999). Efavirenz plus zidovudine and lamivudine, efavirenz plus indinavir, and indinavir plus zidovudine and lamivudine in the treatment of HIV-1 infection in adults. Study 006 Team., *N Engl J Med.* **25**, 1865–1873.

Sterne, J., Hernán, M., Ledergerber, B., Tilling, K., Weber, R., Sendi, P., Rickenbach, M., Robins, J., Egger, M. and the Swiss HIV Cohort Study (2005). Long-term effectiveness of potent antiretroviral therapy in preventing AIDS and death: a prospective cohort study, *The Lancet* **366**, 378–84.

Tarwater, P. M., Gallant, J. E., Mellors, J. W., Gore, M. E., Phair, J. P., Detels, R., Margolick, J. B. and Muñoz, A. (2004). Prognostic value of plasma HIV RNA among highly active antiretroviral therapy users, *AIDS* **18**, 2419–23, URL.

Tarwater, P. M., Mellors, J., Gore, M. E., Margolick, J. B., Phair, J., Detels, R. and Muñoz, A. (2001). Methods to Assess Population Effectiveness of Therapies in Human Immunodeficiency Virus Incident and Prevalent Cohorts, *American Journal of Epidemiology* **154**, 675–681, URL.

Tukey, J. (1965). What part of the sample contains the information?, *Proceedings of the National Academy of Sciences* **53**, 127–134.

Wang, H. and He, X. (2007). Detecting Differential Expressions in GeneChip Microarray Studies: A Quantile Approach, *Journal of the American Statistical Association* **102**, 104–112, URL.

Wei, Y. and He, X. (2006). Conditional growth charts, *The Annal of Statistics* **34**, 2069–2097, URL.

Welsh, A. H. (1987). Kernel estimates of the sparsity function, in *Statistical data*

analysis based on the L_1 norm and related methods, (ed. by Y. Dodge), pp. 369-377, New York: Elsevier.

Whitaker, R. C., Pepe, M. S., Wright, J. A., Seidel, K. D. and Dietz, W. H. (1998). Early Adiposity Rebound and the Risk of Adult Obesity, *Pediatrics* **101**, 5, URL.

Williams, B., Korenromp, E., Gouws, E., Schmid, G., Auvert, B. and Dye, C. (2006). HIV infection, Antiretroviral Therapy, and $CD4^+$ Cell Count Distributions in African Populations, *J Infect Dis* **194**, 1450–1458.

Worsley, K. (1979). On the likelihood ratio test for a shift in location of normal populations, *Journal of the American Statistical Association* **74**, 365–367.

Wu, X. and Yang, Y. (2008). Change-point Estimates in a Longitudinal Binary Data, *Tsinghua Science and Technology* **13**, 553–559.

Curriculum Vitae

Nanshi Sha
722 West 168th Street, R-6
Department of Biostatistics
Columbia University
New York, NY 10032, USA

Voice: (212) 342-1241
Cell: (347) 604-3958
Email: ns2397@columbia.edu

BIOGRAPHICAL DATA

Year and Place of Birth	1983, Shanghai, P. R. China
Citizenship	Citizen of P. R. China

RESEARCH INTEREST

- Change-point problems in longitudinal quantile regression.
- Statistical genetics and adaptive design in clinical trials.

EDUCATION

2006-2011	Graduate studies at Department of Biostatistics, Columbia University in the city of New York. Advisor: Professor Ying Wei. M.S. degree received in October 2007, M.Phil. degree received in February 2010. Ph. D. August 2011.
2002-2006	Undergraduate studies in Department of Statistics and Finance, University of Science and Technology of China, Hefei, P. R. China. B.S. degree received in June 2006.

AWARDS

- Mailman School of Public Health, Student Travel Fund, Columbia University, 2010.
- Mailman School of Public Health, Student Travel Fund, Columbia University, 2009.

- Fellowship for Ph.D. program, Columbia University, 2007.
- Teaching Assistantship, Columbia University, 2007.
- Fellowship for M.S. accelerated predoctoral training program, Columbia University, 2006.
- Outstanding Student Scholarship, USTC, 2005.
- Outstanding Student Scholarship, USTC, 2004.
- Outstanding Student Scholarship, USTC, 2002.

EXPERIENCE

Columbia University, Department of Anesthesiology, New York Presbyterian Hospital, New York, NY
Graduate Research Assistant (January 2010 - present)

- Management of innovative peri-oprative data server (CompuRecord Anesthesia Info. System).
- Fulfillment of departmental data requests by designing and executing SQL queries.
- Provide statistical support for clinical outcome research using large observational datasets.
- Association study of pulmonary hypertension in pediatric patients.
- Evaluation of the surgical Apgar score for non-cardiac procedures.
- Establishment of dynamic bounds of blood pressure for carotiod endarterectomy procedures.
- Design and implement of an working algorithm for desaturation with bradycardia (DWB).

Sanofi-Aventis, Biostatistics and Programming Department, Bridgewater, NJ
Statistician (September 2009 - December 2009)

- Numerical evaluation of an innovative adaptive design applied to an on-going oncology trial.
- Developed a two-part testing procedure with application to a Gd-enhanced T1 lesion volume study.

Merck Research Laboratories, BARDS, Translational Medicine Statistics, Merck & Co. Inc. Rahway, NJ
Biostatistics Summer Intern (June 2009 - August 2009)

- Developed and proposed a new testing methods for pathway activation with pre-specified gene sets with application to an oncology clinical trial.

New York State Psychiatric Institute, Department of Psychiatry, Division Clinical Therapeutics, Columbia University Medical Center, New York, NY
Research Statistician Intern (June 2007 - October 2008)

- Modeling and evaluating early response to antidepressant treatment in bulimia nervosa.

PUBLICATIONS

1. Wan, S., **Sha, N.**, Wong, P., Puig, O. (2010) An Orthogonal Transformed Aggregation Statistic to Test Pathway Activation with Pre-specified Gene Sets. Submitted.

2. Sonty, N., **Sha, N.**, Wald, E. (2011) Acceptance and depression mediate the relationship between pain intensity and pain-related disability. In preparation.

3. Sobel, J., **Sha, N.**, Wunsch, H., Li, G. (2011) Predicting ICU Admission using Surgical Apgar Score. In preparation.

4. Sysko, R., **Sha, N.**, Wang, Y., Duan, N., Walsh, T. (2008) Early Response to Antidepressant Treatment in Bulimia Nervosa. *Psychological Medicine*. Cambridge University Press 15 Sep 2009.

5. Wang, Y., **Sha, N.**, Fang, Y. (2008) Analysis of Genome-Wide Association Data by Large-Scale Bayesian Logistic Regression. *BMC Proceedings* 2009. 3(Suppl 7):S16.

6. Fang, Y., Wang, Y., **Sha, N.** (2008) Armitage's Trend Test for Genomewide Association Analysis: One-sided or Two-sided? *BMC Proceedings* 2009. 3(Suppl 7):S37.

PROFESSIONAL ACTIVITIES

1. Speaker for *Genetic Analysis Workshop 16*, St. Louis, MO, September 2008.

2. Speaker for *Joint Statistical Meetings* 2009, Washington, DC, August 2009.
3. Reviewer for *The International Journal of Biostatistics*, June 2009.
4. Speaker for *ENAR Spring Meetings* 2010, New Orleans, LA, March 2010.
5. Speaker for *Joint Statistical Meetings* 2010, Vancouver, Canada, August 2010.

TECHNICAL SKILLS

- Statistical Packages: SAS, R, S-Plus, STATA, SPSS.
- Mathematical Software: Mathematica.
- Languages: C/C++, SQL.
- Applications: LaTeX, Microsoft SQL Studio, Microsoft SQL Server Management, Windows ODBC database, spreadsheet and presentation software.
- Operating Systems: Unix/Linux, Mac OS X, NT platforms.