Estimation of Change-Points in Spline Models

by

Guangyu Yang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2022

Doctoral Committee:

Professor Min Zhang, Chair
Professor Bhramar Mukherjee
Professor Marie S. O'Neill
Professor Peter X.K. Song

Guangyu Yang

yguangyu@umich.edu

ORCID iD: 0000-0002-4692-1361

ACKNOWLEDGEMENTS

I still vividly remember the day I first came to Ann Arbor, and now I am about to finish all my graduate studies and embark on a new journey in my life. *"Time flies over us, but leaves its shadow behind."* Looking back on six years in the Department of Biostatistics at the University of Michigan, I would like to express my sincere thanks to all people who have accompanied and helped me along the road. My appreciation to you for all you have done for me.

First, I am sincerely grateful to my advisor, who is also my academic role model, Dr. Min Zhang, for her invaluable encouragement and patient guidance in my graduate study, my research, and my life in the United States. I couldn't have a better advisor than her. Throughout my Ph.D. studies, Dr. Zhang inspired my research idea, guided me to develop theoretical biostatistics methodology research topics and methods, and provided valuable opportunities to collaborate with physicians and work with different medical datasets. I would like to extend my sincere thanks to all other members of my dissertation committee, Dr. Bhramar Mukherjee, Dr. Peter X.K. Song, and Dr. Marie S. O'Neill, for their guidance and assistance in my research, life, and job finding in academia.

I would like to thank all my collaborators who belonged to the Michigan Congestive Heart Failure Investigators and the Department of Pediatrics at the University of Michigan. And I am so proud to be a member of these wonderful research teams. During these collaboration experiences, I have learned a lot about medical knowledge

in pediatric respiratory disease, infectious disease, cardiovascular disease, and cardiac surgery. It was these valuable collaborative experiences that made me genuinely aware that biostatistics is not just about theorems and formulas in textbooks but has irreplaceable importance in helping and improving biomedical research and public health. And this pushed me to decide to continue working on biostatistics research after graduation. I especially thank Dr. Donald S. Likosky and Dr. Mary K. Dahmer, who encouraged and helped me a lot with my collaborative research and supported my job finding in academia.

To my friends and family members, thank you for your love and support. I am particularly grateful to my parents, Junfu Yang and Shuchun Zhang, for giving me birth and supporting and loving me unconditionally over the past 28 years. I wouldn't be here today without you.

Finally, I would like to thank all faculties and staff in the Department of Biostatistics at the University of Michigan. Over the previous six years, I learned, grew, and made friends in this warmly big family, the Department of Biostatistics, and now I consider it as my second home. I will pass on what I have learned from here to my future students: rigorous research, caring for students, and pursuit of academic excellence.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

In this dissertation thesis, we present novel, rigorously studied and computationally efficient methods for change-points estimation in different spline models, including linear spline models, generalized linear spline models and constrained spline models.

In Chapter II, we estimate change-points in linear spline models. In this chapter, we study influence functions of regular and asymptotically linear estimators using semiparametric theory. Based on the theoretical development, we propose a novel and simple method to circumvent the nondifferentiability, the key challenge in linear spline models, using the modified derivative idea. Consistency and asymptotic normality are rigorously derived for the proposed estimator. A two-step semismooth Newton-Raphson algorithm is further developed for the proposed method. Simulation studies have shown that the proposed method performs well in terms of both statistical and computational properties and improves over existing methods. For example, the existing smoothing-based method sometimes only has a 60% convergence rate and is sensitive to the initial value of the algorithm. And estimates from the highly cited R package "segmented" sometimes exhibit large outliers and may even have a bimodal distribution with around 99% of the coverage probability. In comparison, our proposed method is more stable in terms of almost 100% convergence rates, more robust to choices of different initial values, and has better coverage probabilities.

In Chapter III, we extend the estimation of change-points from linear spline models to generalized linear spline models. In this chapter, to overcome the nondifferentiability, we follow the idea of modified derivative from which we propose a novel method to

estimate change-points as well as other unknown parameters in generalized linear spline models. Furthermore, we improved the two-step semismooth Newton-Raphson algorithm so that this algorithm is applicable for the proposed method in generalized linear spline models. The statistical properties (consistency, asymptotic normality, and asymptotic efficiency) of the proposed estimator are rigorously studied. Based on simulation studies, the statistical and computational properties for the proposed method performs well.

In Chapter IV and Chapter V, we aim to estimate the threshold in constrained spline models, which assume no effect between the factor of interest and the outcome under or above the unknown threshold according to clinical knowledge. In Chapter IV, using a constrained linear spline model, we estimate the threshold of nadir oxygen delivery level, below which there is an increased risk of postoperative acute kidney injury, during a cardiac surgery. Our proposed method is built upon Chapter III. Through simulation studies, we have shown that the proposed method is more robust and efficient than existing methods. In Chapter V, we extend the constrained linear spline model to the constrained penalized spline model, which is able to account for a flexible pattern after the threshold instead of assuming a linear pattern as in the constrained linear spline model. Using the study of Pregnancy Research on Inflammation, Nutrition, & City Environment: Systematic Analyses, we explore the threshold of exposure to air pollution above which there is an adverse effect in terms of low birth weight for pregnant women.

CHAPTER I

Introduction

The estimation of change-points, also referred to as knots, breakpoints, transition points and thresholds, has widespread and important applications in many fields. For example, in cardiac surgery, it is believed that there is a threshold of nadir oxygen delivery level below which patients are subject to higher risk of postoperative acute kidney injury. But till now, there is no widely accepted threshold in practice and the recommended thresholds in literature are based on ad-hoc data analyses (*Ranucci et al.*, 2005; *De Somer et al.*, 2011). In many applications, estimating and making inferences on the change-points may be of primary interest. However, due to the lack of both rigorously studied and computational stable method for change-points estimation, researchers often have to prespecified change-points based on available prior subject-matter knowledge or chosen change-points in an ad-hoc manner based on exploratory data analyses.

Change-points estimation is a challenging statistical problem due to that the underlying spline model is nondifferentiable. Therefore, traditional methods, such as maximum likelihood estimation, can not be directly applied to estimate change-points. Current available methods can be classified into three broad categories, which are search-based methods, smoothing-based methods and ad-hoc algorithms. The search-based method is intuitive and it is the first solution that has been studied

for the change-points estimation problem in spline models. Early contributions to search-based methods date back to sixty years ago with plenty of studied since then. *Hudson* (1966) proposed a method for finding change-points in linear spline models based on an overall least-squares idea and the consistency and asymptotic properties of this method was later shown by *Feder* (1975a,b). Although the search-based method was proposed around 1960s, existing rigorous studied search-based methods focus only on linear spline models and thus not suitable for use with general (e.g., binary or count) outcomes. Also, in general, search-based methods are computational inefficient, especially for large sample sizes or when multiple change-points exist.

Smoothing, a nature idea for dealing with the nondifferentiability, is another main direction to estimate change-points in spline models. The most recent rigorously studied smoothing-based method in the linear spline model is *Das et al.* (2016), which has demonstrated computational and/or statistical advantages over previous search-based methods. However, due to the involvement of smoothing tuning parameters, the method of *Das et al.* (2016) is sensitive to initial values of change-points. And the extension of smoothing-based methods from linear spline models to generalized linear spline models has not been successful so far. In general, although smoothing-based methods are rigorously studied, these methods also focus only on linear spline models and may fall into computational issues.

A popular ad-hoc algorithm for change-points estimation in generalized linear spline models was developed by *Muggeo* (2003) via linearization technique, implemented in a highly-cited R package (*Muggeo*, 2008). Other estimation strategies and computation algorithms are also studied in the literature in recent years, including Bayesian methods (*Chen et al.*, 2011), algorithms implemented in R packages (*Fong et al.*, 2017) and semismooth methods studied in optimization literature (*Cui et al.*, 2018). However, theoretical and asymptotic properties of these methods were not studied.

To summarize, existing methods are either (1) computationally intensive and slow, (2) sensitive to initial values, or (3) based on ad-hoc algorithms without rigorous study of statistical inferences. A formal and rigorously studied method, equipped with well-developed and computationally efficient algorithm, for estimating and making inference on change-points is greatly needed. In Chapter II and Chapter III, we propose rigorously studied and computationally fast and stable methods for estimation of change-points in linear spline models and generalized linear spline models, respectively. In Chapter IV, we aim to estimate the threshold in constrained linear spline models and constrained penalized spline models, respectively. With the method and theory built upon Chapter III, we estimate the threshold of nadir oxygen delivery level, below which there is an increased risk of postoperative acute kidney injury, during a cardiac surgery in Chapter IV. In Chapter V, we explore the threshold of exposure to air pollution above which there is an adverse effect in terms of low birth weight using the study of Pregnancy Research on Inflammation, Nutrition, & City Environment: Systematic Analyses.

CHAPTER II

Estimation of Knots in Linear Spline Models

2.1 Introduction

Linear regression is by far the most popular model for modeling the relationship of a continuous outcome with other factors. Much of the success and popularity is due to its convenient computation and easy interpretation. In practice, however, the assumption of linearity is often violated. Various ways can be used to model nonlinear relationships including, for example, polynomial regression, parametric nonlinear models, and kernel and spline-based nonparametric methods. These methods are useful for the purpose of prediction or covariate adjustment. However, if the main purpose is to understand and make inference on the relationship between a factor of interest and outcome, these methods such as polynomial regression and nonparametric modeling are often limited by the lack of easy interpretation.

The linear spline model is a piecewise linear model where linear segments are joined at different knots (*Marsh and Cormier*, 2001). While accommodating the overall nonlinear trend, it also allows for an easy interpretation as the linear regression within each segment. That is, within each segment, the strength of association between a factor of interest and the outcome is quantified by the corresponding slope. In the literature, linear spline models are also referred to as segmented models and broken-

line/stick models (*Muggeo*, 2003). Knots, where the slope of linear trend changes, are also referred to as change-points, break-points, transition points, thresholds (*Rigotti*, 2009). In practice, locations of knots are often prespecified based on available prior subject-matter knowledge or chosen in an ad hoc manner based on exploratory data analysis. When locations of knots as known, fitting a linear spline model is essentially the same as fitting the usual linear regression. However, although researchers may have an idea on the overall trend, they may lack the knowledge on the exact location of knots and rely on statisticians to place the knots. Moreover, rather than coefficients/slopes of linear spline models, the locations of knots are the main quantifies of interest in many applications. For example, in cardiac surgery, it is believed that there is a threshold of nadir oxygen delivery level below which patients are subject to higher risk of postoperative acute kidney injury. But till now, there is no widely accepted threshold in practice and the recommended thresholds in literature are based on ad hoc data analyses (*Ranucci et al.*, 2005; *De Somer et al.*, 2011). To summarize, relying on ad hoc ways to place knots is unsatisfying, particularly when locations of knots are of main interest. A formal and rigorously studied method for estimating and making inference on locations of knots is greatly needed.

In a recent article, *Das et al.* (2016) studied knots estimation and gave a nice review on previous work in linear spline models. We refer readers to *Das et al.* (2016) for a detailed review on relevant previous work. The estimation of knots in linear spline models was initially introduced by *Quandt* (1958). *Hudson* (1966) proposed a method for finding knots based on an overall least-squares idea and the consistency and asymptotic properties of this method was later shown by *Feder* (1975b,a). As the search-based method of *Hudson* (1966) was developed during the precomputer era, the search time increased greatly with growing sample sizes, making it a slow method even with a moderate sample size (*Das et al.*, 2016). A direct dynamic programming procedure was developed by *Bellman and Roth* (1969), but this method was still slow

and even slower than *Hudson* (1966). *Muggeo* (2003) developed an estimating strategy and a highly-cited software for estimating unknown knots (*Muggeo*, 2008). However, the theoretical and asymptotic properties of the method were not studied. Bayesian methods for knot estimation were also proposed in the literature, including *Bacon and Watts* (1971), *Carlin et al.* (1992), *Smith and Cook* (1980); see *Chen et al.* (2011) for a detailed description and comparison. A main difficulty in the estimation of knots is that the likelihood function is not differentiable when knots are unknown. Several methods are developed based on the idea of smoothing. *Tishler and Zang* (1981) first proposed to use smoothing to address the nondiffentiability issue and suggested using a quadratic approximation to provide a smoothed version of the likelihood. Inspired by quadratic smoothing, *Chiu et al.* (2002, 2006) proposed the "bent-cable model". They used a quadratic smoothing to approximate the linear spline model in a neighborhood around the knot, with size of the neighborhood treated as an unknown parameter. The recent work of *Das et al.* (2016) was also based on the idea of smoothing and the asymptotic properties of the method was studied rigorously. Also demonstrated by *Das et al.* (2016) in their simulations, their method was much more time-efficient than the early method of *Hudson* (1966).

Unlike smoothing-based methods, we propose a novel and simple method to circumvent the challenge of nondifferentiability. We show that, although not differentiable everywhere in the usual sense, linear spline models are differentiable in quadratic mean. For both parametric and semiparametric linear spline models, we study the influence functions of regular and asymptotically linear estimators in Section 2.3. We derive the efficient influence function via the geometry of Hilbert spaces. By redefining partial derivatives using the concept of differentiable in quadratic mean, we address the nondifferentiability issue and propose a computationally easy, fast and stable method for the estimation of knots as well as other unknown parameters in Section 2.4.1. A two-step estimation algorithm is proposed to facilitate computational

stability. The asymptotic properties of the proposed method are rigorously studied using the empirical process theory in Section 2.4.2. In Section 2.5, we evaluate the proposed method and compare it with the method of *Das et al.* (2016) and the R package "segmented" (*Muggeo*, 2008) through extensive simulation studies. Simulations show that the proposed method performs well in finite samples in both single- and multiple-knot scenarios and is computational fast and stable, making it suitable for use in practice.

2.2 Model and Background

Suppose data are collected on n subjects. For subject i, let Y_i denote the response, X_i denote the factor of interest, and $\boldsymbol{Z}_i$ denote an L-dimensional vector of covariates. The data is summarized as $\boldsymbol{W}_i = \{Y_i, X_i, \boldsymbol{Z}_i\}$, $i = 1, \ldots, n$, independent and identically distributed across subject i. We consider a linear spline model for Y, where the mean effect of X on Y is modeled using a linear spline model with K knots. Specifically, the model is written as

$$Y_i = \mu(\boldsymbol{X}_i^*; \boldsymbol{\theta}) + \epsilon_i = \beta_0 + \beta_1 X_i + \sum_{k=1}^{K} \beta_{1k}(X_i - \tau_k)^+ + \boldsymbol{\eta}^T \boldsymbol{Z}_i + \epsilon_i, \qquad (2.1)$$

where $\boldsymbol{X}_i^* = (X_i, \boldsymbol{Z}_i)$, $E(\epsilon_i | \boldsymbol{X}_i^*) = 0$, and

$$(X_i - \tau_k)^+ = \begin{cases} 0 & \text{if } X_i \leq \tau_k, \\ X_i - \tau_k & \text{if } X_i > \tau_k. \end{cases}$$

The conditional variance is denoted as $E(\epsilon_i^2 | \boldsymbol{X}_i^*) = \sigma^2(\boldsymbol{X}_i^*)$, where $\sigma^2(\boldsymbol{X}_i^*)$ is a function of X_i and $\boldsymbol{Z}_i$. In model (2.1), K is a prespecified number of knots and τ_k is the location of the kth knot ($k = 1, \ldots, K$) that is assumed to be unknown and needs to be estimated. Without loss of generality, we assume that τ_k, $k = 1, \ldots, K$, are

ordered and distinct to ensure model identifiability. That is, for $i < j$, $\tau_i < \tau_j$. We assume that the factor of interest X_i has a bounded domain, denoted as $X_i \in [C_1, C_2]$, and all knots should be within this bounded interval as well. According to this model, the effect of X_i on Y_i is piecewise linear where the slopes in different segments are different. Effects of $\boldsymbol{Z}_i$ are also modeled as linear with coefficients $\boldsymbol{\eta}$. For convenience, we define $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_{11}, ..., \beta_{1K})^T$, $\boldsymbol{\tau} = (\tau_1, ..., \tau_K)^T$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\tau}^T, \boldsymbol{\eta}^T)^T$, where $\boldsymbol{\theta}$ is assumed to belong to a compact set $\boldsymbol{\Theta} \in \mathcal{R}^q$ where $q = 2K + 2 + L$. The true value of $\boldsymbol{\theta}$ is denoted as $\boldsymbol{\theta}^0$, assumed to be an interior point of the compact set $\boldsymbol{\Theta}$.

The key challenge in fitting model (2.1) when knots are unknown is that $(x - \tau_k)^+$ is not differentiable with respect to τ_k. One nature and well-studied method for handling such a challenge is smoothing. A smoothing-based method for estimating knots was rigorously studied by *Das et al.* (2016), which has demonstrated computational and/or statistical advantages over previous methods. We briefly describe the method here. Based on the idea of local smoothing in a shrinking neighborhood $(\tau_k - \varrho_n, \tau_k + \varrho_n)$ around each knot τ_k $(k = 1, \ldots, K)$, they used the bent-cable model (*Chiu et al.*, 2002, 2006) as a smoothing working model to approximate model (2.1), that is, replacing all $(x-\tau_k)^+$ with $q_n(x, \tau_k)$, defined as 0 if $x < \tau_k - \varrho_n$, $\frac{x-\tau_k+\varrho_n}{4\varrho_n}$ if $\tau_k - \varrho_n \leq x \leq \tau_k + \varrho_n$ and $x - \tau_k$ if $x > \tau_k$. With increasing sample sizes, ϱ_n approaches to zero and $q_n(x, \tau_k)$ approaches to $(x - \tau_k)^+$. Treating the quadratic smoothing model as the working model, with the working model being smooth, estimators for unknown parameters can be obtained directly by minimizing the least squares via the Newton-Raphson algorithm. *Das et al.* (2016) showed that this method will lead to $\sqrt{n}$-consistent and asymptotically normal estimators when $\varrho_n = n^{-\alpha}$ for $\alpha > \frac{1}{2}$. One common choose for ϱ_n is $\frac{1}{n}$. Although their estimator has nice statistical properties based on asymptotic theory and certain computational advantages over other existing methods, our simulation studies indicate that it is sensitive to the choice of initial values and not computationally stable, hence not suitable for use in practice.

While smoothing is a natural idea to handle nondifferentiability, we show that it is unnecessarily complicated for this particular problem. As a result, it leads to complexity in computation and unsatisfactory performances, as our simulation studies demonstrate. Alternatively, we propose a novel and simple method to circumvent this challenge by redefining derivatives. This idea is motivated and justified by differentiability in quadratic mean of linear spline models, which we will show later. We study influence functions of linear spline models and derive the parametric and semiparametric efficiency bound via the geometry of influence functions in the next section.

2.3 Influence Functions of Linear Spline Model

2.3.1 Parametric Linear Spline Model

We first consider a parametric linear spline model where the conditional distribution of ϵ given $\boldsymbol{X}^*$ is modeled parametrically. The conditional distribution is denoted as $p_{\epsilon|\boldsymbol{X}^*}(\epsilon|\boldsymbol{x}^*;\boldsymbol{\gamma}_1)$, where $\boldsymbol{\gamma}_1$ is an r_1-dimensional parameter and $p_{\epsilon|\boldsymbol{X}^*}$ is a known function. For example, usually one would assume $p_{\epsilon|\boldsymbol{X}^*}(\epsilon|\boldsymbol{x}^*;\gamma_1)$ is normal with mean zero and constant variance $\gamma_1 = \sigma^2$. Because there is a one-to-one transformation between $(y, \boldsymbol{x}^*)$ and $(\epsilon, \boldsymbol{x}^*)$, it is easy to see that the joint distribution of $(Y, \boldsymbol{X}^*)$ can be written as $p_{Y,\boldsymbol{X}^*}(y, \boldsymbol{x}^*;\boldsymbol{\zeta}) = p_{\epsilon|\boldsymbol{X}^*}\{y - \mu(\boldsymbol{x}^*;\boldsymbol{\theta})|\boldsymbol{x}^*;\boldsymbol{\gamma}_1\}p_{\boldsymbol{X}^*}(\boldsymbol{x}^*;\boldsymbol{\gamma}_2)$, where $p_{\boldsymbol{X}^*}(\boldsymbol{x}^*;\boldsymbol{\gamma}_2)$ denotes the density of $\boldsymbol{X}^*$, and $\boldsymbol{\zeta} = (\boldsymbol{\theta}^T, \boldsymbol{\gamma}^T)^T$ with $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \boldsymbol{\gamma}_2^T)^T$. The dimension of $\boldsymbol{\theta}, \boldsymbol{\gamma}$, and $\boldsymbol{\zeta}$ are q, r and p, respectively. The truth is denoted as $\boldsymbol{\zeta}^0 = (\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)$. In Lemma 1 we show that, although the model $p(\boldsymbol{w};\boldsymbol{\zeta})$ is not differentiable everywhere, it satisfies the condition of differentiable in quadratic mean (DQM).

Lemma 1. *The parametric linear spline model $p(\boldsymbol{w};\boldsymbol{\zeta})$ is differentiable in quadratic mean (DQM) with respect to $\boldsymbol{\zeta}$ if elements $I_{\boldsymbol{\zeta}}(\boldsymbol{W};\boldsymbol{\zeta})$ are well-defined, where $I_{\boldsymbol{\zeta}}(\boldsymbol{W};\boldsymbol{\zeta}) = E[S_{\boldsymbol{\zeta}}(\boldsymbol{W};\boldsymbol{\zeta})S_{\boldsymbol{\zeta}}^T(\boldsymbol{W};\boldsymbol{\zeta})]$, $S_{\boldsymbol{\zeta}}(\boldsymbol{W};\boldsymbol{\zeta}) = \{S_{\boldsymbol{\theta}}^T(\boldsymbol{W};\boldsymbol{\zeta}), S_{\boldsymbol{\gamma}}^T(\boldsymbol{W};\boldsymbol{\zeta})\}^T$, $S_{\boldsymbol{\gamma}}(\boldsymbol{W};\boldsymbol{\zeta}) = \frac{\partial \log\{p(\boldsymbol{W};\boldsymbol{\zeta})\}}{\partial \boldsymbol{\gamma}}$,*

and

$$
\begin{aligned}
S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}) &= \frac{\partial \log\{p(\boldsymbol{W};\boldsymbol{\zeta})\}}{\partial \mu(\boldsymbol{X}^*;\boldsymbol{\theta})} H^T(\boldsymbol{X}^*;\boldsymbol{\theta}) \ \textit{with} \\
H(\boldsymbol{X}^*;\boldsymbol{\theta}) &= (1, X, (X-\tau_1)^+, \ldots, (X-\tau_K)^+, \\
&\quad -\beta_{11} I(X>\tau_1), \ldots, -\beta_{1K} I(X>\tau_K), \boldsymbol{Z})^{1\times q}.
\end{aligned}
$$

Proofs for Lemma 1 as well as other lemmas and theorems presented in the Appendix. According to the definition of DQM in *van der Vaart* (2000), $S_{\boldsymbol{\zeta}}(\boldsymbol{W};\boldsymbol{\zeta})$ and $I_{\boldsymbol{\zeta}}(\boldsymbol{W};\boldsymbol{\zeta})$ defined in Lemma 1 are the score function and Fisher information matrix for $\boldsymbol{\zeta}$ respectively. Specifically, $S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta})$ is the score function of $\boldsymbol{\theta}$. Note in the usual case when $\mu(\boldsymbol{X}^*;\boldsymbol{\theta})$ is differentiable with respect to $\boldsymbol{\theta}$, the score function for $\boldsymbol{\theta}$ would be $\frac{\partial \log\{p(\boldsymbol{W};\boldsymbol{\zeta})\}}{\partial \mu(\boldsymbol{X}^*;\boldsymbol{\theta})}$ times the partial derivative of $\mu(\boldsymbol{X}^*;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. In this sense, $H(\boldsymbol{X}^*;\boldsymbol{\theta})$ defined in Lemma 1 plays the same role as the partial derivative of $\mu(\boldsymbol{X}^*;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ in the usual differentiable case and can be interpreted as a modified partial derivative. In particular, the usual derivative of $(x-\tau_k)^+$ with respect to τ_k is $-I(x>\tau_k)$ if $\tau \neq x$ and undefined if $\tau_k = x$. By the definition of $H(\boldsymbol{X}^*;\boldsymbol{\theta})$, intuitively it is can be viewed that the derivative of $(x-\tau_k)^+$ is redefined as 0 when $\tau_k = x$, that is, the modified partial derivative for $(x-\tau_k)^+$ becomes

$$
\frac{\partial (x-\tau_k)^+}{\partial \tau_k} = -I(x>\tau_k). \tag{2.2}
$$

Similarly, the modified derivative of $I(x>\tau_k)$ with respect to τ_k is defined as 0. As we will see later, this informal interpretation provides a motivation for the proposed method we present in Section 2.4. As an asymptotically linear estimator has a unique influence function almost surely, in Lemma 2 below, we describe all influence functions of regular and asymptotically linear (RAL) estimators for $\boldsymbol{\theta}$.

Lemma 2. *For any parametric model* $\mathcal{P} = \left\{p\{\boldsymbol{w};\boldsymbol{\theta},\boldsymbol{\gamma}\} : \boldsymbol{w} = (y,x,\boldsymbol{z})\right\}$ *with a non-singular Fisher information matrix* $I_{\boldsymbol{\zeta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)$. *Let* $\widehat{\boldsymbol{\theta}}_n$ *be an asymptotically linear estimator with influence function* $\varphi(\boldsymbol{W}) \in \mathcal{H}$, *where* $\mathcal{H}$ *denotes a p-dimensional Hilbert space, such that* $E_{\boldsymbol{\zeta}}\{\varphi(\boldsymbol{W})\}$ *and* $E_{\boldsymbol{\zeta}}\{\varphi(\boldsymbol{W})^T\varphi(\boldsymbol{W})\}$ *exist and are continuous in* $\boldsymbol{\zeta}$ *in a neighborhood of* $\boldsymbol{\zeta}^0$. *Then, if* $\widehat{\boldsymbol{\theta}}_n$ *is regular, this will imply that*

$$\begin{aligned} E\{\varphi(\boldsymbol{W})S^T_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)\} &= I^{q\times q} \text{ and} \\ E\{\varphi(\boldsymbol{W})S^T_{\boldsymbol{\gamma}}(\boldsymbol{W};\boldsymbol{\zeta}^0)\} &= 0^{q\times r}. \end{aligned} \tag{2.3}$$

Furthermore, according to the construction procedure in Chapter 3.3 of *Tsiatis* (2007), any element in the Hilbert space $\mathcal{H}$ satisfying equation (2.3) is the influence function of some RAL estimator. One can find the best estimator (i.e. the one with the smallest asymptotic variance) via the geometry of influence functions. According to Example 25.15 in *van der Vaart* (2000), DQM implies that the tangent space of the parametric model is exactly given by the linear space spanned by the score function of $\boldsymbol{\zeta}$, denoted as $\mathcal{T} = \{B^{q\times p}S_{\boldsymbol{\zeta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)$, for all $q \times p$ matrices $B\}$. Then by Theorem 3.4 of *Tsiatis* (2007), the set of all influence functions, that is, satisfying condition (2.3) in Lemma 2, is the linear variety $\varphi^*(\boldsymbol{W}) + \mathcal{T}^{\perp}$, where $\varphi^*(\boldsymbol{W})$ is any influence function and $\mathcal{T}^{\perp}$ is the space perpendicular to the tangent space $\mathcal{T}$. It is straightforward to show that $\mathcal{T}$ can be written as the direct sum of $\mathcal{T}_{\boldsymbol{\theta}}$ and Λ, where $\Lambda = \{B^{q\times\gamma}S_{\boldsymbol{\gamma}}(\boldsymbol{W};\boldsymbol{\zeta}^0)$, for all $q \times r$ matrices $B\}$ is the nuisance tangent space, and $\mathcal{T}_{\boldsymbol{\theta}} = \{B^{q\times q}S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)$, for all $q \times q$ matrices $B\}$ is the tangent space spanned by $S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)$. By Theorem 3.5 and Corollary 2 in *Tsiatis* (2007), the efficient influence function $\varphi_{eff}(\boldsymbol{W})$ with a non-singular information matrix is

$$\begin{aligned} \varphi_{eff}(\boldsymbol{W}) &= \Gamma I^{-1}_{\boldsymbol{\zeta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0) \\ &= \left[E\{S_{eff}(\boldsymbol{W};\boldsymbol{\zeta}^0)S^T_{eff}(\boldsymbol{W};\boldsymbol{\zeta}^0)\}\right]^{-1}S_{eff}(\boldsymbol{W};\boldsymbol{\zeta}^0), \end{aligned}$$

where $\Gamma = (I^{q\times q}, 0^{q\times r})$, $S_{eff}(\boldsymbol{W};\boldsymbol{\zeta}^0) = S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0) - \Pi[S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)|\Lambda]$ is the efficient score, and $\Pi[h|\Lambda] = E\{hS_{\boldsymbol{\gamma}}^T(\boldsymbol{W};\boldsymbol{\zeta}^0)\}[E\{S_{\boldsymbol{\gamma}}(\boldsymbol{W};\boldsymbol{\zeta}^0)S_{\boldsymbol{\gamma}}^T(\boldsymbol{W};\boldsymbol{\zeta}^0)\}]^{-1}S_{\boldsymbol{\gamma}}^T(\boldsymbol{W};\boldsymbol{\zeta}^0)$ is the projection of h onto space Λ. Therefore, the parametric efficiency bound is $[E\{S_{eff}(\boldsymbol{W};\boldsymbol{\zeta}^0)S_{eff}^T(\boldsymbol{W};\boldsymbol{\zeta}^0)\}]^{-1}$. For example, if $p_{\epsilon|\boldsymbol{X}^*}(\epsilon|\boldsymbol{x}^*;\gamma_1)$ follows a normal distribution with constant variance $\gamma_1 = \sigma^2$, the efficient score is $S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0) = \frac{1}{\sigma^2}\{Y - \mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}H^T(\boldsymbol{X}^*;\boldsymbol{\theta}^0) = S_{eff}(\boldsymbol{W};\boldsymbol{\zeta}^0)$ and the parametric efficiency bound is $\sigma^2V^{-1}(\boldsymbol{\theta}^0)$, where $V(\boldsymbol{\theta}^0) = E[H^T(\boldsymbol{X}^*;\boldsymbol{\theta}^0)H(\boldsymbol{X}^*;\boldsymbol{\theta}^0)]$.

2.3.2 Semiparametric Restricted Moment Linear Spline Model

This section considers a semiparametric restricted moment linear spline model where the conditional distribution of ϵ given $\boldsymbol{X}^*$ in model (2.1) is unspecified. We denote the class of all such densities for a single observation $\boldsymbol{W} = (Y, \boldsymbol{X}^*)$ as

$$\mathcal{P} = \Big\{p(\boldsymbol{w};\boldsymbol{\theta},\boldsymbol{\gamma}), \boldsymbol{w} = (y, x, \boldsymbol{z})\Big\},$$

where $\boldsymbol{\theta}$ is the parameter of interest and $\boldsymbol{\gamma}$ is an infinite-dimensional nuisance parameter. Similarly as in the parametric setting, we can express the density as

$$\begin{aligned} p_{Y,\boldsymbol{X}^*}(y,\boldsymbol{x}^*) &= p_{\epsilon,\boldsymbol{x}^*}\{y - \mu(\boldsymbol{x}^*;\boldsymbol{\theta}),\boldsymbol{x}^*\} \\ &= \delta_1\{y - \mu(\boldsymbol{x}^*;\boldsymbol{\theta}),\boldsymbol{x}^*\}\delta_2(\boldsymbol{x}^*), \end{aligned}$$

where $\delta_1(\epsilon,\boldsymbol{x}^*) = p_{\epsilon|\boldsymbol{X}^*}(\epsilon|\boldsymbol{x}^*)$ and $\delta_2(\boldsymbol{x}^*) = p_{\boldsymbol{X}^*}(\boldsymbol{x}^*)$ are nonnegative functions with the following constraints, that is, for all $\boldsymbol{x}^*$, $\int \delta_1(\epsilon,\boldsymbol{x}^*)d\epsilon = 1$, $\int \epsilon\delta_1(\epsilon,\boldsymbol{x}^*)d\epsilon = 0$ and $\int \delta_2(\boldsymbol{x}^*)dv(\boldsymbol{x}^*) = 1$, where $v(\boldsymbol{x}^*)$ is the dominating measure. The set of functions $\delta_1(\epsilon,\boldsymbol{x}^*)$ and $\delta_2(\boldsymbol{x}^*)$, satisfying the above constraints, are infinite-dimensional and can be used to characterize the semiparametric model. According to similar procedures in Chapter 4 of *Tsiatis* (2007), we can extend the arguments for parametric models in Section 2.3.1 to semiparametric models.

Lemma 3. *The nuisance tangent space Λ for the semiparametric restricted moment linear spline model $\mathcal{P}$ is given by*

$$\Lambda = \Big\{ h^{q\times 1}(\epsilon, \boldsymbol{X}^*),\ \textit{such that } E\{h(\epsilon, \boldsymbol{X}^*)\epsilon | \boldsymbol{X}^*\} = 0^{q\times 1} \Big\}.$$

And the space orthogonal to the nuisance tangent space $\Lambda^{\perp}$ is given by

$$\Lambda^{\perp} = \Big\{ A^{q\times 1}(\boldsymbol{X}^*)\epsilon,\ \textit{for all } q\textit{-dimensional functions } A^{q\times 1}(\boldsymbol{X}^*) \Big\}.$$

Moreover, the semiparametric efficiency bound is given by

$$\mathcal{V} = \Big[E\{S_{eff}(\boldsymbol{W}; \boldsymbol{\zeta}^0) S_{eff}^T(\boldsymbol{W}; \boldsymbol{\zeta}^0)\} \Big]^{-1} = \Big[E\{ \frac{H^T(\boldsymbol{X}^*; \boldsymbol{\theta}^0) H(\boldsymbol{X}^*; \boldsymbol{\theta}^0)}{\sigma^2(\boldsymbol{X}^*)} \} \Big]^{-1},$$

where the efficient score is $S_{eff}(\boldsymbol{W}; \boldsymbol{\zeta}^0) = \{\sigma^2(\boldsymbol{X}^*)\}^{-1} H^T(\boldsymbol{X}^*; \boldsymbol{\theta}^0)\epsilon$.

By Lemma 3 above, combined with Theorem 4.2 in *Tsiatis* (2007), we can construct all influence functions of RAL estimators for the semiparametric linear spline model $\mathcal{P}$. If start with any $q \times 1$ matrix $A(\boldsymbol{X}^*)$, the influence function can be constructed as $\varphi(\boldsymbol{W}) = CA(\boldsymbol{X}^*)\{Y - \mu(\boldsymbol{X}^*; \boldsymbol{\theta}^0)\}$, where $C = \Big[E\{A(\boldsymbol{X}^*)\{Y - \mu(\boldsymbol{X}^*; \boldsymbol{\theta}^0)\} S_{\boldsymbol{\theta}}^T(\boldsymbol{W}; \boldsymbol{\zeta}^0)\} \Big]^{-1}$ is a $q \times q$ normalization constant matrix. Furthermore, the optimal estimator can be obtained by solving the following estimating equation:

$$\sum_{i=1}^{n} \frac{H^T(\boldsymbol{X}_i^*; \boldsymbol{\theta}^0)}{\sigma^2(\boldsymbol{X}_i^*)} \{Y - \mu(\boldsymbol{X}_i^*; \boldsymbol{\theta})\} = 0.$$

For example, when the conditional variance is a constant, that is, $E(\epsilon^2 | \boldsymbol{X}^*) = \sigma^2$, the efficient score is $S_{eff}(\boldsymbol{W}; \boldsymbol{\zeta}^0) = \sigma^{-2} H^T(\boldsymbol{X}^*; \boldsymbol{\theta}^0)\{Y - \mu(\boldsymbol{X}^*; \boldsymbol{\theta}^0)\}$, the semiparametric local efficiency bound is $\mathcal{V} = \sigma^2 \Big[E\{H^T(\boldsymbol{X}^*; \boldsymbol{\theta}^0) H(\boldsymbol{X}^*; \boldsymbol{\theta}^0)\} \Big]^{-1}$ and the optimal

estimator is the solution to the following estimating equation:

$$\sum_{i=1}^{n} H^T(\boldsymbol{X}_i^*;\boldsymbol{\theta}^0)\{Y-\mu(\boldsymbol{X}_i^*;\boldsymbol{\theta})\}=0. \tag{2.4}$$

2.4 Proposed Method

2.4.1 Estimation Algorithm

In this section, based on theoretical results developed in the previous section, we propose a conceptually simple and computationally easy method to estimate knots, as well as other unknown parameters, without using smoothing techniques or tuning parameters. As we demonstrate in our simulation studies, our method leads to a substantial improvement in performance relative to smoothing-based methods. We study and justify asymptotic properties of the proposed estimator rigorously in Section 2.4.2. To simplify notations, we use P to denote the marginal law of observations and $\mathbb{P}_n$ to denote the empirical distribution following the notations in *van der Vaart* (2000). Specifically, $P(f)=Ef(X)=\int f dP$, $\mathbb{P}_n(f)=\frac{1}{n}\sum_{i=1}^n f(X_i)$, and $\mathbb{G}_n(f)$ is the empirical process $\mathbb{G}_n(f)=\sqrt{n}\{\mathbb{P}_n(f)-P(f)\}=\sqrt{n}\{\frac{1}{n}\sum_{i=1}^n f(X_i)-\int f dP\}$.

Suppose locations of knots are known, the ordinary least squares (OLS) method that minimizes the sum of squared residuals is the standard way to fit a linear spline model. Specifically, the sum of squared residuals for model (2.1) is $\mathbb{P}_n(M(\boldsymbol{\theta}))=\frac{1}{n}\sum_{i=1}^n\{y_i-\mu(\boldsymbol{x}_i^*;\boldsymbol{\theta})\}^2$, where $M(\boldsymbol{\theta})=\{y-\mu(\boldsymbol{x}^*;\boldsymbol{\theta})\}^2$. Due to the existence of nondifferentiable terms $(x-\tau_k)^+$, $k=1,\ldots,K$, the function $M(\boldsymbol{\theta})$ is not differentiable with respect to τ_k when $\tau_k=x$. As shown in Lemma 1, linear spline models are differentiable in quadratic mean with respect to $\boldsymbol{\theta}$ and the score function is essentially the usual score function with a modification in the definition of partial derivatives. Motivated by this, with the modified partial derivative defined in equation (2.2), $\mathbb{P}_n(M(\boldsymbol{\theta}))$

can also be treated as a differentiable function. We then propose to proceed as usual, that is, estimating the unknown parameters by solving the estimating equation, $\mathbb{P}_n(Q(\boldsymbol{\theta})) = 0$, where $Q(\boldsymbol{\theta})$ is the modified derivative of $M(\boldsymbol{\theta})$ with respect of $\boldsymbol{\theta}$ as in formula (2.2). Specifically, $Q(\boldsymbol{\theta}) = -2H^T(\boldsymbol{x}^*;\boldsymbol{\theta})\{y - \mu(\boldsymbol{x}^*;\boldsymbol{\theta})\}$. Then the proposed estimator is solution to the following estimating equation:

$$\mathbb{P}_n(Q(\boldsymbol{\theta})) = -\frac{2}{n}\sum_{i=1}^{n} H^T(\boldsymbol{X}_i^*;\boldsymbol{\theta})\{Y - \mu(\boldsymbol{X}_i^*;\boldsymbol{\theta})\} = 0, \tag{2.5}$$

which is equivalent to the estimating equation (2.4) in Section 2.3.2. We denote the solution as $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\beta}}_n^T, \widehat{\boldsymbol{\tau}}_n^T, \widehat{\boldsymbol{\eta}}_n^T)^T$ and we may omit the subscript n when there is no need to emphasize the dependence on n.

The Newton-Raphson (NR) algorithm is a popular method for solving estimating equations and it is also used in the smoothing-based method of *Das et al.* (2016) for estimation of knots. However, for our proposed estimating equation (2.5), the NR algorithm is not applicable because the estimating function $\mathbb{P}_n(Q(\boldsymbol{\theta}))$ is not differentiable in the usual sense. We note that $\mathbb{P}_n(Q(\boldsymbol{\theta}))$ is differentiable everywhere except at a finite number of points, that is, when $x = \tau_k$. Following the same modified derivative idea in formula (2.2) used for showing DQM and for motivating the estimating equation (2.5), we may define a modified derivative for $\mathbb{P}_n(Q(\boldsymbol{\theta}))$ as well, that is, for places where derivatives are not normally defined we redefine it as zero. With these modifications, NR algorithm can be applied. This modified NR algorithm can be rigorously justified by showing that the estimating function $\mathbb{P}_n(Q(\boldsymbol{\theta}))$ is semismooth and the modified derivative we propose is a generalized Jacobian (*Qi and Sun*, 1993). Also a semismooth NR method was studied in *Cui et al.* (2018) for solving general majorization-minimization problems, with an application in continuous piecewise affine regression functions. However, *Cui et al.* (2018) did not study statistical properties.

Our study found that this modified NR algorithm is still not numerically stable. One reason is that the estimation of $\boldsymbol{\tau}$ is entangled with the estimation of other parameters and this algorithm involves a modified Hessian matrix with a relatively large dimension. In fact, all challenges in terms of nondifferentiability discussed previously are only related to estimation of $\boldsymbol{\tau}$. If $\boldsymbol{\tau}$ is known, then model (2.1) becomes a usual linear regression model for which estimators of unknown parameters can easily be obtained via the analytic solution. In consideration of this, we propose a two-step algorithm. In each iteration, we separate updates of $(\boldsymbol{\beta}, \boldsymbol{\eta})$ and $\boldsymbol{\tau}$ into two steps and in each step the update is based on the most recent estimate of the other parameters. In particular the two-step algorithm allows one to take advantage of the analytic solution of the least-square estimator when $\boldsymbol{\tau}$ is known. The extended NR type procedure is only used for updating the estimate of $\boldsymbol{\tau}$, leading to substantial improvement in terms of numerical stability and practical performances.

Denoting the initial value of $\boldsymbol{\tau}$ by $\widehat{\boldsymbol{\tau}}^{(0)}$, the two-step algorithm iterates between updating estimates of other parameters and $\boldsymbol{\tau}$. And the t-th ($t \geq 1$) iteration of the proposed algorithm proceeds as follows.

Step 1. In the tth iteration, update estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ to obtain $\widehat{\boldsymbol{\beta}}^{(t-1)}$, $\widehat{\boldsymbol{\eta}}^{(t-1)}$. Specifically, treating $\widehat{\boldsymbol{\tau}}^{(t-1)}$ as fixed, fit the following linear regression model

$$E(Y|X, Z) = \beta_0 + \beta_1 X + \sum_{k=1}^{K} \beta_{1k}(X - \widehat{\tau}_k^{(t-1)})^+ + \boldsymbol{\eta}^T \boldsymbol{Z},$$

by the OLS method to obtain estimates $\widehat{\boldsymbol{\beta}}^{(t-1)}$, $\widehat{\boldsymbol{\eta}}^{(t-1)}$ and the predicted values $\widehat{\mu}_i^{(t-1)}$, $i = 1, \cdots, n$, from the fitted model.

Step 2. Update $\widehat{\boldsymbol{\tau}}^{(t-1)}$ to obtain $\widehat{\boldsymbol{\tau}}^{(t)}$ by an extended NR type procedure as follows.

Define a $K \times 1$ matrix $U^{(t)}$, where the ℓ-th row of $U^{(t)}$ is

$$U_\ell^{(t)} = \frac{\widehat{\beta}_{1\ell}^{(t-1)}}{n} \sum_{i=1}^{n} (Y_i - \widehat{\mu}_i^{(t-1)}) I(X_i > \widehat{\tau}_\ell^{(t-1)}), \quad \ell \in \{1, \ldots, K\}.$$

Also define a $K \times K$ matrix $J^{(t)}$, where the (ℓ, h)th element of $J^{(t)}$ is

$$J_{\ell,h}^{(t)} = \frac{\widehat{\beta}_{1\ell}^{(t-1)} \widehat{\beta}_{1h}^{(t-1)}}{n} \sum_{i=1}^{n} I(X_i > \widehat{\tau}_\ell^{(t-1)}) I(X_i > \widehat{\tau}_h^{(t-1)}), \quad \ell, h \in \{1, \ldots, K\}.$$

Specifically, $J^{(t)}$ and $U^{(t)}$ are proportional to the modified first-order and second-order derivative (Hessian) of $\mathbb{P}_n(M(\boldsymbol{\theta}))$ with respect to $\boldsymbol{\tau}$, respectively, with $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ fixed at the recent value. Then, based on the NR type procedure, we update $\widehat{\boldsymbol{\tau}}^{(t)} = \widehat{\boldsymbol{\tau}}^{(t-1)} - \{J^{(t)}\}^{-1} U^{(t)}$.

Starting with $t = 1$, the proposed algorithm iterates between Step 1 and Step 2 until the convergence of $\boldsymbol{\tau}$, that is, $\|\widehat{\boldsymbol{\tau}}^{(t)} - \widehat{\boldsymbol{\tau}}^{(t-1)}\| < \xi$, where ξ is a prespecified convergence tolerance value. Then, the final estimator of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are obtained by the OLS method, treating the estimate of $\boldsymbol{\tau}$ as fixed, as in Step 1 of the algorithm. We show later that the two-step algorithm converges and it indeed solves the proposed estimating equation $\mathbb{P}_n(Q(\boldsymbol{\theta})) = 0$.

The Step 2 of the proposed algorithm for updating $\boldsymbol{\tau}$ is an NR-type algorithm, which involves the Hessian matrix. A popular alternative to the NR method is a gradient descent method, where one replaces $\{J^{(t)}\}^{-1}$ by a step size. We expect that, when converged, this two-step gradient descent type algorithm would lead to an estimator with the same asymptotic properties studied in Section 2.4.2. This is because by a similar argument in Appendix, the converged $\boldsymbol{\tau}$ would satisfy $\lim_{t \to \infty} U^{(t)} = 0$ and the converged result would solve the proposed estimating equation. Usually, one needs to use the objective function to be minimized to determine the step size, for example, using backtracking line search. A reasonable objective function to use in determine

the step size is the sum of squares $\mathbb{P}_n(M(\boldsymbol{\theta}))$, as the soluation of $\mathbb{P}_n(Q(\boldsymbol{\theta})) = 0$ is (or approximately is) a local minima of $\mathbb{P}_n(M(\boldsymbol{\theta}))$. In Appendix, we provide a brief justification and some preliminary simulation studies, which show that the gradient descent method leads to estimators with similar statistical properties as those from the NR-type algorithm. Below, we focus on the NR-type algorithm as it is known to have faster convergence rates.

2.4.2 Asymptotic Properties

This section studies asymptotic properties of the proposed estimator. The main results are summarized in the following proposition and two theorems, with proofs available in the appendix. The proof for Proposition 1 makes use of properties of locally Lipschitz continuous functions and semismooth functions as studied in *Qi and Sun* (1993). The proofs for two theorems make heavy use of the empirical process theory studied in *van der Vaart* (2000) and *van der Vaart and Wellner* (1996).

Proposition 1. *The proposed two-step algorithm converges locally and it converges to* $\widehat{\boldsymbol{\theta}}_n$*, which is the solution to the estimating equation (2.5), that is,* $\mathbb{P}_n(Q(\boldsymbol{\theta})) = 0$.

Theorem 1. *Under the semiparametric linear spline model (2.1),* $\widehat{\boldsymbol{\theta}}_n$ *is a consistent estimator for* $\boldsymbol{\theta}^0$*, as* $n \to \infty$.

Theorem 2. *Under the semiparametric linear spline model (2.1),* $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0)$ *converges in distribution to a normal distribution* $\mathcal{N}(\mathbf{0}, V^{-1}(\boldsymbol{\theta}^0)I(\boldsymbol{\theta}^0)V^{-1}(\boldsymbol{\theta}^0))$*, where*

$$\begin{aligned} V(\boldsymbol{\theta}^0) &= P\{H^T(\boldsymbol{X}^*;\boldsymbol{\theta}^0)H(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}, \\ I(\boldsymbol{\theta}^0) &= P\{\sigma^2(\boldsymbol{X}^*)H^T(\boldsymbol{X}^*;\boldsymbol{\theta}^0)H(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}. \end{aligned}$$

The asymptotical variance is of the familiar sandwich variance form, as in the usual

restricted moment model studied in *Tsiatis* (2007). Inference on $\boldsymbol{\theta}$ based on $\widehat{\boldsymbol{\theta}}_n$ can be conducted in the usual way based on Theorem 2. As an example, we provide an explicit variance estimator for the most common setting when $E(\epsilon^2|\boldsymbol{X}^*) = \sigma^2$. Then, it is easy to see that $I(\boldsymbol{\theta}) = \sigma^2 V(\boldsymbol{\theta})$. For constructing confidence intervals and conducting Wald-based inference, one can estimate the asymptotic variance $\sigma^2 V^{-1}(\boldsymbol{\theta}^0)$ by $\widehat{\sigma}^2\widehat{V}^{-1}(\widehat{\boldsymbol{\theta}}_n)$, where $\widehat{V}(\widehat{\boldsymbol{\theta}}_n) = \mathbb{P}_n(H^T(\widehat{\boldsymbol{\theta}}_n)H(\widehat{\boldsymbol{\theta}}_n))$ and $\widehat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \widehat{\mu}_i)^2}{n-q}$, with q being the number of total parameters in model (2.1). The variance of $\widehat{\boldsymbol{\theta}}_n$ can then be consistently estimated by $\widehat{\sigma}^2\widehat{V}^{-1}(\widehat{\boldsymbol{\theta}}_n)/n$. By discussions in Section 2.3.2, when $E(\epsilon^2|\boldsymbol{X}^*) = \sigma^2$, $\widehat{\boldsymbol{\theta}}_n$ achieves the semiparametric local efficiency bound $\sigma^2 V^{-1}(\boldsymbol{\theta}^0)$.

2.5 Simulation Studies

We conducted several Monte Carlo simulation studies to evaluate the proposed method and compared its performance with the method of *Das et al.* (2016) and the popular R package "segmented" (*Muggeo*, 2008). Data were generated from settings where the number of knots was one, two, or four. Settings with one or two knots were the same as *Das et al.* (2016) except that our settings have a covariate in the model. To further illustrate performances of the proposed method in the presence of multiple knots, we additionally considered a linear spline model with four knots. In each setting with $K = 1, 2$, or 4, outcomes were generated under three different data generating scenarios according to model (2.1) and methods were evaluated under sample sizes $n = 200, 500, 1000$, and 2500. In all scenarios, there was a single covariate Z, which was generated from a normal distribution with mean 0 and standard deviation 2 and the corresponding coefficient in the outcome model was set as $\eta^0 = 0.5$. The factor of interest X was generated as $\Phi((V + Z)/\sqrt{5})$, where Φ is the cumulative distribution function of the standard normal distribution and V is standard normal and independent of all other variables. As a result, X followed a uniform (0,1) distribution and was correlated with Z, making Z a confounder for the effect of X on Y.

The error term ϵ was generated from a normal distribution with mean 0 and standard deviation $\sigma = 0.03$. In Table 2.1, we listed the true parameter values, $\boldsymbol{\beta}^0$ and $\boldsymbol{\tau}^0$, for each setup, with the first number in each setup indicating the number of knots.

Table 2.1: True values of parameters in data generating models and initial values of knots.

	$\boldsymbol{\beta}^0$	$\boldsymbol{\tau}^0$	$\widehat{\boldsymbol{\tau}}^{(0)}$
Setup 1.1	$(0.2, 1, 1)^T$	0.6	0.5
Setup 1.2	$(0.3, 1.5, 1)^T$	0.8	0.5
Setup 1.3	$(0.3, 1.5, -1)^T$	0.2	0.5
Setup 2.1	$(0.3, 1, 1, 1)^T$	$(0.2, 0.8)^T$	$(0.1, 0.5)^T$
Setup 2.2	$(0.2, 1, 2, 1)^T$	$(0.4, 0.6)^T$	$(0.2, 0.8)^T$
Setup 2.3	$(0.3, 1, -1, 1)^T$	$(0.2, 0.8)^T$	$(0.1, 0.5)^T$
Setup 4.1	$(0.3, 1, -2, 4, -5, 3)^T$	$(0.2, 0.4, 0.6, 0.8)^T$	$(0.1, 0.3, 0.5, 0.9)^T$
Setup 4.2	$(5, -1, 3, 5, -9, 6)^T$	$(0.3, 0.6, 0.8, 0.9)^T$	$(0.2, 0.5, 0.85, 0.95)^T$
Setup 4.3	$(9, -5, 6, 7, -8, -3)^T$	$(0.1, 0.3, 0.5, 0.7)^T$	$(0.05, 0.45, 0.55, 0.8)^T$

Also listed in Table 2.1, $\widehat{\boldsymbol{\tau}}^{(0)}$ is the initial value specified for $\boldsymbol{\tau}$ in our proposed algorithm and in the R package "segmented" for all setups, and in the method of *Das et al.* (2016) for setups 1.1, 1.2 and 1.3. For reasons that will be explained later, we also implemented the method of *Das et al.* (2016) using the true value $\boldsymbol{\tau}^0$ as the initial value. The method of *Das et al.* (2016) is denoted as "Das et al" when the initial value was set as $\widehat{\boldsymbol{\tau}}^{(0)}$ and as "Das et al$^{\#}$" when the initial value was $\boldsymbol{\tau}^0$. In the method of *Das et al.* (2016), r_n was set as $\frac{1}{n}$, consistent with their simulation studies. In all methods, the tolerance of convergence ξ was set as 10^{-6} and the largest number of iterations was 1000. For both the proposed method and the method of *Das et al.* (2016), when the number of iterations has reached 1000, we relaxed ξ to 10^{-3}. In scenarios of one knot, results were based on $10,000$ Monte Carlo replicates. In all other scenarios, results were based on 1000 Monte Carlo replicates.

Tables 2.2-2.4 contain results for one, two and four-knots scenarios respectively. For

all scenarios, bias, Monte Carlo standard deviation (MCSD), the average of standard errors (AVESE), coverage probabilities (CP) of the 95% confidence intervals (CI) and convergence rates (CR) of the algorithm for estimated knots are reported. For the method of *Das et al.* (2016), in Table 2.2 and Table 2.3, we additionally report the Winsorized bias (WBias), Winsorized standard deviation (WSD) and the Winsorized average of standard error (WSE), where winsorization applies to the top and bottom 5% of the data.

We note that the method of *Das et al.* (2016) seems to have a computational issue. The convergence rate is low even for one-knot scenarios, with convergence rates consistently lower than 50% when the chosen initial values are different from the true values. Due to the computational problem, the Monte Caro standard deviation of the estimator is large without winsorization. With winsorization, the performance of the method of *Das et al.* (2016) in terms of statistical properties is reasonable. When setting the initial value of knots as the true value in the method of *Das et al.* (2016), which is unrealistic in practice, the convergence rates greatly improve and the algorithm converges more than or appropriately 95% of the time. In terms of statistical properties (Bias, MCSD, AVESE and CP), "Das et al$^{\#}$" performs quite similar to the proposed method. The convergence issue of *Das et al.* (2016) is more severe for scenarios with multiple knots. Therefore, in Table 2.3 where $K = 2$, we only report results of the method of *Das et al.* (2016) using the true value as the initial value, that is, "Das et al$^{\#}$". In Table 2.4 where $K = 4$, we only report results for the proposed method and "segmented" method as the method of *Das et al.* (2016) has failed to produce reasonable results even using true values of knots as initial values. In Table 2.3, with true values as the initial values, again the performance of *Das et al.* (2016) is reasonable, although not ideal. Overall, we found that the method of *Das et al.* (2016) is sensitive to choices of initial values, consistent with findings by the original authors through personal communications. Regarding the "segmented" method,

Table 2.2: Simulation results based on $10,000$ Monte Carlo data sets for $K = 1$, where "Das et al", "segmented" and "proposed" denote the method of *Das et al.* (2016), the method in the R package "segmented" and the proposed method respectively, n denotes the sample size, * indicates value $\times 10^{-3}$ and $^{\#}$ indicates the initial value of knots is set at the true value.

Methods	Bias* *(WBias*)*	MCSD* *(WSD*)*	AVESE* *(WSE*)*	CP%	CR%	Bias* *(WBias*)*	MCSD* *(WSD*)*	AVESE* *(WSE*)*	CP%	CR%
					Setup 1.1					
	n=200					**n=500**				
Das et al$^{\#}$	-0.10	10.95	14.04	94.2	98.1	-0.02	5.79	5.61	94.0	98.2
	(-0.04)	*(8.31)*	*(8.98)*			*(-0.03)*	*(5.21)*	*(5.61)*		
Das et al	-0.94	15.26	15.68	93.8	58.0	-0.52	14.39	5.67	93.8	59.3
	(-0.79)	*(8.46)*	*(8.99)*			*(-0.23)*	*(5.28)*	*(5.61)*		
segmented	0.18	10.36	8.92	90.6	100.0	0.05	5.78	5.60	95.4	100.0
proposed	-0.58	9.48	8.98	93.4	100.0	-0.21	5.89	5.61	93.7	100.0
	n=1000					**n=2500**				
Das et al$^{\#}$	0.06	5.65	1.09E+04	94.7	98.3	-0.03	2.55	2.49	94.4	98.6
	(0.03)	*(3.59)*	*(3.95)*			*(-0.03)*	*(2.31)*	*(2.49)*		
Das et al	-1.08	24.52	3.98	94.0	59.3	-1.20	25.87	2.49	94.2	59.7
	(-0.10)	*(3.68)*	*(3.95)*			*(-0.07)*	*(2.32)*	*(2.49)*		
segmented	0.03	3.97	3.94	95.1	100.0	-0.01	2.51	2.48	93.7	100.0
proposed	-0.08	4.05	3.95	94.4	100.0	-0.07	2.56	2.49	94.3	100.0
					Setup 1.2					
	n=200					**n=500**				
Das et al$^{\#}$	-0.45	17.90	11.32	94.3	91.5	-0.24	13.54	8.57	94.5	93.3
	(-0.18)	*(9.76)*	*(11.00)*			*(-0.06)*	*(6.12)*	*(6.83)*		
Das et al	-83.39	230.73	3.49E+11	84.5	12.8	-31.42	153.34	1.26E+10	89.3	13.4
	(-82.66)	*(227.55)*	*(15.48)*			*(-1.15)*	*(7.56)*	*(6.92)*		
segmented	0.16	11.95	10.88	92.7	100.0	-1.08	7.68	6.87	93.0	100.0
proposed	-1.75	12.97	11.02	91.0	98.4	-0.54	7.32	6.84	93.3	100.0
	n=1000					**n=2500**				
Das et al$^{\#}$	-0.40	14.98	4.87	94.3	93.7	-0.01	3.05	3.03	94.4	94.5
	(-0.13)	*(4.38)*	*(4.80)*			*(-0.01)*	*(2.74)*	*(3.03)*		
Das et al	-18.96	119.51	29.68	90.5	13.9	-22.30	130.86	8.63	90.9	14.1
	(-0.91)	*(5.09)*	*(4.82)*			*(-0.45)*	*(3.05)*	*(3.03)*		
segmented	-0.19	4.88	4.80	93.7	100.0	-0.26	3.16	3.02	94.4	100.0
proposed	-0.35	5.06	4.80	93.6	100.0	-0.09	3.12	3.03	94.1	100.0
					Setup 1.3					
	n=200					**n=500**				
Das et al$^{\#}$	0.60	21.26	6.93E+04	94.0	92.7	0.31	13.51	5.99E+04	94.1	94.3
	(0.18)	*(9.90)*	*(11.01)*			*(0.15)*	*(6.20)*	*(6.83)*		
Das et al	57.76	197.92	1.69E+12	87.4	14.4	24.96	138.38	1.71E+10	91.1	14.1
	(56.91)	*(189.41)*	*(13.57)*			*(0.96)*	*(7.27)*	*(6.89)*		
segmented	0.10	12.72	10.85	89.0	100.0	0.49	7.35	6.77	90.5	100.0
proposed	1.75	13.11	11.03	90.8	100.0	0.64	7.35	6.83	93.0	100.0
	n=1000					**n=2500**				
Das et al$^{\#}$	-0.03	5.31	4.80	94.0	94.4	-0.07	3.66	3.02	94.6	95.6
	(-0.02)	*(4.41)*	*(4.80)*			*(-0.05)*	*(2.76)*	*(3.02)*		
Das et al	10.89	100.30	1.41E+11	92.9	14.2	7.16	95.67	8.20E+05	91.3	14.3
	(0.33)	*(4.69)*	*(4.82)*			*(0.13)*	*(3.03)*	*(3.03)*		
segmented	0.19	5.23	4.80	92.2	100.0	0.05	2.89	3.03	95.7	100.0
proposed	0.22	5.08	4.80	93.4	100.0	0.05	3.10	3.03	94.3	100.0

Table 2.3: Simulation results based on 1000 Monte Carlo data setsfor $K = 2$, where "Das et al$^{\#}$" denotes the method of *Das et al.* (2016) with the initial value of knots set at the true value, "segmented" and "proposed" denote the method in the R package "segmented" and the proposed method respectively, n denotes the sample size and * indicates value $\times 10^{-3}$.

Methods	Sample Size	Bias* *(Wbias*)*	MCSD* *(WSD*)*	AVESE* *(WSE*)*	CP%	Bias* *(Wbias*)*	MCSD* *(WSD*)*	AVESE* *(WSE*)*	CP%	CR%
		$\widehat{\tau_1}$				$\widehat{\tau_2}$				
Setup 2.1										
Das et al$^{\#}$	**200**	-0.59	11.37	2.22E+04	93.8	-0.58	22.31	1.34E+04	94.9	83.6
		(-0.65)	*(9.62)*	*(11.13)*		*(-0.05)*	*(9.94)*	*(11.15)*		
segmented		0.88	13.16	11.32	90.3	-0.07	12.52	11.21	92.2	100.0
proposed		-3.09	14.98	11.46	90.3	-2.75	20.05	11.38	91.3	100.0
Das et al$^{\#}$	**500**	0.16	15.12	8.26	93.0	-1.23	26.87	7.64	95.2	83.1
		(-0.03)	*(6.75)*	*(7.04)*		*(-0.22)*	*(6.46)*	*(7.04)*		
segmented		-0.07	7.78	6.99	91.6	-0.95	7.99	7.05	91.8	100.0
proposed		-0.56	7.99	7.06	91.0	-0.74	7.59	7.04	93.6	100.0
Das et al$^{\#}$	**1000**	1.32	26.62	41.58	94.6	0.03	8.82	25.17	96.5	83.8
		(0.09)	*(4.42)*	*(4.94)*		*(-0.09)*	*(4.41)*	*(4.95)*		
segmented		-0.31	5.20	4.93	92.5	-0.34	5.17	4.93	94.0	100.0
proposed		-0.36	5.12	4.94	93.3	-0.48	5.04	4.96	95.4	100.0
Das et.al$^{\#}$	**2500**	-0.04	3.3	3.23	96.6	-0.51	18.95	11.31	94.1	87.8
		(-0.08)	*(2.66)*	*(3.11)*		*(0.13)*	*(2.81)*	*(3.11)*		
segmented		-0.99	0.88	2.99	99.9	-2.44	3.08	3.23	99.8	100.0
proposed		-0.23	3.03	3.11	95.5	0.01	3.17	3.11	93.5	100.0
Setup 2.2										
Das et al$^{\#}$	**200**	-0.44	7.13	69.04	91.8	0.07	19.96	57.72	94.1	86.7
		(-0.60)	*(5.72)*	*(5.96)*		*(0.27)*	*(10.37)*	*(11.80)*		
segmented		-0.16	6.37	6.05	94.0	0.78	14.21	12.03	90.7	100.0
proposed		-1.87	7.49	6.16	89.8	-2.62	15.14	11.93	87.2	91.9
Das et al$^{\#}$	**500**	0.21	5.42	1.10E+04	93.0	-0.28	18.26	1.75E+10	94.1	88.3
		(-0.03)	*(3.62)*	*(3.72)*		*(-0.06)*	*(6.86)*	*(7.51)*		
segmented		0.01	3.79	3.74	94.6	-0.04	8.44	7.48	92.4	100.0
proposed		-0.38	4.13	3.76	92.2	-0.90	8.3	7.48	92.1	99.0
Das et al$^{\#}$	**1000**	0.37	4.76	1454.41	96.2	-1.02	13.41	3.50E+04	93.4	89.3
		(0.09)	*(2.32)*	*(2.61)*		*(-0.06)*	*(4.63)*	*(5.24)*		
segmented		-0.01	2.84	2.62	94.0	0.17	5.71	5.23	93.6	100.0
proposed		-0.04	2.62	2.62	95.7	-0.27	5.49	5.24	92.6	99.9
Das et al$^{\#}$	**2500**	0.19	3.98	1655.64	96.4	-0.92	11.60	4.57E+06	95.2	91.1
		(-0.07)	*(1.47)*	*(1.65)*		*(-0.13)*	*(2.94)*	*(3.30)*		
segmented		-1.32	1.05	1.68	99.9	-0.80	0.70	3.23	99.8	100.0
proposed		-0.16	1.64	1.66	94.9	-0.29	3.42	3.30	93.7	100.0
Setup 2.3										
Das et al$^{\#}$	**200**	2.98	39.80	252.80	94.8	0.16	12.03	40.87	94.6	81.3
		(1.00)	(9.51)	(11.19)		(0.02)	(10.25)	(11.12)		
segmented		0.09	12.16	11.32	93.0	-0.25	13.13	11.22	91.9	100.0
proposed		-0.91	12.57	11.46	91.8	1.36	12.53	11.38	91.7	94.4
Das et al$^{\#}$	**500**	0.26	7.33	2785.05	93.4	-1.86	31.04	2971.20	95.6	83.3
		(0.27)	*(6.42)*	*(7.04)*		*(-0.44)*	*(6.32)*	*(7.02)*		
segmented		0.10	7.53	6.99	92.7	-0.18	7.87	7.02	91.2	100.0
proposed		-0.44	7.94	7.04	91.5	0.34	7.55	7.04	93.5	99.8
Das et al$^{\#}$	**1000**	1.56	31.12	5.56	94.5	-0.64	22.90	66.09	96.5	85.0
		(0.16)	*(4.39)*	*(4.93)*		*(-0.12)*	*(4.52)*	*(4.96)*		
segmented		0.25	5.24	4.93	93.1	-0.29	5.04	4.94	94.1	100.0
proposed		-0.23	5.09	4.94	94.0	0.05	4.98	4.96	95.0	100.0
Das et al$^{\#}$	**2500**	0.77	21.64	679.19	96.1	-0.55	21.89	717.50	93.2	88.3
		(0.16)	*(2.63)*	*(3.12)*		*(0.16)*	*(2.83)*	*(3.11)*		
segmented		0.11	2.97	3.11	95.7	-0.18	3.34	3.11	94.4	100.0
proposed		-0.04	2.94	3.11	96.1	0.20	3.17	3.11	93.3	100.0

Table 2.4: Simulation results based on 1000 Monte Carlo data sets for $K = 4$, where "segmented" and "proposed" denote the method in the R package "segmented" and the proposed method respectively, n denotes the sample size and * indicates value $\times 10^{-3}$.

Methods	n	Bias*	MCSD*	AVESE*	CP%	Bias*	MCSD*	AVESE*	CP%	Bias*	MCSD*	AVESE*	CP%	Bias*	MCSD*	AVESE*	CP%	CR%
Setup 4.1																		
		$\hat{\tau}_1$				$\hat{\tau}_2$				$\hat{\tau}_3$				$\hat{\tau}_4$				
segmented	200	5.24	41.89	7.70	86.9	4.39	32.83	3.77	89.9	1.88	17.42	3.01	90.6	-0.40	12.41	5.07	88.5	100
proposed		-0.50	7.84	7.07	92.0	0.13	3.71	3.52	93.7	-0.28	3.07	2.81	93.2	0.50	5.20	4.69	92.2	98.8
segmented	500	1.96	21.99	4.32	91.1	0.57	26.04	2.22	93.3	0.94	8.30	1.98	94.2	-0.11	4.29	2.99	93.8	100
proposed		-0.03	4.56	4.32	93.3	0.06	2.29	2.16	93.0	-0.11	1.85	1.73	92.6	0.19	3.05	2.89	94.3	100
segmented	1000	0.12	3.13	3.03	94.0	0.02	1.57	1.51	94.3	-0.04	1.23	1.21	94.2	-0.10	2.11	2.01	93.3	100
proposed		-0.21	3.22	3.03	92.9	0.13	1.52	1.51	95.2	-0.04	1.18	1.21	94.7	-0.02	2.13	2.03	93.4	100
segmented	2500	-0.21	2.58	2.02	94.4	0.76	9.88	0.99	93.3	0.36	5.42	0.77	93.7	-0.03	1.37	1.28	90.3	100
proposed		-0.05	1.94	1.91	94.4	-0.01	0.96	0.95	94.3	0.01	0.76	0.76	94.3	0.08	1.27	1.27	95.5	100
Setup 4.2																		
		$\hat{\tau}_1$				$\hat{\tau}_2$				$\hat{\tau}_3$				$\hat{\tau}_4$				
segmented	200	-0.32	8.95	3.88	96.3	-0.70	15.47	2.62	94.5	-0.71	11.40	2.00	92.5	-0.89	11.50	3.86	92.1	98.4
proposed		-0.17	3.88	3.82	94.7	-0.13	2.74	2.55	93.7	0.45	3.15	2.04	90.6	-1.18	5.81	3.51	89.1	99.8
segmented	500	0.05	2.42	2.33	94.5	-0.07	1.59	1.56	94.1	0.02	1.19	1.18	93.8	-0.12	2.16	2.05	92.5	100
proposed		0.08	2.52	2.35	93.4	-0.04	1.65	1.58	93.5	0.07	1.26	1.19	93.8	-0.21	2.16	2.06	92.9	100
segmented	1000	0.05	1.65	1.64	95.7	0.05	1.07	1.10	94.7	0.01	0.82	0.82	95.0	-0.01	1.50	1.43	93.4	100
proposed		0.03	1.74	1.65	93.7	-0.07	1.08	1.10	95.6	0.08	0.80	0.83	95.8	-0.05	1.51	1.44	92.8	100
segmented	2500	-0.14	1.03	1.04	94.4	0.00	0.76	0.69	92.3	0.03	0.55	0.52	93.4	0.05	0.89	0.89	95.8	100
proposed		-0.07	1.06	1.04	94.0	-0.05	0.70	0.70	94.9	0.03	0.53	0.52	93.7	-0.06	0.91	0.90	94.3	100
Setup 4.3																		
		$\hat{\tau}_1$				$\hat{\tau}_2$				$\hat{\tau}_3$				$\hat{\tau}_4$				
segmented	200	1.69	18.05	2.89	90.0	2.15	19.03	2.04	90.9	1.24	15.82	1.82	93.1	1.69	16.00	4.75	94.9	99.2
proposed		-0.45	3.46	2.97	92.3	0.00	2.08	2.01	94.7	-0.03	1.75	1.77	95.4	0.14	4.57	4.26	88.6	99.9
segmented	500	0.05	6.05	1.78	88.0	0.25	6.39	1.21	93.2	0.07	6.56	1.07	94.8	0.34	2.80	2.68	94.8	100
proposed		-0.25	1.84	1.78	93.9	-0.02	1.27	1.24	93.3	0.04	1.10	1.08	94.5	0.12	2.75	2.63	93.5	100
segmented	1000	3.08	22.65	1.24	94.8	3.42	26.11	0.88	92.0	3.05	24.66	0.86	89.7	3.63	29.54	2.59	93.6	100
proposed		-0.11	1.25	1.24	94.9	-0.03	0.91	0.86	94.1	0.04	0.74	0.76	95.5	0.05	1.78	1.85	95.2	100
segmented	2500	-0.02	0.36	0.76	99.2	-0.24	0.29	0.54	99.3	-0.09	0.21	0.47	99.4	0.11	0.47	1.14	99.0	100
proposed		-0.08	0.79	0.78	94.2	-0.06	0.57	0.54	94.6	0.03	0.48	0.48	94.6	0.10	1.15	1.16	95.6	100

when there is only one knot, it performs well and similarly to the proposed method. When there are two or four knots to be estimated, performances of the "segmented" method seems less stable sometimes. For example, in setups 2.1, 2.2, and 4.3 and when $n = 2500$, the coverage probability is close to 1 due to inaccurate standard error estimates, and sometimes the average standard error is more than three times of the Monte Carlo standard deviation (setup 2.1, $n = 2500$ for τ_1). Taking a closer look at the issue through examination of the histograms, we found that estimates from the "segmented" method sometimes exhibit large outliers and may even have bimodal distribution with two modes at each side of the truth. These results are likely due to both computational (e.g., being stuck in local optima) and/or inferential problems. Other than these occasional issues, overall the "segmented" method performs well.

The proposed method performs well in finite samples in terms of both statistical properties and convergence rates in all scenarios considered. Specifically, the bias is close to zero and coverage probabilities are close to the nominal level. Only when

$n = 200$, it sometimes has slightly lower coverage probabilities. However, we comment that this is due to the finite sample effect. These results are as expected, considering that, for example, in four-knots settings, there are 11 parameters to be estimated. Sample size $n = 200$ is a rather extreme setting and we chose to evaluate this setting to illustrate that the proposed method is computationally stable even when n is small. With the chosen convergence criterion, the proposed method converges 100% or almost 100% of the time for all scenarios. In summary, the proposed method is computationally convenient, with a high convergence rate. It performs reasonably well empirically for single and multiple-knots scenarios with different sample sizes. As in all iterative algorithms, being stuck in local optima is a potential problem requiring special attention. This problem is much alleviated in the proposed method compared with the smoothing-based method by using a simple estimating equation without tuning parameters and a two-step algorithm taking advantage of the analytic OLS solution when $\boldsymbol{\tau}$ is fixed. In practice, good initial values that are close to the truth are still important. We suggest using prior knowledge, scientific inputs or data visualization/exploratory analysis to guide selection of initial values. In data analysis, one may also try multiple initial values and choose results based on the likelihood.

2.6 Application

We applied the proposed method to the nefazodone CBASP (Cognitive Behavioral-Analysis System of Psychotherapy) trial on patients with nonpsychotic chronic major depressive disorder (*Keller et al.*, 2000). In this 12-week trial, patients were randomly assigned with equal probability to receive either nefazodone, CBASP, or both treatments. Our outcome of interest is the score on the 24-item Hamilton Rating Scale for Depression (HRSD) at 12 weeks after treatment.Our aim is to understand the relationship between baseline HAMA Psychic Anxiety Score (HPAS) and HRSD after treatment. For both HPAS and HRSD, a larger value indicates a worse condition.

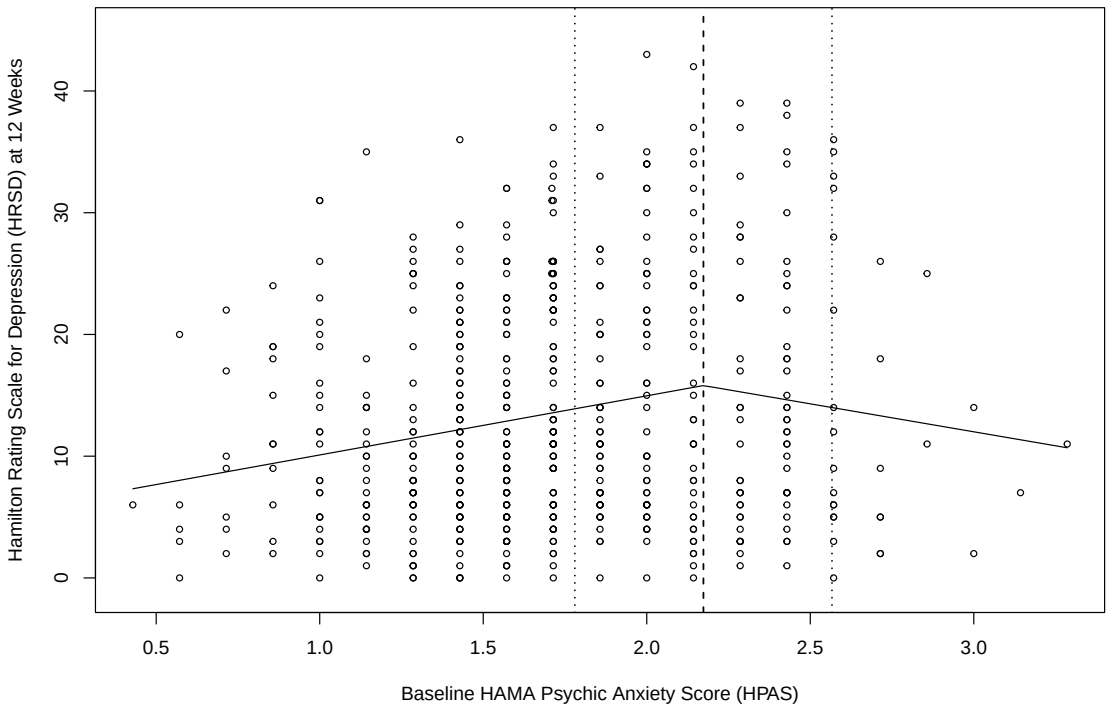

Figure 2.1: The relationship between baseline HAMA Psychic Anxiety Score (HPAS) and Hamilton Rating Scale for Depression (HRSD) at 12 weeks. The solid line is the fitted unadjusted linear spline model. The three dotted vertical lines indicate the estimated knot (middle) and the corresponding 95% confidence interval (left and right).

Our analysis is based on the 577 participants with HRSD score at 12 weeks available. Figure 2.1 shows a scatterplot of HRSD at 12 weeks versus baseline HPAS. It is apparent that the relationship between HPAS and HRSD at 12 weeks is not linear. Instead, HRSD at 12 weeks increases with higher baseline HPAS and then decreases or levels off afterwards. A one-knot linear spline model seems appropriate to model this relationship and we are interested in estimating the value of baseline HPAS at which its effect on HRSD at 12 weeks changes.

The estimated change-point in an unadjusted one-knot linear spline model is 2.17 (95% CI: $1.78, 2.57$). When baseline HPAS is less than 2.17, the HRSD score at 12

weeks increases with HPAS with an estimated slope of 4.86 (p-value < 0.001). When baseline HPAS is greater than 2.17, the HRSD score at 12 weeks decreases with HPAS with an estimated slope of -4.59 (p-value=0.156). However, the decreasing trend is not statistically significant. We comment that hypothesis testing for the difference in slopes is challenging as the corresponding parameter only exists under the alternative hypothesis but not under the null (*Muggeo*, 2016). The p-values reported above are based on standard errors from an usual fitted linear spline model with a knot fixed at 2.17, without accounting for the uncertainty in estimating the knot. More results, including standard errors that account for estimation of the knot using the proposed method are reported in Table 2.5.

Table 2.5: The CBASP trial: results from the fitted linear spline model for the effect of baseline HPAS on HRSD at 12 weeks. Coefficients, $\eta_1, \ldots, \eta_5$ correspond to the effect of age, female (vs. male), white (vs. non-white), single, and widowed/divorced/separated (vs. married). SE_* is the standard error derived from the Theorem 2; SE is the the standard error derived from the linear regression by treating τ fixed at the estimated value. P-values are calculated from the two-sided Wald tests while treating τ fixed at the estimated value.

	Unadjusted One-Knot Model				**Adjusted One-Knot Model**			
	Estimate	**SE_***	**SE**	**P-Value**	**Estimate**	**SE_***	**SE**	**P-Value**
τ	2.173	0.200	-	-	2.138	0.170	-	-
β_0	5.241	1.908	1.810	0.004	5.749	3.105	2.930	0.050
β_1	4.860	1.184	1.097	<0.001	4.805	1.339	1.121	<0.001
β_{11}	-9.454	4.994	3.858	0.015	-8.887	3.886	3.641	0.015
η_1					0.036	0.040	0.040	0.365
η_2					0.532	0.819	0.819	0.516
η_3					-3.504	1.341	1.340	0.009
η_4					2.372	1.009	1.006	0.019
η_5					0.668	0.929	0.927	0.472

Next, we additionally adjust for age, female (vs. male), white (vs. non-white), and marital status (single, or widowed/divorced/separated vs. married) in our model. Results are similar to those based on an unadjusted model. Specifically, the estimated change-point in effect of HPAS is located at 2.14 (95% CI: $1.81, 2.47$). The slopes

before and after the change-point are significantly different (p-value=0.015). Given age, sex, race and marital status, when baseline HPAS is below 2.14, HRSD score at 12 weeks increases with baseline HPAS with a slope of 4.81 (p-value < 0.001), indicating that the outcome at 12 weeks is worse for patients with a severe condition at baseline. When baseline HPAS is greater than 2.14, the estimated slope is -4.08; however, the slope is again not significantly different from zero (p-value = 0.173). Again, reported p-values are bassed on a fitted model with a knot fixed at 2.14 without accounting for the uncertainty in estimating the knot. See Table 2.5 for additional results on the adjusted model.

For the type of nonlinear relationship observed in Figure 2.1, one may alternatively model it using a quadratic model, which is also common and acceptable in practice. However, it is obvious that a quadratic model will not be able to offer the kind of easy and intuitive interpretation as the fitted linear spline models as we show here. In particular, when the interest focuses on where the relationship between HRSD and baseline HPAS changes, the estimation of the change-point based on a quadratic model, that is, the vertex, would be heavily dependent on the assumption of symmetry about the vertex implied by a quadratic model. The assumption does not seem to hold in this particular example, based on clinical knowledge and visual evidences. It is a strong and often invalid assumption in general, likely leading to unreliable estimation of change-points.

2.7 Discussion

Linear spline models are an important class of models that can accommodate non-linear relationships while still allowing easy and intuitive interpretation. Although applications of the linear spline model are already widespread, its use in practice is still hindered by the lack of both rigorously studied and computationally convenient

methods for estimation of knots. As a result, its full potential and flexibility, particularly in estimating change-points and studying threshold effects, have not been completely realized so far.

Most existing rigorously studied methods for knot estimation are based on the natural idea of smoothing, which involves tuning parameters and is unnecessarily complicated, leading to difficulty in computation. Unlike smoothing-based methods, we have taken a fresh angle and proposed a novel and conceptually simple approach to circumvent nondifferentiability based on the idea of modified derivatives, that is, whenever derivatives are needed but do not exist, we redefine it as zero. This modified derivative idea allows us to solve the otherwise challenging nondifferentiabiltiy issue for linear spline models in terms of theory, estimating equation and computation in a unified way. First, in terms of theory, with the proposed modified derivatives we were able to show that the linear spline model is DQM and derive its score function. Our contribution is not limited to a single estimator. We have derived all influence functions of RAL estimators in both parameter and semiparametric settings and studied the efficient influence function and the efficiency bound. The study on influence functions and the efficiency bound fills in a gap in the literature on estimation of knots for linear spline models. The class of influence functions we identified contains that of the smoothing-based method of Das et al. (2016), as their estimator is an RAL estimator. Although none of previous work in the literature on knots estimation uses the concept of DQM, DQM as shown in this paper is in fact the underlying reason for existence of consistent and asymptotic normal estimators.

Second, we have proposed a simple but nontraditional estimating equation approach to estimate knots as well as other parameters, where the estimating function is a modified derivative of the squared error loss. It is simpler than smoothing-based methods as it does not require smoothing or involve tuning parameters to control smooth-

ness. It is nontraditional because the estimating equation itself is nondifferentiable. Finally, we have proposed a new two-step computational algorithm to solve this nontraditional estimating equation, applying the modified derivative idea again to the nondifferentiable estimating function itself. We have shown that the same modified derivative idea used for showing DQM, when applied for solving nondifferentiable equations, is also a generalized Jacobian studied in *Qi and Sun* (1993) for solving semismooth equations in applied mathematics and optimization literature. This two-step algorithm takes advantage of the analytic OLS solution available when knots are treated as known, greatly enhancing numerical stability. As discussed in Section 2.4, our algorithm differs from the NR algorithms studied in *Qi and Sun* (1993) and *Cui et al.* (2018), and these two papers study from the perspective of computation without considering statistical properties. It is interesting that the modified derivative idea we propose in this paper bridges two important concepts in distinct literature, namely the differentiable in quadratic mean studied in statistical literature and the generalized Jacobian for semismooth functions in optimization/computation literature. We expect that this connection can help rigorously study statistical properties of computational and learning methods studied in *Cui et al.* (2018).

In terms of empirical performances, simulation studies have shown that the proposed method greatly improves upon the smoothing-based methods. In terms of computation, the proposed method is reasonably stable and has high rates of convergence. In particular, the proposed method is more insensitive to choices of initial values and able to handle multiple change-points easily. The improvement is due to new developments in terms of the estimating equation and computation discussed above. Asymptotic properties have been studied rigorously using the empirical process theory, which leads to accurate statistical inference for the proposed method. We have demonstrated that the proposed estimator is root-n consistent and asymptotically normal. Furthermore, the proposed estimator achieves the semiparametric local effi-

ciency bound when the conditional variance is constant. Therefore, the computational advantage is obtained without sacrificing statistical efficiency.

In summary, we have proposed a rigorously studied and computationally stable method for estimation of knots in linear spline models. With easy interpretation of the model and convenient implementation of the estimation method, we anticipate linear spline models with unknown knots play a more important role in practice, especially in studying change-points and threshold effects. Finally, we comment that the proposed idea is a generic idea that can be applied to many other settings, for example, in estimating change-points in longitudinal data, in a time-series framework and in a multivariable setup where several variables have change-point effects.

CHAPTER III

Change-Points Estimation in Generalized Linear Spline Models

3.1 Introduction

In biomedical researches, a factor of interest often exhibits a nonlinear effect that should be properly modeled in analysis. A well-known example is the Body Mass Index (BMI), which usually is associated with metabolic and disease outcomes in a nonlinear way since being too low (underweight) or too high (overweight or obese) in BMI negatively affects outcomes. Owing to its easy interpretation and flexibility, the generalized linear spline model is a commonly used approach to model nonlinearity, where the effect on outcomes on a scale determined by the link function is modeled using piecewise linear terms joined at knots/change points. That is, the effect changes at each change point. A practical difficulty in using splines to account for nonlinearity is how to prespecify the knots or change points. Often times, analysts have to rely on subject-matter knowledge to choose change points. In the case of BMI, widely accepted cutoff points are available to categorize a person into underweight, normal weight, overweight and obesity groups. However, perhaps in the majority of cases, for example, when studying the effect of some less studied biomarkers, such a standard way is not available and researchers lack the scientific knowledge to prespecify change

points. The choice of the change points is often made arbitrarily or in an ad-hoc manner. Moreover, rather than the magnitude of association between a factor and the outcome as quantified by the coefficients/slopes, in many applications the change points themselves are the main research interest that can be used for policy making and developing guidelines. In these situations, a formal and rigorous data-driven way to choose the change points becomes critical. Therefore, estimating and making inferences on change points in the generalized linear spline model is an important research question requiring rigorous study.

In our motivating study, interest focuses on understanding the relationship between BMI and glomerular filtration rate (GFR) on in-hospital transfusion or gastrointestinal (GI) bleeding for patients who underwent percutaneous coronary intervention (PCI) at 33 hospitals in Michigan. For both factors, it is expected that the effect is not linear and changes at some change points. We aim to identify these change points empirically from data without relying on prior knowledge. Generalized linear spline models have widespread applications in many other areas. For example, this type of model has been studied in species and habitat relationships in ecological researches (*Francesco Ficetola and Denoël*, 2009; *Eigenbrod et al.*, 2009), in air pollutants and preterm birth in environmental researches (*Llop et al.*, 2010), in heavy rainfall changes in meteorological researches (*Villarini et al.*, 2013), and in heat effects on mortality in epidemiology researches (*Baccini et al.*, 2008).

Efforts for developing statistical methods for change-points estimation date back to more than six decades ago (*Quandt*, 1958), with a lot of research followed since then. Till today, most of the work has focused on linear spline models with continuous outcomes. The main difficulty in change points estimation is that the likelihood function is not differentiable. Search-based and smoothing-based methods are the two most popular approaches. Search-based methods are intuitive and popular. Early

contributions to search-based methods include *Quandt* (1958), *Hudson* (1966), *Feder* (1975b,a) and *Knowles et al.* (1991). Several more recent methods are proposed to estimate change points using grid search methods in optimizing objective functions (*Lerman*, 1980; *Julious*, 2001; *Hansen*, 2000, 2017). A drawback of search-based methods is that they are not computationally efficient for large sample sizes or multiple change points. Another main direction to deal with nondifferentiability is to use smoothing. *Tishler and Zang* (1981) proposed to use a quadratic approximation to provide a smoothed version of the likelihood function. Later, *Chiu et al.* (2002, 2006) proposed the "bent-cable model" as a smoothing version of the linear spline model and the model" was further studied by *Das et al.* (2016). *Hahn et al.* (2017) studied Nesterov smoothing and extended one dimensional linear spline models into multiple dimensions. Other than search- and smoothing-based methods, in the CHAPTER II, we proposed a semi-smooth estimating method by showing the property of differentiable in quadratic mean in linear spline models and developed a two-step semi-smooth Newton-Raphson algorithm.

In general, methods described above are rigorously studied, but they focus only on linear spline models with continuous outcomes. Other estimation strategies and computation algorithms are studied in the literature. *Muggeo* (2003) developed an estimating strategy via linearization technique in generalized linear spline models, implemented in a highly-cited R package (*Muggeo*, 2008). However, theoretical properties of the method were not studied. Combining grid-search and smoothing-based methods, *Fong et al.* (2017) developed an R package to estimate change points in generalized linear spline models. Recently, a semismooth Newton-Raphson method was studied in *Cui et al.* (2018) for solving general majorization-minimization problems, with application in continuous piecewise affine regression functions. However, *Cui et al.* (2018) did not study statistical properties. Bayesian methods for change-points estimation were also proposed in the literature, including *Bacon and Watts* (1971), *Carlin et al.*

(1992), *Smith and Cook* (1980) and *Elliott and Shope* (2003) ; see *Chen et al.* (2011) for a detailed description and comparison.

In this article, we introduce a formal and rigorous method for estimating and making inferences on change points for generalized linear spline models with multidimensional predictors. The model framework and notations are presented in Section 3.2. In Section 3.3, we study influence functions of regular and asymptotic linear estimators and the efficiency bound for parametric and semiparametric generalized linear spline models. In Section 3.4, we propose a semismooth estimating equation and a computationally convenient algorithm. The convergence of the algorithm and statistical properties are rigorously studied. The method is evaluated by simulations in Section 3.5 and illustrated by an application in Section 3.6.

3.2 Generalized Linear Spline Models and Notations

Suppose data are collected on n subjects in the form of $\boldsymbol{W}_i = \{Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i\}$, $i = 1, \ldots, n$, independent and identically distributed across subject i. For each subject i, let Y_i denote the response, $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{Ji})^T$ denote the J-dimensional factor of interest, and $\boldsymbol{Z}_i$ denote an L-dimensional vector of covariates to be adjusted for including, for example, demographics, clinical measurements, biomarkers in biomedical research. Interest focuses on understanding the relationship between the response and the factor of interest $\boldsymbol{X}_i$, adjusting for $\boldsymbol{Z}_i$. For convenience, we denote $\boldsymbol{X}_i^* = (\boldsymbol{X}_i, \boldsymbol{Z}_i)$. We use a generalized linear spline model to model the conditional mean of Y_i on $\boldsymbol{X}_i^*$, denoted as $\mu(\boldsymbol{X}_i^*; \boldsymbol{\theta})$, where effects of $\boldsymbol{X}_i$ are modeled flexibly with linear splines within the convenient framework of generalized linear models (*Dobson and Barnett*, 2018). Specifically, the model is written as

$$g\{\mu(\boldsymbol{X}_i^*; \boldsymbol{\theta})\} = \beta_0 + \sum_{j=1}^{J} \left\{ \beta_j X_{ji} + \sum_{k=1}^{K_j} \beta_{jk} (X_{ji} - \tau_{jk})^+ \right\} + \boldsymbol{\eta}^T \boldsymbol{Z}_i = \xi(\boldsymbol{X}_i^*; \boldsymbol{\theta}), \quad (3.1)$$

where $g(.)$ is a known link function, effects of $\boldsymbol{X}_i$ are modeled using linear spline terms with $(X_{ji}-\tau_{jk})^+ = X_{ji}-\tau_{jk}$ if $X_{ji} > \tau_{jk}$ and 0 otherwise, and effects of $\boldsymbol{Z}_i$ are modeled as linear with coefficients $\boldsymbol{\eta}$ as usual. The conditional variance of Y_i is denoted as $V(Y_i|\boldsymbol{X}_i^*;\boldsymbol{\theta},\phi)$, which is a function of $\boldsymbol{X}_i^*$ and might be related to the parameter of interest $\boldsymbol{\theta}$ and additional parameters ϕ. When change points τ_{jk} ($j = 1,\ldots,J$ and $k = 1,\ldots K_j$) are assumed to be known, model (3.1) is the generalized linear model, where $\mu(\boldsymbol{X}_i^*;\boldsymbol{\theta})$ is related to the linear predictor $\xi(\boldsymbol{X}_i^*;\boldsymbol{\theta})$ through the link function $g(.)$. However, in this article, we do not assume change points are known. Instead, the estimation of change points, as well as the effects of linear spline terms, is the main research interest.

The generalized linear spline model (3.1) dedescribed above is a restricted moment model on the first moment and does not assume a specific distribution for Y given $\boldsymbol{X}_i^*$. Model (3.1) includes the most common situation that $Y|\boldsymbol{x}^*$ follows a distribution in the exponential family. When we additionally assume $Y|\boldsymbol{x}^*$ follows a distribution in the exponential family, the distribution can be written as

$$f(Y = y;\psi,\phi) = \exp\Big\{\frac{y\psi - b(\psi)}{\phi} + c(y;\phi)\Big\},$$

where ψ is the natural parameter and ϕ is the scale parameter. And the conditional variance can be represented as $V(Y_i|\boldsymbol{X}_i^*;\boldsymbol{\theta},\phi) = \phi v\{\mu(\boldsymbol{X}_i^*;\boldsymbol{\theta})\}$, where $v\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta})\} = \partial\mu(\boldsymbol{X}^*;\boldsymbol{\theta})/\partial\psi$. When $g(.)$ is additionally chosen to be the canonical link function, i.e., $\psi = g\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta})\}$, $g(.)$ is equivalent to the inverse function of $\partial b(\psi)/\partial\psi$ and we further have $g(.) = 1/v(.)$.

In model (3.1), for each factor of interest X_{ji} ($j = 1,\ldots,J$), K_j is the pre-specified number of change points. We denote the total number of change-points as $K = K_1+\ldots+K_J$. Without loss of generality, for each j, we assume that τ_{jk} ($k = 1,\ldots,K_j$)

are ordered and distinct to ensure model identifiability. That is, for $m < n$, then $\tau_{jm} < \tau_{jn}$ for all $j = 1, \ldots, J$. We assume that the factor of interest $\boldsymbol{X}_i$ has a bounded domain, i.e. $X_{ji} \in [C_{j1}, C_{j2}]$ for all $j = 1, \ldots, J$. As a result, all change-points are within these bounded intervals as well. We assume the link function $g(.)$ is monotonically increasing, continuous and first-order differentiable. The truth of ϕ is denoted as ϕ^0. And we denote $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_{11}, ..., \beta_{1K_1}, ..., \beta_J, \beta_{J1}, ..., \beta_{JK_J})^T$, $\boldsymbol{\tau} = (\tau_{11}, ..., \tau_{1K_1}, ..., \tau_{J1}, ..., \tau_{JK_J})^T$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\tau}^T, \boldsymbol{\eta}^T)^T$, where $\boldsymbol{\theta}$ is assumed to belong to a compact set $\boldsymbol{\Theta}$ with dimension $q = 2K + J + L + 1$. We also denote the true value of $\boldsymbol{\theta}$ as $\boldsymbol{\theta}^0$, assumed to be an interior point of the compact set $\boldsymbol{\Theta}$.

3.3 Influence Functions of Generalized Linear Spline Model

In this section, we will describe influence functions of regular and asymptotic linear (RAL) estimators, derive the efficient score function, and derive the efficiency bound for both parametric and semiparametric generalized linear spline models. This section roughly follows the idea and proofs studied in *Tsiatis* (2007) and Section 2.4.2.

3.3.1 Parametric Generalized Linear Spline Model

We start by considering the parametric generalized linear spline model, where a parametric model is assumed for the distribution of $Y|\boldsymbol{X}^*$. The joint distribution of $(Y, \boldsymbol{X}^*)$ is denoted as $p_{Y,\boldsymbol{X}^*}(y, \boldsymbol{x}^*; \boldsymbol{\zeta}) = p_{Y|\boldsymbol{X}^*}(y|\boldsymbol{x}^*; \boldsymbol{\gamma}_1) p_{\boldsymbol{X}^*}(\boldsymbol{x}^*; \boldsymbol{\gamma}_2)$, where $p_{Y|\boldsymbol{X}^*}(y|\boldsymbol{x}^*; \boldsymbol{\gamma}_1)$ denotes the the conditional distribution of $Y|\boldsymbol{X}^*$, $p_{\boldsymbol{X}^*}(\boldsymbol{x}^*; \boldsymbol{\gamma}_2)$ denotes the density of $\boldsymbol{X}^*$, and $\boldsymbol{\zeta} = (\boldsymbol{\theta}^T, \boldsymbol{\gamma}^T)^T$ with $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \boldsymbol{\gamma}_2^T)^T$. The dimension of $\boldsymbol{\theta}, \boldsymbol{\gamma}$, and $\boldsymbol{\zeta}$ are q, r and p, respectively. The truth is denoted as $\boldsymbol{\zeta}^0 = (\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0)$. Using similar arguments as in Lemma 1, we can show that the parametric generalized linear spline model $p(\boldsymbol{w}; \boldsymbol{\zeta})$ also satisfies the condition of differentiable in quadratic mean (DQM). Then, based on the definition of DQM in *van der Vaart* (2000), $I_{\boldsymbol{\zeta}}(\boldsymbol{W}; \boldsymbol{\zeta}) = E[S_{\boldsymbol{\zeta}}(\boldsymbol{W}; \boldsymbol{\zeta}) S_{\boldsymbol{\zeta}}^T(\boldsymbol{W}; \boldsymbol{\zeta})]$ is the Fisher information matrix of $\boldsymbol{\zeta}$ and

$S_{\boldsymbol{\zeta}}(\boldsymbol{W};\boldsymbol{\zeta}) = \{S^T_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}), S^T_{\boldsymbol{\gamma}}(\boldsymbol{W};\boldsymbol{\zeta})\}^T$ is the score function of $\boldsymbol{\zeta}$, where $S_{\boldsymbol{\gamma}}(\boldsymbol{W};\boldsymbol{\zeta}) = \frac{\partial \log\{p(\boldsymbol{W};\boldsymbol{\zeta})\}}{\partial \boldsymbol{\gamma}}$, $S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}) = \frac{\partial \log\{p(\boldsymbol{W};\boldsymbol{\zeta})\}}{\partial \xi(\boldsymbol{X}^*;\boldsymbol{\theta})} H^T(\boldsymbol{X}^*;\boldsymbol{\theta})$ and

$$H(\boldsymbol{X}^*;\boldsymbol{\theta}) = \Big\{1, X_1, (X_1-\tau_{11})^+, \ldots, (X_1-\tau_{1K_1})^+, \ldots, X_J, (X_J-\tau_{J1})^+, \ldots, (X_1-\tau_{JK_J})^+, -\beta_{J1} I(X>\tau_{J1}), \ldots, -\beta_{JK_J} I(X>\tau_{JK_J}), \boldsymbol{Z}\Big\}. \quad (3.2)$$

Using the same modified derivative idea in the formula (2.2), $H(\boldsymbol{X}^*;\boldsymbol{\theta})$ function can be viewed as the modified derivative of $\xi(\boldsymbol{X}^*;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. The idea of the modified derivative is to redefine derivatives for places where derivatives do not exist as zero. Specifically, the modified derivative of $(x-\tau)^+ = -I(x>\tau)$ and $I(x>\tau) = 0$.

Similar to Lemma 2, under the same regularity conditions (i.e., $E_{\boldsymbol{\zeta}}\{\varphi(\boldsymbol{W})\}$ and $E_{\boldsymbol{\zeta}}\{\varphi(\boldsymbol{W})^T\varphi(\boldsymbol{W})\}$ exist and are continuous in $\boldsymbol{\zeta}$ in a neighborhood of $\boldsymbol{\zeta}^0$), the influence function $\varphi(\boldsymbol{W})$ of any RAL estimators satisfy the following two equations

$$E\{\varphi(\boldsymbol{W})S^T_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)\} = I^{q\times q} \text{ and } E\{\varphi(\boldsymbol{W})S^T_{\boldsymbol{\gamma}}(\boldsymbol{W};\boldsymbol{\zeta}^0)\} = 0^{q\times r}.$$

Conversely, any element in the Hilbert space $\mathcal{H}$ satisfying the two equations above is the influence function of some RAL estimator. According to Example 25.15 in *van der Vaart* (2000), the tangent space, denoted as $\mathcal{T}$, is given by the linear space spanned by the score function of $\boldsymbol{\zeta}$, i.e., $\mathcal{T} = \{B^{q\times p} S_{\boldsymbol{\zeta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)$, for all $q\times p$ matrices $B\}$. And $\mathcal{T}$ can further be expressed as the direct sum of $\mathcal{T}_{\boldsymbol{\theta}}$ and Λ, where $\Lambda = \{B^{q\times\gamma} S_{\boldsymbol{\gamma}}(\boldsymbol{W};\boldsymbol{\zeta}^0)$, for all $q\times r$ matrices $B\}$ is the nuisance tangent space, and $\mathcal{T}_{\boldsymbol{\theta}} = \{B^{q\times q} S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)$, for all $q\times q$ matrices $B\}$ is the tangent space spanned by $S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)$. Next, according to Theorem 3.5 and Corollary 2 in *Tsiatis* (2007), the efficient influence function

$\varphi_{eff}(\boldsymbol{W})$ with a non-singular information matrix is

$$\begin{aligned}\varphi_{eff}(\boldsymbol{W}) &= \Gamma I_{\boldsymbol{\zeta}}^{-1}(\boldsymbol{W};\boldsymbol{\zeta}^0)S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0) \\ &= \Big[E\{S_{eff}(\boldsymbol{W};\boldsymbol{\zeta}^0)S_{eff}^T(\boldsymbol{W};\boldsymbol{\zeta}^0)\}\Big]^{-1}S_{eff}(\boldsymbol{W};\boldsymbol{\zeta}^0),\end{aligned}$$

where $\Gamma = (I^{q\times q}, 0^{q\times r})$, $S_{eff}(\boldsymbol{W};\boldsymbol{\zeta}^0) = S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0) - \Pi[S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)|\Lambda]$ is the efficient score, and $\Pi[h|\Lambda] = E\{hS_{\boldsymbol{\gamma}}^T(\boldsymbol{W};\boldsymbol{\zeta}^0)\}[E\{S_{\boldsymbol{\gamma}}(\boldsymbol{W};\boldsymbol{\zeta}^0)S_{\boldsymbol{\gamma}}^T(\boldsymbol{W};\boldsymbol{\zeta}^0)\}]^{-1}S_{\boldsymbol{\gamma}}^T(\boldsymbol{W};\boldsymbol{\zeta}^0)$ is the projection of h onto space Λ. Therefore, the parametric efficiency bound is $[E\{S_{eff}(\boldsymbol{W};\boldsymbol{\zeta}^0)S_{eff}^T(\boldsymbol{W};\boldsymbol{\zeta}^0)\}]^{-1}$. For example, when $p_{Y|\boldsymbol{X}^*}(y|\boldsymbol{x}^*;\gamma_1)$ belongs to the exponential family, the parametric efficiency bound is $\phi^0\Big[E[\frac{H^T(\boldsymbol{X}^*;\boldsymbol{\theta}^0)H(\boldsymbol{X}^*;\boldsymbol{\theta}^0)}{v\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}g'\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}^2}]\Big]^{-1}$ and the efficient score can be represented as $S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0) = \frac{\{Y-\mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}H^T(\boldsymbol{X}^*;\boldsymbol{\theta}^0)}{\phi^0 v\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}g'\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}} = S_{eff}(\boldsymbol{W};\boldsymbol{\zeta}^0)$. The optimal estimator is the solution of the following estimating equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{H^T(\boldsymbol{X}_i^*;\boldsymbol{\theta}^0)\{Y_i-\mu(\boldsymbol{X}_i^*;\boldsymbol{\theta})\}}{v\{\mu(\boldsymbol{X}_i^*;\boldsymbol{\theta}^0)\}g'\{\mu(\boldsymbol{X}_i^*;\boldsymbol{\theta}^0)\}} = 0. \tag{3.3}$$

3.3.2 Semiparametric Generalized Linear Spline Model

Next, according to similar procedures in Chapter 4.5 of *Tsiatis* (2007), we extend the arguments to semiparametric generalized linear spline models, i.e., the restricted moment model. As the nondifferentiability problem only occurs on $\boldsymbol{\theta}$, arguments for the nuisance tangent space in *Tsiatis* (2007) can be borrowed. Therefore, the nuisance tangent space Λ for the semiparametric generalized linear spline model (3.1) is

$$\Lambda = \Big\{h^{q\times 1}(Y,\boldsymbol{X}^*),\ \text{such that } E\{h(Y,\boldsymbol{X}^*)\{Y-\mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}|\boldsymbol{X}^*\} = 0^{q\times 1}\Big\}.$$

The space orthogonal to the nuisance tangent space $\Lambda^{\perp}$ is given by

$$\Lambda^{\perp} = \Big\{ A^{q\times 1}(\boldsymbol{X}^*)\{Y - \mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\},\ \text{for all } q\text{-dimensional functions } A^{q\times 1}(\boldsymbol{X}^*) \Big\}.$$

And the projection of arbitrary element $h(Y,X) \in \mathcal{H}$ onto space Λ satisfies

$$\Pi[h(Y,X)|\Lambda^{\perp}] = E[h(Y,X)\{Y - \mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}^T|X]V^{-1}(Y|\boldsymbol{X}^*;\boldsymbol{\theta}^0,\phi^0)\{Y - \mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}.$$

Following similar proofs of Lemma 3 in Chapter II, we can show that $E\{YS^T_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)|\boldsymbol{X}^*\} = H(\boldsymbol{X}^*;\boldsymbol{\theta}^0)/g'\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}$. Thus the efficient score for the semiparametric generalized linear spline model $\mathcal{P}$ is given by

$$\begin{aligned} S_{eff}(\boldsymbol{W};\boldsymbol{\zeta}^0) &= E[S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)\{Y - \mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}^T|\boldsymbol{X}^*]V^{-1}(Y|\boldsymbol{X}^*;\boldsymbol{\theta}^0,\phi^0)\{Y - \mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\} \\ &= \frac{H^T(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\{Y - \mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}}{V(Y|\boldsymbol{X}^*;\boldsymbol{\theta}^0,\phi^0)g'\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}}. \end{aligned}$$

The semiparametric efficiency bound is given by

$$\mathcal{V} = \Big[E\{S_{eff}(\boldsymbol{W};\boldsymbol{\zeta}^0)S^T_{eff}(\boldsymbol{W};\boldsymbol{\zeta}^0)\}\Big]^{-1} = \Big[E\{\frac{H^T(\boldsymbol{X}^*;\boldsymbol{\theta}^0)H(\boldsymbol{X}^*;\boldsymbol{\theta}^0)}{V(Y|\boldsymbol{X}^*;\boldsymbol{\theta}^0,\phi^0)g'\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}^2}\}\Big]^{-1}.$$

If the conditional variance is assumed to be $V(Y|\boldsymbol{X}^*;\boldsymbol{\theta}) = \phi v\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta})\}$, the efficient score $S_{eff}(\boldsymbol{W};\boldsymbol{\zeta}^0)$ and the semiparametric local efficiency bound $\mathcal{V}$ is given by

$$\begin{aligned} S_{eff}(\boldsymbol{W};\boldsymbol{\zeta}^0) &= \frac{H^T(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\{Y - \mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}}{\phi^0 v\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}g'\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}}, \\ \mathcal{V} &= \phi^0\Big[E[\frac{H^T(\boldsymbol{X}^*;\boldsymbol{\theta}^0)H(\boldsymbol{X}^*;\boldsymbol{\theta}^0)}{v\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}g'\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}^2}]\Big]^{-1}. \end{aligned}$$

Thus, the efficient estimator can be obtained from the following equation equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{H^T(\boldsymbol{X}_i^*;\boldsymbol{\theta}^0)\{Y_i - \mu(\boldsymbol{X}_i^*;\boldsymbol{\theta})\}}{v\{\mu(\boldsymbol{X}_i^*;\boldsymbol{\theta}^0)\}g'\{\mu(\boldsymbol{X}_i^*;\boldsymbol{\theta}^0)\}} = 0, \tag{3.4}$$

which is the same as the estimating equation (3.3). If the working model for the conditional variance is wrong, then the estimating equation above is still unbiased. If the working model for the conditional variance is wrong, then the estimating equation above is still unbiased.

3.4 Proposed Method

3.4.1 Estimation Procedure

According to the estimating equations (3.3) and (3.4) that we have derived in Section 3.3, we propose the following estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} Q(\boldsymbol{W}_i;\boldsymbol{\theta}) \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{H^T(\boldsymbol{X}_i^*;\boldsymbol{\theta})\{Y-\mu(\boldsymbol{X}_i^*;\boldsymbol{\theta})\}}{v\{\mu(\boldsymbol{X}_i^*;\boldsymbol{\theta})\}g'\{\mu(\boldsymbol{X}_i^*;\boldsymbol{\theta})\}} = 0, \tag{3.5}$$

where $v(.)$ is a pre-specified function. Specifically, we assume that the working variance function of $V(Y_i|\boldsymbol{X}_i^*;\boldsymbol{\theta},\phi)$ is $\phi v\{\mu(\boldsymbol{X}_i^*;\boldsymbol{\theta})\}$, which may or may not contain the truth. To provide some intuitions for this estimating equation, we note that $Q(\boldsymbol{W};\boldsymbol{\theta})$ in the above estimating function mimics, up to a constant of proportionality, the score function when $Y|\boldsymbol{x}^*$ belongs to an exponential family. The solution to the proposed estimating equation (3.5) is denoted as $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\beta}}_n^T, \widehat{\boldsymbol{\tau}}_n^T, \widehat{\boldsymbol{\eta}}_n^T)^T$, with the subscript n omitted for simplicity sometimes. To justify the validity of $\widehat{\boldsymbol{\theta}}_n$, we study its asymptotic properties rigorously in Section 3.4.3.

3.4.2 Estimation Algorithm

Solving the estimating equation (3.5) is still a challenging problem. First, the estimating function itself in equation (3.5) is not differentiable and non-regular. Therefore, popular computational algorithms, for example, the Newton-Raphson (NR) method, do not directly apply to this case. Second, due to the inherent non-differentiability of the mean function $\mu(\boldsymbol{X}_i^*;\boldsymbol{\theta})$ with respect to $\boldsymbol{\tau}$, numerical algorithms for obtaining

estimates of $\boldsymbol{\tau}$ is prone to the problem of being stuck at local optima and non-convergence, leading to undesirable performances. To solve the two problems, we adapt the idea of a two-step semismooth NR algorithm in Section 2.4.1 for linear spline models with continuous outcomes. In this two-step semismooth NR algorithm, the estimation of $\boldsymbol{\tau}$ is separated from the estimation of other unknown parameters in each iteration. When $\boldsymbol{\tau}$ is fixed, the model is the usual generalized linear model, for which analytical solution or common computationally stable algorithms are available in standard software. When other parameters are treated as known, it simplifies the estimation problem by reducing the dimension of unknown parameters. In this step, we update estimates of $\boldsymbol{\tau}$ using the idea of the modified NR algorithm for solving semismooth equations, recognizing the estimating equation is semismooth for which the generalized Jocobian exists (*Chaney*, 1990; *Scholtes*, 2012). Specifically, denoting the initial value of $\boldsymbol{\tau}$ by $\widehat{\boldsymbol{\tau}}^{(0)}$, the t-th ($t \geq 1$) iteration of the two-step semismooth NR algorithm for generalized linear spline models proceeds as follows.

Step 1. Update estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ to obtain $\widehat{\boldsymbol{\beta}}^{(t-1)}$, $\widehat{\boldsymbol{\eta}}^{(t-1)}$. That is, treating $\widehat{\boldsymbol{\tau}}^{(t-1)}$ as fixed, fit the generalized linear regression model via MLE or quasi-likelihood method to obtain $\widehat{\boldsymbol{\beta}}^{(t-1)}$, $\widehat{\boldsymbol{\eta}}^{(t-1)}$ and predicted values $\widehat{\mu_i}^{(t-1)}$ $(i = 1, \ldots, n)$ for all subjects .

Step 2. Update $\widehat{\boldsymbol{\tau}}^{(t-1)}$ to obtain $\widehat{\boldsymbol{\tau}}^{(t)}$ by a modified NR procedure. Define a $K \times 1$ vector $U^{(t)} = (U_1^{(t)} \ldots U_J^{(t)})^T$, where $U_j^{(t)}$ is a $K_j \times 1$ vector and the p-th element of $U_j^{(t)}$, denoted as $U_{jp}^{(t)}$, is defined as

$$U_{jp}^{(t)} = \frac{\widehat{\beta}_{jp}^{(t-1)}}{n} \sum_{i=1}^{n} \frac{I(X_{ji} > \widehat{\tau}_{jp}^{(t-1)})}{v(\widehat{\mu_i}^{(t-1)}) g'(\widehat{\mu_i}^{(t-1)})} (Y_i - \widehat{\mu_i}^{(t-1)}), \text{ where } p = 1, \ldots, K_j.$$

Also define a $K \times K$ matrix $S^{(t)} = \begin{pmatrix} S_{11}^{(t)} & \dots & S_{1J}^{(t)} \\ \vdots & & \vdots \\ S_{J1}^{(t)} & \dots & S_{JJ}^{(t)} \end{pmatrix}$, where $S_{mn}^{(t)}$ is a $K_m \times K_n$ submatrix, $m, n \in \{1, \dots, J\}$, and the (p,q)-th element of $S_{mn}^{(t)}$, denoted as $S_{mn}^{p,q(t)}$, is defined as

$$S_{mn}^{p,q(t)} = \frac{\widehat{\beta}_{mp}^{(t-1)} \widehat{\beta}_{nq}^{(t-1)}}{n} \sum_{i=1}^{n} \frac{I(X_{mi} > \widehat{\tau}_{mp}^{(t-1)}) I(X_{ni} > \widehat{\tau}_{nq}^{(t-1)})}{v(\widehat{\mu}_i^{(t-1)}) \{g'(\widehat{\mu}_i^{(t-1)})\}^2}, \text{ where}$$

$p = 1, \dots, K_m;\ q = 1, \dots, K_n$. Then update estimate of $\boldsymbol{\tau}$ by $\widehat{\boldsymbol{\tau}}^{(t)} = \widehat{\boldsymbol{\tau}}^{(t-1)} - \{S^{(t)}\}^{-1} U^{(t)}$. From the modified derivative perspective, $U^{(t)}$ and $S^{(t)}$ are, respectively, proportional to the modified score function and modified observed information matrix with respect to $\boldsymbol{\tau}$ only, when $Y|\boldsymbol{x}^*$ is assumed to follow a distribution in the exponential family.

The above algorithm iterates between steps 1 and 2 until $\|\widehat{\boldsymbol{\tau}}^{(t)} - \widehat{\boldsymbol{\tau}}^{(t-1)}\| < \zeta$, where ζ is a pre-specified tolerance value. Once $\widehat{\boldsymbol{\tau}}^{(t)}$ converges, the final estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are obtained via another step 1, treating the final estimate $\widehat{\boldsymbol{\tau}}^{(t)}$ as fixed. The validity of the above algorithm is shown in the following Proposition 1.

Proposition 2. *The two-step semismooth NR algorithm converges locally and it converges to the proposed estimating equation* $\frac{1}{n} \sum_{i=1}^{n} Q(\boldsymbol{W}_i; \boldsymbol{\theta}) = 0$, *i.e.* $\widehat{\boldsymbol{\theta}}_n$.

According to simulation studies in Section 2.5, this two-step semismooth NR algorithm performs well and stable for continuous outcomes, and our simulation studies further support that conclusion for count outcomes. However, our simulation studies show that this two-step semismooth NR algorithm may lead to nonnegligible bias in logistic linear spline regression in some scenarios when the sample size is small and the initial value is not close to the truth. The reason for this unsatisfactory performance is that the two-step semismooth NR algorithm is sensitive to the initial

value of change points, i.e., $\widehat{\boldsymbol{\tau}}^{(0)}$, and the inherent difficulty in fitting a complicated model for a binary outcome where there is less information. To remedy this issue, we propose an objective function, which is the L1-norm of $\sum_{i=1}^{n} Q(\boldsymbol{W}_i; \widehat{\boldsymbol{\theta}}^{(t)})$, where $\widehat{\boldsymbol{\theta}}^{(t)}$ is the final estimate of $\boldsymbol{\theta}$ from the two-step semismooth NR algorithm. Then one is able to fit the model using multiple initial values and choose the one that minimizes the objective function. The likelihood function itself might be a natural objective function to consider and has been used as the selection criteria in the popular R package "segmented" (*Muggeo*, 2008). However, our simulation shows that the maximizer of the likelihood function sometimes is far away from the solution of the estimating equation, i.e., $\frac{1}{n}\sum_{i=1}^{n} Q(\boldsymbol{W}_i; \boldsymbol{\theta}) = 0$, and from the truth. As the proposed estimating function (3.5) is not the derivative of the likelihood function, solving the proposed estimating equation is not equivalent to maximizing the likelihood function. Our simulation studies show that using the likelihood function as the objective function sometimes even introduces more bias and larger variance than just using a fixed initial value , especially when the sample size is small and the outcome is binary.

3.4.3 Asymptotic Properties

This section studies the asymptotic properties of the proposed estimator. The main results are summarized in the following two theorems, with detailed proofs available in the Appendix.

Assumption 1. The link function $g(.)$ is monotonically increasing, continuous and first-order differentiable, and the working variance function $\phi v\{\mu(\boldsymbol{X}_i^*; \boldsymbol{\theta})\}$ is a positive function.

Theorem 3. *Under the semiparametric generalized linear spline model (3.1),* $\widehat{\boldsymbol{\theta}}_n$ *is a consistent estimator for* $\boldsymbol{\theta}^0$*, as* $n \to \infty$*.*

Theorem 4. *Under the semiparametric generalized linear spline model (3.1),* $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n -$

$\boldsymbol{\theta}^0)$ *converges in distribution to* $\mathcal{N}\Big(0, V_1^{-1}(\boldsymbol{\theta}^0)V_2(\boldsymbol{\theta}^0)V_1^{-1}(\boldsymbol{\theta}^0)\Big)$*, where*

$$\begin{aligned} V_1(\boldsymbol{\theta}^0) &= E\Big[\frac{H^T(\boldsymbol{X}^*;\boldsymbol{\theta}^0)H(\boldsymbol{X}^*;\boldsymbol{\theta}^0)}{v\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}g'\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}^2}\Big], \\ V_2(\boldsymbol{\theta}^0) &= E\Big[\frac{H^T(X;\boldsymbol{\theta}^0)V(Y|\boldsymbol{X}^*;\boldsymbol{\theta}^0,\phi^0)H(\boldsymbol{X}^*;\boldsymbol{\theta}^0)}{v\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}^2g'\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}^2}\Big]. \end{aligned}$$

If the working variance function $\phi v\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta})\}$ is correctly specified for the conditional variance $V(Y|\boldsymbol{X}^*;\boldsymbol{\theta},\phi)$, including the situation that $Y|\boldsymbol{x}^*$ arises from an exponential family, the asymptotic variance can be simplified as $\phi^0 V_1^{-1}(\boldsymbol{\theta}^0)$. According to the discussions in Section 3.3.2, the estimator $\widehat{\boldsymbol{\theta}}_n$ achieves the semiparametric local efficiency bound. If we additionally assume that $Y|\boldsymbol{x}^*$ arises from an exponential family with a canonical link, we can further simply the function $V_1(\boldsymbol{\theta}^0) = E[v\{\mu(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\}H^T(\boldsymbol{X}^*;\boldsymbol{\theta}^0)H(\boldsymbol{X}^*;\boldsymbol{\theta}^0)]$.

To make statistical inference, the asymptotic variance in Theorem 2 can be consistently estimated by $\widehat{\phi}\widehat{V}_1^{-1}(\widehat{\boldsymbol{\theta}}_n)\widehat{V}_2(\widehat{\boldsymbol{\theta}}_n)\widehat{V}_1^{-1}(\widehat{\boldsymbol{\theta}}_n)$, where

$$\begin{aligned} \widehat{V}_1(\widehat{\boldsymbol{\theta}}_n) &= \frac{1}{n}\sum_{i=1}^{n}\frac{H^T(\boldsymbol{X}_i^*;\widehat{\boldsymbol{\theta}}_n)H(\boldsymbol{X}_i^*;\widehat{\boldsymbol{\theta}}_n)}{v\{\mu(\boldsymbol{X}_i^*;\widehat{\boldsymbol{\theta}}_n)\}g'\{\mu(\boldsymbol{X}_i^*;\widehat{\boldsymbol{\theta}}_n)\}^2} \\ \widehat{V}_2(\widehat{\boldsymbol{\theta}}_n) &= \frac{1}{n}\sum_{i=1}^{n}\frac{H^T(\boldsymbol{X}_i^*;\widehat{\boldsymbol{\theta}}_n)V(Y|\boldsymbol{X}^*;\widehat{\boldsymbol{\theta}}_n,\widehat{\phi})H(\boldsymbol{X}_i^*;\widehat{\boldsymbol{\theta}}_n)}{v\{\mu(\boldsymbol{X}_i^*;\widehat{\boldsymbol{\theta}}_n)\}^2g'\{\mu(\boldsymbol{X}_i^*;\widehat{\boldsymbol{\theta}}_n)\}^2} \end{aligned}$$

The estimation of ϕ can be obtained via the method of moments estimator $\widehat{\phi} = \sum_{i=1}^{n}\frac{(Y_i-\widehat{\mu}_i)^2}{v(\widehat{\mu}_i)(n-q)}$, where q is the number of total parameters in model (3.1) and $\widehat{\mu}_i = \mu(\boldsymbol{X}_i^*;\widehat{\boldsymbol{\theta}}_n)$. The variance of the proposed estimator $\widehat{\boldsymbol{\theta}}_n$ can then be consistently estimated by $\widehat{\phi}\widehat{V}_1^{-1}(\widehat{\boldsymbol{\theta}}_n)\widehat{V}_2(\widehat{\boldsymbol{\theta}}_n)\widehat{V}_1^{-1}(\widehat{\boldsymbol{\theta}}_n)/n$. If $Y|\boldsymbol{x}^*$ arises from an exponential family, the variance of $\widehat{\boldsymbol{\theta}}_n$ can be consistently estimated by $\widehat{\phi}\widehat{V}_1^{-1}(\widehat{\boldsymbol{\theta}}_n)/n$. In particular, we note for Poisson and logistic regression, ϕ is a constant and is equal to 1. For linear regression with normally distributed outcomes, we can estimate $\phi = \sigma^2$ via

$\widehat{\sigma}^2 = \{\sum_{i=1}^{n}(Y_i - \widehat{\mu}_i)^2\}/\{n-q\}$.

3.5 Simulation Studies

This section reports simulation studies conducted to evaluate the proposed method and to compare it with the method of *Muggeo* (2003). We considered logistic and Poisson linear spline regression for binary and count outcomes respectively, with one or two factors of interest. Data were generated under five models, three for logistic regression and two for Poisson regression. Outcome models and initial values of change points used in algorithms are listed in Table 3.1. In all scenarios, the covariate Z was generated from $\mathcal{N}(0,1)$ with coefficient $\eta = 0.2$. The factors of interest X_1 and X_2 followed uniform $(0,1)$ and uniform $(0,2)$ distributions, respectively, and they both correlated with Z, making Z a confounder for both effects of X_1 and X_2 on Y. Specifically, X_1 were generated as $\Phi((V_1+Z)/\sqrt{2})$ and X_2 as $2\times\Phi((V_2+Z)/\sqrt{5})$, where Φ is the cumulative distribution function of the standard normal distribution and the intermediate variables V_1 and V_2, both independent of all other variables, were from $\mathcal{N}(0,1)$ and $\mathcal{N}(0,4)$ respectively. In all scenarios, methods were evaluated under 1000 Monte Carlo replicates with sample sizes $n = 200, 500, 1000$ or 2500 for logistic linear spline models, and $n = 500, 1000, 2500$ or 5000 for Poisson linear spline models. The method of *Muggeo* (2003) was implemented using the R package "segmented" (*Muggeo*, 2008). In both methods, the tolerance of convergence was set as $\zeta = 10^{-11}$ and the largest number of iterations was set as 500. We used the same fixed initial values of change points (different from the truth) for both the proposed and segmented methods. Fixed initial values of change points are listed in Table 3.1 as ".fix" and simulation results are summarized in Table 3.2 and Table 3.3 as ".fix". For logistic linear spline models where the model fitting is more challenging, we also studied the performance of both methods by using the truth as the fixed initial values with results summarized in Table 3.2 as ".true". Additionally, in the setting of logistic models, we

Table 3.1: Data generating models and the corresponding true parameter values and initial values of change points in simulations.

Logistic 1	$\text{logit}(\mu) = \beta_0 + \beta_1 X_1 + \beta_{11}(X_1 - \tau_{11})^+ + \eta Z$ $(\beta_0, \beta_1, \beta_{11}, \tau_{11}) = (0.1, -5, 9, 0.4)$ Initial values of change point: ".fix" : $\widehat{\tau_{11}}^{(0)} = 0.45$ ".choose": $\widehat{\tau_{11}}^{(0)} = (0.35, 0.4, 0.45)$
Logistic 2	$\text{logit}(\mu) = \beta_0 + \beta_1 X_2 + \beta_{11}(X_2 - \tau_{21})^+ + \eta Z$ $(\beta_0, \beta_1, \beta_{11}, \tau_{11}) = (-1, 3, -10, 1.5)$ Initial values of change point: ".fix" : $\widehat{\tau_{11}}^{(0)} = 1.4$ ".choose": $\widehat{\tau_{11}}^{(0)} = (1.4, 1.5, 1.6)$
Logistic 3	$\text{logit}(\mu) = \beta_0 + \beta_1 X_1 + \beta_{11}(X_1 - \tau_{11})^+ + \beta_2 X_2 + \beta_{21}(X_2 - \tau_{21})^+ + \eta Z$ $(\beta_0, \beta_1, \beta_{11}, \beta_2, \beta_{21}, \tau_{11}, \tau_{21}) = (-0.5, 6, -10, 5, -11, 0.45, 1)$ Initial values of change point: ".fix" : $\widehat{\tau_{11}}^{(0)} = 0.5, \quad \widehat{\tau_{21}}^{(0)} = 0.9$ ".choose": $\widehat{\tau_{11}}^{(0)} = (0.4, 0.45, 0.5), \quad \widehat{\tau_{21}}^{(0)} = (0.9, 1, 1.1)$
Poisson 1	$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_{11}(X_1 - \tau_{11})^+ + \beta_{12}(X_1 - \tau_{12})^+ + \beta_{13}(X_1 - \tau_{13})^+ + \beta_{14}(X_1 - \tau_{14})^+ + \eta Z$ $(\beta_0, \beta_1, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \tau_{11}, \tau_{12}, \tau_{13}, \tau_{14}) = (4, 3, -5, 4, -3, 4, 0.2, 0.4, 0.6, 0.8)$ Initial values of change point: ".fix" : $\widehat{\tau_{11}}^{(0)} = 0.15, \quad \widehat{\tau_{12}}^{(0)} = 0.35, \quad \widehat{\tau_{13}}^{(0)} = 0.55, \quad \widehat{\tau_{14}}^{(0)} = 0.75$
Poisson 2	$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_{11}(X_1 - \tau_{11})^+ + \beta_{12}(X_1 - \tau_{12})^+ + \beta_2 X_2 + \beta_{21}(X_2 - \tau_{21})^+ + \beta_{22}(X_2 - \tau_{22})^+ + \eta Z$ $(\beta_0, \beta_1, \beta_{11}, \beta_{12}, \beta_2, \beta_{21}, \beta_{22}, \tau_{11}, \tau_{12}, \tau_{21}, \tau_{22}) = (3, 2, -5, 4, -2, 6, -8, 0.4, 0.7, 0.8, 1.4)$ Initial values of change point: ".fix" : $\widehat{\tau_{11}}^{(0)} = 0.45, \quad \widehat{\tau_{12}}^{(0)} = 0.8, \quad \widehat{\tau_{21}}^{(0)} = 0.9, \quad \widehat{\tau_{22}}^{(0)} = 1.5$

studied the performance of the proposed method by choosing results from three initial values of change points (denoted as ".choose") via minimizing the objective function, i.e. L1-norm of the estimated proposed estimating equation $\sum_{i=1}^{n} Q(\boldsymbol{W}_i; \widehat{\boldsymbol{\theta}}^{(t)})$, where $\widehat{\boldsymbol{\theta}}^{(t)}$ is the final estimate of $\boldsymbol{\theta}$ from the two-step semismooth NR algorithm.

Results for logistics and Poisson linear spline regression are shown in Table 3.2 and Table 3.3 respectively. For all scenarios, bias, Monte Carlo standard deviation (MCSD), the average of standard errors (AVESE), root mean squared errors (RMSE) and coverage probabilities (CP) of the 95% confidence intervals are reported. First of all, both methods have high convergence rates of the algorithm, both with a 100% convergence rate in all scenarios except for logistic scenario 1 ($n = 200$) and logistic scenario 3 ($n = 200, 500$). Even in these scenarios, the convergence rates are reasonably high. In terms of statistical properties, both methods perform well in terms of biases, with biases close to zero.

Table 3.2: Simulation results for logistic linear spline regression models, where "proposed" denotes the proposed method, "segmented" denotes the method of *Muggeo* (2003), * indicates value $\times 10^{-3}$, ".fix" represents initial values are fixed at values different from the truth, ".truth" represents initial values are fixed at the truth and ".choose" represents initial values are not fixed. MCSD: Monte Caro standard deviation; RMSE: root mean squared error; AVESE: average of standard error; %CP: coverage probability of the 95% confidence intervals.

Method	**n**	**Bias***	**MCSD***	**MSE***	**AVESE***	**%CP**	**Bias***	**MCSD***	**MSE***	**AVESE***	**%CP**
		Logistic 1					**Logistic 2**				
proposed.fix	200	18.60	70.36	72.78	86.09	93.8	-17.10	133.02	134.11	132.77	91.3
segmented.fix		5.12	135.69	135.78	82.90	79.5	-35.50	203.38	206.45	130.18	83.8
proposed.true		1.40	68.24	68.25	84.67	93.9	8.38	123.13	123.41	129.84	91.1
segmented.true		0.16	131.88	131.88	81.91	78.5	-23.39	197.30	198.68	134.69	83.5
proposed.choose		5.57	71.40	71.61	85.46	93.8	1.18	130.61	130.61	214.42	90.7
proposed.fix	500	11.47	51.39	52.65	52.56	93.5	-17.14	81.70	83.47	82.96	94.9
segmented.fix		3.18	70.72	70.79	52.29	87.7	-0.59	93.32	93.32	81.41	89.7
proposed.true		0.15	47.53	47.53	52.25	94.1	-1.46	75.45	75.47	81.69	95.0
segmented.true		2.70	69.97	70.02	52.57	86.8	-11.12	109.85	110.42	82.18	89.4
proposed.choose		3.74	51.39	51.53	52.32	93.8	-9.80	82.01	82.60	82.29	94.6
proposed.fix	1000	6.70	39.45	40.01	37.01	93.5	-7.37	65.42	65.83	57.95	92.8
segmented.fix		2.20	48.81	48.86	36.74	88.1	0.78	65.90	65.91	57.20	90.3
proposed.true		-1.27	35.38	35.40	36.87	94.3	1.78	59.57	59.59	57.46	93.2
segmented.true		-4.64	44.72	44.97	37.48	89.0	1.57	57.40	57.42	57.02	94.7
proposed.choose		0.08	39.28	39.28	36.86	93.5	-2.20	65.06	65.09	57.65	92.5
proposed.fix	2500	3.73	25.78	26.05	23.48	91.6	-4.64	40.39	40.66	36.38	91.9
segmented.fix		-0.84	25.92	25.93	23.36	91.3	1.59	37.07	37.11	36.07	92.6
proposed.true		-0.39	22.65	22.65	23.43	94.3	-0.83	37.25	37.26	36.26	93.6
segmented.true		5.72	20.07	20.87	22.75	99.5	-2.49	40.66	40.74	36.35	92.2
proposed.choose		0.48	25.52	25.53	23.43	91.6	-2.18	40.47	40.53	36.29	91.5
		Logistic 3									
		τ_{11}					τ_{21}				
proposed.fix	200	17.16	120.23	121.45	137.50	90.4	-29.83	194.84	197.11	194.90	91.2
segmented.fix		11.22	212.88	213.17	188.17	71.2	-9.03	228.32	228.50	191.67	86.7
proposed.true		7.97	117.50	117.77	137.78	91.1	-21.00	195.44	196.56	194.65	91.4
segmented.true		-3.87	215.73	215.76	660.63	68.1	-16.48	239.87	240.43	197.28	84.7
proposed.choose		12.28	121.73	122.35	137.83	90.7	-14.48	196.82	197.35	194.57	91.5
proposed.fix	500	4.50	81.21	81.34	87.03	91.9	-5.68	122.46	122.59	122.19	93.7
segmented.fix		1.98	119.19	119.21	84.26	83.9	1.57	129.19	129.20	122.79	92.2
proposed.true		0.83	79.37	79.38	86.77	91.9	-0.16	122.13	122.13	122.11	93.7
segmented.true		-18.56	109.62	111.18	81.03	74.3	-30.43	158.55	161.45	120.11	90.3
proposed.choose		0.73	82.43	82.43	86.56	91.2	-0.47	124.51	124.51	122.03	93.7
proposed.fix	1000	1.76	63.90	63.92	60.81	92.8	-5.84	90.88	91.07	86.20	92.7
segmented.fix		-5.48	82.90	83.08	58.92	81.6	-2.78	89.86	89.91	86.17	93.5
proposed.true		0.57	61.13	61.13	60.87	93.4	-2.00	89.23	89.26	86.20	93.2
segmented.true		-6.74	82.45	82.73	58.94	81.8	-2.00	89.68	89.70	86.11	93.7
proposed.choose		0.75	65.28	65.28	60.80	91.9	-1.56	92.18	92.20	86.15	92.6
proposed.fix	2500	0.43	41.60	41.60	38.54	91.7	-0.60	54.94	54.95	54.87	94.7
segmented.fix		0.99	44.50	44.51	38.04	89.8	-3.42	53.27	53.38	54.82	95.9
proposed.true		3.22	38.57	38.71	38.63	93.5	1.95	53.39	53.42	54.87	95.0
segmented.true		0.81	49.39	49.40	38.34	88.1	-1.00	58.81	58.82	54.84	93.6
proposed.choose		3.29	43.07	43.19	38.58	90.7	1.75	55.29	55.32	54.86	94.4

Differences in performances of the two methods exist. In general, MCSD and RMSE of the proposed method are smaller than the segmented method and the coverage

Table 3.3: Simulation results for Poisson linear spline regression models, where "proposed" denotes the proposed method, "segmented" denotes the method of *Muggeo* (2003) , * indicates value $\times 10^{-3}$ and ".fix" represents initial values are fixed at values different from the truth. MCSD: Monte Caro standard deviation; RMSE: root mean squared error; AVESE: average of standard error; %CP: coverage probability of the 95% confidence intervals.

Method	**n**	**Bias***	**MCSD***	**RMSE***	**AVESE***	**%CP**	**Bias***	**MCSD***	**RMSE***	**AVESE***	**%CP**
						Poisson 1					
				τ_{11}					τ_{12}		
proposed.fix	500	-0.84	9.59	9.63	8.83	92.9	1.72	16.52	16.61	14.38	90.4
segmented.fix		0.13	10.61	10.62	8.88	92.0	-0.68	16.46	16.48	14.35	88.3
proposed.fix	1000	-0.36	6.80	6.81	6.20	92.0	0.98	11.22	11.27	10.07	93.1
segmented.fix		0.91	6.50	6.57	6.17	92.1	-1.36	10.77	10.86	10.02	91.1
proposed.fix	2500	-0.10	3.89	3.89	3.88	95.4	0.38	6.65	6.66	6.32	93.9
segmented.fix		0.61	3.81	3.86	3.88	93.2	-0.92	6.63	6.69	6.30	92.7
proposed.fix	5000	-0.13	2.78	2.79	2.74	94.9	0.38	4.48	4.50	4.47	94.6
segmented.fix		0.03	2.92	2.92	2.75	93.4	0.07	4.76	4.76	4.46	93.9
				τ_{13}					τ_{14}		
proposed.fix	500	0.70	11.66	11.68	11.60	95.2	0.23	4.91	4.92	4.82	94.5
segmented.fix		1.46	10.92	11.02	11.60	95.1	0.20	5.21	5.21	4.83	93.7
proposed.fix	1000	0.12	8.41	8.41	8.17	94.0	0.09	3.43	3.43	3.38	94.1
segmented.fix		0.52	10.07	10.08	8.17	89.3	0.28	3.64	3.65	3.37	90.7
proposed.fix	2500	0.11	5.20	5.20	5.14	94.7	0.04	2.17	2.17	2.12	95.5
segmented.fix		-0.81	6.41	6.46	5.14	83.3	-0.21	2.30	2.31	2.12	93.3
proposed.fix	5000	0.08	3.74	3.74	3.64	94.0	-0.01	1.47	1.47	1.49	95.3
segmented.fix		-0.14	3.70	3.70	3.64	94.1	0.12	1.59	1.60	1.50	93.1
						Poisson 2					
				τ_{11}					τ_{12}		
proposed.fix	500	-1.17	7.17	7.27	6.47	92.8	1.54	9.00	9.13	8.40	91.9
segmented.fix		0.04	7.40	7.40	6.45	90.5	0.31	9.04	9.05	8.43	93.0
proposed.fix	1000	-0.55	4.87	4.90	4.54	92.8	1.01	6.12	6.21	5.92	93.1
segmented.fix		0.05	4.77	4.77	4.55	93.1	-0.31	6.53	6.53	5.93	91.7
proposed.fix	2500	-0.36	3.00	3.02	2.86	93.5	0.75	3.82	3.89	3.72	93.2
segmented.fix		0.02	3.02	3.02	2.88	93.5	-0.41	4.03	4.05	3.74	93.0
proposed.fix	5000	-0.29	2.10	2.12	2.03	93.4	0.49	2.77	2.81	2.64	92.7
segmented.fix		0.13	1.97	1.98	2.02	94.4	0.01	2.76	2.76	2.63	94.5
				τ_{21}					τ_{22}		
proposed.fix	500	-1.74	10.63	10.77	9.63	91.1	0.03	7.60	7.60	7.02	92.4
segmented.fix		-0.13	11.94	11.94	9.70	89.8	-0.05	7.78	7.79	7.03	91.2
proposed.fix	1000	-1.07	7.58	7.66	6.82	90.6	0.32	5.41	5.42	4.95	91.7
segmented.fix		0.35	7.49	7.50	6.81	92.0	-0.22	5.42	5.42	4.95	91.8
proposed.fix	2500	-0.67	4.62	4.67	4.28	92.5	0.23	3.31	3.32	3.12	92.8
segmented.fix		0.25	4.80	4.81	4.32	91.2	0.16	3.51	3.52	3.11	91.0
proposed.fix	5000	-0.33	3.17	3.19	3.03	94.3	0.07	2.31	2.31	2.20	93.4
segmented.fix		-0.38	3.32	3.34	3.02	94.3	-0.09	2.22	2.23	2.20	93.9

probabilities of the proposed method are better than those of the segmented method in most cases, especially in logistic regression. When initial values of change points were chosen at the truth in logistic regression, the performance of the "proposed.true" method is good with small bias and RMSE and accurate statistical inference. How-

ever, "segmented.true" method performs less desirable, especially in terms of MCSD and coverage probability. When the sample size is small, the MCSD of the "segmented.true" method is much larger than the AVESE, and it is almost twice as large as the MCSD of the "proposed.true" method in some cases, for example in logistic scenario 1 with sample size 200. As the sample size increases, the MCSD of the "segmented.true" method becomes more reasonable. Of note, the coverage probability of the "segmented.true" method may be much smaller or larger than the nominal level, even with a large sample size. For example, in logistic scenario 3, when the sample size is $n = 200$, the coverage probability of the "segmented.true" method for estimating τ_{11} is only 68.1%. Even when the sample size is $n = 2500$, the coverage probability is still only 88.1%. In contrast, in logistic scenario 1, when the sample size is $n = 2500$, the coverage probability of the "segmented.true" method for estimating is 99.5%, much higher than the nominal level. As expected, for both methods in logistic models, performances of ".true" are generally similar but better than ".fix" in terms of smaller bias, smaller MSE and better coverage probability. Performances of ".true" represent the ideal case and are not realistic in practice. We note that choosing initial values based on the L-1 norm leads to an improved performance of the "proposed.choose" relative to the "proposed.fix" in logistic scenarios. Overall, "proposed.choose" has smaller bias and smaller MSE than "proposed.fix" method, especially when the sample size is small. For example, in logistic scenario 1, when the sample size is $n = 200$, the bias from the "proposed.fix" method is less than one third of the bias from the "proposed.choose" method. As researchers do not know the truth for change points in real applications, the "proposed.choose" method is a more practical than the "proposed.true".

In settings of Poisson regression, both the proposed and the segmented methods perform well. The issue of under-coverage for the segmented method is less severe for Poisson regression but still exists. For example, in Poisson scenario 1, when $n = 1000$,

the coverage probability for τ_{13} is 83.3%. The lower or higher coverage issue of the segmented method is likely due to finite-sample property, computational issue and/or failure to account for certain variability in the inferential procedure, as the theory behind this method is not rigorously studied.

3.6 Application

We applied the proposed method to adult patients who underwent percutaneous coronary intervention (PCI) from 2007 to 2009 at 33 hospitals in Michigan, using data from the Blue Cross Blue Shield of Michigan Cardiovascular Consortium (*Kline-Rogers et al.*, 2002). We are interested in understanding the effect of baseline glomerular filtration rate (GFR) and body mass index (BMI) on in-hospital transfusion or gastrointestinal (GI) bleeding after PCI. GFR is a measure of kidney function, with a larger value indicating a better condition. According to *Centers for Disease Control and Prevention* (2020), for adults BMI < 18.5, ≥ 18.5 and < 25, ≥ 25 and < 30, and ≥ 30 indicate underweight, normal weight, overweight and obese respectively. Our analysis is based on 63,156 patients with the priority of cardiac status (urgent v.s. nonurgent) available and whose height is between 100 and 250 centimeters. Among all patients, 2602 (4.12%) experienced either in-hospital transfusion or GI bleeding.

We applied the logistic linear spline model to investigate the associations. After a series of model fitting and checking, in the final model BMI was modeled with one change point, τ_{11}, and GFR was modeled with two change points, τ_{21} and τ_{22}. Patient characteristics adjusted in the analysis included age, female (vs. male), white (vs. other race), black (vs. other race), current smoking status, history of hypertension, diabetes, previous PCI, previous coronary artery bypass grafting (CABG), priority of cardiac status: emergent (vs. nonurgent) and priority of cardiac status: urgent (vs. nonurgent). Results on the fitted model are reported in Table 3.4. The concordance

Table 3.4: The PCI trial: results from the logistic linear spline model for the effect of BMI and GFR on in-hospital transfusion/GI-bleeding. The p-value is calculated based on the two-sided Wald test, treating all change-points fixed at the estimated values. Error represents the standard error of log odds ratio.

		Log Odds Ratio	**Std. Error**	**P-Value**
	Intercept	-1.100	0.292	<0.001
Slope	BMI≤ 32.3	-0.017	0.006	0.008
	BMI>32.3	0.053	0.005	<0.001
	GFR≤54.2	-0.031	0.003	<0.001
	54.2<GFR≤114.7	-0.014	0.002	<0.001
	GFR>114.7	-0.007	0.002	<0.001

statistic for the fitted logistic linear spline model is 0.78, indicating a good model fit.

The estimated change-point for the effect of BMI is located at 32.3 with a 95% confidence interval (95%CI: 29.1, 35.4). The slopes before and after the change-point are significantly different (p<0.001). Given all other covariates, when BMI is less than 32.3, the odds ratio of in-hospital transfusion/GI bleeding per 1 unit increase in BMI is 0.98 (p=0.008); when BMI is greater than 32.3, the odds ratio is 1.05 (p<0.001). Roughly and intuitively, these results suggest that, for patients who are not obese, higher BMI is protective in terms of bleeding risk. However, for patients who are obese, the larger the BMI the higher the risk of bleeding. As for the effect of GFR, when GFR is lower than 54.2, which corresponds to patients with moderate or severe loss of kidney function, higher GFR and thus better kidney function are associated with less bleeding (odds ratio=0.97; p<0.001). When GFR is between 54.2 and 114.7, better kidney function is still associated with lower bleeding (odds ratio=0.986; p<0.001) but the effect is relatively smaller. When GFR is greater than 114.7, corresponding to patients with normal and high kidney function, the effect of kidney function is further smaller, but still better kidney function is protective in terms of bleeding (odds ratio=0.993; p<0.001).

3.7 Discussion

Generalized linear spline models have important applications in studying threshold effects and change-points when effects of continuous factors are nonlinear and change at some change-points. Often times, it is the change-points/knots that are of main interest. However, due to the lack of rigorously studied and computationally convenient methods, in practice one has to pre-specify change points in order to fit the generalized linear model and then choose the change points in an ad-hoc manner through model selections. Other than being nonrigorous, this approach does not allow one to quantify the uncertainty in this process or to make inference on change points. Although there has been a lot of methodology developments, in terms of both theory and computational algorithms, for estimation of change points in linear spline models, there is little rigorous study on change-points estimation in generalized linear spline models. In this article, we attempt to fill in this important gap in literature and study estimation of change points, as well as other unknown parameters, in generalized linear spline models with multidimensional predictors, which allows rigorous estimations and inferences on change points.

The lack of rigorously studied methods for change-points estimation in generalized linear spline models is likely due to the inherent difficulty in generalized linear spline models. As discussed in Section 3.1, several smoothing-based methods have been proposed and rigorously studied in the setting of linear spline models. Smoothing is a sound and natural idea, and smoothing-based methods have been shown to be more computational efficient than earlier grid-search type algorithms. However, its extension to generalized linear spline models has not been successful so far. Our own study on trying to adapting the smoothing-based algorithms to generalized linear spline models has shown that it suffers from sensitivity to initial values and difficulty in convergence. In contrast to smoothing, the proposed method is based on a simple idea

of modified derivatives, i.e., redefining the derivative whenever it is not well-defined in the usual sense. The modified derivative idea is first applied to the likelihood or quasi-likelihood function, leading to a simple estimating equation that does not involve any smoothing parameters. The modified derivative idea is then applied again to solve this simple but nonregular and nondifferentiable estimating equation. It is interesting to note that, in the latter case, the proposed modified derivative is a special case of generalized Jacobian studied in optimization literature in solving semismooth equations. As discussed previously, numerical instability has been a major obstacle in estimation of change points for generalized linear spline models. To overcome this, a two-step semismooth NR algorithm is proposed to solve the estimating equation. The algorithm separates updating of change-points and other parameters, greatly improving computational performances in terms of convergence rates and sensitivity to initial values. Our simulation studies have shown that the computational algorithm is fast and convenient, and numerically stable.

The asymptotic properties of the proposed estimator have been rigorously studied using the empirical process theory. We have shown that the proposed estimator is root-n consistent and asymptotically normal. When the model belongs to the exponential family, the proposed estimator is also asymptotically efficient. To the best of our knowledge, there has been no previous work that studied the statistical properties of estimators of change points rigorously in the setting of generalized linear spline models. A popular computational algorithm for change-points estimation for generalized linear spline models was studied in *Muggeo* (2003) and implemented in *Muggeo* (2008). However, they did not study the statistical properties rigorously. In terms of empirical performances, our simulation studies show that the proposed method is comparable or better than the highly-cited method of *Muggeo* (2003, 2008). In particular, the statistical inference of the proposed method is more accurate with better coverage probabilities, especially in logistic regression models for binary outcomes.

The improvement in statistical inferences is due to the rigorous study of asymptotic properties.

In summary, this chapter proposed an estimation method of change-points and all other parameters in the generalized linear spline model with a computationally efficient and easy-implemented two-step algorithm. We note that our study of influence functions using the modified derivative idea is generic that can be applied to many other research topics, such as correlated or clustered data, survival data, longitudinal data, and even some irregular problems beyond change-point estimation.

CHAPTER IV

Modeling and Estimating a Threshold Effect: An Application to Improving Cardiac Surgery Practices

4.1 Introduction

Coronary artery bypass grafting (CABG) is the most common type of cardiac surgery in the United States, with more than 400,000 procedures annually (*Dasta et al.*, 2008). While post-procedural outcomes have improved over time, patients continue to experience a number of adverse sequelae. One of the most recognized complications after CABG is post-procedural acute kidney injury (AKI), which is associated with increased morbidity, mortality and costly long-term treatment (*Robert et al.*, 2010; *Elmistekawy et al.*, 2014; *Shen et al.*, 2017; *Alshaikh et al.*, 2018). Despite numerous studies in recent years, development of evidence-based guidelines for cardiopulmonary bypass (CPB) practice for avoiding post-operative AKI has been limited. In recent years, the effect of lowest/nadir DO2 during CABG has been studied in relation to its influence on post-operative AKI in cardiovascular literature (*Ranucci et al.*, 2005; *De Somer et al.*, 2011; *Magruder et al.*, 2015). Researchers believe that there is a lower threshold of nadir DO2 below which patients are subject to higher risk of post-

operative AKI. A plausible mechanism is that when the nadir DO2 during CPB falls below a critical value, the oxygen supply is insufficient to meet the oxygen demand and this triggers dysoxia and lactic acidosis, leading to impaired post-operative renal function. *Ranucci et al.* (2005) and *De Somer et al.* (2011), two influential data-driven studies in cardiac surgery, suggested a critical value of 272 mL/minute/m^2 and 262 mL/minute/m^2 respectively, in order to reduce post-operative AKI. Their recommended thresholds have been used as guidelines for later studies on CPB trials (*Ranucci et al.*, 2015; *Mukaida et al.*, 2019; *Ranucci et al.*, 2018; *Wahba et al.*, 2020).

Results from *Ranucci et al.* (2005) and *De Somer et al.* (2011) fill in a gap in the literature in management of DO2. The analytical method used in these two papers is a popular approach in practice to identify a threshold. However, from a statistical methodology point of view, whether the analytic approach is appropriate is subject to question and we argue that it can potentially lead to biased results. In this approach, to find the threshold of nadir DO2, one first dichotomizes the continuous variable of DO2 into a binary variable at a certain value, and fits a logistic regression model using this binary variable. Then one repeats the process by dichotomizing DO2 at a series of different values and the threshold of DO2 is identified as the one that leads to the best model fit based on some criterion. This approach may be subject to bias because this model assumes the effect of DO2 on AKI changes in a noncontinuous way (Figure 4.2.1A), which can be unrealistic. That is, it assumes the risk of AKI jumps suddenly once the DO2 is lower than a threshold and stays constant thereafter. As estimating the lower threshold of nadir DO2, estimating a threshold has been a general and important research question arising from many applications. The modeling assumption has important implications on estimation of the threshold, which deserves a more detailed discussion and careful study.

The ad-hoc method described above is widely used in practice for estimating a thresh-

old; however, its statistical property has not been well-studied. More rigorous and general methods that can be adopted for estimating the threshold are studied in the statistical literature. A threshold is a special case of a change-point, at which the effect of a factor of interest changes. A (generalized) linear spline model is a formal statistical model used to model a change-point effect (*Marsh and Cormier*, 2001). In this model, linear segments are joined at a change-point. It is also often referred to as a broken-stick model and a segmented model (Figure 4.2.1B). Estimating change-points in a broken-stick model for a continuous outcome has been rigorously studied (*Quandt*, 1958; *Hudson*, 1966; *Feder*, 1975b,a; *Hansen*, 2000; *Das et al.*, 2016). These methods focus mainly on the theoretical aspects. For general outcomes including, e.g., count and binary outcomes, several algorithms are available to estimate change-points (*Muggeo*, 2003, 2008; *Fong et al.*, 2017). Although the theoretical properties have not been rigorously studied, these methods are popular in practice due to the availability of R packages and computational convenience. Methods described above are based on the broken-stick model, which assumes two linear segments with different slopes are joined together at the change-point without imposing other constraints. In toxicity study, for example, one may believe that there is a minimum tolerance level/threshold below which no one will respond. A family of threshold models, including the linear-plateau model and the hockey-stick model, are studied to model the dose-response relationship in environmental biology and toxicology (*Yanagimoto and Yamamoto*, 1979; *Cox*, 1987; *Hayes and Loomis*, 1996). Unlike the general methods based on splines for estimating change-points, these models are tailored to incorporate the subject-area knowledge. Similarly, in the management of DO2 for patients undergoing CABG, it is believed that only when the oxygen delivery is below a critical level, it is associated with an increased risk of AKI. Therefore, it may be preferable to incorporate this knowledge into modeling to help better estimate the threshold. Other relevant work on modeling and/or estimating change-points or

thresholds include, for example, *Chen et al.* (2011); *Elder and Fong* (2019); *Elliott and Shope* (2003); *Fong et al.* (2017); *Fong* (2019); *Lee* (2021); *Pastor-Barriuso et al.* (2003); *Tapsoba et al.* (2020).

Using the nadir DO2 example as a case study, this paper discusses and compares various models and methods that can be used to estimate a threshold, and demonstrates the potential bias of popular existing methods. In Section 4.2, we describe two existing models, namely, the sudden-jump model and the broken-stick model, and common estimation methods under each model. Then we introduce a constrained broken-stick model and propose a computationally convenient two-step algorithm to estimate the threshold based on the development of Chapter II and Chapter III. Comprehensive simulation studies are reported in Section 4.3 to compare models and methods. Using data from University of Michigan Frankel Cardiovascular Center, we aim to estimate the lower threshold of DO2 during CPB in CABG patients to help develop evidence-based guidelines for improving cardiac surgery practices.

4.2 Models and Methods

4.2.1 Notation

Consider a study with n subjects. For each subject i, $i = 1, ..., n$, let Y_i be the outcome, X_i be the factor of interest, for which a threshold effect is believed to exist, and $\boldsymbol{Z}_i$ be all other covariates to be adjusted for. That is, X_i is assumed to have an effect on the outcome Y_i when X_i is above or below a certain threshold τ. The outcome Y_i can be continuous, binary, count, or of other types. To accommodate different types of outcomes, we assume the outcome $Y_i|X_i, \boldsymbol{Z}_i$ can be modeled using the generalized linear regression model (GLM), and that $Y_i|X_i, \boldsymbol{Z}_i$ follows a distribution from an

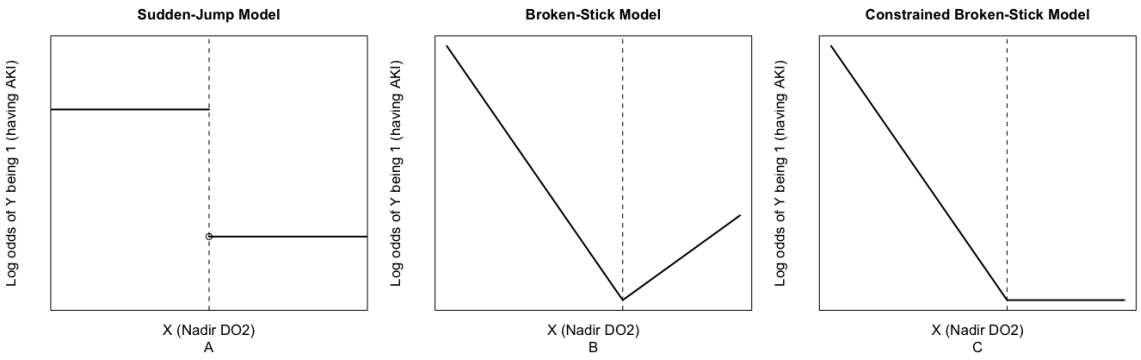

Figure 4.1: Three possible models: sudden-jump model, broken-stick model and constrained broken-stick model with logit link.

exponential family as follows

$$f(Y = y; \psi, \phi) = \exp\Big\{\frac{y\psi - b(\psi)}{\phi} + c(y; \phi)\Big\},$$

where ϕ is the scale parameter and ψ is the natural parameter. The conditional mean and variance of $Y_i|X_i, \boldsymbol{Z}_i$ are denoted as $\mu_i = E(Y_i|X_i, \boldsymbol{Z}_i)$ and $V(Y_i|X_i, \boldsymbol{Z}_i)$ respectively. It is a standard result in GLM that when the conditional mean μ_i is modeled via a canonical link function $g(\mu_i)$, the conditional variance can be expressed as $V(Y_i|X_i, \boldsymbol{Z}_i) = \phi v(\mu_i)$, where $v(\mu_i) = 1/g'(\mu_i)$. For example, $g(\mu_i) = \log(\mu_i)$ and $v(\mu_i) = \mu_i$ in Poisson regression model, $g(\mu_i) = \log(\frac{\mu_i}{1-\mu_i})$ and $v(\mu_i) = \mu_i(1 - \mu_i)$ in logistic regression model, and $g(\mu_i) = \Phi^{-1}(\mu_i)$ and $v(\mu_i) = 1$ in normal regression model.

Interest lies in estimating the threshold τ based on the observed data $(Y_i, X_i, \boldsymbol{Z}_i)$, independent and identically distributed across i. In particular, in our motivating study we are interested in identifying the lower threshold of DO2 below which there is an increased risk of AKI to provide guidance on good intraoperative practice on management of DO2 during a cardiac surgery. In our application, Y_i ($Y_i = 0$ or 1) denotes the binary outcome of whether or not subject i develops an AKI post surgery and X_i is the nadir DO2 level for subject i during the surgery. Below we discuss

different models to model μ_i and their implications.

4.2.2 Sudden-jump model

To learn the threshold from the data, due to its simplicity, perhaps the most popular model used in practice is a regression model including a dichotomous version of X_i as an independent variable, while adjusting for other covariates. Specifically, the conditional mean μ_i is modeled as follows,

$$g(\mu_i) = \beta_0 + \beta_1 I(X_i \leq \tau) + \boldsymbol{\eta}^T \boldsymbol{Z}_i = \begin{cases} (\beta_0 + \beta_1) + \boldsymbol{\eta}^T \boldsymbol{Z}_i & \text{if } X_i \leq \tau \\ \beta_0 + \boldsymbol{\eta}^T \boldsymbol{Z}_i & \text{otherwise.} \end{cases} \tag{4.1}$$

We refer to model (4.1) as a sudden-jump model. In model (4.1), for any given τ and $\boldsymbol{Z}_i$, $g(\mu_i)$ is modeled as a step function of X_i (nadir DO2 level). Specifically, in a logistic sudden-jump model for a binary outcome Y_i (AKI), this model assumes that for a patient with covariate $\boldsymbol{Z}_i$, the risk of developing AKI is a constant when the nadir DO2 is above the threshold τ and, when nadir DO2 reaches the lower threshold, the risk of AKI jumps and remains the same with further decreasing nadir DO2. Figure 4.2.1A depicts such an effect on the probability of AKI for a given subject.

The logistic sudden-jump model (4.1) was adopted in *Ranucci et al.* (2005) and *De Somer et al.* (2011), two influential studies in cardiac surgery, to identify the lower threshold of nadir DO2 using a data-drive approach. As for estimating the threshold τ, *Ranucci et al.* (2005) and *De Somer et al.* (2011) used the receiver operating characteristic (ROC) curve to measure the performance of the fitted model (4.1) under different values of τ, and estimated τ as the value leading to the best performance. Specifically, model (4.1) is fitted under a series of values of τ between $(\min_i(X_i), \max_i(X_i))$ or a possible range determined by clinicians based upon their clinical knowledge. The estimated result for τ is then the value that maximizes a

summary measure of ROC curve. *De Somer et al.* (2011) used the Youden's index, defined as *sensitivity value + (specificity - 1)*, and *Ranucci et al.* (2005) used the Area Under the Curve (AUC) as a summary of the ROC curve. Besides the ROC curve, one may also use the likelihood or equivalently the information-based criteria (AIC/BIC) to assess model fit and estimate τ. In this approach, the estimated result for τ is the value that optimizes the likelihood/AIC/BIC.

Although this model (4.1) is easy to interpret and familiar to applied statisticians and medical researchers, its drawback is also obvious. Clinicians generally believe that nadir DO2 being too low is harmful in terms of the risk of AKI. However, it may not be realistic to assume the risk of AKI jumps suddenly once the DO2 is below the threshold and stays constant thereafter. Clinically, it is more plausible that once the DO2 drops below the threshold, the risk of AKI increases gradually and keeps increasing with decreasing DO2. Although simple statistically, the sudden jump model (4.1) is quite unnatural scientifically.

4.2.3 Broken-stick model

The broken-stick model (i.e., a generalized linear spline model with a single knot) can also be used to model the threshold effect and estimate τ. That is, the conditional mean μ_i is modeled as follows

$$\begin{aligned} g(\mu_i) &= \beta_0 + \beta_1 X_i + \beta_2 (X_i - \tau)_+ + \boldsymbol{\eta}^T \boldsymbol{Z}_i \\ &= \begin{cases} \beta_0 + \beta_1 X_i + \boldsymbol{\eta}^T \boldsymbol{Z}_i & \text{if } X_i \leq \tau \\ (\beta_0 - \beta_2 \tau) + (\beta_1 + \beta_2) X_i + \boldsymbol{\eta}^T \boldsymbol{Z}_i & \text{otherwise,} \end{cases} \end{aligned} \tag{4.2}$$

where $(X_i - \tau)_+ = (X_i - \tau)$ if $X_i > \tau$ and 0 otherwise. Comparing model (4.2) with the sudden-jump model (4.1), $g(\mu_i)$ is modeled as two linear segments with different slopes above and below the threshold as opposed to a step function for X_i in model

(4.2). In a logistic broken-stick model (4.2), the relationship between Y_i and X_i is modeled using two linear segments joined at the threshold τ at the log odds scale. The broken-stick model is scientifically more plausible than the sudden-jump model. However, this model does not incorporate the scientific knowledge that the adverse effect of low X_i only appears when it is below (or above) a certain threshold. Figure 4.2.1B depicts such an effect on the probability of AKI for a given subject.

Although it is straightforward to estimate the other unknown parameters in model (4.2) for a fixed τ, there are not many methods that are rigorously studied or well accepted for estimating τ. *Muggeo* (2003) is a well-cited paper which proposes a method for estimating the change-point τ based on an approximate linearization of first-order Talyor's expansion with respect to τ. We note that one may also adopt this method to estimate the threshold of a factor with a threshold effect. We conduct simulation studies to evaluate its performance and compare it with the usual method based on a jump method and methods based on the constrained broken-stick model studied below. The R package "segmented" (*Muggeo*, 2008) is used for implementing the method in our simulations and the application study presented below.

4.2.3.1 Constrained Broken-stick model

To remedy the drawback of the sudden-jump model and to incorporate the clinical knowledge simultaneously, we propose the constrained broken-stick model to study the threshold effect of nadir DO2. As model (4.1), the constrained broken-stick model assumes there is no effect of DO2 on the risk of AKI when DO2 is above the threshold to reflect the clinical knowledge. However, when nadir DO2 reaches the lower threshold, the risk of AKI increases with decreasing DO2. Specifically, the

constrained broken-stick model is as follows

$$
\begin{aligned}
g(\mu_i) &= \beta_0 + \beta_1 (X_i - \tau)_- + \boldsymbol{\eta}^T \boldsymbol{Z}_i \\
&= \begin{cases} (\beta_0 - \beta_1 \tau) + \beta_1 X_i + \boldsymbol{\eta}^T \boldsymbol{Z}_i & \text{if } X_i \leq \tau \\ \beta_0 + \boldsymbol{\eta}^T \boldsymbol{Z}_i & \text{otherwise,} \end{cases}
\end{aligned} \tag{4.3}
$$

where $(X_i - \tau)_- = X_i - \tau$ if $X_i \leq \tau$ and 0 otherwise. It is easy to see that $g(\mu_i)$ is modeled as a linear spline model with a single knot at the threshold, with the additional constraint that the slope of nadir DO2 is zero when X_i is greater than τ. Figure 4.2.1C depicts such an effect on the probability of AKI for a given subject. This model is scientifically more plausible and strikes a balance between model simplicity and scientific plausibility. Note that model (4.3) can also be written equivalently as $g(\mu_i) = \beta_0^* + \beta_1 \min(X_i, \tau) + \boldsymbol{\eta}^T \boldsymbol{Z}_i$ with $\beta_0^* = \beta_0 - \beta_1 \tau$. This parameterization is perhaps easier to interpret and more intuitive to applied biomedical researchers. The parameterization in model (4.3) and model (4.4) below, however, is more commonly used for linear spline models and broken-stick models.

The constrained broken-stick model in (4.3) is parameterized in a way such that there is an effect of X_i on Y_i when X_i is below a threshold. If instead one believes there is an effect of X_i on Y_i when X_i is greater than a threshold and there is no effect otherwise, i.e., one is interested in finding an upper threshold, then we would parameterize the constrained broken-stick model as follows,

$$
\begin{aligned}
g(\mu_i) &= \beta_0 + \beta_1 (X_i - \tau)_+ + \boldsymbol{\eta}^T \boldsymbol{Z}_i \\
&= \begin{cases} (\beta_0 - \beta_1 \tau) + \beta_1 X_i + \boldsymbol{\eta}^T \boldsymbol{Z}_i & \text{if } X_i \geq \tau \\ \beta_0 + \boldsymbol{\eta}^T \boldsymbol{Z}_i & \text{otherwise,} \end{cases}
\end{aligned} \tag{4.4}
$$

where $(X_i - \tau)_+ = X_i - \tau$ if $X_i \geq \tau$ and 0 otherwise. Again, this model is equivalent

to $g(\mu_i) = \beta_0^* + \beta_1 \max(X_i, \tau) + \boldsymbol{\eta}^T \boldsymbol{Z}_i$ with $\beta_0^* = \beta_0 - \beta_1\tau$. In our application, we believe there exists a threshold of lower DO2 and therefore, our analysis would be based on model (4.3). However, the method we propose below is applicable to both constrained broken-stick model (4.3) and model (4.4). Below we discuss three methods to estimate the threshold τ in constrained broken-stick models. The R package "segmented" (*Muggeo*, 2008) designed for broken-stick model can also be used for the constrained broken-stick model through some simple tricks, however we will not discuss this method in detail.

4.2.3.2 Likelihood Method

Similar to the sudden-jump model, the likelihood or information (AIC/BIC)-based method can be used to estimate τ in a constrained broken-stick model. That is, for each fitted constrained broken-stick regression model at a different value of possible τ in $(\min_i(X_i), \max_i(X_i))$, we evaluate its fit by the likelihood/AIC/BIC and estimate τ as the value that leads to the optimal likelihood/AIC/BIC, that is maximizing the likelihood or minimizing AIC/BIC. Note, all three measures (likelihood, AIC, BIC) are equivalent and will lead to the same optimal value of τ because the number of unknown parameters in our model at different value of τ is the same.

4.2.3.3 ROC Curve Method

When Y_i is binary as in our motivating study, ROC curve method can also be used in the constrained broken-stick model for estimating τ. Similar to methods used in *Ranucci et al.* (2005) and *De Somer et al.* (2011), one can estimate τ as the value that leads to the best ROC curve of the resulting logistic regression model, measured by AUC or the Youden's index. That is, fixing τ at a value in $(\min_i(X_i), \max_i(X_i))$, fit model (4.3) or model (4.4) as appropriate and evaluate its AUC/Youden's index. Then we can estimate τ as the one that leads to the optimal AUC/Youden's index.

4.2.3.4 Proposed Modified Two-step Newton-Raphson Method

To simply notation, we denote $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ and denote all parameters in a constrained broken-stick model (4.3) or model (4.4) as $\boldsymbol{\theta} = (\beta_0, \beta_1, \tau, \boldsymbol{\eta}^T)^T$, and $\boldsymbol{\theta}$ is assumed to belong to a compact set $\boldsymbol{\Theta}$. The true value of $\boldsymbol{\theta}$ is denoted as $\boldsymbol{\theta}^0$, assumed to be an interior point of $\boldsymbol{\Theta}$. We propose to estimate τ by solving the following estimating equation:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{H_i^T(\boldsymbol{\theta})\{Y-\mu_i(\boldsymbol{\theta})\}}{v\{\mu_i(\boldsymbol{\theta})\}g'\{\mu_i(\boldsymbol{\theta})\}}=0, \tag{4.5}$$

where $H_i(\boldsymbol{\theta}) = \{1, (X_i - \tau)_-, -\beta_1 I(X_i < \tau), \boldsymbol{Z}_i\}$ for model (4.3), and $H_i(\boldsymbol{\theta}) = \{1, (X_i - \tau)_+, -\beta_1 I(X_i > \tau), \boldsymbol{Z}_i\}$ for model (4.4). The solution of the proposed estimating equation (4.5) is denoted as $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\beta}}_n^T, \widehat{\tau}_n^T, \widehat{\boldsymbol{\eta}}_n^T)^T$. If we additionally assume that $Y_i|X_i, \boldsymbol{Z}_i$ arises from an exponential family with a canonical link, we have $v\{\mu_i(\boldsymbol{\theta})\} = [g'\{\mu_i(\boldsymbol{\theta})\}]^{-1}$. The validity of the proposed estimating equation (4.5) can be justified through the study of asymptotic properties of $\widehat{\boldsymbol{\theta}}_n$ and results are summarized below, with proofs available in Appendix Section A.4.

Result 1. *$\widehat{\boldsymbol{\theta}}_n$ is a consistent estimator for $\boldsymbol{\theta}^0$, as $n \to \infty$.*

Result 2. *$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0)$ converges in distribution to $\mathcal{N}\Big(0, \phi V^{-1}(\boldsymbol{\theta}^0)\Big)$, where $V(\boldsymbol{\theta}^0) = E\Big[\frac{H^T(\boldsymbol{\theta}^0)H(\boldsymbol{\theta}^0)}{v\{\mu(\boldsymbol{\theta}^0)\}g'\{\mu(\boldsymbol{\theta}^0)\}^2}\Big]$.*

Estimating a threshold/change-point is computationally challenging due to that the model is not differentiable. Note the proposed estimating equation (4.5) is also non-differentiable. Similar to the algorithm proposed in Section 2.4.1, we propose to use a two-step modified Newton-Raphson (NR) algorithm to solve the the proposed estimating equation (4.5). In this algorithm, we update the threshold τ and other unknown parameters separately to improve numerical stability. The motivation for this is that once τ is known then the model is just the usual generalized linear regression model which can be fitted using standard algorithms and software. Denoting the

initial value of τ as $\widehat{\tau}^{(0)}$, the t-th ($t \geq 1$) iteration of the proposed two-step modified NR algorithm proceeds as follows.

Step 1. Update estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$. Specifically, treating $\widehat{\tau}^{(t-1)}$ as fixed, fit the generalized constrained linear regression model by MLE or quasi-likelihood method to obtain estimates $\widehat{\boldsymbol{\beta}}^{(t-1)}$, $\widehat{\boldsymbol{\eta}}^{(t-1)}$ and predicted values $\widehat{\mu}_i{}^{(t-1)}$ $(i = 1, \ldots, n)$.

Step 2. Update the estimate of τ by an extended NR type procedure. Specifically, the threshold τ of the model (4.3) is updated by $\widehat{\tau}^{(t)} = \widehat{\tau}^{(t-1)} - U^{(t)}/S^{(t)}$, where

$$\begin{aligned} U^{(t)} &= \frac{1}{n}\sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i{}^{(t-1)})I(X_i < \widehat{\tau}^{(t-1)})}{v\{\widehat{\mu}_i{}^{(t-1)})\}g'\{\widehat{\mu}_i{}^{(t-1)}\}}, \\ S^{(t)} &= \frac{\widehat{\beta}_1^{(t-1)}}{n}\sum_{i=1}^{n} \frac{I(X_i < \widehat{\tau}^{(t-1)})}{v(\widehat{\mu}_i{}^{(t-1)})\{g'(\widehat{\mu}_i{}^{(t-1)})\}^2}. \end{aligned}$$

If model (4.4) is of interest, then one needs to replace $I(X_i < \widehat{\tau}^{(t-1)})$ in $U^{(t)}$ and $S^{(t)}$ by $I(X_i > \widehat{\tau}^{(t-1)})$ accordingly.

The algorithm starts with $t = 1$ and converges if $\|\widehat{\tau}^{(t)} - \widehat{\tau}^{(t-1)}\| < \zeta$, where ζ is a pre-specified convergence tolerance value. Once the algorithm converges for some t as determined by step 1, we implement step 2 to obtain the final estimate of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$. We show in the Appendix that, when the algorithm converges, the proposed two-step modified NR algorithm solves the proposed estimating equation (4.5).

4.3 Simulations

We conducted simulation studies to compare the various models and methods described above to estimate the threshold in terms of bias, efficiency and robustness. We considered four data-generating scenarios with the outcome being binary in scenarios I and II and being continuous in scenarios III an IV. The data generating models and true values of parameters are listed in Table 4.1. In scenarios I and III,

there is no association between the factor of interest X_i and the outcome Y_i when X_i is greater than the threshold, and there is a linear association at the logit scale (i.e., between X_i and log odds of Y_i) in scenario I and at the original scale in scenario III when X_i is below the threshold. In scenarios II and IV, when X_i is below the threshold, the relationship of X_i and Y_i follows a quadratic form at the logit scale in scenario II and at the original scale in scenario IV. Figure 4.2 plots the relationship between X and Y for all four scenarios. As shown in Figure 4.2, in scenarios I and II where the outcome is binary, the relationships are consistent with the clinical knowledge in our motivating example, i.e., there is no effect of DO2 when it is greater than a threshold, and the risk of AKI increases with decreasing DO2 once the DO2 is below a certain threshold. We note that scenarios I and III satisfy the specified constrained broken-stick model. However, for scenarios II and IV, the constrained broken-stick model, as well as the sudden-jump model and the usual broken-stick model, is misspecified since the true relationship is nonlinear. Scenarios II and IV were designed to evaluate the robustness of the proposed method. In all four scenarios, X_i follows a normal distribution with mean 250 and standard deviation 30, and the covariate Z_i follows a normal distribution with mean 50 and standard deviation 10. In scenarios III and IV, the residual term ϵ_i follows a normal distribution with mean 0 and standard deviation 50. For each scenario, we varied the sample size and considered $n = 500$, 1000, 2500, and 5000. Reported results are based on 1000 Monte Carlo replicates. The true threshold τ is chosen to be 270 in all scenarios, mimicking our motivating data.

In scenarios I and II, we estimated the threshold τ using the following six methods: the AUC method and the Youden's index method based on the sudden-jump model (4.1), the "segmented" method (*Muggeo*, 2008) based on the generalized broken-stick model (4.2), and the three methods based on the constrained broken-stick model (4.3) using the likelihood method, AUC method and the proposed modified Newton-Raphson

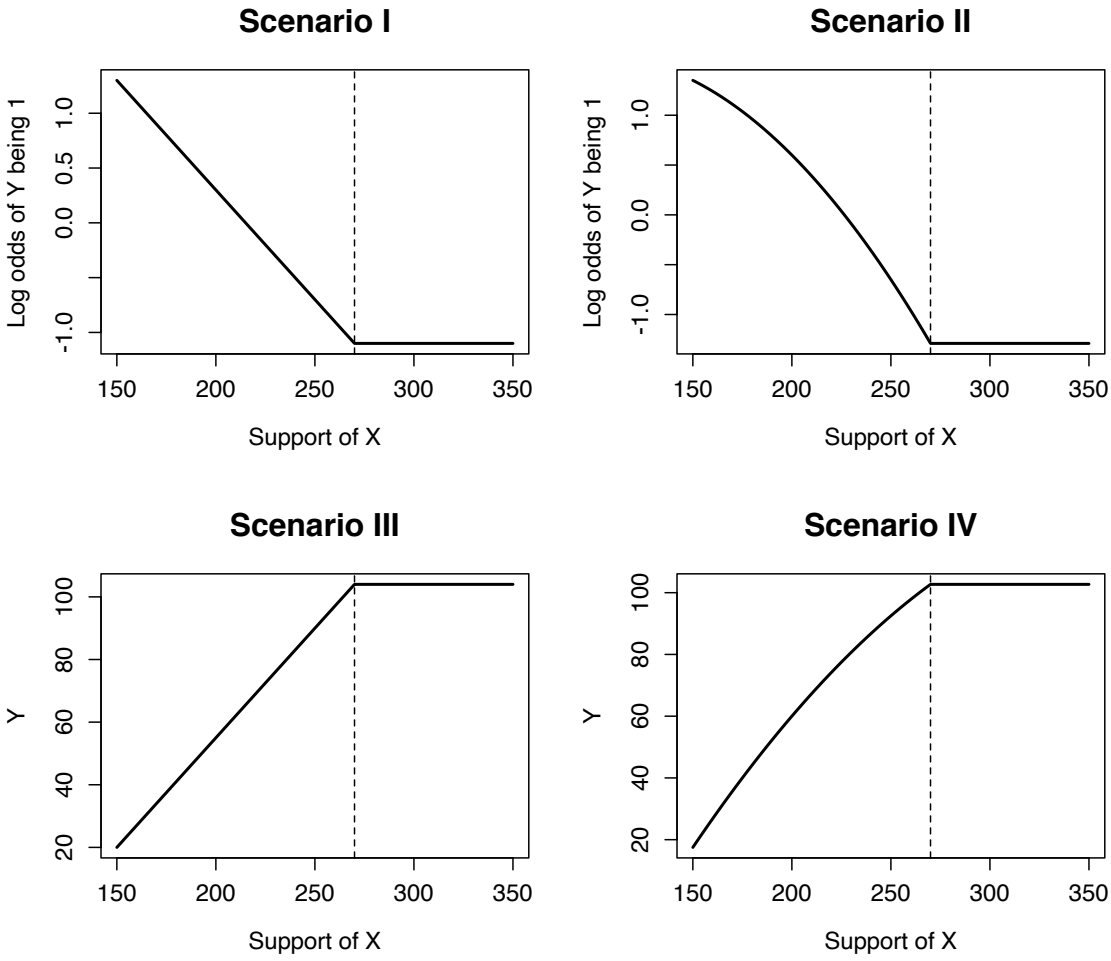

Figure 4.2: Data generating functions of simulations in scenarios I, II, III and IV with $Z = 0$.

Table 4.1: Data generating models and the corresponding true values of parameters in simulation studies.

Scenario I	$\text{logit}(\mu_i) = \beta_0 + \beta_1 \min(X_i, \tau) + \eta Z_i$ $(\beta_0, \beta_1, \eta, \tau) = (4.3, -0.02, 0.05, 270)$
Scenario II	$\text{logit}(\mu_i) = \beta_0 + \beta_1 \min(X_i, \tau) + \beta_2 \min(X_i, \tau)^2 + \eta Z_i$ $(\beta_0, \beta_1, \beta_2, \eta, \tau) = (0.6, 0.02, -0.0001, 0.05, 270)$
Scenario III	$Y_i = \beta_0 + \beta_1 \min(X_i, \tau) + \eta Z_i + \epsilon_i$ $(\beta_0, \beta_1, \eta, \tau) = (-85, 0.7, 8, 270)$
Scenario IV	$Y_i = \beta_0 + \beta_1 \min(X_i, \tau) + \beta_2 \min(X_i, \tau)^2 + \eta Z_i + \epsilon_i$ $(\beta_0, \beta_1, \beta_2, \eta, \tau) = (-170, 1.55, -0.002, 8, 270)$

method respectively. In scenarios III and IV, we estimated the threshold τ based on the three different models using the following four methods: the likelihood method based on the sudden-jump model (4.1), the "segmented" method (*Muggeo*, 2008) based on the generalized broken-stick model (4.2), and the two methods based the constrained broken-stick model (4.3) using the likelihood method and the proposed modified Newton-Raphson method respectively. For each scenario with a specific sample size, we report the bias, standard deviation of estimates across 1000 Monte Carlo replicates (MCSD), mean squared error (MSE) and average of standard error (AVESE) for each methods. Bootstrapping was used to obtain the AVESE using the R package "boot" *Davison and Hinkley* (1997); *Canty and Ripley* (2020) with 200 bootstrap replicates. The R package "pROC" *Robin et al.* (2011) was used to calculate the AUC and the Youden's index , the R function "logLik" was used to calculate the log-likelihood of each fitted regression model in all likelihood methods, and the R function "optimize" was used to optimize the threshold τ in all methods based on AUC, the Youden's index and the likelihood (*R Core Team*, 2020). Results on scenarios I and II are shown in Table 4.2 and on scenarios III and IV are shown in Table 4.3.

We note that the two methods based on a sudden-jump model lead to large biases across all scenarios and the bias persists even when the sample size is very large. In addition, for all scenarios considered here, the direction of bias is negative, indicating that it tends to underestimate the threshold. In the setting of cardiac surgery practices that we consider, a negative bias is worse than a positive bias because the surgeon and perfusionist who manage a patient's DO2 may think a nadir DO2 of 255, say, will not cause harm to a patient but in reality it is causing harm. The bias is a result of model-misspecification. These results demonstrate the consequence of the sudden-jump model when it is misspecified.

Table 4.2: Simulation results based on 1000 Monte Carlo data sets for scenarios I and II. "Bias" is Monte Carlo bias, "MCSD" is Monte Carlo standard deviation, "RMSE" is Monte Carlo root mean squared error, and "AVESE" is average of standard error. "SJ Model" is sudden-jump model, "BS Model" is broken-stick model, and "CBS model" is constrained broken-stick model. The minimum and maximum RMSE are highlighted in bold.

Methods	**Bias**	**MCSD**	**RMSE**	**AVESE**	**Bias**	**MCSD**	**RMSE**	**AVESE**
				Scenario I				
	$n = 500$				$n = 1000$			
SJ Model (AUC)	-26.18	18.83	32.25	19.00	-28.28	15.50	32.25	15.89
SJ Model (Youden)	-24.40	22.68	33.31	22.14	-25.28	20.22	32.37	20.10
BS Model (segmented)	-17.86	41.29	**44.99**	36.36	-10.92	31.07	**32.93**	32.51
CBS Model (NR)	-9.74	16.99	**19.58**	15.44	-5.52	13.38	**14.47**	13.31
CBS Model (AUC)	2.81	26.72	26.87	24.95	4.08	21.68	22.06	20.99
CBS Model (likelihood)	1.51	27.33	27.37	25.75	3.28	21.83	22.08	21.04
	$n = 2500$				$n = 5000$			
SJ Model (AUC)	-28.25	10.38	30.09	11.14	-28.19	8.34	**29.40**	8.61
SJ Model (Youden)	-27.16	13.59	**30.37**	15.20	-26.84	10.70	28.89	11.98
BS Model (segmented)	-5.47	23.62	24.25	25.88	-2.77	15.33	15.58	17.69
CBS Model (NR)	-2.33	10.15	**10.42**	10.16	-1.09	8.00	**8.07**	7.90
CBS Model (AUC)	3.47	15.37	15.75	15.59	1.95	11.28	11.45	11.97
CBS Model (likelihood)	3.11	14.94	15.26	14.79	0.68	8.57	8.60	10.46
				Scenario II				
	$n = 500$				$n = 1000$			
SJ Model (AUC)	-23.43	13.62	27.11	14.25	-23.45	10.77	25.80	11.30
SJ Model (Youden)	-21.18	16.22	26.68	17.26	-22.34	13.15	**25.92**	14.20
BS Model (segmented)	-9.40	36.42	**37.62**	34.51	-2.38	23.48	23.60	29.63
CBS Model (NR)	-3.75	13.76	**14.26**	13.13	0.35	11.03	**11.03**	10.86
CBS Model (AUC)	7.16	20.94	22.13	19.90	6.27	15.73	16.93	15.68
CBS Model (likelihood)	5.73	20.59	21.37	20.43	5.43	14.32	15.31	15.21
	$n = 2500$				$n = 5000$			
SJ Model (AUC)	-23.18	7.31	**24.31**	7.76	-22.85	5.68	**23.55**	5.96
SJ Model (Youden)	-22.52	8.85	24.20	9.76	-21.96	6.56	22.92	7.30
BS Model (segmented)	1.95	13.40	13.54	21.25	5.91	8.60	10.43	9.33
CBS Model (NR)	1.82	6.74	**6.98**	7.58	2.41	4.69	**5.27**	5.11
CBS Model (AUC)	4.21	10.74	11.53	11.21	3.13	6.09	6.85	7.61
CBS Model (likelihood)	3.97	9.22	10.04	9.90	3.13	5.34	6.19	6.30

Table 4.3: Simulation results based on 1000 Monte Carlo data sets for scenarios III and IV. True $\tau = 270$. "Bias" is Monte Carlo bias, "MCSD" is Monte Carlo standard deviation, "RMSE" is Monte Carlo root mean squared error, and "AVESE" is average of standard error. "SJ Model" is sudden-jump model, "BS Model" is broken-stick model, and "CBS model" is constrained broken-stick model. The minimum and maximum RMSE are highlighted in bold.

Methods	**Bias**	**MCSD**	**rMSE**	**AVESE**	**Bias**	**MCSD**	**rMSE**	**AVESE**
	Scenario III							
	$n = 500$				$n = 1000$			
SJ Model (likelihood)	-33.27	12.66	**35.60**	13.08	-32.34	9.24	**33.63**	9.77
BS Model (segmented)	-6.96	29.78	30.58	28.42	-1.12	17.76	17.79	21.69
CBS Model (NR)	-2.77	11.28	**11.61**	10.81	-1.07	8.39	**8.46**	8.59
CBS Model (likelihood)	2.76	15.60	15.84	15.25	1.73	10.49	10.63	11.38
	$n = 2500$				$n = 5000$			
SJ Model (likelihood)	-32.04	6.69	**32.73**	6.83	-31.68	5.38	**32.14**	5.57
BS Model (segmented)	-3.03	5.02	5.86	10.67	-0.23	4.15	4.16	6.05
CBS Model (NR)	-0.30	4.94	**4.95**	5.39	-0.07	3.27	3.27	3.55
CBS Model (likelihood)	0.53	5.41	5.43	6.33	-0.06	3.17	**3.17**	3.75
	Scenario IV							
	$n = 500$				$n = 1000$			
SJ Model (likelihood)	-38.23	14.39	**40.85**	14.95	-36.87	10.86	**38.44**	11.25
BS Model (segmented)	-6.75	23.68	24.62	25.58	-7.80	18.89	20.43	21.69
CBS Model (NR)	-8.56	12.31	**14.99**	11.15	-7.09	9.72	**12.04**	9.16
CBS Model (likelihood)	-2.40	18.04	18.20	17.97	-2.84	12.86	13.17	13.60
	$n = 2500$				$n = 5000$			
SJ Model (likelihood)	-36.37	7.49	**37.13**	7.67	-36.35	5.79	**36.81**	6.03
BS Model (segmented)	-13.73	14.42	19.91	13.46	-8.86	8.93	12.58	9.74
CBS Model (NR)	-5.70	6.39	8.56	6.46	-5.21	4.40	6.82	4.69
CBS Model (likelihood)	-4.30	6.89	**8.12**	7.93	-4.92	4.16	**6.44**	4.77

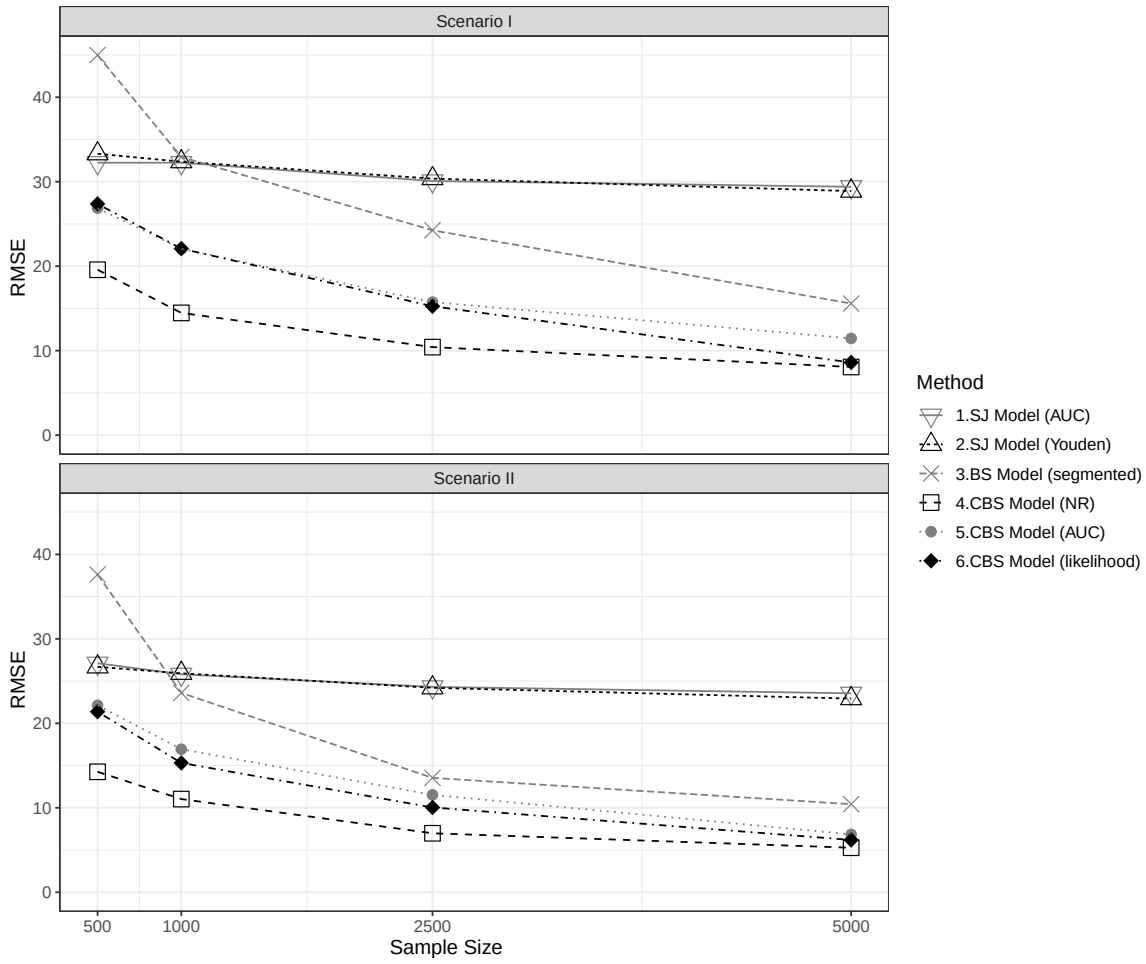

Figure 4.3: Monte Carlo root mean squared error (RMSE) for all methods in simulation scenario I and scenario II, based on 1000 Monte Carlo data sets.

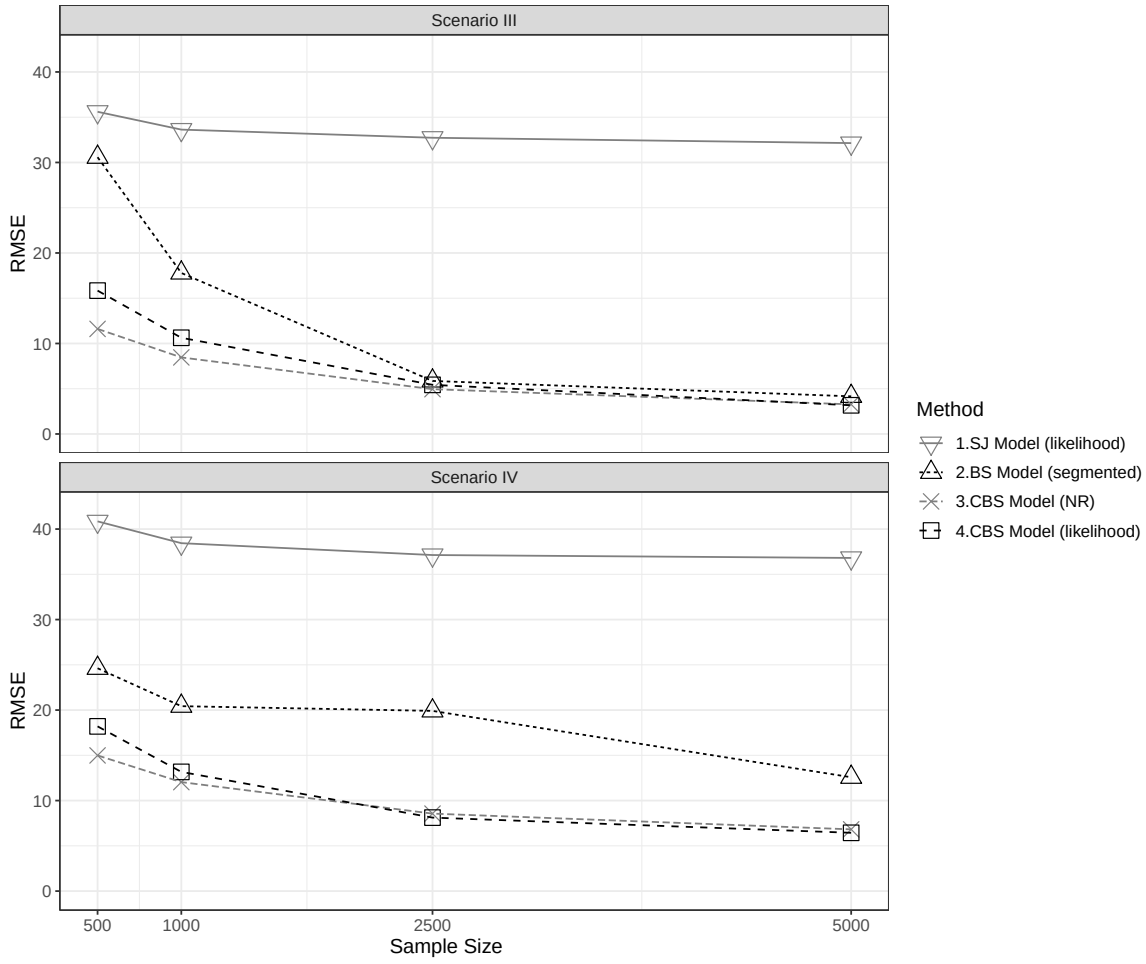

Figure 4.4: Monte Carlo root mean squared error (RMSE) for all methods in simulation scenario III and scenario IV, based on 1000 Monte Carlo data sets.

In scenario I and III, the segmented method is slightly biased when the sample size is small but the bias disappears when the sample size becomes larger (e.g, $n \geq 2500$), suggesting that the bias is a finite sample phenomenon. In general, the biases of the segmented method are smaller than the two methods based on a sudden-jump model. This is expected since the broken-stick model is either correct (scenarios I and III) or less misspecified (scenarios II and IV) than the broken-stick model. Therefore, there is no or less bias due to model misspecification. However, there is a large variability in its estimates in all scenarios, except for scenario III when the sample size goes above 2500. In scenarios I and II, considering the root mean squared error which accounts for both bias and variance, the segmented method performs even worse than the two methods based on a sudden-jump model due to the much larger variability when the sample size is small or finite ($n \leq 2500$ in scenario I and $n = 500$ in scenario II). In scenarios III and IV where the outcomes are continuous, although the root mean squared error is smaller, the MCSD of the segmented method is sometimes almost twice as large as the likelihood method based on a sudden-jump model (e.g., $n = 2500$ in scenario IV).

In all scenarios, all three methods based on a constrained broken-stick model have better performances in terms of both bias and variance than methods under a sudden-jump model or a broken-stick model. Compared with the sudden-jump model, the better performance of the contrained broken-stick model is largely due to the smaller bias as a result of less model-misspecification. As the broken-stick model, the constrained broken-stick model is either correct or mildly misspecified. However, compared with the broken-stick model, the constrained broken-stick model is able to reduce variability of estimation substantially by imposing the constraint that there is no effect when the factor of interest is below/above a certain threshold. These results show that the specified model plays a key role in the performance of estimation of a threshold.

In scenarios I and III, where the constrained broken-stick model is correct, the bias is small for all methods under a constrained broken-stick model. Scenarios II and IV are designed to evaluate robustness of various methods. In these two scenarios, all models including the broken-stick model and the constrained broken-stick model are misspecified. Scenarios II and IV represent situations where there is a threshold but the effect is not strictly linear once the factor of interest is below or above the threshold. This is a plausible relationship in practice. In reality and in our motivating example on DO2 in particular, often times scientific knowledge may suggest there is a monotonic effect below (or above) a threshold but the true relationship may not be necessarily linear. Consistent with other scenarios, in scenarios II and IV, methods based on a sudden-jump model lead to very large negative bias, the segmented method has a large variability when the sample size is small to moderate, and the methods based on a constrained broken-stick model lead to overall much better results in terms of both bias and variance. Although some biases remain due to model misspecification, by capturing the overall monotonic trend, the approximation by a constrained broken-stick model leads to considerable reduction in bias and the remaining bias (e.g., when $n = 500$ in scenario IV, the bias is smaller than 5) is not much clinically relevant in our application.

The above discussion focuses on differences among models. Next, we compare the performance of different methods under a constrained broken-stick model. Under the same constrained broken-stick model, different estimation methods lead to different performances. Overall, the proposed modified Newton-Raphson method has the best performance in terms of small bias and low variability, i.e. smallest root mean squared error, across almost all scenarios and all sample sizes. The likelihood/information-based method has the second best performance in scenarios I and II, and has the best performance when the sample size is large in scenarios III and IV. The AUC method has the worst performance among the three methods. The difference in performance

is more evident in scenarios I and II where the outcome is binary. These are more challenging scenarios for estimation as there is inherently less information contained in a binary outcome than in a continuous outcome. We note that for scenarios I and II, the AUC method is not stable and has large variability. This is expected since the AUC is not a sensitive criterion to evaluate model fit and, as a result, not a good method to base estimation of τ on. The likelihood method performs comparably as the proposed two-step NR method when the sample size is large (e.g. $n = 5000$), however it has larger variability when the sample size is small. The larger variability is mainly due to computation and depends on the the optimization method. In principle, maximizing the likelihood is a sound estimation idea. However, computationally the optimization may be challenging because the underlying likelihood function is not differentiable (*Das et al.*, 2016). In Figure 4.3 and Figure 4.4, we plot the root mean squared error for all methods for scenarios I-IV respectively. It is clear that the proposed modified Newton-Raphson method and the likelihood based method in the constrained broken-stick model lead to overall the most satisfactory results.

4.3.1 Application

This section considers the motivating study introduced in Section 4.1. Data were obtained on 2799 patients undergoing CABG surgery from 07/2007 to 06/2014 at the University of Michigan. This study was approved by the Institutional Review Board of the University of Michigan Health System (HUM00207425). We aim to estimate the low threshold of nadir DO2 that is associated with an increased risk of AKI to guide operative practice during a cardiac surgery. The factor of interest is nadir DO2 during a surgery. The severity of AKI is quantified by the stage ranging from 1 to 3 using the criterion by the Acute Kidney Injury Network. In our analysis, we consider two binary outcomes: any AKI with stage 1, 2 or 3, and moderate or severe AKI defined as AKI with stage 2 or 3. Our analysis adjusts for six covariates determined

by clinical knowledge. Table 4.4 shows the summary statistics for the factor of interest (nadir DO2), outcomes and baseline variables to be adjusted for. The total number of patients used in analyses is 2306, after removing patients with missing nadir DO2 or outcomes.

Table 4.4: Descriptive statistics for the application data set ($n = 2,306$).

Variables	**Mean(SD) or N(%)**
Factor of Interest	
Nadir DO2	443.94 (164.60)
Covariates	
Age	63.64 (13.53)
BMI, kg/m2	29.42 (6.33)
Gender (Male)	1487 (64.48%)
Diabetes	714 (30.96%)
Hypertension	1592 (69.04%)
Smoke	324 (14.05%)
Outcomes	
Any AKI (Stage 1, 2 or 3)	671 (29.10%)
Moderate or severe AKI (Stage 2 or 3)	154 (6.68%)

We carried out the analysis using the aforementioned methods. Estimates of the threshold based on the various methods are listed in Table 4.5 and the details of model fit using each of the methods are shown in Table 4.7. Unlike simulations studies, since we do not know the truth for a real data application it is difficult to judge which method gives the most accurate estimate. But we note the following things. First, the two methods (AUC and Youden's index) based on the sudden-jump model lead to very different answer on when the adverse effect of low nadir DO2 starts to appear. For example, for any AKI, the AUC method based on the sudden-jump model suggests the adverse effect appears when nadir DO2 is below 417; however, the Youden's index method suggests a threshold effect of nadir DO2 when it is below 343. For moderate or severe AKI, the estimated thresholds are 322 and 252 respectively using the AUC and Youden's Index methods in the sudden-jump

model. Second, consistent with our simulation studies, the sudden-jump model seems to underestimate the true threshold as estimates from the sudden-jump model are considerably smaller than estimates from other methods. The sudden-jump model is a popular model often used in practice to estimate thresholds. However, our simulations and data applications demonstrate the danger of using this model and the bias is due to model mispronunciation. Third, for both outcomes, the broken-stick model and the constrained broken-stick model lead to similar estimated thresholds. There are three methods based on the constrained broken-stick model. But we notice the estimate from the AUC method seems different from estimates from the other two methods. Fourth, the segmented method based on a broken-stick model has large variability as shown by the large standard error. The standard error is about twice as much as that from other methods. Methods from the constrained broken-stick lead to much more precise estimates. Finally, the recommended likelihood/information-based method and proposed modified Newton-Raphson method in a constrained broken-stick model lead to seemly reasonable estimates of threshold and results from the two methods are close. These observations are consistent with results observed in our simulation studies. The estimates from modified Newton-Raphson method in a constrained broken-stick model suggest that there is an increased risk of any AKI once nadir DO2 is below 608 and there is an increased risk of moderate or severe AKI if nadir DO2 drops below 494. These results suggest that probably surgeons and perfusionists who manage patient's DO2 during a cardiac surgery should be more conservative than what was recommended in the clinical literature in managing a patient's nadir DO2 to improve AKI related outcomes.

As explained previously, the constrained broken-stick model incorporates the scientific knowledge that the adverse effect of DO2 only appears when DO2 drops below a certain threshold and there is no effect when DO2 is high. To check whether this assumption holds true in our data, we carried out a post hoc analysis, where we fit a

Table 4.5: Threshold estimators of the nadir DO2 in three models (sudden-jump; broken-stick; constrained broken-stick) with two different binary outcomes (any AKI; moderate or severe AKI) for models with covariates.

Model	**Method**	**Any AKI**		**Moderate or Severe AKI**	
		Estimate	Std Error	Estimate	Std Error
SJ Model	AUC	416.53	88.23	322.15	97.84
	Youden	343.13	94.28	251.57	108.08
BS Model	segmented	619.05	244.31	522.86	196.11
CBS Model	AUC	616.66	92.13	527.06	100.75
	AIC	603.86	96.46	494.14	112.64
	NR	608.49	97.61	494.10	84.37

logistic regression model as in broke-stick model (4.2), while fixing τ at the estimated value from different methods. This is, we let the data to estimate the slopes of nadir DO2 below and above each estimated threshold. We check whether the slope of nadir DO2 when nadir DO2 is greater than the threshold is zero or not and whether the slope of nadir DO2 when nadir DO2 is below the threshold is indeed negative. Results based on the threshold value estimated using the segmented method and the three constrained broken-stick model are shown in Table 4.6. Indeed, for all estimates of the threshold from these methods, the slope is close to zero when nadir DO2 is above the threshold and the p-value is not significant. When nadir DO2 is below the threshold, the slope is negative, indicating an increased risk of AKI is associated with decreasing DO2. These results confirm that our clinical knowledge seems reasonable and the adverse effect of nadir DO2 indeed only exists once it drops below a threshold.

4.4 Discussion

In this article we focus on the estimation of a threshold, which has important applications in biomedical research but is relatively less studied. Moreover, popular methods in practice may lead to biased estimation of a threshold or have large variability in estimation. Estimation of a threshold differs from the usual problem of coefficient

Table 4.6: Model fitting results of the broken-stick model with the threshold fitted as the estimated value in Table 4.5. "BS Model" is broken-stick model and "CBS model" is constrained broken-stick model. $*$ represents $\times 10^{-3}$.

Method	Any AKI				Moderate or Severe AKI			
	Slope of nadir DO2	Coefficient*	SE*	P value	Slope of nadir DO2	Coefficient*	SE*	P value
BS Model (segmented)	≤ 619.05	-1.39	0.39	<0.01	≤ 522.86	-2.52	0.79	<0.01
	> 619.05	0.89	1.02	0.39	> 522.86	1.00	1.21	0.41
CBS Model (AUC)	≤ 616.66	-1.39	0.39	<0.01	≤ 527.06	-2.50	0.79	<0.01
	> 616.66	0.86	1.01	0.39	> 527.06	1.04	1.22	0.39
CBS Model (AIC)	≤ 603.86	-1.41	0.40	<0.01	≤ 494.14	-2.65	0.84	<0.01
	> 603.86	0.74	0.96	0.44	> 494.14	0.72	1.11	0.52
CBS Model (NR)	≤ 608.49	-1.40	0.40	<0.01	≤ 494.10	-2.65	0.84	<0.01
	> 608.49	0.78	0.98	0.42	> 494.10	0.72	1.11	0.52

estimation in that the parameter of interest is non-differentiable, which poses challenges in estimation and computation. We discuss three models that can be adopted to estimate a threshold. The sudden-jump model is a popular model in practice, which models the effect of the factor of interest using a step function and assumes the effect has a noncontinuous jump at the threshold. This model is scientifically unrealistic and may lead to severe bias of estimation. Instead of a step function, the broken-stick model models the relationship using piece-wise linear segments joined at a threshold. In this model, the effect changes continuously and does not have sudden jumps. Therefore, it is scientifically more plausible. However, this model does not incorporate the scientific knowledge that, when a threshold effect exists, there is no effect when the factor is below (or above) the threshold, leading to large estimation variability especially for binary outcomes. We propose to use a constrained broken-stick model to estimate the threshold, where a constraint is imposed to model a threshold effect. This model is able to strike a balance between model simplicity and scientific plausibility.

Within each model, various methods are studied to estimate the threshold. In particular, based on a constrained broken-stick model, we have proposed a modified two-step Newton-Raphson algorithm to estimate the threshold. This algorithm separately updates the threshold and the rest of the unknown parameters. Hence, it

Table 4.7: Model fitting results in three models (sudden-jump; broken-stick; constrained broken-stick) for two different binary outcomes (any AKI; moderate or severe AKI) with the threshold fitted as the estimated value in Table 4.5.

Any AKI				Moderate or severe AKI			
Variable	**Coefficient**	**SE**	**P value**	**Variable**	**Coefficient**	**SE**	**P value**
Sudden-Jump Model (AUC)				**Sudden-Jump Model (AUC)**			
intercept	-2.18	0.36	<0.01	intercept	-3.57	0.61	<0.01
Nadir DO2 < 416.53	0.36	0.10	<0.01	Nadir DO2 <322.15	0.55	0.19	<0.01
Age	-2.47E-03	3.83E-03	0.52	Age	-4.60E-03	0.01	0.50
Gender (Male)	0.02	0.10	0.85	Gender (Male)	-0.20	0.18	0.25
BMI	0.03	0.01	<0.01	BMI	0.04	0.01	<0.01
Diabetes	0.52	0.10	<0.01	Diabetes	0.58	0.18	<0.01
Hypertension	0.30	0.12	0.01	Hypertension	-0.11	0.21	0.60
Smoke	-0.08	0.14	0.57	Smoke	-0.19	0.26	0.46
Sudden-Jump Model (Youden's Index)				**Sudden-Jump Model (Youden's Index)**			
intercept	-2.05	0.35	<0.01	intercept	-3.47	0.61	<0.01
Nadir DO2 < 343.13	0.31	0.11	<0.01	Nadir DO2 < 251.57	0.61	0.23	<0.01
Age	-1.67E-03	3.81E-03	0.66	Age	-3.64E-03	0.01	0.60
Gender (Male)	-0.03	0.01	0.73	Gender (Male)	-0.25	0.17	0.15
BMI	0.03	0.10	<0.01	BMI	0.03	0.01	0.01
Diabetes	0.53	0.14	<0.01	Diabetes	0.56	0.18	<0.01
Hypertension	0.30	0.10	0.01	Hypertension	-0.09	0.21	0.66
Smoke	-0.08	0.12	0.58	Smoke	-0.19	0.26	0.48
Broken-Stick Model (segmented)				**Broken-Stick Model (segmented)**			
intercept	-1.43	0.37	<0.01	intercept	-2.51	0.64	<0.01
Nadir DO2	-1.39E-03	3.93E-04	<0.01	Nadir DO2	-2.52E-03	7.92E-04	<0.01
$(\text{Nadir DO2 - 619.05})_+$	2.28E-03	1.21E-03	0.06	$(\text{Nadir DO2 - 522.86})_+$	3.52E-03	1.67E-03	0.03
Age	-2.44E-03	3.84E-03	0.53	Age	-4.60E-03	0.01	0.51
Gender (Male)	0.02	0.10	0.87	Gender (Male)	-0.17	0.18	0.35
BMI	0.03	0.01	<0.01	BMI	0.04	0.01	<0.01
Diabetes	0.52	0.10	<0.01	Diabetes	0.56	0.18	<0.01
Hypertension	0.30	0.12	0.01	Hypertension	-0.10	0.21	0.63
Smoke	-0.08	0.14	0.58	Smoke	-0.19	0.26	0.48
Constrained Broken-Stick Model (AUC)				**Constrained Broken-Stick Model (AUC)**			
intercept	-1.48	0.36	<0.01	intercept	-2.61	0.63	<0.01
min(Nadir DO2, 616.66)	-1.29E-03	3.74E-04	<0.01	min(Nadir DO2, 527.06)	-2.27E-03	7.34E-04	<0.01
Age	-2.66E-03	3.83E-03	0.49	Age	-0.01	0.01	0.44
Gender (Male)	0.02	0.10	0.82	Gender (Male)	-0.14	0.18	0.43
BMI	0.03	0.01	<0.01	BMI	0.04	0.01	<0.01
Diabetes	0.52	0.10	<0.01	Diabetes	0.56	0.18	<0.01
Hypertension	0.30	0.12	0.01	Hypertension	-0.10	0.21	0.61
Smoke	-0.08	0.14	0.57	Smoke	-0.19	0.26	0.46
Constrained Broken-Stick Model (AIC)				**Constrained Broken-Stick Model (AIC)**			
intercept	-1.47	0.36	<0.01	intercept	-2.55	0.64	<0.01
min(Nadir DO2, 603.86)	-1.31E-03	3.80E-04	<0.01	min(Nadir DO2, 494.14)	-2.45E-03	7.83E-04	<0.01
Age	-2.62E-03	3.83E-03	0.49	Age	-0.01	0.01	0.46
Gender (Male)	0.02	0.10	0.82	Gender (Male)	-0.15	0.18	0.40
BMI	0.03	0.01	<0.01	BMI	0.04	0.01	<0.01
Diabetes	0.52	0.10	<0.01	Diabetes	0.56	0.18	<0.01
Hypertension	0.30	0.12	0.01	Hypertension	-0.10	0.21	0.62
Smoke	-0.08	0.14	0.57	Smoke	-0.19	0.26	0.46
Constrained Broken-Stick Model (NR)				**Constrained Broken-Stick Model (NR)**			
intercept	-1.47	0.36	<0.01	intercept	-2.55	0.64	<0.01
min(Nadir DO2, 608.49)	-1.30E-03	3.78E-04	<0.01	min(Nadir DO2, 494.10)	-2.45E-03	7.83E-04	<0.01
Age	-2.63E-03	3.83E-03	0.49	Age	-0.01	0.01	0.46
Gender (Male)	0.02	0.10	0.82	Gender (Male)	-0.15	0.18	0.40
BMI	0.03	0.01	<0.01	BMI	0.04	0.01	<0.01
Diabetes	0.52	0.10	<0.01	Diabetes	0.56	0.18	<0.01
Hypertension	0.30	0.12	0.01	Hypertension	-0.10	0.21	0.62
Smoke	-0.08	0.14	0.57	Smoke	-0.19	0.26	0.46

is able to take advantage of existing popular software available to fit a generalized linear regression model when the threshold is known to stabilize computation. We have discussed the implications of modeling assumptions on threshold estimation and compared various methods/models through extensive simulation studies. Our results show that the popular sudden-jump model leads to large systematic bias. In particular, when an effect is only present when the factor of interest is below a certain threshold, our simulation shows that a sudden-jump model systematically underestimates the threshold. The segmented method under a broken-stick model is less biased in general as the model is not or less misspecified, but has large variability in estimation. In certain cases, for example, when the outcome is binary and the sample size is small, the segmented method in a broken-stick model may even have larger root mean squared errors than the broken-stick model due to its large variability in estimation. Overall, all estimation methods in a constrained broken-stick model have better performance. Among all methods based on a constrained broken-stick model, the proposed modified Newton-Raphson method has overall the best performance in terms of mean squared error, especially for the more challenging case when the outcome is binary, whereas the AUC method has the worst performance.

The lowest nadir DO2 that a CABG patient can expose to during CPB but does not increase the risk of AKI is estimated using the various methods using data obtained from the University of Michigan. Results are consistent with our simulations studies. Notably, the popular sudden jump-model leads to estimates that are much lower than those from other methods. For moderate or severe AKI, our estimates from the sudden-jump model are close to previous published results, i.e., 272 mL/minute/m^2 and 262 mL/minute/m^2, using the same model in the clinical literature (*Ranucci et al.*, 2005; *De Somer et al.*, 2011). Estimates from the sudden-jump model are much lower than those from other methods (about 500 mL/minute/m^2 for moderate or severe AKI), consistent with observations from our simulation studies that the

sudden-jump model in this setting systematically underestimates the threshold. This potential bias has important implications and is worth more attention. Such a negative bias may mislead the surgeons and perfusionists to expose a patient to a harmfully low DO2 since they would have thought an adverse event would only appear if the DO2 is much lower. In fact, for any AKI, based on our estimate from the proposed modified Newton-Raphson algorithm, a nadir DO2 lower than 608mL/minute/m^2 starts to expose patients to a higher risk of AKI. These results suggest that surgeons and perfusionists need to adopt a more conservative strategy in managing a patient's DO2 during a CABG than those recommended in the clinical literature. These results, combined with findings from our simulation studies, highlight that modeling assumptions and the estimation method have a significant impact in estimation of a threshold.

CHAPTER V

Estimation of Threshold in Constrained Continuous Threshold Model

5.1 Introduction

In real applications, nonlinear patterns with single or multiple change-points are consistently observed and should be considered in the analysis. A particular typical case for the change-point is the threshold, where the association between a risk factor and the outcome only occurs before or after the threshold. Moreover, the estimation of the threshold is widely used as a key to answering important research questions in many scientific areas. For example, in the epidemiological study, an interesting and meaningful research question is to explore threshold limits of key air pollutants' exposures for pregnant women that pose adverse health effects in terms of low birth weight, which will lead to a substantial public health impact on perinatal mortality and long-term adverse health consequences for surviving infants (*Ngoc et al.*, 2006; *Behrman et al.*, 2007). Most studies have focused on the threshold effect of air pollution on the general public. In contrast, fetuses appear to be more susceptible than others (*Šrám et al.*, 2005), and existing studies reported that increased levels of air pollutants, such as nitrogen dioxide (NO_2) and sulfur dioxide (SO_2), would contribute to a higher risk of low birth weight (*Lacasana et al.*, 2005; *Sun et al.*, 2016;

Smith et al., 2020).

Approaches to estimating threshold are presented in the literature. In the following, we discuss several representative existing approaches in detail. First, some literature selected the threshold based on a quantile perspective, such as 10% or 25%, directly based on the distribution of their available data (*Lin et al.*, 2004). Although this approach is easy to implement and adopt by some researchers, it relies on the researcher's subjectivity due to the lack of justifiable reasons for choosing the quantile and hence may not provide a reliable threshold. To overcome subjectivity and keep the easy implementation in estimating the threshold, some researchers used the sudden-jump model, which assumes the risk of the outcome jumps suddenly once the factor of interest is lower or higher than a threshold and stays constant after the threshold (*Ranucci et al.*, 2005; *De Somer et al.*, 2011). However, the assumption of the sudden jump may not hold in reality and thus can not provide a reliable threshold. For example, we do not expect that the effect of low birth weight will suddenly jump to another constant once air pollutants achieve the threshold. Since the threshold is a particular case of a change-point, the linear spline model, also called the broken-stick model or the segmented model, has been widely studied in the literature (*Marsh and Cormier*, 2001). Estimating change-points in a linear spline model for a continuous outcome has been rigorously studied (*Quandt*, 1958; *Hudson*, 1966; *Feder*, 1975b,a; *Hansen*, 2000; *Das et al.*, 2016), and some well-developed algorithms are also available (*Muggeo*, 2003, 2008; *Fong et al.*, 2017). Although the linear spline model considers a continuous nonlinear effect for the factor of interest and the outcome, the linear spline model can be improved by adding the constraint for the threshold effect in order to answer some research questions with the threshold, such as the air pollutant problem we mentioned early. To reflect the knowledge of the existence of the threshold, the constrained linear spline model, also called the constrained broken-stick model, is an option to explore the threshold by imposing the assumption of no effect below or

above the threshold and a linear effect above or below the threshold. The estimation method and performances of the constrained linear spline model have been studied extensively in Chapter IV. Although the increasing or decreasing association between the factor of interest and the outcome is still captured by the constrained linear spline model, a linear association assumption is likely to be violated in reality. Higher-order (e.g., quadratic or cubic) constrained spline models are available options for studying threshold effects for researchers who do not believe in a linear association. However, the model assumption for a specific higher-order may still not hold in reality and may sometimes fall into the trouble of overfitting.

In order to balance the model flexibility and overfitting issue, we introduce a constrained penalized spline model and propose a computationally convenient two-step algorithm to estimate the threshold. The constrained penalized spline model and the proposed estimation equation and algorithm are summarized in Section 5.2. Simulation studies are reported in Section 5.3. In Section 5.4, we aim to estimate safe-level thresholds of NO_2 and SO_2 for pregnant women in preventing the outcome of low birth weight, using the study of Pregnancy Research on Inflammation, Nutrition, & City Environment: Systematic Analyses in Mexico City (*Osornio-Vargas et al.*, 2013).

5.2 Method

5.2.1 Constrained Penalized Spline Model and Notations

Consider a study with n subjects in the form of $\{Y_i, X_i, \boldsymbol{Z}_i\}$. For each subject i $(i = 1, ..., n)$, let Y_i be the outcome, X_i be the factor of interest with a a threshold effect, and $\boldsymbol{Z}_i$ be all other covariates to be adjusted for. Specifically, X_i is assumed to have an effect on the outcome Y_i only when X_i is above a certain threshold τ. To learn the threshold from the data, we consider a constrained penalized spline model

as follows,

$$
\begin{aligned}
Y_i &= \mu(X_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) + \epsilon_i \qquad (5.1)\\
&= \beta_0 + \beta_1 \max(X_i, \tau) + \sum_{k=1}^{K} \beta_{1k} \max(X_i, u_k) + \boldsymbol{\eta}^T \boldsymbol{Z}_i + \epsilon_i \text{ with } \sum_{k=1}^{K} \beta_{1k}^2 \leq c,\\
\text{where } \epsilon_i &\sim N(0, \sigma^2), \ \max(X_i, \tau) = \begin{cases} X_i & \text{if } X_i \geq \tau_k \\ \tau & \text{if } X_i < \tau_k, \end{cases} \text{ and } u_k = \tau + \frac{\max(X) - \tau}{K+1} k.
\end{aligned}
$$

In the model (5.1), K is the prespecified number of knots u_k, where u_k is the knot evenly distributed between the threshold τ and the maximum value of X_i, that is $\max(X)$, for $k = 1, \ldots, K$. From the definition above, knots u_k are determined automatically once the data and the threshold τ are given. Effects of $\boldsymbol{Z}_i$ are modeled as linear with coefficients $\boldsymbol{\eta}$. We assume that the factor of interest X_i has a bounded domain and the threshold τ is within this bounded domain. To simplify notations, we denote $\boldsymbol{V}_i = (X_i, \boldsymbol{Z}_i)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_{11}, ..., \beta_{1K})^T$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T, \tau)^T$, where $\boldsymbol{\theta}$ is assumed to belong to a compact set $\boldsymbol{\Theta}$ with dimension $p = K + L + 3$. The true value of $\boldsymbol{\theta}$ is denoted as $\boldsymbol{\theta}^0$, assumed to be an interior point of the compact set $\boldsymbol{\Theta}$.

5.2.2 Proposed Method

Motivated by the ridge regression, we want to minimize the following loss function $L(\boldsymbol{\theta})$ using a Lagrange multiplier argument. Specifically,

$$
\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \left[\frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i - \mu(\boldsymbol{V}_i; \boldsymbol{\theta}) \right\}^2 + \lambda \sum_{k=1}^{K} \beta_{1k}^2 \right], \qquad (5.2)
$$

where λ is the Lagrange multiplier for some $\lambda \geq 0$. The term $\lambda \sum_{k=1}^{K} \beta_{1k}^2$ is called a roughness penalty and the amount of smoothing is controlled by λ, which is also referred as the smoothing parameter. When λ is chosen as zero, there is no penalization on the β_{1k} $(k = 1, \ldots, K)$, which will lead to an overfitting problem. When λ

is large, the parameter β_{1k} $(k = 1, \ldots, K)$ is heavily constrained and the number of degrees of freedom for model (1) will be effectively lower, tending to be $L + 3$.

As the maximum function, that is $\max(x, \tau)$, is not differentiable with respect to τ, we cannot directly solve the unknown parameters $\boldsymbol{\theta}$ from the minimization formula (5.2) by the derivative of $\boldsymbol{\theta}$. To overcome the nondifferentiability problem, we borrow the modified derivative idea as shown in formula (2.2) for the linear spline model. Specifically, we modify the derivative of $\max(x, \tau)$ with the respect of τ as follows,

$$\frac{\partial \max(x, \tau)}{\partial \tau} = I(x < \tau). \tag{5.3}$$

With this modified derivative, the loss function $L(\boldsymbol{\theta})$ can then be treated as a differentiable function with respect to $\boldsymbol{\theta}$ in some sense. We then introduced the function $Q(\boldsymbol{\theta}) = H(\boldsymbol{\theta})\{Y - \mu(\boldsymbol{V}; \boldsymbol{\theta})\} - D_\lambda(\boldsymbol{\theta})$, where

$$\begin{aligned} H(\boldsymbol{\theta}) &= \Big\{1, \max(X, \tau), \max(X, u_1), \ldots, \max(X, u_K), \\ &\qquad \boldsymbol{Z}, \beta_1 I(X < \tau) + \sum_{k=1}^{K} \{\beta_{1k}(1 - \frac{k}{K+1}) I(X < u_k)\}\Big\}^T_{p \times 1}, \\ D_\lambda(\boldsymbol{\theta}) &= \Big\{0, 0, \lambda\beta_{11}, \ldots, \lambda\beta_{1K}, 0, \ldots, 0\Big\}^T_{p \times 1}. \end{aligned}$$

And the $\frac{1}{n}\sum_{i=1}^{n} Q_i(\boldsymbol{\theta})$ function is proportional to the modified derivative of $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ as in formula (5.3). We then propose to proceed as usual, i.e., estimating the unknown parameters $\boldsymbol{\theta}$ by solving the following estimating equation,

$$\frac{1}{n}\sum_{i=1}^{n} Q_i(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} H_i(\boldsymbol{\theta})\{Y_i - \mu(\boldsymbol{V}_i; \boldsymbol{\theta})\} - D_\lambda(\boldsymbol{\theta}) = 0. \tag{5.4}$$

The solution of the above estimating equation (5.4) is denoted as $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\beta}}_n^T, \widehat{\tau}_n^T, \widehat{\boldsymbol{\eta}}_n^T)^T$. Intuitively, $H(\boldsymbol{\theta})$ can be viewed as the modified derivative of $\mu(X, \boldsymbol{Z}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. We note that the minimization of the loss function $L(\boldsymbol{\theta})$ is not the equivalent to

solving the proposed estimating equation (5.4), because the modified derivative idea is rigorous.

To choose the smoothing parameter λ, one standard technique is the generalized cross-validation, or short for GCV. According to Chapter 5.3.2 in *Ruppert et al.* (2003), the smoothing parameter λ can be chosen based on the following GCV criterion $GCV(\lambda)$, where

$$GCV(\lambda) = \sum_{i=1}^{n} \Big(\frac{\{(\boldsymbol{I} - \boldsymbol{S}_\lambda)\boldsymbol{Y}\}_i}{1 - n^{-1} tr(\boldsymbol{S}_\lambda)} \Big)^2$$

and $\boldsymbol{S}_\lambda$ is the smoother matrix associated with the predicted model (1). And we denote the smoothing parameter that minimizes GCV(λ) as $\widehat{\lambda}_{GCV}$. Another method is to use the linear mixed effect model, a special case of the ridge regression, to penalize the constraint term $\lambda \sum_{k=1}^{K} \beta_{1k}^2$. In the linear mixed effect model, $\beta_0 + \beta_1 \max(X_i, \tau) + \boldsymbol{\eta}^T \boldsymbol{Z}_i$ is treated as the fixed part, and $\sum_{k=1}^{K} \beta_{1k} \max(X_i, u_k)$ is treated as the random part. The standard method to fit a linear mixed effect model is the best linear unbiased prediction (BLUP) method.

5.2.3 Proposed Algorithm

This section describes the proposed two-step algorithm in detail to solve the threshold in model (5.1). As the proposed estimating function $Q(\boldsymbol{\theta})$ is also nondifferentiable with respect to $\boldsymbol{\theta}$, we propose the following two-step algorithm. Before starting the algorithm, we specify the initial value of τ, denoted as $\widehat{\tau}^{(0)}$. The t-th ($t \geq 1$) iteration of the algorithm proceeds as follows.

Step 1. Update estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ via treating τ is fixed at $\widehat{\tau}^{(t-1)}$. When τ is fixed, knots u_k ($k = 1, \ldots, K$) are also fixed, which are denoted as $\widehat{u}_k^{(t-1)}$. Specifically, we fit the following ridge regression with the same constraint $\sum_{k=1}^{K} \beta_{1k}^2 \leq c$ as

in model (5.1),

$$E(Y|X,Z) = \beta_0 + \boldsymbol{\eta}^T \boldsymbol{Z} + \beta_1 \max(X, \widehat{\tau}^{(t-1)}) + \sum_{k=1}^{K-1} \beta_{1k} \max(X, \widehat{u}_k^{(t-1)}),$$

and then obtain estimates $\widehat{\boldsymbol{\beta}}^{(t)}$, $\widehat{\boldsymbol{\eta}}^{(t)}$ and the predicted values $\widehat{\mu}_i^{(t)}$, $i = 1, \cdots, n$.

Step 2. Update the estimate of τ via treating $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ fixed at $\widehat{\boldsymbol{\beta}}^{(t)}$ and $\widehat{\boldsymbol{\eta}}^{(t)}$. Specifically, τ is updated by a modified Newton-Raphson procedure $\widehat{\tau}^{(t)} = \widehat{\tau}^{(t-1)} + \{J^{(t)}\}^{-1} U^{(t)}$, where

$$\begin{aligned} U^{(t)} &= \sum_{i=1}^{n} (Y_i - \widehat{\mu}_i^{(t)}) \Big\{ \widehat{\beta}_1^{(t)} I(X_i < \widehat{\tau}^{(t-1)}) + \sum_{k=1}^{K} \widehat{\beta}_{1k}^{(t)} I(X_i < \widehat{u_k}^{(t-1)}) (1 - \frac{k}{K+1}) \Big\}, \\ J^{(t)} &= \sum_{i=1}^{n} \Big\{ \widehat{\beta}_1^{(t)} I(X_i < \widehat{\tau}^{(t-1)}) + \sum_{k=1}^{K} \widehat{\beta}_{1k}^{(t)} I(X_i < \widehat{u_k}^{(t-1)}) (1 - \frac{k}{K+1}) \Big\}^2. \end{aligned}$$

From the modified derivative perspective, $U^{(t)}$ and $J^{(t)}$ are proportional to the modified first-order derivative function and second-order derivative (Hessian) matrix of $L(\boldsymbol{\theta})$ with respect to τ, with $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ fixed at the recent value.

Starting with $t = 1$, the proposed algorithm iterates between step 1 and step 2 until the convergence of τ, i.e. $\|\widehat{\tau}^{(t)} - \widehat{\tau}^{(t-1)}\| < \xi$, where ξ is a pre-specified convergence tolerance value. And the final estimator of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are obtained by another step 1 of the algorithm, fixing the estimate of τ as the final estimate $\widehat{\tau}^{(t)}$.

5.2.4 Estimation of Variance

In this section, we discuss the asymptotic properties and variance estimation methods for the proposed estimator $\widehat{\boldsymbol{\theta}}_n$. This section roughly follows the idea and proofs studied in Section 2.4.2 and *Yu and Ruppert* (2002).

Assumption 2. $\frac{1}{n}\sum_{i=1}^{n}\{\mu(\boldsymbol{V}_i;\boldsymbol{\theta})-\mu(\boldsymbol{V}_i;\boldsymbol{\theta}^*)\}$ converges to some limit function uniformly in $\boldsymbol{\theta}$, $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$, and $\frac{1}{n}\sum_{i=1}^{n}\{\mu(\boldsymbol{V}_i;\boldsymbol{\theta})-\mu(\boldsymbol{V}_i;\boldsymbol{\theta}_0)\}$ has a unique minimum $\boldsymbol{\theta}_0$.

Result 3. *Under assumption 2 and if the smoothing parameter $\lambda = o(1)$, then $\widehat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}_0$.*

Proof. It is obvious to see that the mean function $\mu(\boldsymbol{V};\boldsymbol{\theta})$ is continuous for each fixed $\boldsymbol{V}$ and $\boldsymbol{\theta}$ is assumed to be belonged to the compact set $\boldsymbol{\Theta}$. Therefore, according to Theorem 1 in *Yu and Ruppert* (2002), $\widehat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}_0$. □

Result 4. *Under assumption 2 and if the smoothing parameter $\lambda = o(n^{-1/2})$, then $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)$ converges in distribution to a Normal distribution $N(0,\sigma^2\Omega^{-1}(\boldsymbol{\theta}_0))$, where $\Omega(\boldsymbol{\theta}_0) = E[H(\boldsymbol{\theta}_0)H^T(\boldsymbol{\theta}_0)]$.*

Proof. Although $Q(\boldsymbol{\theta})$ involves maximum functions that are not differentiable, $E(Q(\boldsymbol{\theta}))$ is first-order differentiable with respect to $\boldsymbol{\theta}$ because of the integral of X. By Talyor expansion of $E(Q(\widehat{\boldsymbol{\theta}}_n))$ around $\boldsymbol{\theta}^0$, we have

$$E(Q(\widehat{\boldsymbol{\theta}}_n)) - E(Q(\boldsymbol{\theta}_0)) = R(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0), \tag{5.5}$$

where $R(\boldsymbol{\theta})$ is the first-order derivative matrix of $E(Q(\boldsymbol{\theta}))$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ lies between $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$. With some simple algebra, we can calculate that $R(\boldsymbol{\theta}) = -E\{H^T(\boldsymbol{\theta})H(\boldsymbol{\theta})\}-\lambda I_D$, where I_D is a diagonal matrix with all non-diagonal elements as zero and diagonal elements as $(0,0,1,\ldots,1,0,\ldots,0)_{1\times p}$ where 1 occurred in the order of 3 to $2+K$. And we denote $\Omega(\boldsymbol{\theta}) = E\{H^T(\boldsymbol{\theta})H(\boldsymbol{\theta})\}$. For simplicity, we use the same notation $\mathbb{G}_n(.)$ in *van der Vaart* (2000) to denote the empirical process, i.e., $\mathbb{G}_n(f) = n^{1/2}\{\mathbb{P}_n(f) - P(f)\}$, where P to denote the marginal law of observations

and $\mathbb{P}_n$ to denote the empirical distribution. Equation (5.5) can be re-organized as

$$\begin{aligned}\mathbb{G}_n(Q(\widehat{\boldsymbol{\theta}}_n)) &= \mathbb{G}_n\{Q(\widehat{\boldsymbol{\theta}}_n) - Q(\boldsymbol{\theta}_0)\} + \mathbb{G}_n\{Q(\boldsymbol{\theta}_0)\} \\ &= -\sqrt{n}\{\Omega(\boldsymbol{\theta}^*) + \lambda I_D\}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0). \qquad (5.6)\end{aligned}$$

As $\widehat{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}_0$ in probability and $\boldsymbol{\theta}^*$ lies between $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$, it follows that $\boldsymbol{\theta}^*$ converges to $\boldsymbol{\theta}_0$ in probability when $n \to \infty$. As each component in $\Omega(\boldsymbol{\theta})$ is continuous with respect to $\boldsymbol{\theta}$, by continuous mapping theorem, $\Omega(\boldsymbol{\theta}^*)$ converges to $\Omega(\boldsymbol{\theta}_0)$ as $n \to \infty$. Also because $\lambda = o(n^{-1/2})$, we have $\Omega(\boldsymbol{\theta}^*) + \lambda I_D$ also converges to $\Omega(\boldsymbol{\theta}_0)$. According to the same proof procedure in Theorem 2 of Section 2.4.2, we can show that $Q(\boldsymbol{\theta})$ belongs to the Donsker class, which implies asymptotical equicontinuity, and further we can have $\mathbb{G}_n\{Q(\widehat{\boldsymbol{\theta}}_n) - Q(\boldsymbol{\theta}_0)\} = o_p(1)$. Also, it is easy to check that $\mathbb{G}_n\{Q(\boldsymbol{\theta}_0)\} = \mathbb{G}_n\{Q_H(\boldsymbol{\theta}_0)\}$, where $Q_H(\boldsymbol{\theta}) = H(\boldsymbol{\theta})\{Y - \mu(\boldsymbol{V};\boldsymbol{\theta})\}$. By central limit theorem, $-\mathbb{G}_n(Q(\boldsymbol{\theta}_0))$ converges in distribution to a normal distribution $\mathcal{N}\Big(0, P\{Q_H(\boldsymbol{\theta}_0)Q_H^T(\boldsymbol{\theta}_0)\}\Big)$. With some algebra, we can show that $P\{Q_H(\boldsymbol{\theta}_0)Q_H^T(\boldsymbol{\theta}_0)\} = \sigma^2\Omega(\boldsymbol{\theta}_0)$. Therefore, by Slutsky's theorem, it follows that $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges in distribution to $N(0, \sigma^2\Omega^{-1}(\boldsymbol{\theta}_0))$. □

To make statistical inference, the asymptotic variance can be consistently estimated by $\widehat{V} = \widehat{\sigma}^2\widehat{\Omega}^{-1}(\widehat{\boldsymbol{\theta}}_n)$, where $\widehat{\Omega}^{-1}(\widehat{\boldsymbol{\theta}}_n) = \frac{1}{n}\sum_{i=1}^n H_i^T(\widehat{\boldsymbol{\theta}}_n)H_i(\widehat{\boldsymbol{\theta}}_n)$. According to Chapter 3.14 in *Ruppert et al.* (2003), the estimation of σ^2 can be obtained via $\widehat{\sigma}^2 = \sum_{i=1}^n(Y_i - \widehat{\mu}_i)^2/df_{res}$, where $df_{res} = n - 2tr(\boldsymbol{S}_\lambda) + tr(\boldsymbol{S}_\lambda\boldsymbol{S}_\lambda^T)$.

Because λ goes to 0 sufficiently fast as n goes to infinity, the asymptotic variance $\sigma^2\Omega^{-1}(\boldsymbol{\theta}_0)$ in Result 4 does not involve λ. However, λ may not close enough to zero in a finite sample situation, and thus it is not appropriate to ignore λ in the variance estimation from the numerical perspective. Therefore, we further consider the variance estimation by treating λ as a fixed value. We note that the proposed esti-

mating equation (5.4), that is $\frac{1}{n}\sum_{i=1}^{n} Q_i(\boldsymbol{\theta}) = 0$, is an unbiased estimating equation of $E(Q(\boldsymbol{\theta})) = 0$, whose solution is denoted as $\boldsymbol{\theta}_{\lambda 0}$. Therefore, according to Theorem 1 and Theorem 2, $\widehat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}_{\lambda 0}$ and $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{\lambda 0})$ converges in distribution to $N(0, V_1^{-1}(\boldsymbol{\theta}_{\lambda 0})V_2(\boldsymbol{\theta}_{\lambda 0})V_1^{-1}(\boldsymbol{\theta}_{\lambda 0}))$, where $V_1(\boldsymbol{\theta})$ is the negative modified derivative of $E(Q(\boldsymbol{\theta}))$ with respect to $\boldsymbol{\theta}$ and $V_2(\boldsymbol{\theta}) = E(Q(\boldsymbol{\theta})Q^T(\boldsymbol{\theta}))$. With some algebra, we can calculate that $V_1(\boldsymbol{\theta}_0) = E\{H^T(\boldsymbol{\theta}_0)H(\boldsymbol{\theta}_0)\} + \lambda I_D$ and $V_2(\boldsymbol{\theta}_0) = \sigma^2 E(H(\boldsymbol{\theta}_0)H^T(\boldsymbol{\theta}_0)) - D_\lambda(\boldsymbol{\theta}_0)D_\lambda^T(\boldsymbol{\theta}_0)$, where I_D is a diagonal matrix with all non-diagonal elements as zero and diagonal elements as $(0, 0, 1, \ldots, 1, 0, \ldots, 0)_{1\times p}$ where 1 occurred in the order of 3 to $2 + K$, and $D_\lambda(\boldsymbol{\theta})$ is the same function appears in the proposed estimating equation (5.4). Thus, when λ is treated as fixed, we can provide a sandwich estimator of the covariance matrix $V_{sw} = \widehat{V}_1^{-1}(\widehat{\boldsymbol{\theta}}_n)\widehat{V}_2(\widehat{\boldsymbol{\theta}}_n)\widehat{V}_1^{-1}(\widehat{\boldsymbol{\theta}}_n)$, where $\widehat{V}_1(\widehat{\boldsymbol{\theta}}_n) = \frac{1}{n}\sum_{i=1}^{n} H_i^T(\widehat{\boldsymbol{\theta}}_n)H_i(\widehat{\boldsymbol{\theta}}_n) + \lambda I_D$ and $\widehat{V}_2(\widehat{\boldsymbol{\theta}}_n) = \widehat{\sigma}^2[\frac{1}{n}\sum_{i=1}^{n} H_i^T(\widehat{\boldsymbol{\theta}}_n)H_i(\widehat{\boldsymbol{\theta}}_n)] - D_\lambda(\widehat{\boldsymbol{\theta}}_n)D_\lambda^T(\widehat{\boldsymbol{\theta}}_n)$. Bootstrapping is also an option for the estimation of the variance for the proposed estimator $\widehat{\boldsymbol{\theta}}_n$. And we will compare these three different variance estimation methods in the simulation section.

5.3 Simulation Studies

We conducted a series of simulation studies with either linear or nonlinear patterns above the threshold to evaluate our proposed method for the constrained penalized spline model and compare its performance with the constrained linear spline model. The true generating models for each setup and the choice of λ for the constrained penalized spline model were summarized in Table 5.1, and the true generating models were also visualized in Figure 5.1. In all simulation scenarios, X was assumed to follow a uniform distribution $(0, 50)$, the error term ϵ was assumed to follow a normal distribution with mean zero and standard deviation 12, the true value of the threshold was set at 15, and the initial value of the threshold for the proposed algorithm was set as 30. For all scenarios of the constrained penalized spline model, the number of knots

K was set as 20. Both the constrained penalized spline model and the constrained linear spline model were evaluated under 1000 Monte Carlo replicates with sample sizes varies from $n = 500, 1000, 2500$, and 5000 in all simulation scenarios. And the threshold was estimated through the two-step Newton-Raphson algorithm for both the constrained penalized spline model and the constrained linear spline model.

Table 5.1: True model generating functions and true model coefficients for four simulation setups. λ is only for the constrained penalized spline model.

Setup	**Model**	λ
1	$E(Y\|X) = \frac{1}{6}(X-10)^2 - \frac{25}{6}$ if $x \geq 15; 0$ if $x < 15$	$\frac{250}{n}$
2	$E(Y\|X) = 245 - \frac{1}{5}(X-50)^2$ if $x \geq 15; 0$ if $x < 15$	$\frac{200}{n}$
3	$E(Y\|X) = 7X - 105$ if $x \geq 15; 0$ if $x < 15$	$\frac{250}{n}$
4	$E(Y\|X) = 250$ if $x \geq 40; 10X - 150$ if $x \geq 15$ & $x < 40; 0$ if $x < 15$	$\frac{120}{n}$

For each scenario with a specific sample size, we reported bias, Monte Carlo standard deviation (MCSD), root mean squared errors (RMSE) and the average of standard error derived from the bootstrapping (AVESE_b) for both the constrained penalized spline model and the constrained linear spline model. The bootstrapping was implemented using the R package "boot" (*Davison and Hinkley*, 1997; *Canty and Ripley*, 2020) with 200 bootstrap replicates for all scenarios. To better illustrate the performance of different variance estimation methods from asymptotic variance as described in Chapter 5.2.4, we further provided two average of standard errors, i.e., AVESE_1 and AVESE_2, for the constrained penalized spline model. The AVESE_1 was derived from the asymptotic variance from Result 4, and the AVESE_2 was derived from the sandwich estimation V_{sw} when treating λ as a fixed value.

Simulation results are summarized in Table 5.2. For both the constrained penalized spline model and the constrained linear spline model, the proposed two-step NR algorithm has a 100% convergence rate of the algorithm in all scenarios. For the constrained penalized spline model, the bias is generally small in all scenarios, with

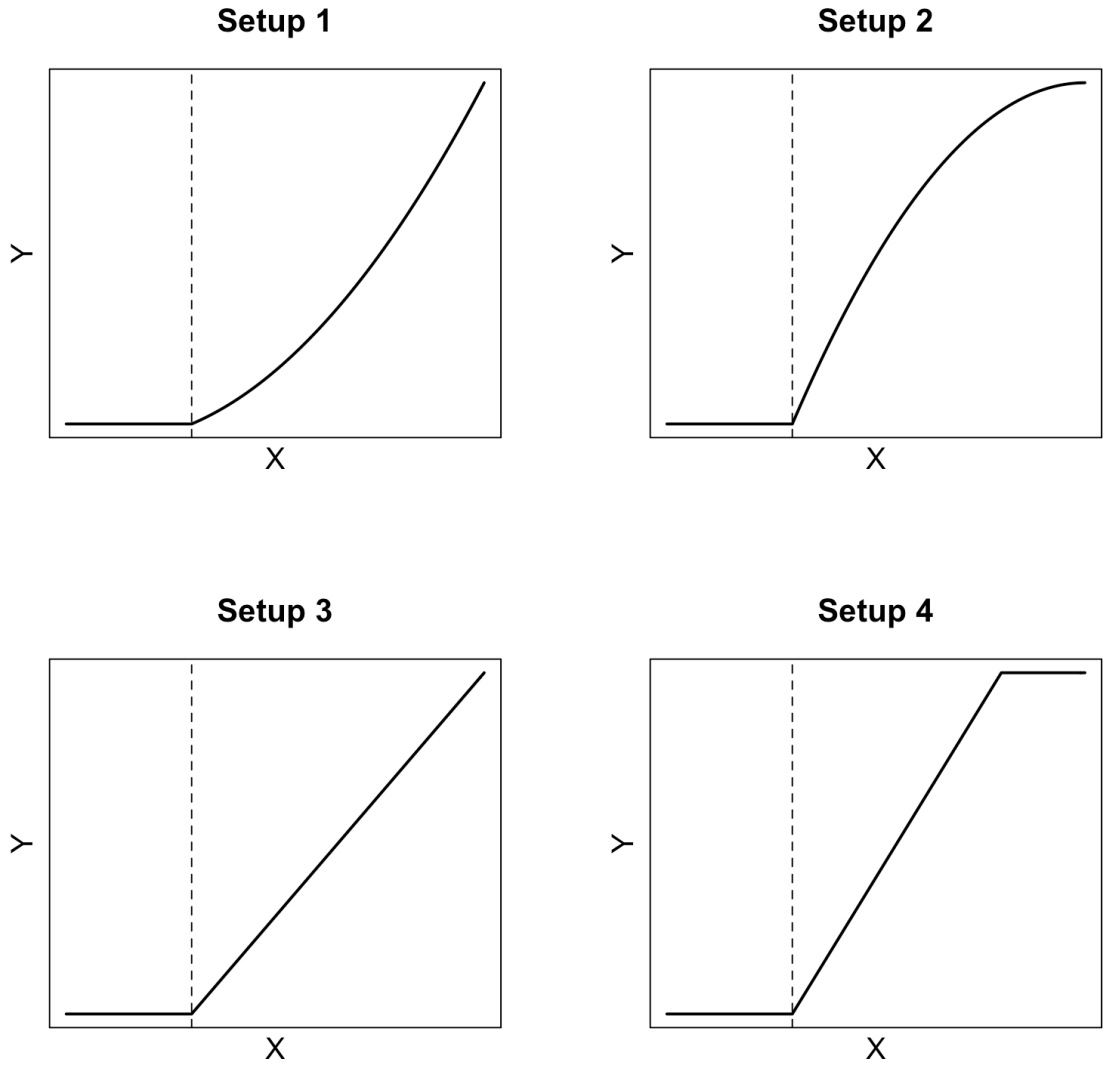

Figure 5.1: Figures of four simulation setups of either linear or non-linear patterns above the threshold.

Table 5.2: Simulation results based on 1000 Monte Carlo data sets for the constrained linear spline model and the constrained penalized spline model. MCSD: Monte Caro standard deviation; RMSE: root mean squared error; AVESE_b: average of standard error derived from the bootstrap; AVESE_1: average of standard error derived from the asymptotic variance in Result 4; AVESE_2: average of standard error derived when λ is treated as fixed.

Model	n	Bias	MCSD	RMSE	AVESE_b	AVESE_1	AVESE_2
				Setup 1			
Constrained Penalized Spline	500	2.002	0.959	2.220	0.928	1.546	0.729
Constrained Linear Spline		7.938	0.422	7.950	0.439	-	-
Constrained Penalized Spline	1000	1.647	0.695	1.787	0.717	1.094	0.583
Constrained Linear Spline		7.930	0.301	7.936	0.309	-	-
Constrained Penalized Spline	2500	1.500	0.452	1.567	0.454	0.681	0.408
Constrained Linear Spline		7.952	0.198	7.955	0.195	-	-
Constrained Penalized Spline	5000	1.475	0.309	1.507	0.322	0.467	0.303
Constrained Linear Spline		7.962	0.140	7.964	0.139	-	-
				Setup 2			
Constrained Penalized Spline	500	-0.158	0.216	0.267	0.220	0.457	0.207
Constrained Linear Spline		-4.373	0.342	4.386	0.342	-	-
Constrained Penalized Spline	1000	-0.116	0.163	0.200	0.163	0.311	0.156
Constrained Linear Spline		-4.366	0.246	4.372	0.244	-	-
Constrained Penalized Spline	2500	-0.074	0.113	0.135	0.113	0.190	0.108
Constrained Linear Spline		-4.377	0.152	4.379	0.155	-	-
Constrained Penalized Spline	5000	-0.050	0.083	0.097	0.087	0.132	0.082
Constrained Linear Spline		-4.379	0.110	4.380	0.109	-	-
				Setup 3			
Constrained Penalized Spline	500	0.035	0.411	0.413	0.423	0.839	0.377
Constrained Linear Spline		0.011	0.236	0.237	0.237	-	-
Constrained Penalized Spline	1000	0.021	0.303	0.304	0.315	0.570	0.286
Constrained Linear Spline		0.007	0.167	0.167	0.166	-	-
Constrained Penalized Spline	2500	-0.004	0.224	0.224	0.217	0.358	0.202
Constrained Linear Spline		0.003	0.105	0.105	0.104	-	-
Constrained Penalized Spline	5000	0.015	0.164	0.164	0.179	0.253	0.155
Constrained Linear Spline		0.004	0.073	0.073	0.074	-	-
				Setup 4			
Constrained Penalized Spline	500	0.055	0.302	0.306	0.310	0.585	0.284
Constrained Linear Spline		-2.222	0.276	2.239	0.282	-	-
Constrained Penalized Spline	1000	0.032	0.226	0.229	0.234	0.398	0.216
Constrained Linear Spline		-2.214	0.196	2.223	0.198	-	-
Constrained Penalized Spline	2500	0.019	0.163	0.164	0.166	0.249	0.152
Constrained Linear Spline		-2.225	0.123	2.228	0.125	-	-
Constrained Penalized Spline	5000	0.015	0.120	0.121	0.128	0.175	0.116
Constrained Linear Spline		-2.221	0.087	2.223	0.088	-	-

a decreasing trend approaching zero when the sample size n increases. Biases in Setup 1 are slightly larger than the rest three setups for the constrained penalized spline model, which can be explained by a more gentle increasing pattern after the threshold. When the pattern is non-linear above the threshold as designed in Setup 1, Setup 2, and Setup 4, there are systematic biases across all sample sizes for the constrained linear spline model. Specifically, the bias is positive for Setup 1, which assumes an increasing convex curve after the threshold. And the bias is negative for Setup 2 and Setup 4, which assume an increasing concave curve and an increasing linear-step pattern after the threshold, respectively. When the true model is the constrained linear spline model, the bias is small and decreases to zero when the sample size increases in Setup 3. For the variability, the constrained penalized spline model has a larger MCSD than the constrained linear spline model in Setup 1, Setup 3, and Setup 4, while less MCSD in Setup 2. When considering the root mean squared error which accounts for both the bias and the variability, the constrained penalized spline model has a better performance than the constrained linear spline model when a non-linear pattern exists above the threshold as in Setup 1, Setup 2, and Setup 4. And the constrained linear spline model only outperforms when it is the true model. In general, we recommend researchers use the constrained penalized spline model to estimate the threshold in real applications unless researchers have the scientific knowledge to support a linear association above the threshold. In terms of the average of standard error for the constrained penalized spline model, the AVESE_1 is generally larger than the MCSD, and AVESE_2 is generally smaller than the MCSD, which is consistent with the observation in *Yu and Ruppert* (2002). Comparing three different approaches to estimate the variance, AVESE_b from the bootstrapping provides the closest estimation of the MCSD. Though bootstrapping requires intensive computation resources and much more computing time than the other two approaches, i.e., AVESE_1 and AVESE_2, based on our simulation studies,

we recommend using bootstrapping to estimate the variance of the threshold in real applications, as researchers usually do not need to replicate their analysis thousands of times as in simulation studies.

5.4 Application

We applied the proposed method to pregnant women residing in Mexico City in 2009 to 2014 participating in the Pregnancy Research on Inflammation, Nutrition, & City Environment: Systematic Analyses (PRINCESA) cohort (*Osornio-Vargas et al.*, 2013). The application research aim was to evaluate the threshold for the impact of sulfur dioxide (SO_2) and nitrogen dioxide (NO_2) on birth weight. Air pollution data, i.e. SO_2 and NO_2, were obtained from the Mexico City Atmospheric Monitoring System and were summarized as the average value of the pollutant values in all the days of pregnancy.

Our analysis is based on 757 women participants with available air pollution data and birth weight data. NO_2 and SO_2 were reported in parts-per-billion (ppb), and birth weight was reported in gram (g). Patient characteristics adjusted in the analysis included maternal age of participant at screening, pre-gestational body mass index (BMI), highest level of education completed, martial status, parity, and baby's gender. Descriptive statistics for all variables in the analysis were presented in Table 5.3. Missing covariates were imputed. Specially, one subject with missing pre-gestational body mass index (BMI) was imputed with the mean of rest subjects, three subjects with baby's gender missing and three subjects with marital status missing were imputed with the highest frequent category respectively, and 75 subjects with highest level of education completed missing and 118 subjects with parity missing were imputed by setting the missing value into the unknown category. The number of knots K was set as 5, and the choice of λ_n was based on the generalized cross-validation

method. The standard error (SE) was obtained by bootstrapping with 200 bootstrap replicates.

Table 5.3: Descriptive statistics for the application data set ($n = 757$).

Outcome	**Mean**	**Std**
Birth Weight (g)	3025.98	509.60
Factor of Interest	**Mean**	**Std**
Nitrogen Dioxide (ppb)	32.65	2.68
Sulfur Dioxide (ppb)	5.14	0.82
Covariates	**Mean**	**Std**
Pre-gestational Body Mass Index	25.72	5.22
Age of participant at Screening	25.06	5.90
	Frequency	**Percent**
Highest Level of Education Completed		
Primary school or no school	75	11.00
Secondary school	311	45.60
Vocational/technical school or 2 year college (associate's degree)	245	35.92
Bachelor's degree	51	7.48
Baby's Gender		
Female	388	51.46
Male	366	48.54
Marital Status		
Single or divorced or widow	196	25.99
Married	164	21.75
Living together but not married	394	52.25
Parity		
0	274	42.88
1	201	31.46
2	122	19.09
$\geq$3	42	6.57

The estimated change-point for the effect of SO_2 is located at 7.17 (SE=0.09), and for the effect of NO_2 is located at 37.1 (SE=0.89) in the univariate analysis, which were shown in Figure 5.2. According to Figure 5.2, there were apparently decreasing pattern of the birth weight after the estimated threshold of both SO_2 and NO_2. This observation was consistent with the knowledge that a higher concentration of air pollutants during the pregnancy would lead to more damage for the fetus, which was reflected as a lower birth weight in this analysis. After adjusting for covariates listed in the above paragraph, estimated change-point for the effect of SO_2 is located at 7.13

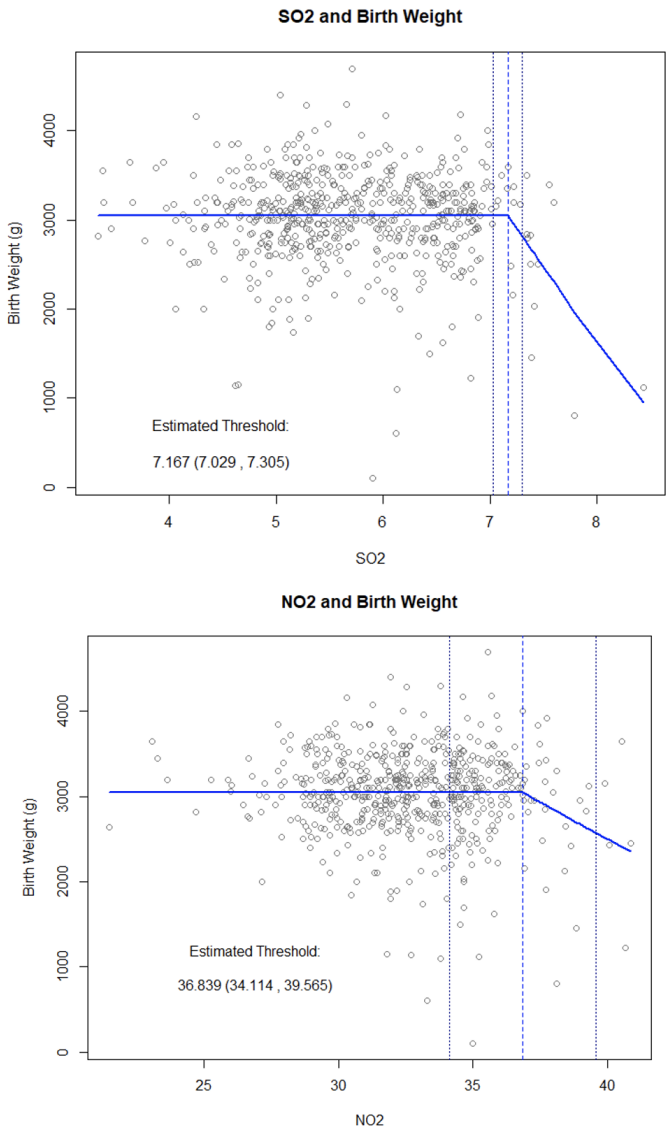

Figure 5.2: Scatterplots of sulfur dioxide (SO2) and nitrogen dioxide (NO2) versus birth weights. The solid line is the fitted regression, and three dotted vertical lines are estimated threshold (middle) and corresponding 95% confidence intervals (left and right).

(SE=0.08), and for the effect of NO_2 is located at 37.0 (SE=0.62). According to a retrospective cohort study by *Lin et al.* (2004), they suggested an significant increase in terms of low birth weight risk exceeding 11.4 ppb during pregnancy compared to low exposure (<7.1 ppb). In the study of *Lin et al.* (2004), the cutoff of low exposure of SO_2 was chosen as the 25th quantile with the number as 7.1 ppb, which is consistent with our estimated threshold 7.13 ppb. Another early study of NO_2 and preterm birth by *Llop et al.* (2010) in Spain during 2003 to 2005 reported a threshold level of NO_2 at 46.2 ppb throughout the entire pregnancy, which was about 10 ppb higher than our estimated threshold. We wanted to point out that our lower NO_2 threshold relative to *Llop et al.* (2010) was reasonable. The reason was that our estimated NO_2 threshold was the start of decreasing birth weight, which was not the clinical defined low birth weight (lower than 2,500 grams), while *Llop et al.* (2010) considered the preterm birth from the clinical definition as less than 37 weeks. However, we emphasised that although not lower than the clinical warning of 2,500 grams in birth weight, it was important to avoid exposing to NO_2 higher than 37 ppb during the entire pregnancy.

5.5 Discussion

In this chapter, we introduce and discuss the estimation of threshold in a constrained penalized spline model, where a constraint is imposed to model the threshold effect, and the unknown association after the threshold is modeled via the penalized spline with knots. Specifically, we introduced knots, which evenly distributed between the threshold and the maximum value of the factor of interest and thus can be determined by the threshold and data automatically, to account for the flexibility of either linear or non-linear association.

Due to the nondifferentiability problem at the threshold, estimation of the threshold in a constrained penalized spline model differs from the usual problem of model fitting

or prediction in penalized spline models. Motivated by the minimization of the loss function (5.2), we introduced the proposed estimating equation (5.4) based on the modified derivative idea. To facilitate computational efficiency, we further extended the two-step NR algorithm to the constrained penalized spline model, which separates the updating procedures of the threshold and the penalized spline in different steps. Specifically, when the threshold is updated, knots are also automatically updated to model the flexibility above the threshold using the penalized spline in each iteration of the two-step NR algorithm. Although the extension of the modified derivative idea looks intuitive, the proposed estimating equation generated from the modified derivative idea may not guarantee to provide a valid estimator. Therefore, besides studying the two-step NR algorithm, we further explored the asymptotic properties of the proposed estimator. Assuming a fixed number of knots, we showed that the proposed estimator is equipped with the property of consistency and asymptotic normality. Furthermore, we discussed three variance estimation methods of the threshold, including (1) estimated from the asymptotic variance, (2) estimated when treating the smoothing parameter λ as a fixed value, and (3) bootstrapping. According to simulation studies, the first two methods tend to slightly over- and under-estimate the variance, which is consistent with the observation from existing literature (*Yu and Ruppert*, 2002), and the bootstrapping method provides the closest variance estimation.

In biomedical research, the estimation of threshold has important applications. Although researchers may have prior knowledge of the threshold effect, they may only have a general idea about the pattern above the threshold, such as an increasing or decreasing pattern. The constrained linear spline model is one option to estimate the threshold and has been discussed thoroughly in Chapter IV. However, the constrained linear spline model assumes a linear pattern above the threshold, which is likely to be violated in reality. In this chapter, we compared the performance of

the proposed constrained penalized spline model with the constrained linear spline model through extensive simulation studies. For either linear or nonlinear patterns, the constrained penalized spline model shows effectiveness and robustness with small biases. However, when the pattern is nonlinear, the constrained linear spline model will lead to systematic biases even with large sample sizes and thus performs worse than the constrained penalized spline model. And the constrained linear spline model only performs better in terms of both bias and variability when the true pattern is linear. Furthermore, the constrained linear spline model performs well in the air pollution application, where the association above the threshold is modeled approximately linear. In real applications, we recommend researchers choose the constrained penalized spline model, as it allows to capture of linear and nonlinear patterns above the threshold.

In summary, we proposed a threshold estimation method in the constrained penalized spline model, which takes advantage of both posing a threshold effect and allows an unknown pattern above the threshold. Although the constrained penalized spline model we studied in this chapter focused on continuous outcomes with one factor of interest of the threshold effect, this idea can be generalized to broader research areas in future studies, such as considering the factor of interest has an interaction threshold effect, general outcomes including binary and count data, correlated or multivariate data, and survival data.

APPENDICES

APPENDIX A

Proofs

A.1 Chapter II: Proofs

A.1.1 Proof of Lemma 1

Proof. Because the map $\boldsymbol{\zeta} \mapsto \sqrt{p(\boldsymbol{w};\boldsymbol{\zeta})}$ is continuous differentiable except at finite single points when $\tau_k = x_k$, for $k = 1, \ldots, K$. At all other points the derivative exits and equals to $R(w;\boldsymbol{\zeta}) = \frac{1}{2} S_{\boldsymbol{\zeta}}(\boldsymbol{w};\boldsymbol{\zeta})\sqrt{p(\boldsymbol{w};\boldsymbol{\zeta})}$. Therefore, the difference of $\sqrt{p(\boldsymbol{w};\boldsymbol{\zeta}+\boldsymbol{h})} - \sqrt{p(\boldsymbol{w};\boldsymbol{\zeta})}$ could still be written as the integral of its derivative, i.e. $\sqrt{p(\boldsymbol{w};\boldsymbol{\zeta}+\boldsymbol{h})} - \sqrt{p(\boldsymbol{w};\boldsymbol{\zeta})} = \int_0^1 \boldsymbol{h}^T R(\boldsymbol{w};\boldsymbol{\zeta}+u\boldsymbol{h})du$. By Jensen's inequality, we can conclude that

$$\begin{aligned}\Big\{\int \frac{\sqrt{p(\boldsymbol{w};\boldsymbol{\zeta}+t\boldsymbol{h}_t)} - \sqrt{p(\boldsymbol{w};\boldsymbol{\zeta})}}{t} dv(\boldsymbol{w})\Big\}^2 &\leq \int\int_0^1 \{\boldsymbol{h}_t^T R(\boldsymbol{w};\boldsymbol{\zeta}+ut\boldsymbol{h}_t)\}^2 du dv(\boldsymbol{w}) \\ &= \frac{1}{4}\int \boldsymbol{h}_t^T I(\boldsymbol{w};\boldsymbol{\zeta}+ut\boldsymbol{h}_t)\boldsymbol{h}_t du.\end{aligned}$$

The last equation follows by Fubini's theorem and the definition of the Fisher information matrix, that is, $I(\boldsymbol{W};\boldsymbol{\zeta}) = E[S_{\boldsymbol{\zeta}}(\boldsymbol{W};\boldsymbol{\zeta})S_{\boldsymbol{\zeta}}^T(\boldsymbol{W};\boldsymbol{\zeta})] = 4\int R(\boldsymbol{w};\boldsymbol{\zeta})R^T(\boldsymbol{w};\boldsymbol{\zeta})dv(\boldsymbol{w})$. And because $S_{\boldsymbol{\zeta}}(\boldsymbol{W};\boldsymbol{\zeta})$ is a continuous function of $\boldsymbol{\zeta}$, thus $I_{\boldsymbol{\zeta}}(\boldsymbol{W};\boldsymbol{\zeta})$ is also a contin-

uous function of $\boldsymbol{\zeta}$. Then by the continuous mapping theorem,

$$\frac{1}{4}\int \boldsymbol{h}_t^T I(\boldsymbol{w};\boldsymbol{\zeta}+ut\boldsymbol{h}_t)\boldsymbol{h}_t du \to \frac{1}{4}\boldsymbol{h}^T I(\boldsymbol{w};\boldsymbol{\zeta})\boldsymbol{h} = \int \{\boldsymbol{h}^T R(\boldsymbol{w};\boldsymbol{\zeta})\}^2 dv(\boldsymbol{w}), \text{ as } \boldsymbol{h}_t \to \boldsymbol{h}.$$

And because the map $\boldsymbol{\zeta} \mapsto \sqrt{p(\boldsymbol{w};\boldsymbol{\zeta})}$ is differentiable almost everywhere in v-measure,

$$\frac{\sqrt{p(\boldsymbol{w};\boldsymbol{\zeta}+t\boldsymbol{h}_t)}-\sqrt{p(\boldsymbol{w};\boldsymbol{\zeta})}}{t} \to \boldsymbol{h}^T R(\boldsymbol{w};\boldsymbol{\zeta}) \text{ } v\text{-almost everywhere.}$$

And thus, the integrand in

$$\int \Big[\frac{\sqrt{p(\boldsymbol{w};\boldsymbol{\zeta}+t\boldsymbol{h}_t)}-\sqrt{p(\boldsymbol{w};\boldsymbol{\zeta})}}{t} - \boldsymbol{h}^T R(\boldsymbol{w};\boldsymbol{\zeta})\Big]^2 dv(\boldsymbol{w})$$

converges pointwise to zero in v-measure. Then by Proposition 2.29 in *van der Vaart* (2000), the above integral converges to 0, that is,

$$\int \Big[\frac{\sqrt{p(w;\boldsymbol{\zeta}+t\boldsymbol{h}_t)}-\sqrt{p(w;\boldsymbol{\zeta})}}{t} - \boldsymbol{h}^T R(w;\boldsymbol{\zeta})\Big]^2 dv(w) = o(1).$$

Denote $\widetilde{\boldsymbol{h}} = t\boldsymbol{h}_t$. Because $\boldsymbol{h}_t \to \boldsymbol{h}$, the above equation becomes

$$\int \Big[\sqrt{p(\boldsymbol{w};\boldsymbol{\zeta}+\widetilde{\boldsymbol{h}})}-\sqrt{p(\boldsymbol{w};\boldsymbol{\zeta})} - \widetilde{\boldsymbol{h}}^T R(\boldsymbol{w};\boldsymbol{\zeta})\Big]^2 dv(\boldsymbol{w}) = o(t^2) = o(||\widetilde{\boldsymbol{h}}||^2).$$

as $t \to 0$, which is equivalent to $\widetilde{\boldsymbol{h}} \to 0$. By definition, $p(\boldsymbol{w};\boldsymbol{\zeta})$ is DQM at $\boldsymbol{\zeta}$. □

A.1.2 Proof of Lemma 2

Proof. Denote $\boldsymbol{\zeta}_n = \boldsymbol{\zeta}^0 + \boldsymbol{h}/\sqrt{n}$ with $\boldsymbol{\zeta}_n = (\boldsymbol{\theta}_n^T, \boldsymbol{\gamma}_n^T)^T$. Following same definitions in *Tsiatis* (2007), we denote $\boldsymbol{V}_n = (\boldsymbol{W}_{1n}, \ldots, \boldsymbol{W}_{nn})$, where $\boldsymbol{W}_{1n}, \ldots, \boldsymbol{W}_{nn}$ are independent and identical distributed random vectors. Also, we denote P_{0n} and P_{1n} be sequences of probability measures with density $p_{on}(\boldsymbol{v}_n) = \prod_{i=1}^n p(\boldsymbol{w}_{in};\boldsymbol{\zeta}^0)$, $\quad p_{1n}(\boldsymbol{v}_n) = \prod_{i=1}^n p(\boldsymbol{w}_{in};\boldsymbol{\zeta}_n)$. By definition, the sequence $p_{1n}(\boldsymbol{v}_n)$ is the local data generating pro-

cess (LDGP) with $n^{1/2}(\boldsymbol{\zeta}_n - \boldsymbol{\zeta}^0) \to c$, where c is a constant. From Lemma 1, we obtain the property of DQM at $\boldsymbol{\zeta}^0$. Then from Theorem 7.2 in *van der Vaart* (2000), DQM implies local asymptotic normality, which implies

$$\log\Big\{\frac{p_{1n}(\boldsymbol{V}_n)}{p_{on}(\boldsymbol{V}_n)}\Big\} \xrightarrow{D(P_{0n})} \mathcal{N}\Big(-\frac{1}{2}\boldsymbol{h}^T I_{\boldsymbol{\zeta}^0}\boldsymbol{h}, \boldsymbol{h}^T I_{\boldsymbol{\zeta}^0}\boldsymbol{h}\Big).$$

Thus, by Lemma 3.1 in *Tsiatis* (2007), the sequence P_{1n} is contiguous to the sequence P_{0n}, that is $o_{P_{0n}}(1)$ implies $o_{P_{1n}}(1)$. Then as $\widehat{\boldsymbol{\theta}}_n$ is a RAL estimator with influence function $\varphi(\boldsymbol{W})$ and the contiguity property, we have $n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = n^{-1/2}\sum_{i=1}^n \varphi(\boldsymbol{w}_{in}) + o_{P_{1n}}(1)$. With some simple algebra, this equation turns into

$$\begin{aligned} n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) &= n^{-\frac{1}{2}}\sum_{i=1}^n [\varphi(\boldsymbol{w}_{in}) - E_{\boldsymbol{\zeta}_n}\{\varphi(\boldsymbol{W})\}] \\ &\quad + n^{\frac{1}{2}} E_{\boldsymbol{\zeta}_n}\{\varphi(\boldsymbol{W})\} - n^{\frac{1}{2}}\{\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\} + o_{P_{1n}}(1). \end{aligned} \tag{A.1}$$

As $\widehat{\boldsymbol{\theta}}_n$ is regular, that is,

$$n^{1/2}\{\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\} \xrightarrow{D(P_{1n})} \mathcal{N}(0^{p\times p}, E_{\boldsymbol{\zeta}^0}\{\varphi(\boldsymbol{W})\varphi^T(\boldsymbol{W})\}). \tag{A.2}$$

Under the probability measure P_{1n}, $[\varphi(\boldsymbol{w}_{in}) - E_{\boldsymbol{\zeta}_n}\{\varphi(\boldsymbol{W})\}]$, $i = 1, \ldots, n$ are independent and identical distributed mean-zero random vectors with variance matrix $E_{\boldsymbol{\zeta}_n}\{\varphi(\boldsymbol{W})\varphi^T(\boldsymbol{W})\} - E_{\boldsymbol{\zeta}_n}\{\varphi(\boldsymbol{W})\}E_{\boldsymbol{\zeta}_n}\{\varphi^T(\boldsymbol{W})\}$. Also because $E_{\boldsymbol{\zeta}}\{\varphi(\boldsymbol{W})\}$ and $E_{\boldsymbol{\zeta}}\{\varphi(\boldsymbol{W})\varphi^T(\boldsymbol{W})\}$ are continuous in $\boldsymbol{\zeta}$ in a neighborhood of $\boldsymbol{\zeta}^0$, by continuous mapping theorem, we have $E_{\boldsymbol{\zeta}_n}\{\varphi(\boldsymbol{W})\varphi^T(\boldsymbol{W})\} \to E_{\boldsymbol{\zeta}^0}\{\varphi(\boldsymbol{W})\varphi^T(\boldsymbol{W})\}$ and $E_{\boldsymbol{\zeta}_n}\{\varphi(\boldsymbol{W})\} \to E_{\boldsymbol{\zeta}^0}\{\varphi(\boldsymbol{W})\} = 0$. Hence, by central limit theorem,

$$n^{-1/2}\sum_{i=1}^n \Big[\varphi(\boldsymbol{w}_{in}) - E_{\boldsymbol{\zeta}_n}\{\varphi(\boldsymbol{W})\}\Big] \xrightarrow{D(P_{1n})} \mathcal{N}\Big(0^{q\times 1}, E_{\boldsymbol{\zeta}^0}\{\varphi(\boldsymbol{W})\varphi^T(\boldsymbol{W})\}\Big). \tag{A.3}$$

Denote $\Gamma = (I^{q\times q}, 0^{q\times r_1})$, which is a $q \times p$ matrix. Hence, $\boldsymbol{\theta}_n - \boldsymbol{\theta}_0 = \Gamma(\boldsymbol{\zeta}_n - \boldsymbol{\zeta}_0)$. Then,

$$n^{1/2}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) = \Gamma\{n^{1/2}(\boldsymbol{\zeta}_n - \boldsymbol{\zeta}_0)\} \to \Gamma c, \text{ as } n \to \infty. \tag{A.4}$$

Because $\varphi(\boldsymbol{W})$ is a real-valued statistic with variance $E_{\boldsymbol{\zeta}^0}\{\varphi(\boldsymbol{W})\varphi^T(\boldsymbol{W})\}$ exists in a neighborhood of $\boldsymbol{\zeta}^0$, and the model is DQM at $\boldsymbol{\zeta}^0$ with score function $S_{\boldsymbol{\zeta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)$ and non-singular Fisher information $I_{\boldsymbol{\zeta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)$, according to Chapter 6.3 in *Pollard* (2000), we have

$$\frac{\partial}{\partial \boldsymbol{\zeta}} E_{\boldsymbol{\zeta}^0}\{\varphi(\boldsymbol{W})\} = E_{\boldsymbol{\zeta}^0}\{\varphi(\boldsymbol{W})S_{\boldsymbol{\zeta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)\}.$$

Therefore, by Taylor expansion, we obtain

$$\begin{aligned} E_{\boldsymbol{\zeta}_n}\{\varphi(\boldsymbol{W})\} &= E_{\boldsymbol{\zeta}^0}\{\varphi(\boldsymbol{W})\} + (\boldsymbol{\zeta}_n - \boldsymbol{\zeta}^0)\frac{\partial}{\partial \boldsymbol{\zeta}^T} E_{\boldsymbol{\zeta}^0}\{\varphi(\boldsymbol{W})\} + o(||\boldsymbol{\zeta}_n - \boldsymbol{\zeta}^0||) \\ &= (\boldsymbol{\zeta}_n - \boldsymbol{\zeta}^0)E_{\boldsymbol{\zeta}^0}\{\varphi(\boldsymbol{W})S_{\boldsymbol{\zeta}}^T(\boldsymbol{W};\boldsymbol{\zeta}^0)\} + o(||\boldsymbol{\zeta}_n - \boldsymbol{\zeta}^0||). \end{aligned} \tag{A.5}$$

Then, from the equation (A.5), we have

$$n^{1/2} E_{\boldsymbol{\zeta}_n}\{\varphi(\boldsymbol{W})\} \to c E_{\boldsymbol{\zeta}^0}\{\varphi(\boldsymbol{W})S_{\boldsymbol{\zeta}}^T(\boldsymbol{W};\boldsymbol{\zeta}^0)\}. \tag{A.6}$$

By equation (A.2), (A.3), (A.4) and (A.6), the limit of equation (A.1) implies that

$$\Big[E_{\boldsymbol{\zeta}^0}\{\varphi(\boldsymbol{W})S_{\boldsymbol{\zeta}}^T(\boldsymbol{W};\boldsymbol{\zeta}^0)\} - \Gamma\Big]c = 0^{q\times 1}.$$

Since c is arbitrary, the above equation implies $E_{\boldsymbol{\zeta}^0}\{\varphi(\boldsymbol{W})S_{\boldsymbol{\zeta}}^T(\boldsymbol{W};\boldsymbol{\zeta}^0)\} = \Gamma$, that is,

$$E\{\varphi(\boldsymbol{W})S_{\boldsymbol{\theta}}^T(\boldsymbol{W};\boldsymbol{\zeta}^0)\} = I^{q\times q} \text{ and } E\{\varphi(\boldsymbol{W})S_{\boldsymbol{\gamma}}^T(\boldsymbol{W};\boldsymbol{\zeta}^0)\} = 0^{q\times r}.$$

□

A.1.3 Proof of Lemma 3

Proof. According to Lemma 1 and Lemma 2, the semiparametric theory can be derived with similar procedures in Chapter 4 in *Tsiatis* (2007). According to Lemma 4.5 in *Tsiatis* (2007), the nuisance tangent space is given by $\Lambda = \Big\{ h^{q\times 1}(\epsilon, \boldsymbol{X}^*)$, such that $E\{h(\epsilon, \boldsymbol{X}^*)\epsilon|\boldsymbol{X}^*\} = 0^{q\times 1}\Big\}$. According to Theorem 4.8 in *Tsiatis* (2007), the space orthogonal to nuisance tangent space is given by $\Lambda^{\perp} = \Big\{ A^{q\times 1}(\boldsymbol{X}^*)\epsilon$, for all q-dimensional functions $A^{q\times 1}(\boldsymbol{X}^*)\Big\}$.

Next, we will show the equation $E\{S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)\epsilon|\boldsymbol{X}^*\} = H(\boldsymbol{X}^*;\boldsymbol{\theta}^0)$. Because the projection of any arbitrary element $h^{q\times 1}(\boldsymbol{X}^*) \in \mathcal{H}$ onto $\Lambda^{\perp}$ satisfies $\Pi[h(\epsilon, \boldsymbol{X}^*)|\Lambda^{\perp}] = E\{h(\epsilon, \boldsymbol{X}^*)\epsilon|\boldsymbol{X}^*\}\{E(\epsilon^2|\boldsymbol{X}^*)\}^{-1}\epsilon$. Therefore, the efficient score, that is, the residual after projecting the score vector with respect to $\boldsymbol{\theta}$ on to the nuisance tangent space, is given by $S_{eff}(\boldsymbol{W}, \boldsymbol{\zeta}^0) = \Pi\{S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)|\Lambda^{\perp}\} = \sigma^{-2}(\boldsymbol{X}^*)H^T(\boldsymbol{X}^*;\boldsymbol{\theta}^0)\epsilon$.

Denote $\delta_{10}(\epsilon, \boldsymbol{x}^*)$ and $\delta_{20}(\boldsymbol{x}^*)$ as the nuisance parameter fixed at the truth. Because only $\delta_{10}\{y-\mu(\boldsymbol{x}^*,\boldsymbol{\theta}), \boldsymbol{x}^*\}$ contains $\mu(\boldsymbol{x}^*,\boldsymbol{\theta})$, we can represent the $S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0)$ as follows,

$$\begin{aligned} S_{\boldsymbol{\theta}}(\boldsymbol{W};\boldsymbol{\zeta}^0) &= \frac{\partial \log\{\delta_{10}(y-\mu(\boldsymbol{x}^*,\boldsymbol{\theta}), \boldsymbol{x}^*)\}}{\partial \mu(\boldsymbol{X}^*,\boldsymbol{\theta})} H^T(\boldsymbol{X}^*;\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} \\ &= \frac{\partial \delta_{10}(\epsilon, \boldsymbol{x}^*)/\partial \mu(\boldsymbol{X}^*,\boldsymbol{\theta})}{\delta_{10}(\epsilon, \boldsymbol{x}^*)} H^T(\boldsymbol{X}^*;\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}. \end{aligned} \tag{A.7}$$

Due to the model restriction, the following equation holds for all $\boldsymbol{x}^*$ and $\boldsymbol{\theta}$,

$$0 = E(\epsilon|\boldsymbol{X}^*) = \int_{-\infty}^{\infty} \{y-\mu(\boldsymbol{X}^*,\boldsymbol{\theta})\}\delta_{10}(y-\mu(\boldsymbol{X}^*,\boldsymbol{\theta}), \boldsymbol{x}^*)dy.$$

Thus, by taking derivative of $\mu(\boldsymbol{X}^*,\boldsymbol{\theta})$ to both sides of above equation, we obtain

$$\frac{\partial}{\partial \mu}\int_{-\infty}^{\infty} \{y-\mu(\boldsymbol{X}^*,\boldsymbol{\theta})\}\delta_{10}(y-\mu(\boldsymbol{X}^*,\boldsymbol{\theta}), \boldsymbol{x}^*)dy\Big|_{\mu=\mu^0} = 0,$$

where $\mu = \mu(\boldsymbol{X}^*, \boldsymbol{\theta})$ and $\mu^0 = \mu(\boldsymbol{X}^*, \boldsymbol{\theta}^0)$. By taking the derivative inside the integral,

$$\int_{-\infty}^{\infty} -\delta_{10}(\epsilon, \boldsymbol{x}^*) + \epsilon \frac{\partial \delta_{10}(\epsilon, \boldsymbol{x}^*)}{\partial \mu(\boldsymbol{X}^*, \boldsymbol{\theta})} d\epsilon = 0$$

for all $\boldsymbol{x}^*$. Next by multiplying $H(\boldsymbol{X}^*; \boldsymbol{\theta}^0)$ to both sides of above equation and applying the equation (A.7) for $S_{\boldsymbol{\theta}}(\boldsymbol{W}; \boldsymbol{\zeta}^0)$, we get

$$\begin{aligned} 0 &= \int_{-\infty}^{\infty} -\delta_{10}(\epsilon, \boldsymbol{x}^*) H(\boldsymbol{X}^*; \boldsymbol{\theta}^0) + \int_{-\infty}^{\infty} \epsilon S_{\boldsymbol{\theta}}^T(\epsilon, \boldsymbol{x}^*; \boldsymbol{\zeta}^0) \delta_{10}(\epsilon, \boldsymbol{x}^*) d\epsilon \\ &= -H(\boldsymbol{X}^*; \boldsymbol{\theta}^0) + \int_{-\infty}^{\infty} \epsilon S_{\boldsymbol{\theta}}^T(\epsilon, \boldsymbol{x}^*; \boldsymbol{\zeta}^0) \delta_{10}(\epsilon, \boldsymbol{x}^*) d\epsilon \end{aligned}$$

for all $\boldsymbol{x}^*$. This is equivalent to $-H(\boldsymbol{X}^*; \boldsymbol{\theta}^0) + E\{\epsilon S_{\boldsymbol{\theta}}^T(\epsilon, \boldsymbol{x}^*; \boldsymbol{\zeta}^0) | \boldsymbol{X}^*\} = 0$. Therefore, we obtain the result $E\{S_{\boldsymbol{\theta}}(\boldsymbol{W}; \boldsymbol{\zeta}^0)\epsilon | \boldsymbol{X}^*\} = H(\boldsymbol{X}^*; \boldsymbol{\theta}^0)$. Then according to Theorem 4.1 in *Tsiatis* (2007), semiparametric efficiency bound is given by $\mathcal{V} = \left[E\{S_{eff}(\boldsymbol{W}, \boldsymbol{\zeta}^0) S_{eff}^T(\boldsymbol{W}, \boldsymbol{\zeta}^0)\}\right]^{-1} = \left[E\{\frac{H^T(\boldsymbol{X}^*; \boldsymbol{\theta}^0) H(\boldsymbol{X}^*; \boldsymbol{\theta}^0)}{\sigma^2(\boldsymbol{X}^*)}\}\right]^{-1}$. □

A.1.4 Proof of Proposition 1

In all following proofs (Proposition 1, Theorem 1 and Theorem 2), for convenience and without loss of generality, we assume the domain of X is positive. Also, for brevity, we denote $X_\tau = (X - \tau)^+$ and $I_\tau = I(X > \tau)$. Similarly, we denote $X_{\widetilde{\tau}} = (X - \widetilde{\tau})^+$, $X_{\widetilde{\tau}^{(t)}} = (X - \widetilde{\tau}^{(t)})^+$, $I_{\widetilde{\tau}} = I(X > \widetilde{\tau})$ and $I_{\widetilde{\tau}^{(t)}} = I(X > \widetilde{\tau}^{(t)})$.

We only give the detailed proof for a simple case $E(Y|X) = \beta_1 X + \beta_{11}(X - \tau)^+$ for Proposition 1. The proof for the multiple knots case is similar but more involved in terms of algebras. First, we will show the local convergence of the proposed algorithm. Next, we will show that when the proposed algorithm converges, it converges to the solution to $\mathbb{P}_n(Q(\boldsymbol{\theta})) = 0$, i.e. $\widehat{\boldsymbol{\theta}}_n = (\widehat{\beta}_1, \widehat{\beta}_{11}, \widehat{\tau})^T$

A.1.4.1 Local Convergence

Proof. Denote $\widehat{\tau}^{(0)}$ as the initial value for τ and $\widehat{\beta}_1^{(0)}$ and $\widehat{\beta}_{11}^{(0)}$ as the OLS estimates of β_1 and β_{11} respectively from step 1 in the proposed algorithm, when treating knot τ fixed at $\widehat{\tau}^{(0)}$. We further denote $\widehat{\boldsymbol{\theta}}^{(0)} = (\widehat{\beta}_1^{(0)}, \widehat{\beta}_{11}^{(0)}, \widehat{\tau}^{(0)})^T$. For brevity, we use $F(\boldsymbol{\theta})$ to denote $\mathbb{P}_n(Q(\boldsymbol{\theta}))$, thus the estimating equation $\mathbb{P}_n(Q(\boldsymbol{\theta})) = 0$ is equivalent to $F(\boldsymbol{\theta}) = 0$. We denote each row of $F(\boldsymbol{\theta})$ as $F_\ell(\boldsymbol{\theta})$, where $\ell = 1, 2, 3$. Also, we denote a symmetric matrix $G(\boldsymbol{\theta})$ as follows

$$G(\boldsymbol{\theta}) = \begin{pmatrix} G_{11}(\boldsymbol{\theta}) & G_{12}(\boldsymbol{\theta}) & G_{13}(\boldsymbol{\theta}) \\ G_{21}(\boldsymbol{\theta}) & G_{22}(\boldsymbol{\theta}) & G_{23}(\boldsymbol{\theta}) \\ G_{31}(\boldsymbol{\theta}) & G_{32}(\boldsymbol{\theta}) & G_{33}(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} \mathbb{P}_n(X^2) & \mathbb{P}_n(XX_\tau) & -\beta_{11}\mathbb{P}_n(XI_\tau) \\ \mathbb{P}_n(XX_\tau) & \mathbb{P}_n(X_\tau^2) & -\beta_{11}\mathbb{P}_n(X_\tau) \\ -\beta_{11}\mathbb{P}_n(XI_\tau) & -\beta_{11}\mathbb{P}_n(X_\tau) & \beta_{11}^2\mathbb{P}_n(I_\tau) \end{pmatrix}.$$

Because $\boldsymbol{\theta}$ belongs to a compact set $\boldsymbol{\Theta}$ and X has a bounded domain, all components in matrix $G(\boldsymbol{\theta})$ can be bounded by a finite constant. By Theorem 3 in *Chaney* (1990), every piecewise differentiable function is locally Lipschitz continuous. And by Corollary 4.1.1 in *Scholtes* (2012), every piecewise differentiable function is semismooth. As each row of $F(\boldsymbol{\theta})$ is obviously a piecewise differentiable function, $F(\boldsymbol{\theta})$ is locally Lipschitz continuous and semismooth. As it is easy to check that $G(\boldsymbol{\theta})$ belongs to $\partial F(\boldsymbol{\theta})$, i.e. the generalized Jacobian, according to Theorem 2.3 in *Qi and Sun* (1993), when $\widehat{\boldsymbol{\theta}}^{(0)} \to \widehat{\boldsymbol{\theta}}_n$, we have

$$\left\| F(\widehat{\boldsymbol{\theta}}_n) - F(\widehat{\boldsymbol{\theta}}^{(0)}) - G(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^{(0)}) \right\| = o(||\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^{(0)}||). \tag{A.8}$$

By definition, $F(\widehat{\boldsymbol{\theta}}_n) = 0$. According to the Step 1 in the proposed algorithm from Chapter II, the OLS method guarantees that $F_1(\widehat{\boldsymbol{\theta}}^{(0)}) = 0$ and $F_2(\widehat{\boldsymbol{\theta}}^{(0)}) = 0$. There-

fore, by the first two rows of equation (A.8), we have the following two results:

$$\begin{aligned} G_{11}(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\beta}_1 - \widehat{\beta}_1^{(0)}) + G_{12}(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\beta}_{11} - \widehat{\beta}_{11}^{(0)}) + G_{13}(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\tau} - \widehat{\tau}^{(0)}) &= o(||\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^{(0)}||), \\ G_{21}(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\beta}_1 - \widehat{\beta}_1^{(0)}) + G_{22}(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\beta}_{11} - \widehat{\beta}_{11}^{(0)}) + G_{23}(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\tau} - \widehat{\tau}^{(0)}) &= o(||\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^{(0)}||). \end{aligned}$$

Re-arranging above two equations, we have

$$\begin{aligned} \widehat{\beta}_1 - \widehat{\beta}_1^{(0)} &= \frac{G_{12}(\widehat{\boldsymbol{\theta}}^{(0)} G_{23}(\widehat{\boldsymbol{\theta}}^{(0)}) - G_{13}(\widehat{\boldsymbol{\theta}}^{(0)}) G_{22}(\widehat{\boldsymbol{\theta}}^{(0)})}{G_{11}(\widehat{\boldsymbol{\theta}}^{(0)}) G_{22}(\widehat{\boldsymbol{\theta}}^{(0)}) - G_{12}^2(\widehat{\boldsymbol{\theta}}^{(0)})} (\widehat{\tau} - \widehat{\tau}^{(0)}) \\ &+ o(||\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^{(0)}||), \qquad \text{(A.9)} \\ \widehat{\beta}_{11} - \widehat{\beta}_{11}^{(0)} &= \frac{G_{12}(\widehat{\boldsymbol{\theta}}^{(0)}) G_{13}(\widehat{\boldsymbol{\theta}}^{(0)}) - G_{11}(\widehat{\boldsymbol{\theta}}^{(0)}) G_{23}(\widehat{\boldsymbol{\theta}}^{(0)})}{G_{11}(\widehat{\boldsymbol{\theta}}^{(0)}) G_{22}(\widehat{\boldsymbol{\theta}}^{(0)}) - G_{12}^2(\widehat{\boldsymbol{\theta}}^{(0)})} (\widehat{\tau} - \widehat{\tau}^{(0)}) \\ &+ o(||\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^{(0)}||). \qquad \text{(A.10)} \end{aligned}$$

According to the last rows of equation (A.8), we have the following equation

$$\begin{aligned} F_3(\widehat{\boldsymbol{\theta}}^{(0)}) &+ G_{31}(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\beta}_1 - \widehat{\beta}_1^{(0)}) + G_{32}(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\beta}_{11} - \widehat{\beta}_{11}^{(0)}) \\ &+ G_{33}(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\tau} - \widehat{\tau}^{(0)}) = o(||\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^{(0)}||). \end{aligned}$$

Applying formulas (A.9) and (A.10) into the above equation, we have

$$\begin{aligned} & F_3(\widehat{\boldsymbol{\theta}}_n) - F_3(\widehat{\boldsymbol{\theta}}^{(0)}) - G_{33}(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\tau} - \widehat{\tau}^{(0)}) \qquad \text{(A.11)} \\ = \; & D(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\tau} - \widehat{\tau}^{(0)}) + o(||\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^{(0)}||), \\ \text{where } D(\boldsymbol{\theta}) = \; & \frac{2G_{12}(\boldsymbol{\theta}) G_{13}(\boldsymbol{\theta}) G_{23}(\boldsymbol{\theta}) - G_{13}^2(\boldsymbol{\theta}) G_{22}(\boldsymbol{\theta}) - G_{11}(\boldsymbol{\theta}) G_{23}^2(\boldsymbol{\theta})}{G_{11}(\boldsymbol{\theta}) G_{22}(\boldsymbol{\theta}) - G_{12}^2(\boldsymbol{\theta})}. \end{aligned}$$

According to Step 2 in the proposed algorithm from Chapter II, the estimate of knot τ is updated via formula $\widehat{\tau}^{(1)} = \widehat{\tau}^{(0)} - G_{33}(\widehat{\boldsymbol{\theta}}^{(0)}) F_3(\widehat{\boldsymbol{\theta}}^{(0)})$. Combing equation (A.11),

we have the following equation

$$\begin{aligned}\widehat{\tau}^{(1)} - \widehat{\tau} &= \widehat{\tau}^{(0)} - \widehat{\tau} - G_{33}(\widehat{\boldsymbol{\theta}}^{(0)})F_3(\widehat{\boldsymbol{\theta}}^{(0)}) \\ &= G_{33}^{-1}(\widehat{\boldsymbol{\theta}}^{(0)})\{F_3(\widehat{\boldsymbol{\theta}}_n) - F_3(\widehat{\boldsymbol{\theta}}^{(0)}) - G_{33}(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\tau} - \widehat{\tau}^{(0)})\} \\ &= G^*(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\tau}^{(0)} - \widehat{\tau}) + o(||\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^{(0)}||), \end{aligned} \tag{A.12}$$

where $G^*(\boldsymbol{\theta}) = -D(\boldsymbol{\theta})/G_{33}(\boldsymbol{\theta})$. We denote the upper left 2×2 matrix of $G(\boldsymbol{\theta})$ as $G_\beta(\boldsymbol{\theta})$, i.e.

$$G_\beta(\boldsymbol{\theta}) = \begin{pmatrix} G_{11}(\boldsymbol{\theta}) & G_{12}(\boldsymbol{\theta}) \\ G_{21}(\boldsymbol{\theta}) & G_{22}(\boldsymbol{\theta}) \end{pmatrix}.$$

Then, we can simplify $G^*(\boldsymbol{\theta})$ by determinant of matrices $G(\boldsymbol{\theta})$ and $G_\beta(\boldsymbol{\theta})$. Specifically,

$$G^*(\boldsymbol{\theta}) = 1 - \frac{\det\{G(\boldsymbol{\theta})\}}{\det\{G_\beta(\boldsymbol{\theta})\}G_{33}(\boldsymbol{\theta})}.$$

Because $G(\boldsymbol{\theta})$, $G_\beta(\boldsymbol{\theta})$ are positive-definite matrices and $G_{33}(\boldsymbol{\theta})$ is positive, according to Fischer's inequality, we have $\det\{G(\boldsymbol{\theta})\} < \det\{G_\beta(\boldsymbol{\theta})\}G_{33}(\boldsymbol{\theta})$. Therefore, $0 < G^*(\boldsymbol{\theta}) < 1$. According to formulas (A.9) and (A.10), it is obvious that $o(||\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^{(0)}||)$ implies $o(||\widehat{\tau} - \widehat{\tau}^{(0)}||)$. Thus, equation (A.12) implies

$$||\widehat{\tau}^{(1)} - \widehat{\tau}|| = G^*(\widehat{\boldsymbol{\theta}}^{(0)})||\widehat{\tau}^{(0)} - \widehat{\tau}|| + o(||\widehat{\tau}^{(0)} - \widehat{\tau}||).$$

Therefore, when choosing an arbitrary small constant $0 < m < 1 - G^*(\widehat{\boldsymbol{\theta}}^{(0)})$, there exists an $r_1 > 0$ such that when $\widehat{\tau}^{(0)} \in B(\widehat{\tau}, r_1)$, we have $o(||\widehat{\tau}^{(0)} - \widehat{\tau}||) < m||\widehat{\tau}^{(0)} - \widehat{\tau}||$ and thus $||\widehat{\tau}^{(1)} - \widehat{\tau}|| < ||\widehat{\tau}^{(0)} - \widehat{\tau}||$. The induction for any step t ($t \geq 1$) is similar. Therefore, the proposed two-step NR algorithm converges.

A.1.4.2 When the proposed algorithm converges, it converges to the solution to $\mathbb{P}_n(Q(\boldsymbol{\theta}))=0$

We denote $\widetilde{\boldsymbol{\theta}}_n$ as the converged result from the proposed algorithm, where $\widetilde{\boldsymbol{\theta}}_n = (\widetilde{\beta}_1, \widetilde{\beta}_{11}, \widetilde{\tau})^T$. At the tth ($t \geq 1$) iteration, the algorithm proceeds as follows.

1. Treating $\widetilde{\tau}^{(t-1)}$ as fixed, $\widetilde{\beta}_1^{(t-1)}$ and $\widetilde{\beta}_{11}^{(t-1)}$ are the OLS estimates of β_1 and β_{11}. Namely, $\widetilde{\beta}_1^{(t-1)}$ and $\widetilde{\beta}_{11}^{(t-1)}$ are the solution of the following equation

$$\mathbb{U}_1^{(t)} = \mathbb{P}_n \begin{pmatrix} X\{Y - \beta_1 X - \beta_{11}(X - \widetilde{\tau}^{(t-1)})^+\} \\ (X - \widetilde{\tau}^{(t-1)})^+\{Y - \beta_1 X - \beta_{11}(X - \widetilde{\tau}^{(t-1)})^+\} \end{pmatrix} = 0.$$

2. Updating $\widetilde{\tau}^{(t-1)}$ into $\widetilde{\tau}^{(t)}$ via $\widetilde{\tau}^{(t)} = \widetilde{\tau}^{(t-1)} - (J^{(t)})^{-1}U^{(t)}$, where $U^{(t)} = \widetilde{\beta}_{11}^{(t-1)}\mathbb{P}_n\{(Y - \widetilde{\beta}_1^{(t-1)}X - \widetilde{\beta}_{11}^{(t)}X_{\widetilde{\tau}^{(t-1)}})I_{\widetilde{\tau}^{(t-1)}}\}$ and $J^{(t)} = \{\widetilde{\beta}_{11}^{(t-1)}\}^2\mathbb{P}_n(I_{\widetilde{\tau}^{(t-1)}})$.

When $\widetilde{\tau}^{(t)}$ converges, we denote $\lim_{t\to\infty}\widetilde{\tau}^{(t)} = \widetilde{\tau}$. By taking limits to the updating procedure of τ above, we have $\lim_{t\to\infty}\widetilde{\tau}^{(t)} = \lim_{t\to\infty}\widetilde{\tau}^{(t-1)} - \lim_{t\to\infty}(J^{(t)})^{-1}U^{(t)}$, which implies

$$\lim_{t\to\infty}(J^{(t)})^{-1}U^{(t)} = \lim_{t\to\infty}\frac{\mathbb{P}_n\{(Y - \widetilde{\beta}_1^{(t-1)}X - \widetilde{\beta}_{11}^{(t-1)}X_{\widetilde{\tau}^{(t-1)}})I_{\widetilde{\tau}^{(t-1)}}\}}{\widetilde{\beta}_{11}^{(t-1)}\mathbb{P}_n(I_{\widetilde{\tau}^{(t-1)}})} = 0. \tag{A.13}$$

According to Step 1, it is clear that when $\widetilde{\tau}^{(t)}$ converges, $\widetilde{\beta}_1^{(t)}$ and $\widetilde{\beta}_{11}^{(t)}$ also converges to $\widetilde{\beta}_1$ and $\widetilde{\beta}_{11}$ respectively, as t goes to infinity. And $\widetilde{\beta}_1$ and $\widetilde{\beta}_{11}$ should be the solution of $\lim_{t\to\infty}\mathbb{U}_1^{(t)} = 0$, that is,

$$\lim_{t\to\infty}\mathbb{U}_1^{(t)} = \mathbb{P}_n \begin{pmatrix} X\{Y - \beta_1 X - \beta_{11}(X - \widetilde{\tau})^+\} \\ (X - \widetilde{\tau})^+\{Y - \beta_1 X - \beta_{11}(X - \widetilde{\tau})^+\} \end{pmatrix} = 0. \tag{A.14}$$

Combining equations (A.13) and (A.14), $\widetilde{\beta}_1$, $\widetilde{\beta}_{11}$ and $\widetilde{\tau}$ are the solution of

$$\mathbb{Q}_1 = \mathbb{P}_n\Big[H_1^T\{Y - \widetilde{\beta}_1 X - \widetilde{\beta}_{11}(X-\widetilde{\tau})^+\}\Big] = 0, \text{ where } H_1 = \Big(X, (X-\widetilde{\tau})^+, I(X > \widetilde{\tau})\Big).$$

From the definition, $\widehat{\boldsymbol{\theta}}_n = (\widehat{\beta}_1, \widehat{\beta}_{11}, \widehat{\tau})^T$ is the solution of the proposed estimating equation $\mathbb{P}_n(Q(\widehat{\boldsymbol{\theta}}_n)) = \mathbb{P}_n\Big[H^T\{Y - \widehat{\beta}_1 X - \widehat{\beta}_{11}(X-\widehat{\tau})^+\}\Big] = 0$, where $H = \Big(X, (X-\tau)^+, -\widehat{\beta}_{11} I(X > \widehat{\tau})\Big)$. It is obvious to see that the solution of $\mathbb{Q}_1 = 0$ and $\mathbb{P}_n(Q(\boldsymbol{\theta})) = 0$ are exactly the same. Hence, when proposed algorithm converges, the converged result is the solution of $\mathbb{P}_n(Q(\boldsymbol{\theta})) = 0$. □

A.1.5 Proof of Theorem 1

In the proofs of Theorem 1 and Theorem 2, we consider the simple model $E(Y|X) = \beta_0 + \beta_1 X + \beta_{11}(X-\tau)^+$. The proof for the multiple knots case is similar but more involved in terms of algebras. First, we will prove that $\boldsymbol{\theta}^0$ is the unique solution to $P(Q(\boldsymbol{\theta})) = 0$. As $P(Q(\boldsymbol{\theta}))$ is a continuous function of $\boldsymbol{\theta}$ in the compact set $\boldsymbol{\Theta}$, and $\boldsymbol{\theta}^0$ is the unique zero of $P(Q(\boldsymbol{\theta})) = 0$, then this unique solution $\boldsymbol{\theta}^0$ is well-separated (*van der Vaart*, 2000). Next, we will show $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\mathbb{P}_n(Q(\boldsymbol{\theta})) - P(Q(\boldsymbol{\theta}))| \xrightarrow{a.s} 0$. Then by Theorem 5.9 in *van der Vaart* (2000), $\widehat{\boldsymbol{\theta}}_n$ converges in probability to $\boldsymbol{\theta}^0$, where $\widehat{\boldsymbol{\theta}}_n$ is the solution to $\mathbb{P}_n(Q(\boldsymbol{\theta})) = 0$.

A.1.5.1 $\boldsymbol{\theta}^0$ is the unique solution of $P(Q(\boldsymbol{\theta})) = 0$

Proof. We consider a simpler model $E(Y|X) = \beta_1 X + \beta_{11}(X-\tau)^+$ by absorbing the intercept into Y. By definition of $Q(\boldsymbol{\theta})$ in Section 2.4.1, $P(Q(\boldsymbol{\theta})) = 0$ is written as

$$P(Q(\boldsymbol{\theta})) = \begin{pmatrix} -2P(XY - \beta_1 X^2 - \beta_{11} X X_\tau) \\ -2P(X_\tau Y - \beta_1 X X_\tau - \beta_{11} X_\tau^2) \\ 2\beta_{11} P(I_\tau Y - \beta_1 X I_\tau - \beta_{11} X_\tau I_\tau) \end{pmatrix} = 0. \tag{A.15}$$

Solving β_1 from the first equation in (A.15) we obtain $\beta_1 = \frac{P(XY)-\beta_{11}P(XX_\tau)}{P(X^2)}$. Substituting it into the second and third equations respectively leads to

$$\beta_{11} = \frac{P(X_\tau Y)P(X^2) - P(XY)P(XX_\tau)}{P(X_\tau^2)P(X^2) - \{P(XX_\tau)\}^2}, \beta_{11} = \frac{P(I_\tau Y)P(X^2) - P(XY)P(XI_\tau)}{P(I_\tau X_\tau)P(X^2) - P(XX_\tau)P(XI_\tau)}. \tag{A.16}$$

It then follows that

$$\frac{P(I_\tau Y)P(X^2) - P(XY)P(XI_\tau)}{P(I_\tau X_\tau)P(X^2) - P(XX_\tau)P(XI_\tau)} = \frac{P(X_\tau Y)P(X^2) - P(XY)P(XX_\tau)}{P(X_\tau^2)P(X^2) - \{P(XX_\tau)\}^2}.$$

Replacing Y by $\beta_1^0 X + \beta_{11}^0 X_{\tau^0} + \epsilon$ and simplifying terms, the above equation becomes

$$\begin{aligned} &P(I_\tau X_{\tau^0})P(X_\tau^2)P(X^2) - P(XX_{\tau^0})P(XI_\tau)P(X_\tau^2) \\ &-P(I_\tau X_{\tau^0})\{P(XX_\tau)\}^2 - P(X_\tau X_{\tau^0})P(X^2)P(I_\tau X_\tau) \\ &+P(I_\tau X_\tau)P(XX_{\tau^0})P(XX_\tau) + P(XX_\tau)P(XI_\tau)P(X_\tau X_{\tau^0}) = 0. \end{aligned} \tag{A.17}$$

For brevity, we denote $a_1 = P(I_\tau X_{\tau^0})P(X_\tau^2)P(X^2), a_2 = P(XX_{\tau^0})P(XI_\tau)P(X_\tau^2), a_3 = P(I_\tau X_{\tau^0})\{P(XX_\tau)\}^2, b_1 = P(X_\tau X_{\tau^0})P(X^2)P(I_\tau X_\tau), b_2 = P(I_\tau X_\tau)P(XX_{\tau^0})P(XX_\tau)$ and $b_3 = P(XX_\tau)P(XI_\tau)P(X_\tau X_{\tau^0})$. Then equation (A.17) $= a_1 - a_2 - a_3 - b_1 + b_2 + b_3 = (a_1 - b_1) + (b_2 - a_2) + (b_3 - a_3) = 0$.

Suppose $\tau \neq \tau^0$ and without loss of generality, we assume $\tau > \tau^0$. It is clear that $I_\tau = I(X > \tau)$ follows a Bernoulli distribution with $p_\tau \equiv Pr(I_\tau = 1) = Pr(X > \tau)$. By the law of total expectation, we have $P(I_\tau X_{\tau^0}) = E\{E(I_\tau(X - \tau^0)|I_\tau)\} = E\{I_\tau E(X - \tau^0|I_\tau)\} = E\{I_\tau(E(X|I_\tau) - \tau^0)\} = p_\tau\{E(X|X > \tau) - \tau^0\}$. Applying the

same idea to each component of (a_1, a_2, a_3, b_1, b_2, b_3), we obtain that

$$\begin{aligned} a_1 - b_1 &= (\tau - \tau^0) Var(X|X > \tau) p_\tau^2 P(X^2), \\ b_2 - a_2 &= -\tau Var(X|X > \tau) p_\tau^2 P(XX_{\tau^0}), \\ b_3 - a_3 &= \tau^0 Var(X|X > \tau) p_\tau^2 P(XX_\tau). \end{aligned}$$

Substituting with the above results, we further simplify equation (A.17) into

$$\{(\tau - \tau^0)P(X^2) - \tau P(XX_{\tau^0}) + \tau^0 P(XX_\tau)\} Var(X|X > \tau) p_\tau^2 = 0.$$

As $Var(X|X > \tau) > 0$ and $p_\tau^2 > 0$, the above equation becomes

$$A \equiv (\tau - \tau^0)P(X^2) - \tau P(XX_{\tau^0}) + \tau^0 P(XX_\tau) = 0. \tag{A.18}$$

Next, we will show that when $\tau > \tau^0$, $A > B > 0$, where $B = (\tau - \tau^0)\{P(X^2) - P(XX_{\tau^0}) - \tau^0 P(XI_{\tau^0})\}$. It is easy to see that $A - B = \tau^0\{P(XX_\tau) - P(XX_{\tau^0}) + (\tau - \tau^0)P(XI_{\tau^0})\} = \tau^0 P\{X(\tau - X)I(\tau^0 < X \leq \tau)\} > 0$. Regarding B, we have $B = (\tau - \tau^0)[P(X^2) - P\{X^2 I(X > \tau^0)\}] = (\tau - \tau^0)P\{X^2 I(X \leq \tau^0)\} > 0$. Therefore, when $\tau > \tau^0$, $A = (\tau - \tau^0)P(X^2) - \tau P(XX_{\tau^0}) + \tau^0 P(XX_\tau) > 0$, which contradicts with equation (A.18). Similarly, $\tau < \tau^0$ will lead to the contradiction. Therefore, $\tau = \tau^0$. Substituting $Y = \beta_1^0 X + \beta_{11}^0 X_{\tau^0} + \epsilon$ to the first equation in (A.16) leads to

$$\frac{\beta_{11}^0}{\beta_{11}} = \frac{P(X_\tau^2)P(X^2) - \{P(XX_\tau)\}^2}{P(X_\tau X_\tau^0)P(X^2) - \{P(XX_\tau)\}^2},$$

which implies that $\beta_{11} = \beta_{11}^0$. Moreover, we have

$$\beta_1 = \frac{P(XY) - \beta_{11}P(XX_\tau)}{P(X^2)} = \beta_1^0 + \frac{\beta_{11}^0 P(XX_{\tau^0}) - \beta_{11}P(XX_\tau)}{P(X^2)} = \beta_1^0.$$

Therefore, $\boldsymbol{\theta}^0$ is the unique solution to $P(Q(\boldsymbol{\theta})) = 0$. □

A.1.5.2 $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|\mathbb{P}_n(Q(\boldsymbol{\theta})) - P(Q(\boldsymbol{\theta}))|\xrightarrow{a.s} 0$

Proof. According to the definition, $Q(\boldsymbol{\theta})$ is as follows

$$Q(\boldsymbol{\theta}) = \begin{pmatrix} -2\{Y - \beta_0 - \beta_1 X - \beta_{11}(X-\tau)^+\} \\ -2\{X(Y - \beta_0 - \beta_1 X - \beta_{11}(X-\tau)^+)\} \\ -2\{(X-\tau)^+(Y - \beta_0 - \beta_1 X - \beta_{11}(X-\tau)^+)\} \\ 2\beta_{11}\{I(X>\tau)(Y - \beta_0 - \beta_1 X - \beta_{11}(X-\tau)^+)\} \end{pmatrix}. \tag{A.19}$$

Denote each row of $Q(\boldsymbol{\theta})$ as $Q_\ell(\boldsymbol{\theta})$, where $\ell = 1,2,3,4$. As $\boldsymbol{\Theta}$ is compact, it is clear that β_0, $\beta_1 X$, $\beta_{11}(X-\tau)^+$, X, $(X-\tau)^+$ and $\beta_{11}I(X>\tau)$ are all uniformly bounded monotone functions of $\boldsymbol{\theta}$ on the real line. Define classes of measurable functions $\mathcal{F}_1 = \{\beta_0 : \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$, $\mathcal{F}_2 = \{\beta_1 X : \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$, $\mathcal{F}_3 = \{\beta_{11}(X-\tau)^+ : \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$, $\mathcal{F}_4 = \{X : \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$, $\mathcal{F}_5 = \{(X-\tau)^+ : \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$ and $\mathcal{F}_6 = \{\beta_{11}I(X>\tau) : \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$. According to Theorem 2.7.5 in *van der Vaart and Wellner* (1996), bounded monotone functions have a bracketing number of order $(1/\epsilon)$, with respect to the $L_1(P)$ norm. Thus, their entropy bracketing number $\log N_{[]}(\epsilon, \mathcal{F}_i, L_1(P)) < \infty$ for all $\epsilon > 0$ $(i = 1,\ldots,6)$. As $\boldsymbol{\Theta}$ is compact, we can assume $|\beta_j| \leq w_j$ $(j = 0,1)$ and $|\beta_{11}| \leq w_2$,where w_j are finite constant. Denote $C = |\max(C_1, C_2)|$. Thus $F_1 = w_0$, $F_2 = w_1 C$, $F_3 = 2w_2 C$, $F_4 = C$, $F_5 = 2C$, and $F_6 = w_2$ are an envelope function for $\mathcal{F}_1,\ldots,\mathcal{F}_6$ respectively. As PF_i, $i = 1,\ldots,6$, are finite, according to Theorem 2.4.3 in *van der Vaart and Wellner* (1996), $\mathcal{F}_i$ belongs to the Glivenko-Cantelli class for $i = 1,\ldots,6$. Define $\mathcal{H}_\ell \equiv \phi_\ell(\mathcal{F}_1,\ldots,\mathcal{F}_6) = \{Q_\ell(\theta) : \theta\in\Theta\}$ for $\ell = 1,2,3,4$, and define $\mathcal{H} \equiv \phi(\mathcal{F}_1,\ldots,\mathcal{F}_6) = \{Q(\theta) : \theta\in\Theta\}$. We also define $H_1 = 2W_Y$, $H_2 = 2CW_Y$, $H_3 = 4CW_Y$ and $H_4 = 2w_2 W_Y$, where $W_Y = |Y| + w_0 + w_1 C + 2w_2 C$. It is easy to see that H_ℓ is an envelope function for $\mathcal{H}_\ell$ and $PH_\ell < \infty$ for $\ell = 1,2,3,4$. As ϕ_ℓ is continuous, then according to Theorem 3 in *van der Vaart and Wellner* (2000), $\mathcal{H}_\ell$ belongs to a Glivenko-Cantelli class. This implies that $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|\mathbb{P}_n(Q_\ell(\boldsymbol{\theta})) - P(Q_\ell(\boldsymbol{\theta}))|\xrightarrow{a.s} 0$. This further implies that $\mathcal{H}$ belongs to a Glivenko-Cantelli class and, as a result,

$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|\mathbb{P}_n(Q(\boldsymbol{\theta})) - P(Q(\boldsymbol{\theta}))| \xrightarrow{a.s} 0.$ □

A.1.6 Proof of Theorem 2

Proof. In the following proof, we still consider $Q(\boldsymbol{\theta})$, defined in formula (A.19). Although $Q(\boldsymbol{\theta})$ involves functions that are not differentiable, $P(Q(\boldsymbol{\theta}))$ is first-order differentiable with respect to $\boldsymbol{\theta}$ because of the integral of X. By Talyor expansion of $P(Q(\widehat{\boldsymbol{\theta}}_n))$ around $\boldsymbol{\theta}^0$,

$$P(Q(\widehat{\boldsymbol{\theta}}_n)) - P(Q(\boldsymbol{\theta}^0)) = 2V(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0), \tag{A.20}$$

where $V(\boldsymbol{\theta}^*)$ is a 4×4 matrix and defined as

$$\frac{1}{2}\Big\{\frac{\partial}{\partial\beta_0}P(Q(\boldsymbol{\theta}^{*(1)})), \frac{\partial}{\partial\beta_1}P(Q(\boldsymbol{\theta}^{*(2)})), \frac{\partial}{\partial\beta_{11}}P(Q(\boldsymbol{\theta}^{*(3)})), \frac{\partial}{\partial\tau}P(Q(\boldsymbol{\theta}^{*(4)}))\Big\}.$$

It is easy to see that $V(\boldsymbol{\theta}^*) = \Big(V_1(\boldsymbol{\theta}^{*(1)}), V_2(\boldsymbol{\theta}^{*(2)}), V_3(\boldsymbol{\theta}^{*(3)}), V_4(\boldsymbol{\theta}^{*(4)})\Big)$, where $\boldsymbol{\theta}^{*(\ell)} = (\beta_0^{*(\ell)}, \beta_1^{*(\ell)}, \beta_{11}^{*}{}^{(\ell)}, \tau^{*(\ell)})^T$ lies between $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}^0$ for $\ell = 1, 2, 3, 4$.

With some simple algebra, we have $V(\boldsymbol{\theta}) = P\{H^T(X;\boldsymbol{\theta})H(X;\boldsymbol{\theta})\}$, where $H(X;\boldsymbol{\theta}) = \{1, X, (X-\tau)^+, -\beta_{11}I(X > \tau_1)\}$. We can define $V(\boldsymbol{\theta}^0)$ accordingly. Because $\widehat{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}^0$ in probability when $n \to \infty$, and $\boldsymbol{\theta}^{*(l)}$ lies between $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}^0$, it follows that $\boldsymbol{\theta}^{*(l)}$ converges to $\boldsymbol{\theta}^0$ in probability when $n \to \infty$, for $\ell = 1, 2, 3, 4$. Also, as each component in matrix $V(\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$, by continuous mapping theorem, $V(\boldsymbol{\theta}^*)$ converges to $V(\boldsymbol{\theta}^0)$ as $n \to \infty$. And for any vector $\boldsymbol{a} = (a_1, a_2, a_3, a_4) \neq (0, 0, 0, 0)$, $\boldsymbol{a}V(\boldsymbol{\theta}^0)\boldsymbol{a}^T = P\{(a_1 + a_2X + a_3X_{\tau^0} - a_4\beta_{11}^0 I_{\tau^0})^2\} > 0$. Therefore, $V(\boldsymbol{\theta}^0)$ is positive definite and hence non-singular. Because $P(Q(\boldsymbol{\theta}^0)) = 0 = \mathbb{P}_n(Q(\widehat{\boldsymbol{\theta}}_n))$, equation (A.20) becomes

$$\mathbb{P}_n(Q(\widehat{\boldsymbol{\theta}}_n)) - P(Q(\widehat{\boldsymbol{\theta}}_n)) = -2V(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0).$$

Denote $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$, and the above equation can be written as

$$\mathbb{G}_n(Q(\widehat{\boldsymbol{\theta}}_n)) = \mathbb{G}_n\{Q(\widehat{\boldsymbol{\theta}}_n) - Q(\boldsymbol{\theta}^{\mathbf{0}})\} + \mathbb{G}_n\{Q(\boldsymbol{\theta}^{\mathbf{0}})\} = -2\sqrt{n}V(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0). \quad \text{(A.21)}$$

Next, we will prove $\mathbb{G}_n\{Q(\widehat{\boldsymbol{\theta}}_n) - Q(\boldsymbol{\theta}^{\mathbf{0}})\} = o_p(1)$. Because $\boldsymbol{\Theta}$ is compact, β_0, $\beta_1 X$, $\beta_{11}(X-\tau)^+$, X, $(X-\tau)^+$ and $\beta_{11}I(X > \tau)$ are all uniformly, bounded monotones functions of $\boldsymbol{\theta}$ on the real line. According to Theorem 2.7.5 in *van der Vaart and Wellner* (1996), the class of all uniformly bounded, monotone functions on the real line is Donsker. By Theorem 2.10.6 in *van der Vaart and Wellner* (1996), addition and multiplication of uniformly bounded, monotone functions preserve the Donsker property. Therefore, $Q_\ell(\boldsymbol{\theta})$ $(\ell = 1, 2, 3, 4)$ is a Donsker class, which implies asymptotical equicontinuity. Because $\widehat{\boldsymbol{\theta}}_n$ converges in probability to $\boldsymbol{\theta}^0$, and $Q_\ell(\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$, by continuous mapping theorem $Q_\ell(\widehat{\boldsymbol{\theta}}_n)$ converges in probability to $Q_\ell(\boldsymbol{\theta}^0)$. That is, $Q_\ell(\widehat{\boldsymbol{\theta}}_n) - Q_\ell(\boldsymbol{\theta}^0) = o_p(1)$, and $\{Q_\ell(\widehat{\boldsymbol{\theta}}_n) - Q_\ell(\boldsymbol{\theta}^0)\}^2 = o_p(1)$. Because $\boldsymbol{\Theta}$ is compact and the domain of X is bounded, $\{Q_\ell(\widehat{\boldsymbol{\theta}}_n) - Q_\ell(\boldsymbol{\theta}^0)\}^2$ is bounded. Then, by the dominated convergence theorem, $P\{Q_\ell(\widehat{\boldsymbol{\theta}}_n) - Q_\ell(\boldsymbol{\theta}^0)\}^2 = o_p(1)$. According to Lemma 19.24 in *van der Vaart* (2000), $\mathbb{G}_n\{Q_\ell(\widehat{\boldsymbol{\theta}}_n) - Q_\ell(\boldsymbol{\theta}^{\mathbf{0}})\} = o_p(1)$ for all $\ell = 1, 2, 3, 4$. That is $\mathbb{G}_n\{Q(\widehat{\boldsymbol{\theta}}_n) - Q(\boldsymbol{\theta}^{\mathbf{0}})\} = o_p(1)$. By equation (A.20), we have

$$2\sqrt{n}V(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0) = -\mathbb{G}_n(Q(\boldsymbol{\theta}^{\mathbf{0}})) + o_p(1).$$

By central limit theorem, $-\mathbb{G}_n(Q(\boldsymbol{\theta}^0))$ converges in distribution to a normal distribution $\mathcal{N}\Big(0, P\{Q(\boldsymbol{\theta}^0)Q^T(\boldsymbol{\theta}^0)\}\Big)$. With some algebra, we can show that

$$P\{Q(\boldsymbol{\theta}^0)Q^T(\boldsymbol{\theta}^0)\} = 4I(\boldsymbol{\theta}^0), \text{ where } I(\boldsymbol{\theta}^0) = P\{\sigma^2(X)H^T(X;\boldsymbol{\theta}^0)H(X;\boldsymbol{\theta}^0)\}.$$

As $-\mathbb{G}_n(Q(\boldsymbol{\theta}^0))$ converges in distribution to $\mathcal{N}(\mathbf{0}, 4I(\boldsymbol{\theta}^0))$, $V(\boldsymbol{\theta}^*)$ converges to $V(\boldsymbol{\theta}^0)$ in probability and $V(\boldsymbol{\theta}^0)$ is a non-singular matrix, by Slutsky's theorem, it follows

that $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0)$ converges in distribution to $\mathcal{N}(\mathbf{0}, V^{-1}(\boldsymbol{\theta}^0)I(\boldsymbol{\theta}^0)V^{-1}(\boldsymbol{\theta}^0))$. □

A.2 Chapter II: Algorithm

A.2.1 Proposed Algorithm For Single Knot Situation

In this section, we provide detailed steps for the proposed algorithm when there is a single knot to be estimated. Specifically, the model is written as $Y_i = \mu(\boldsymbol{X}_i^*; \boldsymbol{\theta}) + \epsilon_i = \beta_0 + \beta_1 X_i + \beta_{11}(X_i - \tau)^+ + \boldsymbol{\eta}^T \boldsymbol{Z}_i + \epsilon_i$. Denoting the initial value of τ by $\widehat{\tau}^{(0)}$, and the t-th ($t \geq 1$) iteration of the proposed algorithm is as follows

Step 1. Update estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ to obtain $\widehat{\boldsymbol{\beta}}^{(t-1)}$, $\widehat{\boldsymbol{\eta}}^{(t-1)}$. Specifically, treating $\widehat{\boldsymbol{\tau}}^{(t-1)}$ as fixed, fit the linear regression model $E(Y|X, Z) = \beta_0 + \beta_1 X + \beta_{11}(X - \widehat{\tau}^{(t-1)})^+ + \boldsymbol{\eta}^T \boldsymbol{Z}$ by the OLS method to obtain estimates $\widehat{\boldsymbol{\beta}}^{(t-1)}$, $\widehat{\boldsymbol{\eta}}^{(t-1)}$ and the predicted values $\widehat{\mu}_i^{(t-1)}$, $i = 1, \cdots, n$, from the fitted model.

Step 2. Update $\widehat{\tau}^{(t-1)}$ to obtain $\widehat{\tau}^{(t)}$ by the extended NR type procedure $\widehat{\tau}^{(t)} = \widehat{\tau}^{(t-1)} - \{J^{(t)}\}^{-1} U^{(t)}$, where $U^{(t)}$ and $J^{(t)}$ are defined as follows

$$\begin{aligned} U^{(t)} &= \frac{\widehat{\beta}_{11}^{(t-1)}}{n} \sum_{i=1}^{n} (Y_i - \widehat{\mu}_i^{(t-1)}) I(X_i > \widehat{\tau}^{(t-1)}), \\ J^{(t)} &= \frac{\{\widehat{\beta}_{11}^{(t-1)}\}^2}{n} \sum_{i=1}^{n} I(X_i > \widehat{\tau}^{(t-1)}). \end{aligned}$$

A.2.2 Brief Justification of the Gradient Descent Type Algorithm

In this section, we provide a brief justification for using the sum of squares $\mathbb{P}_n(M(\boldsymbol{\theta})) = \frac{1}{n}\sum_{i=1}^{n} \{Y_i - \mu(\boldsymbol{X}_i; \boldsymbol{\theta})\}^2$ as the objective function to be minimized in determining the step size in the gradient descent type algorithm. For simplicity, we consider a simple model $E(Y|X) = \beta_0 + \beta_1 X + \beta_{11}(X - \tau)^+$. The solution of $\mathbb{P}_n(Q(\boldsymbol{\theta})) = 0$ is denoted as $\widehat{\boldsymbol{\theta}}_n = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_{11}, \widehat{\tau})^T$.

When $\widehat{\tau} \neq X_i$ for all $i = 1, \ldots, n$, the first and second order derivatives of $M(\boldsymbol{\theta})$ exist at $\widehat{\boldsymbol{\theta}}_n$, since the derivative of $(X_i - \tau)^+$ with respect to τ does not exist only when $\tau = X_i$. In this case, a usual argument can show that $\widehat{\boldsymbol{\theta}}_n$ is a local minima of $\mathbb{P}_n(M(\boldsymbol{\theta}))$ because the first order derivative of $\mathbb{P}_n(M(\boldsymbol{\theta}))$ at $\widehat{\boldsymbol{\theta}}_n$ is zero and the second order derivative (Hessian) is positive definite.

Next, we consider $\widehat{\tau} = X_j$ for some $j \in \{1, \ldots, n\}$. Then we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} M_i(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1, i\neq j}^{n} M_i(\boldsymbol{\theta}) + \frac{1}{n} M_j(\boldsymbol{\theta}) \\
&= \frac{1}{n}[\sum_{i=1, i\neq j}^{n} M_i(\boldsymbol{\theta}) + \{Y_j - \beta_0 - \beta_1 X_j\}^2] + \frac{1}{n}[M_j(\boldsymbol{\theta}) - \{Y_j - \beta_0 - \beta_1 X_j\}^2] \\
&\triangleq \widetilde{M_1}(\boldsymbol{\theta}) + \widetilde{M_2}(\boldsymbol{\theta})
\end{aligned}$$

It is easy to check that $\widehat{\boldsymbol{\theta}}_n$ is a local minima of $\widetilde{M_1}(\boldsymbol{\theta})$, i.e., the first order derivative is zero as $\widehat{\boldsymbol{\theta}}_n$ is the solution to $\mathbb{P}_n(Q(\boldsymbol{\theta})) = 0$ and $\widehat{\tau} = X_j$, and the Hessian matrix is positive definite. The second term $\widetilde{M_2}(\boldsymbol{\theta})$ is a small term relative to the first term and converges to zero when n goes to infinity for any $\boldsymbol{\theta}$. Therefore, $\widehat{\boldsymbol{\theta}}_n$ is also approximately a local minima in this case. The discussion above can be extended to the situation where there exists multiple $j \in (1, \ldots, n)$ such that $X_j = \widehat{\tau}$.

We conducted some preliminary simulation studie for the two-knots setups using the gradient descent method, with the step size chosen by the exact line search method based on the sum of squares as the objective function. Results are shown in Table A.1 below. Based on our simulations, the gradient descent type algorithm leads to estimators with similar statistical performances as the proposed Newton-Raphson type algorithm.

Table A.1: Simulation results based on 1000 Monte Carlo data sets for $K = 2$.

Methods	n	Bias*	MCSD*	AVESE*	CP%	Bias*	MCSD*	AVESE*	CP%	CR%
		$\widehat{\tau}_1$				$\widehat{\tau}_2$				
					Setup 2.1					
proposed	**200**	-3.09	14.98	11.46	90.3	-2.75	20.05	11.38	91.3	100.0
descent		-4.92	14.62	11.41	88.5	-6.33	38.44	11.72	89.8	100.0
proposed	**500**	-0.56	7.99	7.06	91.0	-0.74	7.59	7.04	93.6	100.0
descent		-1.40	7.81	7.00	90.7	-0.87	7.42	7.05	94.4	100.0
proposed	**1000**	-0.36	5.12	4.94	93.3	-0.48	5.04	4.96	95.4	100.0
descent		-0.71	5.34	4.92	92.6	-0.74	5.11	4.96	95.1	100.0
proposed	**2500**	-0.23	3.03	3.11	95.5	0.01	3.17	3.11	93.5	100.0
descent		-0.69	3.14	3.11	95.2	-0.10	3.14	3.11	93.8	100.0
					Setup 2.2					
proposed	**200**	-1.87	7.49	6.16	89.8	-2.62	15.14	11.93	87.2	91.9
descent		1.54	7.05	5.90	89.0	8.66	16.87	12.47	86.1	100.0
proposed	**500**	-0.38	4.13	3.76	92.2	-0.90	8.30	7.48	92.1	99.0
descent		1.08	4.00	3.68	91.2	4.36	8.17	7.63	89.7	100.0
proposed	**1000**	-0.04	2.62	2.62	95.7	-0.27	5.49	5.24	92.6	99.9
descent		0.68	2.71	2.59	93.5	2.96	5.53	5.29	90.2	100.0
proposed	**2500**	-0.16	1.64	1.66	94.9	-0.29	3.42	3.30	93.7	100.0
descent		0.42	1.67	1.64	93.5	2.16	3.51	3.33	90.4	100.0
					Setup 2.3					
proposed	**200**	-0.91	12.57	11.46	91.8	1.36	12.53	11.38	91.7	94.4
descent		4.99	14.47	11.60	89.9	-3.55	21.79	11.50	89.8	100.0
proposed	**500**	-0.44	7.94	7.04	91.5	0.34	7.55	7.04	93.5	99.8
descent		1.76	8.33	7.05	90.3	-0.97	7.52	7.05	93.5	100.0
proposed	**1000**	-0.23	5.09	4.94	94.0	0.05	4.98	4.96	95.0	100.0
descent		0.78	5.34	4.94	92.9	-0.81	5.12	4.96	94.3	100.0
proposed	**2500**	-0.04	2.94	3.11	96.1	0.20	3.17	3.11	93.3	100.0
descent		0.68	3.11	3.12	94.1	-0.13	3.17	3.11	93.6	100.0

Note: "proposed" denotes the proposed Newton-Raphson type algorithm and "descent" denotes the gradient descent type algorithm, n denotes the sample size and * indicates value $\times 10^{-3}$.

A.3 Chapter III: Proofs

We only give detailed proofs for the proposition and two theorems in Chapter III based on the case of $g\{\mu(X;\boldsymbol{\theta})\} = \beta_0 + \beta_1 X + \beta_{11}(X-\tau)^+$. The proof for the case of multiple change-points is similar. For convenience and without loss of generality, we assume the domain of X is positive. The estimating function $Q(\boldsymbol{W};\boldsymbol{\theta})$ is

$$\begin{aligned} Q(\boldsymbol{W};\boldsymbol{\theta}) &= \frac{H^T(X;\boldsymbol{\theta})\{Y-\mu(X;\boldsymbol{\theta})\}}{v\{\mu(X;\boldsymbol{\theta})\}g'\{\mu(X;\boldsymbol{\theta})\}}, \text{ where} \\ H(X;\boldsymbol{\theta}) &= (1, X, (X-\tau)^+, -\beta_{11}I(X>\tau))^T. \end{aligned} \tag{A.22}$$

And we denote each row of $Q(\boldsymbol{W};\boldsymbol{\theta})$ as $Q_\ell(\boldsymbol{W};\boldsymbol{\theta})$, for $\ell = 1,2,3,4$.

A.3.1 Proof of Proposition 2

Proof. Denote $\widehat{\tau}^{(0)}$ as the chosen initial value for τ. Denote $\widehat{\beta}_0^{(0)}$, $\widehat{\beta}_1^{(0)}$ and $\widehat{\beta}_{11}^{(0)}$ as the corresponding estimates from Step 1, when τ is fixed as $\widehat{\tau}^{(0)}$, and $\widehat{\boldsymbol{\theta}}^{(0)} = (\widehat{\beta}_0^{(0)}, \widehat{\beta}_1^{(0)}, \widehat{\beta}_{11}^{(0)}, \widehat{\tau}^{(0)})^T$. For brevity, we define $F(\boldsymbol{\theta}) \equiv -\mathbb{P}_n(Q(\boldsymbol{\theta}))$, with each row of $F(\boldsymbol{\theta})$ denoted as $F_\ell(\boldsymbol{\theta})$. Thus the estimating equation $\mathbb{P}_n(Q(\boldsymbol{\theta})) = 0$ becomes $F(\boldsymbol{\theta}) = 0$. We denote a 4×4 symmetric matrix $G(\boldsymbol{\theta}) = \mathbb{P}_n\Big[\frac{H^T(X;\boldsymbol{\theta})H(X;\boldsymbol{\theta})}{v\{\mu(X;\boldsymbol{\theta})\}g'\{\mu(X;\boldsymbol{\theta})\}^2}\Big]$. We also define two sub-matrices of matrices G, G_β and $G_{\beta\tau}$, as follows

$$G_\beta(\boldsymbol{\theta}) = \begin{pmatrix} G_{11}(\boldsymbol{\theta}) & G_{12}(\boldsymbol{\theta}) & G_{13}(\boldsymbol{\theta}) \\ G_{12}(\boldsymbol{\theta}) & G_{22}(\boldsymbol{\theta}) & G_{23}(\boldsymbol{\theta}) \\ G_{13}(\boldsymbol{\theta}) & G_{23}(\boldsymbol{\theta}) & G_{33}(\boldsymbol{\theta}) \end{pmatrix} \text{ and } G_{\beta\tau}(\boldsymbol{\theta}) = \begin{pmatrix} G_{14}(\boldsymbol{\theta}) \\ G_{24}(\boldsymbol{\theta}) \\ G_{34}(\boldsymbol{\theta}) \end{pmatrix}.$$

Because $\boldsymbol{\theta}$ belongs to a compact set $\boldsymbol{\Theta}$ and X has a bounded domain, all components in matrix $G(\boldsymbol{\theta})$ can be bounded by a finite constant. By Theorem 3 in *Chaney* (1990), every piecewise differentiable function is locally Lipschitz continuous. And by Corollary 4.1.1 in *Scholtes* (2012), every piecewise differentiable function is semismooth.

As each row of $F(\boldsymbol{\theta})$ is a piecewise differentiable function, $F(\boldsymbol{\theta})$ is locally Lipschitz continuous and semismooth. It is easy to check that $G(\boldsymbol{\theta})$ belongs to $\partial F(\boldsymbol{\theta})$, i.e. the generalized gradient. Therefore by Theorem 2.3 in *Qi and Sun* (1993), when the initial value $\widehat{\boldsymbol{\theta}}^{(0)}$ is chosen close enough to the $\widehat{\boldsymbol{\theta}}_n$, i.e., $||\widehat{\boldsymbol{\theta}}^{(0)} - \widehat{\boldsymbol{\theta}}_n|| = o(1)$, we have

$$\left\| F(\widehat{\boldsymbol{\theta}}_n) - F(\widehat{\boldsymbol{\theta}}^{(0)}) - G(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^{(0)}) \right\| = o(||\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^{(0)}||), \tag{A.23}$$

By definition, $F(\widehat{\boldsymbol{\theta}}_n) = 0$, and according to the Step 1, we have $F_1(\widehat{\boldsymbol{\theta}}^{(0)}) = F_2(\widehat{\boldsymbol{\theta}}^{(0)}) = F_3(\widehat{\boldsymbol{\theta}}^{(0)}) = 0$. Therefore, by the first three rows of equation (A.23), we have

$$B = -\{G_\beta(\widehat{\boldsymbol{\theta}}^{(0)})\}^{-1} G_{\beta\tau}(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\tau} - \widehat{\tau}^{(0)}) + o(||\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^{(0)}||), \tag{A.24}$$

where $B = (\widehat{\beta}_0 - \widehat{\beta}_0^{(0)}, \widehat{\beta}_1 - \widehat{\beta}_1^{(0)}, \widehat{\beta}_{11} - \widehat{\beta}_{11}^{(0)})^T$. According to the last row of the equation (A.23), we have

$$F_4(\widehat{\boldsymbol{\theta}}) - F_4(\widehat{\boldsymbol{\theta}}^{(0)}) - G_{44}(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\tau} - \widehat{\tau}^{(0)}) = G_{\beta\tau}^T(\widehat{\boldsymbol{\theta}}^{(0)})B + o(||\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^{(0)}||).$$

Applying formula (A.24) to the above equation, we have

$$\begin{aligned} & F_4(\widehat{\boldsymbol{\theta}}) - F_4(\widehat{\boldsymbol{\theta}}^{(0)}) - G_{44}(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\tau} - \widehat{\tau}^{(0)}) \\ = \; & -G_{\beta\tau}^T(\widehat{\boldsymbol{\theta}}^{(0)})\{G_\beta(\widehat{\boldsymbol{\theta}}^{(0)})\}^{-1} G_{\beta\tau}(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\tau} - \widehat{\tau}^{(0)}) + o(||\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^{(0)}||). \end{aligned} \tag{A.25}$$

According to Step 2 in the proposed algorithm, the estimate of the change-point τ is updated via $\widehat{\tau}^{(1)} = \widehat{\tau}^{(0)} - G_{44}^{-1}(\widehat{\boldsymbol{\theta}}^{(0)})F_4(\widehat{\boldsymbol{\theta}}^{(0)})$. Then by the equation (A.25), we have

$$\begin{aligned} \widehat{\tau}^{(1)} - \widehat{\tau} &= \widehat{\tau}^{(0)} - \widehat{\tau} - G_{44}^{-1}(\widehat{\boldsymbol{\theta}}^{(0)})F_3(\widehat{\boldsymbol{\theta}}^{(0)}) \\ &= G_{44}^{-1}(\widehat{\boldsymbol{\theta}}^{(0)})\{F_3(\widehat{\boldsymbol{\theta}}_n) - F_3(\widehat{\boldsymbol{\theta}}^{(0)}) - G_{44}(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\tau} - \widehat{\tau}^{(0)})\} \\ &= G^*(\widehat{\boldsymbol{\theta}}^{(0)})(\widehat{\tau}^{(0)} - \widehat{\tau}) + o(||\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^{(0)}||), \end{aligned} \tag{A.26}$$

where $G^*(\boldsymbol{\theta}) = G_{44}^{-1}(\boldsymbol{\theta})G_{\beta\tau}^T(\boldsymbol{\theta})\{G_\beta(\boldsymbol{\theta})\}^{-1}G_{\beta\tau}(\boldsymbol{\theta})$. With some algebra, we can simplify $G^*(\boldsymbol{\theta})$ by determinant of matrices $G(\boldsymbol{\theta})$ and $G_\beta(\boldsymbol{\theta})$, i.e. $G^*(\boldsymbol{\theta}) = 1 - \frac{\det\{G(\boldsymbol{\theta})\}}{\det\{G_\beta(\boldsymbol{\theta})\}G_{44}(\boldsymbol{\theta})}$. Because $G(\boldsymbol{\theta})$, $G_\beta(\boldsymbol{\theta})$ are positive-definite matrices and $G_{44}(\boldsymbol{\theta})$ is positive, according to Fischer's inequality, we have $\det\{G(\boldsymbol{\theta})\} < \det\{G_\beta(\boldsymbol{\theta})\}G_{44}(\boldsymbol{\theta})$. Therefore, $0 < G^*(\boldsymbol{\theta}) < 1$. According to formula (A.24), we can see that $o(||\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^{(0)}||)$ implies $o(||\widehat{\tau} - \widehat{\tau}^{(0)}||)$. Thus, equation (A.26) implies

$$||\widehat{\tau}^{(1)} - \widehat{\tau}||= G^*(\widehat{\boldsymbol{\theta}}^{(0)})||\widehat{\tau}^{(0)} - \widehat{\tau}||+o(||\widehat{\tau}^{(0)} - \widehat{\tau}||).$$

Therefore, when choosing an arbitrary small constant $0 < m < 1 - G^*(\widehat{\boldsymbol{\theta}}^{(0)})$, there exists an $r_1 > 0$ such that when $\widehat{\tau}^{(0)} \in B(\widehat{\tau}, r_1)$, we have $o(||\widehat{\tau}^{(0)} - \widehat{\tau}||) < m||\widehat{\tau}^{(0)} - \widehat{\tau}||$ and thus $||\widehat{\tau}^{(1)} - \widehat{\tau}|| < ||\widehat{\tau}^{(0)} - \widehat{\tau}||$. The induction for any step t ($t \geq 1$) is similar. Therefore, $\widehat{\tau}^{(t)}$ converges to $\widehat{\tau}$ when iteration $t \to \infty$.

Next, we will show that $\widehat{\boldsymbol{\theta}}^{(t)}$ converges to $\widehat{\boldsymbol{\theta}}_n$ when t goes to infinity. By similar steps leading to the formula (A.24), for any iteration t, we have

$$B^{(t)} = -\{G_\beta(\widehat{\boldsymbol{\theta}}^{(t)})\}^{-1}G_{\beta\tau}(\widehat{\boldsymbol{\theta}}^{(t)})(\widehat{\tau} - \widehat{\tau}^{(t)}) + o(||\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^{(t)}||), \tag{A.27}$$

where $B^{(t)} = (\widehat{\beta}_0 - \widehat{\beta}_0^{(t)}, \widehat{\beta}_1 - \widehat{\beta}_1^{(t)}, \widehat{\beta}_{11} - \widehat{\beta}_{11}^{(t)})^T$. As it is clear that each element of $G_\beta(\widehat{\boldsymbol{\theta}}^{(t)})$ and $G_{\beta\tau}(\widehat{\boldsymbol{\theta}}^{(t)})$ is finite, all elements of $|\{G_\beta(\widehat{\boldsymbol{\theta}}^{(t)})\}^{-1}G_{\beta\tau}(\widehat{\boldsymbol{\theta}}^{(t)})|$ can be bounded by some positive constant K. Because $\lim_{t\to\infty}\widehat{\tau}^{(t)} = \widehat{\tau}$, for any $\epsilon > 0$, there exists a positive constant T_1, when $t > T_1$, we have $|\widehat{\tau} - \widehat{\tau}^{(t)}| < \epsilon/(2K)$. And there exists a positive constant T_2, when $t > T_2$, we have $|o(||\widehat{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}^{(t)}||)| < \epsilon/2$. Hence, when $t > \max(T_1, T_2)$, according to the first row of the equation (A.27), we have $|\widehat{\beta}_0 - \widehat{\beta}_0^{(t)}| < K|\widehat{\tau} - \widehat{\tau}^{(t)}| + \epsilon/2 < \epsilon$, which implies that $\widehat{\beta}_0^{(t)}$ converges to $\widehat{\beta}_0$ when $t \to \infty$. Similarly, from the second and third row of equation (A.27), we can prove that $\widehat{\beta}_1^{(t)}$ converges to $\widehat{\beta}_1$ and $\widehat{\beta}_{11}^{(t)}$ converges to $\widehat{\beta}_{11}$ when $t \to \infty$. Thus, the proposed algorithm converges

locally and the converged result is the solution to $\mathbb{P}_n(Q(\boldsymbol{\theta})) = 0$. □

A.3.2 Proof of Theorem 3

Proof. First, we will prove that $\boldsymbol{\theta}^0$ is a unique solution of $P(Q(\boldsymbol{W};\boldsymbol{\theta})) = 0$. As $P(Q(\boldsymbol{W};\boldsymbol{\theta}))$ is a continuous function of $\boldsymbol{\theta}$ in the compact set and $\boldsymbol{\theta}^0$ is the unique zero of $P(Q(\boldsymbol{W};\boldsymbol{\theta})) = 0$, then this unique solution $\boldsymbol{\theta}^0$ is well-separated. Next, we will show $\sup_{\boldsymbol{\theta}\in\Theta}|\mathbb{P}_n(Q(\boldsymbol{W};\boldsymbol{\theta})) - P(Q(\boldsymbol{W};\boldsymbol{\theta}))| \xrightarrow{a.s} 0$. According to Theorem 5.9 in *van der Vaart* (2000), $\widehat{\boldsymbol{\theta}}_n$ converges in probability to $\boldsymbol{\theta}^0$.

According to the law of total expectation, we have

$$P(Q(\boldsymbol{W};\boldsymbol{\theta})) = E_X[E\{Q(\boldsymbol{W};\boldsymbol{\theta})|X\}] = E_X\Big[\frac{H^T(X;\boldsymbol{\theta})\{\mu(X;\boldsymbol{\theta}^0) - \mu(X;\boldsymbol{\theta})\}}{v\{\mu(X;\boldsymbol{\theta})\}g'\{\mu(X;\boldsymbol{\theta})\}}\Big].$$

Thus, it is straightforward to see that $\boldsymbol{\theta}^0$ is the solution to $P(Q(\boldsymbol{W};\boldsymbol{\theta})) = 0$. Next, we will show the uniqueness of $\boldsymbol{\theta}^0$. Although $Q(\boldsymbol{W};\boldsymbol{\theta})$ is not first-order differentiable with respect to $\boldsymbol{\theta}$, $P(Q(\boldsymbol{W};\boldsymbol{\theta}))$ is first-order differentiable with respect to $\boldsymbol{\theta}$ because of the integral over X. And it is easy to calculate the first-order derivative of $P(Q(\boldsymbol{W};\boldsymbol{\theta}))$ with respect to $\boldsymbol{\theta}$, denoted as $R(\boldsymbol{W};\boldsymbol{\theta})$, i.e.,

$$R(\boldsymbol{W};\boldsymbol{\theta}) \equiv \frac{\partial P(Q(\boldsymbol{W};\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}^T} = -P\Big[\frac{H^T(X;\boldsymbol{\theta})H(X;\boldsymbol{\theta})}{v\{\mu(X;\boldsymbol{\theta})\}g'\{\mu(X;\boldsymbol{\theta})\}^2}\Big], \tag{A.28}$$

where $H(X;\boldsymbol{\theta})$ is defined as the same in the formula (A.22). For any vector $\boldsymbol{a} = (a_1, a_2, a_3, a_4) \neq (0,0,0,0)$, we can obtain

$$\boldsymbol{a}R(\boldsymbol{W};\boldsymbol{\theta})\boldsymbol{a}^T = -P\Big[\frac{\{a_1 + a_2X + a_3(X-\tau)^+ - a_4\beta_{11}I(X>\tau)\}^2}{v\{\mu(X;\boldsymbol{\theta})\}g'\{\mu(X;\boldsymbol{\theta})\}^2}\Big] < 0,$$

which indicates $R(\boldsymbol{W};\boldsymbol{\theta})$ is negative definite. Therefore, $P(Q(\boldsymbol{W};\boldsymbol{\theta})) = 0$ has a unique solution at $\boldsymbol{\theta}^0$.

Next, we will show that $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|\mathbb{P}_n(Q(\boldsymbol{W};\boldsymbol{\theta}))-P(Q(\boldsymbol{W};\boldsymbol{\theta}))|\xrightarrow{a.s.}0$. Define measurable class of measurable functions $\mathcal{F}=\{Q(\boldsymbol{\theta}):\boldsymbol{\theta}\in\boldsymbol{\Theta}\}$, and $\mathcal{F}_\ell=\{Q_\ell(\boldsymbol{\theta}):\boldsymbol{\theta}\in\boldsymbol{\Theta}\}$, for $\ell=1,2,3,4$. As $\boldsymbol{\Theta}$ is compact, we can assume $|\beta_0|\leq w_0$, $|\beta_1|\leq w_1$ and $|\beta_{11}|\leq w_2$, where w_j are finite constants ($j=0,1,2$). Then, $|\xi|\leq w_0+(w_1+w_2)C_2$. Because the link function is continuous, then the inverse of the link function is also continuous. Therefore $|\mu(X;\boldsymbol{\theta})|=|g^{-1}(\xi)|$ is bounded by some finite constant, and we denote this finite constant as W_1. Also, as $v(.)$ and $g'(.)$ are assumed to be continuous functions, $|v\{\mu(X;\boldsymbol{\theta})\}|$ and $|g'\{\mu(X;\boldsymbol{\theta})\}|$ are also bounded by some finite constants, which are denoted as W_2 and W_3. Then, we can define integrable envelope functions F_ℓ for $\mathcal{F}_\ell$ ($l=\ell,2,3,4$),

$$F_1=\frac{|Y|+W_1}{W_2+W_3};F_2=\frac{C_2(|Y|+W)_1}{W_2+W_3};F_3=\frac{C_2(|Y|+W_1)}{W_2+W_3};F_4=\frac{w_2(|Y|+W_1)}{W_2+W_3}.$$

For each $Q_\ell(\boldsymbol{\theta})$ ($\ell=1,2,3,4$), it is obvious that the map $\boldsymbol{\theta}\mapsto Q_\ell(\boldsymbol{\theta})$ is continuous for every x. Thus, by Example 19.8 in *van der Vaart* (2000), the L_1-bracketing numbers of $\mathcal{F}_\ell$ are finite and hence $\mathcal{F}_\ell$ is Glivenko-Cantelli for $\ell=1,2,3,4$. This implies that $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|\mathbb{P}_n(Q_\ell(\boldsymbol{\theta}))-P(Q_\ell(\boldsymbol{\theta}))|\xrightarrow{a.s.}0$. And then $\mathcal{F}$ belongs to a Glivenko-Cantelli class, and $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|\mathbb{P}_n(Q(\boldsymbol{\theta}))-P(Q(\boldsymbol{\theta}))|\xrightarrow{a.s.}0$. □

A.3.3 Proof of Theorem 4

Proof. Because $P(Q(\boldsymbol{\theta}))$ is first-order differentiable with respect to $\boldsymbol{\theta}$, by Talyor expansion of $P(Q(\hat{\boldsymbol{\theta}}_n))$ around $\boldsymbol{\theta}^0$, we have the following equation

$$P(Q(\widehat{\boldsymbol{\theta}}_n))-P(Q(\boldsymbol{\theta}^0)) \;=\; R(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^0), \tag{A.29}$$

where $R(\boldsymbol{\theta})$ is a negative definite and non-singular matrix defined in (A.28), $\boldsymbol{\theta}^*=(\boldsymbol{\theta}^{*(1)},\ldots,\boldsymbol{\theta}^{*(4)})$ and $\boldsymbol{\theta}^{*(\ell)}=(\beta_0^{*(\ell)},\beta_1^{*(\ell)},\beta_{11}^{*(\ell)},\tau^{*(\ell)})^T$ lies between $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}^0$ for $\ell=1,2,3,4$. As $\widehat{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}^0$ in probability and $\boldsymbol{\theta}^{*(\ell)}$ lies between $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}^0$, $\boldsymbol{\theta}^{*(\ell)}$

converges to $\boldsymbol{\theta}^0$ in probability, for $\ell = 1, 2, 3, 4$. Also, as each component in the matrix $R(\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$, by continuous mapping theorem, $R(\boldsymbol{\theta}^*)$ converges to $R(\boldsymbol{\theta}^0)$ in probability. As $P(Q(\boldsymbol{\theta}^0)) = 0 = \mathbb{P}_n(Q(\widehat{\boldsymbol{\theta}}_n))$, replacing $P(Q(\boldsymbol{\theta}^0))$ in the equation (A.29) by $\mathbb{P}_n(Q(\widehat{\boldsymbol{\theta}}_n))$, we have $\mathbb{P}_n(Q(\widehat{\boldsymbol{\theta}}_n)) - P(Q(\widehat{\boldsymbol{\theta}}_n)) = -R(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0)$, which can be reorganized into the following equation

$$\mathbb{G}_n(Q(\widehat{\boldsymbol{\theta}}_n)) = \mathbb{G}_n\{Q(\widehat{\boldsymbol{\theta}}_n) - Q(\boldsymbol{\theta}^{\mathbf{0}})\} + \mathbb{G}_n\{Q(\boldsymbol{\theta}^{\mathbf{0}})\} = -\sqrt{n}R(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0). \quad \text{(A.30)}$$

Next, we will prove $\mathbb{G}_n\{Q(\widehat{\boldsymbol{\theta}}_n) - Q(\boldsymbol{\theta}^{\mathbf{0}})\} = o_p(1)$. Because $\boldsymbol{\Theta}$ is compact, β_0, $\beta_1 X$, $\beta_{11}(X-\tau)^+$, X, $(X-\tau)^+$ and $\beta_{11}I(X > \tau)$ are all uniformly bounded monotones functions of $\boldsymbol{\theta}$ on the real line. According to Theorem 2.7.5 in *van der Vaart and Wellner* (1996), the class of all uniformly bounded, monotone functions on the real line is Donsker. By Theorem 2.10.6 in *van der Vaart and Wellner* (1996), addition/multiplication of uniformly bounded monotone functions preserves the property of Donsker. Then, $\mathcal{H} = \{\xi = \xi(X;\boldsymbol{\theta}) = \beta_0 + \beta_1 X + \beta_{11}(X-\tau)^+ : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is a Donsker class. Because $\boldsymbol{\Theta}$ is compact, we can assume $|\beta_0| \leq w_0$, $|\beta_1| \leq w_1$ and $|\beta_{11}| \leq w_2$, where w_j are finite constant $(j = 0, 1, 2)$. Then, $|\xi| \leq w_0 + (w_1 + w_2)C_2$. Thus $H = w_0 + (w_1 + w_2)C_2$ is an integrable envelop function of $\mathcal{H}$. Because the link function $g(.)$ is a continuous first-order differentiable function and ξ is bounded, $g^{-1}(\xi)$ is also a first-order differentiable function. As $g^{-1}(\xi)$ is a first-order differentiable function with compact domain, $g^{-1}(\xi)$ is a Lipschitz function. According to Example 19.20 in *van der Vaart* (2000), $\mu(X;\boldsymbol{\theta}) = g^{-1}(\xi)$ belongs to a Donsker class. Because Y is free of $\boldsymbol{\theta}$, $\mathcal{S} = \{Y - \mu(X;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is also a Donsker class. And it is easy to check that $Y - \mu(X;\boldsymbol{\theta})$ is uniformly bounded. Also, because $v(.)$ and $g'(.)$ are first-order differentiable functions with compact domain, $v(\mu)$ and $g'(\mu)$ are also Lipschitz functions and hence belong to a Donsker class. Because $v(\mu)$ and $g'(\mu)$ are uniformly bounded away from 0, $1/v(\mu)$ and $1/g'(\mu)$ belong to a Donsker

class from Example 19.20 in *van der Vaart* (2000). Also, it is simple to show that X, $(X-\tau)^+$ and $\beta_{11}I(X>\tau)$ are all uniformly bounded and belong to a Donsker class. According to Example 19.20 in *van der Vaart* (2000), product of uniformly bounded Donsker classes preserve to be Donsker. Therefore, for $\ell = 1, 2, 3, 4$, $Q_\ell(\boldsymbol{\theta})$ belongs to a Donsker class, which implies asymptotical equicontinuity. Because $\widehat{\boldsymbol{\theta}}_n$ converges in probability to $\boldsymbol{\theta}^0$, and $Q_\ell(\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$, by continous mapping theorem $Q_\ell(\widehat{\boldsymbol{\theta}}_n)$ converges in probability to $Q_\ell(\boldsymbol{\theta}^0)$. That is, $Q_\ell(\widehat{\boldsymbol{\theta}}_n) - Q_\ell(\boldsymbol{\theta}^0) = o_p(1)$, and thus $\{Q_\ell(\widehat{\boldsymbol{\theta}}_n) - Q_\ell(\boldsymbol{\theta}^0)\}^2 = o_p(1)$. Because each row of $Q_\ell(\boldsymbol{\theta}^0)$ is uniformly bounded, by dominated convergence theorem, $P\{Q_\ell(\widehat{\boldsymbol{\theta}}_n) - Q_\ell(\boldsymbol{\theta}^0)\}^2 = o_p(1)$, for $\ell = 1, 2, 3, 4$. According to Lemma 19.24 in *van der Vaart* (2000), $\mathbb{G}_n\{Q_\ell(\widehat{\boldsymbol{\theta}}_n) - Q_\ell(\boldsymbol{\theta}^{\mathbf{0}})\} = o_p(1)$ for all $\ell = 1, 2, 3, 4$, which implies $\mathbb{G}_n\{Q(\widehat{\boldsymbol{\theta}}_n) - Q(\boldsymbol{\theta}^{\mathbf{0}})\} = o_p(1)$.

According to central limit theorem, $\mathbb{G}_n(Q(\boldsymbol{\theta}^0))$ converges in distribution to $N(0, V_2(\boldsymbol{\theta}^0))$, where $V_2(\boldsymbol{\theta}^0) = P\{Q(\boldsymbol{\theta}^0)Q^T(\boldsymbol{\theta}^0)\} = P\Big[\frac{V(Y|X;\boldsymbol{\theta}^0)H^T(X;\boldsymbol{\theta}^0)H(X;\boldsymbol{\theta}^0)}{v\{\mu(\boldsymbol{X}_i^*;\widehat{\boldsymbol{\theta}}_n)\}^2 g'\{\mu(X;\boldsymbol{\theta}^0)\}^2}\Big]$. The last equation can be derived through some algebra. Also, we denote $V_1(\boldsymbol{\theta}) = -R(\boldsymbol{\theta})$. Then, according to the equation (A.30), we have $\sqrt{n}V_1(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0) = \mathbb{G}_n(Q(\boldsymbol{\theta}^{\mathbf{0}})) + o_p(1)$. As $\mathbb{G}_n(Q(\boldsymbol{\theta}^0))$ converges in distribution to $\mathcal{N}(0, I_V(\boldsymbol{\theta}^0))$, $R(\boldsymbol{\theta}^*)$ converges to $R(\boldsymbol{\theta}^0)$ in probability and $R(\boldsymbol{\theta}^0)$ is a non-singular matrix, by Slutsky's theorem $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0)$ converges to $\mathcal{N}\Big(0, \phi V_1^{-1}(\boldsymbol{\theta}^0)V_2(\boldsymbol{\theta}^0)V_1^{-1}(\boldsymbol{\theta}^0)\Big)$ in distribution. □

A.4 Chapter IV: Proofs

We sketch the proofs of the main results below. For simplicity in notations, we only consider a simple case of the constrained broken-stick model (4.3), i.e. $g\{\mu(\boldsymbol{\theta})\} = \beta_0 + \beta_1(X-\tau)_-$. For the more general model (4.4), the proof is similar but more involved in terms of algebras. We denote $\boldsymbol{W} = (Y, X)$ and $Q(\boldsymbol{W};\boldsymbol{\theta}) = \frac{H^T(\boldsymbol{\theta})\{Y-\mu(\boldsymbol{\theta})\}}{v\{\mu(\boldsymbol{\theta})\}g'\{\mu(\boldsymbol{\theta})\}}$, where $H(\boldsymbol{\theta}) = \{1, (X-\tau)_-, -\beta_1 I(X<\tau)\}$. Thus, the proposed estimating equation (4.5) can be reorganized as $\frac{1}{n}\sum_{i=1}^n Q(\boldsymbol{W}_i;\boldsymbol{\theta}) = 0$. We denote each row of $Q(\boldsymbol{W};\boldsymbol{\theta})$

as $Q_\ell(\boldsymbol{W};\boldsymbol{\theta})$, for $\ell = 1, 2, 3$.

A.4.1 Proof of Result 1

As $\widehat{\boldsymbol{\theta}}_n$ is the root of the proposed estimating equation $\frac{1}{n}\sum_{i=1}^{n} Q(\boldsymbol{W}_i;\boldsymbol{\theta}) = 0$, $\widehat{\boldsymbol{\theta}}_n$ is a Z-estimator. Therefore, under regularity conditions, $\widehat{\boldsymbol{\theta}}_n$ converges to the unique solution of $E_{\boldsymbol{W}}(Q(\boldsymbol{W};\boldsymbol{\theta})) = 0$, which is the limit of the estimating equation $\frac{1}{n}\sum_{i=1}^{n} Q(\boldsymbol{W}_i;\boldsymbol{\theta}) = 0$ when n goes to infinity. To prove consistency, i.e. $\widehat{\boldsymbol{\theta}}_n$ converges in probability to the true value $\boldsymbol{\theta}^0$, we will first show that $\boldsymbol{\theta}^0$ is the unique, well-separated solution to $E_{\boldsymbol{W}}(Q(\boldsymbol{W};\boldsymbol{\theta})) = 0$. To guarantee that $\widehat{\boldsymbol{\theta}}_n$ will be close enough to $\boldsymbol{\theta}^0$ with increasing n, we will next show that $Q(\boldsymbol{\theta})$ belongs to a Glivenko-Cantelli class, i.e. $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\frac{1}{n}\sum_{i=1}^{n} Q(\boldsymbol{W}_i;\boldsymbol{\theta}) - E_{\boldsymbol{W}}(Q(\boldsymbol{W};\boldsymbol{\theta}))\right| \xrightarrow{a.s} 0$.

According to the law of total expectation, $E_{\boldsymbol{W}}(Q(\boldsymbol{W};\boldsymbol{\theta})) = E_X[E\{Q(\boldsymbol{W};\boldsymbol{\theta})|X\}] = E_X\left[\frac{H^T(\boldsymbol{\theta})\{\mu(\boldsymbol{\theta}^0)-\mu(\boldsymbol{\theta})\}}{v\{\mu(\boldsymbol{\theta})\}g'\{\mu(\boldsymbol{\theta})\}}\right]$. Therefore, $\boldsymbol{\theta}^0$ is a solution of $E_{\boldsymbol{W}}(Q(\boldsymbol{W};\boldsymbol{\theta})) = 0$. To show uniqueness, we consider the first-order derivative of $E_{\boldsymbol{W}}(Q(\boldsymbol{W};\boldsymbol{\theta}))$ with respect to $\boldsymbol{\theta}$, denoted as $R(\boldsymbol{\theta})$. Specially, we have

$$R(\boldsymbol{\theta}) \equiv \partial E_{\boldsymbol{W}}(Q(\boldsymbol{W};\boldsymbol{\theta}))/\partial\boldsymbol{\theta}^T = -E_{\boldsymbol{W}}\Big[\frac{H^T(\boldsymbol{\theta})H(\boldsymbol{\theta})}{v\{\mu(\boldsymbol{\theta})\}g'\{\mu(\boldsymbol{\theta})\}^2}\Big]. \tag{A.31}$$

We can check that $R(\boldsymbol{\theta})$ is negative definite because, for any vector $\boldsymbol{a} = (a_1, a_2, a_3) \neq (0, 0, 0)$,

$$\boldsymbol{a}R(\boldsymbol{\theta})\boldsymbol{a}^T = -E\Big[\frac{\{a_1 + a_2(X-\tau)_- - a_3\beta_1 I(X<\tau)\}^2}{v\{\mu(\boldsymbol{\theta})\}g'\{\mu(\boldsymbol{\theta})\}^2}\Big] < 0.$$

Thus, $E_{\boldsymbol{W}}(Q(\boldsymbol{W};\boldsymbol{\theta})) = 0$ has a unique solution. As $E_{\boldsymbol{W}}(Q(\boldsymbol{W};\boldsymbol{\theta}))$ is a continuous function of $\boldsymbol{\theta}$ in the compact set $\boldsymbol{\Theta}$ and $\boldsymbol{\theta}^0$ is the unique solution of $E_{\boldsymbol{W}}(Q(\boldsymbol{W};\boldsymbol{\theta})) = 0$, this unique solution $\boldsymbol{\theta}^0$ is well-separated *van der Vaart* (2000). Following similar steps in Appendix A.3.2, we can check that $Q(\boldsymbol{W};\boldsymbol{\theta})$ belongs to a Glivenko-Cantelli class.

We have shown the solution to the proposed estimating equation is consistent. Next we show that when the proposed two-step modified NR algorithm converges, the converged result is the solution to $\frac{1}{n}\sum_{i=1}^{n} Q(\boldsymbol{W}_i;\boldsymbol{\theta}) = 0$. We denote the converged value of the proposed two-step modified NR algorithm as $\widetilde{\boldsymbol{\theta}}_n = (\widetilde{\beta}_0, \widetilde{\beta}_1, \widetilde{\tau})^T$. Consider the t-th ($t \geq 1$) iteration of the algorithm, and denote the updated result of the t-th iteration as $\widetilde{\boldsymbol{\theta}}^{(t)} = (\widetilde{\beta}_0^{(t)}, \widetilde{\beta}_1^{(t)}, \widetilde{\tau}^{(t)})^T$. In Step 1, as $\widetilde{\beta}_1^{(t-1)}$ and $\widetilde{\beta}_{11}^{(t-1)}$ are the MLE or quasi-likelihood estimates, $\widetilde{\beta}_1^{(t-1)}$ and $\widetilde{\beta}_{11}^{(t-1)}$ are the solution of $\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} Q_1(\boldsymbol{W}_i;\beta_1,\beta_{11}|\widetilde{\tau}^{(t-1)}) \\ \frac{1}{n}\sum_{i=1}^{n} Q_2(\boldsymbol{W}_i;\beta_1,\beta_{11}|\widetilde{\tau}^{(t-1)}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. And from Step 2, $U^{(t)}$ is proportional to $\frac{1}{n}\sum_{i=1}^{n} Q_{3,i}(\widetilde{\boldsymbol{\theta}}^{(t)})$. When the proposed two-step NR algorithm converges, we have $\lim_{t\to\infty} \widetilde{\boldsymbol{\theta}}^{(t)} = \widetilde{\boldsymbol{\theta}}_n$, which implies $\lim_{t\to\infty}\{U^{(t)}/S^{(t)}\} = 0$ from Step 2. As a result, we can show that $\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} Q_1(\boldsymbol{W}_i;\widetilde{\boldsymbol{\theta}}_n) \\ \frac{1}{n}\sum_{i=1}^{n} Q_2(\boldsymbol{W}_i;\widetilde{\boldsymbol{\theta}}_n) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ from Step 1 and $\frac{1}{n}\sum_{i=1}^{n} Q_3(\boldsymbol{W}_i;\widetilde{\boldsymbol{\theta}}_n) = 0$ from Step 2. Therefore, $\widetilde{\boldsymbol{\theta}}_n$ is the solution of the proposed estimating equation $\frac{1}{n}\sum_{i=1}^{n} Q(\boldsymbol{W}_i;\boldsymbol{\theta}) = 0$.

A.4.2 Proof of Result 2

According to the Taylor expansion of $E_{\boldsymbol{W}}(Q(\boldsymbol{W};\hat{\boldsymbol{\theta}}_n))$ around $\boldsymbol{\theta}^0$, we have

$$E_{\boldsymbol{W}}(Q(\boldsymbol{W};\hat{\boldsymbol{\theta}}_n)) - E_{\boldsymbol{W}}(Q(\boldsymbol{W};\boldsymbol{\theta}^0)) \;=\; R(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0), \tag{A.32}$$

where $\boldsymbol{\theta}^* = (\boldsymbol{\theta}^{*(1)}, \boldsymbol{\theta}^{*(2)}, \boldsymbol{\theta}^{*(3)})$, $\boldsymbol{\theta}^{*(\ell)} = (\beta_0^{*(\ell)}, \beta_1^{*(\ell)}, \tau^{*(\ell)})^T$ lies between $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}^0$ for $\ell = 1, 2, 3$, and $R(\boldsymbol{\theta})$ is a negative definite matrix defined in the equation (A.31).

After some simple algebra, the equation (A.32) can be reorganized as

$$
\begin{aligned}
R(\boldsymbol{\theta}^*)\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Q(\boldsymbol{W}_i;\boldsymbol{\theta^0}) - \sqrt{n}E_{\boldsymbol{W}}(Q(\boldsymbol{W};\boldsymbol{\theta^0})) \qquad \text{(A.33)}\\
&-\Big[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{Q(\boldsymbol{W}_i;\hat{\boldsymbol{\theta}}_n) - Q(\boldsymbol{W}_i;\boldsymbol{\theta^0})\} - \sqrt{n}\{E_{\boldsymbol{W}}(Q(\boldsymbol{W};\hat{\boldsymbol{\theta}}_n)) - Q(\boldsymbol{W};\boldsymbol{\theta^0}))\}\Big]\\
&\triangleq \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Q(\boldsymbol{W}_i;\boldsymbol{\theta^0}) - \sqrt{n}E_{\boldsymbol{W}}(Q(\boldsymbol{W};\boldsymbol{\theta^0})) - \boldsymbol{\delta},
\end{aligned}
$$

where $\boldsymbol{\delta}$ denotes the formula in (A.33).

By continuous mapping theorem, we have that $R(\boldsymbol{\theta}^*)$ converges to $R(\boldsymbol{\theta}^0)$ as $n \to \infty$. According to central limit theorem, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Q(\boldsymbol{W}_i;\boldsymbol{\theta^0}) - \sqrt{n}E_{\boldsymbol{W}}(Q(\boldsymbol{W};\boldsymbol{\theta^0}))$ converges in distribution to $N(0, \phi V^{-1}(\boldsymbol{\theta}^0))$, where $V(\boldsymbol{\theta}^0) = E_{\boldsymbol{W}}\{Q(\boldsymbol{W};\boldsymbol{\theta}^0)Q^T(\boldsymbol{W};\boldsymbol{\theta}^0)\}$. Following similar steps in Appendix Section A.3.3 , we can check that $Q(\boldsymbol{W};\boldsymbol{\theta})$ belongs to a Donsker class, which further implies that $\boldsymbol{\delta} = o_p(1)$. To satisfy the regularity conditions of $Q(\boldsymbol{W};\boldsymbol{\theta})$ belonging to a Donsker class, we additionally need that g(.) is a continuous and first-order differentiable link function. This holds for distributions in the exponential family. Then by Slutsky's theorem, $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0)$ converges to $\mathcal{N}(0, \phi V^{-1}(\boldsymbol{\theta}^0))$ in distribution and $V(\boldsymbol{\theta}^0)$ can be calculated as $V(\boldsymbol{\theta}^0) = E\Big[\frac{H^T(\boldsymbol{\theta}^0)H(\boldsymbol{\theta}^0)}{v\{\mu(\boldsymbol{\theta}^0)\}g'\{\mu(\boldsymbol{\theta}^0)\}^2}\Big]$.

BIBLIOGRAPHY

BIBLIOGRAPHY

Alshaikh, H. N., N. M. Katz, F. Gani, N. Nagarajan, J. K. Canner, S. Kacker, P. A. Najjar, R. S. Higgins, and E. B. Schneider (2018), Financial impact of acute kidney injury after cardiac operations in the united states, *The Annals of thoracic surgery*, *105*(2), 469–475.

Baccini, M., et al. (2008), Heat effects on mortality in 15 european cities, *Epidemiology*, pp. 711–719.

Bacon, D. W., and D. G. Watts (1971), Estimating the transition between two intersecting straight lines, *Biometrika*, *58*(3), 525–534.

Behrman, R. E., A. S. Butler, et al. (2007), Preterm birth: causes, consequences, and prevention.

Bellman, R., and R. Roth (1969), Curve fitting by segmented straight lines, *Journal of the American Statistical Association*, *64*(327), 1079–1084.

Canty, A., and B. D. Ripley (2020), *boot: Bootstrap R (S-Plus) Functions*, r package version 1.3-25.

Carlin, B. P., A. E. Gelfand, and A. F. Smith (1992), Hierarchical bayesian analysis of changepoint problems, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *41*(2), 389–405.

Centers for Disease Control and Prevention (2020), Healthy weight: Assessing your weight/about adult bmi., `https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html`.

Chaney, R. W. (1990), Piecewise ck functions in nonsmooth analysis, *Nonlinear Analysis: Theory, Methods & Applications*, *15*(7), 649–660.

Chen, C. W., J. S. Chan, R. Gerlach, and W. Y. Hsieh (2011), A comparison of estimators for regression models with change points, *Statistics and Computing*, *21*(3), 395–414.

Chiu, G., R. Lockhart, and R. Routledge (2002), Bent-cable asymptotics when the bend is missing, *Statistics and Probability Letters*, *59*(1), 9–16.

Chiu, G., R. Lockhart, and R. Routledge (2006), Bent-cable regression theory and applications, *Journal of the American Statistical Association*, *101*(474), 542–553.

Cox, C. (1987), Threshold dose-response models in toxicology, *Biometrics*, pp. 511–523.

Cui, Y., J.-S. Pang, and B. Sen (2018), Composite difference-max programs for modern statistical estimation problems, *SIAM Journal on Optimization*, *28*(4), 3344–3374.

Das, R., M. Banerjee, B. Nan, and H. Zheng (2016), Fast estimation of regression parameters in a broken-stick model for longitudinal data, *Journal of the American Statistical Association*, *111*(515), 1132–1143.

Dasta, J. F., S. L. Kane-Gill, A. J. Durtschi, D. S. Pathak, and J. A. Kellum (2008), Costs and outcomes of acute kidney injury (aki) following cardiac surgery, *Nephrology Dialysis Transplantation*, *23*(6), 1970–1974.

Davison, A. C., and D. V. Hinkley (1997), *Bootstrap Methods and Their Applications*, Cambridge University Press, Cambridge, iSBN 0-521-57391-2.

De Somer, F., J. W. Mulholland, M. R. Bryan, T. Aloisio, G. J. Van Nooten, and M. Ranucci (2011), O 2 delivery and co 2 production during cardiopulmonary bypass as determinants of acute kidney injury: time for a goal-directed perfusion management?, *Critical Care*, *15*(4), R192.

Dobson, A. J., and A. G. Barnett (2018), *An introduction to generalized linear models*, CRC press.

Eigenbrod, F., S. J. Hecnar, and L. Fahrig (2009), Quantifying the road-effect zone: threshold effects of a motorway on anuran populations in ontario, canada, *Ecology and Society*, *14*(1).

Elder, A., and Y. Fong (2019), Estimation and inference for upper hinge regression models, *Environmental and Ecological Statistics*, *26*(4), 287–302.

Elliott, M. R., and J. T. Shope (2003), Use of a bayesian changepoint model to estimate effects of a graduated driver's licensing program, *Journal of Data Science*, *1*(1), 43–63.

Elmistekawy, E., B. McDonald, C. Hudson, M. Ruel, T. Mesana, V. Chan, and M. Boodhwani (2014), Clinical impact of mild acute kidney injury after cardiac surgery, *The Annals of thoracic surgery*, *98*(3), 815–822.

Feder, P. I. (1975a), On asymptotic distribution theory in segmented regression problems–identified case, *The Annals of Statistics*, *3*(1), 49–83.

Feder, P. I. (1975b), The log likelihood ratio in segmented regression, *The Annals of Statistics*, pp. 84–97.

Fong, Y. (2019), Fast bootstrap confidence intervals for continuous threshold linear regression, *Journal of Computational and Graphical Statistics*, *28*(2), 466–470.

Fong, Y., Y. Huang, P. B. Gilbert, and S. R. Permar (2017), chngpt: threshold regression model estimation and inference, *BMC bioinformatics*, *18*(1), 454.

Francesco Ficetola, G., and M. Denoël (2009), Ecological thresholds: an assessment of methods to identify abrupt changes in species–habitat relationships, *Ecography*, *32*(6), 1075–1084.

Hahn, G., M. Banergjee, and B. Sen (2017), Parameter estimation and inference in a continuous piecewise linear regression model, *Manuscript, Department of Statistics, Columbia University (December 2016)*, *32*(2), 407–451.

Hansen, B. E. (2000), Sample splitting and threshold estimation, *Econometrica*, *68*(3), 575–603.

Hansen, B. E. (2017), Regression kink with an unknown threshold, *Journal of Business & Economic Statistics*, *35*(2), 228–240.

Hayes, A. W., and T. A. Loomis (1996), *Loomis's essentials of toxicology*, Elsevier.

Hudson, D. J. (1966), Fitting segmented curves whose join points have to be estimated, *Journal of the American Statistical Association*, *61*(316), 1097–1129.

Julious, S. A. (2001), Inference and estimation in a changepoint regression problem, *Journal of the Royal Statistical Society: Series D (The Statistician)*, *50*(1), 51–61.

Keller, M. B., et al. (2000), A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for the treatment of chronic depression, *New England Journal of Medicine*, *342*(20), 1462–1470.

Kline-Rogers, E., et al. (2002), Development of a multicenter interventional cardiology database: the blue cross blue shield of michigan cardiovascular consortium (bmc2) experience, *Journal of interventional cardiology*, *15*(5), 387–392.

Knowles, M., D. Siegmund, and H. Zhang (1991), Confidence regions in semilinear regression, *Biometrika*, *78*(1), 15–31.

Lacasana, M., A. Esplugues, and F. Ballester (2005), Exposure to ambient air pollution and prenatal and early childhood health effects, *European journal of epidemiology*, *20*(2), 183–199.

Lee, C. Y. (2021), Nested logistic regression models and δauc applications: Changepoint analysis, *Statistical Methods in Medical Research*, p. 09622802211022377.

Lerman, P. (1980), Fitting segmented regression models by grid search, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *29*(1), 77–84.

Lin, C.-M., C.-Y. Li, G.-Y. Yang, and I.-F. Mao (2004), Association between maternal exposure to elevated ambient sulfur dioxide during pregnancy and term low birth weight, *Environmental research*, *96*(1), 41–50.

Llop, S., F. Ballester, M. Estarlich, A. Esplugues, M. Rebagliato, and C. Iñiguez (2010), Preterm birth and exposure to air pollutants during pregnancy, *Environmental research*, *110*(8), 778–785.

Magruder, J. T., et al. (2015), Nadir oxygen delivery on bypass and hypotension increase acute kidney injury risk after cardiac operations, *The Annals of thoracic surgery*, *100*(5), 1697–1703.

Marsh, L. C., and D. R. Cormier (2001), *Spline regression models*, vol. 137, Sage.

Muggeo, V. M. (2003), Estimating regression models with unknown break-points, *Statistics in Medicine*, *22*(19), 3055–3071.

Muggeo, V. M. (2008), segmented: an r package to fit regression models with broken-line relationships., *R News*, *8*(1), 20–25.

Muggeo, V. M. (2016), Testing with a nuisance parameter present only under the alternative: a score-based approach with application to segmented modelling, *Journal of Statistical Computation and Simulation*, *86*(15), 3059–3067.

Mukaida, H., S. Matsushita, K. Kuwaki, T. Inotani, Y. Minami, A. Saigusa, and A. Amano (2019), Time–dose response of oxygen delivery during cardiopulmonary bypass predicts acute kidney injury, *The Journal of thoracic and cardiovascular surgery*, *158*(2), 492–499.

Ngoc, N. T. N., et al. (2006), Causes of stillbirths and early neonatal deaths: data from 7993 pregnancies in six developing countries, *Bulletin of the World Health Organization*, *84*, 699–705.

Osornio-Vargas, A., M. A. Buxton, B. N. Sánchez, L. Rojas-Bracho, M. Castillo-Castrejon, I. B. Mordhukovich, D. G. Brown, F. Vadillo-Ortega, et al. (2013), Air pollution, inflammation and preterm birth in mexico city: study design and methods, *Science of the total environment*, *448*, 79–83.

Pastor-Barriuso, R., E. Guallar, and J. Coresh (2003), Transition models for change-point estimation in logistic regression, *Statistics in medicine*, *22*(7), 1141–1162.

Pollard, D. (2000), Asymptopia, *Manuscript, Yale University, Dept. of Statist., New Haven, Connecticut*.

Qi, L., and J. Sun (1993), A nonsmooth version of newton's method, *Mathematical programming*, *58*(1-3), 353–367.

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*(284), 873–880.

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Ranucci, M., F. Romitti, G. Isgrò, M. Cotza, S. Brozzi, A. Boncilli, and A. Ditta (2005), Oxygen delivery during cardiopulmonary bypass and acute renal failure after coronary operations, *The Annals of Thoracic Surgery*, *80*(6), 2213–2220.

Ranucci, M., et al. (2015), Acute kidney injury and hemodilution during cardiopulmonary bypass: a changing scenario, *The Annals of thoracic surgery*, *100*(1), 95–100.

Ranucci, M., et al. (2018), Goal-directed perfusion to reduce acute kidney injury: a randomized trial, *The Journal of thoracic and cardiovascular surgery*, *156*(5), 1918–1927.

Rigotti, T. (2009), Enough is enough? threshold models for the relationship between psychological contract breach and job-related attitudes, *European Journal of Work and Organizational Psychology*, *18*(4), 442–463.

Robert, A. M., et al. (2010), Cardiac surgery-associated acute kidney injury: a comparison of two consensus criteria, *The Annals of thoracic surgery*, *90*(6), 1939–1943.

Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller (2011), proc: an open-source package for r and s+ to analyze and compare roc curves, *BMC Bioinformatics*, *12*, 77.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003), *Semiparametric regression*, 12, Cambridge university press.

Scholtes, S. (2012), *Introduction to piecewise differentiable equations*, Springer Science & Business Media.

Shen, W., R. Aguilar, A. R. Montero, S. J. Fernandez, A. J. Taylor, C. S. Wilcox, M. S. Lipkowitz, and J. G. Umans (2017), Acute kidney injury and in-hospital mortality after coronary artery bypass graft versus percutaneous coronary intervention: a nationwide study, *American journal of nephrology*, *45*(3), 217–225.

Smith, A., and D. Cook (1980), Straight lines with a change-point: A bayesian analysis of some renal transplant data, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *29*(2), 180–189.

Smith, R. B., et al. (2020), Impacts of air pollution and noise on risk of preterm birth and stillbirth in london, *Environment international*, *134*, 105,290.

Šrám, R. J., B. Binková, J. Dejmek, and M. Bobak (2005), Ambient air pollution and pregnancy outcomes: a review of the literature, *Environmental health perspectives*, *113*(4), 375–382.

Sun, X., X. Luo, C. Zhao, B. Zhang, J. Tao, Z. Yang, W. Ma, and T. Liu (2016), The associations between birth weight and exposure to fine particulate matter (pm2. 5) and its chemical constituents during pregnancy: A meta-analysis, *Environmental pollution*, *211*, 38–47.

Tapsoba, J. d. D., C.-Y. Wang, S. Zangeneh, and Y. Q. Chen (2020), Methods for generalized change-point models: with applications to human immunodeficiency virus surveillance and diabetes data, *Statistics in medicine*, *39*(8), 1167–1182.

Tishler, A., and I. Zang (1981), A new maximum likelihood algorithm for piecewise regression, *Journal of the American Statistical Association*, *76*(376), 980–987.

Tsiatis, A. (2007), *Semiparametric theory and missing data*, Springer Science & Business Media.

van der Vaart, A., and J. A. Wellner (2000), Preservation theorems for glivenko–cantelli and uniform glivenko–cantelli classes, in *High Dimensional Probability II*, pp. 115–133, Springer.

van der Vaart, A. W. (2000), *Asymptotic Statistics*, vol. 3, Cambridge University Press.

van der Vaart, A. W., and J. A. Wellner (1996), *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer, New York.

Villarini, G., J. A. Smith, and G. A. Vecchi (2013), Changing frequency of heavy rainfall over the central united states, *Journal of Climate*, *26*(1), 351–357.

Wahba, A., et al. (2020), 2019 eacts/eacta/ebcp guidelines on cardiopulmonary bypass in adult cardiac surgery, *European Journal of Cardio-Thoracic Surgery*, *57*(2), 210–251.

Yanagimoto, T., and E. Yamamoto (1979), Estimation of safe doses: critical review of the hockey stick regression method., *Environmental health perspectives*, *32*, 193–199.

Yu, Y., and D. Ruppert (2002), Penalized spline estimation for partially linear single-index models, *Journal of the American Statistical Association*, *97*(460), 1042–1054.