SEMIPARAMETRIC MODELS FOR MOUTH-LEVEL INDICES IN CARIES RESEARCH

By

Yifan Yang

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Biostatistics - Master of Science

2016

ABSTRACT

SEMIPARAMETRIC MODELS FOR MOUTH-LEVEL INDICES IN CARIES RESEARCH

By

Yifan Yang

For nonnegative count responses in health services research, a large proportion of zero counts are frequently encountered. For such data, the frequency of zero counts is typically larger than its expected counterpart under the classical parametric models, such as Poisson or negative binomial model. In this thesis, a semiparametric zero-inflated regression model is proposed for count data that directly relates covariates to the marginal mean response representing the desired target of inference. The model specifically assumes two semiparametric forms: the log-linear form for the marginal mean and the logistic-linear form for the susceptible probability, in which the fully linear models are replaced with partially linear link functions. A spline-based estimation is proposed for the nonparametric components of the model. Asymptotic properties are discussed for the estimators of the parametric and nonparametric components of the models. Specifically, the estimators are shown to be strong consistent and asymptotically efficient under mild regularity conditions. A bootstrap hypothesis test is performed to evaluate difference involving the nonparametric component. Simulation studies are conducted to evaluate the finite sample performance of the model. Finally, the model is applied to dental caries indices in low income African-American children to evaluate the nonlinear effects of sugar intake on caries development. The conclusion shows that the effect of sugar intake on caries indices is nonlinear, especially among young children under the age of 2. And children whose caregivers are unemployed and have poor oral healthy exhibit higher dental caries rates.

ACKNOWLEDGMENTS

First and foremost, I really appreciate Dr. David Todem's tutelage, encouragement and help throughout the course of this research work. He is available for academic advice whenever I need and gently guides me for a good understanding of biostatistics. I am extremely grateful to Dr. Todem for his unwavering support, as well as his patience and careful planning for graduate study at Michigan State University.

I also would like to take this opportunity to thank my thesis committee members Dr. Qing Lu and Dr. Chenxi Li who have accommodated every request despite their full schedules and provide me with precious feedback. I would like to thank Dr. Joseph Gardiner, Dr. Hao Wang, Dr. Nigel Paneth, Dr. David Barondess, Dr. Zhehui Luo and Dr. Wenjiang Fu for their help and advice during my graduate studies in the Department of Epidemiology and Biostatistics.

Last but definitely not least, I would like to express my deepest gratitude to my beloved parents for their love and support over the years. I also have the pleasure of living in East Lansing with many good friends.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO SYMBOLS

All English and Greek letter-based symbols used for mathematical expressions in this thesis are listed here in alphabetical order. In general, a capital letter refers to a random variable/vector, and the small letter refers to the value/sample corresponding to the random variable/vector.

General Rule 1: English letter-based symbols refer to known data or covariates like X, Y and Z unless otherwise stated;

General Rule 2: Greek letter-based symbols refer to unknown parameters like κ, λ and π unless otherwise stated;

General Rule 3: Ralph Smith's symbols refer to specific classes of functions like $\mathscr{C}$ and $\mathscr{F}$;

General Rule 4: Blackboard bold symbols refer to special operations like the density function $\mathbb{P}$,expectation $\mathbb{E}$ and variance $\mathbb{V}$, or special number sets like the set of real numbers $\mathbb{R}$ and the set of nonnegative natural numbers $\mathbb{N}$;

General Rule 5: a star symbol refers to another point corresponding to the original one which belongs to the same set like both θ and θ^* belong to Θ.

$\mathbb{E}$: the expectation of random variables/vectors or functions of random variables/vectors with respect to specific $\mathbb{P}$;

k_0 and k_1 : the growth rates of number of knots (without containing two ends);

l : the degree of spline basis functions;

$l()$: the log-likelihood function;

L : the likelihood function;

L_2 and L_∞ : the L_2 norm and the L_∞ norm respectively;

m_0 and m_1 : the numbers of knots (without containing two ends);

n, n_0 and n_1 : the sample size of data, the sample size of data when $\Delta = 0$ and the sample size of data when $\Delta = 1$;

N_0 and N_1 : the number of parameters for spline basis functions;

$\mathbb{P}_X$: the probability measures with respect to random vector X;

r_0 and r_1 : the smoothing parameters corresponding to B_1, B_2, B_3 and B_4;

s_i and t_i : the knots on the interval $[0, 1]$;

Φ_n : a mapping that relates a point on parameter space Θ to product space Θ_n;

Γ : the gamma function.

Chapter 1

Introduction

1.1 Oral Health Research

Tooth decay, more academically known as dental caries or cavities, is considered as one of the most prevalent oral diseases, in particular among young children as young as five (Todem, 2012a). It also remains the most common chronic disease among children aged five to eleven and adolescents aged twelve to seventeen years (Dye et al., 2007). Although tooth decay does not significantly decrease school or work performance and social relationships among older adults while it does among children, it dramatically impacts on their chewing abilities, forces them to limit diet selection, and eventually contributes to other overall health problems (Sullivan et al., 1993; Blaum et al., 1995; Ritchie et al., 2000).

To evaluate the severity of dental caries at tooth surface level, many dental epidemiologists conduct the studies using the decayed, missing and filled (DMF) indices (Klein and Palmer, 1938), also called as DMFT indices when applied to all the teeth, or as DMFS index when applied only to tooth surfaces (five per posterior tooth and four per anterior tooth). The integer scores per subject range from 0 to 28/32 in DMFT system while do from 0 to 128/148 in DMFS system (Cappelli and Mobley, 2007).

1.2 Zero-Inflated Models

Investigators frequently encountered integer counts data with a high frequency of zero values in studies like a dental caries study related to DMFS indices. Zero-inflated models, which view data as being generated from a mixture of a point mass at zero and a non-degenerate distribution, have become a popular and interesting tool within the parametric framework (Mullahy, 1986; Farewell and Sprott, 1988; Lambert, 1992; Ridout et al., 2001; Gilthorpe et al., 2009; Wang et al., 2015) to analyze count data with excessive zeros. The non-degenerate distribution can be the Poisson model (Lambert, 1992; Lam et al., 2006; He et al., 2010), the negative binomial model (Yau et al., 2003; Minami et al., 2007; Wang et al., 2015), or other discrete probability distributions like the Conway-Maxwell-Poisson distribution (Shmueli et al., 2005) and so on (Loeys et al., 2012). These zero-inflated models have been applied in many fields, such as the study of length of hospital stay (Atienza et al., 2008; Singh and Ladusingh, 2010), the health care outcomes research (Hur et al., 2002), and the study of pediatric length of stay (Lee et al., 2005).

We believe the zero-inflated negative binomial (ZINB) model is more appropriate to the dental caries data because the variance is assumed to be the same as the mean for the Poisson distribution, which may be violated for real data analysis. The procedure for testing the zero-inflated Poisson (ZIP) model against the ZINB model is discussed in details as well (Ridout et al., 2001).

1.3 Method of Sieves

We consider the zero-inflated model with the semiparametric framework (Xue et al., 2004; Lam et al., 2006; He et al., 2010; Zhang et al., 2010), or called a partial linear model, which is particularly appropriate to the data when the covariate is nonlinearly related to the response. For example, I am interested in evaluating the nonlinear effect of the daily amount of sugar intake (DASI) on caries indices in primary dentition adjusting for important confounders.

It is extremely tricky to handle the nonparametric component in the semiparametric statistical studies. To approximate the nonparametric component, many statistical tools are available, such as piecewise polynomials (Chen, 1988), kernel estimator (Speckman, 1988), M estimator (Härdle and Liang, 2007), profile estimator (Ghosh, 2001), and sieve estimator (Geman and Hwang, 1982; Shen and Wong, 1994; Shen, 1997; Huang and Rossini, 1997). For real data analysis with sieve estimator, the unknown nonparametric component can be approximated by the piecewise linear functions (Xue et al., 2004; Lam et al., 2006; He et al., 2010), the triangle series (Song and Xue, 2000), the small waves functions (Shen and Shi, 2004), and the B-spline basis functions (Zhang et al., 2010). The primary thought is I perform the minimization or maximization within a subset of an infinite parameter space, then let the dimension of this subset grow with the sample size. More details about method of sieve will be given in the Section 2.3.

1.4 Organization of This Thesis

In the rest of this thesis, I propose a semiparametric ZINB model for three goals: (1) evaluating the effect of covariates on the marginal mean response, (2) investigating the nonlinear effect of the DASI on caries indices in primary dentition adjusting for important confounders, and (3) comparing whether the above nonlinear effect may vary in different ages groups. We propose the specific semiparametric ZINB model and sieve maximum likelihood (ML) estimator for both parametric and nonparametric components in Chapter

2. Given necessary assumptions, I derive the asymptotic properties of sieve estimator like strong consistency, rate of convergence, and asymptotic normality in Chapter 3. All proofs of theorems and additional lemmas are present in the appendix. Furthermore, I conduct the bootstrap hypothesis testing for nonparametric components of interest at the end of Chapter 3. In the Chapter 4, I apply the semiparametric ZINB model to real dental caries data after evaluating its advantage in simulation study. We summarize the work in this thesis and discuss its extension to other settings like penalized sieve methods.

Chapter 2

Models and Sieve Estimator

2.1 ZINB Distribution

Let Y denote the count variable that has a zero-inflated negative binomial (ZINB) distribution. Specifically, assume Y have a probability density function

$$\mathbb{P}(Y = y) = \begin{cases} \pi + (1-\pi)(1+\kappa\lambda)^{-\frac{1}{\kappa}}, & y = 0 \\ (1-\pi)\dfrac{\Gamma\left(y+\frac{1}{\kappa}\right)(\kappa\lambda)^y}{\Gamma\left(\frac{1}{\kappa}\right)\Gamma(y+1)(1+\kappa\lambda)^{y+\frac{1}{\kappa}}}, & y > 0 \end{cases} \tag{2.1}$$

where $\kappa \geq 0$ is the dispersion parameter that is assumed not to depend on covariates (Ridout et al., 2001; Wang et al., 2015), $\lambda \geq 0$ is the mean of the underlying negative binomial distribution (Wang et al., 2015), and $0 \leq \pi \leq 1$ is the probability of zero counts. This distribution reduces to the zero-inflated Poisson (ZIP) distribution in the limit $\kappa \to 0$ (Ridout et al., 2001; Minami et al., 2007). The mean and variance of ZINB distribution in (2.1) are given by

$$\mathbb{E}(Y) = (1-\pi)\lambda \tag{2.2}$$

$$\mathbb{V}(Y) = (1-\pi)\lambda(1+\kappa\lambda+\pi\lambda)$$

where $\mathbb{E}$ and $\mathbb{V}$ are the expectation operator and variance operator, respectively. Some scholars present an alternative parameterization of ZINB model as well (Yau et al., 2003).

The dispersion parameter κ can be replaced by its reciprocal called overdispersion parameter $\frac{1}{\kappa}$ in the model (Yau et al., 2003; Minami et al., 2007; Wang et al., 2015). Compared to a nonnegative κ, the representation of overdispersion parameter $\frac{1}{\kappa}$ requires $\kappa \neq 0$, which causes κ belonging an open set $(0, \infty)$. Similarly, the logarithmic function $\log(\lambda)$ with respect and the logistic function $\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$ in the next section require $\lambda \neq 0$ and $\pi \neq 1$, $\pi \neq 0$. Without loss of generality, the domain of κ in overdispersion case is restricted on a subset of the open set like $\kappa \in [\epsilon, \infty)$ and so do λ and π like $\lambda \in [\epsilon, \infty)$ and $\pi \in [\epsilon, 1-\epsilon]$ for any given $\epsilon > 0$, respectively.

The mean of negative binomial distribution λ in (2.1), can be written in terms of mean of ZINB distribution (2.2) representation, that is, $\lambda = \mathbb{E}\left(Y | I_{(Y>0)} = 1\right)$ where the identity variable $I_{(Y>0)}$ takes value 1 if $Y > 0$ and takes value 0 otherwise. Thus, I call $\mathbb{E}(Y)$ in (2.2) as the marginal mean of ZINB distribution corresponding to the conditional mean $\lambda = \mathbb{E}\left(Y | I_{(Y>0)} = 1\right)$.

2.2 Semiparametric ZINB Marginal Mean Regression

In the regression setting, λ and π are called as latent variables, and both $\text{logit}(\pi)$ and $\log(\lambda)$ are assumed to depend on a linear function of covariates for almost all cases (Lambert, 1992; Ridout et al., 2001; Yau et al., 2003; Minami et al., 2007; Wang et al., 2015). Although this latent variables formulation in some settings provides a versatile and useful representation of the data, the implied regression parameterization may fail to provide a clear answer to the question of evaluating the covariate effects on the marginal mean response. Therefore, in this thesis, the latent variable π and the marginal mean $\mathbb{E}(Y)$ are assumed to depend on a linear function of covariates, given by

$$\log(\mathbb{E}(Y)) = \alpha^\top \mathbf{X} \tag{2.3}$$

$$\operatorname{logit}(\pi) = \beta^\top \mathbf{Z} \tag{2.4}$$

where $\mathbf{X} = \left(1, X_1, \ldots, X_{d_1}\right)^\top$ and $\mathbf{Z} = \left(1, Z_1, \ldots, Z_{d_2}\right)^\top$ are $(d_1 + 1) \times 1$ and $(d_2 + 1) \times 1$ covariates vectors, α and β are vectors of unknown regression coefficients, respectively, $\operatorname{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$ and the symbol $^\top$ is the transpose of a vector or matrix. The covariates that affect the marginal mean of outcome may or may not be the same as the covariates that affect the probability of zero counts.

Lam and Xue (2006) considered a semiparametric link function for their ZIP model (Lam et al., 2006), then He and Xue (2010) extended the ZIP model to the doubly semiparametric ZIP model (He et al., 2010). Motivated by these, I extend the parametric ZINB model to a semiparametric one with partially linear link functions for both the marginal mean $\mathbb{E}(Y)$ and the logit of the probability of zeros $\operatorname{logit}(\pi)$, expressed as the following joint model

$$\log(\mathbb{E}(Y)) = \log[(1-\pi)\lambda] = \alpha^\top \mathbf{X} + (1-\Delta) g_0(S) + \Delta g_1(S) \tag{2.5}$$

$$\operatorname{logit}(\pi) = \beta^\top \mathbf{Z} + (1-\Delta) h_0(S) + \Delta h_1(S) \tag{2.6}$$

where $Y \in A_Y \subseteq \mathbb{N}$, $\mathbf{X} = \left(X_1, \ldots, X_{d_1}\right)^\top \in A_X \subseteq \mathbb{R}^{d_1}$ and $\mathbf{Z} = \left(Z_1, \ldots, Z_{d_2}\right)^\top \in A_Z \subseteq \mathbb{R}^{d_2}$ are $d_1 \times 1$ and $d_2 \times 1$ covariates vectors without intercepts, α and β are vectors of unknown regression coefficients as well, $g_0(S), g_1(S), h_0(S)$ and $h_1(S)$ are unknown smoothing functions with respect to continuous covariate $S \in [0, 1]$ that occurs between 0 and 1 for fixed binary variable $\Delta \in \{0,1\}$ that takes value 0 or 1, latent variables satisfy that $\lambda \in A_\lambda \subseteq [0,\infty)$ and $\pi \in A_\pi \subseteq [0,1]$. We call the model (2.1), (2.5) and (2.6) as semiparametric ZINB marginal mean model. For real data analysis, Δ stands for different groups that subjects belong to.

REMARK 2.2.1. Notice the covariates vectors $\mathbf{X}$ in (2.3) and (2.5) have different dimensions, so do $\mathbf{Z}$ in (2.4) and (2.6). We put the intercept terms $X_0 = 1$ and $Z_0 = 1$ into the nonparametric components like other do (Xue et al., 2004; Lam et al., 2006; He et al., 2010)

2.3 Maximum Likelihood Estimation by Method of Sieves

Let $W = \left(Y, \mathbf{X}^\top, \mathbf{Z}^\top, \Delta, S\right)^\top \in \Omega$ be the data vector, where the sample space Ω is given by

$$
\begin{aligned}
\Omega &\triangleq \left\{W : W \in A_Y \times A_X \times A_Z \times \{0,1\} \times [0,1]\right\} \\
&= A_Y \times A_X \times A_Z \times \{0,1\} \times [0,1] \\
&\subseteq \mathbb{N} \times \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \{0,1\} \times [0,1]
\end{aligned}
$$

Let $\theta = \left(\alpha^\top, \beta^\top, \kappa, g_0, g_1, h_0, h_1\right)^\top$ be the vector of all the unknown quantities of interest with $\theta_\mathrm{T} = \left(\alpha_\mathrm{T}^\top, \beta_\mathrm{T}^\top, \kappa_\mathrm{T}, g_{\mathrm{T},0}, g_{\mathrm{T},1}, h_{\mathrm{T},0}, h_{\mathrm{T},1}\right)^\top$ as the unique true value of θ. Assume the parameter space be given by

$$
\begin{aligned}
\Theta &\triangleq \left\{\theta : \alpha \in A_1, \beta \in A_2, \kappa \in A_3, g_0 \in B_1, g_1 \in B_2, h_0 \in B_3, h_1 \in B_4\right\} \\
&= A_1 \times A_2 \times A_3 \times B_1 \times B_2 \times B_3 \times B_4
\end{aligned}
$$

where A_1 and A_2 are compact sets in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, A_3 is a compact set in nonnegative real set $\mathbb{R}_0^+$, and B_1, B_2, B_3 and B_4 are sets of functions that have bounded continuous derivatives on $[0,1]$, and the r_i-th derivative is γ_i-Hölder continuous on $[0,1]$ for $0 < \gamma_i \leq 1$. Specifically, they are given by

$$B_i = \left\{ f \in \mathscr{C}^{r_i}[0,1] : \left\| f^{(j)} \right\|_\infty < \infty, j = 0, 1, \ldots, r_i, \left| f^{(r_i)}(s) - f^{(r_i)}(s^*) \right| \leq C_{r_i,i} \left| s - s^* \right|^{\gamma_i}, \forall s, s^* \in [0,1] \right\} \tag{2.7}$$

where $\mathscr{C}^{r_i}[0,1]$ is the collection of functions with r_i-th continuous derivative on $[0,1]$, $f^{(j)}$ is the j-th derivative of f, constant $C_{r_i,i}$, γ_i and positive integer r_i are given by investigator for $i = 1, 2, 3, 4$. Let r denote $\min\{r_1, r_2, r_3, r_4\}$ in the next section. For any observation w of W, the density function of W is given by

$$\begin{aligned}
&\mathbb{P}_W\left(W = w; \theta\right) \\
&\quad = \mathbb{P}_{Y|\mathbf{X},\mathbf{Z},\Delta,S}\left(Y = y | \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}, \Delta = \delta, S = s\right) \\
&\quad \times \mathbb{P}_{\mathbf{X},\mathbf{Z},\Delta,S}\left(\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}, \Delta = \delta, S = s\right) \\
&\quad = \left[\pi + (1-\pi)(1+\kappa\lambda)^{-\kappa^{-1}}\right]^{\mathbb{I}_{(y=0)}(y)} \\
&\quad \times \left[(1-\pi)\frac{\Gamma\left(y+\kappa^{-1}\right)(\kappa\lambda)^y}{\Gamma\left(\kappa^{-1}\right)\Gamma\left(y+1\right)(1+\kappa\lambda)^{y+\kappa^{-1}}}\right]^{\mathbb{I}_{(y>0)}(y)} \\
&\quad \times \mathbb{P}_{\mathbf{X},\mathbf{Z},\Delta,S}\left(\mathbf{x}, \mathbf{z}, \delta, s\right)
\end{aligned}$$

where

$$\log\left[(1-\pi)\lambda\right] = \alpha^\top \mathbf{x} + (1-\delta)g_0(s) + \delta g_1(s) \tag{2.8}$$

$$\text{logit}(\pi) = \beta^\top \mathbf{z} + (1-\delta)h_0(s) + \delta h_1(s) \tag{2.9}$$

and the identify function $\mathbb{I}_X(x)$ takes value 1 if $x \in X$, takes value 0 otherwise, and $\mathbb{P}_{\mathbf{X},\mathbf{Z},\Delta,S}(\mathbf{x}, \mathbf{z}, \delta, s)$ is the joint density function of $(\mathbf{X}, \mathbf{Z}, \Delta, S)$.

The log-likelihood function can be represented as follows after setting $\mathbb{P}_{\mathbf{X},\mathbf{Z},\Delta,S}(\mathbf{x}, \mathbf{z}, \delta, s)$ aside in the estimation of θ,

$$\begin{aligned}
& l(\theta, w) \\
&= \log \mathbb{P}_W(W = w, \theta) \\
&= \log \left\{ \left[\pi + (1-\pi)(1+\kappa\lambda)^{-\kappa^{-1}}\right]^{\mathbb{I}_{(y=0)}(y)} \right. \\
&\times \left. \left[(1-\pi)\frac{\Gamma\left(y+\kappa^{-1}\right)(\kappa\lambda)^y}{\Gamma\left(\kappa^{-1}\right)\Gamma\left(y+1\right)(1+\kappa\lambda)^{y+\kappa^{-1}}}\right]^{\mathbb{I}_{(y>0)}(y)} \right\} \\
&= \mathbb{I}_{(y=0)}(y) \log\left[\pi + (1-\pi)(1+\kappa\lambda)^{-\kappa^{-1}}\right] \\
&+ \mathbb{I}_{(y>0)}(y) \log\left[(1-\pi)\frac{\Gamma\left(y+\kappa^{-1}\right)(\kappa\lambda)^y}{\Gamma\left(\kappa^{-1}\right)\Gamma\left(y+1\right)(1+\kappa\lambda)^{y+\kappa^{-1}}}\right]
\end{aligned} \tag{2.10}$$

REMARK 2.3.1. According to the definition of B_i in (2.7), any $f \in B_i$ is continuous on [0,1], so f is bounded. Plus, bounded π, λ always belongs to closed sets A_π and A_λ. According to (2.10), given $W = w$, the mapping $l : A_\pi \times A_3 \times A_\lambda \to \mathbb{R}$ is an elementary function with respect to (π, κ, λ), so l is continuous with respect to (π, κ, λ) on $A_\pi \times A_3 \times A_\lambda$. On the other hand, according to (2.8) and (2.9), functions π and λ are continuous with respect to $(\alpha, \beta, g_0, g_1, h_0, h_1)$ on $A_1 \times A_2 \times B_1 \times B_2 \times B_3 \times B_4$. Because the composition of continuous functions is continuous, l is continuous with respect to θ on Θ.

Suppose $\widetilde{W} = (w_1, w_2, \ldots, w_n) \in \Omega^n$ are n independent random samples, and the sample size of $\widetilde{W}|_{\Delta=0}$ and the sample size of $\widetilde{W}|_{\Delta=1}$ are n_0, n_1, respectively. It is true that $n = n_0 + n_1$. Let $0 = s_0 < s_1 < \cdots < s_{m_0+1} = 1$ and $0 = t_0 < t_1 < \cdots < t_{m_1+1} = 1$ be two partitions of $[0, 1]$, where m_0, m_1 are the number of knots without containing 0 and 1, and both m_0 and m_1 are integers that grow at rate $n_0^{k_0}$, $n_1^{k_1}$ for $0 < k_0 < 1$ and $0 < k_1 < 1$ when $\Delta = 0, \Delta = 1$, respectively. In other words, $m_0 = n_0^{k_0}$ and $m_1 = n_1^{k_1}$.

REMARK 2.3.2. There are different definitions of the number of knots on an interval. Some define the number of knots containing two ends of the interval (Huang and Rossini, 1997; Xue et al., 2004; Lam et al., 2006; He et al., 2010), but some define it without two ends of interval (Zhang et al., 2010). In this thesis, I agree with the number of knots without containing two ends of the interval, and always claim it when using the notation.

Given two partitions of $[0, 1]$, it implies there are $N_0 = m_0 + 1 + l$ and $N_1 = m_1 + 1 + l$ normalized uniform B splines basis of l-th degree ($(l + 1)$-th order) (Zhang et al., 2010), and let $\phi_1, \phi_2, \ldots, \phi_{N_0}$ and $\varphi_1, \varphi_2, \ldots, \varphi_{N_1}$ denote the spline basis functions on [0,1], when $\Delta = 0$ and $\Delta = 1$, respectively. Then, the linear spaces spanned by B spline basis $\left\{\phi_1, \phi_2, \ldots, \phi_{N_0}\right\}$ and $\left\{\varphi_1, \varphi_2, \ldots, \varphi_{N_1}\right\}$ with bounded parameters $\tau_{n,1} = (\tau_{1,n,1}, \tau_{2,n,1}, \ldots, \tau_{N_0,n,1})^\top$, $\tau_{n,2} = (\tau_{1,n,2}, \tau_{2,n,2}, \ldots, \tau_{N_1,n,2})^\top$, $\tau_{n,3} = (\tau_{1,n,3}, \tau_{2,n,3}, \ldots, \tau_{N_0,n,3})^\top$,

$\tau_{n,4} = (\tau_{1,n,4}, \tau_{2,n,4}, \ldots, \tau_{N_1,n,4})^\top$ are given by

$$B_{n,1} = \left\{ g_{n,0} : g_{n,0} = \sum_{i=1}^{N_0} \tau_{i,n,1}\phi_i, \max_{1\le i\le N_0} \left\{\left|\tau_{i,n,1}\right|\right\} \le M_{1,2} \right\}$$
$$B_{n,2} = \left\{ g_{n,1} : g_{n,1} = \sum_{i=1}^{N_1} \tau_{i,n,2}\varphi_i, \max_{1\le i\le N_1} \left\{\left|\tau_{i,n,2}\right|\right\} \le M_{2,2} \right\}$$
$$B_{n,3} = \left\{ h_{n,0} : h_{n,0} = \sum_{i=1}^{N_0} \tau_{i,n,3}\phi_i, \max_{1\le i\le N_0} \left\{\left|\tau_{i,n,3}\right|\right\} \le M_{3,2} \right\}$$
$$B_{n,4} = \left\{ h_{n,1} : h_{n,1} = \sum_{i=1}^{N_1} \tau_{i,n,4}\varphi_i, \max_{1\le i\le N_1} \left\{\left|\tau_{i,n,4}\right|\right\} \le M_{4,2} \right\}$$

where $M_{1,2}, M_{2,2}, M_{3,2}, M_{4,2}$ are known constants. We denote the product space $\Theta_n \triangleq A_1 \times A_2 \times A_3 \times B_{n,1} \times B_{n,2} \times B_{n,3} \times B_{n,4}$ depending on n.

REMARK 2.3.3. A specific partition of $[0, 1]$ determines the number of knots without containing 0 and 1. Given the partition of [0, 1] and l, the spline basis functions $\phi_1, \phi_2, \ldots, \phi_{N_0}$ and $\varphi_1, \varphi_2, \ldots, \varphi_{N_1}$ on [0, 1] are determined by the recursion formula, like De Boor's algorithm (De Boor, 1972). Specifically, De Boor's algorithm states, $\phi_{i,1}(u)$ takes value 1 if $u \in [u_i, u_{i+1})$ and takes value 0 otherwise; then, for any $1 < j \le l$, $\phi_{i,j}(u) = \left(\frac{u-u_i}{u_{i+j-1}-u_i}\right)\phi_{i,j-1}(u) + \left(\frac{u_{i+j}-u}{u_{i+j}-u_{i+1}}\right)\phi_{i+1,j-1}(u)$, where $u_1 = \cdots = u_l = s_0 = 0$, $u_{l+1} = s_1, \ldots, u_{N_0-1} = s_{m_0}$ and $u_{N_0} = \cdots = u_{N_0+l} = s_{m_0+1} = 1$; finally, denote ϕ_i instead of $\phi_{i,l}$ for $1 \le i \le N_0$. φ_i is generated in the same way for $1 \le i \le N_1$. Therefore, given the partition of [0, 1], l, the sample size n, $M_{1,2}, M_{2,2}, M_{3,2}, M_{4,2}$, the specific $B_{n,1}, B_{n,2}, B_{n,3}$ and $B_{n,4}$ are determined, which means $g_{n,0}, g_{n,1}, h_{n,0}$ and $h_{n,1}$ can be given by N_0 and N_1 unknowing control parameters. In general, four different partitions of $[0, 1]$ can be used for generating $g_{n,0}, g_{n,1}, h_{n,0}$ and $h_{n,1}$.

Given the data vector W, let $\mathbb{E}_{\mathrm{T}}$ denote the expectation with respect to $\mathbb{P}_{\theta_{\mathrm{T}}}$ where $\mathbb{P}_\theta$ denotes the density of W under the parameter θ. For any two points θ and θ^* in Θ, a (pseudo) distance ρ is defined by

$$\begin{aligned}\rho^2(\theta, \theta^*) &= \mathbb{E}_{\mathrm{T}}\left\{(\alpha - \alpha^*)^\top \mathbf{X} + (1-\Delta)(g_0 - g_0^*) + \Delta(g_1 - g_1^*)\right\}^2 \\ &+ \mathbb{E}_{\mathrm{T}}\left\{(\beta - \beta^*)^\top \mathbf{Z} + (1-\Delta)(h_0 - h_0^*) + \Delta(h_1 - h_1^*)\right\}^2 \\ &+ |\kappa - \kappa^*|^2 \qquad (2.11)\end{aligned}$$

Corollary 1. *Given the definition of B_i, I can choose the specific $B_{n,i}$ such that $\Theta_n \subseteq \Theta$ by modifying the specific $M_{i,2}$.*

Corollary 2. *For any $\theta = (\alpha^\top, \beta^\top, \kappa, g_0, g_1, h_0, h_1)^\top \in \Theta$, there exists a mapping $\Phi_n : \Theta \to \Theta_n$ such that*

$$\Phi_n\theta = (\alpha^\top, \beta^\top, \kappa, g_{n,0}, g_{n,1}, h_{n,0}, h_{n,1})^\top \in \Theta_n \qquad (2.12)$$

and $\rho(\theta, \Phi_n\theta) \to 0$ as $n \to \infty$.

REMARK 2.3.4. We can use the notation $\theta_n \triangleq \Phi_n\theta = (\alpha_n, \beta_n, \kappa_n, g_{n,0}, g_{n,1}, h_{n,0}, h_{n,1})^\top$ where $\alpha_n = \alpha, \beta_n = \beta$ and $\kappa_n = \kappa$ because the mapping Φ_n only functions on nonparametric components while maintaining the parametric components. For simplicity, I remain the notation in (2.12).

Θ_n is a finite dimensional space belonging to $\mathbb{R}^{d_1+d_2+1+2N_0+2N_1}$, then the sieve method is to approximate the infinite space Θ by using a series of finite spaces $\{\Theta_n\}_{n=1}^\infty$, called as the sieve spaces of Θ. In general, it does not require Θ_n is a subset of Θ, but in most cases, $\Theta_n \subseteq \Theta$ (Song and Xue, 2000). In this thesis, I can prove that it is the subset of Θ by using the properties of B spline basis functions in Corollary 1. Corollary 2 motivates us to select Θ_n as a sieve space of Θ. Furthermore, let $\mathbb{P}_n$ denote the empirical measure on sample space Ω. Let $L_n(\theta;\widetilde{W}) \triangleq \mathbb{P}_n(l(\theta;\widetilde{W})) = \frac{1}{n}\sum_{i=1}^{n} l(\theta; w_i)$ be the empirical objective function, then

$$\hat{\theta}_n \triangleq (\hat{\alpha}, \hat{\beta}, \hat{\kappa}, \hat{g}_{n,0}, \hat{g}_{n,1}, \hat{h}_{n,0}, \hat{h}_{n,1})^\top = \underset{\theta\in\Theta_n}{\operatorname{argsup}} L_n(\theta, \widetilde{W})$$

is called as the sieve estimator for θ_{T}. After similar operating in REMARK 2.3.1, L_n is continuous with respect to θ on Θ_n. On the other hand, Θ_n is a boundary closed set, therefore, $\hat{\theta}_n$ must exist.

Chapter 3

Asymptotic Properties and Bootstrap Hypothesis Test

This section provides the general assumptions used through this thesis, asymptotic properties of sieve estimators, and the test for the nonparametric components. In addition to five common Assumptions **C1** - **C5** used in the last section, five more additional mild Assumptions **A1** - **A5** established for the study of asymptotic properties of sieve MLE. Specifically, the sieve estimate $\hat{\theta}_n$ is strong consistent, it converges to the true parameter at an optimal rate $O_p\left(n^{-\frac{r}{1+2r}}\right)$, the asymptotical variance of nonparametric components can be obtained by the estimates of Hessian matrix in a numerical way, the asymptotical variance of parametric components can be determined by the estimated Fisher information matrix in a closed form, or the estimated Hessian matrix. At the end of this section, I build two statistics and conduct a bootstrap hypothesis test for nonparametric components I am interested in.

3.1 Asymptotic Properties of Sieve Estimator

Before deriving the asymptotic results in this thesis, I summarize some necessary assumptions here. The similar assumptions are held for the proportional odds regression model with interval censoring (Huang and Rossini, 1997), the semiparametric regression model with censored data (Xue et al., 2004), the semiparametric ZIP model (Lam et al., 2006), the Cox model with interval censored data (Zhang et al., 2010), and the doubly semiparametric ZIP model (He et al., 2010).

ASSUMPTION C1: The unique true value $\theta_{\mathrm{T}} \in \Theta$, that is, $\alpha_{\mathrm{T}} \in A_1, \beta_{\mathrm{T}} \in A_2, \kappa_{\mathrm{T}} \in A_3, g_{\mathrm{T},0} \in B_1, g_{\mathrm{T},1} \in B_2, h_{\mathrm{T},0} \in B_3$ and $h_{\mathrm{T},1} \in B_4$.

ASSUMPTION C2: A_1 and A_2 are compact sets in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively, and A_3 is a compact subset of the non-negative real number set $\mathbb{R}_0^+$.

ASSUMPTION C3: A_X and A_Z are bounded, that is, there exist constants $M_{1,1}$ and $M_{2,1}$ such that $\mathbb{P}(\|\mathbf{X}\|_2 \leq M_{1,1}) = 1$, $\mathbb{P}(\|\mathbf{Z}\|_2 \leq M_{2,1}) = 1$, while their diameters $M_{1,1}^d$ and $M_{2,1}^d$ such that $\|\mathbf{X} - \mathbf{X}^*\|_2 \leq M_{1,1}^d$, $\|\mathbf{Z} - \mathbf{Z}^*\|_2 \leq M_{2,1}^d$ for any $\mathbf{X}, \mathbf{X}^* \in A_X$, $\mathbf{Z}, \mathbf{Z}^* \in A_Z$.

ASSUMPTION C4: The joint density function $\mathbb{P}_{\mathbf{X},\mathbf{Z},\Delta,S}\left(\mathbf{x}, \mathbf{z}, \delta, s\right)$ does not depend on the unknown parameter θ.

ASSUMPTION C5: Given the partitions $\left\{s_j\right\}_{j=0}^{m_0+1}$ and $\left\{t_j\right\}_{j=0}^{m_1+1}$ of [0,1], $\max_{1\leq j\leq m_0+1}\{s_j - s_{j-1}\} \leq Cn^{-k_0}$ and $\max_{1\leq j\leq m_1+1}\{t_j - t_{j-1}\} \leq Cn^{-k_1}$ for some constant C and $0 < k_0 < 1$ and $0 < k_1 < 1$.

ASSUMPTION A1: θ_{T} is an interior point of Θ.

ASSUMPTION A2: The true nonparametric components $g_{\mathrm{T},0}, g_{\mathrm{T},1}, h_{\mathrm{T},0}$ and $h_{\mathrm{T},1}$ are at least second order continuously differentiable.

ASSUMPTION A3 $\mathbb{E}_{\mathrm{T}}\left[\left(\mathbf{X} - \mathbb{E}_{\mathrm{T}}(\mathbf{X}|S)\right)\left(\mathbf{X} - \mathbb{E}_{\mathrm{T}}(\mathbf{X}|S)\right)^{\top}\right] > 0$, and
$\mathbb{E}_{\mathrm{T}}\left[\left(\mathbf{Z} - \mathbb{E}_{\mathrm{T}}(\mathbf{Z}|S)\right)\left(\mathbf{Z} - \mathbb{E}_{\mathrm{T}}(\mathbf{Z}|S)\right)^{\top}\right] > 0$

ASSUMPTION A4: The joint density function $\mathbb{P}_{\mathbf{X},\mathbf{Z},\Delta,S}\left(\mathbf{x}, \mathbf{z}, \delta, s\right)$ is second order continuously differentiable with respect to S with a bounded derivative.

ASSUMPTION A5: Restrict the partitions such that $\min_{1\leq j\leq m_0+1}\{s_j - s_{j-1}\} = O(n^{-k^*})$ and $\min_{1\leq j\leq m_1+1}\{t_j - t_{j-1}\} = O(n^{-k^*})$, where $k \leq k^* \leq \frac{1-k}{2}$ when $\frac{1}{5} < k < \frac{1}{3}$, $k \leq k^* \leq 2k$ when $\frac{1}{8} < k \leq \frac{1}{5}$, and $k = \min\{k_0, k_1\}$.

REMARK 3.1.1. Assumptions **C1** and **A1** determine the searching space for the true value. Assumption **C2** guarantees that the MLE method is conducted on a finite closed product space and the maximum must exist on it. Assumption **C3** is used for computing the upper boundary of the L_2 norms of nonparametric components in Lemma 2, and is used for counting the covering number in Lemma 5 as well. Assumption **C4** allows us to set aside the nuisance parts when deriving the log-likelihood function. Assumption **C5** and **A5** imply that the number of knots without two ends depend on the sample sizes n_0, n_1 and tunning parameters k_0, k_1. Assumption **A2** and **A4** are required for computing Fisher information matrix in Theorem 3. Assumption **A3** guarantees the parametric components and nonparametric components are separable from the distance given in (2.11).

Theorem 1 (Strong Consistency)**.** *Suppose the Assumptions **C1** - **C5** are held, then* $\rho\left(\hat{\theta}_n, \theta_{\mathrm{T}}\right)$ $\to 0$ *almost surely under* $\mathbb{P}_{\theta_{\mathrm{T}}}$. *Moreover, if the Assumption **A3** is satisfied, then*

$$\|\hat{\alpha}_n - \alpha_{\mathrm{T}}\| \to 0, \|\hat{\beta}_n - \beta_{\mathrm{T}}\| \to 0, |\hat{\kappa}_n - \kappa_{\mathrm{T}}| \to 0,$$
$$\|\hat{g}_{n,0} - g_{\mathrm{T},0}\|_2 \to 0, \|\hat{g}_{n,1} - g_{\mathrm{T},1}\|_2 \to 0,$$
$$\|\hat{h}_{n,0} - h_{\mathrm{T},0}\|_2 \to 0, \|\hat{h}_{n,1} - h_{\mathrm{T},1}\|_2 \to 0$$

almost surely under $\mathbb{P}_{\theta_{\mathrm{T}}}$.

REMARK 3.1.2. The proof of this theorem is similar to arguments used for the doubly semiparametric ZIP model (He et al., 2010), but it is more tricky to build an appropriate distance due to two more nonparametric components and noncovariate parameter κ. The most important part of the proof is to give the upper boundary of the covering number (Pollard, 1990). When using this upper boundary and other lemmas, I can follow the arguments used for the other semiparametric model (Xue et al., 2004; Lam et al., 2006; He et al., 2010).

Theorem 2 (Rate of Convergence)**.** *Suppose the Assumptions **C1** - **C5** and **A3** are held, then*

$$\rho\left(\hat{\theta}_n, \theta_{\mathrm{T}}\right) = O_p\left(\max\left\{n^{-\frac{1-k}{2}}, n^{-rk}\right\}\right)$$

If select $k = \frac{1}{1+2r}$ *for* $r = 1, 2$, *then* $\rho\left(\hat{\theta}_n, \theta_{\mathrm{T}}\right)$ *achieves the optimal nonparametric convergence rate* $O_p\left(n^{-\frac{r}{1+2r}}\right)$.

REMARK 3.1.3. Because the empirical objective function $L_n(\theta, \widetilde{W})$ is bounded, it suffices to verify the three conditions in Shen's theorem 1 (Shen and Wong, 1994). Kullback Leibler distance is greater than the square of the Hellinger distance (Shen and Wong, 1994), which implies the model $\mathscr{P} = \left\{L_n(\theta, \widetilde{W}) : \theta \in \Theta_n\right\}$ is identifiable.

Theorem 3 (Asymptotic Normality and Efficiency)**.** *Under the Assumption **C1** - **C5** and **A1** - **A5**, I have*

$$\begin{aligned}\sqrt{n}\left(\hat{\alpha}_n - \alpha_{\mathrm{T}}, \hat{\beta}_n - \beta_{\mathrm{T}}, \hat{\kappa} - \kappa_{\mathrm{T}}\right)^{\top} &= I^{-1}(\theta_{\mathrm{T}})\sqrt{n}\mathbb{P}_n \widetilde{l}_{\alpha,\beta,\kappa}(\theta_{\mathrm{T}}, W) + o_p(1) \\ &\xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta_{\mathrm{T}}))\end{aligned}$$

where $I^{-1}(\theta_{\mathrm{T}}) = \mathbb{E}_{\mathrm{T}}\left(\widetilde{l}_{\alpha,\beta,\kappa}\widetilde{l}^{\top}_{\alpha,\beta,\kappa}\right) > 0$ *is the Fisher information matrix, and* $\widetilde{l}_{\alpha,\beta,\kappa}$ *is the efficient score function of* (α, β, κ).

REMARK 3.1.4. $I^{-1}(\theta_{\mathrm{T}}) > 0$ guarantees that the likelihood function achieves maximum at θ_{T}.

It is extremely tricky to derive both the efficient score function and the Fisher information matrix in a closed form due to non-covariate parameter κ and more than two nonparametric components, so I ignore the proofs. For the specific formulas of the efficient score function and the Fisher information matrix, please refer to their articles (Xue et al., 2004; He et al., 2010; Huang, 1999; Ma, 2009; Sasieni, 1992). For real data analysis, an alternative way to compute asymptotic variance by Hessian matrix in a numerical way.

In summary, I review the sieve estimator as the following facts:

(1) Given the data $\widetilde{W} \in \Omega^n$, first of all, I choose appropriate compact sets A_1, A_2, A_3, and classes of functions B_1, B_2, B_3 and B_4 with specific r_i corresponding to B_i (like $r_i = 1, 2, 3, \ldots$ for $i = 1, 2, 3, 4$) such that their product space Ω covers the unique (interior point) true value θ_{T} (by Assumptions **C1** and **A1**), denoting $r = \min\{r_1, r_2, r_3, r_4\}$;

(2) Given r, to archive the optimal convergence rate, let $k = k_0 = k_1$ be $\frac{1}{1+2r}$ (by Corollary 1, Theorem 2, Assumption **C5**), and k^* is determined (by Assumption **A5**). In particular, if $r = 2$, then $k = \frac{1}{5}$, and any k^* such that $\frac{1}{5} \le k^* \le \frac{2}{5}$, and if $r = 3$, then $k = \frac{1}{7}$, and any k^* such that $\frac{1}{7} \le k^* \le \frac{2}{7}$. In this thesis, I choose $k^* = k$ for $r = 2, 3$;

(3) Given k and sample sizes n_0 and n_1, the partitions are given by the number of knots without containing 0 and 1 $m_0 = \left\lfloor n_0^k \right\rfloor$ and $m_1 = \left\lfloor n_1^k \right\rfloor$, respectively, where $\lfloor . \rfloor$ is the round function (Zhang et al., 2010). Let $\{s_j\}_{j=0}^{m_0+1}$ and $\{t_j\}_{j=0}^{m_0+1}$ denote the partitions on [0, 1] with containing 0 and 1, respectively. Given B spline basis functions of l-th degree, for example, cubic spline basis functions have 3 degrees, $N_0 = m_0 + 1 + l$ and $N_1 = m_1 + 1 + l$ and B spline basis functions $\left\{\phi_1, \phi_2, \ldots, \phi_{N_0}\right\}$ and $\left\{\varphi_1, \varphi_2, \ldots, \varphi_{N_1}\right\}$ are determined (by REMARK 2.3.3);

(4) Given the B spline basis functions $\left\{\phi_1, \phi_2, \ldots, \phi_{N_0}\right\}$ and $\left\{\varphi_1, \varphi_2, \ldots, \varphi_{N_1}\right\}$, I choose $M_{i,2} \le \min_{0 \le j \le r_i} \left\{ \left\| f^{(j)} \right\|_\infty \right\}$ corresponding to B_i in (2.7), then $B_{n,i}$ are determined. Furthermore, $\Theta_n \subseteq \Theta$ is determined (by Corollary 1);

(5) Given a distance ρ on Θ (by (2.11)), for any point $\theta \in \Theta$, I can find a point $\theta_n \in \Theta_n$

corresponding to θ such that $\rho(\theta, \theta_n) \to 0$ as $n \to \infty$ (by Corollary 1). In other words, for the true value $\theta_\mathrm{T} \in \Theta$, there are a series of $\{\theta_n\}_{n=1}^{\infty}$ such that $\rho(\theta_\mathrm{T}, \theta_n) \to 0$ as $n \to \infty$;

(6) In order to estimate θ_n, I search the maximum of $L_n(\theta, \widetilde{W})$ on Θ_n. The estimator $\hat{\theta}_n = \underset{\theta \in \Theta_n}{\text{argsup}} L_n(\theta, \widetilde{W})$ must exist. On the one hand, θ_T is the maximizer of $L_n(\theta, \widetilde{W})$ almost surely under $\mathbb{P}_{\theta_\mathrm{T}}$ (by REMARK 3.1.4), on the other hand,$\rho(\hat{\theta}_n, \theta_\mathrm{T}) \to 0$ almost surely under $\mathbb{P}_{\theta_\mathrm{T}}$ (by Theorem 1);

(7) At the end, parametric components and nonparametric components are separable from the distance (by Assumption **A3** and Theorem 1).

3.2 Bootstrap Hypothesis Test

For real data analysis, I may be interesting in testing if two certain nonparametric components are significantly different, a general null hypothesis can be formulated as

$$H_0 : u(s) \equiv v(s), s \in [S_*, S^*]$$

where $u(s)$ and $v(s)$ are any two of $g_0(s), g_1(s), h_0(s), h_1(s)$ that are specified by the investigators, and $[S_*, S^*]$ are common boundaries for both $u(s)$ and $v(s)$. Let $\hat{u}(s)$ and $\hat{v}(s)$ denote the estimate of $u(s)$ and $v(s)$ obtained by using the sieve estimator, respectively.

A test statistic T_2 based on L_2 norm is proposed by (Hu et al., 2012),

$$T_2^2 = \int_{s \in [S_*, S^*]} \|u(s) - v(s)\|_2^2 \, ds$$

This integral can be obtained by using Monte Carlo integration. Because it is computing intensive, in order to speed up, I can use L_∞ norm instead of L_2 norm, the T_∞ statistic is given by $T_\infty = \sup_{s \in [S_*, S^*]} |u(s) - v(s)|$.

Generally, it requires the specific sampling distribution of T_2 or T_∞ to calculate the p-value, but I hardly derive the closed form due to the complication in estimating the joint sampling distribution. Thus, an alternative method to compute p-value is by conducting bootstrap approach to approximate the null distribution of T_2 or T_∞ (Hu et al., 2012).

To investigate the convergence and asymptotic normality for bootstrap statistic, it requires the existence of an Edgeworth expansion for its distribution (Hall, 2013), which is beyond the focus of this thesis. So I skip the theoretical development, and more details will be forthcoming in the next section.

Chapter 4

Numerical Results

We illustrate the use of the sieve estimator to evaluate performance of nonparametric components in semiparametric ZINB marginal mean model. Before applying to real data, simulation studies are conducted to demonstrate the importance of the nonparametric modeling. Then, the result from parametric ZIP and ZINB models motivates us to focus on semiparametric ZINB marginal mean model. Finally, I apply the sieve estimator to dental caries data in a semiparametric ZINB model and conduct Bootstrap hypothesis test for comparing the nonparametric components g_0 and g_1 that I am interested in.

4.1 Simulation Studies

In order to show the advantage of using a nonparametric component in a semiparametric ZINB model, I conduct Monte Carlo simulations.

We generate data from the following semiparametric model

$$\begin{aligned}\log\left[(1-\pi)\lambda\right] &= \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + (1-\Delta)g_0(S) + \Delta g_1(S)\\ \log\left(\frac{\pi}{1-\pi}\right) &= \beta_1 Z_1 + \beta_2 Z_2 + (1-\Delta)h_0(S) + \Delta h_1(S)\end{aligned}$$

where X_1, X_2, X_3, Z_1, Z_2, Δ and S are independently drawn from the binomial distribution $\mathcal{B}(1, 0.5)$, the uniform distribution on $[0, 2]$, the normal distribution $\mathcal{N}(1, 2)$, the uniform distribution on $[0,1]$, the normal distribution $\mathcal{N}(0,1)$, the binomial distribution $\mathcal{B}(1,0.5)$, and the uniform distribution on $[0,1]$, respectively, with the regression parameters $\alpha_1 = 0.5$, $\alpha_2 = 1$, $\alpha_3 = -0.5$, $\beta_1 = -1.5$, $\beta_2 = 0.5$, and the nonparametric components $g_0(S) = \sin(\pi S)$, $g_1(S) = 2S^2$, $h_0(S) = \sqrt{S}$, and $h_1(S) = \exp(-2S+1)$. In the main model (2.1), $\kappa = 1$ and Y is generated from $\mathcal{ZINB}(\lambda, \pi, \kappa)$.

The sample sizes $n = 500, 1000, 2000$ are chosen. To investigate whether nonparametric approach is appropriate in above semiparametric model, the following four working models are used to fit the data.

Model 1: all $g_0(S)$, $g_1(S)$, $h_0(S)$ and $h_1(S)$ are modeled nonparametrically and estimated by sieve estimator.

Model 2: all $g_0(S)$, $g_1(S)$, $h_0(S)$ and $h_1(S)$ are modeled linearly and estimated by classical approach.

Model 3: both $g_0(S)$ and $g_1(S)$ are modeled linearly and estimated by classical approach, and both $h_0(S)$ and $h_1(S)$ are modeled nonparametrically and estimated by sieve estimator.

Model 4: both $g_0(S)$ and $g_1(S)$ are modeled nonparametrically and estimated by sieve estimator, and both $h_0(S)$ and $h_1(S)$ are modeled linearly and estimated by classical approach.

When 3-degree (cubic) splines basis functions are used to approximate a nonparametric component in above models, I select the uniform knots on $[0,1]$. Assume the smoothing parameter r be 2, and both n and r determine the optimal convergence rate. The number of knots, m_0 and m_1, can be chosen by the optimal convergence rate (Zhang et al., 2010) or AIC (Lam et al., 2006; He et al., 2010). Monte Carlo sample size is set as 1000.

Table 4.1 presents the relative bias (RB), mean square error (MSE) and standard deviation (SE) of all parametric components in Model 1. Table 4.2 presents the integrated

Table 4.1: Estimates of finite dimensional parameters in working Model 1

	sample size $n = 500$			sample size $n = 1000$			sample size $n = 2000$		
Parameter	RB	MSE	SE	RB	MSE	SE	RB	MSE	SE
α_1	0.003	0.024	0.005	-0.004	0.012	0.004	0.003	0.006	0.002
α_2	0.001	0.019	0.004	-0.001	0.009	0.003	0.005	0.005	0.002
α_3	0.002	0.002	0.001	-0.005	0.001	0.001	-0.001	0.001	0.001
β_1	0.016	0.147	0.012	0.014	0.065	0.008	0.013	0.032	0.006
β_2	0.034	0.015	0.004	0.011	0.006	0.002	0.012	0.003	0.002
κ	-0.117	0.032	0.004	-0.059	0.014	0.003	-0.027	0.006	0.002

Table 4.2: Estimates of infinite dimensional parameters in working Model 1

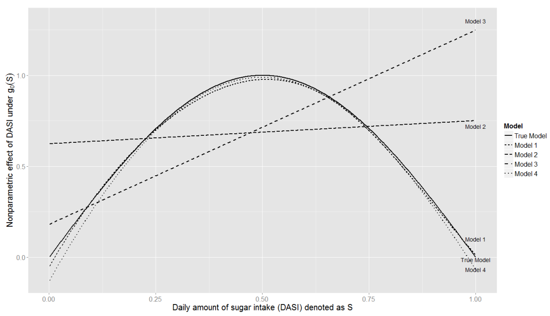
Sample sizes	$g_0(s)$	$g_1(s)$	$h_0(s)$	$h_1(s)$
500	0.048	0.087	0.046	0.046
1000	0.024	0.033	0.019	0.024
2000	0.019	0.019	0.013	0.014

MSE of all nonparametric components in Model 1. Figure 4.1 show that the estimations of nonparametric components in all models when sample size is 2000. The sieve estimators in Model 1 can capture the shapes of true functions reasonably while other models cannot. Both results demonstrates that the model using sieve approach for all nonparametric components is better than less restrictive ones.

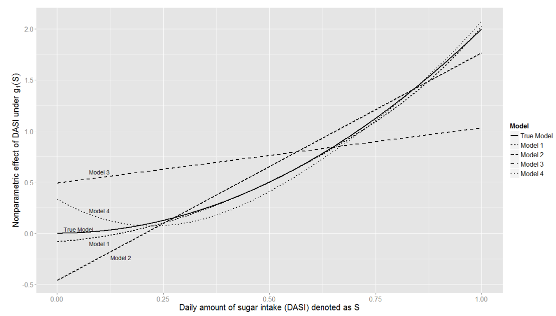
4.2 Real Data: Mouth-Level Indices in Caries Research

In order to evaluate dental caries severity in low-income African American families, a multilevel approach was designed and conducted in Detroit, Michigan (Tellez et al., 2006).

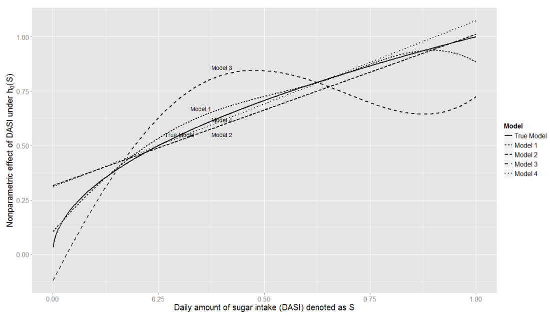
Focusing on dental caries data, this data set contains 874 children's oral health information. The covariates of interest include the standardized caregiver's oral hygiene index, denoted as X_1, Z_1 and ranged from -2.47 to 4.45; the child's age group, let Δ denote this



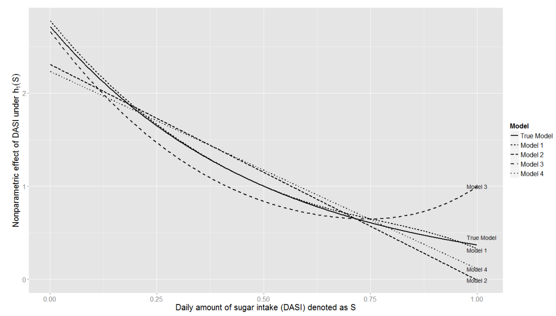
(a) The nonparametric effect of DASI under $g_0(S)$ in simulation study



(b) The nonparametric effect of DASI under $g_1(S)$ in simulation study



(c) The nonparametric effect of DASI under $h_0(S)$ in simulation study



(d) The nonparametric effect of DASI under $h_1(S)$ in simulation study

Figure 4.1: The estimates of nonparametric components in simulation study

binary indicator taking value 1 (392, 45%) if the child's age is less than 2 and taking value 0 (482, 55%) otherwise; the caregiver's employment status, let X_2 denote this binary indicator taking value 0 (344, 39%) if the caregiver has no job and taking value 1 (530, 61%) otherwise; the child's standardized sugar intake, denoted as S and ranged from -1.19 to 5.42.

Let Y denote the response variable, DMFS (number of decayed, missing and filled tooth surfaces) indices, representing the cumulative severity of tooth decay for each surveyed child. The histogram of Y, the negative binomial distribution and Poisson distribution fitted on the data are plotted in Figure 4.2. We encountered a large proportion of zero counts, and this situation motives us to consider zero-inflated model, like zero-inflated negative binomial model in which I define the zero counts are generated from two sources (Todem et al., 2012b; Cao et al., 2014). Therefore, I postulate that the distribution of Y is a zero-inflated negative

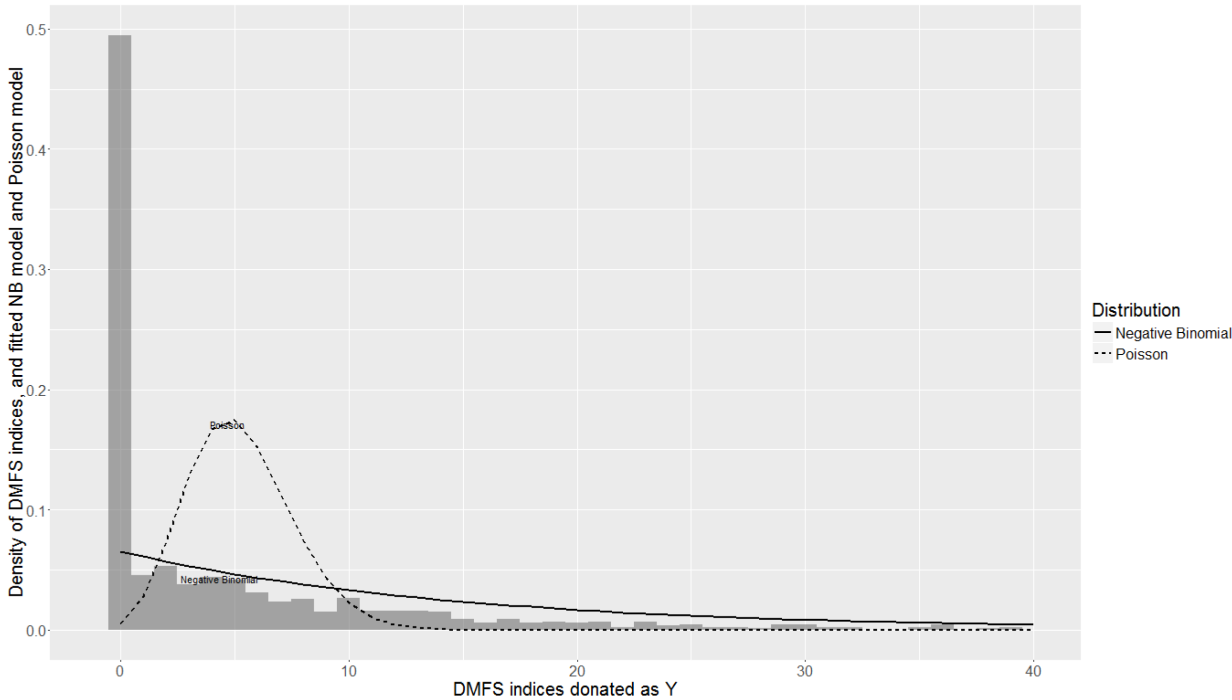


Figure 4.2: The histogram of DMFS indices and fitted NB model and Poisson model

binomial model with the probability of non-zero counts π and the mean of underlying negative binomial distribution λ related to covariates as follows,

$$\log(\pi\lambda) = \alpha_1 X_1 + \alpha_2 X_2 + (1-\Delta)g_0(S) + \Delta g_1(S) \tag{4.1}$$

$$\text{logit}(\pi) = \beta_1 Z_1 + (1-\Delta)h_0(S) + \Delta h_1(S)$$

where $\mathbb{E}(Y) = \pi\lambda$ stands for the marginal mean of ZINB model after adjusting by age of each child and the sugar intake effect S enters the ZINB model nonparametrically.

We use the uniform partition of $[-1.19, 5.42]$, assume the unknown functions be 1-th or 2-th ($r = 1, 2$) derivative in $[-1.19, 5.42]$, choose normalized uniform B splines basis functions of 2-degree ($l = 2$), and let the convergence rate be the optimal rate. The number of knots is chosen by AIC (2 or more knots are better). The estimates and standard errors are summarized in Table 4.3 and the estimates of g_0 and g_1 in the model with 3 knots are

Table 4.3: The estimates and standard deviations of finite dimensional parameters

Parameter	2 knots	3 knots	4 knots	6 knots	9 knots
α_1	0.145(0.050)	0.147(0.050)	0.147(0.050)	0.151(0.050)	0.155(0.050)
α_2	0.307(0.109)	0.310(0.110)	0.311(0.109)	0.310(0.109)	0.311(0.111)
β_1	0.578(0.215)	0.580(0.213)	0.509(0.213)	0.516(0.213)	0.559(0.217)
$\log(\kappa)$	-0.051(0.112)	-0.061(0.111)	-0.072(0.111)	-0.086(0.110)	-0.088(0.111)
AIC	13726.8	13737.1	13749	13770.4	13803.4

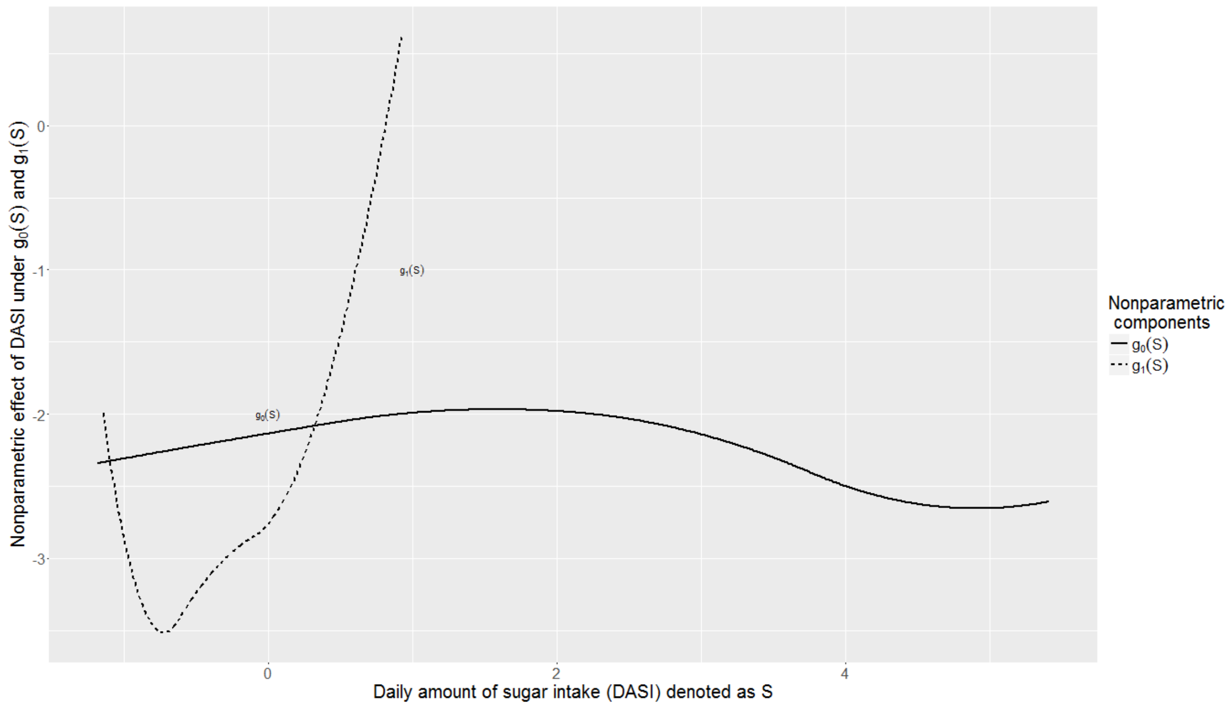


Figure 4.3: The nonparametric effects of DASI under $g_0(S)$ and $g_1(S)$ in real data study

plotted in Fig 4.3.

We are interested in the nonparametric effects of sugar intake on the marginal mean of responses, therefore, I focus on the nonparametric components $g_0(S)$ and $g_1(S)$ for $S \in [S_*, S^*]$, where S_* and S^* are common boundaries. According to the definition, $g_0(S)$ refers to the nonparametric effect of group in which child's age is larger than 2 (corresponding to $\Delta = 0$), and $g_1(S)$ refers to the nonparametric effect of group in which child's age is less than 2 (corresponding to $\Delta = 1$). Figure 4.3 shows the daily amount of sugar intake hardly influences on the marginal mean of responses with respect to $g_0(S)$, while intensively does impact on the marginal mean of responses with respect to $g_1(S)$. In other words, for children who are older than 2 years old, the daily amount of sugar intake almost does not impact on their marginal mean of DMFS indices, while for children who are younger than 2 years old, the daily amount of sugar intake does impact on their marginal mean of DMFS indices. It seems that only the marginal mean of DMFS indices of children aged less than 2 do depend on their daily amount of sugar intake, and there might be some factors rather than DASI will influence the marginal mean of DMFS indices of children aged larger than 2.

We are also interested in whether there is significant difference between nonparametric components $g_0(S)$ and $g_1(S)$, so I conduct a statistical test for the null hypothesis

$$H_0 : g_0(s) \equiv g_1(s), s \in [S_*, S^*].$$

Following the discussion in Section 3.2, I used two test statistics T_2 and T_∞ based on L_2 norm and L_∞ norm, respectively, given by

$$T_2^2 = \int_{s \in [S_*, S^*]} \|g_0(s) - g_1(s)\|_2^2 ds$$

$$T_\infty = \sup_{s \in [S_*, S^*]} |g_0(s) - g_1(s)|$$

In order to compute the p-value of both statistics, it requires us to conduct bootstrap approach to approximated the null distribution of these statistics. In other words, I need to re-sample the data with replacement under null hypothesis. Specifically, if the null hypothesis is true, the formula in (4.1) is equivalent to

$$\begin{aligned}\log(\pi\lambda) &= \alpha_1 X_1 + \alpha_2 X_2 + (1-\Delta)g_0(S) + \Delta g_1(S)\\ &= \alpha_1 X_1 + \alpha_2 X_2 + (1-\Delta)g_0(S) + \Delta g_0(S)\\ &= \alpha_1 X_1 + \alpha_2 X_2 + g_0(S)\end{aligned}$$

which implies $\log(\pi\lambda)$ is independent from variable Δ. Therefore, I set the bootstrap sample size as the same as the original sample size n, re-sample randomly n observations from data $\widetilde{W}$ with replacement, re-arrange each Δ_{boot} value for g_0 and g_1 (not for h_0 and h_1) in bootstrap samples by original proportions ($\Delta_{\text{boot}} = 1(45\%)$, $\Delta_{\text{boot}} = 0(55\%)$) rather than its original Δ value, for example, the Δ_{boot} of first $\lfloor n \times 45\% \rfloor$ observations in bootstrap samples are re-arranged as 1, then the rest of Δ_{boot} of observations in bootstrap samples are re-arranged as 0, where $\lfloor . \rfloor$ is the round function. The other variables in bootstrap samples remain the same as original ones.

We can conduct 1000 times bootstrap re-samplings. In each bootstrap re-sampling, I sample n observations from original data, re-arrange Δ_{boot} values as above, estimate $\hat{g}_0(S)$ and $\hat{g}_1(S)$ using sieve estimator, and eventually, compute $\hat{T}_{2,\text{boot}}$ and $\hat{T}_{\infty,\text{boot}}$ using Monte Carlo integration. Finally, although I do not know the sampling distributions of T_2 and T_∞ statistics, the bootstrap sample lets us estimate the p-values of both T_2 and T_∞ statistics by

$$\text{p-value}_2 = \frac{\#\left\{\hat{T}_{2,\text{boot}} > \hat{T}_{2,\text{original}}\right\}}{1000}$$

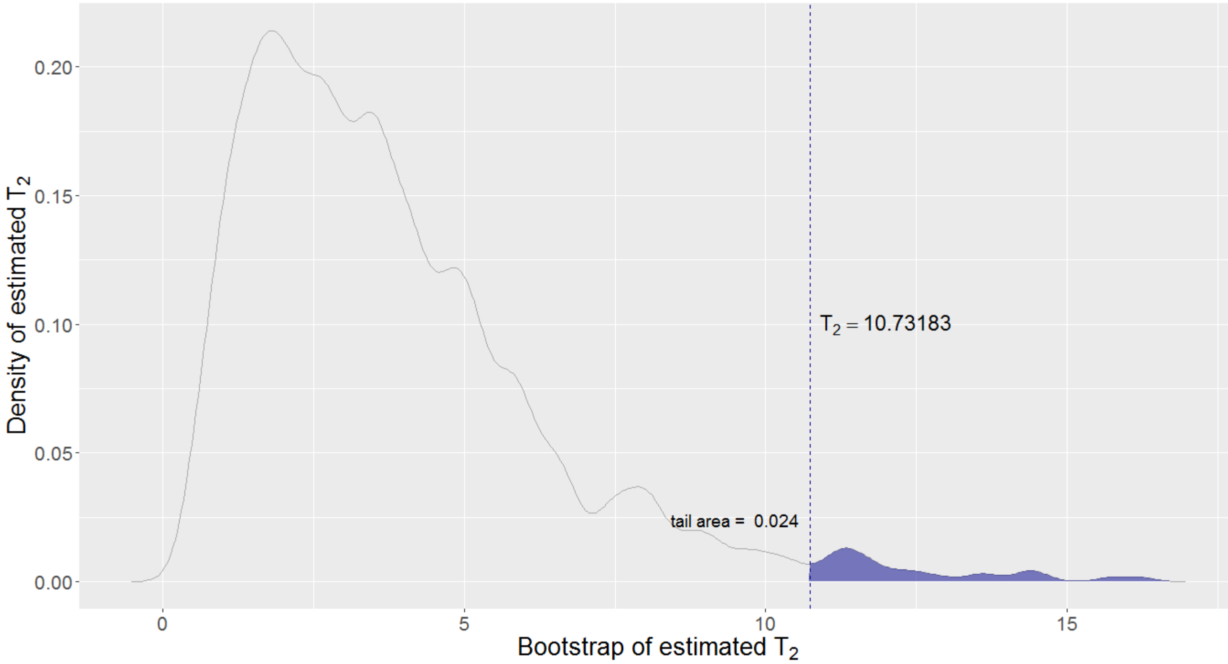
$$\text{p-value}_\infty = \frac{\#\left\{\hat{T}_{\infty,\text{boot}} > \hat{T}_{\infty,\text{original}}\right\}}{1000}$$

where $\#\{.\}$ is the count function, $\hat{T}_{2,\text{original}}$ and $\hat{T}_{\infty,\text{original}}$ are the estimated statistics by

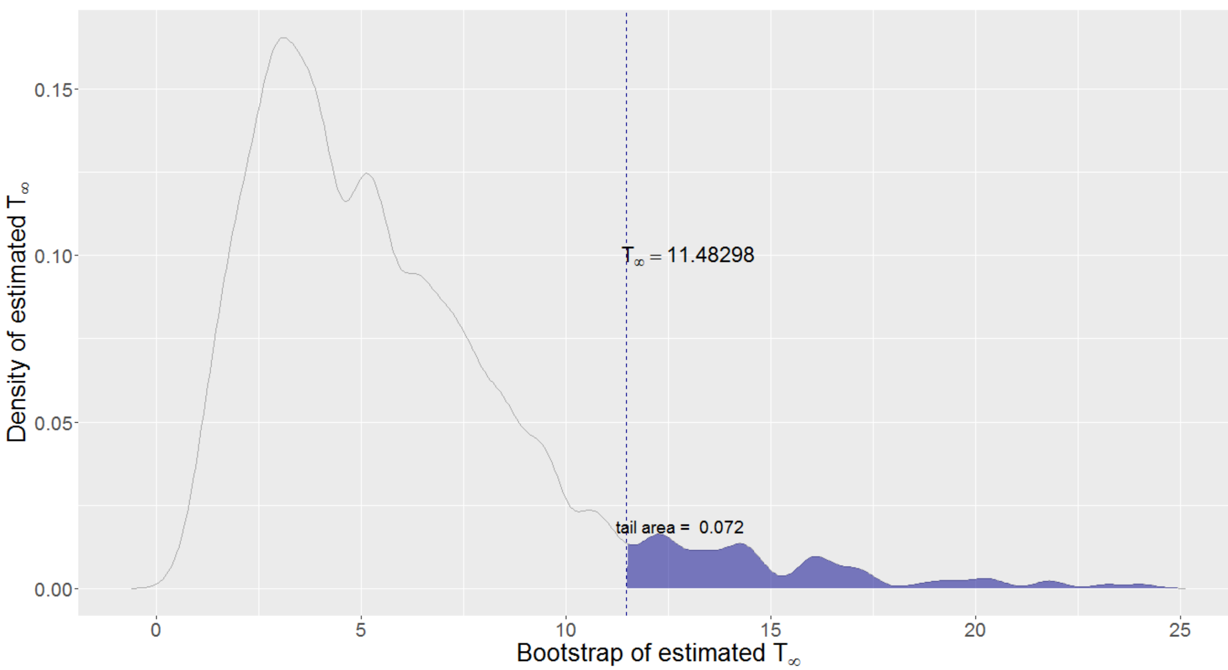
Table 4.4: The result of bootstrap sampling

Statistics	Observed value	#{ bootstrap > observed }	p-value
T_2	10.73183	24	0.024
T_∞	11.48298	72	0.072

original samples. Table 4.4 summaries above result, Figure 4.4 shows the estimated sampling distributions by bootstrap method.



(a) The histogram of estimated T_2



(b) The histogram of estimated T_∞

Figure 4.4: The histogram of both estimated statistics by bootstrap method

Chapter 5

Discussion

In this thesis, a semiparametric zero-inflated marginal mean model is proposed in section 2.1, and B splines based sieve estimator are used to estimate both the parametric components and nonparametric components. I also showed that sieve estimator for the vector of finite dimensional parametric components is strong consistent and asymptotically normally distributed and efficient, given by the specific efficient score function and the Fisher information matrix. Furthermore, a bootstrap hypothesis testing is introduced to any two of nonparametric components. The simulation studies demonstrated that the proposed model has highly satisfactory performance. Given appropriate tunning parameters like smoothing parameter r, growth rate k and the number of knots (without containing two ends) m_0 and m_1, estimation of the nonparametric component has shown to be highly satisfactory in that the overall shape of $g_0(S)$, $g_1(S)$, $h_0(S)$ and $h_1(S)$ can be captured reasonably well. The model can be extended by allowing a penalized likelihood for the smoothing parameters.

For caries research, the semiparametric zero-inflated marginal mean model is applied to the Detroit Dental Caries Study. This model does provide a reasonable representation of data from a homogeneous population, but it is unknown why some children from low-income families would be considered immune to dental caries based on the fact that they distribute as Dirac (pure zero) rather than negative binomial (Todem et al., 2016). Young children may have different oral health outcomes like DMFS due to socio-economic levels of families (Nanayakkara et al., 2013). Some factors are positively associated with child's oral health, like having a dental home, having caregivers with high education and living in a fluoridated community (Chi et al., 2013). In this caries research, child raised by unemployed parents experiences more caries according to estimates of parametric components, which confirms the conclusion of a systematic review (Kumar et al., 2015). According to nonparametric influence of the daily amount sugar intake, children belonging to differential membership of age group are suffered from dental caries differently. Model shows younger children have more pronounced influence than their older counterparts, and there exists a cutoff value (52.32 grams) of sugar intake, above which sugar intake is detrimental for younger children. It is also unclear why the cutoff value of sugar intake exists and why it exists only for younger children. Perhaps children will develop gradually to be "immune" to dental caries as they grow up. Based on these new findings, intervention strategies should inflect consideration of age group, and interventions targeting children aged below 2 are required to associate with the cutoff value of sugar intake.

APPENDIX

Definition

Definition 1 (Envelope). *Suppose $\mathscr{F}$ is a class of functions in $L_p(\mathbb{P})$, that is* $\mathscr{F} = \left\{ f : \int |f|^p d\mathbb{P} < \infty \right\}$. *Call each constant C such that $\|f\|_p \leq C$ for every f in $\mathscr{F}$, an envelope for $\mathscr{F}$.*

Definition 2 (Covering Number). *Suppose $\mathscr{F}$ is a class of functions in $L_p(\mathbb{P})$. For each $\epsilon > 0$ define the covering number $N\left(\epsilon, \mathscr{F}, L_p(\mathbb{P})\right)$ as the smallest number m for which there exist functions $g_1, g_2, \ldots, g_m$ such that $\min_{1 \leq j \leq m} \|f - g_j\|_p \leq \epsilon$ for each f in $\mathscr{F}$.*

Proof of Theorems

Lemma 1. *For any $f \in B_i$, there exists a function $f_n \in B_{n,i}$ and a constant C such that*

$$\sup_{0 \leq s \leq 1} |f_n(s) - f(s)| \leq C n^{-rk}$$

where the constant r and the constant k are known, and k depending on the partition.

Proof. Please refer to Theorem 12.7 (P491) in (Schumaker, 1981). ∎

Corollary 1. *Given the definition of B_i, I can choose the specific $B_{n,i}$ such that $\Theta_n \subseteq \Theta$ by modifying the specific $M_{i,2}$.*

Proof. Please refer to summary in Section 3.1. ∎

Corollary 2. *Under the Assumption **C4**, for any $\theta = (\alpha, \beta, \kappa, g_0, g_1, h_0, h_1) \in \Theta$, there exists*

$$\Phi_n \theta = (\alpha, \beta, \kappa, g_{n,0}, g_{n,1}, h_{n,0}, h_{n,1}) \in A_1 \times A_2 \times A_3 \times B_{n,1} \times B_{n,2} \times B_{n,3} \times B_{n,4}$$

such that $\rho(\theta, \Phi_n\theta) \to 0$ *as* $n \to \infty$.

Proof. For any $\theta \in \Theta$, the nonparametric components g_0, g_1, h_0 and h_1 belong to B_1, B_2, B_3 and B_4, respectively. Using Lemma 1, there exist $g_{n,0} \in B_{n,1}, g_{n,1} \in B_{n,2}, h_{n,0} \in B_{n,3}, h_{n,1} \in B_{n,4}$ and constants C_1, C_2, C_3, C_4 such that

$$\sup_{0\le s\le 1} \left|g_{n,0}(s) - g_0(s)\right| \le C_1 n^{-r_1 k_0}, \sup_{0\le s\le 1} \left|g_{n,1}(s) - g_1(s)\right| \le C_2 n^{-r_2 k_1}$$

$$\sup_{0\le s\le 1} \left|h_{n,0}(s) - h_0(s)\right| \le C_3 n^{-r_3 k_0}, \sup_{0\le s\le 1} \left|h_{n,1}(s) - h_1(s)\right| \le C_4 n^{-r_4 k_1}$$

where k_0 and k_1 are under the Assumption **C4**. Let $\frac{C}{2} = \max\left\{C_1, C_2, C_3, C_4\right\}$ $,r =$

$\min\{r_1, r_2, r_3, r_4\} \geq 1$, and $k = \min\{k_0, k_1\}$, and given W, then

$$
\begin{aligned}
\rho^2(\theta, \Phi_n\theta) &= \mathbb{E}_{\mathrm{T}}\left\{(\alpha-\alpha)^\top \mathbf{X} + (1-\Delta)(g_{n,0}-g_0) + \Delta(g_{n,1}-g_1)\right\}^2 \\
&+ \mathbb{E}_{\mathrm{T}}\left\{(\beta-\beta)^\top \mathbf{Z} + (1-\Delta)(h_{n,0}-h_0) + \Delta(h_{n,1}-h_1)\right\}^2 \\
&+ |\kappa-\kappa|^2 \\
&= \mathbb{E}_{\mathrm{T}}\left[(1-\Delta)(g_{n,0}-g_0) + \Delta(g_{n,1}-g_1)\right]^2 \\
&+ \mathbb{E}_{\mathrm{T}}\left[(1-\Delta)(h_{n,0}-h_0) + \Delta(h_{n,1}-h_1)\right]^2 \\
&= \mathbb{E}_{\mathrm{T}}\left\{[(1-\Delta)(g_{n,0}-g_0)]^2 + [\Delta(g_{n,1}-g_1)]^2\right\} \\
&+ \mathbb{E}_{\mathrm{T}}\left\{[(1-\Delta)(h_{n,0}-h_0)]^2 + [\Delta(h_{n,1}-h_1)]^2\right\} \\
&= \int \left\{[(1-\Delta)(g_{n,0}-g_0)]^2 + [\Delta(g_{n,1}-g_1)]^2\right\} \mathbb{P}_{\Delta,S}(\delta,s)d(\delta,s) \\
&+ \int \left\{[(1-\Delta)(h_{n,0}-h_0)]^2 + [\Delta(h_{n,1}-h_1)]^2\right\} \mathbb{P}_{\Delta,S}(\delta,s)d(\delta,s) \\
&= \int \left|g_{n,0}-g_0\right|^2 d\mathbb{P}_S + \int \left|g_{n,1}-g_1\right|^2 d\mathbb{P}_S + \int \left|h_{n,0}-h_0\right|^2 d\mathbb{P}_S \\
&+ \int \left|h_{n,1}-h_1\right|^2 d\mathbb{P}_S \\
&\leq \int \left\|g_{n,0}-g_0\right\|_\infty^2 d\mathbb{P}_S + \int \left\|g_{n,1}-g_1\right\|_\infty^2 d\mathbb{P}_S \\
&+ \int \left\|h_{n,0}-h_0\right\|_\infty^2 d\mathbb{P}_S + \int \left\|h_{n,1}-h_1\right\|_\infty^2 d\mathbb{P}_S \\
&\leq \left\|g_{n,0}-g_0\right\|_\infty^2 + \left\|g_{n,1}-g_1\right\|_\infty^2 + \left\|h_{n,0}-h_0\right\|_\infty^2 + \left\|h_{n,1}-h_1\right\|_\infty^2 \\
&\leq \left(C_1 n^{-r_1 k_0}\right)^2 + \left(C_2 n^{-r_2 k_1}\right)^2 + \left(C_3 n^{-r_3 k_0}\right)^2 + \left(C_4 n^{-r_4 k_1}\right)^2 \\
&\leq 4\left(\frac{C}{2} n^{-rk}\right)^2
\end{aligned}
$$

Therefore, $\rho(\theta, \Phi_n\theta) \leq Cn^{-rk} \to 0$ as $n \to \infty$. ■

Lemma 2. *Assume **A1** be held,* $\mathbb{E}_{\mathrm{T}}\left[(\boldsymbol{X} - \mathbb{E}_{\mathrm{T}}(\boldsymbol{X}|S))(\boldsymbol{X} - \mathbb{E}_{\mathrm{T}}(\boldsymbol{X}|S))^\top\right]$ *has a minimum* *eigenvalue* $\lambda_{x,\min}$*, and* $\mathbb{E}_{\mathrm{T}}\left[\left(\boldsymbol{Z}-\mathbb{E}_{\mathrm{T}}(\boldsymbol{Z}|S)\right)\left(\boldsymbol{Z}-\mathbb{E}_{\mathrm{T}}(\boldsymbol{Z}|S)\right)^{\top}\right]$ *has a minimum eigenvalue* $\lambda_{z,\min}$*. Then, for any* $\theta, \theta^* \in \Theta$*, I have*

$$\|\alpha-\alpha^*\| \leq \frac{1}{\sqrt{\lambda_{x,\min}}}\rho(\theta,\theta^*), \|\beta-\beta^*\| \leq \frac{1}{\sqrt{\lambda_{z,\min}}}\rho(\theta,\theta^*), |\kappa-\kappa^*| \leq \rho(\theta,\theta^*)$$
$$\|g_0-g_0^*\|_2 \leq \sqrt{2+\frac{2M_{1,1}^2}{\lambda_{x,\min}}}\rho(\theta,\theta^*), \|g_1-g_1^*\|_2 \leq \sqrt{2+\frac{2M_{1,1}^2}{\lambda_{x,\min}}}\rho(\theta,\theta^*)$$
$$\|h_0-h_0^*\|_2 \leq \sqrt{2+\frac{2M_{1,1}^2}{\lambda_{z,\min}}}\rho(\theta,\theta^*), \|h_1-h_1^*\|_2 \leq \sqrt{2+\frac{2M_{1,1}^2}{\lambda_{z,\min}}}\rho(\theta,\theta^*)$$

Proof. For any $\theta, \theta^* \in \Theta$,

$$\begin{aligned}
\rho^2(\theta,\theta^*) &= \mathbb{E}_{\mathrm{T}}\left\{(\alpha-\alpha^*)^{\top}\mathbf{X}+(1-\Delta)(g_0-g_0^*)+\Delta(g_1-g_1^*)\right\}^2 \\
&\quad+\mathbb{E}_{\mathrm{T}}\left\{(\beta-\beta^*)^{\top}\mathbf{Z}+(1-\Delta)(h_0-h_0^*)+\Delta(h_1-h_1^*)\right\}^2 \\
&\quad+|\kappa-\kappa^*|^2 \\
&\geq \mathbb{E}_{\mathrm{T}}\left\{(\alpha-\alpha^*)^{\top}\mathbf{X}+(1-\Delta)(g_0-g_0^*)+\Delta(g_1-g_1^*)\right\}^2 \\
&= \mathbb{E}_{\mathrm{T}}\Big\{(\alpha-\alpha^*)^{\top}\left(\mathbf{X}-\mathbb{E}_{\mathrm{T}}\left(\mathbf{X}|S\right)\right)+(\alpha-\alpha^*)^{\top}\mathbb{E}_{\mathrm{T}}\left(\mathbf{X}|S\right) \\
&\quad+(1-\Delta)(g_0-g_0^*)+\Delta(g_1-g_1^*)\Big\}^2
\end{aligned}$$

Let $J(S)$ denote $(\alpha-\alpha^*)^{\top}\mathbb{E}_{\mathrm{T}}\left(\mathbf{X}|S\right)+(1-\Delta)(g_0-g_0^*)+\Delta(g_1-g_1^*)$, I have

$$\begin{aligned}
\mathbb{E}_{\mathrm{T}}\left(\mathbf{X}-\mathbb{E}_{\mathrm{T}}\left(\mathbf{X}|S\right)\right)\times J(S) &= \mathbb{E}_{\mathrm{T}}\mathbf{X}J(S)-\mathbb{E}_{\mathrm{T}}\mathbb{E}_{\mathrm{T}}(\mathbf{X}|S)J(S) \\
&= \mathbb{E}_{\mathrm{T}}\mathbf{X}J(S)-\mathbb{E}_{\mathrm{T}}\mathbb{E}_{\mathrm{T}}(\mathbf{X}J(S)|S) \\
&= \mathbb{E}_{\mathrm{T}}\mathbf{X}J(S)-\mathbb{E}_{\mathrm{T}}\mathbf{X}J(S) \\
&= 0
\end{aligned}$$

So the interaction term is 0, I have

$$
\begin{aligned}
\rho^2(\theta, \theta^*) &\geq \mathbb{E}_{\mathrm{T}} \left\{ (\alpha - \alpha^*)^\top (\mathbf{X} - \mathbb{E}_{\mathrm{T}}(\mathbf{X}|S)) + J(S) \right\}^2 \\
&= \mathbb{E}_{\mathrm{T}} \left\{ (\alpha - \alpha^*)^\top (\mathbf{X} - \mathbb{E}_{\mathrm{T}}(\mathbf{X}|S)) \right\}^2 + J^2(S) \\
&= \mathbb{E}_{\mathrm{T}} \left\{ (\alpha - \alpha^*)^\top (\mathbf{X} - \mathbb{E}_{\mathrm{T}}(\mathbf{X}|S)) \right\}^2 \\
&+ \mathbb{E}_{\mathrm{T}} \left\{ (\alpha - \alpha^*)^\top \mathbb{E}_{\mathrm{T}}(\mathbf{X}|S) + (1-\Delta)(g_0 - g_0^*) + \Delta(g_1 - g_1^*) \right\}^2 \\
&\geq \mathbb{E}_{\mathrm{T}} \left\{ (\alpha - \alpha^*)^\top (\mathbf{X} - \mathbb{E}_{\mathrm{T}}(\mathbf{X}|S)) \right\}^2 \\
&= (\alpha - \alpha^*)^\top \mathbb{E}_{\mathrm{T}} \left[(\mathbf{X} - \mathbb{E}_{\mathrm{T}}(\mathbf{X}|S)) (\mathbf{X} - \mathbb{E}_{\mathrm{T}}(\mathbf{X}|S))^\top \right] (\alpha - \alpha^*) \\
&\geq \lambda_{x,\min} \|\alpha - \alpha^*\|^2
\end{aligned}
$$

Therefore, I have, $\|\alpha - \alpha^*\| \leq \frac{1}{\sqrt{\lambda_{x,\min}}} \rho(\theta, \theta^*)$, $\|\beta - \beta^*\| \leq \frac{1}{\sqrt{\lambda_{z,\min}}} \rho(\theta, \theta^*)$, and based on the definition of ρ, $|\kappa - \kappa^*| \leq \rho(\theta, \theta^*)$. Then focus on the nonparametric component,

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{T}} \left[(1-\Delta)(g_0 - g_0^*) + \Delta(g_1 - g_1^*) \right]^2 \\
&\quad= \mathbb{E}_{\mathrm{T}} \left[(\alpha - \alpha^*)^\top \mathbf{X} + (1-\Delta)(g_0 - g_0^*) + \Delta(g_1 - g_1^*) - (\alpha - \alpha^*)^\top \mathbf{X} \right]^2 \\
&\quad\leq 2\mathbb{E}_{\mathrm{T}} \left[(\alpha - \alpha^*)^\top \mathbf{x} + (1-\Delta)(g_0 - g_0^*) + \Delta(g_1 - g_1^*) \right]^2 + 2\mathbb{E}_{\mathrm{T}} \left[(\alpha - \alpha^*)^\top \mathbf{X} \right]^2 \\
&\quad\leq 2\rho^2(\theta, \theta^*) + 2M_{1,1}^2 \|\alpha - \alpha^*\|^2 \\
&\quad\leq \left(2 + \frac{2M_{1,1}^2}{\lambda_{x,\min}} \right) \rho^2(\theta, \theta^*)
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{T}}[(1-\Delta)(g_0-g_0^*)+\Delta(g_1-g_1^*)]^2 \\
&\quad=\int[(1-\Delta)(g_0-g_0^*)+\Delta(g_1-g_1^*)]^2\mathbb{P}_{\Delta,S}(\delta,s)d(\delta,s) \\
&\quad=\int(g_0-g_0^*)^2\mathbb{P}_S(s)d(s)+\int(g_1-g_1^*)^2\mathbb{P}_S(s)d(s) \\
&\quad=\|g_0-g_0^*\|_2^2+\|g_1-g_1^*\|_2^2
\end{aligned}
$$

Following the previous discussion, $\|g_0-g_0^*\|_2 \leq \sqrt{2+\frac{2M_{1,1}^2}{\lambda_{x,\min}}\rho(\theta,\theta^*)}$. Similarly, $\|g_1-g_1^*\|_2 \leq \sqrt{2+\frac{2M_{1,1}^2}{\lambda_{x,\min}}\rho(\theta,\theta^*)}$, $\|h_0-h_0^*\|_2 \leq \sqrt{2+\frac{2M_{2,1}^2}{\lambda_{z,\min}}\rho(\theta,\theta^*)}$, and $\|h_1-h_1^*\|_2 \leq \sqrt{2+\frac{2M_{2,1}^2}{\lambda_{z,\min}}\rho(\theta,\theta^*)}$.

■

Lemma 3. *Assume given $\theta,\theta^*\in\Theta_n$, hold the Assumptions XXX, then there exist a constant M_3 such that*

$$
\begin{aligned}
|l(\theta,w)-l(\theta^*,w)| \leq M_3(&\|\alpha-\alpha^*\|_2+\|\beta-\beta^*\|_2 \\
&+\|g_{n,0}-g_{n,0}^*\|_\infty+\|g_{n,1}-g_{n,1}^*\|_\infty+\|h_{n,0}-h_{n,0}^*\|_\infty+\|h_{n,1}-h_{n,1}^*\|_\infty)
\end{aligned}
$$

Proof. Let ξ and ζ denote the joint functions with respect to (θ,w), given by

$$
\begin{aligned}
\xi(\theta,w)&=\alpha^\top\mathbf{X}+(1-\Delta)g_0(S)+\Delta g_1(S) \\
\zeta(\theta,w)&=\beta^\top\mathbf{Z}+(1-\Delta)h_0(S)+\Delta h_1(S)
\end{aligned}
$$

and let Q_1 and Q_2 denote the joint functions with respect to (ξ, ζ, κ), given by

$$Q_1(\xi, \zeta, \kappa) = \mathbb{I}_{(y=0)} \log \left[\pi + (1-\pi)(1+\kappa\pi)^{-\frac{1}{\kappa}} \right]$$
$$Q_2(\xi, \zeta, \kappa) = \mathbb{I}_{(y>0)} \log \left[(1-\pi) \frac{\Gamma\left(y+\frac{1}{\kappa}\right)(\kappa\lambda)^y}{\Gamma\left(\frac{1}{\kappa}\right)\Gamma(y+1)\,(1+\kappa\lambda)^{y+\frac{1}{\kappa}}} \right]$$

Compute the difference,

$$l(\theta, w) - l(\theta^*, w) = Q_1(\xi, \zeta, \kappa) - Q_1(\xi^*, \zeta^*, \kappa^*) + Q_2(\xi, \zeta, \kappa) - Q_2(\xi^*, \zeta^*, \kappa^*)$$

Using Taylor's series expansion, there exists $(\xi^{**}, \zeta^{**}, \kappa^{**})$ satisfying that

$$\begin{aligned} Q_1(\xi, \zeta, \kappa) - Q_1(\xi^*, \zeta^*, \kappa^*) &= \frac{\partial}{\partial \xi} Q_1(\xi^{**}, \zeta^{**}, \kappa^{**})(\xi - \xi^*) \\ &+ \frac{\partial}{\partial \zeta} Q_1(\xi^{**}, \zeta^{**}, \kappa^{**})(\zeta - \zeta^*) \quad + \frac{\partial}{\partial k} Q_1(\xi^{**}, \zeta^{**}, \kappa^{**})(\kappa - \kappa^*) \end{aligned}$$

where $\frac{\partial}{\partial \xi} Q_1(\xi^{**}, \zeta^{**}, \kappa^{**})$, $\frac{\partial}{\partial \zeta} Q_1(\xi^{**}, \zeta^{**}, \kappa^{**})$ and $\frac{\partial}{\partial \kappa} Q_1(\xi^{**}, \zeta^{**}, \kappa^{**})$ are bounded under the Assumption **C2**, and without loss generality, let $M_{1,3}$ denote their common supremum. Therefore,

$$\begin{aligned} |Q_1(\xi, \zeta, \kappa) - Q_1(\xi^*, \zeta^*, \kappa^*)| &= |\frac{\partial}{\partial \xi} Q_1(\xi^{**}, \zeta^{**}, \kappa^{**})(\xi - \xi^*) + \frac{\partial}{\partial \zeta} Q_1(\xi^{**}, \zeta^{**}, \kappa^{**})(\zeta - \zeta^*) \\ &+ \frac{\partial}{\partial k} Q_1(\xi^{**}, \zeta^{**}, \kappa^{**})(\kappa - \kappa^*)| \\ &\leq \|\xi - \xi^*\| \left| \frac{\partial}{\partial \xi} Q_1(\xi^{**}, \zeta^{**}, \kappa^{**}) \right| + \|\zeta - \zeta^*\| \left| \frac{\partial}{\partial \zeta} Q_1(\xi^{**}, \zeta^{**}, \kappa^{**}) \right| \\ &+ |\kappa - \kappa^*| \left| \frac{\partial}{\partial \kappa} Q_1(\xi^{**}, \zeta^{**}, \kappa^{**}) \right| \\ &\leq M_{1,3} \left(\|\xi - \xi^*\| + \|\zeta - \zeta^*\| + |\kappa - \kappa^*| \right) \end{aligned}$$

Therefore, there exists $M^*_{1,3}$ such that

$$
\begin{aligned}
&|Q_1(\xi,\zeta,\kappa) - Q_1(\xi^*,\zeta^*,\kappa^*)| \\
&\leq M^*_{1,3}\Big(\|\alpha-\alpha^*\|_2 + \|\beta-\beta^*\|_2 + |\kappa-\kappa^*| + \|g_{n,0}-g^*_{n,0}\|_\infty + \|g_{n,1}-g^*_{n,1}\|_\infty \\
&+ \|h_{n,0}-h^*_{n,0}\|_\infty + \|h_{n,1}-h^*_{n,1}\|_\infty \Big)
\end{aligned}
$$

The similar operation is applied to $Q_2(\xi,\zeta,\kappa) - Q_2(\xi^*,\zeta^*,\kappa^*)$ and I have $M^*_{2,3}$, let $M_3 = \max\left\{M^*_{1,3}, M^*_{2,3}\right\}$, then,

$$
\begin{aligned}
|l(\theta,w) - l(\theta^*,w)| &= |Q_1(\xi,\zeta,\kappa) - Q_1(\xi^*,\zeta^*,\kappa^*) + Q_2(\xi,\zeta,\kappa) - Q_2(\xi^*,\zeta^*,\kappa^*)| \\
&\leq |Q_1(\xi,\zeta,\kappa) - Q_1(\xi^*,\zeta^*,\kappa^*)| + |Q_2(\xi,\zeta,\kappa) - Q_2(\xi^*,\zeta^*,\kappa^*)| \\
&\leq M^*_{1,3}(\|\alpha-\alpha^*\|_2 + \|\beta-\beta^*\|_2 + |\kappa-\kappa^*| \\
&+ \|g_{n,0}-g^*_{n,0}\|_\infty + \|g_{n,1}-g^*_{n,1}\|_\infty + \|h_{n,0}-h^*_{n,0}\|_\infty + \|h_{n,1}-h^*_{n,1}\|_\infty) \\
&+ M^*_{2,3}(\|\alpha-\alpha^*\|_2 + \|\beta-\beta^*\|_2 + |\kappa-\kappa^*| \\
&+ \|g_{n,0}-g^*_{n,0}\|_\infty + \|g_{n,1}-g^*_{n,1}\|_\infty + \|h_{n,0}-h^*_{n,0}\|_\infty + \|h_{n,1}-h^*_{n,1}\|_\infty) \\
&\leq M_3(\|\alpha-\alpha^*\|_2 + \|\beta-\beta^*\|_2 + |\kappa-\kappa^*| \\
&+ \|g_{n,0}-g^*_{n,0}\|_\infty + \|g_{n,1}-g^*_{n,1}\|_\infty + \|h_{n,0}-h^*_{n,0}\|_\infty + \|h_{n,1}-h^*_{n,1}\|_\infty)
\end{aligned}
$$

■

Lemma 4. *Denote two functional classes, $\mathscr{G} = \{g(.)\}$ and $\mathscr{F} = \{f(.)\}$, where function f satisfies the Lipschitz condtion that for any $f \in \mathscr{F}$, there exists a constant C such that $|f(g) - f(g^*)| \leq C|g - g^*|$ for any g and $g^* \in \mathscr{G}$. Then, for any probability measure $\mathbb{P}$, the* *covering numbers* $N(\epsilon, \mathscr{F}, L_2(\mathbb{P}))$ *and* $N(\epsilon, \mathscr{G}, L_2(\mathbb{P}))$ *of* $\mathscr{F}$ *and* $\mathscr{G}$ *have the property*

$$N(\epsilon, \mathscr{F}, L_2(\mathbb{P})) \leq N\left(\frac{\epsilon}{C}, \mathscr{G}, L_2(\mathbb{P})\right)$$

Proof. According to the definition, there exists a constant C such that $|f(s) - f(s^*)| \leq C|s - s^*|$ for any $f(s), f(s^*) \in \mathscr{F}$. For each $\epsilon > 0$, define $m = N\left(\frac{\epsilon}{C}, \mathscr{G}, L_2(\mathbb{P})\right)$ and then there exist $g_1, g_2, \ldots, g_m$ such that $\min_j \left\|g - g_j\right\|_{L_2(\mathbb{P})} \leq \frac{\epsilon}{C}$ for each g in $\mathscr{G}$. Given $g^* \in \{g_1, g_2, \ldots, g_m\}$, there exist $f \circ g^*$ satisfying

$$\begin{aligned}
\|f \circ g^* - f \circ g\|_{L_2(\mathbb{P})} &= \left(\int (f \circ g^* - f \circ g)^2 \, d\mathbb{P}\right)^{\frac{1}{2}} \\
&\leq \left(\int C^2 (g^* - g)^2 d\mathbb{P}\right)^{\frac{1}{2}} \\
&= C \|g^* - g\|_{L_2(\mathbb{P})} \\
&\leq C \frac{\epsilon}{C} = \epsilon
\end{aligned}$$

for any given $f \circ g \in \mathscr{F}$, which implies $f \circ g_1, f \circ g_2, \ldots, f \circ g_m$ can cover $\mathscr{F}$. Then, $N(\epsilon, \mathscr{F}, L_2(\mathbb{P})) \leq m = N\left(\frac{\epsilon}{C}, \mathscr{G}, L_2(\mathbb{P})\right)$ ■

Lemma 5. *The covering number of the class* $\Lambda_n = \left\{ l(\Phi_n\theta, .) \middle| \Phi_n\theta \in \Theta_n \right\}$ *satisfies*

$$N(\epsilon, \Lambda_n, L_\infty) \leq K \left(\frac{1}{\epsilon}\right)^{d_1 + d_2 + 2N_0 + 2N_1 + 1}$$

where K *is a constant.*

Proof. Let $g_{n,0} = \sum_{i=1}^{N_0} \tau_{i,n,1}\phi_i, g^*_{n,0} = \sum_{i=1}^{N_0} \tau^*_{i,n,1}\phi_i \in B_{n,1}$, then compute

$$\begin{aligned}
\sup_s \left| g_{n,0}(s) - g^*_{n,0}(s) \right| &= \sup_s \left| \sum_{i=1}^{N_0} \tau_{i,n,1}\phi_i(s) - \sum_{i=1}^{N_0} \tau^*_{i,n,1}\phi_i(s) \right| \\
&= \max_{1\le j\le N_0-1} \sup_s \left| \sum_{i=1}^{N_0} \tau_{i,n,1}\phi_i(s) - \sum_{i=1}^{N_0} \tau^*_{i,n,1}\phi_i(s) \right|_{[s_j,s_j+1]} \\
&= \max_{1\le j\le N_0-1} \sup_s \left| \sum_{i=j}^{j+l+1} \tau_{i,n,1}\phi_i(s) - \sum_{i=j}^{j+l+1} \tau^*_{i,n,1}\phi_i(s) \right|_{[s_j,s_j+1]} \\
&\le (l+1) \max_{1\le i\le N_0} \left\{ \left| \tau_{i,n,1} - \tau^*_{i,n,1} \right| \right\} \le (l+1) \left\| \tau_{n,1} - \tau^*_{n,1} \right\|_2
\end{aligned}$$

therefore, using Lemma 4, I have

$$N\left(\epsilon, B_{n,1}, L_\infty\right) \le N\left(\frac{\epsilon}{l+1}, \left\{ \tau_{n,1} : \max_{1\le i\le N_0} \left\{ \left|\tau_{i,n,1}\right| \le M_{1,2} \right\} \right\}, \|.\|_2 \right)$$

According to Lemma 4.1 of (Pollard, 1990),

$$\begin{aligned}
N\left(\epsilon, B_{n,1}, L_\infty\right) &\le N\left(\frac{\epsilon}{l+1}, \left\{ \tau_{n,1} : \max_{1\le i\le N_0} \left\{ \left|\tau_{i,n,1}\right| \right\} \right\} \le M_{1,2}, \|.\|_2 \right) \\
&\le D\left(\frac{\epsilon}{l+1}, \left\{ \tau_{n,1} : \max_{1\le i\le N_0} \left\{ \left|\tau_{i,n,1}\right| \right\} \right\} \le M_{1,2}, \|.\|_2 \right) \\
&\le \left(\frac{3 \times 2M_{1,2}}{\frac{\epsilon}{l+1}} \right)^{N_0} \\
&= \left[\frac{6(l+1)M_{1,2}}{\epsilon} \right]^{N_0}
\end{aligned}$$

where the packing number D (Pollard, 1990) such that $N(\epsilon, F) \leq D(\epsilon, F)$. Similarly, I have

$$N\left(\epsilon, B_{n,1}, L_\infty\right) \leq \left[\frac{6(l+1)M_{1,2}}{\epsilon}\right]^{N_0}, N\left(\epsilon, B_{n,2}, L_\infty\right) \leq \left[\frac{6(l+1)M_{2,2}}{\epsilon}\right]^{N_1}$$
$$N\left(\epsilon, B_{n,3}, L_\infty\right) \leq \left[\frac{6(l+1)M_{3,2}}{\epsilon}\right]^{N_0}, N\left(\epsilon, B_{n,4}, L_\infty\right) \leq \left[\frac{6(l+1)M_{4,2}}{\epsilon}\right]^{N_1}$$

Given a distance $\widetilde{d}$ on Θ_n by

$$\begin{aligned}\widetilde{d}(\theta, \theta^*) &= \|\alpha - \alpha^*\|_2 + \|\beta - \beta^*\|_2 + |\kappa - \kappa^*| \\ &\quad + \|g_{n,0} - g_{n,0}^*\|_\infty + \|g_{n,1} - g_{n,1}^*\|_\infty + \|h_{n,0} - h_{n,0}^*\|_\infty + \|h_{n,1} - h_{n,1}^*\|_\infty\end{aligned}$$

Then, applying Lemma 3 to Polland's section 5,

$$\begin{aligned}N\left(\epsilon, \Lambda_n, L_\infty\right) &\leq N\left(\frac{\epsilon}{M_3}, \Theta_n, \widetilde{d}\right) \\ &\leq N\left(\frac{\epsilon}{7M_3}, A_1, L_2\right) N\left(\frac{\epsilon}{7M_3}, A_2, L_2\right) N\left(\frac{\epsilon}{7M_3}, A_3, L_1\right) \\ &\times N\left(\frac{\epsilon}{7M_3}, B_{n,1}, L_\infty\right) N\left(\frac{\epsilon}{7M_3}, B_{n,2}, L_\infty\right) \\ &\times N\left(\frac{\epsilon}{7M_3}, B_{n,3}, L_\infty\right) N\left(\frac{\epsilon}{7M_3}, B_{n,4}, L_\infty\right) \\ &\leq \left(\frac{21M_3M_{1,1}^d}{\epsilon}\right)^{d_1} \left(\frac{21M_3M_{2,1}^d}{\epsilon}\right)^{d_2} \left(\frac{21M_3M_{3,1}^d}{\epsilon}\right) \\ &\times \left(\frac{42M_3(l+1)M_{1,2}}{\epsilon}\right)^{N_0} \left(\frac{42M_3(l+1)M_{2,2}}{\epsilon}\right)^{N_1} \\ &\times \left(\frac{42M_3(l+1)M_{3,2}}{\epsilon}\right)^{N_0} \left(\frac{42M_3(l+1)M_{4,2}}{\epsilon}\right)^{N_1} \\ &= K\left(\frac{1}{\epsilon}\right)^{d_1+d_2+2N_0+2N_1+1}\end{aligned}$$

where K is a constant. ■

Theorem 1 (Strong Consistency)**.** *Suppose the Assumptions **C1** - **C5** hold, then* $\rho\left(\hat{\theta}_n, \theta_{\mathrm{T}}\right)$ $\to 0$ *almost surely under* $\mathbb{P}_{\theta_{\mathrm{T}}}$*. Moreover, if the condition **A3** are satisfied, then*

$$\|\hat{\alpha}_n - \alpha_{\mathrm{T}}\| \to 0, \|\hat{\beta}_n - \beta_{\mathrm{T}}\| \to 0, |\hat{\kappa}_n - \kappa_{\mathrm{T}}| \to 0,$$

$$\|\hat{g}_{n,0} - g_{\mathrm{T},0}\|_2 \to 0, \|\hat{g}_{n,1} - g_{\mathrm{T},1}\|_2 \to 0,$$

$$\|\hat{h}_{n,0} - h_{\mathrm{T},0}\|_2 \to 0, \|\hat{h}_{n,1} - h_{\mathrm{T},1}\|_2 \to 0$$

almost surely under $\mathbb{P}_{\theta_{\mathrm{T}}}$*.*

Proof. Using Lemma 5 and following the arguments in (Xue et al., 2004). ■

Theorem 2 (Rate of Convergence)**.** *Suppose the Assumptions **C1** - **C5** hold, then*

$$\rho\left(\hat{\theta}_n, \theta_{\mathrm{T}}\right) = O_p\left(\max\left\{n^{-\frac{1-k}{2}}, n^{-rk}\right\}\right)$$

If select $k = \frac{1}{1+2r}$ *for* $r = 1, 2,$ *then* $\rho\left(\hat{\theta}_n, \theta_{\mathrm{T}}\right)$ *achieves the optimal nonparametric convergence rate* $O_p\left(n^{-\frac{r}{1+2r}}\right)$*.*

Proof. Because $L_n(\theta, \widetilde{W})$ is bounded, it suffices to verify the three conditions in theorem 1 in (Shen and Wong, 1994) hold true. Following the authors' notation, I check the conditions one by one:

(1) Similar to the arguments in (Xue et al., 2004), the Kullback-Leibler information is greater than the square of the Hellinger distance (Shen and Wong, 1994), then there exists

a constant C satisfying

$$
\begin{aligned}
&\inf_{\rho(\theta,\theta_{\mathrm{T}})\geq\epsilon,\theta\in\Theta_n} \mathbb{E}_{\mathrm{T}}\left[l(\theta_{\mathrm{T}},W)-l(\theta,W)\right] \\
\geq &\inf_{\rho(\theta,\theta_{\mathrm{T}})\geq\epsilon,\theta\in\Theta_n} C\rho^2(\theta,\theta_{\mathrm{T}}) \geq C\epsilon^2
\end{aligned}
$$

where holding the first condition with $\alpha = 1$ of Shen's.

(2) Combining Lemma 2 and Lemma 3, there exists a constant C satisfying

$$
\begin{aligned}
&\sup_{\rho(\theta,\theta_{\mathrm{T}})\leq\epsilon,\theta\in\Theta_n} \mathbb{V}\left[l(\theta_{\mathrm{T}},W)-l(\theta,W)\right] \\
\leq &\sup_{\rho(\theta,\theta_{\mathrm{T}})\leq\epsilon,\theta\in\Theta_n} \mathbb{E}_{\mathrm{T}}\left[l(\theta_{\mathrm{T}},W)-l(\theta,W)\right]^2 \\
\leq C &\sup_{\rho(\theta,\theta_{\mathrm{T}})\leq\epsilon,\theta\in\Theta_n} \rho^2(\theta,\theta_{\mathrm{T}}) \leq C\epsilon^2
\end{aligned}
$$

where holding the second condition with $\beta = 1$ of Shen's.

(3) Consider the entropy $H(\epsilon, \Lambda_n, \|.\|_\infty) = \log N(\epsilon, \Lambda_n, \|.\|_\infty)$, using Lemma 5, there exists constants C and μ satisfying

$$
\begin{aligned}
&H(\epsilon, \Lambda_n, \|.\|_\infty) \\
\leq &\log K + (d_1 + d_2 + 2N_0 + 2N_1 + 1)\log\frac{1}{\epsilon} \\
\leq &Cn^k \log\frac{1}{\epsilon}
\end{aligned}
$$

where holding the third condition with $2r_0 = k, r = 0^+$ of Shen's. ■

REFERENCES

REFERENCES

N Atienza, J García-Heras, JM Muñoz-Pichardo, and R Villa. An application of mixture distributions in modelization of length of hospital stay. *Statistics in Medicine*, 27(9): 1403–1420, 2008.

CS Blaum, BE Fries, and MA Fiatarone. Factors associated with low body mass index and weight loss in nursing home residents. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 50(3):162–168, 1995.

G Cao, WW Hsu, and D Todem. A score-type test for heterogeneity in zero-inflated models in a stratified population. *Statistics in medicine*, 33(12):2103–2114, 2014.

DP Cappelli and CC Mobley. *Prevention in clinical oral health care*. Elsevier Health Sciences, 2007.

H Chen. Convergence rates for parametric components in a partly linear model. *The Annals of Statistics*, 16(1):136–146, 1988.

Donald L Chi, Katharine C Rossitch, and Elizabeth M Beeles. Developmental delays and dental caries in low-income preschoolers in the usa: a pilot cross-sectional study and preliminary explanatory model. *BMC oral health*, 13(1):1, 2013.

C De Boor. On calculating with b-splines. *Journal of Approximation Theory*, 6(1):50–62, 1972.

BA Dye, S Tan, V Smith, BG Lewis, LK Barker, G Thornton-Evans, PI Eke, ED Beltrán-Aguilar, AM Horowitz, and CH Li. Trends in oral health status: United states, 1988-1994 and 1999-2004. *Vital and health statistics. Series 11, Data from the national health survey*, (248):1–92, 2007.

VT Farewell and DA Sprott. The use of a mixture model in the analysis of count data. *Biometrics*, pages 1191–1194, 1988.

Stuart Geman and Chii-Ruey Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, pages 401–414, 1982.

D Ghosh. Efficiency considerations in the additive hazards model with current status data. *Statistica Neerlandica*, 55(3):367–376, 2001.

MS Gilthorpe, M Frydenberg, Y Cheng, and V Baelum. Modelling count data with excessive zeros: The need for class prediction in zero-inflated models and the issue of data generation in choosing between zero-inflated and generic mixture models for dental caries data. *Statistics in Medicine*, 28(28):3539–3553, 2009.

P Hall. *The bootstrap and Edgeworth expansion*. Springer Science & Business Media, 2013.

W Härdle and H Liang. *Partially linear models*. Springer, 2007.

X He, H Xue, and NZ Shi. Sieve maximum likelihood estimation for doubly semiparametric zero-inflated poisson models. *Journal of Mutivariate Analysis*, 101:2026–2038, 2010.

B Hu, L Li, X Wang, and T Greene. Nonparametric multistate representations of survival and longitudinal data with measurement error. *Statistics in medicine*, 31(21):2303–2317, 2012.

J Huang. Efficient estimation of the partly linear additive cox model. *The annals of Statistics*, 27(5):1536–1563, 1999.

J Huang and AJ Rossini. Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *Journal of the American Statistical Association*, 92(439): 960–967, 1997.

K Hur, D Hedeker, W Henderson, S Khuri, and J Daley. Modeling clustered count data with excess zeros in health care outcomes research. *Health Services and Outcomes Research Methodology*, 3(1):5–20, 2002.

H Klein and CE Palmer. Studies on dental caries: V. familial resemblance in the caries experience of siblings. *Public Health Reports (1896-1970)*, pages 1353–1364, 1938.

Santhosh Kumar, Jyothi Tadakamadla, Jeroen Kroon, and Newell W Johnson. Impact of parent-related factors on dental caries in the permanent dentition of 6-to 12-year-old children: a systematic review. *Journal of dentistry*, 2015.

KF Lam, H Xue, and YB Cheung. Semiparametric analysis of zero-inflated count data. *Biometrics*, 62:996–1003, 2006.

D Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

AH Lee, M Gracey, K Wang, and KKW Yau. A robustified modeling approach to analyze pediatric length of stay. *Annals of Epidemiology*, 15(9):673–677, 2005.

T Loeys, B Moerkerke, O De Smet, and A Buysse. The analysis of zero-inflated count data: Beyond zero-inflated poisson regression. *British Journal of Mathematical and Statistical Psychology*, 65(1):163–180, 2012.

S Ma. Cure model with current status data. *Statistica Sinica*, 19(1):233–233, 2009.

M Minami, CE Lennert-Cody, W Gao, and M Romn-Verdesoto. Modeling shark bycatch: The zero-inflated negative binomial regression model with smoothing. *Fisheries Research*, 84(2):210–221, 2007.

J Mullahy. Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365, 1986.

Vajira Nanayakkara, André Renzaho, Brian Oldenburg, and Lilani Ekanayake. Ethnic and socio-economic disparities in oral health outcomes and quality of life among sri lankan preschoolers: a cross-sectional study. *International journal for equity in health*, 12(1):1–9, 2013.

D Pollard. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages 1–86. JSTOR, 1990.

M Ridout, J Hinde, and C Demetrio. A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, 57:219–223, 2001.

CS Ritchie, K Joshipura, RA Silliman, B Miller, and CW Douglas. Oral health problems and significant weight loss among community-dwelling older adults. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 55(7):366–371, 2000.

P Sasieni. Non-orthogonal projections and their application to calculating the information in a partly linear cox model. *Scandinavian Journal of Statistics*, pages 215–233, 1992.

L Schumaker. *Spline functions: basic theory*. A Wiley Interscience Publication, 1981.

X Shen. On methods of sieves and penalization. *The Annals of Statistics*, pages 2555–2591, 1997.

X Shen and J Shi. Likelihood ratio inference on parametric space. *Science in China Series A Mathematics*, 34(5):584–594, 2004.

X Shen and W Wong. Convergence rate of sieve estimates. *The Annals of Statistics*, pages 580–615, 1994.

G Shmueli, TP Minka, JB Kadane, S Borle, and P Boatwright. A useful distribution for fitting discrete data: revival of the conway–maxwell–poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):127–142, 2005.

CH Singh and L Ladusingh. Inpatient length of stay: a finite mixture modeling analysis. *The European Journal of Health Economics*, 11(2):119–126, 2010.

L Song and H Xue. Asymptotic properties of a kind of sieve mle. *The Annals of Statistics*, 20(3):370–377, 2000.

P Speckman. Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 413–436, 1988.

DH Sullivan, W Martin, N Flaxman, and JE Hagen. Oral health problems and involuntary weight loss in a population of frail elderly. *Journal of the American geriatrics Society*, 41 (7):725–731, 1993.

M Tellez, W Sohn, BA Burt, and AI Ismail. Assessment of the relationship between neighborhood characteristics and dental caries severity among low-income african-americans: A multilevel approach. *Journal of public health dentistry*, 66(1):30–36, 2006.

D Todem. *Statistical models for dental caries data.* INTECH Open Access Publisher, 2012a.

D Todem, WW H, and KM Kim. On the efficiency of score tests for homogeneity in two-component parametric models for discrete data. *Biometrics*, 68(3):975–982, 2012b.

David Todem, KyungMann Kim, and Wei-Wen Hsu. Marginal mean models for zero-inflated count data. *Biometrics*, 2016.

Z Wang, S Ma, and CY Wang. Variable selection for zero-inflated and overdispersed data with application to health care demand in germany. *Biometrical Journal*, 57(5):867–884, 2015.

H Xue, KF Lam, and G Li. Sieve maximum likelihood estimator for semiparametric regression models with current status data. *Journal of the American Statistical Association*, 99 (466):346–356, 2004.

KKW Yau, K Wang, and AH Lee. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal*, 45(4):437–452, 2003.

Y Zhang, L Hua, and J Huang. A spline-based semiparametric maximum likelihood estimation method for the cox model with interval-censored data. *Scandinavian Journal of Statistics*, 37(2):338–354, 2010.