Statistica Sinica Preprint No: SS-2024-0104	
Title	A Semiparametric Quantile Single-Index Model for Zero-Inflated and Overdispersed Outcomes
Manuscript ID	SS-2024-0104
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0104
Complete List of Authors	Zirui Wang and Tianying Wang
Corresponding Authors	Tianying Wang
E-mails	tianyingw0905@outlook.com

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

A Semiparametric Quantile Single-Index Model for Zero-Inflated Outcomes

Zirui Wang

Department of Statistics and Data Science, Tsinghua University

Tianying Wang*

Department of Statistics, Colorado State University

Abstract: We consider the complex data modeling problem motivated by the zero-inflated and overdispersed data from microbiome studies. Analyzing how microbiome abundance is associated with human biological features, such as BMI, is of great importance for host health. Methods based on parametric distributional assumptions, such as zero-inflated Poisson and zero-inflated Negative Binomial regression, have been widely used in modeling such data, yet the parametric assumptions are restricted and hard to verify in real-world applications. We relax the parametric assumptions and propose a semiparametric single-index quantile regression model. It is flexible to include a wide range of possible association functions and adaptable to the various zero proportions across subjects, which relaxes the strong parametric distributional assumptions of most existing zero-inflated data modeling approaches. We establish the asymptotic properties for the index

Corresponding author: Tianying Wang, Department of Statistics, Colorado State University, Fort Collins, CO, 80523, USA. E-mail: Tianying.Wang@colostate.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

coefficients estimator and quantile regression curve estimation. Through extensive simulation studies, we demonstrate the superior performance of the proposed method regarding model fitting.

Key words and phrases: Quantile regression; single-index model; zero-inflation; microbiome count data; profile principle.

1. Introduction

The human microbiota consists of the microorganisms that reside in or on the human body and contribute essential functions to human beings (Cani, 2018). Human microbiome research studies the dynamic interactions among microbiomes, host, and environment (Xia and Sun, 2017). It is of great importance to build more accurate predictive models of taxa and identify the relationship between taxa and clinical parameters (Lloyd-Price et al., 2016). The main challenges in modeling microbiome data are zero inflation and overdispersion (Kaul et al., 2017). It is common that the proportion of zeros in gut microbiota counts can reach 70%-80% (Yatsunenko et al., 2012). Meanwhile, the non-zero counts of the microbiota counts could be as large as thousands and cause overdispersion (McMurdie and Holmes, 2014). The inflated zeros in microbiome data are commonly caused by two reasons: microbes are present in the environment but not detected

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

due to low sequencing depth and sampling variation, or some microbes may be incapable of living in the environment and truly never represented (Zeng et al., 2022). While modeling microbiota counts and testing their relationship with covariates of interest (e.g., lifestyle and disease status), one needs to carefully address the zero inflation and overdispersion challenges in statistical analysis (Zhang et al., 2017; Xia, 2020).

A common strategy to model zero-inflated data is two-part models, which impose a point probability mass at zero and model the positive count data by a parametric distribution, such as zero-inflated Poisson regression, zero-inflated Negative Binomial regression, hurdle models, and many others (Lambert, 1992; Chen and Li, 2016; Jiang et al., 2022). However, those approaches impose strong parametric assumptions, which may be violated in real-world applications and cause problems in downstream analysis (Silverman et al., 2020). Further, most of the aforementioned approaches fail to model the relationship between the proportion of zeros and covariates, and thus, they only capture the effect of covariates partially on the distribution of outcomes (Ling et al., 2022).

Contrary to parametric modeling, quantile regression (Koenker and Bassett, 1978) is a powerful and robust tool to model heterogeneous associations in complex data without any parametric distribution assumptions.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

Further, quantile regression enjoys the merits of flexibly linking covariates to the distribution of response without parametric assumptions and allowing different associations across quantile levels. However, classic quantile regression cannot be directly applied in microbiome data as it assumes a constant probability of observing a positive outcome for all individuals, which is unlikely to hold when the degree of zero inflation varies across subjects. To overcome this challenge, Ling et al. (2022) proposed a zero-inflated linear quantile regression model (denoted as "ZIQ-linear") to relax the parametric distribution assumptions on positive outcomes and applied this method in a carotid plaque data analysis. However, it does not consider either the overdispersion issue or the non-linear relationship between the quantile of microbiome data and the covariates of interest.

We present a motivating example from the gut microbiota count data (De la Cuesta-Zuluaga et al., 2018), which is later analyzed in the real data application. This dataset contains microbiome counts of over 6000 taxa for 411 adults and covariates related to diet, obesity, and cardiometabolic diseases. For illustrative purposes, we present the model-fitting results for one taxon *Clostridiales* with health-related covariates, such as anthropometric measures, glucose metabolism, and blood pressure. A full list of covariates is provided in Section 4. The library size, the sum of all the taxa counts per

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

subject, is adjusted as a covariate in the models below. We compared the observed and predicted count data generated by fitted models from ZIP, ZINB, and ZIQ-linear, respectively. In Figure 1, we observe that ZIP and ZINB fit the data poorly because the parametric assumptions could be violated, and the mass probability imposed on zero is a shared parameter for all subjects in these two approaches rather than modeling different degrees of zeros across subjects. ZIQ-linear models the zero and positive parts well because these two parts are both linked to individual-specific covariates. However, a small proportion of fitted values are negative, against the nature of microbiome counts. The primary reason is that the linear quantile regression model is not flexible enough for overdispersed data.

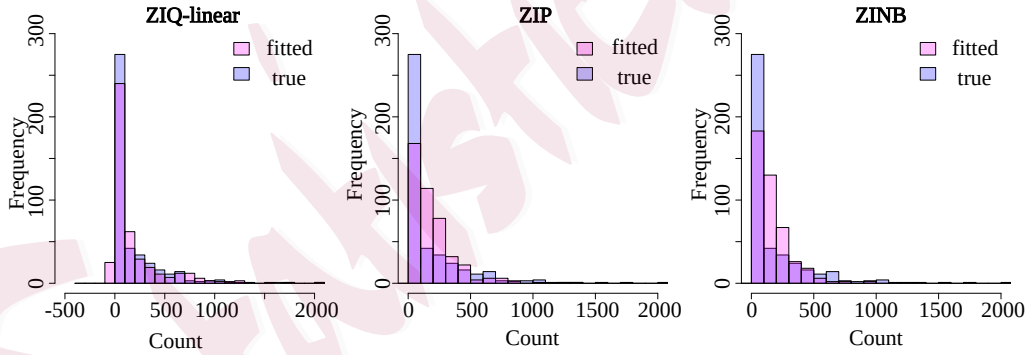


Figure 1: Model fitting results for the taxon *Clostridiales*.

Motivated by the study of zero-inflated and overdispersed outcomes, we consider quantile single-index models to overcome the limitations of the

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

linear quantile regression model while maintaining its robustness and flexibility. Single-index models have been widely used in literature for their merits of handling high-dimensional data while providing interpretable results (Radchenko, 2015; Neykov et al., 2016). Spline-based methods are often preferred for their easy implementation and derivable asymptotic properties (Yu and Ruppert, 2002; Ma and Zhu, 2013). For quantile regression, Ma and He (2016) developed statistical inference for single-index quantile regression models based on the pseudo-profile likelihood approach, yet it cannot be directly applied to the zero-inflated data. Without two-part modeling, the direct quantile single-index models assume a constant chance of observing a positive outcome and ignore the various degrees of zero inflation across subjects.

To this end, we propose a novel two-part modeling approach: the Zero-Inflated Quantile Single-Index model (ZIQSI). The positive part $Y > 0$ is modeled by a quantile single-index model in a semiparametric fashion, which is flexible and general, including a wide range of association functions. Compared to a fully nonparametric model, the proposed method is more interpretable through the index parameter. The probability of being zero (i.e., $P(Y = 0)$) is also linked to covariates, making our approach more adaptable to various zeros across subjects. Our contributions are three-fold.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

Methodologically, we provide a flexible modeling approach of zero-inflated and overdispersed outcomes with less restricted model specifications. The estimation is proceeded by the profile likelihood approach. Theoretically, we derived asymptotic properties for the estimated quantile coefficients, quantile curves, and average quantile effects. Application-wise, we provided a concrete analysis of microbiome data and evaluated the goodness of fit for the proposed method from different perspectives, illustrating its superiority in both distribution-wise modeling and individual-wise coefficient estimation, which could further contribute to personalized medicine.

2. Methods

2.1 Notations and model

Denote Y as a non-negative zero-inflated response variable and $\mathbf{x} = (x_1, ..., x_p)^\top$ be a set of covariates of our interest. Denote the τth quantile of Y as $Q_Y(\tau \mid \mathbf{x})$. To model the distribution of Y, we first decompose the conditional distribution of the zero-inflated outcome Y into the zero part and the positive part: $F(Y \mid \mathbf{x}) = P(Y = 0 \mid \mathbf{x}) + P(Y > 0 \mid \mathbf{x})F(Y \mid \mathbf{x}, Y > 0)$. Then, following the common two-part modeling strategy, we model the two parts, namely $P(Y = 0 \mid \mathbf{x})$ and $F(Y \mid \mathbf{x}, Y > 0)$, separately. We first assume that $P(Y > 0 \mid \mathbf{x})$, the conditional probability of observing a positive

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

Y, follows a logistic regression model,

$$\text{logit}\{P(Y > 0 \mid \mathbf{x})\} = \mathbf{x}^\top \gamma, \tag{2.1}$$

where γ is an unknown parameter. We consider the linear form for the logistic regression model since no compelling evidence suggests that a complicated semiparametric model is necessary for the classification, whereas our motivating example indicates the crucial need to consider a more flexible model for the positive response $Y > 0$. Thus, to ensure the generality of our method, we adopt a semi-parametric approach for the non-zero part $F(Y \mid \mathbf{x}, Y > 0)$. Given a nominal quantile level $\tau_s \in (0, 1)$, the conditional quantile function of Y given $Y > 0$ can be described by a single-index model:

$$Q_Y(\tau_s \mid \mathbf{x}, Y > 0) = G_{\tau_s}(\mathbf{x}^\top \beta_{\tau_s}), \tag{2.2}$$

where $G_{\tau_s}(\cdot)$ is an unknown function, and β_{τ_s} is an unknown parameter. The single-index model (eq (2.2)) is a popular dimensional reduction method for high-dimensional covariates $\mathbf{x}$ with extra flexibility at each quantile level τ_s through the unknown function $G_{\tau_s}(\cdot)$, which is essential for modeling the overdispersion in microbiome data. The method of Ling et al. (2022) can be viewed as a special case of our method, in which $G_{\tau_s}(\mathbf{x}^\top \beta_{\tau_s})$ is set as $\mathbf{x}^\top \beta_{\tau_s}$ for all $\tau_s \in (0, 1)$. To ensure the continuity of this two-part model,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

we assume that for any $\mathbf{x}$, $\lim_{\tau_s \to 0^+} Q_Y(\tau_s \mid \mathbf{x}, Y > 0) = 0$. Thus, when considering Models (2.1)-(2.2) together, the τth conditional quantile of Y given $\mathbf{x}$ can be written as:

$$Q_Y(\tau \mid \mathbf{x}) = I\left\{\tau > 1 - \pi(\mathbf{x}, \gamma)\right\} G_{\tau_s}\left(\mathbf{x}^\top \beta_{\tau_s}\right), \tag{2.3}$$

where $\pi(\mathbf{x}, \gamma) = P(Y > 0 \mid \mathbf{x})$; $I(\cdot)$ is an indicator function; and $\tau_s = \Gamma(\tau; \mathbf{x}, \gamma) = \max\left(\frac{\tau - \{1 - \pi(\gamma, \mathbf{x})\}}{\pi(\gamma, \mathbf{x})}, 0\right)$ maps the target quantile level τ linearly to the quantile level τ_s of $Y \mid Y > 0$ in Model (2.2). Due to the nonparametric nature of G_{τ_s}, we posit the assumptions for model identifiability.

Assumption 1

(1.1) The covariates $\boldsymbol{x}$ satisfies that $\boldsymbol{x} \in \mathcal{C}$, where $\mathcal{C}$ is a compact set.

(1.2) β_{τ_s} belongs to the parameter space $\Theta = \{\beta : \beta \in \mathbb{R}^p, \| \beta\|_2 = 1, \beta_1 \geq 0\}$ for identifiability. We assume p is fixed and shall not increase with n.

(1.3) Support of the function G_{τ_s} is $\left[\inf(\boldsymbol{x}^\top \beta), \sup(\boldsymbol{x}^\top \beta)\right]$, $\forall \boldsymbol{x} \in \mathcal{C}, \beta \in \Theta$.

These assumptions guarantee the identifiability of β_{τ_s} for quantile single-index models (Ma and He, 2016). Compared to Ling et al. (2022), our proposed two-part Zero-Inflated Quantile Single-index model allows more complex nonlinear associations between $\mathbf{x}$ and Y through the functions G_{τ_s}. Compared to other parametric two-part models, such as ZIP and ZINB, it is

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

robust against non-gaussian errors because we do not assume any particular error distributions.

2.2 Estimation

Suppose we have independent and identically distributed random samples $\{(\mathbf{x}_i, y_i); i = 1, 2, ..., n\}$ generated by the conditional quantile regression model (2.3). First, we estimate γ by logistic regression model (2.1):

$$\hat{\gamma}_n = \arg\max_{\gamma} \frac{1}{n} \sum_{i=1}^{n} \left[I(y_i > 0) \log \left\{ \frac{\pi(\gamma, \mathbf{x}_i)}{1 - \pi(\gamma, \mathbf{x}_i)} \right\} + \log\{1 - \pi(\gamma, \mathbf{x}_i)\} \right].$$

With the estimated coefficient $\hat{\gamma}_n$, given $\mathbf{x}, \tau$, we approximate τ_s by

$$\hat{\tau}_s = \Gamma(\tau; \mathbf{x}, \hat{\gamma}_n) = \max \left(\frac{\tau - (1 - \pi(\hat{\gamma}_n, \mathbf{x}))}{\pi(\hat{\gamma}_n, \mathbf{x})}, 0 \right). \tag{2.4}$$

For the quantile regression part, since $G_{\tau_s}(\cdot)$ is unknown, we approximate it by a linear combination of B-spline basis functions as in Wei and He (2006). We first introduce the B-spline basis for estimating the unknown function G_{τ_s}. Denote the total number of positive responses as $n_0 := \sum_{i=1}^{n} I(y_i > 0)$. We denote $a = t_0 < t_1 < ... < t_{N_{n_0}} < b = t_{N_{n_0}+1}$ as a partition of $[a, b]$, where the number of knots N_{n_0} increases with n_0. The partition satisfies $\max_{0 \le j \le N_{n_0}} |t_{j+1} - t_j| / \min_{0 \le j \le N_{n_0}} |t_{j+1} - t_j| \le M$ uniformly in the sample size of positive outcomes n_0 and for some constant $0 < M < \infty$. With m denoted as the order of polynomial splines,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

we denote the normalized B-spline basis of this space (De Boor, 2001), as $B(u) = \{B_j(u) : 1 \leq j \leq J_n\}^\top$, where $J_n = N_{n_0} + m$. In our empirical implementations, for each given β, we use the boundary points, namely $\min_{1\leq i\leq n} \mathbf{x}_i^\top \beta$ and $\max_{1\leq i\leq n} \mathbf{x}_i^\top \beta$, to generate the B-spline basis function $B(u)$. Further, by De Boor (2001), the single-index term $G_{\tau_s}(\mathbf{x}^\top \beta_{\tau_s})$ can be approximated by B-spline as $G_{\tau_s}\left(\mathbf{x}^\top \beta_{\tau_s}\right) \approx B\left(\mathbf{x}^\top \beta_{\tau_s}\right)^\top \theta(\tau_s)$ for some $\theta(\tau_s) \in \mathbb{R}^{J_n}$. Since the true value of τ_s is infeasible, we use its approximation $\hat{\tau}_s$ defined in eq (2.4) to obtain the estimators of the spline coefficients $\theta(\tau_s)$ and the parameter β_{τ_s} by minimizing the pseudo-likelihood function:

$$L_{\hat{\tau}_s,n}(\theta, \beta) = \frac{1}{n_0} \sum_{i=1}^{n} \rho_{\hat{\tau}_s} \left\{ y_i - B(\mathbf{x}_i^\top \beta)^\top \theta \right\} I(y_i > 0), \tag{2.5}$$

where $\rho_\tau(u) = u\,(\tau - I(u < 0))$ is the quantile loss function.

Here, we adopt the profile approach proposed in Ma and He (2016) to estimate β_{τ_s} and $\theta(\tau_s)$ owing to the stable performance showed in the empirical studies of Liang et al. (2010) and Ma and He (2016). We define the profile pseudo-likelihood function of β as

$$\begin{aligned} L^*_{\hat{\tau}_s,n}(\beta) &= \min_{\theta\in\mathbb{R}^{J_n}} L_{\hat{\tau}_s,n}(\beta, \theta) = L_{\hat{\tau}_s,n}\left(\beta, \tilde{\theta}_n(\beta, \hat{\tau}_s)\right) \\ &= \frac{1}{n_0} \sum_{i=1}^{n} \rho_{\hat{\tau}_s} \left\{ y_i - B(\mathbf{x}_i^\top \beta)^\top \tilde{\theta}_n(\beta, \hat{\tau}_s) \right\} I(y_i > 0), \end{aligned} \tag{2.6}$$

where $\tilde{\theta}_n(\beta, \hat{\tau}_s)$ is the minimizer of $L_{\hat{\tau}_s}(\theta, \beta)$ over $\theta \in \mathbb{R}^{J_n}$ for given $\beta \in \Theta$. Thus, the proposed profile likelihood estimation of $\beta \circ \Gamma(\tau; \mathbf{x}, \hat{\gamma}_n)$ is taken

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

to be:

$$\hat{\beta}_{\hat{\tau}_s} = \hat{\beta} \circ \Gamma(\tau; \mathbf{x}, \hat{\gamma}_n) = \arg\min_{\beta \in \Theta} L^*_{\hat{\tau}_s, n}(\beta).$$

Then, the spline estimator of $G_{\tau_s}(u)$ is $\widehat{G}_{\hat{\tau}_s}\left(u, \hat{\beta}_{\hat{\tau}_s}\right) = B(u)^\top \tilde{\theta}_n\left(\hat{\beta}_{\hat{\tau}_s}, \hat{\tau}_s\right)$, where $\tilde{\theta}_n\left(\hat{\beta}_{\hat{\tau}_s}, \hat{\tau}_s\right)$ minimizes $L^{**}_{\hat{\tau}_s, n}(\theta)$ over $\theta \in \mathbb{R}^{J_n}$, and $L^{**}_{\hat{\tau}_s, n}(\theta) = \frac{1}{n_0}\sum_{i=1}^n \rho_{\hat{\tau}_s}\left\{y_i - B(\mathbf{x}_i^\top \hat{\beta}_{\hat{\tau}_s})^\top \theta\right\} I(y_i > 0)$.

For a given $\beta \in \Theta$ and a specific τ_s, we denote:

$$\tilde{\tilde{\theta}}_n(\beta, \tau_s) = \arg\min_{\theta \in \mathbb{R}^{J_n}} \mathbb{E}\{L_{\tau_s, n}(\theta, \beta) \mid \mathbb{X}\}, \tag{2.7}$$

where $L_{\tau_s, n}(\theta, \beta)$ is the score function eq (2.5) and $\mathbb{X}$ are given covariates. We denote $\tilde{\tilde{G}}_{\tau_s}(u, \beta) = B^\top(u)\tilde{\tilde{\theta}}_n(\beta, \tau_s)$, which bridges the estimated $\widehat{G}_{\hat{\tau}_s}(\mathbf{x}^\top \hat{\beta}_{\hat{\tau}_s})$ and true $G_{\tau_s}(\mathbf{x}^\top \beta_{\tau_s})$. We also define

$$E^*(\mathbf{x} \mid \mathbf{x}^\top \beta_{\tau_s}) = \frac{\mathbb{E}\{f_{\epsilon_{\tau_s}}(0 \mid \mathbf{x})\mathbf{x} \mid \mathbf{x}^\top \beta_{\tau_s}\}}{\mathbb{E}\{f_{\epsilon_{\tau_s}}(0 \mid \mathbf{x}) \mid \mathbf{x}^\top \beta_{\tau_s}\}} \quad \text{and} \quad \tilde{\mathbf{x}} = \mathbf{x} - E^*(\mathbf{x} \mid \mathbf{x}^\top \beta_{\tau_s}), \tag{2.8}$$

where $f_{\epsilon_{\tau_s}}(\epsilon \mid \mathbf{x})$ denotes the conditional density of ϵ_{τ_s} given $\mathbf{x}$, and $\epsilon_{\tau_s} = Y - G_{\tau_s}(\mathbf{x}^\top \beta_{\tau_s})$ given $Y > 0$. $E^*(\mathbf{x} \mid \mathbf{x}^\top \beta_{\tau_s})$ and $\tilde{\mathbf{x}}$ are necessary for deducing the asymptotic distribution for the estimated coefficient $\hat{\beta}_{\hat{\tau}_s}$.

2.3 Construction of Quantile Curve and Average Quantile Effect

Given the aforementioned estimators $\hat{\gamma}_n$, $\hat{\beta}_{\hat{\tau}_s}$, and $\tilde{\theta}_n(\hat{\beta}_{\hat{\tau}_s}, \hat{\tau}_s)$, we construct the τth conditional quantile function $\widehat{Q}_Y(\tau \mid \mathbf{x})$ in three regions: (1) $R_{1,n} = \{\tau : 0 < \tau < 1 - \pi(\hat{\gamma}_n, \mathbf{x})\}$, (2) $R_{2,n} = \left\{\tau : 1 - \pi(\hat{\gamma}_n, \mathbf{x}) \le \tau \le 1 - \pi(\hat{\gamma}_n, \mathbf{x}) + n^{-\delta}\right\}$,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

and (3) $R_{3,n} = \left\{1 - \pi(\hat{\gamma}_n, \mathbf{x}) + n^{-\delta} \leq \tau \leq 1\right\}$, where $\delta < 0.5$ is a pre-specified interpolation parameter, and $\pi(\hat{\gamma}_n, \mathbf{x}) = \exp(\mathbf{x}^\top \hat{\gamma}_n)/\{1+\exp(\mathbf{x}^\top \hat{\gamma}_n)\}$ is the estimated probability of observing a positive Y given $\mathbf{x}$. Specifically, $R_{1,n}$ represents the region for a zero Y; $R_{3,n}$ represents the region of the positive part, in which the quantile curve is estimated on the nominal quantile level $\hat{\tau}_s = \Gamma(\tau; \mathbf{x}, \hat{\gamma}_n)$. $R_{2,n}$ is an interpolation region based on the nominal quantile level $\Gamma\left(1 - \pi(\hat{\gamma}_n, \mathbf{x}) + n^{-\delta}; \mathbf{x}, \gamma\right)$. The conditional density of Y given $Y > 0$ goes to zero when the quantile level approaches the change point, which can lead to a large variance if we estimate the quantile directly around the change point. The interpolation region is set for the stability and continuity of the estimated quantile function $\widehat{Q}_Y(\tau \mid \mathbf{x})$. Then, we construct $\widehat{Q}_Y(\tau \mid \mathbf{x})$ as below:

$$\begin{aligned}
\widehat{Q}_Y(\tau \mid \mathbf{x}) \;=\; & 0 \cdot I(\tau \in R_{1,n}) + B\left\{\mathbf{x}^\top \hat{\beta} \circ \Gamma\left(1 - \pi(\hat{\gamma}_n, \mathbf{x}) + n^{-\delta}; \mathbf{x}, \hat{\gamma}_n\right)\right\}^\top \\
& \tilde{\theta}_n\left\{\hat{\beta} \circ \Gamma\left(1 - \pi(\hat{\gamma}_n, \mathbf{x}) + n^{-\delta}; \mathbf{x}, \hat{\gamma}_n\right), \Gamma\left(1 - \pi(\hat{\gamma}_n, \mathbf{x}) + n^{-\delta}; \mathbf{x}, \hat{\gamma}_n\right)\right\} \\
& \cdot \frac{\tau - \{1 - \pi(\hat{\gamma}_n, \mathbf{x})\}}{n^{-\delta}} \cdot I(\tau \in R_{2,n}) + B\left\{\mathbf{x}^\top \hat{\beta} \circ \Gamma(\tau; \mathbf{x}, \hat{\gamma}_n)\right\}^\top \\
& \cdot \tilde{\theta}_n\{\hat{\beta} \circ \Gamma(\tau; \mathbf{x}, \hat{\gamma}_n), \Gamma(\tau; \mathbf{x}, \hat{\gamma}_n)\} \cdot I(\tau \in R_{3,n}). \qquad (2.9)
\end{aligned}$$

Based on $\widehat{Q}_Y(\tau \mid \mathbf{x})$, it is obvious that the covariates $\mathbf{x}$ can be associated with both the probability of observing a positive Y and also the quantile of $Y \mid Y > 0$. As our main focus is predicting quantile curves, it is common

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

that the predicted values are non-integral. One can round the estimation to the nearest integer upon request. The same applies to the following data simulation settings in Section 3.1.

To quantify the effect of the jth covariate (denoted as x_j) on the response Y, we define the model-based average quantile effect (AQE) for our two-part model as below:

$$\Delta_\tau(x_j; u, v) = \mathbb{E}_{\mathbf{x}^{(-j)}} \left\{ Q_Y(\tau \mid x_j = u, \mathbf{x}^{(-j)}) - Q_Y(\tau \mid x_j = v, \mathbf{x}^{(-j)}) \right\} \quad (2.10)$$

where $\mathbf{x}^{(-j)}$ denotes the covariates excluding x_j. AQE is served in an analogous fashion to the average treatment effect in linear models, and it has also been used in Ling et al. (2022). Thus, at a fixed quantile level τ, the importance of the covariate x_j can be estimated by integrating the difference between the conditional quantile of Y, given fixed $\mathbf{x}^{(-j)}$ and different levels of x_j. If x_j represents a continuous variable (e.g., BMI, cholesterol), we may select two levels according to clinical interest as u and v. For example, to assess the quantile effect of BMI, one can set $u \in [18.5, 24.9]$ for the normal group and $v \in [25, 29.9]$ for the overweight group (Weir and Jan, 2019). In particular, if x_j is binary (e.g., sex, treatment), the AQE is the average quantile treatment effect in the source population:

$$\Delta_\tau(x_j; 1, 0) = \mathbb{E}_{\mathbf{x}^{(-j)}} \left\{ Q_Y \left(\tau \mid x_j = 1, \mathbf{x}^{(-j)}\right) - Q_Y \left(\tau \mid x_j = 0, \mathbf{x}^{(-j)}\right) \right\} \quad (2.11)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

A natural sample estimator of eq (2.11) is

$$\widehat{\Delta}_\tau(x_j; u, v) = \frac{1}{n}\sum_{i=1}^{n} \widehat{Q}_Y(\tau \mid x_{i,j} = 1, \mathbf{x}_i^{(-j)}) - \widehat{Q}_Y(\tau \mid x_{i,j} = 0, \mathbf{x}_i^{(-j)}), \quad (2.12)$$

where $\widehat{Q}_Y(\cdot)$ is the estimated conditional quantile function defined in eq (2.9) and $(x_{i,j}, \mathbf{x}_i^{(-j)})$ denote the corresponding covariates of the ith sample. We provide the convergence rate of the AQE in Supplement S1.2.

2.4 Assumptions for asymptotic properties

We introduce some common assumptions on the distribution of zero-inflated data. We denote a_0 and b_0 to be the infimum and supremum of $\mathbf{x}^\top \beta_{\tau_s}$ over $\mathbf{x} \in \mathcal{C}$, where $\mathcal{C}$ is the compact set defined in Assumption 1 above.

Assumption 2

(2.1) Observations $\{(\boldsymbol{x}_i, y_i); i = 1, \cdots, n\}$ *are i.i.d. from a joint distribution* P*, where* $\boldsymbol{x}_i$ *is a* p*-dimensional vector of covariates.*

(2.2) The conditional density $f_Y(Y \mid \boldsymbol{x}, Y > 0)$ *of* Y *given* $X = \boldsymbol{x}$ *and* $Y > 0$ *satisfies the Lipschitz condition of order* 1 *and* $\sup_{\boldsymbol{x},y} f_Y(Y \mid \boldsymbol{x}, Y > 0) < \infty$.

(2.3) The conditional quantile function satisfies $\lim_{\tau \to 0^+} Q_Y(\tau \mid \boldsymbol{x}, Y > 0) = 0$.

(2.4) The quantile coefficient β_{τ_s} *is a differentiable function of* τ_s *with bounded first derivative, i.e.,* $\sup_{\tau_s \in (0,1)} \dot{\beta}_{\tau_s} = \sup_{\tau_s \in (0,1)} \left. \frac{d\beta_t}{dt} \right|_{t=\tau_s} < \infty$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

(2.5) $\forall \boldsymbol{x} \in \mathcal{C}$, *we have* $\|E(\boldsymbol{x}\boldsymbol{x}^\top)\|_\infty < \infty$.

(2.6) The density function of $\boldsymbol{x}^\top\beta$ *is bounded away from zero and infinity on its support, for* β *in a neighborhood of* β_{τ_s}.

Assumption (2.2) is borrowed from Ma and He (2016) to help establish the limiting distribution at the change point $\tau = 1 - \pi(\gamma, \mathbf{x})$. Assumption (2.3) is the continuity assumption stated in Section 2.1. Assumptions (2.4)−(2.5) and Assumption 3 below are necessary for establishing the asymptotic distribution of estimated coefficient $\hat{\beta}_{\hat{\tau}_s}$. With the asymptotic distribution of estimated coefficient $\hat{\beta}_{\hat{\tau}_s}$, Assumption (2.6) and Assumption 3, we can provide the convergence rate of $\widehat{Q}_Y(\tau \mid \mathbf{x})$ for any $\tau > 1 - \pi(\gamma, \mathbf{x})$. The limiting distribution of $\widehat{Q}_Y(\tau \mid \mathbf{x})$ at the change point $\tau = 1 - \pi(\gamma, \mathbf{x})$ is then proved based on the assumptions above and the asymptotic properties of $\widehat{Q}_Y(\tau \mid \mathbf{x})$ when $\tau > 1 - \pi(\gamma, \mathbf{x})$.

For $\widehat{Q}_Y(\tau \mid \mathbf{x})$ given $\tau > 1 - \pi(\gamma, \mathbf{x})$, since our proof concerns nonparametric smoothing literature, we first give some definitions and notations. Let $\mathcal{H}_r$ be the collection of all the functions on $[a_0, b_0]$ such that the mth order derivative satisfies the Hölder condition of order $r - m$, i.e. for each function $\phi \in \mathcal{H}_r$, there exists a constant C_0 s.t. $\left|\phi^{(m)}(u_1) - \phi^{(m)}(u_2)\right| \leq C_0|u_1 - u_2|^{r-m}$, for any $u_1, u_2 \in [a_0, b_0]$. This collection of functions is essential for proving the convergence rate of the spline estimator of G_{τ_s}.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

For given $\beta \in \Theta$ and τ, we denote: $\tilde{\tilde{\theta}}_n(\beta, \tau) = \arg\min_{\theta \in \mathbb{R}^{J_n}} \mathbb{E}\{L_{\tau,n}(\theta, \beta) \mid \mathbb{X}\}$, where $L_{\tau,n}(\theta, \beta)$ is the score function eq (2.5) and $\mathbb{X}$ are given covariates whose corresponding $Y > 0$. We denote $\tilde{\tilde{G}}_{\tau,n}(u, \beta) = B^\top(u)\tilde{\tilde{\theta}}_n(\beta, \tau)$, where $\tilde{\tilde{\theta}}_n(\beta, \tau)$ is from eq (2.7). Now we present assumptions for $\tilde{\tilde{\theta}}_n$ and $G_{\tau_s}(\cdot)$.

Assumption 3

(3.1) There exists $r > \frac{3}{2}$, such that for any $\tau_s \in (0, 1)$ we have $G_{\tau_s} \in \mathcal{H}_r$.

(3.2) There exists a constant $c_0 \in (0, +\infty)$, such that

$$\sup_{\mathbb{X}} \left\| \partial \tilde{\tilde{G}}_{\tau_s,n}(\boldsymbol{x}_i^\top \beta, \beta)/\partial\beta - \partial \tilde{\tilde{G}}_{\tau_s,n}(\boldsymbol{x}_i^\top \beta_{\tau_s}, \beta_{\tau_s})/\partial\beta \right\|_2 \leq c_0 \|\beta - \beta_{\tau_s}\|_2.$$

for any β in the neighborhood of β_{τ_s} and $\tau_s \in (0, 1)$.

(3.3) For fixed $\boldsymbol{x}$, assume $G_{\tau_s}(\boldsymbol{x}^\top \beta_{\tau_s})$ has limited first order derivative with respect to τ_s, i.e., $\sup_{\tau_s \in (0,1)} \left| \frac{\partial G_{\tau_s}(\boldsymbol{x}^\top \beta_{\tau_s})}{\partial \tau_s} \right| < \infty$.

(3.4) For any $\tau_s \in (0, 1)$, $E^*(\boldsymbol{x} \mid \boldsymbol{x}^\top \beta_{\tau_s} = u)$, which is a function of u, has a continuous and bounded first derivative.

Assumption (3.1) is commonly used in the nonparametric smoothing literature (Ma and He, 2016). Assumption (3.2) is a typical assumption in the regression literature, which can be easily satisfied when the dimension of covariates is fixed. Assumption (3.3) is used in the proof of the convergence rate for $\tau > 1 - \pi(\mathbf{x}; \gamma)$. Assumption (3.1)-(3.2) with Assumption 2 provides

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

the constraints of the asymptotic property of normal spline estimator provided in Ma and He (2016). Also, Assumption 3 together with Assumption (2.4)-(2.5) ensures that the following matrices exist and positive definite:

$$\begin{aligned}\Lambda_{1,\tau_s} &= \mathbb{E}[\pi(\gamma,\mathbf{x})f_{\epsilon_{\tau_s}}\{G_{\tau_s}(\mathbf{x}^\top\beta_{\tau_s}) \mid \mathbf{x}\}\{G^{(1)}_{\tau_s}(\mathbf{x}^\top\beta_{\tau_s})\tilde{\mathbf{x}}\}^{\otimes 2}],\\ \Omega_{\tau_s} &= \mathbb{E}[\{G^{(1)}_{\tau_s}(\mathbf{x}^\top\beta_{\tau_s})\tilde{\mathbf{x}}\}^{\otimes 2}], \quad D_{1,\gamma} = \mathbb{E}[\pi(\gamma,\mathbf{x})\{1-\pi(\gamma,\mathbf{x})\}\mathbf{x}\mathbf{x}^\top],\end{aligned}$$

where $\tilde{\mathbf{x}} = \mathbf{x} - E^*(\mathbf{x} \mid \mathbf{x}^\top\beta_{\tau_s})$, and $A^{\otimes 2} = AA^\top$, and $G^{(1)}_{\tau_s}(\cdot)$ is the first derivative of $G_{\tau_s}(\cdot)$. Here the matrices Ω_{τ_s} and Λ_{1,τ_s} are constructed to approximate the variance-covariance matrix for the parameter $\hat{\beta}_{\hat{\tau}_s}$.

2.5 Asymptotic properties for estimation

First, we provide the asymptotic normality for the individual estimated single-index coefficient $\hat{\beta}\circ\Gamma(\tau;\mathbf{x},\hat{\gamma}_n)$ using the property of B-spline estimator in Ma and He (2016) and the property of logistic regression. Denote the Moore-Penrose inverse of a matrix A as A^+.

Theorem 1 *Suppose $n\to\infty$ and $n_0/n\to b_0$ with $0<b_0<1$. Under the Assumptions 1-3, for all $\boldsymbol{x}\in\mathcal{C}$, when $\tau > 1-\pi(\gamma,\boldsymbol{x})$, we have:*

$$\sqrt{n}\left\{\hat{\beta}_{\hat{\tau}_s}-\beta_{\tau_s}\right\} = \sqrt{n}\left\{\hat{\beta}\circ\Gamma(\tau;\boldsymbol{x},\hat{\gamma}_n)-\beta\circ\Gamma(\tau;\boldsymbol{x},\gamma)\right\} \xrightarrow{d} N(0,\Sigma_1+\Sigma_2),$$

where $\Sigma_1 = b_0^{-1/2}\Gamma(\tau;\boldsymbol{x},\gamma)\{1-\Gamma(\tau;\boldsymbol{x},\gamma)\}\Lambda^+_{1,\Gamma(\tau;\boldsymbol{x},\gamma)}\Omega_{\Gamma(\tau;\boldsymbol{x},\gamma)}\Lambda^+_{1,\Gamma(\tau;\boldsymbol{x},\gamma)}$, $\Sigma_2 =$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

$\{1-\Gamma(\tau;\boldsymbol{x})\}^2\{1-\pi(\gamma,\boldsymbol{x})\}^2\boldsymbol{x}^\top D_{1,\gamma}^{-1}\boldsymbol{x}\boldsymbol{x}^\top\dot{\beta}\circ\Gamma(\tau;\boldsymbol{x},\gamma)\dot{\beta}\circ\Gamma(\tau;\boldsymbol{x},\gamma)^\top\boldsymbol{x}$, *and* $\dot{\beta}\circ\Gamma(\tau;\boldsymbol{x},\gamma)=\left.\dfrac{d\beta_\tau}{d\tau}\right|_{\Gamma(\tau;\boldsymbol{x},\gamma).}$

The covariance matrices, Σ_1 and Σ_2, are constructed using B-spline-based single-index quantile regression and logistic regression, respectively, and are then combined through the delta method. Both Λ_{1,τ_s} and Ω_{τ_s} are evaluated conditional on $Y>0$ and adjusted for the individual zero-inflation rate, $\pi(\gamma,\mathbf{x})$. That is, $\pi(\gamma,\mathbf{x})$ can be viewed as the propensity score to adjust for the covariance matrix since only the positive Y's are considered to fit the quantile regression model. Then, we construct the asymptotic consistency for $\widehat{Q}_Y(\tau\mid\mathbf{x})$ in Theorem 2.

Theorem 2 *Suppose* $n\to\infty$ *and* $n_0/n\to b_0$ *with* $0<b_0<1$. *Under the Assumptions 1-3, we have* $\widehat{Q}_Y(\tau\mid\boldsymbol{x})\xrightarrow{p}Q_Y(\tau\mid\boldsymbol{x})$.

Next, we provide the asymptotic properties for the limiting distribution of $\widehat{Q}_Y(\tau\mid\mathbf{x})$ in Theorem 3. For $\tau<1-\pi(\gamma,\mathbf{x})$, $\widehat{Q}_Y(\tau\mid\mathbf{x})$ converges to 0 super-efficiently due to the property of logistic regression. For the change point $\tau=1-\pi(\gamma,\mathbf{x})$, $\widehat{Q}_Y(\tau\mid\mathbf{x})$ has different convergence conditions based on the parameter δ. When $\tau>1-\pi(\gamma,\mathbf{x})$, the usage of the B-spline basis function makes it infeasible to establish the asymptotic distribution for $\widehat{Q}_Y(\tau\mid Y>0,\mathbf{x})$ as the number of knots N_{n_0} increases with n_0. Thus, we provide the global convergence rate for $\widehat{Q}_Y(\tau\mid Y>0,\mathbf{x})$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

Theorem 3 *Under the conditions of Theorem 1-2, given* $\boldsymbol{x}$ *and* τ*, we have the asymptotic convergence for the estimated quantile function as follows:*

(i) when $\tau < 1-\pi(\gamma,\boldsymbol{x})$*, we have* $\sqrt{n}\left\{\widehat{Q}_Y(\tau \mid \boldsymbol{x}) - 0\right\} \xrightarrow{p} 0$;

(ii) when $\tau = 1-\pi(\gamma,\boldsymbol{x})$*, we denote* $Q'_Y(0 \mid \boldsymbol{x}, Y>0)$ *as the right derivative and* $Z_0 \sim N(0,1)$*, then we have:*

(a) when $\delta = 0.25$*,*

$$\sqrt{n}\left\{\widehat{Q}_Y(\tau \mid Y>0,\boldsymbol{x}) - 0\right\} \xrightarrow{d} \{1-\pi(\gamma,\boldsymbol{x})\}\sqrt{\boldsymbol{x}^\top D_{1,\gamma}^{-1}\boldsymbol{x}}\, Q'_Y(0 \mid \boldsymbol{x}, Y>0) Z_0 I(Z_0>0);$$

(b) when $0.25 < \delta < 0.5$*,* $\widehat{Q}_Y(\tau \mid Y>0,\boldsymbol{x}) - 0 = O_P\left(J_n^{\frac{1}{2}} n^{-\frac{1}{2}} + J_n^{-r}\right)$;

(iii) when $\tau > 1-\pi(\gamma,\boldsymbol{x})$*, we have the global optimal convergence rate as*

$$\widehat{Q}_Y(\tau \mid Y>0,\boldsymbol{x}) - Q_Y(\tau \mid Y>0,\boldsymbol{x}) = O_P\left(J_n^{\frac{1}{2}} n^{-\frac{1}{2}} + J_n^{-r}\right),$$

i.e., $B\left(\boldsymbol{x}^\top \hat{\beta}_{\hat{\tau}_s}\right)^\top \tilde{\theta}_n\left(\hat{\beta}_{\hat{\tau}_s}, \hat{\tau}_s\right) - G_{\Gamma(\tau;\boldsymbol{x},\gamma)}\left\{\boldsymbol{x}^\top \beta \circ \Gamma(\tau;\boldsymbol{x},\gamma)\right\} = O_P\left(J_n^{\frac{1}{2}} n^{-\frac{1}{2}} + J_n^{-r}\right)$,
where r *is defined in Assumption* (3.1).

The asymptotic property at the change point mainly depends on the interpolation region $R_{2,n}$ with length $n^{-\delta}$, in which the threshold for δ is determined based on the convergence condition at $\tau = 1-\pi(\gamma,\mathbf{x})$. When $\delta \le 0.25$, the variance from quantile regression at the change point is controlled by n^{δ}, allowing $\sqrt{n}$ convergence, but the slow convergence of the

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

interpolation region leads to noticeable bias. For $\delta \in (0.25, 0.5)$, we achieve faster convergence of the interpolation region while keeping variance within a reasonable range. When $\delta \geq 0.5$, similar to the proof of Theorem 3 (ii)(b), the convergence rate at the change point slows, and larger δ values result in growing variance and unstable estimates. In numerical studies, we set $\delta = 0.499$ as in Ling et al. (2022) for a fair comparison. We also provide results with $\delta = 0.250$ in Supplement S2.5 and S3.1, which suggests that the choice of δ does not affect the estimation results very much.

From Theorem 3 (*iii*), we have the following corollary directly.

Corollary 1 *When $\tau > 1 - \pi(\gamma, \boldsymbol{x})$, under the conditions of Theorem 3, we have $\frac{1}{n}\sum_{i=1}^{n} \widehat{Q}(\tau \mid \boldsymbol{x}_i) - Q(\tau \mid \boldsymbol{x}_i) = O_P\left(J_n^{\frac{1}{2}} n^{-\frac{1}{2}} + J_n^{-r}\right)$, i.e.,*

$$\frac{1}{n}\sum_{i=1}^{n} B\left(\boldsymbol{x}_i^\top \hat{\beta}_{\hat{\tau}_s}\right)^\top \tilde{\theta}_n\left(\hat{\beta}_{\hat{\tau}_s}, \hat{\tau}_s\right) - G_{\Gamma(\tau;\boldsymbol{x}_i,\gamma)}\left\{\boldsymbol{x}_i^\top \beta \circ \Gamma(\tau; \boldsymbol{x}_i, \gamma)\right\} = O_P\left(J_n^{\frac{1}{2}} n^{-\frac{1}{2}} + J_n^{-r}\right).$$

The proofs for the theorems above are provided in Supplement S1.1.

2.6 Implementation Details

Here, we discuss how to select the nuisance parameters, i.e., the interpolation parameter δ and the number of knots N_{n_0}, in the proposed ZIQSI method. As shown in the proof of Theorem 3, a larger δ is preferred for a faster convergence rate of the interpolation region, yet it may lead to a large

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

variance at the change point. Note that our primary focus is constructing entire quantile curves rather than predicting conditional quantiles at a single τ. Estimating the entire curve is generally insensitive to the choice of δ, and we recommend $\delta = 0.499$ for simplicity. For prediction at a specific τ, cross-validation can optimize δ for better performance (Ling et al., 2022).

To estimate β_{τ_s}, which is required for estimating the quantile curve, we use equally spaced knots for the order m B-spline with $N_{n_0} = \lfloor C n_0^{1/(2m+1)} \rfloor + 1$, where $\lfloor a \rfloor$ denotes the integer part of a number, $C > 0$ is a constant, and n_0 is the number of positive outcomes. The choice of C does not change the estimation much in a reasonable range (Ma and He, 2016). In our numerical studies, we set $C = 1$ and choose N_{n_0} by finding the first local minimum of the following BIC criterion: $\mathrm{BIC}(N_{n_0}) = \log\left\{L^{**}_{\hat{\tau}_s,n}(\theta)\right\} + \frac{\log(n_0)}{2n_0}\left(N_{n_0} + m\right)$.

3. Simulations

We present numerical experiments to assess the performance of the proposed ZIQSI, the method of Ling et al. (2022) (denoted as "ZIQ-linear"), and the method of Ma and He (2016) (denoted as "Quantile Single-index"). We mainly focus on quantile-regression-based methods because Ling et al. (2022) already showed the superiority of their method compared to methods that require specific parametric assumptions (e.g., ZIP and ZINB), classic

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

linear quantile regression without two-part modeling, and the Hurdle regression model. ZIQ-linear can be viewed as a special case of ZIQSI by setting the function $G_{\tau_s}(\cdot)$ as an identity link function. Quantile Single-index performs similarly to the positive part of ZIQSI without adjusting τ by taking into account logistic regression. The Quantile Single-index model assumes the outcome to be continuous, and its estimation algorithm often fails to converge when the data contains a probability mass at zero. To circumvent this numerical difficulty, we added a small perturbation ($N(0, 10^{-10})$) to the zero-valued outcomes and applied their method to the perturbed data. For a fair comparison to ZIQ-linear, we use $\delta = 0.499$ for ZIQSI. Additional simulation results suggest that using a more minor δ, such as $\delta = 0.250$, does not cause a significant difference in estimation (see Supplement S2.5).

Though ZIQSI provides estimates for both linear index β_{τ_s} and the function $G_{\tau_s}(\cdot)$, namely $\hat{\beta}_{\hat{\tau}_s}$ and $\widehat{G}_{\hat{\tau}_s}(\cdot)$, they are subject-specific and not comparable, as $\hat{\tau}_s = \Gamma(\tau; \mathbf{x}, \hat{\gamma})$ is a function of $\mathbf{x}$. Therefore, we estimate quantile functions for 12 individuals, whose health-related covariates $\mathbf{x}$ are representative in real data (Table S2.1 in Supplement S2.1).

To compare the performance of the three methods, we assess the estimated quantile curves $\widehat{Q}_Y(\tau \mid \mathbf{x})$ by the relatively integrated mean squared error (RIMSE), the relatively integrated bias-squared (RIBIAS), and the

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

relatively integrated variance (RIVAR) defined as follows:

$$(1)\text{RIMSE} = \int \mathbb{E}\left\{\widehat{Q}_Y(\tau \mid \mathbf{x}) - Q_Y(\tau \mid \mathbf{x})\right\}^2 d\tau \Big/ \int Q_Y(\tau \mid \mathbf{x})^2 d\tau,$$

$$(2)\text{RIBIAS} = \int \left\{\mathbb{E}\widehat{Q}_Y(\tau \mid \mathbf{x}) - Q_Y(\tau \mid \mathbf{x})\right\}^2 d\tau \Big/ \int Q_Y(\tau \mid \mathbf{x})^2 d\tau,$$

$$(3)\text{RIVAR} = \int \mathbb{E}\left\{\widehat{Q}_Y(\tau \mid \mathbf{x}) - \mathbb{E}\widehat{Q}_Y(\tau \mid \mathbf{x})\right\}^2 d\tau \Big/ \int Q_Y(\tau \mid \mathbf{x})^2 d\tau.$$

All three measurements are based on fixed $\mathbf{x}$ and standardized by the squared scale of the quantile curve integrated through the entire process $\tau \in (0, 1)$. The integrals in the three measurements are numerically approximated by the Riemann sums on $\tau = 0.01, 0.02, \cdots, 0.99$.

3.1 Simulation settings

The dataset is simulated to mimic the real microbiome count data, with Y being the read counts and $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)^\top$ being covariates, according to the distribution from real data. For the covariates $\mathbf{x}$, we generate $x_1 \sim Bernoulli(0.5)$ for medicament, $x_2 \sim N(28, 2^2)$ for BMI, $x_3 \sim N(92.5, 13^2)$ for waist circumference, $x_4 \sim N(80, 12^2)$ for diastolic blood pressure, and $x_5 \sim N(124, 18.5^2)$ for systolic blood pressure. For each dataset, we generated $(\mathbf{x}_i, y_i)$ for $i = 1, \cdots, n$ with the sample size $n = 500$, similar to the sample size of the real data application. For the ith subject $\mathbf{x}_i = (x_{i,1}, \cdots, x_{i,5})^\top$, we first randomly simulated a vari-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

able $\tau_i \sim Unif(0,1)$ representing the quantile level of the ith individual and D_i from a *Bernoulli* distribution with a success probability defined as $P(D_i = 1 \mid \mathbf{x}_i) = \pi(\gamma, \mathbf{x}_i) = \frac{\exp(\gamma_0 + \sum_{j=1}^{5} \gamma_j x_{i,j})}{1+\exp(\gamma_0 + \sum_{j=1}^{5} \gamma_j x_{i,j})}$, where the parameter $\gamma = (-0.4, -0.480, -0.022, 0.021, 0.015, -0.009)^\top$ were set to control the proportion of zeros in outcomes. Then, we set $y_i = 0$ if $D_i = 0$. If $D_i = 1$, we generated the microbial count from the following quantile function: $Q_Y(\tau_i \mid \mathbf{x}_i, Y_i > 0) = G_{\tau_i}\left(\beta_0(\tau_i) + \mathbf{x}_i^\top \beta(\tau_i)\right)$, where the two sets of the true coefficients $\beta(\tau) = (\beta_1(\tau), \beta_2(\tau), \beta_3(\tau), \beta_4(\tau), \beta_5(\tau))^\top$ and the quantile functions $G_\tau(\cdot)$ are simulated to mimic the distributions of a taxon in our real data analysis (see Supplement S2.1): $\beta_0(\tau) = -147.7\tau - 50\tau^2 - 20$, $\beta_1(\tau) = 0.6\sqrt{\tau} - 2\tau$, $\beta_2(\tau) = 2.2\tau^2$, $\beta_3(\tau) = \frac{2}{3}\tau^2 - \frac{1}{3}\tau + 0.4$, $\beta_4(\tau) = -0.1\sin(2\pi\tau)$, $\beta_5(\tau) = -0.6\tau^2 + 2\tau$, and $G_\tau(x) = \frac{1}{6}\tau x^4 \times 10^{-5} + \frac{1}{15}\tau x^2$. We provide the comparison between the distributions of the read count generated by our simulation setting and the read count of one real taxon count in Supplement S2.1 (Figure S2.1). Simulation results are presented based on 500 Monte Carlo replicates. Our method takes approximately 30 seconds to estimate the quantile regression model on a grid of nominal levels $\tau = 0.01, \cdots, 0.99$ on a macOS machine with an Apple M2 chip.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

3.2 Results for model fitting

We use the samples $(\mathbf{x}_i, y_i)$ for $i = 1, \cdots, 500$ and the measurements above. From Table 1, we observe that the proposed ZIQSI method has a significantly smaller bias (RIBIAS) compared to both ZIQ-linear and Quantile Single-index. The RIVAR and RIMSE of ZIQSI and Quantile Single-index are comparable, and ZIQ-linear could have surprisingly large RIMSE due to large RIBIAS (e.g., subjects 1 and 2) as well as RIVAR (e.g., subjects 9 and 10). In general, ZIQ-linear performs worse than Quantile Single-index because the linear assumption on the quantile function for the positive part $Y > 0$ is violated. We also assess the simulation results where $G_\tau(x) = \tau x$ is a simple linear function in Supplement S2.7, and the result is consistent with our expectations.

We further report the average proportion of negative predicted counts over the quantile process $\tau \in (0, 1)$ for three methods in Table S2.2 in Supplement S2.2. Based on Table 1 and Table S2.2 in Supplement S2.2, we observe that ZIQ-linear has a small portion of negative predictions but severe bias; meanwhile, Quantile Single-index suffers from a large portion of negative predictions, though the estimation bias is moderate. To eliminate the effect of results below zero, we present the results truncated at

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

Table 1: Summary of RIMSE(%), RIBIAS(%), RIVAR(%) of the estimated conditional quantile functions by ZIQSI, ZIQ-linear(ZIQ), and Quantile Single-index(QSI).

	RIBIAS			RIVAR			RIMSE		
ID	ZIQSI	ZIQ	QSI	ZIQSI	ZIQ	QSI	ZIQSI	ZIQ	QSI
1	0.19	21.18	1.14	3.20	5.25	2.81	3.39	26.43	3.95
2	0.07	21.29	0.44	3.96	6.19	3.94	4.03	27.48	4.38
3	0.24	4.07	1.10	1.54	1.63	1.49	1.78	5.70	2.59
4	0.04	4.17	0.13	1.67	1.66	1.82	1.71	5.83	1.95
5	0.10	2.53	0.76	3.31	1.01	3.62	3.41	3.54	4.38
6	0.04	2.34	0.13	3.80	1.20	3.80	3.84	3.54	3.93
7	0.34	1.23	0.84	3.09	2.00	2.95	3.43	3.23	3.79
8	0.12	1.27	0.65	3.54	2.27	3.55	3.66	3.54	4.20
9	0.13	19.12	1.01	1.57	4.30	2.36	2.70	23.42	3.37
10	0.02	18.88	0.15	3.01	4.83	3.14	3.03	23.71	3.29
11	0.02	9.04	0.99	1.98	2.22	1.53	2.00	11.26	2.54
12	$6.26e^{-5}$	9.93	0.12	2.25	2.55	2.42	2.25	12.48	2.54

zero in Supplement S2.3 (see Table S2.3), where our method remains its advantages. The proposed method ZIQSI shows its superiority regarding the smallest integrated bias and a reasonably small portion of predictions below zero.

For each subject, we also visualize the estimation performance of each method across the quantile process $\tau \in (0, 1)$. We reported the average estimated quantile curves and their 95% confidence intervals based on the 500 estimations above. The confidence interval is constructed based on the percentile of the empirical distribution of $\widehat{Q}_Y(\tau \mid \mathbf{x})$ at a given τ. We show subject 11 in Figure 2 and present others in Supplement S2.4. We observe

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

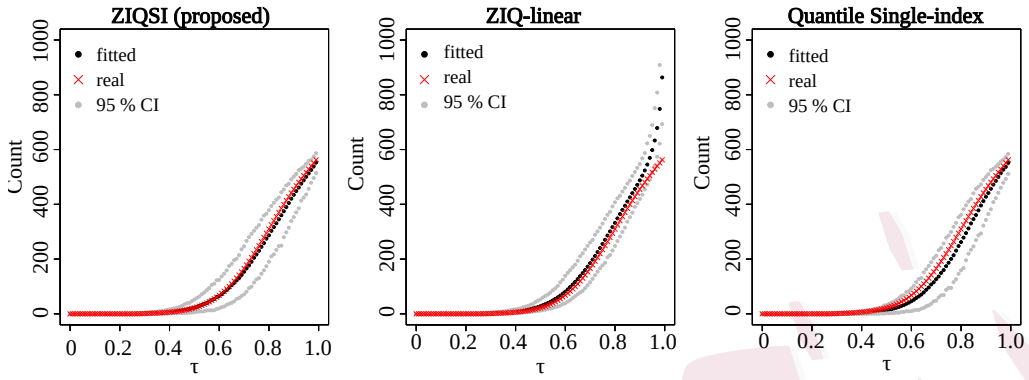


Figure 2: Quantile curves based on 500 times estimations (Subject 11).

that both Quantile Single-index and ZIQ-linear have an obvious bias, and a larger estimation bias of ZIQ-linear is observed at upper quantiles. Similar patterns are observed for other individuals (Supplement S2.4). The simulation results for AQE indicate that ZIQSI provides the most accurate and stable estimation compared to the other two methods (Supplement S2.6).

4. Application

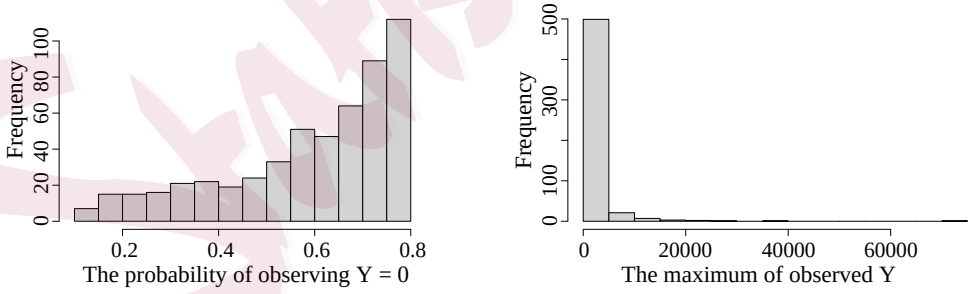
In this section, we illustrate the performance of our ZIQSI method by the study of Columbian's Gut (De la Cuesta-Zuluaga et al., 2018; Gonzalez et al., 2018), with the dataset publicly available at `https://qiita.ucsd.edu/`. We compare the proposed method (ZIQSI) with the method of Ling et al. (2022) (ZIQ-linear) and the method of Ma and He (2016) (Quantile Single-index) by assessing model fitting from the population and individual

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

perspectives. As in Section 3, we use $\delta = 0.499$ for a fair comparison with ZIQ-linear. The results of using a smaller δ are similar and are provided in Supplement S3.1. We have also developed an `R` package implementing our method, which is available at `https://github.com/tianyingw/ZIQSI/`.

4.1 Data description

The dataset contains microbiome counts of over 6000 taxa for 441 adults, along with covariates related to diet, obesity, and cardiometabolic diseases. We consider taxa with observed zero proportions less than 0.8, as a larger percentage of zeros commonly leads to unreliable results (Wadsworth et al., 2017; Jiang et al., 2021; Zhang and Yi, 2020). From Figure 3, we observe that a large number of taxa are heavily zero-inflated, and the observed counts are overdispersed.



(a) Zero proportion per taxon. (b) Maximum count per taxon.

Figure 3: Histogram for microbiome counts.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

Following the study of De la Cuesta-Zuluaga et al. (2018), we analyzed taxa counts with health-related covariates as follows: anthropometric measures (age, BMI, sex, waist circumstance), lipid profile (adiponectin, total cholesterol, HDL, LDL, triglycerides), glucose metabolism (glucose, glycosylated hemoglobin, insulin), blood pressure (diastolic blood pressure, systolic blood pressure), city, medicament, and macronutrient consumption (fiber, percentage of animal protein, carbohydrates, monounsaturated fat, polyunsaturated fat, saturated fat, total fat, protein). Among them, categorical variables, such as sex, medicament, and city, are treated as dummy variables. We further removed 3 subjects for missing values and 2 subjects for extremely high values of triglycerides over 800 mg/dL, resulting in 436 samples in our analysis. We analyzed 535 taxa with observed zero proportions in the range of 0.1-0.8. As a common practice in other microbiome studies (Xia et al., 2018), we adjust for the library size, which is the sum of all 535 taxa counts per person.

4.2 Goodness-of-fit

To provide a thorough analysis, we used a representative taxon, namely *Slackia*, which has the third highest abundance out of 97 taxa in the *Coriobacteriaceae* family from the co-abundance groups *Prevotella* based

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

on hierarchical clustering with Ward's linkage (Claesson et al., 2012). We assessed model fitting and quantile curve estimation from different perspectives. We also analyzed taxa with varying degrees of zeros and provided results in Supplement S3.2.

To assess the goodness of fit for a model, we adopt the measurement used in Ling et al. (2022) and Heyman et al. (1991) to compare the distribution of the observed data and the predicted values from fitted models. We first fit the model based on the aforementioned three methods. Then, a quantile level τ is randomly drawn from $Unif(0,1)$, and $\widehat{Q}_Y(\tau \mid X)$ is reported as the fitted microbiome counts given observed covariates. The computation time for estimating the quantile single-index models on the nominal quantile levels $\tau = 0.01, \cdots, 0.99$ is around 32 seconds on a macOS machine with an Apple M2 chip.

From Figure 4, we observe that the proposed ZIQSI method better fits the taxon *Slackia* compared to the other two methods, especially at two tails. On the contrary, both ZIQ-linear and Quantile Single-index predicted counts below zero, which is against the non-negativity of the number of microorganisms. Model fitting results for other taxa also suggest a similar pattern (see Supplement S3.2). Though ZIQ-linear commonly has a smaller

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

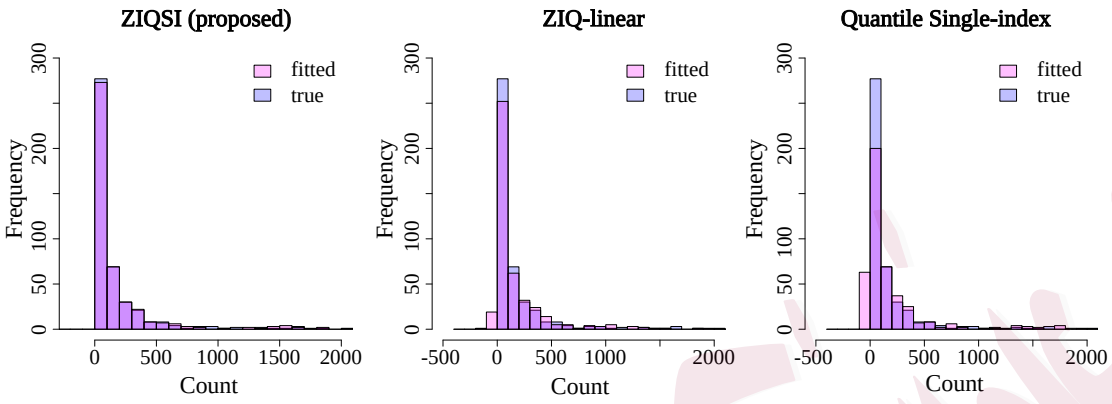


Figure 4: Histogram plot of *Slackia*.

proportion of negative predicted counts compared to Quantile Single-index, the values could be as small as -500 (Supplement S3.2 Figure S3.8). Quantile Single-index often has many negative predicted counts, which is consistent with the simulation results. To investigate the lack of goodness of fit for ZIQ-linear and Quantile Single-index methods, we provide detailed discussions from the population and individual perspectives below.

4.3 Estimated quantile curves

As the quantile effect is caused by the logistic and quantile single-index components, visualizing it is more complicated than simply presenting $\hat{\beta}_\tau$ or $\widehat{G}_\tau(\mathbf{x}^\top \hat{\beta}_\tau)$. It needs to be highlighted that the effect of covariates in the logistic regression also plays a role through $\Gamma(\tau; \mathbf{x}, \hat{\gamma}_n)$ (i.e., $\hat{\tau}_s$). That is,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

given a fixed τ and $\hat{\gamma}_n$ estimated from logistic regression, $\hat{\beta}_{\hat{\tau}_s} = \hat{\beta}_{\Gamma(\tau;\mathbf{x},\hat{\gamma}_n)}$ is a function of $\mathbf{x}$. Thus, we visualize the quantile effects for covariates by fixing τ while changing $\mathbf{x}$, or vice versa.

First, we present how $\widehat{Q}_Y(\tau;\mathbf{x})$ changes with $\mathbf{x}$ at given τ. For illustration, we consider the distinct variable systolic blood pressure (denoted as "systolic bp") as the target covariate and take the other covariates fixed, since systolic bp has a significant effect on the abundance of *Slackia* (De la Cuesta-Zuluaga et al., 2018). Specifically, the continuous covariates are fixed at their average levels, and we take binary/categorical covariates "sex" as female, "medicament" as 1, and "city" as Cali. Using the microbiota *Slackia* as an example again, we present the estimated quantile curves $\widehat{G}_\tau(\mathbf{x}^\top\beta(\tau))$ regarding different levels of systolic bp at the nominal quantile levels $\tau = \{0.5, 0.6, 0.7\}$ in Figure 5. Of note, though the nominal quantile level τ is fixed, τ_s, adjusted by the logistic regression, changes with different levels of systolic bp. Thus, the points in Figure 5 do not align well, and we provided B-spline fitted curves based on the estimated points. We observe that ZIQ-linear has the quantile crossing issue when the systolic bp is larger than 160. That is, the estimated counts at $\tau = 0.6$ are lower than the ones at $\tau = 0.5$ and higher than the ones at $\tau = 0.7$, which violates

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

the monotonic nature of quantiles. Also, the predicted counts at $\tau = 0.6$ with ZIQ-linear are negative with large systolic bp values, which explains the negative predicted values we observed in the histogram (Figure 4).

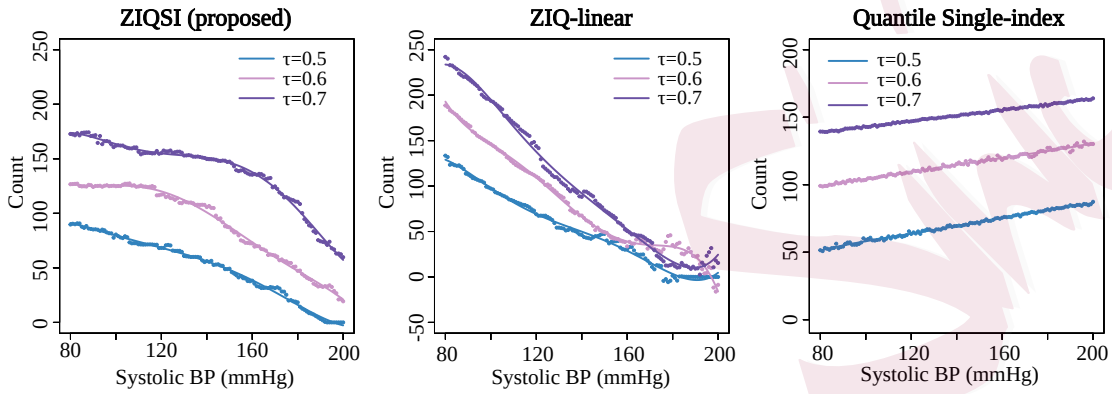
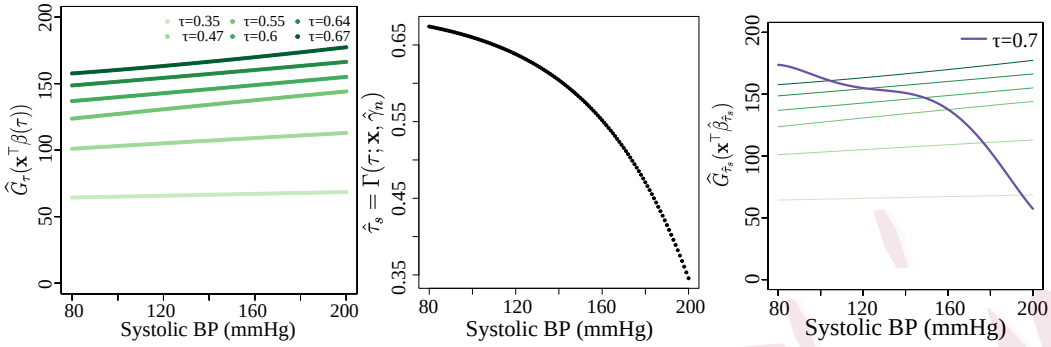


Figure 5: Predicted counts for taxon *Slackia* with the change of systolic bp (other covariates are fixed).

From Figure 5, we observed that the predicted counts from ZIQ-linear and ZIQSI have a similar decreasing pattern with the increase of systolic bp, though ZIQ-linear has some unreasonable predictions. The estimated quantile curves by the Quantile Single-index method, however, showed a different trend as it does not adjust the quantile level τ and assumes the probability of observing a zero outcome is the same for every subject. Thus, we further illustrate the difference between the pre-fixed quantile level τ and its adjusted version $\hat{\tau}_s$. In Figure 6(a), the quantile curves estimated

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104



(a) (b) (c)

Figure 6: Compare the estimated quantile curve with original and adjusted quantile levels. (a): Estimated quantile curves from ZIQSI with unadjusted τ. (b): Mapping $\tau = 0.7$ to $\hat{\tau}_s$ through $\Gamma(\tau, \mathbf{x}, \hat{\gamma}_n)$ with the change of systolic bp ($\mathbf{x}$). (c): Estimated quantile curves from ZIQSI with adjusted $\hat{\tau}_s$ (purple curve), while the quantile level is fixed at $\tau = 0.7$.

by ZIQSI with the fixed quantile levels τ have similar trends as the curves estimated by the Quantile Single-index method (Figure 5 (right)). Then, when we consider a quantile level $\tau = 0.7$, its mapped quantile level $\hat{\tau}_s$ is decreasing with the increase of systolic bp owing to its negative effect (Figure 6(b)), as systolic bp has a negative estimated coefficient in the logistic regression, which means a higher systolic bp level can lead to a lower nominal $\hat{\tau}_s$. Naturally, the change of $\hat{\tau}_s$ results in the accelerated decreasing curve $\widehat{G}_{\hat{\tau}_s}(\mathbf{x}^\top\hat{\beta}_{\hat{\tau}_s})$ (Figure 6(c)), which is consistent with the

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

results presented in Figure 5 (left). For ZIQ-linear, the trend of its estimated quantile curves is similar to ZIQSI, as the adjustment for τ through logistic regression (i.e., $\hat{\gamma}_n$) remains the same.

Then, we show the estimated quantile curve $\widehat{Q}_Y(\tau; \mathbf{x})$ for a specific subject. Among the subjects whose predicted counts are negative, we randomly select one sample and present the fitted quantile curves (Figure 7). The proposed ZIQSI method reasonably estimates the entire quantile curve, while ZIQ-linear showed a non-monotone curve with the increase of τ, which is counter-intuitive and against the nature of quantiles. Further, ZIQ-linear has negative predictions, which is counter-intuitive as the response is required to be non-negative. We also present the effect of a specific covariate by comparing the AQE based on each method (see Supplement S3.3).

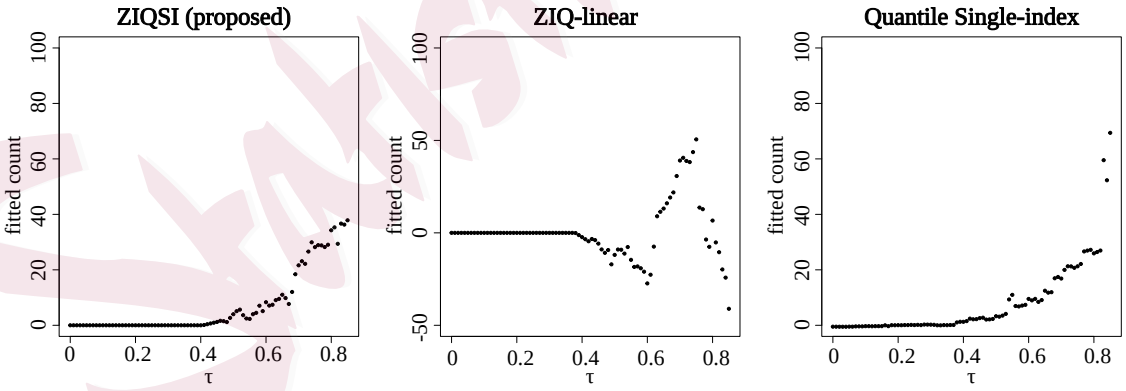


Figure 7: Predicted quantile curve of subject X11993.MI385H.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

5. Discussion

In this paper, we focus on statistical modeling for zero-inflated and overdispersed microbiome data. To relax parametric assumptions in existing two-part modeling approaches and provide more flexibility in handling complex associations, we propose a novel semiparametric single-index quantile regression model that first extends single-index quantile regression models to zero-inflated and overdispersed outcomes. Both the theoretical and empirical works suggest that this method outperformed in modeling zero-inflated and overdispersed outcomes.

Several interesting topics warrant further investigation. First, current quantile regression methods for zero-inflated data, including our ZIQSI method, do not enforce non-negativity, potentially leading to negative predictions due to numerical issues. Adding a non-negativity constraint to the link function G_τ could address this; see (Cannon, 2018) for an example in composite quantile regression. Next, while our method accommodates high-dimensional covariates via single-index models, it may struggle with high-dimensional data. Incorporating regularization (Li and Yin, 2008; Peng and Huang, 2011) or using semiparametric dimension reduction (Ma and Zhu, 2012) could improve performance, though this would require refining

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

the asymptotic theory. Challenges also arise when the number of covariates grows with sample size, complicating the nonparametric estimation of $G_\tau(\cdot)$ and the linear index. Lastly, our method focuses on normal quantile levels, but estimating tail quantiles is particularly challenging, especially as τ_n approaches 1 at the rate $n(1-\tau_n) \to c > 0$ (Xu et al., 2022). Extending the tail single-index model (Xu et al., 2022) to zero-inflated data presents a promising avenue, as zero inflation further complicates extreme quantile estimation by reducing the effective sample size.

Supplementary Material:

The online Supplementary Material contains the proofs of the theorems and the additional results for simulation and application.

Acknowledgments

We thank the editor, associate editor, and two referees for their valuable comments and constructive suggestions.

References

Cani, P. D. (2018). Human gut microbiome: hopes, threats and promises. *Gut 67*(9), 1716–1725.

Cannon, A. J. (2018). Non-crossing nonlinear regression quantiles by monotone composite

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

quantile regression neural network, with application to rainfall extremes. *Stochastic environmental research and risk assessment 32*(11), 3207–3225.

Chen, E. Z. and H. Li (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics 32*(17), 2611–2617.

Claesson, M. J., I. B. Jeffery, S. Conde, S. E. Power, and E. M. e. a. O'connor (2012). Gut microbiota composition correlates with diet and health in the elderly. *Nature 488*(7410), 178–184.

De Boor, C. (2001). Revised edition. applied mathematical sciences.

De la Cuesta-Zuluaga, J., V. Corrales-Agudelo, E. P. Velásquez-Mejía, J. A. Carmona, J. M. Abad, and J. S. Escobar (2018). Gut microbiota is associated with obesity and cardiometabolic disease in a population in the midst of westernization. *Scientific reports 8*(1), 1–14.

Gonzalez, A., J. A. Navas-Molina, T. Kosciolek, D. McDonald, Y. Vázquez-Baeza, and G. A. et al. (2018). Qiita: rapid, web-enabled microbiome meta-analysis. *Nature Methods 15*, 796–798.

Heyman, D., A. Tabatabai, and T. Lakshman (1991). Statistical analysis and simulation study of video teleconference traffic in atm networks. In *IEEE Global Telecommunications Conference GLOBECOM'91: Countdown to the New Millennium. Conference Record*, pp. 21–27. IEEE.

Jiang, R., X. Zhan, and T. Wang (2022). A flexible zero-inflated poisson-gamma model with

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

application to microbiome read counts. *arXiv preprint arXiv:2207.07796*.

Jiang, S., G. Xiao, A. Y. Koh, J. Kim, Q. Li, and X. Zhan (2021). A bayesian zero-inflated negative binomial regression model for the integrative analysis of microbiome data. *Biostatistics 22*(3), 522–540.

Kaul, A., S. Mandal, O. Davidov, and S. D. Peddada (2017). Analysis of microbiome data in the presence of excess zeros. *Frontiers in microbiology 8*, 2114.

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

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

Li, L. and X. Yin (2008). Sliced inverse regression with regularizations. *Biometrics 64*(1), 124–131.

Liang, H., X. Liu, R. Li, and C.-L. Tsai (2010). Estimation and testing for partially linear single-index models. *Annals of statistics 38*(6), 3811.

Ling, W., B. Cheng, Y. Wei, J. Z. Willey, and Y. K. Cheung (2022). Statistical inference in quantile regression for zero-inflated outcomes. *Statistica Sinica 32*(3), 1411.

Lloyd-Price, J., G. Abu-Ali, and C. Huttenhower (2016). The healthy human microbiome. *Genome medicine 8*(1), 1–11.

Ma, S. and X. He (2016). Inference for single-index quantile regression models with profile optimization. *The Annals of Statistics 44*(3).

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

REFERENCES

Ma, Y. and L. Zhu (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association 107*(497), 168–179.

Ma, Y. and L. Zhu (2013). Doubly robust and efficient estimators for heteroscedastic partially linear single-index models allowing high dimensional covariates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(2), 305–322.

McMurdie, P. J. and S. Holmes (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS computational biology 10*(4), e1003531.

Neykov, M., J. S. Liu, and T. Cai (2016). L1-regularized least squares for support recovery of high dimensional single index models with gaussian designs. *The Journal of Machine Learning Research 17*(1), 2976–3012.

Peng, H. and T. Huang (2011). Penalized least squares for single index models. *Journal of Statistical Planning and Inference 141*(4), 1362–1379.

Radchenko, P. (2015). High dimensional single index models. *Journal of Multivariate Analysis 139*, 266–282.

Silverman, J. D., K. Roche, S. Mukherjee, and L. A. David (2020). Naught all zeros in sequence count data are the same. *Computational and structural biotechnology journal 18*, 2789–2798.

Wadsworth, W. D., R. Argiento, M. Guindani, J. Galloway-Pena, and S. A. e. a. Shelburne (2017). An integrative bayesian dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC bioinformatics 18*(1), 1–12.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

Wei, Y. and X. He (2006). Conditional growth charts. *The Annals of Statistics 34*(5), 2069–2097.

Weir, C. B. and A. Jan (2019). Bmi classification percentile and cut off points.

Xia, Y. (2020). Correlation and association analyses in microbiome study integrating multiomics in health and disease. *Progress in Molecular Biology and Translational Science 171*, 309–491.

Xia, Y. and J. Sun (2017). Hypothesis testing and statistical analysis of microbiome. *Genes & diseases 4*(3), 138–148.

Xia, Y., J. Sun, D.-G. Chen, Y. Xia, J. Sun, and D.-G. Chen (2018). Modeling zero-inflated microbiome data. *Statistical analysis of microbiome data with R*, 453–496.

Xu, W., H. J. Wang, and D. Li (2022). Extreme quantile estimation based on the tail single-index model. *Statistica Sinica 32*(2), 893–914.

Yatsunenko, T., F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, and M. e. a. Contreras (2012). Human gut microbiome viewed across age and geography. *nature 486*(7402), 222–227.

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.

Zeng, Y., J. Li, C. Wei, H. Zhao, and W. Tao (2022). mbdenoise: microbiome data denoising using zero-inflated probabilistic principal components analysis. *Genome Biology 23*(1),

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0104

1–29.

Zhang, X., H. Mallick, Z. Tang, L. Zhang, X. Cui, and A. K. e. a. Benson (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC bioinformatics 18*(1), 1–10.

Zhang, X. and N. Yi (2020). Fast zero-inflated negative binomial mixed modeling approach for analyzing longitudinal metagenomics data. *Bioinformatics 36*(8), 2345–2351.

Department of Statistics and Data Science, Tsinghua University

E-mail: wzr23@mails.tsinghua.edu.cn

Department of Statistics, Colorado State University

E-mail: Tianying.Wang@colostate.edu