Statistica Sinica **32** (2022), 1411-1433
doi:https://doi.org/10.5705/ss.202020.0368

STATISTICAL INFERENCE IN QUANTILE REGRESSION FOR ZERO-INFLATED OUTCOMES

Wodan Ling[1], Bin Cheng[2], Ying Wei[2],
Joshua Z. Willey[2] and Ying Kuen Cheung[2]

[1]*Fred Hutchinson Cancer Research Center and* [2]*Columbia University*

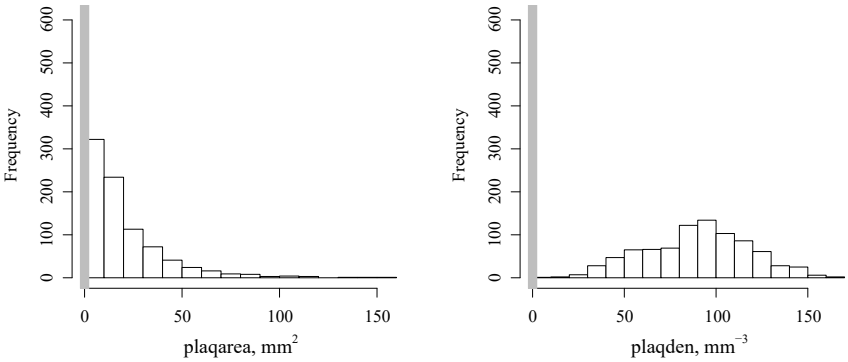
Abstract: An extension of quantile regression is proposed to model zero-inflated outcomes, which have become increasingly common in biomedical studies. The method is flexible enough to depict complex and nonlinear associations between the covariates and the quantiles of the outcome. We establish the theoretical properties of the estimated quantiles, and develop inference tools to assess the quantile effects. Extensive simulation studies indicate that the novel method generally outperforms existing zero-inflated approaches and the direct quantile regression in terms of the estimation and inference of the heterogeneous effect of the covariates. The approach is applied to data from the Northern Manhattan Study to identify risk factors for carotid atherosclerosis, measured by the ultrasound carotid plaque burden.

Key words and phrases: Constrained post-estimation smoothing, nonnormal asymptotic distribution, quantile regression, zero-inflated outcomes.

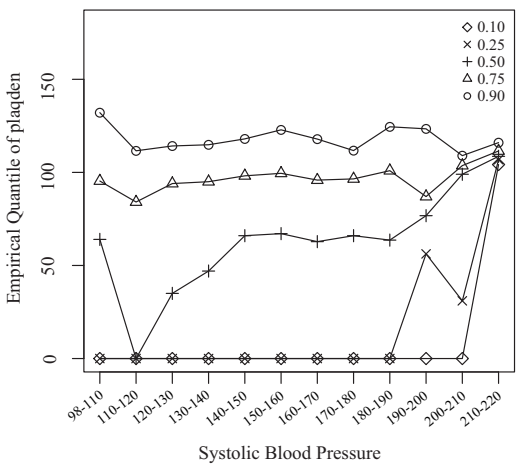
1. Introduction

Zero-inflated outcomes are common in disease etiology studies. One such example is carotid plaque (thickening of part of the artery wall), which measures carotid atherosclerosis, a proximate risk factor for stroke and cardiovascular diseases. Figure 1a shows the frequency histograms of two carotid plaque features, namely plaque area (plaqarea, in mm^2) and plaque echodensity (plaqden, in mm^{-3}), measured using high-resolution ultrasounds in 1,462 participants of the NOrthern MAnhattan Study (NOMAS) (Cheung et al. (2017)). Specifically, the plaque area measures the size of the plaque, and the echodensity indicates the texture of the plaque. When an individual does not have detectable plaque, both variables are zero. One objective of the study is to understand how the potential determinants of cardiovascular risks, including demographics, health behaviors, and medical conditions, are associated with the natural progress of carotid atherosclerosis.

Corresponding author: Bin Cheng, Department of Biostatistics, Mailman School of Public Health, Columbia University, NY 10032, USA. E-mail: bc2159@cumc.columbia.edu.



(a) Frequency histograms of plaque area (plaqarea, left) and echodensity (plaqden, right) in carotid plaque data.



(b) Empirical quantiles of plaque echodensity (plaqden) against systolic blood pressure. The relationship is nonlinear because the proportion of zeros changes with systolic blood pressure.

Figure 1. Plots of carotid plaque data.

A typical modeling approach for zero-inflated outcomes assumes that the distribution of the outcome is a mixture of a degenerated distribution at zero and another parametric distribution(s), such as the zero-inflated Poisson (ZIP) regression (Lambert (1992)) or generalized ZIP (GZIP) mixture regression (Lim, Li and Philip (2014)). More generally, Jorgensen (1987) considered a compound Poisson-gamma (CPG) distribution within the generalized linear model framework. However, these parametric methods often impose strong assumptions on

the outcome distributions, likely leading to biased results and invalid inferences.

The standard quantile regression (Koenker and Bassett (1978)) is more robust, because it avoids parametric specifications, and more versatile at describing heterogeneous effects at different quantile levels. However, it cannot be applied directly to model zero-inflated outcomes, for two reasons. First, the feasibility of the estimation and the validity of the inference for quantile regression models are based on the assumption that the conditional distribution of the outcome is absolutely continuous, which is violated with the presence of zero inflation. Second, the direct quantile regression implicitly assumes a constant chance of observing a positive outcome, which is unlikely because the degree of zero inflation varies across subjects. Furthermore, because the probability of the outcome taking the value zero varies according to the covariates, the quantile function of the outcome depends on the covariates in a nonlinear fashion, which is not readily depicted in a regular quantile regression model. To illustrate this point, Figure 1b plots the τth empirical quantiles of echodensity (plaqden) by subgroups based on systolic blood pressure, where $\tau = 0.10, 0.25, 0.50, 0.75, 0.90$. It shows that individuals with lower systolic blood pressure are associated with a greater proportion of zeros, resulting in a nonlinear relationship between the quantiles of echodensity and systolic blood pressure. This characteristic is not captured by a linear quantile regression that ignores the point mass at zero.

In this paper, we propose a two-part modeling strategy that uses a logistic regression to model the probability of being positive, and a linear quantile regression to model the positive part, with the quantile levels adjusted by subject-specific zero inflation rates. The model generalizes the parametric two-part regression approach of Duan et al. (1983) and the hurdle regression model of Mullahy (1986). Although conceptually straightforward, obtaining a valid estimation and inference of the proposed two-part quantile regression model is challenging, for two reasons. First, the neighboring quantile estimation around the change point from zero to positive may have an unbounded variance. The variance of the unadjusted estimated quantile process is inversely proportional to the local data density. When approaching the change point, the local density of positive data could go to zero, in which case, the variance is pushed to infinity. Second, estimations and inferences of the quantile covariate effect are complicated. The quantile effect is a composite of the logistic and quantile regression components, and it depends on the values of the covariate of interest and other covariates in the two-part model.

To address these challenges, we develop an algorithm to achieve a consistent estimation of the conditional quantiles, while circumventing the unbounded variance at the quantile level where the conditional quantile changes from zero to

positive. The consistency and asymptotic distribution of the resulting estimated conditional quantile function are established. To facilitate inferences, we define marginal quantile treatment effects, and develop inference tools to determine their statistical significance. Similar applications of two-part quantile regression models are used in Heras, Moreno and Vilar-Zanón (2018) to estimate actuarial profiles and provide new insights for actuarial science. The work, however, does not include any theoretical validation or development. To the best of our knowledge, our work provides the first theoretically valid estimation and inference of two-part quantile regression models for zero-inflated outcomes.

The rest of the paper is organized as follows. Section 2 presents our proposed model, the model-based conditional quantile estimation, and its asymptotic properties. The model-based inference tool for the quantile treatment effects and model-based predictions are discussed in Section 2.3. We compare the finite-sample performance of the proposed method with that of the uncorrected direct quantile regression and parametric approaches for zero-inflated outcomes using simulation studies in Section 3. Section 4 presents a real application of the proposed method, studying the effect of risk factors for carotid atherosclerosis, measured by the ultrasound carotid plaque burden, comparing the results with those competing methods. Section 5 concludes the paper.

2. Proposed Methods

2.1. Model

Suppose Y is a nonnegative, zero-inflated outcome, and $\boldsymbol{X}$ is a vector of covariates that may be associated with the quantiles of Y. Throughout the paper, we denote $Q_Y(\tau|\boldsymbol{X})$ as the τth conditional quantile of Y given $\boldsymbol{X}$.

To estimate the distribution of the zero-inflated Y, we decompose its conditional distribution as

$$F(Y|\boldsymbol{X}) = P(Y = 0|\boldsymbol{X}) + F(Y|\boldsymbol{X}, Y > 0)\, P(Y > 0|\boldsymbol{X}),$$

and then model the two components $F(Y|\boldsymbol{X}, Y > 0)$ and $P(Y > 0|\boldsymbol{X})$ separately. We first assume that the probability of observing a positive Y, $P(Y > 0|\boldsymbol{X})$, follows a logistic regression model,

$$\text{logit}\{P(Y > 0|\boldsymbol{X})\} = \boldsymbol{X}^\top \boldsymbol{\gamma}, \tag{2.1}$$

where $\boldsymbol{\gamma}$ is the true coefficient, such that $P(Y > 0|\boldsymbol{X}) = \exp(\boldsymbol{X}^\top\boldsymbol{\gamma})/\{1 + \exp(\boldsymbol{X}^\top\boldsymbol{\gamma})\}$. Next, we assume that for any nominal quantile level $\tau \in (0, 1)$,

the conditional quantile of Y given $Y > 0$ is a linear function of $\boldsymbol{X}$,

$$Q_Y(\tau|\boldsymbol{X}, Y > 0) = \boldsymbol{X}^\top \boldsymbol{\beta}(\tau). \tag{2.2}$$

The model implies that the conditional quantile function $\boldsymbol{X}^\top \boldsymbol{\beta}(\tau)$ is nonnegative, while neither the covariate X nor the coefficient function $\boldsymbol{\beta}(\tau)$ is required to be positive. In addition, we assume that for any $\boldsymbol{X}$,

$$\lim_{\tau \to 0^+} Q_Y(\tau|\boldsymbol{X}, Y > 0) = 0, \tag{2.3}$$

which ensures that the quantile function $Q_Y(\tau|\boldsymbol{X})$ is continuous at zero. Note that, in practice, different subsets of the covariate profile $\boldsymbol{X}$ can be used as the two covariates in Models (2.1) and (2.2).

Under Models (2.1) and (2.2) and Assumption (2.3), the τth conditional quantile of Y given the covariates $\boldsymbol{X}$ can be written as

$$Q_Y(\tau|\boldsymbol{X}) = I\{\tau > 1 - \pi(\boldsymbol{\gamma}, \boldsymbol{X})\} \cdot \boldsymbol{X}^\top \boldsymbol{\beta} \circ \Gamma(\tau; \boldsymbol{X}, \boldsymbol{\gamma}), \tag{2.4}$$

where $\pi(\boldsymbol{\gamma}, \boldsymbol{X}) = P(Y > 0|\boldsymbol{X})$ is the probability of observing a positive Y given the covariates $\boldsymbol{X}$. The function $\Gamma(\tau; \boldsymbol{X}, \boldsymbol{\gamma}) : (1 - \pi(\boldsymbol{\gamma}, \boldsymbol{X}), 1) \to (0, 1)$ maps the target quantile level τ of Y to the nominal quantile level τ_s of $Y|Y > 0$ in Model (2.2). Specifically,

$$\boldsymbol{\beta} \circ \Gamma(\tau; \boldsymbol{X}, \boldsymbol{\gamma}) = \boldsymbol{\beta}(\tau_s), \text{ and } \tau_s = \Gamma(\tau; \boldsymbol{X}, \boldsymbol{\gamma}) = \max\left(\frac{\tau - \{1 - \pi(\boldsymbol{\gamma}, \boldsymbol{X})\}}{\pi(\boldsymbol{\gamma}, \boldsymbol{X})}, 0\right). \tag{2.5}$$

Equation (2.5) is derived from the fact that, for a $\tau > 1 - \pi(\boldsymbol{\gamma}, \boldsymbol{X})$,

$$\begin{aligned} \tau &= P\{Y \le Q_Y(\tau \mid \boldsymbol{X})|\boldsymbol{X}\} \\ &= \{1 - \pi(\boldsymbol{\gamma}, \boldsymbol{X})\} + \pi(\boldsymbol{\gamma}, \boldsymbol{X}) P\{Y \le Q_Y(\tau_s \mid \boldsymbol{X}, Y > 0)|\boldsymbol{X}, Y > 0\}. \end{aligned}$$

The proposed quantile model for the zero-inflated outcome (2.4) is flexible enough to accommodate nonlinear heterogeneous quantile associations and a wide range of outcome distributions by linear models only.

Suppose we have independent and identically distributed (i.i.d.) random samples $\{(\boldsymbol{x}_i, y_i); i = 1, \ldots, n\}$ following the conditional quantile model (2.4). We can estimate the coefficients $\boldsymbol{\gamma}$ using a logistic regression (i.e., regress $I\{y_i > 0\}$ against $\boldsymbol{x}_i$), and then estimate the quantile coefficient function $\boldsymbol{\beta}(\tau)$ by regressing the positive y_i against $\boldsymbol{x}_i$ using a quantile regression at a sequence of quantile

levels. Specifically, $\widehat{\boldsymbol{\gamma}}_n$ and $\widehat{\boldsymbol{\beta}}_n(\tau)$ are solutions to the objective functions,

$$\widehat{\boldsymbol{\gamma}}_n = \underset{\boldsymbol{\gamma}}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n}\left[I(y_i>0)\log\left\{\frac{\pi(\boldsymbol{\gamma},\boldsymbol{x}_i)}{1-\pi(\boldsymbol{\gamma},\boldsymbol{x}_i)}\right\}+\log\{1-\pi(\boldsymbol{\gamma},\boldsymbol{x}_i)\}\right],$$

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_n(\tau) &= \underset{\boldsymbol{\beta}(\tau)}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\rho_\tau\{y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta}(\tau)\}I(y_i>0)\\ &= \underset{\boldsymbol{\beta}(\tau)}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\{y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta}(\tau)\}\{\tau-I(y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta}(\tau)<0)\}I(y_i>0).\end{aligned}$$

However, owing to the change point at $\tau = 1-\pi(\boldsymbol{\gamma},\boldsymbol{x})$ and the fact that $\operatorname{Var}\{\widehat{\boldsymbol{\beta}}_n(\tau)\}\to\infty$ when $\tau\to 0^+$, it is nontrivial to combine $\widehat{\boldsymbol{\gamma}}_n$ and $\widehat{\boldsymbol{\beta}}_n(\tau)$ to obtain a consistent estimation of $Q_Y(\tau|\boldsymbol{x})$ with a bounded variance around the change point. In Section 2.2, we propose a piecewise estimator for the conditional quantiles and establish its consistency and asymptotic distribution.

2.2. Estimation of $Q_Y(\tau|\boldsymbol{x})$

Recall that $\widehat{\boldsymbol{\gamma}}_n$ and $\widehat{\boldsymbol{\beta}}_n(\tau)$ are the estimated coefficients from Models (2.1) and (2.2). The procedure to estimate the conditional quantile function $Q_Y(\tau|\boldsymbol{x})$ is implemented as follows:

Step 1. Estimate the probability of observing a positive Y given the covariates $\boldsymbol{x}$,

$$\pi(\widehat{\boldsymbol{\gamma}}_n,\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^\top\widehat{\boldsymbol{\gamma}}_n)}{1+\exp(\boldsymbol{x}^\top\widehat{\boldsymbol{\gamma}}_n)}.$$

Step 2. Let δ be a constant in $(0,1/2)$. Divide the support of the target quantile levels $(0,1)$ of Y into three sub-intervals A_n, B_n, and C_n, such that $(0,1) = A_n\cup B_n\cup C_n$, and

$$\begin{aligned}A_n &= \Big\{\tau: 0<\tau<1-\pi(\widehat{\boldsymbol{\gamma}}_n,\boldsymbol{x})\Big\},\\ B_n &= \Big\{\tau: 1-\pi(\widehat{\boldsymbol{\gamma}}_n,\boldsymbol{x})\le\tau\le 1-\pi(\widehat{\boldsymbol{\gamma}}_n,\boldsymbol{x})+n^{-\delta}\Big\},\\ C_n &= \Big\{\tau: 1-\pi(\widehat{\boldsymbol{\gamma}}_n,\boldsymbol{x})+n^{-\delta}<\tau<1\Big\}.\end{aligned}$$

Step 3. Estimate the quantile coefficients $\widehat{\boldsymbol{\beta}}_n$ at the nominal quantile level $\Gamma(1-\pi(\widehat{\boldsymbol{\gamma}}_n,\boldsymbol{x})+n^{-\delta};\boldsymbol{x},\widehat{\boldsymbol{\gamma}}_n)$ and perform an interpolation if the target quantile level τ of Y belongs to B_n. If τ is in C_n, directly estimate $\widehat{\boldsymbol{\beta}}_n$ at $\Gamma(\tau;\boldsymbol{x},\widehat{\boldsymbol{\gamma}}_n)$. The estimator $\widehat{Q}_Y(\tau|\boldsymbol{x})$, as shown in Figure 2, is then a piecewise function

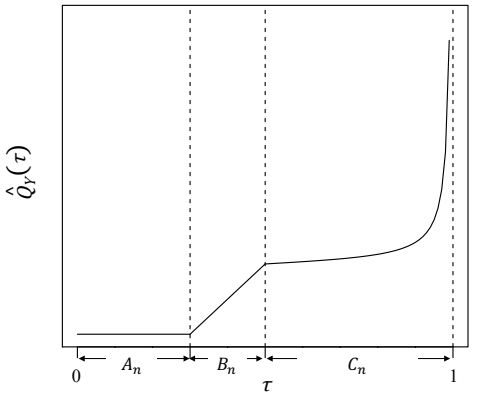


Figure 2. Piecewise estimator of the conditional quantile function $Q_Y(\tau|\boldsymbol{x})$.

defined by

$$\begin{aligned}\widehat{Q}_Y(\tau|\boldsymbol{x}) = {} & 0 \cdot I\{\tau \in A_n\} \\ & + \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}_n \circ \Gamma(1 - \pi(\widehat{\boldsymbol{\gamma}}_n, \boldsymbol{x}) + n^{-\delta}; \boldsymbol{x}, \widehat{\boldsymbol{\gamma}}_n) \\ & \cdot \frac{\tau - \{1 - \pi(\widehat{\boldsymbol{\gamma}}_n, \boldsymbol{x})\}}{n^{-\delta}} \cdot I\{\tau \in B_n\} \\ & + \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}_n \circ \Gamma(\tau; \boldsymbol{x}, \widehat{\boldsymbol{\gamma}}_n) \cdot I\{\tau \in C_n\}. \end{aligned} \tag{2.6}$$

The first and third pieces of the estimator (2.6) correspond to the two parts in (2.4), while the second piece is a linear interpolation between zero and the conditional quantile $\boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}_n \circ \Gamma(1 - \pi(\widehat{\boldsymbol{\gamma}}_n, \boldsymbol{x}) + n^{-\delta}; \boldsymbol{x}, \widehat{\boldsymbol{\gamma}}_n)$. The width of the interpolation window, $n^{-\delta}$, is designed to converge more slowly than the convergence rate of $\widehat{\boldsymbol{\gamma}}_n$, so that we do not need to estimate at the problematic change point, $1 - \pi(\widehat{\boldsymbol{\gamma}}_n, \boldsymbol{x})$. In Section 2.2.1, we establish the asymptotic properties of the estimator $\widehat{Q}_Y(\tau|\boldsymbol{x})$ in (2.6).

2.2.1. Asymptotic properties of $\widehat{Q}_Y(\tau|\boldsymbol{x})$

In this subsection, we establish the asymptotic properties of $\widehat{Q}_Y(\tau|\boldsymbol{x})$, where $\boldsymbol{x}$ denotes a placeholder. We first make the following assumptions:

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

Assumption 2. *The conditional distribution function $F_Y(\cdot|\boldsymbol{x}, Y > 0)$ is absolutely continuous with a positive continuous density $f_{Y|Y>0}(\cdot|\boldsymbol{x})$ on $[0, \infty)$.*

Assumption 3. *The conditional quantile function has the property*

$$\lim_{\tau \to 0^+} Q_Y(\tau|\boldsymbol{x}, Y > 0) = 0.$$

Assumption 4. *The quantile coefficient function* $\boldsymbol{\beta}(\tau)$ *is differentiable at* $\forall\, \tau \in (0,1)$*, with a bounded first derivative that* $\sup_{\tau\in(0,1)} \dot{\boldsymbol{\beta}}(\tau) = \sup_{\tau\in(0,1)}(\mathrm{d}\boldsymbol{\beta}(t)/\mathrm{d}t) \,|_{t=\tau} < \infty$.

Assumption 5. $\|\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top)\|_\infty < \infty$.

Assumption 2 is borrowed from Theorem 4.1 of Koenker (2005) to ensure the validity of the linear quantile regression on the positive part, and it incorporates Theorem 1 on Page 640 of Shorack and Wellner (1986) to help establish the limiting distribution at the change point. Assumption 3 is the connectivity constraint stated in (2.3). Assumptions 2 and 3 together ensure a nonnormal limiting distribution of $\widehat{Q}_Y(\tau|\boldsymbol{x})$ at the special quantile level $\tau = 1 - \pi(\boldsymbol{\gamma}, \boldsymbol{x})$. Assumption 5 ensures that the following matrices exist and are positive definite:

$$\boldsymbol{D}_{1,\boldsymbol{\beta}(\tau)} = \mathbb{E}\Big[\pi(\boldsymbol{\gamma}, \boldsymbol{X}) f_{Y|Y>0}\{\boldsymbol{X}^\top \boldsymbol{\beta}(\tau) \mid \boldsymbol{X}\}\boldsymbol{X}\boldsymbol{X}^\top\Big], \tag{2.7}$$

$$\boldsymbol{D}_0 = \mathbb{E}\Big\{\pi(\boldsymbol{\gamma}, \boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^\top\Big\}, \tag{2.8}$$

$$\boldsymbol{D}_{1,\boldsymbol{\gamma}} = \mathbb{E}\Big[\pi(\boldsymbol{\gamma}, \boldsymbol{X})\{1 - \pi(\boldsymbol{\gamma}, \boldsymbol{X})\}\boldsymbol{X}\boldsymbol{X}^\top\Big]. \tag{2.9}$$

Assumptions 4 and 5 both impose constraints on the quantities involved in the theory of $\widehat{Q}_Y(\tau|\boldsymbol{x})$, for any $\tau > 1 - \pi(\boldsymbol{\gamma}, \boldsymbol{x})$, ensuring a normal asymptotic distribution. Then, we have Theorem 1, the proof of which is provided in the online Supplementary Material.

Theorem 1. *Under Model* (2.4) *and Assumptions* 1–5*, for any given* $\tau \in (0,1)$*, we have*

(i) $\widehat{Q}_Y(\tau|\boldsymbol{x})$ *is a consistency estimator; that is, as* $n \to \infty$*,*

$$\widehat{Q}_Y(\tau|\boldsymbol{x}) \underset{p}{\to} Q_Y(\tau|\boldsymbol{x}).$$

(ii) $\widehat{Q}_Y(\tau|\boldsymbol{x})$ *has different limiting distributions, given different relationships between* τ *and* $\pi(\boldsymbol{\gamma}, \boldsymbol{x})$*:*

(a) If $\tau < 1 - \pi(\boldsymbol{\gamma}, \boldsymbol{x})$*,* $\widehat{Q}_Y(\tau|\boldsymbol{x})$ *is super-efficient; that is, as* $n \to \infty$*,*

$$\sqrt{n}\Big(\widehat{Q}_Y(\tau|\boldsymbol{x}) - 0\Big) \underset{p}{\to} 0.$$

(b) If $\tau = 1 - \pi(\boldsymbol{\gamma}, \boldsymbol{x})$*, denote* $Q'_Y(0|\boldsymbol{x}, Y > 0)$ *as the right derivative, which is well defined because* $\boldsymbol{\beta}(\tau)$ *is right differentiable at zero. Then as* $n \to \infty$*,*

$$\sqrt{n}\Big(\widehat{Q}_Y(\tau|\boldsymbol{x}) - 0\Big) \underset{d}{\to} \{1 - \pi(\boldsymbol{\gamma}, \boldsymbol{x})\}\sqrt{\boldsymbol{x}^\top \boldsymbol{D}_{1,\boldsymbol{\gamma}}^{-1}\boldsymbol{x}}\; Q'_Y(0|\boldsymbol{x}, Y > 0) Z_0 I\{Z_0 > 0\},$$

where $\boldsymbol{D}_{1,\boldsymbol{\gamma}}$ *is defined in* (2.9) *and* $Z_0 \sim N(0,1)$*.*

(c) If $\tau > 1 - \pi(\boldsymbol{\gamma}, \boldsymbol{x})$*, as* $n \to \infty$*,*

$$\sqrt{n}\Big(\widehat{Q}_Y(\tau|\boldsymbol{x}) - Q_Y(\tau|\boldsymbol{x})\Big) \underset{d}{\to} N(0,\, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2),$$

where

$$\begin{aligned}
\boldsymbol{\Sigma}_1 &= \Gamma(\tau; \boldsymbol{x}, \boldsymbol{\gamma})\big\{1 - \Gamma(\tau; \boldsymbol{x}, \boldsymbol{\gamma})\big\}\boldsymbol{x}^\top \boldsymbol{D}^{-1}_{1,\boldsymbol{\beta}\circ\Gamma(\tau;\boldsymbol{x},\boldsymbol{\gamma})} \boldsymbol{D}_0 \boldsymbol{D}^{-1}_{1,\boldsymbol{\beta}\circ\Gamma(\tau;\boldsymbol{x},\boldsymbol{\gamma})}\boldsymbol{x}, \\
\boldsymbol{\Sigma}_2 &= \big\{1 - \Gamma(\tau; \boldsymbol{x}, \boldsymbol{\gamma})\big\}^2\big\{1 - \pi(\boldsymbol{\gamma}, \boldsymbol{x})\big\}^2 \boldsymbol{x}^\top \boldsymbol{D}^{-1}_{1,\boldsymbol{\gamma}}\boldsymbol{x} \\
&\quad \cdot \boldsymbol{x}^\top \dot{\boldsymbol{\beta}} \circ \Gamma(\tau; \boldsymbol{x}, \boldsymbol{\gamma})\, \dot{\boldsymbol{\beta}} \circ \Gamma(\tau; \boldsymbol{x}, \boldsymbol{\gamma})^\top \boldsymbol{x},
\end{aligned}$$

and $\boldsymbol{D}_{1,\boldsymbol{\beta}(\tau)}, \boldsymbol{D}_0$*, and* $\boldsymbol{D}_{1,\boldsymbol{\gamma}}$ *are defined in* (2.7)*,* (2.8)*, and* (2.9)*, respectively.*

Note that at the change point $\tau = 1 - \pi(\boldsymbol{\gamma}, \boldsymbol{x})$, $\widehat{Q}_Y(\tau|\boldsymbol{x})$ follows a zero-inflated half-normal limiting distribution with a variance determined by the variation from the logistic regression and the right derivative of the conditional quantile at zero. For $\tau > 1 - \pi(\boldsymbol{\gamma}, \boldsymbol{x})$, the asymptotic distribution is normal, while the two components of the asymptotic variance, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, are composites of variations in the logistic and quantile regression models. In contrast to the standard asymptotic results of the linear quantile regression, the Hessian matrix, $\boldsymbol{D}_{1,\boldsymbol{\beta}(\tau)}$ in (2.7) is evaluated on the conditional density given $Y > 0$, and adjusted using the individual zero inflation rate, $\pi(\boldsymbol{\gamma}, \boldsymbol{X})$. The Jacobian matrix, $\boldsymbol{D}_0$ in (2.8), is also adjusted using the subject-specific zero inflation. Because only the positive y contribute to the quantile regression model fitting, $\pi(\boldsymbol{\gamma}, \boldsymbol{X})$ can be regarded as a propensity score, and adjusts the contribution of each observation when estimating the covariance matrix.

2.2.2. Choice of δ

The estimation of the conditional quantile function $\widehat{Q}_Y(\tau|\boldsymbol{x})$ involves a nuisance parameter δ. An inappropriate choice of δ could introduce bias into the

conditional quantile estimation via the linear interpolation on B_n. A large δ that approaches 1/2 is preferred, because a faster convergent interpolation area induces a smaller bias. However, a δ that is too large would inflate the variance of the estimated quantile around the change point, leading to unstable estimation. Thus, we recommend choosing $\delta = 0.499$.

If predicting future outcome values is of interest, we can perform a cross-validation on a grid of potential δ to determine the optimal choice. Details of prediction methods based on the proposed model and the corresponding measure of prediction quality are discussed in Section 2.3.2.

2.2.3. Constrained post-estimation smoothing

The piecewise estimator outlined in (2.6) guarantees a consistent estimation of the quantile function. The estimated function, however, is nonsmoothing. To achieve a smooth estimation, one can take advantage of the constrained B-spline smoothing (COBS) introduced by He and Ng (1999).

We propose estimating the linear quantile model (2.2) on a sequence of k_n evenly spaced quantile levels $[1/(k_n+1), k_n/(k_n+1)]$, where $k_n = o(n^{1/2})$, which is slightly finer than $n^{-\delta}$. Then, we construct $\widehat{\boldsymbol{\beta}}_n(\tau)$ as the linear spline expanded from the estimated quantile coefficients. As shown by Wei and Carroll (2009), $\widehat{\boldsymbol{\beta}}_n(\tau)$ is a uniformly consistent estimator of $\boldsymbol{\beta}(\tau)$. Next, we apply COBS to the resulting $\tilde{Q}_Y(\tau|\boldsymbol{x}, Y > 0) = \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}_n(\tau)$ to obtain the smoothed $\widehat{Q}_Y(\tau|\boldsymbol{x}, Y > 0)$. After matching the k_n nominal quantile levels to the target quantile levels based on the estimated probability of observing a positive Y, we can obtain the final estimate $\widehat{Q}_Y(\tau|\boldsymbol{x})$.

The asymptotic properties of the smooth estimator are not discussed in this paper. Its finite sample performance is not inferior to that of the nonsmooth version, as shown in the simulation studies in Section 3 and the real data application in Section 4.

2.3. Model-based inference and prediction

2.3.1. Average quantile effect and its estimation

Under the proposed two-part model, the covariates $\boldsymbol{X}$ could influence the conditional quantiles of Y in two ways: changing the probability of observing a positive Y, and changing the quantiles of $Y|Y > 0$. Consequently, as Model (2.4) shows, the quantile effect of a covariate X_j depends on the actual value of X_j, and also varies with the levels of other covariates $\boldsymbol{X}^{(-j)}$, where $\boldsymbol{X}^{(-j)}$ stands for the covariates excluding X_j. Hence, we define the average quantile effect (AQE)

of the covariate X_j by

$$\Delta_\tau(X_j;u,v)=\mathbb{E}_{\boldsymbol{X}^{(-j)}}\left\{Q_Y(\tau|X_j=u,\boldsymbol{X}^{(-j)})-Q_Y(\tau|X_j=v,\boldsymbol{X}^{(-j)})\right\}. \quad (2.10)$$

The AQE, $\Delta_\tau(X_j;u,v)$, is the marginal change of the τth quantile of Y due to the change of X_j from v to u. When X_j is the treatment assignment, coded as one for the treatment and zero for the placebo, we have the average quantile treatment effect (AQTE),

$$\Delta_\tau(X_j;1,0)=\mathbb{E}_{\boldsymbol{X}^{(-j)}}\left\{Q_Y(\tau|X_j=1,\boldsymbol{X}^{(-j)})-Q_Y(\tau|X_j=0,\boldsymbol{X}^{(-j)})\right\}, \quad (2.11)$$

which is the expected quantile treatment effect in the target population.

A natural sample estimator of the AQE is

$$\widehat{\Delta}_\tau(X_j;u,v)=\frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{Q}_Y(\tau|X_j=u,\boldsymbol{x}_i^{(-j)})-\widehat{Q}_Y(\tau|X_j=v,\boldsymbol{x}_i^{(-j)})\right\}, \quad (2.12)$$

where $\widehat{Q}_Y(\cdot)$ is the estimated conditional quantile function defined in (2.6). In what follows, we provide the asymptotic properties of $\widehat{\Delta}_\tau(X_j;u,v)$. We first make the following assumption:

Assumption 6. The coefficient functions $\boldsymbol{\beta}\circ\Gamma(\tau;X_j,\boldsymbol{X}^{(-j)},\boldsymbol{\gamma})$ are smooth functions of $\boldsymbol{X}^{(-j)}$ with compact supports.

Then, we have Theorem 2, the proof of which is deferred to the Supplementary Material.

Theorem 2. *At a given quantile level $\tau\in(0,1)$, $\widehat{\Delta}_\tau(X_j;u,v)$ is the estimator constructed in* (2.12) *for the AQE defined in* (2.10)*. Under Assumptions 1–6, there exists a tight process $G(\boldsymbol{X}^{(-j)})$, indexed by $\boldsymbol{X}^{(-j)}$, such that*

$$\sqrt{n}\Big(\widehat{\Delta}_\tau(X_j;u,v)-\Delta_\tau(X_j;u,v)\Big)\underset{d}{\to}\int G(\boldsymbol{X}^{(-j)})dP_{\boldsymbol{X}^{(-j)}}.$$

If the distribution $P_{\boldsymbol{X}^{(-j)}}$ of $\boldsymbol{X}^{(-j)}$ is absolutely continuous w.r.t. the Lebesgue measure, we have

$$\sqrt{n}\Big(\widehat{\Delta}_\tau(X_j;u,v)-\Delta_\tau(X_j;u,v)\Big)\underset{d}{\to}\int G(\boldsymbol{X}^{(-j)})dP_{\boldsymbol{X}^{(-j)}}=N(0,\sigma^2),$$

where

$$\sigma^2=\int\int \mathrm{Cov}\{G(\boldsymbol{X}^{(-j)}),G(\boldsymbol{X}^{*(-j)})\}\,d\,\mathbb{P}_{\boldsymbol{X}^{(-j)}}\,d\,\mathbb{P}_{\boldsymbol{X}^{*(-j)}},$$

with $\boldsymbol{X}^{(-j)}$ *and* $\boldsymbol{X}^{*(-j)}$ *i.i.d. under* $\mathbb{P}_{\boldsymbol{X}^{(-j)}}$.

Although the asymptotic variance, σ^2, can be decomposed into tractable components, estimating it directly is complicated. In practice, we use the paired bootstrap to numerically construct a bootstrap percentile interval, and then conduct hypothesis testing for the marginal covariate effect accordingly.

2.3.2. Prediction

Accurate clinical predictions are of great importance and interest in medical applications. Owing to its percentile interpretation, a conditional quantile function can be conveniently used to construct prediction intervals. Let $\boldsymbol{x}_{\text{new}}$ be the covariate profile of a new patient. We can construct the $(1-\alpha)\times 100\%$ level prediction interval of his/her outcome as

$$\left[\widehat{Q}_Y\left(\frac{\alpha}{2}\middle|\boldsymbol{x}_{\text{new}}\right),\ \widehat{Q}_Y\left(1-\frac{\alpha}{2}\middle|\boldsymbol{x}_{\text{new}}\right)\right],$$

where $\widehat{Q}_Y(\cdot)$ is the estimated conditional quantile function defined in (2.6). In addition, we propose using the conditional median,

$$\widehat{m}_{Y|\boldsymbol{x}_{\text{new}}} = \widehat{Q}_Y(0.5|\boldsymbol{x}_{\text{new}}),$$

as the predicted value. Conventionally, the predicted value is defined as the estimated conditional mean given $\boldsymbol{x}_{\text{new}}$, which can be estimated by integrating the conditional quantile function, that is, $\widehat{\mu}_{Y|\boldsymbol{x}_{\text{new}}} = \int_0^1 \widehat{Q}_Y(\tau|\boldsymbol{x}_{\text{new}})\,d\tau$. However, owing to the zero-inflated nature of the outcome, the conditional median would be a better choice. For example, if a subject has over 80% chance of obtaining a zero outcome, given his/her covariate profile, zero (the conditional median) could be a more sensible prediction than the mean-based one. Consequently, to achieve an optimal prediction, we use cross-validations with some chosen measures to select the covariates and nuisance parameters δ.

3. Simulation

3.1. Simulation settings

In this section, we present a numerical study to illustrate the finite-sample performance of the proposed methods, and to compare this with that of the direct quantile regression and existing parametric models for zero-inflated outcomes. We generate simulated data in the context of the carotid plaque data, with echodensity (plaqden) as the outcome, and male and systolic blood pressure (systolic) as

covariates. For each sample, we first generate the discrete covariate, male, from Bernoulli(0.5), and the continuous covariate, systolic, from $N(150, 15^2)$. We then generate a binary indicator D from a Bernoulli trial with the success probability

$$P(D = 1|\boldsymbol{X}) = \pi(\gamma, \boldsymbol{X}) = \frac{\exp(-1.92 + 0.19\,\text{male} + 0.02\,\text{systolic})}{1 + \exp(-1.92 + 0.19\,\text{male} + 0.02\,\text{systolic})},$$

where $\boldsymbol{X} = (\text{male}, \text{systolic})^\top$, and the parameters $\boldsymbol{\gamma} = (-1.92, 0.19, 0.02)^\top$ were estimated based on the carotid plaque echodensity of the NOMAS data. For a sample with $D = 1$, we generate plaqden from the conditional quantile function

$$Q_{\text{plaqden}}(\tau|\boldsymbol{X}, \text{plaqden} > 0) = \beta_0(\tau) + \beta_1(\tau)\,\text{male} + \beta_2(\tau)\,\text{systolic},$$

where the true coefficient functions, $\boldsymbol{\beta}(\tau)$, are estimated based on echodensity again, and plotted in Figure S1 of the Supplementary Material. Specifically, we randomly draw a variable U from $U(0, 1)$, and then generate a positive value of plaqden as $\beta_0(U) + \beta_1(U)\,\text{male} + \beta_2(U)\,\text{systolic}$. For a sample with $D = 1$, we assign the value zero to plaqden. We generate $n = 500$ random samples in one data set, and repeat the simulation process 1,000 times.

We compare the proposed methods to the following existing approaches: (1) direct quantile regression, (2) ZIP regression, (3) hurdle regression, and (4) CPG regression. The direct quantile regression assumes the outcome is absolutely continuous. When the data contain a probability mass at zero, the estimation algorithm often fails to converge. To circumvent this numerical difficulty, we add a small perturbation ($\sim N(0, 10^{-14})$) to the zero-valued outcomes, and apply the linear quantile regression to the perturbed data directly. To use the ZIP/hurdle models designed for count data, we round the outcomes to integers before estimation. Though the semi-continuous setting does not favor ZIP/hurdle models, we select them as comparisons because they represent the widely applied parametric mixture/two-part model, and the rounding does not substantially affect their results.

3.2. Estimation of conditional quantile functions

In this section, we compare the estimation accuracy of the conditional quantiles by the various methods. We estimate the quantile functions given 10 covariate profiles, which are formed by male $\in \{0, 1\}$ and systolic $\in \{130.78, 139.88, 150.00, 160.12, 169.22\}$ (the 0.10th, 0.25th, 0.50th, 0.75th, and 0.90th empirical quantiles of systolic blood pressure in the NOMAS data). We consider three

measures to assess the estimation performance:

$$\mathrm{RIMSE}_{\widehat{Q}} = \frac{\int \mathbb{E}\{\widehat{Q}_Y(\tau|\boldsymbol{X}) - Q_Y(\tau|\boldsymbol{X})\}^2 d\tau}{\int Q_Y(\tau|\boldsymbol{X})^2 d\tau},$$
$$\mathrm{RIBias}^2_{\widehat{Q}} = \frac{\int \{\mathbb{E}\,\widehat{Q}_Y(\tau|\boldsymbol{X}) - Q_Y(\tau|\boldsymbol{X})\}^2 d\tau}{\int Q_Y(\tau|\boldsymbol{X})^2 d\tau},$$
$$\mathrm{RIVar}_{\widehat{Q}} = \frac{\int \mathbb{E}\{\widehat{Q}_Y(\tau|\boldsymbol{X}) - \mathbb{E}\,\widehat{Q}_Y(\tau|\boldsymbol{X})\}^2 d\tau}{\int Q_Y(\tau|\boldsymbol{X})^2 d\tau},$$

where $\mathrm{RIMSE}_{\widehat{Q}}$ is the relative integrated mean squared error, $\mathrm{RIBias}^2_{\widehat{Q}}$ is the relative integrated bias squared, and $\mathrm{RIVar}_{\widehat{Q}}$ is the relative integrated variance.

Table 1 reports $\mathrm{RIMSE}_{\widehat{Q}}$, $\mathrm{RIBias}^2_{\widehat{Q}}$, and $\mathrm{RIVar}_{\widehat{Q}}$ of the estimated conditional quantile functions using the proposed estimation with smoothing, proposed estimation without smoothing, direct quantile regression, and competing parametric approaches, based on the 10 sets of covariate values. In general, the proposed methods have much smaller biases than that of the direct quantile regression. The reduction in bias by the nonsmooth estimation is 0.24%, on average, across the 10 cases (0.10% vs. 0.84%, 0.05% vs. 0.33%,..., 0.03% vs. 0.33%). By the smooth estimation, the mean reduction of the bias is also 0.24% (0.09% vs. 0.84%, 0.04% vs. 0.33%, ..., 0.04% vs. 0.33%). Note that the proposed methods show more noticeable advantages when systolic is assumed to contain more extreme values. For example, with (male, systolic)=(0, 130.78), the bias of the nonsmooth proposed estimator is 0.10%, while that of the direct method is 0.84%, yielding a reduction in bias of 0.74%. However, with (male, systolic)=(1, 150.00), the reduction is only 0.08% (0.03% vs. 0.11%). Furthermore, while the proposed nonsmooth estimation leads to larger variances, the additional post-estimation smoothing can reduce the variance by approximately 0.08%, on average (1.07% vs. 1.19, 0.64% vs. 0.73%, ..., 0.44% vs. 0.52%).

We also note that the bias of the direct method is more evident around the change point, that is, the very τ where the quantile function $Q_Y(\tau|\boldsymbol{X})$ changes from zero to positive. To investigate this in detail, we evaluate $\mathrm{RIMSE}_{\widehat{Q}}$, $\mathrm{RIBias}^2_{\widehat{Q}}$, and $\mathrm{RIVar}_{\widehat{Q}}$ in an interval of half-length of 0.1 around the change point. The results are summarized in Table S1 in the Supplementary Material. As shown in Table S1, the bias in the neighborhood of the change point is remarkably reduced by the proposed methods compared to the direct approach, especially for the covariates with more extreme values.

In general, the parametric approaches perform poorly, even worse than the direct quantile regression (Table 1 and S1). This is because the performance of

Table 1. Summary of RIMSE(%), RIBias^2(%), and RIVar(%) of the estimated conditional quantile functions of echodensity on the entire $Q_Y(\tau|\boldsymbol{X})$.

	Proposed (smooth)			Proposed (nonsmooth)			Direct		
(gender, systolic)	RIMSE	RIBias^2	RIVar	RIMSE	RIBias^2	RIVar	RIMSE	RIBias^2	RIVar
$(0, 130.78)$	1.16	0.09	1.07	1.28	0.10	1.19	1.45	0.84	0.61
$(0, 139.88)$	0.68	0.04	0.64	0.77	0.05	0.73	0.79	0.33	0.45
$(0, 150.00)$	0.48	0.03	0.45	0.55	0.03	0.52	0.56	0.18	0.39
$(0, 160.12)$	0.48	0.03	0.45	0.56	0.03	0.53	0.57	0.14	0.43
$(0, 169.22)$	0.58	0.04	0.54	0.68	0.04	0.64	0.81	0.27	0.54
$(1, 130.78)$	0.91	0.07	0.84	0.99	0.07	0.92	0.85	0.28	0.57
$(1, 139.88)$	0.55	0.04	0.51	0.61	0.04	0.57	0.61	0.21	0.40
$(1, 150.00)$	0.39	0.03	0.36	0.44	0.03	0.41	0.43	0.11	0.32
$(1, 160.12)$	0.39	0.03	0.36	0.45	0.03	0.42	0.47	0.12	0.35
$(1, 169.22)$	0.47	0.04	0.44	0.56	0.03	0.52	0.79	0.33	0.46
	ZIP			Hurdle			CPG		
(gender, systolic)	RIMSE	RIBias^2	RIVar	RIMSE	RIBias^2	RIVar	RIMSE	RIBias^2	RIVar
$(0, 130.78)$	5.94	3.65	2.29	5.94	3.65	2.29	6.21	5.84	0.37
$(0, 139.88)$	5.43	3.81	1.61	5.43	3.81	1.61	5.59	5.32	0.27
$(0, 150.00)$	5.18	3.90	1.27	5.18	3.90	1.27	5.16	4.94	0.22
$(0, 160.12)$	5.09	3.83	1.26	5.09	3.83	1.26	4.97	4.74	0.23
$(0, 169.22)$	5.11	3.73	1.38	5.11	3.73	1.38	4.99	4.69	0.30
$(1, 130.78)$	6.88	4.85	2.03	6.88	4.85	2.03	4.59	4.23	0.36
$(1, 139.88)$	6.49	5.05	1.45	6.49	5.04	1.45	4.10	3.83	0.27
$(1, 150.00)$	6.31	5.18	1.13	6.31	5.18	1.13	3.77	3.55	0.21
$(1, 160.12)$	6.25	5.16	1.09	6.25	5.16	1.09	3.63	3.40	0.23
$(1, 169.22)$	6.26	5.10	1.17	6.26	5.10	1.17	3.71	3.40	0.30

the parametric methods depends on whether the model assumptions are satisfied. Neither the mixture (or combination) of zeros and the Poisson distribution nor the CPG distribution is an appropriate model for echodensity in this simulation.

3.3. Point and interval estimations of AQTE

In this section, we compare the point and interval estimates of the average quantile treatment effect (AQTE) of being male by the various methods. With each simulated data set, the point estimate of AQTE is computed as stated in (2.12). Next, in each of the 1,000 simulation runs, we conduct 500 paired bootstraps and construct the $(1-\alpha)\times 100\%$-level bootstrap percentile confidence interval of the estimated AQTE

$$\left[\widehat{\Delta}_\tau^{(B)}(\text{male}; 1, 0)_{\alpha/2}, \widehat{\Delta}_\tau^{(B)}(\text{male}; 1, 0)_{1-\alpha/2}\right],$$

with the estimated AQTE based on each of the bootstrapped data sets

$$\widehat{\Delta}_{\tau}^{(B)}(\text{male}; 1, 0) \tag{3.1}$$
$$= \frac{1}{n}\sum_{i=1}^{n}\{\widehat{Q}_{\text{plaqden}}^{(B)}(\tau|\text{male} = 1, \text{systolic}_i^{(B)}) - \widehat{Q}_{\text{plaqden}}^{(B)}(\tau|\text{male} = 0, \text{systolic}_i^{(B)})\}.$$

Note that Theorem 2 in Section 2.3.1 guarantees a normal limiting distribution of the estimated AQTE because systolic follows a continuous distribution. Here, we set $\alpha = 0.10$ and use a grid of representative quantile levels, $\tau = 0.10, 0.25, 0.50, 0.75, 0.90$. For the parametric approaches, we can estimate any quantity of the conditional distribution based on the estimated parameters, together with the instance (male=0 or 1, $\text{systolic}_i^{(B)}$). Therefore, using the ZIP, hurdle, and CPG regression models, we can also estimate the conditional quantiles and estimate the AQTE, as stated in (3.1). We use three measures to evaluate the inference performance for the AQTE: (1) the bias of the average estimate of the AQTE, (2) the coverage rate of the 90% bootstrap percentile confidence interval, and (3) the average length of the confidence interval.

As Table 2 shows, the proposed methods provide the most accurate estimates of the AQTEs on all five quantiles. In addition, their coverage rates are all close to the nominal level, 90%. Though the direct quantile regression gives the best estimate, zero, at $\tau = 0.10$, its coverage rate is 0%. This reflects the fact that the direct method cannot capture the different levels of zero inflation between different covariate profiles at lower quantiles of the outcome. The coverage rates of the ZIP and hurdle regressions at higher quantiles are remarkably lower than 90%, which signifies their limitation in describing the extreme tails of outcome distributions. The CPG model produces the worst coverage rates when $\tau = 0.25, 0.50$. Although the coverage of the ZIP and hurdle regressions at lower quantiles and that of the CPG model at extreme quantiles are close to the nominal rate, the average lengths of the intervals are much wider than those of the proposed approaches. Thus, for making inferences based on the AQTE, the proposed quantile regression model outperforms the direct quantile regression and existing parametric methods in all respects.

4. Analysis of the Carotid Plaque Data

In this section, we apply the proposed method to analyze the motivating carotid plaque data, NOMAS, presented in Section 1, to examine how various risk factors affect carotid atherosclerosis. The risk factors considered include high-density lipoprotein, triglyceride, low-density lipoprotein, race and ethnicity,

Table 2. Summary of the average estimate, bias, coverage rate (%) of the 90% bootstrap percentile confidence interval, and the average length of the interval for the AQTE of male on the τth quantile of echodensity.

τ	AQTE	Measure	Proposed (smooth)	Proposed (nonsmooth)	Direct	ZIP	Hurdle	CPG
		Estimate	0.04	0.04	0.00	0.35	0.35	-1.42
0.10	0.0036	Bias	0.04	0.04	0.00	0.35	0.35	-1.42
		Coverage	85.80	85.80	0.00	86.70	86.70	86.30
		Length	0.67	0.77	0.00	4.36	4.36	15.23
		Estimate	6.52	6.51	5.94	13.72	13.72	-0.81
0.25	7.5189	Bias	-1.00	-1.01	-1.58	6.20	6.20	-8.33
		Coverage	89.10	89.30	89.00	89.80	89.80	34.30
		Length	36.34	39.49	39.47	64.37	64.37	8.89
		Estimate	-6.48	-6.53	-7.00	-4.36	-4.36	-1.19
0.50	-6.6304	Bias	0.15	0.10	-0.37	2.27	2.27	5.44
		Coverage	90.60	90.70	91.60	82.10	82.10	62.40
		Length	15.86	15.49	14.11	11.16	11.16	13.12
		Estimate	-3.22	-3.26	-3.23	-5.14	-5.14	-1.57
0.75	-2.9669	Bias	-0.25	-0.29	-0.26	-2.17	-2.17	1.40
		Coverage	91.60	92.60	93.10	81.10	81.10	83.50
		Length	11.73	11.33	11.74	10.58	10.58	16.75
		Estimate	-1.38	-1.55	-1.57	-5.46	-5.46	-1.88
0.90	-0.9258	Bias	-0.45	-0.62	-0.64	-4.53	-4.53	-0.95
		Coverage	90.20	93.40	93.80	60.10	60.10	89.20
		Length	13.21	16.55	17.03	10.89	10.89	19.86

diabetes, blood pressure, smoking status, higher education, and body mass index. We present the estimated AQEs of the risk factors in Section 4.1, the model fitness in Section 4.2, and the prediction performance in Section 4.3.

4.1. Estimated AQEs

We apply the proposed quantile regression (without smoothing) to study the carotid plaque echodensity (plaqden) and to estimate the risk factors' AQEs following (2.12) in Section 2.3.1. The estimated AQEs for individual covariates are plotted in Figure 3. The black solid and red dashed lines represent the quantile functions of echodensity given two distinctive covariate values of interest. The gray area indicates the range of quantile levels where the corresponding AQE reaches the 95% pointwise significance. As shown in Figure 3, race and ethnicity, diabetes, systolic and diastolic blood pressure, smoking status, body mass index, and glomerular filtration rate significantly impact echodensity across all quantile

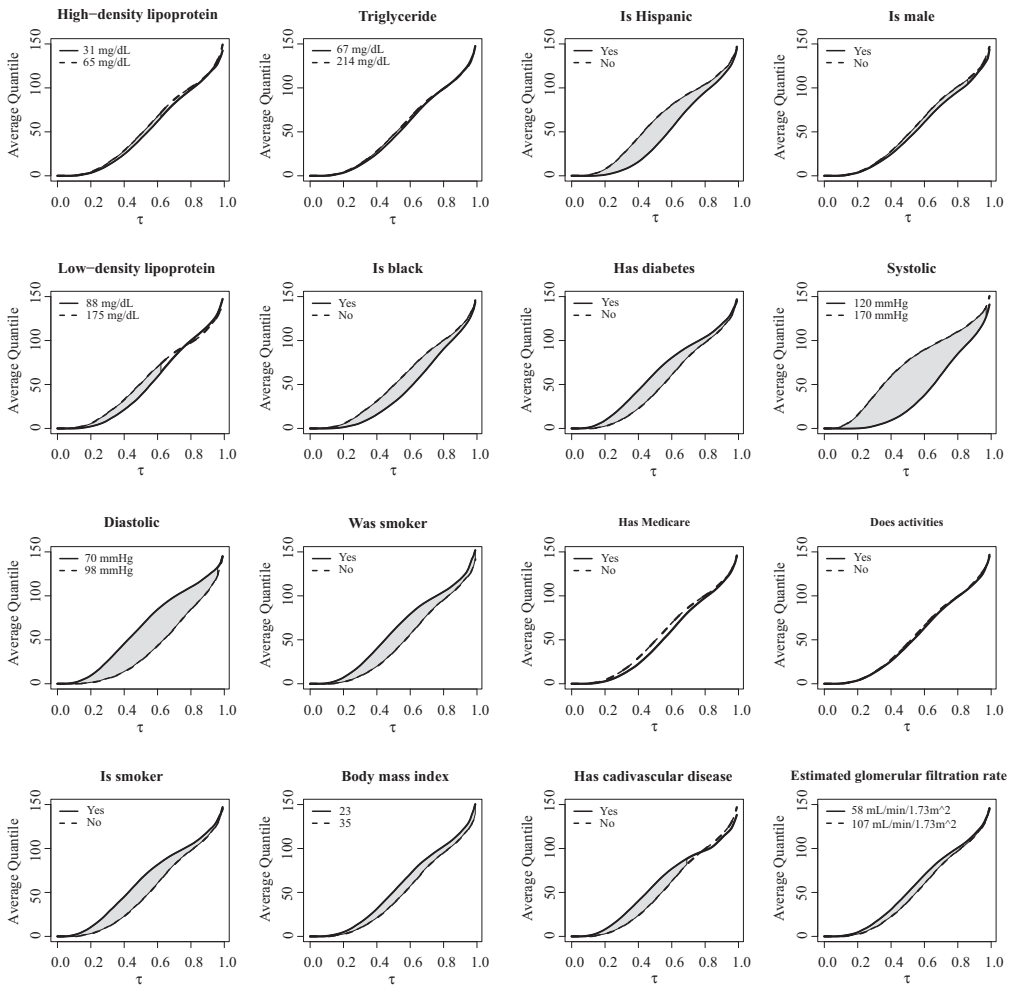


Figure 3. Estimated AQEs of selected covariates by the proposed method (without smoothing) on all quantiles of echodensity, which are presented as the differences between the dashed and solid lines. Significant AQEs are highlighted by the shaded area.

levels. Moreover, the effects vary with the quantile level.

For example, although it is well known that a higher systolic blood pressure is associated with a higher value of plaque echodensity, the quantile-specific effects provide a better understanding of how systolic blood pressure affects the texture of the plaque. The most significant difference between the two levels (120 mmHg vs. 170 mmHg) is around the median. For patients with a systolic blood pressures of 120 mmHg, 50% have their echodensity controlled below 20, while 50% of the patients with a systolic blood pressure of 170 mmHg have echodensity

over 80. The difference becomes smaller as the quantile level increases, suggesting that the risk of an extreme plaque burden is comparable between the two levels. Our analysis also reveals that the risk of having a positive echodensity is different between the two levels. Among individuals with a systolic blood pressure of 170 mmHg, there is a substantial likelihood of having positive plaque, while 20% of those with a systolic blood pressure of 120 mmHg are expected to have zero plaque. The direct quantile regression would miss this difference. To illustrate, we plot the estimated AQEs of systolic blood pressures using the proposed method and the direct quantile method (Figure S2 in the Supplementary Material). As expected, the direct quantile method shows no difference in the risk of positive plaque and, consequently, underestimates the risk of a plaque burden among patients with a systolic blood pressure of 170 mmHg. On the other hand, though the parametric methods distinguish the risk of taking positive plaque, they provide a biased approximation of the tail events. As presented in Figure S2, the ZIP and hurdle models underestimate the risk of systolic blood pressure for severe patients with more accumulated plaque than 70% of individuals in the 120 and 170 mmHg groups, while the CPG model consistently overestimates the risk. To validate the inference by the proposed method, we check the model fitness in the next section.

4.2. Goodness-of-fit

To measure the goodness-of-fit of a model, we simulate the outcomes based on the estimated model, and compare the simulated outcomes with those observed from the data using histograms and Q-Q plots. If a model fits the data well, we expect the distributions of the simulated and the observed outcomes to be comparable. Such a visual goodness-of-fit assessment has been used in Heyman, Tabatabai and Lakshman (1992), and is an effective way to illustrate the goodness-of-fit of quantile models.

Figure S3a in the Supplementary Material shows that the proposed methods, with and without smoothing, provide the best fit to the echodensity data. The distributions of the simulated outcomes under parametric models are very different from the observed outcomes. Although the direct quantile regression provides a proper fit at the upper tail of the outcome distribution, it misses the lower tail (where the outcome tends to take the zero value). Interestingly, when the plaque area is the outcome (Figure S3b in the Supplementary Material), the direct quantile regression also misses the higher tail. This difference indicates that the two features, plaque echodensity and plaque area, have quite different distributions, and that the proposed method is robust and advantageous, regardless of

the outcome distributions.

4.3. Prediction

In this section, we use five-fold cross-validation to compare the prediction performance of the various methods. As outlined in Section 2.3.2, we predict an outcome using the estimated conditional median given the covariate profile, and construct the 95% prediction upper bound using the 0.95th conditional quantile. We use three measures to assess the prediction performance: (1) the correctly predicted rate for zero outcomes, (2) the coverage rate of the 95% prediction upper bound, and (3) its average length for positive outcomes. Because of the small perturbation added to the zero values, the direct quantile regression may predict negative values. We then treat the negative predictions as zeros.

As shown by Table S2 in the Supplementary Material, the proposed methods outperform the direct quantile regression in terms of predicting the zero outcomes of the two plaque burden features. The performance of the parametric approaches strongly relies on how well the parametric assumptions hold for the data. Although ZIP and hurdle regressions work as well as the proposed methods in predicting zeros, their 95% upper bounds have actual coverages of 82% and 84% for echodensity and plaque area, respectively. The CPG model correctly predicts most of the zeros for plaque area (Table S2, bottom), but it demonstrates poor performance in predicting the zeros for echodensity (Table S2, top). In conclusion, the proposed methods deliver reasonably good predictions for both zero and positive outcomes of echodensity and plaque area.

5. Conclusion

We have developed a model-based estimation algorithm, proposed a conditional quantile inference tool, and derived theoretical results of the estimation and inference for a two-part quantile regression model addressing zero-inflated outcomes. Although the model has been used previously in several applications, this study provides the first effort to investigate its theoretical properties and develop valid inference tools.

The piecewise estimation of the conditional quantile proposed in this paper involves a data-driven interpolation window around the change point from zero to positive. It is capable of handling features that are either zero or greater than a certain value. If that threshold T is known by some domain knowledge, we can interpolate to connect the endpoint value T and the estimated quantile $\boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}_n \circ \Gamma(1 - \pi(\widehat{\boldsymbol{\gamma}}_n, \boldsymbol{x}) + n^{-\delta}; \boldsymbol{x}, \widehat{\boldsymbol{\gamma}}_n)$. If T is unknown, it is robust to regard the

value at the left end as zero for interpolation, and the resulting estimator in (2.6) is asymptotically consistent at any fixed quantile level, by Theorem 1. When the τ of interest is close to the problematic change point where the conditional quantile changes from zero to positive, the limiting distribution is a zero-inflated half-normal, and the inference requires special care. If the target τ is beyond the change point, the estimated conditional quantile is asymptotically normal, and the inference can be made using standard methods. While it is not straightforward to make inferences about covariate effects in a two-part model, our AQE provides an effective way of quantifying the quantile treatment effects, conducting corresponding hypothesis testing, and constructing confidence intervals. In addition, we have provided tools for model-based predictions.

Using simulation studies and an analysis of carotid plaque data, we have shown that the proposed methods provide more accurate and robust estimations, better goodness-of-fit, more accurate predictions, and more accurate and comprehensive inferences than those of the direct quantile regression and existing parametric zero-inflated methods.

With a limited sample size, one possible concern is that the estimated conditional quantile function $\boldsymbol{X}^T\boldsymbol{\beta}(\tau)$ could be negative or nonmonotone for some $\boldsymbol{X}$ values and some quantile levels, especially when they are outlying in the covariate space, contradicting the fact that Y is nonnegative. In this case, one could follow a similar approach to that of Chernozhukov, Fernández-Val and Galichon (2010) to "rearrange" the estimated quantiles to ensure monotonicity and nonnegativity. Chernozhukov, Fernández-Val and Galichon (2010) has shown that, owing to the root-n convergence of the quantile estimates, such post-estimation rearrangement does not affect the asymptotic behaviors of the quantile estimations under fairly mild conditions. In applications where nonnegativity must be ensured for all $\boldsymbol{X}$, one could assume a linear quantile regression model on $\log(Y)$. However, the theories need to be carefully re-derived for such transformation quantile regressions. To ensure connectivity between zero and positive quantiles for any $\boldsymbol{X}$, $\boldsymbol{\beta}(\tau)$ should go to negative infinity around the change point, making the inference challenging. In fact, with adequate samples, even though one does not use a post-estimation rearrangement or model $\log(Y)$, the proposed model ensures that the resulting conditional quantile function is nonnegative almost surely. We have shown theoretically and numerically that the estimated quantile function converges to the true value as the sample size increases.

There are various interesting directions in which to extend the proposed methods. Although the main focus of this study is zero-inflated outcomes, the proposed methods can be easily extended to model outcomes with point masses at

multiple values. In addition, while the inference tools developed here are based on the parametric logistic model and the linear quantile model for the positive part, it would be interesting to examine inferences with the two models being replaced with semiparametric or nonparametric models. Finally, the interval estimation in this study is based on a pointwise inference. However, a simultaneous inference is possible by incorporating a minimum p-value procedure (Lee, Wu and Lin (2012)) that determines statistical significance based on the smallest p-value across all quantile levels and uses a resampling procedure to derive the threshold. One can also construct a joint χ^2 test statistic to test whether the logistic coefficients and quantile coefficients at multiple quantile levels are simultaneously equal to zero.

Supplementary Material

The online Supplementary Material contains the proofs of Theorems 1 and 2 and additional figures and tables.

Acknowledgments

This work was supported by the National Institutes of Health grant R01 HG008980 and by the National Science Foundation DMS-1953527.

References

Chernozhukov, V., Fernández-Val, I. and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica* **78**, 1093–1125.

Cheung, Y. K., Moon, Y. P., Kulick, E. R., Sacco, R. L., Elkind, M. S. and Willey, J. Z. (2017). Leisure-time physical activity and cardiovascular mortality in an elderly population in northern Manhattan: A prospective cohort study. *Journal of General Internal Medicine* **32**, 168–174.

Duan, N., Manning, W. G., Morris, C. N. and Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics* **1**, 115–126.

He, X. and Ng, P. (1999). Cobs: Qualitatively constrained smoothing via linear programming. *Computational Statistics* **14**, 315–337.

Heras, A., Moreno, I. and Vilar-Zanón, J. L. (2018). An application of two-stage quantile regression to insurance ratemaking. *Scandinavian Actuarial Journal* **2018**, 753–769.

Heyman, D. P., Tabatabai, A. and Lakshman, T. (1992). Statistical analysis and simulation study of video teleconference traffic in ATM networks. In *IEEE Transactions on Circuits and Systems for Video Technology* **2**, 49–59.

Jorgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)* **49**, 127–162.

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

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

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

Lee, S., Wu, M. C. and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775.

Lim, H. K., Li, W. K. and Philip, L. (2014). Zero-inflated poisson regression mixture model. *Computational Statistics & Data Analysis* **71**, 151–158.

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

Shorack, G. and Wellner, J. (1986). Empirical processes with applications to statistics. *John Wiley & Sons, New York*.

Wei, Y. and Carroll, R. J. (2009). Quantile regression with measurement error. *Journal of the American Statistical Association* **104**, 1129–1143.

Wodan Ling

Public Health Sciences Division, Fred Hutchinson Cancer Research Center, WA 98109, USA.

E-mail: wling@fredhutch.org

Bin Cheng

Department of Biostatistics, Mailman School of Public Health, Columbia University, NY 10032, USA.

E-mail: bc2159@cumc.columbia.edu

Ying Wei

Department of Biostatistics, Mailman School of Public Health, Columbia University, NY 10032, USA.

E-mail: yw2148@cumc.columbia.edu

Joshua Z. Willey

Department of Neurology, College of Physicians and Surgeons, Columbia University, NY 10032, USA.

E-mail: jzw2@cumc.columbia.edu

Ying Kuen Cheung

Department of Biostatistics, Mailman School of Public Health, Columbia University, NY 10032, USA.

E-mail: yc632@cumc.columbia.edu

(Received October 2020; accepted December 2020)