

Statistica Sinica Preprint No: SS-2020-0368

Title	STATISTICAL INFERENCE IN QUANTILE REGRESSION FOR ZERO-INFLATED OUTCOMES
Manuscript ID	SS-2020-0368
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0368
Complete List of Authors	Wodan Ling, Bin Cheng, Ying Wei, Joshua Willey and Ying Kuen Cheung
Corresponding Author	Bin Cheng
E-mail	bc2159@cumc.columbia.edu
Notice: Accepted version subject to English editing.	

STATISTICAL INFERENCE IN QUANTILE REGRESSION FOR ZERO-INFLATED OUTCOMES

Wodan Ling¹, Bin Cheng^{*,2}, Ying Wei², Joshua Z. Willey², and Ying Kuen Cheung²

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 non-linear associations between covariates and quantiles of the outcome. Theoretical properties of the estimated quantiles are established, and inference tools for assessment of quantile effects are developed. Extensive simulation studies indicate that the novel method generally outperforms the existing zero-inflated approaches and direct quantile regression in estimation and inference of the heterogeneous effect of covariates. The approach is applied to the Northern Manhattan Study (NOMAS) data to identify risk factors for carotid atherosclerosis measured by ultrasound carotid plaque burden.

Key words and phrases: Constrained post-estimation smoothing, Non-normal asymptotic distribution, Quantile regression, Zero-inflated outcomes.

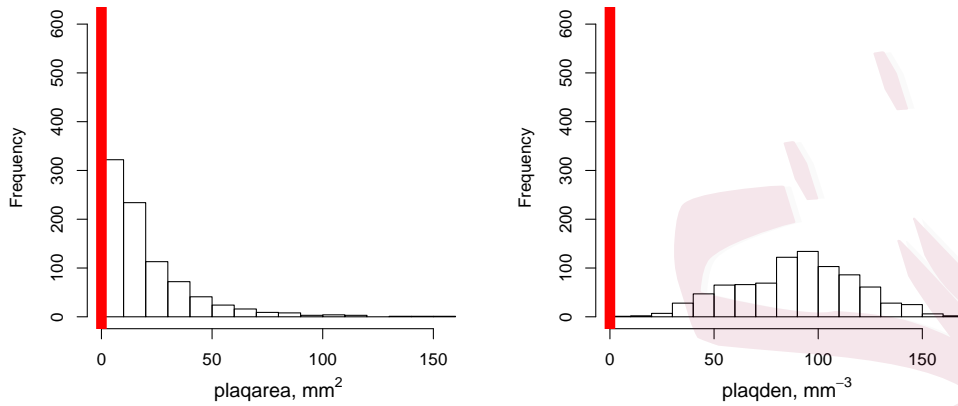
1. INTRODUCTION

1. Introduction

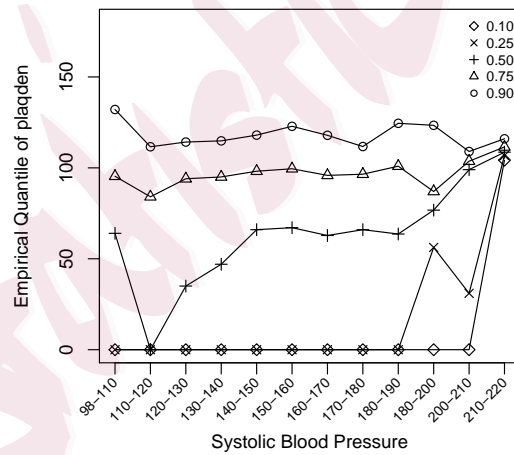
Zero-inflated outcomes are common in disease etiology studies. One such example is carotid plaque (an area of artery wall thickening), which measures carotid atherosclerosis, a proximate risk factor for stroke and cardiovascular diseases. Figure 1a shows the frequency histograms of two carotid plaque features, namely plaque area (plaqarea, in mm^2) and plaque echodensity (plaqden, in mm^{-3}), measured by high-resolution ultrasounds in 1462 participants of the NORthern MANhattan Study (Cheung et al. 2017). Specifically, plaque area measures the size of the plaque, and echodensity indicates the texture of the plaque. When an individual does not have detectable plaque, both of these variables would be 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.

A typical modeling approach for the 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 zero-inflated Poisson (ZIP) regression (Lambert 1992) and generalized ZIP (GZIP) mixture regression (Lim et al. 2014). More generally, Jorgensen (1987) considered a compound Poisson-gamma (CPG) distribution within the gener-

1. INTRODUCTION



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



(b) Empirical quantiles of plaque echodensity (plaqden) against systolic blood pressure. The relationship is non-linear as the proportion of zeros changes with systolic blood pressure.

Figure 1: Plots of carotid plaque data.

1. INTRODUCTION

alized linear model framework. However, these parametric methods often impose strong assumptions on the outcome distributions, which likely leads to biased results and invalid inference.

The standard quantile regression (Koenker and Bassett 1978), though more robust by avoiding parametric specifications and more versatile at describing heterogeneous effects at different quantile levels, is not directly applicable to model zero-inflated outcomes for two reasons. First, the feasibility of estimation and the validity of inference of quantile regression 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, direct quantile regression implicitly assumes a constant chance to observe a positive outcome, which is unlikely as the degree of zero-inflation varies across subjects. Further, as the probability that the outcome takes the value of zero varies according to the covariates, the quantile function of the outcome depends on the covariates in a non-linear fashion, which is not readily depicted in a regular quantile regression model. To illustrate this point, Figure 1b plots the τ th 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,

1. INTRODUCTION

resulting in a non-linear relationship between the quantiles of echodensity and systolic blood pressure. This characteristic will not be captured by a linear quantile regression that ignores the point mass at zero.

In this paper, we propose to use a two-part modeling strategy that uses logistic regression to model the probability of being positive, and 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 by Duan et al. (1983) and the hurdle regression model by Mullahy (1986). Although conceptually straightforward, obtaining valid estimation and inference of the proposed two-part quantile regression model is challenging for the two following reasons. First, the neighboring quantile estimation around the change point from zero to positive may have unbounded variance. The variance of the unadjusted estimated quantile process is inverse-proportional to local data density. When approaching the change point, the local density of positive data could go to zero, then the variance will be pushed to infinity. Second, estimation and inference of the quantile covariate effect are complicated. The quantile effect is a composite from the logistic and quantile regression components, and it depends on both the value of the covariate of interest and other covariates in the two-part model.

1. INTRODUCTION

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 have been used in Heras et al. (2018) to estimate actuarial profiles and provide new insights for actuarial science. The work, however, does not involve any theoretical validation and development. To the best of our knowledge, our work in this article 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 provides the details of our proposed model, the model-based conditional quantile estimation, and its asymptotic properties. The model-based inference tool for quantile treatment effects and the model-based prediction are discussed in Section 2.3. We compare the finite sample performance of the proposed method with uncorrected direct quantile regression and parametric approaches for zero-inflated outcomes through simulation studies in Section 3.

2. PROPOSED METHODS

Section 4 presents a real application of the proposed method, studying the effect of risk factors for carotid atherosclerosis measured by ultrasound carotid plaque burden, compared with the same set of competing methods.

We close the paper with some concluding remarks in Section 5.

2. Proposed methods

2.1 Model

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

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

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

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

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

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

2. PROPOSED METHODS

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

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

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

$$\lim_{\tau \rightarrow 0^+} Q_Y(\tau|\mathbf{X}, Y > 0) = 0, \quad (2.3)$$

which ensures that the quantile function $Q_Y(\tau|\mathbf{X})$ is continuous at 0. Note that different subsets of the covariate profile \mathbf{X} can be used as the two covariates in Models (2.1) and (2.2) in practice.

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

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

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

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

2. PROPOSED METHODS

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

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

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

Suppose we have i.i.d. random samples $\{(\mathbf{x}_i, y_i); i = 1, \dots, n\}$ following the conditional quantile model (2.4). We can estimate the coefficients $\boldsymbol{\gamma}$ by logistic regression (i.e., regress $I\{y_i > 0\}$ against \mathbf{x}_i), and estimate the quantile coefficient function $\boldsymbol{\beta}(\tau)$ by regressing the positive y_i 's against \mathbf{x}_i using quantile regression at a sequence of quantile levels. Specifically, $\hat{\boldsymbol{\gamma}}_n$ and $\hat{\boldsymbol{\beta}}_n(\tau)$ are solutions to the objective functions,

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_n &= \operatorname{argmax}_{\boldsymbol{\gamma}} \frac{1}{n} \sum_{i=1}^n \left[I(y_i > 0) \log \left\{ \frac{\pi(\boldsymbol{\gamma}, \mathbf{x}_i)}{1 - \pi(\boldsymbol{\gamma}, \mathbf{x}_i)} \right\} + \log\{1 - \pi(\boldsymbol{\gamma}, \mathbf{x}_i)\} \right], \\ \hat{\boldsymbol{\beta}}_n(\tau) &= \operatorname{argmin}_{\boldsymbol{\beta}(\tau)} \frac{1}{n} \sum_{i=1}^n \rho_{\tau}\{y_i - \mathbf{x}_i^{\top} \boldsymbol{\beta}(\tau)\} I(y_i > 0) \\ &= \operatorname{argmin}_{\boldsymbol{\beta}(\tau)} \frac{1}{n} \sum_{i=1}^n \{y_i - \mathbf{x}_i^{\top} \boldsymbol{\beta}(\tau)\} \{\tau - I(y_i - \mathbf{x}_i^{\top} \boldsymbol{\beta}(\tau) < 0)\} I(y_i > 0).\end{aligned}$$

However, due to the change point at $\tau = 1 - \pi(\boldsymbol{\gamma}, \mathbf{x})$, and the fact that $\operatorname{Var}\{\hat{\boldsymbol{\beta}}_n(\tau)\} \rightarrow \infty$ when $\tau \rightarrow 0^+$, it is nontrivial to combine $\hat{\boldsymbol{\gamma}}_n$ and $\hat{\boldsymbol{\beta}}_n(\tau)$ to obtain a consistent estimation of $Q_Y(\tau | \mathbf{x})$ with bounded variance around

2. PROPOSED METHODS

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 $Q_Y(\tau|\mathbf{x})$

Recall that $\hat{\gamma}_n$ and $\hat{\beta}_n(\tau)$ are estimated coefficients from Models (2.1) and (2.2). The procedure to estimate the conditional quantile function $Q_Y(\tau|\mathbf{x})$ is implemented by the following steps:

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

$$\pi(\hat{\gamma}_n, \mathbf{x}) = \exp(\mathbf{x}^\top \hat{\gamma}_n) / \{1 + \exp(\mathbf{x}^\top \hat{\gamma}_n)\}.$$

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

$$\begin{aligned} A_n &= \left\{ \tau : 0 < \tau < 1 - \pi(\hat{\gamma}_n, \mathbf{x}) \right\}, \\ B_n &= \left\{ \tau : 1 - \pi(\hat{\gamma}_n, \mathbf{x}) \leq \tau \leq 1 - \pi(\hat{\gamma}_n, \mathbf{x}) + n^{-\delta} \right\}, \\ C_n &= \left\{ \tau : 1 - \pi(\hat{\gamma}_n, \mathbf{x}) + n^{-\delta} < \tau < 1 \right\}. \end{aligned}$$

Step 3. Estimate the quantile coefficients $\hat{\beta}_n$ at the nominal quantile level $\Gamma(1 - \pi(\hat{\gamma}_n, \mathbf{x}) + n^{-\delta}; \mathbf{x}, \hat{\gamma}_n)$ and do interpolation if the target quantile

2. PROPOSED METHODS

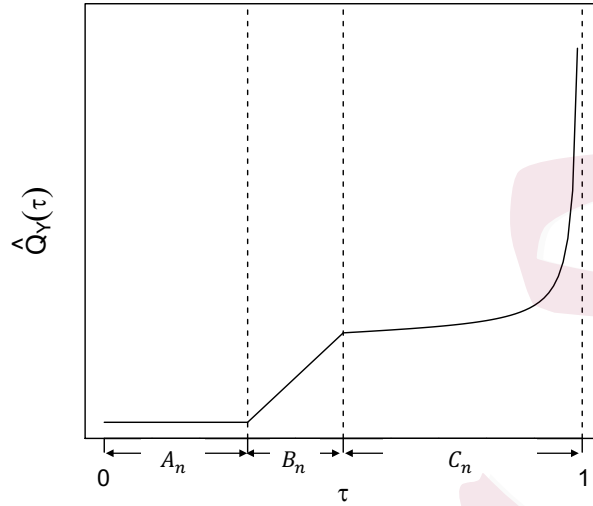


Figure 2: piecewise estimator of the conditional quantile function $Q_Y(\tau|\mathbf{x})$.

level τ of Y belongs to B_n . While if τ is in C_n , directly estimate $\hat{\beta}_n$ at $\Gamma(\tau; \mathbf{x}, \hat{\gamma}_n)$. The estimator $\hat{Q}_Y(\tau|\mathbf{x})$, as shown in Figure 2, is then a piecewise function defined by

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

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 $\mathbf{x}^\top \hat{\beta}_n \circ \Gamma(1 - \pi(\hat{\gamma}_n, \mathbf{x}) + n^{-\delta}; \mathbf{x}, \hat{\gamma}_n)$. The

2. PROPOSED METHODS

width of the interpolation window, $n^{-\delta}$, is designed to converge slower than the convergence rate of $\widehat{\gamma}_n$, so that we do not need to estimate at the problematic change point, $1 - \pi(\widehat{\gamma}_n, \mathbf{x})$. In Section 2.2.1, we establish the asymptotic properties of the estimator $\widehat{Q}_Y(\tau|\mathbf{x})$ in (2.6).

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

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

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

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

Assumption 3. The conditional quantile function has the property

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

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

2. PROPOSED METHODS

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

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

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

$$\mathbf{D}_0 = \mathbb{E}\left\{\pi(\boldsymbol{\gamma}, \mathbf{X})\mathbf{X}\mathbf{X}^\top\right\}, \quad (2.8)$$

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

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

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

2. PROPOSED METHODS

(i) $\widehat{Q}_Y(\tau|\mathbf{x})$ is a consistency estimator, i.e., as $n \rightarrow \infty$,

$$\widehat{Q}_Y(\tau|\mathbf{x}) \rightarrow_p Q_Y(\tau|\mathbf{x}).$$

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

(a) If $\tau < 1 - \pi(\boldsymbol{\gamma}, \mathbf{x})$, $\widehat{Q}_Y(\tau|\mathbf{x})$ is super-efficient, i.e., as $n \rightarrow \infty$,

$$\sqrt{n}(\widehat{Q}_Y(\tau|\mathbf{x}) - 0) \rightarrow_p 0.$$

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

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

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

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

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

where

$$\boldsymbol{\Sigma}_1 = \Gamma(\tau; \mathbf{x}, \boldsymbol{\gamma}) \{1 - \Gamma(\tau; \mathbf{x}, \boldsymbol{\gamma})\} \mathbf{x}^\top \mathbf{D}_{1,\beta \circ \Gamma(\tau; \mathbf{x}, \boldsymbol{\gamma})}^{-1} \mathbf{D}_0 \mathbf{D}_{1,\beta \circ \Gamma(\tau; \mathbf{x}, \boldsymbol{\gamma})}^{-1} \mathbf{x},$$

2. PROPOSED METHODS

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

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

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

2. PROPOSED METHODS

2.2.2 Choice of δ

The estimation of the conditional quantile function $\widehat{Q}_Y(\tau|\mathbf{x})$ 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. Practically, we recommend choosing $\delta = 0.499$.

If predicting future outcome values is of interest, we can do cross-validation on a grid of potential δ 's 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 non-smoothing. To achieve a smooth estimation, one can take advantage of the constrained B-spline smoothing (COBS) introduced by He and Ng (1999).

We propose to estimate 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 $k_n =$

2. PROPOSED METHODS

$o(n^{1/2})$, 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 $\tilde{Q}_Y(\tau|\mathbf{x}, Y > 0) = \mathbf{x}^\top \widehat{\boldsymbol{\beta}}_n(\tau)$ and obtain the smoothed $\widehat{Q}_Y(\tau|\mathbf{x}, Y > 0)$. 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 $\widehat{Q}_Y(\tau|\mathbf{x})$.

Asymptotic properties of the smooth estimator are not discussed in this paper. Its finite sample performance is not inferior to the non-smooth version, which is shown in the simulation studies in Sections 3 and the real data application in Sections 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 \mathbf{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 by the levels of other covariates $\mathbf{X}^{(-j)}$, where $\mathbf{X}^{(-j)}$ stands for the covariates excluding X_j . Hence we define the

2. PROPOSED METHODS

average quantile effect (AQE) of the covariate X_j by

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

The AQE, $\Delta_\tau(X_j; u, v)$, is the marginal change of the τ th quantile of Y due to the change of X_j from v to u . When X_j is the treatment assignment, coded as 1 for treatment and 0 for placebo, we have the average quantile treatment effect (AQTE),

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

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

A natural sample estimator of AQE is

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

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

2. PROPOSED METHODS

Then, we have the following Theorem 2, the proof of which is deferred to the Online Supplementary Materials.

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 $G(\mathbf{X}^{(-j)})$ indexed by $\mathbf{X}^{(-j)}$ such that*

$$\sqrt{n} \left(\widehat{\Delta}_\tau(X_j; u, v) - \Delta_\tau(X_j; u, v) \right) \rightarrow_d \int G(\mathbf{X}^{(-j)}) dP_{\mathbf{X}^{(-j)}}.$$

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

$$\sqrt{n} \left(\widehat{\Delta}_\tau(X_j; u, v) - \Delta_\tau(X_j; u, v) \right) \rightarrow_d \int G(\mathbf{X}^{(-j)}) dP_{\mathbf{X}^{(-j)}} = N(0, \sigma^2),$$

where

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

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

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

2. PROPOSED METHODS

2.3.2 Prediction

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

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

where $\widehat{Q}_Y(\cdot)$ is the estimated conditional quantile function defined in (2.6).

In addition, we propose to use conditional median,

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

as the predicted value. Conventionally, the predicted value is often defined as the estimated conditional mean given \mathbf{x}_{new} , which can be estimated by integrating the conditional quantile function, i.e., $\widehat{\mu}_{Y|\mathbf{x}_{\text{new}}} = \int_0^1 \widehat{Q}_Y(\tau|\mathbf{x}_{\text{new}}) d\tau$.

However, due to the zero-inflated nature of the outcome, 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 the optimal prediction, we can use cross-validations with some chosen measures to select covariates and the nuisance

3. SIMULATION

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 in comparison with direct quantile regression and existing parametric models for zero-inflated outcomes. We generate the simulated data in the context of the carotid plaque data with echodensity (plaqden) as 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

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

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

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

3. SIMULATION

where the true coefficient functions, $\beta(\tau)$, are estimated based on echodensity again and plotted in Figure S1 of Online Supplementary Materials. Specifically, we randomly draw a variable U from $U(0, 1)$, and then generate a positive value of plaqden by $\beta_0(U) + \beta_1(U)$ male + $\beta_2(U)$ systolic. For a sample with $D = 1$, we assign the value of plaqden to be 0. We generate $n = 500$ random samples in one dataset, and repeat the simulation process 1000 times.

We compare the proposed methods to the following existing approaches — (1) direct quantile regression, (2) zero-inflated Poisson regression (ZIP), (3) hurdle regression, and (4) compound Poisson-gamma regression (CPG). Direct quantile regression assumes the outcome to be absolutely continuous. When the data contains 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 linear quantile regression to the perturbed data directly. To use ZIP/hurdle model designed for count data, we round the outcomes to integers before estimation. Though the semi-continuous setting does not favor ZIP/hurdle model, we select them as comparisons since they represent the widely-applied parametric mixture/two-part model, and the rounding won't substantially affect their results.

3.2 Estimation of conditional quantile functions

In this section, we compare the estimation accuracy of conditional quantiles by the various methods. We estimate the quantile functions given 10 covariate profiles, which are formed by $\text{male} \in \{0, 1\}$ and $\text{systolic} \in \{130.78, 139.88, 150.00, 160.12, 169.22\}$ (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:

$$\begin{aligned}\text{RIMSE}_{\hat{Q}} &= \int \mathbb{E}\{\hat{Q}_Y(\tau|\mathbf{X}) - Q_Y(\tau|\mathbf{X})\}^2 d\tau / \int Q_Y(\tau|\mathbf{X})^2 d\tau, \\ \text{RIBias}_{\hat{Q}}^2 &= \int \{\mathbb{E}\hat{Q}_Y(\tau|\mathbf{X}) - Q_Y(\tau|\mathbf{X})\}^2 d\tau / \int Q_Y(\tau|\mathbf{X})^2 d\tau, \\ \text{RIVar}_{\hat{Q}} &= \int \mathbb{E}\{\hat{Q}_Y(\tau|\mathbf{X}) - \mathbb{E}\hat{Q}_Y(\tau|\mathbf{X})\}^2 d\tau / \int Q_Y(\tau|\mathbf{X})^2 d\tau;\end{aligned}$$

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

Table 1 reports $\text{RIMSE}_{\hat{Q}}$, $\text{RIBias}_{\hat{Q}}^2$, and $\text{RIVar}_{\hat{Q}}$ 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. Generally, the proposed methods have much smaller biases than direct quantile regression. The reduction in bias by the non-smooth estimation

3. SIMULATION

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 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 with more extreme values. For example, with (male, systolic)=(0, 130.78), the bias of the non-smooth proposed estimator is 0.10%, while that of the direct method is 0.84%, then the reduction in bias is 0.74%. However, with (male, systolic)=(1, 150.00), the reduction is only 0.08% (0.03% vs 0.11%). Another remark is that while the proposed non-smooth estimation leads to larger variances, the additional post-estimation smoothing can reduce the variance by approximately 0.08% in 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, i.e., the very τ where the quantile function $Q_Y(\tau|\mathbf{X})$ changes from 0 to positive. To investigate it in detail, we evaluate $\text{RIMSE}_{\hat{Q}}$, $\text{RIBias}_{\hat{Q}}^2$, and $\text{RIVar}_{\hat{Q}}$ in an interval of half-length of 0.1 around the change point. The results are summarized in Table S1 in the Supplementary Materials. As shown in the 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.

3. SIMULATION

Table 1: Summary of RIMSE(%), RIBias²(%) and RIVar(%) of the estimated conditional quantile functions of echodensity on the entire $Q_Y(\tau|\mathbf{X})$.

(gender, systolic)	Proposed (smooth)			Proposed (non-smooth)			Direct			
	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	

3. SIMULATION

The parametric approaches have a generally poor performance, which is even worse than that of the direct quantile regression (Table 1 and S1). This is because the performance of parametric methods depends on whether the model assumptions are satisfied. Neither the mixture (or combination) of zeros and Poisson distribution nor the compound Poisson-gamma 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 dataset, the point estimate of AQTE is computed as stated in (2.12). Next, in each of the 1000 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 datasets

$$\begin{aligned} & \widehat{\Delta}_{\tau}^{(B)}(\text{male}; 1, 0) & (3.13) \\ = & \frac{1}{n} \sum_{i=1}^n \{ \widehat{Q}_{\text{plaqden}}^{(B)}(\tau | \text{male} = 1, \text{systolic}_i^{(B)}) - \widehat{Q}_{\text{plaqden}}^{(B)}(\tau | \text{male} = 0, \text{systolic}_i^{(B)}) \}. \end{aligned}$$

3. SIMULATION

Note that Theorem 2 in Section 2.3.1 guarantees a normal limiting distribution of the estimated AQTE since 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, $(\text{male}=0 \text{ or } 1, \text{systolic}_i^{(B)})$. Therefore, by ZIP, hurdle, and CPG regression models, we can also estimate the conditional quantiles and estimate the AQTE as stated in (3.13). We use three measures to evaluate the inference performance regarding AQTE: (1) the bias of the average estimate of 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 AQTEs on all of the five quantiles. In addition, their coverage rates are all close to the nominal level, 90%. Though the direct quantile regression gives the best estimate, 0, 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 ZIP and hurdle regression at higher quantiles are remarkably lower than 90%, which signifies their limitation in

3. SIMULATION

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 τ th quantile of echodensity.

τ	AQTE	Measure	Proposed (smooth)	Proposed (non-smooth)	Direct	ZIP	Hurdle	CPG
0.10	0.0036	Estimate	0.04	0.04	0.00	0.35	0.35	-1.42
		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
0.25	7.5189	Estimate	6.52	6.51	5.94	13.72	13.72	-0.81
		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
0.50	-6.6304	Estimate	-6.48	-6.53	-7.00	-4.36	-4.36	-1.19
		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
0.75	-2.9669	Estimate	-3.22	-3.26	-3.23	-5.14	-5.14	-1.57
		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
0.90	-0.9258	Estimate	-1.38	-1.55	-1.57	-5.46	-5.46	-1.88
		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

4. ANALYSIS OF THE CAROTID PLAQUE DATA

describing the extreme tails of outcome distributions. CPG model produces the worst coverage rates when $\tau = 0.25, 0.50$. Although the coverage of ZIP and hurdle regression at lower quantiles and that of 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, to make inference based on AQTE, the proposed quantile regression model outperforms 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 analyzing 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, diabetes, blood pressure, smoking status, higher education, body mass index, etc. We present the estimated average quantile effects of the risk factors in Section 4.1, the model fitness in Section 4.2, and the prediction performance in Section 4.3.

4. ANALYSIS OF THE CAROTID PLAQUE DATA

4.1 Estimated AQEs

We apply the proposed quantile regression (without smoothing) to study the carotid plaque echodensity (plaqden) and 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 levels. Moreover, the effects vary with the quantile level.

For example, although it is well-known that 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 considerable difference between the two levels (120 mmHg vs. 170 mmHg) is around the median. For the patients whose systolic blood pressures are 120 mmHg, 50% of them have their echodensity controlled below 20, while 50% of the patients with systolic blood pressure 170 mmHg have echodensity over 80. Such a difference

4. ANALYSIS OF THE CAROTID PLAQUE DATA

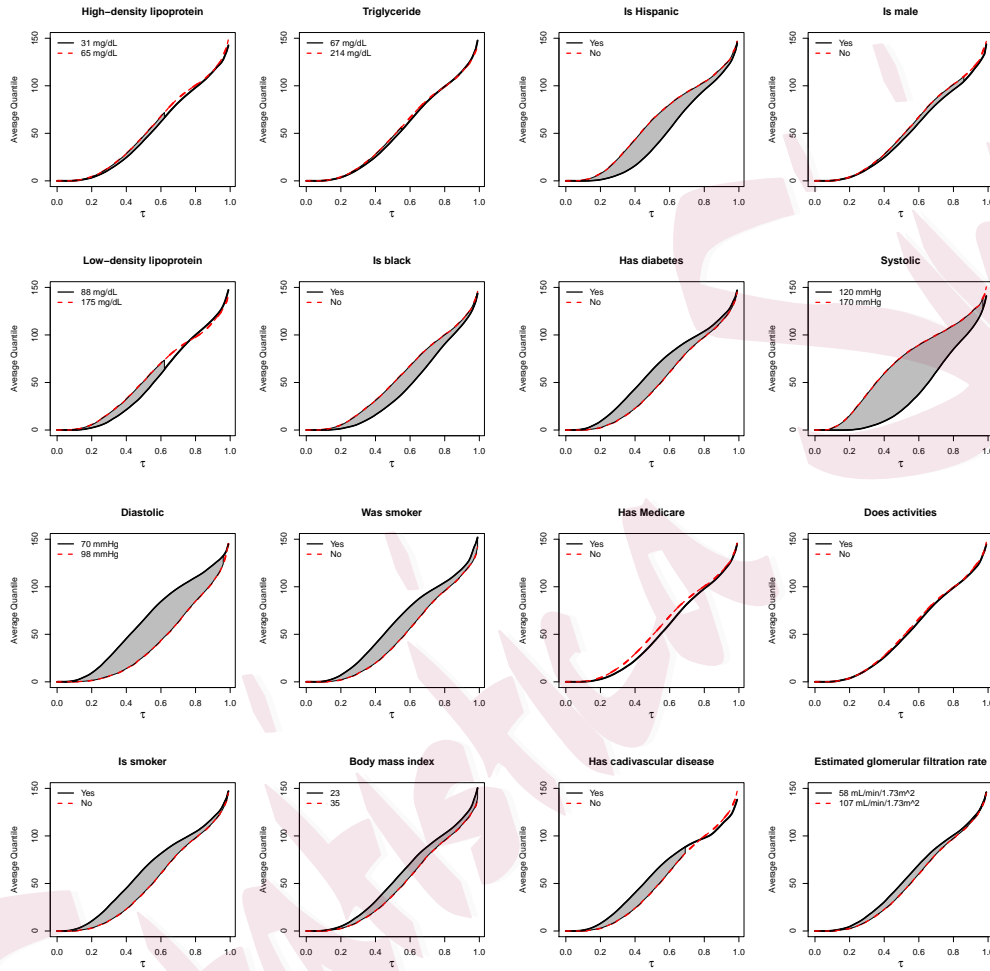


Figure 3: Estimated AQEs of selected covariates by the proposed method (without smoothing) on all quantiles of echodensity, which are presented as the differences between the red dashed and black solid lines. Significant AQEs are highlighted by shaded area.

4. ANALYSIS OF THE CAROTID PLAQUE DATA

becomes smaller as the quantile level increases, suggesting that the risk of 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 systolic blood pressure 170 mmHg, there is a substantial likelihood of having positive plaque, while 20% of those with systolic blood pressure 120 mmHg are expected to have zero plaque. Direct quantile regression would miss such a difference. To illustrate it, we plot the estimated AQEs of systolic blood pressures using the proposed method and the direct quantile method, respectively (Figure S2 in Supplementary Materials). As expected, the direct quantile method shows no difference in the risk of positive plaque and consequently underestimates the risk of plaque burden among the patients with systolic blood pressure 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 by Figure S2, ZIP and hurdle model underestimate the risk of systolic blood pressure for the severe patients who have plaque accumulated more than 70% individuals in the 120 or 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. ANALYSIS OF THE CAROTID PLAQUE DATA

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 the observed ones in data using histograms and Q-Q plots. If a model fits the data well, we would expect the distributions of the simulated outcomes and the observed ones to be comparable. Such a visual goodness-of-fit assessment has been used in Heyman et al. (1992), and it is an effective way to illustrate the goodness-of-fit of quantile models.

Figure S3a in Supplementary Materials 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. 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, regarding the plaque area as the outcome (Figure S3b in Supplementary Materials), 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. ANALYSIS OF THE CAROTID PLAQUE DATA

4.3 Prediction

In this section, we use 5-fold cross-validation to compare the prediction performance of the various methods. As outlined in Section 2.3.2, we predict an outcome by the estimated conditional median given the covariate profile and construct the 95% prediction upper bound by 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 95% prediction upper bound, and (3) its average length for positive outcomes. Because of the small perturbation added to the zero values, direct quantile regression may predict negative values. We then treat the negative predictions as zeros.

As shown by Table S2 in Supplementary Materials, the proposed methods outperform direct quantile regression in 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 regression work equally 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. CPG model correctly predicts a majority of the zeros for plaque area (Table S2, bottom), but it demonstrates poor performance in predicting the

5. CONCLUSION AND DISCUSSION

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 and discussion

This paper developed a model-based estimation algorithm, proposed a conditional quantile inference tool, and derived the estimation and inference's theoretical results for a two-part quantile regression model addressing zero-inflated outcomes. Though the model has been used in several applications, this paper 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 $\mathbf{x}^\top \widehat{\boldsymbol{\beta}}_n \circ \Gamma(1 - \pi(\widehat{\boldsymbol{\gamma}}_n, \mathbf{x}) + n^{-\delta}; \mathbf{x}, \widehat{\boldsymbol{\gamma}}_n)$. If T is unknown, it is robust to regard the value at the left end to be 0 for interpolation, and the resulting estimator in (2.6) is asymptotic consistent at any fixed quantile level by Theorem 1. When the τ of interest is close to the problematic

5. CONCLUSION AND DISCUSSION

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 in standard ways. While it is not straightforward to make inference about covariate effects in a two-part model, our average quantile effect (AQE) provides an effective way to quantify the quantile treatment effects, conduct corresponding hypothesis testing, and construct confidence intervals. In addition, we provided tools for model-based prediction.

Through simulation studies and analysis of the carotid plaque data, we have shown that the proposed methods provide more accurate and robust estimation, better goodness-of-fit, more accurate prediction, and more accurate and comprehensive inference, compared with 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 $\mathbf{X}^T \boldsymbol{\beta}(\tau)$ could be negative or non-monotone for some \mathbf{X} values and some quantile levels, especially when they are outlying in the covariate space, contradicting the fact that Y is non-negative. When it occurs, one could follow similar approaches in Chernozhukov et al. (2010) to rearrange the estimated quantiles to ensure monotonicity and non-

5. CONCLUSION AND DISCUSSION

negativity. Chernozhukov et al. (2010) has shown that, due to the root- n convergence of quantile estimates, such post-estimation rearrangement does not affect the asymptotic behaviors of quantile estimations under fairly mild conditions. In applications where non-negativity must be ensured for all \mathbf{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 \mathbf{X} , $\beta(\tau)$ should go to negative infinity around the change point, making the inference challenging. In fact, with adequate samples, even though one does not use post-estimation rearrangement or model $\log(Y)$, the proposed model ensures that the resulting conditional quantile function is non-negative almost surely. We have theoretically and numerically shown that the estimated quantile function will converge 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 paper 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 based on the parametric logistic model and the linear quantile model for the positive part, it will be of interest to examine inference with the two

ZERO-INFLATED QUANTILE

models being replaced with semi-parametric or complete non-parametric models. Finally, interval estimation in this article is based on pointwise inference, while simultaneous inference is possible by incorporating a minimum p-value procedure (Lee et al. 2012), which determines statistical significance based on the smallest p-value across all the quantile levels, and uses a resampling procedure to derive the threshold. One can also construct a joint χ^2 test statistics to test whether the logistic coefficients and quantile coefficients at multiple quantile levels equal to zero simultaneously.

Supplementary Materials

The Online Supplementary Materials contain the proofs of Theorems 1 and 2, and additional figures and tables.

Acknowledgements

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**(3): 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

REFERENCES

- 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**: 315–338.
- Heras, A., Moreno, I. and Vilar-Zanón, J. L. (2018). An application of two-stage quantile regression to insurance ratemaking, *Scandinavian Actuarial Journal* **2018**(9): 753–769.
- Heyman, D. P., Tabatabai, A. and Lakshman, T. (1992). Statistical analysis and simulation study of video teleconference traffic in atm networks, *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* **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

REFERENCES

association studies, *Biostatistics* **13**(4): 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, Inc: New York* .

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

Public Health Sciences Division,

Fred Hutchinson Cancer Research Center, WA 98109, USA E-mail: (wling@fredhutch.org)

Department of Biostatistics, Mailman School of Public Health,

Columbia University, NY 10032, USA E-mail: (bc2159@cumc.columbia.edu)

Phone: (212-342-1238), Fax: (212-305-9408)

Department of Biostatistics, Mailman School of Public Health,

Columbia University, NY 10032, USA E-mail: (yw2148@cumc.columbia.edu)

Department of Neurology, College of Physicians and Surgeons,

Columbia University, NY 10032, USA E-mail: (jzw2@cumc.columbia.edu)

Department of Biostatistics, Mailman School of Public Health,

Columbia University, NY 10032, USA E-mail: (yc632@cumc.columbia.edu)