The Annals of Statistics
2018, Vol. 46, No. 6B, 3707–3740
https://doi.org/10.1214/17-AOS1673

EXTREMAL QUANTILE TREATMENT EFFECTS

BY YICHONG ZHANG

Singapore Management University

This paper establishes an asymptotic theory and inference method for quantile treatment effect estimators when the quantile index is close to or equal to zero. Such quantile treatment effects are of interest in many applications, such as the effect of maternal smoking on an infant's adverse birth outcomes. When the quantile index is close to zero, the sparsity of data jeopardizes conventional asymptotic theory and bootstrap inference. When the quantile index is zero, there are no existing inference methods directly applicable in the treatment effect context. This paper addresses both of these issues by proposing new inference methods that are shown to be asymptotically valid as well as having adequate finite sample properties.

1. Introduction. The sign and magnitude of treatment effects vary depending on their place in the overall distribution of outcomes, a heterogeneity captured by quantile treatment effects (QTEs). In many empirical applications, the populations of interest, such as infants with low birth weights or students with low scores, are located in the tail of the outcome distribution. Thus researchers encounter not only the usual missing counterfactual, but also data sparsity because there are not many observations in the tails. While previous literature has considered the two problems separately, how to cope with both at the same time while conducting proper statistical inferences has yet to be addressed.

This paper establishes a new asymptotic theory and inference method for an estimator of the QTE for low-rank populations. To resolve the usual missing counterfactual problem, it assumes unconfoundedness and rely on the propensity score (i.e., the conditional probability of an individual being treated) to identify QTEs. To address the data sparsity, it models a small quantile index τ as a drifting object with sample size n; that is, $\tau := \tau_n \to 0$ as $n \to \infty$. Then, it uses the modeling of extremal quantiles to derive a new asymptotic approximation for the finite sample distribution of the QTE estimator when the quantile index τ is close to zero.

This paper establishes the asymptotic properties for extremal QTE estimators when $\tau_n \to 0$. It finds that there are two asymptotic distributions of the estimator of τ_nth QTE, depending on how rapidly τ_n approaches zero. Following the terminology used in [15], τ_n is called *intermediate* when $\tau_n \to 0$ and $\tau_n n \to \infty$. In this case, this paper shows that the asymptotic distribution for the proposed estimator of QTE is still Gaussian. Again, following [15], when $\tau_n \to 0$, $\tau_n n \to k$, for some

Received February 2017; revised November 2017.
MSC2010 subject classifications. Primary 62E20, 62G32; secondary 62P20.
Key words and phrases. Extreme quantile, intermediate quantile.

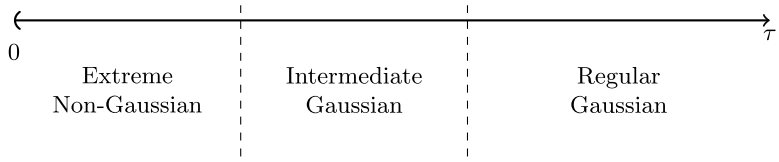

FIG. 1. *Asymptotic distribution over the quantile index.*

$k > 0$, τ_n is called *extreme*. In this case, the paper shows that the asymptotic distribution is non-Gaussian. For completeness, a quantile index is called *regular* if it is fixed strictly between zero and one. In this case, [28] showed that the QTE estimator is asymptotically normal. Figure 1 summarizes the asymptotic behaviors of the estimator of QTE.

This paper builds on the established literature on treatment effects and extremal quantiles. For the treatment effect literature addressing the missing counterfactual problem, it adapts the unconfoundedness assumption proposed in [45, 46] and [47]. For the extremal quantile literature addressing the data sparsity problem, [15, 16, 27, 38, 43], and [48] assumed that the conditional quantile is linear, while [7, 10], and [24] investigated extremal percentiles. See [17] for a recent overview.

When there are no covariates except the constant term and the treatment status is randomly assigned, the QTE and quantile treatment effect on treated (QTT) studied in this paper are equal to each other and reduce to the difference of two percentiles from the treatment and control groups. This difference is equal to the linear coefficient obtained by a quantile regression of observed outcome on the treatment status. In this case, the distributional and inferential theories established in this paper reduce to those of extremal percentiles as well as the extremal quantile regressions established in [15] and [16].

On the other hand, when there are covariates, the models, parameters of interest, estimation methods, distributional theories, and inferences in this paper are different from those in [15] and [16] as well as the extremal percentile literature. For the model in this paper, both the conditional quantiles of the potential outcomes and the propensity score are nonparametrically specified, which is different from the linear quantile regression model assumed in [15] and [16]. For the parameters, QTE and QTT studied in this paper are unconditional objects, which are different from the linear coefficient in the *conditional* quantile regression studied in [15] and [16]. The QTE is also distinct from the difference of the conventional percentiles of the treatment and control groups because the data are not missing at random. For estimation, the propensity score as an infinite dimensional nuisance parameter is estimated nonparametrically and used to correct for the selection bias. For the distributional theories, this paper establishes uniform results for the quantile index in both intermediate and extreme regions. The new theories rely on a high-level condition on the behavior of covariates as the quantile index approaches zero,

which is further revealed to depend on the structure of the conditional boundaries of the potential outcomes given covariates.

For inference, when the quantile index is intermediate, the paper shows that the standard bootstrap confidence interval (CI) for the QTE estimator controls size. For the extreme-order quantile case, the paper proves that the conventional bootstrap CI does not control size. It then proposes a resampling method with or without replacement that controls size uniformly over a range of quantile indices. Lastly, by considering a linear combination of extreme QTE estimators with carefully chosen weights, one can construct a consistent CI for the 0th QTE without imposing additional restrictions or extrapolating.

To choose among different categories of quantile index, this paper proposes a quantile-order-category-selection procedure similar to the identification-category-selection procedure used in [4]. The difference here is that I have two thresholds while they only have one. When the quantile index is smaller than the first threshold, the extreme-order quantile asymptotic distribution is expected to approximate the finite sample distribution of the QTE estimator better than the normal approximation. In this case, the new resampling CI developed in this paper can be used to conduct inference. The simulations in this paper examine the performance of this threshold in various designs with small, moderate and large size samples. In all cases, it is found that when the criterion is satisfied, the new resampling CI controls size while the standard bootstrap CI undercovers (i.e., over-rejects) by as much as 18 absolute percentage points. When the quantile index is greater than the second threshold, the paper proves that the standard bootstrap CI controls size. Last, when the quantile index is between the first and second threshold, the paper suggests using a conservative critical value.

The modeling of extremal quantiles in this paper is related to the concept of drifting sequence asymptotics. This concept goes back to [42] using Pitman drift to characterize power functions. Recently, the concept has been used in the context of weak instruments by, for example, [49] and [50], and other various models by [4, 5, 14] and [37].

The rest of the paper is organized as follows. Section 2 defines the parameters of interest, introduces additional notation and provides relevant background on extreme value theory. Sections 3 and 4 consider the asymptotic properties of the estimators for intermediate and extreme QTEs, respectively. Section 5 establishes the inference theory and provides a step-by-step description of implementation. A supplement [53] gathers additional theoretical results on estimating the extreme value (EV) index and conducting two-sample inference, numerical examples of limiting distributions, empirical and simulation results of the inference methods proposed in the paper and all theoretical proofs.

2. Definition, extreme value theory and notation. First, denote the potential outcomes for treated and control groups as Y_1 and Y_0, respectively. The treatment status is denoted as D, where $D = 1$ means treated and $D = 0$ means untreated.

The researcher can only observe (Y, X, D) where $Y = Y_1 D + Y_0(1 - D)$, and X is a collection of confounders. The propensity score $P(D = 1|X = x)$ is denoted as $\Pi(x)$. The parameters of interest are the τth QTE defined as

$$q(\tau) := q_1(\tau) - q_0(\tau)$$

and the τth QTT defined as

$$q_{|D=1}(\tau) := q_{1|D=1}(\tau) - q_{0|D=1}(\tau),$$

in which $q_j(\tau)$ and $q_{j|D=1}(\tau)$ denote the τth quantile of random variables Y_j and $Y_j|D = 1$, $j = 0, 1$, respectively.

As originally defined by [25] and [40], the QTE, for a fixed percentile, corresponds to the horizontal difference between the marginal distributions of two potential outcomes and is called the shift function in [25]. If an individual can maintain his ranking in the potential outcome distributions regardless of his treatment status, the QTE is equal to the quantile of the treatment effect $(Y_1 - Y_0)$. Without the rank preservation, the QTE still summarizes interesting distributional aspects of the treatment effect which complements the average treatment effect. Similarly, the QTT, for a fixed percentile, is the horizontal difference between the two potential outcome distributions of the individuals in the treatment group. In the program evaluation, the treatment effect for the group of treated individuals is of particular interest. QTT provides summary statistics to the distributional aspect of this effect.

Next, I introduce some extreme value theory, which will be used when I characterize the asymptotic theories in Sections 3 and 4. The cumulative distribution function (CDF) F *belongs to the domain of attraction of generalized extreme value distributions* if there exist sequences $(\alpha_n)_{n\in\mathbb{N}}$, $(\beta_n)_{n\in\mathbb{N}}$ and a CDF G indexed by a parameter ξ, such that, for any independent draws $(U_1, \ldots, U_n)$ from F, $\alpha_n(\min(U_1, \ldots, U_n) - \beta_n)$ converges in distribution to G. Here, F belongs to the domain of attraction of generalized extreme value distributions with a parameter ξ called the extreme value (EV) index. Define $A(z) := F(z)/F'(z)$ for some $z > s_l$ as the auxiliary function, in which s_l is the lower end point of the support of U. In addition, for two generic functions $f_1(\cdot)$ and $f_2(\cdot)$, denote $f_1(z) \sim f_2(z)$ if

$$\frac{f_1(z)}{f_2(z)} \to 1 \qquad \text{as } z \to s_l.$$

Then based on the value of ξ, F has three types of tails:

$$\begin{aligned}
&\text{type 1 tail } (\xi = 0): \quad \text{as } z \to s_l, \qquad F\big(z + vA(z)\big) \sim F(z)e^v \qquad \forall v \in \mathbb{R},\\
&\text{type 2 tail } (\xi > 0): \quad \text{as } z \to -\infty, \qquad F(vz) \sim v^{-1/\xi} F(z) \qquad \forall v > 0,\\
&\text{type 3 tail } (\xi < 0): \quad \text{as } z \to 0, \qquad F(s_l + vz) \sim v^{-1/\xi} F(s_l + z) \qquad \forall v > 0.
\end{aligned}$$

For example, normal, T and Beta distributions have type 1, 2 and 3 tails, respectively.

Finally, I provide two weak convergence concepts this paper will rely on. I use $U_n \rightsquigarrow U$ to indicate weak convergence as defined by [51]. When U_n and U are k-dimensional elements, the space of the sample path is $\mathbb{R}^k$ equipped with the Euclidean metric. When U_n and U are stochastic processes, the space of the sample path will be specified later in each different context. For this paper, the space is either $l^\infty(\{v \in \mathbb{R} : |v| < B\})$, for some positive B equipped with the sup norm or the Skorohod space $\mathcal{D}([-B, B])$, for some positive B equipped with the Skorohod metric.[1]

3. Intermediate quantile treatment effects. Recall the setup in Section 2. I further assume the following.

ASSUMPTION 1.

(1) $\{Y_i, D_i, X_i\}_{i=1}^n$ is i.i.d.
(2) $(Y_1, Y_0) \perp\!\!\!\perp D|X$.
(3) X is r-dimensional. The support of X, $\text{Supp}(X)$, is compact. For some $c > 0$, $c < \Pi(x) < 1 - c$, $\forall x \in \text{Supp}(X)$.

Assumption 1(1) is stronger than necessary. Since the parameters of interest in this paper are all tail objects, I only require the conditional tails of (Y_1, Y_0) given X to be the same across individuals, which allows the middle and the upper tail of the distributions to be heterogeneous.[2] Assumption 1(2) is the unconfoundedness assumption, which states that the potential outcomes are independent of the treatment status conditional on additional covariates X. Although strong, this assumption has been widely used in both theoretical investigations and empirical studies; see, for example, [12, 18, 28, 34, 45]. For extremal QTEs, it is appropriate to start with this unconfoundedness condition. When the quantile index is regular, that is, bounded away from 0 and 1, papers such as [1, 19, 20] and [30] extend the assumption to allow for endogenous treatment status and rely on an instrumental variable to correct the selection bias. Similar strategies can be applied here to the extremal quantile case. While important, I leave the problem of establishing the corresponding asymptotic theory to future research.

ASSUMPTION 2. The quantile index τ_n is intermediate, that is, (1) $\tau_n \to 0$ as $n \to \infty$, (2) $\tau_n n \to \infty$ as $n \to \infty$.

[1] To differentiate, D is reserved for the binary treatment status and $\{\mathcal{D}_{i,j}\}_{i=1}^\infty$, $j = 0, 1$ are the sets of random variables defined in the limiting objective function in Section 4.

[2] I thank the referee for pointing this out.

Under Assumption 1 and the fact that Y is continuously distributed, [28] found that the four quantiles $q_1(\tau)$, $q_0(\tau)$, $q_{1|D=1}(\tau)$ and $q_{0|D=1}(\tau)$ for any $\tau \in (0,1)$ are identified based on the following four moment equalities:

$$\mathbb{E}\left[\frac{D}{\Pi(X)}\big(\tau - \mathbb{1}\{Y \le q_1(\tau)\}\big)\right] = 0,$$
$$\mathbb{E}\left[\left(\frac{1-D}{1-\Pi(X)}\right)\big(\tau - \mathbb{1}\{Y \le q_0(\tau)\}\big)\right] = 0,$$
$$\mathbb{E}\big[D\big(\tau - \mathbb{1}\{Y \le q_{1|D=1}(\tau)\}\big)\big] = 0$$

and

$$\mathbb{E}\left[\frac{(1-D)\Pi(X)}{1-\Pi(X)}\big(\tau - \mathbb{1}\{Y \le q_{0|D=1}(\tau)\}\big)\right] = 0,$$

respectively.

Define $\hat{q}(\tau_n)$, the estimator of the τ_nth QTE, as $\hat{q}(\tau_n) := \hat{q}_1(\tau_n) - \hat{q}_0(\tau_n)$ and $\hat{q}_{|D=1}(\tau_n)$, the estimator of τ_nth QTT, as $\hat{q}_{|D=1}(\tau_n) := \hat{q}_{1|D=1}(\tau_n) - \hat{q}_{0|D=1}(\tau_n)$. Despite the extremal feature of the quantile index, the natural sample estimator $\hat{q}_1(\tau_n)$ for the τ_nth quantile of Y_1 can be computed through an inverse propensity score weighted quantile regression:

$$\hat{q}_1(\tau_n) := \arg\min_{q\in\mathbb{R}} \sum_{i=1}^{n} \frac{D_i}{\widehat{\Pi}(X_i)}(Y_i - q)\big(\tau_n - \mathbb{1}\{Y_i \le q\}\big), \tag{3.1}$$

in which $\widehat{\Pi}(\cdot)$ is an estimator of $\Pi(\cdot)$ to be defined later. Similarly, $\hat{q}_0(\tau_n)$, an estimator of the τ_nth quantile of Y_0, can be computed as

$$\hat{q}_0(\tau_n) := \arg\min_{q\in\mathbb{R}} \sum_{i=1}^{n} \frac{1-D_i}{1-\widehat{\Pi}(X_i)}(Y_i - q)\big(\tau_n - \mathbb{1}\{Y_i \le q\}\big). \tag{3.2}$$

For estimating the QTT, $\hat{q}_{1|D=1}(\tau_n)$ and $\hat{q}_{0|D=1}(\tau_n)$ can be computed as

$$\hat{q}_{1|D=1}(\tau_n) := \arg\min_{q\in\mathbb{R}} \sum_{i=1}^{n} D_i(Y_i - q)\big(\tau_n - \mathbb{1}\{Y_i \le q\}\big)$$

and

$$\hat{q}_{0|D=1}(\tau_n) := \arg\min_{q\in\mathbb{R}} \sum_{i=1}^{n} \frac{(1-D_i)\widehat{\Pi}(X_i)}{1-\widehat{\Pi}(X_i)}(Y_i - q)\big(\tau_n - \mathbb{1}\{Y_i \le q\}\big),$$

respectively.

Under some suitable conditions specified later, the propensity score estimator $\widehat{\Pi}(\cdot)$ is strictly between 0 and 1 with probability approaching one. This implies that the objective functions for estimating $(\hat{q}_j(\tau_n), \hat{q}_{j|D=1}(\tau_n))$, $j = 0, 1$ are all

convex. One can also directly compute $(\hat{q}_j(\tau_n), \hat{q}_{j|D=1}(\tau_n))$, $j = 0, 1$ based on the subgradient conditions of the minimization problems.[3]

Following [28] and [34], $\Pi(X)$, the propensity score, is estimated by the sieve method of fitting a logistic model. The method does not require the true propensity score to be correctly specified as a logistic model. I denote the logistic CDF by $L(a) := \exp(a)/(1 + \exp(a))$ and the number of sieve bases by h_n, which depends on the sample size n and can grow to infinity as $n \to \infty$. Let $H_{h_n}(x) := (b_{1n}(x), \ldots, b_{h_n n}(x))'$, where $\{b_{hn}\}_{h=1}^{h_n}$ are h_n bases of a linear sieve space $\mathcal{B}$. Given all r elements of X are continuously distributed, one can construct the linear sieve space $\mathcal{B}$ as follows:

1. For each element $X^{(l)}$ of X, $l = 1, \ldots, r$, let $\mathcal{B}_l$ be the univariate sieve space of dimension J_n. For example, $\mathcal{B}_l$ is a linear span of J_n dimensional power series, that is,

$$\mathcal{B}_l = \left\{ \sum_{k=0}^{J_n} \alpha_k x^k, x \in \operatorname{Supp}\big(X^{(l)}\big), \alpha_k \in \Re \right\}$$

or a linear span of third-order splines with J_n nodes, that is,

$$\mathcal{B}_l = \left\{ \sum_{k=0}^{2} \alpha_k x^k + \sum_{j=1}^{J_n} b_j \big[\max(x - t_j, 0)\big]^2, x \in \operatorname{Supp}\big(X^{(l)}\big), \alpha_k, b_j \in \Re \right\},$$

where $-\infty = t_0 \le t_1 \le \cdots \le t_{J_n} \le t_{J_n+1} = \infty$ partition $\operatorname{Supp}(X^{(l)})$ into $J_n + 1$ subsets $I_j = [t_j, t_{j+1}) \cap \operatorname{Supp}(X^{(l)})$, $j = 1, \ldots, J_n - 1$, $I_0 = (t_0, t_1) \cap \operatorname{Supp}(X^{(l)})$, and $I_{J_n} = (t_{J_n}, t_{J_n+1}) \cap \operatorname{Supp}(X^{(l)})$.

2. Let $\mathcal{B}$ be the tensor product of $\{\mathcal{B}_l\}_{l=1}^r$, which is defined as a linear space spanned by functions $\prod_{l=1}^r g_l$, where $g_l \in \mathcal{B}_l$. The dimension of $\mathcal{B}$ is then $h_n := rJ_n$.

Denote $\widehat{\Pi}(x) := L(H_{h_n}(x)'\hat{\pi}_n)$ with

$$\hat{\pi}_n := \operatorname*{arg\,max}_{\pi \in \mathbb{R}^{h_n}} \sum_{i=1}^{n} \big(D_i \log L\big(H_{h_n}(X_i)'\pi\big) + (1 - D_i) \log\big(1 - L\big(H_{h_n}(X_i)'\pi\big)\big)\big).$$

[3]The subgradient conditions for $\hat{q}_j(\tau_n)$, $j = 0, 1$ imply $\hat{q}_j(\tau_n) = Y_{h_j}$ for some positive integers h_j, $j = 0, 1$ such that $D_{h_j} = j$,

$$\tau_n n - \frac{1}{\widehat{\Pi}(X_{h_1})} \le \sum_{i \ne h_1} \frac{D_i}{\widehat{\Pi}(X_i)} \mathbb{1}\{Y_i < Y_{h_1}\} \le \tau_n n$$

and

$$\tau_n n - \frac{1}{1 - \widehat{\Pi}(X_{h_0})} \le \sum_{i \ne h_0} \frac{1 - D_i}{1 - \widehat{\Pi}(X_i)} \mathbb{1}\{Y_i < Y_{h_0}\} \le \tau_n n.$$

Both h_j, $j = 0, 1$ are uniquely defined as long as the above two conditions do not hold in equality, which is usually the case in practice.

There are other methods to estimate the *regular* QTE in addition to the inverse propensity weighting method used in this paper. [2] established the large sample properties of the matching estimator and showed that bootstrap inference is invalid. [36] considered using propensity score to balance the covariates in the treatment and control groups. [29], [31] and [44] studied the use of doubly robust moment conditions to estimate the regular QTE. It is still an open question how these estimators behave as the quantile index approaches 0.

For brevity, the rest of the paper only considers the estimation of $\hat{q}_1(\tau_n)$, $\hat{q}_0(\tau_n)$ and $\hat{q}(\tau_n)$. The asymptotic results for $\hat{q}_{1|D=1}(\tau_n)$, $\hat{q}_{0|D=1}(\tau_n)$, and $\hat{q}_{|D=1}(\tau_n)$ can be derived in a similar manner.

Furthermore, for the intermediate case, instead of only one quantile index τ_n, I focus on a range of them. That is, $k\tau_n$, $k \in [\kappa_1, \kappa_2]$ for some fixed and known constants κ_1 and κ_2 such that $0 < \kappa_1 < \kappa_2 < \infty$. I then aim to derive a uniform asymptotic theory for the process $\{(\hat{q}_1(k\tau_n), \hat{q}_0(k\tau_n)) : k \in [\kappa_1, \kappa_2]\}$, where for each k,

$$\hat{q}(k\tau_n) := \hat{q}_1(k\tau_n) - \hat{q}_0(k\tau_n),$$

$$\hat{q}_1(k\tau_n) := \arg\min_{q\in\mathbb{R}} \sum_{i=1}^{n} \frac{D_i}{\widehat{\Pi}(X_i)}(Y_i - q)\big(k\tau_n - \mathbb{1}\{Y_i \le q\}\big)$$

and

$$\hat{q}_0(k\tau_n) := \arg\min_{q\in\mathbb{R}} \sum_{i=1}^{n} \frac{1 - D_i}{1 - \widehat{\Pi}(X_i)}(Y_i - q)\big(k\tau_n - \mathbb{1}\{Y_i \le q\}\big).$$

The following sufficient regularity conditions are adapted from Assumptions A.1 and A.2 of [28].

ASSUMPTION 3.

(1) The density of X is bounded above and bounded away from 0 over its support.

(2) The propensity score $\Pi(x)$ is s-times continuously differentiable with all the derivatives bounded.

(3) The conditional expectation $\mathbb{E}(k\tau_n - \mathbb{1}\{Y_j \le q_j(k\tau_n)\}|x)$ is t-times continuously differentiable in x with all derivatives bounded by M_n uniformly over $(x, k) \in \mathrm{Supp}(X) \times [\kappa_1, \kappa_2]$.

(4) Let $\zeta(h_n) = \sup_{x\in\mathrm{Supp}(X)} \|H_{h_n}(x)\|$.[4] Then, $\frac{\zeta(h_n)^2 h_n}{\sqrt{n}} \to 0$, $\frac{\tau_n \zeta(h_n)^{10} h_n}{n} \to 0$, $n\tau_n \zeta(h_n)^6 h_n^{-s/r} \to 0$ and $\frac{nM_n}{\tau_n h_n^{t/r}} \to 0$ where r is the dimension of X.

Assumptions 3(1) and 3(2) are common in the sieve estimation literature, for example, [13, 28] and [34]. Assumptions 3(3) and 3(4) are tailored to fit the special

[4]For an arbitrary vector or matrix A, denote $\|A\|$ as $\sqrt{\mathrm{tr}(A^T A)}$.

case in which the quantile index is intermediate and the derivative of the quantile varies with the sample size. In fact, the magnitude of M_n depends on the tail behavior of Y_j conditional on X. When the density of $Y_j|X$ vanishes in its lower tail, M_n decreases to zero. When the density of $Y_j|X$ diverges in its lower tail (such as a Beta distribution with the first shape parameter less than 1), M_n diverges to infinity. Assumptions 3(3) and 3(4) implicitly deal with the case that all r elements of X are continuous. If some elements of X are discrete, dimension r is interpreted as the dimension of continuous covariates and Assumptions 3(3) and 3(4) can be extended in a conceptually straightforward manner by using the continuous covariates estimator within samples that are homogeneous in discrete covariates, at the expense of additional notation. Furthermore, based on the standard sieve estimation results, $\zeta(h_n) = O(h_n^{1/2})$ and $\zeta(h_n) = O(h_n)$ for B-splines and power series, respectively. Therefore, if B-splines are used to form the sieve bases and $h_n = Cn^c$ for some positive constants C and c, then Assumption 3(4) is equivalent to $c < \frac{1}{4}$, $\tau_n n^{6c-1} \to 0$, $\tau_n n^{1+c(3-s/r)} \to 0$ and $M_n n^{1-t/r}/\tau_n \to 0$. Given sufficient smoothness, if $c \leq 1/6$, Assumption 3(4) holds. In addition, the convergence rate of $\widehat{\Pi}(\cdot)$ to $\Pi(\cdot)$ is of order $(\frac{h_n}{n})^{1/2} + h_n^{-\frac{s}{r}}$. Therefore, Assumption 3(4) requires the convergence rate of $\widehat{\Pi}(\cdot)$ to be faster than $n^{5/12}$. This rate is faster than the usual $n^{1/4}$ rate in semiparametric estimations, which suggests Assumption 3 may be relaxed. On the other hand, the convergence rate of the intermediate QTE estimator is not $n^{1/2}$. This implies the slowest rate allowed for the propensity score estimator may not be $n^{1/4}$. Estimation of the intermediate QTE under minimal requirement on the convergence rate of the propensity score is left as a potential area for future research. Last, Assumptions 3(3) and 3(4) are expected to be further relaxed by using the doubly robust estimation method as illustrated in [29].

Next, I impose regularity conditions on the tails of Y_1 and Y_0.

ASSUMPTION 4. For $j = 0, 1$:

(1) Y_j, $Y_j|X$ are continuously distributed with density $f_j(\cdot)$ and $f_j(\cdot|X)$, respectively.

(2) The density $f_j(\cdot)$ is monotone in its lower tails.

(3) The CDF of Y_j belongs to the domain of attraction of generalized EV distributions with the EV index ξ_j.

These restrictions are mild. Assumption 4(1) is common in the quantile regression literature. Assumption 4(2) refers to the tail of the distribution, which is satisfied by most well-known continuous distributions. Assumption 4(3) is a standard condition in extreme value theory and is satisfied by almost all continuous distributions.

Before stating the first main theoretical result of the paper, I introduce the normalizing factor $\lambda_{j,n}(k)$ for $\hat{q}_j(k\tau_n)$:

$$\lambda_{j,n}(k) := \sqrt{\frac{n}{k\tau_n}} f_j\big(q_j(k\tau_n)\big) \qquad \text{for } j = 0, 1 \text{ and } k \in [\kappa_1, \kappa_2]. \tag{3.3}$$

Recall that for the estimator of the τth percentile in which τ is regular, the convergence rate is $\sqrt{n}$ and the asymptotic variance is $\frac{\tau(1-\tau)}{f_j^2(q_j(\tau))}$. By moving the asymptotic standard deviation to the same side of the convergence rate, we obtain a normalizing factor

$$\sqrt{\frac{n}{\tau(1-\tau)}} f_j\big(q_j(\tau)\big).$$

Then letting $\tau := \tau_n \to 0$, we heuristically obtain the normalizing factor for the intermediate-order quantile estimators defined in (3.3) with $k = 1$.

THEOREM 3.1. *If Assumptions 1–4 hold, then*

$$\big(\lambda_{1,n}(k)\big(\hat{q}_1(k\tau_n) - q_1(k\tau_n)\big), \lambda_{0,n}(k)\big(\hat{q}_0(k\tau_n) - q_0(k\tau_n)\big)\big)$$

as a two-dimensional stochastic process indexed by k is asymptotically tight under the uniform metric. In addition, if there exist functions $H_1(k_1, k_2)$, $H_0(k_1, k_2)$, $H_{01}(k_1, k_2)$ and $H_{10}(k_1, k_2)$ on $(k_1, k_2) \in [\kappa_1, \kappa_2] \times [\kappa_1, \kappa_2]$ such that, as $\tau_n \to 0$,

$$\begin{aligned}
&\frac{1}{\tau_n}\mathbb{E}\bigg[\frac{P(Y_1 \le q_1(\min(k_1, k_2)\tau_n)|X)}{\Pi(X)} \\
&\qquad - \frac{1-\Pi(X)}{\Pi(X)} P\big(Y_1 \le q_1(k_1\tau_n)|X\big) P\big(Y_1 \le q_1(k_2\tau_n)|X\big)\bigg] \to H_1(k_1, k_2), \\
&\frac{1}{\tau_n}\mathbb{E}\bigg[\frac{P(Y_0 \le q_0(\min(k_1, k_2)\tau_n)|X)}{1-\Pi(X)} \\
&\qquad - \frac{\Pi(X)}{1-\Pi(X)} P\big(Y_0 \le q_0(k_1\tau_n)|X\big) P\big(Y_0 \le q_0(k_2\tau_n)|X\big)\bigg] \to H_0(k_1, k_2), \\
&\frac{1}{\tau_n}\mathbb{E} P\big(Y_0 \le q_0(k_1\tau_n)|X\big) P\big(Y_1 \le q_1(k_2\tau_n)|X\big) \to H_{01}(k_1, k_2)
\end{aligned}$$

and

$$\frac{1}{\tau_n}\mathbb{E} P\big(Y_1 \le q_1(k_1\tau_n)|X\big) P\big(Y_0 \le q_0(k_2\tau_n)|X\big) \to H_{10}(k_1, k_2),$$

then for $k \in [\kappa_1, \kappa_2]$,

$$\big(\lambda_{1,n}(k)\big(\hat{q}_1(k\tau_n) - q_1(k\tau_n)\big), \quad \lambda_{0,n}(k)\big(\hat{q}_0(k\tau_n) - q_0(k\tau_n)\big)\big) \rightsquigarrow \mathcal{B}(k),$$

where $\mathcal{B}(k)$ is a Brownian bridge with covariance kernel:

$$\mathcal{H}(k_1,k_2) := \begin{pmatrix} \dfrac{H_1(k_1,k_2)}{\sqrt{k_1k_2}} & \dfrac{H_{10}(k_1,k_2)}{\sqrt{k_1k_2}} \\ \dfrac{H_{01}(k_1,k_2)}{\sqrt{k_1k_2}} & \dfrac{H_0(k_1,k_2)}{\sqrt{k_1k_2}} \end{pmatrix}.$$

Theorem 3.1 shows that the asymptotic distribution of the intermediate QTE estimator is still Gaussian, just as when the quantile index is regular. Intuitively, this is because for $j = 0, 1$, $\hat{q}_j(\tau_n)$ can be interpreted as a cutoff for which the numbers of $\{Y_{i,j}\}_{i=1}^n$ below and above the cutoff are of the same order of $n\tau_n$ and $n(1-\tau_n)$, respectively. When τ_n is intermediate, both orders diverge to infinity, which is the same as the case in which τ is regular. Thus the shapes of asymptotic distributions under regular and intermediate-order quantile indices are the same.

Based on [28], the influence function for $\hat{q}_1(\tau)$ with regular τ is

$$\begin{aligned} &\frac{1}{f_1(q_1(\tau))}\Bigg[\frac{D_i}{\Pi(X_i)}\big(\tau - \mathbb{1}\{Y_{i,1} \le q_1(\tau)\}\big) \\ &\quad - \frac{\mathbb{E}((\tau - \mathbb{1}\{Y_{i,1} \le q_1(\tau)\})|X_i)}{\Pi(X_i)}\big(D_i - \Pi(X_i)\big)\Bigg]. \end{aligned} \tag{3.4}$$

Theorem 3.1 shows the influence function for $\hat{q}_j(\tau_n)$ is

$$\phi_{i,1,n} := \frac{1}{\sqrt{\tau_n}}\Bigg[\frac{D_i}{\Pi(X_i)}T_{i,1,n} - \frac{\mathbb{E}(T_{i,1,n}|X_i)}{\Pi(X_i)}\big(D_i - \Pi(X_i)\big)\Bigg], \tag{3.5}$$

where

$$T_{i,1,n} := \tau_n - \mathbb{1}\{Y_{i,1} \le q_1(\tau_n)\}.$$

Comparing (3.4) and (3.5), we find that the two influence functions are the same up to a deterministic sequence $\frac{f_1(q_1(\tau_n))}{\sqrt{\tau_n}}$ when $k = 1$ and τ in (3.4) is replaced by τ_n.

In the influence function, the first term represents the estimation error when the propensity score is known and the second term is the information gain by nonparametrically estimating the propensity score. For the regular case, both the first and second terms in (3.4) contribute to the asymptotic variance of the estimator. However, for the intermediate case, the second term in (3.5) may be asymptotically negligible, implying that there is no information gain by nonparametrically estimating the propensity score. This is in contrast to the regular case. To see this, note

$$\frac{1}{\tau_n}\mathbb{E}\bigg(\frac{\mathbb{E}(T_{i,1,n}|X_i)}{\Pi(X_i)}\big(D_i - \Pi(X_i)\big)\bigg)^2 = \mathbb{E}\frac{P^2(Y_1 \le q_1(\tau_n)|X)(1-\Pi(X))}{\tau_n\Pi(X)} + o(1).$$

Since $P(Y_1 \le q_1(\tau_n)|X) = O_p(\tau_n)$, under some suitable integrability condition, the second term in the influence function vanishes as $\tau_n \to 0$.

There are two difficulties to derive the asymptotic theory of $\hat{q}(\tau_n) := \hat{q}_1(\tau_n) - \hat{q}_0(\tau_n)$. First, since the density $f_j(\cdot)$ is unknown, the normalizing factors proposed in Theorem 3.1 are not feasible. Second, as $\tau_n \to 0$, $f_j(q_j(\tau_n))$ may decay to zero (e.g., normal distribution, T distribution, Beta distribution with the first parameter greater than 1) or diverge to infinity (e.g., Beta distribution with the first parameter less than 1). Therefore, due to the difference of tail behaviors of Y_1 and Y_0, the convergence rates of $\hat{q}_1(\tau_n)$ and $\hat{q}_0(\tau_n)$ are not necessarily the same. To address the first point, I follow [15] and build a feasible normalizing factor. To address the second point, I use the following assumption.

ASSUMPTION 5. Let m be some spacing parameter that is greater than 1. Then

$$\frac{q_1(m\tau_n) - q_1(\tau_n)}{q_0(m\tau_n) - q_0(\tau_n)} \to \rho \in [0, +\infty].$$

Assumption 5 aims to bridge the normalizing factors of $\hat{q}_1(\tau_n)$ and $\hat{q}_0(\tau_n)$ by considering the ratio of the differences of quantiles of Y_1 and Y_0 at quantile indices $m\tau_n$ and τ_n. Given other assumptions in the paper, if Assumption 5 holds for one m, then it holds for any positive value of m as well. For the analytical inference, researchers need to choose m. The choice of m is discussed in Section 5.5. In addition, for the intermediate case, the bootstrap inference will be shown to be valid and does not require that the value of m to be specified. When $\rho = 0$, the convergence rate of $\hat{q}_0$ is slower so the estimation error of $\hat{q}_1(\tau_n)$ is asymptotically negligible. On the other hand, if $\rho = \infty$, $\hat{q}_0(\tau_n)$ has a faster convergence rate than $\hat{q}_1(\tau_n)$, and thus can be treated as known. Finally, when $\rho \in (0, \infty)$, the convergence rates of $\hat{q}_1(\tau_n)$ and $\hat{q}_0(\tau_n)$ are the same. For analytical inference, when τ_n is intermediate, ρ can be estimated by

$$\hat{\rho} = \frac{\hat{q}_1(m\tau_n) - \hat{q}_1(\tau_n)}{\hat{q}_0(m\tau_n) - \hat{q}_0(\tau_n)}.$$

Under Assumption 5, I define the feasible normalizing factor for $\hat{q}(\tau_n)$ as

$$\hat{\lambda}_n(k) := \frac{\sqrt{nk\tau_n}}{\max\{(\hat{q}_1(mk\tau_n) - \hat{q}_1(k\tau_n)), (\hat{q}_0(mk\tau_n) - \hat{q}_0(k\tau_n))\}}.$$

The next theorem shows that the intermediate QTE estimator is asymptotically normal with the feasible normalizing factor $\hat{\lambda}_n(k)$.

THEOREM 3.2. *Let* $\rho(k) = k^{\xi_0 - \xi_1}\rho$, $C_1(\rho, m, k) := (\frac{1-m^{-\xi_1}}{\xi_1})^{-1} \frac{\rho(k)}{\max(1, \rho(k))}$, $C_0(\rho, m, k) := (\frac{1-m^{-\xi_0}}{\xi_0})^{-1} \frac{1}{\max(\rho(k), 1)}$,[5] *and* $\mathcal{B}(\cdot)$ *be as defined in Theorem 3.1. If Assumptions in Theorem 3.1 and Assumption 5 hold, then uniformly over*

[5]Here I adapt the convention that $\frac{c}{\infty} = 0$, $\frac{c}{0} = \text{sign}(c)\infty$ for any real number c, and $\frac{1-m^{-\xi}}{\xi} = \log(m)$ when $\xi = 0$.

$k \in [\kappa_1, \kappa_2]$,

$$\hat{\lambda}_n(k)\big(\hat{q}(k\tau_n) - q(k\tau_n)\big) \rightsquigarrow \big[C_1(\rho, m, k), -C_0(\rho, m, k)\big]\mathcal{B}'(k).$$

The next theorem shows that the standard bootstrap inference for the intermediate QTE controls size. It is worth noting that the propensity score has to be re-estimated for every bootstrap sample. Let $\hat{q}^\dagger(\tau_n)$ be the estimator computed using the bootstrap sample and $\widetilde{C}_a^{nn}(\tau_n)$ be the ath quantile of $\hat{q}^\dagger(\tau_n) - \hat{q}(\tau_n)$ conditional on data. The two-sided $(1-a)$th bootstrap CI for any $a \in (0,1)$ can be written as

$$\mathrm{CI}^{\mathrm{boot}}(\tau_n) = \big(\hat{q}(\tau_n) - \widetilde{C}_{1-a/2}^{nn}(\tau_n), \hat{q}(\tau_n) - \widetilde{C}_{a/2}^{nn}(\tau_n)\big).$$

THEOREM 3.3. *If Assumptions 1–5 hold, then*

$$\lim_{n\to\infty} P\big(q(\tau_n) \in \mathrm{CI}^{\mathrm{boot}}(\tau_n)\big) = 1 - a.$$

The key advantage of using bootstrap inference is that it does not require estimation of either the normalizing factor $\hat{\lambda}_n$ or the asymptotic variance-covariance matrix. [26] has already proven the validity of bootstrap inference for the intermediate-order percentiles. For the regression case, [22] pointed out that the bootstrap inference is valid for linear intermediate-order quantile regressions. Recently, [23] proved that the bootstrap inference for the intermediate-order quantile regression is valid in sample selection models. This paper shows that the bootstrap inference is also valid for the intermediate-order QTE estimator.

4. Extreme quantile treatment effects. Section 4.1 establishes asymptotic theory for the τ_nth QTE when τ_n is extreme. It serves as the foundation for the inference theory in Section 5. Section 4.2 considers the asymptotic distribution of the extreme QTE estimator with a feasible normalizing factor, which permits inference through a resampling method proposed in Section 5.2.

4.1. *The main result.*

ASSUMPTION 6. Assume τ_n is extreme; that is, (1) $\tau_n \to 0$ as $n \to \infty$, (2) $\tau_n n \to k$ for some positive constant k as $n \to \infty$.[6]

[6]The use of k in this section is slightly different from the use in the intermediate case. In the intermediate case, the limiting distribution is established for $\tau_n k$, in which τ_n serves as the anchor quantile index and k is a positive multiplier. In the extreme case here, k is the limit of $\tau_n n$. However, we can set the anchor quantile index τ_n to $1/n$. Then k can still be interpreted as the multiplier and the limiting distribution will be derived for $\tau_n k$.

Define the estimator $\hat{q}(\tau_n)$ of the τ_nth QTE $q(\tau_n)$ as

$$\hat{q}(\tau_n) := \hat{q}_1(\tau_n) - \hat{q}_0(\tau_n), \tag{4.1}$$

where $\hat{q}_1(\tau_n)$ and $\hat{q}_0(\tau_n)$ are computed from (3.1) and (3.2), respectively.

I use the same objective functions as those used to compute the regular and intermediate QTE, although, as will be shown later, the asymptotic behavior of $\hat{q}_j(\tau_n)$ is no longer normal compared to the ones with intermediate and regular quantile indices. This is because the number of observations below $q_j(\tau_n)$ is of the same order of magnitude of $\tau_n n$, which does not diverge to infinity (Assumption 6). I also need the propensity score estimator $\widehat{\Pi}(\cdot)$ to be uniformly consistent.

ASSUMPTION 7. The estimator of the propensity score is uniformly consistent, that is, $\sup_{x\in \mathrm{Supp}(X)} |\widehat{\Pi}(x) - \Pi(x)| = o_p(1)$.

This assumption does not require that the convergence rate for the nonparametric propensity score estimator be faster than $n^{1/4}$, as usually assumed; see, for example, [41]. The reason is similar to the nonnormality of the limiting distribution: there are only a finite number of observations below the estimator of $\hat{q}_j(\tau_n)$, which are thus counted in the summations in (3.1) and (3.2). Summing over a finite number of observations prevents the accumulation of the first-order approximation error $\widehat{\Pi}(X_i) - \Pi(X_i)$.

The next high-level assumption determines the shape of the asymptotic distribution of the extreme QTE estimator.

ASSUMPTION 8. For $j = 0, 1$:

(1) Let $P(X \in \cdot|Y_j = y)$ denote the conditional distribution of X given $Y_j = y$. Then $P(X \in \cdot|Y_j = y)$ weakly converges to the CDF of a random vector $\mathcal{X}_j$ as $y \to q_j(0)$.

(2) Let $P_j^+(\mathcal{X}_j \in \cdot|Y_j = q_j(0))$ be the CDF of $\mathcal{X}_j$. Then $P_j^+(\mathcal{X}_j \in \cdot|Y_j = q_j(0))$ has finite mass points.

(3) Let $\mathcal{S}$ be the discontinuity of the function $x \mapsto \Pi(x)$. Then $P_j^+(\mathcal{X}_j \in \mathcal{S}|Y_j = q_j(0)) = 0$.

Assumption 8(1) is high-level. Section C in the supplement provides primitive sufficient conditions for Assumption 8(1) to hold. Section D in the supplement contains more numerical illustrations. In general, $P_j^+(\mathcal{X}_j \in \cdot|Y_j = q_j(0))$ depends on the conditional boundary of Y_j given X. The phenomenon that the asymptotic distribution depends on boundary conditions is common in nonregular estimations; see, for example, [21, 35] and [39]. For Assumption 8(2), the number of mass points depends on the number of discrete minimizers of the conditional boundary of Y_j given X which is usually finite. Also, Assumption 8(2) holds when $\mathcal{X}_j$ is

continuous, in which case there is no mass point. Assumption 8(3) is mild as in many parametric models, for example, probit and logit, $\mathcal{S} = \varnothing$.

Theorem 4.1, the main theoretical result of this section, establishes the joint asymptotic distribution of $\hat{q}_j(\tau_n)$, $j = 0, 1$ by showing that a normalized version of $\hat{q}_j(\tau_n)$, $j = 0, 1$ weakly converges to the minimizer of an asymptotic objective function. I first state the normalized version of $\hat{q}_j(\tau_n)$, $j = 0, 1$.

For $j = 0, 1$, the normalized versions of $\hat{q}_j(\tau_n)$ with or without centering are

$$\hat{Z}^c_{j,n}(k) := \alpha_{j,n}\big(\hat{q}_j(\tau_n) - q_j(\tau_n)\big)$$

and

$$\hat{Z}_{j,n}(k) := \alpha_{j,n}\big(\hat{q}_j(\tau_n) - a_j - \beta_{j,n}\big),$$

respectively, where a_j is an auxiliary constant so that $U_j = Y_j - a_j$ has lower endpoint 0 or $-\infty$. In particular, if $q_j(0) > -\infty$, then $a_j = q_j(0)$; otherwise, a_j is arbitrary. The normalizing constants $(\alpha_{j,n}, \beta_{j,n})$ for $j = 0, 1$ are given by

$$\begin{aligned}
&\text{type 1 tails } (\xi_j = 0): \quad \alpha_{j,n} = 1/\big(A\big(F^{-1}_{U_j}(1/n)\big)\big), \qquad \beta_{j,n} = F^{-1}_{U_j}(1/n),\\
&\text{type 2 tails } (\xi_j > 0): \quad \alpha_{j,n} = -1/\big(F^{-1}_{U_j}(1/n)\big), \qquad \beta_{j,n} = 0,\\
&\text{type 3 tails } (\xi_j < 0): \quad \alpha_{j,n} = 1/\big(F^{-1}_{U_j}(1/n)\big), \qquad \beta_{j,n} = 0,
\end{aligned}$$

in which F_{U_j} is the CDF of U_j and $A(\cdot) := \frac{F_{U_j}(\cdot)}{F'_{U_j}(\cdot)}$ is the auxiliary function defined in Section 2.

The asymptotic objective function of the local parameter z is

$$-kz + \sum_{i=1}^{\infty} W_j(\mathcal{D}_{i,j}, \Pi(\mathcal{X}_{i,j}))l_\delta(\mathcal{J}_{i,j}, z), \tag{4.2}$$

in which $W_1(d, \pi) = \frac{d}{\pi}$ and $W_0(d, \pi) = \frac{1-d}{1-\pi}$. To see the meaning of each term in (4.2), I denote, for $j = 0, 1$,

$$\begin{aligned}
&\text{type 1 tails } (\xi_j = 0): \quad h_j(l) = \exp(l) \qquad \text{for } l \in \mathbb{R}, \qquad \eta_j(k) = \log(k),\\
&\text{type 2 tails } (\xi_j > 0): \quad h_j(l) = (-l)^{-1/\xi_j} \qquad \text{for } l < 0, \qquad \eta_j(k) = (-k)^{-\xi_j},\\
&\text{type 3 tails } (\xi_j < 0): \quad h_j(l) = l^{-1/\xi_j} \qquad \text{for } l > 0, \qquad \eta_j(k) = k^{-\xi_j}.
\end{aligned}$$

Then $\{\mathcal{E}_{i,j}, \mathcal{D}_{i,j}, \mathcal{X}_{i,j}\}$ is an i.i.d. sequence such that

$$\{\mathcal{E}_{i,1}, \mathcal{D}_{i,1}, \mathcal{X}_{i,1}\} \perp\!\!\!\perp \{\mathcal{E}_{i,0}, \mathcal{D}_{i,0}, \mathcal{X}_{i,0}\}$$

and for $j = 0, 1$, $\mathcal{X}_{i,j}$ is governed by the law $P^+_j(\mathcal{X}_j \in \cdot | Y_j = q_j(0))$, $\mathcal{D}_{i,j}$ is Bernoulli distributed with success probability $\Pi(\mathcal{X}_{i,j})$ conditional on $\mathcal{X}_{i,j}$, and $\mathcal{E}_{i,j}$ is standard exponentially distributed independently of both $(\mathcal{X}_{i,j}, \mathcal{D}_{i,j})$. In addition, $\mathcal{J}_{i,j} := h_j^{-1}(\sum_{l=1}^{i} \mathcal{E}_{l,j})$ and $l_\delta(u, v) := \mathbb{1}\{u < v\}(v-u) - \mathbb{1}\{u \le -\delta\}(-\delta - u)$ for an arbitrary $\delta > 0$. The same function of $l_\delta(u, v)$ is first used in [15].

ASSUMPTION 9. For both $j = 0, 1$ and a generic fixed constant $k > 0$,

$$-kz + \sum_{i=1}^{\infty} W_j(\mathcal{D}_{i,j}, \Pi(\mathcal{X}_{i,j}))l_\delta(\mathcal{J}_{i,j}, z)$$

has a unique minimizer almost surely.

Assumption 9 indicates that the asymptotic objective function has a unique minimizer which is needed for applying the argmin theory. This type of assumption is common in nonregular estimation literature; see, for example, [16, 21, 32] and [39]. Lemma G.6 provides a sufficient condition for this assumption to hold. In general, the assumption holds when $\mathcal{X}_j$ is absolutely continuous. If $\mathcal{X}_j$ has a unique mass point at x_0, the sufficient condition requires that $k\Pi(x_0)$ is not an integer, where $\Pi(\cdot)$ is the propensity score. Since integers are sparse on the real line, this sufficient condition is mild. In addition, [32] considered the nonparametric estimation of the conditional boundary and encountered the problem that their asymptotic objective function does not have a minimizer. This problem is mainly due to the fact that their estimator may not be tight.[7] In contrast, the tightness of the extreme QTE estimator is established in the supplement. Last, it is possible to regularize the linear program in the subgradient condition to guarantee a unique minimizer for every n, and thus, relax Assumption 9.[8]

THEOREM 4.1. *If Assumptions 1, 4 and 6–8 hold, there exist κ_1 and κ_2 such that $0 < \kappa_1 < \kappa_2 < \infty$ and (κ_1, κ_2) satisfy Assumption 9, then $(\hat{Z}_{1,n}(k), \hat{Z}_{0,n}(k)) \rightsquigarrow (Z_{1,\infty}(k), Z_{0,\infty}(k))$ in $\mathcal{D}^2([\kappa_1, \kappa_2])$, where*

$$\begin{aligned}&\big(Z_{1,\infty}(k), Z_{0,\infty}(k)\big)\\ &\quad := \operatorname*{arg\,min}_{(z_1,z_0)\in\mathbb{R}^2} \sum_{j=0,1}\left[-kz_j + \sum_{i=1}^{\infty} W_j(\mathcal{D}_{i,j}, \Pi(\mathcal{X}_{i,j}))l_\delta(\mathcal{J}_{i,j}, z_j)\right].\end{aligned}$$

In addition, in $\mathcal{D}^2([\kappa_1, \kappa_2])$,

$$\big(\hat{Z}^c_{1,n}(k), \hat{Z}^c_{1,n}(k)\big) \rightsquigarrow \big(Z^c_{1,\infty}(k), Z^c_{0,\infty}(k)\big) := \big(Z_{1,\infty}(k) - \eta_1(k), Z_{0,\infty}(k) - \eta_0(k)\big).$$

The immediate corollary of Theorem 4.1 is the finite dimensional convergence. Due to the lack of continuity of the sample path of $(Z_{1,\infty}(\cdot), Z_{0,\infty}(\cdot))$, the projection mapping is only continuous when the index k is not at the discontinuity.

[7] To be more precise, the distributions of Hall and Van Keilegom's [32] estimator as a sequence indexed by the sample size n may not be tight in the sense of [11].

[8] I thank the referee for this point.

COROLLARY 4.1. *If the assumptions in Theorem 4.1 hold and Assumption 9 is satisfied for* $k \in \{k_l\}_{l=1}^L$, *then*

$$\big(\hat{Z}_{1,n}(k_l), \hat{Z}_{0,n}(k_l)\big)_{l=1}^L \rightsquigarrow \big(Z_{1,\infty}(k_l), Z_{0,\infty}(k_l)\big)_{l=1}^L$$

such that

$$\begin{aligned}&\big(Z_{1,\infty}(k_l), Z_{0,\infty}(k_l)\big)_{l=1}^L \\ &\quad := \underset{(z_{1,l}, z_{0,l})_{l=1}^L}{\arg\min} \sum_{j=0,1} \sum_{l=1}^{L} \left\{ -k_l z_{j,l} + \sum_{i=1}^{\infty} W_j(\mathcal{D}_{i,j}, \Pi(\mathcal{X}_{i,j})) l_\delta(\mathcal{J}_{i,j}, z_{j,l}) \right\}\end{aligned}$$

and

$$\begin{aligned}\big(\hat{Z}^c_{1,n}(k_l), \hat{Z}^c_{0,n}(k_l)\big)_{l=1}^L &\rightsquigarrow \big(Z^c_{1,\infty}(k_l), Z^c_{0,\infty}(k_l)\big)_{l=1}^L \\ &:= \big(Z_{1,\infty}(k_l) - \eta_1(k_l), Z_{0,\infty}(k_l) - \eta_0(k_l)\big)_{l=1}^L.\end{aligned}$$

REMARK 1. Theorem 4.1 and Theorem 3.1 (for the intermediate-order quantile), along with Theorem 1 in [28] (for the regular quantile), characterize the asymptotic distribution of the QTE estimator when the quantile index ranges from 0 to 1. Starting with the regular quantile, the asymptotic distribution is normal. Estimating the unknown propensity score provides additional information. When the quantile index is intermediate, the shape of the asymptotic distribution remains normal, but the additional information from estimating the propensity score becomes asymptotically negligible. When the quantile index moves even closer to the origin so that it is extreme, the shape of the asymptotic distribution becomes non-Gaussian, but the information from estimating the propensity score is still asymptotically negligible.

REMARK 2. I do not impose any parametric restriction on the conditional quantile of Y_j given X, in contrast to [15], which considered linear extreme-order quantile regressions. The parameters considered in linear quantile regressions are conditional objects, while QTEs in this paper are unconditional objects. In order to deal with conditional quantiles, [15] proposed an innovative solution: use the asymptotic independence between residuals and covariates X in the tails in addition to linearity to regulate the conditional tail behavior. On the other hand, Assumption 8 does not nest, nor is nested by, the linearity and asymptotic independence condition in [15]. In particular, linearity is not imposed by Assumption 8. Section C in the supplement verifies Assumption 8 under three different conditional boundary conditions.

REMARK 3. Theorem 4.1 has shown that $\hat{q}_1(\tau_n)$ and $\hat{q}_0(\tau_n)$ are asymptotically independent because, by construction,

$$\{\mathcal{J}_{i,1}, \mathcal{X}_{i,1}, \mathcal{D}_{i,1}\}_{i\geq 1} \perp\!\!\!\perp \{\mathcal{J}_{i,0}, \mathcal{X}_{i,0}, \mathcal{D}_{i,0}\}_{i\geq 1}.$$

Thus the joint asymptotic distribution of $(\hat{q}_0(\tau_n), \hat{q}_1(\tau_n))$ is fully characterized by the marginals. In Section D of the Supplementary Material I compute the marginal distribution of $\hat{q}_1(\tau)$ under various boundary conditions.

REMARK 4. Directly computing the critical value of the asymptotic distribution of $\hat{q}(\tau_n)$ is infeasible. Note that the ultimate parameter of interest is $q(\tau_n) := q_1(\tau_n) - q_0(\tau_n)$. Although the joint asymptotic distribution of $(\hat{q}_0(\tau_n), \hat{q}_1(\tau_n))$ has been established by Theorem 4.1, the convergence rates depend on the tails of Y_1 and Y_0 and are difficulty to estimate consistently. Furthermore, the asymptotic distributions of $\hat{q}_0(\tau_n)$ and $\hat{q}_1(\tau_n)$ are complicated and depend on unknown boundary conditions. In Section 5, I propose to use a modified b out of n bootstrap with or without replacement to construct a CI and draw inferences.

REMARK 5. As pointed out in Remark 1, the shape of the asymptotic distribution changes as the quantile index moves from the intermediate region to the extreme region. Therefore, the extreme-order quantile asymptotic distributions proposed in Theorem 4.1 are valid only if $k = \tau_n n$ is not large, that is, $\tau_n \leq \tau_n^{(1)}$ for some $\tau_n^{(1)}$ which will be defined in Section 5.3.

4.2. *Feasible normalizing factor.* This section considers the next missing piece needed for the resampling inference: the feasible normalizing factor. Following [16], I propose a feasible normalizing factor and establish the corresponding asymptotic theory.

The normalizing factor for the τ_nth QTE estimator when τ_n is extreme has not been obvious. Note that the estimator of τ_nth QTE is $\hat{q}(\tau_n) := \hat{q}_1(\tau_n) - \hat{q}_0(\tau_n)$. Due to the different tail behaviors, the normalizing factors for $\hat{q}_1(\tau_n)$ and $\hat{q}_0(\tau_n)$ are not necessarily the same. In addition, by Theorem 4.1, the normalizing factors for $\hat{q}_1(\tau_n)$ and $\hat{q}_0(\tau_n)$ depend on first-order statistics that are unknown and difficult to estimate.

I propose the following feasible normalizing factor:

$$\hat{\alpha}_n := \frac{\sqrt{\tau_{n,0} n}}{\max\{\hat{q}_1(m\tau_{n,0}) - \hat{q}_1(\tau_{n,0}), \hat{q}_0(m\tau_{n,0}) - \hat{q}_0(\tau_{n,0})\}}, \tag{4.3}$$

where m is a spacing parameter that is greater than one and $\tau_{n,0}$ is an extreme quantile index. How to choose m and $\tau_{n,0}$ will be discussed in Section 5.5. The feasible normalizing factor uses the smaller of the two factors for $\hat{q}_1(\tau_n)$ and $\hat{q}_0(\tau_n)$. In addition, the proposed factor has the same order of magnitude as, but is not a consistent estimator of, the infeasible normalizing factor $\alpha_{j,n}$, which is made possible by the following assumption.

ASSUMPTION 10.

(1) $\tau_{n,0} n \to k_0 \in (0, \infty)$.
(2) k_0 satisfies the condition in Lemma G.7 as well as Assumption 9.

Similar to Assumption 5, I have to bridge the two normalizing factors.

ASSUMPTION 11. Let m be the spacing parameter in (4.3). Then

$$\frac{q_1(\frac{mk_0}{n}) - q_1(\frac{k_0}{n})}{q_0(\frac{mk_0}{n}) - q_0(\frac{k_0}{n})} \to \rho \in [0, \infty].$$

Assumption 11 rules out the case in which the ratio of two normalizing factors oscillates and neither converges to a finite number nor diverges to infinity. Since ρ can be 0 and ∞, the assumption incorporates the case in which one convergence rate dominates another.

The next theorem characterizes the weak convergence of the extreme QTE estimator with the feasible normalizing factor. Let

$$\chi(\xi, m, k) = \begin{cases} k_0^{-\xi}(m^{-\xi} - 1) & \text{if } \xi \neq 0, \\ \log(m) & \text{if } \xi = 0. \end{cases}$$

THEOREM 4.2. *The assumptions in Theorem 4.1 and Assumptions 10 and 11 hold. Denote*

$$\tilde{\rho} := \frac{\chi(\xi_1, m, k_0)}{\chi(\xi_0, m, k_0)\rho} \quad \text{and} \quad \hat{Z}_n^c(k) := \hat{\alpha}_n\big(\hat{q}(\tau_n) - q(\tau_n)\big)$$

for any $\tau_n n \to k$. Then, for k_0 fixed,

$$\hat{Z}_n^c(k) \rightsquigarrow Z_\infty^c(k) \qquad \text{in } \mathcal{D}[\kappa_1, \kappa_2],$$

in which

$$Z_\infty^c(k) := \frac{\sqrt{k_0}(Z_{1,\infty}^c(k) - \tilde{\rho} Z_{0,\infty}^c(k))}{\max\{Z_{1,\infty}(mk_0) - Z_{1,\infty}(k_0), \tilde{\rho}(Z_{0,\infty}(mk_0) - Z_{0,\infty}(k_0))\}}.$$

An immediate corollary from the above theorem is the weak convergence of a linear combination of $\hat{Z}_n^c(k)$'s. In Section 5.4, I use the linear combination of extreme QTE estimators to construct a point estimator and a CI for the 0th QTE. Corollary 4.2 establishes the theoretical foundation for this construction.

ASSUMPTION 12. Let $\{\hat{r}_l\}_{l=1}^L$ be a set of weights that can be random, and:

(1) $\sum_{l=1}^L \hat{r}_l = 1$,
(2) $\hat{r}_l \xrightarrow{p} r_l$ for all $l = 1, \ldots, L$ and $\{r_l\}_{l=1}^L$ a set of constant real numbers,
(3) $\tau_{n,l} n \to k_l$ where $\{k_l\}_{l=1}^L$ satisfy Assumption 9.

COROLLARY 4.2. *If the assumptions in Theorem 4.2 and Assumption 12 hold, then*

$$\hat{\alpha}_n \left(\sum_{l=1}^L \hat{r}_l \hat{q}(\tau_{n,l}) - \sum_{l=1}^L r_l q(\tau_{n,l}) \right) \rightsquigarrow \sum_{l=1}^L r_l Z_\infty^c(k_l).$$

5. Inference. This section establishes inference theories for extreme QTE estimators. Section 5.1 shows the conventional bootstrap CI does not control size. Section 5.2 establishes a new confidence band that controls size uniformly over a range of quantile indices. Section 5.3 considers a robust confidence interval over different categories of quantile indices. Section 5.4 proposes to infer the 0th QTE by combining a set of extreme QTE estimators.

5.1. *The standard bootstrap inference.* First define the bootstrap estimator with proper normalizations:

$$\begin{aligned}
&(\hat{Z}_{1,n}^\dagger(k), \hat{Z}_{0,n}^\dagger(k)) \\
&\quad := \operatorname*{arg\,min}_{(z_1,z_0)\in\mathbb{R}^2} \sum_{j=0,1} \left\{ -\sum_{i=1}^n \left(\sum_{l=1}^n \mathbb{1}\{I_l = i\} \right) W_j(D_i, \widehat{\Pi}(X_i)) \tau_n z_j \right. \\
&\qquad \left. + \sum_{i=1}^n \left(\sum_{l=1}^n \mathbb{1}\{I_l = i\} \right) W_j(D_i, \widehat{\Pi}(X_i)) l_\delta(\alpha_{j,n}(U_{i,j} - \beta_{j,n}), z_j) \right\}
\end{aligned}$$

in which $\hat{Z}_{j,n}^\dagger(k) := \alpha_{j,n}(\hat{q}_{j,n}^\dagger(\tau_n) - a_j - \beta_{j,n})$ for $\tau_n n \to k$, a_j and $\beta_{j,n}$ are defined in Section 4.1, and $\hat{q}_{j,n}^\dagger(\tau_n)$ is the point estimator computed from (3.1) and (3.2) using the bootstrap sample. Similarly, $\hat{Z}_{j,n}^{c\dagger}(k) := \alpha_{j,n}(\hat{q}_{j,n}^\dagger(\tau_n) - q_j(\tau_n))$. Here, $(I_{n,1}, I_{n,2}, \ldots, I_{n,n})$ is a multinomial vector with parameter n and probabilities $(\frac{1}{n}, \ldots, \frac{1}{n})$. The data is denoted as Φ_n and

$$(I_{n,1}, I_{n,2}, \ldots, I_{n,n}) \perp\!\!\!\perp \Phi_n.$$

Let $\{\mathcal{J}_{i,j}, \mathcal{D}_{i,j}, \mathcal{X}_{i,j}\}_{i\geq 1, j=0,1}$ be the same as the ones in Theorem 4.1 and $\{\Gamma_{i,j}\}_{i\geq 1}$ is a sequence of i.i.d. Poisson random variables with unit mean such that

$$\{\Gamma_{i,j}\}_{i\geq 1, j=0,1} \perp\!\!\!\perp \{\mathcal{J}_{i,j}, \mathcal{D}_{i,j}, \mathcal{X}_{i,j}\}_{i\geq 1, j=0,1}$$

and $\Gamma_{i,1} \perp\!\!\!\perp \Gamma_{i,0}$.

THEOREM 5.1. *If the Assumptions in Theorem 4.1 hold, then in* $\mathcal{D}^2([\kappa_1, \kappa_2])$,

$$\big(\hat{Z}^{\dagger}_{1,n}(k), \hat{Z}^{\dagger}_{0,n}(k)\big) \rightsquigarrow \big(Z^{\dagger}_{1,\infty}(k), Z^{\dagger}_{0,\infty}(k)\big),$$

in which

$$\begin{aligned}&\big(Z^{\dagger}_{1,\infty}(k), Z^{\dagger}_{0,\infty}(k)\big)\\ &\quad:= \operatorname*{arg\,min}_{(z_1,z_0)\in\mathbb{R}^2} \sum_{j=0,1}\left[-kz_j + \sum_{i=1}^{\infty}\Gamma_{i,j}W_j\big(\mathcal{D}_{i,j}, \Pi(\mathcal{X}_{i,j})\big)l_\delta(\mathcal{J}_{i,j}, z_j)\right]\end{aligned}$$

and

$$\big(\hat{Z}^{c\dagger}_{1,n}(k), \hat{Z}^{c\dagger}_{1,n}(k)\big) \rightsquigarrow \big(Z^{c\dagger}_{1,\infty}(k), Z^{c\dagger}_{0,\infty}(k)\big) := \big(Z^{\dagger}_{1,\infty}(k) - \eta_1(k), Z^{\dagger}_{0,\infty}(k) - \eta_0(k)\big).$$

The asymptotic distribution of the bootstrap estimator of extreme QTE is different from the original estimator. Compared with the limiting process in Theorem 4.1, there is an additional Poisson random variable term. Since the asymptotic objective function is not quadratic, $Z^{\dagger}_{j,\infty}(k)$, $j = 0, 1$ are not linear in $\Gamma_{i,j}$ which implies the standard bootstrap inference does not control size. Furthermore, due to the lack of linear expansion of the estimator, $\hat{Z}^{\dagger}_{j,n}(k) - \hat{Z}_{j,n}(k)$ does not share the same limiting distribution with $\hat{Z}_{j,n}(k)$.

The intuition behind the invalidity of standard bootstrap is similar to the case of order statistics. When there are no missing counterfactuals or the data are fully missing at random, the extreme-order quantile estimator considered in this paper reduces to an order statistic. However, Bickel and Freedman [9] have already shown that the standard n out of n bootstrap inference does not work for order statistics. More generally, Zarepour and Knight [52] pointed out the usual bootstrap fails asymptotically in cases for which there exists a Poisson point process in the limit.[9]

5.2. *The modified b out of n bootstrap inference.* Let the quantile index for the subsample be τ_b. The key insight for the modified b out of n bootstrap inference is to align $\tau_b b$ with $\tau_n n$. Theorem 4.2 shows that the asymptotic distribution of the τ_nth QTE is indexed by k. Letting $\tau_b b = \tau_n n = k$ ensures that the sample distribution of the subsample estimator can mimic the asymptotic distribution of the full sample estimator.

I consider the modified b out of n bootstrap inference for extreme QTEs both with and without replacement. Not allowing for replacement (subsampling), Bertail et al. [7] studied the validity of inference for extreme-order statistics without covariates. Chernozhukov and Fernández-Val [16] considered a similar inference procedure in linear extreme-order quantile regressions. Allowing for replacement, Bickel and Sakov [10] considered the modified b out of n bootstrap inference

[9] I thank the referee for this reference.

in extreme-order statistics without covariates. Theorem 5.2 proves that the modified b out of n bootstrap inferences both with and without replacement control size when inferring the extreme QTE.[10]

Before stating the main theorem of this section, I introduce the resampling version of the feasible normalizing factor for the subsample

$$\hat{\alpha}_b^* := \frac{\sqrt{\tau_{b,0} b}}{\max\{\hat{q}_1^*(m\tau_{b,0}) - \hat{q}_1^*(\tau_{b,0}), \hat{q}_0^*(m\tau_{b,0}) - \hat{q}_0^*(\tau_{b,0})\}},$$

where $\tau_{b,0} b = \tau_{n,0} n$ and $(\tau_{n,0}, m)$ are the same as the ones used to compute $\hat{\alpha}_n$ in Theorem 4.2. Then the normalized estimator for the subsample is

$$\hat{Z}_n^{c*}(k) := \hat{\alpha}_b^* \big(\hat{q}^*(\tau_b) - \hat{q}(\tau_b)\big).$$

In the above two equations, $\hat{q}^*(\tau) := \hat{q}_1^*(\tau) - \hat{q}_0^*(\tau)$ where $\hat{q}_j^*(\tau)$ is computed by (3.1) and (3.2) with τ_n replaced by $\tau = \tau_b$ or $\tau_{b,0}$ and using only the data from the subsample, which is generated either with or without replacement. Without the star symbol, $\hat{q}(\tau_b) := \hat{q}_1(\tau_b) - \hat{q}_0(\tau_b)$ where $\hat{q}_j(\tau_b)$ is computed by (3.1) and (3.2) with τ_n replaced by τ_b and using the full sample.

THEOREM 5.2. *If the assumptions in Theorem 4.2 hold and as $n \to \infty$, $\frac{b}{n} \to 0$ and $b \to \infty$ polynomial in n, then $\hat{Z}_n^{c*}(k) \rightsquigarrow Z_\infty^c(k)$ in $\mathcal{D}([\kappa_1, \kappa_2])$.*

Theorem 5.2 builds the theoretical foundation for constructing the uniform confidence band for the extreme QTE over $k \in [\kappa_1, \kappa_2]$, in which κ_1, κ_2 are not at the discontinuity of the limiting process with probability 1. Next, I want to studentize the process $\hat{Z}_n^{c*}(k)$. When the limiting process is Gaussian, it is common to first studentize the process by the pointwise standard deviation and then construct the uniform confidence band. Here I consider the same studentization in the non-Gaussian case. Let $S_n(k)$ and $\sigma(k)$ be the feasible and infeasible studentizing factors.

ASSUMPTION 13. For a (random) scale function $S_n(k)$, there exists $\sigma(k) > 0$, a deterministic function of k, such that

$$\sup_{k \in [\kappa_1, \kappa_2]} \left| \frac{S_n(k)}{\sigma(k)} - 1 \right| = o_p(1).$$

In addition, with probability approaching one, $\sigma(k)$ and $S_n(k)$ are both continuous in k and uniformly bounded and bounded away from zero over $k \in [\kappa_1, \kappa_2]$.

[10] I suggest using the modified b out of n bootstrap with replacement because it performs better in simulations.

Note $S_n(k)$ can be $S_n(k) := 1$ or $S_n(k) := k^{-\hat{\xi}_1} + k^{-\hat{\xi}_0}$ with corresponding $\sigma(k) := 1$ or $\sigma(k) := k^{-\xi_1} + k^{-\xi_0}$, respectively. In the latter case, ξ_j, $j = 0, 1$ are unknown. I replace them by their consistent estimators $\hat{\xi}_j$, $j = 0, 1$. I refer readers to Section A in the supplement for the detail of these estimators. The choice of studentizing factors will not affect the asymptotic size of the uniform confidence band, but will rather affect its power. Unlike the case with Gaussian limit in which letting $\sigma(k)$ be the pointwise standard deviation is natural, the best choice for the studentizing factor in this non-Gaussian case is still an open question and should be the focus of future research.

COROLLARY 5.1. *Let $\widehat{C}_{1-a}$ denote the $(1-a)$th quantile of*

$$\max_{k\in[\kappa_1,\kappa_2]} \left|\hat{Z}_n^{c*}(k)/S_n(k)\right|.$$

If the assumptions in Theorem 5.2 and Lemma G.8 *as well as Assumption 13 hold, then*

$$P\left(q\left(\frac{k}{n}\right) \in \left[\hat{q}\left(\frac{k}{n}\right) - S_n(k)\widehat{C}_{1-a}/\hat{\alpha}_n, \hat{q}\left(\frac{k}{n}\right) + S_n(k)\widehat{C}_{1-a}/\hat{\alpha}_n\right] : k \in [\kappa_1, \kappa_2]\right) \to 1-a.$$

Let $\{k_l\}_{l=1}^L$ be a fine grid, $\tau_{n,l} = \frac{k_l}{n}$, $\tau_{b,l} = \frac{k_l}{b}$, $\tau_{n,0} = \frac{k_0}{n}$ and $\tau_{b,0} = \frac{k_0}{b}$. The number of subsamples is B_n, which is as large as computationally possible. Researchers can compute the uniform confidence band (CB_α) based on the following procedure:

1. Compute $\hat{q}(\tau_{n,l})$ and $\hat{q}(\tau_{b,l})$ as in (4.1). Compute $\hat{\alpha}_n$, $S_n(k)$ and the propensity score $\widehat{\Pi}(\cdot)$ using the full sample.
2. For the ith subsample, compute $\hat{q}^*(\tau_{b,l})$ for $l = 0, \ldots, L$ as in (4.1). Denote

$$\hat{\alpha}_b^* := \frac{\sqrt{\tau_{b,0}b}}{\max\{\hat{q}_1^*(m\tau_{b,0}) - \hat{q}_1^*(\tau_{b,0}), \hat{q}_0^*(m\tau_{b,0}) - \hat{q}_0^*(\tau_{b,0})\}},$$

where for $j = 0, 1$, $\hat{q}_j^*(\cdot)$ is computed as in (3.1) and (3.2), respectively, using the subsample data and the propensity score estimated in the first step. Denote

$$\widehat{V}_{i,b}^* := \max_{l=1,\ldots,L} \hat{\alpha}_b^* \left|\left(\hat{q}^*(\tau_{b,l}) - \hat{q}(\tau_{b,l})\right)/S_n(k)\right|.$$

3. Repeat step 2 for $i = 1, \ldots, B_n$.[11] Compute $\widehat{C}_{1-a}$ as the $(1-a)$th quantile of the $\{\widehat{V}_{i,b}^*\}_{i=1}^{B_n}$.

[11]Note that step 1 is not repeated, which means the propensity score only needs to be estimated once.

4. Construct

$$\mathrm{CB}_\alpha = \left\{\left[\hat{q}\left(\frac{k}{n}\right) - S_n(k)\widehat{C}_{1-a}/\hat{\alpha}_n, \hat{q}\left(\frac{k}{n}\right) + S_n(k)\widehat{C}_{1-a}/\hat{\alpha}_n\right] : k \in [\kappa_1, \kappa_2]\right\}.$$

Next, I consider the modified b out of n inference for a linear combination of extreme QTEs. Let C_a be the ath quantile of $\sum_{l=1}^{L} \gamma_r Z_\infty^c(k_l)$ and $\widehat{C}_a$ be the ath quantile of

$$\hat{\alpha}_b^*\left(\sum_{l=1}^{L} \hat{\gamma}_l \hat{q}^*(\tau_{b,l}) - \sum_{l=1}^{L} \hat{\gamma}_l \hat{q}(\tau_{b,l})\right).$$

Given that $\sum_{l=1}^{L} \gamma_r Z_\infty^c(k_l)$ is continuously distributed,[12] Proposition 5.1 shows that $\widehat{C}_a$ is a consistent estimator of C_a. Denote

$$\sum_{l=1}^{L} \hat{r}_l \hat{q}(\tau_{n,l}) - \widehat{C}_{0.5}/\hat{\alpha}_n \quad \text{and}$$

$$\left[\sum_{l=1}^{L} \hat{r}_l \hat{q}(\tau_{n,l}) - \widehat{C}_{1-a/2}/\hat{\alpha}_n, \sum_{l=1}^{L} \hat{r}_l \hat{q}(\tau_{n,l}) - \widehat{C}_{a/2}/\hat{\alpha}_n\right]$$

the median-unbiased estimator and a $(1-a)\times 100\%$ CI, respectively.

PROPOSITION 5.1. *Under the assumptions in Theorem 5.2 and Assumption 12, I have*

$$\hat{\alpha}_b^*\left(\sum_{l=1}^{L} \hat{\gamma}_l \hat{q}^*(\tau_{b,l}) - \sum_{l=1}^{L} \hat{\gamma}_l \hat{q}(\tau_{b,l})\right) \rightsquigarrow \sum_{l=1}^{L} \gamma_r Z_\infty^c(k_l), \tag{5.1}$$

$$\lim_{n\to\infty} P\left(\sum_{l=1}^{L} \hat{r}_l \hat{q}(\tau_{n,l}) - \widehat{C}_{0.5}/\hat{\alpha}_n \le \sum_{l=1}^{L} r_l q(\tau_{n,l})\right) = 0.5 \tag{5.2}$$

and

$$\begin{aligned}\lim_{n\to\infty} P\Bigg(&\sum_{l=1}^{L} \hat{r}_l \hat{q}(\tau_{n,l}) - \widehat{C}_{1-a/2}/\hat{\alpha}_n \\ &\le \sum_{l=1}^{L} r_l q(\tau_{n,l}) \le \sum_{l=1}^{L} \hat{r}_l \hat{q}(\tau_{n,l}) - \widehat{C}_{a/2}/\hat{\alpha}_n\Bigg) \\ &= 1-a.\end{aligned} \tag{5.3}$$

[12] This is shown in Lemma G.7 in the supplement.

In Proposition 5.1, (5.1) shows the weak convergence of the linear combination of extreme QTE estimators, (5.2) shows the median-unbiased estimator is asymptotically median-unbiased, and (5.3) implies the CI asymptotically controls size.

To implement, let B_n denote the number of subsamples. I use the following steps to compute $\widehat{C}_a$:

1. Compute $\{\hat{r}_l, \hat{q}(\tau_{b,l}), \hat{q}(\tau_{n,l})\}_{l=1}^{L}$ and the propensity score estimator $\widehat{\Pi}(x)$ using the full sample.
2. For the ith subsample, compute $\hat{q}^*_{i,b}(\tau_{b,l})$ for $l = 0, \ldots, L$ as in (4.1). Denote

$$\hat{\alpha}^*_b := \frac{\sqrt{\tau_{b,0} b}}{\max\{\hat{q}^*_1(m\tau_{b,0}) - \hat{q}^*_1(\tau_{b,0}), \hat{q}^*_0(m\tau_{b,0}) - \hat{q}^*_0(\tau_{b,0})\}},$$

where for $j = 0, 1$, $\hat{q}^*_j(\tau_b)$ is computed as in (4.1) for each subsample. Denote

$$\widehat{V}^*_{i,b} := \hat{\alpha}^*_b \left[\sum_{l=1}^{L} \hat{r}_l \big(\hat{q}^*(\tau_{b,l}) - \hat{q}(\tau_{b,l})\big) \right].$$

3. Repeat the second step for $i = 1, \ldots, B_n$. Compute $\widehat{C}_{1-a}$ as the $(1-a)$th quantile of the $\{\widehat{V}^*_{i,b}\}_{i=1}^{B_n}$.

When $L = 1$, one can use this procedure to construct the pointwise CI for the τ_nth QTE. The finite sample performance of the CI is examined in Section E of the supplement.

5.3. *A robust confidence interval.* The inference methods for intermediate and extreme QTE estimators are different. This difference raises the practical issue of how to choose the inference method in a given dataset with a small but given quantile index. Note that for $a \in (0, 1)$, any two-sided $(1-a)$th CI can be written as

$$\text{CI} = \big(\hat{q}(\tau_n) - \widetilde{C}_{1-\frac{a}{2}}(\tau_n), \hat{q}(\tau_n) - \widetilde{C}_{\frac{a}{2}}(\tau_n)\big), \tag{5.4}$$

where $\widetilde{C}_a(\tau_n)$ is some critical value. However, the choice of $\widetilde{C}_a(\tau_n)$ depends on the order of τ_n.

Ideally, I want to use different critical values for different quantile index orders. For the extreme-order quantile index,

$$\widetilde{C}_a(\tau_n) = \widetilde{C}^{bn}_a(\tau_n) := \widehat{C}_a(\tau_n)/\hat{\alpha}_n,$$

where $\widehat{C}_a(\tau_n)$ is the critical value computed by a modified b out of n bootstrap procedure for τ_n. The corresponding CI is called BN-CI. For the intermediate- and regular-order quantile indices, $\widetilde{C}_a(\tau_n) = \widetilde{C}^{nn}_a(\tau_n)$ where $\widetilde{C}^{nn}_a(\tau_n)$ is the critical value computed by a standard bootstrap procedure. The corresponding CI is called NN-CI. But in practice, it is impossible to determine the order of any quantile index because the size of the dataset is finite. The ideal procedure is not feasible.

Andrews and Cheng [4] faced a similar problem because the model they considered can be either weakly, semi-strongly or strongly identified. What they propose is an identification-category-selection (ICS) procedure based on the strength of identification. Similarly, I propose an order-category-selection (OCS) procedure based on the quantile index of interest and construct a robust CI.

Let $\tau_n^{(1)} := \min(\frac{40}{n}, \frac{0.1b}{n})$, $\tau_n^{(2)} = \frac{b}{n\sqrt{\log(n)}}$, and for any $a \in (0, 1)$,

$$\widetilde{C}_{a/2}^{lf}(\tau_n) = \max\bigl(\widetilde{C}_{a/2}^{bn}(\tau_n), \widetilde{C}_{a/2}^{nn}(\tau_n)\bigr)$$

and

$$\widetilde{C}_{1-a/2}^{lf}(\tau_n) = \min\bigl(\widetilde{C}_{1-a/2}^{bn}(\tau_n), \widetilde{C}_{1-a/2}^{nn}(\tau_n)\bigr).$$

The robust CI is constructed based on a hybrid critical value $\widetilde{C}_a^h(\tau_n)$ defined as follows:

$$\widetilde{C}_a^h(\tau_n) = \begin{cases} \widetilde{C}_a^{bn}(\tau_n) & \text{if } \tau_n \le \tau_n^{(1)}, \\ \widetilde{C}_a^{lf}(\tau_n) & \text{if } \tau_n \in \bigl(\tau_n^{(1)}, \tau_n^{(2)}\bigr), \\ \widetilde{C}_a^{nn}(\tau_n) & \text{if } \tau_n \ge \tau_n^{(2)}. \end{cases}$$

In general, $\tau_n^{(1)}$ takes the form of $\min(\frac{C_1}{n}, \frac{C_2 b}{n})$, where C_1 and C_2 are two positive constants. If $k := \tau n$ is large, the approximation error from estimating the propensity score will contaminate the asymptotic approximation. To prevent this contamination, I require $n\tau \le C_1$. Chernozhukov [15] and Chernozhukov and Fernández-Val [16] suggest using $C_1 \in [40, 80]$. To be cautious, I choose $C_1 = 40$.

Second, the modified b out of n bootstrap method with subsample size b is only valid if the quantile index used in the subsample, $\tau_b := \frac{k}{b} = \frac{\tau n}{b}$, is close to zero, which leads to the second requirement that $\tau_b \le C_2$. Based on the simulations, the quantile index τ_b is small enough if it is less than $C_2 = 0.1$. Combining these two requirements, I obtain $\tau_n^{(1)}$.

For n large enough, $\tau_n^{(1)} = \frac{40}{n}$. If $\tau \le \tau_n^{(1)}$, $n\tau \le 40 < \infty$. For such τ, it is expected that the extreme-order asymptotic distribution can approximate the finite distribution of the τth QTE estimator better than the standard normal distribution. In this case, the robust CI equals the BN-CI.

On the other hand, if $\tau \ge \tau_n^{(2)}$,

$$\tau n \ge \frac{b}{\sqrt{\log(n)}} \to \infty$$

because $b \to \infty$ polynomially in n. For such τ, it is expected that the finite sample distribution of the τth QTE estimator is well approximated by the intermediate or regular-order quantile asymptotic distribution. In such a case, the standard bootstrap CI controls size and the robust CI is just the NN-CI.

When $\tau \in (\tau_n^{(1)}, \tau_n^{(2)})$, it is unclear whether normal or EV approximation works better. In this case, the robust CI uses the least favorable critical value, which is conservative.

The OCR procedure is different from the ICS procedure used in [4] because here I have two thresholds and when the quantile index is less than the first threshold, the asymptotic size is exact, while in [4], they only have one threshold and when the strength of identification is less than the threshold, their asymptotic size is conservative.

Let

$$\Gamma_{\text{ex}} := \{\{\tau_n\}_{n\geq 1} : \tau_n \to 0, n\tau_n \to k \in (0,\infty), k \text{ satisfies Assumption 9}\},$$

$$\Gamma_{\text{int}} := \{\{\tau_n\}_{n\geq 1} : \tau_n \to 0, n\tau_n \to \infty\}$$

and

$$\Gamma_{\text{reg}} := \{\{\tau_n\}_{n\geq 1} : \tau_n = \tau \in (0,1)\}$$

denote the collections of extreme-, intermediate- and regular-order sequences of quantile indices. Then let $\Gamma := \Gamma_{\text{ex}} \cup \Gamma_{\text{int}} \cup \Gamma_{\text{reg}}$.

THEOREM 5.3. *Assumptions 1, 3–5 and 7–8 hold. Subsample size $b \to \infty$ polynomially in n and $\frac{b}{n} \to 0$. Then, for any $a \in (0,1)$,*

$$\inf_{\{\tau_n\}_{n\geq 1}\in\Gamma} \lim_{n\to\infty} P\big(q(\tau_n) \in \big(\hat{q}(\tau_n) - \widetilde{C}^h_{1-\frac{a}{2}}(\tau_n), \hat{q}(\tau_n) - \widetilde{C}^h_{\frac{a}{2}}(\tau_n)\big)\big) = 1 - a.$$

Theorem 5.3 shows that the robust confidence interval controls size uniformly over different types of quantile indices.

5.4. *Inference theory for the 0th QTE.* I use a linear combination of extreme QTE estimators to infer the 0th QTE so that the estimation bias is canceled out. To see the source of bias, I need to assume Y_1 and Y_0 have type 3 tails. This implies that $a_j = q_j(0)$ and $\beta_{j,n} = 0$. Hence,

$$\hat{q}(\tau_n) - \big(q_1(0) - q_0(0)\big) = \hat{q}(\tau_n) - q(\tau_n) + \frac{k^{-\xi_1} + o(1)}{\alpha_{1,n}} - \frac{k^{-\xi_0} + o(1)}{\alpha_{0,n}}. \tag{5.5}$$

I can approximate the critical value of the asymptotic distribution for $\hat{q}(\tau_n) - q(\tau_n)$ based on the procedure after Proposition 5.1. The second term on the RHS of (5.5) is the bias caused by the fact that the parameter of interest is $q(0)$, instead of $q(\tau_n)$.

To get rid of this bias, I propose a feasible estimator $\hat{q}(0) := \sum_{l=1}^L \hat{r}_l \hat{q}(\tau_{n,l})$ in which the weights $\{\hat{r}_l\}_{l=1}^L$ solve the following system of equations:

$$\sum_{l=1}^L \hat{r}_l = 1, \qquad \sum_{l=1}^L \hat{r}_l k_l^{-\hat{\xi}_1} = 0, \qquad \sum_{l=1}^L \hat{r}_l k_l^{-\hat{\xi}_0} = 0. \tag{5.6}$$

Here, $(\hat{\xi}_0, \hat{\xi}_1)$ are consistent estimators of (ξ_0, ξ_1) studied in Section A of the supplement.

To implement, I compute $\hat{q}(0)$ using only three different values of $\hat{\tau}_{n,l}$, that is, $L = 3$. The reason is twofold: (i) I do not have a selection rule for choosing among solutions of weights that satisfies (5.6) if the solution is not unique; and (ii) within a fixed range, the more quantile indices I use, the higher the absolute values of the weights, which will widen the implied CI.

PROPOSITION 5.2. *Let $\hat{\xi}_j$ be consistent estimates of ξ_j for $j = 0, 1$, $L = 3$, $(\hat{r}_1, \hat{r}_2, \hat{r}_3)$ be computed as in (5.6), $\hat{q}(0) := \sum_{l=1}^{L} \hat{r}_l \hat{q}(\tau_{n,l})$ and $\widehat{C}_a$ be computed as in the procedure after Proposition 5.1. If the assumptions in Theorem 4.2 hold and Y_j has a type 3 lower tail, for $j = 0, 1$, then*

$$\lim_{n\to\infty} P\big(\hat{q}(0) - \widehat{C}_{1-a/2}/\hat{\alpha}_n \le q(0) \le \hat{q}(0) - \widehat{C}_{a/2}/\hat{\alpha}_n\big) = 1 - a.$$

Type 3 tails are also called Pareto-type tails, which are prevalent in economic data such as wealth and incomes, as argued in Section 2.2 of [16]. Second, Y_j has a type 3 lower tail if and only if the EV index is negative, which is testable based on Theorem A.1. In practice, it implies that the CDF of the two potential outcomes decay or diverge polynomially as $\tau \to 0$.

Since the lower boundaries of Y_1 and Y_0 are bounded, Y_1 and Y_0 cannot have type 2 tails. I still need to assume away type 1 tails when inferring the 0th quantile. This is because, for type 1 tails, the location normalizing factor $\beta_{j,n} \neq 0$. Then (5.5) becomes

$$\begin{aligned}&\hat{q}(\tau_n) - \big(q_1(0) - q_0(0)\big)\\ &\quad = \hat{q}(\tau_n) - q(\tau_n) + \frac{\log(k) + \alpha_{1,n}\beta_{1,n} + o(1)}{\alpha_{1,n}} - \frac{\log(k) + \alpha_{0,n}\beta_{0,n} + o(1)}{\alpha_{0,n}}.\end{aligned}$$

The extra terms cannot be canceled by the proposed method.

There are two alternative methods to infer the 0th QTE, each of which has its own restriction. The first alternative is to analytically compute $\frac{k^{-\xi_1}}{\alpha_{1,n}} - \frac{k^{-\xi_0}}{\alpha_{0,n}}$, the leading term of the bias in (5.5), which requires the estimation of the infeasible convergence rate $\alpha_{j,n}$. However, computing an estimator $\tilde{\alpha}_{j,n}$ of $\alpha_{j,n}$ such that $\frac{\tilde{\alpha}_{j,n}}{\alpha_{j,n}} \to 1$ is harder than simply estimating the EV index ξ_j. Usually, in order to compute $\tilde{\alpha}_{j,n}$, distributional assumptions such as $\alpha_{j,n} = C_j n^{\xi_j}$ for some constant C_j are imposed. See, for example, the discussion in [16] on the distributional assumption and [8] on the point of conducting subsampling inference when the convergence rate is unknown. These distributional assumptions are not needed in Proposition 5.2.

The second alternative is to ignore the finite sample bias as it is asymptotically negligible. To be more specific, combining Theorems 4.1 and 4.2, it is clear that for $\tau_n n \to k$,

$$\hat{\alpha}_n\big(\hat{q}(\tau_n) - q(0)\big)$$

converges weakly to a nondegenerate limiting distribution. I can then approximate the critical value of the limiting distribution by computing

$$\hat{Z}_n^*(k) := \hat{\alpha}_b^*(\hat{q}^*(\tau_b) - \hat{q}(\tau_n))$$

for $\tau_b b = \tau_n n$. Comparing $\hat{Z}_n^*(k)$ with $\hat{Z}_n^{c*}(k)$ in (5.2), the only difference is that the subsample estimator $\hat{q}^*(\tau_b)$ is now centered at $\hat{q}(\tau_n) := \hat{q}_1(\tau_n) - \hat{q}_0(\tau_n)$, the full sample QTE estimator at τ_n, instead of $\hat{q}(\tau_b)$. The reason is that for the subsample, $\hat{q}(\tau_b)$ and $\hat{q}(\tau_n)$ can be viewed as proxies for $q(\tau_b)$ and $q(0)$, respectively. Then, after obtaining an estimator of the critical value of the limiting distribution of $\hat{Z}_n^*(k)$ by a similar b out of n bootstrap procedure, one can construct a median-unbiased estimator and a consistent CI for $q(0)$. This method works because the bias of using $\hat{q}(\tau_n)$ as a proxy of $q(0)$ vanishes asymptotically. However, researchers have no control of the magnitude of the bias in a finite sample. The properties of the implied CI in finite samples can be sensitive to both the choice of $k = \tau_n n$ and the subsample size b. Therefore, this method is less robust than the one proposed in Proposition 5.2.

5.5. *Tuning parameters*. On the one hand, it is almost inevitable to use tuning parameters to infer extremal QTEs due to the nonparametric and nonregular nature of the problem. On the other hand, to determine the optimal tuning parameters and establish a data-driven method to select them requires further analysis of higher-order properties of the extremal QTE estimators, which is closely related to this paper. This important topic will be the subject of future research. In this section, I provide rules of thumb of selecting tuning parameters, based on either the previous literature on nonparametric sieve estimation and extremal quantile regressions or my own simulation experience. Detailed simulation evidence based on the choice of tuning parameters discussed here can be found in Sections E and H in the supplement.

The number of sieve bases h_n. For the intermediate QTE, given sufficient smoothness and B-spline sieve space, Assumption 3 boils down to $h_n = Cn^c$ where $\tau_n n^{6c-1} \to 0$ and τ_n is the intermediate quantile index researchers are interested in. In particular, I require $\tau_n n^{6c-1} \leq 0.1$, which leads to $h_n \leq C(\frac{0.1n}{\tau_n})^{1/6}$. For $n = 5000$, $\tau_n = 0.2$ and $C \in [0.5, 2]$, $(\frac{0.1n}{\tau_n})^{1/6} \approx 3.6$, which indicates that $h_n = 2, \ldots, 7$ are reasonable choices for the number of sieve bases. Alternatively, if the power series are used, then Assumption 3 boils down to $\tau_n n^{11c-1} \to 0$. By the same reasoning, I require $\tau_n n^{11c-1} \leq 0.1$, which leads to $h_n \leq C(\frac{0.1n}{\tau_n})^{1/11}$. For $n = 5000$, $\tau_n = 0.2$ and $C \in [0.5, 2]$, $(\frac{0.1n}{\tau_n})^{1/11} \approx 2$, which indicates that $h_n = 1, \ldots, 4$ are reasonable choices for the number of sieve bases. This type of heuristic calibration of h_n was also considered in [3]. Recently, [33] proposed to use cross-validation to determine h_n, which also works here. For estimating the extreme QTE, only consistency of the propensity score estimator is required, which indicates the estimation and inference of extreme QTE is less sensitive to the

choice of h_n compared to the intermediate case. Last, when there are a rich set of covariates (or sieve bases), [6] proposed to use the L_1 penalized logistic regression to estimate the propensity score and provided an algorithm to compute the penalty loadings. Given the penalty loadings, the number of covariates (or sieve bases) is determined by data. The exact same procedure can be applied to this paper. All the simulation results in the Supplementary Material use $h_n = 4$, which performs quite well in all sixteen simulation designs with small, moderate and large sample sizes.

The subsample size b. In Sections E and H, I use $b = (120, 300, 1000)$ for sample sizes $(300, 1000, 5000)$, respectively. By linear interpolation, I suggest the formula for b as a function of n as follows:

$$b = \left\lfloor 0.4n - \frac{1}{7}(n-300)^+ - \frac{2.3}{28}(n-1000)^+ - \frac{7}{40}\left(1 - \frac{\log(5000)}{\log(n)}\right)(n-5000)^+ \right\rfloor,$$

where $x^+ = \max(0, x)$. Based on the formula, when $n \geq 5000$,

$$b = 1000 + \frac{7\log(5000)}{40\log(n)}(n - 5000),$$

which implies $b \to \infty$ polynomially in n and $\frac{b}{n} \to 0$. The simulation results also indicate that the coverages of BN-CI are quite stable across $b \in (100, 200)$, $(150, 500)$ and $(500, 1500)$ for $n = 300, 1000, 5000$, respectively.

The spacing parameter m and $\tau_{n,0}$ in the feasible normalizing factor. Theoretically, the choice of $\tau_{n,0}$ in $\hat{\alpha}_n$ does not impact the asymptotic validity of the normalizing factor. However, in finite samples, this choice involves a trade-off between bias and variance. If $n\tau_{n,0}$ is small, there are fewer observations used for estimating $\hat{q}_j(\tau_{n,0})$, which produces a large variance. But, if $n\tau_{n,0}$ is large, it can introduce bias in two ways. First, as $n\tau_{n,0}$ increases, the estimation error of the propensity score will accumulate and contaminate the CI. In addition, since I use a modified b out of n bootstrap method with subsample size b to construct the CI, if $n\tau_{n,0}/b$ is large, then this quantile index cannot be interpreted as extreme-order. Both imply that the extreme-order asymptotic approximation is not suitable. To address all the above issues, I choose the index $\tau_{n,0}$ as $\tau_{n,0} = \min(\frac{10}{n}, \frac{0.1b}{n})$ and $\tau_{n,0} = \tau_n^{(1)} = \min(\frac{40}{n}, \frac{0.1b}{n})$ with $\tau_n \leq \tau_n^{(1)}$ and $\tau_n > \tau_n^{(1)}$, respectively. The simulation study in Section E in the supplement shows this rule performs well in finite samples.

For m, I follow [16] and use

$$m = 1 + \frac{1 + sp}{k_0},$$

where $k_0 = \tau_{n,0} n$ and sp ranges from 2 to 20.[13] [16] reported the inference performances of extreme quantile regressions are quite stable across $sp \in [2, 20]$. For simplicity, I choose $sp = 9$ which implies $m = 1 + \frac{10}{k_0}$. In simulations, I find the finite sample performance of BN-CI is insensitive to the choice of $sp \in [2, 20]$ also.

The quantile indices $\{\tau_{n,l}\}_{l=1}^3$ *used to infer the* 0*th QTE.* Let $k_l = n\tau_{n,l}$, $l = 1, 2, 3$. As has already been discussed in Proposition 5.2, k_l should be distant from each other. In addition, to ensure that $\{\tau_{n,l}\}_{l=1}^3$ are extreme-order, k_l should be less than 40. With this two rules of thumb, I use $(k_1, k_2, k_3) = (5, 17.5, 30)$. Its finite sample performances are satisfying, as illustrated in Sections E and H in the supplement. In addition, Sections A and E.5 in the supplement contain the estimators of the EV index and some implementation details, respectively.

6. Conclusion. By addressing the issues of missing data and data sparsity simultaneously, this paper establishes asymptotic theory and inference procedures for an estimator of the unconditional QTE when the quantile index is close to or equal to zero.

Acknowledgments. I am deeply grateful to my co-advisors Shakeeb Khan and Arnaud Maurel, as well as my committee members Federico Bugni and Matt Masten for their guidance and encouragement. I also thank the Editor, the Associate Editor, three anonymous referees, Xavier D'Haultfœuille, Jia Li and participants in the microeconometrics lunch at Duke University and the 11th World Congress of the Econometric Society for their comments.

SUPPLEMENTARY MATERIAL

Supplement to "Extremal quantile treatment effects" (DOI: 10.1214/17-AOS1673SUPP; .pdf). This supplement contains all the proofs, two empirical applications, and simulation results.

REFERENCES

[1] ABADIE, A., ANGRIST, J. and IMBENS, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica* **70** 91–117. MR1926256

[2] ABADIE, A. and IMBENS, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74** 235–267.

[3] AI, C. and CHEN, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* **71** 1795–1843. MR2015420

[13] The original formula of [16] is $m = 1 + \frac{d+sp}{k_0}$ where d is the dimension of X in their linear quantile regression. In my paper, the object of interest is the unconditional quantile. Therefore, I set $d = 1$.

[4] ANDREWS, D. W. and CHENG, X. (2012). Estimation and inference with weak, semi-strong, and strong identification. *Econometrica* **80** 2153–2211.
[5] ANDREWS, D. W. and CHENG, X. (2013). Maximum likelihood estimation and uniform inference with sporadic identification failure. *J. Econometrics* **173** 36–56.
[6] BELLONI, A., CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. and HANSEN, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica* **85** 233–298. MR3611771
[7] BERTAIL, P., HAEFKE, C., POLITIS, D. N. and WHITE, H. (2004). Subsampling the distribution of diverging statistics with applications to finance. *J. Econometrics* **120** 295–326.
[8] BERTAIL, P., POLITIS, D. N. and ROMANO, J. P. (1999). On subsampling estimators with unknown rate of convergence. *J. Amer. Statist. Assoc.* **94** 569–579.
[9] BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217. MR0630103
[10] BICKEL, P. J. and SAKOV, A. (2008). On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statist. Sinica* **18** 967–985. MR2440400
[11] BILLINGSLEY, P. (1999). *Convergence of Probability Measures* **493**, 2nd ed. Wiley, Hoboken, NJ.
[12] BITLER, M., GELBACH, J. and HOYNES, H. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *Am. Econ. Rev.* **96** 988–1012.
[13] CHEN, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handb. Econom.* **6** 5549–5632.
[14] CHEN, X., PONOMAREVA, M. and TAMER, E. (2014). Likelihood inference in some finite mixture models. *J. Econometrics* **182** 87–99.
[15] CHERNOZHUKOV, V. (2005). Extremal quantile regression. *Ann. Statist.* **33** 806–839. MR2163160
[16] CHERNOZHUKOV, V. and FERNÁNDEZ-VAL, I. (2011). Inference for extremal conditional quantile models, with an application to market and birthweight risks. *Rev. Econ. Stud.* **78** 559–589. Supplementary data available online. MR2808129
[17] CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. and KAJI, T. (2016). Extremal quantile regression: An overview. ArXiv Preprint ArXiv:1612.06850.
[18] CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. and MELLY, B. (2013). Inference on counterfactual distributions. *Econometrica* **81** 2205–2268. MR3138546
[19] CHERNOZHUKOV, V. and HANSEN, C. (2005). An IV model of quantile treatment effects. *Econometrica* **73** 245–261.
[20] CHERNOZHUKOV, V. and HANSEN, C. (2008). Instrumental variable quantile regression: A robust inference approach. *J. Econometrics* **142** 379–398. MR2408741
[21] CHERNOZHUKOV, V. and HONG, H. (2004). Likelihood estimation and inference in a class of nonregular econometric models. *Econometrica* **72** 1445–1480.
[22] CHERNOZHUKOV, V. V. (2000). Conditional extremes and near-extremes: Concepts, asymptotic theory, and economic applications. Ph.D. thesis, Stanford Univ., Stanford, CA.
[23] D'HAULTFOEUILLE, X., MAUREL, A. and ZHANG, Y. (2018). Extremal quantile regressions for selection models and the black-white wage gap. *J. Econometrics*. **203** 129–142.
[24] DEKKERS, A. L. M. and DE HAAN, L. (1989). On the estimation of the extreme-value index and large quantile estimation. *Ann. Statist.* **17** 1795–1832. MR1026314
[25] DOKSUM, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *Ann. Statist.* **2** 267–277. MR356350
[26] FALK, M. (1991). A note on the inverse bootstrap process for large quantiles. *Stochastic Process. Appl.* **38** 359–363.
[27] FEIGIN, P. D. and RESNICK, S. I. (1994). Limit distributions for linear programming time series estimators. *Stochastic Process. Appl.* **51** 135–165.

[28] FIRPO, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica* **75** 259–276.
[29] FIRPO, S. and ROTHE, C. (2014). Semiparametric estimation and inference using doubly robust moment conditions. Technical Report, IZA Discussion Paper.
[30] FRÖLICH, M. and MELLY, B. (2013). Unconditional quantile treatment effects under endogeneity. *J. Bus. Econom. Statist.* **31** 346–357. MR3173686
[31] HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66** 315–331.
[32] HALL, P. and VAN KEILEGOM, I. (2009). Nonparametric "regression" when errors are positioned at end-points. *Bernoulli* **15** 614–633. MR2555192
[33] HANSEN, B. E. (2014). Nonparametric sieve regression: Least squares, averaging least squares, and cross-validation. In *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics* 215–248. Oxford Univ. Press, Oxford. MR3306927
[34] HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189.
[35] HIRANO, K. and PORTER, J. R. (2003). Asymptotic efficiency in parametric structural models with parameter-dependent support. *Econometrica* **71** 1307–1338.
[36] IMAI, K. and RATKOVIC, M. (2014). Covariate balancing propensity score. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 243–263.
[37] KHAN, S. and NEKIPELOV, D. (2013). On uniform inference in nonlinear models with endogeneity. Economic Research Initiatives at Duke (ERID) Working Paper 153.
[38] KNIGHT, K. (2001). Limiting distributions of linear programming estimators. *Extremes* **4** 87–103.
[39] LEE, S. and SEO, M. H. (2008). Semiparametric estimation of a binary response model with a change-point due to a covariate threshold. *J. Econometrics* **144** 492–499.
[40] LEHMANN, E. L. (1974). *Nonparametrics*: *Statistical Methods Based on Ranks*. Holden-Day, San Francisco, CA.
[41] NEWEY, W. K. and MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. *Handb. Econom.* **4** 2111–2245.
[42] PITMAN, E. J. G. (1949). *Lecture Notes on Nonparametric Statistical Inference*. Columbia Univ.
[43] PORTNOY, S. and JUREČKOVÁ, J. (1999). On extreme regression quantiles. *Extremes* **2** 227–243. MR1781938
[44] ROBINS, J. M. and ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* **90** 122–129.
[45] ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.
[46] RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592.
[47] RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR472152
[48] SMITH, R. L. (1994). Nonregular regression. *Biometrika* **81** 173–183.
[49] STAIGER, D. and STOCK, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica* **65** 557–586. MR1445622
[50] STOCK, J. H. and YOGO, M. (2005). Testing for weak instruments in linear IV regression. In *Identification and Inference for Econometric Models* 80–108. Cambridge Univ. Press, Cambridge. MR2232140
[51] VAN DER VAART, A. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.
[52] ZAREPOUR, M. and KNIGHT, K. (1999). Bootstrapping point processes with some applications. *Stochastic Process. Appl.* **84** 81–90.

[53] ZHANG, Y. (2018). Supplement to "Extremal quantile treatment effects." DOI:10.1214/17-AOS1673SUPP.

SINGAPORE MANAGEMENT UNIVERSITY
90 STAMFORD ROAD
SINGAPORE 178903
SINGAPORE
E-MAIL: yczhang@smu.edu.sg