Nonparametric inference on quantile marginal effects

David M. Kaplan*
Department of Economics, University of Missouri

August 21, 2014; first version January 23, 2014

Abstract

We propose a nonparametric method to construct confidence intervals for quantile marginal effects (i.e., derivatives of the conditional quantile function). Under certain conditions, a quantile marginal effect equals a causal (structural) effect in a general nonseparable model, or equals an average thereof within a particular subpopulation. The high-order accuracy of our method is derived. Simulations and an empirical example demonstrate the new method's favorable performance and practical use. Code for the new method is provided.

KEYWORDS: fractional order statistics, high-order accuracy.
JEL: C21.

1 Introduction

Quantile regression readily provides analysis of economic topics like heterogeneity and inequality. Using quantiles, heterogeneity in treatment effects has been increasingly examined in papers such as Bitler et al. (2006) for welfare reform in the U.S., Djebbari and Smith (2008) for the Mexican conditional cash transfer program PROGRESA, and Jackson and Page (2013) for class-size effects in education. Inequality in the U.S. wage structure has been analyzed by, among others, Buchinsky (1994), Angrist et al. (2006), and Kopczuk et al. (2010). The seminal work of Koenker and Bassett (1978) also discusses at length cases where the variance of the sample median is smaller than the variance of the sample mean, which is infinite for distributions like the Cauchy. In some cases, certain quantiles are explicitly

[*] Email: kaplandm@missouri.edu. Thanks to the University of Missouri for financial support through a Summer Research Fellowship.

of interest, such as with low birthweight (for example Abrevaya, 2001; Chernozhukov and Fernández-Val, 2011) or value-at-risk (for example Chernozhukov and Umantsev, 2001).

Nonparametric quantile regression provides additional robustness by obviating the functional form assumption and provides additional heterogeneity by allowing the regression function's slope to vary arbitrarily with the regressors (as well as quantile). The slope also relates to causal effects in nonseparable models. Hoderlein and Mammen (2007), and Sasaki (2012), among others, relate the derivative of the conditional quantile function (a.k.a. the quantile marginal effect) to derivatives of a nonseparable structural function. Under certain conditions, this object is equal to a causal (structural) effect or an average of causal effects among a specific subpopulation. See §2 below for more mathematical details.

This paper concerns inference on the quantile marginal effect. In related work, Kaplan (2011), Goldman and Kaplan (2014a), and Goldman and Kaplan (2014b, hereafter GK) discuss high-order accurate nonparametric inference on unconditional and conditional quantiles, as well as inference on multiple quantiles jointly, on linear combinations of quantiles (e.g. interquantile ranges), and on quantile treatment effects for a binary treatment variable. If a treatment variable is discrete (but not binary) and the marginal effect going from, e.g., $T = 2$ to $T = 3$ is of interest, then their existing framework is sufficient: a new binary variable $W = 1\{T = 3\}$ is the "treatment," and only observations with $T = 2$ or $T = 3$ are used, where $1\{\cdot\}$ is the indicator function. Here, we propose a new method for inference on the quantile marginal effect with respect to a continuous variable.

Our new method relies on the fact that a function's derivative may be approximated by a linear combination of function values near the point of interest. This is the same idea behind computing numerical derivatives. As described in Definition 3, this approximation and local smoothing reduce the problem to inference on an unconditional quantile treatment effect. There are three sources of coverage probability error: the derivative approximation, the smoothing of continuous variable(s), and the application of the unconditional method. These are all precisely characterized, and then the optimal bandwidth and coverage properties are derived. The key assumption is on sampling (A1); mild smoothness assumptions are also required.

The primary contribution of this paper is the new method for constructing QME confidence intervals and the characterization of its properties. The approach differs greatly from bootstrap or normality-based confidence intervals based on local polynomials. This includes a new feasible (plug-in) bandwidth that targets coverage probability accuracy rather than estimation precision. As in many cases, the bandwidth minimizing the mean squared error of a local polynomial estimator does *not* lead to even first-order accurate inference because such a bandwidth equates the estimator's bias with its standard deviation (in order of magnitude). The truly optimal bandwidth arguably lies in between since there is (within this range) a tradeoff between coverage accuracy and confidence interval length, though "optimal" depends on one's objective function. In practice, we suggest (and implement in our code) shifting toward the larger bandwidth rate as the sample size grows.

The new method has advantages over the local polynomial approaches. Under the assumptions in this paper, the new method's optimal coverage probability error is smaller than a local polynomial's, $O(n^{-12/25})$ versus $O(n^{-3/8})$, although by assuming smoothness approaching infinity and fitting a polynomial of degree approaching infinity the latter theoretically approaches $O(n^{-1/2})$. Practically, this means the advantage of the local polynomial is in the case where there is sufficient data to fit higher-degree local polynomials and where the underlying function is volatile enough to restrict the new method's bandwidth. In simulations, performance is quite similar in many cases, both with local linear and cubic estimators. However, the analytic normality-based intervals (from Chaudhuri, 1991) can have severe under-coverage or over-coverage. The bootstrapped local polynomial intervals only suffer significant under-coverage in more extreme cases, but they have the usual bootstrap drawbacks of relying on randomization and taking longer to compute. Bootstrap intervals can be longer in some cases, too.

Sections 2 and 3 detail our setup, procedure, and theoretical results. Section 4 contains a simulation study, and Section 5 contains an empirical application. Notationally, $\doteq$ should be read as "is equal to, up to smaller-order terms"; $\asymp$ as "has exact (asymptotic) rate/order of" (same as "big theta" Bachmann–Landau notation, $\Theta(\cdot)$); and $A_n = O(B_n)$ as usual, $\exists k < \infty$ s.t. $|A_n| \leq B_n k$ for sufficiently large n. Acronyms used are those for cumulative distribution function (CDF), confidence interval (CI), coverage probability (CP), coverage probability error (CPE), mean squared error (MSE), probability density function (PDF), quantile marginal effect (QME), and quantile treatment effect (QTE). Vectors are column vectors unless otherwise noted, and $'$ denotes transpose. Proofs absent in the main text are in the appendix. Code is available from the author's website.

2 Setup

We adopt the definitions of "structural marginal effect" and "quantile marginal effect" (QME) from Sasaki (2012). He defines a nonparametric, nonseparable structural (causal) function $Y = g(X, U)$, where Y is the scalar outcome of interest, X is the regressor (taken to be a scalar for simplicity), and $U \in \mathbb{R}^M$ is a vector of unobserved determinants of Y. The non-structural conditional τ-quantile function of Y given $X = x$ is denoted $Q_{Y|X}(\tau \mid x) \equiv \inf\{y : F_{Y|X}(y \mid x) \geq \tau\}$, where Y and X are as before and $\tau \in (0, 1)$ is the quantile of interest, such at $\tau = 0.5$ for the median. The following are respectively Definitions 1 and 2 in Sasaki (2012), with only slight notational differences.

Definition 1 (structural marginal effect)**.** Given a structural function $g(\cdot, \cdot)$, the structural marginal effect at $(X = x_0, U = u_0)$ is given by $\beta(x_0, u_0) \equiv \frac{\partial}{\partial x} g(x, u_0)\big|_{x=x_0}$.

Definition 2 (quantile marginal effect)**.** Given a quantile regression $Q_{Y|X}(\cdot \mid \cdot)$, the quantile marginal effect (QME) at $X = x_0$ is defined by $\frac{\partial}{\partial x} Q_{Y|X}(\tau \mid x)\big|_{x=x_0}$ for each quantile $\tau \in (0, 1)$.

The QME equals the structural marginal effect if X is exogenous, U is scalar, and $g(\cdot, \cdot)$ is monotone in U. They are also equal (Sasaki, 2012, Cor. 1) if monotonicity is weakened to "local monotonicity," where the value u^* s.t. $y = g(x_0, u^*)$ is unique.

Under much weaker conditions, Sasaki (2012) shows, "the quantile marginal effect identifies a weighted average of structural marginal effects among the subpopulation of individuals at the conditional quantile of interest." For example, if $\tau = 0.9$, $x_0 = 0$, and $Q_{Y|X}(\tau = 0.9 \mid X = 0) = 1.3$, then the average is taken over different values of $U \in \{u : g(x = 0, u) = 1.3\}$. If $U = (U_1, U_2)$ and $g(X, U) = X + U_1 + U_2$, then in this example the average is over $\{u : u_1 + u_2 = 1.3\}$. Sasaki (2012, Cor. 2) states that when $\frac{\partial^2}{u_i u_j} g(x, u) = 0$ for $i \neq j$, the weights in the average are proportional to the PDF of U.

Let outcome $Y \in \mathbb{R}$ and regressor of interest $X \in \mathbb{R}$ both have continuous distributions. Let W be a vector of other control variables. The population object of interest is the QME at $X = x_0$ and $W = w_0$ for quantile of interest $\tau \in (0, 1)$:

$$\frac{\partial}{\partial x} Q_{Y|X,W}(\tau \mid x, w_0)\bigg|_{x=x_0} . \tag{1}$$

If W contains a fixed number of discrete variables with $P(W = w_0) = p > 0$, then the method can be run on the subset of observations with $W = w_0$. Since this approach does not affect the asymptotic properties,[1] discrete W are omitted hereafter for notational simplicity. Nonetheless, it is possible in finite samples to have no observations with $W = w_0$. While formal analysis of such cases is beyond the present paper's scope, the choice then is either to decide the data contain negligible information about $W = w_0$ or to assume that values close to w_0 are similar enough to yield reasonable inferences about w_0, which depends on the economic meaning of W and how similar nearby values are.

If W contains continuous variables, then $P(W = w_0) = 0$ and smoothing is required. As shown in Goldman and Kaplan (2014a), this will not increase the order of the bias beyond that caused by X being continuous, but it will decrease the growth rate of the local sample size and thus affect the theoretical coverage.

[1] The probability that the number of observations with $W = w_0$ is *outside* a given range of fixed proportions of the overall sample size n, i.e. outside $[npc_1, npc_2]$ for $0 < c_1 < 1 < c_2$, asymptotically decays to zero at an exponential rate, which by the Borel–Cantelli Lemma implies it occurs only finitely often; so the local sample size with $W = w_0$ is asymptotically the same order of magnitude as the overall sample size n.

Our new method is based on the derivative approximation $f'(x) \approx [f(x+h) - f(x-h)]/(2h)$. The controls W are temporarily ignored. For some bandwidth $h \to 0$, only observations within the windows $[x_0, x_0 + 2h]$ and $[x_0 - 2h, x_0)$ are considered. A confidence interval (CI) for the difference between the two windows' τ-quantiles can be computed. In principle, any method for quantile treatment effect (QTE) inference would suffice; we use GK and Kaplan (2011) for their high-order accuracy and fast computation. Up to bias, the QTE is approximately $Q_{Y|X}(\tau \mid x_0 + h) - Q_{Y|X}(\tau \mid x_0 - h)$, so dividing by $2h$ yields an approximation of the QME. Definition 3 enumerates these steps.

Definition 3 (QME inference method)**.** The steps to compute our method are:

(i) Choose a desired coverage probability (CP) $1 - \alpha$, quantile of interest $\tau \in (0, 1)$, and point of interest $(x_0, w_0')'$.

(ii) Normalize each continuously distributed element of W_i to have the same sample variance as X_i (since a common bandwidth is used).

(iii) Select a bandwidth $h > 0$ possessing the optimal rate from Theorem 4, such as the plug-in bandwidth suggested immediately after Theorem 4.

(iv) Define the "lower local sample" to be observed values of Y_i for which $(X_i, W_i')' \in C_{h-}$: $X_i \in [x_0 - 2h, x_0)$, continuously distributed elements of W_i are within the corresponding elements of $[w_0 - h, w_0 + h]$, and discretely distributed elements of W_i are equal to the corresponding elements of w_0. Define the "upper local sample" to be almost the same, but instead for $X_i \in [x_0, x_0 + 2h]$.

(v) Construct a CI for the QME at the point of interest: first, apply the QTE inference of GK or Kaplan (2011) with the upper local sample as the "treatment" sample and the lower local sample as the "control" sample; second, divide the endpoint values by $2h$.

The key remaining steps are to characterize the order of magnitude of each source of coverage probability error (CPE) and then solve for the optimal bandwidth rate. This in turn determines the overall CPE.

We maintain the following assumptions and definitions.

Assumption A1. For continuous scalars Y_i and X_i, and vector of controls W_i whose first $d-1$ elements are continuously distributed and whose fixed number of remaining elements are discretely distributed with finite support: vector $(Y_i, X_i, W_i')'$ is sampled iid from its population distribution for $i = 1, \ldots, n$. The point of interest x_0 lies in the interior of the support of X, and similarly for the $d-1$ continuous elements of w_0 and their corresponding supports.

Assumption A2. The joint density of X and W, $f_{X,W}(\cdot, \cdot)$, satisfies $f_{X,W}(x_0, w_0) > 0$ and has a Lipschitz continuous partial derivative in each of the d dimensions in a neighborhood of $(x_0, w_0')'$.For example, for constant $c > 0$ and small enough h to be within the neighborhood, $\left|\frac{\partial}{\partial x} f_{X,W}(x_0 + h, w_0) - \frac{\partial}{\partial x} f_{X,W}(x_0, w_0)\right| \leq c|h|$.

Assumption A3. For all u in a neighborhood of τ and all $\tilde{w}$ in a neighborhood of w_0 (or simply $\tilde{w} = w_0$ if $d-1 = 0$), $Q_{Y|X,W}(u \mid x, \tilde{w})$ has at least two Lipschitz continuous derivatives in x.

Assumption A4. For the bandwidth h, as $n \to \infty$, (i) $h \to 0$, (ii) $nh^d/[\log(n)]^2 \to \infty$.

Assumption A5. The conditional density of Y given X and W is positive at the point of interest: $f_{Y|X,W}\big(Q_{Y|X,W}(\tau \mid x_0, w_0) \mid x_0, w_0\big) > 0$.

Assumption A6. For all y in a neighborhood of $Q_{Y|X,W}(\tau \mid x_0, w_0)$, all x in a neighborhood of x_0, and all w in a neighborhood of w_0 (with the discrete elements simply the same as w_0), $f_{Y|X,W}(y \mid x, w)$ has at least two Lipschitz continuous derivatives in its first argument (y).

Assumption A1 is stronger than necessary for first-order accuracy; see Fan and Liu (2012). Assumptions A5, A6, and A4(ii) satisfy the requirements of GK and Kaplan (2011). Assumptions A2, A3, and A4(i) control the bias. Assumptions A2, A3, and A6 are equivalent to $s_x \geq 2$, $s_Q \geq 3$, and $s_Y \geq 3$ in the notation of Goldman and Kaplan (2014a).

Definition 4 (local sample)**.** The local samples are the Y_i values whose X_i and W_i are within one of the two windows defined by the bandwidth h. When W contains only discrete variables, the two windows and local sample sizes are

$$C_{h+} \equiv \{(x, w_0')' : x \in [x_0, x_0 + 2h]\}, \qquad C_{h-} \equiv \{(x, w_0')' : x \in [x_0 - 2h, x_0)\}, \tag{2}$$

$$N_{n+} \equiv \#(\{Y_i : (X_i, W_i')' \in C_{h+}\}), \qquad N_{n-} \equiv \#(\{Y_i : (X_i, W_i')' \in C_{h-}\}). \tag{3}$$

Additionally, the τ-quantile of Y conditional on $(X, W')' \in C_{h+}$ or C_{h-} is

$$Q_{Y|C_{h+}}(\tau) \equiv \inf\{y : \tau \le P(Y \le y \mid (X, W')' \in C_{h+})\}, \tag{4}$$

$$Q_{Y|C_{h-}}(\tau) \equiv \inf\{y : \tau \le P(Y \le y \mid (X, W')' \in C_{h-})\}. \tag{5}$$

There are two effects of using the local samples: first, the number of observations used, $N_{n+} + N_{n-}$, is of smaller order than the overall sample size n; second, there is bias from including observations with $X_i \neq x_0$. The first effect causes accuracy to increase with h, while the second effect causes accuracy to decrease with h. This tension helps determine the optimal bandwidth. The common order of magnitude of N_{n+} and N_{n-} (ensured by A1, A2, and A4) will be denoted N_n.

3 Theoretical results

The object of interest is the QME defined in (1). We approximate

$$\begin{aligned}
&\frac{Q_{Y|C_{h+}}(\tau) - Q_{Y|C_{h-}}(\tau)}{2h} \\
&\quad = \mathrm{QME}(\tau, x_0, w_0) + \mathrm{ApproxErr} + \frac{\mathrm{Bias}_{h+} - \mathrm{Bias}_{h-}}{2h}, && (6)\\
&\mathrm{Bias}_{h+} = Q_{Y|C_{h+}}(\tau) - Q_{Y|X,W}(\tau \mid x_0 + h, w_0), && (7)\\
&\mathrm{Bias}_{h-} = Q_{Y|C_{h-}}(\tau) - Q_{Y|X,W}(\tau \mid x_0 - h, w_0), && (8)\\
&\mathrm{ApproxErr} = \frac{Q_{Y|X,W}(\tau \mid x_0 + h, w_0) - Q_{Y|X,W}(\tau \mid x_0 - h, w_0)}{2h} - \mathrm{QME}(\tau, x_0, w_0). && (9)
\end{aligned}$$

Lemma 1 characterizes the orders of magnitude of the two error terms in (6).

Lemma 1. *Under Assumptions A1–A6, Definition 4, and the QME definition in* (1)*,*

$$\frac{Q_{Y|C_{h+}}(\tau) - Q_{Y|C_{h-}}(\tau)}{2h} = QME(\tau, x_0, w_0) + O(h^2). \tag{10}$$

The goal of our method is accurate coverage probability (CP). Specifically, let coverage probability error (CPE) be

$$\text{CPE} \equiv P\Big(\text{QME}(\tau, x_0, w_0) \in \widehat{\text{CI}}\Big) - (1-\alpha), \tag{11}$$

where $1-\alpha$ is the nominal CP (e.g., 0.95) and the hat over CI is a reminder that the CI endpoints are random variables. In this section, we characterize the CPE order of magnitude for our new method, starting with the CPE due to bias.

Lemma 2 (CPE from bias)**.** *Under Assumptions A1–A6 and Definition 4, for the method in Definition 3, the CPE due to the bias in Lemma 1 is* $\mathit{CPE}_{\mathit{Bias}} = O(h^3\sqrt{N_n})$*.*

In addition to CPE_{Bias}, CPE_{QTE} denotes the CPE from invoking the unconditional QTE inference method. While there is no "treatment" per se, the numerator of the left-hand side of (10) is equivalent to a τ-QTE: it is the difference in τ-quantile between (sub)populations C_{h+} and C_{h-}, from which we have independent draws of Y.

The rate-limiting source of CPE for both unconditional QTE methods is PDF estimation error in the two-sided case, so contribute $O(N_n^{-2/3})$ CPE. In Kaplan (2011), the PDF estimates enter the standard error directly. In GK, they enter only through nuisance parameter γ, the ratio of "treatment" and "control" PDFs evaluated at the quantile of interest.

If γ is known, then the GK CPE is $O(N_n^{-1})$. (For one-sided CIs, it is $O(N_n^{-1/2})$ either way, for either method.) Here, $\gamma = 1 + O(h)$:

$$\begin{aligned} f_{Y|X}(Q_{Y|X}(\tau \mid x_0+h) \mid x_0+h) &= f_{Y|X}(Q_{Y|X}(\tau \mid x_0+h) \mid x_0) + O(h) \\ &= f_{Y|X}(Q_{Y|X}(\tau \mid x_0) \mid x_0) + O(h), \end{aligned}$$

using the smoothness in (and implied by) A3 and A6. The same holds at $x_0 - h$, so the ratio is $1 + O(h)$. The CPE in GK from estimation of γ is the same order of magnitude as the bias of $\hat{\gamma}$ plus its variance. If the "estimator" $\hat{\gamma} = 1$ is used, then the bias is $O(h)$ and the variance is zero, so CPE is $O(h + N_n^{-1})$.

Lemma 3. *Under Assumptions A1–A6 and Definition 4, the CPE of the two-sided CI for*

$$\frac{Q_{Y|C_{h+}}(\tau) - Q_{Y|C_{h-}}(\tau)}{2h}$$

is $\mathit{CPE}_{QTE} = O\Big(\min\{h + N_n^{-1}, N_n^{-2/3}\}\Big)$ *using GK or* $O(N_n^{-2/3})$ *using Kaplan (2011). Both methods have one-sided* $\mathit{CPE}_{QTE} = O\Big(N_n^{-1/2}\Big)$.

For one-sided CIs, the optimal h balances $h^3\sqrt{N_n}$ and $N_n^{-1/2}$, which yields $h \asymp n^{-1/(3+d)}$ and overall CPE $O(n^{-3/(6+2d)})$. For two-sided CIs, due to the different possible $\hat{\gamma}$, there are two local minima of CPE as a function of h. The results are summarized in Table 1.

Table 1: Effect of h on CPE for two-sided QME CIs, under Assumptions A1–A6.

	Dominant		
h	CPE term	CPE	Notes
$(n^{-1/d}, n^{-1/(1+d)})$	N_n^{-1}	$o(1)$	$O(1)$ CPE with $h \asymp n^{-1/d}$
$n^{-1/(1+d)}$	N_n^{-1}, h	$O(n^{-1/(1+d)})$	local CPE min
$(n^{-1/(1+d)}, n^{-2/(3+2d)}]$	h	$O(h)$	
$(n^{-2/(3+2d)}, n^{-7/(18+7d)})$	$N_n^{-2/3}$	$O(N_n^{-2/3})$	
$n^{-7/(18+7d)}$	$N_n^{-2/3}$, $h^3\sqrt{N_n}$	$O(n^{-12/(18+7d)})$	local CPE min
$(n^{-7/(18+7d)}, n^{-1/(3+d)})$	$h^3\sqrt{N_n}$	$O(h^3\sqrt{N_n})$	
$[n^{-1/(3+d)}, n^{-1/(6+d)})$	$h^3\sqrt{N_n}$	$O(h^3\sqrt{N_n})$	same CPE as normality
$n^{-1/(6+d)}$	$h^3\sqrt{N_n}$	$O(1)$	MSE-optimal h

Theorem 4 (optimal bandwidth and CPE)**.** *For the method in Definition 3 constructing a two-sided CI for the QME defined in* (1)*, under Assumptions A1–A6, the CPE-optimal bandwidth rate is* $h \asymp n^{-7/(18+7d)}$*, and corresponding CPE is* $O(n^{-12/(18+7d)})$*. With* $d = 1$ *and GK, using* $\hat{\gamma} = 1$ *and bandwidth* $h \asymp n^{-1/2}$ *slightly improves CPE, to* $O(n^{-1/2})$ *from* $O(n^{-12/25})$*. For one-sided CIs, the CPE-optimal bandwidth rate is* $h \asymp n^{-1/(3+d)}$*, resulting in* $O(n^{-3/(6+2d)})$ *CPE.*

For a plug-in bandwidth, we suggest taking the plug-in bandwidth from Goldman and Kaplan (2014a) for two-sided inference on $Q_{Y|X,W}(\tau \mid x_0, w_0)$, which is proportional to $n^{-1/(2+d)}$, and multiplying by $n^{4/[(18+7d)(2+d)]}$ to match the two-sided rate in Theorem 4. Since the overall window width is $4h$ here, as opposed to $2h$ in Goldman and Kaplan (2014a), we also divide by two. For one-sided intervals, multiplying instead by $n^{1/[(2+d)(3+d)]}$ gives the correct rate. There is still room for improvement since, for example, the factors influencing single-quantile bias are different than those for QME bias. This is left for future refinement since the current proposal performs reasonably in simulations.

The CPE in Theorem 4 may be compared that for asymptotic normality or for basic bootstraps that have the same CPE as normality. Let $s_Q = k_Q + \gamma_Q$ denote the local smoothness of the conditional quantile function, whose k_Qth (partial) derivative is Hölder continuous with exponent γ_Q. In (Chaudhuri, 1991, p. 764, Step 2), the QME is $h^{-1}\hat{\beta}_1$, so the asymptotic bias is $h^{-1}B_n = h^{s_Q-1}$ (with B_n in Prop. 4.1), the asymptotic variance is $h^{-2}N_n^{-1}$ (Prop. 4.2), and the Bahadur remainder[2] is $h^{-1}N_n^{-1}$, ignoring log terms. To minimize CPE, h equates the orders of CPE from bias and from the Bahadur remainder. The CPE order is the bias (or remainder) times the height of the PDF of the CI endpoint, which is inversely proportional to the standard deviation. Solving $h^{s_Q-1}hN_n^{1/2} = h^{-1}N_n^{-1}hN_n^{1/2}$ gives $h^*_{\text{CPE}} \asymp n^{-1/(s_Q+d)}$ and overall CPE of $(nh^d)^{-1/2} = n^{-s_Q/(2s_Q+2d)}$.

With $s_Q = 3$ as in our A3, the asymptotic normality CPE is $O(n^{-3/(6+2d)})$. This is the same as the one-sided CPE in Theorem 4, but larger than the two-sided $O(n^{-12/(18+7d)})$ CPE. For example, with $d = 1$ (i.e., X is continuous and W is entirely discrete), our two-sided CPE is $O(n^{-12/25}) \approx O(n^{-0.48})$, whereas the CPE using asymptotic normality or bootstrap is $O(n^{-3/8}) \approx O(n^{-0.38})$. With $d = 2$, CPE is $O(n^{-0.38})$ for Theorem 4 and $O(n^{-0.3})$ for normality. For general d, our CPE is smaller by a factor of n^k for $k = (3d+18)/[(7d+18)(2d+6)]$.

With a large enough sample size n and higher degree of smoothness s_Q, a local polynomial

[2] This is smaller than Theorem 3.3(ii) in Chaudhuri (1991) in light of the recent result in Portnoy (2012).

of degree $k_Q = \lceil s_Q \rceil - 1$ can reduce the bias component of CPE for asymptotic normality. Since our method relies on a second-order (uniform) kernel, the bias cannot be reduced with a higher-order kernel. This is the most significant disadvantage of our approach.

4 Simulation study

The simulation study compares our QME CIs with those based on local polynomials, in terms of coverage probability (CP) and interval length (median over simulation replications). The symmetric local polynomial CIs use the conventional normality-based formula, $\hat{\theta} \pm z_{1-\alpha/2}\mathrm{SE}(\hat{\theta})$, with either estimated or bootstrapped standard errors. Code for the new method and the simulations is available from the author's website.

Our suggested plug-in bandwidth is used for the new method ("New" in graph legends). If the point of interest x_0 is near the minimum value of the data sample, then the bandwidth is truncated to $x_0 - X_{\min}$ to reduce risk of bias, and similarly for x_0 near the maximum. For the local polynomials, this bandwidth is then multiplied by a constant so that the CI lengths are all comparable when $n = 1000$ with a flat conditional quantile function and Gaussian error terms, as shown in Figure 1. Local linear ("Linear") and local cubic ("Cubic") methods are shown; local quadratic had consistently worse performance. For the local polynomial, once the bandwidth and thus local sample is determined, function `rq` in the `quantreg` package (Koenker, 2012) in R (R Core Team, 2013) is used to estimate the slope at x_0. For the bootstrap ("boot"), standard errors are computed by `summary.rq` with 299 replications. For the analytic standard error ("C91"), the dispersion matrix Q from Chaudhuri (1991) is used along with an estimate of the conditional PDF of Y given $X = x_0$ evaluated at the quantile of interest.

Three conditional quantile functions are considered. The first is based on Ruppert et al. (2003, §17.5.1), which is also used in the `rqss` vignette as part of the `quantreg` package (Koenker, 2012) in R (R Core Team, 2013). The function as a whole is not intended to be

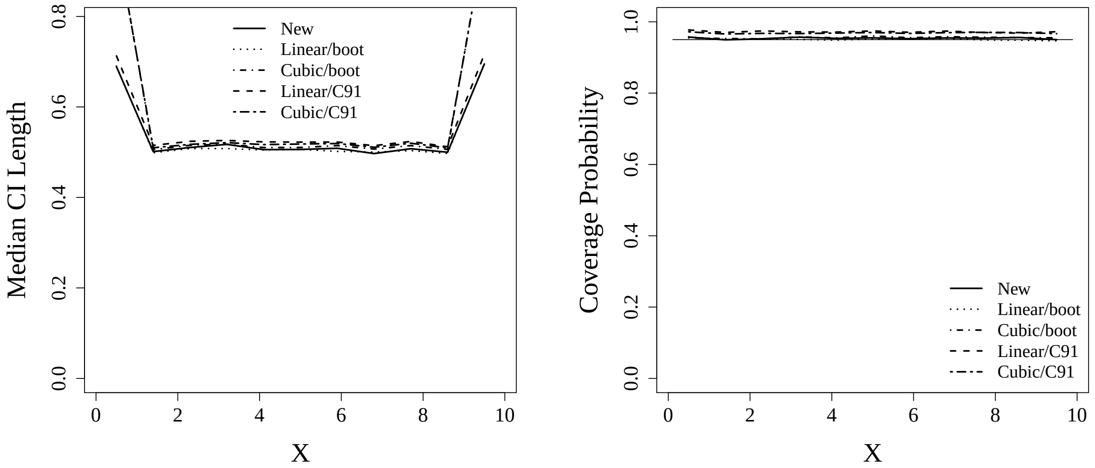

Figure 1: Median CI lengths (left) calibrated to be equal (except local cubic near boundary) when $n = 1000$, $Q_{Y|X}(0.5 \mid X = x) = 0$, $U_i \sim N(0, 1)$, $\sigma(x) = 0.2$, $X_i \sim \text{Unif}(0, 10)$. Right: coverage probabilities in same setup, nominal 0.95.

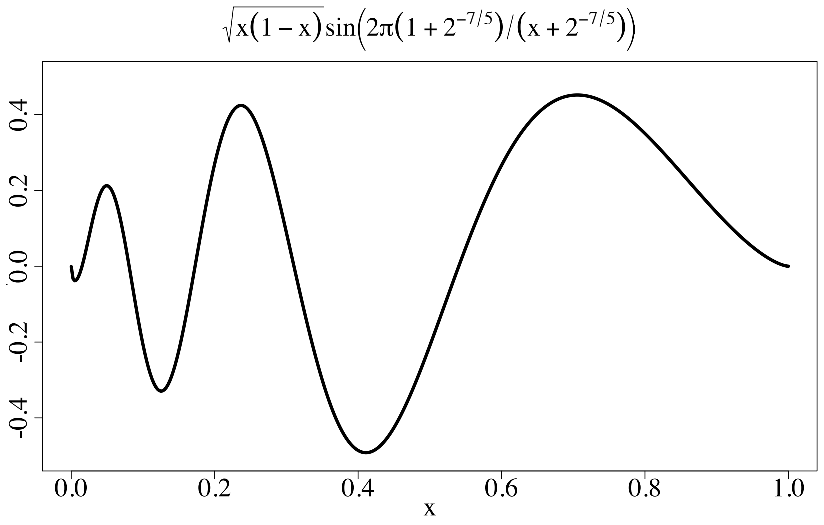

Figure 2: One of three conditional quantile functions for simulations, scaled to unit interval.

realistic for economics, but rather locally incorporate different realistic features at different x_0. Overall, the function becomes smoother as x increases, making inference less difficult. Let $h(x) = \sqrt{x(1-x)} \sin\Bigl(\frac{2\pi(1+2^{-7/5})}{x+2^{-7/5}}\Bigr)$, which is plotted in Figure 2. Scalar $X_i \overset{iid}{\sim} \text{Unif}(0, 10)$, and $Y_i = h(X_i/10) + \sigma(X_i/10)U_i$, where the U_i are iid with conditional τ-quantile equal to zero, and $\sigma(x) = 0.2$ (homoskedastic) or $\sigma(x) = 0.2(1+x)$ (heteroskedastic). Each simulation is 5000 replications. If some x_0 in some replication has local sample size too small to compute the new method (given τ and α), it is discarded for all methods, to keep the comparison fair. In such cases, the local polynomial approach is unlikely to be reliable either, and a method based on extremal quantile regression such as in Chernozhukov and Fernández-Val (2011) is more appropriate. The second setup is the same but with $h(x) = \ln(x)$, and the third has $h(x) = 0$. The 11 points of interest x_0 are evenly spaced between the smallest value (the 0.05-quantile of X) and the largest (0.95-quantile) unless otherwise noted. In the figures, interpolating lines are drawn to ease visual comparison.

Figure 3 shows results with some of the setups from the `rqss` vignette (as-is, other than scaling the X-axis up to 10 instead of one), with normal and Cauchy conditional distributions. (With t_3, results are in between normal and Cauchy, and with χ^2_3 the five methods perform most similarly; heteroskedasticity doesn't change the qualitative comparisons.) The analytic method is somewhat conservative with the normal and much longer with the Cauchy. Although nearly identical to bootstrap at some of the x_0, the new method has the shortest CI lengths.

In some cases, the performance of all methods is similar. This is particularly true of the new method and the bootstraps; Figures 3, 4, and 5 show their similarity even when the analytic CIs are less inaccurate. Figure 1 shows one example where all five methods are quite similar (except the local cubic lengths near the boundaries). Most of the results in this section highlight cases where there are notable differences among the methods, without meaning to suggest that this is always the case.

The analytic local polynomial CIs can have significant under-coverage in cases where the

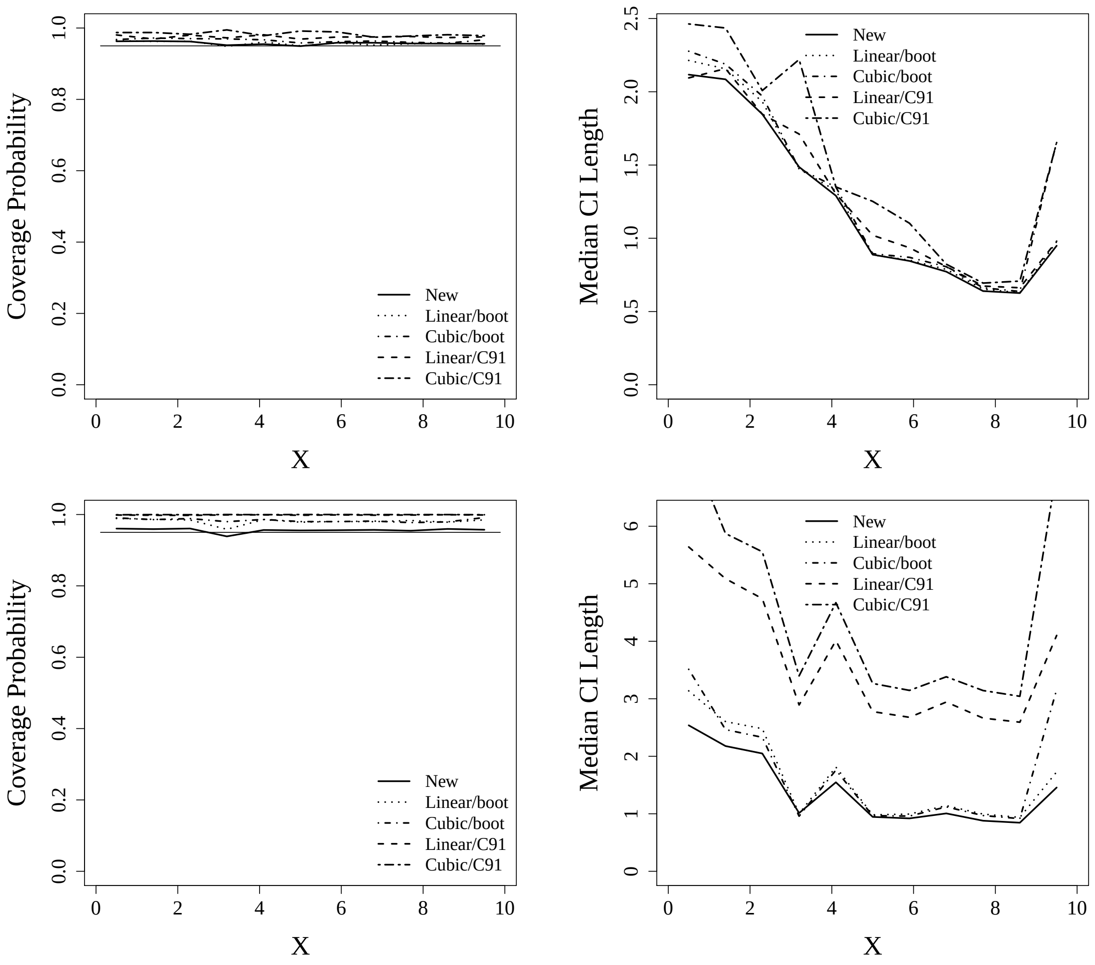

Figure 3: Coverage probability (left; nominal 0.95) and median CI length (right); $n = 400$, $Q_{Y|X}(\tau = 0.5 \mid x) = h(x/10)$ with $h(x) = \sqrt{x(1-x)}\sin\left(\frac{2\pi(1+2^{-7/5})}{x+2^{-7/5}}\right)$, uniform X_i, homoskedastic normal (top) or Cauchy (bottom) errors.

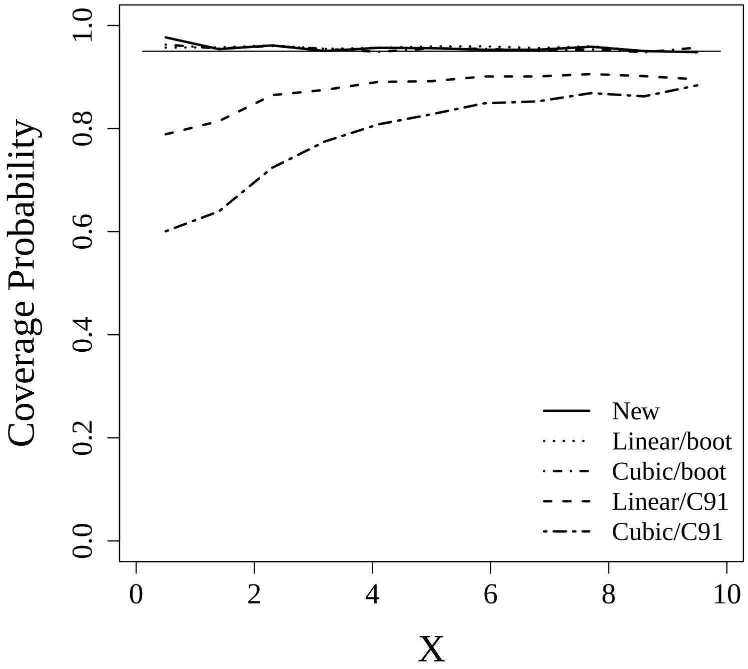

Figure 4: Coverage probability, nominal 0.95; $n = 5000$, $Q_{Y|X}(\tau = 0.05 \mid x) = \ln(x/10)$, homoskedastic t_3 errors, $X_i \sim \text{Unif}(0, 10)$.

new method does not. For example, this is true at all x_0 in Figure 4, for linear and cubic alike. Figure 6 shows other examples.

Other times, the analytic CIs are much longer than the new method's CIs. For example, with a flat conditional median and Cauchy errors, their median length is around four times longer than the others'. Figure 3 shows another example.

While uncommon, there are cases where the bootstrap suffers under-coverage. Figure 6 shows examples. The under-coverage is not severe (lowest CP is 87.9%) but still noticeable, and the new method does not under-cover in any of these. That said, these are cases where the local sample sizes are "small," and our method is slightly conservative, even at points where the bootstrap has correct coverage. ("Small" depends on τ and α; e.g., if $\alpha = 0.05$, $N_{n+} = N_{n-} = 50$ is too small to use GK if $\tau = 0.05$ but not if $\tau = 0.5$.) This is partly from the new method using the QTE CI of Kaplan (2011) when that of GK cannot be computed, which is true in the vast majority of replications in these particular simulations. In these

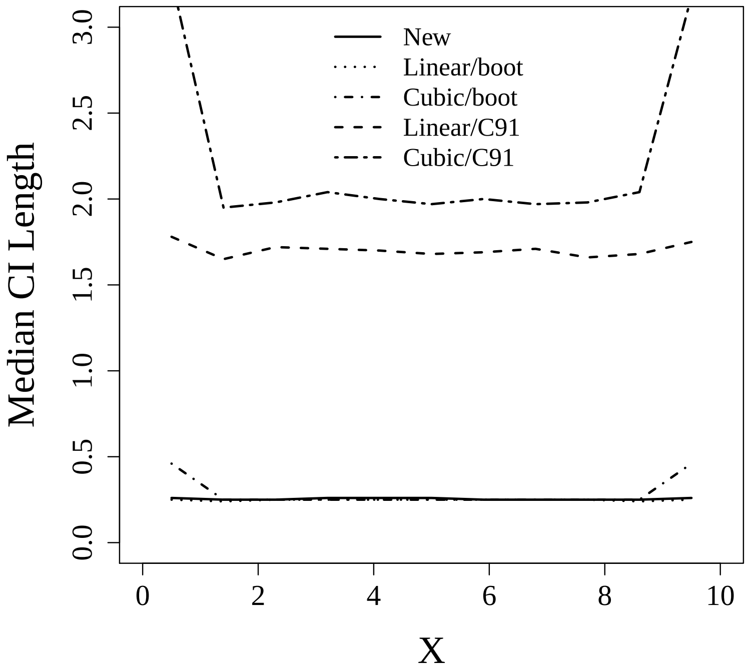

Figure 5: Median CI length; $n = 6000$, $Q_{Y|X}(\tau = 0.5 \mid x) = 0$, homoskedastic Cauchy errors, $X_i \sim \text{Unif}(0, 10)$.

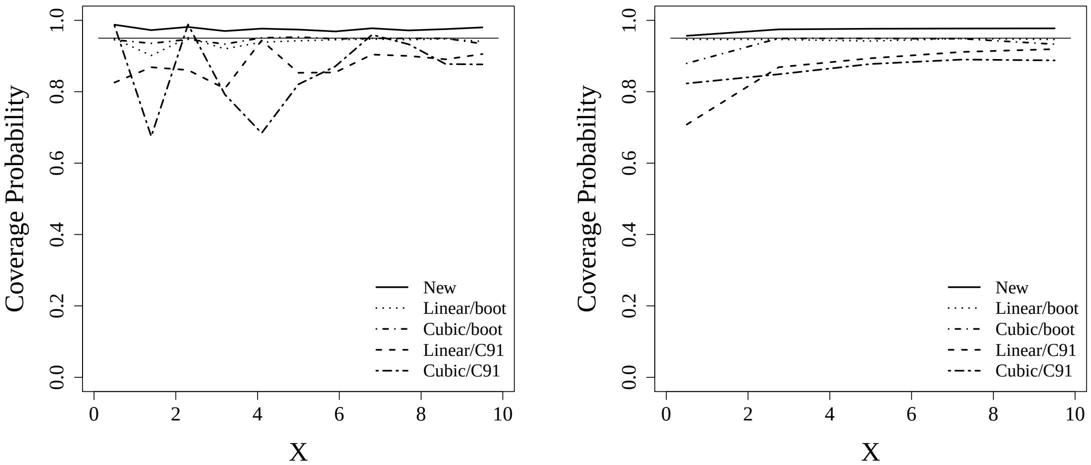

Figure 6: Coverage probability, nominal 0.95; $n = 1000$, homoskedastic normal errors, uniform X_i, and $Q_{Y|X}(\tau = 0.95 \mid x) = h(x/10)$ with $h(x) = \sqrt{x(1-x)}\sin\left(\frac{2\pi(1+2^{-7/5})}{x+2^{-7/5}}\right)$ (left) or $h(x) = \ln(x)$ (right).

cases, there is simply a trade-off between coverage and length when comparing our method with the bootstraps.

There are also cases when the bootstrap produces longer CIs than the new method. They are not multiple times longer like the analytic CIs can be, but the difference can be significant. Examples of this are found in Figure 3.

When x_0 is close enough to the edge of the data that the local sample is truncated (to keep it symmetric about x_0, as noted), the local cubic CIs are often much longer. To obtain comparable length to the new method, the cubic methods must use bigger local samples; when x_0 is near the edge of the data, this is not possible. For example, note the lengths of the local cubic CIs (both analytic and bootstrap) at the smallest and largest x_0 shown in Figures 1 and 3.

Table 2: Computation time in seconds, including bandwidth selection. X_i and Y_i both iid Unif$(0, 1)$, $X_i \perp\!\!\!\perp Y_i$, $\tau = 0.5$; n and number of x_0 (spaced between lower and upper quartiles of X) shown in table. Run (with no other applications open) on 3.2GHz Intel i5 processor with 8GB RAM.

		sample size, n		
Method	$\#x_0$	10^4	10^5	10^6
Bootstrap	1	0.4	6.7	377.6
New	1	0.2	1.6	14.8
Bootstrap	10	2.2	72.8	6056.4
New	10	0.5	2.5	25.0
Bootstrap	100	17.6	707.2	49338.6
New	100	3.4	13.0	128.2

The new method is generally faster to compute than the bootstrap. The bootstrap's disadvantage is somewhat mitigated by only the *local* sample requiring resampling. Still, our new method is often many times faster, and it scales better with the number of x_0, too. For example, with $n = 10^5$ and 10 x_0 points, the local cubic bootstrap takes over a minute, whereas the new method takes three seconds. With 100 different x_0, $n = 10^4$ takes the new method three seconds instead of 18 for bootstrap, and $n = 10^5$ takes 13 seconds instead of

12 minutes. Even $n = 10^6$ is feasible for the new method, taking under a minute with 10 x_0 instead of over an hour for bootstrap. See Table 2 for other examples.

5 Empirical application

We apply our new method to quantile expenditure elasticities,

$$\eta_\tau(y) \equiv \frac{\partial Q_{\ln(q) \mid \ln(y)}(\tau \mid \ln(y))}{\partial \ln(y)},$$

for household total expenditure y, quantity consumed q, and quantile τ. Let $w = pq/y$ be the budget share. Then

$$\frac{\partial w}{\partial y} = p\left[\frac{\partial q}{\partial y} y^{-1} - q y^{-2}\right], \quad \frac{\partial w/w}{\partial y/y} = (p/w)\frac{\partial q}{\partial y} - \frac{pq}{wy} = \frac{\partial q/q}{\partial y/y} - 1,$$
$$\eta_\tau(y) = 1 + \frac{\partial Q_{\ln(w) \mid \ln(y)}(\tau \mid \ln(y))}{\partial \ln(y)}.$$

We examine four “goods” in the annual 2001–2012 U.K. Living Costs and Food Surveys[3] (Office for National Statistics and Department for Environment, Food and Rural Affairs, 2012): food and non-alcoholic beverages (“food” for short; excludes restaurants), alcohol, transportation, and restaurants and hotels. These respectively map to categories 01, 02.1, 07, and 11 from the United Nations Classification of Individual Consumption According to Purpose (COICOP).[4] Code is available online for replication from the raw data. Expenditure amounts are adjusted using annual CPI data.[5] Only households with one or two adults under age 60 (and no children) are used; there are 21,220 observations. Pointwise 90% CIs for the elasticities are constructed at the empirical deciles of $\ln(y)$. There are some qualitative similarities with the nonparametric quantile elasticity estimates in Beatty (2009) for food at home and food away from home in Canadian data. The addition of our robust CIs (he only presented estimates, not CIs) helps distinguish the patterns strongly supported by the data

[3] This succeeded the Expenditure and Food Survey, which succeeded the Family Expenditure Survey.

[4] `http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=5&Lg=1&Top=1`

[5] `http://www.ons.gov.uk/ons/datasets-and-tables/data-selector.html?cdid=D7BT&dataset=mm23&table-id=1.1`

from those that may simply be noise.

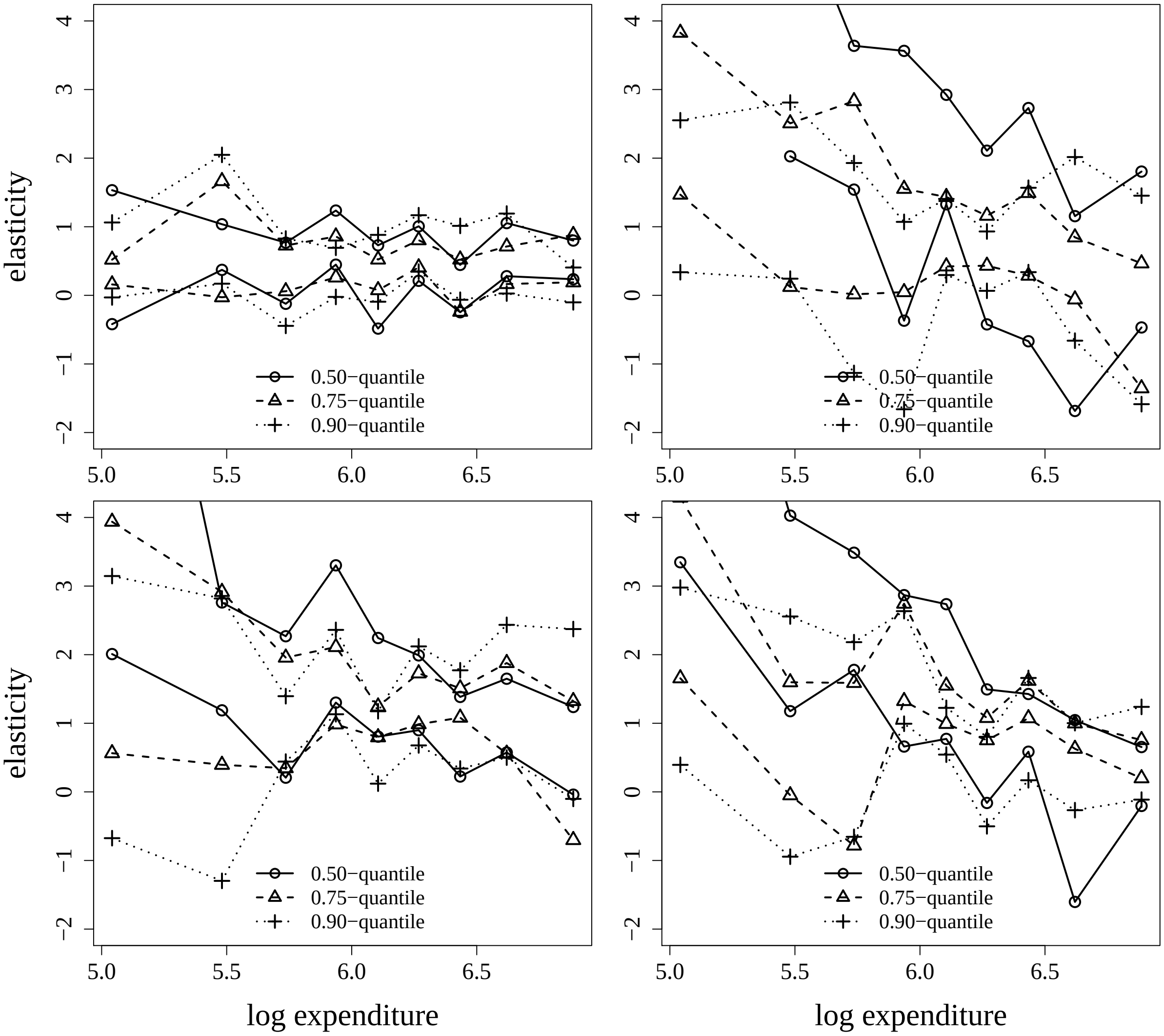

Figure 7: Pointwise 90% confidence intervals for expenditure elasticities: food and non-alcoholic beverages (top left), alcohol (top right), transportation (bottom left), and restaurants and hotels (bottom right).

The results for food confirm that it is most likely a normal good at all levels of expenditure and quantiles. The differences across quantiles and expenditure appear small: of the 27 confidence intervals (9 expenditure levels, 3 quantiles each), 25 contain the value 0.5 (or 0.4). However, note that prior suggestions that food's budget share is a linear (affine) function of log expenditure would imply elasticity decreases with expenditure since $\eta = \frac{\partial w/w}{\partial \ln(y)}$ and w decreases with y. This seems less likely, although these are pointwise intervals, not joint (which would be wider). Most of the intervals at the third through ninth expenditure deciles are relatively precise, but the policy implications of the upper endpoints may still be very different than those of the lower endpoints. Ideally, these could be made more precise by incorporating information like shape constraints.

The intervals for alcohol are less precise than for food since alcohol is (usually) a much smaller budget share than food. The shortest intervals are for $\tau = 0.75$ and $\tau = 0.9$ at the fifth through seventh expenditure deciles, including roughly the range $\eta_\tau(y) = \eta_\tau \in [0.5, 1.2]$. While precise values are mostly elusive, the differences across expenditure and across quantiles appear statistically significant. For example, the elasticity at the 0.9-quantile is roughly constant, while the elasticity at the conditional median (or 0.75-quantile) is decreasing with expenditure. (This matches the descriptive statistics: the 0.9-quantile of the alcohol budget share is 14% in the first decile of expenditure and decreases gradually to just under 9% in the tenth decile, whereas the median is 0% in the first decile and increases to 2.8% in the sixth decile before declining.)

Transportation includes cars and airplane tickets, so makes sense that the lower endpoints would hover around one at most expenditure levels. Again, since there are many zeroes and near-zeroes at the lowest total expenditure decile, the elasticity is quite high, but also imprecisely estimated. The elasticity at the median declines as expenditure rises. In contrast, the elasticity at higher quantiles appears to oscillate around a constant, and there exist constant functions that go through the confidence intervals at all levels of expenditure, e.g. $\eta_{0.75}(y) = 1.2$ and $\eta_{0.9}(y) = 1.3$.

There is an even more pronounced decline in elasticity at the median for restaurants and hotels. The higher quantiles' confidence intervals no longer fit around a constant elasticity. This is particularly clear for the shorter intervals at $\tau = 0.75$: there is an initial fall and rise of elasticity at lower total expenditure levels before a nonlinear decrease from above 1.3 to below 0.7.

6 Conclusion

We propose a new method for nonparametric inference on quantile marginal effects. These objects are structural marginal effects, or weighted averages thereof, in a general nonseparable model. Both new theoretical results and simulations show advantages over normality-based inference with a local polynomial estimator. Future refinement to the bandwidth may improve the performance even more.

References

Abrevaya, J. (2001). The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics*, 26(1):247–257.

Angrist, J., Chernozhukov, V., and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica*, 74(2):539–563.

Beatty, T. K. M. (2009). Semiparametric quantile Engel curves and expenditure elasticities: a penalized quantile regression spline approach. *Applied Economics*, 41(12):1533–1542.

Bitler, M. P., Gelbach, J. B., and Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, 96(4):988–1012.

Buchinsky, M. (1994). Changes in the U.S. wage structure 1963–1987: Application of quantile regression. *Econometrica*, 62(2):405–458.

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. *Annals of Statistics*, 19(2):760–777.

Chernozhukov, V. and Fernández-Val, I. (2011). Inference for extremal conditional quantile models, with an application to market and birthweight risks. *The Review of Economic Studies*, 78(2):559–589.

Chernozhukov, V. and Umantsev, L. (2001). Conditional value-at-risk: Aspects of modeling and estimation. *Empirical Economics*, 26(1):271–292.

Djebbari, H. and Smith, J. (2008). Heterogeneous impacts in PROGRESA. *Journal of Econometrics*, 145(1):64–80.

Fan, Y. and Liu, R. (2012). A direct approach to inference in nonparametric and semiparametric quantile regression models. Working paper.

Goldman, M. and Kaplan, D. M. (2014a). Fractional order statistic approximation for nonparametric conditional quantile inference. Working paper, available at `http://web.missouri.edu/~kaplandm/personalResearch.html`.

Goldman, M. and Kaplan, D. M. (2014b). Nonparametric inference on conditional quantile treatment effects and other objects using L-statistics. Working paper, available at `http://web.missouri.edu/~kaplandm/personalResearch.html`.

Hoderlein, S. and Mammen, E. (2007). Identification of marginal effects in nonseparable models without monotonicity. *Econometrica*, 75(5):1513–1518.

Jackson, E. and Page, M. E. (2013). Estimating the distributional effects of education reforms: A look at project STAR. *Economics of Education Review*, 32:92–103.

Kaplan, D. M. (2011). Improved quantile inference via fixed-smoothing asymptotics and Edgeworth expansion. Working paper, available at `http://web.missouri.edu/~kaplandm/personalResearch.html`.

Koenker, R. (2012). *quantreg: Quantile Regression.* R package version 4.81.

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

Kopczuk, W., Saez, E., and Song, J. (2010). Earnings inequality and mobility in the United States: evidence from Social Security data since 1937. *The Quarterly Journal of Economics*, 125(1):91–128.

Office for National Statistics and Department for Environment, Food and Rural Affairs (2012). Living Costs and Food Survey. 2nd Edition. Colchester, Essex: UK Data Archive. `http://dx.doi.org/10.5255/UKDA-SN-7472-2`.

Portnoy, S. (2012). Nearly root-n approximation for regression quantile processes. *The Annals of Statistics*, 40(3):1714–1736.

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Sasaki, Y. (2012). What do quantile regressions identify for general structural functions? Working paper, available at `http://www.econ2.jhu.edu/People/Sasaki`.

A Mathematical proofs

Lemma 1 proof

For the approximation error, consider a general scalar function $f(\cdot)$ where $f''(\cdot)$ is Lipschitz continuous. For small $h>0$, Taylor approximation yields

$$f(x_0+h) = f(x_0) + hf'(x_0) + (1/2)h^2 f''(\tilde{x}), \tag{12}$$
$$f(x_0-h) = f(x_0) - hf'(x_0) + (1/2)h^2 f''(\bar{x}), \tag{13}$$

where $x_0 - h \le \bar{x} \le x_0 \le \tilde{x} \le x_0 + h$. Subtracting (13) from (12) and dividing by $2h$ yields

$$\frac{f(x_0+h)-f(x_0-h)}{2h} = f'(x_0) + (1/2)h^2[f''(\tilde{x}) - f''(\bar{x})]/(2h) = f'(x_0) + O(h^2) \tag{14}$$

since the Lipschitz continuity implies $|f''(\tilde{x}) - f''(\bar{x})| = O(h)$. Given A3, the approximation error defined in (9) and appearing in (6) is $O(h^2)$.

The bias may be derived from Goldman and Kaplan (2014a, Lem. 5). In our notation and omitting W,

$$\begin{aligned}\text{Bias}_{h+} &\equiv Q_{Y|C_{h+}}(\tau) - Q_{Y|X}(\tau \mid x_0+h) \\ &= -h^2 \frac{f_X(x_0+h)F^{(0,2)}_{Y|X}(\xi_{\tau+} \mid x_0+h) + 2f'_X(x_0+h)F^{(0,1)}_{Y|X}(\xi_{\tau+} \mid x_0+h)}{6f_X(x_0+h)f_{Y|X}(\xi_{\tau+} \mid x_0+h)}\end{aligned}$$

$$
\begin{aligned}
&\quad + O(h^3), && (15)\\
&\xi_{\tau+} \equiv Q_{Y|X}(\tau \mid x_0 + h),\\
F_{Y|X}^{(0,1)}(y \mid x_0 + h) &\equiv \frac{\partial}{\partial x} F_{Y|X}(y \mid x)\bigg|_{x=x_0+h}, \quad F_{Y|X}^{(0,2)}(y \mid x_0 + h) \equiv \frac{\partial^2}{\partial x^2} F_{Y|X}(y \mid x)\bigg|_{x=x_0+h},
\end{aligned}
$$

and similarly for Bias_{h-}.[6] Since sufficient smoothness is assumed, (15) can be rewritten with the functions evaluated at x_0 instead of $x_0 + h$ without changing the error. Thus, first defining ξ_τ,

$$
\begin{aligned}
\xi_\tau &\equiv Q_{Y|X}(\tau \mid x_0),\\
\text{Bias}_{h+} &= -h^2 \frac{f_X(x_0) F_{Y|X}^{(0,2)}(\xi_\tau \mid x_0) + 2 f'_X(x_0) F_{Y|X}^{(0,1)}(\xi_\tau \mid x_0)}{6 f_X(x_0) f_{Y|X}(\xi_\tau \mid x_0)} + O(h^3), && (16)\\
\text{Bias}_{h-} &= -h^2 \frac{f_X(x_0) F_{Y|X}^{(0,2)}(\xi_\tau \mid x_0) + 2 f'_X(x_0) F_{Y|X}^{(0,1)}(\xi_\tau \mid x_0)}{6 f_X(x_0) f_{Y|X}(\xi_\tau \mid x_0)} + O(h^3), && (17)\\
\frac{\text{Bias}_{h+} - \text{Bias}_{h-}}{2h} &= O(h^2). && (18)
\end{aligned}
$$

Plugging (14) and (18) into (6) yields Lemma 1.

Lemma 2 proof

CPE arises when the CI contains the true QME but not the biased QME, or vice-versa. The probability of the CI endpoint being between the true value and the biased value is approximated by the distance between the two values times the height of the PDF of the CI endpoint. For example, if lower one-sided CI endpoint $\hat{E}$ provides 95% CP for the biased value $\xi + B$, then $95\% = P(\hat{E} > \xi + B)$. The true CP is $P(\hat{E} > \xi)$. If $B > 0$, then

$$P(\hat{E} > \xi) - P(\hat{E} > \xi + B) = P(\xi < \hat{E} < \xi + B) = O(B) O(f_{\hat{E}}(\xi))$$

under some continuity of the PDF of $\hat{E}$ in a neighborhood of ξ (assuming $B \to 0$).

The endpoints from either method we use are asymptotically (first-order) equivalent to those from derived from asymptotic normality of a sample quantile, so the PDF is proportional to $h\sqrt{N_n}$ (Chaudhuri, 1991). Multiplying the bias from Lemma 1 by $h\sqrt{N_n}$ yields a CPE contribution of $O(h^3\sqrt{N_n})$ from bias.

[6] The original proof showed an $o(h^3)$ remainder with an extra derivative and $o(h^2)$ when weakening the Lipschitz continuity in our A2 and A3 to Hölder continuity with any exponent $\gamma > 0$; the $O(h^3)$ remainder when $\gamma = 1$ is readily seen from the original derivation.