Quantile Regression Kink Designs

Harold D. Chiang* Yuya Sasaki†‡

Johns Hopkins University

PRELIMINARY AND INCOMPLETE. COMMENTS ARE APPRECIATED.

April 7, 2016

Abstract

This paper discusses quantile regression kink designs (QRKD) for identification and estimation of heterogeneous treatment effects. We first develop causal interpretations of the QRKD estimand. Second, we propose a sample counterpart QRKD estimator, and develop its asymptotic properties for statistical inference of heterogeneous treatment effects. Applying our methods to the Continuous Wage and Benefit History Project (CWBH) data, we find significantly heterogeneous positive moral hazard effects of unemployment insurance benefits on unemployment durations in Louisiana between 1981 and 1982.

Keywords: causal inference, quantile regression kink designs, treatment effects.

JEL Classification: C14, C21

*hchiang7@jhu.edu. Department of Economics, Johns Hopkins University.

†sasaki@jhu.edu. Department of Economics, Johns Hopkins University.

‡We would like to thank Patty Anderson and Bruce Meyer for kindly agreeing to our use of the CWBH data. We benefited from very useful comments by Blaise Melly, Jungmo Yoon, and seminar participants at Penn State and Pittsburgh. All remaining errors are ours.

1 Introduction

Recent empirical papers, including Nielsen, Sørensen and Taber (2010), Landais (2015), Simonsen, Skipper and Skipper (2015) and Card, Lee, Pei and Weber (2016), conduct causal inference via the regression kink design (RKD). A natural extension with a flavor of treatment heterogeneity is the quantile RKD (QRKD), which is the object that we explore in this paper. Specifically, consider the quantile derivative Wald ratio of the form

$$QRKD(\tau) = \frac{\lim_{x \downarrow x_0} \frac{\partial}{\partial x} Q_{Y|X}(\tau \mid x) - \lim_{x \uparrow x_0} \frac{\partial}{\partial x} Q_{Y|X}(\tau \mid x)}{\lim_{x \downarrow x_0} \frac{d}{dx} b(x) - \lim_{x \uparrow x_0} \frac{d}{dx} b(x)} \tag{1.1}$$

at a design point x_0 of a running variable x, where $Q_{Y|X}(\tau|x) := \inf\{y : F(y|x) \geq \tau\}$ defines the τ-th conditional quantile function of Y given $X = x$, and b is a policy function. Note that it is analogous to the estimand of Card, Lee, Pei and Weber (2016):

$$RKD = \frac{\lim_{x \downarrow x_0} \frac{\partial}{\partial x} \mathrm{E}[Y \mid X = x] - \lim_{x \uparrow x_0} \frac{\partial}{\partial x} \mathrm{E}[Y \mid X = x]}{\lim_{x \downarrow x_0} \frac{d}{dx} b(x) - \lim_{x \uparrow x_0} \frac{d}{dx} b(x)}, \tag{1.2}$$

except that the conditional expectations in the numerator are replaced by the corresponding conditional quantiles. While the QRKD estimand (1.1) is of potential interest in the empirical literature for assessment of heterogeneous treatment effects (e.g., Landais, 2011), little seems known about econometric theories of identification, estimation, and inference. This paper develops causal interpretation (identification) and estimation theories for the QRKD estimand (1.1). Consequently, we also propose methods of inference for heterogeneous treatment effects based on the QRKD.

Making causal interpretations of the QRKD estimand (1.1) is perhaps more challenging than the mean RKD estimand (1.2) because the differentiation operator $\frac{d}{dx}$ and the conditional quantile do not 'swap.' For the mean RKD estimand (1.2), the interchangeability of the differentiation operator and the expectation (integration) operator allows each term of the

numerator in (1.2) to be additively decomposed into two parts, namely the causal effects and the endogeneity effects. Taking the difference of two terms in the numerator then cancels out the endogeneity effects, leaving only the causal effects. This trick allows the mean RKD estimand (1.2) to have causal interpretations in the presence of endogeneity. Due to the lack of such interchangeability for the case of quatiles, this trick is not straightforwardly inherited by the quantile counterpart (1.1). Having said this, we show in Section 2 (and more formally in Appendix A.1) that a similar decomposition is possible for the QRKD estimand (1.1) by applying Sasaki (2015), and therefore argue that its causal interpretations are possible.

For estimation of the causal effects, we propose sample-counterpart estimators for the QRKD estimand (1.1) in Sections 3. To derive their asymptotic properties, we take advantage of the existing literature on uniform Bahadur representations for quantile-type loss functions, including Kong, Linton and Xia (2010), Guerre and Sabbah (2012; 2014), and Qu and Yoon (2015a). Qu and Yoon (2015b) apply the results of Qu and Yoon (2015a) to develop methods of statistical inference with quantile regression discontinuity designs (QRDD), which is closely related to our QRKD framework. We take a similar approach with suitable modifications to derive asymptotic properties of our QRKD estimator. Weak convergence results for the estimator as a quantile process are derived. Applying the weak convergence results, we propose procedures for testing treatment significance and treatment heterogeneity following Koenker and Xiao (2002), Chernozhukov and Fernández-Val (2005) and Qu and Yoon (2015b). In Section 4, we conduct Monte Carlo experiments. The results of the experiments support our theoretical properties.

Literature: The model and method studied in this paper fall in the broad framework of design-based causal inference methods, including RDD and RKD. There is an extensive body of literature on RDD by now – see for example the special issue of *Journal of Econometrics*

edited by Imbens and Lemieux (2008) and the literature review by Lee and Lemieux (2010). The first extension to quantile treatment effects in the RDD framework was made by Frandsen, Frölich and Melly (2012). More recently, Qu and Yoon (2015b) develop uniform inference methods with QRDD that empirical researchers can use to test a variety of important empirical questions on heterogeneous treatment effects. While the RDD has a rich set of empirical and theoretical results including these quantile extensions, the RKD method that developed more recently does not have a quantile counterpart in the literature yet, despite potential demands for it by empirical researchers (e.g., Landais, 2011). Our paper can be seen as either a quantile extension to Card, Lee, Pei and Weber (2016) or a RKD counterpart of Frandsen, Frölich and Melly (2012) and Qu and Yoon (2015b).

2 Causal Interpretation of the QRKD Estimand

The causal relation of interest is represented by the structural equation

$$y = g(b, x, \epsilon).$$

The outcome y is determined through the structural function g by two observed factors, $b \in \mathbb{R}$ and $x \in \mathbb{R}$, and a scalar unobserved factor, $\epsilon \in \mathbb{R}$. We assume that g is monotone increasing in ϵ, effectively imposing the rank invariance; causal interpretations in a more general setup with non-monotone g and/or multivariate ϵ is discussed in Appendix A.1. The factor b is a treatment input, and is in turn determined by the running variable x through the structural equation

$$b = b(x)$$

for a known policy function b. We say that b has a kink at x_0 if $b'(x_0^+) := \lim_{x \to x_0^+} \frac{db(x)}{dx} \neq \lim_{x \to x_0^-} \frac{db(x)}{dx} =: b'(x_0^-)$ is true, where $x \to x_0^+$ and $x \to x_0^-$ mean $x \downarrow x_0$ and $x \uparrow x_0$, respectively.

Throughout this paper, we assume that the location, x_0, of the kink is known from a policy-based research design, as is the case with Card, Lee, Pei and Weber (2016).

Assumption 1. *$b'(x_0^+) \neq b'(x_0^-)$ holds, and b is continuous on $\mathbb{R}$ and differentiable on $\mathbb{R} \setminus \{x_0\}$.*

The structural partial effects are $g_1(b,x,\epsilon) := \frac{\partial}{\partial b} g(b,x,\epsilon)$, $g_2(b,x,\epsilon) := \frac{\partial}{\partial x} g(b,x,\epsilon)$ and $g_3(b,x,\epsilon) := \frac{\partial}{\partial \epsilon} g(b,x,\epsilon)$. In particular, a researcher is interested in g_1 which measures heterogeneous partial effects of the treatment intensity b on an outcome y. While the structural partial effect g_1 is of interest, it is not clear if the QRKD estimand (1.1) provides any information about g_1. In this section, we argue that (1.1) does have a causal interpretation in the sense that it measures the structural causal effect $g_1(b(x_0),x_0,\epsilon)$ at the τ-th conditional quantile of ε given $X = x_0$.

Under regularity conditions (see Appendix A.1), an application of Lemma 1 of Sasaki (2015) to the current model yields the decomposition

$$\frac{\partial}{\partial x} Q_{Y|X}(\tau \mid x) = g_1(b(x),x,\epsilon) \cdot b'(x) + g_2(b(x),x,\epsilon) - \frac{\int_{-\infty}^{\epsilon} \frac{\partial}{\partial x} f_{\varepsilon|X}(e \mid x) de}{f_{\varepsilon|X}(\epsilon \mid x)} \cdot g_3(b(x),x,\epsilon), \quad (2.1)$$

where $\tau = F_{\varepsilon|X}(\epsilon \mid x)$. The first term on the right-hand side is the partial effect of the running variable x on the outcome y through the policy function b. The second term is the direct partial effect of the running variable on the outcome y. The third term measures the effect of endogeneity in the running variable x. We can see that this third term is zero under exogeneity, $\frac{\partial}{\partial x} f_{\varepsilon|X} = 0$. In order to get the causal effect $g_1(b(x),x,\epsilon)$ of interest through the QRKD estimand (1.1), therefore, we want to remove the last two terms in (2.1).

Suppose that the designed kink condition of Assumption 1 is true, but all the other functions, g_1, g_2, g_3, $1/f_{\varepsilon|X}$ and $\frac{\partial}{\partial x} f_{\varepsilon|X}$, in the right-hand side of (2.1) are continuous in (b,x) at $(b(x_0),x_0)$. Then, (2.1) yields

$$\frac{\frac{\partial}{\partial x} Q_{Y|X}(\tau \mid x_0^+) - \frac{\partial}{\partial x} Q_{Y|X}(\tau \mid x_0^-)}{b'(x_0^+) - b'(x_0^-)} = g_1(b(x_0),x_0,\epsilon),$$

showing that the QRKD estimand (1.1) measures the structural causal effect $g_1(b(x_0), x_0, \epsilon)$ of b on y for the subpopulation of individuals at the τ-th conditional quantile of ε given $X = x_0$. This section provides only an informal argument for ease of exposition, but Appendix A.1 provides a formal mathematical argument under a general setup without the rank invariance assumption. Furthermore, we provide a result for the case of fuzzy QRKD in Appendix A.2.

3 Estimation and Inference

We propose to estimate the QRKD estimand (1.1) by the sample counterpart

$$\widehat{QRKD}(\tau) \;=\; \frac{\hat{\beta}^+(\tau) - \hat{\beta}^-(\tau)}{\lim_{x\downarrow x_0} b'(x) - \lim_{x\uparrow x_0} b'(x)} \tag{3.1}$$

with the two terms in the numerator given by the one-sided local linear quantile smoothers

$$\hat{\beta}^+(\tau) = \iota_2' \arg\min_{\alpha,\beta} \sum_{i=1}^{n} d_i^+ K(\frac{x_i - x_0}{h_{n,\tau}}) \rho_\tau(y_i - \alpha - \beta(x_i - x_0)) \quad \text{and}$$
$$\hat{\beta}^-(\tau) = \iota_2' \arg\min_{\alpha,\beta} \sum_{i=1}^{n} d_i^- K(\frac{x_i - x_0}{h_{n,\tau}}) \rho_\tau(y_i - \alpha - \beta(x_i - x_0)),$$

for $\tau \in T$, where $T \subset (0,1)$ is a closed interval, K is a kernel function, $\rho_\tau(u) = u(\tau - \mathbb{1}\{u < 0\})$, $d_i^+ = \mathbb{1}\{x_i \geq x_0\}$, $d_i^- = \mathbb{1}\{x_i \leq x_0\}$, and $\iota_2 = [0,1]'$. A researcher observing a sample $\{y_i, x_i\}_{i=1}^n$ of n observations can compute (3.1) explicitly to estimate (1.1).

In the remainder of this section, we first develop Bahadur representations for the component estimators, $\hat{\beta}^+(\tau)$ and $\hat{\beta}^-(\tau)$, uniformly in τ over T. Second, weak convergence results are developed for quantile processes for $\hat{\beta}^+(\tau)$ and $\hat{\beta}^-(\tau)$, which in turn yield weak convergence results for quantile processes for the QRKD estimator of treatment effects. Third, using the weak convergence results, we propose some tests of hypotheses concerning heterogeneous treatment effects.

We fix some notations for convenience of our analysis. Although they are not direct objects of interest, the level estimators are denoted by $\hat{\alpha}^+(\tau) = \iota_1' \arg\min_{\alpha,\beta} \sum_{i=1}^n d_i^+ K(\frac{x_i - x_0}{h_{n,\tau}}) \rho_\tau(y_i - \alpha - \beta(x_i - x_0))$ and $\hat{\alpha}^-(\tau) = \iota_1' \arg\min_{\alpha,\beta} \sum_{i=1}^n d_i^- K(\frac{x_i - x_0}{h_{n,\tau}}) \rho_\tau(y_i - \alpha - \beta(x_i - x_0))$, where $\iota_1 = [1, 0]'$. We define the kernel-dependent constant matrices $N^+(\tau) = \int_0^\infty (1, u)'(1, u) K(u) du$ and $N^-(\tau) = \int_{-\infty}^0 (1, u)'(1, u) K(u) du$. Some transformations of data points are succinctly written as $z_{i,n,\tau}' = (1, (x_i - x_0)/h_{n,\tau})$ and $K_{i,n,\tau} = K(\frac{x_i - x}{h_{n,\tau}})$. Define the linear extrapolation error $e_i = \big[Q(\tau|x_0^+) + (x_i - x_0)\frac{\partial Q(\tau|x_0^+)}{\partial x}\big] - Q(\tau|x_i)$ and the estimation errors $\hat{\phi}(\tau) = \sqrt{n h_{n,\tau}}[\hat{\alpha}^+(\tau) - Q(\tau|x_0^+),\ h_{n,\tau}(\hat{\beta}^+(\tau) - \frac{\partial Q(\tau|x_0^+)}{\partial x})]'$. We make the following assumptions.

Assumption 2. *There exist $\bar{x} > x_0$ and $\underline{x} < x_0$ such that the following conditions are satisfied:*

(i) (a) The density function $f_X(\cdot)$ exists and is continuously differentiable in a neighborhood of x_0 and $0 < f_X(x_0) < \infty$. (b) $\{(y_i, x_i)\}_{i=1}^n$ is an i.i.d. sample of n observations of the bivariate random vector (Y, X).

(ii) (a) $f_{Y|X}(y|x)$ is Lipschitz continuous on $[\inf_{(\tau,x)\in T\times(x_0,\bar{x}]} Q(\tau|x),\ \sup_{(\tau,x)\in T\times(x_0,\bar{x}]} Q(\tau|x)] \times (x_0, \bar{x}]$ and $[\inf_{(\tau,x)\in T\times[\underline{x},x_0)} Q(\tau|x),\ \sup_{(\tau,x)\in T\times[\underline{x},x_0)} Q(\tau|x)] \times [\underline{x}, x_0)$. (b) There exist finite constants $f_L > 0$, $f_U > 0$ and $\epsilon > 0$, such that $f_{Y|X}(Q(\tau|x) + \eta|x)$ lies between f_L and f_U for all $\tau \in T$, $|\eta| \le \epsilon$ and $x \in [\underline{x}, \bar{x}]$.

(iii) (a) $Q(\tau|x_0^+)$, $\partial Q(\tau|x_0^+)/\partial\tau$,$Q(\tau|x_0^-)$, and $\partial Q(\tau|x_0^-)/\partial\tau$ exist and are Lipschitz continuous in τ on T. (b) $\partial Q(\tau|x)/\partial x$ and $\partial Q^2(\tau|x)/\partial x^2$ exist and are Lipschitz continuous on $\{(x, \tau)|x \in (x_0, \bar{x}], \tau \in T\}$ and $\{(x, \tau)|x \in [\underline{x}, x_0), \tau \in T\}$.

(iv) The kernel K is compactly supported, Lipschitz, differentiable, and satisfying $K(\cdot) \ge 0$, $\int K(u)du = 1$, $\int uK(u)du = 0$. Also, $\int_0^\infty u^k K(u)du$ and $\int_{-\infty}^0 u^k K(u)du$ are finite for $k = 1, 2, 3$. The matrices N^+ and N^- are positive definite.

(v) The bandwidths satisfy $h_{n,\tau} = c(\tau)h_n$, where $nh_n^3 \to \infty$ and $h_n = o(n^{-1/5})$ as $n \to \infty$, and

$c(\cdot)$ is Lipschitz continuous satisfying $0 < \underline{c} \leq c(\tau) \leq \overline{c} < \infty$ for all $\tau \in T$.

Part (i) (a) requires smoothness of the density of the running variable. This can be interpreted as the design requirement for absence of endogenous sorting across the kink point x_0. The i.i.d assumption in part (i) (b) is usually considered to be satisfied for micro data of random samples. Part (ii) concerns about regularities of the conditional density function of Y given X. It requires sufficient smoothness, but does not rule out quantile regression kinks at x_0, which is the main crucial assumption for our identification argument. Part (iii) concerns about regularities of the conditional quantile function of Y given X. Like part (ii), it does not rule out quantile regression kinks at x_0. Part (iv) prescribes requirements for kernel functions to be chosen by users. In Section 4 for Monte Carlo experiments, we propose an example of such a choice to satisfy this requirement. Finally, part (v) specifies the rate at which the bandwidth parameters diminish as the sample size becomes large. It obeys the standard rate for a first-order derivative estimation, but we also require its uniformity over quantiles τ in T. While $h_n = o(n^{-1/5})$ is required for a valid inference without bias reduction. This requirement can be relaxed to $h_n = O(n^{-1/5})$ if one is willing to take an additional step of bias reduction – see Appendix B. Under this set of assumptions, we obtain uniform Bahadur representations for the component estimators, $\hat{\beta}^+(\tau)$ and $\hat{\beta}^-(\tau)$, of our interest.

Lemma 1. *Under Assumption 2, we have*

$$\hat{\phi}(\tau) = \begin{bmatrix} \sqrt{nh_{n,\tau}}(\hat{\alpha}^+(\tau) - Q(\tau|x_0^+)) \\ \sqrt{nh_{n,\tau}^3}(\hat{\beta}^+(\tau) - \frac{\partial Q(\tau|x_0^+)}{\partial x}) \end{bmatrix} = (nh_{n,\tau}^5)^{\frac{1}{2}} \frac{(N^+)^{-1}}{2} \int_0^\infty u^2 \frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}(1,u)' K(u) du$$
$$+ \frac{(N^+)^{-1}(nh_{n,\tau})^{-\frac{1}{2}} \sum_{i=1}^n (\tau - \mathbb{1}\{y_i \leq Q(\tau|x_i)\}) z_{i,n,\tau} K_{i,n,\tau} d_i^+)}{f_X(x_0) f_{Y|X}(Q(\tau|x_0^+)|x_0^+)} + o_p(1)$$

uniformly in $\tau \in T$. *In particular, we have*

$$\begin{aligned}&\sqrt{nh_{n,\tau}^3}\Big(\hat{\beta}^+(\tau) - \frac{\partial Q(\tau|x_0^+)}{\partial x} - h_{n,\tau}\frac{\iota_2'(N^+)^{-1}}{2}\int_0^\infty u^2\frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}(1,u)'K(u)du\Big)\\ =&\frac{\iota_2'(N^+)^{-1}(nh_{n,\tau})^{-\frac{1}{2}}\sum_{i=1}^n(\tau - \mathbb{1}\{y_i \le Q(\tau|x_i)\})z_{i,n,\tau}K_{i,n,\tau}d_i^+)}{f_X(x_0)f_{Y|X}(Q(\tau|x_0^+)|x_0^+)} + o_p(1)\end{aligned}$$

uniformly in $\tau \in T$. *Similar results hold for* $\hat{\beta}^-(\tau)$.

A proof is provided in Appendix D.2 – auxiliary results of uniform consistency that are used to prove this theorem are also provided in Appendix D.1. The Bahadur representation obtained in this lemma is uniform in quantiles $\tau \in T$ for a fixed location of the running variable x. Kong, Linton and Xia (2010) derived a Bahadur representation that is uniform in x for a fixed quantile level. Guerre and Sabbah (2012) derived a result that is uniform in both τ and x for interior points – see also Guerre and Sabbah (2014). Since we are interested in a representation at the boundary point x_0 of the truncated distribution, and since we did not require the uniformity in x, and we developed our approach more closely following Qu and Yoon (2015a).

Applying the uniform Bahadur representation in Lemma 1, we now establish weak convergence results for our component estimators. We focus on $\hat{\beta}^+$, but a similar result follows for $\hat{\beta}^-$. Furthermore, since these right and left estimators use two mutually exclusive sides of an i.i.d. sample on the running variable, their asymptotic distributions are mutually independent processes. This independence allows the asymptotic distribution of the QRKD estimator (3.1) to be easily constructed using the asymptotic distributions of the two quantile processes of $\hat{\beta}^+$ and $\hat{\beta}^-$.

Theorem 1. *Under Assumptions 2, we have the weak convergence*

$$\begin{aligned}&\sqrt{nh_{n,\tau}^3 f_X(x_0)}f_{Y|X}(Q(\tau|x_0^+)|x_0^+)\times\\ &\Big(\hat{\beta}^+(\tau) - \frac{\partial Q(\tau|x_0^+)}{\partial x} - h_{n,\tau}\frac{\iota_2'(N^+)^{-1}}{2}\int_0^\infty u^2\frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}(1,u)'K(u)du\Big) \Rightarrow G^+(\tau),\end{aligned}$$

for the zero mean Gaussian process $G^+(\tau)$ *defined over* T *with covariance function*

$$E(G^+(r)G^+(s)) = (\kappa(r)\kappa(s))^{-1/2}(r \wedge s - rs)\iota_2'(N^+)^{-1}T^+(r,s)(N^+)^{-1}\iota_2,$$

for each $r, s \in T$, *where* $T^+(r,s) = \int_0^\infty \begin{bmatrix} 1 & \frac{u}{\kappa(r)} \end{bmatrix}' K(\frac{u}{\kappa(r)}) \begin{bmatrix} 1 & \frac{u}{\kappa(s)} \end{bmatrix} K(\frac{u}{\kappa(s)})du$ *with* $\kappa(\tau) = h_{n,\tau}/h_{n,1/2} = \frac{c(\tau)}{c(1/2)} \geq (\underline{c}/\overline{c}) > 0$. *A similar result follows for* $\hat{\beta}^-(\tau)$.

A proof is provided in Appendix D.3. While we write the above weak convergence result explicitly accounting for the finite-sample bias term $h_{n,\tau}\frac{\iota_2'(N^+)^{-1}}{2}\int_0^\infty u^2 \frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}(1,u)'K(u)du$, it goes away in large sample as $h_{n,\tau}$ goes to zero uniformly in $\tau \in T$. In other words, this bias term can be considered to be absent in the above equation. With this said, in case one wish to reduce this finite-sample bias, we propose in Appendix B how to estimate this bias term.

We are now ready to present a weak convergence result for our QRKD estimator (3.1). By Section 2.1 of Giné and Nickl (2015), the sum of two independent Gaussian processes is a Gaussian process with the mean (respectively, the covariance) being the sum of the means (respectively, the covariances).

Corollary 1. *Under Assumptions 1 and 2 we have the weak convergence*

$$\begin{aligned} &\sqrt{nh_{n,\tau}^3}\Big(\widehat{QRKD}(\tau) - QRKD(\tau)\Big) \\ \Rightarrow Y(\tau) &= \frac{1}{\sqrt{f_X(x_0)}\left(b'(x_0^+) - b'(x_0^-)\right)}\left[\frac{G^+(\tau)}{f_{Y|X}(Q(\tau|x_0^+)|x_0^+)} - \frac{G^-(\tau)}{f_{Y|X}(Q(\tau|x_0^-)|x_0^-)}\right]. \end{aligned}$$

The random process $Y(\cdot)$ has mean zero, as $G^+(\tau)$ and $G^-(\tau)$ do. For any pair r, $s \in T$ of quantiles, the covariance can be computed by:

$$\begin{aligned} E[Y(r)Y(s)] =& \frac{1}{f_X(x_0)(b'(x_0^+) - b'(x_0^-))^2} \\ &\times \left[\frac{EG^+(r)G^+(s)}{f_{Y|X}(Q(r|x_0^+)|x_0^+)f_{Y|X}(Q(s|x_0^+)|x_0^+)} + \frac{EG^-(r)G^-(s)}{f_{Y|X}(Q(r|x_0^-)|x_0^-)f_{Y|X}(Q(s|x_0^-)|x_0^-)}\right]. \end{aligned}$$

This uniform convergence result is applicable to many purposes, such as to compute uniform confidence bands for the QRKD. Of particular interest may be the empirical tests of the following hypotheses among others.

$$\text{Treatment Significance} \quad H_0^S: \quad QRKD(\tau) = 0 \quad \text{for all } \tau \in T.$$

$$\text{Treatment Heterogeneity} \quad H_0^H: \quad QRKD(\tau) = QRKD(\tau') \quad \text{for all } \tau, \tau' \in T.$$

They are both considered in Koenker and Xiao (2002), Chernozhukov and Fernández-Val (2005) and Qu and Yoon (2015b), among others. Following the approach of these preceding papers, the two hypotheses, H_0^S and H_0^H, may be tested using the statistics

$$\begin{aligned} WS_n(T) &= \sup_{\tau \in T} |\widehat{QRKD}(\tau)| \qquad \text{and} \\ WH_n(T) &= \sup_{\tau \in T} \left| \widehat{QRKD}(\tau) - \int_T \widehat{QRKD}(\tau') d\tau' \right|, \end{aligned}$$

respectively. Consequence of Corollary 1 are the following asymptotic distributions of these test statistics, a proof of which is provided in Appendix D.4.

Corollary 2. *Under Assumptions 1 and 2, we have*

(i) $\sqrt{nh_{n,\tau}^3} WS_n(T) \Rightarrow \sup_{\tau \in T} |Y(\tau)|$ *under the null hypothesis* H_0^S*; and*

(ii) $\sqrt{nh_{n,\tau}^3} WH_n(T) \Rightarrow \sup_{\tau \in T} |\phi'_{QRKD}(Y)(\tau)|$ *under the null hypothesis* H_0^H*, where* $\phi'_{QRKD}(g)(\tau) = g(\tau) - \int_T g(\tau') d\tau'$ *for all* $g \in L_m^\infty(T)$*, the space of all bounded, measurable, real-valued functions defined on* T*.*

4 Monte Carlo Experiments

Consider the following policy function with a kink at $x_0 = 0$.

$$b(x) = \begin{cases} 2x & \text{if } x \leqslant 0 \\ 0.5x & \text{if } x > 0 \end{cases}$$

For convenience of assessing the performance of our estimator for homogeneous treatment effects and heterogeneous treatment effects, we consider the following three outcome structures.

$$\begin{aligned}
&\text{Structure 0:} \quad && g(b,x,\epsilon) = 0.0b + 0.5x + 0.05x^2 + \epsilon \\
&\text{Structure 1:} \quad && g(b,x,\epsilon) = 0.5b + 0.5x + 0.05x^2 + \epsilon \\
&\text{Structure 2:} \quad && g(b,x,\epsilon) = 0.5[0.5 + F_\varepsilon(\epsilon)]b + 0.5x + 0.05x^2 + \epsilon
\end{aligned}$$

where F_ε denotes the CDF of ε. Note that Structures 0 and 1 entail homogeneous treatment effects, while Structure 2 entails heterogeneous treatment effects across quantiles τ as follows.

$$\begin{aligned}
&\text{Structure 0:} \quad && g_1(b,x,Q_\varepsilon(\tau)) = 0.0 \\
&\text{Structure 1:} \quad && g_1(b,x,Q_\varepsilon(\tau)) = 0.5 \\
&\text{Structure 2:} \quad && g_1(b,x,Q_\varepsilon(\tau)) = 0.5[0.5 + \tau]
\end{aligned}$$

Allowing for endogeneity, we generate the primitive data according to

$$\begin{pmatrix} X_i \\ \varepsilon_i \end{pmatrix} \overset{i.i.d.}{\sim} N\left(0, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_\varepsilon \\ \rho\sigma_X\sigma_\varepsilon & \sigma_\varepsilon^2 \end{pmatrix}\right),$$

where $\sigma_X = \sigma_\varepsilon = \rho = 0.5$. For estimation, we use the tricube kernel defined by

$$K(u) = \frac{70}{81}\left(1 - |u|\right)^3 \mathbb{1}\{|u| < 1\}.$$

The bandwidths are selected with our choice rule based on the MSE minimization – see Appendices B and E for details.

Figures 1 and 2 show Monte Carlo distributions of the QRKD estimates under Structure 1 and Structure 2, respectively. In each of the two figures, the left column lists results with no bias reduction, and the right column lists results with bias reduction (based on Appendix B). The top row, the middle row and the bottom row list results for the sample sizes $N = 1,000$,

2,000 and 4,000, respectively. In each graph, the horizontal axis measures quantiles τ, while the vertical axis measures the QRKD. The true QRKD are indicated by solid gray lines. The other broken curves indicate the 5-th, 10-th, 50-th, 90-th, and 95-th percentiles of the Monte Carlo distributions of our estimates based on 5,000 Monte Carlo iterations. We observe the following four points in these figures. First, the displayed distribution shrinks for each structure at each quantile τ as the sample size N increases. Second, the results in the left column with no bias reduction are biased. Third, the results in the right column with bias reduction indeed improve the biases. Fourth, even the results in the left column with no bias reduction have the bias diminishing in N, evidencing the asymptotic unbiasedness. These aspects of the results confirm that our methods work as in theory.

In order to more quantitatively analyze the finite sample pattern, we summarize some basic statistics for the Monte Carlo distributions in Tables 1 and 2 for Structure 1 and Structure 2, respectively. In each table, the four column groups list the absolute biases (MC Bias), the standard deviations (MC SD), root mean squared errors (MC RMSE), and the rejection frequencies for point-wise 5%-level t-tests for the null hypotheses of the true QMTE values (MC 5% Size). The upper half of each table displays statistics for the results with no bias reduction (No BR), and the lower half of each table displays statistics for the results with bias reduction (With BR). For each structure at each quantile τ, we again observe that SD and RMSE decrease as the sample size N increases. We also observe that the results with bias reduction have close-to-zero biases whereas those without bias reduction are biased. However, the results with bias reduction tend to produce slightly larger SD and RMSE than those without bias reduction. These patterns are consistent with our previous discussions on Figures 1 and 2. Observe that the MC 5% sizes are reasonably accurate at each quantile for each structure with bias reduction, while they can explode without bias reduction. We remark that the latter fact

does not contradict with our theory, because the practical choice of optimal bandwidth based on an approximate MSE fails to under-smooth the estimates while our theory of asymptotic distribution without bias reduction requires an under-smoothing. As such, we recommend the bias reduction in practice when one chooses to use the optimal bandwidth.

While these sizes concern about point-wise inference, we also provide uniform inference results. Table 3 shows rejection probabilities for the 95% level uniform test of significance (panel A) and the 95% level uniform test of heterogeneity (panel B) based on 1,000 Monte Carlo iterations. Panel A shows that the rejection probability for the test of the null hypothesis of insignificance does not increase in the sample size for Structure 0, while it increases in the sample size for each of Structure 1 and Structure 2. Panel B shows that the rejection probability for the test of the null hypothesis of homogeneity does not increase in the sample size for Structure 0 and Structure 1, while it increases in the sample size for Structure 2. These results are consistent with the construction of Structure 0, Structure 1, and Structure 2.

5 An Empirical Illustration

In labor economics, causal effects of the unemployment insurance (UI) benefits on the duration of unemployment are of interest from policy perspectives. The elasticity of labor supply with respect to changes in unemployment insurance is an intertwining result of two endogenous forces – the liquidity effects and the moral hazard effects. Landais (2015) demonstrates a reinterpretation of these forces in terms of the traditional framework of dynamic labor supply, and shows how the moral hazard effects of UI on search efforts can be explained by the Frisch elasticity concept, i.e., responses of search efforts to changes in benefits keeping the marginal utility of wealth constant. He then proposes an empirical strategy using the RKD to identify

the moral hazard effects of UI. Using the data set of the Continuous Wage and Benefit History Project (CWBH – see Moffitt, 1985), Landais estimates the effects of benefit amounts on the duration of unemployment. In this section, we apply our QRKD methods, and aim to discover potential heterogeneity in these causal effects.

In all of the states in the United States, a compensated unemployed individual receives a weekly benefit amount b that is determined as a fraction τ_1 of his or her highest earning quarter x in the base period (the last four completed calendar quarters immediately preceding the start of the claim) up to a fixed maximum amount b_{max}, i.e. $b = \min\{\tau_1 \cdot x,\ b_{max}\}$. The both parameters, τ_1 and b_{max}, of the policy rule vary from state to state. Furthermore, the ceiling level b_{max} changes over time within a state. For these reasons, empirical analysis needs to be conducted for each state for each restricted time period. The potential duration of benefits is determined in a somewhat more complicated manner. Yet, it also can be written as a piecewise linear and kinked function of a fraction of a running variable x in the CWBH data set.

Following Landais (2015), we make our QRKD empirical illustration by using the CWBH data for Louisiana. The data cleaning procedure is conducted in the same manner as in Landais. As a result of the data processing, we obtain the same descriptive statistics (up to deflation) as those in Landais for those variables that we use in our analysis. For the dependent variable y, we consider both the claimed number of weeks of UI and the actually paid number of weeks. For the running variable x, we use the highest quarter wage in the based period. The treatment intensity b is computed by using the formula $b(x) = \min\{(1/25) \cdot x,\ b_{max}\}$, with a kink where the maximum amount is $b_{max} = \$4,575$ for the period between September 1981 and September 1982 and $b_{max} = \$5,125$ for the period between September 1982 and December 1983.

Table 4 summarizes empirical results for the time period between September 1981 and September 1982. Table 5 summarizes empirical results for the time period between September

1982 and December 1983. In each table, we display the RKD results by Landais (2015) for a reference. In the following rows, the QRKD estimates are reported with respective standard errors in parentheses for quantiles $\tau \in \{0.10, \cdots, 0.90\}$. At the bottom of each table, we report the p-values for the test of significance and the test of heterogeneity.

We can observe the following patterns in these result tables. First, the estimated causal effects have positive signs throughout all the quantiles, implying that higher benefit amounts cause longer unemployment durations consistently across all the outcome levels. Second, these causal effects are smaller and insignificant at lower quantiles, while they are larger and significantly different from zero at middle and higher quantiles. This pattern implies that unemployed individuals who have longer unemployment durations tend to have larger unemployment elasticities with respect to benefit levels. This result is consistent with a simple intuitive story; individuals with longer unemployment durations are associated with lower abilities and are therefore more likely to show moral hazard. The extent of this increase of the causal effects in quantiles is more prominent for the results in Table 4 (1981–1982) than in Table 5 (1982–1983). Third, the causal effects are very similar between the results for claimed UI as the outcome and the results for paid UI as the outcome variable. The respective standard errors are almost the same between these two outcome variables, but they are not exactly the same. Fourth, the uniform tests show that the causal effects are significantly different from zero for the both time periods. Lastly, the uniform tests show that the causal effects are significantly heterogeneous in Table 4 (1981–1982), while the heterogeneity is insignificant in Table 5 (1982–1983).

6 Summary

Economists have taken advantage of policy irregularities to assess causal effects of endogenous treatment intensities. A new approach along this line is the regression kink design (RKD) used by recent empirical papers, including Nielsen, Sørensen and Taber (2010), Landais (2015), Simonsen, Skipper and Skipper (2015) and Card, Lee, Pei and Weber (2016). While the prototypical framework is only able to assess the average treatment effect at the kink point, inference for heterogeneous treatment effects using the RKD is of potential interest by empirical researchers (e.g., Landais, 2011). In this light, this paper develops econometric tools for the quantile regression kink design (QRKD).

We first develop causal interpretations of the QRKD estimand. It is shown that the QRKD estimand measures the marginal effect of the treatment variable on the outcome variable at the conditional quantile of the outcome given the design point of the running variable. Second, we propose a sample counterpart QRKD estimator, and develop its asymptotic properties for statistical inference of heterogeneous treatment effects. Using uniform Bahadur representations, we derive a weak consistency result for the QRKD estimator. Applying the weak consistency result, we propose procedures for statistical tests of treatment significance and treatment heterogeneity. We also discuss finite-sample bias reduction and bandwidth selection. Monte Carlo experiments support our theoretical results. Applying our methods to the Continuous Wage and Benefit History Project (CWBH) data, we find significantly heterogeneous causal effects of unemployment insurance benefits on unemployment durations in the state of Louisiana for the period between September 1981 and September 1982. Finally, while the main text mostly focuses on the sharp QRKD that is relevant to our empirical illustration, we remark that identification and estimation results for the fuzzy QRKD are also available in Appendices A.2 and

C for completeness.

References

Angrist, Joshua D. and Guido W. Imbens (1995) "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association,* Vol. 90, No. 430, pp. 431–442.

Bashtannyk, David M., and Rob J. Hyndman (2001) "Bandwidth selection for kernel conditional density estimation," *Computational Statistics and Data Analysis,* Vol. 36, No. 3, pp. 279–298.

Card, David, David Lee, Zhuan Pei, and Andrea Weber (2016) "Inference on Causal Effects in a Generalized Regression Kink Design," *Econometrica,* Vol. 83, No. 6, pp. 2453–2483.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato (2014) "Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-Dimensional Random Vectors," *The Annals of Statistics,* Vol. 41, No. 6, pp. 2786–2819.

Chernozhukov, Victor and Ivan Fernández-Val (2005) "Subsampling Inference on Quantile Regression Processes," *Sankhya: The Indian Journal of Statistics,* Vol. 67, No. 2, pp. 253–276.

Einmahl, Uwe and David M. Mason (2005) "Uniform in Bandwidth Consistency of Kernel-Type Function Estimators," *Annals of Statistics,* Vol. 33, No. 3, pp. 1380–1403.

Fan, Jianqing and Irène Gijbels (1996) "Local Polynomial Modeling and Its Applications," Chapman & Hall/CRC: London.

Fan, Jianqing, Tien-Chung Hu, and Young K. Truong (1994) "Robust Non-Parametric Function Estimation," *Scandinavian Journal of Statistics,* Vol. 21, No. 4, pp. 433–446.

Frandsen, Brigham R., Markus Frölich and Blaise Melly (2012) "Quantile Treatment Effects in the Regression Discontinuity Design," *Journal of Econometrics,* Vol. 168, No.2 pp. 382-395.

Giné, Evarist and Richard Nickl (2015) "Mathematical Foundations of Infinite-Dimensional Statistical Models," Cambridge University Press: Cambridge.

Guerre, Emmanuel and Camille Sabbah (2012) "Uniform Bias Study and Bahadur Representation for Local Polynomial Estimators of the Conditional Quantile Function," *Econometric Thoery,* Vol. 26, No. 5, pp. 1529-1564.

Guerre, Emmanuel and Camille Sabbah (2014) "Uniform Confidence Bands for Local Polynomial Quantile Estimators," *ESAIM: Probability and Statistics,* Vol. 18, pp. 265-276.

Imbens, Guido and Thomas Lemieux (2008) "Special Issue Editors' Introduction: The Regression Discontinuity Design – Theory and Applications," *Journal of Econometrics,* Vol. 142, No. 2, pp. 611–614.

Koenker, Roger (2005) "Quantile Regression," Cambridge University Press: Cambridge.

Koenker, Roger and Zhijie Xiao (2002) "Inference on the quantile regression process," *Econometrica,* Vol. 70, No. 4, pp.1583–1612.

Kong, Efang, Oliver B. Linton, and Yingcun Xia (2010) "Uniform Bahadur Representation for Local Polynomial Estimates of M-Regression and its Application to the Additive Model?," *Econometric Thoery,* Vol. 26, No. 5, pp. 1529-1564.

Landais, Camille (2011) "Heterogeneity and Behavioral Responses to Unemployment Benefits over the Business Cycle," Working Paper, LSE.

Landais, Camille (2015) "Assessing the Welfare Effects of Unemployment Benefits Using the Regression Kink Design," *American Economic Journal: Economic Policy*, Vol. 7, No. 4, pp. 243–278.

Lee, David S., and Thomas Lemieux (2010) "Regression Discontinuity Designs in Economics," *Journal of Economic Literature,* Vol. 48, No. 2, pp. 281–355.

Moffitt, Robert (1985) "The Effect of the Duration of Unemployment Benefits on Work Incentives: An Analysis of Four Datasets," Unemployment Insurance Occasional Papers 85-4, U.S. Department of Labor, Employment and Training Administration.

Nielsen, Helena Skyt, Torben Sørensen, and Christopher Taber (2010) "Estimating the Effect of Student Aid on College Enrollment: Evidence from a Government Grant Policy Reform," *American Economic Journal: Economic Policy,* Vol. 2, No. 2, pp. 185–215.

Pagan, Adrian and Aman Ullah (1999) "Nonparametric Econometrics," Cambridge University Press: Cambridge.

Qu, Zhongjun and Jungmo Yoon (2015a) "Nonparametric Estimation and Inference on Conditional Quantile Processes," *Journal of Econometrics,* Vol. 185, No.1 pp. 1-19.

Qu, Zhongjun and Jungmo Yoon (2015b) "Uniform Inference on Quantile Effects under Sharp Regression Discontinuity Designs," *Working Paper,* 2015.

Sasaki, Yuya (2015) "What Do Quantile Regressions Identify for General Structural Functions?," *Econometric Thoery,* Vol. 31, No. 5, pp. 1102-1116.

Silverman, Bernard W. (1986) "Density Estimation for Statistics and Data Analysis," Chapman & Hall/CRC: London.

Simonsen, Marianne, Lars Skipper, and Niels Skipper (2015) "Price sensitivity of demand for prescription drugs: Exploiting a regression kink design," *Journal of Applied Econometrics,* Forthcoming.

van der Vaart, Aad W. (1998) "Asymptotic Statistics," Cambridge University Press: Cambridge.

van der Vaart, Aad W. and Jon A. Wellner (1996) "Weak Convergence and Empirical Processes," Springer-Verlag: New York.

A Causal Interpretation in General Settings

A.1 Causal Interpretation without Rank Invariance

In this appendix, we inherit the basic settings from Section 2 except that the unobserved factors ϵ are now allowed to be M-dimensional, as opposed to be a scalar. As such, we can consider general structural functions g without the rank invariance. Define the lower contour set of ϵ evaluated by $g(b(x), x, \cdot)$ below a given level of y as follows:

$$V(y, x) = \{\epsilon \in \mathbb{R}^M | g(b(x), x, \epsilon) \leq y\}.$$

Its boundary is denoted by $\partial V(y, x)$. Furthermore, the velocities of the boundary $\partial V(y, x)$ at u with respect to a change in y and a change in x are denoted by $\partial \upsilon(y, x; u)/\partial y$ and $\partial \upsilon(y, x; u)/\partial x$, respectively. Σ denotes an $(M-1)$-dimensional rectangle. For a short hand notation, we write $h(x, \epsilon) = g(b(x), x, \epsilon)$ and $h_1(x, \epsilon) = \frac{\partial h(x,\epsilon)}{\partial x}$. Let m^M and H^{M-1} denote the Lebesgue measure on $\mathbb{R}^M$ and the Hausdorff measure on $\partial V(y, x)$, respectively.[1] Letting $\mathcal{X} = supp(X)$, we make the following assumptions.

Assumption 3. *(i) $h(\cdot, \epsilon)$ is continuously differentiable on $\mathcal{X} \setminus \{x_0\}$ for all $\epsilon \in \mathcal{E}$ and $h(x, \cdot)$ is continuously differentiable for all $x \in \mathcal{X}$. (ii) $\|\nabla_\epsilon h(x, \cdot)\| \neq 0$ on $\partial V(y, x)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. (iii) The conditional distribution of ε given X is absolutely continuous with respect to m^M, $f_{\varepsilon|X}$ is continuously differentiable, and $f_{\varepsilon|X} \in C^1(\mathcal{X}; L^1(\mathbb{R}^M))$ is true. (iv) $\int_{\partial V(y,x)} f_{\varepsilon|X}(\epsilon \mid x) dH^{M-1}(\epsilon) > 0$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. (v) $\partial V(y, \cdot) \in C^1(\Sigma \times \mathcal{X}; \mathbb{R}^M)$ holds for all $y \in \mathcal{Y}$ and $\partial V(\cdot, x) \in C^1(\Sigma \times \mathcal{Y}; \mathbb{R}^M)$ holds for all $x \in \mathcal{X}$. (vi) $\partial \upsilon(y, \cdot\ ; \cdot)/\partial x \in C^1(\mathcal{X}; L^1(\Sigma))$ holds for all $y \in \mathcal{Y}$ and $\partial \upsilon(\cdot, x\ ; \cdot)/\partial y \in C^1(\mathcal{Y}; L^1(\Sigma))$ holds for all $x \in \mathcal{X}$.*

[1] We obtain the $(M-1)$-dimensional Hausdorff measure by the restriction of the function $H^{M-1} : 2^{\mathbb{R}^M} \to \mathbb{R}$ defined by $H^{M-1}(S) = \sup_{\delta>0} \inf \left\{ \sum_{i=1}^{\infty} (\text{diam} U_i)^{M-1} \mid \cup_{i=1}^{\infty} S_i \supset S, \text{ diam} S_i < \delta \right\}$ to the collection of Carathéodory-measurable sets.

Assumption 4. *Let* $\gamma(x,\epsilon) := \|\nabla_\epsilon h(x,\epsilon)\|^{-1}$. *There exist values* $p \geqslant 1$ *and* $q \geqslant 1$ *satisfying* $p^{-1} + q^{-1} = 1$ *such that* $\|\gamma(x,\ \cdot\ )\|_{L^p(\partial V(y,x),H^{M-1})} < \infty$ *and* $\|f_\varepsilon\|_{L^q(\partial V(y,x),H^{M-1})} < \infty$ *hold for all* $(x,y) \in \mathcal{X} \times \mathcal{Y}$.

Assumption 5. *There exists a function* $w \in L^1(\partial V(y,x), H^{M-1})$ *such that* $|\gamma(x,\epsilon)h_x(x,\epsilon)f_{\varepsilon|X}(\epsilon|x)| \leq w(\epsilon)$ *for all* $\epsilon \in \times V(y,x)$ *for all* $x \in \mathcal{X}$.

Assumption 6. $lim_{x\to x_0^+}\frac{\partial}{\partial x}Q_{Y|X}(\tau|x)$ *and* $lim_{x\to x_0^-}\frac{\partial}{\partial x}Q_{Y|X}(\tau|x)$ *exist.*

Assumptions 3 and 4 are used to derive a structural decomposition of the quantile partial derivative – see Sasaki (2015) for detailed discussions of these assumptions. The regularity conditions in Assumptions 5 and 6 together facilitate the dominated convergence theorem to make a structural sense of the QRKD estimand (1.1). With $\mathcal{B}(y,x)$ denoting the collection of Borel subsets of $\partial V(y,x)$, we define the function $\mu_{y,x} : \mathcal{B}(y,x) \to \mathbb{R}$ by

$$\mu_{y,x}(S) := \frac{\int_s \frac{1}{\|\nabla_\epsilon h(x,\epsilon)\|} f_{\varepsilon|X}(\epsilon|x) dH^{M-1}(\epsilon)}{\int_{\partial V(y,x)} \frac{1}{\|\nabla_\epsilon h(x,\epsilon)\|} f_{\varepsilon|X}(\epsilon|x) dH^{M-1}(\epsilon)} \qquad \text{for all } S \in \mathcal{B}(y,x).$$

The next theorem claims that this is a probability measure and gives weights with respect to which the QRKD estimand (1.1) measures the average structural causal effect of the treatment intensity b on an outcome y for those individuals at the τ-th conditional quantile of Y given $X = x_0$.

Theorem 2. *Suppose that Assumptions 1, 3, 4, 5 and 6 hold. Then,* $\mu_{y,x}$ *is a probability measure on* $\partial V(y,x)$ *for all* $(x,y) \in \mathcal{X} \times \mathcal{Y}$, *and we have*

$$QRKD(\tau) \ = \ E_{\mu_{y,x_0}}\left[g_1(b(x_0), x_0, \varepsilon)\right] \tag{A.1}$$

where $\tau = F_{Y|X}(y \mid x_0)$.

A proof is provided in Appendix D.7. We may derive a similar causal interpretation for the case of fuzzy QRKD – see Appendix A.2. As is often the case in the treatment literature (e.g.,

Angrist and Imbens, 1995), this theorem shows a causal interpretation in terms of a weighted average. Specifically, (A.1) shows that the QRKD estimand (1.1) measures a weighted average of the heterogeneious causal effects $g_1(b(x_0), x_0, \epsilon)$ displayed on the right-hand side of (A.1). The weights given in the definition of μ_{y,x_0} are proportional to $f_{\varepsilon|X}(\epsilon|x_0)/\|\nabla_\epsilon h(x_0, \epsilon)\|$ which is positive on the conditional support of ε given $X = x_0$. In other words, the QRKD estimand measures a strict convex combination of the ceteris paribus causal effects of b on y for those individuals at the τ-th conditional quantile of Y given $X = x_0$. One may worry about the obscurity of the causal interpretations under the 'weighted' averages. There are two special cases where the QRKD estimand allows for causal interpretations in terms of purely unweighted averages, i.e., $\|\nabla_\epsilon h(x_0, \epsilon)\|$ is constant in ϵ. One example is the case where the structural function $g(b, x, \cdot)$ is monotone in a scalar unobservable ϵ, which is the special case discussed in Section 2. The other example is the more general case where the structure exhibits partial additivity, e.g., $g(b, x, \epsilon) = \sum_{m=1}^{M} \epsilon_m g^m(b, x)$. When an empirical practitioner is reluctant to make either of these assumptions, the QRKD estimand can be still interpreted as a weighted average measurement of the treatment effects among the subpopulation of individuals at the τ-th conditional quantile of Y given $X = x_0$. In either of these cases, heterogeneity in values of the QRKD estimand across quantiles τ can be used as evidence for heterogeneity in treatment effects. Therefore, we can still conduct statistical inference for heterogeneous treatment effects based on the weak convergence results obtained in Section 3 and Appendix B.

A.2 Causal Interpretation of the Fuzzy QRKD

While Appendix A.1 focuses on the case of sharp QRKD, our identification result (Theorem 2) can be extended to the case of fuzzy QRKD in an analogous manner to the corresponding extension in Card, Lee, Pei and Weber (2016). In the current appendix section, we show the

causal interpretation for the fuzzy QRKD.

The fuzzy QRKD estimand reads as

$$\frac{\lim_{x\downarrow x_0}\frac{\partial}{\partial x}Q_{Y|X}(\tau \mid x) - \lim_{x\uparrow x_0}\frac{\partial}{\partial x}Q_{Y|X}(\tau \mid x)}{\lim_{x\downarrow x_0}\frac{\partial}{\partial x}E[B \mid X = x] - \lim_{x\uparrow x_0}\frac{\partial}{\partial x}E[B|X = x]},$$

where B denotes the random variable for the treatment intensity. Unlike the sharp case, it is not deterministically controlled by the running variable. With the M-dimensional unobservables $\epsilon = (\epsilon_1, \epsilon_2)$ decomposed into two parts, we specify the relevant causal structure by

$$\begin{aligned} y &= g(b, x, \epsilon_2) \\ b &= b(x, \epsilon_1) \end{aligned}$$

For short-hand notations, we write $b_1(x, \epsilon_1) = \frac{\partial}{\partial x}b(x, \epsilon_1)$ and $h(x, \epsilon) = g(b(x, \epsilon_1), x, \epsilon_2)$. With these notations under the above setup, we make the following assumption.

Assumption 7. *(a) $b(\,\cdot\,, \epsilon_1)$ is continuous on $\mathcal{X}$ and continuously differentiable on $\mathcal{X} \setminus \{x_0\}$ for all ϵ_1. (b) There exist an absolutely integrable function that envelops $b_1(x,\ \cdot)$ for all x. (c) $E[(b_1(x_0^+, \varepsilon_1) - b_1(x_0^-, \varepsilon_1))|X = x_0]$ and $\int(b_1(x_0^+, \epsilon_1) - b_1(x_0^-, \epsilon_1))d\mu_{y,x_0}(\epsilon)$ exist and are finite and nonzero, where μ_{y,x_0} is defined as in Section A.1.*

Under this assumption, together with the basic assumptions from Section A.1, we obtain the following causal interpretation of the fuzzy QRKD estimand by similar lines of proof to those of Theorem 2.

Theorem 3. *Suppose that Assumptions 3 (with the modified definitions of g, h and ϵ in the current appendix section), 4, 5, 6 and 7 are satisfied. For each $y \in \mathcal{Y}$, we have*

$$\frac{\lim_{x\downarrow x_0}\frac{\partial}{\partial x}Q_{Y|X}(\tau \mid x) - \lim_{x\uparrow x_0}\frac{\partial}{\partial x}Q_{Y|X}(\tau \mid x)}{\lim_{x\downarrow x_0}\frac{\partial}{\partial x}E[B \mid X = x] - \lim_{x\uparrow x_0}\frac{\partial}{\partial x}E[B|X = x]} = E_{\psi_{y,x_0}}[g_1(b(x_0, \varepsilon_1), x_0, \varepsilon_2)]$$

where $\tau = F_{Y|X}(y \mid x_0)$ and $\psi_{y,x_0}(S) = \frac{\int_S(b_1(x_0^+,\epsilon_1)-b_1(x_0^-,\epsilon_1))d\mu_{y,x_0}(\epsilon)}{\int(b_1(x_0^+,\epsilon_1)-b_1(x_0^-,\epsilon_1))d\mu_{y,x_0}(\epsilon)}$ for all $S \in \mathcal{B}(y, x)$.

B Bias Reduction and Bandwidth Selection

While the bias term $h_{n,\tau}\frac{\iota_2'(N^+)^{-1}}{2}\int_0^\infty u^2\frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}(1,u)'K(u)du$ in Theorem 1 is asymptotically negligible, some users may wish to mitigate this finite sample bias by explicitly estimating it. Such a reduction may make a difference especially when the uderlying quantile regressions exhibit large curvatures. The second derivative $\frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}$ can be consistently estimated by the one-sided local quadratic quantile smoother

$$\hat{\lambda}^+(\tau) = \iota_3' \arg\min_{\alpha,\beta,\lambda} \sum_{i=1}^n d_i^+ K(\frac{x_i - x_0}{\bar{h}_{n,\tau}})\rho_\tau(y_i - \alpha - \beta(x_i - x_0) - \lambda(x_i - x_0)^2),$$

where $\bar{h}_{n,\tau}$ denotes a bandwidth parameter and $\iota_3' = (0,0,1)$. The left version of the estimator $\lambda^-(\tau)$ can be similarly defined. To ensure effective bias correction with these estimators, we first obtain a uniform Bahadur representation for the second derivative estimator $\hat{\lambda}^+$ in a similar manner to Lemma 1, and then derive its asymptotic properties in a similar manner to Theorem 1. To this goal, we introduce additional short-hand notations. Some transformations of data points are denoted by $\bar{z}_{i,n,\tau}' = (1, (x_i - x_0)/\bar{h}_{n,\tau}, (x_i - x_0)^2/\bar{h}_{n,\tau}^2))$ and $\bar{K}_{i,n,\tau} = K((x_i - x_0)/\bar{h}_{n,\tau})$. We let $\bar{N}^+$ denote the 3-by-3 matrix with the (i,j)-th element given by $\mu_{i+j-2}^+ = \int_0^\infty u^{i+j-2}K(u)du$. With these notations, we make the following assumption.

Assumption 8.

(i) $\partial Q^3(\tau|x)/\partial x^3$ is finite and Lipschitz continuous on $T \times ([\underline{x}, \bar{x}] \setminus \{x_0\})$.

(ii) $\partial Q^3(\tau|x_0^+)/\partial x^3$ and $\partial Q^3(\tau|x_0^-)/\partial x^3$ are finite and Lipschitz continuous in $\tau \in T$.

(iii) The bandwidths satisfy $\bar{h}_{n,\tau} = c(\tau)\bar{h}_n$, where $n\bar{h}_n^5 \to \infty$ and $\bar{h}_n = o(n^{-1/7})$ as $n \to \infty$, and $c(\cdot)$ is Lipschitz continuous with $0 < \underline{c} \leq c(\tau) \leq \bar{c} < \infty$ for all $\tau \in T$.

(iv) $h_n^3\bar{h}_n^{-5} = o(1)$.

Following a similar argument to the proof of Lemma 3 by Qu and Yoon (2015b) and using our Lemma 5 in Appendix D.1 yield the following Bahadur representation result for $\hat{\lambda}^+$.

Lemma 2. *Under Assumptions 2 and 8 (i)–(iii), we have*

$$\begin{aligned}&\sqrt{n\bar{h}_{n,\tau}^5}\Big(\hat{\lambda}^+(\tau)-\frac{1}{2}\frac{\partial Q^2(\tau|x_0^+)}{\partial x^2}\Big)\\=&\frac{\iota_3'(\bar{N}^+)^{-1}(n\bar{h}_{n,\tau})^{-\frac{1}{2}}\sum_{i=1}^n(\tau-\mathbb{1}\{y_i\le Q(\tau|x_i)\})\bar{z}_{i,n,\tau}d_i^+\bar{K}_{i,n,\tau}}{f_X(x_0)f_{Y|X}(Q(\tau|x_0^+)|x_0^+)}+o_p(1)\end{aligned}$$

uniformly in $\tau\in T$. A similar result holds for $\hat{\lambda}^-(\tau)$.

With a similar reasoning as the proof of Theorem 1, this representation yields the following asymptotic property.

Lemma 3. *Under Assumptions 2 and 8 (i)–(iii), we have the weak convergence*

$$\sqrt{n\bar{h}_{n,\tau}^5 f_X(x_0)}f_{Y|X}(Q(\tau|x_0^+)|x_0^+)\Big(\hat{\lambda}^+(\tau)-\frac{1}{2}\frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}\Big)\Rightarrow\bar{G}^+(\tau),$$

for the zero mean Gaussian process $\bar{G}^+(\tau)$ defined over T with covariance function

$$E(\bar{G}^+(r)\bar{G}^+(s))=(\kappa(r)\kappa(s))^{-1/2}(r\wedge s-rs)\iota_3'(\bar{N}^+)^{-1}\bar{T}^+(r,s)(\bar{N}^+)^{-1}\iota_3,$$

for each $r,s\in T$, where $\bar{T}^+(r,s)=\int_0^\infty\begin{bmatrix}1&\frac{u}{\kappa(r)}&\frac{u^2}{\kappa(r)^2}\end{bmatrix}'K(\frac{u}{\kappa(r)})\begin{bmatrix}1&\frac{u}{\kappa(s)}&\frac{u^2}{\kappa(s)^2}\end{bmatrix}K(\frac{u}{\kappa(s)})du$ with $\kappa(\tau)=\bar{h}_{n,\tau}/\bar{h}_{n,1/2}=\frac{c(\tau)}{c(1/2)}\ge(\underline{c}/\overline{c})>0$. A similar result follows for $\hat{\lambda}^-(\tau)$.

Proof. The proof is almost the same as the proof of Theorem 1. The only difference is that we now work with the Bahadur representation from Lemma 2 instead of Lemma 1. □

With Lemmas 1–3, we can now derive a weak convergence result for the slope estimator $\hat{\beta}^+$ with a uniform bias correction as follows.

Theorem 4. *Under Assumptions 2 and 8, we have*

$$\begin{aligned}&\sqrt{nh_{n,\tau}^3 f_X(x_0)}f_{Y|X}(Q(\tau|x_0^+)|x_0^+)\bigg(\hat{\beta}^+(\tau)-\frac{\partial Q(\tau|x_0^+)}{\partial x}-\hat{\lambda}^+(\tau)h_{n,\tau}\iota_2(N^+)^{-1}\int_0^\infty u^2(1,u)'K(u)du\bigg)\\&\Rightarrow G^+(\tau),\end{aligned}$$

where $G^+(\tau)$ is defined in Theorem 1.

Proof. From Lemmas 1 and 2, we can write

$$\sqrt{nh_{n,\tau}^3 f_X(x_0)} f_{Y|X}(Q(\tau|x_0^+)|x_0^+)\left(\hat{\beta}^+(\tau) - \frac{\partial Q(\tau|x_0^+)}{\partial x} - \hat{\lambda}^+(\tau) h_{n,\tau}\iota_2 (N^+)^{-1}\int_0^\infty u^2(1,u)'K(u)du\right)$$
$$= A_{n,\tau} + \left(\frac{h_{n,\tau}^3}{\bar{h}_{n,\tau}^5}\right)^{1/2} B_{n,\tau} + o_p(1)$$

where

$$A_{n,\tau} = \frac{\iota_2'(N^+)^{-1}(nh_{n,\tau})^{-\frac{1}{2}}\sum_{i=1}^n(\tau - \mathbb{1}\{y_i \le Q(\tau|x_i)\})z_{i,n,\tau}d_i^+ K_{i,n,\tau})}{\sqrt{f_X(x_0)}}$$
$$B_{n,\tau} = -\frac{\iota_3'(\bar{N}^+)^{-1}(n\bar{h}_{n,\tau})^{-\frac{1}{2}}\sum_{i=1}^n(\tau - \mathbb{1}\{y_i \le Q(\tau|x_i)\})\bar{z}_{i,n,\tau}d_i^+ \bar{K}_{i,n,\tau}}{\sqrt{f_X(x_0)}}$$

The term $A_{n,\tau}$ weakly converges to G^+ by Theorem 1. Assumptions 2 and 8 imply that the second term is $o_p(1)$ since $B_{n,\tau}$ is $O_p(1)$ by Lemma 3. Therefore, the desired result follows. □

We can define a version of the QRKD estimator with bias reduction by

$$\widehat{QRKD}_{BR}(\tau) = \frac{(\hat{\beta}^+(\tau) - \hat{\beta}^-(\tau)) - \left(h_{n,\tau}\iota_2(\hat{\lambda}^+(\tau)(N^+)^{-1}R^+ - \hat{\lambda}^-(\tau)(N^-)^{-1}R^-)\right)}{b'(x_0^+) - b'(x_0^-)},$$

where $R^+ = \int_0^\infty u^2\bar{u}K(u)du$ and $R^- = \int_{-\infty}^0 u^2\bar{u}K(u)du$. By similar lines of argument to Corollary 1, we obtain the following weak convergence result for this estimator with bias reduction.

Corollary 3. *Under Assumptions 1, 2, 8, we have the weak convergence*

$$\sqrt{nh_{n,\tau}^3}\left(\widehat{QRKD}_{BR}(\tau) - QRKD(\tau)\right)$$
$$\Rightarrow Y(\tau) = \frac{1}{\sqrt{f_X(x_0)}\left(b'(x_0^+) - b'(x_0^-)\right)}\left[\frac{G^+(\tau)}{f_{Y|X}(Q(\tau|x_0^+)|x_0^+)} - \frac{G^-(\tau)}{f_{Y|X}(Q(\tau|x_0^-)|x_0^-)}\right].$$

Finally, we discuss bandwidth choices. We derive the first-order optimal bandwidths in terms of mean square errors (MSE) in finite sample.

Corollary 4. *Under Assumption 2, the approximate optimal choice of $h_{n,\tau}$ is*

$$h_{n,\tau}^* = \left(\frac{6}{(\iota_2'(N^+)^{-1}D^+)^2}\frac{\tau(1-\tau)\iota_2'(N^+)^{-1}T^+(N^+)^{-1}\iota_2}{f_X(x_0)(f_{Y|X}(Q(\tau|x_0^+)|x_0^+))^2}\right)^{\frac{1}{5}} n^{-\frac{1}{5}},$$

where $D^+ = \int_0^\infty u^2\frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}(1,u)'K(u)du$, $T^+ = \int_0^\infty (1,u)'(1,u)K(u)^2du$.

Corollary 5. *Under Assumptions 2 and 8, the approximate optimal choice of* $\bar{h}_{n,\tau}$ *is*

$$\bar{h}^*_{n,\tau} = \left(\frac{90}{(\iota_3'(\bar{N}^+)^{-1}\bar{D}^+)^2} \frac{\tau(1-\tau)\iota_3'(\bar{N}^+)^{-1}\bar{T}^+(\bar{N}^+)^{-1}\iota_3}{f_X(x_0)(f_{Y|X}(Q(\tau|x_0^+)|x_0^+))^2} \right)^{\frac{1}{7}} n^{-\frac{1}{7}},$$

where $\bar{N} = \int_0^\infty [1,u,u^2]'[1,u,u^2]K(u)du$, $\bar{T}^+ = \int_0^\infty [1,u,u^2]'[1,u,u^2]K(u)^2du$, *and* $\bar{D}^+ = \int_0^\infty u^3 \frac{\partial^3 Q(\tau|x_0^+)}{\partial x^3}[1,u,u^2]'K(u)du$.

Proofs are provided in Appendix D.5 and D.6. Note that these two corollaries prescribing the approximate MSE-optimal bandwidth choices involve unknown densities, f_X and $f_{Y|X}$, as well as the unknown conditional quantile function Q. We suggest to plug-in preliminary estimates, $\hat{f}_X$ $\hat{f}_{Y|X}$ and $\hat{Q}$, where bandwidth choices for these preliminary estimates in turn can be conducted by existing rule-of-thumb or data-driven methods. See Appendix E for a guide to practice in a bandwidth choice procedure.

C Estimation and Asymptotics for the Fuzzy QRKD

The main text focuses on the sharp QRKD. In this appendix, we provide an estimator for the fuzzy QRKD estimand developed in Appedix A.2 and its asymptotic properties. The conditional mean of the policy with errors, $b(X,\varepsilon_1)$, is written as $m(x) = E[b(X,\varepsilon_1)|X=x]$. The regression is also represented by the canonical decomposition $B = m(X) + U$, where the error U satisfies $E[U|X=0] = 0$ and $V(U|X=x) = \sigma^2(x)$. The fuzzy QRKD estimand can then be estimated by

$$\widehat{QRKD}_f(\tau) \;=\; \frac{\hat{\beta}^+(\tau) - \hat{\beta}^-(\tau)}{\hat{m}'(x_0^+) - \hat{m}'(x_0^-)},$$

where the numerator is the same as the one in Section 3. For denominator, we use the local derivative estimator

$$\hat{m}'(x_0^+) = \frac{-\frac{1}{nh_n}\sum_{i=1}^n b_i K'(\frac{x_i - x_0}{h_n})d_i^+ - \hat{m}(x_0)\big(-\frac{1}{nh_n}\sum_{i=1}^n K'(\frac{x_i-x_0}{h_n})d_i^+\big)}{\frac{1}{nh_n}\sum_{i=1}^n K(\frac{x_i-x_0}{h_n})d_i^+}$$

as in Equation (4.14) of Pagan and Ullah (1999). The left counterpart, $\hat{m}'(x_0^-)$, can be defined analogously. Notice that we are using h_n as the bandwidth for $\hat{m}'(x_0^+)$. This is reasonable for the asymptotic argument because, as we will see, $\hat{m}'(x_0^+)$ and $\hat{\beta}^+(\tau)$ have the same rate of convergence. We make the following assumptions.

Assumption 9.

(i) The partial derivatives of f_X, m and $\int b^2 f_{BX}(b, \cdot)db$ exist up to the third order and are bounded.

(ii) There exists a $\delta > 0$ such that $E[|U|^{2+\delta} \mid X = x_0] < \infty$ and $\int |K(u)|^{2+\delta} < \infty$.

(iii) $\|K'\|_\infty < \infty$.

(iv) f_{YXB} exists and is continuous in x for each $(y, b) \in \mathbb{R}^2$. Also, there exists $a > 0$ such that $|m'(x_0^+) - m'(x_0^-)| > a$.

(v) There exists an $\bar{x}$ such that $Q(\tau|\cdot)$ is monotone on $(x_0, \bar{x})$.

Theorem 5. *Under Assumptions 2, 7, and 9, we have*

$$\sqrt{nh_n^3}(\widehat{QRKD}_f(\tau) - QRKD_f(\tau)) \Rightarrow \frac{(m'(x_0^+) - m'(x_0^-))G_\Delta(\tau) - (\beta^+(\tau) - \beta^-(\tau))G_\Delta(b)}{(m'(x_0^+) - m'(x_0^-))^2},$$

where G_Δ is a Gaussian process with mean zero and covariance function as following: for any given r, s, $\tau \in T$ and let b stand for the dimension of $\hat{m}'(x)$, $Cov(G_\Delta(b), G_\Delta(b)) = \sigma_{b,b}^+ + \sigma_{b,b}^-$, $Cov(G_\Delta(\tau), G_\Delta(b)) = \sigma_{\tau,b}^+ + \sigma_{\tau,b}^-$, $Cov(G_\Delta(r), G_\Delta(s)) = \sigma_{r,s}^+ + \sigma_{r,s}^-$, where

$$\begin{aligned}
\sigma_{b,b}^+ =& \frac{\sigma^2(x_0)}{f_X(x_0)} \int K'(v)^2 dv, \\
\sigma_{\tau,b}^+ =& \frac{\iota_2'(N^+)^{-1}}{\sqrt{c(\tau)} f_X(x_0)^2 f_{Y|X}(Q(\tau|x_0^+)|x_0^+)} \int_{\mathbb{R}} \int_{\mathbb{R}} \int_{(0,\infty)} (\tau - \mathbb{1}\{y \le Q(\tau|x_0)\})(1, \frac{v}{c(\tau)})' \\
& \times K(\frac{v}{c(\tau)}) K'(v)(b - E[b(X, \mathcal{E}_1)|X = x_0]) f_{YXB}(y, x_0 + h_n v, b) dv dy db, \\
\sigma_{r,s}^+ =& \frac{c(1/2)}{c(r)c(s) f_X(x_0) f_{Y|X}(Q(r|x_0^+)|x_0) f_{Y|X}(Q(s|x_0^+)|x_0)} (r \wedge s - rs) \iota_2'(N^+)^{-1} T^+(r, s)(N^+)^{-1} \iota_2,
\end{aligned}$$

and the left counterparts are defined analogously.

A proof of this theorem is provided in Appendix D.8

D Mathematical Appendix

D.1 Auxiliary Lemmas: Uniform Consistency

In this appendix section, we develeop the following auxiliary results that show uniform convergences of some useful local sample moments over T. While it is stated for right observations only, we remark that similar results hold for left observations too.

Lemma 4. *Under Assumption 2, we have*

(i) $(nh_{n,\tau})^{-1}\sum_{i=1}^n K_{i,n,\tau}z_{i,n,\tau}z'_{i,n,\tau}d_i^+ \xrightarrow{P} f_X(x_0)N^+$ *uniformly in* $\tau\in T$*;*

(ii) $(nh_{n,\tau})^{-1}\sum_{i=1}^n f_{Y|X}(\tilde{y}_i|x_i)K_{i,n,\tau}z_{i,n,\tau}z'_{i,n,\tau}d_i^+ \xrightarrow{P} f_{Y|X}(Q(\tau|x_0^+)|x_0^+)f_X(x_0)N^+$ *uniformly in* $\tau\in T$ *with any* $\tilde{y}_i$ *lying between* $Q(\tau|x_i)$ *and* $Q(\tau|x_i)+e_i(\tau)+(nh_{n,\tau})^{-1/2}z'_{i,n,\tau}d_i^+\hat{\phi}(\tau)$ *for each* i*;*

(iii) $(nh_{n,\tau}^3)^{-1}\sum_{i=1}^n \frac{1}{2}\left(\frac{x_i-x_0}{h_{n,\tau}}\right)^2\frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}h_{n,\tau}^2 z_{i,n,\tau}K_{i,n,\tau}d_i^+ \xrightarrow{P} \frac{f_X(x_0)}{2}\int_0^\infty u^2\frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}(1,u)'K(u)du$ *uniformly in* $\tau\in T$*.*

Proof. (i): We claim $E\left[(nh_{n,\tau})^{-1}\sum_{i=1}^n K_{i,n,\tau}z_{i,n,\tau}z'_{i,n,\tau}d_i^+\right]\to f_X(x_0)N^+$ uniformly in $\tau\in T$ and $(nh_{n,\tau})^{-1}\sum_{i=1}^n K_{i,n,\tau}z_{i,n,\tau}z'_{i,n,\tau}d_i^+ \xrightarrow{p} E\left[(nh_{n,\tau})^{-1}\sum_{i=1}^n K_{i,n,\tau}z_{i,n,\tau}z'_{i,n,\tau}d_i^+\right]$ uniformly in $\tau\in T$. We provide a proof for only $\iota_2'(nh_{n,\tau})^{-1}\sum_{i=1}^n K_{i,n,\tau}z_{i,n,\tau}z'_{i,n,\tau}\iota_2 = (nh_{n,\tau})^{-1}\sum_{i=1}^n\left(\frac{x_i-x_0}{h_{n,\tau}}\right)^2 K\left(\frac{x_i-x_0}{h_{n,\tau}}\right)$. Similar arguments apply for the other simpler entries.

First note that, by Assumption 2 (i)(a), (i) (b), (iv), and (v),

$$E\iota_2'(nh_{n,\tau})^{-1}\sum_{i=1}^n K_{i,n,\tau}z_{i,n,\tau}z'_{i,n,\tau}d_i^+\iota_2 = \int_0^\infty u^2K(u)\{f_X(x_0)+f'(x_0)uh_{n,\tau}+o(uh_{n,\tau})\}du.$$

Assumption 2 (i) (a) implies that $f'_X(x)$ is bounded in a neighbourhood of x_0. Assumption 2 (iv) and (v) then implies that the right-hand side in the equation above is $f_X(x_0)\iota_2'N^+\iota_2+O(h_{n,\tau})$.

Since $O(h_{n,\tau}) = O(c(\tau)h_n) = O(\bar{c}h_n)$, the convergence is uniform in $\tau \in T$.

In the remainder, we show the uniform convergence of the stochastic part using empirical process theories. For notational convenience, we denote $\sup\{supp(K)\} = \bar{k}$. Let $\mathscr{F} = \{x_i \mapsto \frac{a^2(x_i-x_0)^2}{c(\tau)^3}\mathbb{1}\{K(\frac{a(x_i-x_0)}{c(\tau)}) > 0\}K\big(\frac{a(x_i-x_0)}{c(\tau)}\big)\mathbb{1}\{x_i \geq x_0\} : (\tau, a) \in T \times [0,\infty)\}$. Because each $f \in \mathscr{F}$ is right continuous under (iv) of Assumption 2, this family is a point-wise measurable class – see Einmahl and Mason (2005) and Section 2.3 of van der Vaart and Wellner (1996). For each $(\tau, a) \in \{T \times [0,\infty),\ x_i \mapsto \frac{a^2(x_i-x_0)^2}{c(\tau)^2}\mathbb{1}\{K(\frac{a(x_i-x_0)}{c(\tau)}) > 0\}$ is monotone on its support and bounded by $\bar{k}^2$ under (iv) of Assumption 2. Meanwhile, $c(\tau)$ is finite and bounded away from 0 uniformly in $\tau \in T$ under (v) of Assumption 2, and so is $(1/c(\tau))$. Therefore, $x_i \mapsto \frac{a^2(x_i-x_0)^2}{c(\tau)^3}\mathbb{1}\{K(\frac{a(x_i-x_0)}{c(\tau)}) > 0\}$ is of bounded variation. $x_i \mapsto \mathbb{1}\{x_i \geq x_0\}$ is trivially of bounded variation. Putting them together, we have that each element in $\mathscr{F}$ is of bounded variation with a measurable envelope $F(x_i) = \frac{\bar{k}^2\|K\|_\infty}{\underline{c}}$, where the constant $\underline{c}$ is from (v) of Assumption 2. Since F is a finite constant, $\|F\|_2 = \big(\int |F|^2 f_X(x_i)dx_i\big)^{(1/2)} < \infty$. Without loss of generality as it is bounded, we can assume $F \leq 1$. For a function of bounded variation being the difference of two monotone functions, by Theorem 2.7.5 of van der Vaart and Wellner (1996), there exists a constant $k < \infty$ such that $\log N_{[\,]}(\epsilon \|F\|_2, \mathscr{F}, L_2(P)) \leq \frac{k}{\epsilon\|F\|_2}$ for all $\epsilon > 0$ and for all probability measures P supported on $supp(X)$.

Now, for every finite δ, we have $J(\delta, \mathscr{F}, F) = \int_0^\delta \sup_P \sqrt{1 + \log N_{[\,]}(\epsilon \|F\|_2, \mathscr{F}, L_2(P))}d\epsilon \leq \int_0^\delta \sqrt{1 + \frac{k}{\epsilon\|F\|_2}}d\epsilon < \infty$. Since $F \in L_2(P)$, with any constant $\sigma^2 \in [\sup_{f\in\mathscr{F}} Pf^2, \|F\|_2]$, $\delta = \sigma/\|F\|_2$ and $M = \max_{1\leq i\leq n} F(X_i) < \infty$, we can apply Theorem 5.2 of Chernozhukov, Chetverikov and Kato (2014) to obtain

$$E\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n f(X_i) - \int f dP\right|\right] \leq J(\delta, \mathscr{F}, F)\|F\|_2 + \frac{\|M\|_2 J(\delta, \mathscr{F}, F)^2}{\delta^2\sqrt{n}} < \infty.$$

Multiplying both sides by $(\sqrt{n}h_n)^{-1}$ yields

$$E\Bigg[\sup_{\tau\in T}\Bigg|(nh_{n,\tau})^{-1}\sum_{i=1}^{n}\big(\frac{x_i-x_0}{h_{n,\tau}}\big)^2K\big(\frac{x_i-x_0}{h_{n,\tau}}\big)d_i^+ - E\big[(nh_{n,\tau})^{-1}\sum_{i=1}^{n}\big(\frac{x_i-x_0}{h_{n,\tau}}\big)^2K\big(\frac{x_i-x_0}{h_{n,\tau}}\big)d_i^+\big]\Bigg|\Bigg]$$
$$\leq \frac{1}{\sqrt{n}h_{n,\tau}}\big(J(\delta,\mathscr{F},F)\,\|F\|_2 + \frac{\|M\|_2\,J(\delta,\mathscr{F},F)^2}{\delta^2\sqrt{n}}\big).$$

Thus, the right hand side goes to 0 if $nh_{n,\tau}^2\to\infty$ as $n\to\infty$. Finally, Markov inequality gives the desired result.

(ii): As in the proof of (i), we show $E\left[(nh_{n,\tau})^{-1}\sum_{i=1}^n f_{Y|X}(\tilde{y}_i|x_i)K_{i,n,\tau}z_{i,n,\tau}z'_{i,n,\tau}d_i^+\right]\to f_{Y|X}(Q(\tau|x_0^+)|x_0^+)f_X(x_0)N^+$ uniformly in $\tau\in T$, and then we show $(nh_{n,\tau})^{-1}\sum_{i=1}^n f_{Y|X}(\tilde{y}_i|x_i)K_{i,n,\tau}z_{i,n,\tau}z'_{i,n,\tau}d_i^+\xrightarrow{p}E\left[(nh_{n,\tau})^{-1}\sum_{i=1}^n f_{Y|X}(\tilde{y}_i|x_i)K_{i,n,\tau}z_{i,n,\tau}z'_{i,n,\tau}d_i^+\right]$ uniformly in $\tau\in T$.

First we bound $|Q(\tau|x_i)-(Q(\tau|x_i)+e_i(\tau)+(nh_{n,\tau})^{-1/2}z'_{i,n,\tau}\hat{\phi}(\tau))|\mathbb{1}\{K(\frac{x_i-x_0}{h_{n,\tau}})>0\}=|e_i(\tau)+(nh_{n,\tau})^{-1/2}z'_{i,n,\tau}\hat{\phi}(\tau)|\mathbb{1}\{K(\frac{x_i-x_0}{h_{n,\tau}})>0\}$. Following part 1 of the proof of Theorem 1 in Qu and Yoon (2015a), and using part (i) of our Lemma 4, we have $\hat{\phi}(\tau)\leq\sqrt{\log nh_n}$ with probability approaching one uniformly in $\tau\in T$. Therefore, under Assumption 2 (iv) and (v), we have $(nh_{n,\tau})^{-1/2}z'_{i,n,\tau}\hat{\phi}(\tau)\mathbb{1}\{K(\frac{x_i-x_0}{h_{n,\tau}})>0\}=O_p(\sqrt{\frac{\log nh_n}{nh_n}})$ uniformly in i and $\tau\in T$.

We next bound $e_i(\tau)\mathbb{1}\{K(\frac{x_i-x_0}{h_{n,\tau}})>0\}=\{\big[Q(\tau|x_0^+)+(x_i-x_0)\frac{\partial Q(\tau|x_0^+)}{\partial x}\big]-Q(\tau|x_i)\}\mathbb{1}\{K(\frac{x_i-x_0}{h_{n,\tau}})>0\}$. By the mean value expansion of $Q(\tau|x_i)$ at $x=x_0+\delta$ and let $x\to x_0^+$, we have $e_i(\tau)\mathbb{1}\{K(\frac{x_i-x_0}{h_{n,\tau}})>0\}=[\frac{\partial Q(\tau|x_0^+)}{\partial x}-\frac{\partial Q(\tau|\tilde{x})}{\partial x}](x_i-x_0)\mathbb{1}\{K(\frac{x_i-x_0}{h_{n,\tau}})>0\}\leq M\,\|(\tau,x_0)-(\tau,\tilde{x})\|\,(x_i-x_0)\mathbb{1}\{K(\frac{x_i-x_0}{h_{n,\tau}})>0\}\leq M\,\|(\max T,x_0)-(\min T,\bar{x})\|\,O(h_n)=O(h_n)$ uniformly in $\tau\in T$ for some constant M by Lipschitz continuity and properties of bandwidth from Assumption 2 (iii) (a), (iii) (b) and (iv).

Similarly, we can bound $|Q(\tau|x_0^+)-Q(\tau|x_i)|\mathbb{1}\{K(\frac{x_i-x_0}{h_{n,\tau}})>0\}$. By the joint Lipschitz continuity from Assumption 2 (iii)(a)(b), we have $|Q(\tau|x_0^+)-Q(\tau|x_i)|\mathbb{1}\{K(\frac{x_i-x_0}{h_{n,\tau}})>0\}\leq M\,\|(x_0,\tau)-(x_i,\tau)\|=O(h_n)$ uniformly in $\tau\in T$ for some constant M.

Combing the auxiliary results above, we have $|Q(\tau|x_i)-\tilde{y}_i|\mathbb{1}\{K(\frac{x_i-x_0}{h_{n,\tau}})>0\}\leq|Q(\tau|x_i)-$

$(Q(\tau|x_i)+e_i(\tau)+(nh_{n,\tau})^{-1/2}z'_{i,n,\tau}\hat{\phi}(\tau))|\mathbb{1}\{K(\frac{x_i-x_0}{h_{n,\tau}})>0\} = O_p(\sqrt{\frac{\log nh_n}{nh_n}})+O(h_n)$ and $|Q(\tau|x_0^+)-Q(\tau|x_i)|\mathbb{1}\{K(\frac{x_i-x_0}{h_{n,\tau}})>0\}\leq 1\} = O(h_n)$ uniformly in i and in $\tau\in T$. By the triangle inequality, $|Q(\tau|x_0^+)-\tilde{y}_i|\mathbb{1}\{K(\frac{x_i-x_0}{h_{n,\tau}})>0\}\leq O_p(\sqrt{\frac{\log nh_n}{nh_n}})+O(h_n)$ uniformly in i and $\tau\in T$.

Using Assumption 2 (i) (a), (i) (b), (ii) (a) and (iv) along with the asymptotic bounds obtained above,

$$E\iota_2'(nh_{n,\tau})^{-1}\sum_{i=1}^{n} f_{Y|X}(\tilde{y}_i|x_i)K_{i,n,\tau}z_{i,n,\tau}z'_{i,n,\tau}d_i^+\iota_2$$

$$=f_{Y|X}(Q(\tau|x_0^+)|x_0^+)f_X(x_0)\iota_2'N^+\iota_2+O_p((\log nh_n/nh_n)^{1/2})+O(h_n)$$

holds uniformly in $\tau\in T$. Convergence of the other entries follows similarly.

For the second part, let $\mathscr{F} = \{\ x_i\mapsto \frac{a^2(x_i-x_0)^2 f_{Y|X}(b|x_i)}{c(\tau)^3}\mathbb{1}\{K\big(\frac{a(x_i-x_0)}{c(\tau)}\big)>0\}K\big(\frac{a(x_i-x_0)}{c(\tau)}\big)\mathbb{1}\{x_i\geq x_0\}\ :\ (\tau,a,b)\in T\times[0,\infty)\times[\inf_{(\tau,x)\in T\times(x_0,\bar{x}]}Q(\tau|x),\sup_{(\tau,x)\in T\times(x_0,\bar{x}]}Q(\tau|x)]\ \}$. Notice that the interval of infimum and supremum is bounded by Assumption 2 (iii) (b). An argument similar to the proof of (i) shows that each element in $\mathscr{F}$ is of bounded variation with a measurable envelope $F(x_i) = \frac{\bar{k}^2\|K\|_\infty \sup_{(y,x)} f_{Y|X}}{\underline{c}}$, where the supremum is taken over $(x,y)\in(x_0,\bar{x}]\times[\inf_{(\tau,x)\in T\times(x_0,\bar{x}]}Q(\tau|x),\sup_{(\tau,x)\in T\times(x_0,\bar{x}]}Q(\tau|x)]$, under parts (ii), (iv) and (v) of Assumption 2. Applying the same inequality from Chernozhukov, Chetverikov and Kato (2014) provides a bound for expectation that goes to zero if $nh_n^2\to 0$. Markov inequality then gives the desired result.

(iii): We focus on the entry $\iota_2'(nh_{n,\tau}^3)^{-1}\sum_{i=1}^n \frac{1}{2}\Big(\frac{x_i-x_0}{h_{n,\tau}}\Big)^2\frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}h_{n,\tau}^2 z_{i,n,\tau}K_{i,n,\tau}d_i^+\iota_2$. Similar arguments apply to the other entries. The process is similar to (i). The deterministic part can be shown by computing the expectation. For the stochastic part, let $\mathscr{F} = \{x_i\mapsto \frac{a^3(x_i-x_0)^3}{c(\tau)^4}\frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}\mathbb{1}\{K(\frac{a(x_i-x_0)}{c(\tau)}>0\}K\big(\frac{a(x_i-x_0)}{c(\tau)}\big)\ \mathbb{1}\{x_i\geq x_0\} : (\tau,a)\in T\times[0,\infty)\}$ is a VC type class with a measurable envelope $F(x_i) = \frac{\bar{k}^3 m\|K\|_\infty}{\underline{c}}$ where $m = \sup_T \frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2} < \infty$ under Assumption 2 (iii)(b) and T is compact. Then the same uniform consistency argument applies

under parts (iii) (a) (b), (iv) and (v) of Assumption 2. The same inequality from Chernozhukov, Chetverikov and Kato (2014) and Markov inequality give the desired result. □

Following similar reasoning, we also have the corresponding results for uniform convergences of some local moments for local quadratic regression.

Lemma 5. *Under Assumption 2, 8, we have*

(i) $(n\bar{h}_{n,\tau})^{-1}\sum_{i=1}^{n}\bar{K}_{i,n,\tau}\bar{z}_{i,n,\tau}\bar{z}'_{i,n,\tau}d_i^+ \xrightarrow{P} f_X(x_0)\bar{N}^+$ *uniformly in* $\tau\in T$*;*

(ii) $(n\bar{h}_{n,\tau})^{-1}\sum_{i=1}^{n}f_{Y|X}(\tilde{y}_i|x_i)\bar{K}_{i,n,\tau}\bar{z}_{i,n,\tau}\bar{z}'_{i,n,\tau}d_i^+ \xrightarrow{P} f_{Y|X}(Q(\tau|x_0^+)|x_0^+)f_X(x_0)\bar{N}^+$ *uniformly in* $\tau\in T$ *with any* $\tilde{y}_i$ *lying between* $Q(\tau|x_i)$ *and* $Q(\tau|x_i)+\bar{e}_i(\tau)+(n\bar{h}_{n,\tau})^{-1/2}\bar{z}'_{i,n,\tau}d_i^+\hat{\bar{\phi}}(\tau)$ *for each* i*, where* $\hat{\bar{\phi}}(\tau)=\sqrt{n\bar{h}_{n,\tau}}[\hat{\alpha}^+(\tau)-Q(\tau|x_0^+),\ \bar{h}_{n,\tau}(\hat{\beta}^+(\tau)-\frac{\partial Q(\tau|x_0^+)}{\partial x}),\ \bar{h}_{n,\tau}^2(\hat{\lambda}^+(\tau)-\frac{1}{2}\frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2})]'$ *and* $\bar{e}_i(\tau)=[Q(\tau|x_0^+)+(x_i-x_0)\frac{\partial Q(\tau|x_0^+)}{\partial x}+(x_i-x_0)^2\frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}]-Q(\tau|x_i)$ *;*

(iii) $(n\bar{h}_{n,\tau}^4)^{-1}\sum_{i=1}^{n}\frac{1}{3!}\left(\frac{x_i-x_0}{\bar{h}_{n,\tau}}\right)^3\frac{\partial^3 Q(\tau|x_0^+)}{\partial x^3}\bar{h}_{n,\tau}^3\bar{z}_{i,n,\tau}\bar{K}_{i,n,\tau}d_i^+ \xrightarrow{P} \frac{f_X(x_0)}{3!}\int_0^\infty u^3\frac{\partial^3 Q(\tau|x_0^+)}{\partial x^3}[1,u,u^2]'K(u)du$ *uniformly in* $\tau\in T$*.*

D.2 Proof of Lemma 1

Proof. For this lemma, we mostly follow the proof of Theorem 1in Qu and Yoon (2015a). The major difference is that we focus on the second coordinate of $\hat{\phi}$ instead of the first one. By step 1 of the proof of Theorem 1 in Qu and Yoon (2015a) which is applicable under parts (i), (ii) (a), (ii) (b), (iii) (a), (iii) (b), (iv) and (v) of our Assumption 2, we have $\sup_{\tau\in T}\|\hat{\phi}(\tau)\|\leq(\log nh_n)^{1/2}$ with probability approaching one as $n\to\infty$. Asymptotically, therefore, we only need to focus on studying the behavior of the subgradient

$$(subgradient)=\sum_{i=1}^{n}\{\tau-\mathbb{1}(u_i^0(\tau)\leq e_i(\tau)+(nh_{n,\tau})^{-1/2}z'_{i,n,\tau}\hat{\phi}(\tau))\}z_{i,n,\tau}K_{i,n,\tau}d_i^+$$

on the set $\Phi_n = \{(\tau, \phi(\tau)) : \tau \in T, \|\phi(\tau)\| \leq \log^{1/2}(nh_n)\}$, where $u_i^0(\tau) = y_i - Q(\tau|x_i)$ and $e_i(\tau) = \big[Q(\tau|x_0^+) + (x_i - x_0)\frac{\partial Q(\tau|x_0^+)}{\partial x}\big] - Q(\tau|x_i)$. Denote

$$S_n(\tau, \phi(\tau), e_i(\tau)) = (nh_n)^{-1/2} \sum_{i=1}^{n} \{P((u_i^0(\tau) \leq e_i(\tau) + (nh_{n,\tau})^{-1/2} z'_{i,n,\tau}\phi(\tau))|x_i)$$
$$-\mathbb{1}(u_i^0(\tau) \leq e_i(\tau) + (nh_{n,\tau})^{-1/2} z'_{i,n,\tau}\phi(\tau))\} z_{i,n,\tau} K_{i,n,\tau} d_i^+.$$

Theorem 2.1 of Koenker (2005) and Assumption 2 (iv) imply $(nh_n)^{-1/2} \cdot (subgradient) = O_p((nh_n)^{-1/2})$ uniformly in $\tau \in T$.

Following Qu and Yoon (2015a), we can rewrite the subgradient (scaled by $(nh_n)^{-1/2}$) as

$$(nh_n)^{-1/2} \sum_{i=1}^{n} \{\tau - \mathbb{1}(u_i^0(\tau) \leq e_i(\tau) + (nh_{n,\tau})^{-1/2} z'_{i,n,\tau}\hat{\phi}(\tau))\} z_{i,n,\tau} K_{i,n,\tau} d_i^+$$
$$= \{S_n(\tau, \hat{\phi}(\tau), e_i(\tau)) - S_n(\tau, 0, e_i(\tau))\} + \{S_n(\tau, 0, e_i(\tau)) - S_n(\tau, 0, 0)\} + S_n(\tau, 0, 0)$$
$$+ (nh_n)^{-1/2} \sum_{i=1}^{n} \{\tau - P((u_i^0(\tau) \leq e_i(\tau) + (nh_{n,\tau})^{-1/2} z'_{i,n,\tau}\hat{\phi}(\tau)|x_i)\} z_{i,n,\tau} K_{i,n,\tau} d_i^+$$

The differences inside the first two pairs of curly brackets are of order $o_p(1)$ on the set Φ_n by Lemma B5 of Qu and Yoon (2015a), which is applicable under parts (i), (ii) (a), (ii) (b), (iii) (a), (iii) (b), (iv) and (v) of our Assumption 2. The $S_n(\tau, 0, 0)$ term is $O_p(1)$ under Assumption 2 (i) (a), (iv), (v). The conditional probability in the last term is a conditional CDF of $Y|X$. Applying the first order mean value expansion to the last term at $y = Q(\tau|x_i)$ yields

$$(nh_n)^{-1/2} \sum_{i=1}^{n} \{\tau - P((u_i^0(\tau) \leq e_i(\tau) + (nh_{n,\tau})^{-1/2} z'_{i,n,\tau}\hat{\phi}(\tau)|x_i)\} z_{i,n,\tau} K_{i,n,\tau} d_i^+$$
$$= -(nh_n)^{-1/2} \sum_{i=1}^{n} f_{Y|X}(\tilde{y}_i|x_i) e_i(\tau) z_{i,n,\tau} K_{i,n,\tau} d_i^+$$
$$- (nh_n)^{-1/2}(nh_{n,\tau})^{-1/2} \Big(\sum_{i=1}^{n} f_{Y|X}(\tilde{y}_i|x_i) K_{i,n,\tau} z_{i,n,\tau} z'_{i,n,\tau} d_i^+ \Big) \hat{\phi}(\tau),$$

where $\tilde{y}_i$ lies between $Q(\tau|x_i)$ and $Q(\tau|x_i) + e_i(\tau) + (nh_{n,\tau})^{-1/2} z'_{i,n,\tau} d_i^+ \hat{\phi}(\tau)$.

Taking the above auxiliary results together, we can now rewrite subgradient (scaled by

$(nh_n)^{-1/2}$) as

$$S_n(\tau,0,0) - (nh_n)^{-1/2}\sum_{i=1}^{n} f_{Y|X}(\tilde{y}_i|x_i)e_i(\tau)z_{i,n,\tau}K_{i,n,\tau}d_i^+$$
$$- (nh_n)^{-1/2}(nh_{n,\tau})^{-1/2}\big(\sum_{i=1}^{n} f_{Y|X}(\tilde{y}_i|x_i)K_{i,n,\tau}z_{i,n,\tau}z'_{i,n,\tau}d_i^+\big)\hat{\phi}(\tau).$$

Recall that this subgradient (scaled by $(nh_n)^{-1/2}$) is $o_p(1)$ uniformly in $\tau \in T$.

By Lemma 4 (ii), $(nh_{n,\tau})^{-1}\sum_{i=1}^{n} f_{Y|X}(\tilde{y}_i|x_i)K_{i,n,\tau}z_{i,n,\tau}z'_{i,n,\tau}d_i^+ \xrightarrow{P} f_{Y|X}(Q(\tau|x_0^+)|x_0^+)f_X(x_0)N^+$ uniformly in $\tau \in T$ and so

$$S_n(\tau,0,0) - (nh_n)^{-1/2}\sum_{i=1}^{n} f_{Y|X}(\tilde{y}_i|x_i)e_i(\tau)z_{i,n,\tau}K_{i,n,\tau}d_i^+$$
$$=(\frac{h_{n,\tau}}{h_n})^{1/2}\big[f_{Y|X}(Q(\tau|x_0^+)|x_0^+)f_X(x_0)N^+ + o_p(1)\big]\hat{\phi}(\tau) + o_p(1)$$

uniformly in $\tau \in T$. Since N^+ is positive definite and $f_{Y|X}(Q(\tau|x_0^+)|x_0^+)f_X(x_0) > 0$ by parts (i), (ii) (b) and (iv) of Assumption 2, we obtain

$$\hat{\phi}(\tau) =\big(f_X(x_0)f_{Y|X}(Q(\tau|x_0^+)|x_0^+)N^+ + o_p(1)\ \big)^{-1}\times$$
$$\left[\left(\frac{h_n}{h_{n,\tau}}\right)^{1/2} S_n(\tau,0,0) - (nh_{n,\tau})^{-1/2}f_{Y|X}(Q(\tau|x_0^+)|x_0^+)\sum_{i=1}^{n} e_i(\tau)z_{i,n,\tau}K_{i,n,\tau}d_i^+ + o_p(1)\right] \tag{D.1}$$

uniformly in $\tau \in T$.

Under Assumption 2 (iii) (a), (iii) (b) the Taylor expansion and $e_i(\tau) = [Q(\tau|x_0^+) + (x_i - x_0)\frac{\partial Q(\tau|x_0^+)}{\partial x}] - Q(\tau|x_i)$ suggest that for any $x_i \geq x_0$ such that $(x_i - x_0)/h_{n,\tau} \in supp(K)$, we have

$$e_i(\tau) = -\frac{1}{2}\Big(\frac{x_i - x_0}{h_{n,\tau}}\Big)^2\frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}h_{n,\tau}^2 + o(h_{n,\tau}^2)$$

uniformly in $\tau \in T$. By Lemma 4 (iii), we have $-(h_{n,\tau})^{-2}(nh_{n,\tau})^{-1}\sum_{i=1}^{n} e_i(\tau)z_{i,n,\tau}K_{i,n,\tau}d_i^+ \xrightarrow{P}$ $\frac{f_X(x_0)}{2}\int_0^\infty u^2\frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}(1,u)'K(u)du$ uniformly in $\tau \in T$. Substitute this and $S_n(\tau,0,0)$ with its

definition into equation (D.1), we have

$$\begin{aligned}\hat{\phi}(\tau) =& \frac{(N^+)^{-1}(nh_{n,\tau})^{-\frac{1}{2}}\sum_{i=1}^{n}(\tau - \mathbb{1}\{y_i \leq Q(\tau|x_i)\})z_{i,n,\tau}K_{i,n,\tau}d_i^+)}{f_X(x_0)f_{Y|X}(Q(\tau|x_0^+)|x_0^+)} \\ &+ (nh_{n,\tau}^5)^{\frac{1}{2}}\frac{(N^+)^{-1}}{2}\int_0^\infty u^2\frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}(1,u)'K(u)du + o_p(1)\end{aligned}$$

uniformly in $\tau \in T$. □

D.3 Proof of Theorem 1

Proof. By Theorem 18.14 in van der Vaart (1998), it suffices to show the finite dimensional convergence in distribution and the tightness. The tightness follows from the boundedness assumptions of f_X and $f_{Y|X}$, and by Lemma B3 in Qu and Yoon (2015a) which is applicable under our Assumptions 2 (i), (ii) (a), (ii) (b), (iii) (a), (iii) (b), (iv) and (v). Specifically, the denominator is bounded away from zero by Assumption 2 (i), and the numerator is tight by their lemma.

For finite dimensional convergence in distribution, We introduce a couple of additional short-hand notations. For each $\tau \in T$, let

$$Z_{n,i}^+(\tau) = \frac{1}{\sqrt{n}}\frac{\iota_2'(N^+)^{-1}(\tau - \mathbb{1}\{y_i \leq Q(\tau|x_i)\})z_{i,n,\tau}d_i^+K_{i,n,\tau}}{\sqrt{f_X(x_0)h_{n,\tau}}}.$$

For any finite set $\{\tau_1, ..., \tau_k\} \subset T$ of quantiles, we write $W_{n,i}^+(\tau_1, ..., \tau_k) = (Z_{n,i}^+(\tau_1), ..., Z_{n,i}^+(\tau_k))'$.

Note that

$$\begin{aligned}E\left(\frac{\iota_2'(N^+)^{-1}(r - \mathbb{1}\{y_i \leq Q(r|x_i)\})z_{i,n,r}d_i^+K_{i,n,r}}{\sqrt{f_X(x_0)h_{n,r}}}\right) =& E\left(\frac{\iota_2'(N^+)^{-1}E[(r - \mathbb{1}\{y_i \leq Q(r|x_i)\})|X]z_{i,n,r}d_i^+K_{i,n,r}}{\sqrt{f_X(x_0)h_{n,r}}}\right) \\ =& E\left(\frac{\iota_2'(N^+)^{-1}(r-r))z_{i,n,r}d_i^+K_{i,n,r}}{\sqrt{f_X(x_0)h_{n,r}}}\right) = 0\end{aligned}$$

holds for each $n \in \mathbb{N}$ and $r \in T$. Since it is n-invariant, let $\sum_{i=1}^n CovW_{n,i}^+(\tau_1, ..., \tau_k) = \Sigma_{\{\tau_1,...,\tau_k\}}$. The entry of the covariance matrix $\Sigma_{\{\tau_1,...,\tau_k\}}$ corresponding to the pair $r, s \in T$ of quantiles is

given by

$$E\left(\frac{\iota_2'(N^+)^{-1}(r-\mathbb{1}\{y_i\leq Q(r|x_i)\})z_{i,n,r}d_i^+K_{i,n,r}}{\sqrt{f_X(x_0)h_{n,r}}}\right)\left(\frac{\iota_2'(N^+)^{-1}(s-\mathbb{1}\{y_i\leq Q(s|x_i)\})z_{i,n,s}d_i^+K_{i,n,s}}{\sqrt{f_X(x_0)h_{n,s}}}\right)'$$
$$=E\frac{\iota_2'(N^+)^{-1}z_{i,n,r}K_{i,n,s}K_{i,n,s}z_{i,n,s}'(N^+)^{-1}\iota_2 d_i^+(r\wedge s-rs)}{f_X(x_0)\sqrt{h_{n,r}h_{n,s}}}$$
$$=(\kappa(r)\kappa(s))^{-1/2}(r\wedge s-rs)\iota_2'(N^+)^{-1}\int_0^\infty\begin{bmatrix}1\\ \frac{u}{\kappa(r)}\end{bmatrix}K(\frac{u}{\kappa(r)})\begin{bmatrix}1 & \frac{u}{\kappa(s)}\end{bmatrix}K(\frac{u}{\kappa(s)})du(N^+)^{-1}\iota_2.$$

We remark that the last line is invariant from $h_{n,\tau}$ because they cancel out through change of variables and by $h_{n,\tau}=c(\tau)h_n$ in Assumption 2 (v). This is finite because Assumption 2 (iv) and (v) imply $|\int_0^\infty u^kK(\frac{u}{\kappa(r)})K(\frac{u}{\kappa(s)})du|\leq\|K\|_\infty\int_0^\infty u^kK(\frac{u}{\kappa(r)})du|<\infty$ for $k=0,1,2$ and all other parts are finite.

Secondly, we show that the moment condition of Lindeberg-Feller is satisfied. Write $(N^+)^{-1}=\left[\begin{smallmatrix}a & b\\ b & c\end{smallmatrix}\right]$, fix any finite set $\{\tau_1,...,\tau_k\}\subset T$ of quantiles. Under Assumptions 2 (i)(a)(b), (iv) and (v), we have for any $\epsilon>0$,

$$\begin{aligned}
&\sum_{i=1}^n E\left\|W_{n,i}^+(\tau_1,...,\tau_k)\right\|^2\mathbb{1}(\left\|W_{n,i}^+\{\tau_1,...,\tau_k)\right\|>\epsilon\}\\
=&\sum_{i=1}^n E\Big[\sum_{j=1}^k Z_{n,i}^+(\tau_j)^2\mathbb{1}\{\sum_{j=1}^k Z_{n,i}^+(\tau_j)^2>\epsilon^2\}\Big]\\
=&\sum_{i=1}^n E\Big[\sum_{j=1}^k\left(\frac{\iota_2'(N^+)^{-1}(r-\mathbb{1}\{y_i\leq Q(\tau_j|x_i)\})z_{i,n,\tau_j}d_i^+K_{i,n,\tau_j}}{\sqrt{f_X(x_0)nh_{n,\tau_j}}}\right)^2\\
&\quad\times\mathbb{1}\{\sum_{j=1}^k\left(\frac{\iota_2'(N^+)^{-1}(r-\mathbb{1}\{y_i\leq Q(\tau_j|x_i)\})z_{i,n,\tau_j}d_i^+K_{i,n,\tau_j}}{\sqrt{f_X(x_0)nh_{n,\tau_j}}}\right)^2>\epsilon^2\}\Big]\\
\leq&E\Big[k\sup_{\tau\in T}\left(\frac{[b+c(\frac{x_i-x_0}{h_{n,\tau}})]^2d_i^+K(\frac{x_i-x_0}{h_{n,\tau}})^2}{f_X(x_0)h_{n,\tau}}\right)\mathbb{1}\{k\sup_{\tau\in T}\left(\frac{[b+c(\frac{x_i-x_0}{h_{n,\tau}})]^2d_i^+K(\frac{x_i-x_0}{h_{n,\tau}})^2}{f_X(x_0)nh_{n,\tau}}\right)>\epsilon^2\}\Big]\\
\leq&\int_0^\infty m\Big([b+c\bar{k}]^2\left\|K\right\|_\infty K(u)\Big)\mathbb{1}\{\frac{[b+c\bar{k}]^2\left\|K\right\|_\infty^2}{nh_n}>\underline{c}\epsilon^2/m\}\big(f_X(x_0)+O(uh_n)\big)du\\
=&\int_0^\infty m_1K(u)\mathbb{1}\{\frac{1}{nh_n}>m_2\epsilon^2\}duf_X(x_0)+O(h_n)
\end{aligned}$$

for some finite non-negative constant m, m_1, m_2 and $\bar{k} = \sup supp(K)$. By applying Dominated Convergence Theorem, the last equation goes to zero since $nh_n \to \infty$ implies the indicator goes to zero and all other terms are finite.

Therefore, by Lindeberg-Feller's Central Limit Theorem, we have

$$\left[\sqrt{nh_{n,\tau}^3 f_X(x_0)} f_{Y|X}(Q(\tau|x_0^+)|x_0^+)\Big(\hat{\beta}^+(\tau) - \frac{\partial Q(\tau|x_0^+)}{\partial x} - h_{n,\tau}\frac{\iota_2'(N^+)^{-1}}{2}\int_0^\infty u^2 \frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}(1,u)'K(u)du\Big)\right]_{\tau\in\{\tau_1,\ldots,\tau_k\}}$$
$$= \left[\sqrt{n}\frac{\iota_2'(N^+)^{-1}(h_{n,\tau})^{-\frac{1}{2}}\sum_{i=1}^n(\tau - \mathbb{1}(y_i \le Q(\tau|x_i)))d_i^+ z_{i,n,\tau}K_{i,n,\tau})}{n\sqrt{f_X(x_0)}}\right]_{\tau\in\{\tau_1,\ldots,\tau_k\}} \xrightarrow{d} N(0, \Sigma_{\{\tau_1,\ldots,\tau_k\}})$$

as $n \to \infty$. □

D.4 Proof of Corollary 2

Proof. The first part of the corollary follows immediately from Corollary 1. The second part follows by an application of the functional delta method (van der Vaart,1998; Theorem 20.8). It suffices to show that the linear functional $\phi : g \mapsto g - \int_T g d\tau$ is Hadamard differentiable at $QRKD$ tangentially to $L_m^\infty(T)$. The linearity of ϕ'_{QRKD} is trivial and the continuity is implied by its boundedness as $\left\|\phi'_{QRKD}(g)\right\| \le \|g\| \, |1 + diam(T)|$ for all $g \in L_m^\infty(T)$. We want to show that for $g_n \to g \in L_m^\infty(T)$ and $t_n \to 0$

$$\frac{\phi(QRKD + t_n g_n) - \phi(QRKD)}{t_n} - \phi'_{QRKD}(g) \to 0 \qquad \text{in } L_m^\infty(T).$$

The left hand side is equal to $g_n - \int_T g_n d\tau - \phi'_{QRKD}(g)$. By the bounded convergence theorem, it converges to 0. □

D.5 Proof of Corollary 4

Proof. By Theorem 1, we have

$$MSE(h) = \frac{h^2}{4}\left(\iota_2'(N^+)^{-1}D^+\right)^2 + \frac{\tau(1-\tau)\iota_2'(N^+)^{-1}T^+(N^+)^{-1}\iota_2}{nh^3 f_X(x_0)(f_{Y|X}(Q(\tau|x_0^+)|x_0^+))^2} + o_p(\frac{1}{nh^3}).$$

By the first order condition with the two leading terms, we obtain the desired result. □

D.6 Proof of Corollary 5

Proof. From Lemma 3, we have

$$MSE(\bar{h}) = \frac{\bar{h}^2}{36}\left(\iota_3'(\bar{N}^+)^{-1}\bar{D}^+\right)^2 + \frac{\tau(1-\tau)\iota_3'(\bar{N}^+)^{-1}\bar{T}^+(\bar{N}^+)^{-1}\iota_3}{n\bar{h}^5 f_X(x_0)(f_{Y|X}(Q(\tau|x_0^+)|x_0^+))^2} + o_p(\frac{1}{n\bar{h}^5}).$$

By the first order condition with the two leading terms, we obtain the desired result. □

D.7 Proof of Theorem 2

Proof. The first result that $\mu_{y,x}$ is a probability measure on $\partial V(y,x)$ follows from Lemma 2 of Sasaki (2015) under Assumption 4. Next, by Lemma 1 of Sasaki (2015) under Assumptions 3 and 4, the QPD $\frac{\partial}{\partial x}Q_{Y|X}(\tau \mid x)$ exists and

$$\begin{aligned}\frac{\partial}{\partial x}Q_{Y|X}(\tau \mid x) &= \frac{\int_{\partial V(y,x)} \frac{h_x(x,\epsilon)}{\|\nabla_\epsilon h(x,\epsilon)\|}\frac{f_{\varepsilon|X}(\epsilon|x)\cdot M\pi^{(M-1)/2}}{2^{M-1}\Gamma(\frac{M+1}{2})}dH^{M-1}(\epsilon) - \int_{V(y,x)}\frac{\partial}{\partial x}f_{\varepsilon|X}(\epsilon \mid x)dm^M(\epsilon)}{\int_{\partial V(y,x)}\frac{1}{\|\nabla_\epsilon h(x,\epsilon)\|}\frac{f_{\varepsilon|X}(\epsilon|x)\cdot M\pi^{(M-1)/2}}{2^{M-1}\Gamma(\frac{M+1}{2})}dH^{M-1}(\epsilon)} \\ &= \mathrm{E}_{\mu_{y,x}}[h_x(x,\varepsilon)] - A(y,x),\end{aligned}$$

where Γ is the Gamma function and A is defined by

$$A(y,x) := \frac{\int_{V(y,x)}\frac{\partial}{\partial x}f_{\varepsilon|X}(\epsilon \mid x)dm^M(\epsilon)}{\int_{\partial V(y,x)}\frac{1}{\|\nabla_\epsilon h(x,\epsilon)\|}\frac{f_{\varepsilon|X}(\epsilon|x)\cdot M\pi^{(M-1)/2}}{2^{M-1}\Gamma(\frac{M+1}{2})}dH^{M-1}(\epsilon)}$$

Note that $g_2 = \frac{\partial g}{\partial x}$ is continuous in x by Assumption 3 (i). Also, $\mu_{y,x}(\epsilon)$ is continuous in x for each fixed y according to parts (i), (ii) and (iii) of Assumption 3. Furthermore, Assumption 3 (i), (ii), (iii) and (iv) imply that $A(y,x)$ is well-defined and is continuous in x for all $y \in \mathcal{Y}$. Therefore, applying the dominated convergence theorem under Assumptions 5 and 6 yields

$$\begin{aligned}\lim_{x\to x_0^+}\frac{\partial}{\partial x}Q_{Y|X}(\tau \mid x) &= \lim_{x\to x_0^+}\int\{h_x(x,\epsilon)\}d\mu_{y,x}(\epsilon) - \lim_{x\to x_0^+}A(y,x) \\ &= \int\lim_{x\to x_0^+}\frac{\partial}{\partial x}\{g(b(x),x,\epsilon)\}d\mu_{y,x}(\epsilon) - A(y,x_0) \\ &= \int\lim_{x\to x_0^+}\{g_1(b(x),x,\epsilon)b'(x) + g_2(b(x),x,\epsilon)\}d\mu_{y,x}(\epsilon) - A(y,x_0) \\ &= \int\{g_1(b(x_0),x_0,\epsilon)b'(x_0^+) + g_2(b(x_0),x_0,\epsilon)\}d\mu_{y,x_0}(\epsilon) - A(y,x_0)\end{aligned}$$

Similarly, taking the limit from the left, we have

$$\lim_{x \to x_0^-} \frac{\partial}{\partial x} Q_{Y|X}(\tau \mid x) = \int \{g_1(b(x_0), x_0, \epsilon) b'(x_0^-) + g_2(b(x_0), x_0, \epsilon)\} d\mu_{y,x_0}(\epsilon) - A(y, x_0).$$

Taking the difference of the right and left limits eliminates $\int g_2(b(x_0), x_0, \epsilon) d\mu_{y,x_0}(\epsilon) - A(y, x_0)$, and thus produces

$$\lim_{x \to x_0^+} \frac{\partial}{\partial x} Q_{Y|X}(\tau \mid x) - \lim_{x \to x_0^-} \frac{\partial}{\partial x} Q_{Y|X}(\tau \mid x) = [b'(x_0^+) - b'(x_0^-)] E_{\mu_{y,x_0}} \left[g_1(b(x_0), x_0, \varepsilon) \right].$$

Finally, note that Assumption 1 has $b'(x_0^+) - b'(x_0^-) \neq 0$, and hence we can divide both sides of the above equality by $b'(x_0^+) - b'(x_0^-)$. This gives the desired result. □

D.8 Proof of Theorem 5

Proof. From the proof of theorem 4.2 of Pagan and Ullah (1999),

$$\sqrt{nh_n^3}(\hat{m}'(x_0^+) - E[\hat{m}'(X)|X = x_0^+]) = \sum_{i=1}^{n} \frac{K'_{n,i} u_i d_i^+}{\sqrt{nh_n} f_X(x_0)} + o_p(1)$$

where $K'_{n,i} = \frac{\partial}{\partial v} K(v)\big|_{v = \frac{x_i - x_0}{h_n}}$. We denote $T_{n,i}^+ = \frac{K'_{n,i} u_i d_i^+}{\sqrt{nh_n} f_X(x_0)}$. We define additional shorthand notations for this proof: let $W_{n,i}^+(\tau_1, ...\tau_k) = (T_{n,i}^+, Z_{n,i}^+(\tau_1), ..., Z_{n,i}^+(\tau_k))$, where $Z_{n,i}^+(\tau)$ is defined as in the proof of Theorem 1. Define

$$H_n^+ = (\sqrt{nh_n^3}(\hat{m}'(x_0^+) - E[\hat{m}'(X)|X = x_0^+]), \{\sqrt{nh_n^3}(\hat{\beta}^+(\tau) - \beta_n^+(\tau)) : \tau \in T\}) = (b_n^+, A_n^+),$$

where $\beta_n^+(\tau) = \frac{\partial Q(\tau|x_0^+)}{\partial x} + h_{n,\tau} \frac{\iota_2'(N^+)^{-1}}{2} \int_0^\infty u^2 \frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}(1, u)' K(u) du$. We also write $A_n^+(\tau)$ for $\sqrt{nh_{n,\tau}^3}(\hat{\beta}^+(\tau) - \beta_n^+(\tau))$.

First we show the finite dimensional convergence of the process. The covariance of any combination of coordinates that does not involve b_n^+ is the same as the one in Theorem 1.

Under Assumption 9 (iv), (v), the covariance of a coordinate of A_n^+ with $\tau \in T$ and b_n^+ is

$$\sum_{i=1}^{n} Cov\left(\frac{\iota_2'(N^+)^{-1} \sum_{i=1}^{n} (\tau - \mathbb{1}\{y_i \leq Q(\tau|x_i)\}) z_{i,n,\tau} K_{i,n,\tau} d_i^+)}{\sqrt{nh_{n,\tau}} f_X(x_0) f_{Y|X}(Q(\tau|x_0^+)|x_0^+)}, \frac{K_{i,n}' u_i}{\sqrt{nh_n} f_X(x_0)} \right)$$

$$\to \sigma_{\tau,b}^+ = \frac{\iota_2'(N^+)^{-1}}{\sqrt{c(\tau)} f_X(x_0)^2 f_{Y|X}(Q(\tau|x_0^+)|x_0^+)} \int_{\mathbb{R}} \int_{\mathbb{R}} \int_{(0,\infty)} (\tau - \mathbb{1}\{y \leq Q(\tau|x_0)\})(1, \frac{v}{c(\tau)})'$$

$$\times K(\frac{v}{c(\tau)}) K'(v)(b - E[b(X, \mathcal{E}_1)|X = x_0]) f_{YXB}(y, x_0, b) dv dy db$$

as $n \to \infty$ by the dominated convergence theorem. This is finite under Assumption 9 (ii). Finally, as in Theorem 4.2 of Pagan and Ullah (1999), the asymptotic variance of b_n^+ is $\frac{\sigma^2(x_0)}{f_X(x_0)} \int K'(v)^2 dv < \infty$. Thus the covariance matrix is finite for any given finite dimensions of A_n^+.

We now show that the moment condition of Lindeberg-Feller is satisfied. For any finite set $\{\tau_1, ..., \tau_k\} \subset T$ of quantiles. Under Assumptions 2 (i) (a), (i) (b), (iv), and (v), and Assumption 9 (iii), we have

$$\sum_{i=1}^{n} E \left\| W_{n,i}^+(\tau_1, ..., \tau_k) \right\|^2 \mathbb{1}(\left\| W_{n,i}^+\{\tau_1, ..., \tau_k) \right\| > \epsilon\}$$

$$= \sum_{i=1}^{n} E\left[(\sum_{j=1}^{k} Z_{n,i}^+(\tau_j)^2 + (T_{n,i}^+)^2) \mathbb{1}\{\sum_{j=1}^{k} Z_{n,i}^+(\tau_j)^2 + (T_{n,i}^+)^2 > \epsilon^2\} \right]$$

$$\leq \int (m_1 K(v) + m_2 K'(v) u^2) \mathbb{1}\{u^2 > nh_n \epsilon^2 + m_3\} dF_{UX}(u, x_0 + vh_n)$$

For some constants, m_1, m_2 and m_3, for any $\epsilon > 0$ given a fixed n. Applying Fubini's theorem under Assumptions 2 (iv) and Assumption 9 (iii), the last line above becomes

$$\int_{(0,\infty)} \int_{\mathbb{R}} (m_1 K(v) + m_2 K'(v) u^2) \mathbb{1}\{u^2 > nh_n \epsilon^2 + m_3\} dF_{U|X}(u|x_0 + vh_n) f_X(x_0 + vh_n) dv$$

We denote the first and second terms of the above expression by (1) and (2), respectively.

Hölder's inequality implies for any $1 < p, q < \infty$ such that $\frac{1}{p} + \frac{1}{q} = 1$,

$$(2) \leq \int_{(0,\infty)} m_2 K(v) E[U^{2p}|X = x_0 + vh_n]^{1/p} P(U^2 > nh_n\epsilon^2 + m_3|X = x_0 + vh_n)^{1/q} f_X(x_0 + vh_n)dv$$

$$\leq ess\ sup_z E[U^{2p}|Z]^{1/p} \left(\int_{(0,\infty)} m_1 K(v) P(U^2 > nh_n\epsilon^2 + m_3|X = x_0 + vh_n)^{1/q} f_X(x_0)dv + O(h_n) \right)$$

Assumption 9 (ii) implies the essential supremum term is finite. Since $P(U^2 > nh_n\epsilon^2 + m_3|X = x_0 + vh_n) \to 0$ as $n \to \infty$, applying the dominated convergence theorem gives us that $(2) \to 0$ as $n \to \infty$. It can be shown that $(1) \to 0$ following a similar reasoning. This shows that the moment condition of Lindeberg-Feller is satisfied. Together with the covariance condition, we have established the finite dimensional convergence of the process H_n^+.

The tightness of all but $\hat{m}'(x_0^+)$ dimensions is shown by the proof of Theorem 1 and $\hat{m}'(x_0^+)$ is trivially tight since it's one dimensional. By Lemma 1.4.3 of van der Vaart and Wellner (1996), H_n^+, the product of them, is tight. Applying theorem 18.14 of van der Vaart (1998), we now have $H_n^+ \Rightarrow G_f^+$, a Gaussian process with zero mean and covariance function as specified above. Using a similar argument, we also have $H_n^- \Rightarrow G_f^-$.

Since H_n^+ and H_n^- are based on an i.i.d. sample from two different sides of the kink, continuous mapping theorem implies

$$\sqrt{nh_n^3} \left(\begin{bmatrix} \hat{m}'(x_0^+) - \hat{m}'(x_0^-) \\ \hat{\beta}^+(\tau) - \hat{\beta}^-(\tau) \end{bmatrix}_{\tau \in T} - \begin{bmatrix} m'(x_0^+) - m'(x_0^-) \\ \beta^+(\tau) - \beta^-(\tau) \end{bmatrix}_{\tau \in T} \right) \Rightarrow G_\Delta = G_f^+ - G_f^+$$

Finally, to derive the asymptotic distribution of $\widehat{QRKD}_f(\tau)$, we apply the uniform version of functional delta method – see theorem 3.9.5 of van der Vaart and Wellner (1996). This version required here because $\beta_n^+(\tau) - \beta_n^-(\tau)$ depends on n. Note $\beta^+(\tau) = \lim_n \beta_n^+(\tau)$ and $\beta^-(\tau) = \lim_n \beta_n^-(\tau)$, the existence is implied by (iii) and (v) of Assumption 2. Define $\Phi : L_m^\infty(T) \times [a, \infty) \to L_m^\infty(T)$, $a > 0$, by $\Phi(A(\tau), b) = \frac{A(\tau)}{b}$. We show that Φ Hadamard differentiable at (A, b) tangentially to $L_m^\infty(T) \times (a, \infty)$. Since for any $(g_n, c_n) \in L_m^\infty(T) \times [a, \infty)$ such that $g_n \to g$

and $c_n \to c$ and any $t_n \to 0$,

$$\frac{\Phi(A + t_n g_n, b + t_n c_n) - \Phi(A, b)}{t_n} \to \Phi'_{(A,b)}(g, c) = \frac{bg - cA}{b^2}.$$

The linearity of $\Phi'_{(A,b)}(g, c)$ is obvious, and its continuity is implied by its boundedness as $\left\|\Phi'_{(A,b)}(g, c)\right\| \leq M \max\{\|g\|_\infty, |c|\} = M \|(g, c)\|$. Since such Hadamard derivative exists at every $(A, b) \in L_m^\infty(T) \times (a, \infty)$, Φ is uniformly differentiable.

By the uniform functional delta method, we have the weak convergence result for the fuzzy QRKD estimator

$$\sqrt{nh_n^3}(\widehat{QRKD}_f(\tau) - QRKD_f(\tau)) \Rightarrow \frac{(m'(x_0^+) - m'(x_0^-))G_\Delta(\tau) - (\beta^+(\tau) - \beta^-(\tau))G_\Delta(b)}{(m'(x_0^+) - m'(x_0^-))^2},$$

as desired. □

E Practical Recipe on Bandwidth Choice

This section provides a guide to practice in bandwidth choices. Corollaries 4 and 5 prescribe the approximate MSE-optimal bandwidth choices. The two corollaries suggest

$$h^*_{n,\tau} = \left(\frac{6}{(\iota_2'(N^+)^{-1}D^+)^2} \frac{\tau(1-\tau)\iota_2'(N^+)^{-1}T^+(N^+)^{-1}\iota_2}{f_X(x_0)(f_{Y|X}(Q(\tau|x_0^+)|x_0^+))^2} \right)^{\frac{1}{5}} n^{-\frac{1}{5}},$$

where $D^+ = \int_0^\infty u^2 \frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}(1, u)' K(u) du$, and

$$\bar{h}^*_{n,\tau} = \left(\frac{90}{(\iota_3'(\bar{N}^+)^{-1}\bar{D}^+)^2} \frac{\tau(1-\tau)\iota_3'(\bar{N}^+)^{-1}\bar{T}^+(\bar{N}^+)^{-1}\iota_3}{f_X(x_0)(f_{Y|X}(Q(\tau|x_0^+)|x_0^+))^2} \right)^{\frac{1}{7}} n^{-\frac{1}{7}},$$

where $\bar{N} = \int_0^\infty [1, u, u^2]'[1, u, u^2] K(u) du$, $\bar{T} = \int_0^\infty [1, u, u^2]'[1, u, u^2] K(u)^2 du$, and $\bar{D}^+ = \int_0^\infty u^3 \frac{\partial^3 Q(\tau|x_0^+)}{\partial x^3}[1, u, u^2]' K(u) du$. Qu and Yoon (2015a) use a very large value for $\bar{h}^*_{n,\tau}$, which is effectively assuming to have $\frac{\partial^3 Q}{\partial x^3} = 0$. Once we compute $\hat{\lambda}^+(\tau)$ based on the choice of $b*_{n,\tau}$, we can in turn substitute $2\hat{\lambda}^+(\tau)$ for $\frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}$ in the definition of D^+ in order to choose $h^*_{n,\tau}$.

In the above formulas, the unknown densities, f_X and $f_{Y|X}$, and the unknown conditional quantile function Q need to be replaced by the respective non-parametric estimates $\hat{f}_X$, $\hat{f}_{Y|X}$ and $\hat{Q}$. Bandwidth choices for the preliminary estimates, $\hat{f}_X$ $\hat{f}_{Y|X}$ and $\hat{Q}$, in turn can be conducted by existing rule-of-thumb or data-driven methods. Because we are only using the observations to right of the kink point, we confine ourselves to the observations $\{(y_j, x_j)\}_{j\in I^+}$ with $I^+ = \{j \in \{1, 2, ..., n\} : x_j \geq x_0\}$. Write $n^+ = |I^+|$, the number of observations to the right of the kink. Write the bandwidths used for estimating $\hat{f}_X$ $\hat{f}_{Y|X}$ and $\hat{Q}$ as $h^x_{n^+}$, $(\bar{h}^y_{n^+}, \bar{h}^x_{n^+})'$ and $h^q_{n^+,\tau}$, respectively.

First, for the standard kernel density estimator, h^x_n may be obtained by minimizing approximate mean integrated square errors: $h^x_n = \big(\int u^2 K(u)du\big)^{-2/5}\big(\int K(u)^2 du\big)^{1/5}\big(\frac{3}{8\sqrt{\pi}}\sigma_X^{-5}\big)^{-1/5} n^{-1/5}$, where σ_X can be estimated by sample variance of X. See session 3.3 and 3.4 of Silverman's (1986). Second, for the standard kernel conditional density estimator, Bashtannyk and Hyndman (2001) suggest that, based on normal approximation of the marginal of X and heteroskedasicity of $Y|X$, $(\bar{h}^y_{n^+}, \bar{h}^x_{n^+})'$ may be obtained by

$$(\bar{h}^y_{n^+}, \bar{h}^x_{n^+})' = \left(\Big(\frac{d^2 v}{2.85\sqrt{2\pi}\sigma^5_{X+}}\Big)^{1/4}\bar{h}^x_{n^+}, \Big(\frac{32R^2(K)\sigma^5_{Y+}(260\pi^9\sigma^{58}_{X+})^{1/8}}{n^+\sigma^4_K d^{5/2}v^{3/4}[v^{1/2} + d(16.25\pi\sigma^{10}_{X+})^{1/4}]}\Big)^{1/6}\right)'$$

where $R(K) = \int K^2(u)du$, $v = 0.95\sqrt{2\pi}\sigma^3_{X+}(3d^2\sigma^2_{X+} + 8\sigma^2_{Y+}) - 32\sigma^2_{X+}\sigma^2_{Y+}e^{-2}$, and d is the slope of an OLS of y_i on $[1, x_i]'$ computed with observations $i \in I^+$. σ^2_{X+} and σ^2_{Y+} can be computed by sample variances of x_i and y_i with $i \in I^+$. σ^2_K is the variance of the kernel K. Third, for the local linear conditional function estimator, the bandwidth $h^q_{n^+,\tau}$ for $\hat{Q}$ can be set by the Yu and Jones' (1998) rule of thumb based on the normality assumption of $f_{Y|X}$:

$$h^q_{n^+,\tau} = \left[2\pi^{-1}\tau(1-\tau)\phi(\Phi^{-1}(\tau))^{-2}\right]^{1/5} h^q_{n^+,1/2},$$

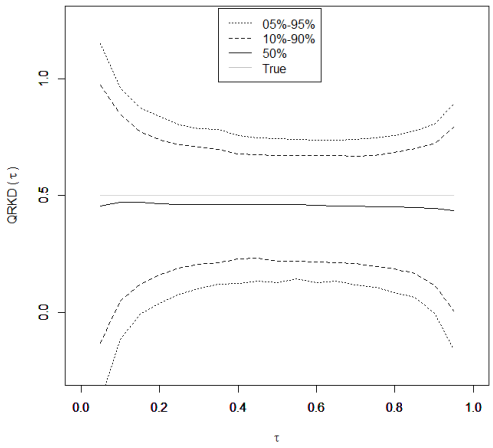
where ϕ and Φ denote the PDF and CDF for the standard normal distribution, respectively, and $h^q_{n^+,1/2}$ can be set to be equal to $h_{n^+,mean} = \left(\frac{\int K(u)^2 du \sigma^2(x)}{n^+(\int u^2 K(u)du)^2\{m''(x)\}^2 f_X(x)}\right)^{1/5}$. The functions,

$m(x)$ and $\sigma^2(x)$, denote the conditional mean and the conditional variance of Y given X. The second-derivative $m''(x)$ can be estimated by the coefficient of the square term of the OLS y_i on $[1, x_i, x_i^2]$ with $i \in I^+$. The skedastic function $\sigma^2(x)$ can be estimated by the sample counterpart of $E[Y^2|X] - (E[Y|X])^2$ that can be computed by using the OLS of y_i^2 on $[1, x_i]$ and y_i on $[1, x_i]$ with $i \in I^+$.

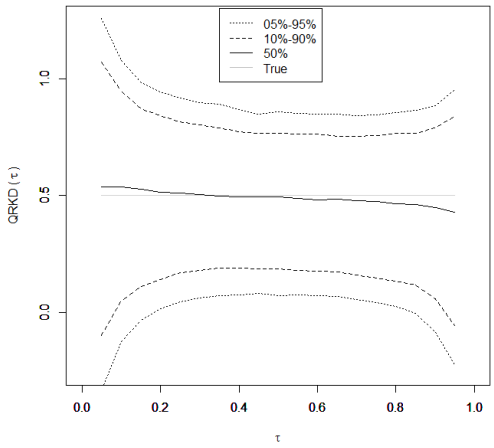
Figures and Tables

Figure 1: Monte Carlo distributions of QRKD estimates under Structure 1.

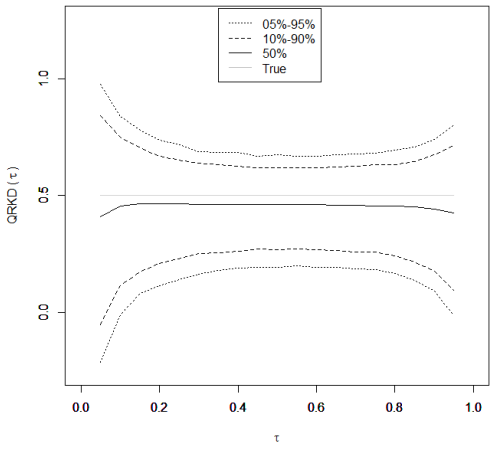
No Bias Reduction; $N = 1,000$



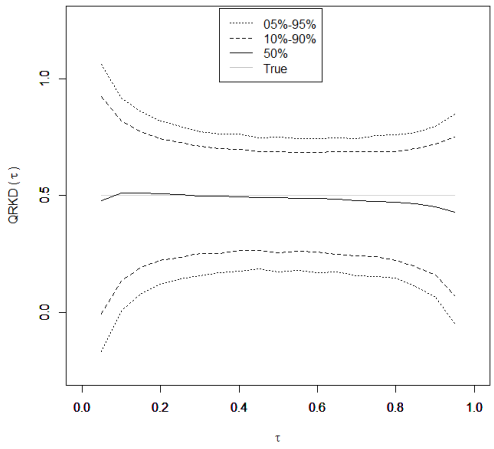
With Bias Reduction; $N = 1,000$



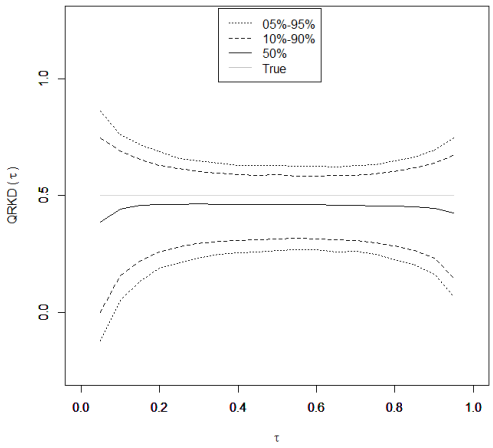
No Bias Reduction; $N = 2,000$



With Bias Reduction; $N = 2,000$



No Bias Reduction; $N = 4,000$



With Bias Reduction; $N = 4,000$

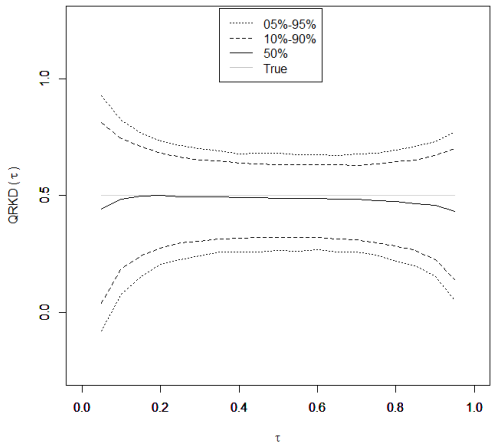
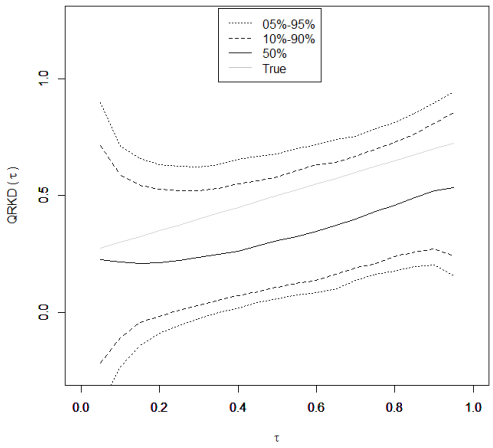
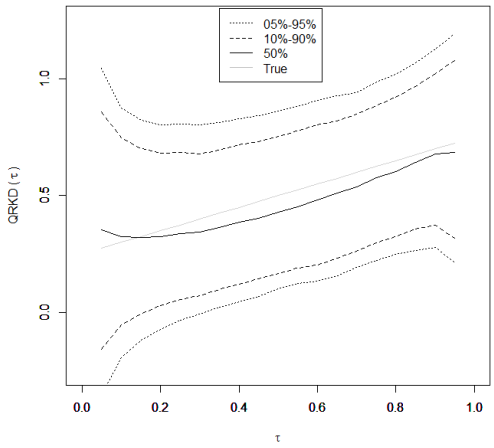


Figure 2: Monte Carlo distributions of QRKD estimates under Structure 2.

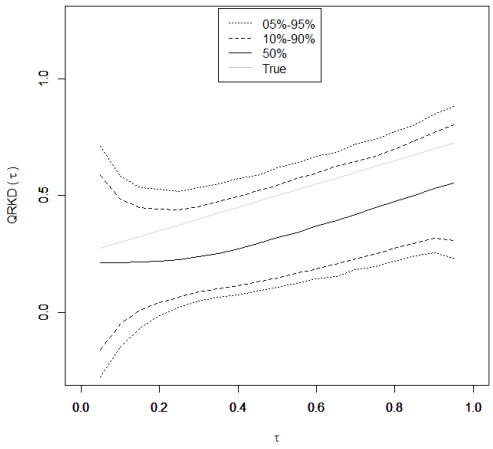
No Bias Reduction; $N = 1,000$



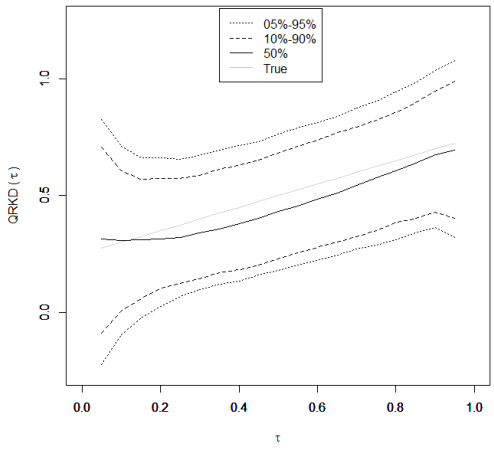
With Bias Reduction; $N = 1,000$



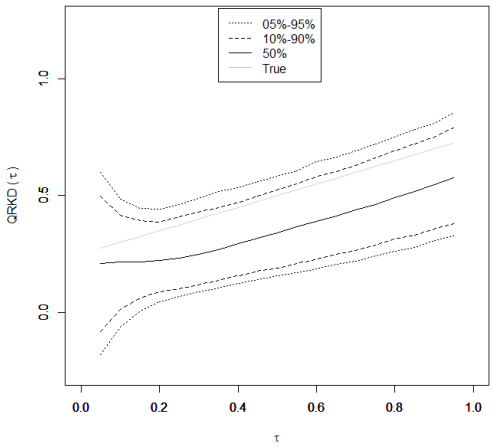
No Bias Reduction; $N = 2,000$



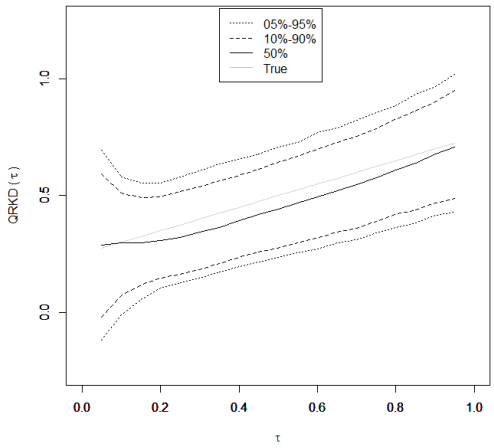
With Bias Reduction; $N = 2,000$



No Bias Reduction; $N = 4,000$



With Bias Reduction; $N = 4,000$



No BR	MC Bias			MC SD			MC RMSE			MC 5% Size		
$N =$	1000	2000	4000	1000	2000	4000	1000	2000	4000	1000	2000	4000
$\tau = 0.10$	0.05	0.06	0.07	0.33	0.26	0.22	0.33	0.26	0.23	0.16	0.17	0.19
$\tau = 0.20$	0.04	0.05	0.05	0.24	0.19	0.15	0.25	0.20	0.16	0.09	0.11	0.12
$\tau = 0.30$	0.04	0.05	0.05	0.21	0.16	0.13	0.21	0.17	0.14	0.07	0.08	0.09
$\tau = 0.40$	0.05	0.05	0.05	0.19	0.15	0.12	0.20	0.16	0.12	0.05	0.07	0.08
$\tau = 0.50$	0.05	0.05	0.04	0.19	0.14	0.11	0.20	0.15	0.12	0.04	0.05	0.06
$\tau = 0.60$	0.05	0.05	0.04	0.19	0.14	0.11	0.19	0.15	0.12	0.03	0.05	0.05
$\tau = 0.70$	0.05	0.05	0.05	0.19	0.15	0.11	0.20	0.16	0.12	0.03	0.04	0.04
$\tau = 0.80$	0.06	0.06	0.05	0.21	0.16	0.13	0.21	0.17	0.14	0.02	0.03	0.03
$\tau = 0.90$	0.07	0.07	0.06	0.25	0.19	0.16	0.26	0.21	0.17	0.02	0.02	0.03
With BR	MC Bias			MC SD			MC RMSE			MC 5% Size		
$N =$	1000	2000	4000	1000	2000	4000	1000	2000	4000	1000	2000	4000
$\tau = 0.10$	0.01	0.01	0.02	0.36	0.28	0.23	0.36	0.28	0.23	0.20	0.19	0.20
$\tau = 0.20$	0.00	0.01	0.01	0.28	0.21	0.16	0.28	0.21	0.16	0.15	0.15	0.14
$\tau = 0.30$	0.01	0.01	0.01	0.25	0.19	0.14	0.25	0.19	0.14	0.12	0.12	0.11
$\tau = 0.40$	0.01	0.01	0.02	0.24	0.18	0.13	0.24	0.18	0.13	0.10	0.10	0.09
$\tau = 0.50$	0.02	0.02	0.02	0.24	0.17	0.13	0.24	0.17	0.13	0.09	0.09	0.08
$\tau = 0.60$	0.02	0.02	0.02	0.24	0.17	0.12	0.24	0.17	0.13	0.08	0.08	0.07
$\tau = 0.70$	0.03	0.03	0.02	0.24	0.18	0.13	0.24	0.18	0.13	0.07	0.07	0.05
$\tau = 0.80$	0.04	0.04	0.03	0.26	0.19	0.14	0.26	0.19	0.15	0.05	0.05	0.05
$\tau = 0.90$	0.07	0.05	0.05	0.30	0.22	0.18	0.31	0.23	0.18	0.04	0.03	0.04

Table 1: Monte Carlo finite-sample statistics of the QRKD estimates under Structure 1.

No BR	MC Bias			MC SD			MC RMSE			MC 5% Size		
$N =$	1000	2000	4000	1000	2000	4000	1000	2000	4000	1000	2000	4000
$\tau = 0.10$	0.07	0.08	0.08	0.29	0.22	0.16	0.30	0.24	0.19	0.12	0.13	0.13
$\tau = 0.20$	0.11	0.12	0.12	0.22	0.17	0.12	0.25	0.20	0.17	0.14	0.18	0.23
$\tau = 0.30$	0.14	0.14	0.14	0.20	0.15	0.12	0.25	0.21	0.18	0.18	0.24	0.33
$\tau = 0.40$	0.16	0.16	0.14	0.20	0.15	0.13	0.25	0.22	0.19	0.19	0.27	0.31
$\tau = 0.50$	0.17	0.17	0.15	0.19	0.16	0.13	0.26	0.23	0.20	0.17	0.24	0.28
$\tau = 0.60$	0.18	0.17	0.15	0.19	0.16	0.14	0.27	0.23	0.21	0.14	0.19	0.25
$\tau = 0.70$	0.18	0.17	0.16	0.19	0.17	0.14	0.27	0.24	0.21	0.10	0.14	0.21
$\tau = 0.80$	0.18	0.17	0.15	0.20	0.17	0.15	0.26	0.24	0.21	0.07	0.09	0.15
$\tau = 0.90$	0.17	0.16	0.15	0.21	0.18	0.16	0.27	0.24	0.21	0.06	0.06	0.09
With BR	MC Bias			MC SD			MC RMSE			MC 5% Size		
$N =$	1000	2000	4000	1000	2000	4000	1000	2000	4000	1000	2000	4000
$\tau = 0.10$	0.03	0.01	0.01	0.33	0.24	0.18	0.33	0.24	0.18	0.17	0.15	0.12
$\tau = 0.20$	0.01	0.02	0.04	0.27	0.19	0.14	0.27	0.19	0.14	0.16	0.14	0.13
$\tau = 0.30$	0.03	0.04	0.05	0.25	0.18	0.14	0.25	0.18	0.15	0.16	0.16	0.18
$\tau = 0.40$	0.05	0.05	0.05	0.24	0.18	0.14	0.24	0.19	0.15	0.15	0.16	0.16
$\tau = 0.50$	0.05	0.05	0.05	0.23	0.18	0.14	0.24	0.19	0.15	0.12	0.12	0.13
$\tau = 0.60$	0.06	0.05	0.05	0.24	0.18	0.15	0.24	0.19	0.16	0.10	0.09	0.11
$\tau = 0.70$	0.05	0.05	0.05	0.23	0.19	0.15	0.24	0.19	0.16	0.07	0.06	0.08
$\tau = 0.80$	0.03	0.04	0.04	0.24	0.19	0.16	0.24	0.19	0.16	0.04	0.04	0.06
$\tau = 0.90$	0.01	0.02	0.02	0.26	0.21	0.17	0.26	0.21	0.17	0.04	0.03	0.03

Table 2: Monte Carlo finite-sample statistics of the QRKD estimates under Structure 2.

(A) Rejection Probabilities for the 95% Level Test of Significance

	No Bias Reduction				With Bias Reduction			
	$1,000$	$2,000$	$3,000$	$4,000$	$1,000$	$2,000$	$3,000$	$4,000$
Structure 0	0.096	0.093	0.100	0.095	0.156	0.157	0.144	0.127
Structure 1	0.271	0.506	0.692	0.794	0.388	0.625	0.770	0.843
Structure 2	0.227	0.545	0.811	0.918	0.536	0.855	0.966	0.993

(B) Rejection Probabilities for the 95% Level Test of Heterogeneity

	No Bias Reduction				With Bias Reduction			
	$1,000$	$2,000$	$3,000$	$4,000$	$1,000$	$2,000$	$3,000$	$4,000$
Structure 0	0.108	0.087	0.089	0.080	0.157	0.145	0.130	0.125
Structure 1	0.094	0.098	0.120	0.125	0.117	0.115	0.136	0.126
Structure 2	0.080	0.088	0.141	0.200	0.124	0.165	0.221	0.297

Table 3: Rejection probabilities for the 95% level uniform test of significance (panel A) and the 95% level uniform test of heterogeneity (panel B) based on 1,000 Monte Carlo replications.

September 1981 – September 1982

Dependent Variable		UI Claimed		UI Paid	
RKD (Landais, 2015)		0.038	(0.009)	0.040	(0.009)
QRKD	$\tau = 0.10$	0.019	(0.030)	0.034	(0.030)
	$\tau = 0.20$	0.036	(0.037)	0.038	(0.038)
	$\tau = 0.30$	0.054	(0.041)	0.065	(0.041)
	$\tau = 0.40$	0.067	(0.044)	0.069	(0.044)
	$\tau = 0.50$	0.081	(0.044)	0.086	(0.044)
	$\tau = 0.60$	0.109	(0.043)	0.105	(0.043)
	$\tau = 0.70$	0.115	(0.041)	0.112	(0.041)
	$\tau = 0.80$	0.161	(0.037)	0.150	(0.037)
	$\tau = 0.90$	0.167	(0.030)	0.191	(0.030)
Test of Significance	p-Value	0.000		0.000	
Test of Heterogeneity	p-Value	0.000		0.004	

Table 4: Empirical estimates and inference for the causal effects of UI benefits on unemployment durations based on the RKD and QRKD. The period of data is from September 1981 to September 1982. The numbers in parentheses indicate standard errors.

September 1982 – December 1983

Dependent Variable		UI Claimed		UI Paid	
RKD (Landais, 2015)		0.046	(0.006)	0.042	(0.006)
QRKD	$\tau = 0.10$	0.023	(0.028)	0.024	(0.028)
	$\tau = 0.20$	0.049	(0.034)	0.053	(0.034)
	$\tau = 0.30$	0.067	(0.038)	0.065	(0.038)
	$\tau = 0.40$	0.086	(0.040)	0.080	(0.040)
	$\tau = 0.50$	0.108	(0.041)	0.107	(0.041)
	$\tau = 0.60$	0.092	(0.040)	0.097	(0.040)
	$\tau = 0.70$	0.111	(0.038)	0.110	(0.038)
	$\tau = 0.80$	0.074	(0.034)	0.082	(0.034)
	$\tau = 0.90$	0.073	(0.027)	0.070	(0.027)
Test of Significance	p-Value	0.026		0.021	
Test of Heterogeneity	p-Value	0.265		0.276	

Table 5: Empirical estimates and inference for the causal effects of UI benefits on unemployment durations based on the RKD and QRKD. The period of data is from September 1982 to December 1983. The numbers in parentheses indicate standard errors.