# Uniform Inference on Quantile Effects under Sharp Regression Discontinuity Designs*

Zhongjun Qu† Boston University

Jungmo Yoon‡ Hanyang University

October 24, 2017

**Abstract**

This study develops methods for conducting uniform inference on quantile treatment effects for sharp regression discontinuity designs. We develop a score test for the treatment significance hypothesis and Wald-type tests for the hypotheses related to treatment significance, homogeneity, and unambiguity. The bias from the nonparametric estimation is studied in detail. In particular, we show that under some conditions, the asymptotic distribution of the score test is unaffected by the bias, without under-smoothing. For situations where the conditions can be restrictive, we incorporate a bias correction into the Wald tests and account for the estimation uncertainty. We also provide a procedure for constructing uniform confidence bands for quantile treatment effects. As an empirical application, we use the proposed methods to study the effect of cash-on-hand on unemployment duration. The results reveal pronounced treatment heterogeneity, and also emphasize the importance of considering the long-term unemployed.

**Keywords**: Heterogeneity, quantile regression, regression discontinuity, treatment effect, unemployment duration.

**JEL classification**: C14, C21.

*First version: July 25, 2014. We thank the seminar participants at Iowa, UC-Irvine, USC, UC-Davis, and the 2013 North American Summer Meeting of the Econometric Society for their useful comments and suggestions, and Youming Liu and Andres Sagner for their research assistance. We thank the co-editor, the associate editor, and two referees for their constructive comments that have substantially improved the paper.

†Department of Economics, Boston University, 270 Bay State Rd., Boston, MA, 02215 (qu@bu.edu).

‡College of Economics and Finance, 222 Wangsimni-Ro, Seongdong-Gu, Seoul, Korea (jmyoon@hanyang.ac.kr).

## 1 Introduction

The regression discontinuity (RD) design (Thistlethwaite and Campbell, 1960) has emerged as an important methodology for identifying and estimating causal effects from observational data. Under this design, the assignment to a treatment depends on whether a covariate exceeds a certain threshold. As such, the treatment effects can be identified and estimated by comparing individuals positioned just above this threshold with those just below. Since the late 1990s, RD designs have been applied in a wide range of fields (see Lee and Lemieux, 2010), including education, labor markets, political economy, health, crime, and environmental issues. Thus far, the majority of studies in the RD literature have focused on the average treatment effect (ATE).

However, treatment effects can be heterogeneous. The average effect, although important, sometimes reveals only a partial picture. For example, it cannot directly measure whether the treatment has altered the dispersion of the outcome distribution. In addition, it is usually silent on whether the effects are stronger in some quantiles than in others. Such distributional aspects can be important and arise naturally, for example, when evaluating effects of unionization on wage inequality (Freeman, 1980, and Card, 1996) and effects of government training programs on lower quantiles of the earning distribution (Lalonde, 1995, and Abadie, Angrist, and Imbens, 2002).

Quantile treatment effects (QTE; Lehmann, 1975, and Doksum, 1974) can be effective for documenting such heterogeneity. Recent studies on QTE outside the RD literature include Heckman, Smith, and Clements (1997), Abadie, Angrist, and Imbens (2002), Chernozhukov and Hansen (2005), and Firpo (2007). Within the RD literature, such studies have remained sparse. The work of Frandsen, Frölich, and Melly (2012) is a notable contribution, where the authors propose an estimator for QTE by inverting conditional distributions of potential outcomes. They consider both sharp and fuzzy designs, and present confidence intervals that are pointwise with respect to the quantile index. In spite of such progress, econometric methods related to quantile effects under the RD design have not been fully developed in relation to the following issues: (1) How can we test whether the treatment is significant at some unknown quantiles within a pre-specified quantile range (Treatment Significance)? (2) How can we test whether the treatment effects are homogeneous across the quantiles (Treatment Homogeneity)? (3) How can we test whether the effects are always nonnegative within the quantile range (Treatment Unambiguity)? (4) How can we construct a uniform confidence band for the quantile effects? These issues all involve a range of quantiles, and addressing them requires methods that are uniform with respect to the quantiles.

This study examines these four issues and provides empirically useful methods. To facilitate

implementation, the theory and empirical application are discussed together, whenever possible. When a theoretical result or procedure is introduced, it is immediately complemented by an empirical discussion to make it practical. In addition, an R package is developed and made available. Researchers can replace the current data file with theirs, and carry out analyses parallel to those described in this paper. In addition, the choices and trade-offs that an empirical researcher may face (e.g., bandwidth choice and bias estimation) are discussed in detail. They are incorporated into the R package as options so that the researcher can experiment and evaluate the sensitivity of the result.

More specifically, this study uses conditional quantile processes to develop methods for conducting uniform inference on QTEs under the sharp RD design. The methods rely on the following results of Qu and Yoon (2015): (i) The quantile processes can be estimated nonparametrically using a family of local linear regressions, with quantile monotonicity maintained through linear inequality constraints or rearrangement (see Chernozhukov, Fernández-Val, and Galichon (2010) for the study of rearrangement as a generic method for estimating monotone probability and quantile curves). (ii) The resulting estimators converge weakly to continuous Gaussian processes, the critical values of which can be obtained from simulations. Building on these results, the analysis in this study proceeds as follows.

First, we develop Score and Wald-type tests for hypotheses related to treatment significance, homogeneity, and unambiguity. The score test treats the cut-off as an interior point, which is useful for testing the treatment significance hypothesis. Because its implementation does not require estimating conditional density, it can be particularly attractive when the sample size is small. The Wald-type tests treat the cut-off value as a boundary point. These tests can be used for all three hypotheses. We derive the limiting distributions of both types of tests and also give procedures for tabulating their critical values.

As is typical with nonparametric regressions, the quantile process estimator is affected by a bias term. Here, we analyze two possibilities and provide two options for addressing this important issue. First, we provide sufficient conditions under which the score test is asymptotically unaffected by the bias term without under-smoothing. Second, for situations where the conditions may not hold, we provide Wald tests that explicitly correct for the biases, while accounting for the estimation uncertainty. The second option is inspired by Calonico, Cattaneo, and Titiunik (2014). However, there is an important difference. In our case, the object of interest is a quantile process, not a finite dimensional parameter. Thus, we exploit the conditional pivotality of the subgradient process, following the insights of Parzen, Wei, and Ying (1994) and Chernozhukov, Hansen, and Jansson

(2009). This approach is applicable to nonparametric inference beyond the RD design setting.

Then, we provide procedures for constructing uniform confidence bands for the quantile effects. This covers situations where the biases have no effect and provides extensions that implement bias correction. The procedure is a direct application of the theory of Qu and Yoon (2015). However, this is the first time such uniform bands have been derived in the RD literature.

The proposed methods are applied to study the effects of cash-on-hand on unemployment duration using the dataset of Card, Chetty, and Weber (2007). This application is introduced in Section 2, and then continued as theoretical results are developed. The data have two potential discontinuities, due to eligibility for severance pay and for extended unemployment insurance benefits. The results show that both discontinuities are statistically significant and that substantial treatment heterogeneity exists in both cases. Interestingly, at the 80th percentile, the two estimated effects on unemployment duration are close to the respective benefit levels. This leads to an intriguing hypothesis that, for the long-term unemployed, the liquidity effect can be a dominating factor in determining the duration of a job search. This hypothesis mirrors the statement by Card, Chetty, and Weber (2007) that a substantial share of behavioral responses to more generous unemployment benefits is attributable to a liquidity effect.

There is a growing body of literature developing econometric methods for RD designs. Hahn, Todd, and Van der Klaauw (2001) provide an influential contribution that develops a framework for identifying and estimating the average treatment effect under weak functional form restrictions. More recent contributions include those on distributional treatment effects (Frölich and Melly, 2010, Frandsen, Frölich, and Melly, 2012, and Shen and Zhang, 2016), alternative estimation procedures (Porter, 2003, Lee, Moretti, and Butler, 2004, and Otsu, Xu, and Matsushita, 2015), graphical methods (Calonico, Cattaneo, and Titiunik, 2015), methods for bandwidths selections (Ludwig and Miller, 2007, Imbens and Kalyanaraman, 2012), robust inference (Calonico, Cattaneo, and Titiunik, 2014), specification analysis (McCrary, 2008, and Lee and Card, 2008), and kink designs (Dong, 2012, Card, Lee, Pei, and Weber, 2015, and Chiang and Sasaki, 2016). A comprehensive survey can be found in Imbens and Lemieux (2008). This study contributes to the literature by developing uniform procedures for documenting treatment heterogeneity for sharp RD designs. Some methods developed here are also informative for fuzzy designs. Section 2 has further discussions.

The remainder of the paper proceeds as follows. Section 2 presents the framework. Section 3 describes two estimation procedures. Section 4 develops a score test for the treatment significance hypothesis. Section 5 develops Wald tests for the three hypotheses related to treatment significance, homogeneity, and unambiguity. Section 6 presents uniform confidence bands. Section 7

shows how various robustness checks can be carried out in practice. Section 8 examines the finite sample properties of the testing procedures. Lastly, Section 9 concludes the paper. All proofs and extensions are included in the online appendix. An R package reproduces the empirical findings, including more details on the implementation.

## 2 Quantile effects under sharp RD designs

Here, we introduce the issues to be studied, and then discuss these in the context of an application.

### 2.1 Issues to be studied

Let $d_i$ denote the treatment status of an individual $i$. Under the sharp RD design, $d_i$ is a deterministic function of some scalar variable $x_i$ (i.e., $d_i = 1\{x_i \geq x_0\}$), where $x_0$ is a known cut-off. Let $F_{Y|X}(\cdot|x)$ denote the cumulative distribution of the outcome variable $Y$, given $X = x$, and let $Q(\tau|x)$ be its conditional quantile at $\tau \in (0,1)$: $Q(\tau|x) = F^{-1}_{Y|X}(\tau|x) = \inf\{s : F_{Y|X}(s|x) \geq \tau\}$. Following Lehmann (1975) and Doksum (1974), the treatment effect at the $\tau$-th percentile can be defined as

$$\delta(\tau) = \lim_{x \downarrow x_0} Q(\tau|x) - \lim_{x \uparrow x_0} Q(\tau|x).$$

To simplify the notation, let $Q(\tau|x_0^+)$ and $Q(\tau|x_0^-)$ denote $\lim_{x\downarrow x_0} Q(\tau|x)$ and $\lim_{x\uparrow x_0} Q(\tau|x)$, respectively. Throughout this paper, the conditional quantile functions of potential outcomes are assumed to be continuous at $x_0$.

We treat $Q(\tau|x)$ as a general nonlinear function of $x$ and $\tau$, and $\delta(\tau)$ is treated as a process indexed by $\tau \in \mathcal{T}$, where $\mathcal{T} = [\lambda_1, \lambda_2]$ with $0 < \lambda_1 \leq \lambda_2 < 1$. In practice, $\mathcal{T}$ can be chosen flexibly, depending on the treatment. For example, if the treatment target is the low part of the distribution, then we can choose $\mathcal{T} = [\varepsilon, 0.5]$, with $\varepsilon$ being a small positive number.

The analysis focuses on estimating and conducting inference on $\delta(\tau)$. The objectives are to develop methods for testing various hypotheses related to $\delta(\tau)$ and to obtain a confidence band that covers this process with a desired probability.

**Treatment significance.** Under the null hypothesis ($H_0$), the treatment has no effect within $\mathcal{T}$. Under the alternative hypothesis ($H_1$), the treatment is significant at some unknown quantiles within $\mathcal{T}$. Formally: $H_0 : \delta(\tau) = 0$ for all $\tau \in \mathcal{T}$, and $H_1 : \delta(\tau) \neq 0$ for some $\tau \in \mathcal{T}$.

**Treatment homogeneity.** The null and alternative hypotheses are given as follows: $H_0 : \delta(\tau)$ is constant over $\tau \in \mathcal{T}$, and $H_1 : \delta(\tau) \neq \delta(s)$ for some $\tau, s \in \mathcal{T}$.

**Treatment unambiguity.** Under the null hypothesis, the effect is unambiguously beneficial at all quantiles within $\mathcal{T}$. Under the alternative hypothesis, it is detrimental at some quantiles. Formally, $H_0 : \delta(\tau) \geq 0$ for all $\tau \in \mathcal{T}$, and $H_1 : \delta(\tau) < 0$ for some $\tau \in \mathcal{T}$.

**Uniform confidence band.** Let $0 < \alpha < 1$ be a coverage level. We want to construct $L_\alpha(\tau)$ and $U_\alpha(\tau)$ such that $\liminf_{n\to\infty} P\left(L_\alpha(\tau) \leq \delta(\tau) \leq U_\alpha(\tau) \text{ for all } \tau \in \mathcal{T}\right) \geq \alpha$.

We now discuss interpretations of $\delta(\tau)$ and the three hypotheses, with and without making the rank invariance assumption. Consider a hypothetical individual, with $X = x_0$. Let $Y^1$ denote the potential outcome if the individual receives the treatment, and $Y^0$ if not. Denote the distributions of $Y^1$ and $Y^0$ by $F_1(\cdot)$ and $F_0(\cdot)$, respectively. Define $\xi(\tau) = F_1^{-1}(\tau) - F_0^{-1}(\tau)$. Then, the rank invariance assumption is equivalent to assuming $F_1(Y^1) = F_0(Y^0)$. In the current framework, if this assumption holds, then $\delta(\tau) = \xi(\tau)$. As a result, the three hypotheses and the confidence band can all be formulated equivalently, with $\xi(\tau)$ replacing $\delta(\tau)$. If this assumption is relaxed, then there are two effects. First, $\xi(\tau)$ will become set-identified (i.e., multiple values are consistent with the two marginal distributions $F_1(\cdot)$ and $F_0(\cdot)$), and $\delta(\tau)$ will represent one value in this set. Second, the three hypotheses will remain informative about $\xi(\tau)$. In particular, if we reject the treatment significance hypothesis, then we also reject the hypothesis of $\xi(\tau) = 0$ for all $\tau \in (0, 1)$ at the same significance level. This is because if the two potential outcome distributions are the same, then the actual outcome distributions of the treated and the untreated must also be identical. Similarly, if we reject $\delta(\tau) = \delta(s)$ for all $\tau, s \in \mathcal{T}$, or reject $\delta(\tau) \geq 0$ for all $\tau \in \mathcal{T}$, then we also reject $\xi(\tau) = \xi(s)$ for all $\tau, s \in (0, 1)$, or $\xi(\tau) \geq 0$ for all $\tau \in (0, 1)$, respectively, at the same significance level. For a more thorough discussion of the rank invariance assumption, along with other assumptions on the identification of heterogeneous treatment effects, see Heckman, Smith, and Clements (1997).

In addition to the above four issues, the methods developed can be used to study the following. First, we can compare the QTEs between subgroups defined by a covariate $z$:

$$\delta^d(\tau) = \left\{ \lim_{x\downarrow x_0} Q(\tau|x, z = z_1) - \lim_{x\uparrow x_0} Q(\tau|x, z = z_1) \right\} - \left\{ \lim_{x\downarrow x_0} Q(\tau|x, z = z_2) - \lim_{x\uparrow x_0} Q(\tau|x, z = z_2) \right\}.$$

For example, $z_1$ and $z_2$ can correspond to males and females, or to two different age groups. Second, we can examine how the QTEs change between two periods, $t_0$ and $t_1$:

$$\delta^d(\tau) = \left\{ \lim_{x\downarrow x_0} Q(\tau|x, t = t_0) - \lim_{x\uparrow x_0} Q(\tau|x, t = t_0) \right\} - \left\{ \lim_{x\downarrow x_0} Q(\tau|x, t = t_1) - \lim_{x\uparrow x_0} Q(\tau|x, t = t_1) \right\}.$$

This can be useful when there is a confounding policy at the cutoff (Grembi, Nannicini, and Troiano, 2016), or if we need to check the robustness of the original results (Landais, 2015). Finally, we can

study the QTEs at two distinct discontinuity points, $x_0$ and $x_1$ (see Angrist and Lavy, 1999, and Van der Klaauw, 2002 for the average effect):

$$\delta^d(\tau) = \left\{ \lim_{x \downarrow x_0} Q(\tau|x) - \lim_{x \uparrow x_0} Q(\tau|x) \right\} - \left\{ \lim_{x \downarrow x_1} Q(\tau|x) - \lim_{x \uparrow x_1} Q(\tau|x) \right\}.$$

The online supplement contains details on how to test hypotheses about $\delta^d(\tau)$ and how to construct a uniform confidence band for it. See Section S.2.

Some methods we develop are also applicable to fuzzy designs. To illustrate this, consider a standard fuzzy design (as in Section 2.3 of Imbens and Lemieux, 2008) that allows for *compliers* (i.e., individuals who receive the treatment if and only if they are eligible), *never takers* (individuals who always refuse the treatment) and *always takers* (individuals who receive the treatment, irrespective of eligibility). Let $Y^1$ and $Y^0$ denote potential outcomes. Let $Q_{Y^1|x_0,c}(\tau)$ and $Q_{Y^0|x_0,c}(\tau)$ be the $\tau$-th quantiles of $Y^1$ and $Y^0$, respectively, conditional on a complier with $X = x_0$. Suppose the object of interest is the QTE for compliers, defined as $\delta_c(\tau) = Q_{Y^1|x_0,c}(\tau) - Q_{Y^0|x_0,c}(\tau)$. Then, as observed in Shen and Zhang (2016), we have the following: (i) If $\delta_c(\tau) = 0$ for all $\tau \in (0,1)$, then $\delta(\tau) = 0$ for all $\tau \in (0,1)$; (ii) If $\delta_c(\tau) \geq 0$ for all $\tau \in (0,1)$, then $\delta(\tau) \geq 0$ for all $\tau \in (0,1)$. The relationships (i) and (ii) imply that, under fuzzy designs, we can still apply our tests for the treatment significance hypothesis or the treatment unambiguity hypothesis, as if the design were sharp. If a rejection occurs, then this implies that we also reject the hypothesis $\delta_c(\tau) = 0$ for all $\tau \in (0,1)$ or $\delta_c(\tau) \geq 0$ for all $\tau \in (0,1)$ at the same significance level. Testing the homogeneity hypothesis about $\delta_c(\tau)$ and constructing a uniform confidence band for it are more challenging. For them, one needs to consider $\delta_c(\tau)$ directly. A procedure that inverts the conditional quantile functions $Q_{Y^1|x_0,c}(\tau)$ and $Q_{Y^0|x_0,c}(\tau)$, and then applies reweighting, in the spirt of Frandsen, Frölich, and Melly (2012), is worth considering. This is left for future research.

We assume that $x_0$ is known. It would be interesting to test whether the QTE is significant at some unknown $\tau$ and $x$. This would require establishing approximations to $\hat{Q}(\tau|x)$ that are uniform with respect to both $\tau$ and $x$. The techniques in Guerre and Sabbah (2012) and Lee, Linton, and Whang (2009) can be valuable for such an analysis.

## 2.2 Application: Cash-on-hand and unemployment duration

Does disposable income ("cash-on-hand") substantially affect the duration of a job search? Card, Chetty, and Weber (2007) study this issue by considering discontinuities generated by severance pay and by extended unemployment benefits in the Austrian labor market. Austrian law requires that workers who are laid off after three years of service at the same company receive a lump sum

amount equal to two months' salary. This implies that an employee's eligibility for severance pay jumps from 0% to 100% as his or her job tenure reaches 36 months. This leads to a sharp RD design. In addition, job seekers with sufficient work history can receive unemployment benefits that vary discontinuously. Those who worked for more than 12, but less than 36 months (at any company) in the past five years are eligible for 20 weeks of unemployment benefits, while those who worked for 36 months or more can receive an extra 10 weeks of benefits. This also leads to a sharp RD design.

In this application, the treatment significance hypothesis asks whether the treatment (severance pay or extended benefits) affects the distribution of the unemployment duration within the quantile range $\mathcal{T}$. Testing this hypothesis can be more informative than testing the ATE if the latter turns out to be small. The homogeneity hypothesis asks whether the treatment shifts the duration by approximately the same magnitude. This is important, because the treatment may affect the long- and short-term unemployed very differently. A test of the treatment unambiguity hypothesis asks whether the effects are always non-negative. Finally, the uniform confidence band quantifies how the effect changes as $\tau$ moves from the lower to the upper tail of the conditional distribution. This band gives detailed information on magnitudes and significance of the effects for all quantiles in $\mathcal{T}$.

We may gain valuable information from the tests that is not obtainable from the ATE. Such information can be useful for informing policy or for discriminating between different models of job seekers' behavior. For example, suppose the test for the treatment unambiguity hypothesis rejects the null hypothesis. Then, this suggests that some job seekers have utilized the increased resources to make their search more efficient, for example, by broadening the search diameter or by paying for childcare in order to gain more time. If this is the case, then follow-up studies can be conducted to identify these efficiency channels to facilitate their usage. As another example, suppose the treatment effect is significant and the treatment homogeneity test does not reject the null hypothesis. Then, this suggests that moral hazard may play an important role in determining the job search. Instead, suppose the treatment homogeneity test rejects the null hypothesis, and that the QTE is found to be strongly increasing with respect to $\tau$. Then, this suggests that an important share of the behavioral responses may be attributable to a liquidity effect. Of course, further studies would be needed to confirm or falsify these tentative conclusions. Meanwhile, if we do not examine distributional effects, we may not gain such information in the first place.

## 3 Estimating quantile treatment effects

This section first estimates $\delta(\tau)$ by applying the procedures of Qu and Yoon (2015) to the RD setting. Then, the estimator is applied to the unemployment duration application.

### 3.1 Estimation procedures

Qu and Yoon (2015) study two estimation procedures in a general nonparametric setting. The procedures share two features. First, they are based on local linear regressions, the suitability of which for RD designs is discussed in Hahn, Todd, and Van der Klaauw (2001) and Porter (2003). Second, they both allow the bandwidth to vary across quantiles. The procedures differ in how the quantile monotonicity is achieved. The first procedure uses linear inequality constraints, while the second applies rearrangement, following Chernozhukov, Fernández-Val, and Galichon (2010). The finite sample properties of the two procedures are shown to be quite similar. Below, we outline how to implement the two procedures in the RD setting. Let $\{(x_i, y_i)\}_{i=1}^n$ denote a sample of $n$ observations. Let $h_{n,\tau}$ be a bandwidth parameter that can differ between quantiles, $K(\cdot)$ be a kernel function, and $\rho_\tau(\cdot)$ be the check function evaluated at the $\tau$-th percentile. Define $d_i = 1\{x_i \geq x_0\}$.

**The first procedure.** STEP 1: Partition $\mathcal{T}$ into an even grid $\{\tau_1, ..., \tau_m\}$ and solve

$$\min_{\{\alpha^+(\tau_j),\beta^+(\tau_j)\}_{j=1}^m} \sum_{j=1}^{m}\sum_{i=1}^{n} \rho_{\tau_j}\left(y_i - \alpha^+(\tau_j) - \beta^+(\tau_j)(x_i - x_0)\right) d_i K\left(\frac{x_i - x_0}{h_{n,\tau_j}}\right), \tag{1}$$

subject to $\alpha^+(\tau_j) \leq \alpha^+(\tau_{j+1})$ for all $j = 1, ..., m-1$. Denote the estimates by $\widetilde{\alpha}^+(\tau_j)$ and $\widetilde{\beta}^+(\tau_j)$ with $j$=1,...,$m$. STEP 2: Compute

$$\begin{aligned} \hat{\alpha}^+(\tau) &= \gamma(\tau)\,\widetilde{\alpha}^+(\tau_j) + (1-\gamma(\tau))\,\widetilde{\alpha}^+(\tau_{j+1}) \\ \hat{\beta}^+(\tau) &= \gamma(\tau)\,\widetilde{\beta}^+(\tau_j) + (1-\gamma(\tau))\,\widetilde{\beta}^+(\tau_{j+1}), \end{aligned} \tag{2}$$

for any $\tau \in [\tau_j, \tau_{j+1}]$, where $\gamma(\tau) = (\tau_{j+1} - \tau)/(\tau_{j+1} - \tau_j)$ and $j$ = 1,..., $m-1$. Set $\hat{Q}(\tau|x_0^+) = \hat{\alpha}^+(\tau)$.

**The second procedure.** STEP 1: Solve (1) without the inequality constraints. Denote the estimates by $\widetilde{\alpha}^+(\tau_1), ..., \widetilde{\alpha}^+(\tau_m)$ and $\widetilde{\beta}^+(\tau_1), ..., \widetilde{\beta}^+(\tau_m)$. STEP 2: Apply (2), and then obtain $\hat{\alpha}^{*+}(\tau) = \inf\left\{y \in \mathbb{R} : \int_{\mathcal{T}} 1\left(\hat{\alpha}^+(u) \leq y\right) du \geq \tau - \lambda_1\right\}$, where $\lambda_1$ is the lower limit of $\mathcal{T}$. Set $\hat{Q}(\tau|x_0^+) = \hat{\alpha}^{*+}(\tau)$.

In the first procedure, the monotonicity constraints involve only the intercepts in the local linear approximation. They are independent of the data (such as their support) and the bandwidths. In the second procedure, the rearrangement is applied after the linear interpolation in order to be

consistent with the theoretical analysis in Chernozhukov, Fernández-Val and, Galichon (2010). As their paper suggests, to facilitate the implementation, it can also be applied directly to $\hat{\alpha}^+(\tau_j)$ $(j = 1, ..., m)$, provided that $m$ is sufficiently large. The linear interpolation can then be applied to the monotonized estimate to obtain the final estimate.

After obtaining $\hat{Q}(\tau|x_0^+)$ using one of the two procedures, we can obtain $\hat{Q}(\tau|x_0^-)$ by replacing $d_i$ in (1) with $1 - d_i$ and "+" with "-". Finally, $\delta(\tau)$ can be estimated as

$$\hat{\delta}(\tau) = \hat{Q}(\tau|x_0^+) - \hat{Q}(\tau|x_0^-) \text{ for any } \tau \in \mathcal{T}. \tag{3}$$

An important step for the estimation is the bandwidth selection. As in Qu and Yoon (2015), we first determine the bandwidth at the median, and then relate it to other quantiles using the link function of Yu and Jones (1998): $h_{n,\tau} = \left\{2\tau(1-\tau)/[\pi\phi(\Phi^{-1}(\tau))^2]\right\}^{1/5} h_{n,0.5}$, where $\phi(\cdot)$ and $\Phi^{-1}(\cdot)$ are the density and quantile functions, respectively, of a standard normal distribution. We consider three selectors for the bandwidth at the median. The first selector uses leave-one-out cross validation. The second uses the minimum MSE bandwidth formula of Qu and Yoon (2015). The third is an adaptation of Imbens and Kalyanaraman's (2012) selector from the conditional mean to the conditional quantile setting. All three selectors treat $x_0$ as a boundary point. Denote the chosen bandwidths by $h_{n,\tau}^{cv}, h_{n,\tau}^{bdy}$, and $h_{n,\tau}^{ik}$, respectively. A description of these selectors can be found in Section S.1 of the online supplement. They are all implemented in the empirical application.

### 3.2 Empirical application (cont'd)

We estimate the effects of the two treatments on unemployment duration. The analysis uses the restricted sample of Card, Chetty, and Weber (2007). This sample excludes individuals who worked for only one firm in the past five years because, for these individuals, the two treatments are perfectly collinear. Thus, including them can potentially lead to positive biases for both effects. Later, in Section 7, we conduct a sensitivity analysis using another subsample.

The unemployment duration $Y$ is defined as the number of days from the end of the last job to the beginning of the next job. The spells are typically relatively short: over one-half of job losers find a new job within 20 weeks, and over three-quarters do so within a year. The resulting distribution is heavily skewed, with a long right tail. This constitutes a serious concern when estimating the average effect, as in Card, Chetty, and Weber (2007). For this reason, for some of the analysis (see their Figures V and VIII(a)), they exclude observations with spells longer than two years, with the latter roughly corresponding to the 85th percentile of the unconditional distribution. We choose not to exclude these observations, because, arguably, the long-term unemployed are the most

important. The quantile regression framework enables us to do so, because the estimation is local in the sense that the estimation of the lower quantiles is little affected, or even unaffected by the higher quantiles. However, in order not to be affected by those who may have left the labor force, we restrict our attention to distributional effects up to the 80th percentile. That is, we consider the quantile range $[0.2, 0.8]$. This leaves us with 457,615 observations; 72,014 of these are eligible for severance pay and 337,069 are eligible for extended benefits. For severance pay, the running variable $X$ is the job tenure at the firm from which the individual is laid off. For the extended benefits, the variable is the number of months worked in the past five years. All the estimates are obtained using the first estimation procedure; the second procedure returns essentially the same results.

The estimates $\hat{\delta}(\tau)$ are reported as dashed lines in Figures 1(a) (for severance pay) and 2(a) (for extended benefits). The $x$-axis is the quantile index and the $y$-axis is the unemployment duration. All estimates are computed with the bandwidth at the median set to 4.5. This represents a conservative choice, because the bandwidth values chosen by $h^{cv}_{0.5}, h^{bdy}_{0.5}$, and $h^{ik}_{0.5}$ for the two treatments are equal to (5.0, 4.5, 6.8) and $(9.0, 6.5, 6.9)$, respectively. Later, in Section 7, we report estimates with the median bandwidth set to 6.5 to examine the sensitivity of the result.

Two findings emerge from Figure 1(a). First, the effect of severance pay on unemployment duration is economically significant. Second, the estimates show substantial heterogeneity. At the 30th percentile, the effect is barely positive at a value of 4.00 days. At the median, the effect increases to 20.17 days. At the 70th and 80th percentiles, they reach 36.25 days and 71.00 days, respectively. To put these values in perspective, Card, Chetty, and Weber (2007) report an average effect of 10 days, while excluding observations with unemployment spells longer than two years. Under the rank invariance assumption, the heterogeneity implies that for individuals who would accept the next job relatively quickly without the treatment (e.g. those at the 30th percentile), the severance pay would have little effect (extending the unemployment duration only by four days). However, for individuals who would search for a longer time, the effects would be substantial. Interestingly, the point estimate at the 80th percentile is close to two months, while the severance pay is equal to two months' salary.

Now, consider extended benefits in Figure 2(a). The effects again exhibit heterogeneity, as in Figure 1(a). Further, at the 30th, 50th, 70th, and 80th percentiles, the estimates are equal to 7.00, 13.08, 27.50, and 36.33 days, respectively. The values at lower percentiles are close to the severance pay case, while those at the higher percentiles are smaller. Because the extended benefits amount to about 1.4 months' (i.e., around 42 days) salary, the estimate at the 80th percentile is again close

to the benefit level.

In summary, we find economically significant effects for both treatments. They tend to become stronger as $\tau$ increases. Next, we study them further using hypothesis tests and confidence bands.

## 4 A score test

This section first develops a score test for the treatment significance hypothesis, and then applies the test to the empirical application.

### 4.1 The test statistic

The intuition behind the score test is as follows. Under the null hypothesis of no treatment effect, $Q(\tau|x_0^+) = Q(\tau|x_0^-) = Q(\tau|x_0)$ for all $\tau \in \mathcal{T}$. Therefore, the observations around $x_0$ can be pooled to obtain a consistent estimate of $Q(\tau|x_0)$. In practice, this can be done using one of the two procedures in the previous section, while dropping $d_i$ from the objective function (1). Denote the estimate by $\hat{Q}(\tau|x_0)$. Under the alternative hypothesis, $Q(\tau|x_0^+) \neq Q(\tau|x_0^-)$ for some $\tau \in \mathcal{T}$. As a result, $\hat{Q}(\tau|x_0)$ will differ from $Q(\tau|x_0^+)$ or $Q(\tau|x_0^-)$, even asymptotically. The localized subgradient (the directional derivative of (1) with respect to the intercept), when evaluated at $\hat{Q}(\tau|x_0)$ using the observations on one side of $x_0$, should also be distinct from zero asymptotically.

Define

$$R_n(\tau) = (nh_{n,\tau})^{-1/2} \sum_{i=1}^{n} (\tau - 1(\hat{u}_i(\tau) \leq 0))\, d_i K_{i,\tau}, \tag{4}$$

where

$$\hat{u}_i(\tau) = y_i - \hat{\alpha}(\tau) - (x_i - x_0)\hat{\beta}(\tau) \text{ and } K_{i,\tau} = K\left((x_i - x_0)/h_{n,\tau}\right).$$

Here, $\hat{\alpha}(\tau)$ and $\hat{\beta}(\tau)$ are estimates in (1) after dropping $d_i$. The test statistic is

$$R_n\left(\mathcal{T}\right) = \sup_{\tau \in \mathcal{T}} \left|R_n(\tau)\right|. \tag{5}$$

The test does not require estimating the marginal density of $X$ or the conditional density of $Y$. It is constructed using only the observations on the right side of the cut-off. Those on the left side are redundant. To see this, let $R_n^-(\tau)$ denote the counterpart of $R_n(\tau)$ constructed using the observations on the left side of $x_0$: $R_n^-(\tau) = (nh_{n,\tau})^{-1/2} \sum_{i=1}^{n} (\tau - 1(\hat{u}_i(\tau) \leq 0))\,(1 - d_i) K_{i,\tau}$. Then, by the subgradient condition for the pooled sample, $R_n^-(\tau) + R_n(\tau) = o_p(1)$ uniformly over $\mathcal{T}$. This implies $R_n^-(\tau) = -R_n(\tau) + o_p(1)$ uniformly over $\mathcal{T}$, which leads to $\sup_{\tau\in\mathcal{T}} |R_n(\tau)| = \sup_{\tau\in\mathcal{T}} |R_n^-(\tau)| + o_p(1)$. This holds under both the null and the alternative hypotheses.

The following conditions are assumed to hold under the null hypothesis of no effect. They correspond to Assumptions 1–5 in Qu and Yoon (2015), but specifically for an interior point.

Let $B(x_0)$ be an open neighborhood of $x_0$. Let $f_X$ and $f_{Y|X}$ be the marginal density of $X$ and conditional density of $Y$, given $X$. $f_{Y|X}(Q(\tau|x)|x)$ is abbreviated as $f_{Y|X}(\tau|x)$, unless confusion may arise.

**Assumption 1** *The density $f_X$ is continuously differentiable at $x_0$, satisfying $0 < f_X(x_0) < \infty$.*

**Assumption 2** *(i) $f_{Y|X}(\tau|x_0)$ is Lipschitz continuous with respect to $\tau$ over $\mathcal{T}$. (ii) There exist finite positive constants $f_L, f_U$, and $\epsilon$, such that $f_L \leq f_{Y|X}(\tau+\eta|s) \leq f_U$ for all $|\eta| \leq \epsilon, s \in B(x_0)$, and $\tau \in \mathcal{T}$.*

**Assumption 3** *(i) $Q(\tau|x_0)$ and $\partial Q(\tau|x_0)/\partial\tau$ are finite and Lipschitz continuous over $\mathcal{T}$. (ii) $\partial^2 Q(\tau|s)/\partial s^2$ is finite and Lipschitz continuous over the set $\{(s,\tau)$: $s \in B(x_0)$ and $\tau \in \mathcal{T}\}$.*

**Assumption 4** *$K(\cdot)$ is compactly supported, symmetric, and bounded. It has finite first-order derivatives and satisfies $K(\cdot) \geq 0, \int K(u)du = 1, \int uK(u)du = 0$, and $\int u^2 K(u)du = \mu_2 < \infty$.*

**Assumption 5** *The bandwidth $h_{n,\tau}$ satisfies $h_{n,\tau} = c(\tau)h_n$, where $h_n = O(n^{-1/5})$, $nh_n \to \infty$, and $c(\tau)$ is Lipschitz continuous, with $0 < \underline{c} \leq c(\tau) \leq \bar{c} < \infty$ for all $\tau \in \mathcal{T}$.*

Assumption 1 is fairly standard. Assumptions 2 and 3 are local. That is, they involve only neighborhoods surrounding $x_0$ and $\mathcal{T}$. For example, if $\mathcal{T} = [0.5, 0.8]$, then the lower part of the conditional distribution is left unrestricted. Because $\partial Q(\tau|x_0)/\partial\tau = 1/f_{Y|X}(\tau|x_0)$, Assumption 2 implies 3(i), provided that $Q(\tau|x)$ is differentiable with respect to $x$ at $x_0$. Technically, under Assumption 3(ii), using a local quadratic estimator with an optimally chosen bandwidth can achieve a faster rate of convergence than that of a local linear estimator. Nevertheless, we choose to use the local linear estimator in order to derive simple procedures by applying the results in Qu and Yoon (2015). Assumption 5 requires the bandwidth parameter be a smooth function of $\tau$. This is needed to ensure stochastic equicontinuity. It is not restrictive, and is satisfied by the optimal bandwidth that minimizes the asymptotic MSE; see the discussion in Qu and Yoon (2015).

Define

$$\mu_j^+ = \int_0^\infty u^j K(u)\, du \text{ and } \mu_j^- = \int_{-\infty}^0 u^j K(u)\, du \text{ for } j = 0, 1, 2, 3.$$

Because of the symmetry of $K(\cdot)$, $\mu_j^+ = \mu_j^-$ when $j$ is even. The next result provides an approximation to $R_n(\tau)$ that holds uniformly over $\tau \in \mathcal{T}$.

**Lemma 1** *Let Assumptions 1–5 hold and assume* $m/(nh_n)^{1/4} \to \infty$ *as* $n \to \infty$. *In addition, let* $m/(nh_n)^{1/2} \to 0$ *hold for the first estimation procedure and* $(nh_{n,\tau}^5)^{1/2} \to h(\tau) < \infty$ *hold for the second procedure. Then, under the null hypothesis of no treatment effects,*

$$R_n(\tau) = (nh_{n,\tau})^{-1/2} \sum_{i=1}^{n} \left(\tau - 1(u_i^0(\tau) \leq 0)\right) \left\{ d_i - \frac{1}{2} - \left( \frac{x_i - x_0}{h_{n,\tau}} \right) \frac{\mu_1^+}{\mu_2} \right\} K_{i,\tau} + o_p(1), \qquad (6)$$

*where* $u_i^0(\tau) = y_i - Q(\tau|x_i)$ *and the order* $o_p(1)$ *holds uniformly over* $\mathcal{T}$.

The conditions on $m$ are the same as in Qu and Yoon (2015). They ensure that the monotonicity constraints and the rearrangement have no first-order effect on $\hat{Q}(\tau|x_0)$. The existence of a finite $h(\tau)$ is needed to satisfy Assumption 2 in Chernozhukov, Fernández-Val, and Galichon (2010). The simulation in Qu and Yoon (2015) shows that the finite sample properties of the estimators are not sensitive to the choice of $m$, provided that its value is not too small (say at least 10). In (6), the constants $\mu_1^+$ and $\mu_2$ are known once the kernel function is specified. When using the Epanechnikov kernel, $R_n(\tau) = (nh_{n,\tau})^{-1/2} \sum_{i=1}^{n} \left(\tau - 1(u_i^0(\tau) \leq 0)\right) \{d_i - 0.5 - (15/16)(x_i - x_0)/h_{n,\tau}\} K_{i,\tau} + o_p(1)$.

As a somewhat unexpected feature, the usual bias term does not enter the leading term in the approximation (6). This holds even when the bandwidth is of order $O(n^{-1/5})$, that is, without requiring under-smoothing. To see the intuition behind this, note that the key element in the test statistic $R_n(\mathcal{T})$ is $\hat{u}_i(\tau)$, which can be decomposed as

$$\begin{aligned} \hat{u}_i(\tau) &= \{y_i - Q(\tau|x_i)\} \\ &\quad + \left\{ Q(\tau|x_i) - Q(\tau|x_0) - (x_i - x_0) \frac{\partial Q(\tau|x_0)}{\partial x} \right\} \\ &\quad + \left\{ Q(\tau|x_0) + (x_i - x_0) \frac{\partial Q(\tau|x_0)}{\partial x} - \hat{\alpha}(\tau) - (x_i - x_0)\hat{\beta}(\tau) \right\}. \end{aligned} \qquad (7)$$

The first term on the right-hand side is equal to $u_i^0(\tau)$, the second term represents the remainder when replacing $Q(\tau|x_i)$ with a local linear approximation, and the third term reflects the effect of the parameter estimation. The last two terms both contain a bias, because they both depend on $\partial^2 Q(\tau|x_0)/\partial x^2$. For example, if $Q(\tau|x)$ is concave at $x_0$, then the biases in these two terms will be negative and positive, respectively. In aggregate, they cancel out each other, making $\hat{u}_i(\tau)$ bias free.

**Remark 1** *For the bias cancellation to occur,* $Q(\tau|x)$ *needs to be second-order differentiable at* $x_0$ *under the null hypothesis. If this assumption is too strong in a particular application, then the procedures in Section 5 can be used instead.*

**Proposition 1** *Under the same conditions as in Lemma 1, we have:* $R_n(\mathcal{T}) \Rightarrow \sup_{\tau\in\mathcal{T}}|G(\tau)|$, *where* $G(\tau)$ *is a zero-mean continuous Gaussian process with a covariance function that satisfies*

$$E[G(t)G(s)] = f_X(x_0)(\kappa(t)\kappa(s))^{-1/2}(t\wedge s - ts) \int_{-\infty}^{\infty}\left(1(z\geq 0)-\frac{1}{2}-\frac{\mu_1^+}{\mu_2}\frac{z}{\kappa(t)}\right)\left(1(z\geq 0)-\frac{1}{2}-\frac{\mu_1^+}{\mu_2}\frac{z}{\kappa(s)}\right)K\left(\frac{z}{\kappa(t)}\right)K\left(\frac{z}{\kappa(s)}\right)dz,$$

*where* $\kappa(\tau)=h_{n,\tau}/h_{n,1/2}=c(\tau)/c(1/2)$, *with* $c(\tau)$ *defined as in Assumption 5.*

The distribution of $\sup_{\tau\in\mathcal{T}}|G(\tau)|$ depends on the model only through the marginal density $f_X(x_0)$. The relevant critical values can be estimated by simulating from the following distribution:

$$\sup_{\tau\in\mathcal{T}}\left|(nh_{n,\tau})^{-1/2}\sum_{i=1}^{n}\left(\tau-1(u_i^0(\tau)\leq 0)\right)\left\{d_i-\frac{1}{2}-\left(\frac{x_i-x_0}{h_{n,\tau}}\right)\frac{\mu_1^+}{\mu_2}\right\}K_{i,\tau}\right|. \tag{8}$$

A general algorithm is given in Qu and Yoon (2015). It can be implemented in three steps. First, fix $x_i, d_i, K_{i,\tau}$, and $h_{n,\tau}$, and draw $(u_1,\ldots,u_n)\overset{iid}{\sim}\text{Uniform}(0,1)$. Second, evaluate the above expression, with $u_i^0(\tau)$ replaced by $u_i-\tau$ for all $\tau\in\mathcal{T}$. Third, repeat the above steps a large number of times. Sort the resulting values and use the corresponding percentiles as critical values.

For the score test, we implement the following two bandwidth selectors that explicitly treat $x_0$ as an interior point. The first selector determines the bandwidth at the median using leave-one-out cross validation. The second selector uses the minimum MSE bandwidth formula of Qu and Yoon (2015) for an interior point. The bandwidth at the median is then related to other quantiles using the formula of Yu and Jones (1998). Denote the chosen bandwidths by $h_{n,\tau}^{cvi}$ and $h_{n,\tau}^{ini}$, respectively. More details can be found in Section S.1 of the online supplement.

### 4.2 Empirical application (cont'd)

For severance pay, the two bandwidth selectors return $h_{n,0.5}^{cvi}=3.0$ and $h_{n,0.5}^{ini}=7.8$. The score test is equal to 0.102 and 0.107, respectively, using these two bandwidth values. The $p$-values are both less than 0.0001. For the extended benefits, the bandwidth selectors return $h_{n,0.5}^{cvi}$=4.0 and $h_{n,0.5}^{ini}$=6.6. The Score test is equal to 0.072 and 0.090, respectively. The two $p$-values are again both less than 0.0001. Therefore, there is strong statistical evidence against the null hypothesis of no treatment effects.

## 5 Wald tests

This section develops Wald tests for the treatment significance, homogeneity and, unambiguity hypotheses, allowing $\partial^2Q(\tau|x_0^+)/\partial x^2\neq\partial^2Q(\tau|x_0^-)/\partial x^2$. The analysis that maintains the restriction $\partial^2Q(\tau|x_0^+)/\partial x^2=\partial^2Q(\tau|x_0^-)/\partial x^2$ for all $\tau\in\mathcal{T}$ can be found in Section S.3 of the online appendix.

### 5.1 Test statistics

Let

$$W_n^R(\tau) = \sqrt{nh_{n,\tau}}\hat{f}_{Y|X}(\tau|x_0)\left(\hat{\delta}(\tau) - h_{n,\tau}^2(\hat{d}_\tau^+ - \hat{d}_\tau^-)\right), \tag{9}$$

where $\hat{\delta}(\tau)$ is defined in (3), $h_{n,\tau}^2\hat{d}_\tau^+$ and $h_{n,\tau}^2\hat{d}_\tau^-$ are estimates of the biases in $\hat{Q}(\tau|x_0^+) - Q(\tau|x_0^+)$ and $\hat{Q}(\tau|x_0^-) - Q(\tau|x_0^-)$, respectively, and $\hat{f}_{Y|X}(\tau|x_0) = [\hat{f}_{Y|X}(\tau|x_0^+) + \hat{f}_{Y|X}(\tau|x_0^-)]/2$, with $\hat{f}_{Y|X}(\tau|x_0^+)$ and $\hat{f}_{Y|X}(\tau|x_0^-)$ being estimates of $\lim_{x\downarrow x_0} f_{Y|X}(\tau|x)$ and $\lim_{x\uparrow x_0} f_{Y|X}(\tau|x)$, respectively. The multiplication by $\hat{f}_{Y|X}(\tau|x_0)$ ensures that the distribution of $W_n^R(\tau)$ depends asymptotically on the data only through $f_X(x_0)$ under the null hypothesis. The constructions of $\hat{d}_\tau^+, \hat{d}_\tau^-, \hat{f}_{Y|X}(\tau|x_0^+)$ and $\hat{f}_{Y|X}(\tau|x_0^-)$ are discussed in Subsection 5.2.

**Treatment significance.** This hypothesis can be tested using a Kolmogorov−Smirnov type test:

$$WS_n^R(\mathcal{T}) = \sup_{\tau\in\mathcal{T}}\left|W_n^R(\tau)\right|,$$

**Treatment homogeneity.** This hypothesis can be tested by measuring the deviation of $W_n^R(\tau)$ from the average of $W_n^R(\tau)$ over $\mathcal{T}$ :

$$WH_n^R(\mathcal{T}) = \sup_{\tau\in\mathcal{T}}\left|W_n^R(\tau) - \frac{\sqrt{nh_{n,\tau}}\hat{f}_{Y|X}(\tau|x_0)}{\int_{s\in\mathcal{T}}\sqrt{nh_{n,s}}\hat{f}_{Y|X}(s|x_0)ds}\int_{\tau\in\mathcal{T}}W_n^R(\tau)d\tau\right|.$$

**Treatment unambiguity.** To test this hypothesis, we determine whether the treatment can be detrimental at some unknown quantiles, using

$$WA_n^R(\mathcal{T}) = \sup_{\tau\in\mathcal{T}}\left|1\left(W_n^R(\tau)\leq 0\right)W_n^R(\tau)\right|.$$

Koenker and Xiao (2002) and Chernozhukov and Fernández-Val (2005) considered similar test statistics, but for parametric conditional quantile models. Note that the treatment unambiguity hypothesis can also be tested using $RA_n(\mathcal{T}) = \sup_{\tau\in\mathcal{T}}|1(R_n(\tau)\leq 0)R_n(\tau)|$, where $R_n(\tau)$ is given by (4). (This was pointed out by a referee). Under $\delta(\tau) = 0$, this statistic converges weakly to $\sup_{\tau\in\mathcal{T}}|G(\tau)1(G(\tau)\leq 0)|$ if $\partial^2 Q(\tau|x_0^+)/\partial x^2 = \partial^2 Q(\tau|x_0^-)/\partial x^2$ for all $\tau\in\mathcal{T}$.

### 5.2 Implementation

The three test statistics require estimating conditional densities. The density $f_{Y|X}(\tau|x_0^+)$ can be estimated using a kernel method:

$$\hat{f}_{Y|X}(\tau|x_0^+) = \int\frac{1}{h_{yx}}K\left((z-y)/h_{yx}\right)d\hat{F}(y|x_0^+), \tag{10}$$

where $z = \hat{Q}(\tau|x_0^+)$ and $\hat{F}(y|x_0^+) = \sup\{\tau \in (0,1)|\hat{Q}(\tau|x_0^+) \leq y\}$. To implement (10), we first obtain $\hat{Q}(\tau|x_0^+)$ using the bandwidth $h_{n,\tau}^{cv}$. Then, we sample from the distribution $\hat{F}(y|x_0^+)$ and apply kernel smoothing to the draws with bandwidth $h_{yx} = 2\widetilde{h}_{yx}$, where $\widetilde{h}_{yx}$ is determined using Silverman's rule of thumb formula. Alternatively, $f_{Y|X}(\tau|x_0^+)$ can be estimated using the difference quotient (see Koenker (2005) for a more detailed discussion about this estimator):

$$\hat{f}_{Y|X}\left(\tau|x_0^+\right) = \frac{2\delta_{n,\tau}}{\hat{Q}\left(\tau + \delta_{n,\tau}|x_0^+\right) - \hat{Q}\left(\tau - \delta_{n,\tau}|x_0^+\right)}, \tag{11}$$

where $\delta_{n,\tau}$ is a bandwidth parameter. Because $\sqrt{nh_{n,\tau}}(\hat{Q}(\tau|x_0^+) - Q(\tau|x_0^+))=O_p(1)$ uniformly over $\mathcal{T}$, (11) converges uniformly to $f_{Y|X}\left(\tau|x_0^+\right)$ if $\delta_{n,\tau} \to 0$ and $\delta_{n,\tau}(nh_{n,\tau})^{1/2} \to \infty$ uniformly over $\mathcal{T}$. The density $f_{Y|X}(\tau|x_0^-)$ can be estimated similarly using observations on the left side of $x_0$.

We estimate the bias using local quadratic regressions in three steps. First, minimize

$$\sum_{j=1}^{m}\sum_{i=1}^{n} \rho_{\tau_j}(y_i - \alpha^+(\tau_j) - \beta^+(\tau_j)(x_i - x_0) - \lambda^+(\tau_j)(x_i - x_0)^2)d_i K\left(\frac{x_i - x_0}{b_{n,\tau_j}}\right) \tag{12}$$

with respect to $\{\alpha^+(\tau_j), \beta^+(\tau_j), \lambda^+(\tau_j)\}_{j=1}^m$, where $b_{n,\tau_j}$ are bandwidth parameters that can vary across quantiles. Next, apply linear interpolation to obtain $\hat{\lambda}^+(\tau) = \gamma(\tau)\hat{\lambda}^+(\tau_j) + (1 - \gamma(\tau))\hat{\lambda}^+(\tau_{j+1})$, where $\hat{\lambda}^+(\tau_j)$ are the solutions to (12) and $\gamma(\tau) = (\tau_{j+1} - \tau)/(\tau_{j+1} - \tau_j)$. Finally, compute

$$\hat{d}_\tau^+ = \Gamma\hat{\lambda}^+(\tau) \quad \text{with } \Gamma = \frac{\left(\mu_2^+\right)^2 - \mu_1^+\mu_3^+}{\mu_0^+\mu_2^+ - \left(\mu_1^+\right)^2}. \tag{13}$$

The estimate $\hat{d}_\tau^-$ can be obtained in the same way after replacing $d_i$ in (12) with $1 - d_i$.

### 5.3 Asymptotic properties

We first separately study $\hat{\delta}(\tau)$ and $\hat{d}_\tau^+ - \hat{d}_\tau^-$, and then combine the results to obtain asymptotic approximations for the three test statistics.

The next two assumptions relax Assumptions 2 and 3 to allow $Q(\tau|x)$ to be discontinuous at $x_0$. Note that they allow $\partial^2 Q(\tau|x_0^+)/\partial x^2 \neq \partial^2 Q(\tau|x_0^-)/\partial x^2$.

**Assumption 6** *(i) The densities $f_{Y|X}(\tau|x_0^+)$ and $f_{Y|X}(\tau|x_0^-)$ are Lipschitz continuous in $\tau$ over $\mathcal{T}$. (ii) There exist finite constants $f_L > 0$, $f_U > 0$, $\epsilon > 0$, and $c > 0$, such that $f_{Y|X}(\tau + \eta|s)$ and $f_{Y|X}(\tau + \eta|s')$ lie between $f_L$ and $f_U$ for all $|\eta| \leq \epsilon$, $s \in [x_0 - c, x_0)$, $s' \in (x_0, x_0 + c]$, and $\tau \in \mathcal{T}$.*

**Assumption 7** *(i) $Q(\tau|x_0^+)$, $Q(\tau|x_0^-)$, $\partial Q(\tau|x_0^+)/\partial\tau$, and $\partial Q(\tau|x_0^-)/\partial\tau$ are finite and Lipschitz continuous in $\tau$ over $\mathcal{T}$. (ii) $\partial Q(\tau|x)/\partial x$ and $\partial^2 Q(\tau|x)/\partial x^2$ are finite and Lipschitz continuous over $\{(x,\tau)$: $x \in (x_0, x_0 + c], \tau \in \mathcal{T}\}$ and $\{(x,\tau)$: $x \in [x_0 - c, x_0), \tau \in \mathcal{T}\}$ for some $c > 0$.*

To present the asymptotic approximation to $\hat{\delta}(\tau)$, define

$$d_\tau^+ = \frac{1}{2}\Gamma\frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2},\ d_\tau^- = \frac{1}{2}\Gamma\frac{\partial^2 Q(\tau|x_0^-)}{\partial x^2}, \Xi_{i,\tau}^+ = \frac{\mu_2^+ - \left(\frac{x_i - x_0}{h_{n,\tau}}\right)\mu_1^+}{\mu_0^+\mu_2^+ - \left(\mu_1^+\right)^2},\ \Xi_{i,\tau}^- = \frac{\mu_2^- - \left(\frac{x_i - x_0}{h_{n,\tau}}\right)\mu_1^-}{\mu_0^-\mu_2^- - \left(\mu_1^-\right)^2}.$$

Let

$$W_{n,c}(\tau) = \sqrt{nh_{n,\tau}}\left\{\hat{\delta}(\tau) - \delta(\tau) - h_{n,\tau}^2\left(d_\tau^+ - d_\tau^-\right)\right\}. \tag{14}$$

**Lemma 2** *Let Assumptions 1, 4, 5, 6, and 7 hold. Assume $m/(nh_n)^{1/4} \to \infty$ as $n \to \infty$. In addition, let $m/(nh_n)^{1/2} \to 0$ hold for the first estimation procedure and $(nh_{n,\tau}^5)^{1/2} \to h(\tau) < \infty$ hold for the second procedure. Then, uniformly over $\tau \in \mathcal{T}$,*

$$W_{n,c}(\tau) = D_1^+(\tau) - D_1^-(\tau) + o_p(1),$$

*where*

$$D_1^+(\tau) = \frac{(nh_{n,\tau})^{-1/2}\sum_{i=1}^n\left(\tau - 1(u_i^0(\tau) \le 0)\right)\Xi_{i,\tau}^+ d_i K_{i,\tau}}{f_X(x_0)\, f_{Y|X}\left(\tau|x_0^+\right)},$$

*and $D_1^-(\tau)$ is equal to $D_1^+(\tau)$ with $d_i$, $f_{Y|X}(\tau|x_0^+)$, and $\Xi_{i,\tau}^+$ replaced by $(1-d_i)$, $f_{Y|X}(\tau|x_0^-)$, and $\Xi_{i,\tau}^-$, respectively.*

We now study $\hat{d}_\tau^+ - \hat{d}_\tau^-$. Under conventional asymptotics, if $\hat{d}_\tau^+$ and $\hat{d}_\tau^-$ satisfy $\hat{d}_\tau^+ - d_\tau^+ = o_p(1)$ and $\hat{d}_\tau^- - d_\tau^- = o_p(1)$ uniformly over $\tau \in \mathcal{T}$, then $\sqrt{nh_{n,\tau}}\{\hat{\delta}(\tau) - \delta(\tau) - h_{n,\tau}^2(\hat{d}_\tau^+ - \hat{d}_\tau^-)\}$ will converge weakly to the same Gaussian process as $W_{n,c}(\tau)$ does. In practice, however, $\hat{d}_\tau^+$ and $\hat{d}_\tau^-$ can exhibit high variability and can have large effects on the finite sample distributions of the tests. This phenomenon is well documented in the nonparametric inference literature. Below, we derive an alternative approximation that explicitly accounts for the effect of the bias estimation. This approximation is inspired by work of Calonico, Cattaneo, and Titiunik (2014), who study the inference on the average treatment effect under RD designs. Their key contribution is a novel variance formula for the average treatment effect with additional components coming from the bias estimation. Relative to their setting, the inference here is on the quantile process, rather than on a finite dimensional parameter. Thus it might not be clear whether an analogous development is possible. We show that this is the case using the conditionally pivotal property of the relevant expressions.

**Assumption 8** *(i) $\partial^3 Q(\tau|x)/\partial x^3$ is finite and Lipschitz continuous over $\{(x,\tau)$: $x \in (x_0, x_0 + c], \tau \in \mathcal{T}\}$ and also over $\{(x,\tau)$: $x \in [x_0 - c, x_0), \tau \in \mathcal{T}\}$, where $c$ is some positive constant. (ii) $\partial^3 Q(\tau|x_0^+)/\partial x^3$ and $\partial^3 Q(\tau|x_0^-)/\partial x^3$ are finite and Lipschitz continuous over $\mathcal{T}$.*

**Assumption 9** *The bandwidth $b_{n,\tau}$ satisfies $b_{n,\tau} = c(\tau)b_n$, where $b_n = o(n^{-1/7})$ and $nb_n \to \infty$ and $c(\tau)$ is Lipschitz continuous, satisfying $0 < \underline{c} \leq c(\tau) \leq \bar{c} < \infty$ for all $\tau \in \mathcal{T}$. The values of $c(\tau), \underline{c}$, and $\bar{c}$ can be different from those in Assumption 5.*

Assumption 8 is analogous to Assumption 7(ii). It ensures that the remainder term from the local quadratic approximation is uniformly small. The third-order derivatives on the two sides of the threshold can be different. Assumption 9 plays a similar role to that of Assumption 5 in the local linear regression. The bandwidth $b_{n,\tau}$ can be of the same or higher order than $h_{n,\tau}$. To present the result, let $\iota_3' = [0, 0, 1]$, $\bar{z}_{i,\tau}' = [1 \;\; (x_i - x_0)/b_{n,\tau} \;\; (x_i - x_0)^2/b_{n,\tau}^2]$, $\bar{K}_{i,\tau} = K\left((x_i - x_0)/b_{n,\tau}\right)$, and $\bar{N}^+$ be a 3-by-3 matrix, where the $(i,j)$-th element is given by $\mu_{i+j-2}^+ = \int_0^\infty u^{i+j-2} K(u)\, du$. Let $\bar{N}^-$ equal $\bar{N}^+$, with $\mu_{i+j-2}^+$ replaced by $\mu_{i+j-2}^- = \int_{-\infty}^0 u^{i+j-2} K(u)\, du$.

**Lemma 3** *Let Assumptions 1, 4, 5, 6, 7, 8, and 9 hold. Assume $m/(nb_{n,\tau}^5)^{1/4} \to \infty$ as $n \to \infty$. Then, uniformly over $\tau \in \mathcal{T}$ :*

$$\sqrt{nb_{n,\tau}^5}\left(\hat{d}_\tau^+ - \hat{d}_\tau^- - \left(d_\tau^+ - d_\tau^-\right)\right) = D_2^+(\tau) - D_2^-(\tau) + o_p(1),$$

*where*

$$D_2^+(\tau) = \Gamma \frac{\iota_3'(\bar{N}^+)^{-1}(nb_{n,\tau})^{-1/2}\sum_{i=1}^n \left\{\tau - 1\left(u_i^0(\tau) \leq 0\right)\right\}\bar{z}_{i,\tau} d_i \bar{K}_{i,\tau}}{f_{Y|X}\left(\tau|x_0^+\right) f_X(x_0)},$$

*and $D_2^-(\tau)$ is equal to $D_2^+(\tau)$, with $d_i$, $f_{Y|X}(\tau|x_0^+)$, and $\bar{N}^+$ replaced by $(1 - d_i)$, $f_{Y|X}(\tau|x_0^-)$, and $\bar{N}^-$, respectively.*

Define

$$G_*^R(\tau) = \hat{f}_{Y|X}(\tau|x_0)\left\{\left[D_1^+(\tau) - D_1^-(\tau)\right] - \left(\frac{\sqrt{nh_{n,\tau}^5}}{\sqrt{nb_{n,\tau}^5}}\right)\left[D_2^+(\tau) - D_2^-(\tau)\right]\right\}. \tag{15}$$

Combining Lemmas 2 and 3 leads to the following approximations for the three test statistics.

**Proposition 2** *Let the conditions in Lemmas 2 and 3 hold. Then:*

1. *Under $\delta(\tau) = 0$ for all $\tau \in \mathcal{T}$, $WS_n^R(\mathcal{T}) - \sup_{\tau\in\mathcal{T}}\left|G_*^R(\tau)\right| = o_p(1)$.*

2. *Under $\delta(\tau) = \delta$ for all $\tau \in \mathcal{T}$ for some $\delta \in \mathbb{R}$,*

$$WH_n^R(\mathcal{T}) - \sup_{\tau\in\mathcal{T}}\left|G_*^R(\tau) - \frac{\sqrt{nh_{n,\tau}}\hat{f}_{Y|X}(\tau|x_0)}{\int_{s\in\mathcal{T}}\sqrt{nh_{n,s}}\hat{f}_{Y|X}(s|x_0)ds}\int_{\tau\in\mathcal{T}} G_*^R(\tau)d\tau\right| = o_p(1).$$

*3. Under the least favorable null hypothesis of $\delta(\tau) = 0$ for all $\tau \in \mathcal{T}$,*

$$WA_n^R(\mathcal{T}) - \sup_{\tau\in\mathcal{T}} \left|1\left(G_*^R(\tau) \leq 0\right) G_*^R(\tau)\right| = o_p(1).$$

**Remark 2** *Conditional on $\{x_i\}_{i=1}^n$, the randomness in the four components $D_1^+(\tau), D_2^+(\tau), D_1^-(\tau)$, and $D_2^-(\tau)$ all comes from the same source: $\left\{\tau - 1(u_i^0(\tau) \leq 0)\right\}_{i=1}^n$. As a result, $G_*^R(\tau)$ can be simulated by: (i) obtaining $(u_1, \ldots, u_n) \overset{iid}{\sim} Uniform(0,1)$, (ii) evaluating the four components with $\{\tau - 1(u_i - \tau \leq 0)\}_{i=1}^n$ replacing $\left\{\tau - 1(u_i^0(\tau) \leq 0)\right\}_{i=1}^n$, (iii) repeating this a large number of times. The simulated $G_*^R(\tau)$ can then be substituted into the expressions in Proposition 2 to obtain the critical values. The validity of this procedure is proved in the online supplement. See Section S.5.*

**Remark 3** *Proposition 2 can be extended to allow us to conduct inference on conditional quantile processes in a general nonparametric setting, with $x_0 \in R^k$ being an interior or a boundary point.*

**Remark 4** *The local power properties of the score and Wald tests are analyzed in the online appendix. There, the restriction $\partial^2 Q(\tau|x_0^+)/\partial x^2 = \partial^2 Q(\tau|x_0^-)/\partial x^2$ is imposed in order to make the comparison meaningful. Interestingly, the score and Wald tests for the treatment significance hypothesis have the same local asymptotic power against the sequence considered there. In addition, the results show that the tests can have nontrivial power against alternatives of order $(nh_n)^{-1/2}$. Finally, what matters for power is not only the difference $Q(\tau|x_0^+) - Q(\tau|x_0^-)$, but also the conditional density and the bandwidth. Everything else being equal, the power is higher if the departure from the null occurs in a dense region or at a place where the bandwidth is wider. See Section S.4 for details.*

### 5.4 An extension

The above analysis allows $d_\tau^+ - d_\tau^-$ to vary freely across quantiles. In practice, there can be situations where we expect $d_\tau^+$ and $d_\tau^-$ to be different, but their difference remains relatively constant across the quantiles. The following procedure incorporates this information into the tests:
STEP 1. Solve (12) to obtain $\hat{d}_\tau^+$ and $\hat{d}_\tau^-$. STEP 2. Compute $\bar{d} = \int_{\mathcal{T}} (\hat{d}_\tau^+ - \hat{d}_\tau^-) d\tau$ and

$$W_n^E(\tau) = \sqrt{nh_{n,\tau}}\hat{f}_{Y|X}(\tau|x_0)\left(\hat{\delta}(\tau) - h_{n,\tau}^2 \bar{d}\right). \tag{16}$$

STEP 3. Construct $WS_n^E(\mathcal{T}), WH_n^E(\mathcal{T})$, and $WA_n^E(\mathcal{T})$ in the same way as $WS_n^R(\mathcal{T}), WH_n^R(\mathcal{T})$, and $WA_n^R(\mathcal{T})$, but with $W_n^E(\tau)$ replacing $W_n^R(\tau)$.

To obtain the critical values, define

$$\bar{D}_2^+(\tau) = \sqrt{nh_{n,\tau}^5}\int_{\mathcal{T}} (nb_{n,s}^5)^{-1/2} D_2^+(s)\, ds \quad \text{and} \quad \bar{D}_2^-(\tau) = \sqrt{nh_{n,\tau}^5}\int_{\mathcal{T}} (nb_{n,s}^5)^{-1/2} D_2^-(s)\, ds.$$

In addition, let $G_*^E(\tau) = \hat{f}_{Y|X}(\tau|x_0)\{[D_1^+(\tau) - \bar{D}_2^+(\tau)] - [D_1^-(\tau) - \bar{D}_2^-(\tau)]\}$. Then, the tests $WS_n^E(\mathcal{T}), WH_n^E(\mathcal{T})$, and $WA_n^E(\mathcal{T})$ satisfy the same formulae as in Proposition 2, with $G_*^E(\tau)$ replacing $G_*^R(\tau)$. The critical values can be obtained in the same way as in Remark 2.

### 5.5 Empirical application (cont'd)

We use the tests assuming $\partial^2 Q(\tau|x_0^+)/\partial x^2 = \partial^2 Q(\tau|x_0^-)/\partial x^2$ as the benchmark, and then examine the effect of the bias correction. The bandwidth at the median is set to 4.5.

First, consider severance pay. The Wald test for the treatment significance hypothesis is equal to 74.50 and the $p$-value is less than 0.0001. This confirms the finding of the score test. The Wald test for the treatment homogeneity hypothesis is equal to 60.43. The $p$-value is again less than 0.0001. Therefore, there is strong evidence supporting treatment heterogeneity. The test for the treatment unambiguity hypothesis equals 0.00, with the $p$-value being 0.8810. This suggests that the treatment effects are uniformly non-negative.

When applying the quantile-by-quantile bias correction, the Wald tests for the three hypotheses equal 56.40, 30.67, and 0.00, respectively. The $p$-values are 0.0025, 0.0065, and 0.8764, respectively. Therefore, the conclusions are the same as before. Finally, under the constrained bias estimation, the Wald tests equal 52.94, 63.56, and 40.61, respectively. The $p$-values are 0.0003, 0.0001, and 0.0465, respectively. The first two tests lead to the same conclusions as before. However, the unambiguity test now rejects the null hypothesis at the 5% level. This is because the bias correction makes the estimates at lower quantiles slightly negative, which is interpreted by the test as evidence supporting negative treatment effects.

For extended benefits, the Wald test for the treatment significance hypothesis is equal to 61.22. The $p$-value is less than 0.0001. Therefore, the treatment is statistically significant. The test for the treatment homogeneity hypothesis is equal to 37.54. The $p$-value is less than 0.0001. There is clear evidence of heterogeneity. The test for the treatment unambiguity hypothesis is equal to 0.00, with a $p$-value of 0.8796. This is consistent with the effects being uniformly positive. Applying the bias correction confirms these conclusions. Specifically, under the quantile-by-quantile bias estimation, the Wald tests for the three hypotheses equal 65.80, 45.01, and 0.00, while the $p$-values are less than 0.0001 for the first two tests, and 0.8772 for the last test. Under the constrained bias estimation, the three Wald tests equal 63.41, 37.47, and 0.00, with $p$-values of 0.0001, 0.0009, and 0.8123, respectively.

In summary, the tests suggest that the two treatments are statistically significant, heterogeneous, and uniformly positive. The only exception is in the case of severance pay under the

constrained bias estimation. There, the effects at the 30th percentile and below appear slightly negative.

## 6 Uniform confidence bands for the quantile effects

We first develop uniform confidence bands for $\delta(\tau)$, and then apply them to the empirical application.

### 6.1 Confidence bands with and without bias estimation

For now, assume $d_\tau^+ - d_\tau^-$ is known. Then, by Lemma 2,

$$\left|\sqrt{nh_{n,\tau}}\hat{f}_{Y|X}\left(\tau|x_0\right)\left(\hat{\delta}(\tau)-\delta(\tau)-h_{n,\tau}^2\left(d_\tau^+-d_\tau^-\right)\right)\right| = \hat{f}_{Y|X}(\tau|x_0)\left|D_1^+\left(\tau\right)-D_1^-\left(\tau\right)\right|+o_p\left(1\right).$$

The following procedure leads to an asymptotic $100(1-p)\%$ confidence band for $\delta(\tau)$:

Step 1. Simulate the supremum of $\hat{f}_{Y|X}(\tau|x_0)\left|D_1^+\left(\tau\right)-D_1^-\left(\tau\right)\right|$ over $\tau\in\mathcal{T}$.

Step 2. Compute the $(1-p)$-th percentile of the resulting empirical distribution. Call it $c_p(x_0)$.

Step 3. Compute the confidence band for $\delta(\tau)$ over $\tau\in\mathcal{T}$ as

$$\left(\hat{\delta}(\tau)-h_{n,\tau}^2\left(d_\tau^+-d_\tau^-\right)\right)\pm\frac{c_p(x_0)}{\sqrt{nh_{n,\tau}}\hat{f}_{Y|X}\left(\tau|x_0\right)}. \tag{17}$$

It follows from Lemma 2 and Corollary 2 in Qu and Yoon (2015) that this band covers $Q(\tau|x_0^+)-Q(\tau|x_0^-)$ over $\tau\in\mathcal{T}$ asymptotically, with probability $(1-p)$. The band is usually wider near the tails of the conditional distribution. This is because $\hat{f}_{Y|X}\left(\tau|x_0\right)$ tends to get smaller near the tails. This effect typically dominates the corresponding increase in $h_{n,\tau}$ as $\tau$ approaches the tails.

Now, we adapt the above procedure to the following three circumstances: (i) $d_\tau^+-d_\tau^-=0$ for all $\tau\in\mathcal{T}$; (ii) $d_\tau^+-d_\tau^-\neq 0$ for some $\tau\in\mathcal{T}$, with the possibility of $d_\tau^+-d_\tau^-\neq d_s^+-d_s^-$ for some $\tau,s\in\mathcal{T}$; and (iii) $d_\tau^+-d_\tau^-\neq 0$ for some $\tau\in\mathcal{T}$, and $d_\tau^+-d_\tau^-=d_s^+-d_s^-$ for all $\tau,s\in\mathcal{T}$.

Under (i), we have $(d_\tau^+-d_\tau^-)=0$. The confidence band (17) reduces to

$$\hat{\delta}(\tau)\pm\frac{c_p(x_0)}{\sqrt{nh_{n,\tau}}\hat{f}_{Y|X}\left(\tau|x_0\right)}. \tag{18}$$

Under (ii), a uniform confidence band is given by

$$\left(\hat{\delta}(\tau)-h_{n,\tau}^2(\hat{d}_\tau^+-\hat{d}_\tau^-)\right)\pm\frac{c_p^R(x_0)}{\sqrt{nh_{n,\tau}}\hat{f}_{Y|X}\left(\tau|x_0\right)}, \tag{19}$$

where $\hat{d}_\tau^+$ and $\hat{d}_\tau^-$ are the estimates of $d_\tau^+$ and $d_\tau^-$ given by (13), and $c_p^R(x_0)$ is the $(1-p)$-th percentile of $\sup_{\tau\in\mathcal{T}}\left|G_*^R(\tau)\right|$, without imposing $f_{Y|X}(\tau|x_0^-)=f_{Y|X}(\tau|x_0^+)$. Under (iii), a uniform

band is given by

$$\left(\hat{\delta}(\tau) - h_{n,\tau}^2 \bar{d}\right) \pm \frac{c_p^E(x_0)}{\sqrt{nh_{n,\tau}}\hat{f}_{Y|X}\left(\tau|x_0\right)}, \tag{20}$$

where $c_p^E(x_0)$ is the $(1-p)$-th percentile of $\sup_{\tau\in\mathcal{T}}\left|G_*^E(\tau)\right|$, without imposing $f_{Y|X}(\tau|x_0^-) = f_{Y|X}(\tau|x_0^+)$.

Our results are related to those of two other studies: Frandsen, Frölich, and Melly (2012) and Shen and Zhang (2016). Frandsen, Frölich, and Melly (2012) show that the conditional quantile processes, obtained by inverting conditional distributions of potential outcomes, converge weakly to Gaussian processes. One advantage of a distribution function based approach is that it allows a discrete outcome variable. Although it has not been done, their result might also lead to tests and uniform confidence bands for the quantile effects. However, because their analysis is based on a different methodology, the implementation and theory for inference would both differ from those presented here. Shen and Zhang (2016) consider hypothesis tests for distributional treatment effects of the form $F_{Y|X}(y|x_0^+) - F_{Y|X}(y|x_0^-) = 0$ or $\geq 0$ $(\leq 0)$ for $y \in R$. Their tests are of the Kolmogorov–Smirnov type, based on comparing CDFs on two sides of the cut-off. Their study differs from ours in three ways. First, the objects being studied are different. Shen and Zhang (2016) study distributional effects, while we study quantile effects. Although the two effects are always consistent in signs under sharp RD designs, their magnitudes can convey very different information. For example, a homogeneous quantile effect (i.e., $Q(\tau|x_0^+) - Q(\tau|x_0^-) = \delta$ for all $\tau \in \mathcal{T}$) can correspond to non-monotonic distributional effects (e.g., $F_{Y|X}(y|x_0^+) - F_{Y|X}(y|x_0^-)$ increasing, and then decreasing as $y$ increases). Therefore, distributional and quantile effects are complementary, rather than substitutable. This is the most important difference between the two studies. Second, the scopes are different. Shen and Zhang (2016) use the clever observation that, under fuzzy designs, the signs of local distributional treatments are the same as that of $F_{Y|X}(y|x_0^+) - F_{Y|X}(y|x_0^-)$. This makes their tests applicable to fuzzy designs. We focus on sharp designs. However, the procedure can be used to study the treatment homogeneity hypothesis, which is beyond the scope of Shen and Zhang (2016). Finally, the inference is also different. Shen and Zhang (2016) report pointwise confidence intervals, while we develop and implement uniform confidence bands.

## 6.2 Empirical application (cont'd)

For each treatment, we use the confidence band without bias correction as the benchmark, and then study the difference caused by the bias estimation. We use the same bandwidth as before.

Consider severance pay. The shaded area in Figure 1(a) represents the 90% uniform band without bias correction, calculated from (18). The values of the band at the 30th, 50th, 70th, and 80th percentiles equal $[-1.88, 9.88]$, $[12.99, 27.35]$, $[21.10, 51.40]$, and $[27.34, 114.66]$, respectively. Clearly, substantial heterogeneity is present.

Figures 1(b) and 1(c) report the results with bias correction. The dashed lines represent the bias adjusted estimates and the shaded areas are 90% uniform confidence bands. Under quantile-by-quantile bias correction, the estimates show the same tendency as in Figure 1(a), although the extent of the heterogeneity is less pronounced. This is partly because the confidence bands are substantially wider, but also because the point estimates take on smaller values. Under constrained bias estimation, the heterogeneity is comparable to that in Figure 1(a). In summary, the results consistently show that the treatment effect tends to increase as the quantile index increases. At the same time, there is uncertainty about the exact magnitude of the effect, which depends on the assumption about the second-order derivative of the conditional quantile function at the cut-off.

Now, consider extended benefits. The shaded area in Figure 2(a) corresponds to a confidence band without bias correction. The effects exhibit pronounced heterogeneity, as in Figure 1(a). Further, at the 30th, 50th, 70th, and 80th percentiles, the values of the 90% confidence band equal $[1.35, 12.65]$, $[6.26, 19.91]$, $[12.53, 42.47]$, and $[-10.04, 82.70]$, respectively. The point estimates in Figures 2(b) and 2(c) are similar to those in Figure 2(a). The confidence band in Figure 2(b) is again noticeably wider. Overall, consistent evidence for heterogeneity is detected, irrespective of whether or not we implement bias correction.

## 7 Sensitivity analysis and summary of findings

This section repeats the analysis using a subsample. It also considers an alternative bandwidth.

### 7.1 A different subsample

Considering a subsample is motivated by the following concern. Although the sample above excludes workers who worked at only one company in the past five years, it includes workers who worked for one to four months at a different company. For these individuals, it is possible that they become eligible for extended benefits within a four-month window prior to becoming eligible for severance pay. Because the bandwidth we use is equal to 4.5, this can potentially lead to over-estimating the effect of the extended benefits, while under-estimating that of the severance pay. To determine whether this has happened, we repeat the estimation using a subsample that only includes workers who worked for more than four months at a firm different from the one where they were laid off.

The estimates and their 90% confidence bands are reported in Figures 3 and 4. The bandwidth at the median is still equal to 4.5. The results are fairly close to those reported in Figures 1 and 2. In particular, for severance pay, the estimates and the confidence bands at the 30, 50, 70, and 80th percentiles equal 4.00, $[-1.93, 9.93]$; 20.17, $[12.89, 27.44]$; 33.33, $[18.16, 48.51]$; and 57.00, $[12.16, 101.84]$, respectively. For extended benefits, the values are 6.00, $[0.10, 11.90]$; 13.33, $[6.15, 20.52]$; 29.08, $[13.18, 44.99]$; and 31.75, $[-17.37, 80.87]$, respectively. Therefore, there is no evidence of biased estimates.

### 7.2 An alternative bandwidth

The analysis has used a conservative bandwidth to avoid a large bias. However, using a larger bandwidth can potentially yield sharper results if the bias remains adequately accounted for. To examine this further, we obtain the estimates and the confidence bands using $h_{n,0.5} = 6.5$. The results are reported in Figures 5 and 6. Qualitatively, they are similar to those reported in Figures 1 and 2. Quantitatively, the bands for severance pay under the quantile-by-quantile bias correction are now narrower, and the effects are closer to those without the bias correction. In summary, the conclusions remain the same after carrying out the two sensitivity analyses.

### 7.3 Summary

The analysis finds significant heterogeneity for both treatments. This raises an interesting question: what economic models are consistent with such behaviors? Card, Chetty, and Weber (2007) argue that their estimates are inconsistent with the prediction of a simple permanent income model, and are also inconsistent with naive "rule of thumb" behavior. Simply put, the effect is too large compared with the former, but too small compared with the latter. Our quantile based estimates lead to values that lie on both sides of the Card, Chetty, and Weber estimates. Under the rank-invariance assumption, it appears some individuals (the short-term unemployed) behave as if they follow the permanent income hypothesis, while others (the long-term unemployed) behave myopically. On the one hand, this implies that Card, Chetty, and Weber's conclusion remains relevant when viewed through the lens of quantile regressions. On the other hand, it implies that the heterogeneity must be accounted for and can also be a key channel explaining the behavior of job searchers.

The results demonstrate that it is important to consider the long-term unemployed (those in the upper quantiles of the distribution). Such individuals are more liquidity constrained. In addition, under the rank-invariance assumption, the treatment effects are strongest for these individuals.

Interestingly, our results show that at the 80th percentile, the estimated effects are close to the respective levels of the benefits. This suggests an intriguing hypothesis that for the long-term unemployed, the liquidity effect may be playing a dominating role in determining the length of the job search. This reinforces the conclusion of Card, Chetty, and Weber that a substantial share of the behavioral responses to longer unemployment benefits is attributable to a liquidity effect.

## 8 Monte Carlo

This section studies four issues: (1) whether the bandwidth selectors are informative, (2) the performance of the score test relative to that of the Wald test for the treatment significance hypothesis, and (3) the size and power properties of the Wald tests, with and without bias correction.

We consider four data generating processes (DGPs). Under the null hypothesis of no treatment effects, their conditional quantile functions $Q(\tau|x)$ are given by

$$\begin{aligned}
&\text{Model 1:} && 1 + x + (0.5 + 0.3x)\, Q_\varepsilon(\tau), \\
&\text{Model 2:} && 0.5 + x + x^2 + \sin(\pi x - 1) + (x + 1.25)\, Q_\varepsilon(\tau), \\
&\text{Model 3:} && \begin{cases} 0.48 + 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5 + 0.1295 Q_\varepsilon(\tau) & \text{if } x < 0 \\ 0.48 + 0.84x - 3x^2 + 7.99x^3 - 9.01x^4 + 3.16x^5 + 0.1295\, Q_\varepsilon(\tau) & \text{if } x \geq 0 \end{cases}, \\
&\text{Model 4:} && \begin{cases} 3x^2 + 0.1295\, Q_\varepsilon(\tau) & \text{if } x < 0 \\ 4x^2 + 0.1295 Q_\varepsilon(\tau) & \text{if } x \geq 0 \end{cases}.
\end{aligned} \tag{21}$$

The first two models satisfy $\partial^2 Q(\tau|x_0^+)/\partial x^2 = \partial^2 Q(\tau|x_0^-)/\partial x^2$. Model 1 is linear in $x$, while Model 2 exhibits significant curvature. Models 3 and 4 are taken from Imbens and Kalyanaraman (2012). They both satisfy $\partial^2 Q(\tau|x_0^+)/\partial x^2 \neq \partial^2 Q(\tau|x_0^-)/\partial x^2$. Other aspects of the DGPs are as follows. For Models 1 and 2, the values of $x$ are independent realizations from $\mathrm{U}(-1, 1)$. For Models 3 and 4, following Imbens and Kalyanaraman (2012), the values of $x$ are independent realizations from $(2\mathrm{Beta}(2, 4) - 1)$. The errors $\varepsilon_i$ are always generated i.i.d. from $N(0, 1)$, such that $Q_\varepsilon(\tau) = \Phi^{-1}(\tau)$.

The quantile range $\mathcal{T}$ is set to $[0.2, 0.8]$. The sample sizes are $n = 500, 1000, 2000$. Because the nonparametric estimation involves conditional quantiles at $\tau = 0.2$ and 0.8, $n = 500$ can be viewed as a relatively small sample size. The cut-off is $x_0 = 0$, and the treatment assignment satisfies $d_i = 1(x_i \geq x_0)$. The nominal level for the tests is set to 10%. All the results reported are based on 2000 replications, and are obtained using the first estimation procedure. Subsection 8.1 considers Models 1 and 2, while Subsection 8.2 considers Models 3 and 4.

### 8.1 Models 1 and 2

**The selected bandwidth.** Table 1 summarizes the means and standard deviations of the selected bandwidths at $\tau = 0.5$. The bandwidths are divided into two groups, depending on whether or not they impose the null hypothesis of no effects. Their values are restricted to fall between 0.1 and 0.5. The lower bound safeguards against using too few observations (e.g., it corresponds to about 25 observations when $n = 500$ and the bandwidth is selected under the alternative hypothesis), while the upper bound allows us to use approximately half the observations on either side of $x_0$.

The results show that the bandwidth selectors are informative. In particular, the bandwidths for Model 1 are consistently wider than those for Model 2. Between the two bandwidth selectors that impose the null hypothesis, $h^{int}_{0.5}$ tends to return greater bandwidths than $h^{cvi}_{0.5}$ does. Among the three selectors that operate under the alternative hypothesis, the bandwidths chosen by $h^{cv}_{0.5}$ tend to be the largest. The Imbens and Kalyanaraman bandwidth exhibits the smallest standard deviation owning to the effects of the regularization terms $r_-$ and $r_+$. Below, the score test is computed using the bandwidths $h^{cvi}_{n,\tau}$ and $h^{int}_{n,\tau}$ and the Wald tests are computed using $h^{cv}_{n,\tau}$, $h^{bdy}_{n,\tau}$, and $h^{ik}_{n,\tau}$.

**Rejection frequencies under null hypotheses.** The results are reported in Tables 2−4. Here, 'Wald', 'Wald Robust' and 'Wald Robust EC' correspond to tests that assume $\partial^2 Q(\tau|x_0^+)/\partial x^2 = \partial^2 Q(\tau|x_0^-)/\partial x^2$, tests that allow the above difference to vary freely across the quantiles, and tests that impose the constraint that this difference is constant across the quantiles, respectively.

The rejection frequencies of the score test are close to 10%. The maximum size distortion across all cases is only 0.023. This follows because the test does not require estimating the conditional density. The Wald tests show moderate over-rejections when $n = 500$, with the distortions falling between 0.007 and 0.098 in Tables 2 to 4. Their sizes improve uniformly as the sample size increases, with the maximum distortion reduced to 0.046 when $n = 1000$. We also repeat the above simulation using $n = 500$, with $\hat{f}_{Y|X}(\tau|x_0)$ replaced by its true value. The rejection frequencies of the three Wald tests are then between 0.090 and 0.110 for the two models across all bandwidth choices. This confirms that the over-rejection is indeed due to the estimation of the conditional densities. Finally, in all cases, the rejection frequencies of the three Wald tests are comparable. This suggests that the bias correction does not impose extra costs, in terms of size, in these two models. This is also found below when considering Models 3 and 4.

**Rejection frequencies under alternative hypotheses.** We consider alternatives where the treatment effects are significant and heterogeneous. Specifically, the conditional quantile functions $Q(\tau|x)$ for $x \geq 0$ equal (the functions for $x < 0$ remain the same as (21))

$$\begin{aligned}
\text{Model 1} \quad &: \quad 1 + x + (0.5 + 0.3x)\{Q_\varepsilon(\tau) + c_h\sigma_1 \arctan(4\pi\tau - 4)\}, \\
\text{Model 2} \quad &: \quad 0.5 + x + x^2 + \sin(\pi x - 1) + (x + 1.25)\{Q_\varepsilon(\tau) + c_h\sigma_2 \arctan(4\pi\tau - 4)\},
\end{aligned}$$

where "arctan" denotes the arctangent function, which is used here to generate effects that decrease smoothly from positive to negative levels as $\tau$ decreases. The sign change occurs at approximately the 32th percentile. We set $\sigma_1 = 1.43$ and $\sigma_2 = 0.57$, such that the maximum effect is equal to $c_h$ times the standard deviation of the conditional distribution at $x_0^-$. We consider $c_h = 0.3, 0.6, 1.0,$ and 2.0.

Table 5 reports the rejection frequencies when testing for the treatment significance hypothesis. There, the power of the score test and the 'Wald' test are comparable, with their differences mainly reflecting the variations in the bandwidth. To examine this further, we fix the bandwidths for the two models at 0.4 and apply them to both the score and the Wald tests. Then, for n=1000, the rejection frequencies under $c_h = 0.3, 0.6, 1.0,$ and 2.0 are as follows. For Model 1, the values are $(0.265, 0.663, 0.948, 1.000)$ for the score test and $(0.293, 0.648, 0.922, 1.000)$ for the Wald test. For Model 2, the values are $(0.386, 0.741, 0.975, 1.000)$ for the score test and $(0.336, 0.693, 0.941, 1.000)$ for the Wald test. The values are close to each other in both cases.

The effects of the bias correction on the Wald tests are clearly visible. For Model 1, the maximum power difference between 'Wald' and 'Wald Robust' is 0.280, while for Model 2 it is 0.312. When the equality constraint is imposed, the gap is reduced, but remains nonnegligible. The maximum differences are 0.192 and 0.244, respectively, for the two models. Table 6 reports the results for the treatment homogeneity hypothesis. There, the comparison between 'Wald' and 'Wald Robust' shows a pattern similar to that in Table 5, while the powers of 'Wald' and 'Wald Robust EC' are now close to each other owning to a cancellation effect. Table 7 reports results for the treatment unambiguity hypothesis. There, the overall pattern is similar to that in Table 5, although the power is lower. This is because the effects are positive, except when $\tau$ is below 0.32.

Therefore, whether to implement a bias correction can imply a substantial difference in power. For Models 1 and 2, the difference in power is a pure loss, because the second-order derivatives are continuous. We further study the trade-off using Models 3 and 4.

### 8.2 Models 3 and 4

We conduct the same analyses as those for Models 1 and 2, to further evaluate the effect of the discontinuity in the second order derivative on the three Wald tests.

**The selected bandwidth.** See Table 8. For Model 3, $h_{0.5}^{int}$ tends to be higher than $h_{0.5}^{cvi}$, while $h_{0.5}^{cv}$ tends to be the highest among the three selectors that operate under the alternative hypothesis. This pattern is the same as those in Models 1 and 2. For Model 4, the bandwidths are closer to each other and such a pattern is not visible. In both models, the $h_{0.5}^{ik}$ bandwidths still exhibit the least variation among the three bandwidths that operate under the alternative hypothesis.

**Rejection frequencies under null hypotheses.** Table 9 reports rejection frequencies of the tests with bias correction. As in Tables 2 to 4, the robust tests here show mild size distortions when $n = 500$, and the over-rejection is reduced when the sample size is increased. This suggests that the bias adjustments are effective in controlling the size, provided that the sample size is not small. Next, we turn to the non-robust tests.

Consider Model 3. For the treatment significance hypothesis, when $n = 500$, the rejection frequencies for the score test are 0.159 (using the bandwidth $h_{0.5}^{cvi}$) and 0.330 ($h_{0.5}^{int}$), and the values for the Wald tests are 0.308 ($h_{n,\tau}^{cv}$), 0.260 ($h_{n,\tau}^{bdy}$), and 0.282 ($h_{n,\tau}^{ik}$). The rejection frequencies deteriorate further when the sample size is increased to 2000, with the maximum values reaching 0.673 and 0.474 for the two tests. When testing for the treatment unambiguity hypothesis, size distortions are also present, except that the rejection frequencies are now below the nominal level. At $n = 500$, the rejection frequencies for the Wald tests are 0.024 ($h_{n,\tau}^{cv}$), 0.036 ($h_{n,\tau}^{bdy}$), and 0.044 ($h_{n,\tau}^{ik}$). They further decrease to 0.006 ($h_{n,\tau}^{cv}$), 0.007 ($h_{n,\tau}^{bdy}$), and 0.002 ($h_{n,\tau}^{ik}$) when $n$ is increased to 2000. Finally, when testing the treatment homogeneity hypothesis, the results are quite different. The rejection frequencies are now 0.134 ($h_{n,\tau}^{cv}$), 0.162 ($h_{n,\tau}^{bdy}$), and 0.158 ($h_{n,\tau}^{ik}$) when $n = 500$, and 0.110 ($h_{n,\tau}^{cv}$), 0.110 ($h_{n,\tau}^{bdy}$), and 0.104 ($h_{n,\tau}^{ik}$) when $n = 2000$. This reflects the same cancellation effect observed in Models 1 and 2. For Model 4, the pattern is similar to Model 3, although the magnitudes are much less pronounced. The details are omitted.

Therefore, ignoring discontinuity in the second order derivative can lead to substantial size distortions. This suggests that the robust tests can be valuable even though their power is lower.

**Rejection frequencies under alternative hypotheses.** We consider alternatives similar to those used for Models 1 and 2. Specifically, the conditional quantile functions $Q(\tau|x)$ for $x \geq 0$

equal (the functions for $x < 0$ remain the same as (21)):

$$\text{Model 3} \quad : \quad 0.48 + 0.84x - 3x^2 + 7.99x^3 - 9.01x^4 + 3.16x^5 + 0.1295\,\{Q_\varepsilon(\tau) + c_h\sigma_3 \arctan(4\pi\tau - 4)\},$$

$$\text{Model 4} \quad : \quad 4x^2 + 0.1295\,\{Q_\varepsilon(\tau) + c_h\sigma_4 \arctan(4\pi\tau - 4)\},$$

where $\sigma_3 = \sigma_4 = 5.55$, such that the maximum effect will equal $c_h$ times the standard deviation of the conditional distribution at $x_0^-$. We continue to consider $c_h = 0.3, 0.6, 1.0$, and 2.0. Table 10 reports the results for the robust tests. As in Tables 5 to 7, imposing the equality constraint brings mild power gains, relative to the unconstrained bias estimation when testing for the treatment significance and unambiguity hypotheses. The power gain is significantly higher when testing for the treatment homogeneity hypothesis. These two features apply to both Models 3 and 4.

### 8.3 Summary

This section has reported a relatively comprehensive study of the size and power properties of the proposed test statistics. The tests show desirable size and power properties. The score test is an attractive option for testing the treatment significance hypothesis when the sample size is small. The results also show that allowing the second-order derivative to change at the cut-off can make a substantial difference. Because these derivatives are often difficult to estimate, in practice, it can be useful to obtain results with and without bias correction, compare them, and then provide a full disclosure of the results. Finally, when applying the bias correction, imposing the equality constraint increases the power in all the specifications considered, sometimes substantially so. This method requires the assumption that *changes* in the second-order derivatives are constant across the quantiles. However, we feel that this can be a reasonable assumption in certain situations and that the resulting tests warrant consideration.

## 9 Conclusion

This study developed a framework for conducting uniform inference on quantile treatment effects for sharp RD designs. It proposes two sets of statistics that can be used to test hypotheses related to treatment significance, homogeneity, and unambiguity. It also suggests a procedure for constructing uniform confidence bands for quantile treatment effects. We conjecture that the methods can serve as useful complements to the standard ATE analysis to uncover and document potential treatment effect heterogeneity.

## References


Abadie, A., J. Angrist, and G. Imbens (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica 70*(1), pp. 91–117.

Angrist, J. D. and V. Lavy (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics 114*(2), 533–575.

Billingsley, P. (1986). *Probability and Measure* (Second ed.). New York: Wiley.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica 82*(6), 2295–2326.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2015). Optimal data-driven regression discontinuity plots. *Journal of the American Statistical Association 110*(512), 1753–1769.

Card, D. (1996). The effect of unions on the structure of wages: A longitudinal analysis. *Econometrica 64*(4), 957–979.

Card, D., R. Chetty, and A. Weber (2007). Cash-on-hand and competing models of intertemporal behavior: New evidence from the labor market. *The Quarterly Journal of Economics 122*(4), 1511–1560.

Card, D., D. Lee, Z. Pei, and A. Weber (2015). Nonlinear policy rules and the identification and estimation of causal effects in a generalized regression kink design. *Econometrica 83*(6), 2453–2483.

Chernozhukov, V. and I. Fernández-Val (2005). Subsampling inference on quantile regression processes. *Sankhya: The Indian Journal of Statistics 67*(2), pp. 253–276.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010). Quantile and probability curves without crossing. *Econometrica 78*(3), 1093–1125.

Chernozhukov, V. and C. Hansen (2005). An IV model of quantile treatment effects. *Econometrica 73*(1), pp. 245–261.

Chernozhukov, V., C. Hansen, and M. Jansson (2009). Finite sample inference for quantile regression models. *Journal of Econometrics 152*(2), pp. 93–103.

Chiang, H. D. and Y. Sasaki (2016). Quantile regression kink designs. *Working Paper, Johns Hopkins University*.

DiNardo, J. and D. S. Lee (2004). Economic impacts of new unionization on private sector employers: 1984-2001. *The Quarterly Journal of Economics* (4), 1383–1441.

Doksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The Annals of Statistics 2*(2), pp. 267–277.

Dong, Y. (2012). Jumpy or kinky? regression discontinuity without the discontinuity. *Working Paper, University of California at Irvine*.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica 75*(1), pp. 259–276.

Frandsen, B. R., M. Frolich, and B. Melly (2012). Quantile treatment effects in the regression discontinuity design. *Journal of Econometrics 168*(2), 382 – 395.

Freeman, R. B. (1980). Unionism and the dispersion of wages. *Industrial & Labor Relations Review 34*(1), 3–23.

Frolich, M. and B. Melly (2010). Quantile treatment effects in the regression discontinuity design: Process results and gini coefficient. *IZA Discussion Paper No. 4993*.

Grembi, V., T. Nannicini, and U. Troiano (2016). Do fiscal rules matter? *American Economic Journal: Applied Economics 8*(3), 1–30.

Guerre, E. and C. Sabbah (2012). Uniform bias study and bahadur representation for local polynomial estimators of the conditional quantile function. *Econometric Theory 28*(01), pp. 87–129.

Hahn, J. (1995). Bootstrapping quantile regression estimators. *Econometric Theory 11*(1), 105–121.

Hahn, J., P. Todd, and W. Van der Klaauw (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica 69*(1), 201–209.

Heckman, J. J., J. Smith, and N. Clements (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies 64*(4), pp. 487–535.

Imbens, G. and K. Kalyanaraman (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies 79*(3), 933–959.

Imbens, G. W. and T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics 142*(2), 615–635.

Koenker, R. and Z. Xiao (2002). Inference on the quantile regression process. *Econometrica 70*(4), pp. 1583–1612.

LaLonde, R. J. (1995). The promise of public sector-sponsored training programs. *The Journal of Economic Perspectives 9*(2), pp. 149–168.

Landais, C. (2015). Assessing the welfare effects of unemployment benefits using the regression kink design. *American Economic Journal: Economic Policy 7*(4), 243–78.

Lee, D. S. and D. Card (2008). Regression discontinuity inference with specification error. *Journal of Econometrics 142*(2), 655–674.

Lee, D. S. and T. Lemieux (2010). Regression discontinuity designs in economics. *Journal of Economic Literature 48*, 281–355.

Lee, D. S., E. Moretti, and M. J. Butler (2004). Do voters affect or elect policies? evidence from the us house. *The Quarterly Journal of Economics*, 807–859.

Lee, S., O. Linton, and Y. J. Whang (2009). Testing for stochastic monotonicity. *Econometrica 77*(2), 585–602.

Lehmann, E. L. (1975). *Nonparametrics: statistical methods based on ranks*. Holden-Day, San Francisco.

Ludwig, J. and D. L. Miller (2007). Does head start improve children's life chances? evidence from a regression discontinuity design. *The Quarterly Journal of Economics 122*(1), 159–208.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics 142*(2), 698–714.

Otsu, T., K.-L. Xu, and Y. Matsushita (2015). Empirical likelihood for regression discontinuity design. *Journal of Econometrics 186*(1), pp. 94–112.

Parzen, M. I., L. J. Wei, and Z. Ying (1994). A resampling method based on pivotal estimating functions. *Biometrika 81*(2), pp. 341–350.

Politis, D. N. and J. P. Romano (1994). The stationary bootstrap. *Journal of the American Statistical Association 89*(428), 1303–1313.

Porter, J. (2003). Estimation in the regression discontinuity model. *Working Paper, Department of Economics, University of Wisconsin at Madison*.

Qu, Z. and J. Yoon (2015). Nonparametric estimation and inference on conditional quantile processes. *Journal of Econometrics 185*(1), 1–19.

Shen, S. and X. Zhang (2016). Distributional tests for regression discontinuity: Theory and empirical examples. *Forthcoming in the Review of Economics and Statistics*.

Thistlethwaite, D. L. and D. T. Campbell (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology 51*(6), 309–317.

Van Der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression discontinuity approach. *International Economic Review 43*(4), 1249–1287.

Yu, K. and M. C. Jones (1998). Local linear quantile regression. *Journal of the American Statistical Association 93*(441), pp. 228–237.

Table 1: Summary Statistics of Bandwidths at the Median (Models 1 & 2).

| | **Model 1** | | | **Model 2** | | |
|---|---|---|---|---|---|---|
| | n=500 | n=1000 | n=2000 | n=500 | n=1000 | n=2000 |
| **Imposing $H_0$** | | | | | | |
| $h_{0.5}^{cvi}$ | 0.337 | 0.347 | 0.346 | 0.244 | 0.221 | 0.195 |
| | (0.138) | (0.137) | (0.136) | (0.094) | (0.076) | (0.061) |
| $h_{0.5}^{int}$ | 0.500 | 0.500 | 0.500 | 0.327 | 0.286 | 0.248 |
| | (0.000) | (0.000) | (0.000) | (0.018) | (0.012) | (0.007) |
| **Allowing $H_1$** | | | | | | |
| $h_{0.5}^{cv}$ | 0.470 | 0.475 | 0.479 | 0.455 | 0.443 | 0.414 |
| | (0.053) | (0.047) | (0.044) | (0.061) | (0.065) | (0.067) |
| $h_{0.5}^{bdy}$ | 0.428 | 0.433 | 0.435 | 0.396 | 0.383 | 0.347 |
| | (0.069) | (0.066) | (0.065) | (0.070) | (0.066) | (0.052) |
| $h_{0.5}^{ik}$ | 0.301 | 0.304 | 0.304 | 0.301 | 0.302 | 0.301 |
| | (0.018) | (0.016) | (0.016) | (0.018) | (0.017) | (0.018) |

Note. The values are averages over 2000 replications. Standard deviations are in the parentheses. $h_{n,\tau}^{cvi}$ and $h_{n,\tau}^{int}$ denote the cross validation and MSE optimal bandwidths treating the cutoff as an interior point. They will be used for the score test. $h_{n,\tau}^{cv}$, $h_{n,\tau}^{bdy}$ and $h_{n,\tau}^{ik}$ denote cross validation, MSE optimal and Imbens and Kalyanaraman (2012) bandwidths by treating the cutoff as a boundary point. They will be applied with the Wald tests.

Table 2: The Size of Tests for the Treatment Significance Hypothesis (Models 1 & 2).

| **Tests** | **Model 1** | | | **Model 2** | | |
|---|---|---|---|---|---|---|
| | n=500 | n=1000 | n=2000 | n=500 | n=1000 | n=2000 |
| **Score** | | | | | | |
| $h_{0.5}^{cvi}$ | 0.099 | 0.088 | 0.086 | 0.108 | 0.103 | 0.098 |
| $h_{0.5}^{int}$ | 0.116 | 0.108 | 0.107 | 0.123 | 0.116 | 0.114 |
| **Wald** | | | | | | |
| $h_{0.5}^{cv}$ | 0.160 | 0.110 | 0.094 | 0.172 | 0.146 | 0.126 |
| $h_{0.5}^{bdy}$ | 0.166 | 0.122 | 0.100 | 0.182 | 0.138 | 0.120 |
| $h_{0.5}^{ik}$ | 0.188 | 0.136 | 0.112 | 0.198 | 0.146 | 0.112 |
| **Wald Robust** | | | | | | |
| $h_{0.5}^{cv}$ | 0.132 | 0.099 | 0.086 | 0.132 | 0.102 | 0.081 |
| $h_{0.5}^{bdy}$ | 0.138 | 0.106 | 0.088 | 0.161 | 0.111 | 0.090 |
| $h_{0.5}^{ik}$ | 0.164 | 0.113 | 0.090 | 0.173 | 0.123 | 0.088 |
| **Wald Robust EC** | | | | | | |
| $h_{0.5}^{cv}$ | 0.150 | 0.109 | 0.102 | 0.154 | 0.115 | 0.097 |
| $h_{0.5}^{bdy}$ | 0.154 | 0.116 | 0.094 | 0.168 | 0.118 | 0.090 |
| $h_{0.5}^{ik}$ | 0.170 | 0.124 | 0.098 | 0.180 | 0.129 | 0.096 |

Note. The table reports rejection frequencies at the 10% nominal level over 2000 replications. "Wald", "Wald Robust" and "Wald Robust EC" denote tests that assume a continuous second order derivative at the cutoff, tests that allow a discontinuous second order derivative whose magnitude of discontinuity can vary freely across the quantiles, and tests that allow a discontinuous second order derivative whose magnitude of discontinuity remains constant across the quantiles.

Table 3: The Size of Tests for the Treatment Homogeneity Hypothesis (Models 1 & 2).

| **Tests** | **Model 1** | | | **Model 2** | | |
|---|---|---|---|---|---|---|
| | n=500 | n=1000 | n=2000 | n=500 | n=1000 | n=2000 |
| **Wald** | | | | | | |
| $h_{0.5}^{cv}$ | 0.138 | 0.107 | 0.091 | 0.142 | 0.114 | 0.102 |
| $h_{0.5}^{bdy}$ | 0.142 | 0.108 | 0.099 | 0.153 | 0.118 | 0.100 |
| $h_{0.5}^{ik}$ | 0.154 | 0.120 | 0.099 | 0.164 | 0.120 | 0.110 |
| **Wald Robust** | | | | | | |
| $h_{0.5}^{cv}$ | 0.123 | 0.090 | 0.082 | 0.126 | 0.096 | 0.083 |
| $h_{0.5}^{bdy}$ | 0.123 | 0.098 | 0.080 | 0.128 | 0.102 | 0.084 |
| $h_{0.5}^{ik}$ | 0.142 | 0.112 | 0.086 | 0.139 | 0.109 | 0.090 |
| **Wald Robust EC** | | | | | | |
| $h_{0.5}^{cv}$ | 0.122 | 0.098 | 0.082 | 0.126 | 0.104 | 0.092 |
| $h_{0.5}^{bdy}$ | 0.120 | 0.094 | 0.090 | 0.135 | 0.107 | 0.090 |
| $h_{0.5}^{ik}$ | 0.139 | 0.108 | 0.090 | 0.146 | 0.110 | 0.097 |

Note. See Table 2 for the definitions of the test statistics.

Table 4: The Size of Tests for the Treatment Unambiguity Hypothesis (Models 1 & 2).

| **Tests** | **Model 1** | | | **Model 2** | | |
|---|---|---|---|---|---|---|
| | n=500 | n=1000 | n=2000 | n=500 | n=1000 | n=2000 |
| **Wald** | | | | | | |
| $h_{0.5}^{cv}$ | 0.136 | 0.104 | 0.088 | 0.107 | 0.070 | 0.062 |
| $h_{0.5}^{bdy}$ | 0.140 | 0.105 | 0.098 | 0.126 | 0.086 | 0.082 |
| $h_{0.5}^{ik}$ | 0.162 | 0.110 | 0.102 | 0.156 | 0.103 | 0.094 |
| **Wald Robust** | | | | | | |
| $h_{0.5}^{cv}$ | 0.120 | 0.096 | 0.084 | 0.118 | 0.094 | 0.080 |
| $h_{0.5}^{bdy}$ | 0.131 | 0.098 | 0.086 | 0.133 | 0.096 | 0.084 |
| $h_{0.5}^{ik}$ | 0.136 | 0.101 | 0.087 | 0.140 | 0.096 | 0.086 |
| **Wald Robust EC** | | | | | | |
| $h_{0.5}^{cv}$ | 0.126 | 0.101 | 0.093 | 0.138 | 0.096 | 0.094 |
| $h_{0.5}^{bdy}$ | 0.132 | 0.098 | 0.094 | 0.142 | 0.102 | 0.096 |
| $h_{0.5}^{ik}$ | 0.146 | 0.103 | 0.094 | 0.146 | 0.108 | 0.092 |

Note. See Table 2 for the definitions of the test statistics.

Table 5: Power of Tests for the Treatment Significance Hypothesis (Models 1 & 2)

| **Tests** | **Model 1** | | | | **Model 2** | | | |
|---|---|---|---|---|---|---|---|---|
| | $c_h = 0.3$ | 0.6 | 1.0 | 2.0 | $c_h = 0.3$ | 0.6 | 1.0 | 2.0 |
| **Score** | | | | | | | | |
| $h_{0.5}^{cvi}$ | 0.196 | 0.469 | 0.690 | 0.951 | 0.179 | 0.395 | 0.658 | 0.947 |
| $h_{0.5}^{int}$ | 0.299 | 0.746 | 0.986 | 1.000 | 0.250 | 0.572 | 0.904 | 1.000 |
| **Wald** | | | | | | | | |
| $h_{0.5}^{cv}$ | 0.302 | 0.680 | 0.950 | 1.000 | 0.368 | 0.742 | 0.961 | 1.000 |
| $h_{0.5}^{bdy}$ | 0.300 | 0.662 | 0.940 | 1.000 | 0.332 | 0.678 | 0.930 | 1.000 |
| $h_{0.5}^{ik}$ | 0.262 | 0.560 | 0.857 | 1.000 | 0.292 | 0.586 | 0.872 | 1.000 |
| **Wald Robust** | | | | | | | | |
| $h_{0.5}^{cv}$ | 0.198 | 0.422 | 0.729 | 0.993 | 0.202 | 0.430 | 0.733 | 0.991 |
| $h_{0.5}^{bdy}$ | 0.186 | 0.424 | 0.705 | 0.991 | 0.196 | 0.420 | 0.686 | 0.990 |
| $h_{0.5}^{ik}$ | 0.179 | 0.354 | 0.577 | 0.964 | 0.187 | 0.373 | 0.608 | 0.970 |
| **Wald Robust EC** | | | | | | | | |
| $h_{0.5}^{cv}$ | 0.226 | 0.502 | 0.814 | 0.996 | 0.220 | 0.498 | 0.819 | 0.997 |
| $h_{0.5}^{bdy}$ | 0.214 | 0.484 | 0.789 | 0.996 | 0.222 | 0.474 | 0.760 | 0.996 |
| $h_{0.5}^{ik}$ | 0.198 | 0.414 | 0.665 | 0.988 | 0.210 | 0.428 | 0.683 | 0.990 |

Note. See Table 2 for the definitions of the test statistics. The bandwidths are defined as in Table 1. They are now computed from the data generated under the alternative hypothesis.

Table 6: Power of Tests for the Treatment Homogeneity Hypothesis (Models 1 & 2)

| **Tests** | **Model 1** | | | | **Model 2** | | | |
|---|---|---|---|---|---|---|---|---|
| | $c_h = 0.3$ | 0.6 | 1.0 | 2.0 | $c_h = 0.3$ | 0.6 | 1.0 | 2.0 |
| **Wald** | | | | | | | | |
| $h_{0.5}^{cv}$ | 0.382 | 0.782 | 0.968 | 0.998 | 0.359 | 0.752 | 0.960 | 0.997 |
| $h_{0.5}^{bdy}$ | 0.364 | 0.762 | 0.957 | 0.997 | 0.333 | 0.715 | 0.935 | 0.992 |
| $h_{0.5}^{ik}$ | 0.306 | 0.646 | 0.897 | 0.985 | 0.312 | 0.647 | 0.901 | 0.983 |
| **Wald Robust** | | | | | | | | |
| $h_{0.5}^{cv}$ | 0.221 | 0.515 | 0.806 | 0.978 | 0.225 | 0.509 | 0.798 | 0.974 |
| $h_{0.5}^{bdy}$ | 0.226 | 0.502 | 0.782 | 0.966 | 0.228 | 0.474 | 0.748 | 0.952 |
| $h_{0.5}^{ik}$ | 0.200 | 0.403 | 0.671 | 0.925 | 0.211 | 0.412 | 0.672 | 0.929 |
| **Wald Robust EC** | | | | | | | | |
| $h_{0.5}^{cv}$ | 0.364 | 0.762 | 0.962 | 0.998 | 0.339 | 0.734 | 0.955 | 0.996 |
| $h_{0.5}^{bdy}$ | 0.344 | 0.740 | 0.946 | 0.996 | 0.319 | 0.689 | 0.927 | 0.990 |
| $h_{0.5}^{ik}$ | 0.283 | 0.620 | 0.883 | 0.982 | 0.292 | 0.620 | 0.887 | 0.981 |

Note. See Table 2 for the definitions of the test statistics. The bandwidths are defined as in Table 1. They are now computed from the data generated under the alternative hypothesis.

Table 7: Power of Tests for the Treatment Unambiguity Hypothesis (Models 1 & 2)

| **Tests** | **Model 1** | | | | **Model 2** | | | |
|---|---|---|---|---|---|---|---|---|
| | $c_h = 0.3$ | 0.6 | 1.0 | 2.0 | $c_h = 0.3$ | 0.6 | 1.0 | 2.0 |
| **Wald** | | | | | | | | |
| $h_{0.5}^{cv}$ | 0.108 | 0.220 | 0.425 | 0.752 | 0.078 | 0.177 | 0.354 | 0.709 |
| $h_{0.5}^{bdy}$ | 0.106 | 0.222 | 0.406 | 0.733 | 0.097 | 0.190 | 0.354 | 0.677 |
| $h_{0.5}^{ik}$ | 0.110 | 0.199 | 0.336 | 0.635 | 0.109 | 0.203 | 0.323 | 0.637 |
| **Wald Robust** | | | | | | | | |
| $h_{0.5}^{cv}$ | 0.093 | 0.149 | 0.266 | 0.514 | 0.093 | 0.146 | 0.262 | 0.514 |
| $h_{0.5}^{bdy}$ | 0.091 | 0.160 | 0.254 | 0.504 | 0.096 | 0.156 | 0.258 | 0.486 |
| $h_{0.5}^{ik}$ | 0.090 | 0.152 | 0.222 | 0.428 | 0.091 | 0.150 | 0.229 | 0.438 |
| **Wald Robust EC** | | | | | | | | |
| $h_{0.5}^{cv}$ | 0.099 | 0.178 | 0.318 | 0.602 | 0.096 | 0.178 | 0.312 | 0.595 |
| $h_{0.5}^{bdy}$ | 0.094 | 0.173 | 0.308 | 0.592 | 0.107 | 0.174 | 0.292 | 0.577 |
| $h_{0.5}^{ik}$ | 0.095 | 0.162 | 0.254 | 0.500 | 0.104 | 0.166 | 0.254 | 0.516 |

Note. See Table 2 for the definitions of the test statistics. The bandwidths are defined as in Table 1. They are now computed from the data generated under the alternative hypothesis.

Table 8: Summary Statistics of Bandwidths at the Median (Models 3 & 4).

| | **Model 3** | | | **Model 4** | | |
|---|---|---|---|---|---|---|
| | n=500 | n=1000 | n=2000 | n=500 | n=1000 | n=2000 |
| **Imposing $H_0$** | | | | | | |
| $h_{0.5}^{cvi}$ | 0.258 | 0.211 | 0.164 | 0.129 | 0.116 | 0.107 |
| | (0.131) | (0.108) | (0.070) | (0.027) | (0.019) | (0.012) |
| $h_{0.5}^{int}$ | 0.447 | 0.411 | 0.360 | 0.126 | 0.109 | 0.100 |
| | (0.057) | (0.062) | (0.053) | (0.007) | (0.004) | (0.000) |
| **Allowing $H_1$** | | | | | | |
| $h_{0.5}^{cv}$ | 0.395 | 0.356 | 0.308 | 0.222 | 0.192 | 0.169 |
| | (0.082) | (0.081) | (0.070) | (0.038) | (0.030) | (0.024) |
| $h_{0.5}^{bdy}$ | 0.282 | 0.248 | 0.216 | 0.240 | 0.215 | 0.191 |
| | (0.053) | (0.031) | (0.019) | (0.038) | (0.025) | (0.016) |
| $h_{0.5}^{ik}$ | 0.283 | 0.284 | 0.285 | 0.278 | 0.274 | 0.267 |
| | (0.019) | (0.017) | (0.016) | (0.024) | (0.023) | (0.025) |

Note. See Table 1 for the definitions of the bandwidths.

Table 9: The Size of Robust Tests in Models 3 & 4.

| **Tests** | **Model 3** | | | **Model 4** | | |
|---|---|---|---|---|---|---|
| | n=500 | n=1000 | n=2000 | n=500 | n=1000 | n=2000 |
| **Treatment Significance:** | | | | | | |
| **Wald Robust** | | | | | | |
| $h_{0.5}^{cv}$ | 0.166 | 0.143 | 0.152 | 0.182 | 0.125 | 0.106 |
| $h_{0.5}^{bdy}$ | 0.157 | 0.117 | 0.092 | 0.181 | 0.115 | 0.098 |
| $h_{0.5}^{ik}$ | 0.168 | 0.125 | 0.110 | 0.164 | 0.106 | 0.081 |
| **Wald Robust EC** | | | | | | |
| $h_{0.5}^{cv}$ | 0.190 | 0.160 | 0.161 | 0.192 | 0.136 | 0.118 |
| $h_{0.5}^{bdy}$ | 0.170 | 0.130 | 0.102 | 0.194 | 0.128 | 0.108 |
| $h_{0.5}^{ik}$ | 0.180 | 0.139 | 0.125 | 0.176 | 0.118 | 0.092 |
| **Treatment Homogeneity:** | | | | | | |
| **Wald Robust** | | | | | | |
| $h_{0.5}^{cv}$ | 0.114 | 0.086 | 0.090 | 0.142 | 0.118 | 0.104 |
| $h_{0.5}^{bdy}$ | 0.128 | 0.094 | 0.093 | 0.145 | 0.111 | 0.096 |
| $h_{0.5}^{ik}$ | 0.124 | 0.092 | 0.093 | 0.130 | 0.100 | 0.080 |
| **Wald Robust EC** | | | | | | |
| $h_{0.5}^{cv}$ | 0.116 | 0.082 | 0.098 | 0.149 | 0.128 | 0.108 |
| $h_{0.5}^{bdy}$ | 0.144 | 0.100 | 0.103 | 0.152 | 0.118 | 0.100 |
| $h_{0.5}^{ik}$ | 0.138 | 0.099 | 0.096 | 0.138 | 0.114 | 0.102 |
| **Treatment Unambiguity:** | | | | | | |
| **Wald Robust** | | | | | | |
| $h_{0.5}^{cv}$ | 0.060 | 0.048 | 0.043 | 0.126 | 0.100 | 0.098 |
| $h_{0.5}^{bdy}$ | 0.087 | 0.078 | 0.064 | 0.119 | 0.089 | 0.088 |
| $h_{0.5}^{ik}$ | 0.084 | 0.073 | 0.048 | 0.110 | 0.074 | 0.061 |
| **Wald Robust EC** | | | | | | |
| $h_{0.5}^{cv}$ | 0.061 | 0.046 | 0.044 | 0.125 | 0.108 | 0.099 |
| $h_{0.5}^{bdy}$ | 0.086 | 0.078 | 0.068 | 0.124 | 0.103 | 0.092 |
| $h_{0.5}^{ik}$ | 0.098 | 0.072 | 0.052 | 0.107 | 0.074 | 0.070 |

Note. See Table 2 for the definitions of the test statistics.

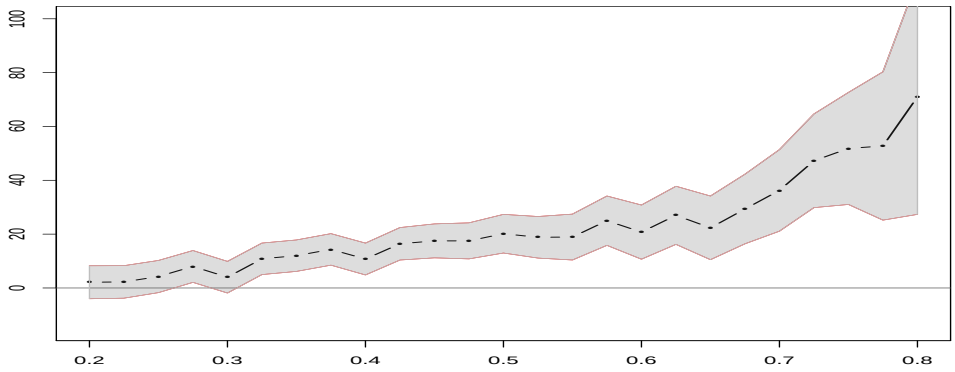
Table 10: Power of Robust Tests in Models 3 & 4

| **Tests** | **Model 3** | | | | **Model 4** | | | |
|---|---|---|---|---|---|---|---|---|
| | $c_h = 0.3$ | 0.6 | 1.0 | 2.0 | $c_h = 0.3$ | 0.6 | 1.0 | 2.0 |
| **Treatment Significance: Wald Robust** | | | | | | | | |
| $h_{0.5}^{cv}$ | 0.307 | 0.514 | 0.779 | 0.988 | 0.190 | 0.356 | 0.568 | 0.928 |
| $h_{0.5}^{bdy}$ | 0.211 | 0.388 | 0.622 | 0.975 | 0.202 | 0.370 | 0.599 | 0.951 |
| $h_{0.5}^{ik}$ | 0.230 | 0.421 | 0.671 | 0.976 | 0.207 | 0.412 | 0.657 | 0.973 |
| **Wald Robust EC** | | | | | | | | |
| $h_{0.5}^{cv}$ | 0.338 | 0.560 | 0.828 | 0.994 | 0.218 | 0.409 | 0.621 | 0.963 |
| $h_{0.5}^{bdy}$ | 0.237 | 0.450 | 0.698 | 0.988 | 0.236 | 0.426 | 0.655 | 0.979 |
| $h_{0.5}^{ik}$ | 0.259 | 0.474 | 0.739 | 0.992 | 0.245 | 0.458 | 0.725 | 0.990 |
| **Treatment Homogeneity: Wald Robust** | | | | | | | | |
| $h_{0.5}^{cv}$ | 0.220 | 0.480 | 0.754 | 0.955 | 0.190 | 0.361 | 0.591 | 0.879 |
| $h_{0.5}^{bdy}$ | 0.190 | 0.399 | 0.641 | 0.905 | 0.186 | 0.388 | 0.616 | 0.891 |
| $h_{0.5}^{ik}$ | 0.207 | 0.422 | 0.672 | 0.917 | 0.198 | 0.423 | 0.665 | 0.919 |
| **Wald Robust EC** | | | | | | | | |
| $h_{0.5}^{cv}$ | 0.352 | 0.743 | 0.931 | 0.990 | 0.257 | 0.556 | 0.814 | 0.957 |
| $h_{0.5}^{bdy}$ | 0.292 | 0.623 | 0.865 | 0.976 | 0.274 | 0.587 | 0.840 | 0.964 |
| $h_{0.5}^{ik}$ | 0.311 | 0.659 | 0.893 | 0.980 | 0.289 | 0.637 | 0.880 | 0.974 |
| **Treatment Unambiguity: Wald Robust** | | | | | | | | |
| $h_{0.5}^{cv}$ | 0.048 | 0.096 | 0.181 | 0.406 | 0.098 | 0.138 | 0.200 | 0.380 |
| $h_{0.5}^{bdy}$ | 0.078 | 0.123 | 0.198 | 0.387 | 0.096 | 0.133 | 0.208 | 0.383 |
| $h_{0.5}^{ik}$ | 0.064 | 0.113 | 0.201 | 0.389 | 0.075 | 0.122 | 0.195 | 0.398 |
| **Wald Robust EC** | | | | | | | | |
| $h_{0.5}^{cv}$ | 0.048 | 0.103 | 0.194 | 0.480 | 0.094 | 0.143 | 0.218 | 0.428 |
| $h_{0.5}^{bdy}$ | 0.076 | 0.128 | 0.219 | 0.450 | 0.098 | 0.140 | 0.226 | 0.430 |
| $h_{0.5}^{ik}$ | 0.065 | 0.121 | 0.212 | 0.454 | 0.080 | 0.126 | 0.218 | 0.456 |

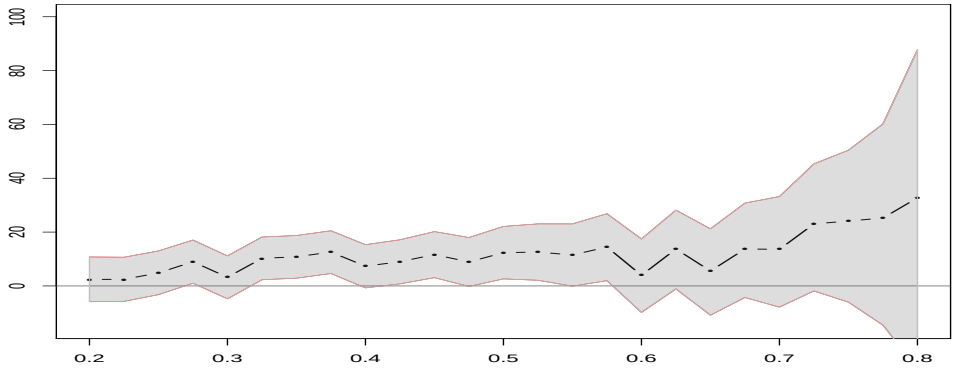
Note. See Table 2 for the definitions of the test statistics.

Figure 1: Quantile effects of Severance Pay on Unemployment Duration

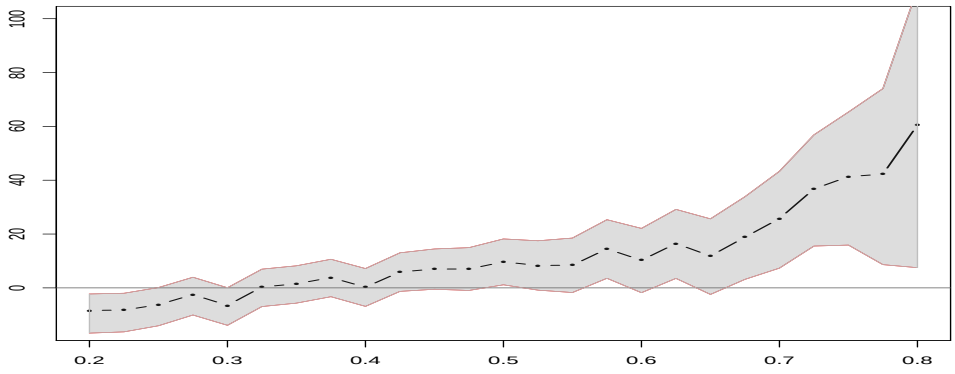
(a) Without bias adjustment




(b) With quantile-by-quantile bias estimation



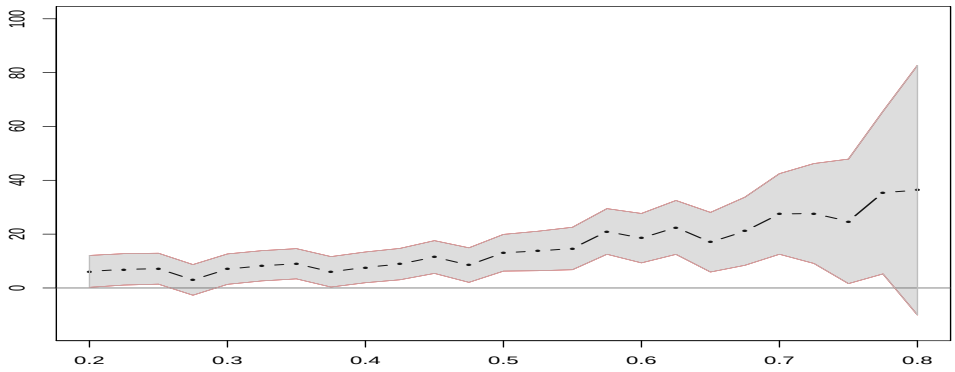

(c) With constrained bias estimation



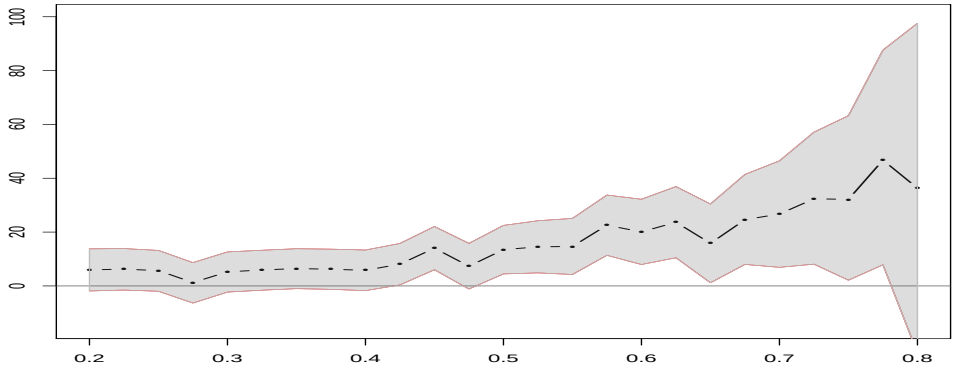

The results are produced using the restricted sample of Card, Chetty, and Weber (2007). The three figures contains point estimates (the dashed lines) and uniform confidence bands (the shaded areas) obtained from the following procedures: (a) assuming a continuous second order derivative at the cutoff, (b) allowing a discontinuous second order derivative whose magnitude of discontinuity can vary freely across the quantiles, and (c) allowing a discontinuous second order derivative whose magnitude of discontinuity remains constant across the quantiles. The bandwidth at the median is 4.5.

Figure 2: Quantile effects of Extended Benefits on Unemployment Duration

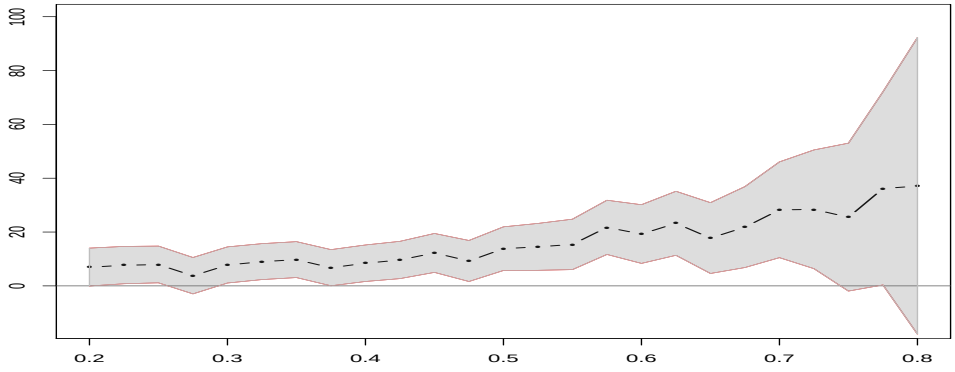
(a) Without bias adjustment




(b) With quantile-by-quantile bias estimation



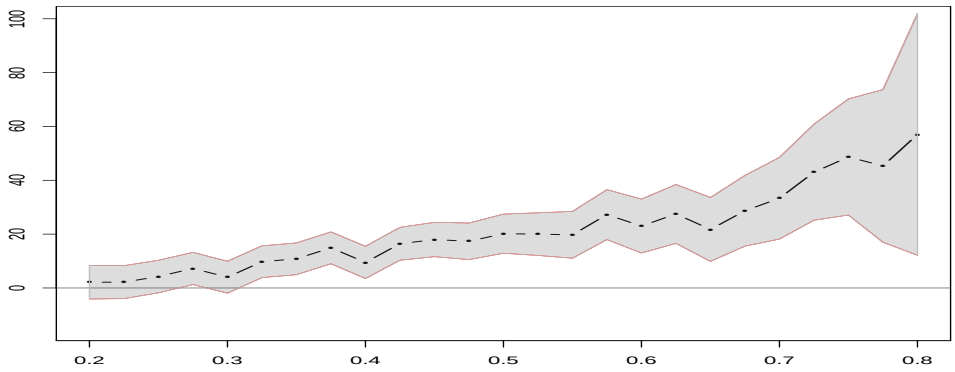

(c) With constrained bias estimation



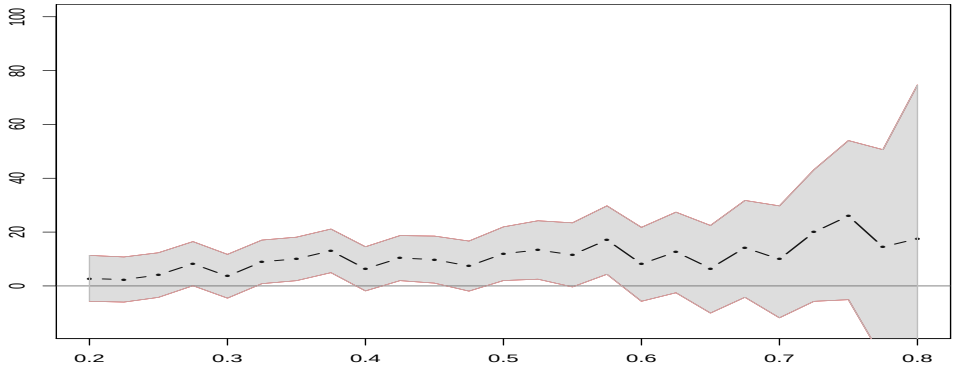

The results are produced using the restricted sample of Card, Chetty, and Weber (2007). The three figures contains point estimates (the dashed lines) and uniform confidence bands (the shaded areas) obtained from the following procedures: (a) assuming a continuous second order derivative at the cutoff, (b) allowing a discontinuous second order derivative whose magnitude of discontinuity can vary freely across the quantiles, and (c) allowing a discontinuous second order derivative whose magnitude of discontinuity remains constant across the quantiles. The bandwidth at the median is 4.5.

Figure 3: Quantile Effects of Severance Pay estimated using a subsample

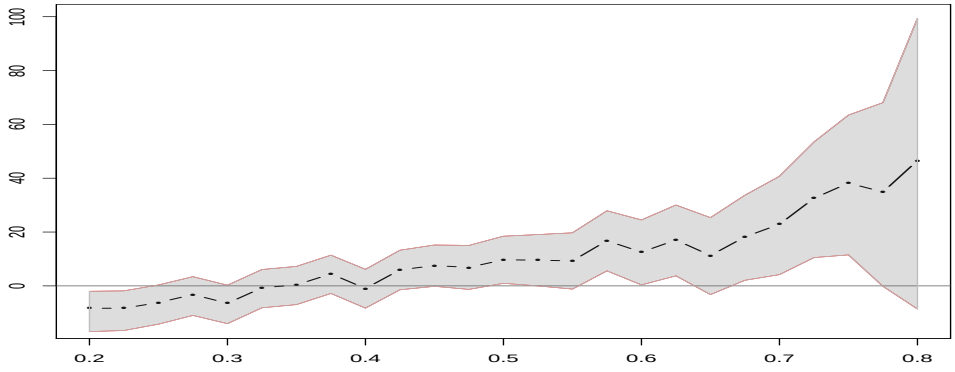
(a) Without bias estimation




(b) With quantile-by-quantile bias estimation



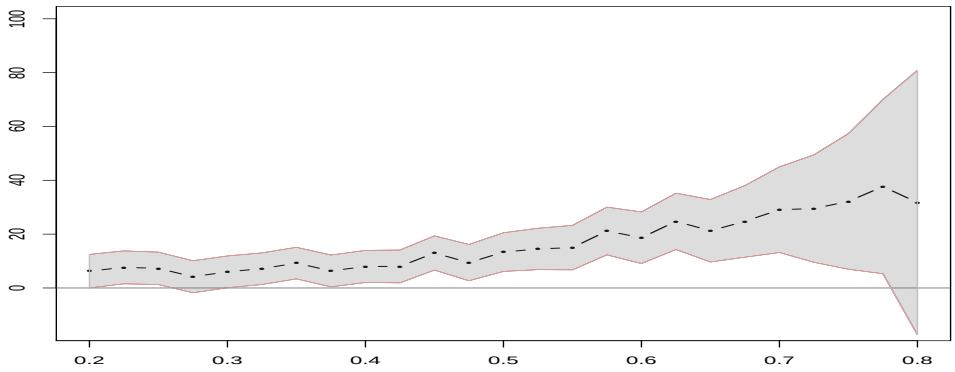

(c) With constrained bias estimation



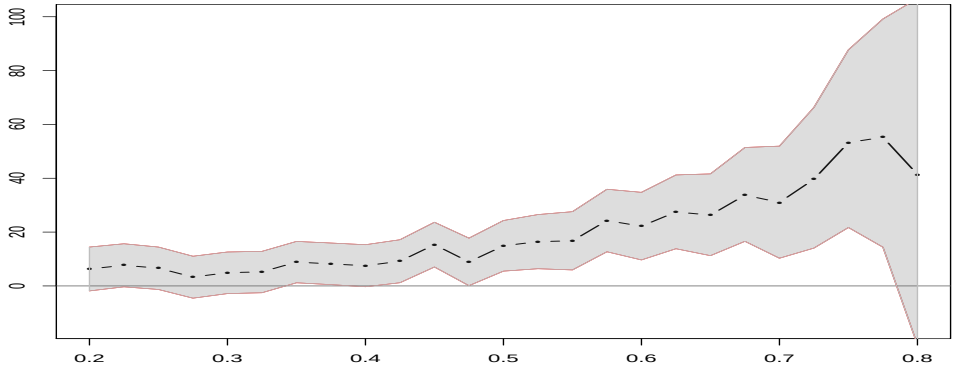

The results are produced using a subsample that include only workers who worked more than 4 months at a firm different from the one where they got laid off. Other specifications, including the bandwidth, are the same as Figure 1.

Figure 4: Quantile effects of Extended Benefits estimated using a subsample

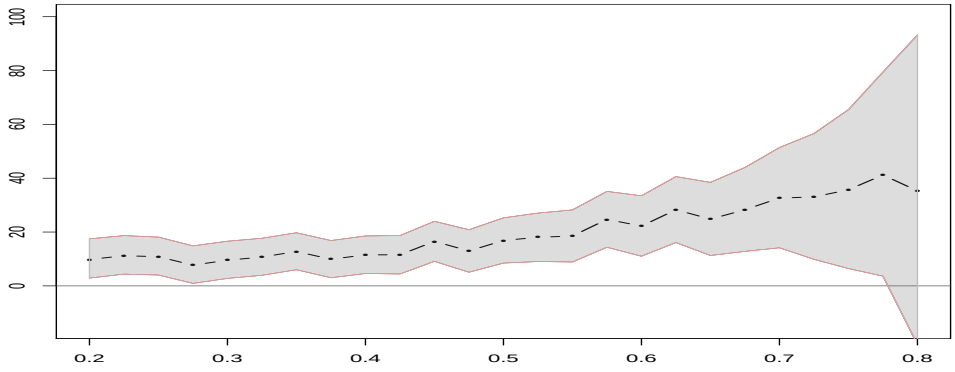
(a) Without bias estimation




(b) With quantile-by-quantile bias estimation



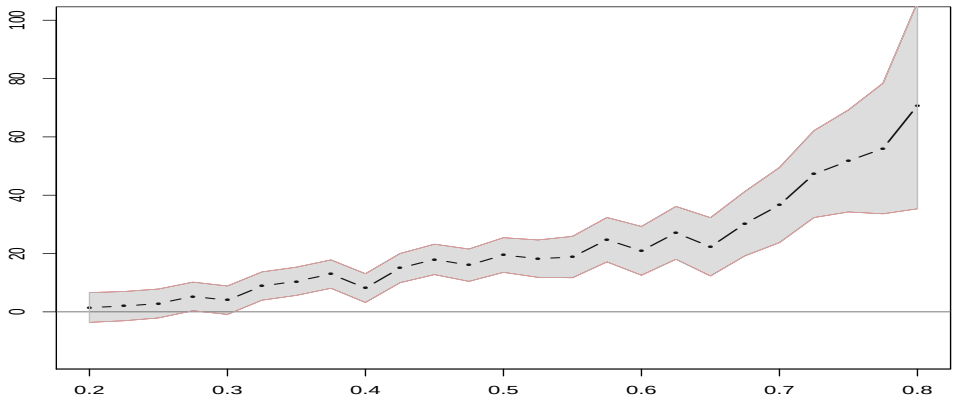

(c) With constrained bias estimation



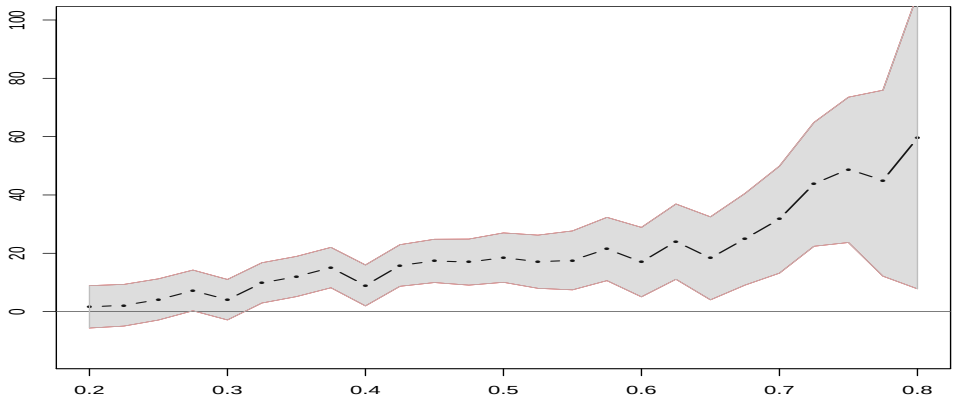

The results are produced using a subsample that include only workers who worked more than 4 months at a firm different from the one where they got laid off. Other specifications, including the bandwidth, are the same as Figure 2.

Figure 5: Quantile Effects of Severance Pay estimated using a different bandwidth

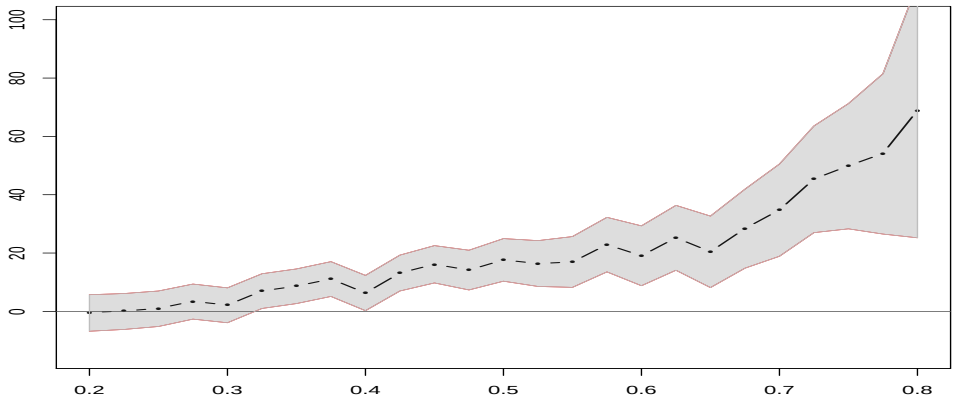
(a) Without bias estimation




(b) With quantile-by-quantile bias estimation



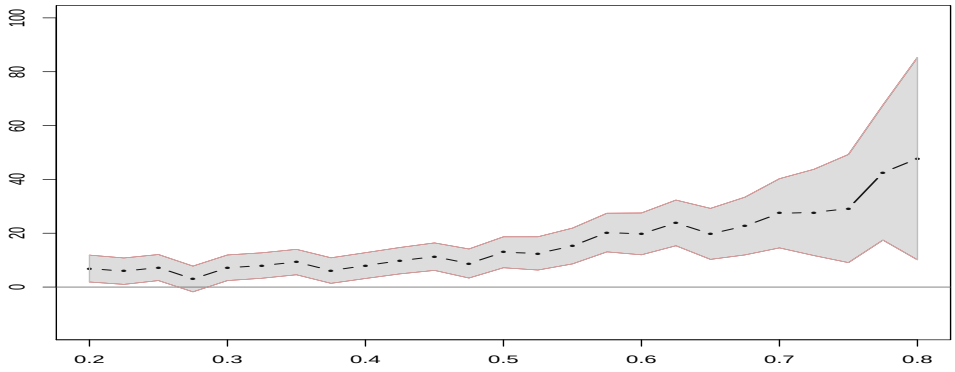

(c) With constrained bias estimation



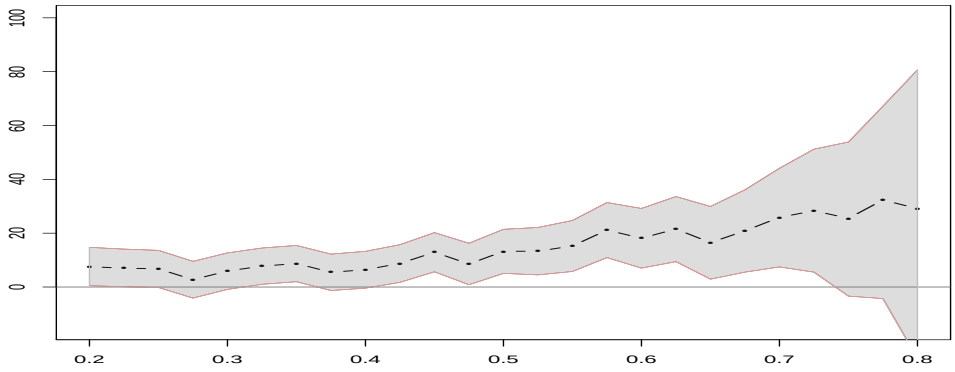

The bandwidth at the median equals 6.5. Other specifications are the same as in Figure 1.

Figure 6: Quantile effects of Extended Benefits estimated using a different bandwidth

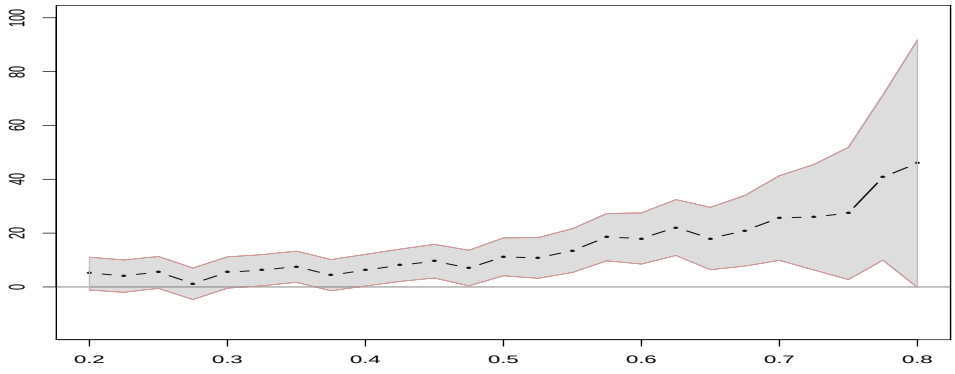
(a) Without bias estimation




(b) With quantile-by-quantile bias estimation




(c) With constrained bias estimation




The bandwidth at the median equals 6.5. Other specifications are the same as in Figure 2.

# Supplementary Appendix:
# Bandwidth Selection, Additional Results, and Proofs

This appendix is structured as follows. Section S.1 explains the bandwidth selectors used in the paper. Section S.2 discusses the hypothesis tests and the confidence bands related to $\delta^d(\tau)$ introduced in Section 2. Section S.3 develops Wald tests for the three hypotheses, assuming the second-order derivative of the conditional quantile function is continuous at the cut-off. Section S.4 provides a local asymptotic power analysis for the score and Wald tests. Section S.5 includes the proofs of the results given in the paper. Section S.6 reports additional simulation results, followed by several tables.

## S.1 Bandwidth selection

This section discusses how the five selectors determine $h_{n,0.5}$, the bandwidth at the median. Then, bandwidths at other quantiles are computed using the link function of Yu and Jones (1998):

$$h_{n,\tau} = \left\{2\tau(1-\tau)/[\pi\phi(\Phi^{-1}(\tau))^2]\right\}^{1/5} h_{n,0.5},$$

where $\phi(\cdot)$ and $\Phi^{-1}(\cdot)$ are the density and quantile functions of a standard normal distribution.

The first two bandwidth selectors are based on the leave-one-out cross validation. They are simple modifications of the methods given in Ludwig and Miller (2007) and Imbens and Lemieux (2008), originally designed for the average treatment effect. Specifically, for a given candidate bandwidth $h$, we estimate the conditional median at $x_i$ using the local linear regression, while leaving out $(x_i, y_i)$, and denote the estimate by $\hat{Q}_h(0.5|x_i)$. Then, we compute $CV(h) = k^{-1}\sum_{i=1}^{k} |y_i - \hat{Q}_h(0.5|x_i)|$ and determine the bandwidth as $h_{n,0.5} = \arg\min_h CV(h)$. Because the focus here is on the responses near $x_0$, observations far from $x_0$ are less relevant. Therefore, following Imbens and Lemieux (2008), we use only half the observations that are closest to $x_0$ as evaluation points. These two selectors differ in terms of whether $x_0$ is treated as an interior or a boundary point. The first selector treats $x_0$ as an interior point, that is, utilizing observations on both sides of $x_i$ when estimating the conditional median at $x_i$. This can be viewed as selecting the bandwidth by imposing the null hypothesis of no treatment effects. We denote the chosen bandwidth as $h_{n,\tau}^{cvi}$. The second selector treats $x_0$ as a boundary point. For example, if $x_i < x_0$, then only observations to the left of $x_i$ are used when estimating the conditional median at $x_i$. This can be viewed as selecting the bandwidth under the alternative hypothesis. We denote the chosen bandwidth by $h_{n,\tau}^{cv}$.

The third bandwidth selector uses the minimum MSE bandwidth formula of Qu and Yoon (2015), while treating $x_0$ as an interior point. This leads to

$$h_{n,0.5} = \left( \frac{\int_{-\infty}^{\infty} K(v)^2 \, dv}{4\mu_2^2 f_X(x_0) f_{Y|X}(0.5|x_0)^2 \left(\frac{\partial^2 Q(0.5|x_0)}{\partial x^2}\right)^2} \right)^{1/5} n^{-1/5}. \tag{S.1}$$

The densities $f_X(x_0)$ and $f_{Y|X}(0.5|x_0)$ are estimated as follows. The marginal density estimate is $\hat{f}_X(x_0) = (nh_x)^{-1}\sum_{i=1}^{n} K((x_i - x_0)/h_x)$, where $h_x$ is a bandwidth parameter. The conditional density estimate is

$$\hat{f}_{Y|X}(z|x_0) = \int \frac{1}{h_{yx}} K((z-y)/h_{yx}) \, d\hat{F}(y|x_0), \tag{S.2}$$

where $h_{yx}$ is another bandwidth parameter, and $\hat{F}(y|x_0) = \sup\{\tau \in (0,1) | \hat{Q}(\tau|x_0) \leq y\}$ is the inverse function of $\hat{Q}(\tau|x_0)$, which is the estimated conditional quantile with the cross validation bandwidth. To implement (S.2), we draw samples from $\hat{F}(y|x_0)$ and apply the kernel density estimator to the sample with kernel $K(\cdot)$ and bandwidth $h_{yx}$. In practice, the bandwidth $h_{yx}$ is

set to $2\widetilde{h}_{yx}$, with $\widetilde{h}_{yx}$ and $h_x$ being the bandwidths determined using Silverman's rule of thumb formula. Finally, the second-order derivative $\partial^2 Q(0.5|x_0)/\partial x^2$ is estimated using the local cubic median regression. Its bandwidth will be set to 1.0 throughout the simulations (note that, in this case, the support of $x$ is $[-1,1]$). We denote the resulting bandwidth by $h_{n,\tau}^{int}$.

The fourth bandwidth selector also uses the formula of Qu and Yoon (2015), but treats $x_0$ as a boundary point. This leads to the following bandwidth for $Q(0.5|x_0^+)$ :

$$h_{n,0.5}^{+}=\left(\frac{\iota_1' N^{-1} M N^{-1} \iota_1}{4 f_X\left(x_0\right) f_{Y|X}\left(0.5|x_0^+\right)^2\left(\frac{\partial^2 Q(0.5|x_0^+)}{\partial x^2}\right)^2\left(\iota_1' N^{-1} L\right)^2}\right)^{1/5} n^{-1/5}, \tag{S.3}$$

where $\iota_1=(1\ \ 0)'$, $N$ and $M$ are 2-by-2 matrices with the $(i,j)th$ elements given by $\int_0^\infty u^{i+j-2}K\left(u\right)du$ and $\int_0^\infty u^{(i+j-2)}K(u)^2du$ and $L=[\ \int_0^\infty uK\left(u\right)du\ \ \int_0^\infty u^2K\left(u\right)du]'$. In the implementation, the derivative $\partial^2 Q(0.5|x_0^+)/\partial x^2$ is estimated in the same way as for the third bandwidth selector, but now uses only observations on the right side of $x_0$. The MSE optimal bandwidth for estimating $Q(0.5|x_0^-)$ satisfies the same expression as (S.3), but with $\int_{-\infty}^0$ replacing $\int_0^\infty$ and $x_0^-$ replacing $x_0^+$. The one-sided conditional density $f_{Y|X}\left(0.5|x_0^+\right)$ uses the same formula as in (S.2), except that $\hat{F}(y|x_0)$ is replaced by $\hat{F}(y|x_0^+)$, which is computed by inverting $\hat{Q}(\tau|x_0^+)$. Finally, after obtaining estimates for $h_{n,0.5}^+$ and $h_{n,0.5}^-$, we use the smaller of the two to implement the tests. The motivation is that using a smaller bandwidth, although sacrificing some efficiency, will not erroneously introduce a large bias. We denote the bandwidth by $h_{n,\tau}^{bdy}$.

The fifth bandwidth selector is an adaptation of the Imbens and Kalyanaraman (2012) selector from the conditional mean to the conditional quantile setting. Instead of minimizing the MSEs associated with the conditional mean functions, Imbens and Kalyanaraman (2012) suggested minimizing the MSE associated with estimating their difference. For quantile treatment effects, calculations lead to the following bandwidth formula:

$$h_{n,0.5}=\left(\frac{\iota_1' N^{-1} M N^{-1} \iota_1\left(\frac{1}{f_{Y|X}\left(0.5|x_0^+\right)^2}+\frac{1}{f_{Y|X}\left(0.5|x_0^-\right)^2}\right)}{4\left(\iota_1' N^{-1} L\right)^2 f_X\left(x_0\right)\left(\left(\frac{\partial^2 Q(0.5|x_0^+)}{\partial x^2}-\frac{\partial^2 Q(0.5|x_0^-)}{\partial x^2}\right)^2+\left(r_-+r_+\right)\right)}\right)^{1/5} n^{-1/5}, \tag{S.4}$$

where $r_+$ and $r_-$ are regularization terms that equal three times the variances of $\partial^2\hat{Q}(0.5|x_0^+)/\partial x^2$ and $\partial^2\hat{Q}(0.5|x_0^-)/\partial x^2$, respectively. Their purpose is to stabilize the bandwidth in situations where the second-order derivatives do not change at $x_0$, or when they are imprecisely estimated. The quantities $r_-$ and $r_+$ depend on the following three factors for obtaining $\partial^2\hat{Q}(0.5|x_0^+)/\partial x^2$ and $\partial^2\hat{Q}(0.5|x_0^-)/\partial x^2$: the order of the local regressions, the kernel used, and the bandwidths. In simulations, we consider local quadratic regressions, the Epanechnikov kernel, and the bandwidth

$h_r = 0.5$. This leads to

$$r^+ = \frac{3}{nh_r^5} \frac{\iota_3' \bar{N}^{-1} \bar{M} \bar{N}^{-1} \iota_3'}{f_{Y|X}\left(0.5|x_0^+\right)^2 f_X\left(x_0\right)}, \tag{S.5}$$

where $\iota_3 = (0 \quad 0 \quad 1)'$, and $\bar{N}$ and $\bar{M}$ are 3-by-3 matrices, with the $(i,j)$-th elements given by $\int_0^\infty u^{i+j-2} K\left(u\right) du$ and $\int_0^\infty u^{i+j-2} K(u)^2 du$. The expression of $r^-$ is the same as (S.5), but with $\int_{-\infty}^0$ and $x_0^-$ replacing $\int_0^\infty$ and $x_0^+$, respectively. In the implementation, we rewrite (S.5) as

$$\frac{3}{nh_r^5} \frac{\iota_3' \left(f_X\left(x_0\right) \bar{N}\right)^{-1} \left[f_X\left(x_0\right) \bar{M}\right] \left(f_X\left(x_0\right) \bar{N}\right)^{-1} \iota_3'}{f_{Y|X}\left(0.5|x_0^+\right)^2}.$$

Then, the relevant quantities can be estimated using $(nh_r)^{-1} \sum_{i=1}^n \bar{z}_{i,\tau} \bar{z}_{i,\tau}' d_i K_{i,\tau} \to^p f_X\left(x_0\right) \bar{N}$ and $(nh_r)^{-1} \sum_{i=1}^n \bar{z}_{i,\tau} \bar{z}_{i,\tau}' d_i K_{i,\tau}^2 \to^p f_X\left(x_0\right) \bar{M}$. We denote the bandwidth by $h_{n,\tau}^{ik}$. We also experiment with estimating $\partial^2 Q(0.5|x_0^+)/\partial x^2$ and $\partial^2 Q(0.5|x_0^-)/\partial x^2$ using local cubic rather than quadratic regressions. Then, $(r_- + r_+)$ tends to take on substantially higher values than when using the local quadratic regression, often dominating the term $\left[\partial^2 Q(0.5|x_0^+)/\partial x^2 - \partial^2 Q(0.5|x_0^-)/\partial x^2\right]^2$. For this reason, we choose to use the quadratic regressions in the simulations and the empirical application.

Among the five selections, $h_{n,\tau}^{cvi}$ and $h_{n,\tau}^{int}$ are consistent with the principle of the score test because they impose the null hypothesis of no treatment effects. In addition, $h_{n,\tau}^{cv}$, $h_{n,\tau}^{bdy}$, and $h_{n,\tau}^{ik}$ are consistent with the principle of the Wald test. We use these pairings in the experimentations.

Finally, when implementing the tests with the bias estimation, we need additional bandwidth parameters for the regressions in (12). Motivated by the results in Calonico, Cattaneo, and Titiunik (2014), we let these equal the bandwidths for the local linear regressions (i.e., $b_{n,\tau} = h_{n,\tau}$ for all $\tau \in \mathcal{T}$) throughout the experimentations.

## S.2 Hypothesis tests and confidence bands related to $\delta^d(\tau)$

This subsection shows how to test the hypotheses and to construct uniform confidence bands for $\delta^d(\tau)$. For any of the three specifications of $\delta^d(\tau)$ in Section 2, let $\delta_1(\tau)$ denote the quantity inside the first parentheses and $\delta_2(\tau)$ be the quantity inside the second parentheses. Then,

$$\delta^d(\tau) = \delta_1(\tau) - \delta_2(\tau).$$

The three null hypotheses of interest are: (i) $H_0^1 : \delta^d(\tau) = 0$ for any $\tau \in \mathcal{T}$; (ii) $H_0^2 : \delta^d(\tau) = c$ for some $c \in R$ for all $\tau \in \mathcal{T}$; and (iii) $H_0^3 : \delta^d(\tau) \geq 0$ for all $\tau \in \mathcal{T}$. Let $n_1$ and $n_2$ be the sample sizes, and $h_{n_1,\tau}$ and $h_{n_2,\tau}$ be the bandwidths when estimating $\delta_1(\tau)$ and $\delta_2(\tau)$. Let $f_{Y|X,j}(\tau|x_0^+)$, $f_{Y|X,j}(\tau|x_0^-)$, $d_{\tau,j}^+$, and $d_{\tau,j}^-$ $(j = 1, 2)$ be the respective conditional densities and biases. (In the case with two cut-offs, interpret $f_{Y|X,1}(\tau|x_0^+)$ as $f_{Y|X}(\tau|x_0^+)$ and $f_{Y|X,2}(\tau|x_0^+)$ as $f_{Y|X}(\tau|x_1^+)$.) Define $f_{Y|X}(\tau|x_0) = (f_{Y|X,1}(\tau|x_0^+) + f_{Y|X,1}(\tau|x_0^-) + f_{Y|X,2}(\tau|x_0^+) + f_{Y|X,2}(\tau|x_0^-))/4$.

### S.2.1 Testing hypotheses assuming continuous second-order derivatives at the cutoffs

Define

$$W_n^d(\tau) = \sqrt{n_1 h_{n_1,\tau}}\hat{f}_{Y|X}(\tau|x_0)\left(\hat{\delta}_1(\tau) - \hat{\delta}_2(\tau)\right),$$

and consider the following test statistics.

$$\begin{aligned}
\text{For } H_0^1 \quad &: \quad WS_n^d(\mathcal{T}) = \sup_{\tau\in\mathcal{T}}\left|W_n^d(\tau)\right|, \\
\text{For } H_0^2 \quad &: \quad WH_n^d(\mathcal{T}) = \sup_{\tau\in\mathcal{T}}\left|W_n^d(\tau) - \frac{\sqrt{n_1 h_{n_1,\tau}}\hat{f}_{Y|X}(\tau|x_0)}{\int_{s\in\mathcal{T}}\sqrt{n_1 h_{n_1,s}}\hat{f}_{Y|X}(s|x_0)ds}\int_{\tau\in\mathcal{T}} W_n^d(\tau)d\tau\right|, \\
\text{For } H_0^3 \quad &: \quad WA_n^d(\mathcal{T}) = \sup_{\tau\in\mathcal{T}}\left|1\left(W_n^d(\tau)\le 0\right)W_n^d(\tau)\right|.
\end{aligned}$$

To present the limiting distributions of the test statistics, let $G_*^j(\tau)$ $(j=1,2)$ be two mutually independent Gaussian processes that are the limits of

$$\frac{1}{f_{X,j}(x_0)\sqrt{n_j h_{n_j,\tau}}}\sum_{i=1}^{n}\left(\tau - 1\left(u_i^0(\tau)\le 0\right)\right)\left\{\frac{f_{Y|X}(\tau|x_0)}{f_{Y|X,j}(\tau|x_0^+)}\Xi_{i,\tau,j}^+ d_i - \frac{f_{Y|X}(\tau|x_0)}{f_{Y|X,j}(\tau|x_0^-)}\Xi_{i,\tau,j}^-(1-d_i)\right\}K_{i,\tau,j},$$

where $\Xi_{i,\tau,j}^+$, $\Xi_{i,\tau,j}^-$, and $K_{i,\tau,j}$ are computed with bandwidth $h_{n_j,\tau}$. Let $\kappa(\tau)$ be the quantity defined in the Proposition below and $G_*^d(\tau) = G_*^1(\tau) - \kappa(\tau)G_*^2(\tau)$.

**Proposition 3** *Assume the conditions in Lemma 2 hold for j=1,2 with* $\partial^2 Q(\tau|x_0^+)/\partial x^2 = \partial^2 Q(\tau|x_0^-)/\partial x^2$ *for all* $\tau\in\mathcal{T}$. *Assume* $\sqrt{n_1 h_{n_1,\tau}}/\sqrt{n_2 h_{n_2,\tau}} \to \kappa(\tau) > 0$. *Then:*

1. *Under* $\delta^d(\tau) = 0$ *for all* $\tau\in\mathcal{T}$, $WS_n^d(\mathcal{T}) \Rightarrow \sup_{\tau\in\mathcal{T}}\left|G_*^d(\tau)\right|$.

2. *Under* $\delta^d(\tau) = \delta$ *for all* $\tau\in\mathcal{T}$ *for some* $\delta\in\mathbb{R}$,

$$WH_n^d(\mathcal{T}) \Rightarrow \sup_{\tau\in\mathcal{T}}\left|G_*^d(\tau) - \frac{\sqrt{n_1 h_{n_1,\tau}}f_{Y|X}(\tau|x_0)}{\int_{s\in\mathcal{T}}\sqrt{n_1 h_{n_1,s}}f_{Y|X}(s|x_0)ds}\int_{\tau} G_*^d(\tau)d\tau\right|.$$

3. *Under the least favorable null hypothesis of* $\delta^d(\tau) = 0$ *for all* $\tau\in\mathcal{T}$,

$$WA_n^d(\mathcal{T}) \Rightarrow \sup_{\tau\in\mathcal{T}}\left|1\left(G_*^d(\tau)\le 0\right)G_*^d(\tau)\right|.$$

### S.2.2 Testing hypotheses allowing discontinuous second-order derivatives at the cutoffs

Define

$$W_n^{R,d}(\tau) = \sqrt{n_1 h_{n_1,\tau}}\hat{f}_{Y|X}(\tau|x_0)\left(\hat{\delta}_1(\tau) - \hat{\delta}_2(\tau) - h_{n_1,\tau}^2(\hat{d}_{\tau,1}^+ - \hat{d}_{\tau,1}^-) + h_{n_2,\tau}^2(\hat{d}_{\tau,2}^+ - \hat{d}_{\tau,2}^-)\right),$$

where $\hat{d}^+_{\tau,j}$ and $\hat{d}^-_{\tau,j}$ are estimated with local quadratic regressions with bandwidth $h_{n_j,\tau}$ $(j = 1, 2)$. The tests are:

$$\begin{aligned}
\text{For } H_0^1 \quad &: \quad WS_n^{R,d}(\mathcal{T}) = \sup_{\tau\in\mathcal{T}} \left| W_n^{R,d}(\tau) \right|, \\
\text{For } H_0^2 \quad &: \quad WH_n^{R,d}(\mathcal{T}) = \sup_{\tau\in\mathcal{T}} \left| W_n^{R,d}(\tau) - \frac{\sqrt{n_1 h_{n_1,\tau}}\hat{f}_{Y|X}(\tau|x_0)}{\int_{s\in\mathcal{T}} \sqrt{n_1 h_{n_1,s}}\hat{f}_{Y|X}(s|x_0)ds} \int_{\tau\in\mathcal{T}} W_n^{R,d}(\tau)d\tau \right|, \\
\text{For } H_0^3 \quad &: \quad WA_n^{R,d}(\mathcal{T}) = \sup_{\tau\in\mathcal{T}} \left| 1\left(W_n^{R,d}(\tau) \leq 0\right) W_n^{R,d}(\tau) \right|.
\end{aligned}$$

Let $G_*^{R,j}(\tau)$ (j=1,2) be two independent copies of $G_*^R(\tau)$; see (15) in Section 5.2. Note that in $G_*^{R,j}(\tau)$, the bandwidth is equal to $h_{n_j,\tau}$. Define $G_*^{R,d}(\tau) = G_*^{R,1}(\tau) - \kappa(\tau)G_*^{R,2}(\tau)$.

**Proposition 4** *Let the conditions in Lemma 2 and Lemma 3 hold for j=1,2. Assume* $\sqrt{n_1 h_{n_1,\tau}}/\sqrt{n_2 h_{n_2,\tau}} \to \kappa(\tau) > 0$. *Then:*

1. *Under* $\delta(\tau) = 0$ *for all* $\tau \in \mathcal{T}$, $WS_n^{R,d}(\mathcal{T}) - \sup_{\tau\in\mathcal{T}} \left| G_*^{R,d}(\tau) \right| = o_p(1)$.

2. *Under* $\delta(\tau) = \delta$ *for all* $\tau \in \mathcal{T}$ *for some* $\delta \in \mathbb{R}$,

$$WH_n^{R,d}(\mathcal{T}) - \sup_{\tau\in\mathcal{T}} \left| G_*^{R,d}(\tau) - \frac{\sqrt{nh_{n,\tau}}\hat{f}_{Y|X}(\tau|x_0)}{\int_{s\in\mathcal{T}} \sqrt{nh_{n,s}}\hat{f}_{Y|X}(s|x_0)ds} \int_{\tau\in\mathcal{T}} G_*^{R,d}(\tau)d\tau \right| = o_p(1).$$

3. *Under the least favorable null hypothesis of* $\delta(\tau) = 0$ *for all* $\tau \in \mathcal{T}$,

$$WA_n^{R,d}(\mathcal{T}) - \sup_{\tau\in\mathcal{T}} \left| 1\left(G_*^{R,d}(\tau) \leq 0\right) G_*^{R,d}(\tau) \right| = o_p(1).$$

The relevant critical values can be obtained using simulations.

### S.2.3 Uniform confidence bands for $\delta^d(\tau)$

A uniform band can be obtained by inverting the Wald tests for the hypothesis $H_0^1$. In the case with continuous second-order derivatives, let $c_p^d$ be the $(1-p)$ percentile of the distribution of $\sup_{\tau\in\mathcal{T}} \left|G_*^d(\tau)\right|$. The confidence band for $\delta^d(\tau)$ is then given by

$$\hat{\delta}_1(\tau) - \hat{\delta}_2(\tau) \pm \frac{c_p^d}{\sqrt{n_1 h_{n_1,\tau}}\hat{f}_{Y|X}(\tau|x_0)}.$$

When discontinuous second-order derivatives are allowed, let $c_p^{R,d}$ be the $(1-p)$ percentile of the distribution of $\sup_{\tau\in\mathcal{T}} \left|G_*^{R,d}(\tau)\right|$. The uniform band is given by

$$\hat{\delta}_1(\tau) - \hat{\delta}_2(\tau) - h_{n_1,\tau}^2(\hat{d}^+_{\tau,1} - \hat{d}^-_{\tau,1}) + h_{n_2,\tau}^2(\hat{d}^+_{\tau,2} - \hat{d}^-_{\tau,2}) \pm \frac{c_p^{R,d}}{\sqrt{n_1 h_{n_1,\tau}}\hat{f}_{Y|X}(\tau|x_0)}.$$

## S.3 Wald tests assuming $\partial^2 Q(\tau|x_0^+)/\partial x^2 = \partial^2 Q(\tau|x_0^-)/\partial x^2$

Define

$$W_n(\tau) = \sqrt{nh_{n,\tau}}\hat{f}_{Y|X}(\tau|x_0)\hat{\delta}(\tau), \tag{S.6}$$

**Treatment significance.** This hypothesis can be tested using a Kolmogorov−Smirnov-type test:

$$WS_n(\mathcal{T}) = \sup_{\tau\in\mathcal{T}} |W_n(\tau)| .$$

**Treatment homogeneity.** This hypothesis can be tested by measuring the deviation of $W_n(\tau)$ from the average of $W_n(\tau)$ over $\mathcal{T}$ :

$$WH_n(\mathcal{T}) = \sup_{\tau\in\mathcal{T}} \left| W_n(\tau) - \frac{\sqrt{nh_{n,\tau}}\hat{f}_{Y|X}(\tau|x_0)}{\int_{s\in\mathcal{T}} \sqrt{nh_{n,s}}\hat{f}_{Y|X}(s|x_0)ds} \int_{\tau\in\mathcal{T}} W_n(\tau)d\tau \right| .$$

**Treatment unambiguity.** To test this hypothesis, we determine whether the treatment can be detrimental at some unknown quantiles, using

$$WA_n(\mathcal{T}) = \sup_{\tau\in\mathcal{T}} |1(W_n(\tau) \le 0) W_n(\tau)| .$$

Let $G_1(\tau)$ be a zero-mean continuous Gaussian process with a covariance function that satisfies

$$E[G_1(t)G_1(s)] = \frac{(t\wedge s - ts)}{f_X(x_0)(\mu_0^+\mu_2^+ - (\mu_1^+)^2)^2 (\kappa(t)\kappa(s))^{1/2}} \int_{-\infty}^{\infty} H(t)H(s)K\left(\frac{u}{\kappa(t)}\right) K\left(\frac{u}{\kappa(s)}\right) du, \tag{S.7}$$

where

$$H(\tau) = \frac{f_{Y|X}(\tau|x_0)}{f_{Y|X}(\tau|x_0^+)} \left(\mu_2^+ - \left(\frac{u}{\kappa(\tau)}\right)\mu_1^+\right) I(u\ge 0) - \frac{f_{Y|X}(\tau|x_0)}{f_{Y|X}(\tau|x_0^-)} \left(\mu_2^- - \left(\frac{u}{\kappa(\tau)}\right)\mu_1^-\right)(1 - I(u\ge 0)) .$$

**Proposition 5** *Assume the same conditions as in Lemma 2 hold, with* $\partial^2 Q(\tau|x_0^+)/\partial x^2 = \partial^2 Q(\tau|x_0^-)/\partial x^2$ *for all* $\tau\in\mathcal{T}$*. Then:*

1. *Under* $\delta(\tau) = 0$ *for all* $\tau\in\mathcal{T}$, $WS_n(\mathcal{T}) \Rightarrow \sup_{\tau\in\mathcal{T}} |G_1(\tau)|$ .
2. *Under* $\delta(\tau) = \delta$ *for all* $\tau\in\mathcal{T}$ *for some* $\delta\in\mathbb{R}$,

$$WH_n(\mathcal{T}) \Rightarrow \sup_{\tau\in\mathcal{T}} \left| G_1(\tau) - \frac{\sqrt{nh_{n,\tau}}f_{Y|X}(\tau|x_0)}{\int_{s\in\mathcal{T}} \sqrt{nh_{n,s}}f_{Y|X}(s|x_0)ds} \int_{\tau} G_1(\tau)d\tau \right| .$$

3. *Under the least favorable null hypothesis of* $\delta(\tau) = 0$ *for all* $\tau\in\mathcal{T}$ *(this is explained in the proof),*

$$WA_n(\mathcal{T}) \Rightarrow \sup_{\tau\in\mathcal{T}} |1(G_1(\tau)\le 0) G_1(\tau)| .$$

**Proof of Proposition 5.** In all three results, the effects are homogeneous across quantiles. This implies $f_{Y|X}(\tau|x_0^+) = f_{Y|X}(\tau|x_0^-)$ and, consequently,

$$W_{n,c}(\tau) = \frac{1}{f_X(x_0)\sqrt{nh_{n,\tau}}} \sum_{i=1}^{n} \left(\tau - I\left(u_i^0(\tau) \leq 0\right)\right) \frac{(2d_i - 1)\mu_2^+ - \left(\frac{x_i - x_0}{h_\tau}\right)\mu_1^+}{\mu_0^+\mu_2^+ - \left(\mu_1^+\right)^2} K_{i,\tau} + o_p(1).$$

The results then follow from the same arguments as in the proof of Proposition 1. For Case 3, the reason why "$\delta(\tau) = 0$ for all $\tau \in \mathcal{T}$ " is the least favorable null for the treatment unambiguity hypothesis is as follows. Define $M(\tau) = \sqrt{nh_{n,\tau}}\hat{f}_{Y|X}(\tau|x_0)\delta(\tau)$. Then, for any $\delta(\tau)$ satisfying the null hypothesis (i.e., $\delta(\tau) \geq 0$ for all $\tau \in \mathcal{T}$), the following two inequalities always hold because $M(\tau) \geq 0$:

$$\begin{aligned} |1\left(W_n(\tau) \leq 0\right) W_n(\tau)| &\leq |1\left(W_n(\tau) \leq 0\right)\left(W_n(\tau) - M(\tau)\right)| \\ &\leq |1\left(W_n(\tau) - M(\tau) \leq 0\right)\left(W_n(\tau) - M(\tau)\right)|. \end{aligned} \tag{S.8}$$

The term $W_n(\tau) - M(\tau)$ is equal to $W_{n,c}(\tau)$, defined in (14). Therefore, it satisfies the approximation given in Lemma 2 for any $\delta(\tau) \geq 0$. As a result, the supremum of (S.8) converges to $\sup_{\tau\in\mathcal{T}} |1\left(G_1(\tau) \leq 0\right) G_1(\tau)|$ under $\delta(\tau) \geq 0$. This shows that the test may be conservative if $\delta(\tau) \geq 0$ but $\delta(\tau)$ is not always zero. The test will not over-reject the null hypothesis. This completes the proof.

## S.4 Local asymptotic power analysis

The local alternatives are specified as follows. When testing for the treatment significance and unambiguity hypotheses, let

$$Q(\tau|x_0^+) - Q(\tau|x_0^-) = (nh_n)^{-1/2}\eta(\tau), \tag{S.9}$$

with $|\eta(\tau)| < +\infty$ for all $\tau \in \mathcal{T}$. When testing for the treatment homogeneity hypothesis, let

$$Q(\tau|x_0^+) - Q(\tau|x_0^-) = \delta + (nh_n)^{-1/2}\eta(\tau), \tag{S.10}$$

with $|\delta| < +\infty$ and $|\eta(\tau)| < +\infty$ for all $\tau \in \mathcal{T}$ . The bandwidth $h_n$ satisfies Assumption (5). The quantities $\delta$ and $\eta(\tau)$ are fixed as $n \to \infty$.

**Proposition 6** *Assume the same conditions as in Lemma 2 hold with $\partial^2 Q(\tau|x_0^+)/\partial x^2 = \partial^2 Q(\tau|x_0^-)/\partial x^2$ for all $\tau \in \mathcal{T}$. Let*

$$\widetilde{G}_1(\tau) = G_1(\tau) + c(\tau)^{1/2} f_{Y|X}(\tau|x_0)\eta(\tau),$$

*where $c(\tau)$ is defined in Assumption (5). Then:*

1. *Under (S.9),* $R_n(\mathcal{T}) \Rightarrow f_X(x_0)\left(\frac{\mu_0^+\mu_2^+-(\mu_1^+)^2}{2\mu_2^+}\right)\sup_{\tau\in\mathcal{T}}\left|\widetilde{G}_1(\tau)\right|$.

2. *Under (S.9),* $WS_n(\mathcal{T}) \Rightarrow \sup_{\tau\in\mathcal{T}}\left|\widetilde{G}_1(\tau)\right|$.

3. *Under (S.10),*

$$WH_n(\mathcal{T}) \Rightarrow \sup_{\tau\in\mathcal{T}}\left|\widetilde{G}_1(\tau) - \frac{\sqrt{nh_{n,\tau}}f_{Y|X}(\tau|x_0)}{\int_{s\in\mathcal{T}}\sqrt{nh_{n,s}}f_{Y|X}(s|x_0)ds}\int_\tau \widetilde{G}_1(\tau)d\tau\right|.$$

4. *Under (S.9) with* $\eta(\tau) < 0$ *for all* $\tau \in \mathcal{T}$,

$$WA_n(\mathcal{T}) \Rightarrow \sup_{\tau\in\mathcal{T}}\left|1\left(\widetilde{G}_1(\tau)\leq 0\right)\widetilde{G}_1(\tau)\right|.$$

The proof uses the same arguments as that of Lemmas 1 and 2. It is omitted. Interestingly, the first two results show that the score and Wald tests for the treatment significance hypothesis have the same local asymptotic power against the sequence (S.9). This follows after noting that, under the null hypothesis, their covariance functions satisfy

$$E(G(t)G(s)) = f_X(x_0)^2\left(\frac{\mu_0^+\mu_2^+ - (\mu_1^+)^2}{2\mu_2^+}\right)^2 E(G_1(t)G_1(s)).$$

In addition, the four results show that the tests can have nontrivial power against alternatives of order $(nh_n)^{-1/2}$. Finally, what matters for power is not only the difference $Q(\tau|x_0^+) - Q(\tau|x_0^-)$, but also the conditional density and the bandwidth. Everything else being equal, the power is higher if the departure from the null occurs in a dense region or at a place where the bandwidth is wider.

## S.5 Proofs of results in the paper

**Proof of Lemma 1.** For any $\alpha(\tau) \in \mathbb{R}$ and $\beta(\tau) \in \mathbb{R}$, define

$$\begin{aligned} e_i(\tau) &= Q(\tau|x_0) + (x_i - x_0)'\frac{\partial Q(\tau|x_0)}{\partial x} - Q(\tau|x_i), \\ \phi(\tau) &= \sqrt{nh_{n,\tau}}\begin{pmatrix} \alpha(\tau) - Q(\tau|x_0) \\ h_{n,\tau}\left(\beta(\tau) - \frac{\partial Q(\tau|x_0)}{\partial x}\right)\end{pmatrix} \quad \text{and} \quad z_{i,\tau}' = \left(1, \frac{x_i - x_0}{h_{n,\tau}}\right). \end{aligned}$$

Applying (7), we can write

$$u_i(\tau) = u_i^0(\tau) - e_i(\tau) - (nh_{n,\tau})^{-1/2}z_{i,\tau}'\phi(\tau).$$

Consequently,

$$R_n(\tau) = (nh_{n,\tau})^{-1/2}\sum_{i=1}^n\left\{\tau - 1[u_i^0(\tau) \leq (nh_{n,\tau})^{-1/2}z_{i,\tau}'\phi(\tau) + e_i(\tau)]\right\}d_iK_{i,\tau}.$$

To establish the asymptotic property of $R_n(\tau)$, we need to analyze both the effect of the parameter estimation and that of the local linear approximation. To this end, define

$$\begin{aligned} S_n\left(\tau, \phi(\tau), e_i\left(\tau\right)\right) &= (nh_{n,\tau})^{-1/2} \sum_{i=1}^{n} \left\{ P\left[ u_i^0(\tau) \leq (nh_{n,\tau})^{-1/2} z_{i,\tau}' \phi(\tau) + e_i(\tau) \middle| x_i \right] \right. \\ &\quad \left. - 1\left[ u_i^0(\tau) \leq (nh_{n,\tau})^{-1/2} z_{i,\tau}' \phi(\tau) + e_i(\tau) \right] \right\} d_i K_{i,\tau}. \end{aligned}$$

Let $\hat{\phi}(\tau)$ and $S_n(\tau, \hat{\phi}(\tau), e_i\left(\tau\right))$ equal $\phi(\tau)$ and $S_n\left(\tau, \phi(\tau), e_i\left(\tau\right)\right)$, but evaluated at $\hat{\alpha}(\tau)$ and $\hat{\beta}(\tau)$. Then, by adding and subtracting terms:

$$\begin{aligned} R_n(\tau) &= S_n\left(\tau, 0, 0\right) && \text{(Term 1)} \\ &\quad + \left\{ S_n\left(\tau, 0, e_i\left(\tau\right)\right) - S_n\left(\tau, 0, 0\right) \right\} && \text{(Term 2)} \\ &\quad + \left\{ S_n(\tau, \hat{\phi}(\tau), e_i\left(\tau\right)) - S_n\left(\tau, 0, e_i\left(\tau\right)\right) \right\} && \text{(Term 3)} \\ &\quad + (nh_{n,\tau})^{-1/2} \sum_{i=1}^{n} \left\{ \tau - P(\, u_i^0(\tau) \leq e_i\left(\tau\right) + (nh_{n,\tau})^{-1/2} z_{i,\tau}' \hat{\phi}\left(\tau\right) \middle| x_i) \right\} d_i K_{i,\tau} && \text{(Term 4).} \end{aligned}$$

Term 1 depends only on the data generating process. Term 2 depends on the remainder term from the local linear approximation. Terms 3 and 4 are affected by the parameter estimation. By Theorems 2 and 3 in Qu and Yoon (2015), the inequality constraints (or rearrangement) have no first-order effect on $\hat{\phi}(\tau)$. Therefore, we can treat $\hat{\phi}(\tau)$ as the estimator obtained by applying quantile-by-quantile local linear regressions without imposing any constraints (or rearrangement). Further, Qu and Yoon (2015, Step 1 in the proof of Theorem 1) show that $\Pr(\sup_{\tau \in \mathcal{T}} ||\widehat{\phi}(\tau)|| \leq \log^{1/2}(nh_{n,\tau})) \to 1$. Therefore, it suffices to consider the set $\{\phi(\tau) : ||\phi(\tau)|| \leq \log^{1/2}(nh_{n,\tau})\}$ when analyzing $R_n(\tau)$.

We study Terms 1 to 4 separately. By Lemma B.5 in Qu and Yoon (2015), $\sup_{\tau \in \mathcal{T}} ||(\text{Term } 2)|| = o_p\left(1\right)$ and $\sup_{\tau \in \mathcal{T}} ||(\text{Term } 3)|| = o_p\left(1\right)$.[1] Apply the mean value theorem:

$$\begin{aligned} (\text{Term } 4) &= -(nh_{n,\tau})^{-1/2} \sum_{i=1}^{n} f_{Y|X}(\widetilde{y}_i|x_i)\, e_i(\tau) d_i K_{i,\tau} - \left( \frac{1}{nh_{n,\tau}} \sum_{i=1}^{n} f_{Y|X}(\widetilde{y}_i|x_i) K_{i,\tau} d_i z_{i,\tau}' \right) \hat{\phi}(\tau) \\ &= A_{n,1}(\tau) + A_{n,2}(\tau) \hat{\phi}(\tau), \end{aligned}$$

where $\widetilde{y}_i$ lies between $Q(\tau|x_i)$ and $Q(\tau|x_i) + e_i(\tau) + (nh_{n,\tau}^d)^{-1/2} z_{i,\tau}' \hat{\phi}$. To analyze $A_{n,1}(\tau)$, note that

$$e_i(\tau) = -\frac{1}{2} h_{n,\tau}^2 \left( \frac{x_i - x_0}{h_{n,\tau}} \right)^2 \frac{\partial Q^2(\tau|x_0)}{\partial x^2} + o(h_{n,\tau}^2) \text{ uniformly over } \tau \in \mathcal{T}.$$

[1] Lemma B.5 focuses on Term 3 while establishing the order of Term 2 as an intermediate result; see the second term on the right-hand side of (B.8) on page 18.

Therefore, uniformly over $\tau \in \mathcal{T}$,

$$\begin{aligned} A_{n,1}(\tau) &= \frac{1}{2}(nh_{n,\tau}^5)^{1/2}\frac{\partial^2 Q(\tau|x_0)}{\partial x^2}\left\{\frac{1}{nh_{n,\tau}}\sum_{i=1}^n f_{Y|X}(\widetilde{y}_i|x_i)\left(\frac{x_i-x_0}{h_{n,\tau}}\right)^2 d_i K_{i,\tau}\right\}+o_p(1) \\ &= \frac{1}{2}(nh_{n,\tau}^5)^{1/2}\frac{\partial^2 Q(\tau|x_0)}{\partial x^2}f_{Y|X}(\tau|x_0)\left\{\frac{1}{nh_{n,\tau}}\sum_{i=1}^n \left(\frac{x_i-x_0}{h_{n,\tau}}\right)^2 d_i K_{i,\tau}\right\}+o_p(1) \\ &= \frac{1}{2}(nh_{n,\tau}^5)^{1/2}f_{Y|X}(\tau|x_0)f_X(x_0)\frac{\partial^2 Q(\tau|x_0)}{\partial x^2}\mu_2^+ + o_p(1), \end{aligned}$$

where the second equality holds because $x_i$ is in a vanishing neighborhood of $x_0$, and the third equality is by the uniform law of large numbers. By similar arguments, $A_{n,2}(\tau) = -f_{Y|X}(\tau|x_0)f_X(x_0)(\mu_0^+ \mu_1^+) + o_p(1)$. Finally, for $\hat{\phi}(\tau)$, apply Theorem 1 of Qu and Yoon (2015, see (A4) on page 15):

$$\begin{aligned} \hat{\phi}(\tau) &= \frac{1}{f_{Y|X}(\tau|x_0)\,f_X(x_0)}\begin{pmatrix}1 & 0\\ 0 & \mu_2^{-1}\end{pmatrix}\left\{(nh_{n,\tau})^{-1/2}\sum_{i=1}^n\left(\tau - 1\left(u_i^0(\tau)\le 0\right)\right)z_{i,\tau}K_{i,\tau}\right. \\ &\quad \left. + \frac{1}{2}\sqrt{nh_{n,\tau}^5}f_{Y|X}(\tau|x_0)\,f_X(x_0)\frac{\partial^2 Q(\tau|x_0)}{\partial x^2}\begin{pmatrix}\mu_2\\ \mu_3\end{pmatrix}\right\}+o_p(1). \end{aligned}$$

The results for $A_{n,1}(\tau), A_{n,2}(\tau)$, and $\hat{\phi}(\tau)$ jointly imply, uniformly over $\tau \in \mathcal{T}$ :

$$\begin{aligned} & A_{n,1}(\tau) + A_{n,2}(\tau)\hat{\phi}(\tau) \\ = \; & -(nh_{n,\tau})^{-1/2}\sum_{i=1}^n\left(\tau - 1(u_i^0(\tau)\le 0)\right)\left(\mu_0^+ + \frac{\mu_1^+}{\mu_2}\left(\frac{x_i-x_0}{h_{n,\tau}}\right)\right)K_{i,\tau} \\ - \; & \frac{1}{2}\sqrt{nh_{n,\tau}^5}f_{Y|X}(\tau|x_0)f_X(x_0)\frac{\partial Q^2(\tau|x_0)}{\partial x^2}\left(\mu_0^+\mu_2 + \frac{\mu_1^+\mu_3}{\mu_2} - \mu_2^+\right)+o_p(1). \end{aligned}$$

Combining the results for Terms 1 to 4, we have

$$\begin{aligned} R_n(\tau) &= (nh_{n,\tau})^{-1/2}\sum_{i=1}^n\left(\tau - 1(u_i^0(\tau)\le 0)\right)\left(d_i - \mu_0^+ - \frac{\mu_1^+}{\mu_2}\left(\frac{x_i-x_0}{h_{n,\tau}}\right)\right)K_{i,\tau} \\ &+ \frac{1}{2}\sqrt{nh_{n,\tau}^5}f_{Y|X}(\tau|x_0)f_X(x_0)\frac{\partial Q^2(\tau|x_0)}{\partial x^2}\left(\mu_2^+ - \mu_0^+\mu_2 - \frac{\mu_1^+\mu_3}{\mu_2}\right)+o_p(1). \end{aligned}$$

Because the kernel is symmetric, $\mu_3 = 0, \mu_0^+ = 1/2$ and $\mu_2^+ = 0.5\mu_2$. As a result, $\mu_2^+ - \mu_0^+\mu_2 - \frac{\mu_1^+\mu_3}{\mu_2} = 0$. This completes the proof.

**Proof of Proposition 1.** It suffices to consider the leading term on the right-hand side in Lemma 1. For any fixed $\tau \in \mathcal{T}$, this term satisfies the central limit theorem. Its stochastic equicontinuity with respect to $\tau$ is implied by Lemma B3 in Qu and Yoon (2015). The result follows because the supremum operator is continuous when taken over a compact set.

**Proof of Lemma 2.** By Theorem 1 in Qu and Yoon (2015),

$$\begin{aligned}
&\sqrt{nh_{n,\tau}}\left(\hat{Q}(\tau|x_0^+) - Q(\tau|x_0^+)\right) \qquad (S.11)\\
= \;& \sqrt{nh_{n,\tau}^5}d_\tau^+ + \frac{\iota_1' N^{-1}\left(nh_{n,\tau}\right)^{-1/2}\sum_{i=1}^n\left(\tau - 1(u_i^0(\tau)\le 0)\right)d_i z_{i,\tau}K_{i,\tau}}{f_X(x)\, f_{Y|X}\left(\tau|x_0^+\right)} + o_p(1),
\end{aligned}$$

where $\iota_1 = (1,0)', u\in\mathbb{R}, \bar{u} = (1,u')', d_\tau^+ = \frac{1}{2}\iota_1' N^{-1}\frac{\partial^2 Q(\tau|x_0^+)}{\partial x^2}\int_0^\infty u^2\bar{u}K(u)\,du$, and $N = \int_0^\infty \bar{u}\bar{u}'K(u)\,du$. Because $\iota_1' N^{-1} = (\mu_0^+\mu_2^+ - (\mu_1^+)^2)^{-1}[\mu_2^+ \quad -\mu_1^+]$, the first term on the right side of (S.11) is equal to

$$\frac{1}{2}\sqrt{nh_{n,\tau}^5}\frac{\partial Q^2(\tau|x_0^+)}{\partial x^2}\frac{(\mu_2^+)^2 - \mu_1^+\mu_3^+}{\mu_0^+\mu_2^+ - (\mu_1^+)^2}, \qquad (S.12)$$

while the second term is equal to

$$\frac{\left(nh_{n,\tau}\right)^{-1/2}\sum_{i=1}^n\left(\tau - 1(u_i^0(\tau)\le 0)\right)\Xi_{i,\tau}^+ d_i K_{i,\tau}}{f_X(x)\, f_{Y|X}\left(\tau|x_0^+\right)}. \qquad (S.13)$$

Applying the same arguments to $\hat{Q}(\tau|x_0^-)$, we have

$$\begin{aligned}
\sqrt{nh_{n,\tau}}\left(\hat{Q}(\tau|x_0^-) - Q(\tau|x_0^-)\right) &= \frac{1}{2}\sqrt{nh_{n,\tau}^5}\frac{\partial Q^2(\tau|x_0^-)}{\partial x^2}\frac{(\mu_2^-)^2 - \mu_1^-\mu_3^-}{\mu_0^-\mu_2^- - (\mu_1^-)^2} \qquad (S.14)\\
&\quad + \frac{\left(nh_{n,\tau}\right)^{-1/2}\sum_{i=1}^n\left(\tau - 1(u_i^0(\tau)\le 0)\right)\Xi_{i,\tau}^-(1-d_i)K_{i,\tau}}{f_X(x)\, f_{Y|X}\left(\tau|x_0^-\right)} + o_p(1).
\end{aligned}$$

Combining (S.12), (S.13) and (S.14) leads to the desired result.

**Proof of Lemma 3.** It suffices to show that $\sqrt{nb_{n,\tau}^5}\left(\hat{d}_\tau^+ - d_\tau^+\right) = D_2^+(\tau) + o_p(1)$ uniformly over $\tau\in\mathcal{T}$. The proof is similar to that of Lemma 1. To reflect this, we define the notation analogously. Let

$$\begin{aligned}
\bar{u}_i(\tau) &= y_i - \alpha^+(\tau) - (x_i - x)\beta^+(\tau) - (x_i - x)^2\gamma^+(\tau),\\
\bar{e}_i(\tau) &= \left[Q(\tau|x_0^+) + (x_i - x_0)\frac{\partial Q(\tau|x_0^+)}{\partial x} - (x_i - x_0)^2\frac{1}{2}\frac{\partial Q^2(\tau|x_0^+)}{\partial x^2}\right] - Q(\tau|x_i),\\
\bar{\phi}(\tau) &= \sqrt{nb_{n,\tau}}\begin{pmatrix} \alpha^+(\tau) - Q(\tau|x_0^+)\\ b_{n,\tau}(\beta^+(\tau) - \frac{\partial Q(\tau|x_0^+)}{\partial x})\\ b_{n,\tau}^2(\gamma^+(\tau) - \frac{1}{2}\frac{\partial Q^2(\tau|x_0^+)}{\partial x^2})\end{pmatrix}, \text{ and } \bar{z}_{i,\tau}' = \begin{bmatrix} 1\\ (x_i - x_0)/b_{n,\tau}\\ (x_i - x_0)^2/b_{n,\tau}^2\end{bmatrix}.
\end{aligned}$$

Define

$$\begin{aligned}
\bar{S}_n\left(\tau, \bar{\phi}(\tau), \bar{e}_i(\tau)\right) &= (nb_n)^{-1/2}\sum_{i=1}^n\left\{P\left[u_i^0(\tau) \le (nb_{n,\tau})^{-1/2}\bar{z}_{i,\tau}'\bar{\phi}(\tau) + \bar{e}_i(\tau)\middle| x_i\right]\right.\\
&\quad \left. - 1\left[u_i^0(\tau)\le (nb_{n,\tau})^{-1/2}\bar{z}_{i,\tau}'\bar{\phi}(\tau) + \bar{e}_i(\tau)\right]\right\} d_i\bar{K}_{i,\tau}.
\end{aligned}$$

Note that $\bar{e}_i(\tau)$ satisfies

$$\bar{e}_i(\tau) = -\frac{1}{3!}\left(\frac{x_i - x}{b_{n,\tau}}\right)^3 \frac{\partial^3 Q(\tau|x_0^+)}{\partial x^3} b_{n,\tau}^3 + o\left(b_{n,\tau}^3\right). \tag{S.15}$$

Let $\widehat{\bar{\phi}}(\tau)$ equal $\bar{\phi}(\tau)$, but with $\alpha^+(\tau), \beta^+(\tau)$, and $\gamma^+(\tau)$ replaced by the estimates from the local quadratic regression. Applying the subgradient condition, we have

$$(nb_n)^{-1/2}\sum_{i=1}^n 1\left[u_i^0(\tau) \le (nb_{n,\tau})^{-1/2}\bar{z}_{i,\tau}'\widehat{\bar{\phi}}(\tau) + e_i(\tau)\right]\Big\} d_i\bar{K}_{i,\tau} = o_p(1)$$

uniformly over $\tau \in \mathcal{T}$. This implies

$$\begin{aligned}
&\{\bar{S}_n(\tau, \widehat{\bar{\phi}}(\tau), \bar{e}_i(\tau)) - \bar{S}_n(\tau, 0, \bar{e}_i(\tau))\} \\
&+ \left\{\bar{S}_n(\tau, 0, \bar{e}_i(\tau)) - \bar{S}_n(\tau, 0, 0)\right\} + \bar{S}_n(\tau, 0, 0) \\
&+(nb_n)^{-1/2}\sum_{i=1}^n \left\{\tau - P(u_i^0(\tau) \le \bar{e}_i(\tau) + (nb_{n,\tau})^{-1/2}\bar{z}_{i,\tau}'\widehat{\bar{\phi}}(\tau)\Big| x_i)\right\} d_i\bar{z}_{i,\tau}\bar{K}_{i,\tau} = o_p(1).
\end{aligned} \tag{S.16}$$

The terms in the first two curly brackets are $o_p(1)$ uniformly. Applying a first-order Taylor expansion to the last term, we obtain:

$$-(nb_n)^{-1/2}\sum_{i=1}^n f_{Y|X}(\widetilde{y}_i|x_i)\,\bar{e}_i(\tau)\,d_i\bar{z}_{i,\tau}\bar{K}_{i,\tau} - (nb_n)^{-1/2}(nb_{n,\tau})^{-1/2}\left(\sum_{i=1}^n f_{Y|X}(\widetilde{y}_i|x_i)\,d_iK_{i,\tau}\bar{z}_{i,\tau}\bar{z}_{i,\tau}'\right)\widehat{\bar{\phi}}(\tau),$$

where $\widetilde{y}_i$ lies between $Q(\tau|x_i)$ and $Q(\tau|x_i) + e_i(\tau) + (nb_{n,\tau})^{-1/2}d_i\bar{z}_{i,\tau}'\widehat{\bar{\phi}}(\tau)$. As a result,

$$\begin{aligned}
\widehat{\bar{\phi}}(\tau) \;=\; & \left(f_{Y|X}\left(\tau|x_0^+\right) f_X(x_0)\,\bar{N}^+\right)^{-1} \\
& \left\{\left(\frac{b_n}{b_{n,\tau}}\right)^{1/2}\bar{S}_n(\tau, 0, 0) - (nb_{n,\tau})^{-1/2} f_{Y|X}\left(\tau|x_0^+\right)\sum_{i=1}^n \bar{e}_i(\tau)\,d_i\bar{z}_{i,\tau}\bar{K}_{i,\tau}\right\} + o_p(1).
\end{aligned}$$

The term involving $\bar{e}_i(\tau)$ is negligible because $nb_{n,\tau}^7 = o(1)$. Therefore,

$$\widehat{\bar{\phi}}(\tau) = \left(f_{Y|X}\left(\tau|x_0^+\right) f_X(x_0)\,\bar{N}^+\right)^{-1}(b_n/b_{n,\tau})^{1/2}\,\bar{S}_n(\tau, 0, 0) + o_p(1).$$

Multiplying both sides by $\Gamma\iota_3'$ leads to the desired result.

**Proof of Proposition 2.** By Lemma 2 and Lemma 3,

$$W_n^R(\tau) = G_*^R(\tau) + o_p(1)$$

uniformly over $\mathcal{T}$. Then, the proof can be completed by applying the same arguments as those in the proof of Proposition 1.

**Proof of the validity of the procedure in Remark 2.** The proof is given in four steps, using similar arguments as in Politis and Romano (1994) and Hahn (1995). It allows for two

possible bandwidth sequences: (i) $b_{n,\tau}/h_{n,\tau} \to \infty$ for all $\tau \in \mathcal{T}$. This corresponds to using a larger bandwidth for the local quadratic regression than the local linear regression. (ii) $b_{n,\tau}/h_{n,\tau} = r(\tau)$ with $0 < r(\tau) < \infty$ for all $\tau \in \mathcal{T}$. This corresponds to using a comparable bandwidth for the local quadratic relative to the local linear regression. Note that under the three null hypotheses, $f_{Y|X}\left(\tau|x_0^+\right) = f_{Y|X}\left(\tau|x_0^+\right) = f_{Y|X}\left(\tau|x_0\right)$.

*Step 1. We verify that $G_*^R(\tau)$ converges weakly to a continuous Gaussian process over $\mathcal{T}$ under both bandwidth sequences.*

Under bandwidth sequence (i), $(\sqrt{nh_{n,\tau}^5}/\sqrt{nb_{n,\tau}^5})\left(D_2^+(\tau) - D_2^-(\tau)\right)$ converges weakly to 0 over $\tau \in \mathcal{T}$. Therefore,

$$G_*^R(\tau) = \hat{f}_{Y|X}(\tau|x_0)\{D_1^+(\tau) - D_1^-(\tau)\} + o_p(1) \Rightarrow G_1(\tau) \text{ over } \tau \in \mathcal{T},$$

where $G_1(\tau)$ is the Gaussian process defined in (S.7).

Under bandwidth sequence (ii), the limit of $\hat{f}_{Y|X}(\tau|x_0)(D_1^+(\tau) - D_1^-(\tau))$ is still given by $G_1(\tau)$. The limit of $\hat{f}_{Y|X}(\tau|x_0)(\sqrt{nh_{n,\tau}^5}/\sqrt{nb_{n,\tau}^5})(D_2^+(\tau) - D_2^-(\tau))$, denoted by $G_2(\tau)$, is a zero-mean Gaussian process with covariance function

$$E\left[G_2(t)G_2(s)\right] = \frac{(t \wedge s - ts)\Gamma^2}{f_X(x_0)\left(\kappa(t)\kappa(s)\right)^{1/2}\left(r(t)r(s)\right)^{5/2}} \int_{-\infty}^{\infty} H_2(t)H_2(s)K\left(\frac{u}{\kappa(t)}\right)K\left(\frac{u}{\kappa(s)}\right)du,$$

where

$$g(\tau)' = \left[1 \;\; \frac{u}{\kappa(\tau)} \;\; \frac{u^2}{\kappa(\tau)^2}\right], r(\tau) = b_{n,\tau}/h_{n,\tau}, \kappa(\tau) = b_{n,\tau}/b_{n,1/2},$$

and

$$H_2(\tau) = \iota_3'\left\{\frac{f_{Y|X}(\tau|x_0)}{f_{Y|X}\left(\tau|x_0^+\right)}(\bar{N}^+)^{-1}I(u \geq 0) - \frac{f_{Y|X}(\tau|x_0)}{f_{Y|X}\left(\tau|x_0^-\right)}(\bar{N}^-)^{-1}\left(1 - I(u \geq 0)\right)\right\}g(\tau).$$

Therefore,

$$G_*^R(\tau) \Rightarrow G_1(\tau) - G_2(\tau) \text{ over } \tau \in \mathcal{T}.$$

*Step 2. Denote the simulated version of $G_*^R(\tau)$ by $\hat{S}_*^R(\tau)$. We prove that, if some convergences hold, then $\hat{S}_*^R(\tau)$ converges weakly to the same Gaussian process as given in Step 1, conditionally on the original observations.*

We first establish some general results, and then apply them to the two bandwidth sequences (i) and (ii). It is useful to write out the expression of $\hat{S}_*^R(\tau)$ explicitly:

$$\hat{S}_*^R(\tau) = [\widehat{S}_1^+(\tau) - \widehat{S}_1^-(\tau)] - [\widehat{S}_2^+(\tau) - \widehat{S}_2^-(\tau)],$$

where

$$\begin{aligned}
\widehat{S}_1^+(\tau) &= \frac{\hat{f}_{Y|X}(\tau|x_0) f_X(x_0)}{\hat{f}_{Y|X}\left(\tau|x_0^+\right) \hat{f}_X(x_0)} S_1^+(\tau), \qquad \text{(S.17)}\\
\widehat{S}_2^+(\tau) &= r(\tau)^{5/2} \frac{\sqrt{nh_{n,\tau}^5}}{\sqrt{nb_{n,\tau}^5}} \frac{\hat{f}_{Y|X}(\tau|x_0) f_X(x_0)}{\hat{f}_{Y|X}\left(\tau|x_0^+\right) \hat{f}_X(x_0)} S_2^+(\tau),\\
S_1^+(\tau) &= \frac{(nh_{n,\tau})^{-1/2} \sum_{i=1}^n (\tau - 1(u_i - \tau \leq 0)) \Xi_{i,\tau}^+ d_i K_{i,\tau}}{f_X(x_0)},\\
S_2^+(\tau) &= r(\tau)^{-5/2} \Gamma \frac{\iota_3' (\bar{N}^+)^{-1} (nb_{n,\tau})^{-1/2} \sum_{i=1}^n \{\tau - 1(u_i - \tau \leq 0)\} d_i \bar{z}_{i,\tau} \bar{K}_{i,\tau}}{f_X(x_0)},
\end{aligned}$$

and $\widehat{S}_1^-(\tau)$ and $\widehat{S}_2^-(\tau)$ are defined in the same way as $\widehat{S}_1^+(\tau)$ and $\widehat{S}_2^+(\tau)$, except using observations on the left side of the cut-off. The parameter $r(\tau)$ is defined only under bandwidth sequence (ii). It can be set to any finite positive value under bandwidth sequence (i).

For now, assume the following three convergences hold for the sample sequence $(x_1, y_1), (x_2, y_2), ...$ :

(C1) $\hat{f}_{Y|X}(\tau|x_0^+) \to f_{Y|X}(\tau|x_0^+)$, $\hat{f}_{Y|X}(\tau|x_0^-) \to f_{Y|X}(\tau|x_0^-)$, and $\hat{f}_{Y|X}(\tau|x_0) \to f_{Y|X}(\tau|x_0)$ uniformly over $\tau \in \mathcal{T}$. In addition, $\hat{f}_X(x_0) \to f_X(x_0)$.

(C2) For any $t, s \in \mathcal{T}$,

$$\frac{t \wedge s - ts}{n (h_{n,t} h_{n,s})^{1/2}} \sum_{i=1}^n \frac{(\Xi_{i,t}^+ d_i K_{i,t} - \Xi_{i,t}^- (1-d_i) K_{i,t})(\Xi_{i,s}^+ d_i K_{i,s} - \Xi_{i,s}^- (1-d_i) K_{i,s})}{f_X(x_0)^2} \to E[G_1(t) G_1(s)]$$

(C3) Under bandwidth sequence (ii), for any $t, s \in \mathcal{T}$,

$$\begin{aligned}
\frac{t \wedge s - ts}{n (b_{n,t} b_{n,s})^{1/2}} \sum_{i=1}^n \Bigg\{ & \frac{\Gamma^2}{r(t)^{5/2} r(s)^{5/2} f_X(x_0)^2} \iota_3' \left[(\bar{N}^+)^{-1} d_i - (\bar{N}^-)^{-1} (1-d_i)\right] \times \\
& \bar{z}_{i,t} \bar{K}_{i,t} \bar{K}_{i,s} \bar{z}_{i,s}' \left[(\bar{N}^+)^{-1} d_i - (\bar{N}^-)^{-1} (1-d_i)\right]' \iota_3 \Bigg\} \to E[G_2(t) G_2(s)].
\end{aligned}$$

We claim that, for every sequence $(x_1, y_1), (x_2, y_2), ...$ that satisfies (C1)-(C3), the following two results always hold:

(R1) The process $S_1^+(\tau) - S_1^-(\tau)$, conditional on $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, converges weakly to $G_1(\tau)$ over $\mathcal{T}$.

(R2) Under bandwidth sequence (ii), the process $S_2^+(\tau) - S_2^-(\tau)$, conditional on $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, converges weakly to $G_2(\tau)$ over $\mathcal{T}$.

Claim (R1) can be proved as below. First, the finite dimensional convergence of $S_1^+(\tau) - S_1^-(\tau)$ follows by applying the Cramer-Wold device conditionally, and then applying (C2). Note that the left-hand side of (C2) equals the covariance of $S_1^+(t)$ and $S_1^+(s)$, conditional on $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. Second, the stochastic equicontinuity of $S_1^+(\tau) - S_1^-(\tau)$ can be verified by applying the arguments in Lemma B.3 in Qu and Yoon (2015) conditionally, and then applying (C2). Claim (R1) then follows by combining these two results. Claim (R2) can be proved in the same way, using the condition (C3) instead of (C2).

Now, we apply (C1)−(C3) and (R1)−(R2) to the two bandwidth sequences. Under bandwidth sequence (i), (C1) and (R2) imply $\widehat{S}_2^+(\tau) - \widehat{S}_2^-(\tau)$ converges weakly to 0 conditionally. Further, (C1) and (R1) imply $\widehat{S}_1^+(\tau) - \widehat{S}_1^-(\tau)$ converges weakly to $G_1(\tau)$ conditionally. Therefore, $\hat{G}_*^R(\tau)$ converges weakly to $G_1(\tau)$, conditional on $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. Under bandwidth sequence (ii), (C1) and (R2) imply $\widehat{S}_2^+(\tau) - \widehat{S}_2^-(\tau)$ converges weakly to $G_2(\tau)$ conditionally. Further, (C1) and (R1) imply $\widehat{S}_1^+(\tau) - \widehat{S}_1^-(\tau)$ converges weakly to $G_1(\tau)$ conditionally. Therefore, $\hat{G}_*^R(\tau)$ converges weakly to $G_1(\tau) - G_2(\tau)$, conditional on $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$.

*Step 3. We show that (C1)−(C3) hold in probability for the original sample sequence* $(x_1, y_1), (x_2, y_2), ....$

For (C1), $\hat{f}_{Y|X}(\tau|x_0^+) \xrightarrow{p} f_{Y|X}(\tau|x_0^+)$ uniformly over $\mathcal{T}$, because $\sqrt{nh_{n,\tau}}(\hat{Q}(\tau|x_0^+) - Q(\tau|x_0^+)) = O_p(1)$ uniformly over $\mathcal{T}$, see (11). Similarly, $\hat{f}_{Y|X}(\tau|x_0^-) \xrightarrow{p} f_{Y|X}(\tau|x_0^-)$ uniformly over $\mathcal{T}$. By the continuous mapping theorem, $\hat{f}_{Y|X}(\tau|x_0)$ converges in probability to $f_{Y|X}(\tau|x_0)$ uniformly over $\mathcal{T}$. Finally, $\hat{f}_X(x_0) \to^p f_X(x_0)$ because $\hat{f}_X(x_0)$ is a standard kernel density estimator.

To prove (C2), it suffices to verify that the expectation of the left-hand side of (C2) converges to $E[G_1(t)G_1(s)]$, and that its variance converges to 0. Because the summands are i.i.d., the expectation of the left hand side is equal to

$$\frac{t \wedge s - ts}{(h_{n,t}h_{n,s})^{1/2}} E\left\{ \frac{(\Xi_{i,t}^+ d_i K_{i,t} - \Xi_{i,t}^-(1-d_i)K_{i,t})(\Xi_{i,s}^+ d_i K_{i,s} - \Xi_{i,s}^-(1-d_i)K_{i,s})}{f_X(x_0)^2} \right\}. \tag{S.18}$$

Consider the following component of (S.18):

$$\begin{aligned} & \frac{t \wedge s - ts}{(h_{n,t}h_{n,s})^{1/2}} E\left\{ \frac{\Xi_{i,t}^+ d_i K_{i,t} \Xi_{i,s}^-(1-d_i)K_{i,s}}{f_X(x_0)^2} \right\} \\ = & \frac{t \wedge s - ts}{(h_{n,t}h_{n,s})^{1/2} f_X(x_0)^2 \left(\mu_0^+\mu_2^+ - (\mu_1^+)^2\right)^2} \int_{-\infty}^{\infty} \left(\mu_2^+ - \left(\frac{x-x_0}{h_{n,t}}\right)\mu_1^+\right) I(x \geq x_0) \\ & \times K\left(\frac{x-x_0}{h_{n,t}}\right)\left(\mu_2^- - \left(\frac{x-x_0}{h_{n,s}}\right)\mu_1^-\right)(1 - I(x \geq x_0)) K\left(\frac{x-x_0}{h_{n,s}}\right) f_X(x)dx. \end{aligned}$$

Let $u = (x - x_0)/h_{n,0.5}$ and apply the mean value theorem. Then, the preceding display converges

to

$$\frac{t \wedge s - ts}{f_X(x_0)(\mu_0^+\mu_2^+ - (\mu_1^+)^2)^2 (\kappa(t)\kappa(s))^{1/2}} \int_{-\infty}^{\infty} \left(\mu_2^+ - \left(\frac{u}{\kappa(t)}\right)\mu_1^+\right) I(u \geq 0) \tag{S.19}$$
$$\times K\left(\frac{u}{\kappa(t)}\right)\left(\mu_2^- - \left(\frac{u}{\kappa(s)}\right)\mu_1^-\right)(1 - I(u \geq 0))\, K\left(\frac{u}{\kappa(s)}\right) du.$$

The remaining components of (S.18) can be analyzed in the same way. Combining these results, it follows that (S.18) converges to $E[G_1(t)G_1(s)]$. To show the variance of the left hand side of (C2) converges to zero, it is sufficient to prove that

$$\frac{(t \wedge s - ts)^2}{n^2 (h_{n,t}h_{n,s})} \sum_{i=1}^{n}\sum_{j=1}^{n} E\left(\frac{(\Xi_{i,t}^+ d_i K_{i,t} - \Xi_{i,t}^-(1-d_i)K_{i,t})(\Xi_{i,s}^+ d_i K_{i,s} - \Xi_{i,s}^-(1-d_i)K_{i,s})}{f_X(x_0)^2}\right)$$
$$\times \left(\frac{(\Xi_{j,t}^+ d_j K_{j,t} - \Xi_{j,t}^-(1-d_j)K_{j,t})(\Xi_{j,s}^+ d_j K_{j,s} - \Xi_{j,s}^-(1-d_j)K_{j,s})}{f_X(x_0)^2}\right) \to (E[G_1(t)G_1(s)])^2,$$

which holds because of the arguments (S.18)-(S.19), the independence of the summands when $i \neq j$, and $nh_{n,0.5} \to \infty$. The convergence in (C3) holds for the same reason as that in (C2); the detail is omitted.

*Step 4. We apply a subsequence argument to show that the simulation procedure is weakly consistent.*

First, from Step 3, any subsequence of $(x_1, y_1), (x_2, y_2), ...$ contains a further subsequence, such that (C1)−(C3) holds with probability 1, by Theorem 20.5 in Billingsley (1986). Second, from Step 2, conditional on any of such further subsequence, the simulated process $\hat{S}_*^R(\tau)$ converges weakly to the same limit as $G_*^R(\tau)$ does. Therefore, $\hat{S}_*^R(\tau)$, conditional on the original sequence $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, converges weakly to the desired limit in probability. This implies that the simulation procedure is weakly consistent.

## S.6 Additional simulation results

This subsection considers three issues. First, it compares the power of the score and Wald tests under an ideal simulation setting. Second, it evaluates the effect of estimating the conditional density on the size of the Wald tests. Third, it reports the rejection rates at the 5% nominal level for Models 1−4 considered in Section 8.

**Power comparison.** Tests may display different power if they use distinct bandwidths or different bias correction methods. To exclude such effects, we compare the score test and the Wald test for the treatment significance hypothesis using the same bandwidth without bias estimation. The sample size is 1000 and the bandwidth at the median is 0.4. The data generating processes are

Models 1 and 2. The rejection frequencies at the 10% level are reported in Table A1. The values are comparable between the two tests for all the cases considered. This confirms the results from the local power analysis.

**Conditional density estimation.** Section 8 documents that the Wald tests have less stable size properties compared to those of the score test, especially when the sample size is small. Here, we examine the extent to which this is because the Wald tests require estimating the conditional density. We repeat the same procedures as in Section 8, but using the true conditional densities instead of the estimated densities. The data generating processes are Models 1 and 2 for which all three tests are valid. The sample size is $n = 500$ and the nominal level is 10%. Table A2 shows the results. Compared with Tables 2−4, the values are now consistently close to the nominal level. They are also comparable to those of the score test. Therefore, estimating conditional densities accounts for most of the size distortions in small samples.

**Empirical sizes at the 5% nominal level.** Section 8 only reports sizes at the 10% level. To complement these results, here we also report the sizes the 5% level. The results are shown in Tables A3−A6. Overall, the same patterns as in Tables 2−4 and 9 are observed. The conclusions therefore remain the same.

Table A1: Power of Score and Wald tests using the same bandwidth (10%).

| **Test** | **Model 1** | | | | **Model 2** | | | |
|---|---|---|---|---|---|---|---|---|
| | $c_h$=0.3 | 0.6 | 1.0 | 2.0 | $c_h$=0.3 | 0.6 | 1.0 | 2.0 |
| Score | 0.265 | 0.663 | 0.948 | 1.000 | 0.386 | 0.741 | 0.975 | 1.000 |
| Wald | 0.293 | 0.648 | 0.922 | 1.000 | 0.336 | 0.693 | 0.941 | 1.000 |

Empirical rejection frequencies based on 2000 repetitions. The sample size $n = 1000$ and the bandwidth at the median is fixed at 0.4.

Table A2: The Size of Wald tests using true conditional density functions (10%).

| **Methods** | **Model 1** | | | **Model 2** | | |
|---|---|---|---|---|---|---|
| | **TS** | **TH** | **TU** | **TS** | **TH** | **TU** |
| **Wald** | | | | | | |
| $h_{0.5}^{cv}$ | 0.102 | 0.108 | 0.104 | 0.101 | 0.101 | 0.076 |
| $h_{0.5}^{bdy}$ | 0.105 | 0.100 | 0.111 | 0.110 | 0.096 | 0.090 |
| $h_{0.5}^{ik}$ | 0.105 | 0.106 | 0.106 | 0.102 | 0.098 | 0.102 |
| **Wald Robust** | | | | | | |
| $h_{0.5}^{cv}$ | 0.094 | 0.094 | 0.098 | 0.094 | 0.089 | 0.097 |
| $h_{0.5}^{bdy}$ | 0.090 | 0.090 | 0.092 | 0.094 | 0.076 | 0.100 |
| $h_{0.5}^{ik}$ | 0.096 | 0.081 | 0.098 | 0.100 | 0.074 | 0.096 |
| **Wald Robust EC** | | | | | | |
| $h_{0.5}^{cv}$ | 0.106 | 0.100 | 0.098 | 0.100 | 0.090 | 0.103 |
| $h_{0.5}^{bdy}$ | 0.103 | 0.096 | 0.106 | 0.108 | 0.082 | 0.104 |
| $h_{0.5}^{ik}$ | 0.100 | 0.096 | 0.102 | 0.100 | 0.090 | 0.102 |

Empirical rejection frequencies based on 2000 repetitions. The sample size is $n = 500$. **TS**, **TH**, and **TU** stand for the treatment significance/homogeneity/unambiguity hypotheses.

Table A3: The Size of Tests for the Treatment Significance Hypothesis (Models 1 & 2).

| **Tests** | **Model 1** | | | **Model 2** | | |
|---|---|---|---|---|---|---|
| | n=500 | n=1000 | n=2000 | n=500 | n=1000 | n=2000 |
| **Score** | | | | | | |
| $h_{0.5}^{cvi}$ | 0.054 | 0.042 | 0.048 | 0.057 | 0.050 | 0.048 |
| $h_{0.5}^{int}$ | 0.054 | 0.058 | 0.056 | 0.062 | 0.062 | 0.058 |
| **Wald** | | | | | | |
| $h_{0.5}^{cv}$ | 0.090 | 0.067 | 0.056 | 0.098 | 0.094 | 0.072 |
| $h_{0.5}^{bdy}$ | 0.101 | 0.074 | 0.058 | 0.116 | 0.084 | 0.062 |
| $h_{0.5}^{ik}$ | 0.130 | 0.084 | 0.059 | 0.132 | 0.092 | 0.061 |
| **Wald Robust** | | | | | | |
| $h_{0.5}^{cv}$ | 0.081 | 0.051 | 0.045 | 0.084 | 0.060 | 0.041 |
| $h_{0.5}^{bdy}$ | 0.085 | 0.059 | 0.042 | 0.091 | 0.062 | 0.042 |
| $h_{0.5}^{ik}$ | 0.102 | 0.074 | 0.042 | 0.109 | 0.074 | 0.040 |
| **Wald Robust EC** | | | | | | |
| $h_{0.5}^{cv}$ | 0.088 | 0.060 | 0.054 | 0.097 | 0.068 | 0.050 |
| $h_{0.5}^{bdy}$ | 0.094 | 0.064 | 0.052 | 0.112 | 0.070 | 0.051 |
| $h_{0.5}^{ik}$ | 0.113 | 0.074 | 0.049 | 0.122 | 0.078 | 0.048 |

Note. The table reports rejection frequencies at the **5 percent** nominal level based on 2000 replications. "Wald", "Wald Robust" and "Wald Robust EC" denote tests constructed assuming a continuous second order derivative at the cutoff, allowing a discontinuous second order derivative whose magnitude of discontinuity can vary freely across the quantiles, and allowing a discontinuous second order derivative whose magnitude of discontinuity remains constant across the quantiles. See the footnote of Table 1 in the main text for the definitions of the bandwidth parameters.

Table A4: The Size of Tests for the Treatment Unambiguity Hypothesis (Models 1 & 2).

| **Tests** | **Model 1** | | | **Model 2** | | |
|---|---|---|---|---|---|---|
| | n=500 | n=1000 | n=2000 | n=500 | n=1000 | n=2000 |
| **Wald** | | | | | | |
| $h_{0.5}^{cv}$ | 0.086 | 0.052 | 0.046 | 0.062 | 0.044 | 0.031 |
| $h_{0.5}^{bdy}$ | 0.088 | 0.057 | 0.047 | 0.078 | 0.052 | 0.040 |
| $h_{0.5}^{ik}$ | 0.100 | 0.064 | 0.055 | 0.102 | 0.062 | 0.047 |
| **Wald Robust** | | | | | | |
| $h_{0.5}^{cv}$ | 0.070 | 0.049 | 0.040 | 0.064 | 0.046 | 0.037 |
| $h_{0.5}^{bdy}$ | 0.077 | 0.053 | 0.046 | 0.088 | 0.053 | 0.044 |
| $h_{0.5}^{ik}$ | 0.082 | 0.053 | 0.047 | 0.087 | 0.058 | 0.042 |
| **Wald Robust EC** | | | | | | |
| $h_{0.5}^{cv}$ | 0.078 | 0.054 | 0.052 | 0.078 | 0.052 | 0.044 |
| $h_{0.5}^{bdy}$ | 0.082 | 0.056 | 0.048 | 0.092 | 0.058 | 0.042 |
| $h_{0.5}^{ik}$ | 0.088 | 0.058 | 0.046 | 0.091 | 0.057 | 0.044 |

Note. The nominal level is **5 percent**. See Table A3.

Table A5: The Size of Tests for the Treatment Homogeneity Hypothesis (Models 1 & 2).

| **Tests** | **Model 1** | | | **Model 2** | | |
|---|---|---|---|---|---|---|
| | n=500 | n=1000 | n=2000 | n=500 | n=1000 | n=2000 |
| **Wald** | | | | | | |
| $h_{0.5}^{cv}$ | 0.084 | 0.060 | 0.046 | 0.088 | 0.064 | 0.052 |
| $h_{0.5}^{bdy}$ | 0.082 | 0.062 | 0.048 | 0.094 | 0.071 | 0.050 |
| $h_{0.5}^{ik}$ | 0.102 | 0.073 | 0.052 | 0.110 | 0.068 | 0.057 |
| **Wald Robust** | | | | | | |
| $h_{0.5}^{cv}$ | 0.077 | 0.049 | 0.042 | 0.077 | 0.048 | 0.038 |
| $h_{0.5}^{bdy}$ | 0.078 | 0.053 | 0.038 | 0.080 | 0.052 | 0.040 |
| $h_{0.5}^{ik}$ | 0.096 | 0.066 | 0.043 | 0.102 | 0.070 | 0.046 |
| **Wald Robust EC** | | | | | | |
| $h_{0.5}^{cv}$ | 0.072 | 0.053 | 0.042 | 0.074 | 0.056 | 0.044 |
| $h_{0.5}^{bdy}$ | 0.064 | 0.054 | 0.041 | 0.076 | 0.062 | 0.042 |
| $h_{0.5}^{ik}$ | 0.088 | 0.064 | 0.046 | 0.098 | 0.059 | 0.045 |

Note. The nominal level is **5 percent**. See Table A3.

Table A6: The Size of Robust Tests in Models 3 & 4.

| **Tests** | **Model 3** | | | **Model 4** | | |
|---|---|---|---|---|---|---|
| | n=500 | n=1000 | n=2000 | n=500 | n=1000 | n=2000 |
| **Treatment Significance:** | | | | | | |
| **Wald Robust** | | | | | | |
| $h_{0.5}^{cv}$ | 0.110 | 0.094 | 0.086 | 0.125 | 0.075 | 0.054 |
| $h_{0.5}^{bdy}$ | 0.106 | 0.063 | 0.042 | 0.124 | 0.060 | 0.046 |
| $h_{0.5}^{ik}$ | 0.110 | 0.062 | 0.055 | 0.105 | 0.058 | 0.036 |
| **Wald Robust EC** | | | | | | |
| $h_{0.5}^{cv}$ | 0.123 | 0.100 | 0.100 | 0.133 | 0.078 | 0.060 |
| $h_{0.5}^{bdy}$ | 0.122 | 0.072 | 0.051 | 0.136 | 0.079 | 0.060 |
| $h_{0.5}^{ik}$ | 0.121 | 0.084 | 0.064 | 0.120 | 0.064 | 0.041 |
| **Treatment Unambiguity:** | | | | | | |
| **Wald Robust** | | | | | | |
| $h_{0.5}^{cv}$ | 0.029 | 0.026 | 0.020 | 0.078 | 0.052 | 0.052 |
| $h_{0.5}^{bdy}$ | 0.048 | 0.038 | 0.028 | 0.076 | 0.050 | 0.048 |
| $h_{0.5}^{ik}$ | 0.052 | 0.036 | 0.018 | 0.066 | 0.036 | 0.031 |
| **Wald Robust EC** | | | | | | |
| $h_{0.5}^{cv}$ | 0.032 | 0.027 | 0.019 | 0.082 | 0.061 | 0.057 |
| $h_{0.5}^{bdy}$ | 0.047 | 0.046 | 0.034 | 0.083 | 0.056 | 0.048 |
| $h_{0.5}^{ik}$ | 0.057 | 0.042 | 0.022 | 0.067 | 0.038 | 0.032 |
| **Treatment Homogeneity:** | | | | | | |
| **Wald Robust** | | | | | | |
| $h_{0.5}^{cv}$ | 0.072 | 0.044 | 0.050 | 0.104 | 0.068 | 0.061 |
| $h_{0.5}^{bdy}$ | 0.080 | 0.052 | 0.054 | 0.096 | 0.066 | 0.054 |
| $h_{0.5}^{ik}$ | 0.080 | 0.048 | 0.050 | 0.077 | 0.066 | 0.044 |
| **Wald Robust EC** | | | | | | |
| $h_{0.5}^{cv}$ | 0.068 | 0.052 | 0.048 | 0.098 | 0.082 | 0.064 |
| $h_{0.5}^{bdy}$ | 0.088 | 0.052 | 0.060 | 0.100 | 0.074 | 0.058 |
| $h_{0.5}^{ik}$ | 0.086 | 0.056 | 0.045 | 0.091 | 0.066 | 0.050 |

Note. The nominal level is **5 percent**. See Table A3.