Quantile-based Test for Heterogeneous Treatment Effects

EunYi Chung[†]
Department of Economics
University of Illinois, Urbana–Champaign
eunyi@illinois.edu

Mauricio Olivares
Department of Statistics
LMU Munich
m.olivares@lmu.de

March 1, 2023

Abstract

We introduce a permutation test for heterogeneous treatment effects based on the quantile process. However, tests based on the quantile process often suffer from estimated nuisance parameters that jeopardize their validity, even in large samples. To overcome this problem, we use Khmaladze's martingale transformation. We show that the permutation test based on the transformed statistic controls size asymptotically. Numerical evidence asserts the good size and power performance of our test procedure compared to other popular quantile-based tests. We discuss a fast implementation algorithm and illustrate our method using experimental data from a welfare reform.

Keywords: Permutation Test, Quantile Treatment Effects, Heterogeneous Treatment Effects.
JEL Classification: C12, C14, C46.

[†]This paper supersedes the paper by the authors which was circulated with the title *Comment on "Can Variation in Subgroups' Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment."* Mauricio Olivares acknowledges support from the European Research Council (Starting Grant No. 852332). We are grateful to MDRC for granting access to the experimental data we use in this paper. All errors are our own.

1 Introduction

The study of heterogeneous treatment effects (HTE) plays an important role in program evaluation. A popular approach to studying HTE involves a form of subgroup analysis: divide the sample into subgroups defined by covariates, and then estimate the average treatment effects (ATEs) across subgroups. Then, this approach detects HTE by checking whether the ATEs vary significantly across subgroups.[1]

While the ATE subgroup analysis might detect some HTE, it has important limitations because it focuses on mean impacts (Bitler, Gelbach, and Hoynes, 2006). Therefore, we take an alternative route to HTE analysis based on the comparison of quantile treatment effect (QTE) at different quantiles. Under this framework, we test for HTE by checking if the QTEs are constant across quantiles, *i.e.*, whether the QTEs are equal to an unknown constant γ for all quantiles. Thus, by looking at the entire distribution and not only the mean, the QTE approach complements the ATE subgroup analysis by giving the researcher more tools to investigate HTE.

However, tests of HTE based on QTEs come with their own challenges. In particular, when we estimate the nuisance parameter γ to compute the test statistic, the estimation error influences the limit behavior of the test statistic. In fact, its asymptotic distribution becomes intractable because of the dependence on the (unknown) probability distribution generating the data. This phenomenon is called the Durbin problem in the literature.

In this paper, we introduce a new permutation test for HTE based on the quantile process. Importantly, we establish the asymptotic validity of our method despite the presence of the nuisance parameter. To this end, we apply Khmaladze's (1981) martingale transformation of the quantile process. Simply put, the Khmaladze transformation removes the estimation effects by residualizing the quantile process. This transformation yields an asymptotically pivotal statistic, *i.e.*, a statistic whose limiting distribution does not depend on the fundamentals. Our main theoretical result shows the permutation distribution of the transformed statistic and

[1] According to a survey of papers in top economics journals, 40% of the studies report at least one treatment effect this way (Chernozhukov et al., 2018). This practice sparked the development of tests for HTE. For example, Crump et al. (2008) test whether the ATEs conditional on covariates are identical for all subgroups.

the true sampling distribution are the same in large samples, thus restoring the asymptotic validity of the permutation test with estimated parameters. To complement our theoretical result and ease implementing our permutation test, we also provide free software in the `RATest` R package, available on CRAN, and discuss a fast computational implementation based on the preprocessing algorithm of Portnoy and Koenker (1997).

To illustrate our method, we revisit the Connecticut's Jobs First welfare reform. We begin by showing how to apply our method to test for HTE within subgroups, for a family of subgroups formed by pre-treatment characteristics. To do so, we cast our problem as a multiple testing problem. This multiple testing approach allows us to detect for which subgroups there is evidence of HTE. When applied to our empirical application, we provide strong evidence against the null hypothesis of constant treatment effects for a series of subgroups. Our conclusion aligns with the heterogeneous predictions of the static labor model as in Bitler, Gelbach, and Hoynes (2017, BGH). Even though our results are not qualitatively different from BGH, our test differs from theirs in two important ways. First, our method compares quantiles as opposed to distribution functions (CDFs). While this difference may seem innocuous, it has substantial implications for estimation and inference. For example, the calculation of the quantile process requires a uniformly consistent estimator of the density, unlike a test statistic based on empirical CDFs (e.g. (6) below). Second, the estimated constant treatment effect enters the test statistic in both cases, so neither is immune to the Durbin problem. However, BGH's inference method overlooks this issue when testing for HTE because they wrongly assumed the asymptotic validity of their permutation test with a plug-in estimator. On the other hand, we provide an asymptotically valid permutation test that overcomes the Durbin problem when testing for HTE. We elaborate on this second point in Section 7.1 and in the online Appendix IV.

The present paper falls under the umbrella of a literature that has addressed inference for HTE using tests based on the quantile or the empirical process. As we argued before, this approach to testing for HTE suffers from the Durbin problem. From this angle, we can classify the literature into two branches. One approach seeks to restore large-sample pivotality and use an asymptotic test, while the other uses a resampling technique to construct valid critical

values. Noteworthy examples of the former include Durbin (1973, 1975, 1985), Khmaladze (1981, 1993), Koenker and Xiao (2002). Meanwhile, examples of the latter include tests based on subsampling the quantile process, as in Chernozhukov and Fernández-Val (2005), or the permutation tests of Ding, Feller, and Miratrix (2016).

Our paper combines ideas from the two modeling approaches we discussed before. This notion is best explained by comparing our proposed test with Koenker and Xiao's (2002). In their paper, the authors show that the Khmaladze transformation of the quantile process renders an asymptotically pivotal statistic, and therefore valid inference is possible by simulating the asymptotic distribution, often depending on user-specific parameters. Our proposed method goes one step further. We show that the permutation distribution of the Khmaladze transformed statistic mimics the true sampling distribution of the test statistic. Thus, our permutation test offers an off-the-shelf way to generate data-dependent, asymptotically valid critical values without simulating the limiting distribution. In addition, we show in Section 4.2 that the asymptotic power of our permutation test against contiguous alternatives is identical to Koenker and Xiao's (2002) test, so there is no loss in power when using the permutation-based critical values. For the sake of exposition, we also compare the two methods in a Monte Carlo experiment in Section 6. We find in our numerical simulations that their approach leads to a more conservative test procedure than ours across the specifications we considered.

The idea behind using an asymptotically pivotal statistic as the input for a permutation test is not new, dating back at least to the pioneer works of Neuhaus (1993) and Janssen (1997). Chung and Romano (2013) generalized this principle, sparking multiple applications of this method ever since (see Chung and Romano (2016) and their references). In this spirit, our paper relates closely to Chung and Olivares (2021), who also test for HTE using a Khmaladze transformed statistic. However, their method is based on the comparison of CDFs as opposed to quantile functions as we do here. Though CDFs and quantiles are logically connected, they are conceptually different. We highlight three crucial differences between these approaches that stem from this fact. First, we argue the test in this paper is more relevant for applications because distributional effects are more widely studied in terms of quantiles than CDFs (Bitler,

Gelbach, and Hoynes, 2006; Khandker, Koolwal, and Samad, 2009; Frölich and Sperlich, 2019). To see why, think of the policymaker in our empirical application. This policymaker may want to determine whether the welfare reform affects the lower or upper tail more than, say, the center of the earnings distribution. In this reasonable scenario, the differences in quantile functions may be a more natural object (estimand) than differences in CDFs because inspecting QTEs across quantiles is more intuitive for this task. Second, our proposed method takes advantage of interior point methods applicable in quantile regression that make the calculation of the quantile process computationally efficient and, therefore, attractive from a practitioner's point of view; we discuss these algorithms in Section 5. Lastly, we provide numerical evidence in the online Appendix III showing that a test based on quantiles exhibits better size control than one based on CDF comparisons. Thus, we can see the permutation test in this paper is a more intuitive complement rather than a substitute.

We organize the rest of the paper as follows. In the next section, we introduce our general setup, including a formal description of the statistical environment. We begin by providing our hypothesis of interest and the test statistic based on the quantile process and then turn our attention to the classical construction of the permutation test. As mentioned previously, the hypothesis of constant treatment effects involves nuisance parameters whose estimation affects the limiting distribution of the test statistic. In Section 3, we examine these effects. In particular, we show that the permutation test that ignores the Durbin problem fails to control the type I error, even asymptotically. Our main result is the content of Section 4. First, we introduce the Khmaladze transformation of the quantile process and then we establish the asymptotic validity of a permutation test based on the transformed test statistic in Section 4.2. We discuss a fast implementation of our test using Portnoy and Koenker's (1997) preprocessing algorithm in Section 5. In Section 6 we examine the finite-sample performance of our permutation test via a Monte Carlo study, and compare its behavior with other popular quantile-based methods, such as Koenker and Xiao (2002); Chernozhukov and Fernández-Val (2005); Linton, Maasoumi, and Whang (2005). Finally, in Section 7, we apply our inference method to re-examine the treatment effect variation of a welfare program on earnings using experimental

data from Connecticut's Jobs First. We collect the proofs of the main results and auxiliary lemmas in the online Appendix. Similarly, we leave additional simulation results, numerical implementation details, and additional discussion with regard to our empirical application in the online supplementary appendix.

2 Statistical Environment

Suppose that Y is a real outcome of interest and D is a treatment or policy indicator taking values 1 if treated, and 0 otherwise. The observed outcome is linked to the potential outcomes through the relationship $Y = Y(1)D + (1 - D)Y(0)$. The object of interest is individual i's treatment effect given by $\delta_i = Y_i(1) - Y_i(0)$, and we seek whether the treatment effect varies across $i = 1, \ldots, N$ individuals. More formally, the null hypothesis of constant treatment effects states that

$$H_0^s : \delta_i = \delta \quad \text{for some } \delta, \ \forall \ i = 1, \ldots, N \ . \tag{1}$$

Hypotheses like (1) are not directly testable in practice because typically δ is unknown in practice and we never observe both potential outcomes for the same experimental unit. Motivated by this limitation, one might consider testing a weaker null hypothesis instead and work under the Doksum–Lehmann model (Doksum, 1974; Lehmann, 1974). In this model, we test heterogeneity in the treatment effect by checking whether the treatment impact varies *across quantiles* (e.g. Koenker and Xiao, 2002; Chernozhukov and Fernández-Val, 2005).

More formally, let $F_1(\cdot)$ and $F_0(\cdot)$ denote the distribution functions of $Y(1)$ and $Y(0)$, respectively. The QTE is given by

$$\gamma(\tau) = F_1^{-1}(\tau) - F_0^{-1}(\tau), \ \forall \, \tau \in [0, 1] \, ,$$

where $F_d^{-1}(\tau) = \inf\{y \ : \ F_d(y) \geq \tau\}$, $d \in \{0, 1\}$. Throughout this paper, we will focus on quantiles τ in $\mathscr{T}$, a closed subinterval of $(0, 1)$. Therefore, the testable hypothesis of constant treatment effect in this paper is given by

$$H_0 : \gamma(\tau) = \gamma \ \text{ for some } \ \gamma, \ \ \forall \ \tau \in \mathscr{T}, \tag{2}$$

and the alternative is the hypothesis of heterogeneous effects, that is $\gamma(\tau)$ varies across $\tau \in \mathscr{T}$. We note that (1) implies (2), so a test that rejects H_0 will reject the more restrictive sharp null of constant treatment effects.[2]

One natural candidate for a test statistic for hypothesis (2) is to compare the empirical quantile functions based on two independent random samples from F_1 and F_0. More formally, let $Y_{1,1}, \ldots, Y_{1,m}$ and $Y_{0,1}, \ldots, Y_{0,n}$ be two independent random samples having distribution functions $F_1(\cdot)$ and $F_0(\cdot)$, respectively.[3] Let $N = m+n$ and collect these outcomes in one vector as $\boldsymbol{Z} = (Z_1, \ldots, Z_N) = (Y_{1,1}, \ldots, Y_{1,m}, Y_{0,1}, \ldots, Y_{0,n})$. Observe that the number of observations m and n are deterministic. Thus, the first m entries in $\boldsymbol{Z}$ are drawn from F_1 and the last n from F_0. For the case with a random number of observations, see Bertanha and Chung (2022).

The QTE, $\gamma(\tau)$, is estimable in the two-sample setting by

$$\hat{\gamma}(\tau; \boldsymbol{Z}) = \hat{F}_1^{-1}(\tau) - \hat{F}_0^{-1}(\tau), \ \ \tau \in \mathscr{T}, \tag{3}$$

where $\hat{F}_1^{-1}(\tau)$ denotes the empirical quantile function based on $Y_{1,1}, \ldots, Y_{1,m}$, and analogously, $\hat{F}_0^{-1}(\tau)$ is the empirical quantile function based on $Y_{0,1}, \ldots, Y_{0,n}$. If we formulate a quantile regression model for the binary treatment, then we may estimate the QTE in (3) by the individual coefficient associated with the policy indicator in the conditional quantile regression model. More specifically, let $F_{Y|D}^{-1}(\tau) = \inf\{y : F_{Y|D}(y) \geq \tau\}$ be the τth quantile of the conditional CDF of Y given D, and formulate the quantile regression model as,

$$F_{Y|D}^{-1}(\tau) = \alpha(\tau) + \gamma(\tau)D, \ \ \tau \in \mathscr{T},$$

then, we can estimate the QTE directly by solving the quantile regression problem

$$\{\hat{\alpha}(\tau; \boldsymbol{Z}), \hat{\gamma}(\tau; \boldsymbol{Z})\} = \underset{a,b \in \mathbb{R}}{\arg\min} \sum_{i=1}^{N} \rho_\tau \left(Y_i - a - bD_i\right), \ \ \tau \in \mathscr{T}, \tag{4}$$

[2] To see why, suppose (1) holds, then $Y(0) = Y(1) - \delta$. Since $Y(0)$ is an affine transformation of $Y(1)$, a simple application of the change of variable theorem implies that $F_1(y+\delta) = F_0(y)$. Take an arbitrary τ and observe that the change of variable $y \mapsto F_0^{-1}(\tau)$ implies $\delta = F_1^{-1}(\tau) - F_0^{-1}(0)$, so $\gamma = \delta$, as desired. From this, one could define the null hypothesis in terms of the distribution functions $F_1(\cdot)$ and $F_0(\cdot)$ as $H_0 : F_1(y+\delta) = F_0(y)$, for some δ (e.g. Ding, Feller, and Miratrix, 2016; Chung and Olivares, 2021).

[3] That is, $Y_{1,i} = Y_i$ among the treated, and $Y_{0,i} = Y_i$ among the non-treated. Throughout, we assume complete randomization; see Zhang and Zheng (2020) for a more detailed discussion about the estimation and inference for QTE under covariate-adaptive randomization.

where ρ_τ is the check function defined as $\rho_\tau(u) = u(\tau - \mathbb{1}_{\{u<0\}})$ and $\mathbb{1}_{\{\cdot\}}$ is the indicator function.

The test statistic we consider in this paper is the two-sample Kolmogorov–Smirnov statistic based on the quantile process (2SKSQ):

$$K_{\text{N}}(\boldsymbol{Z}) = \sup_{\tau\in\mathscr{T}} |\hat{\upsilon}_{\text{N}}(\tau; \boldsymbol{Z})| \ , \tag{5}$$

where $\hat{\upsilon}_{\text{N}}(\cdot\,; \boldsymbol{Z})$ is the standardized quantile regression process given by

$$\hat{\upsilon}_{\text{N}}(\tau; \boldsymbol{Z}) = \sqrt{\frac{mn}{N}}\hat{\varphi}(\tau; \boldsymbol{Z}) \left\{\hat{\gamma}(\tau; \boldsymbol{Z}) - \hat{\gamma}\right\}, \ \ \tau \in \mathscr{T}, \tag{6}$$

$\hat{\gamma}$ is an estimate of the nuisance parameter γ, $\hat{\varphi}(\tau; \boldsymbol{Z})$ is an estimate of $\varphi(\tau) = f_0(F_0^{-1}(\tau))$, and f_0 is the (strictly positive) density of F_0. In what follows, we drop the dependency on $\boldsymbol{Z}$ and write $\hat{\gamma}(\tau)$ and $\hat{\varphi}(\tau)$ to shorten notation when it is clear from the context. Following Shorack and Wellner (2009, Chp. 18), we normalize the quantile process by $\hat{\varphi}(\tau)$ because, if the nuisance parameter γ was known, the test statistic in (5) would be asymptotically pivotal (see also Van der Vaart and Wellner (1996, Example 3.9.24)). We relegate the technical details and assumptions about all these quantities to Section 3.

Remark 1. Observe that $\hat{\gamma}$ enters (6) because the nuisance parameter γ is unknown and therefore we need to estimate it. In this paper, we estimate γ by the OLS estimator of a regression of Y on D. Alternatively, let $\kappa > 0$ and set $\mathscr{T} = [\kappa, 1-\kappa]$ so that $\mathscr{T}$ is a closed subinterval of $[0, 1]$. Therefore, we can estimate γ by $\frac{1}{1-2\kappa}\int_\kappa^{1-\kappa} \hat{\gamma}(\tau)d\tau$ as well. Even though the latter method yields approximately the mean treatment effect as estimated by the associated OLS, one should be cautious about this interpretation in the presence of outliers. ■

Remark 2. Observe that we are recentering the quantile process in (6) by the estimated nuisance parameter $\hat{\gamma}$. To gain more intuition of the recentering embedded in (6), let us show an equivalent formulation of it. Denote

$$\tilde{\boldsymbol{Z}} = (\tilde{Z}_1, \ldots, \tilde{Z}_N) = (Y_{1,1} - \hat{\gamma}, \ldots, Y_{1,m} - \hat{\gamma}, Y_{0,1}, \ldots, Y_{0,n}) \ , \tag{7}$$

where $\hat{\gamma}$ is given as before (see Remark 1). That is, we are shifting the first m entries in $\boldsymbol{Z}$ by the estimated nuisance parameter. Let $\tilde{F}_1^{-1}(\tau)$ be the sample analog of the τ-quantile

function based on the first m entries of $\tilde{\boldsymbol{Z}}$, and $\hat{F}_0^{-1}(\tau)$ be defined as before. Therefore, we can equivalently write the standardized quantile regression process (6) as

$$\hat{v}_{\text{N}}(\tau;\tilde{\boldsymbol{Z}}) = \sqrt{\frac{mn}{N}}\hat{\varphi}(\tau)\left\{\tilde{F}_1^{-1}(\tau) - \hat{F}_0^{-1}(\tau)\right\}, \quad \tau \in \mathscr{T}, \tag{8}$$

where $\hat{\varphi}(\tau)$ is unaffected by the recentering because it only depends on the sample from F_0. Similarly, whenever we write $K_{\text{N}}(\tilde{\boldsymbol{Z}})$, we mean the 2SKSQ based on (8). ■

2.1 Adding Covariates

In practice, we typically observe a vector of baseline covariates besides D. However, adding covariates is not innocuous in our context because the interpretation of conditional and unconditional quantiles is different. To see why, suppose the outcome of interest is earnings, as in the empirical illustration of Section 7. We can argue that the 0.9 unconditional quantile of the earnings distribution may be quite different from the 0.9 quantile of the earnings distribution conditional on education. For example, the high earners within each education bracket (*no high-school diploma*, *high-school diploma*, or *more than high-school diploma*) might not be the high earners overall.[4]

Following the ongoing discussion, we will consider a null hypothesis based on the *conditional* QTEs in this Section. To describe how to handle the baseline covariates in our analysis, we introduce additional notation. Let $\boldsymbol{X} = (X_1, \boldsymbol{X}_2) = (D, \boldsymbol{X}_2) \in \mathbb{R}^l$ denote the vector of covariates, where $\boldsymbol{X}_2$ contains the pre-treatment characteristics. In the spirit of Abadie, Angrist, and Imbens (2002), we impose the linear quantile regression model given by $F_{Y|X}^{-1}(\tau) = \boldsymbol{X}'\beta(\tau) = \beta_1(\tau)D + \boldsymbol{X}_2'\beta_2(\tau)$ for all τ. Then, the QTE at τ conditional on $\boldsymbol{X}_2 = \boldsymbol{x}$ is $\beta_1^x(\tau)$. It is given by the difference in conditional τ-quantiles of Y_1 and Y_0:

$$\beta_1^x(\tau) = F_{Y|D=1,X_2=x}^{-1}(\tau) - F_{Y|D=0,X_2=x}^{-1}(\tau) \ .$$

[4]Observe that this problem does not arise when we are dealing with averages because the unconditional mean is the average of conditional means by the law of iterated expectations.

The hypothesis of interest for $\boldsymbol{X}_2 = \boldsymbol{x}$ now becomes

$$H_0^x : \beta_1^x(\tau) = \beta_1^x, \quad \text{for some} \quad \beta_1^x, \quad \tau \in \mathscr{T} \ . \tag{9}$$

As is standard in the quantile regression literature, we can estimate the parameters $\beta(\tau)$ by

$$\hat{\beta}(\tau; \boldsymbol{X}) = \underset{b \in \mathbb{R}^l}{\arg\min} \sum_{i=1}^{N} \rho_\tau \left(Y_i - \boldsymbol{X}_i' b \right), \quad \tau \in \mathscr{T} \ .$$

Therefore, we can mirror the testing procedure described in the previous Sections and consider covariates $\boldsymbol{X}_2$ for testing H_0^x in (9), with $\beta_1^x(\tau)$ and β_1^x playing the role of $\gamma(\tau)$ and γ, provided some standard regularity conditions on $\boldsymbol{X}_2$ hold (e.g. Koenker and Machado, 1999, Assumption 2). We refer to the reader to Koenker and Xiao (2002, Theorems 2 and 3), and Chernozhukov and Fernández-Val (2005, Proposition 1) for further details. We will focus on the case with no covariates from now on.

2.2 Permutation Test based on the Quantile Process

Before turning to the theoretical results, we first show how the construction of a permutation test to assess (2) works. To facilitate the exposition, we will illustrate this construction using the test statistic in eq. (8). To do so, we introduce further notation. Let $\mathbf{G}_N$ be the set of all permutations π of $\{1, \ldots, N\}$, so $|\mathbf{G}_N| = N!$. Fix a nominal level $\alpha \in (0, 1)$, and set $k = N! - \lfloor N! \, \alpha \rfloor$. We can describe the construction of the permutation test in four steps as follows:

Step 1. Calculate $\hat{\gamma}$ as in Remark 1 and recenter the units from F_1, *i.e.*, $\{Y_{1,i} - \hat{\gamma} \ : \ 1 \leq i \leq m\}$. At the end of this step, we are left with $\tilde{\boldsymbol{Z}}$ in (7). See also Remark 2 for more intuition.

Step 2. Given $\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{z}}$, calculate the test statistic $K_{\text{N}}(\tilde{\boldsymbol{z}})$ based on (8).

Step 3. Recompute $K_{\text{N}}(\tilde{\boldsymbol{z}})$ for all permutations $\pi \in \mathbf{G}_N$. Denote by $K_{\text{N}}^{(1)}(\tilde{\boldsymbol{z}}) \leq K_{\text{N}}^{(2)}(\tilde{\boldsymbol{z}}) \leq \cdots \leq K_{\text{N}}^{(N!)}(\tilde{\boldsymbol{z}})$ the ordered values of $\{K_{\text{N}}(\tilde{\boldsymbol{z}}_\pi) \ : \ \pi \in \mathbf{G}_N\}$, where $\tilde{\boldsymbol{z}}_\pi$ denotes the action of $\pi \in \mathbf{G}_N$ on $\tilde{\boldsymbol{z}}$, *i.e.*, a permutation of the recentered data. Denote $K_{\text{N}}^{(k)}(\tilde{\boldsymbol{z}})$ as the "critical value."

Step 4. Define numbers $M^+(\tilde{\boldsymbol{z}})$ and $M^0(\tilde{\boldsymbol{z}})$

$$M^+(\tilde{\boldsymbol{z}}) = \left|\{1 \leq j \leq N! : K_{\mathrm{N}}^{(j)}(\tilde{\boldsymbol{z}}) > K_{\mathrm{N}}^{(k)}(\tilde{\boldsymbol{z}})\}\right|$$
$$M^0(\tilde{\boldsymbol{z}}) = \left|\{1 \leq j \leq N! : K_{\mathrm{N}}^{(j)}(\tilde{\boldsymbol{z}}) = K_{\mathrm{N}}^{(k)}(\tilde{\boldsymbol{z}})\}\right| .$$

Step 5. The permutation test is given by

$$\phi(\tilde{\boldsymbol{z}}) = \begin{cases} 1 & K_{\mathrm{N}}(\tilde{\boldsymbol{z}}) > K_{\mathrm{N}}^{(k)}(\tilde{\boldsymbol{z}}) \\ a(\boldsymbol{z}) & K_{\mathrm{N}}(\tilde{\boldsymbol{z}}) = K_{\mathrm{N}}^{(k)}(\tilde{\boldsymbol{z}}) \\ 0 & K_{\mathrm{N}}(\tilde{\boldsymbol{z}}) < K_{\mathrm{N}}^{(k)}(\tilde{\boldsymbol{z}}) \end{cases}, \quad \text{where} \quad a(\tilde{\boldsymbol{z}}) = \frac{N!\,\alpha - M^+(\tilde{\boldsymbol{z}})}{M^0(\tilde{\boldsymbol{z}})} . \tag{10}$$

Thus, the permutation test $\phi(\tilde{\boldsymbol{z}})$ rejects the hypothesis (2) if $K_{\mathrm{N}}(\tilde{\boldsymbol{z}})$ exceeds the "critical value" $K_{\mathrm{N}}^{(k)}(\tilde{\boldsymbol{z}})$, does not reject when $K_{\mathrm{N}}(\tilde{\boldsymbol{z}}) < K_{\mathrm{N}}^{(k)}(\tilde{\boldsymbol{z}})$, and will randomize the decision with probability $a(\tilde{\boldsymbol{z}})$ when $K_{\mathrm{N}}(\tilde{\boldsymbol{z}}) = K_{\mathrm{N}}^{(k)}(\tilde{\boldsymbol{z}})$.

Remark 3. The calculation of the permutation test is computationally prohibitive for moderately large N, which is typically the case in practice. In these scenarios, it is possible to rely on a stochastic approximation without affecting the permutation test's theoretical properties by sampling permutations π from $\mathbf{G}_N$ with or without replacement. More formally, let $\hat{\mathbf{G}}_N = \{\pi_1, \ldots, \pi_M\}$, where π_1 is the identity permutation and $\pi_2, \ldots, \pi_M$ are i.i.d. uniform on $\mathbf{G}_N$. The same construction follows if we replace $\mathbf{G}_N$ with $\hat{\mathbf{G}}_N$, and the approximation is arbitrarily close for M sufficiently large (Romano, 1989, Section 4). From now on we focus on $\mathbf{G}_N$ while in practice we fall back on $\hat{\mathbf{G}}_N$ (e.g. the implementation in Algorithm 2 below). ■

The previous construction yields an exact level α test in finite samples provided that the distributions of $\tilde{\boldsymbol{Z}}$ and $\tilde{\boldsymbol{Z}}_\pi$ are the same under the null hypothesis for any permutation $\pi \in \mathbf{G}_N$ (e.g. Lehmann and Romano, 2022, Theorem 17.2.1). This would be the case if, for example, the researcher knows γ. In such a case, the null hypothesis is sharp, equation (7) becomes $(Y_{1,1} - \gamma, \ldots, Y_{1,m} - \gamma, Y_{0,1}, \ldots, Y_{0,n})$, and it follows that $\tilde{\boldsymbol{Z}}$ and $\tilde{\boldsymbol{Z}}_\pi$ have the same distribution under the null hypothesis.

However, as we argued before, knowing γ is infeasible in most empirically relevant scenarios.

Therefore, we need to resort to large-sample approximations and consider a permutation test with asymptotic rejection probability equal to α for the more general null hypothesis (2). To facilitate the study of the limiting behavior of the permutation test $\phi(\tilde{\boldsymbol{z}})$, it is useful to consider the permutation distribution of the 2SKSQ, defined as follows:

$$\hat{R}_{\mathrm{N}}^{K}(t) = \frac{1}{N!} \sum_{\pi \in \mathbf{G}_N} \mathbb{1}_{\{K_{\mathrm{N}}(\tilde{z}_{\pi(1)},\ldots,\tilde{z}_{\pi(N)}) \leq t\}} \,. \tag{11}$$

Roughly speaking, the permutation test rejects the null hypothesis (2) if $K_{\mathrm{N}}(\tilde{\boldsymbol{z}})$ exceeds the upper α quantile of the permutation distribution. In the next sections, we will study the large-sample behavior of the permutation distribution (11).

3 Asymptotic Results

We now introduce three standard assumptions in the quantile regression literature, which are relevant throughout the paper:

A. 1. *Let* $0 < a < b < 1$. F_0 *is continuously differentiable on the interval* $[F_0^{-1}(a) - \varepsilon, F_0^{-1}(b) + \varepsilon]$ *for some* $\varepsilon > 0$, *with strictly positive derivative* f_0, *and analogously for* F_1.

A. 2. *Let* $N = n + m$, $n \to \infty$, $m \to \infty$, *and* $p_m = m/N \to p \in (0,1)$ *with* $p_m - p = \mathcal{O}(N^{-1/2})$.

A. 3. *There exists estimators of* γ *and* $\varphi(\tau)$, *denoted* $\hat{\gamma}$ *and* $\hat{\varphi}$, *satisfying i)* $\sqrt{N}\{\hat{\gamma} - \gamma\} = \mathcal{O}_p(1)$, *and ii)* $\sup_{\tau \in \mathscr{T}} |\hat{\varphi}(\tau) - \varphi(\tau)| = o_p(1)$.

Several remarks are in place. We would replace Assumption A.2 by the typical full-rank condition (e.g. Koenker and Machado (1999, Assumption A.2)) if we consider covariates as in Section 2.1. However, without covariates, the full-rank condition simplifies to A.2. Moreover, the convergence rates in Assumption A.2 play a key role when we investigate the asymptotic behavior of the permutation distribution (11). Assumption A.3 guarantees we can replace the unknown quantities, γ and $\varphi(\tau)$, with estimates satisfying general assumptions. For example, suppose $\hat{\gamma}$ is given by the OLS estimator of Y on D (see Remark 1). If $\sigma_d^2 \equiv \mathbb{V}(Y_{d,i}) < \infty$, $d \in \{0,1\}$, then $\hat{\gamma}$ satisfies A.3 (i) by the central limit theorem for i.i.d. random variables.

Lastly, it is generally easy to find estimates $\hat{\varphi}$ satisfying A.3 (ii) for F_0 satisfying assumption A.1. For example, the so-called kernel class of estimates has this property (Wand and Jones, 1994). In this paper, we will consider the adaptive estimation and kernel-smoothing methods studied in Portnoy and Koenker (1989). Besides producing an estimate $\hat{\varphi}$ satisfying the uniformity condition A.3 (ii) (see Portnoy and Koenker, 1989, Lemma 3.2), this approach has useful implications for the calculation of the Khmaladze transformation, as we discuss in Section 4.

It is worth mentioning that the asymptotic properties of the permutation test remain unaffected even if we estimate $\varphi(\tau)$, as long as assumption A.3 holds. However, we will see that the previous statement is not true for estimated γ, even if it satisfies A.3 (i). Indeed, the estimated nuisance parameter $\hat{\gamma}$ renders the limiting distribution of $K_{\text{N}}(\boldsymbol{Z})$ intractable because of the dependence on the unknown probability distribution generating the data. Thus, the corresponding permutation test fails to control the Type I error even asymptotically. We formalize these ideas in the next subsection.

3.1 Limiting Null Distribution of $K_{\text{N}}(\boldsymbol{Z})$

The following result is a special case of Koenker and Xiao (2002, Theorem 2) applied to the HTE testing problem in this paper. We include it here as a lemma for completeness. This lemma establishes the asymptotic behavior of the quantile process and 2SKSQ, eqs. (6) and (5) respectively, under the null hypothesis (2). To ease exposition, we collect the definitions of all the processes and their covariance functions, as well as the proof, in the online appendix.

Lemma 1. *Under assumptions A.1–A.3, the process* $\{\hat{v}_{\text{N}}(\tau;\boldsymbol{Z}) : \tau \in \mathscr{T}\}$ *converges weakly in* $\ell^{\infty}(\mathscr{T})$*—the space of bounded functions on* $\mathscr{T}$ *equipped with the uniform norm—to a Gaussian process* $v(\cdot)+\xi(\cdot)$ *under the null hypothesis (2). Here,* $v(\cdot)+\xi(\cdot)$ *has zero mean and covariance function* $\mathbb{C}(v(\tau_1),\xi(\tau_2))$*, given in the online appendix. Furthermore,* $K_N(\boldsymbol{Z})$ *given by (5) converges in distribution to* $K \equiv \sup_{\tau\in\mathscr{T}}|v(\tau)+\xi(\tau)|$ *with CDF given by* $J(t) \equiv Pr\{K \leq t\}$*.*

Several remarks are in order. First, the limit process $v(\cdot)+\xi(\cdot)$ consists of two parts: $v(\cdot)$, a Brownian bridge, and $\xi(\cdot)$, a Gaussian process with zero mean and covariance function

$\mathbb{C}(\xi(\tau_1), \xi(\tau_2)) = \varphi(\tau_1)\varphi(\tau_2)\sigma_0^2$. We can show that if γ was known, but otherwise under the same hypotheses of Lemma 1, the process $\hat{\upsilon}_{\text{N}}(\cdot; \boldsymbol{Z})$ would converge to a Brownian bridge. Yet, the estimation of the nuisance parameter introduced the extra component $\xi(\cdot)$. Second, the covariance function $\mathbb{C}(\upsilon(\tau_1), \xi(\tau_2))$ depends on $f_0(\cdot)$ and $F_0(\cdot)$, which are generally unknown (see eq. (II.3) in the online appendix). Therefore, we cannot simulate the limiting distribution of the test statistic $K_N(\boldsymbol{Z})$, which makes it difficult to use this test in empirical work.

In the next theorem, we establish the asymptotic behavior of the permutation test based on 2SKSQ. We show that the corresponding permutation distribution of the test statistic does not approximate the unconditional distribution of the test statistic. Note that we do *not* assume the null hypothesis (2).

Theorem 1. *Consider testing the hypothesis* (2) *based on the test statistic* (5)*. Under assumptions A.1–A.3, the permutation distribution* (11) *based on the 2SKSQ statistic* $K_{\text{N}}(\boldsymbol{Z})$ *is such that*

$$\sup_{t\in\mathbb{R}}\left|\hat{R}^K_{\text{N}}(t) - G(t)\right| \xrightarrow{\text{p}} 0 \ , \tag{12}$$

where $G(\cdot)$ *is the CDF of* $K^* \equiv \sup_{\tau\in\mathscr{T}}|\upsilon(\tau)|$*, and* $\upsilon(\cdot)$ *is a Brownian bridge process.*

From Theorem 1 we see that the permutation test based on $K_{\text{N}}(\boldsymbol{Z})$ fails to achieve the asymptotic rejection probability of α because the limiting behavior of the sampling and permutation distribution are different. In the next section, we provide a new permutation test based on a martingale transformation of (6).

4 Asymptotically Valid Permutation Test

4.1 Limiting Null Distribution of $\tilde{K}_{\text{N}}(\boldsymbol{Z})$

We present a brief discussion about the martingale transformation of Khmaladze (1981) in this section. For a more detailed discussion, we refer the reader to Koenker and Xiao (2002) and Bai (2003). Let $g(s) = [g_1(s), g_2(s)]' = [s, \varphi(s)]'$ on $[0, 1]$, where $\varphi(\tau) = f_0(F_0^{-1}(\tau))$. Then, the

derivative of g, denoted by $\dot{g}(\cdot)$, is defined to be $\dot{g}(s) = [\dot{g}_1(s), \dot{g}_2(s)]' = [1, (\dot{f}_0/f_0)(F_0^{-1}(s))]'$. The function g is closely connected to the score function. Indeed, we can show that g is the integrated score function of the model (see Bai (2003, Section IV)).

Let $D[0,1]$ be the space of càdlàg functions on $[0,1]$, and denote by $\psi_g(h)(\cdot)$ the compensator of h, $\psi_g : D[0,1] \to D[0,1]$ given by

$$\psi_g(h)(t) = \int_0^t \left[\dot{g}(s)' C(s)^{-1} \int_s^1 \dot{g}(r) dh(r)\right] ds \ , \tag{13}$$

where $C(s) = \int_s^1 \dot{g}(t)\dot{g}(t)' dt$. We can think of $\psi_g(h)(\cdot)$ as the functional analog of the fitted values in the regression context, where the extended score $\dot{g}(r)$ acts as the regressor, and $C(s)^{-1} \int_s^1 \dot{g}(r) dh(r)$ as the orthogonal projection of $dh(r)$ onto $\dot{g}(r)$ over the interval $(s,1]$. Following this interpretation, the residuals at s are given by $dh(s) - \left[\dot{g}(s)' C(s)^{-1} \int_s^1 \dot{g}(r) dh(r)\right] ds$. As explained in Bai (2003, Section III.A), the residuals in the previous display are recursive residuals because we perform a projection at each point s. Therefore, the cumulative sum of the recursive residuals over $[0,t]$ will lead to a Brownian motion. We will show this for our case in Lemma 2 below. This is why Bai (2003) calls the martingale transform a continuous-time detrending operation. We provide more intuition about this interpretation and the numerical calculation of the compensator (13) in more detail in the online appendix (see Section VI).

Following the ongoing discussion, define the two-sample martingale-transformed quantile process of (6) as

$$\tilde{\upsilon}_{\text{N}}(\tau, \boldsymbol{Z}) = \hat{\upsilon}_{\text{N}}(\tau; \boldsymbol{Z}) - \psi_g(\hat{\upsilon}_{\text{N}})(\tau; \boldsymbol{Z}) \ , \tag{14}$$

and the resulting martingale transformation of the 2SKSQ as

$$\tilde{K}_{\text{N}}(\boldsymbol{Z}) = \sup_{\tau \in \mathscr{T}} |\tilde{\upsilon}_{\text{N}}(\tau; \boldsymbol{Z})| \ . \tag{15}$$

Similar to $\varphi(\tau)$, we typically do *not* know function $\dot{g}(s)$ in practice, so we need to estimate it. However, the estimation of $\dot{g}(s)$ will not affect the asymptotic properties of the martingale transformation under the following technical condition holds.

A. 4. *There exists an estimator, $\dot{g}_{\text{N}}(\tau)$, such that* $\sup_{\tau \in \mathscr{T}} |\dot{g}_{\text{N}}(\tau) - \dot{g}(\tau)| = o_p(1)$.

There are several ways to estimate score functions like $\dot{g}(s)$, e.g., smoothing splines (Cox, 1985). In this paper, we employ the adaptive kernel method of Portnoy and Koenker (1989). The reason is two-fold. First, their estimate of the function $\dot{g}(s)$ automatically delivers an estimate of the density $\varphi(s)$. Second, these estimates satisfy the uniformity conditions in A.4 and A.3 (ii), respectively (Portnoy and Koenker, 1989, Lemma 3.2). For alternative estimators with similar uniform convergence properties, see Bhattacharya (1967) and Schuster (1969).

The following result states the asymptotic behavior of the martingale-transformed 2SKSQ statistic. As in the case of Lemma 1, it is a particular case of Koenker and Xiao (2002, Theorem 3) applied to our testing problem. We include it here as a lemma for the sake of exposition. See the online appendix for a proof of this result.

Lemma 2. *Under assumptions A.1–A.4, we have that the process* $\{\tilde{v}_N(\tau) : \tau \in \mathscr{T}\}$ *converges weakly in* $\ell^\infty(\mathscr{T})$ *to* $\zeta(\cdot)$ *under the null hypothesis* (2)*. Here,* $\zeta(\cdot)$ *denotes the standard Brownian motion. Furthermore, the test statistic* $\tilde{K}_N(\boldsymbol{Z})$*, defined in* (15)*, converges in distribution to* $\tilde{K} \equiv \sup_{\tau\in\mathscr{T}}|\zeta(\tau)|$ *with CDF given by* $H(t) \equiv Pr\{\tilde{K} \leq t\}$*.*

Lemma 2 demonstrates that the Khmaladze transformation renders a test statistic whose limiting distribution does not depend on the fundamentals. Therefore, it is possible to carry on valid inference in large samples by simulating the limiting distribution of (15). To gain further intuition as to why the asymptotic shift $\xi(\cdot)$ in Lemma 1 is no longer affecting the asymptotic behavior of the test statistic, we note the compensator (13) *i)* is a linear mapping with respect to h, and *ii)* satisfies $\psi_g(h)(cg) = cg$ for a constant or random variable c (Bai, 2003). Combine these properties with the asymptotic representation in the proof of Lemma 1 and write

$$\begin{aligned}
\tilde{v}_{\text{N}}(\tau;\boldsymbol{Z}) &= \hat{v}_{\text{N}}(\tau;\boldsymbol{Z}) - \psi_g\left(\hat{v}_{\text{N}}\right)(\tau;\boldsymbol{Z}) \\
&= v_{\text{N}}(\tau;\boldsymbol{Z}) + \xi_{\text{N}}(\tau;\boldsymbol{Z}) - \psi_g\left(v_{\text{N}} + \xi_{\text{N}}\right)(\tau;\boldsymbol{Z}) + o_p(1) \\
&= v_{\text{N}}(\tau;\boldsymbol{Z}) - \psi_g(v_{\text{N}})(\tau;\boldsymbol{Z}) + o_p(1)\ .
\end{aligned}$$

From here, we can see the Khmaladze transformation renders the transformed process asymptotically distribution-free (see the proofs in Section II in the online appendix for more details).

Even though the Khmaladze transformation yields an asymptotically pivotal statistic, the

limiting distribution of (15) depends on the norm, the pre-specified $\mathscr{T}$, and the number of covariates. For example, Koenker and Xiao (2002, Appendix A) approximate $\zeta(\cdot)$ by a Gaussian random walk with $20,000$ replications and use the ℓ_1 norm to simulate $\tilde{K}$ and its critical values. In the next Section, we formally introduce our permutation test and show how it offers an off-the-shelf way to generate data-dependent, asymptotically valid "critical values" without simulating the limiting distribution.

4.2 Main Result

We now turn to our main theoretical result—the permutation test based on the martingale-transformed statistic behaves asymptotically like the sampling distribution. Let $\hat{R}_{\mathrm{N}}^{\tilde{K}}$ be the permutation distribution defined in (11) with K_{N} replaced by $\tilde{K}_{\mathrm{N}}$. The following theorem establishes the limiting behavior of $\hat{R}_{\mathrm{N}}^{\tilde{K}}$ and its upper α-quantile $\hat{r}_{\mathrm{N}}(1-\alpha) = \inf\{t : \hat{R}_{\mathrm{N}}^{\tilde{K}}(t) \geq 1-\alpha\}$. Note that we do *not* impose the null hypothesis (2) when deriving the results.

Theorem 2. *Consider testing the hypothesis (2) at level $\alpha \in (0,1)$ based on the test statistic (15). Under assumptions A.1–A.4, the permutation distribution (11) based on the the Khmaladze transformed statistic $\tilde{K}_{\mathrm{N}}$ is such that*

$$\sup_{t\in\mathbb{R}}\left|\hat{R}_{\mathrm{N}}^{\tilde{K}}(t) - H(t)\right| \xrightarrow{\mathrm{p}} 0 \ , \tag{16}$$

where $H(\cdot)$ is the CDF of $\tilde{K}$ defined in Lemma 2. Moreover, if $r(1-\alpha) = \inf\{t : H(t) \geq 1-\alpha\}$, then $\hat{r}_{\mathrm{N}}(1-\alpha) \xrightarrow{\mathrm{p}} r(1-\alpha)$.

Theorem 2 states that the permutation distribution $\hat{R}_{\mathrm{N}}^{\tilde{K}}(\cdot)$ asymptotically approximates the (unconditional) limit distribution of the Khmaladze transformed statistic $\tilde{K}_{\mathrm{N}}$, which is the supremum of a Brownian motion process by Lemma 2. Therefore, the permutation test of the null (2) based on $\tilde{K}_{\mathrm{N}}$ exhibits asymptotic size control. The second part in the conclusion of Theorem 2 states that we can use the upper-α quantiles $\hat{r}_{\mathrm{N}}$ of the permutation distribution $\hat{R}_{\mathrm{N}}^{\tilde{K}}(\cdot)$ as "critical values." This follows from the fact that the distribution of the norm of a Brownian motion is absolutely continuous with a positive density (see online Appendix for more details I.2). We summarize the ongoing discussion in the following corollary:

Corollary 1. *Let $\phi(\boldsymbol{Z})$ be the permutation test as described in (10) with K_N replaced by $\tilde{K}_N$. Under assumptions A.1–A.4, it follows by Theorem 2 that $\mathbb{E}[\phi(Z)] \to \alpha$ under H_0.*

Since the conclusion of Theorem 2 holds irrespective of whether the null hypothesis (2) holds, we can talk about the power properties of the permutation test. Indeed, the permutation test has the same limiting local power as the asymptotic test based on $\tilde{K}_N(\boldsymbol{Z})$ for contiguous alternatives. To see why, observe the asymptotic test rejects the null hypothesis when $\tilde{K}_N(\boldsymbol{Z}) > r(1-\alpha)$. Suppose that $\tilde{K}_N(\boldsymbol{Z})$ converges in distribution to some law $H'(\cdot)$ under some sequence of alternatives that are contiguous to a distribution satisfying the null hypothesis. Then, the power of the test approaches $1 - H'(H^{-1}(1-\alpha))$. Observe that by Theorem 2 the "critical values" from the permutation distribution, $\hat{r}_N$, are such such that $\hat{r}_N \xrightarrow{p} H^{-1}(1-\alpha)$. The same result follows under a sequence of contiguous alternatives, implying the same limiting local power of the test based on $\tilde{K}_N(\boldsymbol{Z})$. Therefore, there is no loss in power when using critical values set by the permutation test.

5 Algorithms and Numerical Implementation

The permutation test we introduce in this paper relies on the whole quantile process so we need to estimate several conditional quantile models as an ensemble, e.g. Section 6. Moreover, this process is repeated for permutations $\pi \in \mathbf{G}_N$ of the data. Therefore, the calculation of our test can be computationally expensive when N is large.

In this section, we cover some algorithmic aspects for estimation with many τ's and π's based on the preprocessing idea of Portnoy and Koenker (1997). Preprocessing substantially reduces the computation burden of our calculations while delivering the same numerical estimates as the standard estimation procedures.[5] We can think of preprocessing in a simple way as follows. Suppose that we have a preliminary solution at some τ^*, e.g., an estimate based on a random subsample of the whole sample. Then, we might use the residuals from this quantile regression to inform the sign of the residuals in the whole sample. We then collect the pseudo-observations

[5]For more complexity results based on worst-case analysis, see Portnoy and Koenker (1997, Sec 5).

with either "'large" negative or "'large" positive residuals into a sample Portnoy and Koenker coined as *glob*. As it turns out, we can remove the "globbed" sample from the optimization problem, thus reducing the effective sample size.

To formalize the ongoing discussion, we borrow notation from Portnoy and Koenker (1997). Fix τ, and suppose that we "knew" that some subset J_L of the observations fall below the hyperplane defined by the check function ρ_τ, and that another subset J_H fall above. Then, the quantile regression problem in (4) yields the same solution as the following revised problem

$$\{\hat{\alpha}(\tau), \hat{\gamma}(\tau)\} = \underset{a,b\in\mathbb{R}}{\arg\min} \sum_{i\notin(J_L\cup J_H)} \rho_\tau\left(Y_i - a - bD_i\right) + \rho_\tau\left(Y_L - a - bD_L\right) + \rho_\tau\left(Y_H - a - bD_H\right) \ , \tag{17}$$

where $D_k = \sum_{i\in J_k} D_k$ for $k \in \{L, H\}$, and Y_L and Y_H can be chosen arbitrarily small or large to ensure that the signs of their corresponding residuals remain negative and positive, respectively. We can now define the globbed observations as (Y_k, D_k), $k \in \{L, H\}$. As we argued before, the revised problem (17) reduces the effective sample size because we withdraw the $\#\{J_L \cup J_H\} - 2$ observations in the globs.

The next algorithm outlines the implementation of preprocessing for a single quantile τ. See Koenker (2020) and Chernozhukov, Fernández-Val, and Melly (2020) for more information about the R and Stata implementations.

Algorithm 1 (Portnoy and Koenker (1997))

1. *Solve for the model* (4) *using a subsample of size* $N_0 = (2N)^{2/3}$. *Denote* $\{\tilde{\alpha}(\tau), \tilde{\gamma}(\tau)\}$ *as the quantile regression estimate of* $\alpha(\tau)$ *and* $\gamma(\tau)$ *based on* N_0.

2. *Calculate the residuals* $\tilde{\varepsilon}_i$ *and a conservative estimate of their standard errors, denoted* $\tilde{s}_i$. *Calculate the* $\tau \pm 1/(2N) \times \theta N_0$ *quantiles of* $\tilde{\varepsilon}/\tilde{s}$. *The parameter* θ *can be taken conservatively to be approximately* 1.

3. *Define the globs by collecting the observations below* $\tau - 1/(2N) \times \theta N_0$ *into* J_L, *and the observations above* $\tau + 1/(2N) \times \theta N_0$ *into* J_H. *Keep the observations between these two quantiles for the next step.*

4. *Solve the revised problem* (17) *and obtain* $\{\hat{\alpha}(\tau), \hat{\gamma}(\tau)\}$.

5. *Verify that all the observations in globs* J_L *and* J_H *have the anticipated residual signs. If all the signs agree with those predicted by the confidence bands: return the optimal solution. If less than* $0.1 \times \theta N_0$ *incorrect signs: adjust the globs by re-introducing these observations into the new globed observations and resolve as in Step 4. If more than* $0.1 \times \theta N_0$ *incorrect signs: go back to step 1 and increase* N_0 *(e.g., double the size).*

Building upon Algorithm 1, Chernozhukov, Fernández-Val, and Melly (2020, Algorithm 2) show how to extend the preprocessing algorithm to many quantiles $\tau_1 < \tau_2 < \cdots < \tau_T$. In a nutshell, their algorithm recursively globs adjacent quantiles, yielding estimates $\{(\hat{\alpha}(\tau_t), \hat{\gamma}(\tau_t)) : 1 \leq t \leq T\}$ for a grid of evenly spaced quantiles $\tau \in \{\tau_1, \ldots, \tau_T\}$. We borrow their insights and show how to apply this idea to accelerate the calculation of a permutation test based on the quantile process in the next subsection.

5.1 Preprocessing for Permutation-based Inference

Suppose we are interested in estimating T quantile regressions for a grid of evenly spaced quantiles $\tau \in \{\tau_1, \ldots, \tau_T\}$ for each permutation $\pi \in \mathbf{G}_N$ of the data. Let $\boldsymbol{Z}_\pi = (Z_{\pi(1)}, \ldots, Z_{\pi(N)})$ be the permuted data for a permutation $\pi \in \mathbf{G}_N$. The next algorithm describes how to estimate $\{(\hat{\alpha}^{\pi_j}(\tau_t), \hat{\gamma}^{\pi_j}(\tau_t)) : 1 \leq t \leq T, 1 \leq j \leq M\}$ for T quantiles and M permutations π of the data (see Remark 3).

Algorithm 2

For each permutation $j = 1, \ldots, M$*, including the identity,*

1. *Take a random permutation of data, denoted by* $\{(Y_{\pi_j(i)}, D_{\pi_j(i)}), 1 \leq i \leq N\}$.

2. *Using the permuted data, estimate* $(\hat{\alpha}^{\pi_j}(\tau_1), \hat{\beta}^{\pi_j}(\tau_1))$ *as in Algorithm 1.*

3. *For each quantile* τ_t*,* $t = 2, \ldots, T$*,*

   a) *Use* $(\hat{\alpha}^{\pi_j}(\tau_{t-1}), \hat{\beta}^{\pi_j}(\tau_{t-1}))$ *as a preliminary estimate.*

b) Calculate the residuals $\hat{\varepsilon}_{\pi_j(i)} = Y_{\pi_j(i)} - \hat{\alpha}^{\pi_j}(\tau_{t-1}) - \hat{\gamma}^{\pi_j}(\tau_{t-1})D_{\pi_j(i)}$*, as well as a conservative estimate of their standard errors, denoted* $\hat{s}_{\pi_j}$*. Calculate the* $\tau \pm 1/(2N) \times \theta(2N)^{1/2}$ *quantiles of* $\hat{\varepsilon}_{\pi_j}/\hat{s}_{\pi_j}$*, where the parameter* θ *is currently set to* 3*.*

c) Define the globs by collecting the observations below $\tau - 1/(2N) \times \theta(2N)^{1/2}$ *into* J_L*, and the observations above* $\tau + 1/(2N) \times \theta(2N)^{1/2}$ *into* J_H*. Keep the observations between these two quantiles for the next step.*

d) Solve the revised problem (17) for permuted data $\{(Y_{\pi_j(i)}, D_{\pi_j(i)}), 1 \leq i \leq N\}$ *and obtain* $(\hat{\alpha}^{\pi_j}(\tau_t), \hat{\gamma}^{\pi_j}(\tau_t))$*.*

e) Verify that all the observations in globs have the anticipated residual signs. If all the signs agree with those predicted by the confidence bands: return $(\hat{\alpha}^{\pi_j}(\tau_t), \hat{\gamma}^{\pi_j}(\tau_t))$*. If less than* $0.1 \times \theta(2N)^{1/2}$ *incorrect signs: adjust the globs by re-introducing these observations into the new globed observations, and resolve as in Step 3.d). If more than* $0.1 \times \theta(2N)^{1/2}$ *incorrect signs: go back to step 3.b) and double* θ*.*

At first glance, it is difficult to see where the speed gains come from. As we argued at the end of the previous subsection, preprocessing with many τ's boils down to globbing data for adjacent quantiles. Therefore, the residuals at τ_{t-1} should be a reasonable predictor for the residuals at τ_t if τ_{t-1} and τ_t are "close." As Chernozhukov, Fernández-Val, and Melly (2020) point out, we can formalize this notion of closeness by assuming $\sqrt{N}(\tau_t - \tau_{t-1}) = \mathcal{O}_p(1)$. Thus, under this closeness assumption, we only need to keep a sample proportional to $N^{1/2}$ in Step 3.c) above, as opposed to $N^{2/3}$ like in Algorithm 1.

6 Monte Carlo Experiments

This section examines the finite sample performance of the proposed test compared to other methods based on the quantile regression process. We adhere to the design in Koenker and Xiao (2002) and Chernozhukov and Fernández-Val (2005). For $1 \leq i \leq N$, we generate the potential outcomes according to $Y_i(0) = \varepsilon_i$ and $Y_i(1) = \delta_i + Y_i(0)$, where $\delta_i = \gamma + \sigma_\gamma Y_i(0)$.

The parameter σ_γ denotes the different levels of heterogeneity, with $\sigma_\gamma = 0$ inducing a constant treatment effect.

In each specification we consider, ε_i, $1 \leq i \leq N$ are i.i.d. according to standard normal, lognormal, and Student's t distribution with 5 degrees of freedom. For the calculation of the quantile process, we consider an equally spaced grid of quantiles $\tau \in \{0.1, 0.15, \ldots, 0.9\}$ and $N \in \{100, 400, 1000\}$. We set $\Pr\{D = 1\} = 0.4$ for Table 1 and $\Pr\{D = 1\} = 0.5$ for Table 2.

We implement our permutation test using R package `RATest`, available on CRAN. To estimate the density and score functions, we used the univariate adaptive kernel density estimation in Portnoy and Koenker (1989), where we adopted the normal kernel and set Silverman's local bandwidth to control the degree of smoothness of the estimate (see Portnoy and Koenker (1989, Section 4) for more details). `RATest` estimates these quantities using the `akj` function from R package `quantreg`. As we argued in Sections 3 and 4, these estimates satisfy the uniformity conditions A.3 (ii) and A.4. Finally, in all of the tests we consider in this section, we calculate $\hat{\gamma}$ as the slope parameter in a linear regression of Y on D (see Remark 1).

In the simulation results in Tables 1–2, we compare the proposed permutation test based on (15)—which we denote **mtPermTest**—against five other alternative tests.

Classical: This is the permutation test based on the 2SKSQ with the true values $\varphi(\tau)$ and γ. Even though this is an infeasible test, we present it as a benchmark case.

Naive KS: This is the 2SKSQ test. We call it *naive* because it ignores the effect that $\hat{\gamma}$ has on the limiting distribution. Thus, this test *naively* relies on the asymptotic critical values simulated from the distribution function of the supremum of a Brownian bridge.

mtQR: This is Koenker and Xiao's (2002) test. We estimate the martingale-transformed test statistic using R package `quantreg`. The asymptotic test uses simulated critical values. See the online appendix in Koenker and Xiao (2002).

Subsampling: This test, proposed by Chernozhukov and Fernández-Val (2005), is based on subsampling the recentered inference process $\sqrt{N}\sup_{\tau\in\mathcal{T}}|\{\hat{\gamma}(\tau) - \gamma(\tau)\} - \{\hat{\gamma} - \gamma\}|$, where $\{\hat{\gamma}(\tau) - \hat{\gamma}\}$ is used itself to "estimate" $\{\gamma(\tau) - \gamma\}$. Arguing as in Chernozhukov and Fernández-

Val (2005, Section 3.4), we set subsampling block size $b = 20 + N^{1/4}$, and 250 bootstrap repetitions within each simulation.

Bootstrap: This test is based on the full-sample bootstrap approximation of the sampling distribution of the 2SKSQ statistic. To do so, we borrow the insights from Linton, Maasoumi, and Whang (2005, Section 6) and apply them to the present context. Arguing as in Ding, Feller, and Miratrix (2016), we recenter treatment and control groups, and sample with replacement from the pooled vector of residuals.

Table 1 reports rejection probabilities under the null hypothesis (2) with $\gamma = 1$. Across specifications, our permutation test exhibits remarkable performance in terms of size control even though we estimate the score and density functions. As expected by the theory, the **Classical** case has empirical rejection probabilities close to the nominal level across specifications. However, we note that **Naive**, **mtQR** and **Subsampling** tests yield rejection probabilities substantially below the nominal level, though subsampling yields rejection rates closer to the nominal level in the normal case as N increases. On the other hand, the **Bootstrap** test shows considerable size distortions across specifications.

Table 2 reports the rejection probabilities under the alternative hypothesis, *i.e.*, $\sigma_\gamma > 0$. We compare the performance of our proposed test with **Subsampling** and **mtQR** for several levels of heterogeneity σ_γ and $\gamma = 1$. We no longer consider the other tests because they are either infeasible or invalid. In all the alternatives we consider, our permutation test is considerably more powerful than **Subsampling** and **mtQR**.

7 Empirical Application

As we emphasized in the introduction, one popular approach to investigating HTE consists of calculating ATEs for a series of subgroups defined by covariates. Under the assumption of constant treatment effects within subgroups, this type of subgroup analysis concludes the existence of HTE if the ATEs vary significantly across subgroups. Thus, we will refer to this type of analysis as the constant subgroup treatment effect (CSTE) model.

Table 1: Size of $\alpha = 0.05$ tests H_0 : Constant Treatment Effect ($\gamma = 1$).

		Distributions		
N	Method	Normal	Lognormal	t_5
	Classical	0.0551	0.0520	0.0478
	Naive	0.0018	0.0008	0.0038
$N = 100$	mtQR	0.0012	0.0004	0.0000
	Subsampling	0.0212	0.0132	0.0192
	Bootstrap	0.0840	0.0838	0.0708
	mtPermTest	0.0478	0.0492	0.0455
	Classical	0.0458	0.0490	0.0548
	Naive	0.0004	0.0010	0.0032
$N = 400$	mtQR	0.0012	0.0074	0.0000
	Subsampling	0.0422	0.0043	0.0136
	Bootstrap	0.0820	0.0862	0.0840
	mtPermTest	0.0480	0.0424	0.0508
	Classical	0.0502	0.0512	0.0514
	Naive	0.0004	0.0004	0.0016
$N = 1000$	mtQR	0.0010	0.0090	0.0000
	Subsampling	0.0474	0.0080	0.0101
	Bootstrap	0.0814	0.0806	0.0818
	mtPermTest	0.0500	0.0526	0.0482

The rejection probabilities based on 5000 replications for the five tests defined in the text, three data generating processes, and three different sample sizes. We use 1000 permutations for the stochastic approximation of the permutation distribution.

Table 2: Power of $\alpha = 0.05$ tests for several levels of heterogeneity σ_γ, and $\gamma = 1$

N	mtQR			Subsampling			mtPermTest		
$n = m$	$\sigma_\gamma = 0$	$\sigma_\gamma = 0.2$	$\sigma_\gamma = 0.5$	$\sigma_\gamma = 0$	$\sigma_\gamma = 0.2$	$\sigma_\gamma = 0.5$	$\sigma_\gamma = 0$	$\sigma_\gamma = 0.2$	$\sigma_\gamma = 0.5$
Normal Outcomes									
100	0.009	0.053	0.497	0.0212	0.054	0.302	0.0472	0.1388	0.4844
400	0.023	0.412	0.997	0.0422	0.308	0.951	0.0480	0.4190	0.9720
800	0.041	0.792	1	0.0384	0.614	1	0.0500	0.7100	1
Lognormal Outcomes									
100	0.0004	0.0322	0.1878	0.0132	0.057	0.302	0.0492	0.1420	0.5122
400	0.0074	0.1844	0.8840	0.0043	0.304	0.970	0.0424	0.4350	0.975
800	0.0092	0.4382	1	0.0320	0.579	1	0.0560	0.7160	1

The rejection probabilities based on 5000 replications for three data generating processes, and three different sample sizes. We use 1000 permutations for the stochastic approximation of the permutation distribution.

Even though the CSTE model is easy to implement and may capture some of the treatment effect variation, it has important limitations. The reason is straightforward: experimental groups may vary in ways beyond the mean. Thus, we may have evidence of ATEs that do not differ across subgroups and yet have HTE, e.g., the groups may differ in higher moments.[6]

The next section describes how to use our method to test the underlying assumption behind the CSTE model, namely, whether the treatment effect is constant within subgroups. More generally, we explain how to use the proposed permutation test to detect for which subgroups, if any, there is heterogeneity in the treatment effect. Given we are interested in simultaneously testing for HTE for a family of subgroups, we account for multiple testing using Holm's step-down procedure. To illustrate the mechanics and motivate our procedure, we analyze experimental data from a welfare reform in Connecticut: Jobs First. We conclude in Section 7.3 that, for most of the families of subgroups we study, the CSTE model's underlying assumption is violated, so the CSTE model does not apply in those cases.

7.1 Multiple Testing for Subgroup Heterogeneity

Our goal is to detect for which subgroups, if any, there is evidence of HTE, *i.e.*, we are interested in simultaneously testing null hypotheses $H_{0,s}$ $(s = 1, \ldots, \mathcal{S})$, where each $H_{0,s}$ represents the null hypothesis of constant treatment effect within subgroup $s \in \{1, \ldots, \mathcal{S}\}$. For simplicity, we assume that the researcher defines these subgroups and we take them as given from now on.

To formally define the problem, we introduce more notation. Let $Y_{0,s}$ and $Y_{1,s}$ denote the control and treatment outcomes for subgroup s, respectively, with corresponding CDFs $F_{0,s}(\cdot)$

[6] Looking at the ATEs across subgroups often cannot capture other forms of treatment effect variation beyond the mean. Bitler, Gelbach, and Hoynes (2006) provide an in-depth account of this phenomenon. In their study, they document how the CSTE model misses detecting the heterogeneous effects of a welfare reform as predicted by a static labor model. For example, this economic model implies effects of opposing signs, but the ATEs obscure this relationship by averaging them together.

and $F_{1,s}(\cdot)$.[7] The QTE for subgroup s is given by

$$\gamma_s(\tau) = F_{1,s}^{-1}(\tau) - F_{0,s}^{-1}(\tau), \ \forall\, \tau \in \mathscr{T} \ . \tag{18}$$

The individual null hypothesis $H_{0,s}$ for s in the family $1 \leq s \leq \mathcal{S}$ is

$$H_{0,s} : \gamma_s(\tau) = \gamma_s \ \forall\, \tau, \ \text{ for some } \ \gamma_s \ ,$$

where γ_s is the unknown subgroup-specific QTE that we need to estimate. For example, we can estimate γ_s using the same ideas we discussed in Remark 1 but applied to data in subgroup s.

Using the data from each mutually exclusive subgroup, we can calculate the p-values for each individual hypothesis $H_{0,s}$—we simply apply the permutation test from Section 4.2 to those data in subgroup s. Suppose that $\hat{p}_s$ is the p-value for testing $H_{0,s}$ using the permutation test based on $\tilde{K}_{\text{N}}$ applied to data belonging to subgroup s. Denote the ordered p-values as $\hat{p}_{(1)} \leq \cdots < \hat{p}_{(\mathcal{S})}$ with associated subgroup-specific ordered null hypotheses $H_{0,(1)}, \ldots, H_{0,(\mathcal{S})}$. We control for the family-wise error rate using the Holm adjustment as follows:

Step 1. If $\hat{p}_{(1)} > \alpha/\mathcal{S}$, do not reject $H_{0,(1)}, \ldots, H_{0,(\mathcal{S})}$ and stop. If $\hat{p}_{(1)} \leq \alpha/\mathcal{S}$, reject $H_{0,(1)}$ and test the remaining $\mathcal{S} - 1$ hypotheses at level $\alpha/(\mathcal{S} - 1)$.
Step 2. If $\hat{p}_{(1)} \leq \alpha/\mathcal{S}$ but $\hat{p}_{(2)} > \alpha/(\mathcal{S}-1)$, do not reject $H_{0,(2)}, \ldots, H_{0,(\mathcal{S})}$ and stop. If $\hat{p}_{(1)} \leq \alpha/\mathcal{S}$ and $\hat{p}_{(2)} \leq \alpha/(\mathcal{S} - 1)$, reject $H_{0,(1)}$ and $H_{0,(2)}$, and test the remaining $\mathcal{S} - 2$ hypotheses at level $\alpha/(\mathcal{S} - 2)$.
And so on.[8]

The Holm step-down procedure allows us to determine for which subgroups there is evidence of heterogeneity in the treatment effect. This follows mechanically from the way the step-down procedure works. We further illustrate this method with the empirical analysis in Section 7.3.

[7] Following BGH, we test the adequacy of the CSTE model using the earnings distribution for those subjects with participation-adjusted, positive earnings. The reason is two-fold. First, it rules out potential violations of assumption A.1 that could render our approach inapplicable. Secondly, CSTE models with different mass points between experimental groups trivially reject the null hypotheses $H_{0,s}$ or (20). For example, suppose $P(Y > 0|D = 1) = p_1$ and $P(Y > 0|D = 0) = p_0$. Assume w.l.o.g. that $p_0 > p_1 > 0$, therefore the QTEs are 0 for quantiles $\tau^* \leq p_1$ whereas the QTEs for quantiles $p_1 < \tau^* \leq p_0$ are generally different than 0.

[8] The Holm procedure only depends on the p-values of the individual tests $s = 1, \ldots, \mathcal{S}$ without imposing dependence assumptions among them. See Chapter 9.1.3 in Lehmann and Romano (2022) for more details.

Remark 4. We can test the underlying assumption behind the CSTE model as a by-product of the proposed multiple testing procedure. Indeed, the CSTE model for a family of subgroups $\{1, \ldots, \mathcal{S}\}$ would apply as long as we fail to reject the joint null $\mathbf{H}_0^{joint}$ given by:

$$\mathbf{H}_0^{joint} : \bigcap_{1 \leq s \leq \mathcal{S}} H_{0,s} \ , \tag{19}$$

where $H_{0,s}$ are defined as before. Observe we would reject the joint hypothesis (19) if any one of the null hypotheses $H_{0,s}$ for subgroup $s \in \{1, \ldots, \mathcal{S}\}$ is rejected. Therefore, we reject the joint null hypothesis (19) if $\hat{p}_{(1)} \leq \alpha/\mathcal{S}$, and fail to reject otherwise (see *Step 1* above). ■

Remark 5. Even though it is not our goal in this paper, we can improve upon Holm's method by incorporating the dependence structure of the individual tests. In our context, for example, one could approximate the distribution of a max-type statistic, where the maximum is taken over the mutually exclusive subgroups $s = 1, \ldots, \mathcal{S}$. See Romano and Wolf (2005) for more details. ■

Remark 6. BGH test a similar, but different set of hypotheses:

$$H_{0,s}^{cdf} : F_{1,s}(y) = F_{0,s}(y - \delta_s) \ , \text{ for some } \delta_s \ , \tag{20}$$

where δ_s is the subgroup-specific treatment effect that we need to estimate (e.g., we can use $\hat{\gamma}_s$). Thus, we reject a joint hypothesis similar to (19) where the individual hypotheses $H_{0,s}$ are replaced by (20). However, we argue that BGH's heuristic justification to establish the asymptotic validity of their test is incorrect. See the online appendix for more details. ■

7.2 Jobs First

During the 1990s, the U.S. passed a series of welfare reforms to boost employment and reduce welfare dependency. Under the new regime, states replaced their Aid to Families with Dependent Children (AFDC) programs with the Temporary Assistance for Needy Families (TANF). In this section, we use data from Connecticut's welfare reform, Jobs First, collected by the Manpower Demonstration Research Corporation (MDRC) and the Connecticut Department of Social Services. The data from this program has two important advantages for our purposes.

First, it summarizes the main features of the welfare reforms (time limits to welfare assistance, financial work incentives, work requirements, and sanctions). Second, nearly 5000 single-parent families from disadvantaged backgrounds with at least one child under age 18 were randomly assigned to the Jobs First (treatment) or to the former AFDC program (control). See Bloom et al. (2002) for a full report on this welfare reform initiative.

As Bitler, Gelbach, and Hoynes (2006) point out, static labor supply predicts heterogeneous responses to the reform. First, the earnings distribution will have a mass point at 0 in both experimental groups. Second, for women with positive earnings, earnings will be greater under the reform over some range of the earnings distribution. However, women on the higher end of the distribution may experience a reduction or no effect on earnings as a result of the reform.

These heterogeneous responses depend on a series of baseline characteristics (prior to the intervention). The MDRC collected data that proxy these individual characteristics, e.g., education, earnings, welfare history, age, marital status, ages of the youngest child,[9] and earnings history. Following BGH, we used these socio-demographic characteristics' levels and their interactions to form the subgroups (see also Table 3 below).

7.3 Results

Table 3 displays the results of our empirical analyses. Each row indicates the covariate's levels, or interaction of covariates' levels, giving rise to a family of subgroups. Thus, each row displays the results of applying our permutation test for different families of subgroups. Column 2 shows the total number of subgroups within each family. With the exception of the first row, all the subgroups consist of the interaction between quarters after random assignment and a demographic characteristic. Therefore, one should understand each row as a family of subgroups defined by quarter-specific covariates' levels. For example, the second row (Education) considers the quarterly levels of *No high-school diploma*, *High-school diploma*, and *More than high-school diploma*, generating the family of 21 subgroups that we find in column 2. The first row in Table 3 (Full sample) has only seven subgroups because in that case we only look at the cross-section

[9] Women are eligible to receive benefits as long as their youngest child is under 18 years of age.

of individuals for each of the seven quarters after the reform.

Columns 3 and 4 in Table 3 indicate the number of individual hypotheses $H_{0,s}$ that our test rejects at 10% and 5%, respectively. Following Section 7.1, we account for multiple testing across subgroup configurations using the Holm adjustment (for completeness, we also include Bonferroni corrections in the online Appendix).[10] Lastly, following Remark 4, column 5 displays whether our proposed test rejects (19) at 5%.[11] Thus, from a practitioner's point of view, column 5 indicates whether the CSTE model applies (✗) or not (✓).

Table 3: Testing for Heterogeneity in the Treatment Effect by Subgroups, Time-varying mean treatment effects by subgroup with participation adjustment.

Subgroup	Number of Tests	Number of Rejections at 10%	Number of Rejections at 5%	mtPermTest Rejects H_0^{joint} at 5%
Full Sample	7	4	3	✓
Education	21	3	1	✓
Age of youngest child	21	4	3	✓
Marital status	21	6	2	✓
Earnings level seventh Q pre-RA	21	0	0	✗
Number of pre-RA Q with earnings	21	1	0	✗
Welfare receipt seventh Q pre-RA	14	2	2	✓
Education subgroups interacted with				
Age of youngest child	49	6	5	✓
Marital status	35	4	2	✓
Earnings level seventh Q pre-RA	63	0	0	✗
Number of pre-RA Q with earnings	63	2	1	✓
Welfare receipt seventh Q pre-RA	42	1	0	✗
Age of youngest child interacted with				
Marital status	35	3	1	✓
Earnings level seventh Q pre-RA	63	2	0	✗
Number of pre-RA Q with earnings	49	3	1	✓
Welfare receipt seventh Q pre-RA	42	1	0	✗
Marital status subgroup interacted with				
Earnings level seventh Q pre-RA	63	1	0	✗
Number of pre-RA Q with earnings	63	2	0	✗
Welfare receipt seventh Q pre-RA	42	2	1	✓
Earnings level seventh Q pre-RA subgroups interacted with				
Number of pre-RA Q with earnings	49	2	0	✗
Welfare receipt seventh Q pre-RA	42	1	1	✓
Number of quarters any earnings pre-RA subgroup interacted with				
Welfare receipt seventh Q pre-RA	42	2	1	✓

We report the results after adjusting for the multiplicity of tests within the families of subgroups using the Holm adjustment. For the calculation of the quantile process, we consider an equally spaced grid of quantiles $\tau \in \{0.1, 0.15, \ldots, 0.85, 0.9\}$. For the calculation of our test, we estimate the density and score functions using the univariate adaptive kernel density estimation. The stochastic approximation of the permutation distribution is based on 1000 permutations.

[10] We note that the stepwise Holm procedure can be conservative. This is so because Holm's adjustment does not take into account the dependence structure of the individual p-values, i.e., it assumes the worst-case dependence structure(Romano and Wolf, 2005). See Remark 5.

[11] Recall that our test rejects the joint hypothesis (19) if any of the null hypotheses $H_{0,s}$ is rejected.

Table 3 provides evidence against the joint null hypothesis (19) for many of the families of subgroups in our analysis. The first seven rows in Table 3 constitute the main covariates in our empirical analysis. As discussed by BGH, these covariates are the main drivers of heterogeneity as predicted by the economic theory. The rest of the rows in the table are simply interactions of the covariates in the first seven rows. For example, the first row represents the time effect of the welfare reform (number of quarters after the reform) and gives rise to a family of 7 subgroups. In this case, our test rejects 4 (3) of the 7 individual hypotheses $H_{0,s}$ at 10% (5%), so we have evidence against the joint null hypothesis (19), as we display in column 5. Analogously, if we consider the family of subgroups by the quarterly levels of education, we can see that our proposed method rejects 3 (1) out of 21 individual hypotheses at the 10% (5%), so we have evidence against (19). A similar conclusion follows if we apply our test to the families of subgroups defined by quarterly levels of the age of the youngest child, marital status, and welfare receipt seven quarters before the intervention (rows 3, 4, and 7 in Table 3). However, observe that when we apply our test to the subgroups defined by the number of quarters with earnings prior to the random assignment (row 6), we only reject one of the individual hypotheses at 10%, so we only reject (19) at 10% but not at 5%. Lastly, we note that when we apply our test to the family of subgroups defined by the earnings level seven months prior to the Jobs First welfare reform, we reject none of the individual hypotheses. Therefore, in this case, we cannot reject the joint null hypothesis (19).

The remainder of the table displays the results of our test applied to a series of quarterly subgroups by interacting the covariates' levels for the baseline characteristics of interest. It is not surprising that the number of subgroups increases with the number of interactions, yielding in some cases families with 63 subgroups (e.g., seven quarters by three earnings levels by three education levels). However, the general conclusion is qualitatively the same as before, *i.e.*, we reject the joint hypothesis (19).

Table 3 has another important implication for practitioners. By rejecting (19) for many of the families of subgroups we consider, our permutation test shows strong evidence of a systematic violation of the fundamental assumption behind the CSTE model for many of the

families of subgroups under consideration. Therefore, despite its appeal, one should be more careful when applying the CSTE model to investigating treatment effect variation.

8 Conclusions

The permutation test we introduced here provides a means of conducting asymptotically valid inference for HTE using QTE. Our test procedure relies on a modified version of the quantile process to handle the Durbin problem. Numerical evidence in this paper indicates that our permutation test outperforms the alternative quantile-based test. We provide easy-to-implement free software and discuss its fast implementation using the preprocessing algorithm.

On the empirical side, we illustrate our test using experimental data from Connecticut's Jobs First. In this application, we challenge a common practice in empirical work that seeks to investigate treatment effect variation by estimating ATEs across subgroups defined by covariates' levels. Our empirical findings provide evidence against the underlying assumption behind this practice, limiting its applicability in such cases.

References

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

Bai, J. (2003). Testing parametric conditional distributions of dynamic models. *Review of Economics and Statistics*, 85(3):531–549.

Bertanha, M. and Chung, E. (2022). Permutation tests at nonparametric rates. *Journal of the American Statistical Association*, (Forthcoming):1–34.

Bhattacharya, P. K. (1967). Estimation of a probability density function and its derivatives. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 373–382.

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

Bitler, M. P., Gelbach, J. B., and Hoynes, H. W. (2017). Can variation in subgroups' average treatment effects explain treatment effect heterogeneity? evidence from a social experiment. *Review of Economics and Statistics*, 99(4):683–697.

Bloom, D., Scrivener, S., Michalopoulos, C., Morris, P., Hendra, R., Adams-Ciardullo, D., and Walter, J. (2002). Jobs first: Final report on connecticut's welfare reform initiative. *Manpower Demonstration Research Corporation*.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernandez-Val, I. (2018). Generic machine learning inference on heterogenous treatment effects in randomized experiments. Technical report, National Bureau of Economic Research.

Chernozhukov, V. and Fernández-Val, I. (2005). Subsampling inference on quantile regression processes. *Sankhyā*, pages 253–276.

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2020). Fast algorithms for the quantile regression process. *Empirical Economics*, pages 1–27.

Chung, E. and Olivares, M. (2021). Permutation test for heterogeneous treatment effects with a nuisance parameter. *Journal of Econometrics*, 225(2):148–174.

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507.

Chung, E. and Romano, J. P. (2016). Multivariate and multiple permutation tests. *Journal of Econometrics*, 193(1):76–91.

Cox, D. D. (1985). A penalty method for nonparametric estimation of the logarithmic derivative of a density function. *Annals of the Institute of Statistical Mathematics*, 37(2):271–288.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405.

Ding, P., Feller, A., and Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

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

Durbin, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *The Annals of Statistics*, pages 279–290.

Durbin, J. (1975). Kolmogorov-smirnov tests when parameters are estimated with applications to tests of exponentiality and tests on spacings. *Biometrika*, pages 5–22.

Durbin, J. (1985). The first-passage density of a continuous gaussian process to a general boundary. *Journal of Applied Probability*, 22(1):99–122.

Frölich, M. and Sperlich, S. (2019). *Distributional Policy Analysis and Quantile Treatment Effects*. Cambridge University Press.

Janssen, A. (1997). Studentized permutation tests for non-iid hypotheses and the generalized behrens-fisher problem. *Statistics & probability letters*, 36(1):9–21.

Khandker, S. R., Koolwal, G. B., and Samad, H. A. (2009). *Handbook on impact evaluation: quantitative methods and practices*. World Bank Publications.

Khmaladze, E. V. (1981). Martingale approach in the theory of goodness-of-fit tests. *Theory of Probability & Its Applications*, 26(2):240–257.

Khmaladze, E. V. (1993). Goodness of fit problem and scanning innovation martingales. *The Annals of Statistics*, 21(2):798–829.

Koenker, R. (2020). Quantile regression methods: An r vinaigrette.

Koenker, R. and Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association*, 94(448):1296–1310.

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

Lehmann, E. L. (1974). *Nonparametrics: statistical methods based on ranks*. San Francisco: Holden-Day.

Lehmann, E. L. and Romano, J. P. (2022). *Testing statistical hypotheses*. Springer.

Linton, O., Maasoumi, E., and Whang, Y.-J. (2005). Consistent testing for stochastic dominance under general sampling schemes. *The Review of Economic Studies*, 72(3):735–765.

Neuhaus, G. (1993). Conditional rank tests for the two-sample problem under random censorship. *The Annals of Statistics*, 21(4):1760–1779.

Portnoy, S. and Koenker, R. (1989). Adaptive l-estimation for linear models. *The Annals of Statistics*, pages 362–381.

Portnoy, S. and Koenker, R. (1997). The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, 12(4):279–300.

Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *The Annals of Statistics*, pages 141–159.

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

Schuster, E. F. (1969). Estimation of a probability density function and its derivatives. *The Annals of Mathematical Statistics*, 40(4):1187–1195.

Shorack, G. R. and Wellner, J. A. (2009). *Empirical processes with applications to statistics*. SIAM.

Van der Vaart, A. W. and Wellner, J. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.

Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing*. CRC press.

Zhang, Y. and Zheng, X. (2020). Quantile treatment effects and bootstrap inference under covariate-adaptive randomization. *Quantitative Economics*, 11(3):957–982.