Quantile-based Test for Heterogeneous Treatment Effects

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

Mauricio Olivares
Department of Economics
University College London
mauricio.olivares@ucl.ac.uk

February 25, 2022

Abstract

One way to look at the distributional effects of a policy intervention comprises estimating the quantile treatment effect at different quantiles. We exploit this idea and develop a new permutation test for heterogeneous treatment effects based on a modified quantile process. To establish asymptotic validity of our test, we transform the test statistic using a martingale transformation so that its limit behavior is distribution free. Numerical evidence shows our permutation test outmatches other popular quantile-based tests in terms of size and power performance. 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."* The second author 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

A large part of the literature on program evaluation examines heterogeneous treatment effects (HTE) using subgroup analysis. This approach, which we refer to as the constant subgroup treatment effect (CSTE) model, estimates average treatment effects (ATEs) that differ across subgroups defined by covariates, but remain constant within them. Thus, the CSTE model implicitly assumes constant treatment effects within subgroups and treatment effect variation follows from ATEs varying significantly across subgroups. According to a survey of published papers in top economics journals, 40% of the studies report at least one treatment effect this way (Chernozhukov et al., 2018). This practice has motivated the development of tests for HTE in various settings. For example, Crump et al. (2008) test for HTE by checking whether the ATEs conditional on covariates are identical for all subgroups.

However, the CSTE model has been under scrutiny. 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.

An alternative way to look at the distributional effects of a policy comprises estimating the quantile treatment effect (QTE) at different quantiles (Doksum, 1974; Lehmann, 1974). Thus, we can test for HTE by checking whether the QTEs are constant across quantiles, *i.e.*, equal to an unknown constant γ across quantiles. However, 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 the so-called Durbin problem (Durbin, 1973).

In this paper, we introduce a new permutation test for HTE based on the quantile process. Even though our setup contains unknown nuisance parameters, we establish the asymptotic validity of our method under weak conditions. 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 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 provide strong evidence in favor of HTE across subgroups defined by pre-treatment characteristics. Indeed, we reject the null hypothesis of constant treatment effects for a series of these subgroups, thus challenging the methodological soundness of the CSTE to account for the treatment effect variation. This 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, our permutation test formally addressed the Durbin problem, whereas BGH's inference method does this only heuristically. We elaborate on this second point in Section 7.1 and in Appendix II.

The present paper comes 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 therefore classify the literature into two branches. The main difference between them is that 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). For the second approach, one could use bootstrap methods like those presented in Linton, Maasoumi, and Whang (2005) and apply them to our context. Another possibility comprises 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 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). Second, we can incorporate baseline covariates for conditional quantile estimation straightforwardly by using standard quantile regression models (see Section 2.1). Third, 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. 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 no longer controls 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 in the Appendix. We leave the proofs of the auxiliary lemmas and additional discussion with regards our empirical application to 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 the treatment effect given by $\delta_i = Y_i(1) - Y_i(0)$, and we are interested in testing the hypothesis of whether the treatment effect varies across units. More formally, the null hypothesis of constant treatment effects states that

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

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 units in the treatment and control groups, respectively. The QTE is given by

$$\gamma(\tau) = F_1^{-1}(\tau) - F_0^{-1}(\tau), \;\; \forall \, \tau \in \mathscr{T} \;\; ,$$

where $\mathscr{T}$ is a closed subinterval of $(0, 1)$ and $F_d^{-1}(\tau) = \inf\{y \; : \; F_d(y) \leq y\}$, $d \in \{0, 1\}$. Then, the testable hypothesis of constant treatment effect is

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

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) without any other assumption, so a test that rejects H_0 will reject the more restrictive sharp null of constant treatment effects.[1]

One natural candidate for a test statistic for hypothesis (2) is to compare the empirical quantile functions based on two independent random samples drawn 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.[2] Collect all these outcomes in one vector as $\boldsymbol{Z} = (Y_{1,1}, \ldots, Y_{1,m}, Y_{0,1}, \ldots, Y_{0,n}) = (Z_1, \ldots, Z_N)$. Then, in the two-sample setting, $\gamma(\tau)$ is estimable by

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

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 $Y_{0,1}, \ldots, Y_{0,n}$.

We can estimate the difference in quantiles (3) by estimating the individual coefficient associated with the policy indicator in the conditional quantile regression model

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

Then, we may base our analysis on the quantile regression estimates given by

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

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.

Remark 1. Observe that the nuisance parameter γ in (2) is unknown and thus one needs to estimate it. We can estimate γ in this case by integrating the individual coefficient, *i.e.* $\hat{\gamma} = \int_{\mathscr{T}} \hat{\gamma}(\tau)$, or by the OLS estimator of a regression of Y on D. Even though the integrated coefficients yield approximately the mean effects as estimated by the associated least squares

[1] Alternatively, one might 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).

[2] That is, $Y_i^1 = Y_i$ among the treated, and $Y_i^0 = Y_i$ 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.

estimates, one should be cautious about this interpretation, e.g. in the presence of outliers. ■

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

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

where

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

is the standardized quantile regression process for heterogeneous treatment effects, $\hat{\varphi}(\tau)$ is an estimate of $\varphi(\tau) = f_0(F_0^{-1}(\tau))$ satisfying assumption A.3 below, and $\hat{\gamma}$ is an estimator of the nuisance parameter γ, e.g., those we describe in Remark 1.

2.1 Adding Covariates

In practice, we typically observe a vector of baseline covariates besides D. In this case, we can still test the hypothesis of heterogeneous treatment effects (2) while allowing the outcome to depend on other pre-treatment characteristics. To describe how to handle the baseline covariates in our analysis, we introduce additional notation. Let $X \in \mathbb{R}^d = (D, X_{-1})$ denote the vector of covariates, where X_{-1} contains the pre-treatment characteristics.

The linear quantile regression model is now given by

$$F^{-1}_{Y|X}(\tau) = X'\beta(\tau) = \gamma(\tau)D + X'_{-1}\alpha(\tau) \ , \ \ \tau \in \mathscr{T} \ .$$

We consider a linear hypothesis of the form

$$\boldsymbol{R}(\tau)\beta(\tau) - \boldsymbol{r}(\tau) = 0, \, \tau \in \mathscr{T} \ ,$$

where $\boldsymbol{R}(\tau)$ denotes a $q \times d$ matrix, $q \leq d$, and $\boldsymbol{r}(\tau)$ is a $q \times 1$ vector. The quantile regression estimates are then

$$\hat{\beta}(\tau) = \arg\min_{b \in \mathbb{R}^d} \sum_{i=1}^{N} \rho_\tau \left(Y_i - X_i' b\right) \ , \ \tau \in \mathscr{T} \ .$$

Note that (2) is equivalent to setting $\boldsymbol{R} = [1, 0, \dots, 0]$ and $\boldsymbol{r}(\tau) = \gamma$ in this context. Thus, the theoretical results in sections 3–4 remain the same in the presence of covariates X_{-1} for these values $\boldsymbol{R}$ and $\boldsymbol{r}(\tau)$, provided some standard regularity conditions on X_{-1} hold (e.g. Koenker and Machado, 1999, Assumption 2). See Koenker and Xiao (2002, Theorems 2 and 3), and Chernozhukov and Fernández-Val (2005, Proposition 1) for further details. Therefore, we will focus on the case with no covariates when deriving the asymptotic properties of our proposed method for the sake of simplicity, while keeping in mind that adding them is straightforward.

2.2 Permutation Test based on the Quantile Process

Before turning to the theoretical results, we first show how the construction of a permutation tests to asses (2) works. To define the test, we introduce further notation. Let $\mathbf{G}_N$ be the set of all permutations π of $\{1, \dots, N\}$, with $|\mathbf{G}_N| = N!$. Given $\boldsymbol{Z} = \boldsymbol{z}$, recompute $K_{\text{N}}(\boldsymbol{z})$ for all permutations $\pi \in \mathbf{G}_N$ and denote by $K_{\text{N}}^{(1)}(\boldsymbol{z}) \leq K_{\text{N}}^{(2)}(\boldsymbol{z}) \leq \cdots \leq K_{\text{N}}^{(N!)}(\boldsymbol{z})$ the ordered values of $\{K_{\text{N}}(\boldsymbol{z}_\pi) : \pi \in \mathbf{G}_N\}$, where $\boldsymbol{z}_\pi$ denotes the action of $\pi \in \mathbf{G}_N$ on $\boldsymbol{z}$.

Let $k = N! - \lfloor N!\, \alpha \rfloor$ and define

$$\begin{aligned} M^+(\boldsymbol{z}) &= \left| \{1 \leq j \leq N! : K_{\text{N}}^{(j)}(\boldsymbol{z}) > K_{\text{N}}^{(k)}(\boldsymbol{z})\} \right| \\ M^0(\boldsymbol{z}) &= \left| \{1 \leq j \leq N! : K_{\text{N}}^{(j)}(\boldsymbol{z}) = K_{\text{N}}^{(k)}(\boldsymbol{z})\} \right| \ . \end{aligned}$$

Using this notation, the permutation test is given by

$$\phi(\boldsymbol{z}) = \begin{cases} 1 & K_{\text{N}}(\boldsymbol{z}) > K_{\text{N}}^{(k)}(\boldsymbol{z}) \\ a(\boldsymbol{z}) & K_{\text{N}}(\boldsymbol{z}) = K_{\text{N}}^{(k)}(\boldsymbol{z}) \\ 0 & K_{\text{N}}(\boldsymbol{z}) < K_{\text{N}}^{(k)}(\boldsymbol{z}) \end{cases} , \quad \text{where} \quad a(\boldsymbol{z}) = \frac{N!\, \alpha - M^+(\boldsymbol{z})}{M^0(\boldsymbol{z})} \ . \tag{7}$$

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

Remark 2. The computation 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 under a group invariance assumption. For example, suppose the researcher knows δ. In this case, we would be able to input all the missing values in the science table using the observed outcomes and the null hypothesis (1). That is, H_0^s becomes a sharp null, and therefore we can show the permutation test $\phi(\boldsymbol{z})$ is size α (e.g. Lehmann and Romano, 2005, Theorem 15.2.1).

However, as we argued before, knowing δ is infeasible in most empirically relevant scenarios. Therefore, we need to resort to large-sample approximations and instead 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 $\phi(\boldsymbol{z})$, it is useful to consider the permutation distribution of the 2SKSQ, defined as follows:

$$\hat{R}_{\mathrm{N}}^{K}(t) = \frac{1}{N!} \sum_{\pi \in \mathbf{G}_N} \mathbb{1}_{\{K_{\mathrm{N}}(z_{\pi(1)}, \ldots, z_{\pi(N)}) \leq t\}} \ . \tag{8}$$

Roughly speaking, the permutation test rejects the null hypothesis (2) if $K_{\mathrm{N}}(\boldsymbol{z})$ exceeds the upper α quantile of the permutation distribution. In the next sections, we will use (8) to derive the large-sample properties of the permutation test.

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 if we consider* F_1.

A. 2. *Let* $n \to \infty$, $m \to \infty$, *with* $N = n + m$, $p_m = m/N$, *and* $p_m \to p \in (0, 1)$ *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)* $\sup_{\tau\in\mathscr{T}}|\hat{\varphi}(\tau) - \varphi(\tau)| = o_p(1)$.

Assumption A.2 replaces the typical full-rank condition of the design matrix in the quantile regression literature, e.g., Koenker and Machado (1999, Assumption A.2). Also, these convergence rates play a key role when we investigate the asymptotic behavior of the permutation distribution (8). Assumption A.3 guarantees we can replace the unknown quantities, γ and $\varphi(\tau)$, with estimates satisfying general assumptions. For example, if $\sigma_d^2 \equiv \mathbb{V}(Y_{d,i}) < \infty$, $d \in \{0, 1\}$, then $\hat{\gamma}$ as in Remark 1 satisfies A.3 (i) by the central limit theorem for i.i.d. variables.

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 we cannot say the same about γ. Indeed, the estimated nuisance parameter $\hat{\gamma}$ renders the limiting distribution of $K_{\mathrm{N}}(\boldsymbol{Z})$ 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 $K_{\mathrm{N}}(\boldsymbol{Z})$

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

Several remarks are in order. First, the limit process $\upsilon(\cdot) + \xi(\cdot)$ consists of two parts: $\upsilon(\cdot)$, 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 $\hat{\upsilon}_{\mathrm{N}}(\cdot; \boldsymbol{Z})$ 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. (I.3) in the online appendix). Therefore, we cannot simulate 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.

Theorem 1. *Consider testing the hypothesis* (2) *based on the test statistic* (5)*. Under assumptions A.1–A.4, the permutation distribution* (8) *based on the 2SKSQ statistic* $K_{\mathrm{N}}(\boldsymbol{Z})$ *is such that*

$$\sup_{0\leq t\leq 1}\left|\hat{R}^K_{\mathrm{N}}(t) - G(t)\right| \xrightarrow{\mathrm{p}} 0 \ , \tag{9}$$

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 $K_{\mathrm{N}}(\boldsymbol{Z})$ fails to achieve the asymptotic rejection probability of α as the limiting behavior of the sampling distribution and the permutation distribution are different. In the next section, we provide a new permutation test based on a martingale transformation of (6). This transformation ensures that the modified statistic is asymptotically distribution free. We demonstrate that a permutation test based on the modified statistic achieves asymptotic rejection probability equal to α.

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); Bai (2003); Chung and Olivares (2021). Let $g(s) = [g_1(s), g_2(s)] = [s, \varphi(s)]'$ on $[0,1]$, and $\dot{g}(s) = [\dot{g}_1(s), \dot{g}_2(s)]'$ so that $\dot{g}$ is the derivative of g. Then $\dot{g}(s) = [1, (\dot{f}_0/f_0)(F_0^{-1}(s))]'$. Function g previously defined is closely connected with the score function. Indeed, it can be shown that g is the integrated score function of the model (see remarks after assumption A2 in Bai (2003)).

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

$$\psi_g(h)(t) = \int_0^t \left[\dot{g}(s)'C(s)^{-1} \int_s^1 \dot{g}(r)dh(r)\right] ds \ , \tag{10}$$

where $C(s) = \int_s^1 \dot{g}(t)\dot{g}(t)'dt$. We can think of $\psi_g(h)(\cdot)$ as the functional equivalent of the fitted values in a linear regression, where the extended score $\dot{g}(s)$ acts as the regressor, and $C(s)^{-1} \int_s^1 \dot{g}(r)dh(r)$ as the OLS estimator. See Chung and Olivares (2021, Sections 3.3 and 3.4) for more details about the numerical calculation of the Khmaladze transformation.

The two-sample martingale-transformed quantile process (6) and 2SKSQ are given by

$$\tilde{\upsilon}_{\text{N}}(\tau, \boldsymbol{Z}) = \hat{\upsilon}_{\text{N}}(\tau; \boldsymbol{Z}) - \psi_g(\hat{\upsilon}_{\text{N}})(\tau; \boldsymbol{Z}) \ , \quad \text{and} \tag{11}$$

$$\tilde{K}_{\text{N}}(\boldsymbol{Z}) = \sup_{\tau \in \mathscr{T}} |\tilde{\upsilon}_{\text{N}}(\tau; \boldsymbol{Z})| \ , \tag{12}$$

respectively. 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 if the following technical condition holds.

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

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.

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 (12), 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 (12). However, the limiting distribution 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.

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. We seek the limiting behavior of $\hat{R}^{\tilde{K}}_{m,n}$ and its upper α-quantile, which we now denote $\hat{r}_{m,n}(1-\alpha)$, where $\hat{r}_{m,n}(1-\alpha)=\inf\{t:\hat{R}^{\tilde{K}}_{m,n}(t)\geq 1-\alpha\}$.

The following theorem shows that the proposed test is asymptotically valid, *i.e.*, the permutation distribution based on $\tilde{K}_{\mathrm{N}}(\boldsymbol{Z})$ behaves like the supremum of a standard Brownian motion process. Consequently, the α-upper quantiles $\hat{r}_{m,n}$ can be used as "critical values" for the modified test statistic. See Appendix A.2 for a proof. Note that (2) is not assumed.

Theorem 2. *Consider testing the hypothesis* (2) *based on the test statistic* (12)*. Under assumptions A.1–A.4, the permutation distribution* (8) *based on the the Khmaladze transformed statistic $\tilde{K}_{\mathrm{N}}$ is such that*

$$\sup_{0 \le t \le 1} \left| \hat{R}_{\mathrm{N}}^{\tilde{K}}(t) - H(t) \right| \xrightarrow{\mathrm{P}} 0 \ , \tag{13}$$

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 $\hat{r}_{m,n}(1-\alpha) \xrightarrow{\mathrm{P}} r(1-\alpha)$.

This result states that the permutation distribution asymptotically approximates the sampling distribution of the Khmaladze transformed statistic. Therefore, the proposed permutation test exhibits asymptotic size control, as we summarize in the following corollary

Corollary 1. *Consider testing the hypothesis* (2) *based on a permutation test $\phi(Z)$ which rejects when $\tilde{K}_{\mathrm{N}} > \hat{r}_{m,n}(1-\alpha)$, does not reject when $\tilde{K}_{\mathrm{N}} < \hat{r}_{m,n}(1-\alpha)$, and possibly randomizes when $\tilde{K}_{\mathrm{N}} = \hat{r}_{m,n}(1-\alpha)$. Hence, it follows that Theorem 2 implies $\mathbb{E}\left[\phi(Z)\right] \to \alpha$.*

Remark 3. The asymptotic uniform equivalence in Theorem 2 implies the permutation test has the same limiting local power as the asymptotic test based on $\tilde{K}_{\mathrm{N}}(\boldsymbol{Z})$ for contiguous alternatives. Therefore, there is no loss in power when using the permutation-based critical values. ■

The permutation test we propose has two additional features that make it useful in empirical applications. First, the asymptotic validity of the proposed permutation test prevails even though we need to estimate the density and score functions to calculate the Khmaladze transformation. Second, our testing procedure offers an off-the-shelf way to generate data-dependent, asymptotically valid critical values without simulating the limiting distribution that depends on user-specific tuning parameters. Summing up, our permutation test delivers a robust method to carry on asymptotically valid inference for (2) in the presence of estimated nuisance parameters.

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.[3] 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 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, (4) yields the same solution as the following revised problem

$$\arg\min_{a,b\in\mathbb{R}} \sum_{i\notin(J_L\cup J_H)} \rho_\tau\left(Y_i - a - bD_i\right) + \rho_\tau\left(Y_L - a - bD_L\right) + \rho_\tau\left(Y_H - a - bD_H\right) \ , \tag{14}$$

where $D_k = \sum_{i\in J_k} D_i$ 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 (14) reduces the effective sample size because we withdraw the $\#\{J_L \cup J_H\} - 2$

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

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* $\{\tilde{\alpha}(\tau), \tilde{\beta}(\tau)\}$ *as the quantile regression estimate of* $\alpha(\tau)$ *and* $\gamma(\tau)$ *based on* N_0.

2. *Calculate the residuals* $\hat{\varepsilon}_i$ *and a conservative estimate of their standard errors, denoted* $\hat{s}_i$. *Calculate the* $\tau \pm \; (1/2N) \times \theta(2N)^{2/3}$ *quantiles of* $\hat{\varepsilon}/\hat{s}$. *The parameter* θ *can be taken conservatively to be approximately* 1.

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

4. *Solve the revised problem* (14) *and obtain* $\{\hat{\alpha}(\tau), \hat{\beta}(\tau)\}$.

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* $0.1 \times \theta(2N)^{2/3}$ *incorrect signs: adjust the globs by re-introducing these observations into the new globed observations, and resolve as in Step 4. If more than* $0.1 \times \theta(2N)^{2/3}$ *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 many τ's and permutations π of the data

Algorithm 2

1) *Estimate $\{(\hat{\alpha}(\tau_t), \hat{\gamma}(\tau_t)) : 1 \leq t \leq T\}$ as in Chernozhukov, Fernández-Val, and Melly (2020, Algorithm 2). Then, for $t = 1, \ldots, T$, do*

   a) *Take a random permutation of data $\boldsymbol{Z}_{\pi_j}$.*

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

   c) *Define the globs as in Algorithm 1, Step 3.*

   d) *Solve the revised problem (14) for permuted data $\boldsymbol{Z}_{\pi_j}$ and obtain $\hat{\alpha}^{\pi_j}(\tau), \hat{\gamma}^{\pi_j}(\tau)$.*

   e) *Verify that all the observations in globs J_L and J_H have the anticipated residual signs as in Step 5 in Algorithm 1.*

   f) *Repeat Steps a)–e) above for $j = 1, \ldots, M$.*

At first glance, it is difficult to see where the speed gains come from. This is embedded in the first step of Algorithm 2. 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 $N^{1/2}$ as opposed to $N^{2/3}$, like in Algorithm 1.

6 Monte Carlo Experiments

We present numerical evidence to examine 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 calculate our permutation test using R package `RATest` from CRAN. We estimate the density and score functions using the univariate adaptive kernel density estimation *á la Silverman*, which satisfies the uniformity requirements in A.3 (ii) and A.4 (Portnoy and Koenker, 1989, Lemma 3.2). We estimate γ by OLS in all of the tests we consider in this section.

In the simulation results in Tables 1–2, we compare the proposed permutation test based on (12)—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 Koenker and Xiao's (2002) test. We estimate the martingale-transformed test statistic using R package `quantreg`.

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 an application of Linton, Maasoumi, and Whang (2005, Section 6). It is based on the full-sample bootstrap approximation of the sampling distribution of the 2SKS statistic. 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 a 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

One popular approach to investigating treatment effect heterogeneity is to estimate ATEs that vary across the subgroups but remain constant within the subgroups. Thus, if the ATEs are significantly different across subgroups, then one claims evidence of HTE. This is what we referred to as the constant subgroup treatment effect (CSTE). Even though the CSTE model is easy to implement and may capture some of the treatment effect variation, it also has

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.

important limitations. The reason is straightforward: experimental groups may vary in ways that go beyond the average. Thus, we may have evidence of ATEs that do not differ across subgroups and yet have HTE, e.g., the groups may differ in their variances or higher moments.

We illustrate in this section how we can apply our method to test for within-group heterogeneity across subgroups defined by covariates. We then test the joint hypothesis of heterogeneity if we reject any of the individual subgroup-specific null hypotheses. In the next subsections, we argue our approach provides evidence against the underlying assumption behind the CSTE using experimental data from a welfare reform in Connecticut.

7.1 Subgroup Heterogeneity

We are interested in jointly testing the null hypotheses that treatment effects are constant *within* mutually exclusive subgroups while allowing them to be different *across* subgroups. Here the subgroups are defined by different combinations of baseline covariates' levels. We assume that the researcher defines these subgroups $1 \leq s \leq \mathcal{S}$, and we take as given from now on.

Let $Y_{0,s}$ and $Y_{1,s}$ denote the control and treatment outcomes for subgroup s, respectively, with corresponding CDFs $F_{0,s}(\cdot)$ and $F_{1,s}(\cdot)$.[4] The QTE for subgroup s is given by

$$\gamma_s(\tau) = F_{1,s}^{-1}(\tau) - F_{0,s}^{-1}(\tau), \ \forall \, \tau \in \mathscr{T} \ . \tag{15}$$

Thus, we simultaneously test a finite number of hypotheses

$$\mathbf{H}_0^{joint} : \bigcap_{1 \leq s \leq \mathcal{S}} H_{0,s} \ , \tag{16}$$

where each individual hypothesis is given by $H_{0,s} : \gamma_s(\tau) = \gamma_s$ for some γ_s, where γ_s is the unknown subgroup-specific QTE we need to estimate. For example, we can estimate γ_s as the

[4] 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 (17). 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.

mean difference between experimental groups within subgroup s (see Remark 1).

We reject the joint hypothesis (16) if any one of the null hypotheses $H_{0,s}$ is rejected. We control the family-wise error rate using the Holm adjustment. Needless to say, if γ_s were known, a permutation test for $H_{0,s}$ would retain the finite-sample validity though this case is infeasible in practice.

Remark 4. BGH test a similar, but different set of hypotheses:

$$H_{0,s}^{cdf} : F_{1,s}(y) = F_{0,s}(y - \delta_s) \ , \text{ for some } \delta_s \ , \tag{17}$$

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 (16) where the individual hypotheses $H_{0,s}$ are replaced by (17). However, we argue the heuristic justification of BGH's test procedure is insufficient to establish asymptotic validity of their test. 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) ones. 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 a series of 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 earnings distribution may experience a reduction on earnings or no effect 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,[5] and earnings history. We use these socio-demographic characteristics' levels and their interactions to form the $s = 1, \ldots, S$ groups we describe in the previous section (see also Table 3 below).

7.3 Results

Table 3 displays the results of our empirical analysis. Each row represents the results of applying our permutation test for different families of subgroups, and the total number of subgroups within each family equals the numbers in Column 2. We form the groups by quarter-specific covariates' levels, for each of the seven quarters we study here. For example, the second row (Education) considers the quarterly levels of *No high-school diploma*, *High-school diploma*, and *More than high-school diploma*, giving rise to the 21 subgroups in column 2. However, the first row of Table 3 (Full sample) has only seven subgroups because we look at the cross section of individuals for each of the seven quarters. As we argued in Section 7.1, our proposed test rejects the joint hypothesis (16) if any of the null hypotheses $H_{0,s}$ is rejected. Therefore, we report whether our proposed test rejects (16) at 5% in column 5.

One of the byproducts of the multiple testing approach we presented in Section 7.1 is that it allows us to detect for which subgroups we have evidence in favor of HTE. Therefore, we also display the number of individual hypotheses that our test results of our test procedure in columns 3 and 4. We account for multiple testing across subgroup configurations using the

[5] Women are eligible to receive benefits as long as their youngest child is under 18 years of age, among other criteria.

Holm adjustment since it is less conservative than the Bonferroni correction (for completeness, we include Bonferroni corrections in the online Appendix).[6]

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.

We begin by focusing on the primary subgroups (first 7 rows in Table 3) since they proxy the main determinants of the treatment effect heterogeneity, and because they form the basis for all the remaining comparisons in the table. The first row shows the quarterly effect of the welfare reform. We observe that our test rejects 3 (4) of the 21 individual hypotheses $H_{0,s}$ at 5% (10%), so we have evidence against the joint null hypothesis (16), as we display in column

[6] However, we note that the stepwise Holm procedure can be very conservative as well. 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, roughly speaking (Romano and Wolf, 2005).

5. Similarly, if we consider the quarterly levels of education, we can see that our proposed method rejects 3 (1) out of 21 individual hypotheses at the 10% (5%). The same conclusion follows if we apply our quantile-based permutation test for the families of subgroups defined by age of youngest child, marital status, welfare receipt seven quarters before the intervention, and number of quarters with earnings prior to random assignment, (rows 3, 4, 6 and 7 in Table 3). However, our test reaches a different conclusion for the family of groups defined by the earnings level seven months prior to the Jobs First welfare reform. More specifically, the proposed permutation test fails to reject any of the individual null hypotheses $H_{0,s}$ of constant QTE, so we cannot reject the joint null hypothesis (16).

The remainder of the table displays a series of quarterly subgroups we generate by interacting the covariates' levels for the baseline characteristics of interest. Therefore, it is not surprising that the number of subgroups increases, yielding in some cases up to 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 (16).

Table 3 has another important implication for practitioners. By rejecting (16) for many of the families of subgroups we consider, our permutation test provides strong evidence against the use of the CSTE model. Recall that the fundamental assumption behind the CSTE model states that the treatment effects are constant within subgroups, and the only source of variation comes from the fact that the ATEs vary significantly across subgroups. Thus, the CSTE model's underlying assumption is violated whenever our proposed permutation test rejects (16), disproving the CSTE model as a whole. However, we could still investigate treatment effect variation using the CSTE model in the handful of cases when our test fails to reject (16).

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 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 that this common practice is a poor account of the heterogeneity in the treatment effect.

A Appendix

Throughout we adopt the following notation, not necessarily introduced in the main text. If ξ is a random variable defined on a probability space $(\Omega, \mathscr{B}, P)$, it is assumed that $\xi_1, \dots, \xi_N$ are coordinate projections on the product space $(\Omega^N, \mathscr{B}^N, P^N)$, and the expectations are computed for P^N. If auxiliary variables—independent of the ξs—are involved, we use a similar convention. In that case, the underlying probability space is assumed to be of the form $(\Omega^N, \mathscr{B}^N, P^N) \times (\mathscr{Z}, \mathscr{C}, Q)$, with $\xi_1, \dots, \xi_N$ equal to the coordinate projections on the first N coordinates and the additional variables depending only on the $N + 1$st coordinate.

We view the empirical processes here as random maps into $\ell^\infty(\mathscr{T})$—the space of all bounded functions equipped with the uniform norm—and weak convergence is understood as convergence in distribution in $\ell^\infty(\mathscr{T})$. We assume that the class $\mathscr{T}$ is pointwise measurable (Van der Vaart and Wellner, 1996, Example 2.3.4), ruling out measurability problems with regards to suprema.

Independent of the Zs, let $(\pi(1), \dots, \pi(N))$ and $(\pi'(1), \dots, \pi'(N))$ be two independent random permutations of $\{1, \dots, N\}$. Denote $\boldsymbol{Z}_\pi = (Z_{\pi(1)}, \dots, Z_{\pi(N)})$; $\boldsymbol{Z}_{\pi'}$ is defined the same way with π replaced by π'

A.1 Proof of Theorem 1

We seek to show the asymptotic behavior of $R^{K}_{\mathrm{N}}(\dot{})$, the permutation distribution based on the 2SKSQ. Since $\hat{\upsilon}_{\mathrm{N}}(\tau; \boldsymbol{Z})$ is a continuous mapping by the arguments in the proof of Lemma 2, it suffices by the continuous mapping theorem (CMT) for randomization distributions (Chung and Romano, 2016, Lemma A.6) to establishes the asymptotic behavior of the permutation distribution based on $\hat{\upsilon}_{\mathrm{N}}(\tau; \boldsymbol{Z})$.

To this end, we begin by recentering the $\{Y_{1,i} : 1 \le i \le m\}$ by $\hat{\gamma}$ and collecting them in $\tilde{\boldsymbol{Z}}$ given by $\tilde{\boldsymbol{Z}}_{\mathrm{N}} = (Y_{1,1} - \hat{\gamma}, \dots, Y_{1,m} - \hat{\gamma}, Y_{0,1}, \dots, Y_{0,n})$. Independent of the $\tilde{Z}$s, let $(\pi(1), \dots, \pi(N))$ and $(\pi'(1), \dots, \pi'(N))$ be two independent random permutations of $\{1, \dots, N\}$. Denote $\tilde{\boldsymbol{Z}}_\pi = (\tilde{Z}_{\pi(1)}, \dots, \tilde{Z}_{\pi(N)})$; $\tilde{\boldsymbol{Z}}_{\pi'}$ is defined the same way with π replaced by π'.

We can show that $\hat{\upsilon}_{\mathrm{N}}(\tau, \boldsymbol{Z})$ is numerically equivalent to $\hat{\upsilon}_{\mathrm{N}}(\tau, \tilde{\boldsymbol{Z}})$ given by

$$\hat{\upsilon}_{\mathrm{N}}(\tau, \tilde{\boldsymbol{Z}}) = \sqrt{\frac{mn}{N}}\hat{\varphi}(\tau)\{\tilde{F}_1^{-1}(\tau) - \hat{F}_0^{-1}(\tau)\} ,$$

where $\tilde{F}_1^{-1}(\cdot)$ is the sample quantile function based on the first m entries in $\tilde{\boldsymbol{Z}}$. Therefore, we will establish the asymptotic behavior of the permutation distribution based on $\hat{\upsilon}_{\mathrm{N}}(\tau, \tilde{\boldsymbol{Z}})$.

By Lehmann and Romano (2005, Theorem 15.2.3), it suffices to that

$$\left\{\left(\hat{\upsilon}_{\mathrm{N}}(\tau; \tilde{\boldsymbol{Z}}_{\pi}), \hat{\upsilon}_{\mathrm{N}}(\tau; \tilde{\boldsymbol{Z}}_{\pi'})\right) : \tau \in \mathscr{T}\right\} \tag{A.1.1}$$

converges weakly to a tight process $\left\{\left(\upsilon(\tau), \upsilon'(\tau)\right) : \tau \in \mathscr{T}\right\}$, where $\left(\upsilon(\cdot), \upsilon'(\cdot)\right)$ are independent Brownian bridges. Observe that assumption A.1 implies that the inverse map is Hadamard-differentiable by Van der Vaart and Wellner (1996, Lemma 3.9.23) and that $\sup\left|\hat{\varphi(\tau)}\right| < \infty$. Therefore, it follows by Chung and Olivares (2021, Theorem 2) and the Delta-method that the process (A.1.1) converges weakly to a tight process $\left\{\left(\upsilon(\tau), \upsilon'(\tau)\right) : \tau \in \mathscr{T}\right\}$, where $\left(\upsilon(\cdot), \upsilon'(\cdot)\right)$ are independent Brownian bridges, as desired. This finishes the proof.

A.2 Proof of Theorem 2

We seek to show the asymptotic behavior of $R_{\mathrm{N}}^{\tilde{K}}$, the permutation distribution based on the two-sample martingale-transformed quantile process $\tilde{\upsilon}_{\mathrm{N}}(\tau; \boldsymbol{Z})$. Since $\tilde{\upsilon}_{\mathrm{N}}(\tau; \boldsymbol{Z})$ is a continuous mapping by the arguments in the proof of Lemma 2, then it suffices by the continuous mapping theorem (CMT) for randomization distributions (Chung and Romano, 2016, Lemma A.6) to establish the asymptotic behavior of the permutation distribution based on $\tilde{\upsilon}_{\mathrm{N}}(\tau; \boldsymbol{Z})$, given by

$$\tilde{\upsilon}_{\mathrm{N}}(\tau; \boldsymbol{Z}) = \hat{\upsilon}_{\mathrm{N}}(\tau; \boldsymbol{Z}) - \psi_g\left(\hat{\upsilon}_{\mathrm{N}}\right)(\tau; \boldsymbol{Z}) \ .$$

Thus, it siffices to prove by Lehmann and Romano (2005, Theorem 15.2.3) that

$$\left\{\left(\hat{\upsilon}_{\mathrm{N}}(\tau; \boldsymbol{Z}_{\pi}) - \psi_g(\hat{\upsilon}_{\mathrm{N}})(\tau; \boldsymbol{Z}_{\pi}), \hat{\upsilon}_{\mathrm{N}}(\tau; \boldsymbol{Z}_{\pi'}) - \psi_g(\hat{\upsilon}_{\mathrm{N}})(\tau; \boldsymbol{Z}_{\pi'})\right) : \tau \in \mathscr{T}\right\} \tag{A.2.1}$$

converges weakly to a tight process $\Big\{\Big(\zeta(\tau), \zeta'(\tau)\Big) : \tau \in \mathscr{T}\Big\}$, where $(\zeta(\tau), \zeta'(\tau))$ denotes a vector of two independent Brownian motion processes given by $\upsilon(\tau) - \varphi(\upsilon)(\tau)$, where $\upsilon(\cdot)$ is a tight Brownian bridge (and the same is true if we replace $\upsilon(\tau)$ by $\upsilon'(\tau)$).

We proved in Theorem 1 that (A.1.1) converges weakly to $\Big\{\Big(\upsilon(\tau), \upsilon'(\tau)\Big) : \tau \in \mathscr{T}\Big\}$, where $\Big(\upsilon(\cdot), \upsilon'(\cdot)\Big)$ are independent Brownian bridges. Furthermore, the continuity of $\psi_g(\cdot)$ implies that $\Big\{\Big(\psi_g(\hat{\upsilon}_{\mathrm{N}})(\tau; \boldsymbol{Z}_\pi), \psi_g(\hat{\upsilon}_{\mathrm{N}})(\tau; \boldsymbol{Z}_{\pi'})\Big) : \tau \in \mathscr{T}\Big\}$ converges weakly to $\Big(\psi_g(\upsilon), \psi_g(\upsilon')\Big)(\cdot)$ by the CMT for randomization distributions (Chung and Romano, 2016, Lemma A.6). Here, continuity follows by noting ψ_g is a Fredholm operator on a Banach space, hence a bounded operator. But an operator between normed spaces is bounded if and only if it is a continuous operator (Abramovich and Aliprantis, 2002). Then, the weak limit of (A.2.1) follows by Slutsky's theorem for randomization distributions (Chung and Romano, 2013, Theorem 5.2). This finishes the proof of our claim and the first part of the theorem.

For the second part of the theorem, we note that the distribution of $\tilde{K}$, *i.e.*, the distribution of the norm of a tight Brownian motion process, is strictly increasing and absolutely continuous with a positive density (Beran and Millar, 1986, Proposition 2). Thus, under the conditions of the theorem, $\hat{r}_{m,n}(1-\alpha) \xrightarrow{\mathrm{P}} r(1-\alpha) = \inf\{t : H(t) \geq 1-\alpha\}$ by Lehmann and Romano (2005, Lemma 11.2.1 (ii)), concluding the proof.

A.3 Proof of Corollary 1

From the construction of the permutation test based on $\tilde{K}_{\mathrm{N}}(\boldsymbol{Z})$, we have

$$\Pr\Big\{\tilde{K}_{\mathrm{N}} > \hat{r}_{m,n}(1-\alpha)\Big\} \leq \mathbb{E}\left[\phi(Z)\right] \leq \Pr\Big\{\tilde{K}_{\mathrm{N}} \geq \hat{r}_{m,n}(1-\alpha)\Big\} \ .$$

Hence, Theorem 2 implies $\mathbb{E}\left[\phi(Z)\right] \to 1 - H(r(1-\alpha) = \alpha$, as desired.

References

Abramovich, Y. A. and Aliprantis, C. D. (2002). *An Invitation to Operator Theory*, volume 1. American Mathematical Soc.

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

Beran, R. and Millar, P. (1986). Confidence sets for a multivariate distribution. *The Annals of Statistics*, pages 431–443.

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.

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.

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. (2005). *Testing statistical hypotheses*. Springer Science & Business Media.

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.

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

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