Supplement for *Comment on "Can Variation in Subgroups' Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment"*

EunYi Chung[†] Mauricio Olivares[†]

We organize this supplementary appendix as follows. In Section 1, we present empirical and simulation evidence showing BGH's is not asymptotically valid with estimated parameters (see Chung and Olivares (2021) for the theoretical justification in this section). Section 2 outlines our statistical environment. We establish the asymptotic behavior of the test statistics we consider in Section 3. Section 4 contains our main theoretical contribution. We show that the permutation distribution of the martingale-transformed statistic and the true limiting distribution of the test statistic are asymptotically equal. We derive a fast algorithm for the numerical implementation of our proposed method in Section 5, and conclude with Monte Carlo experiments to compare the finite-sample performance of our permutation test and other quantile-based tests in Section 6.

Throughout this appendix, we focus on a simplified scenario with one group only. The generalization to many subgroups is straightforward and therefore omitted. Thus, we focus on the substance of our method while minimizing notational clutter.

1 The Invalidity of BGH's approach

To illustrate that BGH's test procedure can lead to distorted inference in empirically relevant settings, we begin by investigating the performance of their method in a simple Monte Carlo experiment. We introduce some notation first. Suppose that Y is the observed outcome of interest and D is a treatment or policy indicator taking values 1 if treated, and 0 otherwise. Y is linked to the potential outcomes through the relationship

[†]Department of Economics, University of Illinois at Urbana–Champaign. All errors are our own.

$Y = Y(1)D + (1-D)Y(0)$. 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 of the following specifications, ε_i, $1 \leq i \leq N$ are i.i.d. according to: standard normal, lognormal, Student's t distribution with 5 degrees of freedom, and the standard exponential distribution. To illustrate the effect that the estimated parameter has on the permutation test, we include a permutation test in the infeasible case when γ is known so that we do not have the estimated-parameter problem. In fact, the permutation test controls the size in finite samples in this case. We call this case the classical 2SKS.

Table 1 reports the rejection probabilities under the null hypothesis of equality of simulated and actual outcomes distributions with $\gamma = 1$. This simple Monte Carlo experiment illustrates how BGH's test procedure delivers empirical rejection probabilities that are considerably far from the nominal level. For example, in the Normal and t distributions, BGH's permutation test is hyper-conservative. More problematic is the lognormal and exponential cases, when BGH's method fails to control the rejection probabilities.

Table 1: Size of $\alpha = 0.05$ tests H_0 : Equality of distributions.

		Distributions			
N	Method	Normal	Lognormal	Student's t	Exponential
$N = 13$	Classical 2SKS	0.0494	0.0482	0.0522	0.0522
$n = 8$, $m = 5$	BGH	0.0000	0.0298	0.0002	0.0014
$N = 50$	Classical 2SKS	0.0528	0.0506	0.0460	0.0460
$n = 30$, $m = 20$	BGH	0.0002	0.3116	0.0014	0.0764
$N = 80$	Classic 2SKS	0.0452	0.0516	0.0510	0.0510
$n = 50$, $m = 30$	BGH	0.0000	0.3244	0.0016	0.0872
$N = 200$	Classic 2SKS	0.0472	0.0548	0.0486	0.0486
$n = 120$, $m = 80$	BGH	0.0004	0.3912	0.0032	0.1532

Rejection probabilities computed using 5000 replications across specifications.

Not surprisingly, Table 1 also reveals that when γ is known, the permutation test based on the classical 2SKS delivers rejection probabilities under the null hypothesis that are very close to the nominal level for all specifications we consider. We expected this since the permutation test attains the exactness in finite samples in this case.

Table 2 displays the results from BGH's empirical analysis. As in BGH, each row represents the results for a particular set of subgroups, and the total number of subgroups equals the numbers in Column 2. Columns 3–4 contain the empirical results from BGH's Table 2, which we include verbatim for a fair comparison. We note that a joint test of the family of null hypotheses across these subgroups rejects if we reject any one of the subgrop-specific null hypotheses (see Bitler, Gelbach, and Hoynes, 2017, Sec. V). To underscore the detrimental effects of estimated parameters, we also include the results if we had considered an asymptotically valid permutation test for the same joint hypothesis instead. This benchmark corresponds to Chung and Olivares (2021) in columns 5–6.

Table 2: Testing for Heterogeneity in the Treatment Effect by Subgroups, Time-varying mean treatment effects by subgroup with participation adjustment

		BGH's Permutation Test		Asymptotically Valid Permutation Test	
Subgroup	Number of Tests	Number of Reject at 10%	Number of Reject at 5%	Number of Reject at 10%	Number of Reject at 5%
Full Sample	7	4	4	7	7
Education	21	3	1	9	9
Age of youngest child	21	3	1	11	10
Marital status	21	2	1	14	14
Earnings level seventh Q pre-RA	21	2	1	17	16
Number of pre-RA Q with earnings	21	1	0	17	16
Welfare receipt seventh Q pre-RA	14	3	3	14	14
Education subgroups interacted with					
Age of youngest child	49	1	0	14	14
Marital status	35	3	3	18	17
Earnings level seventh Q pre-RA	63	1	0	15	14
Number of pre-RA Q with earnings	63	0	0	13	11
Welfare receipt seventh Q pre-RA	42	1	0	15	14
Age of youngest child interacted with					
Marital status	35	1	1	17	15
Earnings level seventh Q pre-RA	63	0	0	17	14
Number of pre-RA Q with earnings	49	1	1	14	12
Welfare receipt seventh Q pre-RA	42	1	0	14	13
Marital status subgroup interacted with					
Earnings level seventh Q pre-RA	63	2	1	14	11
Number of pre-RA Q with earnings	63	0	0	15	11
Welfare receipt seventh Q pre-RA	42	1	0	14	13
Earnings level seventh Q pre-RA subgroups interacted with					
Number of pre-RA Q with earnings	49	0	0	16	15
Welfare receipt seventh Q pre-RA	42	1	1	17	15
Number of quarters any earnings pre-RA subgroup interacted with					
Welfare receipt seventh Q pre-RA	42	0	0	14	13

All reported results account for multiple testing using Bonferroni adjustment. We use 1000 permutations for the stochastic approximation of the permutation distribution.

Compared to BGH's, columns 5–6 report a higher number of rejections of the null of equality of the within-subgroup distributions after the Bonferroni adjustment. While both tests reject the null of equality of distributions for many of the families of subgroups in Table 2, there are some crucial discrepancies between the two tests. For example, when they create subgroups based on the number of quarters with any earnings before random assignment, BGH cannot reject at the 5% level after adjustment. Meanwhile, Chung and Olivares's (2021) test rejects 16 of 21 tests at the 5% . We document this type of discrepancy between the tests in the remainder of the table for several families of subgroups.

2 Statistical Environment

Let $F_1(\cdot)$ and $F_0(\cdot)$ denote the distribution functions of units in the treatment and control groups, respectively. The quantile treatment effect 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 null hypothesis of constant treatment effect is

$$H_0^q : \gamma(\tau) = \gamma \quad \text{for some} \quad \gamma, \quad \forall \ \tau \in \mathscr{T} \quad , \tag{1}$$

whereas the alternative represents heterogeneous effects, *i.e.*, $\gamma(\tau)$ varies across $\tau \in \mathscr{T}$. Observe that the nuisance parameter γ in (1) is unknown so we need to estimate it. One may estimate γ by $\hat{\gamma} = \int_{\mathscr{T}} \hat{\gamma}(\tau)$, or by the OLS estimator of a regression of Y on D.

One natural candidate for a test statistic for hypothesis (1) is to compare the empirical quantile functions based on two independent random samples drawn from F_1 and F_0. We can estimate the difference in quantiles by the individual coefficient associated with D in the conditional quantile regression model $F_{Y|D}^{-1}(\tau) = \alpha(\tau) + \gamma(\tau)D$, $\tau \in \mathscr{T}$. Then, we base our analysis on the quantile regression estimates given by

$$\{\hat{\alpha}(\tau), \hat{\gamma}(\tau)\} = \underset{a,b\in\mathbb{R}}{\arg\min} \sum_{i=1}^{N} \rho_\tau \left(Y_i - a - bD_i\right) \ , \ \tau \in \mathscr{T} \ , \tag{2}$$

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

Before introducing the test statistic, we introduce more notation. Let $Y_1^1, \ldots, Y_m^1$ and $Y_1^0, \ldots, Y_n^0$ be two independent random samples having distribution functions $F_1(\cdot)$ and

$F_0(\cdot)$, respectively.[1] Collect all these outcomes in one vector as $\boldsymbol{Z} = (Z_1, \ldots, Z_N) = (Y_1^1, \ldots, Y_m^1, Y_1^0, \ldots, Y_n^0)$. The test statistic we consider in this paper is the two-sample Kolmogorov–Smirnov statistic based on the quantile process (2SKSQ):

$$K_{\mathrm{N}}(\boldsymbol{Z}) = \sup_{\tau \in \mathscr{T}} |\hat{\upsilon}_{\mathrm{N}}(\tau; \boldsymbol{Z})| \ , \tag{3}$$

where

$$\hat{\upsilon}_{\mathrm{N}}(\tau; \boldsymbol{Z}) = \sqrt{\frac{mn}{N}} \hat{\varphi}(\tau) \left\{ \hat{\gamma}(\tau) - \hat{\gamma} \right\}, \ \ \tau \in \mathscr{T} \tag{4}$$

is the quantile regression process, and $\hat{\varphi}(\tau)$ is an estimate of $\varphi(\tau) = f_0(F_0^{-1}(\tau))$.

We now show how the construction of a permutation tests for H_0^q works. To define the test, we introduce further notation. Let $\mathbf{G}_N$ be the set of all permutations π of $\{1, \ldots, N\}$, with $|\mathbf{G}_N| = N!$. Given $\boldsymbol{Z} = \boldsymbol{z}$, recompute $K_{\mathrm{N}}(\boldsymbol{z})$ for all permutations $\pi \in \mathbf{G}_N$ and denote by $K_{\mathrm{N}}^{(1)}(\boldsymbol{z}) \leq K_{\mathrm{N}}^{(2)}(\boldsymbol{z}) \leq \cdots \leq K_{\mathrm{N}}^{(N!)}(\boldsymbol{z})$ the ordered values of $\{K_{\mathrm{N}}(\boldsymbol{z}_\pi) : \pi \in \mathbf{G}_N\}$, where $\boldsymbol{z}_\pi$ denotes the action of $\pi \in \mathbf{G}_N$ on $\boldsymbol{z}$.

The permutation test is $\phi(\boldsymbol{z}) = \mathbb{1}\left\{K_{\mathrm{N}}(\boldsymbol{z}) > K_{\mathrm{N}}^{(k)}(\boldsymbol{z})\right\} + a(\boldsymbol{z})\mathbb{1}\left\{K_{\mathrm{N}}(\boldsymbol{z}) = K_{\mathrm{N}}^{(k)}(\boldsymbol{z})\right\}$, where $a(\boldsymbol{z})$ is defined as in Lehmann and Romano (2005, p. 632). Alternatively, the permutation test rejects (1) if $K_{\mathrm{N}}(\boldsymbol{z})$ exceeds the $1 - \alpha$ quantile of the permutation distribution:

$$\hat{R}_{\mathrm{N}}^K(t) = \frac{1}{N!} \sum_{\pi \in \mathbf{G}_N} \mathbb{1}_{\{K_{\mathrm{N}}(z_{\pi(1)}, \ldots, z_{\pi(N)}) \leq t\}} \ . \tag{5}$$

The construction of (5) is computationally prohibitive for moderately large N, so we rely on a stochastic approximation as follows. 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 approximation does not affect the test's theoretical properties if we replace $\mathbf{G}_N$ with $\hat{\mathbf{G}}_N$, and the

[1] 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.

approximation is arbitrarily close for M sufficiently large (Romano, 1989, Section 4). Thus, we focus on $\mathbf{G}_N$ from now on while in practice we fall back on $\hat{\mathbf{G}}_N$.

2.1 Adding Covariates

In practice, we typically observe a vector of baseline covariates besides D, e.g., those in Table 2. In this case, we can still test the hypothesis of heterogeneous treatment effects (1) 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 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 $\boldsymbol{R}(\tau)\beta(\tau) - \boldsymbol{r}(\tau) = 0$, $\tau \in \mathscr{T}$, 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

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

Observe that (1) is equivalent to setting $\boldsymbol{R} = [1, 0, \ldots, 0]$ and $\boldsymbol{r}(\tau) = \gamma$ in this context. Thus, the conclusions of Theorems 1–3 remain the same in the presence of covariates 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 more details.

3 Limiting Distributions of the Test Statistics

We now introduce three assumptions, standard in the quantile regression literature, that 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). Moreover, these convergence rates also play a key role when we investigate the asymptotic behavior of the permutation distribution (5).

Assumption A.3 guarantees that we can replace the unknown quantities, γ and $\varphi(\tau)$, with estimates satisfying general assumptions. While the asymptotic results we present below remain valid if we estimate $\varphi(\tau)$ such that A.3 (ii) holds, we will see that the estimation of γ does have a detrimental effect on the asymptotic behavior of the 2SKSQ.

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

The following theorem is a special case of Koenker and Xiao (2002, Theorem 2) applied to the heterogeneous-treatment-effect testing problem in this paper. This theorem establishes the asymptotic behavior of the quantile process and 2SKSQ under (1).

Theorem 1. *Under assumptions A.1–A.3, the process $\{\hat{\upsilon}_{\mathrm{N}}(\tau; \boldsymbol{Z}) - \hat{\varphi}(\tau)\xi_{\mathrm{N}}(\boldsymbol{Z}) : \tau \in \mathscr{T}\}$, where $\xi_{\mathrm{N}}(\boldsymbol{Z}) = \mathcal{O}_p(1)$, converges weakly in $\ell^\infty(\mathscr{T})$—the space of all bounded functions on $\mathscr{T}$ equipped with the uniform norm—to a standard Brownian bridge process, denoted by $\upsilon(\cdot)$, under the null hypothesis (1). Furthermore, the test statistic $K_N(\boldsymbol{Z})$, defined in (3), converges in distribution to $K \equiv \sup_{\tau \in \mathscr{T}} |\upsilon(\tau)|$ with CDF given by $J(t) \equiv Pr\{K \leq t\}$.*

Proof. We are interested in showing the asymptotic behavior of the test statistic under the null hypothesis. We begin by rewriting (4) as

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

where the last term in (6) is zero under the null hypothesis. Develop further to obtain

$$\begin{aligned}\hat{\upsilon}_{\text{N}}(\tau;\boldsymbol{Z}) &= \sqrt{\frac{mn}{N}}\varphi(\tau)\left\{\hat{\gamma}(\tau)-\gamma(\tau)\right\} - \sqrt{\frac{mn}{N}}\varphi(\tau)\left\{\hat{\gamma}-\gamma\right\} \\ &\quad + \sqrt{\frac{mn}{N}}\left[\hat{\varphi}(\tau)-\varphi(\tau)\right]\left\{\hat{\gamma}(\tau)-\gamma(\tau)\right\} - \sqrt{\frac{mn}{N}}\left[\hat{\varphi}(\tau)-\varphi(\tau)\right]\left\{\hat{\gamma}-\gamma\right\} \\ &= \sqrt{\frac{mn}{N}}\varphi(\tau)\left\{\hat{\gamma}(\tau)-\gamma(\tau)\right\} - \sqrt{\frac{mn}{N}}\varphi(\tau)\left\{\hat{\gamma}-\gamma\right\} + o_p(1) \\ &= \upsilon_{\text{N}}(\tau;\boldsymbol{Z}) + \varphi(\tau)\xi_{\text{N}}(\boldsymbol{Z}) + o_p(1) \ ,\end{aligned} \tag{7}$$

where the $o_p(1)$ term holds uniformly over $\mathscr{T}$ by Assumption A.3 (ii). Under Assumptions A.1 and A.2, $\{\upsilon_{\text{N}}(\tau;\boldsymbol{Z}) : \tau \in \mathscr{T}\}$ converges weakly in $\ell^\infty(\mathscr{T})$ to a Brownian bridge process $\upsilon(\cdot)$ by Shorack and Wellner (2009, Theorem 2, Ch. 18).

By Assumption A.3 (i), the second term on the right-hand side of (7) is in $\ell^\infty(\mathscr{T})$ if and only if $\sup_{\mathscr{T}}|\varphi(\tau)| < \infty$. But this follows by A.1, which implies F_0 is Lipschitz continuous and therefore $\sup_{\mathscr{T}}|\varphi(\tau)| < \infty$. This finishes the first part of the proof.

Note that the maps $\upsilon \to \|\upsilon\|$ from $\ell^\infty(\mathscr{T})$ into $\mathbb{R}$ are continuous with respect to the supremum norm. Then, a direct application of the continuous mapping theorem (CMT) yields the final result. This finishes the proof. □

Theorem 1 shows that the presence of an estimated nuisance parameter breaks the asymptotically distribution-freeness of the 2SKSQ. To see why, note that $\hat{\upsilon}_{\text{N}}(\cdot;\boldsymbol{Z})$ behaves asymptotically like a Brownian bridge plus an extra component whose limit distribution depends on γ. Koenker and Xiao (2002) showed that we can eliminate $\xi_{\text{N}}(\cdot;\boldsymbol{Z})$ in (7) via Khmaladze's (1981) martingale transformation. This is the content of the next section.

3.2 Limiting Distribution of $\tilde{K}_{\mathrm{N}}(\boldsymbol{Z})$

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

$$\psi_g(h)(t) = \int_0^t \left[\dot{g}(s)'C(s)^{-1} \int_s^1 \dot{g}(r)dh(r)\right] ds \ , \tag{8}$$

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 (4) and 2SKSQ are given by

$$\tilde{\upsilon}_{\mathrm{N}}(\tau, \boldsymbol{Z}) = \hat{\upsilon}_{\mathrm{N}}(\tau; \boldsymbol{Z}) - \psi_g(\hat{\upsilon}_{\mathrm{N}})(\tau; \boldsymbol{Z}) \ , \quad \text{and} \tag{9}$$

$$\tilde{K}_{\mathrm{N}}(\boldsymbol{Z}) = \sup_{\tau \in \mathscr{T}} |\tilde{\upsilon}_{\mathrm{N}}(\tau; \boldsymbol{Z})| \ , \tag{10}$$

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 2SKS statistic. As in the case of Theorem 1, it is a particular case of Koenker and Xiao

(2002, Theorem 3) applied to our testing problem.

Theorem 2. *Under assumptions A.1–A.4, we have that the process* $\{\tilde{\upsilon}_N(\tau) : \tau \in \mathscr{T}\}$ *converges weakly in* $\ell^\infty(\mathscr{T})$ *to* $\zeta(\cdot)$ *under the null hypothesis* (1). *Here,* $\zeta(\cdot)$ *denotes the standard Brownian motion. Furthermore, the test statistic* $\tilde{K}_N(\boldsymbol{Z})$*, defined in* (10)*, converges in distribution to* $\tilde{K} \equiv \sup_{\tau\in\mathscr{T}}|\zeta(\tau)|$ *with CDF given by* $H(t) \equiv Pr\{\tilde{K} \leq t\}$.

Proof. Recall the martingale-transformation of the quantile process (9) is given by

$$\begin{aligned}\tilde{\upsilon}_{\mathrm{N}}(\tau;\boldsymbol{Z}) &= \hat{\upsilon}_{\mathrm{N}}(\tau;\boldsymbol{Z}) - \psi_g\left(\hat{\upsilon}_{\mathrm{N}}\right)(\tau;\boldsymbol{Z}) \\ &= \upsilon_{\mathrm{N}}(\tau;\boldsymbol{Z}) + \varphi(\tau)\xi_{\mathrm{N}}(\boldsymbol{Z}) - \psi_g\left(\hat{\upsilon}_{\mathrm{N}}\right)(\tau;\boldsymbol{Z}) + o_p(1)\ ,\end{aligned} \tag{11}$$

where the second equality follows by the asymptotic expansion (7) and the $o_p(1)$ term holds uniformly over $\mathscr{T}$. By properties of the compensator ψ_g and (7), we have that

$$\psi_g\left(\hat{\upsilon}_{\mathrm{N}}\right)(\tau;\boldsymbol{Z}) = \psi_g\left(\upsilon_{\mathrm{N}}\right)(\tau;\boldsymbol{Z}) + \varphi(\tau)\xi_{\mathrm{N}}(\boldsymbol{Z}) + o_p(1)\ . \tag{12}$$

Plugging (12) into (11) yields $\tilde{\upsilon}_{\mathrm{N}}(\tau;\boldsymbol{Z}) = \upsilon_{\mathrm{N}}(\tau;\boldsymbol{Z}) - \psi_g\left(\upsilon_{\mathrm{N}}\right)(\tau;\boldsymbol{Z}) + o_p(1)$, and by Khmaladze (1981, 4.3), $\{\tilde{\upsilon}_{\mathrm{N}}(\tau;\boldsymbol{Z}) : \tau \in \mathscr{T}\}$ converges weakly to $\zeta(\cdot)$, the standard Brownian motion. This finishes the proof of th fist part of the theorem. A direct application of the CMT as in the proof of Theorem 1 finishes the proof. □

4 Asymptotically Valid Permutation Test

We now turn to our main theoretical result—the permutation test based on the martingale-transformed statistic behaves asymptotically like the true unconditional limiting sampling distribution. We seek the limiting behavior of $\hat{R}^{\tilde{K}}_{m,n}$ and its upper α-quantile, which we now denote $\hat{r}_{m,n}$, 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. Note that (1) is not assumed.

Theorem 3. *Consider testing the hypothesis* (1) *based on the test statistic* (10). *Under assumptions A.1–A.4, the permutation distribution* (5) *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 Theorem 2. Moreover, if $r(1-\alpha)=\inf\{t: H(t)\ge 1-\alpha\}$, then $\hat{r}_{m,n}(1-\alpha)\xrightarrow{\mathrm{p}}r(1-\alpha)$.

Proof. Independent of the Zs, let $(\pi(1),\ldots,\pi(N))$ and $(\pi'(1),\ldots,\pi'(N))$ be two independent random permutations of $\{1,\ldots,N\}$. Denote $\boldsymbol{Z}_{\pi}=(Z_{\pi(1)},\ldots,Z_{\pi(N)})$; $\boldsymbol{Z}_{\pi'}$ is defined the same way with π replaced by π'. We seek to show that $\left(\tilde{K}_{\mathrm{N}}(\boldsymbol{Z}_{\pi}),\tilde{K}_{\mathrm{N}}(\boldsymbol{Z}_{\pi'})\right)\xrightarrow{\mathrm{d}}\left(\tilde{K},\tilde{K}'\right)$, where $\tilde{K}$ and $\tilde{K}'$ are independent with common CDF $H(\cdot)$. Then Lehmann and Romano (2005, Theorem 15.2.3) implies (13), completing the proof of the first part of the theorem. To show $\left(\tilde{K}_{\mathrm{N}}(\boldsymbol{Z}_{\pi}),\tilde{K}_{\mathrm{N}}(\boldsymbol{Z}_{\pi'})\right)\xrightarrow{\mathrm{d}}\left(\tilde{K},\tilde{K}'\right)$, it suffices to prove the asymptotic behavior of

$$\left\{(\tilde{\upsilon}_{\mathrm{N}}(\tau;\boldsymbol{Z}_{\pi}),\tilde{\upsilon}_{\mathrm{N}}(\tau;\boldsymbol{Z}_{\pi'})):\tau\in\mathscr{T}\right\}\ , \tag{14}$$

by the usual CMT. In the following, we show the asymptotic behavior of (14) in 4 steps.

Step 1. We show that—regardless of the null hypothesis—we can write $\tilde{\upsilon}_{\mathrm{N}}(\cdot;\boldsymbol{Z})$ as

$$\tilde{\upsilon}_{\mathrm{N}}(\tau;\boldsymbol{Z})=\upsilon_{\mathrm{N}}(\tau;\boldsymbol{Z})-\psi_g\left(\upsilon_{\mathrm{N}}\right)(\tau;\boldsymbol{Z})+o_p(1)\ , \tag{15}$$

where the $o_p(1)$ term holds uniformly over $\mathscr{T}$, and the same is true if we replace π with π'. To see why, we write (6) as

$$\hat{\upsilon}_{\mathrm{N}}(\tau;\boldsymbol{Z})=\sqrt{\frac{mn}{N}}\varphi(\tau)\Big(\{\hat{\gamma}(\tau)-\gamma(\tau)\}-\{\hat{\gamma}-\gamma\}+\{\gamma(\tau)-\gamma\}\Big)+o_p(1)\ , \tag{16}$$

where the $o_p(1)$ term holds uniformly over $\mathscr{T}$ by Assumption A.3 (ii). Then (15) follows by properties of $\psi_g(\cdot)$.

Step 2. We now establish the asymptotic behavior of $\tilde{\upsilon}_{\mathrm{N}}(\cdot;\boldsymbol{Z})$ based on the asymptotic representation (15). We begin by considering the first two terms on the right-hand side,

$$\left\{\left(\upsilon_{\mathrm{N}}(\tau;\boldsymbol{Z}_\pi)-\psi_g(\upsilon_{\mathrm{N}})(\tau;\boldsymbol{Z}_\pi),\upsilon_{\mathrm{N}}(\tau;\boldsymbol{Z}_{\pi'})-\psi_g(\upsilon_{\mathrm{N}})(\tau;\boldsymbol{Z}_{\pi'})\right):\tau\in\mathscr{T}\right\}\ . \tag{17}$$

For notational convenience, abbreviate $\mathbb{Z}_{\mathrm{N}}=\left\{\left(\upsilon_{\mathrm{N}}(\tau;\boldsymbol{Z}_\pi),\upsilon_{\mathrm{N}}(\tau;\boldsymbol{Z}_{\pi'})\right):\tau\in\mathscr{T}\right\}$. We first claim that the process $\mathbb{Z}_{\mathrm{N}}$ converges weakly to a tight process $\mathbb{Z}\equiv\left\{\left(\upsilon(\tau),\upsilon'(\tau)\right):\tau\in\mathscr{T}\right\}$, where $\left(\upsilon(\cdot),\upsilon'(\cdot)\right)$ denotes a vector of two independent standard Brownian bridges. To see why, we apply the coupling construction of Chung and Romano (2013). More specifically, couple data $\boldsymbol{Z}$ with an auxiliary sample of N i.i.d. observations $\bar{\boldsymbol{Z}}=(\bar{Z}_1,\ldots,\bar{Z}_N)$ from the mixture distribution with CDF $\bar{P}(y)=pF_1(y+\gamma)+(1-p)F_0(y)$, where $p=\lim_{m\to\infty}m/N$. In view of the arguments in the proof of Chung and Romano (2013, Lemma 5.1), we can determine the behavior of $\mathbb{Z}_{\mathrm{N}}$ by verifying the following sufficient conditions

i) The process $\left\{\left(\upsilon_{\mathrm{N}}(\tau;\bar{\boldsymbol{Z}}_\pi),\upsilon_{\mathrm{N}}(\tau;\bar{\boldsymbol{Z}}_{\pi'})\right):\tau\in\mathscr{T}\right\}$ converges weakly to $\mathbb{Z}$.

ii) $\upsilon_{\mathrm{N}}(\tau;\bar{\boldsymbol{Z}}_{\pi,\pi_0})-\upsilon_{\mathrm{N}}(\tau;\boldsymbol{Z}_\pi)\xrightarrow{\mathrm{p}}0$ uniformly, where the permutation π_0 is defined in Chung and Romano (2013).

Assumption A.1 implies that the inverse map is Hadamard-differentiable by Van der Vaart and Wellner (1996, Lemma 3.9.23). Hence, i)–ii) follow by lemmas B.1–B.2 in Chung and Olivares (2021) combined with the Delta-method (Van der Vaart and Wellner, 1996, Theorem 3.9.4).

The continuity of $\psi_g(\cdot)$ implies that $\left\{\left(\psi_g(\upsilon_{\mathrm{N}})(\tau;\boldsymbol{Z}_\pi),\psi_g(\upsilon_{\mathrm{N}})(\tau;\boldsymbol{Z}_{\pi'})\right):\tau\in\mathscr{T}\right\}$ converges weakly to $\left(\psi_g(\upsilon),\psi_g(\upsilon')\right)(\cdot)$ by the CMT for randomization distributions, Chung and Romano (2016, Lemma A.6). Continuity follows by noting ψ_g is a Fredholm operator on a Banach space (Koenker and Xiao, 2002), 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, (17) follows by Slutsky's theorem for randomization distributions (Chung and Romano, 2013, Theorem 5.2).

Step 3 We now prove that the o_p term in (15) holds under permutations, *i.e.*,

$$\tilde{\upsilon}_{\mathrm{N}}(\tau; \boldsymbol{Z}_\pi) - \upsilon_{\mathrm{N}}(\tau; \boldsymbol{Z}_\pi) - \psi_g\left(\upsilon_{\mathrm{N}}\right)(\tau; \boldsymbol{Z}_\pi) \xrightarrow{\mathrm{P}} 0 \ . \tag{18}$$

In view of the contiguity result in Chung and Romano (2013, Lemma 5.3), we can deduce (18) from the basic assumption of how it behaves under i.i.d. observations from the mixture distribution $\bar{P}$. But we showed this in **Step 3**, so the desired conclusion follows.

Step 4 Combine the conclusions of ***Step 2*** and ***Step 3*** with the Slutsky's theorem for randomization distributions (Chung and Romano, 2013, Theorem 5.2) once again and conclude (13). 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. □

Remark 1. From the construction of the permutation test based on $\tilde{K}_{\mathrm{N}}(\boldsymbol{Z})$, we have $\Pr\left\{\tilde{K}_{\mathrm{N}} > \hat{r}_{m,n}\right\} \leq \mathbb{E}\left[\phi(Z)\right] \leq \Pr\left\{\tilde{K}_{\mathrm{N}} \geq \hat{r}_{m,n}\right\}$. Hence, Theorem 3 implies $\mathbb{E}\left[\phi(Z)\right] \to \alpha$. See Lehmann and Romano (2005, Section 15.2.2). ■

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. Consequently, 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.[2] We can think of preprocessing in a simple way. 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 that result from this quantile regression to inform the sign of the residuals in the whole sample. Collect these pseudo-observations with either "'large" negative or "'large" positive residuals, a sample Portnoy and Koenker coined as *glob*. Thus, we can remove the "globbed" sample from the optimization problem, 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, (2) yields the same solution as the following revised problem

$$\arg\min_{a,b\in\mathbb{R}} \sum_{i\notin(J_L\cup J_H)} \rho_\tau\left(Y_i - a - bD_i\right) + \rho_\tau\left(Y_L - a - bD_L\right) + \rho_\tau\left(Y_H - a - bD_H\right)\ , \tag{19}$$

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), $k \in \{L, H\}$. Observe that the revised problem (19) has a reduction in the effective sample size—it has $\#\{J_L \cup J_H\} - 2$ fewer observations, the number of 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 (2) 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, de-*

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

noted $\hat{s}_i$. 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* (19) *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 $\tau \in \{\tau_1, \ldots, \tau_T\}$. Intuitively, 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 by assuming $\sqrt{N}(\tau_t - \tau_{t-1}) = \mathcal{O}_p(1)$, thus implying we only need to keep a sample proportional to $N^{1/2}$ as opposed to $N^{2/3}$, like in Algorithm 1.

5.1 Preprocessing for Permutation-based Inference

As in the previous section, 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 (19) *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 $j = 1, \ldots, M$*.*

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 Section 1, which follows from Koenker and Xiao (2002) and Chernozhukov and Fernández-Val (2005). For this exercise, we focus our attention to the following probability distributions: standard normal, lognormal, 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 3 and $\Pr\{D = 1\} = 0.5$ for Table 4.

We perform the numerical calculation of 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* (e.g., Portnoy and Koenker, 1989), 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 3–4, we compare the proposed permutation test based on (10)—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 3 reports rejection probabilities under the null hypothesis (1) 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 3: Size of $\alpha = 0.05$ tests H_0 : Constant Treatment Effect ($\gamma = 1$).

N	Method	Distributions		
		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 4 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**.

Table 4: 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.

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. (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.

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. *forthcoming in the Journal of Econometrics*, pages 1–27.

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.

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

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

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. 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.

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.

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.

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