Online Supplement for *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

We organize this online appendix as follows. Section I contains the proofs of Lemmas 1 and 2. These lemmas establish the asymptotic null behavior of the 2SKSQ and the Khmaladze transformed 2SKSQ, respectively. In Section II, we describe Bitler, Gelbach, and Hoynes's (2017, BGH) simulated outcomes approach in more detail. We emphasize how their approach is not immune to the Durbin problem and what is the source of the problem. Even though their heuristic approach to the Durbin problem yields to a correct conclusion (inadequacy of the CSTE model), we argue the theoretical reasoning behind the simulated outcomes approach does not formally address the problem and, therefore, we cannot claim the asymptotic validity of their permutation test. The second part of this appendix, Section III, contains the results of BGH's empirical exercise. We include them here verbatim to highlight how their results are qualitatively the same as ours, despite the presence of an estimated nuisance parameter.

[†]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.

I Proofs of Lemmas

I.1 Proof of Lemma 1

We are interested in showing the asymptotic behavior of the test statistic under the null hypothesis. We begin by rewriting (8) as

$$\hat{\upsilon}_{\text{N}}(\tau;\boldsymbol{Z}) = \sqrt{\frac{mn}{N}}\hat{\varphi}(\tau)\left\{\hat{\gamma}(\tau)-\gamma(\tau)\right\} - \sqrt{\frac{mn}{N}}\hat{\varphi}(\tau)\left\{\hat{\gamma}-\gamma\right\} + \sqrt{\frac{mn}{N}}\hat{\varphi}(\tau)\left\{\gamma(\tau)-\gamma\right\} \ . \tag{I.1}$$

where the last term in (I.1) is zero under the null hypothesis. Develop further to obtain

$$\begin{aligned}\hat{\upsilon}_{\text{N}}(\tau;\boldsymbol{Z}) &= \sqrt{\frac{mn}{N}}\varphi(\tau)\left\{\hat{\gamma}(\tau)-\gamma(\tau)\right\} - \sqrt{\frac{mn}{N}}\varphi(\tau)\left\{\hat{\gamma}-\gamma\right\} \\ &\quad + \sqrt{\frac{mn}{N}}\left[\hat{\varphi}(\tau)-\varphi(\tau)\right]\left\{\hat{\gamma}(\tau)-\gamma(\tau)\right\} - \sqrt{\frac{mn}{N}}\left[\hat{\varphi}(\tau)-\varphi(\tau)\right]\left\{\hat{\gamma}-\gamma\right\} \\ &= \sqrt{\frac{mn}{N}}\varphi(\tau)\left\{\hat{\gamma}(\tau)-\gamma(\tau)\right\} - \sqrt{\frac{mn}{N}}\varphi(\tau)\left\{\hat{\gamma}-\gamma\right\} + o_p(1) \\ &= \upsilon_{\text{N}}(\tau;\boldsymbol{Z}) + \xi_{\text{N}}(\tau;\boldsymbol{Z}) + o_p(1) \ ,\end{aligned} \tag{I.2}$$

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 (I.2) 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 $\sup_{\mathscr{T}}|\varphi(\tau)| < \infty$. Therefore, $\{\xi_{\text{N}}(\tau;\boldsymbol{Z}) : \tau \in \mathscr{T}\}$ converges weakly in $\ell^{\infty}(\mathscr{T})$ to a mean zero Gaussian process $\xi(\cdot)$ with covariance function $\mathbb{C}(\xi(\tau_1),\xi(\tau_2)) = \varphi(\tau_1)\varphi(\tau_2)\sigma_0^2$. Thus, the limit process $\upsilon(\cdot)+\xi(\cdot)$ is a Gaussian process with zero mean and covariance function

$$\begin{aligned}\mathbb{C}(\upsilon(\tau_1),\xi(\tau_2)) &= \min\{\tau_1,\tau_2\} - \tau_1\tau_2 + \varphi(\tau_1)\varphi(\tau_2)\sigma_0^2 \\ &\quad + \tau_1(1-\tau_1)\varphi(\tau_2)\{\mathbb{E}(Y_0|Y_0 \le F_0^{-1}(\tau_2)) - \mathbb{E}(Y_0|Y_0 > F_0^{-1}(\tau_2))\} \ .\end{aligned} \tag{I.3}$$

This finishes the first part of the proof. Note that the maps $v \to \|v\|$ 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.

I.2 Proof of Lemma 2

Recall the martingale-transformation of the quantile process (15) is given by

$$\begin{aligned}\tilde{v}_{\mathrm{N}}(\tau;\boldsymbol{Z}) &= \hat{v}_{\mathrm{N}}(\tau;\boldsymbol{Z}) - \psi_g\left(\hat{v}_{\mathrm{N}}\right)(\tau;\boldsymbol{Z}) \\ &= v_{\mathrm{N}}(\tau;\boldsymbol{Z}) + \varphi(\tau)\xi_{\mathrm{N}}(\boldsymbol{Z}) - \psi_g\left(\hat{v}_{\mathrm{N}}\right)(\tau;\boldsymbol{Z}) + o_p(1) \ ,\end{aligned} \tag{I.4}$$

where the second equality follows by the asymptotic expansion (I.2) and the $o_p(1)$ term holds uniformly over $\mathscr{T}$. By properties of the compensator ψ_g and (I.2), we have that

$$\psi_g\left(\hat{v}_{\mathrm{N}}\right)(\tau;\boldsymbol{Z}) = \psi_g\left(v_{\mathrm{N}}\right)(\tau;\boldsymbol{Z}) + \varphi(\tau)\xi_{\mathrm{N}}(\boldsymbol{Z}) + o_p(1) \ . \tag{I.5}$$

Plugging (I.5) into (I.4) yields

$$\tilde{v}_{\mathrm{N}}(\tau;\boldsymbol{Z}) = v_{\mathrm{N}}(\tau;\boldsymbol{Z}) - \psi_g\left(v_{\mathrm{N}}\right)(\tau;\boldsymbol{Z}) + o_p(1) \ , \tag{I.6}$$

and therefore, $\{\tilde{v}_{\mathrm{N}}(\tau;\boldsymbol{Z}) : \tau \in \mathscr{T}\}$ converges weakly to $\zeta(\cdot)$, the standard Brownian motion by Khmaladze (1981, 4.3). This finishes the proof of the fist part of the lemma. A direct application of the CMT as in the proof of Lemma 1 finishes the proof.

II BGH and The Durbin Problem

BGH apply a Fisher-randomization test using the plug-in method for the individual hypotheses (21) (Bitler, Gelbach, and Hoynes, 2017, Section V.B, p. 694). To formalize the ongoing discussion, we need more notation. Let the observed data for each mutually exclusive subgroup be given by $\boldsymbol{Z}^s = \left(Z_1^s \ldots, Z_{N_s}^s\right) = \left(Y_{s,1}^1, \ldots, Y_{s,m_s}^1, Y_{s,1}^0, \ldots, Y_{s,n_s}^0\right)$ for all $1 \leq s \leq \mathcal{S}$, where every subgroup $\boldsymbol{Z}^s$ has $N_s = m_s + n_s$ observations.

Under the assumptions of the CSTE model, $Y_{1,s} = Y_{0,s} + \delta_s$ for all s. Then, one might shift the observations in the control group, $Y_{0,s}$, by adding these subgroup-specific δ_s, that we can estimate as the ATEs within subgroup s, to the actual outcome. Let $\hat{Y}_{0,s} = Y_{0,s} + \hat{\delta}_s$ be the simulated outcomes within subgroup s. This simple transformation across subgroups yields a simulated outcome under treatment, $\{\hat{Y}_{0,s} : 1 \leq s \leq \mathcal{S}\}$.

BGH argue that if the CSTE model is a good representation of HTE, then the distribution of the simulated outcomes should be "close" in some sense to $F_{1,s}(\cdot)$, the distribution of the observed outcomes under treatment. Therefore, BGH's permutation test is based on the two-sample Kolmogorov–Smirnov test statistic (2SKS) for each subgroup,

$$K^s_{\mathrm{N},\hat{\delta}}(\boldsymbol{Z}^s) = \sup_{y \in \mathbb{R}} \left| V^s_{\mathrm{N}}(y, \hat{\delta}_s; \boldsymbol{Z}^s) \right| , \tag{II.1}$$

where

$$V^s_{\mathrm{N}}(y, \hat{\delta}_s; \boldsymbol{Z}^s) = \sqrt{\frac{m_s n_s}{N_s}} \left\{ \hat{F}_{1,s}(y) - \hat{F}_{0,s}(y - \hat{\delta}_s) \right\} . \tag{II.2}$$

From the previous test statistic, one can define BGH's permutation test as follows. Let $\boldsymbol{G}_s$ be the set of all permutations π of $\{1, \ldots, N_s\}$. For a fixed subgroup s, BGH's permutation test based on 2SKS rejects the individual hypothesis (21) if the observed (II.1) exceeds the $1 - \alpha/\mathcal{S}$ quantile of the permutation distribution:

$$\hat{R}^{K(\hat{\delta})}_{\mathrm{N},s}(t) = \frac{1}{N_s!} \sum_{\pi \in \mathbf{G}_s} \mathbb{1}_{\left\{ K^s_{\mathrm{N},\hat{\delta}}\left(Z^s_{\pi(1)}, \ldots, Z^s_{\pi(N_s)} \right) \leq t \right\}} . \tag{II.3}$$

II.1 Commentary on BGH's Approach

BGH state that the critical values derived from (II.3) are asymptotically valid even in the presence of estimated parameters (Section V, p. 694). However, this claim is false. To establish the asymptotic validity of the permutation test, we need to show that the permutation distribution (II.3) approximates the true unconditional sampling distribution of (II.1). Under relatively weak assumptions, we can show that (II.3) behaves like the distribution of (the supremum of) a Brownian bridge (Chung and Olivares, 2021, Theorem 2). In contrast, the limiting distribution of the 2SKS is given by the distribution of (the supremum of) a different Gaussian process; it has mean 0 and a covariance structure that depends on unknown parameters as proved in Ding, Feller, and Miratrix (2016, Theorem 4). Therefore, the permutation test based on the 2SKS fails to control the type 1 error asymptotically.

Then, how do we make sense of BGH's claims? The authors' justification relies on a result by Præstgaard (1995) that states the permutation empirical process converges weakly to a Brownian bridge corresponding to a mixture measure. Indeed, the testing problem in BGH's environment satisfies the premises in Præstgaard's (1995), so (II.3) asymptotically does behave like the process in Præstgaard (1995).

In other words, the asymptotic behavior of the permutation distribution (II.3) does not change when the estimated δ_s enter the test statistic instead of the known value δ_s. However, it is not the case for the true unconditional limiting distribution of the test statistic. The asymptotic behavior of the 2SKS statistic does change in the presence of estimated parameters. Therefore, the permutation distribution does not mimic the true unconditional distribution of the test statistic in large samples when δ_s is being estimated, invalidating BGH's permutation test. Will BGH's approach ever be valid? Yes, but only in the infeasible scenario when we know the subgroup-specific δ_s. In fact, in this case, the permutation test achieves the finite sample exactness. Therefore, BGH's claim that their method yields asymptotically valid inference is only well-grounded in this extraordinary case and incorrect otherwise.

III Empirical Results from BGH

Table 1 displays the results from BGH empirical analysis and the proposed permutation test based on the quantile process in Section 4.2. Columns 3–4 contain the empirical results from BGH's Table 2, which we include verbatim for a fair comparison. Meanwhile, columns 5–6 contain the results from our proposed method. We note that a joint test of the family of null hypotheses across these subgroups rejects if we reject any one of the subgroup-specific null hypotheses (see BGH's Section V).

Table 1: 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	3	3
Education	21	3	1	3	1
Age of youngest child	21	3	1	3	3
Marital status	21	2	1	3	2
Earnings level seventh Q pre-RA	21	2	1	0	0
Number of pre-RA Q with earnings	21	1	0	1	0
Welfare receipt seventh Q pre-RA	14	3	3	2	2
Education subgroups interacted with					
Age of youngest child	49	1	0	6	5
Marital status	35	3	3	4	2
Earnings level seventh Q pre-RA	63	1	0	0	0
Number of pre-RA Q with earnings	63	0	0	2	1
Welfare receipt seventh Q pre-RA	42	1	0	1	0
Age of youngest child interacted with					
Marital status	35	1	1	3	1
Earnings level seventh Q pre-RA	63	0	0	1	0
Number of pre-RA Q with earnings	49	1	1	3	1
Welfare receipt seventh Q pre-RA	42	1	0	1	0
Marital status subgroup interacted with					
Earnings level seventh Q pre-RA	63	2	1	1	0
Number of pre-RA Q with earnings	63	0	0	2	0
Welfare receipt seventh Q pre-RA	42	1	0	2	1
Earnings level seventh Q pre-RA subgroups interacted with					
Number of pre-RA Q with earnings	49	0	0	2	0
Welfare receipt seventh Q pre-RA	42	1	1	1	1
Number of quarters any earnings pre-RA subgroup interacted with					
Welfare receipt seventh Q pre-RA	42	0	0	2	1

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

References

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.

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

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.

Præstgaard, J. T. (1995). Permutation and bootstrap kolmogorov-smirnov tests for the equality of two distributions. *Scandinavian Journal of Statistics*, pages 305–322.

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