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

EunYi Chung[†]
Department of Economics
UIUC
eunyi@illinois.edu

Mauricio Olivares
Department of Economics
UIUC
lvrsgnz2@illinois.edu

August 2, 2021

Abstract

A common empirical practice for testing heterogeneous treatment effects compares the means of subgroups. To challenge the soundness of this practice, Bitler, Gelbach, and Hoynes (2017, BGH) introduce a permutation test that is based on a test statistic with estimated parameters. In this comment, we argue that BGH's permutation test fails to control the type 1 error asymptotically unless the means are known. To fix this problem, we provide an asymptotically valid permutation test based on a modified quantile process. Numerical evidence shows that our method outperforms alternative quantile-based procedures.

Keywords: Permutation Test, Quantile Regression, Heterogeneous Treatment Effects, Connecticut's Jobs First.

JEL Classification: C12, C14, C46, I38.

[†] We would like to thank Marianne Bitler, Jonah Gelbach, and Hilary Hoynes for their useful comments and generous feedback that led to the improvement of the paper as well as for making their code for replication publicly available. We have also benefited from discussions with Dan Bernhardt and Roger Koenker during this project. Lastly, we are grateful to MDRC for granting access to the experimental data used in this paper. All errors are our own.

1 Introduction

Empirically oriented researchers and policymakers are interested in knowing to what extent the effect of an experiment or policy intervention varies across subgroups formed by pre-treatment characteristics. One popular approach to accomplish this goal is to estimate average treatment effects that vary across the subgroups but remain constant within the subgroups. Indeed, according to a survey of published papers in top economic journals, 40% of the studies report at least one treatment effect in this fashion (Chernozhukov et al., 2018).

While this type of subgroup analysis is easy to implement and may capture some form of variation, its inadequacy to fully assess treatment effect heterogeneity has been realized and discussed. In a significant contribution, Bitler, Gelbach, and Hoynes (2017, BGH) introduce a permutation test to evaluate the suitability of this standard approach to assessing heterogeneity in the treatment effect. Their suggested testing procedure relies on the following observation. If the conventional approach were correct, one might shift the observations in the control group by adding these subgroup-specific average treatment effects to the actual outcome. This simple transformation across subgroups yields a simulated outcome under treatment. Thus, if the standard approach is a good representation of the heterogeneity in the treatment effect, then the distribution of the simulated outcomes should be close in some sense to the distribution of the observed outcomes under treatment.

We argue in the next section that BGH's proposed permutation test to verify the adequacy of the mean impacts approach is invalid and potentially misleading. The main challenge for BGH's testing procedure is that when constructing the distribution of simulated outcomes, the subgroup-specific average treatment effects are unknown and we need to estimate them. However, when we plug the estimated average treatment effects into the test statistic, the large-sample behavior of the so-called permutation distribution, defined in (4), does not mimic the true sampling distribution, thus invalidating BGH's permutation test.

In this comment, we explore one path to restore the asymptotic validity of the permutation test with estimated nuisance parameters based on a modified quantile process. We motivate our quantile-based method by exploiting two facts. First, on a purely theoretical level, we can formulate BGH's testing problem of comparing cumulative distribution functions (CDFs) into one of comparing quantile treatment effects (QTE). The second fact responds to BGH's empirical findings—their work shows how QTE capture the heterogeneity in the treatment effect as predicted by the economic theory. Thus, we see our approach as a more direct, intuitive, and unifying way to represent BGH's idea.

Our strategy consists of applying 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 that the permutation distribution of the newly modified statistic and the true sampling distribution are asymptotically the same, thus restoring the asymptotic validity of the permutation test in the presence of estimated parameters.

To complement our theoretical result and ease the implementation of our permutation test, we also provide free software in the `RATest` R package, available on CRAN. Moreover, we discuss in the online appendix a fast computational implementation based on the preprocessing algorithm of Portnoy and Koenker (1997). This way, our proposed method not only corrects the flaws in BGH's testing procedure, but also offers a reliable course of action to perform asymptotically valid permutation-based inference via quantiles at a low cost.

The following two sections illustrate the invalidity of BGH's proposed method and how we can overcome its main difficulties. We collect all the formal statements, proofs, simulations, empirical applications, and technically nuanced discussions in a separate appendix for brevity.

2 The Invalidity of the Simulated Outcomes Approach

BGH are interested in testing the null hypothesis that treatment effects are constant within subgroups $s \in \mathcal{S}$ for those with positive earnings. To set the scene, let Y_s^1 and Y_s^0 denote the outcome variable for those in the treatment and control groups, respectively, with CDFs $F_{1,s}(\cdot)$ and $F_{0,s}(\cdot)$, for subgroup $1 \leq s \leq \mathcal{S}$. The goal is to simultaneously test a finite number of hypotheses $H_{0,s}$ $(s = 1, \dots \mathcal{S})$, where each individual hypothesis is given by

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

and δ_s is the unknown subgroup-specific treatment effect that we need to estimate (e.g., as the difference in sample means between both groups). We reject the joint hypothesis $H_{0,1}, \dots, H_{0,\mathcal{S}}$ if any one of the null hypotheses (1) is rejected.[1] Needless to say, if δ_s were known, a permutation test for (1) would retain the finite-sample validity though this case is infeasible in practice.

BGH apply a Fisher-randomization test using the plug-in method (FRT-PI) for the individual hypotheses (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 $\boldsymbol{Z}^s = \left(Z_1^s \dots, Z_{N_s}^s\right) = \left(Y_{s,1}^1, \dots, Y_{s,m_s}^1, Y_{s,1}^0, \dots, Y_{s,n_s}^0\right)$ for all $1 \leq s \leq \mathcal{S}$, where every subgroup $\boldsymbol{Z}^s$ has $N_s = m_s + n_s$ observations. BGH's permutation test is based on the two-sample Kolmogorov–Smirnov test statistic (2SKS) for each subgroup,

$$K_{\text{N},\hat{\delta}}^s(\boldsymbol{Z}^s) = \sup_{y \in \mathbb{R}} \left| V_{\text{N}}^s(y, \hat{\delta}_s; \boldsymbol{Z}^s) \right| \ , \tag{2}$$

[1] To control the family-wise error rate, BGH use the Bonferroni adjustment. Alternatively, we may consider a stepwise multiple testing procedure e.g. the Westfall–Young algorithm or Holm's method. For more details, see the setup and implementation algorithms in Chung and Olivares (2021, Section 4 and Appendix D).

where

$$V^s_{\mathrm{N}}(y, \hat{\delta}_s; \boldsymbol{Z}^s) = \sqrt{\frac{m_s n_s}{N_s}} \left\{ \hat{F}_{1,s}(y) - \hat{F}_{0,s}(y - \hat{\delta}_s) \right\} \ . \tag{3}$$

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 (1) if the observed (2) exceeds the $1 - \alpha/\mathcal{S}$ quantile of the permutation distribution:

$$\hat{R}^{K(\hat{\delta})}_{\mathrm{N},s}(t) = \frac{1}{N_s!} \sum_{\pi \in \mathbf{G}_s} \mathbb{1}_{\left\{ K^s_{\mathrm{N},\hat{\delta}}\left(Z^s_{\pi(1)}, \ldots, Z^s_{\pi(N_s)}\right) \leq t \right\}} \ . \tag{4}$$

BGH state that the critical values derived from (4) 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 (4) approximates the true unconditional sampling distribution of (2). Under relatively weak assumptions, we show that (4) 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 (2) 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 (4) asymptotically does behave like the process in Præstgaard (1995). In other words, the asymptotic behavior of the permutation distribution (4) 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.

To confirm our point, and to highlight the detrimental effects of estimated δ_s on inference, we conduct a Monte Carlo study in Section 2 in the appendix. Our exercise considers BGH's testing procedure with known and estimated δ_s. When δ_s are known, BGH's permutation test delivers empirical rejection rates close to the nominal level, as predicted by the theory. However, when δ_s are estimated, BGH's permutation test delivers rejection probabilities under the null that are considerably different from α across all the specifications we consider (see Table 1 in the appendix). These size distortions caution against BGH's approach, a fact we also document by revisiting their empirical application (e.g., Table 2 in appendix).

Then, how can we restore the asymptotic validity of the testing procedure with estimated parameters? There are two leading approaches in the literature. The first one entails abandoning the asymptotically distribution-free nature of the test based on (3) and adopting a resampling strategy that yields asymptotically-valid critical values. For instance, one could use bootstrap methods like those presented in Linton, Maasoumi, and Whang (2005) and apply them to this context. Another possibility in this line of attack comprises tests based on subsampling the quantile process, e.g., Chernozhukov and Fernández-Val (2005).

Alternatively, one could modify the test statistic so that it becomes asymptotically pivotal, effectively removing the estimation effects, e.g., Durbin (1973); Khmaladze (1981); Koenker and Xiao (2002). We stick to this approach. More specifically, we will capitalize on the QTE found by BGH and define an asymptotically pivotal test statistic based on a modified quantile process. Then, following the methodology advocated by Neuhaus (1993); Janssen (1997); Chung and Romano (2013, 2016), we show that the permutation distribution based on the modified statistic asymptotically behaves like the true unconditional distribution of the new test statistic, reviving the asymptotic validity of the permutation test[2].

3 An Alternative Quantile-Based Permutation Test

At the core, BGH's critique originates from showing how QTE unmask the heterogeneous labor supply effects of Jobs First (JF), Connecticut's welfare reform. Concretely, BGH show that while the QTE are consistent with the predictions of the neoclassical labor supply model, these patterns are primarily missed by solely looking at the mean impacts.[3] Thus, it makes it more intuitive to work with a test statistic that directly relies upon the QTE. To this end, we cast BGH's testing problem (1) in terms of quantiles instead of CDFs.[4]

To follow up on the previous discussion, we formulate the individual hypotheses (1) for each subgroup s in terms of the quantile functions of the treatment and control groups,

$$H^q_{0,s} : F^{-1}_{1,s}(\tau) - F^{-1}_{0,s}(\tau) = \gamma_s \ \forall\, \tau \in [0, 1], \quad \text{for some} \quad \gamma_s \ , \tag{5}$$

[2] As another option, one could follow Ding, Feller, and Miratrix (2016) and implement their permutation test. This procedure is valid and works without modifying the test statistic or resorting to asymptotic approximations, though this approach is generally conservative; see Chung and Olivares (2021, Table 1).

[3] The standard neoclassical labor supply model predicts a welfare program changes the incentives for individuals to locate above or below the eligibility point, and this behavioral response sets off work incentives and program enrollment. See Moffitt (2002) for more on this.

[4] Chung and Olivares (2021, Section 3) employ the Khmaladze transformation of the empirical process (3) to introduce an asymptotically valid permutation test for hypotheses based on CDFs, like (1), as opposed to quantiles like we do in this section.

and consider a test statistic based on the quantile process instead (see Chernozhukov and Fernández-Val, 2005, Example 2). However, the typical quantile-based KS-type statistic for (5) is not immune to the estimated-parameter problem either—γ_s is unknown and we need to estimate it. Analogously to the 2SKS statistic case, the limiting distribution of the quantile-based test statistic for (5) is not asymptotically distribution-free (see Theorem 1 in the appendix).

We sidestep the negative effects of estimated γ_s in two steps. First, we modify the test statistic so that its limiting distribution does not depend on the fundamentals. To do so, we use the Khamaladze transformation of the quantile process suggested by Koenker and Xiao (2002). Then, we use the newly-modified test statistic as the input for the permutation distribution (4).

More specifically, let $\tilde{\upsilon}_{\mathrm{N}}(\tau, \boldsymbol{Z}^s)$, $\tau \in \mathscr{T}$, be the martingale transformation of the quantile process as in Koenker and Xiao (2002, eq. (4.6)), where $\mathscr{T}$ is a closed subinterval of $(0, 1)$. This gives rise to the martingale-transformed test statistic

$$\tilde{K}^s_{\mathrm{N}}(\boldsymbol{Z}^s) = \sup_{\tau \in \mathscr{T}} |\tilde{\upsilon}_{\mathrm{N}}(\tau; \boldsymbol{Z}^s)| \ . \tag{6}$$

Under some regularity assumptions that hold for the problem at hand, Koenker and Xiao (2002) showed that (6) converges in distribution to a continuous CDF $H(\cdot)$. Let us now establish the asymptotic behavior of the permutation distribution. To this end, let $\hat{R}^{\tilde{K}}_{\mathrm{N},s}$ denote the permutation distribution (4) based on (6). We show in Theorem 3 in the appendix that, under the same regularity conditions, we have

$$\sup_{0 \le t \le 1} \left| \hat{R}^{\tilde{K}}_{\mathrm{N},s}(t) - H(t) \right| \xrightarrow{\mathrm{p}} 0 \ . \tag{7}$$

This result states that the permutation distribution asymptotically approximates the true unconditional limiting distribution of (6). Hence, our permutation test has limiting rejection probability equal to α. Moreover, (7) implies that the permutation test has the same limiting local power as the test based on (6) for contiguous alternatives. Further, it is also worth pointing out that the asymptotic validity prevails even though we need to estimate the density and score functions to calculate the Khmaladze transformation. Summing up, our permutation test delivers a robust method to carry on asymptotically valid inference for (5) in the presence of estimated nuisance parameters.

To highlight the finite sample performance of our method, we conduct a Monte Carlo experiment in the appendix. Our proposed permutation test controls size remarkably well across multiple specifications, outperforming other popular quantile-based tests that handle estimated parameters, most notably Koenker and Xiao (2002); Chernozhukov and Fernández-Val (2005); Linton, Maasoumi, and Whang (2005). Our numerical exercise shows that our proposed method also has greater power than subsampling and Koenker and Xiao's (2002) test for all the alternatives we examine. We summarize these results in Section 7 in the appendix. For the numerical calculation of the test, we provide free software available in the `RATest` `R` package.

4 Conclusions

While we subscribe to BGH's message that estimating subgroup-specific average treatment effects for subgroups defined by covariates is an inadequate account of the heterogeneity in the treatment effect, we must stress that a correct assessment of this standard practice using a permutation test is just as critical. This comment provides a reliable course of action to perform asymptotically valid permutation-based inference for testing heterogeneous treatment effects.

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.

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.

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)*.

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

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

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

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.

Moffitt, R. A. (2002). Welfare programs and labor supply. *Handbook of public economics*, 4:2393–2430.

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

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