AN EXACT t-TEST

GUILLAUME A. POULIOT
Harris School of Public Policy, University of Chicago

For most recent version, please click here.

Multivariate linear regression and randomization-based inference are two essential methods in statistics and econometrics. Nevertheless, the problem of producing a randomized test for the value of a single regression coefficient that is exactly valid when errors are exchangeable, and which is asymptotically valid for the best linear predictor, has remained elusive. In this paper, we produce a test that is exactly valid with exchangeable errors and which allows for general covariate designs; covariates may be continuous as well as discrete and may be correlated. The test is asymptotically valid when the errors are not exchangeable, in particular in the presence of conditional heteroskedasticity.

KEYWORDS: Randomization inference, linear regression.

1. INTRODUCTION

Consider the linear regression model

$$Y_i = \beta_0 + X_{i1}\beta_1 + \cdots + X_{ip}\beta_p + \varepsilon_i,\ i = 1, \ldots, n, \tag{1}$$

where $\{Y_i, X_i, \varepsilon_i\}_{i=1}^n$ is drawn from an unknown distribution $\mathcal{P}_n$. For ease of exposition, we will consider the vector representation of the bivariate model,

$$\mathbf{Y} = \mathbf{1}_n\beta_0 + \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \boldsymbol{\varepsilon}, \tag{2}$$

where $\mathbf{1}_n$ is a $n \times 1$ vector of 1's, $\mathbf{X}_j = (X_{1j}, \ldots, X_{nj})^T$, $j = 1, 2$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$. All results and manipulations can be done with minor modifications for matrices of general dimensions. In particular, the reader may think of $\mathbf{X}_2$ as being in $\mathbb{R}^{p-1}$ throughout without any formal incongruities, and all technical results are proved for that case.

Our interest lies in testing the value of the first slope coefficient, i.e., in testing the null,

$$H_0 : \beta_1 = \beta_1^0. \tag{3}$$

Importantly, inverting such a test produces a marginally valid confidence interval for β_1. Of course, the first covariate is chosen without loss of generality since the order of the covaraites does not impact the value of their estimate.

The standard, perhaps ubiquitous, approach for testing (3) is to use the classical t-test (Student, 1908, Wooldridge, 2010). This test is asymptotically valid under standard conditions on the first and second moments. In the event that the errors are independent and identically distributed Gaussian random variables, the t-test will be *exact* in the sense that its probability

---

Guillaume A. Pouliot: guillaumepouliot@uchicago.edu

This project is the product of long-ongoing conversations with Gary Chamberlain, Jean-Marie Dufour, James Heckman and Azeem Shaikh. The manuscript has benefited from insightful comments from Stéphane Bonhomme, Matias Cattaneo, Max Farell, Michael Leung, Lihua Lei, Matt Masten, Mauricio Olivares, Mikkel Plagborg-Møller, Kirill Ponomarev, Whitney Newey, Joseph Romano, Max Tabord-Meehan and Alexander Torgovitsky. Deniz Dutz and Samuel Higbee have read an early version of the document and provided very helpful feedback. I thank the participants of the Duke University Statistics and Data Science seminar, the Munich Econometrics Seminar, the CEMFI econometrics seminar, and the Paris Econometrics Seminar for helpful comments. Ian Xu has provided outstanding research assistance. All errors are my own.

of false rejection in finite samples will not exceed the user-prescribed, nominal probability $\alpha \in [0,1]$. In fact, the t-test will have probability of rejection exactly α under the null; we say that it has *exact nominal size*.

Another classical approach to testing is randomization. Indeed, in applications such as one- or two-sample tests for the mean, or tests of economic inequality (Dufour et al., 2019), and many others (Romano, 1990, Lehmann and Romano, 2005), randomization-based approaches deliver exact tests without making assumptions about the specific distribution of the data. To the extent that such tests have large-sample performance comparable to that of standard tests under comparable assumptions, they may consist in a desirable alternative given their different, sometimes better adapted small-sample properties and weaker assumptions.

The question of producing an exactly valid, randomization-based test for (3) that is asymptotically valid under general conditions thus arises very naturally.

However, and rather remarkably, Lei and Bickel (2021) provide an extensive literature review substantiating the claim that, for null hypotheses such as (3) , *"there is no test that is exact under reasonably general assumptions*". Their notion of "general assumptions" rules out covariate designs where all nuisance covariates are discrete. They presumably rule out $\max$-p procedures as too conservative,[1] or are restricting themselves to tests of exact nominal size.

To the best of my knowledge, Lei and Bickel (2021) produce the first such exact test for general linear hypotheses in the linear model (1). They use a construction that cleverly builds the required invariances into the test statistic.

Wen et al. (2022) later produced another such test which is better adapted to high-dimensional regression and which, although it does not have nominal size—i.e., it rejects under the null hypothesis with probability lower than the nominal level α— in some cases displays better power properties.

To the best of my knowledge, Wen et al. (2022) are the first to use the idea of somehow explicitly "orthogonalizing" permuted nuisance regressors. This idea, while it is instantiated quite differently here, plays a prominent role in the proposed method.

While these procedures closed an important mathematical statistics question, neither of these are shown to be even asymptotically valid when exchangeability does not avail, for instance under heteroskedasticity —indeed, they are not, see Section 3. This is critical for applications since in practice we typically cannot assume that the conditional expectation is linear, but are content with estimating its best approximation, the best linear predictor $E^*[Y_i|X_i] = X_i^T\beta$, $X_i = (1, X_{i1}, \ldots, X_{ip})$, with errors $\varepsilon_i = Y_i - E^*[Y_i|X_i]$ satisfying orthogonality conditions $E[\varepsilon_i] = 0$ and $E[\varepsilon_i \cdot X_{ij}] = 0$, $j = 1, \ldots, p$. It is then a rather cavalier assumption to make about $\mathcal{P}_n$ that the thus produced errors retain no dependence with the covariates. Heterogeneous treatment effects, for instance, bring about heteroskedasticity (Breusch and Pagan, 1979). To be sure, this is an assumption we have to make to produce exactly valid tests, but it is not an assumption we have to make in order to produce asymptotically valid tests—considering, for instance, the classical heteroskedasticity robust t-test.

As it pertains to producing such an asymptotically robust randomization test, the case in which the tested covariate is independent of other covariates is covered in DiCiccio and Romano (2017) and substantially generalized in Young (2023) who also provides improved asymptotic robustness guarantees.

To the best of my knowledge, the first and only exactly valid procedure allowing for dependent covariates that is also asymptotically valid for the best linear predictor under heteroskedas-

[1] Given a procedure producing a p-value $p(\beta_1, \beta_2, \ldots, \beta_p)$ for null hypotheses on the full vector $(\beta_1, \beta_2, \ldots, \beta_p)$, we speak of $\max_{\beta_2,\ldots,\beta_p} p(\beta_1, \beta_2, \ldots, \beta_p)$ as the max-p statistic, a conservative yet valid p-value corresponding to the projection method.

ticity was produced by D'Haultfœuille and Tuvaandorj (2024). Their work builds on the studentization strategy of DiCiccio and Romano (2017) and others cited therein. Their method applies to the case in which all nuisance covariates are discrete and each support point of the nuisance covariates vector is sampled an asymptotically growing number of times. They state that "*constructing a permutation test for subvectors that is both exact under independence and asymptotically heteroskedasticity-robust for any design*" is yet unresolved.

We propose the following test statistic, which will be shown to produce a randomized test that is exact when errors are exchangeable, and to be asymptotically valid for the best linear predictor, allowing in particular for conditional heteroskedasticity. Let $\mathbf{G}_n$ be a group of linear transformations (e.g., any subgroup of the group of permutations) of $\{1, \ldots, n\}$ onto itself. Define, for each $g \in \mathbf{G}_n$, the statistic

$$t_g\left(\beta_1^0\right) = \frac{\bar{\mathbf{X}}_1^T\left(\mathbf{Y} - \mathbf{X}_1\beta_1^0\right)_g}{\hat{\sigma}_g}, \tag{4}$$

where

$$\hat{\sigma}_g^2 = \frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{X}}_{i1}^2\hat{\varepsilon}_{g(i)}^2, \tag{5}$$

$\bar{\mathbf{X}}_1 = \mathcal{Q}\mathbf{X}_1$, $\mathcal{Q}$ is the orthogonal projection onto the orthogonal complement of the span of $\{g\mathbf{X}_2\}_{g\in\mathbf{G}_n}$, and the residuals $\hat{\varepsilon}_i$ obtain from the regression of $\mathbf{Y}$ on an intercept term, $\{g\mathbf{X}_1\}_{g\in\mathbf{G}_n}$ and $\{g\mathbf{X}_2\}_{g\in\mathbf{G}_n}$. As detailed in Section 2, we suggest to use a group of non-overlapping block permutations as the set of group actions $\mathbf{G}_n$.[2]

To the best of my knowledge, the randomization test using the above statistic is the first exact randomization test of nominal size[3] for (3) that is asymptotically valid under conditional heteroskedasticity.

From a practical standpoint, the most natural alternative to the suggested procedure is the classical t-test. The proposed test is exactly valid under strictly weaker assumptions than the classical t-test. Indeed, as illustrated with power plots in Section 3, while the classical t-test may outperform the suggested procedure in small samples with normally distributed errors, it is invalid in small samples with non-normally —in particular, skewed— distributed errors and in samples of any size with conditionally heteroskedastic errors. The proposed test, on the other hand, is exactly valid with any exchangeable errors, and asymptotically valid with conditionally heteroskedastic errors. The classical heteroskedasticity-robust t-test is observed to be invalid in small samples, even with Gaussian data, while the proposed procedure, which remains exactly valid with exchangeable data, furthermore appears to approach its power in large-samples, suggesting it can be a competitive choice if the data analysts wants a heteroskedasticity-robust test.

Finally, the more general case of testing a general linear hypothesis $R\beta = 0$ for any given full rank matrix R, which Lei and Bickel (2021) show is tantamount to testing a that a subvector of β is equal to zero, can be handled by an immediate extension of the results herein and using the quadratic form analog of (4).

[2]Keeping with the literature, we use the group action notation g such that its explicit form is clear from context. In particular, $g\mathbf{Z}$ may read as applying a $n \times n$ permutation matrix g to the $n \times 1$ vector $\mathbf{Z}$, and that is equal to $\mathbf{Z}_g = (\mathbf{Z}_{g(1)}, \ldots, \mathbf{Z}_{g(n)})^T$ where g reads as a function from and to $\{1, \ldots, n\}$ corresponding to the same permutation.

[3]Or, to put it more transparently, the first test statistic that is "inherently" invariant with respect to the nuisance coefficient β_2. Because it takes a worst case over nuisance parameters, we do not consider the $\max$-p statistic to be "inherently" invariant. We consider this a reasonable criteria since it excludes procedures that are intrinsically conservative.

The remainder of this paper is divided as follows. In Section 2, I develop the proposed test in more detail. In Section 3, I characterize the power properties, in small and large samples, of the proposed test as compared with some available alternatives. Section 4 discusses and concludes.

2. METHODOLOGY

Consider the proposed statistic (4) with explicit parametrization,

$$t_g\left(\beta_1^0\right) = t\left(\beta_1^0 : \mathbf{X}_1, \mathbf{X}_2, g\left(\mathbf{Y} - \mathbf{X}_1\beta_1^0\right)\right),$$

where we emphasize its arguments and that the group action is only applied to the "short residuals" $\mathbf{Y} - \mathbf{X}_1\beta_1^0$.

We recall the standard mechanics of the construction of a randomization test. Suppose $M = |\mathbf{G}_n|$, the cardinality of $\mathbf{G}_n$. Let $\mathbf{W} = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y} - \mathbf{X}_1\beta_1^0)$ and use the shorthand $g\mathbf{W} = (\mathbf{X}_1, \mathbf{X}_2, g(\mathbf{Y} - \mathbf{X}_1\beta_1^0))$. For every $\mathbf{W} \in \mathrm{supp}(\mathbf{W})$, let

$$t^{(1)}(\mathbf{W}) \leq t^{(2)}(\mathbf{W}) \leq \cdots \leq t^{(M)}(\mathbf{W})$$

be the ordered values of $t(\beta_1^0 : g\mathbf{W})$ as g varies over $\mathbf{G}_n$ and write $t^{(i)}$ in lieu of $t^{(i)}(\mathbf{W})$, for any given i, when no confusion arise. Given a nominal level α, let k be defined as

$$k = M - \lfloor M\alpha \rfloor,$$

where $\lfloor M\alpha \rfloor$ denotes the largest integer less than or equal to $M\alpha$.

Let $\hat{t} := t(\beta_1^0 : \mathbf{W})$ be the realized test statistic, and define the test function[4]

$$\phi(\mathbf{W}) = \begin{cases} 1, & \hat{t} > t^{(k)} \\ a, & \hat{t} = t^{(k)} \\ 0, & \hat{t} < t^{(k)} \end{cases},$$

with $a = (M\alpha - M^+)/M^0$, where M^+ and $M^{(0)}$ are the number of values $t^{(j)}$, $j = 1, ..., M$, that are greater than $t^{(k)}$ and equal to $t^{(k)}$, respectively.

Intuitively, if the distribution of the test statistic is invariant under group actions $g \in \mathbf{G}_n$, then the empirical distribution $\{t(\beta_1^0 : g\mathbf{W})\}_{g \in \mathbf{G}_n}$ may be used as a randomization distribution with respect to which the quantile of the observed test statistic $\hat{t}$ is a valid p-value.

2.1. *Exact validity*

Consider the model (2). The *strong null* we are willing to make in order to obtain exact validity is that the errors are exchangeable. We are not, however, willing to assume that the tested and nuisance covariates are exchangeable with respect to each other.

ASSUMPTION 2—Strong Null: *For any* $g \in \mathbf{G}_n$,

$$(\mathbf{X}_1, \mathbf{X}_2, \boldsymbol{\varepsilon}) \sim (\mathbf{X}_1, \mathbf{X}_2, g\boldsymbol{\varepsilon}).$$

[4]In practice, one may use the slightly conservative but simpler *nonrandomized test* with $k = \lceil (1-\alpha) M \rceil$ and

$$\phi(\mathbf{W}) = \begin{cases} 1, & \hat{t} > t^{(k)} \\ 0, & \hat{t} \leq t^{(k)} \end{cases}.$$

Note that the symbol "$\sim$" used in $A \sim B$ denotes that A and B have the same distribution. When Assumption 2 is satisfied for a given $\mathcal{P}_n$ and a choice of $\mathbf{G}_n$, the proposed procedure is exactly valid.

THEOREM 1—Validity of Exact t-Test: *Suppose Assumption 2 holds and* $\mathbf{Y} - \mathbf{X}_1\beta_1^0$ *is not in the column span of* $\{g(\mathbf{X}_1, \mathbf{X}_2)\}_{g\in\mathbf{G}_n}$. *Then, under the strong null,*

$$E_{\mathcal{P}}\left[\phi\left(\mathbf{Y}, \mathbf{X}_1, \mathbf{X}_2\right)\right] = \alpha. \tag{6}$$

PROOF OF THEOREM 1: We omit the subscript and use $\mathbf{G} = \mathbf{G}_n$ for ease of notation. First, observe that, under the null and invoking Lemma A.1 , the numerator rewrites as

$$\mathbf{X}_1^T\left(\mathcal{Q}g\left(\mathbf{Y} - \mathbf{X}_1\beta_1^0\right)\right) = \mathbf{X}_1^T\left(g\mathcal{Q}\left(\mathbf{Y} - \mathbf{X}_1\beta_1^0\right)\right) = \mathbf{X}_1^T\mathcal{Q}\left(g\boldsymbol{\varepsilon}\right), \tag{7}$$

which is a function of $(\mathbf{X}_1, \mathbf{X}_2, g\boldsymbol{\varepsilon})$ alone.

Second, let $\hat{\mathcal{Q}}$ be the orthogonal projection onto $\mathcal{C}^{\perp}\left(\mathbf{1}_n, \{g\mathbf{X}_1\}_{g\in\mathbf{G}}, \{g\mathbf{X}_2\}_{g\in\mathbf{G}}\right)$, the orthogonal complement of the span of $\mathbf{1}_n$, $\{g\mathbf{X}_1\}_{g\in\mathbf{G}}$ and $\{g\mathbf{X}_2\}_{g\in\mathbf{G}}$. Consider the permuted fitted residuals $\hat{\boldsymbol{\varepsilon}}(g) \equiv \hat{\boldsymbol{\varepsilon}}\left(\mathbf{X}_1, \mathbf{X}_2, g\left(\mathbf{Y} - \mathbf{X}_1\beta_1^0\right)\right)$, the residuals from the regression of $g\left(\mathbf{Y} - \mathbf{X}_1\beta_1^0\right)$ onto $\mathbf{1}_n$, $\{g\mathbf{X}_1\}_{g\in\mathbf{G}}$ and $\{g\mathbf{X}_2\}_{g\in\mathbf{G}}$. Observe that $\hat{\boldsymbol{\varepsilon}}(g)$ rewrites as

$$\hat{\mathcal{Q}}g\left(\mathbf{Y} - \mathbf{X}_1\beta_1^0\right) = g\hat{\mathcal{Q}}\left(\mathbf{Y} - \mathbf{X}_1\beta_1^0\right) = g\hat{\mathcal{Q}}\boldsymbol{\varepsilon} = \hat{\mathcal{Q}}g\boldsymbol{\varepsilon}, \tag{8}$$

where the first and third equality invoke Lemma A.1. In particular, $\hat{\mathcal{Q}}\boldsymbol{\varepsilon} = \hat{\boldsymbol{\varepsilon}}(\mathrm{id})$ and we deduce that $g\hat{\boldsymbol{\varepsilon}}(\mathrm{id}) \equiv (\hat{\varepsilon}_{g(1)}, \ldots, \hat{\varepsilon}_{g(n)})$ is equal to $\hat{\boldsymbol{\varepsilon}}(g)$ which is a function of $(\mathbf{X}_1, \mathbf{X}_2, g\boldsymbol{\varepsilon})$ alone.

Standard arguments then apply to establish (6). Specifically, consider the explicit parametrization of the test

$$\phi\left(\mathbf{W}, \{g\mathbf{W}'\}_{g\in\mathbf{G}}\right) = \begin{cases} 1, & \hat{t}\left(\mathbf{W}\right) > t^{(k)}\left(\mathbf{W}'\right) \\ a, & \hat{t}\left(\mathbf{W}\right) = t^{(k)}\left(\mathbf{W}'\right) \\ 0, & \hat{t}\left(\mathbf{W}\right) < t^{(k)}\left(\mathbf{W}'\right) \end{cases},$$

where $\hat{t}\left(\mathbf{W}\right) = t\left(\beta_1^0 : \mathbf{W}\right)$ and $t^{(k)}\left(\mathbf{W}'\right)$ is the k^{th} order statistic of $\{t(\beta_1^0 : g\mathbf{W}')\}_{g\in\mathbf{G}}$. In particular, $\phi\left(\mathbf{W}\right) = \phi\left(\mathbf{W}, \{g\mathbf{W}\}_{g\in\mathbf{G}}\right)$.

By construction,

$$\sum_{g\in\mathbf{G}}\phi\left(g\mathbf{W}\right) = \sum_{g\in\mathbf{G}}\phi\left(g\mathbf{W}, \{g'g\mathbf{W}\}_{g'\in\mathbf{G}}\right) = \sum_{g\in\mathbf{G}}\phi\left(g\mathbf{W}, \{g\mathbf{W}\}_{g\in\mathbf{G}}\right) = M\alpha,$$

where the second equality holds because $\mathbf{G}$ is a group. Therefore,

$$M\alpha = E\left[\sum_{g\in\mathbf{G}}\phi\left(g\mathbf{W}\right)\right] = \sum_{g\in\mathbf{G}}E\left[\phi\left(g\mathbf{W}\right)\right].$$

Because $\phi\left(g\mathbf{W}\right)$ is only a function of $g\mathbf{W}$ through $t\left(\beta_1^0 : g\mathbf{W}\right)$ and $\{t\left(\beta_1^0 : g'g\mathbf{W}\right)\}_{g'\in\mathbf{G}}$, each of which have the same distribution for every $g \in \mathbf{G}$, it obtains that

$$M\alpha = \sum_{g\in\mathbf{G}}E\left[\phi\left(\mathbf{W}\right)\right] = ME\left[\phi\left(\mathbf{W}\right)\right].$$

Q.E.D.

REMARK 1: We relied on the operator commutativity

$$g\mathcal{Q} = \mathcal{Q}g, \tag{9}$$

guaranteed by Lemma A.1, in order to establish (7) and (8). Alternatively, we may for instance have observed that, by construction, $\mathcal{Q}(g\mathbf{X}_2) = 0$, for any $g \in \mathbf{G}_n$. This is in line with the intuition behind the construction of the test statistic. Invariance of the test statistic with respect to β_2 comes from orthogonalizing away not just $\mathbf{X}_2$, but from orthogonalizing away $g\mathbf{X}_2$ for all $g \in \mathbf{G}_n$.

The assumption on the kernel of $\{g(\mathbf{X}_1, \mathbf{X}_2)\}_{g \in \mathbf{G}_n}$ insures the denominator is not zero and the test statistic is well defined. It can be verified in practice and will typically hold with probability one.

The choice of $\mathbf{G}_n$, of course, is critical. The group of linear transformations must satisfy three criteria.

First, it must be large—and thus granular—enough to produce accurate p-values and not forego "too much" statistical power.

Second, it must be small enough to leave explanatory variation in $\mathcal{Q}\mathbf{X}_1$ so that the test retains statistical power. For instance, if $\mathbf{G}_n$ were the set of all permutations from $\{1, \ldots, n\}$ onto itself, then many typical data generating processes for X_2—e.g., continuous—would tend to produce $\mathbf{X}_2 \in \mathbb{R}^n$ such that $\operatorname{rank}\left((g\mathbf{X}_2)_{g \in \mathbf{G}_n}\right) = n$, meaning $\mathcal{Q}\mathbf{X}_1 = 0_n$, and the test would have no power whatsoever.

Third, it must be structured enough that a central limit theorem for the randomization distribution can be worked out.

We suggest to pick $\mathbf{G}_n$ to be the set of *non-overlapping block permutations*. For instance, for $n = 9$ and number of blocks $n_{\text{blocks}} = 3$, some $g \in \mathbf{G}_n$ gives

$$g(1,2,3,4,5,6,7,8,9) = (4,5,6,7,8,9,1,2,3).$$

With respect to our first criteria, this choice seems satisfying. Indeed, when using this group in the univariate case, the loss of power from decreased block size is not traded off for an increase in power from retaining more explanatory variation in $\mathcal{Q}\mathbf{X}_1$, and is thus isolated. We find in simulations a very moderate loss in power from using blocks. See Appendix A.1 for details and an illustration.

With respect to our second criteria, theory avails that allows us to quantify the dimension of the space that is annihilated by applying the orthogonal projection $\mathcal{Q}$ corresponding to a certain $\mathbf{G}_n$. We collect such a theorem.

THEOREM 2: *If $n \geq n_{\text{blocks}}^2$, then*

$$\max_{\mathbf{X} \in \mathbb{R}^n} \operatorname{rank}(\mathbf{G}\mathbf{X}) = n_{\text{blocks}} \cdot (n_{\text{blocks}} - 2) + 2,$$

where $\mathbf{G}\mathbf{X} = \left(g^{(1)}\mathbf{X}, \ldots, g^{(M)}\mathbf{X}\right)$.

Remarkably, the block structure makes it such that the column rank of $\mathbf{G}\mathbf{X}$ increases much slower than the number of columns of $\mathbf{G}\mathbf{X}$. This is nicely illustrated in Table I.

The proof of Theorem A.1 for the general case of a matrix argument $\mathbf{X} \in \mathbb{R}^{n \times p}$ is given in Appendix A.2, and may be of independent interest as a nice interplay of linear algebra and combinatorics.

TABLE I

n_{blocks}	2	3	4	5	6	7	8	9	10
Number of columns	2	6	24	130	720	5,040	40,320	362,880	3,628,800
Rank	2	5	10	17	26	37	50	65	82

Note: Number of columns and maximal rank for a matrix whose columns are $g^{(1)}\mathbf{X}, \ldots, g^{(M)}\mathbf{X}$, allowing a conformable n.

Theorem 2 conveys in a principled way why the method can work in the first place. Indeed, while the number of columns of the matrix $\left(g^{(1)}\mathbf{X}, \ldots, g^{(M)}\mathbf{X}\right)$ increases exponentially fast in the number of blocks, its rank, and thus that of the kernel of $\mathcal{Q}$, increases like a second-order polynomial.

Note that Theorem 2 is not only presented as theory to justify and motivate our choice of group actions. It is important in the construction of the test statistic because it facilitates the construction of $\mathcal{Q}$. It is furthermore important for theory, and is indeed relied upon in the proof of Theorem A.3 below.

2.2. *Large-sample validity*

Exact validity of the proposed test was established in Theorem 1, under a strong null implied, for instance, by correct specification of the conditional expectation as being linear and the observations being independently and identically distributed. Exact validity also obtained for the tests proposed by Lei and Bickel (2021), D'Haultfœuille and Tuvaandorj (2024), Brown and Maritz (1982) and Wen et al. (2022). However, this can be appreciated as an answer to a mathematical statistics question, or as an "additional property" for a test that ought to be valid in a weaker sense under assumptions more typically encountered by data analysts.

Here, the analogy with the classical t-test is instructive. While the exact validity —and pivotality— of the t-test when data is normally distributed is both an edifying mathematical fact and somewhat of a reassurance when, in practice, the data analyst is confronted with a small sample believed to be approximately normally distributed, it is the much more general validity of the test in large samples that makes it such a ubiquitous tool.

Specifically, while the proposed test is exactly valid for the best linear predictor when the errors are exchangeable, assumptions made directly about the errors produced as the difference between observations $\mathbf{Y}$ and the best linear approximation to a nonlinear conditional expectation may be hard to assess—or to entertain seriously. The generic case allows dependence between X and ε.

As per the classical t-test, while the exact validity of the permutation test under exchangeability of the blocks of errors—which is already weaker than requiring their being independently and identically normally distributed—is an attractive property, asymptotic validity under weaker assumptions—comparable to those required for asymptotic validity of the t-test—is crucial for the practical relevance of the method.

Moving from analogy to comparison, it is interesting to consider that, power concerns notwithstanding, the proposed exact test is asymptotically robust to heteroskedasticity, while the classical t-test is not. On the other hand, the heteroskedasticity robust t-test asymptotically delivers its eponymous robustness but is not exactly valid in small samples, even with independent Gaussian errors, while the proposed test is both asymptotically heteroskedasticity-robust and exactly valid in small samples with exchangeable errors, Gaussian or not.

Thus motivated, we produce theory guaranteeing that our exactly valid procedure remains at least asymptotically valid when the strong null fails, for instance under conditional heteroskedasticity. Remark that, amongst the aforementioned exact tests for individual regression coefficients, only D'Haultfœuille and Tuvaandorj (2024) produced a provably asymptotically heteroskedasticity robust procedure, in the case of discrete control variables $\mathbf{X}_2$.

The key to producing the asymptotically robust randomized test is to studentize. This idea was first put forward by Neuhaus (1993) and was used in a similar context by D'Haultfœuille and Tuvaandorj (2024) and DiCiccio and Romano (2017); these also provide a more complete and historical set of references pertaining to studentizing. A randomization test is asymptotically valid if the randomization distribution of its test statistic coincides, in arbitrarily large samples, with its resampling distribution. The randomized test then inherits, in such large samples, the validity of the resampling procedure. In standard settings, the resampling and permutational limit distributions of a test statistic are normal, but their variances may disagree. However, if there exists a variance estimate that converges to the correct, possibly different asymptotic variances in each asymptotic regime, then by studentizing the test statistic with that variance estimate, both limits of the studentized test statistic will be standard normal and asymptotic validity will obtain.

Specifically, in our notation, the numerator of (4) converges in distribution to $N\left(0, E\left[\varepsilon^2 \cdot \bar{X}_1^{*2}\right]\right)$, where $\bar{X}_1^*$ is population analog of the entries of $\bar{\mathbf{X}}_1$. Under a permutational central limit theorem, the permutations "enforce" independence and the numerator of (4) converges in distribution to $N\left(0, E\left[\varepsilon^2\right] \cdot E\left[\bar{X}_1^{*2}\right]\right)$. However, the variance estimate $\hat{\sigma}_g^2$ converges to $E\left[\varepsilon^2 \cdot \bar{X}_1^{*2}\right]$ for $g = \text{id}$ under the standard frequentist central limit theorem, and converges to $E\left[\varepsilon^2\right] \cdot E\left[\bar{X}_1^{*2}\right]$ under the permutational central limit theorem. Consequently, the studentized statistic (4) converges to a standard normal under both regimes. The randomization procedures thus matches the resampling procedure and inherits its validity.

We collect key assumptions.

ASSUMPTIONS M—Moments:
M.1 $E\left[\varepsilon_i\right] = 0$ *and* $E\left[\varepsilon_i X_i\right] = 0$, $\forall\, i$,
M.2 $E\left[X_{i,j}^4\right], E\left[Y_i^4\right] < \infty$, $\forall\, i, j$.

Assumption M.1 defines the best linear predictor, while M.2 requires bounds on moments which may be considered as standard, see for instance DiCiccio and Romano (2017).

ASSUMPTIONS S—Spectral Conditions: *For* $\overrightarrow{\mathbf{X}} \in \mathbb{R}^{n \times \overrightarrow{p}}$ *a full rank matrix such that* $\overrightarrow{p} = \text{rank}\left(\mathbf{GX_{-1}}\right)$ *where* $\mathbf{GX_{-1}} = \left(g^{(1)}\mathbf{X}_2, g^{(2)}\mathbf{X}_2, \ldots, g^{(M)}\mathbf{X}_p\right)$, *and* $\check{\mathbf{X}} \in \mathbb{R}^{n \times \check{p}}$ *a full rank matrix such that* $\check{p} = \text{rank}\left(\mathbf{GX}\right)$ *where* $\mathbf{GX} = \left(g^{(1)}\mathbf{X}_1, ..., g^{(M)}\mathbf{X}_1, g^{(1)}\mathbf{X}_2, \ldots, g^{(M)}\mathbf{X}_p\right)$, *the maximal eigenvalues of the inverse Gram matrices satisfy* $\lambda_{\max}\left(\left(\frac{\overrightarrow{\mathbf{X}}^T\overrightarrow{\mathbf{X}}}{n}\right)^{-1}\right) = O_P\left(\overrightarrow{p}\right)$ *and* $\lambda_{\max}\left(\left(\frac{\check{\mathbf{X}}^T\check{\mathbf{X}}}{n}\right)^{-1}\right) = O_P\left(\overrightarrow{p}\right)$.

Assumption S is analogous to assumptions made in the "large number of moments" literature, see for instance Koenker and Machado (1999). They are required to insure that the residuals have standard behavior even though the number of regressors may be growing. See further discussion below.

ASSUMPTIONS D—Design: $X_{ij}, \forall\, i, j$, *are uniformly bounded in probability.*

Assumptions D restricts the support of the design matrix and, heuristically speaking, requires that arbitrarily large values of the covariates occur with arbitrarily small probability.

We can now produce the main theorem which states that even if the strong null does not hold, if the *weak null* consisting of the null hypothesis of interest (3) and assumptions including M, S and D hold, then the test ϕ is asymptotically valid.

THEOREM 3: *Suppose* $(Y_i, X_i) \sim \mathcal{P}$ *independently, and that Assumptions M, S and D hold. Suppose that S holds and* $n_{\text{blocks}}^8/n \to 0$*. Further suppose that* $\varepsilon_i^2 = o_P(n)$*, uniformly, and that* $n_{\text{blocks}}/b^\zeta \to \infty$*, for some* $0 < \zeta < 1$*. Then, under the null hypothesis* H_0*, we have that*

$$E_{\mathcal{P}}[\phi(\mathbf{Y}, \mathbf{X}_1, \mathbf{X}_2)] \to \alpha,$$

where $(\mathbf{Y}, \mathbf{X}_1, \mathbf{X}_2) \in \mathbb{R}^{n \times (2+p)}$.

PROOF OF THEOREM 3: Let $J_{t_{\text{id}},n}(\cdot)$ be the sampling distribution of $t_{\text{id}}(\beta_1^0)$, whose dependence on n is implicit. By Theorem A.2, we have that

$$\lim_{n \to \infty} \sup_{s \in \mathbb{R}} |J_{t_{\text{id}},n}(s) - \Phi(s)| = 0,$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Let $\hat{R}_{t_g,n}$ be the permutation distribution of $t_g(\beta_1^0)$. By Theorem A.4, we have that

$$\hat{R}_{t_g,n}(s) \xrightarrow{P} \Phi(s),$$

for all points of continuity of s.

Invoking Lemma 11.2.1 of Lehmann and Romano (2005), as they do in their Theorem 15.2.3, we find that since $\Phi(\cdot)$ is everywhere strictly increasing and continuous,

$$\hat{R}_{t_g,n}^{-1}(1-\alpha) \xrightarrow{P} \Phi^{-1}(1-\alpha).$$

Because t_{id} is asymptotically normal, it follows by Slutsky's Theorem that

$$\mathrm{E}_{\mathcal{P}}[\phi(\mathbf{Y}, \mathbf{X}_1, \mathbf{X}_2)] = \mathcal{P}\left(t_{\text{id}}(\beta_1^0) > \hat{R}_{t_g,n}^{-1}(1-\alpha)\right) + o(1) \to \mathcal{P}(Z > \Phi^{-1}(1-\alpha)) = \alpha,$$

where Z denotes a standard normal random variable.

In particular, by comparing the realized test statistic to the quantiles of the randomization distribution, the randomization test asymptotically accepts or rejects according to the standard resampling test.

Q.E.D.

The requirement that the blocks don't grow too fast in the sense that $n_{\text{blocks}}/b^\zeta \to \infty$ is a rather innocuous technicality since ζ can be taken as small as desired. Likewise, we find that requiring $\varepsilon_i^2 = o_P(n)$ uniformly is not restrictive, it is for instance satisfied and easy to verify in the case of normal random variables.

Assumption S is certainly stronger than required. Simulation studies and intuition suggest the rate $O_P(1)$ for the maximal eigenvalues. However, to the best of my knowledge, we do not have the random matrix theory required in order to derive this rate in terms of simple conditions on $\mathbf{X}$. I therefore opted for a conservative but noncontroversial low-level condition. One implication of the conservative rate on the maximal eigenvalue is a stronger requirement on the growth rate of the number of blocks.

Indeed, the requirement that $n_{\text{blocks}}^8/n \to 0$ is very strong, but must be qualified. First, the requirement that the blocks be very small and "grow slowly" does not hinder power, as discussed

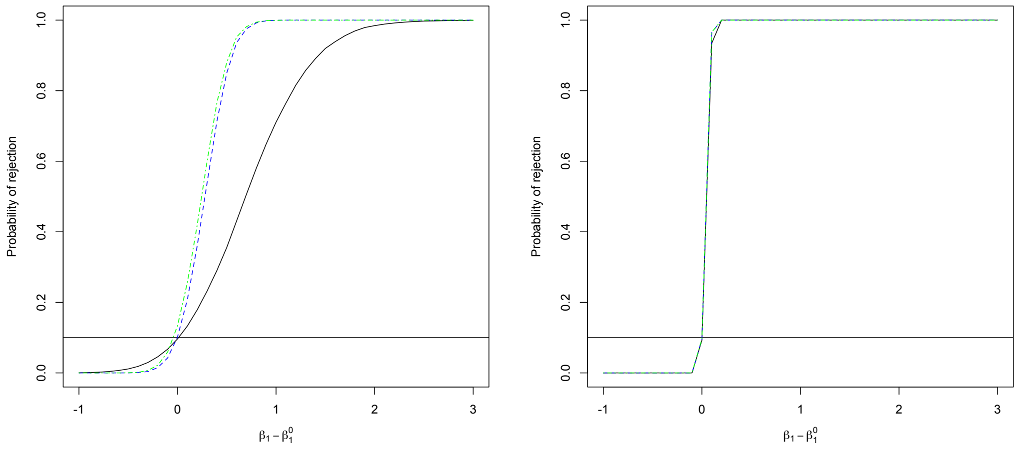

FIGURE 1.—Comparison of classical t-tests with proposed exact t-test when errors are normally distributed. The proposed exact t-test is in black bold, the classical t-test in gray, and the classical robust t-test in dashed grey. Left-hand side plot is for $n = 25$ and $n_{\text{blocks}} = 5$ and the right-hand side plot is for $n = 1000$ and $n_{\text{blocks}} = 10$.

above and in Appendix A. In a sense, the practically constraining regularity condition is that they grow at all. Second, simulations suggest that the true required rate is much lower. Theory is also suggestive. Even under Assumption S, if one assumes that the errors have conditional mean zero –even allowing for heteroskedasticity– then standard arguments such as those used in Donald and Newey (1994) apply and require $n_{\text{blocks}}^4/n \to 0$.

3. SIMULATIONS AND APPLICATIONS

Because the key property of the proposed test is its validity, both exactly in small sample under the strong null and asymptotically under the weak null, the objective of this section is not so much to provide a thorough characterization of power properties across a vast and exotic array of data generating processes but rather to qualitatively suggest that the proposed test has reasonable power even though validity has been extended. In order to do that with simple yet compelling illustrations, we focus on *ex ante* expected data generating processes to impress upon the reader that the simulation designs were not cherry picked. Specifically, we consider standard normal errors for the "Gaussian" design, Gamma variables with shape and rate parameter equal to 0.01 and 1 and rescaled to have variance 1 for the "skewed" design, the "exponential" design corresponds to error distributed according to an exponential distribution with parameter 1, and normal errors with variance $\sigma_\epsilon^2(X) \propto \sqrt{|X_1|}$ for the "heteroskedastic" design.

Unless otherwise specified, the covariates (X_{i1}, X_{i2}) are drawn from a bivariate normal with marginal variances 1 and covariance 0.15.

We can see from Figure 1 that, as expected, the classical t-tests apparently dominates the proposed test in small samples with Gaussian data. However, the performance of the proposed test appears to be competitive asymptotically. This is quantified in Table II, which documents the seeming convergence of the power of the classical t-test and proposed test with Gaussian errors. The same convergence is observed for the heteroskedasticity robust t-test and the proposed test.

TABLE II

RELATIVE POWER OF PROPOSED TEST ALONG A PITMAN SEQUENCE.

n	25	50	100	1000
n_{blocks}	5	5	5	10
rel. to t-test, Gaussian errors	0.37	0.82	0.99	0.99
rel. to robust t-test, heteroskedastic errors	0.88	0.93	0.91	0.97

Note: The rejection probability of the proposed exact test divided by the rejection probability of the standard t-test. The tested null is $\beta_1^0 = \beta_1 + \sqrt{25/n}$. Errors are standard normal. Note that the power of the heteroskedasticity robust procedures (second row) may not be meaningful for $n \leq 100$ because these procedures are only asymptotically valid.

TABLE III

ASSESSING VALIDITY

n	n_{blocks}	Error distribution	t-test	robust t-test	exact robust t-test
25	5	Gaussian	0.10	0.16	0.10
		skewed	0.14	0.05	0.10
		heteroskedastic	0.14	0.03	0.09
250	10	Gaussian	0.10	0.11	0.10
		skewed	0.10	0.11	0.10
		heteroskedastic	0.17	0.11	0.10

Note: $n = 250$ and $n_{\text{blocks}} = 10$.

TABLE IV

ASSESSING POWER

$\beta_1^0 - \beta_1$	Error distribution	t-test	robust t-test	exact robust t-test
0.1	Gaussian	0.62	0.62	0.45
	skewed	0.76	0.81	0.70
	heteroskedastic	*	0.51	0.41
0.2	Gaussian	0.97	0.97	0.84
	skewed	0.94	0.94	0.87
	heteroskedastic	*	0.94	0.79

Note: $n = 250$ and $n_{\text{blocks}} = 10$.

Tables II, III and IV may be considered together. Table IV suggests that while $n = 25$ is too small a sample, in our specific but realistic simulation design, for the hetereoskedasticity-robust methods to be valid, $n = 250$ suffices. It also reminds us that the classical t-test is not valid, even asymptotically, under heteroskedasticity. Table IV suggests that, while both the hetereoskedasticity-robust t-test and the proposed procedure are valid for $n = 250$, the robust t-test seems to have better power. In yet larger samples, as detailed in Table II, specifically $n = 1000$, they are however observed to have comparable power.

This paints a nice picture for the proposed test. The classical t-test is exact under Gaussian errors but invalid, even asymptotically, under heteroskedasticity. The heteroskedasticity

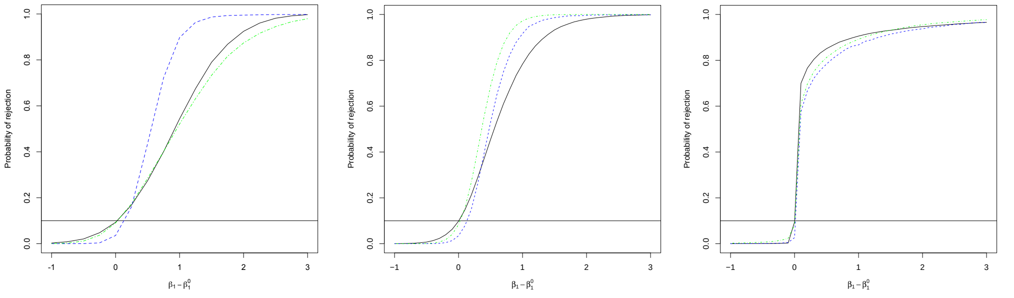

FIGURE 2.—Comparison of exact t-test with alternative methods in small sample. The bold black line corresponds to the proposed test, the blue dashed line corresponds to RPT, and the green dot dashed line corresponds to CPT. The left-hand side plot corresponds to Gaussian errors, the middle plot corresponds to exponential errors, and the right-hand side corresponds to the "skewed" design. Throughout, $n = 25$ and $n_{\text{blocks}} = 5$.

TABLE V

ASSESSING POWER AND VALIDITY

$\beta_1^0 - \beta_1$	Error distribution	exact robust t-test	CPT	RPT
0	Gaussian	0.10	0.10	0.04
	Exp	0.10	0.08	0.03
	Gamma	0.10	0.08	0.03
	heteroskedastic	0.09	0.19	0.02
1	Gaussian	0.54	0.51	0.88
	Exp	0.80	0.97	0.92
	Gamma	0.91	0.89	0.87

Note: $n = 25$ and $n_{\text{blocks}} = 5$.

robust t-test is invalid in small samples, even with Gaussian errors. The proposed test is valid both in small samples with exchangeable data—e.g., iid Gaussian— and asymptotically under heteroskedasticity. Yet, it displays competitive power properties.

In Figure 2 and in the last three columns of Table V, we see that neither the proposed procedure, the circular permutation test (CPT) of Lei and Bickel (2021), nor the residual permutation test (RPT) of Wen et al. (2022) dominates or is dominated by the others.

In the first three columns of Table V, we see that all exact tests are indeed valid in small samples with exchangeable errors. None of the tests are exactly valid with heteroskedastic errors. In Table V, only CPT is observed to over-reject, but all three tests can be made to over-reject under heteroskedasticity. Simulations not presented here confirm that the CPT and RPT procedures –as well as the unstudentized exact t-test– remain invalid even when n is large.

Remark that the CPT procedure has exact nominal size, the mild undercoverage is attributable to the small number of permutations and ensuing coarseness of the randomization distribution in this very small sample.

As illustrated in the left-hand side of Figure 3, the test loses power when the correlation between the tested and nuisance covariate increases. This is to be expected, since we explicitly orthogonalize nuisance covariates when producing $\bar{\mathbf{X}}_1$, whose explanatory variation—specifically, covariation with $\mathbf{X}_1$— gives power to the test. This can be compared with

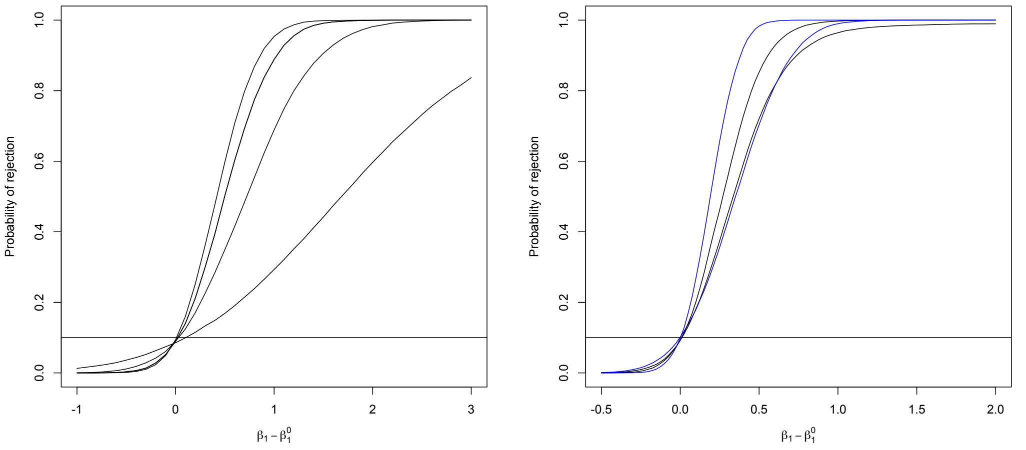

FIGURE 3.—Power of robust exact t-test for different correlations between covariates. The left figure plots the power curve for $\rho = Cov\,(X_1, X_2)$ at values $\rho = 0, 0.25, 0.5, 0.75$. Errors are Gaussian, $n = 25$ and $n_{\text{blocks}} = 5$. The right figure plots the power curve in the Gaussian design for p=2 and p=35 covariates, $n = 50$ and $n_{\text{blocks}} = 5$. The proposed test is in black and the standard t-test is in blue, power is lesser when p is greater.

TABLE VI

REANALYZING EXAMPLE 4.3 OF WOOLDRIDGE (2020).

	$\hat{\beta}$	randomized	p-val	t-test	p-val	robust t	p-val
HS	0.412	(0.057, 0.596)	0.065	(0.257, 0.566)	0.000	(0.247, 0.577)	0.000
ACT	0.015	(0.005, 0.04)	0.048	(-0.003, 0.032)	0.164	(-0.004, 0.033)	0.196
skip	-0.083	(-0.123, -0.005)	0.067	(-0.126,-0.04)	0.002	(-0.128, -0.039)	0.002

Note: Critical level for confidence intervals is $\alpha = 0.1$, $n = 140$ and $n_{\text{blocks}} = 5$.

the standard t-test which likewise forgoes power when the tested covariate is correlated with nuisance covariates. We do observe that the loss of power is moderate for reasonable correlations, even up to $\rho = 0.5$, but is quite large for $\rho = 0.75$.

The right-hand side of Figure 3 illustrates how including more nuisance regressors brings about a decrease in power of the proposed test commensurate to that observed with the classical t-test.

Simulations and power plots are convenient for the analysis of power, but must be complemented with examples. We consider two examples from the econometrics textbook Wooldridge (2020).

We may be interested in determinants of college GPA, and to that effect carry out regression analysis of college GPA on a measure of high school GPA, a measure of ACT scores, and a measure the extent to which the student skipped classes. These estimates and their confidence intervals are presented in Table VI.

We see that the confidence interval on the effect of high school GPA produced by the classical t-tests are about half the width of that produced by our method, and they decidedly reject the null hypothesis that the coefficient is 0, while we do not reject at confidence level $\alpha = 0.1$. For the coefficient on ACT scores however, the intervals are of comparable length, and only

TABLE VII

REANALYZING EXAMPLE 15.5 OF WOOLDRIDGE (2020).

$\hat{\beta}_{\mathrm{educ}}$	randomized	p-val	classical	p-val	HCO	p-val
0.063	(0.011, 0.141)	0.09	(0.008, 0.117)	0.06	(0.004, 0.121)	0.08

Note: Critical level for confidence intervals is $\alpha = 0.1$, $n = 424$ and $n_{\text{blocks}} = 8$.

the proposed method rejects the null hypothesis that the coefficient is 0. For the coefficient on skipping classes, the confidence intervals produced by all compared methods are very similar.

Our method naturally extends to regression analysis with instrumental variables. Wooldridge (2020) reanalyzes Mroz (1987). He analyzes how $Y = \log(\text{wage})$ is affected by $X_1 = \text{education}$, instrumenting with $(Z_1, Z_2) = (\text{mother education}, \text{father education})$, controlling for $X_2 = \text{experience}$ and its square. Because there are two instruments, the natural test statistic is

$$t_g\left(\beta_1^0\right) = T_g\left(\beta_1^0\right)^T \hat{\Sigma}_g^{-1} T_g\left(\beta_1^0\right)$$

where $T_g\left(\beta_1^0\right) = (g\mathbf{Z})^T \mathcal{Q}\left(\mathbf{Y} - \mathbf{X}_1\beta_1^0\right)$ and $\hat{\Sigma}_g = \frac{1}{n}\sum_{i=1}^n \bar{\mathbf{Z}}_{g(i)}\bar{\mathbf{Z}}_{g(i)}^T\hat{\varepsilon}_i^2$.

Remark the distinction with the instrumental variables analysis of Imbens and Rosenbaum (2005). They produce a Fisher test using the instrument assignment as a source of randomization and thus provide inference that is valid for the fixed, observed sample. In contrast, we rely on invariance under the null to produce inference that is valid unconditionally on the observed sample.

The confidence intervals produced by the proposed method are presented in Table VII. They are tangibly but not dramatically wider than those produced by the standard inference methods for two-stage least-squares. This may be found, on a case-by-case basis, to be a reasonable loss in power to trade off against greater validity. Treatment effect heterogeneity, which brings about heteroskedasticity, may be such a motivation. Further note that the method is robust to weak identification of the tested coefficient. Of course, the method is also amenable to the projection method if robustness to weak identification of some nuisance coefficients is required.

4. CONCLUSION

The proposed test is presented as an alternative to the classical t-test. While the classical t-test is exactly valid when errors are independent and identically distributed Gaussian but invalid under conditional heteroskedasticity, even asymptotically, the heteroskedasticity robust t-test is not exactly valid in small samples, even with Gaussian errors. Meanwhile, the proposed test is both exactly valid with exchangeable errors, which is a weaker requirement than independent Gaussian errors, and is asymptotically valid under misspecification of linearity and conditional heteroskedasticity. While we may not expect the proposed test to have better power than the t-test when it is valid, the good power properties exhibited above suggest that it may nevertheless be an attractive alternative if the data analyst is concerned about validity.

The power of the proposed procedure is comparable to that of the general exact methods of Lei and Bickel (2021) and Wen et al. (2022), but these are not asymptotically valid under conditional heteroskedasticity. In the special case of all discrete nuisance covariates, we recommend using the tailor made method of D'Haultfœuille and Tuvaandorj (2024).

We picked the group $\mathbf{G}_n$ so to satisfy the three criteria discussed in Section 2. One could push the inquiry and try to select a different group $\mathbf{G}_n$ out of power considerations, or to adapt to specificities of the problem at hand and its design matrix.

An important topic of future research is to find, if it exists, a test for individual linear quantile regression coefficients that is exactly valid under a strong null requiring not much more than correct specification of the regression function and that is asymptotically valid under misspecification. This is particularly important since the regression rankscore test (Gutenbrunner and Jurecková, 1992, Koenker, 1994, 2005) was originally motivated as a regression extension of classical rank tests.

An immediate benefit of the presented and cited results is that they may allow conclusions drawn in –conditional– Fisher type tests, such as in Berry and Fowler (2018) or Imbens and Rosenbaum (2005), to be extended to population parameters by interpreting them as –unconditional– randomization tests.

More generally, the connections between Fisher tests and randomization tests constitute, in my opinion, a fascinating territory for research. For instance, the same way "robust" randomization tests are asymptotically valid under a "weak null" guaranteeing robustness to failure of the "strong null", or randomization hypothesis, exact randomization tests may themselves provide a more robust interpretation for Fisher tests. For instance, if an experimenter ran a Fisher test but "forgot" the treatment assignment mechanism, and one of the possible mechanisms constitutes uniform assignment over a group support, then implementing the Fisher test according to that mechanism will guarantee unconditional validity even if the treatment assignment is incorrect, and conditional validity therefore does not obtain. I believe many more such connections exist.

APPENDIX A: IMPACT ON POWER OF USING BLOCK PERMUTATIONS IN THE UNIVARIATE REGRESSION MODEL

The univariate case is

$$\mathbf{Y} = \beta_0 \mathbf{1}_n + \mathbf{X}\beta_1 + \boldsymbol{\varepsilon},$$

where $\mathbf{X} \in \mathbb{R}^n$ is the univariate regressor.

The test statistic is

$$t_g\left(\beta_1^0\right) = \frac{1}{\sqrt{n}} \left(g\mathbf{X}\right)^T \left(\mathbf{Y} - \mathbf{X}\beta_1^0\right).$$

The key is that this framework allows us to investigate the loss of power, or lack thereof, from using a smaller group in a way that is “isolated”. Indeed, in a general set-up, smaller groups also increase power through a different channel, i.e., they conserve more explanatory variation in $\mathcal{Q}\mathbf{X}_1$. But in the univariate setting, there are no covariates to orthogonalize. Consequently, potential loss of power from using fewer blocks in not counterbalanced by a gain of power due to weaker orthogonalization of nuisance covariates, and can be studied on its own.

APPENDIX B: MATHEMATICAL APPENDIX

B.1. *Mathematical Appendix*

LEMMA A.1: *Let $\mathbf{G}_n$ be a group of non-overlapping block permutation from $\{1, \ldots, n\}$ onto itself. Let $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_p) \in \mathbb{R}^{n \times p}$ be a given matrix. Let $\mathcal{Q}$ be the orthogonal projection onto $\mathcal{C}^{\perp}\left(g^{(1)}\mathbf{Z}_1, \ldots, g^{(M)}\mathbf{Z}_1, g^{(1)}\mathbf{Z}_2, \ldots, g^{(M)}\mathbf{Z}_p\right)$. Then, for any $\mathbf{X} \in \mathbb{R}^n$,*

$$g\mathcal{Q}\mathbf{X} = \mathcal{Q}g\mathbf{X},\ \forall\, g \in \mathbf{G}_n. \tag{10}$$

PROOF OF LEMMA A.1: Denote, for short, the column space $\mathcal{C}\left(\mathbf{GZ}\right) = \mathcal{C}\left(g^{(1)}\mathbf{Z}_1, \ldots, g^{(M)}\mathbf{Z}_1, g^{(1)}\mathbf{Z}_2, \ldots, g^{(M)}\mathbf{Z}_p\right)$ and its orthogonal complement $\mathcal{C}^{\perp}\left(\mathbf{GZ}\right)$.

TABLE A.I

PROBABILITY OF REJECTION OF THE NULL AT NOMINAL LEVEL $\alpha = 0.1$

Block size	1	10	20	50	1	10	20	50	1	10	20	50
$n = 100$	0.1	0.1	0.1	0	0.23	0.23	0.20	0	1	1	1	0
$n = 200$	0.1	0.1	0.1	0.1	0.38	0.38	0.36	0.29	1	1	1	1
$n = 400$	0.1	0.1	0.1	0.1	0.69	0.67	0.66	0.65	1	1	1	1
$n = 800$	0.1	0.1	0.1	0.1	0.88	0.88	0.87	0.88	1	1	1	1

Note: Columns are block size. Errors of the posited nulls are $\beta_1 - \beta_1^0 = 0$ (left), $\beta_1 - \beta_1^0 = 0.1$ (center), $\beta_1 - \beta_1^0 = 1$ (right).

Observe that $\mathcal{Q} = I - \mathcal{M}$, where $\mathcal{M}$ is the orthogonal projection onto $\mathcal{C}(\mathbf{GZ})$. Because (10) holds if and only if $g\mathcal{M}\mathbf{X} = \mathcal{M}g\mathbf{X}$, it suffices to show the latter.

For any $\mathbf{X} \in \mathbb{R}^n$, we may produce the decomposition

$$\mathbf{X} = u + v,\ u \in \mathcal{C}(\mathbf{GZ}),\ v \in \mathcal{C}^{\perp}(\mathbf{GZ}).$$

We collect two observations. First, observe that $u \in \mathcal{C}(\mathbf{GZ})$ implies that $u = \sum_{j=1}^{p}\sum_{i=1}^{M} a_{i,j} g^{(i)} Z_j$, for some real numbers $a_{i,j}$, $j = 1, \dots, p$, $i = 1, \dots, M$. Hence, for any $g \in \mathbf{G}_n$,

$$gu = g\sum_{j=1}^{p}\sum_{i=1}^{M} a_{i,j} g^{(i)} Z_j = \sum_{j=1}^{p}\sum_{i=1}^{M} a_{i,j} \underbrace{gg^{(i)}}_{\in \mathbf{G}_n} Z_j \in \mathcal{C}(\mathbf{GZ}).$$

Second, observe that $v \in \mathcal{C}^{\perp}(\mathbf{GZ})$ can (for a general v) equivalently be stated as $v \perp g^{(1)}\mathbf{Z}_1, \dots, g^{(M)}\mathbf{Z}_1, g^{(1)}\mathbf{Z}_2, \dots, g^{(M)}\mathbf{Z}_p$. Hence,

$$(gv)^T \left(g^{(i)}\mathbf{Z}_j\right) = v^T \left(\underbrace{g^{-1}g^{(i)}}_{\in \mathbf{G}_n}\mathbf{Z}_j\right) = 0,$$

for any i and j. Which is to say, $gv \in \mathcal{C}^{\perp}(\mathbf{GZ})$.

This means that

$$g\mathcal{M}\mathbf{X} = g\mathcal{M}(u + v) = g(\mathcal{M}u + \mathcal{M}v) = g\mathcal{M}u = gu,$$

and

$$\mathcal{M}g\mathbf{X} = \mathcal{M}g(u + v) = \mathcal{M}gu + \mathcal{M}gv \overset{(a)}{=} \mathcal{M}gu \overset{(b)}{=} gu,$$

where (2) holds because of the second observation, and (b) holds because of the first observation. *Q.E.D.*

THEOREM A.1—General Case: *If $n \geq p \cdot n_{\text{blocks}}^2$, then*

$$\max_{\mathbf{X} \in \mathbb{R}^{n \times p}} \operatorname{rank}(\mathbf{GX}) = p \cdot (n_{\text{blocks}} \cdot (n_{\text{blocks}} - 2) + 2),$$

where $\mathbf{GX} = \left(g^{(1)}\mathbf{X}, \dots, g^{(M)}\mathbf{X}\right)$.

We present the proof of Theorem A.1 in two parts. As a Lemma, and to simplify the exposition of the general case, we first prove the case with $p = 1$.

PROOF OF THEOREM A.1, THE $p=1$ CASE: We can provide a proof by induction. For any admissible n and n_{blocks}, designate the blocks of the n-tuple $\mathbf{X}$ by $A, B, \dots \in \mathbb{R}^b$, where $A = (X_1, \dots, X_b)^T$, $B = (X_{b+1}, \dots, X_{2b})$, and so on.

In the base case with $n_{\text{blocks}} = 2$, we have

$$\mathbf{G}_{b\cdot 2}\mathbf{X} = \left(\begin{pmatrix} A \\ B \end{pmatrix}, \begin{pmatrix} B \\ A \end{pmatrix} \right),$$

which obviously has $\max_{\mathbf{X}\in\mathbb{R}^n} \operatorname{rank}(\mathbf{G}_{b\cdot 2}\mathbf{X}) = 2$.

Now consider any given number of blocks n_{blocks}. Consider splitting all possible block permutations into sets according to their first block. Specifically, consider

$$\underbrace{\begin{pmatrix} A \\ B \\ \vdots \\ Y \\ Z \end{pmatrix}, \begin{pmatrix} A \\ B \\ \vdots \\ Z \\ Y \end{pmatrix}, \dots, \begin{pmatrix} A \\ Z \\ \vdots \\ \vdots \\ B \end{pmatrix}}_{\mathcal{S}_1}, \underbrace{\begin{pmatrix} B \\ A \\ \vdots \\ Y \\ Z \end{pmatrix}, \begin{pmatrix} B \\ A \\ \vdots \\ Z \\ Y \end{pmatrix}, \dots, \begin{pmatrix} B \\ Z \\ \vdots \\ \vdots \\ A \end{pmatrix}}_{\mathcal{S}_2}, \dots, \underbrace{\begin{pmatrix} Z \\ A \\ \vdots \\ W \\ Y \end{pmatrix}, \begin{pmatrix} Z \\ A \\ \vdots \\ Y \\ W \end{pmatrix}, \dots, \begin{pmatrix} Z \\ Y \\ \vdots \\ \vdots \\ A \end{pmatrix}}_{\mathcal{S}_{n_{\text{blocks}}}},$$

where the set $\mathcal{S}_1$ has block A in the first block of rows, and lists all permutations of the other blocks $B, \dots, Z$ over all its vector. The set $\mathcal{S}_2$ has block B in its first block of rows and lists all permutations of the other blocks $A, C, \dots, Z$ over all its vectors, and so on. To be sure, Z is the $n_{\text{blocks}}^{\text{th}}$ block, not necessarily the 26^{th} block.

For exposition purposes, consider the following reordering of the columns of $\mathbf{GX}$,

$$\underbrace{\begin{pmatrix} A \\ B \\ \vdots \\ Y \\ Z \end{pmatrix}, \begin{pmatrix} A \\ B \\ \vdots \\ Z \\ Y \end{pmatrix}, \dots, \begin{pmatrix} A \\ Z \\ \vdots \\ \vdots \\ B \end{pmatrix}}_{\mathcal{S}_1}, \underbrace{\begin{pmatrix} B \\ A \\ \vdots \\ Y \\ Z \end{pmatrix}, \begin{pmatrix} B \\ C \\ A \\ \vdots \\ \vdots \end{pmatrix}, \dots, \begin{pmatrix} B \\ C \\ \vdots \\ \vdots \\ A \end{pmatrix}, \dots, \begin{pmatrix} B \\ Y \\ \vdots \\ \vdots \\ C \end{pmatrix}}_{\mathcal{S}_2}, \dots, \underbrace{\begin{pmatrix} Z \\ A \\ \vdots \\ W \\ Y \end{pmatrix}, \begin{pmatrix} Z \\ A \\ \vdots \\ Y \\ W \end{pmatrix}, \dots, \begin{pmatrix} Z \\ Y \\ \vdots \\ \vdots \\ A \end{pmatrix}}_{\mathcal{S}_{n_{\text{blocks}}}},$$

where we reordered the columns of the set $\mathcal{S}_2$ to emphasize that we can have the first $n_{\text{blocks}} - 1$ vectors of $\mathcal{S}_2$ figuring block A in the second, third,..., and $n_{\text{blocks}}^{\text{th}}$ block of rows, respectively.

By inductive hypothesis, the rank of $\mathcal{S}_1$ attains $(n_{\text{blocks}} - 1) \cdot (n_{\text{blocks}} - 3) + 2$. Now the key is that, since $n \geq n_{\text{blocks}}^2$, which is to say, $b \geq n_{\text{blocks}}$, we can always pick an X such the block A cannot be produced as a linear combination of the $n_{\text{blocks}} - 1$ blocks $B, C \dots, Z$. Appending one-by-one to $\mathcal{S}_1$ vectors for which A shows up in a block of in a block of rows for the first time, we must each time increase the rank by 1. Hence, appending S_2 -and thus its first $n_{\text{blocks}} - 1$ vectors in the reordering above—to $\mathcal{S}_1$, we must increase the rank by at least $n_{\text{blocks}} - 1$. By the same logic, considering the first block of rows, appending one-by-one to $(\mathcal{S}_1, \mathcal{S}_2)$ the sets $\mathcal{S}_3, \dots, \mathcal{S}_{n_{\text{blocks}}}$ must increase the rank by at least 1 each time. Therefore,

$$\begin{aligned} \max_{\mathbf{X}\in\mathbb{R}^n} \operatorname{rank}(\mathbf{GX}) &\geq (n_{\text{blocks}} - 1) \cdot (n_{\text{blocks}} - 3) + 2 + n_{\text{blocks}} - 1 + n_{\text{blocks}} - 2 \\ &= n_{\text{blocks}} \cdot (n_{\text{blocks}} - 2) + 2. \end{aligned}$$

To complete the argument, it suffices to show that any column other than the ones invoked above to lower bound the rank—in fact, any column—can be produced as a linear combination of the said used columns. Specifically, it suffices to show that

$$\underbrace{\begin{pmatrix} A \\ B \\ \vdots \\ Y \\ Z \end{pmatrix}, \begin{pmatrix} A \\ B \\ \vdots \\ Z \\ Y \end{pmatrix}, \dots, \begin{pmatrix} A \\ Z \\ \vdots \\ \vdots \\ B \end{pmatrix}}_{\mathcal{S}_1}, \underbrace{\begin{pmatrix} B \\ A \\ \vdots \\ Y \\ Z \end{pmatrix}, \begin{pmatrix} B \\ C \\ A \\ \vdots \\ \vdots \end{pmatrix}, \dots, \begin{pmatrix} B \\ C \\ \vdots \\ \vdots \\ A \end{pmatrix}}_{n_{\text{blocks}}-1 \text{ vectors from } \mathcal{S}_2}, \underbrace{\begin{pmatrix} C \\ A \\ B \\ \vdots \\ Z \end{pmatrix}}_{1 \text{ vector from } \mathcal{S}_3} \dots, \underbrace{\begin{pmatrix} Z \\ A \\ B \\ \vdots \\ Z \end{pmatrix}}_{1 \text{ vector from } \mathcal{S}_{n_{\text{blocks}}}},$$

span all columns of $\mathbf{GX}$. Note that the columns of $\mathcal{S}_1$ are not linearly independent—-we could instead extract the first $(n_{\text{blocks}} - 1) \cdot (n_{\text{blocks}} - 3)$ columns in this ordering, they are linearly independent—but we know the rank of the set.

As a first step, we show that all columns in the second set $\mathcal{S}_2$ are in the span of $\mathcal{S}_1$ and the $n_{\text{blocks}} - 1$ vectors selected from $\mathcal{S}_2$. Pick any vector from $\mathcal{S}_2$; it has block B in the first block of rows. Without loss of generality, say A is in the second block of entries. I.e., the vector has form

$$\begin{pmatrix} B \\ A \\ A_3 \\ \vdots \\ A_{n_{\text{blocks}}} \end{pmatrix}, \tag{11}$$

for some $A_3, \dots, A_{n_{\text{blocks}}}$. In the vectors selected from $\mathcal{S}_2$, we have

$$\begin{pmatrix} B \\ A \\ A_3' \\ \vdots \\ A_{n_{\text{blocks}}}' \end{pmatrix}, \tag{12}$$

for some $A_3', \dots, A_{n_{\text{blocks}}}'$. Remark that $\{A_3, \dots, A_{n_{\text{blocks}}}\} = \{A_3', \dots, A_{n_{\text{blocks}}}'\} = \{C, \dots, Z\}$. Further remark that, in $\mathcal{S}_1$, we have the vectors

$$\begin{pmatrix} A \\ B \\ A_3 \\ \vdots \\ A_{n_{\text{blocks}}} \end{pmatrix}, \begin{pmatrix} A \\ B \\ A_3' \\ \vdots \\ A_{n_{\text{blocks}}}' \end{pmatrix},$$

which span the difference

$$\begin{pmatrix} 0_b \\ 0_b \\ A_3 - A_3' \\ \vdots \\ A_{n_{\text{blocks}}} - A_{n_{\text{blocks}}}' \end{pmatrix}.$$

Adding this vector to (12) produces (11).

Now consider any column from the third to $n_{\mathrm{blocks}}^{\mathrm{th}}$ set. Without loss of generality, take it from the third set $\mathcal{S}_3$. The vector has the form

$$\begin{pmatrix} C \\ A_2 \\ A_3 \\ \vdots \\ A_{n_{\mathrm{blocks}}} \end{pmatrix} \tag{13}$$

for some $A_2, A_3, \ldots, A_{n_{\mathrm{blocks}}}$. In the candidate spanning set, we have the vector

$$\begin{pmatrix} C \\ A'_2 \\ A'_3 \\ \vdots \\ A'_{n_{\mathrm{blocks}}} \end{pmatrix} \tag{14}$$

for some $A'_2, A'_3, \ldots, A'_{n_{\mathrm{blocks}}}$. Remark that $\{A_2, A_3, \ldots, A_{n_{\mathrm{blocks}}}\} = \{A'_2, A'_3, \ldots, A_{n_{\mathrm{blocks}}}\} = \{B, D, \ldots, Z\}$. Without loss of generality, let $A_2 = A$ and $A_3 = B$. I.e., we are trying to produce

$$\begin{pmatrix} C \\ A \\ B \\ \vdots \\ A_{n_{\mathrm{blocks}}} \end{pmatrix} \tag{15}$$

form the candidate spanning set.

We need to consider cases. First, consider the case in which $A'_2 = A$. We can produce

$$\begin{pmatrix} 0 \\ 0 \\ B - A'_3 \\ \vdots \\ A_{n_{\mathrm{blocks}}} - A'_{n_{\mathrm{blocks}}} \end{pmatrix}$$

from set $\mathcal{S}_1$, add it to (14), and produce (15).

Second, consider the case in which $A'_3 = A$. From set $\mathcal{S}_1$, you can produce the vector

$$\begin{pmatrix} 0 \\ C - A'_2 \\ B - C \\ A_4 - A'_4 \\ \vdots \\ \vdots \\ A_{n_{\mathrm{blocks}}} - A'_{n_{\mathrm{blocks}}} \end{pmatrix}. \tag{16}$$

From set $\mathcal{S}_2$, you can produce the vector

$$\begin{pmatrix} 0 \\ A-C \\ C-A \\ \vdots \\ 0' \end{pmatrix}. \tag{17}$$

Adding (16) and (17) to (14), we produce (15).

Finally, consider the third case in which $A'_j = A$, $j \in \{4, \ldots, n_{\text{blocks}}\}$. Without loss of generality, suppose $A'_4 = A$. From set $\mathcal{S}_1$, you can produce the vector

$$\begin{pmatrix} 0 \\ C-A'_2 \\ B-A'_3 \\ A_4-C \\ A_5-A'_5 \\ \vdots \\ A_{n_{\text{blokcs}}}-A'_{n_{\text{blokcs}}} \end{pmatrix}. \tag{18}$$

From the set $\mathcal{S}_2$, you can produce the vector

$$\begin{pmatrix} 0 \\ A-C \\ 0 \\ C-A'_4 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ A-C \\ 0 \\ C-A \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{19}$$

Adding (18) and (19) to (14), we produce (15).

Q.E.D.

We now prove the general case for $p < n$.

PROOF OF THEOREM A.1, THE $p < n$ CASE: We can provide a proof by induction. For any admissible n and n_{blocks} For each $j = 1, \ldots, p$, designate the blocks of the n-tuple $\mathbf{X}_j$ by $A_j, B_j, \dots \in \mathbb{R}^b$, where $A = (X_{1,j}, \ldots, X_{b,j})^T$, $B = (X_{b+1,j}, \ldots, X_{2b,j})$, and so on.

In the base case with $n_{\text{blocks}} = 2$, we have

$$\mathbf{G}_{n \cdot 2}\mathbf{X} = \left(\begin{pmatrix} A_1 \\ B_1 \end{pmatrix}, \begin{pmatrix} B_1 \\ A_1 \end{pmatrix}, \begin{pmatrix} A_2 \\ B_2 \end{pmatrix}, \cdots, \begin{pmatrix} B_p \\ A_{1p} \end{pmatrix} \right),$$

which obviously has $\max_{\mathbf{X} \in \mathbb{R}^n} \text{rank}\left(\mathbf{G}_{n \cdot 2}\mathbf{X}\right) = 2p$.

Now consider any given number of blocks n_{blocks}. For each $j = 1, \dots, p$, consider splitting all possible block permutations of $\mathbf{X}_j$ into sets according to their first block. Specifically, consider

$$\underbrace{\begin{pmatrix} A_j \\ B_j \\ \vdots \\ Y_j \\ Z_j \end{pmatrix}, \begin{pmatrix} A_j \\ B_j \\ \vdots \\ Z_j \\ Y_j \end{pmatrix}, \dots, \begin{pmatrix} A_j \\ Z_j \\ \vdots \\ \vdots \\ B_j \end{pmatrix}}_{\mathcal{S}_{1,j}}, \underbrace{\begin{pmatrix} B_j \\ A_j \\ \vdots \\ Y_j \\ Z_j \end{pmatrix}, \begin{pmatrix} B_j \\ A_j \\ \vdots \\ Z_j \\ Y_j \end{pmatrix}, \dots, \begin{pmatrix} B_j \\ Z_j \\ \vdots \\ \vdots \\ A_j \end{pmatrix}}_{\mathcal{S}_{2,j}}, \dots, \underbrace{\begin{pmatrix} Z_j \\ A_j \\ \vdots \\ W_j \\ Y_j \end{pmatrix}, \begin{pmatrix} Z_j \\ A_j \\ \vdots \\ Y_j \\ W_j \end{pmatrix}, \dots, \begin{pmatrix} Z_j \\ Y_j \\ \vdots \\ \vdots \\ A_j \end{pmatrix}}_{\mathcal{S}_{n_{\text{blocks}},j}},$$

where the set $\mathcal{S}_{1,j}$ has block A_j in the first block of rows, and lists all permutations of the other blocks $B_j, \dots, Z_j$ over all its vector. The set $\mathcal{S}_{2,j}$ has block B_j in its first block of rows and lists all permutations of the other blocks $A_j, C_j, \dots, Z_j$ over all its vectors, and so on. To be sure, Z_j is the $n_{\text{blocks}}^{\text{th}}$ block, not necessarily the 26^{th} block.

For exposition purposes, consider the following reordering of the columns of $\mathbf{G}\mathbf{X}_j$, for each $j = 1, \dots, p$, as

$$\underbrace{\begin{pmatrix} A_j \\ B_j \\ \vdots \\ Y_j \\ Z_j \end{pmatrix}, \begin{pmatrix} A_j \\ B_j \\ \vdots \\ Z_j \\ Y_j \end{pmatrix}, \dots, \begin{pmatrix} A_j \\ Z_j \\ \vdots \\ \vdots \\ B_j \end{pmatrix}}_{\mathcal{S}_{1,j}}, \underbrace{\begin{pmatrix} B_j \\ A_j \\ \vdots \\ Y \\ Z \end{pmatrix}, \begin{pmatrix} B_j \\ C_j \\ A_j \\ \vdots \\ \vdots \end{pmatrix}, \dots, \begin{pmatrix} B_j \\ C_j \\ \vdots \\ \vdots \\ A_j \end{pmatrix}, \dots, \begin{pmatrix} B_j \\ Y_j \\ \vdots \\ \vdots \\ C_j \end{pmatrix}}_{\mathcal{S}_{2,j}}, \dots, \underbrace{\begin{pmatrix} Z_j \\ A_j \\ \vdots \\ W_j \\ Y_j \end{pmatrix}, \begin{pmatrix} Z_j \\ A_j \\ \vdots \\ Y_j \\ W_j \end{pmatrix}, \dots, \begin{pmatrix} Z_j \\ Y_j \\ \vdots \\ \vdots \\ A_j \end{pmatrix}}_{\mathcal{S}_{n_{\text{blocks}},j}},$$

where we reordered the columns of each set $\mathcal{S}_{2,j}$ to emphasize that we can have the first $n_{\text{blocks}} - 1$ vectors of each $\mathcal{S}_{2,j}$ figuring block A_j in the second, third,..., and $n_{\text{blocks}}^{\text{th}}$ block of rows, respectively.

Consider

$$\mathcal{S}_1 := \mathcal{S}_{1,1} \cup \mathcal{S}_{1,2} \cup \cdots \cup \mathcal{S}_{1,p}.$$

Because $n \geq n_{\text{blocks}}^2 \cdot p$, we have that $b \geq n_{\text{blocks}} \cdot p$ and thus, for any j, A_j does not obtain as a linear combination of vectors in $\{A_l\}_{l \neq j} \bigcup \{B_j, C_j, \dots\}_{j=1}^{p}$ and thus, invoking the induction hypothesis, we see that

$$\begin{aligned} \operatorname{rank}(\mathcal{S}_1) &= \operatorname{rank}(\mathcal{S}_{1,1}) + \operatorname{rank}(\mathcal{S}_{1,2}) + \cdots + \operatorname{rank}(\mathcal{S}_{1,p}) \\ &= p \cdot ((n_{\text{blocks}} - 1) \cdot (n_{\text{blocks}} - 3) + 2) \end{aligned}$$

In the argument below, we invoke vectors and subvectors of $\mathbf{X}$ being linearly independent if the cardinality of the set they make up is less than the dimensionality of the vectors or subvectors. Repeatedly invoking should not raise concerns about existence; an $\mathbf{X}$ whose entries are drawn from a distribution that is uniformly continuous with respect to Lebesgue measure will satisfy all of these this with probability one.

Now consider adding the first column of the reordered set $\mathcal{S}_{2,1}$. A_1 does not appear in the second block of rows of any columns of $\mathcal{S}_1$ and A_1 cannot be produced as a linear combination of the second block of rows for any columns in $\mathcal{S}_1$ since there are $p \cdot (n_{\text{blocks}} - 1)$ distinct blocks in the second block of rows, and $b > p \cdot (n_{\text{blocks}} - 1)$. Appending the first column of the reordered set $\mathcal{S}_{2,1}$ must this increase the maximum rank by 1.

Let $\mathcal{S}$ denote $\mathcal{S}_1$ appended with the first column of the reordered set $\mathcal{S}_{2,1}$. Now consider adding the second column of the reordered set $\mathcal{S}_{2,1}$ to $\mathcal{S}$. A_1 does not appear in the third block of rows of any columns of $\mathcal{S}$ with and A_1 cannot be produced as a linear combination of the second block of rows for any columns in $\mathcal{S}_1$ since there are $p \cdot (n_{\text{blocks}} - 1)$ distinct blocks in the second block of rows, and $b > p \cdot (n_{\text{blocks}} - 1)$.

The same argument holds analogously for the 3^{rd} to $n_{\text{blocks}}^{\text{th}}$ vector of the reordered set $\mathcal{S}_{2,1}$. Hence, appending $S_{2,1}$ to $\mathcal{S}_1$ must increase the rank by $n_{\text{blocks}} - 1$. In fact, the same argument holds for the first—or any—vector of $S_{3,1}$,...,$S_{n_{\text{blocks}},1}$. Hence, appending $S_{3,1} \cup \cdots \cup S_{n_{\text{blocks}},1}$ to $\mathcal{S}$ must increase the rank by $n_{\text{blocks}} - 2$. This establishes the lower bound

$$\begin{aligned}\operatorname{rank}(\mathbf{GX}) &\geq p \cdot ((n_{\text{blocks}} - 1) \cdot (n_{\text{blocks}} - 3) + 2) + p \cdot (n_{\text{blocks}} - 1 + n_{\text{blocks}} - 2) \\ &= p \cdot (n_{\text{blocks}} \cdot (n_{\text{blocks}} - 2) + 2).\end{aligned}$$

To see that the bound may be attained, apply the argument of the $p = 1$ proof to show that within each subset $\mathcal{S}_{i_{\text{blocks}},j}$, $i_{\text{blocks}} = 1, \ldots, n_{\text{blocks}}$, we can generate the full set from $(n_{\text{blocks}} - 1) \cdot (n_{\text{blocks}} - 3) + 2$ linearly independent column vectors identified above to lower bound its rank. *Q.E.D.*

B.2. *Large Sample Theory*

The model is

$$\mathbf{Y} = \mathbf{1}_n\beta_0 + \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \cdots + \mathbf{X}_p\beta_p + \boldsymbol{\varepsilon}.$$

Let $\bar{\mathbf{X}}_1 = \mathcal{Q}\mathbf{X}_1$, $\mathcal{Q}$ is the orthogonal projection onto the orthogonal complement of the span of $\{g\mathbf{X}_2, \ldots, g\mathbf{X}_p\}_{g\in\mathbf{G}_n}$, and the residuals $\hat{\varepsilon}_i$ obtain from the regression of $\mathbf{Y}$ on an intercept term, $\{g\mathbf{X}_1\}_{g\in\mathbf{G}_n}$ and $\{g\mathbf{X}_2, \ldots, g\mathbf{X}_p\}_{g\in\mathbf{G}_n}$.

For the resampling asymptotic theory, we are interested in the statistic

$$t_{\text{id}}\left(\beta_1^0\right) = \frac{\mathbf{X}_1^T\mathcal{Q}\left(\mathbf{Y} - \mathbf{X}_1\beta_1^0\right)}{\hat{\sigma}_{\text{id}}} = \frac{\left(\mathcal{Q}\mathbf{X}_1\right)^T\left(\mathbf{Y} - \mathbf{X}_1\beta_1^0\right)}{\hat{\sigma}_{\text{id}}}, \tag{20}$$

where

$$\hat{\sigma}_{\text{id}}^2 = \frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{X}}_{i1}^2\hat{\varepsilon}_i^2, \tag{21}$$

with $\bar{\mathbf{X}}_1 = \mathcal{Q}\mathbf{X}_1$.

Recall that we use $\mathcal{C}(A)$ to designate the space span by the columns of a matrix A.

THEOREM A.2: *Let $J_{t_{\text{id}},n}(\cdot)$ be the sampling distribution of $t_{\text{id}}(\beta_1^0)$. Suppose $(Y_i, X_i) \sim \mathcal{P}$ independently, and that Assumptions M, S and D hold. Suppose that $n_{\text{blocks}}^8/n \to 0$. Then, under the null hypothesis (3),*

$$\lim_{n\to\infty}\sup_{s\in\mathbb{R}}\left|J_{t_{\text{id}},n}(s) - \Phi(s)\right| = 0,$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

PROOF OF THEOREM A.2: Let $\bar{X}_{i1}^*$ be the population analog of $\bar{X}_{i1}$. The numerator of (4) satisfies

$$\frac{1}{\sqrt{n}}\bar{\mathbf{X}}_1^T\boldsymbol{\varepsilon} \xrightarrow{d} N\left(0, E\left(\bar{X}_{i1}^{*2}\varepsilon_i^2\right)\right). \tag{22}$$

To see this explicitly, construct a full rank matrix $\overrightarrow{\mathbf{X}} \in \mathbb{R}^{n\times\overrightarrow{p}}$ satisfying $\mathcal{C}\left(\overrightarrow{\mathbf{X}}\right) = \mathcal{C}\left(\mathbf{G}\mathbf{X}_{-1}\right)$, where $\mathbf{G}\mathbf{X}_{-1} = \left(g^{(1)}\mathbf{X}_2, ..., g^{(M)}\mathbf{X}_2, g^{(1)}\mathbf{X}_3, \ldots, g^{(M)}\mathbf{X}_p\right)$. Assume without loss of generality that the maximum rank is attained and $\overrightarrow{p} = (p-1)\left(n_{\text{blocks}} \cdot (n_{\text{blocks}} - 2) + 2\right)$.

Let $\bar{\mathbf{X}}_1 = (I - \mathbf{M})\,\mathbf{X}_1 =: \mathbf{X}_1 - \overrightarrow{\mathbf{X}}\hat{\gamma}$, where $\mathbf{M}$ is the orthogonal projection onto $\mathcal{C}\left(\overrightarrow{\mathbf{X}}\right)$. Let $\bar{\mathbf{X}}_1^* = \mathbf{X}_1 - E^*\left[\mathbf{X}_1 | \overrightarrow{\mathbf{X}}\right]$, where $E^*\left[\mathbf{X}_1 | \overrightarrow{\mathbf{X}}\right] =: \overrightarrow{\mathbf{X}}\gamma$ stands for the best linear predictor of $\mathbf{X}_1$ given $\overrightarrow{\mathbf{X}}$.[5] Note that the numerator of (4) satisfies

$$\frac{1}{\sqrt{n}}\bar{\mathbf{X}}_1^T\boldsymbol{\varepsilon} = \frac{1}{\sqrt{n}}\bar{\mathbf{X}}_1^{*T}\boldsymbol{\varepsilon} - \frac{1}{\sqrt{n}}\bar{\mathbf{X}}_1^{*T}\mathbf{M}\boldsymbol{\varepsilon}.$$

The first term satisfies

$$\frac{1}{\sqrt{n}}\bar{\mathbf{X}}_1^{*T}\boldsymbol{\varepsilon} \xrightarrow{d} N\left(0, E\left(\bar{X}_1^{*2}\varepsilon^2\right)\right), \tag{23}$$

by Assumption M.

The second term has order of magnitude at most $O_P\left(\overrightarrow{p}^2/\sqrt{n}\right)$ because

$$\frac{1}{\sqrt{n}}\left|\frac{\bar{\mathbf{X}}_1^{*T}\overrightarrow{\mathbf{X}}}{\sqrt{n}}\left(\frac{\overrightarrow{\mathbf{X}}^T\overrightarrow{\mathbf{X}}}{n}\right)^{-1}\frac{\overrightarrow{\mathbf{X}}^T\boldsymbol{\varepsilon}}{\sqrt{n}}\right| \leq \frac{1}{\sqrt{n}}\left\|\frac{\bar{\mathbf{X}}_1^{*T}\overrightarrow{\mathbf{X}}}{\sqrt{n}}\right\|_2 \lambda_{\max}\left(\left(\frac{\overrightarrow{\mathbf{X}}^T\overrightarrow{\mathbf{X}}}{n}\right)^{-1}\right)\left\|\frac{\overrightarrow{\mathbf{X}}^T\boldsymbol{\varepsilon}}{\sqrt{n}}\right\|_2, \tag{24}$$

where $\lambda_{\max}\left(\left(\frac{\overrightarrow{\mathbf{X}}^T\overrightarrow{\mathbf{X}}}{n}\right)^{-1}\right) = O_P\left(\overrightarrow{p}\right)$, $\left\|\bar{\mathbf{X}}_1^{*T}\overrightarrow{\mathbf{X}}/\sqrt{n}\right\|_2 = O_P\left(\sqrt{\overrightarrow{p}}\right)$, and $\left\|\overrightarrow{\mathbf{X}}^T\boldsymbol{\varepsilon}/\sqrt{n}\right\|_2 = O_P\left(\sqrt{\overrightarrow{p}}\right)$.

The rate on the norm $\left\|\bar{\mathbf{X}}_1^{*T}\overrightarrow{\mathbf{X}}/\sqrt{n}\right\|_2$ obtains by Markov inequality because

$$P\left(\left\|\frac{\bar{\mathbf{X}}_1^{*T}\overrightarrow{\mathbf{X}}}{\sqrt{n}}\right\|_2^2/\overrightarrow{p} \geq a\right) \leq \frac{\sum_{j=1}^{\overrightarrow{p}} E\left[\left(\frac{\bar{\mathbf{X}}_1^{*T}\overrightarrow{\mathbf{X}}_j}{\sqrt{n}}\right)^2\right]}{a\overrightarrow{p}} \leq C\frac{E\left[\left(\frac{\bar{\mathbf{X}}_1^{*T}\mathbf{X}_2}{\sqrt{n}}\right)^2\right]}{a}, \tag{25}$$

where $C = \max\left\{1, E\left[\bar{X}_{i1}^{*2}\right]E\left[X_{i2}^2\right]/E\left[\bar{X}_{i1}^{*2}X_{i2}^2\right]\right\}$, and the moment on the right-hand side term converges by Assumption M and the von Bahr Theorem (Von Bahr, 1965, see also Theorem 6.3, DasGupta, 2008). The analogous argument holds for the norm $\left\|\overrightarrow{\mathbf{X}}^T\boldsymbol{\varepsilon}/\sqrt{n}\right\|_2$.

Taken together, we obtain (22). By Polya's lemma, the convergence is uniform.

Hence, by Slutsky's theorem, it suffices to show that $\hat{\sigma}^2$ converges in probability to $\sigma^2 = E\left(\bar{X}_{1i}^{*2}\varepsilon_i^2\right)$. Consider the decomposition

$$\hat{\sigma}^2 - \sigma^2 = \frac{1}{n}\sum_{i=1}^n \bar{X}_{i1}^2\hat{\varepsilon}_i^2 - E\left(\bar{X}_{i1}^{*2}\varepsilon_i^2\right)$$

[5]The vector best linear predictor is well defined and in this specific case, even though rows of $\overrightarrow{\mathbf{X}}$ are not independent, we have $\left(E^*\left[\mathbf{X}_1 | \overrightarrow{\mathbf{X}}\right]\right)_i = E^*\left[\mathbf{X}_1 | \mathbf{X}_2\right]_i = E^*\left[\mathbf{X}_{i,1} | \mathbf{X}_{i,2}\right]$.

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\bar{X}_{i1}^{2}-\bar{X}_{i1}^{*2}\right)\left(\hat{\varepsilon}_{i}^{2}-\varepsilon_{i}^{2}\right)+\frac{1}{n}\sum_{i=1}^{n}\left(\bar{X}_{i1}^{2}-\bar{X}_{i1}^{*2}\right)\varepsilon_{i}^{2}$$
$$+\frac{1}{n}\sum_{i=1}^{n}\bar{X}_{i1}^{*2}\left(\hat{\varepsilon}_{i}^{2}-\varepsilon_{i}^{2}\right)+\frac{1}{n}\sum_{i=1}^{n}\bar{X}_{i1}^{*2}\varepsilon_{i}^{2}-E\left(\bar{X}_{i1}^{*2}\varepsilon_{i}^{2}\right).$$

The first term satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\left(\bar{X}_{i1}^{2}-\bar{X}_{i1}^{*2}\right)\left(\hat{\varepsilon}_{i}^{2}-\varepsilon_{i}^{2}\right)\leq\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\bar{X}_{i1}^{2}-\bar{X}_{i1}^{*2}\right)^{2}}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\varepsilon}_{i}^{2}-\varepsilon_{i}^{2}\right)^{2}}.$$

Notice that

$$\frac{1}{n}\sum_{i=1}^{n}\left(\bar{X}_{i1}^{2}-\bar{X}_{i1}^{*2}\right)^{2}=\frac{1}{n}\sum_{i=1}^{n}\bar{X}_{i1}^{4}-2\frac{1}{n}\sum_{i=1}^{n}\bar{X}_{i1}^{*2}\bar{X}_{i1}^{2}+\frac{1}{n}\sum_{i=1}^{n}\bar{X}_{i1}^{*4}$$
$$\leq\frac{1}{n}\sum_{i=1}^{n}\bar{X}_{i1}^{4}-2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\bar{X}_{i1}^{*4}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\bar{X}_{i1}^{4}}+\frac{1}{n}\sum_{i=1}^{n}\bar{X}_{i1}^{*4}\xrightarrow{P}0,$$

by M.2 and the law of large numbers. Specifically, $\frac{1}{n}\sum_{i=1}^{n}\bar{X}_{i1}^{4}\xrightarrow{P}E\left[\bar{X}_{i1}^{*4}\right]$ by M.2 and since $\left\|\hat{\gamma}-\gamma\right\|_{2}=O_{P}\left(\sqrt{\frac{\vec{p}^{3}}{n}}\right)$, which obtains by manipulations analogous to those in (24).

By a similar argument, the second term satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\left(\bar{X}_{i1}^{2}-\bar{X}_{i1}^{*2}\right)\varepsilon_{i}^{2}\leq\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\bar{X}_{i1}^{2}-\bar{X}_{i1}^{*2}\right)^{2}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{i}^{4}}\xrightarrow{P}0.$$

By the symmetric argument, $\frac{1}{n}\sum_{i=1}^{n}\bar{X}_{i1}^{*2}\left(\hat{\varepsilon}_{i}^{2}-\varepsilon_{i}^{2}\right)\xrightarrow{P}0$. Finally, $\frac{1}{n}\sum_{i=1}^{n}\bar{X}_{i1}^{*2}\varepsilon_{i}^{2}-E\left(\bar{X}_{i1}^{*2}\varepsilon_{i}^{2}\right)\xrightarrow{P}0$ by M.2 and the law of large numbers.

Q.E.D.

We can now turn to the large sample theory of the randomized test. We opt for the Hoeffding approach as our proof technique. The distinction with the fully conditional approach is discussed in Pouliot (2023) and Chung and Romano (2013).

THEOREM A.3: *Let g and g' be independent and uniformly distributed over $\mathbf{G}_n$, i.e., $g,g'\sim U\left[\mathbf{G}_n\right]$, and be independent of the data $\mathbf{W}_n$. Suppose, under P_n,*

$$\left(T_n\left(g\mathbf{W}_n\right),T_n\left(g'\mathbf{W}_n\right)\right)\xrightarrow{d}\left(T,T'\right),\tag{26}$$

where T and T' are independent, each with common cumulative distribution function $R_T(\cdot)$. Then, for all continuity points S of $R_T(\cdot)$,

$$\hat{R}_{T,n}(S)\xrightarrow{P}R_T(S).\tag{27}$$

Conversely, if (27) holds for some limiting cumulative distribution function $R_T(\cdot)$ whenever S is a continuity point, then (26) holds.

The sufficiency extension of Hoeffding's Theorem is due to Chung and Romano (2013).

THEOREM A.4: *Suppose* $(Y_i, X_i) \sim \mathcal{P}$ *independently, and that Assumptions M, S and D hold. Suppose that* $n_{\text{blocks}}^8 / n \to 0$*. Further suppose that* $\varepsilon_i^2 = o_P(n)$*, uniformly, and that* $\frac{n_{\text{blocks}}}{b^\zeta} \to \infty$*, for some* $0 < \zeta < 1$*. Then, under the null hypothesis (3), the distribution* $\hat{R}_{t_g, n}$ *satisfies*

$$\hat{R}_{t_g,n}(s) \xrightarrow{P} \Phi(s),$$

for all points of continuity of s, where $\Phi(\cdot)$ *is the standard normal cumulative distribution.*

PROOF OF THEOREM A.4: Let

$$T_g = T\left((\mathbf{X}_1, \mathbf{X}_2), g(\mathbf{Y} - \mathbf{X}_1\beta_1)\right) = \frac{1}{\sqrt{n}} \bar{\mathbf{X}}_1^T g(\mathbf{Y} - \mathbf{X}_1\beta_1),$$

which under the null takes on the value

$$\frac{1}{\sqrt{n}} \bar{\mathbf{X}}_1^T (g\boldsymbol{\varepsilon}).$$

First, we want to verify Hoeffding's condition for T_g. Let $g, g' \sim U[\mathbf{G}_n]$, drawn independently, and consider the random pair $(T_g, T_{g'})$. We are interested in the limiting distribution of this bivariate random variable. We use the Cramér-Wold device. Take any $(a_1, a_2) \in \mathbb{R}^2$, under the null hypothesis,

$$\begin{aligned} a_1 T_g + a_2 T_{g'} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \bar{X}_{i1} \left(a_1 \varepsilon_{g(i)} + a_2 \varepsilon_{g'(i)}\right) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \bar{X}_{i1}^* \left(a_1 \varepsilon_{g(i)} + a_2 \varepsilon_{g'(i)}\right) + o_P(1), \end{aligned}$$

by the same argument as in the proof of Theorem A.2.

We establish the CLT by verifying the main condition of Lemma 11.3.3 of Lehmann and Romano (2005), i.e.,

$$\frac{\max_{i=1\ldots n} \frac{1}{n} \left(a_1 \varepsilon_{g(i)} + a_2 \varepsilon_{g'(i)}\right)^2}{\frac{1}{n} \sum_{i=1}^n \left(a_1 \varepsilon_{g(i)} + a_2 \varepsilon_{g'(i)}\right)^2} \to 0 \text{ as } n \to \infty. \tag{28}$$

In order to both check (28) and to establish the asymptotic variance of T_g, we compute the probability limit of the denominator. Condition on g and g' being such that $\#\{i : g(i) = g'(i)\} < n_{\text{eq}}$, where n_{eq} is some small value depending on n which we will pick later, and which is required to satisfy $\frac{n_{\text{eq}}}{n} \to 0$. Then,

$$\frac{1}{n} \sum_{i=1}^n \left(a_1 \varepsilon_{g(i)} + a_2 \varepsilon_{g'(i)}\right)^2$$

$$= a_1^2 \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{g(i)}^2 + a_2^2 \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{g'(i)}^2 + 2a_1 a_2 \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{g(i)} \varepsilon_{g'(i)}$$

$$= a_1^2 \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{g(i)}^2 + a_2^2 \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{g'(i)}^2 + 2a_1 a_2 \frac{1}{n} \sum_{i:g(i) \neq g'(i)} \varepsilon_{g(i)} \varepsilon_{g'(i)} + 2a_1 a_2 \frac{1}{n} \sum_{i:g(i) = g'(i)} \varepsilon_{g(i)} \varepsilon_{g'(i)}$$

$$\to a_1^2 E\left[\varepsilon_i^2\right] + a_2^2 \left[\varepsilon_i^2\right] + 0 + 2a_1 a_2 E\left[\varepsilon_i^2\right] \cdot \lim_{n\to\infty} \frac{n_{\text{eq}}}{n} = a_1^2 E\left[\varepsilon_i^2\right] + a_2^2 E\left[\varepsilon_i^2\right],$$

relying on Assumption M.

We used that

$$\frac{1}{n} \sum_{i:g(i)=g'(i)} \varepsilon_{g(i)} \varepsilon_{g'(i)} = \frac{n_{\text{eq}}}{n} \frac{1}{n_{\text{eq}}} \sum_{i:g(i)=g'(i)} \varepsilon_{g(i)} \varepsilon_{g'(i)},$$

and

$$\frac{1}{n_{\text{eq}}} \sum_{i:g(i)=g'(i)} \varepsilon_{g(i)} \varepsilon_{g'(i)} \to E\left[\varepsilon_i^2\right],$$

by the law of large numbers.

The probability that the conditioned upon event does not realize is

$$P\left(\#\left\{i : g(i) = g'(i)\right\} \geq n_{\text{eq}}\right) \leq \frac{E\left[\#\left\{i : g(i) = g'(i)\right\}\right]}{n_{\text{eq}}} = \frac{b}{n_{\text{eq}}}.$$

We can pick need n_{eq} such that $\frac{n_{\text{eq}}}{n} \to 0$ and such that $\frac{b}{n_{\text{eq}}} \to 0$, as $n \to \infty$. For instance, pick $n_{\text{eq}} = b^{1+\zeta}$.

In particular, for any $c > 0$ and $\epsilon > 0$, for n large enough,

$$P\left(\#\left\{i : g(i) = g'(i)\right\} \geq n_{\text{eq}}\right) < \epsilon/2 \tag{29}$$

and

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(a_1\varepsilon_{g(i)} + a_2\varepsilon_{g'(i)}\right)^2 - \left(a_1^2 E\left[\varepsilon_i^2\right] + a_2^2 E\left[\varepsilon_i^2\right]\right)\right| > c \right. \\ \left. : \#\left\{i : g(i) = g'(i)\right\} < n_{\text{eq}}\right) < \epsilon/2. \tag{30}$$

The probability limit thus holds unconditionally, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n}\left(a_1\varepsilon_{g(i)} + a_2\varepsilon_{g'(i)}\right)^2 \xrightarrow{P} a_1^2 E\left[\varepsilon_i^2\right] + a_2^2 E\left[\varepsilon_i^2\right].$$

Then, the denominator of (28) converges in probability and, because $\varepsilon_i^2/n = o_P(1)$ uniformly, its numerator is $o_p(1)$. Therefore, by Slutsky's Theorem, the quotient must be $o_P(1)$, hence verifying the condition. By Lemma 11.3.3 of Lehmann and Romano (2005) and Slutsky's Theorem, we have that

$$a_1 T_g + a_2 T_{g'} \xrightarrow{d} N\left(0, E\left[\bar{X}_{i1}^{*2}\right]\left(a_1^2 E\left[\varepsilon_i^2\right] + a_2^2 E\left[\varepsilon_i^2\right]\right)\right),$$

thus verifying Hoeffding's condition. In particular, we find that the asymptotic covariance between T_g and $T_{g'}$ is zero. By Hoeffding's Theorem, we then have that

$$\hat{R}_{T,n}(S) \xrightarrow{P} R_T(S), \tag{31}$$

for all points of continuity S, where $R_T(S)$ is the distribution function of $N\left(0, E\left[\bar{X}_{i1}^2\right] E\left[\varepsilon_i^2\right]\right)$.

Second, we verify that the studentizing term $\hat{\sigma}_g$ is consistent. Observe that

$$\begin{aligned}
\hat{\sigma}_g - E\left[\bar{X}_{i1}^{*2}\right] E\left[\varepsilon_i^2\right] &= \frac{1}{n}\sum_{i=1}^{n} \bar{X}_{g(i),1}^2 \hat{\varepsilon}_i^2 - E\left[\bar{X}_{i1}^{*2}\right] E\left[\varepsilon_i^2\right] \\
&= \frac{1}{n}\sum_{i:i\neq g(i)}^{n} \bar{X}_{g(i),1}^2 \hat{\varepsilon}_i^2 - E\left[\bar{X}_{i1}^{*2}\right] E\left[\varepsilon_i^2\right] + \frac{1}{n}\sum_{i:i=g(i)}^{n} \bar{X}_{g(i),1}^2 \hat{\varepsilon}_i^2 \\
&= \frac{1}{n}\sum_{i:i\neq g(i)}^{n} \bar{X}_{g(i),1}^2 \hat{\varepsilon}_i^2 - E\left[\bar{X}_{i1}^{*2}\right] E\left[\varepsilon_i^2\right] + o_P(1) \\
&= O(1)\left(\frac{1}{n-m}\sum_{i:i\neq g(i)} \bar{X}_{g(i),1}^2 \varepsilon_i^2 - E\left[\bar{X}_{i1}^{*2}\right] E\left[\varepsilon_i^2\right]\right) + o_P(1),
\end{aligned}$$

where the number of match is $m = \#\{i : i \neq g(i)\}$ and $m/n = o_P(1)$, and which goes to zero in probability by an argument analogous to that used in the proof of Theorem A.2.

We must then have that

$$\hat{\sigma}_g \xrightarrow{P} E\left[\bar{X}_{i1}^{*2}\right] E\left[\varepsilon_i^2\right]. \tag{32}$$

Third, we find by Slutsky's Theorem for randomization distributions, Theorem 5.2 of Chung and Romano (2013), that (31) and (32) imply

$$\hat{R}_{t_g,n}(s) \xrightarrow{P} R_{t_g}(s),$$

for all points of continuity of s, where $R_{t_g}(s)$ is the distribution function of $N(0,1)$. *Q.E.D.*

REMARK 2: We analyzed the case for a null hypothesis on a scalar coefficient. The large-sample theory for the multivariate case obtains analogously.

REFERENCES

BERRY, CHRISTOPHER AND ANTHONY FOWLER (2018): "Leadership or Luck? Randomization Inference for Leader Effects (RIFLE)," . [15]

BREUSCH, TREVOR S AND ADRIAN R PAGAN (1979): "A simple test for heteroscedasticity and random coefficient variation," *Econometrica: Journal of the econometric society*, 1287–1294. [2]

BROWN, BRUCE M AND JS MARITZ (1982): "Distribution-Free Methods in Regression 1," *Australian Journal of Statistics*, 24 (3), 318–331. [7]

CHUNG, EUNYI AND JOSEPH P ROMANO (2013): "Exact and asymptotically robust permutation tests," *The annals of statistics*. [24, 25, 27]

DASGUPTA, ANIRBAN (2008): *Asymptotic theory of statistics and probability*, vol. 180, Springer. [23]

D'HAULTFŒUILLE, XAVIER AND PUREVDORJ TUVAANDORJ (2024): "A robust permutation test for subvector inference in linear regressions," *Quantitative Economics*, 15 (1), 27–87. [3, 7, 8, 14]

DICICCIO, CYRUS J AND JOSEPH P ROMANO (2017): "Robust permutation tests for correlation and regression coefficients," *Journal of the American Statistical Association*, 112 (519), 1211–1220. [2, 3, 8]

DONALD, STEPHEN G AND WHITNEY K NEWEY (1994): "Series estimation of semilinear models," *Journal of Multivariate Analysis*, 50 (1), 30–40. [10]

DUFOUR, JEAN-MARIE, EMMANUEL FLACHAIRE, AND LYNDA KHALAF (2019): "Permutation tests for comparing inequality measures," *Journal of Business & Economic Statistics*, 37 (3), 457–470. [2]

GUTENBRUNNER, CHRISTOPH AND J JURECKOVÁ (1992): "Regression rank scores and regression quantiles," *The Annals of Statistics*, 305–330. [15]

IMBENS, GUIDO W AND PAUL R ROSENBAUM (2005): "Robust, accurate confidence intervals with a weak instrument: quarter of birth and education," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 168 (1), 109–126. [14, 15]

KOENKER (2005): *Quantile Regression (Econometric Society Monographs; No. 38)*, Cambridge university press. [15]

KOENKER, ROGER (1994): "Confidence intervals for regression quantiles," in *Asymptotic statistics*, Springer, 349–359. [15]

KOENKER, ROGER AND JOSÉ AF MACHADO (1999): "GMM inference when the number of moment conditions is large," *Journal of Econometrics*, 93 (2), 327–344. [8]

LEHMANN, EL AND JOSEPH P ROMANO (2005): "Testing statistical hypotheses," *Springer texts in statistics*. [2, 9, 25, 27]

LEI, LIHUA AND PETER J BICKEL (2021): "An assumption-free exact test for fixed-design linear models with exchangeable errors," *Biometrika*, 108 (2), 397–412. [2, 3, 7, 12, 14]

MROZ, THOMAS A (1987): "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, 55 (4), 765. [14]

NEUHAUS, GEORG (1993): "Conditional rank tests for the two-sample problem under random censorship," *The Annals of Statistics*, 1760–1779. [8]

POULIOT, GUILLAUME (2023): "Optimization-Conscious Econometrics," PhD Course Manuscript. [24]

ROMANO, JOSEPH P (1990): "On the behavior of randomization tests without a group invariance assumption," *Journal of the American Statistical Association*, 85 (411), 686–692. [2]

STUDENT (1908): "The probable error of a mean," *Biometrika*, 6 (1), 1–25. [1]

VON BAHR, BENGT (1965): "On the convergence of moments in the central limit theorem," *The Annals of Mathematical Statistics*, 808–818. [23]

WEN, KAIYUE, TENGYAO WANG, AND YUHAO WANG (2022): "Residual Permutation Test for High-Dimensional Regression Coefficient Testing," *arXiv preprint arXiv:2211.16182*. [2, 7, 12, 14]

WOOLDRIDGE, JEFFREY M (2010): *Econometric Analysis of Cross Section and Panel Data*, MIT press. [1]

——— (2020): *Introductory Econometrics: A Modern Approach (7e)*, Cengage learning. [13, 14]

YOUNG, ALWYN (2023): "Asymptotically Robust Permutation-based Randomization Confidence Intervals for Parametric OLS Regression," *European Economic Review*, 104644. [2]

[Name Surname; will be inserted later]