Invariant Inference via Residual Randomization

Panos Toulis
University of Chicago,
Booth School of Business

July 4, 2022

Abstract

The dominant paradigm in statistical inference relies on a structure of i.i.d. data from a hypothetical infinite population. Despite its unquestionable success, this framework is inflexible under complex data structures, even in those cases where it is clear what the infinite population represents. In this paper, we explore an alternative framework whereby the basis of inference is only an invariance assumption on the model errors, such as exchangeability or sign symmetry. As a general method for this problem of *invariant inference*, we propose a randomization-based procedure. To test significance, the procedure repeatedly calculates a suitable test statistic over transformations of the residuals according to the invariant. Inversion of the test can produce confidence intervals. We prove general conditions for asymptotic validity of this procedure, and illustrate in many data structures, including clustered errors in one-way and two-way layouts. We find that invariant inference via residual randomization offers three main advantages over classical inference: (1) It can be valid under weaker conditions than classical inference, allowing for problems with heavy-tailed data, finite clustering, and even some high-dimensional settings. (2) It has better finite-sample performance as it does not rely on the regularity conditions needed for classical asymptotics. (3) It addresses the problem of inference in a unified way that adapts to the data structure. Classical procedures like OLS or bootstrap, on the other hand, presuppose the i.i.d. structure, and need to be modified whenever the actual problem structure is different. This mismatch in the classical framework has led to a multitude of robust error techniques and bootstrap variants, which frequently confounds applied research. We corroborate these findings with extensive empirical evaluations. Residual randomization performs favorably against many alternatives, including robust error methods, bootstrap variants, and hierarchical models.

1 Introduction

Suppose data $(y_1, x_1), \dots, (y_n, x_n) \in \mathbb{R} \times \mathbb{R}^p$ and a model f with fixed parameters $\beta \in \mathbb{R}^p$,

$$y_i = f(x_i, \beta) + \varepsilon_i, \tag{1}$$

where ε_i is unobserved error. Define $X = (x_1, \dots, x_n)$, $y = (y_1, \dots, y_n)$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$. The goal of this paper is to study the conditions of valid testing and inference on β when (y_i, x_i) are possibly non-i.i.d., but the errors satisfy an invariance assumption conditional on X,

$$\varepsilon \overset{\mathrm{d}}{=} \mathrm{g}\varepsilon \mid X, \text{ for all } \mathrm{g} \in \mathcal{G}_n, \; X \in \mathbb{X}^{(n)} \subseteq \mathbb{R}^{n \times p}. \tag{2}$$

In (2), $\mathcal{G}_n$ is the "inferential primitive", an algebraic group of $\mathbb{R}^n \to \mathbb{R}^n$ linear maps. We call this problem *invariant inference* to distinguish it from classical inference —OLS, bootstrap, or likelihood-based— that typically relies on i.i.d. data from a hypothetical infinite population.

In this setting, a simple null hypothesis on β, say $H_0 : \beta = \beta_0$, can be tested through the randomization method (Fisher, 1935; Lehmann and Romano, 2005). Consider a test statistic, T_n, such that $T_n = t_n(\varepsilon)$ under the null hypothesis, for a known measurable function $t_n \in \mathbb{R}^n \to \mathbb{R}$. The (one-sided) randomization test then rejects or accepts the null hypothesis as follows.

$$\phi_n^*(y, X) = \mathbb{1}\{T_n > c_{n,\alpha}(y - f(X, \beta_0))\}. \tag{3}$$

Here, $f(X, \beta_0)$ is understood component-wise, and $c_{n,\alpha} : \mathbb{R}^n \to \mathbb{R}$ is the critical value function,

$$c_{n,\alpha}(u) = \inf\left\{z \in \mathbb{R} : \frac{1}{|\mathcal{G}_n|} \sum_{\mathrm{g} \in \mathcal{G}_n} \mathbb{1}\{t_n(\mathrm{g}u) \le z\} \ge 1 - \alpha\right\}. \tag{4}$$

The randomization test in (3) has some highly desirable properties. First, it is valid for any finite $n > 0$, following standard randomization theory (Lehmann and Romano, 2005, Theorem 15.2.3). Second, it does not rely on any assumptions about the distribution or asymptotics of (X, ε) as long as (2) holds. And, third, through an appropriate choice of $\mathcal{G}_n$ the randomization test remains valid even when the errors have a complex structure. However, if the null hypothesis is partial, then the randomization test is not applicable. This includes hypotheses of significance, $H_0 : a'\beta = a_0$ or $H_0 : R\beta = 0$, where $a \in \mathbb{R}^p, a_0 \in \mathbb{R}$, $R \in \{0,1\}^{k \times p}$, $k < p$, are all fixed. Under these nulls, the error component cannot be identified uniquely, and so the critical value in (3) cannot be calculated.

In this paper, we address this issue by using error residuals in the randomization test in place of the true errors. That is, we first consider an estimator of β, say $\hat{\beta}_n$, preferably calculated under the null. Then, we plug in the residuals, $\widehat{\varepsilon} = y - f(X, \hat{\beta}_n)$, to form a residual randomization test:

$$\phi_n(y, X) = \mathbb{1}\{T_n > c_{n,\alpha}(y - f(X, \hat{\beta}_n))\}. \tag{5}$$

Of course, the randomization test in (5) is approximate, and thus no longer valid in finite samples.

It is then natural to ask: Are there conditions for which this approximate test is valid asymptotically? Moreover, does the approximate test inherit any of the robustness properties of the original randomization test, and how does this relate to the structure of $\mathcal{G}_n$?

1.1 Overview of results and Examples

In Section 2, we show that the main condition for (asymptotic) validity of (5) has an elegant form:

$$\frac{\mathbb{E}[(t_n(G\widehat{\varepsilon}) - t_n(G\varepsilon))^2]}{\mathbb{E}[(t_n(G'\varepsilon) - t_n(G''\varepsilon))^2]} \to 0, \tag{6}$$

where $G, G', G'' \sim \text{Unif}(\mathcal{G}_n)$ i.i.d. The proof of this result is non-asymptotic, and so it "preserves" the robustness properties of the original randomization method. Condition (6) does not require, for example, that $\hat{\beta}_n$ has particular asymptotics, or that it is even consistent. The condition only stipulates that the variation in the approximation error from the feasible residual-based test (numerator) is dominated by the natural variation of the infeasible test using the true errors ε (denominator). In Theorem 2, we show that the rate at which (6) is satisfied also determines a finite-sample bound on the over-rejection of the feasible test. Theorem 3 shows a similar result on the test's power. Under regular conditions, this rate is $O(n^{-1/3})$, which reveals a robustness-efficiency trade-off in the randomization method that we discuss throughout the paper.

In Section 3, we specialize (6) in the linear setting. To review these results, and also clarify the scope of invariant inference via residual randomization, we proceed with some applied examples.

Example 1 (Exchangeability)**.** *Consider a setting with n patients, where y denotes a health outcome and x denotes usage of a medical device. We fit a simple linear model:*

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \tag{7}$$

In classical inference, we assume that (x_i, y_i) or (x_i, ε_i) are i.i.d. from an infinite population, and then proceed with asymptotic inference on β. This assumption can be unrealistic, however, because of sample bias. Moreover, it is often unclear even what the population concept really represents—it may be a "fantasy world" (Freedman and Lane, 1983b, Sec. 1). Alternatively, we may view ε_i as measurement errors, or other unobserved shocks, common to all patients in our sample:

$$\varepsilon_i = g_{0,n}(X) + g_{1,n}(X)\epsilon_i, \quad \epsilon_i \ \sim i.i.d. \tag{8}$$

In this error structure, $g(\cdot)$ denotes a common effect on all patients, and ϵ_i are idiosyncratic errors. Even though these functions are unknown, invariance (2) holds with $\mathcal{G}_n$ as the permutation group on n elements, allowing for invariant inference. In this context, we show that Condition (6) reduces to the remarkably simple $p/n = o(1)$—see Theorem 4. This condition does not restrict the asymptotics of (X, ε), and even allows certain high-dimensional regimes with $p < n$ and $p \to \infty$. Of course, inference on β_0 is precluded because the intercept is not identifiable under only exchangeability. ■

Example 2 (Sign symmetry; double invariance). *We could account for heterogeneity in model (7) by assuming an error structure of the form:*

$$\varepsilon_i = g_{0,n}(X) + g_{i,n}(X)\epsilon_i, \quad \epsilon_i \overset{\mathrm{d}}{=} -\epsilon_i. \tag{9}$$

Here, $g_i(\cdot)$ denotes an effect on patient i that may depend on i's characteristics. Even though the $g_i(\cdot)$s are unknown, the constant term, $g_{0,n}$, can be subsumed by the intercept, and so invariance (2) holds with $\mathcal{G}_n$ as the group of $n \times n$ random sign matrices. In this context, we show that Condition (6) reduces to simple conditions on the leverage structure of X, ε (Theorem 5). In Theorem 6, we extend these results to a double invariance combining exchangeability and sign symmetry. ■

Example 3 (One-way clustering). *Consider an ecological study (Section 5.1) where y measures a bacterium's spore density in a hive, and x is infection status of the hive. There are $\mathrm{c} = 1, \ldots, J$ total hives, and $k = 1, \ldots K$ repeated measurements per hive. With such a clustered structure (per hive), it is common to index a datapoint as $i = (\mathrm{c}, k)$, and employ a linear mixed model of the form $y_{\mathrm{c}k} = \beta_0 + \beta_1 x_{\mathrm{c}k} + \eta_{\mathrm{c}} + \epsilon_{\mathrm{c}k}$. This is also known as a "random effects one-way layout" (Lehmann and Casella, 2006, 3.5), and includes η_{c} as cluster (random) effects, and $\epsilon_{\mathrm{c}k}$ as independent noise. Standard inference with this model requires assumptions on the design balance (Lehmann and Casella, 2006, Example 5.1), and normality assumptions for asymptotic analysis (McCulloch and Searle, 2004).*

An alternative is to adopt invariant inference by considering a clustered error structure,

$$\varepsilon_i \equiv \varepsilon_{\mathrm{c}k} = \eta_{\mathrm{c}} + \epsilon_{\mathrm{c}k}, \quad i = (\mathrm{c}, k), \tag{10}$$

where $\eta = (\eta_{\mathrm{c}})$ and $\epsilon = (\epsilon_{\mathrm{c}k})$ are mutually independent. Eq. (10) could be made more general by including additional factors like the "$g(\cdot)$" terms in Examples 1-2. Here, we omitted such terms to keep the notation simple.

(a) If η_{c} are exchangeable, and ϵ are i.i.d., then ε are exchangeable within each cluster. Thus, (2) holds with $\mathcal{G}_n$ as the (sub)group of permutations within clusters.

(b) If η_{c} and ϵ are sign symmetric, then ε are also sign symmetric across clusters. Thus, (2) holds with $\mathcal{G}_n$ as the (sub)group of $n \times n$ random sign matrices with an appropriately blocked diagonal.

In Theorems 7-8 we analyze these structures, and show that Condition (6) is satisfied as long as X, ε are not too highly leveraged across clusters, while also allowing for large p. Our method is also valid when (a) and (b) are combined—see remarks of Theorem 8. ■

Example 4 (Two-way clustering). *In trade data (Section 5.2), y usually measures trade flows between two countries, and x denotes their economic characteristics. The model may be written as $y_{\mathrm{r}\mathrm{c}k} = \beta_0 + \beta_1 x_{\mathrm{r}\mathrm{c}k} + \xi_{\mathrm{r}} + \eta_{\mathrm{c}} + \epsilon_{\mathrm{r}\mathrm{c}k}$, and includes $\xi_{\mathrm{r}}/\eta_{\mathrm{c}}$ as receiver/sender country effects, and k as the replication index as in the previous example. This is also known as a"random effects two-way layout" (Lehmann and Casella, 2006, Example 5.2), and standard methods require normality assumptions similar to the one-way layout of Example 3.*

Under invariant inference, we can consider the following error structure.

$$\varepsilon_i \equiv \varepsilon_{\mathrm{rc}k} = \xi_{\mathrm{r}} + \eta_{\mathrm{c}} + \epsilon_{\mathrm{rc}k}, \quad i = (\mathrm{r}, \mathrm{c}, k), \tag{11}$$

where $\eta = (\eta_{\mathrm{c}})$, $\xi = (\xi_{\mathrm{r}})$ and $\epsilon = (\epsilon_{\mathrm{rc}k})$ are mutually independent, and ϵ are i.i.d.

(a) (Generic 2-way). If $\eta_{\mathrm{c}}, \xi_{\mathrm{r}}$ are exchangeable, then ε have a 'partial exchangeability' property (Aldous, 1981) in the following sense. Suppose we arrange the errors in a 2-dim array by (r, c), such that cell (r, c) contains $\{\varepsilon_{\mathrm{rc}k} : k = 1, \ldots, K\}$. Then, this array is invariant to any permutation of its rows, or columns, or any permutation within cells. Thus, (2) holds with $\mathcal{G}_n$ as the (sub)group of row-wise, column-wise or cell-wise permutations in this array representation.

(b) (Dyadic data). If, in addition, $\eta = \xi$ and the data are dyadic (one observation per pair), then (2) holds with $\mathcal{G}_n$ as in (a) but the row-wise and column-wise permutations must be coupled.

In Theorems 9 and 11 we analyze these structures, and show that Condition (6) is satisfied under generally mild leverage conditions on X, ε, while also allowing for large p. ■

Example 5 (Panel data)**.** *In some financial applications, y measures a firm's stock returns and x denotes estimates of the firm's value (e.g., "book-to-market" ratios). The data have a two-way clustering structure as in Example 4, but there are no repeated measurements ($K = 1$) and $\eta_{\mathrm{c}} = \eta_t$ now denotes time trends. The error structure may be written as*

$$\varepsilon_i \equiv \varepsilon_{\mathrm{r}t} = \xi_{\mathrm{r}} + \eta_t + \epsilon_{\mathrm{r}t}, \quad i = (\mathrm{r}, t). \tag{12}$$

In most applications, time trends cannot be considered i.i.d. or exchangeable, and so the concept of partial exchangeability described in (a) and (b) of Example 4 is not applicable. One approach is to de-trend the regression model by centering the data (y, X) at every time t. This produces the errors, $\varepsilon'_{\mathrm{r}t} = \xi_r - \bar{\xi} + \epsilon_{\mathrm{r}t} - \bar{\epsilon}_{.t}$, where $\bar{\xi} = (1/N)\sum_{\mathrm{r}} \xi_{\mathrm{r}}, \bar{\epsilon}_{.t} = (1/N)\sum_{\mathrm{r}} \epsilon_{\mathrm{r}t}$ denote cross-section averages at t across all N firms. The transformed errors, ε', are exchangeable wrt r whenever ξ_r are exchangeable. Now, we may apply the invariance described in Example 4(a), but only using row-wise permutations. In Theorem 10, we show that Condition (6) is satisfied, under this setting, as long as the number of firms, N, grows at a faster rate than a leverage quantity of X, ε. Again, our results allow for large p. ■

These examples highlight an appealing feature of the residual randomization method. As every one of these examples corresponds to a unique invariance, $\mathcal{G}_n$, we can employ the exact same procedure (formally defined in Section 2) for every example, each time using the appropriate $\mathcal{G}_n$ in a "plug-and-play" fashion. The problem structure therefore dictates the analysis in a simple and unified manner. In contrast, classical inference is built upon the i.i.d. assumption, and so standard methods need to be modified whenever the actual problem structure is different. This has led to a large number of robust standard error techniques in OLS inference, and a similar number of bootstrap variants, which all too frequently confounds applied researchers.

1.2 Related work and Contributions

The idea of statistical inference based on invariance assumptions can be traced back to, at least, the works of Fisher (1935) and Pitman (1937). Fisher (1935), in particular, proposed several permutation tests for nonparametric significance testing in randomized experiments. This type of tests are robust and valid in finite samples, and so they have been widely adopted in experimental design and causal inference (Cox and Hinkley, 1979, Ch. 6); (Rubin, 1980; Kempthorne, 1992; Ernst, 2004; Higgins, 2004; Hinkelmann and Kempthorne, 2007; Edgington and Onghena, 2007; Gerber and Green, 2012; Good, 2013); (Rosenbaum, 2002, Ch. 2); (Imbens and Rubin, 2015, Ch. 5). These disparate methods were also brought under a common, rigorous framework —collectively known as the *randomization method*— in the seminal work of (Lehmann and Romano, 2005, Ch. 15).

Extending these ideas to observational settings was initiated by Freedman and Lane (1983b,a). These authors started from a criticism of classical inference, and its reliance on what they thought as a poorly defined concept of infinite population. Freedman and Lane instead advocated the use of data invariances as the basis of inference —hence, invariant inference— even in observational studies, a position that we adopt in this paper. Perhaps as a twist of luck, with the advent of the bootstrap method, Freedman and his collaborators focused on developing bootstrap versions of this idea (Freedman, 1981; Bickel et al., 1981; Bickel and Freedman, 1983; Freedman and Peters, 1984; Peters and Freedman, 1984), and did not consider more complex invariances beyond exchangeability. In another line of work, researchers have considered the use of permutation tests for inference in regression models, or the closely related task of testing "weak nulls" (Neuhaus, 1993; Janssen, 1997; Anderson and Robinson, 2001; Chung and Romano, 2013; DiCiccio and Romano, 2017; Ding, 2017; Zhao and Ding, 2021). A key insight of this literature is that the permutation distribution of a correctly studentized test statistic approximates, in the limit, the true sampling distribution. This is generally interpreted as a robustness property of randomization tests, where the tests can be both finite-sample valid for a "sharp null" and asymptotically valid for a weak null.

Our paper shares the common goal of invariant inference in regression models as prior work, but it differs in two crucial ways. The first difference is technical but it has important practical ramifications. While prior work has largely focused on an asymptotic analysis of randomization tests, our main analysis here is nonasymptotic. Our results are sharp enough to characterize the finite-sample robustness properties of our test in Theorem 2, and these properties also carry over to the linear setting through Theorems 4-11. Our paper thus contributes to what appears to be a limited list of results on the finite-sample robustness of approximation randomization tests. Canay et al. (2017) have also studied approximate randomization tests, but their setup is different as it assumes a test statistic that, in the limit, satisfies an invariance such as (2). Their analysis was asymptotic, and so our results could be considered complementary. Chernozhukov et al. (2021) also derived finite-sample bounds on a residual-based randomization test. Their analysis also showed that consistency of $\hat{\beta}_n$ is not necessary for their test under certain regularity conditions. The key differences between our setting and (Chernozhukov et al., 2021) is that they worked only on panel data, under a single choice of $\mathcal{G}_n$ (permutations), and a linear test statistic.

This last point summarizes a second key contribution of this paper. While earlier work has largely focused on permutation tests, this paper considers a large collection of invariance structures, as illustrated in Examples 1-5. For the linear model, in particular, Theorems 4-11 derive simple, interpretable conditions of validity according to the structure encoded in $\mathcal{G}_n$. This allows us to employ the exact same procedure (5) in various data settings, each time applying the appropriate $\mathcal{G}_n$ in a "plug-and-play" fashion. Moreover, the technical foundation of these results (Lemmas 9 and 10 in the supplement) can be re-used to analyze other invariance structures in future work.

As a practical matter, our main procedure (5) sometimes leads to resampling schemes that have an almost identical bootstrap counterpart. Our paper therefore contributes to the general bootstrap literature as well. For instance, when $\mathcal{G}_n$ denotes permutations, our method resamples residuals without replacement but is otherwise identical to the residual bootstrap (Freedman, 1981). This bootstrap variant is asymptotically valid when ε_i are i.i.d. with constant variance, and $X^\top X/n$ converges; see Assumptions (1.3) & (1.4) in (Freedman, 1981). Our analysis proves validity under much simpler conditions because Condition (6) does not depend on the asymptotics of $\hat{\beta}_n$. We make similar comparisons with other bootstrap variants throughout the paper. However, we caution against a direct comparison between bootstrap and residual randomization. These methods rely on fundamentally different assumptions, and no framework is a special case of the other. In a sense, bootstrap and residual randomization could be considered complementary.

Indeed, there are settings where procedure (5) leads to novel resampling schemes and theory. One example is the setting with one-way clustered errors. In this setting, "cluster wild bootstrap" (Cameron and Miller, 2015) can be valid asymptotically with a finite number of clusters, but under a somewhat strict homogeneity condition on X (Canay et al., 2018). In Section 3.4, we show that this condition is not required when the errors are exchangeable within clusters. Residual randomization suggests a resampling procedure that simply permutes the residuals in clusters. This procedure is valid in a wide range of regimes, including a finite number of clusters, heavy tailed data and large p (Theorem 7). Another important example is the setting with doubly clustered errors of Example 4. Recent work has developed important bootstrap variants for this setting (Cameron et al., 2011; Menzel, 2021; Davezies et al., 2018; MacKinnon et al., 2021), but these methods require that both clusters increase in size. This may be unrealistic in practice. The randomization method suggests a new resampling procedure that permutes residuals in a manner mimicking partial exchangeability as described in Example 4(a). In Theorem 9, we show that this procedure is asymptotically valid under mild conditions even with finite clusters. In Section 5.2, we apply our method in a model linking trade flows to membership in a currency union. We show that some results in prior work do not hold up under a partial exchangeability structure.

Finally, in Section 6, we discuss extensions of our main methodology. These extensions include results for quadratic test statistics, which are useful in hypotheses tests of significance for a subset of parameters. In Section 6.2, we discuss a novel concept of 'reflection symmetry' around the time axis, which may be used under autocorrelated errors. In a simulated study, the residual randomization test based on reflection symmetry performs favorably against standard "HAC" robust errors.

2 Main Method and Results

Our goal is to test a partial null hypothesis, H_0, on parameters β of model (1). Let T_n be a test statistic such that $T_n = t_n(\varepsilon)$, whenever the null hypothesis holds. This condition is satisfied for many test problems in practice. For example, in the linear setting of Section 3, we use $T_n = a'\hat{\beta}_n - a_0$, where $\hat{\beta}_n$ is the OLS estimator and $H_0 : a'\beta = a_0$ is the null being tested. Then, $T_n = a'(X^\top X)^{-1}X^\top \varepsilon = t_n(\varepsilon)$, under the null. We list more examples of this kind in Appendix A.

To apply the residual randomization method we need the critical value in (4). This enumerates the entire set of invariances, $\mathcal{G}_n$, which can be computationally hard. A more practical approach is to use $m > 0$ uniformly sampled transformations from $\mathcal{G}_n$. Let $G_r \sim \text{Unif}(\mathcal{G}_n)$ i.i.d., for $r = 1 \ldots, m$, and define $S_{n,\varepsilon} = \{t_n(\varepsilon)\} \cup \{t_n(G_r\varepsilon) : r = 1, \ldots, m\}$ as the corresponding set of test statistic values using the true errors. Let $c_{n,\alpha}(S) = \inf\{z \in \mathbb{R} : \frac{1}{|S|}\sum_{s\in S} \mathbb{1}\{s \leq z\} \geq 1 - \alpha\}$ be a modified critical value function, where S is a set of scalars. Define $\phi_n^*(y, X) = \mathbb{1}\big\{T_n > c_{n,\alpha}(S_{n,\varepsilon})\big\}$ as the one-sided idealized test. In Appendix B.2, we prove that this test is valid in finite samples, for any $n, m > 0$. Of course, this test remains infeasible because it uses the true unobserved errors.

To construct the feasible test, we simply plug in the residuals in the set of randomized values. That is, we define $S_{n,\widehat{\varepsilon}} = \{T_n\} \cup \{t_n(G_r\widehat{\varepsilon}) : r = 1, \ldots, m\}$ and

$$\phi_n(y, X) = \mathbb{1}\big\{T_n > c_{n,\alpha}(S_{n,\widehat{\varepsilon}})\big\}. \tag{13}$$

As usual, a two-sided test can be defined as a combination of two one-sided tests at level $\alpha/2$. We are now ready to define the main residual randomization procedure studied in this paper.

Procedure 1: Residual randomization method for invariant inference

1. Compute the test statistic, T_n, and $\widehat{\varepsilon}_i = y_i - f(x_i, \hat{\beta}_n)$, the residuals from some estimator $\hat{\beta}_n$.
2. For fixed $m > 0$, sample $G_r \sim \text{Unif}(\mathcal{G}_n)$, $r = 1, \ldots, m$, i.i.d. Compute $S_{n,\widehat{\varepsilon}}$ as above.
3. Use the one-sided test (13), or a derivative two-sided test.
4. **Inference.** To do inference on β_j, for some $j \in \{1, \ldots, p\}$, we can invert the test. Let $d_{j,\alpha}(b) \in \{0, 1\}$ denote the test decision for the null $H_0 : \beta_j = b$ at level α, and define:

$$\text{CI}_j = \big\{x \in \mathbb{R} : \min(L_{j,\alpha}) \leq x \leq \max(L_{j,\alpha})\big\}, \text{where } L_{j,\alpha} = \{x \in \mathbb{R} : d_{j,\alpha}(x) = 0\}. \tag{14}$$

   Then, CI_j is the $100(1-\alpha)\%$ residual randomization-based confidence interval for β_j.

Remarks on Procedure 1. The test inversion in step 4 is routinely used in the randomization literature to translate test decisions into inference (Rosenbaum, 2002, Section 2.6.2). While the inversion procedure can be computationally challenging, the problem is also "embarrassingly parallelizable" as all tests can be executed independently. In the empirical evaluations of Sections 4-5, we employ a Slurm-based computing cluster when necessary. ■

Our theoretical analysis of Procedure 1 requires two assumptions throughout the paper:

$$X \text{ is non-stochastic, and } \mathbb{E}(\varepsilon), \mathbf{V}_\varepsilon = \mathbb{E}(\varepsilon\varepsilon') \text{ are both finite.} \tag{A1}$$

$$\gamma_{0n} = \max_{\mathrm{g},\mathrm{g}' \in \mathcal{G}_n, \mathrm{g} \neq \mathrm{g}'} P\big\{t_n(\mathrm{g}\varepsilon) = t_n(\mathrm{g}'\varepsilon)\big\} = o(1), \text{ and } 1/|\mathcal{G}_n| = o(1). \tag{A2}$$

Assumption (A1) requires that the first two error moments exist. These moments need not converge, and may even be asymptotically unbounded. Assumption (A2) has two parts. First, it requires that the invariant set grows with n, which it typically does. For example, if $\mathcal{G}_n$ is the set of permutations of n elements, then $|\mathcal{G}_n| = n! > (n/e)^n$, by Stirling's approximation. The second part of (A2) aims to preclude pathological cases where the test statistic becomes degenerate. To see how (A2) can fail, suppose $\varepsilon = \varepsilon_0\mathbf{1}$ —i.e., all error components are identical— and $\mathcal{G}_n$ denotes permutations. Then, $\mathrm{g}\varepsilon = \varepsilon_0\mathrm{g}\mathbf{1} = \varepsilon_0\mathbf{1} = \varepsilon$, for any $\mathrm{g} \in \mathcal{G}_n$. That is, all test statistic values are identical, and so the randomization distribution is degenerate. Barring such cases, (A2) is satisfied with $\gamma_{0n} = 0$ whenever the error distribution is continuous.

Under (A1) and (A2), the following limit becomes the key condition for validity of our test (13):

$$\frac{\mathbb{E}[(t_n(G\widehat{\varepsilon}) - t_n(G\varepsilon))^2]}{\mathbb{E}[(t_n(G'\varepsilon) - t_n(G''\varepsilon))^2]} = \zeta_n^2 \to 0, \text{ where } G, G', G'' \sim \mathrm{Unif}(\mathcal{G}_n). \tag{C1}$$

This condition does not require "$\sqrt{n}$-asymptotics" on $\hat{\beta}_n$, or even consistency of the residuals, as generally required in classical inference. It only stipulates that the variation in the error from the feasible test is dominated by the natural variation of the idealized test. This is in line with the results of Chernozhukov et al. (2021), who also showed that consistency of $\hat{\beta}_n$ is not necessary for asymptotic validity of their residual randomization test. The crucial difference between our setting and (Chernozhukov et al., 2021) is that they work exclusively with panel data under a single choice of $\mathcal{G}_n$ (permutations), and a single choice of a linear test statistic.

Theorem 1. *Suppose that Assumptions (A1)-(A2) hold for model (1) with $\mathcal{G}_n$ as the inferential primitive. Under Condition (C1), the rejection probability of the residual randomization test in Procedure 1 satisfies $\mathbb{E}(\phi_n) \leq \alpha + O(1/m) + o(1)$, whenever the null hypothesis holds.*

Remarks on Theorem 1. (a) The proof of the theorem uses the following crucial robustness property of the residual randomization test (13). Because (13) only compares the order of T_n in $S_{n,\widehat{\varepsilon}}$, the decision of the feasible test is *identical* to the decision of the idealized test whenever the maximum approximation error in the feasible test is smaller than the minimum spacing between the test statistic values in the idealized test. Formally, $\phi_n = \phi_n^*$, if $\max_r |t_n(G_r\widehat{\varepsilon}) - t_n(G_r\varepsilon)| < \min_{r \neq r'} |t_n(G_r\varepsilon) - t_n(G_{r'}\varepsilon)|$. The proof then connects these extrema to the random variables $\Delta_n = t_n(G\widehat{\varepsilon}) - t_n(G\varepsilon)$ and $\Lambda_n = t_n(G'\varepsilon) - t_n(G''\varepsilon)$, as they appear in (C1). Intuitively, Δ_n is the "corruption" introduced in the idealized test from using the residuals, and Λ_n describes the spacings in the idealized test. Condition (C1) guarantees that the variation in the spacings of the idealized test can "withstand" the level of corruption introduced from using the residuals.

(b) The m^{-1} term in the theorem is a product of an analysis where the idealized test is corrupted in a worst-case manner. In practice, we "correct" the test decision by a factor $\alpha/[\alpha+(m+1)^{-1}] = 1-O(1/m)$ to obtain an (asymptotically) α-level test. See Appendix F.1 for details. Because m is usually very large, we will ignore this complication in the main text. ■

2.1 Finite sample analysis: Validity and power

In this section, we provide a finite-sample analysis of Procedure 1. To do so, we need some additional assumptions on the test statistic. In particular, let $\bar{\Lambda}_n = \Lambda_n/\sigma_{\Lambda_n}$ be the standardized spacing variable, and $F_{\bar{\Lambda}_n}$ is its cdf. We will assume that this cdf satisfies a γ-Hölder continuity property,

$$F_{\bar{\Lambda}_n}(\epsilon) - F_{\bar{\Lambda}_n}(-\epsilon) = O(\epsilon^\gamma) + O(\gamma_{0n} + 1/|\mathcal{G}_n|), \text{ for some } \gamma > 0, \text{and any sufficiently small } \epsilon > 0. \tag{A3}$$

Theorem 2. *Suppose that Assumptions (A1)-(A3) hold for model (1) with $\mathcal{G}_n$ as the inferential primitive. Then, under Condition (C1), and for any $n > 0$,*

$$\mathbb{E}(\phi_n) \leq \alpha + O(1/m) + O(m\gamma_{0n}) + O(m^2/|\mathcal{G}_n|) + A(\gamma, m)\cdot O\big(\zeta_n^{2\gamma/(2+\gamma)}\big),$$

whenever the null hypothesis holds. Here, $A(\gamma,m) = 8(1+\gamma)m^{(4+\gamma)/(2+\gamma)} = O(m^2)$.

Remarks on Theorem 2. (a) Parameter γ describes the tails of the spacings variable, Λ_n. For instance, if Λ_n is asymptotically normal under regular conditions, then γ is approximately 1. Values of γ smaller than 1 are the result of heavy-tailed data. In Appendix B.3, we numerically estimate γ in a simulation study that includes heavy-tailed data. In our study, γ takes values in the range $[0.3, 1.3]$, and seems to depend more on the tails of ε rather than X.

(b) The rate at which (C1) is satisfied also determines the finite-sample bound, $\zeta_n^{2\gamma/(2+\gamma)}$, on the level of the residual randomization test. This bound is reminiscent of minimax convergence rates in nonparametric regression (Tsybakov, 2009, Section 1.6). If $\bar{\Lambda}_n$ is asymptotically normal, under regularity conditions, then $\zeta_n^2 = 1/n$ and $\gamma = 1$, implying the rate $O(n^{-1/3})$. In that sense, residual randomization is not "first-order correct", and this may be viewed as the trade-off for its robustness. Heavier-tailed data may "slow down" this asymptotic rate even more. ■

Theorem 3. *Suppose that Assumptions (A1)-(A2) and Condition (C1) hold for model (1). Consider a sequence of alternatives, $H_1 : T_n = t_n(\varepsilon) + \tau_n$, and let $\sigma_n^2 = \mathbb{E}[t_n^2(\varepsilon)]$.*

(3i.) Procedure 1 has asymptotic power one against any sequence satisfying $\tau_n/\sigma_n \to \infty$.

(3ii.) Suppose, in addition, that (A3) holds, and there exists a nondecreasing function $q : \mathbb{R}^+ \to \mathbb{R}^+$ such that $F_{\bar{\Lambda}_n}(l) - F_{\bar{\Lambda}_n}(-l) \geq 1 - e^{-q(l)}$, for any $l > 0$. Then, for any $n > 0$,

$$\mathbb{E}(\phi_n) \geq 1 - O(m^2)e^{-q(\tau_n/2\sqrt{2}\sigma_n)} - A(\gamma, m)\cdot O(\zeta_n^{2\gamma/(2+\gamma)}).$$

The expression for the two-sided test uses $|\tau_n|/\sigma_n$, but is otherwise identical.

REMARKS ON THEOREM 3. The residual randomization test is consistent as long as the signal, τ_n, dominates the natural variation in the idealized test. The exact rates will vary depending on the setting. Consider, for example, a linear setting with $H_0 : \beta_j = 0$. Let $H_1 : \beta_j = \tau$ be the alternative, and let $T_n = a'\hat{\beta}_n$ be the test statistic, where $a \in \{0,1\}^p$ and is nonzero only at $a_j = 1$. In this setting, $t_n(\varepsilon) = a'(X^\top X)^{-1}X^\top\varepsilon$ and regular CLT conditions imply that $\text{var}^{1/2}(t_n(\varepsilon)) = O(1/\sqrt{n})$ and $q(l) = O(l^2)$. The power expression would therefore involve the familiar term $1 - e^{-\tau^2 n}$. ■

3 Linear Model

In this section, we specialize the generic model (1) into a linear model,

$$y_i = x_i'\beta + \varepsilon_i. \tag{15}$$

Here, we assume that x_i has one as the first element, and β_0 is the intercept. We focus on testing a linear hypothesis, $H_0 : a'\beta - a_0$, where $||a|| = 1$ without loss of generality. We will use $T_n = a'\hat{\beta}_n - a_0$ as the test statistic, where $\hat{\beta}_n$ is the regular OLS estimator. Under the null, $T_n = t_n(\varepsilon)$, with $t_n(\varepsilon) = a'Q_{\mathrm{x}}\varepsilon$, $Q_{\mathrm{x}} = (X^\top X)^{-1}X^\top$. To ensure that $X^\top X$ is always invertible we also update (A1):

$$X \text{ is non-stochastic, } \lambda_{\min}(X^\top X) > 0, \text{ and } \mathbb{E}(\varepsilon), \mathbf{V}_\varepsilon = \mathbb{E}(\varepsilon\varepsilon') \text{ are both finite.} \tag{A1'}$$

Our analysis will consider two choices for the residuals in step 1 of Procedure 1. One choice is to simply use the regular OLS residuals, $\hat{\varepsilon} = y - X\hat{\beta}_n$. A second choice is to use the restricted OLS residuals, $\widehat{\varepsilon^{\mathrm{o}}} = y - X\hat{\beta}_n^{\mathrm{o}}$, where $\hat{\beta}_n^{\mathrm{o}} = \hat{\beta}_n - (X^\top X)^{-1}a(a'\hat{\beta}_n - a_0)/a'(X^\top X)^{-1}a$ is the constrained OLS estimator calculated under the null hypothesis. The restricted residuals perform better in small samples, and should be favored in practice. However, both methods are asymptotically equivalent, and so we will analyze them both.

3.1 Exchangeable errors

We begin with exchangeability, perhaps the simplest invariance. When the errors are exchangeable, their distribution remains invariant to any permutation of their labels. The inferential primitive may be written as $\mathcal{G}_n = \mathcal{G}_n^{\mathsf{p}} = \left\{\sum_{i=1}^n \mathbf{1}_i\mathbf{1}_{\pi(i)}' : \pi \in \Pi_n\right\}$. Here, $\mathbf{1}_i \in \{0,1\}^n$ and is nonzero only at ith component, and Π_n is the permutations group of n elements.

Theorem 4. *For the linear model (15) let $\varepsilon \overset{\mathrm{d}}{=} \mathrm{g}\varepsilon \mid X$ for all $\mathrm{g} \in \mathcal{G}_n^{\mathsf{p}}$ and $X \in \mathbb{X}^{(n)}$, $n > 0$, such that Assumptions (A1')-(A2) are satisfied. Suppose also that*

(4i.) $p/n = o(1)$ and $a_1 = 0$; or
(4ii.) $p/n = o(1)$ and $\bar{\varepsilon} = 0$ w.p. 1.

Then, Condition (C1) is satisfied under regular OLS residuals, and the resulting residual randomization test in Procedure 1 is asymptotically valid. With restricted OLS residuals, (4ii) is also sufficient for asymptotic validity.

Remarks on Theorem 4. (a) Under exchangeability, the intercept, β_0, cannot be identified. The role of the conditions $a_1 = 0$ or $\bar{\varepsilon} = 0$ is therefore to preclude testing for β_0. The second condition, $\bar{\varepsilon} = 0$, can be achieved by centering the columns of X, which removes the intercept.

(b) The theorem allows certain high-dimensional cases where $p < n$ but $p \to \infty$. No further restrictions are placed on (X, ε). This shows that exchangeability is a "strong assumption", in line with the concept of exchangeability as a "mixture of i.i.d. sequences" from de Finetti's theorem.

(c) Operationally, the residual randomization procedure here repeatedly calculates the test statistic on permuted residuals, as in residual bootstrap (Freedman, 1981). The crucial difference is that the residual bootstrap aims to approximate the full sampling distribution of $\hat{\beta}_n$, and so it needs additional regularity conditions to ensure convergence in distribution of $\hat{\beta}_n$ at an appropriate rate ("$\sqrt{n}$-asymptotics"). This explains why the known theoretical conditions for the residual bootstrap are substantially stricter than Theorem 4, requiring i.i.d. errors with constant variance and a limit for $X^\top X/n$; see Assumptions (1.3) & (1.4) in (Freedman, 1981).

(d) Freedman and Lane (1983a) proposed a bootstrap test for $H_0 : \beta_S = 0$ on a subset S of parameters, which also resamples residuals—see (DiCiccio and Romano, 2017; Zhao and Ding, 2021) for extensions and a discussion. The crucial difference with our method is that these papers assume i.i.d. data and impose conditions on the population distribution. As a result, they cannot easily extend to complex error structures. See also Section 6.1 for an empirical comparison. ■

3.2 Sign symmetric errors

When the errors are sign symmetric their distribution remains invariant to arbitrary sign flips. This form of invariance is useful under heteroskedasticity, as discussed in Example 2. The inferential primitive may be written as $\mathcal{G}_n = \mathcal{G}_n^{\mathsf{s}} = \left\{ \sum_{i=1}^n s_i \mathbf{1}_i \mathbf{1}_i' : s_i = \pm 1 \right\}$.

Theorem 5. *For the linear model in (15) let $\varepsilon \overset{\mathrm{d}}{=} \mathrm{g}\varepsilon \mid X$ for all $\mathrm{g} \in \mathcal{G}_n^{\mathsf{s}}$ and $X \in \mathbb{X}^{(n)}$, $n > 0$, such that Assumptions (A1')-(A2) are satisfied. Let h_{ii} be the i-th diagonal element of the "hat matrix" $\mathbf{H}_{\mathrm{x}} = X(X^\top X)^{-1}X^\top$, and define $\bar{\lambda}_n = \max_{i\in[n]} h_{ii}/(\sum_{j=1}^n h_{jj}/n) = \max_i h_{ii}/(p/n)$. Suppose that*

$$\frac{\max_{i\in[n]} \mathbb{E}(\varepsilon_i^2)}{\min_{j\in[n]} \mathbb{E}(\varepsilon_j^2)}\ \bar{\lambda}_n p/n = o(1).$$

Then, Condition (C1) is satisfied under either regular or restricted OLS residuals, and the residual randomization test in Procedure 1 is asymptotically valid.

Remarks on Theorem 5. (a) The term h_{ii} is the leverage score of datapoint i, and so $\bar{\lambda}_n$ is a global summary of leverage; e.g., if $x_i \sim N(0,1)$ i.i.d. then $\bar{\lambda}_n = O_P(\log n)$ (Embrechts et al., 2013, Sec. 3). The "max / min" term is equal to $\kappa(\mathbf{V}_\varepsilon)$, the condition number of the error variance-covariance matrix. Thus, the theorem allows for certain highly-leveraged or high-dimensional settings. This compares favorably to classical "HAC" errors (White et al., 1980; Eicker et al., 1963), which are known to underperform in small samples, and especially when the design is highly leveraged (MacKinnon and White, 1985; Davidson and Flachaire, 2008; Godfrey and Orme, 2001).

(b) Operationally, the residual randomization test with $\mathcal{G}_n^{\mathsf{s}}$ as the primitive repeatedly calculates the test statistic on residuals with their signs randomly flipped, similar to some variants of the "wild bootstrap" (Wu et al., 1986; Liu et al., 1988; Mammen et al., 1993). Again, the crucial difference is that in the bootstrap framework we need to impose regularity conditions on $\hat{\beta}_n$ to ensure "$\sqrt{n}$-asymptotics". Such conditions are not necessary with residual randomization. ■

3.3 Double invariance: Exchangeability and sign symmetry

When the errors are both exchangeable and sign symmetric, we can write the primitive as $\mathcal{G}_n = \mathcal{G}_n^{\mathsf{p+s}} = \left\{\sum_{i=1}^n s_i \mathbf{1}_i \mathbf{1}_{\pi(i)} : s_i = \pm 1, \pi \in \Pi_n\right\}$. Operationally, the residual randomization test repeatedly calculates the test statistic on permuted residuals with their signs randomly flipped.

Theorem 6. *For the linear model in (15) let $\varepsilon \overset{\mathrm{d}}{=} \mathrm{g}\varepsilon \mid X$ for all $\mathrm{g} \in \mathcal{G}_n^{\mathsf{p+s}}$ and $X \in \mathbb{X}^{(n)}$, $n > 0$, such that Assumptions (A1')-(A2) are satisfied. Suppose also that $p/n = o(1)$. Then, Condition (C1) is satisfied, and the residual randomization test is asymptotically valid under either regular or restricted OLS residuals.*

Remarks on Theorem 6. It is an interesting general question under what conditions two different invariances, which individually produce valid tests, can be combined into one "bigger" valid test. The proof of the theorem discusses some ideas using Theorem 6 as an example. ■

3.4 Clustered errors: one-way clustering

Empirical observations often occur in clusters, such as when student scores are observed by school or classroom. In such settings, it is common to report "clustered standard errors" (Liang and Zeger, 1986; Arellano, 1987; Cameron and Miller, 2015; MacKinnon et al., 2022). This motivates us to consider a family of cluster invariances on the errors. We will not consider the question of whether one should cluster the errors or not—we refer to (Abadie et al., 2017) for such a discussion.

To proceed we need some additional notation. C^n will denote a clustering of $[n] = \{1, \dots, n\}$. In the main text, C^n will be a partition of $[n]$, but this is not necessary for our theory. $J_n = |\mathrm{C}^n|$ denotes the number of clusters in C^n, and $i \in [\mathrm{c}]$ denotes that datapoint i is in cluster $\mathrm{c} \in \{1, \dots, J_n\}$. Let $\mathbf{1}_\mathrm{c} = \sum_{i \in [\mathrm{c}]} \mathbf{1}_i$ be the n-length binary vector that is nonzero only at units in c. We use $n_\mathrm{c} = \mathbf{1}'\mathbf{1}_\mathrm{c} = \sum_{i \in [\mathrm{c}]} 1$ to denote the size of cluster c. Define $\mathbf{I}_\mathrm{c} = \sum_{i \in [\mathrm{c}]} \mathbf{1}_i \mathbf{1}_i'$ and $\mathbf{U}_\mathrm{c} = \mathbf{1}_\mathrm{c}\mathbf{1}_\mathrm{c}'$, and let $s_\mathrm{c}^2(z) = \frac{1}{n_\mathrm{c}-1} z'(\mathbf{I}_\mathrm{c} - \mathbf{U}_\mathrm{c}/n_\mathrm{c})z$ denote the sample variance of $z \in \mathbb{R}^n$ in cluster c.

To facilitate our discussion, we will refer to the cluster error structure from the one-way layout model described in Example 3:

$$\varepsilon_i \equiv \varepsilon_{\mathrm{c}k} = \eta_\mathrm{c} + \epsilon_{\mathrm{c}k}, \quad i = (\mathrm{c}, k). \tag{16}$$

Throughout this section, we will assume that the cluster random effects, $\eta = (\eta_\mathrm{c})$, and the idiosyncratic errors, $\epsilon = (\epsilon_{\mathrm{c}k})$, are mutually independent.

3.4.1 Exchangeability within clusters

We begin with a cluster variant of exchangeability. In (16), this invariance holds whenever η_{c} are exchangeable, and $\epsilon_{\mathrm{c}k}$ are exchangeable within any cluster c. The inferential primitive may be written as $\mathcal{G}_n = \mathcal{G}_n^{\mathsf{p}}(\mathrm{C}^n) = \left\{\sum_{i=1}^n \mathbf{1}_i\mathbf{1}'_{\pi(i)} : \pi \in \Pi(\mathrm{C}^n)\right\}$, where $\Pi(\mathrm{C}^n)$ is the (sub)group of within-cluster permutations; i.e., $i \in [\mathrm{c}] \Rightarrow \pi(i) \in [\mathrm{c}]$ for any $\pi \in \Pi(\mathrm{C}^n)$.

For our analysis, the following cluster variant of leverage will be useful:

$$\bar{\lambda}_1(\mathrm{C}^n) = \frac{\max_{\mathrm{c}\in[J_n]} \ (1/n_{\mathrm{c}}) \sum_{i\in[\mathrm{c}]} h_{ii}}{(1/n)\sum_{j=1}^n h_{jj}}. \tag{17}$$

This quantity plays a role in the asymptotic validity of the test, as shown below.

Theorem 7. *For the linear model in (15) let $\varepsilon \overset{\mathrm{d}}{=} \mathrm{g}\varepsilon \mid X$ for all $\mathrm{g} \in \mathcal{G}_n^{\mathsf{p}}(\mathrm{C}^n)$ and $X \in \mathbb{X}^{(n)}$, $n > 0$, such that Assumptions (A1')-(A2) are satisfied. Suppose X is centered within each cluster, and*

$$\frac{\max_{\mathrm{c}\in[J_n]} \mathbb{E}[s_{\mathrm{c}}^2(\varepsilon)]}{\min_{\mathrm{c}'\in[J_n]} \mathbb{E}[s_{\mathrm{c}'}^2(\varepsilon)]} \ \bar{\lambda}_1(\mathrm{C}^n) \ p/n = o(1).$$

Then, Condition (C1) is satisfied and the residual randomization test in Procedure 1 is asymptotically valid under either regular or restricted OLS residuals.

Remarks on Therorem 7. As with simple invariance structures, the residual randomization test is valid as long as the problem is not too highly leveraged (in the cluster sense), while also allowing $p \to \infty$. The condition of centered X is necessary because cluster effects are not identifiable under only cluster exchangeability. In the special case with $J_n = 1$ (one cluster), the overall condition in the theorem reduces to $p/n = o(1)$, which exactly matches result (4ii) of Theorem 4.

(b) Procedure 1 with $\mathcal{G}_n^{\mathsf{p}}(\mathrm{C}^n)$ as the primitive repeatedly calculates the test statistic under permutations of the residuals within the clusters in C^n. Operationally, this is similar to some variants of the randomized cluster bootstrap (Davison and Hinkley, 1997). See (Field and Welsh, 2007) for a related discussion. ■

3.4.2 Sign symmetry across clusters

We continue with a cluster variant of sign symmetry. In (16), this invariance holds whenever η_{c} are sign symmetric, and $\epsilon_{\mathrm{c}k}$ are symmetric within each cluster c. This accounts for certain types of heterogeneity within clusters. The inferential primitive is $\mathcal{G}_n = \mathcal{G}_n^{\mathsf{s}}(\mathrm{C}^n) = \left\{\sum_{\mathrm{c}\in[J_n]} s_{\mathrm{c}} \sum_{i\in[\mathrm{c}]} \mathbf{1}_i\mathbf{1}'_i : s_{\mathrm{c}} = \pm 1\right\}$.

To analyze this procedure, we will need the following variant of cluster-level leverage:

$$\bar{\lambda}_2(\mathrm{C}^n) = \frac{\max_{\mathrm{c}\in[J_n]} \ ||X^\top \mathbf{I}_{\mathrm{c}} X||_F^2 \ /n_{\mathrm{c}}^2}{||X^\top X||_F^2 \ /n^2}. \tag{18}$$

The technical difference from (17) is that $\bar{\lambda}_2(\mathrm{C}^n)$ describes the cluster leverage structure on the fourth moment of X, while $\bar{\lambda}_1(\mathrm{C}^n)$ describes the leverage structure on the second moment.

This setting is much richer than the settings we have considered so far. Our most general result imposes mild leverage conditions on (X, ε), but requires the number of clusters to grow ($J_n \to \infty$). An interesting situation arises when the covariates have a cluster uniformity property in a sense described in the theorem below. In this case, the residual randomization test using restricted OLS residuals is in fact *exact* in finite samples. When the uniformity property holds in the limit, then the same test is valid asymptotically even when the number of clusters stays finite ($J_n < \infty$). The following theorem summarizes our results for these three scenarios.

Theorem 8. *For the linear model in (15) let $\varepsilon \overset{\mathrm{d}}{=} \mathrm{g}\varepsilon \mid X$ for all $\mathrm{g} \in \mathcal{G}_n^{\mathrm{s}}(\mathrm{C}^n)$ and $X \in \mathbb{X}^{(n)}$, $n > 0$, under Assumptions (A1')-(A2). Let $\kappa(A)$ be the condition number of A, and define $\kappa_n = \kappa(X^\top X)$.*

*(8i.) ($J_n \to \infty$). Suppose that $p^3 \kappa_n^2 \kappa(Q_{\mathrm{x}} \mathbf{V}_\varepsilon Q_{\mathrm{x}}^\top)\ \bar{\lambda}_2(\mathrm{C}^n) \sum_{\mathrm{c}=1}^{J_n} n_{\mathrm{c}}^2/n^2 = o(1)$.
Then, Condition (C1) is satisfied, and the residual randomization test is asymptotically valid under either regular or restricted OLS.*

(8ii.) ($J_n < \infty$). Suppose that for every cluster $\mathrm{c} = 1, \ldots, J_n$, there exist scalars r_{c} for which

$$(X^\top X)^{-1}(X^\top \mathbf{I}_{\mathrm{c}} X) = r_{\mathrm{c}} \mathbf{I} + E_{n,\mathrm{c}}, \tag{CU}$$

such that $\max_{\mathrm{c}} ||E_{n,\mathrm{c}}|| = o(1)$ and $pJ_n \kappa(Q_{\mathrm{x}} \mathbf{V}_\varepsilon Q_{\mathrm{x}}^\top) = O(1)$. Then, the residual randomization test using restricted OLS residuals is valid asymptotically.

(8iii.) (Any J_n, any $n > 0$). The residual randomization test is exact in finite samples whenever (CU) holds in the sample, such that $E_{n,\mathrm{c}} = 0$ for every cluster c and any $n > 0$.

Remarks on Theorem 8. (a) The main condition in (8i) is $\sum_{\mathrm{c}} n_{\mathrm{c}}^2/n^2 \to 0$. This requires an increasing number of clusters, and also that no cluster dominates the others in terms of size. To work under a finite number of clusters, our analysis needs a uniformity condition, namely (CU). The term $\kappa(Q_{\mathrm{x}} \mathbf{V}_\varepsilon Q_{\mathrm{x}}^\top)$ in (8i) and (8ii) cannot be simplified without additional assumptions. For example, if $\mathbf{V}_\varepsilon$ has a block diagonal structure, then $\kappa(Q_{\mathrm{x}} \mathbf{V}_\varepsilon Q_{\mathrm{x}}^\top) \leq \kappa_n \max_{\mathrm{c},\mathrm{c}'} s_{\mathrm{c}}^2(\varepsilon)/s_{\mathrm{c}'}^2(\varepsilon)$, which is exactly the error term that appears in Theorem 7.

(b) Operationally, the residual randomization procedure repeatedly calculates the test statistic on residuals that have their signs randomly jointly flipped on the cluster level. This is similar to the "cluster wild bootstrap" (Cameron et al., 2008; Cameron and Miller, 2015). Several authors (Cameron et al., 2008; Cameron and Miller, 2015; Imbens and Kolesar, 2016) have argued that this bootstrap variant improves upon the sometimes poor finite-sample performance of standard asymptotic "cluster-robust inference" (Moulton, 1986; Arellano, 1987; White, 1984; Hansen, 2007; Carter et al., 2017; Ibragimov and Müller, 2010, 2016; Bester et al., 2011; Conley and Taber, 2011). Another point of similarity is condition (CU), which is equivalent to the uniformity condition required in the best known results on the validity of cluster wild bootstrap with a finite number of clusters (Canay et al., 2018, Assumption 2iii). As with previous instances, the bootstrap approach requires additional regularity conditions for $\sqrt{n}$-asymptotic convergence of $\hat{\beta}_n$, which are not necessary in residual randomization.

(c) If the uniformity condition (CU) holds in the sample, then the randomization test is in fact exact. This condition is, of course, strong but it does not preclude us from choosing a suitable clustering C^n to satisfy the condition. In Section E.1 of the supplement, we analyze one striking instance of the Behrens–Fisher problem based on a simulation setup by Angrist and Pischke (2009). There is no analog of finite-sample exactness in the bootstrap literature to our best knowledge.

(d) As in Section 3.3, we could combine the two cluster invariants of this section, $\mathcal{G}_n^{\mathsf{p}}(\mathrm{C}^n)$ and $\mathcal{G}_n^{\mathsf{s}}(\mathrm{C}^n)$, into a larger invariant, say $\mathcal{G}_n^{\mathsf{p+s}}(\mathrm{C}^n)$. This test would be valid as long as either the conditions of Theorem 7 or the conditions of Theorem 8 hold, thus combining the best of both worlds. This ability to "mix-and-match" different invariance structures highlights the flexibility of the residual randomization method. ■

3.5 Two-way clustered errors

In this section, we consider doubly clustered error structures. To facilitate our discussion, we will refer to the error structure from the two-way layout model of Example 4, extended to include a common random effect (μ):

$$\varepsilon_i \equiv \varepsilon_{\mathrm{rc}k} = \mu + \xi_{\mathrm{r}} + \eta_{\mathrm{c}} + \epsilon_{\mathrm{rc}k}, \quad i = (\mathrm{r}, \mathrm{c}, k),\ \mathrm{r} \in \mathrm{R}^n,\ \mathrm{c} \in \mathrm{C}^n. \tag{19}$$

Here, R^n denotes the "row" cluster, and C^n the "column" cluster; k denotes the replication within a "cell" (r, c). As before, $\mu, \eta = (\eta_{\mathrm{c}}), \xi = (\xi_{\mathrm{c}})$ and $\epsilon = (\epsilon_{\mathrm{rc}k})$, are assumed mutually independent. Any structural assumptions on these variables will thus dictate which particular structural invariance on the errors to analyze.

3.5.1 Partial exchangeability

In many settings, the data come as observations on pairs of units, which motivates the use of doubly clustered errors. For example, in international trade data, an observation could be the trade flow between two countries. In an influential paper, Cameron et al. (2011) proposed certain standard error estimates under two-way clustering. Fafchamps and Gubert (2007); Conley (1999); Aronow et al. (2015) have studied sandwich estimators inspired from spatial models (Conley, 1999). The theoretical properties of these procedures, however, have not been well understood. It is also known that these procedures can be numerically unstable, often producing invalid variance estimates.

To address these issues, recent work adopted an invariance concept known as "partial exchangeability", first introduced in the seminal work of Aldous (1981) and Hoover (1979). This structure is particularly challenging for inference. McCullagh (2000) has shown, for example, that no straightforward bootstrap procedure exists even for such simple statistics as the sample average. Recent work has built upon these insights and has developed bootstrap procedures for the sample average (Menzel, 2021), generalized method of moments estimation (Davezies et al., 2018), and linear regression (MacKinnon et al., 2021). Our approach adopts the same concept of partial exchangeability of Aldous (1981); Hoover (1979), but weakens some conditions in earlier work.

In our framework, the formulation in (19) suggests a special type of invariance, which extends the concept of partial exchangeability. Let $\mathcal{E} = (\varepsilon_{\mathrm{rc}k})$ be a set-valued $|\mathrm{R}^n| \times |\mathrm{C}^n|$ matrix, where the (r, c)-element in $\mathcal{E}$ is equal to $\{\varepsilon_{\mathrm{rc}k} : k = 1, \ldots, K\}$, the repeated measurements in cell (r, c). We refer to this matrix as the "$\mathcal{E}$-representation" of ε. Then, the distribution of $\mathcal{E}$ is invariant to any permutation of its rows or columns, and to any permutations within cells, whenever $\eta_\mathrm{c}, \xi_\mathrm{r}$ are exchangeable and $(\epsilon_{\mathrm{rc}k})$ are exchangeable within cells. This invariance is identical to the partial exchangeability of Aldous (1981) without replications. We illustrate the invariance with an example.

Example 6 ($\mathcal{E}$-representation)**.** *Let* $\mathcal{E} = \begin{pmatrix} \{\varepsilon_1, \varepsilon_2\} & \{\varepsilon_5, \varepsilon_6\} \\ \{\varepsilon_3, \varepsilon_4\} & \{\varepsilon_7, \varepsilon_8\} \end{pmatrix} \equiv \begin{pmatrix} \{\varepsilon_{11(1)}, \varepsilon_{11(2)}\} & \{\varepsilon_{12(1)}, \varepsilon_{12(2)}\} \\ \{\varepsilon_{21(1)}, \varepsilon_{21(2)}\} & \{\varepsilon_{22(1)}, \varepsilon_{22(2)}\} \end{pmatrix}$, *where we set* $|\mathrm{R}^n| = |\mathrm{C}^n| = K = 2$*. Then, the following invariances hold on* $\mathcal{E}$*:*

$$\mathcal{E} \overset{\mathrm{d}}{=} \begin{pmatrix} \{\varepsilon_{21(1)}, \varepsilon_{21(2)}\} & \{\varepsilon_{22(1)}, \varepsilon_{22(2)}\} \\ \{\varepsilon_{11(1)}, \varepsilon_{11(2)}\} & \{\varepsilon_{12(1)}, \varepsilon_{12(2)}\} \end{pmatrix} \overset{\mathrm{d}}{=} \begin{pmatrix} \{\varepsilon_{22(1)}, \varepsilon_{22(2)}\} & \{\varepsilon_{21(1)}, \varepsilon_{21(2)}\} \\ \{\varepsilon_{12(1)}, \varepsilon_{12(2)}\} & \{\varepsilon_{11(1)}, \varepsilon_{11(2)}\} \end{pmatrix} \overset{\mathrm{d}}{=} \begin{pmatrix} \{\varepsilon_{22(2)}, \varepsilon_{22(1)}\} & \{\varepsilon_{21(1)}, \varepsilon_{21(2)}\} \\ \{\varepsilon_{12(1)}, \varepsilon_{12(2)}\} & \{\varepsilon_{11(2)}, \varepsilon_{11(1)}\} \end{pmatrix}.$$

In the first step, we permuted the rows of $\mathcal{E}$*, and in the second step we permuted its columns. In the third step, we permuted the observations within the diagonal cells of* $\mathcal{E}$*. In our framework, partial exchangeability implies that the distribution of* $\mathcal{E}$ *remains invariant throughout all these steps.*

Formally, let $r(i) \in \mathrm{R}^n$ denote the row cluster of datapoint i, and $c(i) \in \mathrm{C}^n$ its column cluster in the $\mathcal{E}$-array representation; e.g., in Example 6, $r(5) = 1$ and $c(5) = 2$. Let $\Pi(\mathrm{R}^n, \mathrm{C}^n)$ be the (sub)group of permutations of $\{1, \ldots, n\}$ that preserves the rows/columns of the elements in $\mathcal{E}$. That is, $r(i) = r(j) \Rightarrow r(\pi(i)) = r(\pi(j))$ and $c(i) = c(j) \Rightarrow c(\pi(i)) = c(\pi(j))$ for any $\pi \in \Pi(\mathrm{R}^n, \mathrm{C}^n)$. Then, the invariant may be defined as $\mathcal{G}_n = \mathcal{G}_n^{\mathsf{p}}(\mathrm{R}^n, \mathrm{C}^n) = \left\{ \sum_{i=1}^n \mathbf{1}_i \mathbf{1}'_{\pi(i)} : \pi \in \Pi(\mathrm{R}^n, \mathrm{C}^n) \right\}$.

Theorem 9. *For the linear model in (15) let* $\varepsilon \overset{\mathrm{d}}{=} \mathrm{g}\varepsilon \mid X$ *for all* $\mathrm{g} \in \mathcal{G}_n^{\mathsf{p}}(\mathrm{R}^n, \mathrm{C}^n)$ *and* $X \in \mathbb{X}^{(n)}$, $n > 0$, *such that Assumptions (A1')-(A2) are satisfied. Suppose that* X *has been centered, and*

$$p^4 \kappa_n^2 \ \kappa(Q_\mathrm{x} \mathbf{V}_\varepsilon Q_\mathrm{x}^\top) \ \big(1/|\mathrm{R}^n| + 1/|\mathrm{C}^n|\big) = o(1).$$

Then, Condition (C1) is satisfied, and the residual randomization test is asymptotically valid under either regular or restricted OLS. If $|\mathrm{R}^n| < \infty$ *(or* $|\mathrm{C}^n| < \infty$*), then the test is still valid asymptotically if* X *is centered row-wise (or column-wise).*

Remarks on Theorem 9. (a) The residual randomization test is valid as long as the number of row and column clusters grows, and the usual leverage conditions hold. This offers an appealing alternative to standard mixed models (McCulloch and Searle, 2004), which require a linear specification like (19) under normality assumptions. Moreover, residual randomization allows certain high-dimensional settings with large p. The procedure remains valid even when the number of clusters in a given dimension stays finite, as long as the covariates have been centered along that dimension. To our best knowledge, this kind of result is novel in the literature of partially exchangeable errors. The result can also be extended to multiway clustering simply by adding the corresponding terms in the main condition. See the proof of the theorem for details.

(b) Recent papers on bootstrapping partially exchangeable data impose smoothness conditions (Menzel, 2021, A.2.2); (Davezies et al., 2018, A.3), or independence structures in the DGP (Davezies et al., 2018, A.1.2); (MacKinnon et al., 2021, A.4), which are not necessary under residual randomization. Crucially, all these methods operate under asymptotic regimes where both clusters increase in size, whereas residual randomization can work with finite clusters. Another important distinction between these methods and residual randomization is how they can handle missing data—see Section 3.5.4 that follows.

(c) Operationally, the residual randomization method first transforms the residuals in the "$\mathcal{E}$-representation" illustrated in Example 6, randomly permutes the residuals row-wise, column-wise, or within cells, and then transforms back to vector format. This procedure is related to the "quadratic assignment procedure" (QAP) (Hubert, 1986; Hubert and Schultz, 1976; Krackhardt, 1988). This procedure is used in network analysis to test independence in network data (Christakis and Fowler, 2013). In a regression model, QAP permutes the rows and vectors of the dependent variable and refits the model. As such, a QAP test acts as an F-test, but sometimes is misused to assess significance of individual parameters. More appropriate for significance is the residual randomization test based on $\mathcal{G}_n^{\mathsf{p}}(\mathrm{R}^n, \mathrm{C}^n)$. Dekker et al. (2007) have made a similar point based on their adaptation of the residual method of Freedman and Lane (1983a). See also (Snijders, 2011) for a comprehensive review of similar statistical problems in social network science.

(d) The linear form in (19) is not necessary. When there are no replications, Aldous's result implies that if the errors are partially exchangeable then we can write $\varepsilon_{\mathrm{rc}} = q(\mu, \eta_{\mathrm{c}}, \xi_{\mathrm{r}}, \epsilon_{\mathrm{rc}})$, for some measurable function q. This representation implies the same type of row-wise and column-wise permutation invariance captured by $\mathcal{G}_n^{\mathsf{p}}(\mathrm{R}^n, \mathrm{C}^n)$, and illustrated in Example 6. ■

3.5.2 Panel data

A special case of a two-way clustered structure appears in panel data. In this setting, the column cluster indicates time and there is usually only one observation per cell. Thus, a datapoint may be indexed by unit (r) —e.g., a firm— and time (t). We may write the model as in Example 5,

$$y_{\mathrm{r}t} = x_{\mathrm{r}t}'\beta + \eta_{\mathrm{r}} + \xi_t + \epsilon_{\mathrm{r}t}, \quad \mathrm{r} = 1, \ldots, N, \quad t = 1, \ldots, T, \;\; n = NT. \tag{20}$$

In this formulation, η_{r} may be viewed as firm effects (fixed or random), ξ_t as time trends, and $\epsilon_{\mathrm{r}t}$ as random noise. The error component is $\varepsilon_i \equiv \varepsilon_{\mathrm{r}t} = \eta_{\mathrm{r}} + \xi_t + \epsilon_{\mathrm{r}t}$. For simplicity, let us assume that $\eta = (\eta_r)$ and $\epsilon = (\epsilon_{\mathrm{r}t})$ are identically distributed and mutually independent—simple exchangeability would also work, as shown in previous examples. If we can also assume that the time trends, $\xi = (\xi_t)$, are identically distributed, then we could immediately apply the results of Section 3.5.1 , and use a randomization test that doubly permutes residuals across firms and across time. However, this assumption is unrealistic here because t denotes time. One simple approach is to de-trend the data, and only exploit a partial exchangeability invariance on the row dimension of the "$\mathcal{E}$-representation" of the errors.

Specifically, if we average out the time trends, we obtain the model $y_{\mathrm{rt}} - \bar{y}_{\mathrm{r.}} = (x_{\mathrm{rt}} - \bar{x}_{\mathrm{r.}})'\beta + (\eta_{\mathrm{r}} - \bar{\eta}) + (\epsilon_{\mathrm{rt}} - \bar{\epsilon}_{\mathrm{r.}})$, and the resulting error structure is

$$\varepsilon'_{\mathrm{rt}} = (\eta_{\mathrm{r}} - \bar{\eta}) + (\epsilon_{\mathrm{rt}} - \bar{\epsilon}_{\mathrm{r.}}).$$

The de-trended errors, ε', are now exchangeable wrt r. One idea is thus to perform a residual randomization test by permuting the residuals row-wise in their "$\mathcal{E}$-representation"; i.e., permute the de-trended residuals "across firms". We illustrate this idea with an example.

Example 7. *Consider observations on 2 firms and T time periods. Then, the de-trended errors, $\varepsilon'_{\mathrm{rt}} = \varepsilon_{\mathrm{rt}} - \bar{\varepsilon}_{\mathrm{r.}}$ are exchangeable wrt* r*. That is, $\mathcal{E}' = \begin{pmatrix} \varepsilon'_{11} & \varepsilon'_{12} & \dots & \varepsilon'_{1T} \\ \varepsilon'_{21} & \varepsilon'_{22} & \dots & \varepsilon'_{2T} \end{pmatrix} \overset{\mathrm{d}}{=} \begin{pmatrix} \varepsilon'_{21} & \varepsilon'_{22} & \dots & \varepsilon'_{2T} \\ \varepsilon'_{11} & \varepsilon'_{12} & \dots & \varepsilon'_{1T} \end{pmatrix}$, where we simply permuted between the two rows.*

Formally, let $\Pi(\mathrm{R}^n)$ be the set of permutations of $\{1, \dots, n\}$ that preserves the rows of the elements in $\mathcal{E}$; i.e., $r(i) = r(j) \Rightarrow r(\pi(i)) = r(\pi(j))$ for any $\pi \in \Pi_n$. Then, $\Pi(\mathrm{R}^n)$ is a subgroup of Π_n, and the invariant may be written as $\mathcal{G}_n = \mathcal{G}_n^{\mathsf{p}}(\mathrm{R}^n) = \left\{ \sum_{i=1}^n \mathbf{1}_i \mathbf{1}'_{\pi(i)} : \pi \in \Pi(\mathrm{R}^n) \right\}$.

Theorem 10. *For the linear model in (15) let $\varepsilon \overset{\mathrm{d}}{=} \mathrm{g}\varepsilon \mid X$ for all $\mathrm{g} \in \mathcal{G}_n^{\mathsf{p}}(\mathrm{R}^n)$ and $X \in \mathbb{X}^{(n)}$, $n > 0$, such that Assumptions (A1')-(A2) are satisfied. Define $\bar{\lambda}_{n,2} = \max_{i \in [n]} \frac{||x_{i.}||^2}{||X||_F^2/n}$, and suppose that the model has been de-trended, and that*

$$p^4 \kappa_n^2 \; \kappa(Q_{\mathrm{x}} \mathbf{V}_{\varepsilon} Q_{\mathrm{x}}^{\top}) \; \bar{\lambda}_{n,2}/N = o(1).$$

Then, Condition (C1) is satisfied, and the residual randomization test in Procedure 1 is asymptotically valid under either regular or restricted OLS.

Remarks on Theorem 10. (a) The leverage term, $\bar{\lambda}_{n,2}$, is similar to $\bar{\lambda}_n$ in Theorem 5. In fact, it is straightforward to show that $\bar{\lambda}_{n,2}/\bar{\lambda}_n \in [1/\kappa_n, \kappa_n]$. The leverage condition precludes that one single "firm-time" observation can have an out-sized leverage score compared to other observations. As in Theorem 8, the term $\kappa(Q_{\mathrm{x}} \mathbf{V}_{\varepsilon} Q_{\mathrm{x}}^{\top})$ cannot be simplified without additional assumptions. For instance, if $\mathbf{V}_{\varepsilon}$ is diagonal, then $\kappa(Q_{\mathrm{x}} \mathbf{V}_{\varepsilon} Q_{\mathrm{x}}^{\top}) \leq \kappa_n$. Thus, the overall condition in the theorem requires that the design is not too leveraged, and that the row dimension, N, (e.g., number of firms) increases. Again, this allows for large p. ■

3.5.3 Dyadic exchangeability

In many applications with two-way clustered data, there is an additional dyadic structure, where the rows and columns correspond to the same type of clustering, and there is only one observation per cell. An example of this structure is cross-sectional trade data where only the aggregate tradeflow between a pair of countries is observed. In this setting, there is no distinction between rows and columns, and the invariance concept needs to be adapted accordingly. We refer to this concept as dyadic partial exchangeability, and formally define it under the following conditions.

With respect to structure (19), it holds:

(i) $\mathrm{R}^n = \mathrm{C}^n$ and $K = 1$. Let $|\mathrm{R}^n| = |\mathrm{C}^n| = N$ be the number of the common clusters.

(ii) $y_{\mathrm{rc}} = y_{\mathrm{cr}}$, and so there are $n = N(N-1)/2$ total observations.

(iii) $\xi = \eta$ such that $\varepsilon_i \equiv \varepsilon_{\mathrm{rc}} = q(\mu, \eta_{\mathrm{r}}, \eta_{\mathrm{c}}, \epsilon_{\mathrm{rc}})$, and q is symmetric in the η-arguments.

(iv) The diagonal values, (ε_{jj}), are undefined (or they are identically 0).

These conditions put some constraints in the array representation of the errors, $\mathcal{E} = (\varepsilon_{\mathrm{rc}})$, described in Section 3.5.1 and Example 6. First, $\mathcal{E}$ needs to be symmetric with an empty diagonal, due to (ii)-(iv). Second, due to (ii), the errors can be arranged only in a half-part of $\mathcal{E}$, say, in its lower triangular part. Finally, due to (iii) we cannot permute the rows and columns of $\mathcal{E}$ independently, but these permutations need to be coupled. We illustrate with an example.

Example 8. *Let $N = 4$ (countries), and $n = 6$. The $\mathcal{E}$-representation of errors is defined as*

$$\mathcal{E} = \begin{pmatrix} * & \varepsilon_1 & \varepsilon_2 & \varepsilon_3 \\ \varepsilon_1 & * & \varepsilon_4 & \varepsilon_5 \\ \varepsilon_2 & \varepsilon_4 & * & \varepsilon_6 \\ \varepsilon_3 & \varepsilon_5 & \varepsilon_6 & * \end{pmatrix}. \textit{ Then, } \mathcal{E}' = \begin{pmatrix} * & \varepsilon_5 & \varepsilon_3 & \varepsilon_6 \\ \varepsilon_5 & * & \varepsilon_1 & \varepsilon_4 \\ \varepsilon_3 & \varepsilon_1 & * & \varepsilon_2 \\ \varepsilon_6 & \varepsilon_4 & \varepsilon_2 & * \end{pmatrix} \overset{\mathrm{d}}{=} \mathcal{E}.$$

Here, $\mathcal{E}'$ was obtained by applying the same permutation, $(4\ 2\ 1\ 3)$, on both the rows and columns of $\mathcal{E}$. Transforming back to ε implies the invariance: $(\varepsilon_1, \ldots, \varepsilon_6) \overset{\mathrm{d}}{=} (\varepsilon_5, \varepsilon_3, \varepsilon_6, \varepsilon_1, \varepsilon_4, \varepsilon_2)$. We will use $\mathcal{G}_n^{\mathsf{dyad}}$ to denote all permutations of ε induced by such coupled row/column permutations on $\mathcal{E}$.

Theorem 11. *For the linear model in (15) let $\varepsilon \overset{\mathrm{d}}{=} \mathrm{g}\varepsilon \mid X$ for all $\mathrm{g} \in \mathcal{G}_n^{\mathsf{dyad}}$ and $X \in \mathbb{X}^{(n)}$, $n > 0$, such that Assumptions (A1')-(A2) are satisfied. Let δ_{ij} be a binary indicator such that $\delta_{ij} = 1$ only if ε_j is in the same row or column with ε_i in their dyadic $\mathcal{E}$-representation. Let $\boldsymbol{\Delta} = (\delta_{ij}) \in \{0,1\}^{n \times n}$, and define $\lambda_n = \frac{|\mathbb{E}(\varepsilon'\boldsymbol{\Delta}\varepsilon)|/2(N-2)n}{\mathbb{E}(\varepsilon'\mathbf{I}\varepsilon)/n}$ and $\omega_n = \frac{|a'Q_{\mathrm{x}}\boldsymbol{\Delta}Q_{\mathrm{x}}^\top a|}{a'Q_{\mathrm{x}}\mathbf{I}Q_{\mathrm{x}}^\top a}$. Suppose that X has been centered, and the following conditions also hold:*

(10i.) $||\mathbf{H}_{\mathrm{x}}||_{\max} n^{3/4} = o(1)$, *and* $p/n = o(1)$.

(10ii.) $\lambda_n = \Theta(N^{-1+\epsilon_1})$ *and* $\omega_n \sim \Theta(N^{-1+\epsilon_2})$, *with* $\epsilon_1 \in (0,1)$ *and* $\epsilon_2 \in (0,2]$.

(10iii.) $\min(\omega_n\lambda_n, 1/\omega_n\lambda_n) \leq 1 - \delta$, *for some fixed* $\delta > 0$.

Then, Condition (C1) is satisfied, and the residual randomization test in Procedure 1 is asymptotically valid under either regular or restricted OLS.

Remarks on Theorem 11. (a) Condition (10i) aims to control the leverage of the design, as we have seen in previous results. For instance, if $\mathbf{H}_{\mathrm{x}}$ has some uniformity in its elements, then $||\mathbf{H}_{\mathrm{x}}||_{\max} = O(1/n)$ since the rows (or columns) of $\mathbf{H}_{\mathrm{x}}$ need to sum to one.

(b) Variable λ_n is an F-statistic comparing the within-dyad error variance with the sample variance. Variable ω_n captures a similar quantity, but for the covariates. Condition (10ii) ensures that the dyads are not too dissimilar from the sample. Under regularity conditions, $\lambda_n = O(1/N)$ and $\omega_n = O(N)$, which satisfies (10ii). The min term in condition (10iii) is always smaller than one, and the condition aims only to avoid extreme situations where the randomization distribution becomes degenerate. This might violate (A2), so (10iii) could potentially be relaxed.

(c) The asymptotics require that N increases. For inference with finite N, we could use repeated observations within every cell if these are available. In this setting, we could treat every cell as a single cluster, and leverage the theory of cluster exchangeability developed in Section 3.4.1. ■

3.5.4 Missing data

Frequently, there are missing values in the data, which may be important for inference. Here, we discuss how residual randomization can handle missing data, and briefly compare with alternatives. The subject of missing data is, of course, extensive (Little and Rubin, 2019); (Wooldridge, 2015, Section 9.5), so we will restrict our attention to a setting with dyadic exchangeability (Section 3.5.3). We use the trade application of Section 5.2 as a motivation, with a fixed set of countries, $\mathbb{N} = \{1, \ldots, J\}$. The data are cross-sectional trade flows between country dyads as in Section 3.5. Let $\mu_{ij} = \{0, 1\}$ denote whether pair (i, j) is observed ($\mu_{ij} = 1$), and define $\mathbf{M} = (\mu_{ij}) \in \{0, 1\}^{J \times J}$.

Any dependence between $\mathbf{M}$ and the data generating mechanism could significantly affect the analysis. However, the standard error methods of Conley (1999); Cameron et al. (2011); Fafchamps and Gubert (2007) do not consider the missing data issue. The bootstrap methods of Davezies et al. (2018); Menzel (2021); MacKinnon et al. (2021) consider $\mathbf{M}$ as a random variable, and impose on $\mathbf{M}$ the same partial exchangeability structure as the observations. This is probably an unrealistic assumption in practice because missingness is usually dyad-specific; e.g., whether the trade flow between two countries is missing or not depends on their geographic and cultural affinity, both of which are not partially exchangeable.

In contrast, the residual randomization method remains valid as long as the dyadic exchangeability property holds conditional on $\mathbf{M}$. Formally, let P denote a clustering of $\mathbb{N}$ such that the trade flow between any dyad (i, j) in the same cluster is observed. Let $\mathbb{P}(\mathbf{M})$ denote the set of all such clusterings. Let $\mathcal{G}_n^{\mathsf{P}}$ denote a subgroup of $\mathcal{G}_n^{\mathsf{dyad}}$, such that any $\mathrm{g} \in \mathcal{G}_n^{\mathsf{P}}$ doubly-permutes the residuals as described in Section 3.5, but is restricted only to units in the same cluster $\mathsf{P} \in \mathbb{P}(\mathbf{M})$.

With these definitions, suppose that

$$\varepsilon \overset{\mathrm{d}}{=} \mathrm{g}\varepsilon \mid X, \mathbf{M}, \text{ for any } \mathrm{g} \in \mathcal{G}_n^{\mathsf{P}} \subseteq \mathcal{G}_n^{\mathsf{dyad}}, \mathbf{M}. \tag{21}$$

Then, the residual randomization test of Section 3.5 using $\mathcal{G}_n^{\mathsf{P}}$ as the primitive is valid asymptotically under Theorem 9. Assumption (21) is satisfied whenever the missingness mechanism is independent of the errors, which appears to be a reasonable assumption. One practical matter with this procedure, however, is how to calculate P. One approach is to treat $\mathbf{M}$ as an undirected graph and calculate a clique cover, P^*, of $\mathbf{M}$. By definition, $\mathsf{P}^* \in \mathbb{P}(\mathbf{M})$, and so we could use a residual randomization test with $\mathcal{G}_n^{\mathsf{P}^*}$ as the invariant—we follow this approach in Section 5.2. Ideally, we would use the minimum clique cover of M. Even though this problem is NP-complete (Karp, 1972; Hartmanis, 1982), certain efficient approximations exist (Cai et al., 2013).

4 Numerical Examples

We begin our empirical evaluation of residual randomization with clustered error structures. Exact algorithmic details pertaining to implementation can be found in Appendix F.

4.1 One-way clustered data

We use the one-way layout simulation inspired by Cameron et al. (2008); Hansen (2018); Canay et al. (2018). The data generating model is

$$y_i = \beta_0 + x_i\beta + \varepsilon_i, \quad x_i = \mathrm{x}_{\mathrm{c}} + \mathrm{x}_{i\mathrm{c}}, \quad \varepsilon_i = \eta_{\mathrm{c}} + \epsilon_{i\mathrm{c}}, \quad i \in [\mathrm{c}]. \tag{22}$$

We consider the following simulation settings.

- $\eta_{\mathrm{c}} = 0$ or $\eta_{\mathrm{c}} \sim N(0,1)$; and $\epsilon_{ic} \sim N(0,1)$, all i.i.d..
- $\mathrm{x}_{\mathrm{c}} \sim N(0,1)$ or $\mathrm{x}_{\mathrm{c}} \sim .5LN(0,1)$, the log-normal distribution (produces high leverage points).
- For heteroskedasticity, we scale the errors by $3|x_i|$.
- $J_n \in \{10, 15, 20\}$ with 30 units per cluster; thus, $n \in \{300, 450, 600\}$.
- $(\beta_0, \beta_1) = (0, 0)$ with homoskedasticity; and $(\beta_0, \beta_1) = (1, 0)$ with heteroskedasticity.

We consider two versions of the residual randomization test. One employs the cluster sign symmetry described in Section 3.4.2, and is always valid. The other test employs both cluster exchangeability and sign symmetry, as described in Remark (d) of Theorem 8. This double invariance is valid under homoskedasticity, but is invalid in the heteroskedastic setting. Table 1 reports the rejection frequencies over 5,000 data replications at 5% level. The results include regular OLS errors as a strawman, and cluster robust error estimates through the `vcovCL` function from the `sandwich` `R` package (Zeileis, 2004). The randomization tests use $2,000$ resamples ($m = 2,000$).

In Panel (A) with homoskedastic errors, we see that standard OLS achieves the nominal level when there is no cluster correlation ($\eta_c = 0$), but significantly over-rejects when there is clustering. The cluster robust errors perform better but are also far from the nominal level in general. This is likely a small sample problem since rejection rates for this method improve as the number of clusters grows. The residual randomization tests works well across settings: their rejection rates are around 5% even when the covariates are log-normal, as shown in the columns marked as (2).

In Panel (B) with heteroskedastic errors, standard OLS fails severely across all settings. The cluster robust errors work roughly as well as before, but they tend to over-reject even more under log-normal covariates. For example, the rejection rates for this method are around 14% with log-normal covariates ($J_n = 10$), almost 3 times the nominal level. In contrast, the residual cluster sign test shows remarkable robustness. It mostly achieves the nominal rate with normal covariates. And with log-normal covariates, its worst rejection rate is around 8% when $J_n = 10$, but otherwise remains close to 5%. We also see that the double invariance test over-rejects in this setting, as expected, at a level that is higher but still comparable to the clustered error method.

(A) Homoskedastic												
	cluster effect (η_c)											
	$\eta_c = 0$ (no clustered effects)						$\eta_c \sim N(0,1)$ (clustered effects)					
	number of clusters (J)											
	$J = 10$		$J = 15$		$J = 20$		$J = 10$		$J = 15$		$J = 20$	
	cluster covariate (x_c)											
	N(0,1)	LN(0,1)										
OLS	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Standard	0.057	0.051	0.054	0.051	0.050	0.053	0.490	0.382	0.480	0.394	0.493	0.421
Cluster robust	0.086	0.090	0.076	0.079	0.061	0.077	0.103	0.110	0.081	0.089	0.081	0.090
Randomizations												
RR-cluster-sign	0.059	0.047	0.054	0.049	0.047	0.054	0.053	0.055	0.056	0.048	0.055	0.050
RR-cluster-double	0.061	0.054	0.056	0.052	0.051	0.056	0.055	0.052	0.054	0.046	0.051	0.050
(B) Heteroskedastic												
	$\eta_c = 0$						$\eta_c \sim N(0,1)$					
	$J = 10$		$J = 15$		$J = 20$		$J = 10$		$J = 15$		$J = 20$	
OLS	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Standard	0.228	0.249	0.244	0.264	0.239	0.286	0.278	0.274	0.301	0.303	0.301	0.309
Cluster robust	0.095	0.140	0.085	0.116	0.074	0.114	0.100	0.126	0.091	0.124	0.075	0.121
Randomizations												
RR-cluster-sign	0.055	0.084	0.055	0.072	0.052	0.072	0.049	0.065	0.059	0.071	0.056	0.072
RR-cluster-double	0.205	0.194	0.198	0.174	0.183	0.170	0.166	0.167	0.168	0.163	0.150	0.155

Table 1: Rejection rates under the null, $H_0 : \beta_1 = 0$ for the simulation of Section 4.1. Method "*RR-cluster-sign*" is the residual randomization test in Procedure 1 using $\mathcal{G}_n^{\mathsf{s}}(\mathrm{C}^n)$ as the invariant defined in Section 3.4.2. "*RR-cluster-double*" is the residual randomization test in Procedure 1 using $\mathcal{G}_n^{\mathsf{s+p}}(\mathrm{C}^n)$ discussed in Theorem 8. Column (1) corresponds to normal covariates; and (2) corresponds to lognormal covariates.

The power results under an alternative $H_1 : \beta_1 = 0.1$ are shown in Table 7 of Section E.3 of the supplement. We can see that the cluster sign randomization test pays a price for its robustness, but achieves about 85%-95% of the power of robust OLS across all settings. In the simulation results, we also include the power calculations for the double invariance test. This test achieves a significant improvement on power over the simple cluster-sign test. It is an interesting open question under what conditions employing a "richer" invariance can lead to such power improvements.

4.2 Dyadic regression

Here, we conduct a simulated study on dyadic regression inspired by simulations in (Cameron et al., 2011; Aronow et al., 2015). The data generating model is defined as follows:

$$y_i = \beta_0 + \beta_1 |x_\mathrm{r} - x_\mathrm{c}| + \varepsilon_i, \quad \varepsilon_i = \eta_\mathrm{r} + \eta_\mathrm{c} + \epsilon_{\mathrm{rc}}, \quad i = (\mathrm{r}, \mathrm{c}).$$

Here, $i = (\mathrm{r}, \mathrm{c})$ denotes an observation on a dyad of units; $\mathrm{r} \in \{1, \ldots, N\}$ is the "row" unit of the dyad, and $\mathrm{c} \in \{1, \ldots, N\}$ is the "column" unit of the dyad. Thus, $n = N(N-1)/2$. We consider the following simulation settings.

- $\epsilon_{jj'} \sim N(0,1)$; and $\eta_j \sim N(0,1)$ or $\eta_j \sim .5N(-1,.25^2) + .5N(1,.25^2)$, $j = 1,\ldots,N$.
- $x_j \sim N(0,1)$ or $x_j \sim LN(0,1)$, the standard log-normal distribution, $j = 1,\ldots,N$.
- $N \in \{10, 20, 35\}$ so that $n \in \{45, 190, 595\}$ (increases quadratically with N).
- $(\beta_0, \beta_1) = (1,1)$. We test $H_0 : \beta_1 = 1$ (true) and $H_0 : \beta_1 = 1.3$ (false).

Under this setup, we see that the errors satisfy the dyadic exchangeability invariance discussed in Section 3.5.3. We adopt this invariance for the residual randomization method. Table 2 reports results over 40,000 replications on four methods:

(I) A HC2 robust error method as a strawman.

(II) A standard two-way clustered error method, implemented with function `vcovCL` from the `sandwich` `R` package. In this method, we include dyad fixed effects.

(III) A linear mixed model with dyad random effects through function `lmer` in the `lme4` `R` package.

(IV) The residual randomization test based on the dyadic exchangeability, $\mathcal{G}_n^{\mathsf{dyad}}$, of Section 3.5.3.

In Table 2, we see that the HC2 method performs badly overall. This is expected because this method ignores the doubly clustered structure of the errors. The two-way clustered error method is, of course, an improvement. However, it tends to over-reject with a range being between 8%-13%. The problem is more pronounced with small samples and heavy-tailed data. Another problem with the two-way method is that it gives invalid estimates about 10% of the time. We discarded those cases from our results. The random effects model, on the other hand, performs substantially better. It tends to over-reject in small samples, but it approaches the nominal level in larger samples. For instance, with normal data, its rejection rate falls from 9.4% ($n = 45$) to 5.1% ($n = 595$). This method is also negatively impacted from heavy tailed data, but only through the covariates. For instance, in Panel (A), the random effects method has a 5.63% rejection rate with lognormal errors and normal covariates, while it has a 7.68% rejection rate with normal errors and lognormal covariates. In contrast, the residual randomization test remains closer to the nominal level than all other methods. Its performance is also remarkably robust across sample sizes and data distributions. In Panel (A), the rejection rate of the randomization test is always around 5%, with the worst rate being 5.17% with lognormal errors.

In Panel (B), we show rejection rates under a false null hypothesis. We compare only the randomization test with the random effects model because the other methods tend to substantially over-reject. With normal data and $n = 595$, the random effects model rejects 98.1% of the time, whereas the randomization test rejects only 34.8%, almost 3 times smaller. In the same setting, under the null hypothesis, the rejection rates of these two methods are, respectively, 5.8% and 5.09%. When the errors are lognormal this ratio is smaller as the rejection rates are 98.2% and 58%, respectively, with null rejection rates at 5.63% and 5.04%, respectively. These results show that the randomization is robust but, pays a price in efficiency under normal data.

Panel (A). $H_0 : \beta_1 = 1.0$

	Error-covariate, (ε_i, x_i)											
	(normal, normal)			(normal, lognormal)			(lognormal, normal)			(lognormal, lognormal)		
	Sample size, n											
Method	45	190	595	45	190	595	45	190	595	45	190	595
(I) HC2	18.07	29.21	41.50	22.90	40.16	53.65	14.03	25.87	38.14	15.77	30.27	46.06
(II) 2-way clustered	11.32	9.62	8.20	13.55	12.34	10.83	11.27	9.71	7.93	13.28	12.19	11.00
(III) random effects	9.37	6.07	5.80	11.91	8.71	7.68	9.00	6.72	5.63	11.13	8.72	7.45
(IV) RR-dyadic	5.11	4.63	5.09	5.00	5.09	4.91	4.89	5.17	5.04	4.85	4.94	4.97

Panel (B). $H_0 : \beta_1 = 1.3$

	45	190	595	45	190	595	45	190	595	45	190	595
random-effects	25.57	67.05	98.12	25.46	49.68	84.51	28.61	70.11	98.24	29.12	55.20	85.70
RR-dyadic	12.31	21.93	34.78	11.80	20.31	31.46	19.68	42.19	58.02	19.15	42.84	61.44

Table 2: Rejection rates (%) for HC2 errors, two-way cluster robust errors, a random effects model, and residual randomization ("RR-dyadic") under dyadic exchangeability using 2,500 resamples.

5 Empirical Examples

In this section, we present some empirical applications of residual randomization. We focus on complex error structures. One example with a simple error structure based on the hormone data of Efron and Tibshirani (1986, Section 9.3) is presented in Appendix G.1.

5.1 American foulbrood disease of honey bees

American foulbrood (AFB) is the most destructive of the honey bee brood diseases. It is caused by *Paenibacillus larvae*, a bacterium that produces spores infecting bee larvae. The disease is highly infectious and destructive to the extent that in many places, including the EU and the US, the law requires all infected hives to be burnt completely. It is therefore important to detect outbreaks early. One effective method is to monitor the hive for spores of the bacterium (Hornitzky and Karlovskis, 1989). Here, we use a dataset analyzed by Zuur et al. (2009) containing spore density observations from 24 hives. There are 3 observations per hive, with a total of $n = 72$ observations. The data also contain auxiliary information about each hive, including infection status and size.

The goal is to understand the relationship between spore density and infection status, controlling for the other variables, through the model

$$\log(\text{spore\_density}_{ck}) = \beta_0 + \beta_1 \text{Infection}_{ck} + \beta_2' X_{ck} + \varepsilon_{ck}.$$

Here, c denotes the hive and $k \in \{1, 2, 3\}$ denotes measurement in a hive. Due to the hierarchical nature of the data, Zuur et al. (2009) use various combinations of generalized least squares and mixed models (Lindstrom and Bates, 1990) utilizing functions `gls` and `lme` in `R`. These models require assumptions on linearity and normality of the hierarchical random effects, and on the error correlation structure, which may be unrealistic in practice. The residual randomization method, on the other hand, requires only assumptions on the error invariance structure. In this setting, we can assume that ε_{ck} have a one-way clustering structure per hive, and that these errors are sign symmetric across clusters. Alternatively, we could assume that the errors are both exchangeable within clusters and sign symmetric across clusters.

Our results in this dataset are shown in the table below.

	95% CI for β_1	
Method	lower	upper
OLS errors	2.259	3.545
Mixed model (RE by hive)	1.769	4.034
RR-cluster-sign	1.495	3.859
RR-cluster-double	1.484	3.926

In the first two rows, the table reports the intervals from regular OLS, and the mixed model selected based on AIC by Zuur et al. (2009). These results differ, probably because OLS tends to be anti-conservative with hierarchical data. The next two rows report the inverted intervals from the randomization tests. These methods don't differ a lot because there are only three observations per cluster. Overall, we see that the intervals from residual randomization are somewhat "in-between" OLS and the mixed model: They are wider than OLS but narrower than the mixed model. One benefit of residual randomization shown here is its simplicity and flexibility in using hierarchical error structures over the use of complex software packages.

5.2 International trade and currency unions

In an influential paper, Rose and Engel (2002) studied whether currency unions are associated with increased economic integration. The data are trade flows measured on dyads of countries. The working model is known as the gravity model of trade, and may be written as

$$\log(\mathrm{TRADE_{rc}}) = \beta_0 + \beta_1 \mathrm{CU_{rc}} + \beta_2 \log(\mathrm{GDP_r}) + \beta_3 \log(\mathrm{GDP_c}) + \beta_4 \mathrm{LANG_{rc}} + ... + \varepsilon_{\mathrm{rc}}.$$

Here, $\mathrm{TRADE_{rc}}$ is the trade flow between country r and country c measured in dollars; $\mathrm{CU_{rc}} \in \{0, 1\}$ indicates whether r and c are in the same currency union; GDP_j is the GDP of country j, and $\mathrm{LANG_{rc}} \in \{0, 1\}$ indicates whether r and c have their official language in common. The model can have additional control variables, such as GDP per capita, distance between countries, and so on. The parameter of interest is β_1. In this context, Cameron and Miller (2014) discuss several standard error estimates: regular OLS, HC, two-way clustered, and dyadic. The standard errors increase as more structure is imposed, but their analysis maintains the main result of Rose and Engel (2002) that β_1 is positive and significant. These error estimates, however, may be problematic for the reasons discussed in Section 3.5. Moreover, the data have a substantial amount of missingness as only 58.6% out of all possible unique country-pair values are observed.

To demonstrate the residual randomization method, we follow the missing data method described in Section 3.5.4. First, we calculate a clique cover of $\mathbf{M}$ (the missingness indicator matrix) using the `igraph` package in `R`. This produces a clustering, P^*, of the countries. The important property of P^* is that there are no missing values for any country-dyad in any cluster in P^*. We can then perform a dyadic permutation using $\mathcal{G}_n^{\mathsf{P}^*}$ as in Example 8. Second, to study the sensitivity of our inference, we can restrict P^* to consider only countries that satisfy certain specifications; e.g., being in the same continent. The results of this analysis are shown in Table 3.

In Panel (A), we see that the inferences from standard OLS errors and one-way clustered errors roughly agree. The estimated slope is positive and significant at the 5% level. This finding remains robust across specifications and additional clustering configurations not shown in the table, but available in the supplementary code. In Panel (B), the results from various residual randomization procedures are mixed. When we employ standard, non-clustered invariances we get similar results with OLS. However, these error structures are implausible in the gravity model that has by definition a dyadic structure. A more plausible structure is the dyadic exchangeability, $\mathcal{G}_n^{\mathsf{P}*}$, discussed above and in Section 3.5.4. Under this invariance, the slope coefficient is largely not significant. This finding remains robust across several specifications where we restrict the partial exchangeability condition to include only countries in the same continent, or speaking the same language, or a combination of these factors—see the “Filter” column in Table 3 for these specifications.

How can we resolve these conflicting conclusions? Here, we discuss some additional evidence suggesting that $\beta_1 = 0$ cannot be rejected with confidence— this supports the findings of residual randomization under dyadic exchangeability.

(a) Only 21/4615 dyads (4.5%) belong to a currency union (CU=1), implying a highly leveraged design. The average leverage scores of these countries is more than 10 times larger than the global average leverage score. There is also one country (Cameroon, CMR) that participates in roughly half of these dyads (11/21) dyads in the currency union set (CU=1), further skewing the design’s leverage.

(b) The countries in a currency union set are quite distinct from the population. They have, on average, an observed trade flow in the lower 25% quartile of the entire dataset, and generally have lower GDP per capita as most are in Africa. These countries are also closer to each other than the average. As a result, CU is strongly correlated with other variables. Indeed, a conditional permutation test based on the “propensity score” model that predicts the probability that CU=1 based on other characteristics (Candes et al., 2018; Shaikh and Toulis, 2021) also results in a non-significant β_1. A residual permutation test based on the method of DiCiccio and Romano (2017) also indicates a non-significant β_1.

(c) The residuals from the standard OLS regressions in Panel (A) show a clear geographic clustering. This can be visualized by plotting the eigenvectors of row/column residuals (Volfovsky and Hoff, 2015, Figure 2). As another diagnostic, a regression of these residuals against continent (say of country 1 in the dyad) yields a highly significant model (p-value = 0, global F-test).

Specification/Model				95% CI	for β_1
Panel (A)			estimate	lower	upper
(Rose and Engel, 2002)					
OLS			1.054 (0.38)	0.300	1.808
OLS, centered X			1.038 (0.33)	0.392	1.684
(Cameron and Miller, 2014)					
OLS, clustered by country 1			1.484 (0.28)	0.931	2.038
OLS, clustered by country 2			1.484 (0.62)	0.262	2.706
Panel (B)					
Res. Randomization	Invariant	Filter			
RR-perms.	$\mathcal{G}_n^{\mathsf{p}}$	{}		0.126	1.828
RR-signs	$\mathcal{G}_n^{\mathsf{s}}$	$*$		0.139	1.955
RR-double	$\mathcal{G}_n^{\mathsf{p+s}}$	$*$		0.076	1.941
RR-dyadic	$\mathcal{G}_n^{\mathsf{p}^*}$	$*$		0.0519	0.2660
$*$	$*$	{continent}		-0.0626	0.3617
$*$	$*$	{language}		-0.0814	0.2303
$*$	$*$	{continent, language}		-0.0438	0.1965

Table 3: Inference on the trade example of Section 5.2. Panel (A) reports OLS-based results. Panel (B) reports results from residual randomization. "RR-perms" uses full exchangeability; "RR-signs" uses sign symmetry; "RR-double" uses both invariances. "RR-dyadic" is the permutation test under dyadic partial exchangeability with missing data described in Section 3.5.4. The "filter" indicates that the dyadic permutation test uses only units sharing the same value for the variables included in the filter.

(d) The OLS models in Panel (A) do not adequately account for substantial data heteroskedasticity (p-value $<$ 1e-16, Breusch-Pagan test). In fact, β_1 is also not significant under a weighted least squares model with weights $w_i \propto 1/(1 + \mathrm{CU}_i)$ at the i-th datapoint (dyad). This finding remains robust in a substantial range of weight specifications.

All these additional results are made available in the supplementary code.

6 Extensions

Here, we discuss extensions of the linear setting of Section 3. Our focus is to demonstrate the potential of residual randomization to unify even broader applied settings than what has been considered so far. The theoretical content of this section will therefore be limited.

6.1 Generalized null hypotheses

Consider testing $H_0 : \beta_S = 0$, for a subset $S \subset \{1, \dots, n\}$ of parameters, assuming only exchangeable errors ($\mathcal{G}_n = \mathcal{G}_n^{\mathsf{p}}$). The hypothesis can also be written as $H_0 : R\beta = 0$, where $R \in \{0, 1\}^{|S| \times p}$ and $R_{ij} = 1$ if and only if $i = j$ and both $i, j \in S$.

A natural choice for the test statistic here is a Wald-type statistic, $T_n = n\hat{\beta}_S' V_S^{-1} \hat{\beta}_S$, where $\hat{\beta}_S = R\hat{\beta}_n$ is the OLS estimate for the components in S, and $V_S = nR\mathbf{S}R^\top$ is the corresponding sub-matrix of $\mathbf{S} = (X^\top X)^{-1}$. Thus, $T_n = n\hat{\beta}_n' R^\top (nR\mathbf{S}R^\top)^{-1} R\hat{\beta}_n$ and so, under the null hypothesis,

$$T_n \overset{\mathrm{H_0}}{=} \varepsilon' X \mathbf{S} \Omega_n \mathbf{S} X^\top \varepsilon = t_n(\varepsilon), \quad \Omega_n = R^\top (R\mathbf{S}R^\top)^{-1} R. \tag{23}$$

The test statistic is a known function of ε, but t_n is quadratic and so the theory of Section 3 is not applicable. To adapt our main theory, we will need the following additional assumptions.

$$\frac{\mathrm{Kurt}(\varepsilon_1)}{\lambda^2_{\mathrm{min}}(X^\top X)} = O(1). \tag{A3}$$

$$p, \kappa_n, \kappa(\Omega_n) = O(1). \tag{A4}$$

Assumption (A3) restricts the kurtosis in the error distribution. This condition is mild and probably violated only in extreme cases, since $\lambda_{\mathrm{min}}(X^\top X) \to \infty$ under regular conditions. Assumption (A4) imposes the usual constraints on the condition number of $X^\top X$ as in the linear model. The added restriction is $p < \infty$. This precludes certain high-dimensional settings that were allowed in the linear model.

Theorem 12. *For the linear model in (15) let $\varepsilon \overset{\mathrm{d}}{=} \mathrm{g}\varepsilon \mid X$ for all $\mathrm{g} \in \mathcal{G}_n^{\mathsf{p}}$ and $X \in \mathbb{X}^{(n)}$, $n > 0$, such that Assumptions (A1')-(A4) are satisfied. Suppose also that X has been centered. Then, the residual randomization test using (23) as the test statistic is asymptotically valid for $H_0 : \beta_S = 0$, under either regular or restricted OLS residuals.*

Remarks on Theorem 12. In this setting, an alternative to our approach is to repeatedly permute X_S and then recalculate the test statistic. DiCiccio and Romano (2017) have shown that this procedure has a robustness property. It is exact in finite samples when X_S is independent of other X and the errors, and is asymptotically valid under standard OLS assumptions. We compare our test with this method in the following example developed by DiCiccio and Romano (2017).

Example 9. *Let $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, $i = 1, \ldots, n$. We set $(\beta_0, \beta_1) = (1, 0)$, and consider two simulation settings: one with normal x_2, ε and uniform x_1, and another with all variables being t_5. The results are shown in Table 4. In the normal scenario, we see that the permutation test is quick to achieve the nominal level. The randomization test using restricted OLS has a larger rejection rate (e.g., 8.6% at $n = 10$) but it becomes comparable to the permutation test when $n > 25$. The situation is different in the t_5-scenario. The permutation test tends to over-reject, whereas residual randomization appears faster to achieve the nominal level. The table also shows that using restricted OLS residuals has a better finite-sample performance than regular OLS, but the difference diminishes as n grows.*

	$x_1 \sim U(1,4)$, $x_2, \varepsilon \sim N(0,1)$			$x_1, x_2, \varepsilon \sim \mathrm{t}_5$		
		residual randomization			residual randomization	
n	permutation	regular OLS	restricted OLS	permutation	regular OLS	restricted OLS
10	0.0488	0.1549	0.0866	0.0803	0.1560	0.0837
25	0.0487	0.0840	0.0631	0.0880	0.0821	0.0611
50	0.0505	0.0646	0.0548	0.0791	0.0648	0.0528
100	0.0532	0.0587	0.0539	0.0740	0.0567	0.0523
200	0.0502	0.0532	0.0509	0.0718	0.0529	0.0509

Table 4: Simulation results of Example 9. The results for the permutation method are taken from Table 4 of (DiCiccio and Romano, 2017).

6.2 Autocorrelated errors

Consider a simple linear time series model,

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \tag{24}$$

where t indexes time. In this setting, the errors may have a serial dependence. To employ residual randomization we thus need a novel concept of invariance. One interesting option would be to assume that the errors are symmetric around the time axis. We will refer to this property as reflection symmetry. This symmetry holds, for instance, when ε_t are observations from a continuous-time autoregressive process of order one (Jones, 1981; Belcher et al., 1994). This process is especially useful for modeling non-stationary or irregularly sampled data.

Under reflection symmetry, the residual randomization method implies a resampling scheme where the residuals are reflected around the time axis at appropriate "knots". Formally, let $\mathrm{C}_+^n(\varepsilon)$ denote a clustering of the errors under the following constraints: (i) every cluster in $\mathrm{C}_+^n(\varepsilon)$ may contain only sequential errors of the same sign; and (ii) adjacent clusters in $\mathrm{C}_+^n(\varepsilon)$ have different signs. As an example, suppose $\varepsilon = (1.5, 0.5, -0.1, -0.2, -0.05, 0.3, 0.6)$, then the implied clustering is $\mathrm{C}_+^n(\varepsilon) = \{\{1,2\},\{3,4,5\},\{6,7\}\}$. Reflection symmetry implies that, conditional on $\mathrm{C}_+^n(\varepsilon)$,

$$(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5, \varepsilon_6, \varepsilon_7) \overset{\mathrm{d}}{=} (\pm(\varepsilon_1, \varepsilon_2), \pm(\varepsilon_3, \varepsilon_4, \varepsilon_5), \pm(\varepsilon_6, \varepsilon_7)).$$

In other words, under reflection symmetry, the errors are cluster sign symmetric with respect to $\mathrm{C}_+^n(\varepsilon)$, and residual randomization can be applied using $\mathcal{G}_n^{\mathsf{s}}(\mathrm{C}_+^n(\varepsilon))$ as the invariant. This test is a special type of the cluster sign test in Section 3.4.2. The key difference is that $\mathrm{C}_+^n(\varepsilon)$ is random, as it depends on ε, while the clustering in Section 3.4.2 was fixed for any given $n > 0$.

While it is straightforward to show that the idealized test is exact under reflection symmetry, the theoretical analysis of the feasible, residual-based randomization test is more complicated. Crucially, the feasible test needs to cope with the noisy clustering, $\mathrm{C}_+^n(\widehat{\varepsilon})$, produced from using the residuals. Theorem 8 could in principle be applied under reflection symmetry, but it is unclear under what assumptions on the underlying stochastic process would $\mathrm{C}_+^n(\widehat{\varepsilon})$ satisfy the conditions of the theorem. On a more basic level, it is challenging to even characterize the asymptotics of the number of clusters, $J_n = |\mathrm{C}_+^n(\widehat{\varepsilon})|$, or the individual cluster sizes, $\{n_\mathrm{c} : \mathrm{c} = 1, \ldots, J_n\}$, which also appear in Theorem 8. For these reasons, the theoretical analysis of residual randomization based on reflection symmetry is best left for future work.

In the following example, we empirically demonstrate the potential of the proposed randomization test in settings with autocorrelated errors.

Example 10. *Consider model (24) with an AR(1) error structure, $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$. We set:*

- $\rho \in \{0, 0.3, 0.5, 0.8\}$, *and* $n = 100$ *datapoints;* $(\beta_0, \beta_1) = (-1, 1)$.
- $u_t \sim N(0, 1)$, *or a mixture* $5N(-1, .25^2) + .5N(1, .25^2)$, $t = 1, \ldots, 100$.
- *(i)* $x_t \sim N(0, 1)$, *(ii)* $x_t \sim LN(0, 1)$, *(iii)* $x_t = \rho x_{t-1} + N(0, 1)$, *or (iv)* $x_t = \rho x_{t-1} + LN(0, 1)$.

Further implementation details, and a power study, are in E.4 of the supplement.

The results from this simulation are shown in Table 5. We include results from standard OLS and HAC errors (White et al., 1980; MacKinnon and White, 1985; Andrews, 1991) implemented by the `vcovHAC` function in R's `sandwich` package. In Panel (A) with $\rho = 0.3$, we see that HAC errors are sometimes worse than standard OLS, probably due to small samples. HAC errors are also strongly affected by heavy-tailed, lognormal covariates. For instance, rejection rates from HAC errors jump from 6.24% to 11.42% as covariates become autocorrelated. In Panel (B) with $\rho = 0.8$, HAC errors perform much better but they are still affected by lognormal and highly autocorrelated covariates. Their rejection rates can be more than 2 times the nominal level. The residual randomization test under reflection invariance has superior performance. The test achieves the nominal level in nearly all settings. Its highest over-rejection is 5.86% when the errors have a mixture distribution and covariates are i.i.d. Despite these over-rejections, the test largely remains robust even with heavy-tailed data. Interestingly, while all other methods tend to significantly over-reject whenever there is autocorrelation in the covariates, the randomization test is just more conservative. ■

6.3 High-dimensional regression

In a high-dimensional setting we need to work with different estimators than OLS, but the principle behind residual randomization may still be applicable. Consider, for example, the linear hypothesis of Section 3, and suppose we use the ridge estimator, $\hat{\beta}^{\mathrm{ridge}} = (X^\top X + \lambda I)^{-1} X^\top y$, in the test statistic, $T_n = a'\hat{\beta}^{\mathrm{ridge}} - a_0$. Under the linear hypothesis, $T_n = a' P_\lambda^{-1} X^\top \varepsilon - \lambda a' P_\lambda^{-1} \beta$. Recall that our main method requires that, under the null hypothesis, $T_n = t_n(\varepsilon)$ for a known, measurable function t_n. Here, this property does not hold because of an additional term, $\lambda a' P_\lambda^{-1} \beta$, that is unknown. To proceed, we could use a plug-in estimate for β, such as ridge or lasso, and then run the residual randomization test as usual. We illustrate empirically this idea in the supplement with good empirical results. Another choice would be to directly aim to reduce the effect of the unknown term in the test statistic through debiasing techniques (Wang et al., 2021).

Panel (A): $\rho = 0.3, H_0 : \beta_1 = 1.0$ (true)								
	Error $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, $u_t = \ldots$							
	normal				*mixture*			
	Covariates x_t							
	iid		*autocorrelated*		*iid*		*autocorrelated*	
Method	(i)	(ii)	(iii)	(iv)	(i)	(ii)	(iii)	(iv)
OLS	5.06	4.98	7.17	7.23	5.09	4.55	7.08	7.00
HAC	6.70	11.31	7.13	11.21	6.88	14.32	7.20	13.82
RR-reflection	5.42	5.15	3.91	3.44	5.43	5.40	3.31	3.12

Panel (B): $\rho = 0.8, H_0 : \beta_1 = 1.0$ (true)								
	(i)	(ii)	(iii)	(iv)	(i)	(ii)	(iii)	(iv)
OLS	5.04	4.96	34.67	34.07	5.04	4.90	34.43	34.55
HAC	5.68	9.60	11.44	13.44	5.88	9.98	11.34	13.88
RR-reflection	5.78	5.53	3.72	3.22	5.86	5.58	3.93	3.77

Table 5: Rejection rates (%) over 100,000 replications for standard OLS errors, HAC errors, and the residual randomization test based on reflection symmetry in the simulated study of Section 6.2.

7 Concluding remarks

The extensions of residual randomization discussed in Section 6 pose some interesting open problems ahead. Another open question is to understand how to combine residual randomization tests that are individually valid. This requires us to extend the theory in this paper along two fronts. First, we need to understand how to combine test statistics, possibly in a non-linear way. In Theorem 12 we achieved such a result with a quadratic test statistic, but the proof uses a local Lipschitz condition that is specific to the quadratic form. Second, we need to be able to combine different invariance structures, $\mathcal{G}_n$, and understand how this affects validity and power. The proof of Theorem 6 gives a practical example of this idea, but the result is specific to the combination of exchangeability and sign symmetry. The power question is particularly challenging, and would require an analysis of how each invariance structure affects the residual randomization distribution.

Acknowledgements. I would like to thank Peter Bickel, Ivan Canay, Dean Eckles, Max Farrell, Avi Feller, Chris Hansen, Guido Imbens, Ed Kao, James MacKinnon, Lihua Lei, Paul Rosenbaum, and Azeem Shaikh for valuable feedback and discussions. I am also grateful for the John E. Jeuck Fellowship at Booth.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. (2017). When should you adjust standard errors for clustering? Technical report, National Bureau of Economic Research.

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598.

Anderson, M. J. and Robinson, J. (2001). Permutation tests for linear models. *Australian & New Zealand Journal of Statistics*, 43(1):75–88.

Andrews, D. (1991). Heteroskedasticity and autocorrelation consistent covariant matrix estimation. *Econometrica*, 59(3):817–858.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

Arellano, M. (1987). Practitioner's corner: Computing robust standard errors for within-groups estimators. *Oxford bulletin of Economics and Statistics*, 49(4):431–434.

Aronow, P., Cyrus, S., and Assenova, V. A. (2015). Cluster–robust variance estimation for dyadic data. *Political Analysis*, 23(4):564–577.

Baksalary, J. K. and Pordzik, P. (1990). A note on comparing the unrestricted and restricted least-squares estimators. *Linear Algebra and Its Applications*, 127:371–378.

Belcher, J., Hampton, J., and Wilson, G. T. (1994). Parameterization of continuous time autoregressive models for irregularly sampled time series data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):141–155.

Bester, C. A., Conley, T. G., and Hansen, C. B. (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, 165(2):137–151.

Bickel, P. J. and Freedman, D. A. (1983). Bootstrapping regression models with many parameters. *Festschrift for Erich L. Lehmann*, pages 28–48.

Bickel, P. J., Freedman, D. A., et al. (1981). Some asymptotic theory for the bootstrap. *The annals of statistics*, 9(6):1196–1217.

Cai, S., Su, K., Luo, C., and Sattar, A. (2013). Numvc: An efficient local search algorithm for minimum vertex cover. *Journal of Artificial Intelligence Research*, 46:687–716.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2):238–249.

Cameron, A. C. and Miller, D. L. (2014). Robust inference for dyadic data. *Unpublished manuscript, University of California-Davis.*

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372.

Canay, I. A., Romano, J. P., and Shaikh, A. M. (2017). Randomization tests under an approximate symmetry assumption. *Econometrica*, 85(3):1013–1030.

Canay, I. A., Santos, A., Shaikh, A. M., et al. (2018). The wild bootstrap with a" small" number of" large" clusters.

Candes, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577.

Carter, A. V., Schnepel, K. T., and Steigerwald, D. G. (2017). Asymptotic behavior of at-test robust to cluster heterogeneity. *Review of Economics and Statistics*, 99(4):698–709.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021). An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116(536):1849–1864.

Christakis, N. A. and Fowler, J. H. (2013). Social contagion theory: examining dynamic social networks and human behavior. *Statistics in medicine*, 32(4):556–577.

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507.

Conley, T. G. (1999). Gmm estimation with cross sectional dependence. *Journal of econometrics*, 92(1):1–45.

Conley, T. G. and Taber, C. R. (2011). Inference with difference in differences with a small number of policy changes. *The Review of Economics and Statistics*, 93(1):113–125.

Cox, D. R. and Hinkley, D. V. (1979). *Theoretical statistics*. Chapman and Hall/CRC.

Davezies, L., D'Haultfoeuille, X., and Guyonvarch, Y. (2018). Asymptotic results under multiway clustering. *arXiv preprint arXiv:1807.07925*.

Davidson, R. and Flachaire, E. (2008). The wild bootstrap, tamed at last. *Journal of Econometrics*, 146(1):162–169.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Number 1. Cambridge university press.

Dekker, D., Krackhardt, D., and Snijders, T. A. (2007). Sensitivity of mrqap tests to collinearity and autocorrelation conditions. *Psychometrika*, 72(4):563–581.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: confidence intervals, p-values and r-software hdi. *Statistical science*, pages 533–558.

DiCiccio, C. J. and Romano, J. P. (2017). Robust permutation tests for correlation and regression coefficients. *Journal of the American Statistical Association*, 112(519):1211–1220.

Ding, P. (2017). A paradox from randomization-based causal inference. *Statistical science*, pages 331–345.

Edgington, E. and Onghena, P. (2007). *Randomization tests.* Chapman and Hall/CRC.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75.

Eicker, F. et al. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *The Annals of Mathematical Statistics*, 34(2):447–456.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (2013). *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media.

Ernst, M. D. (2004). Permutation methods: a basis for exact inference. *Statistical Science*, pages 676–685.

Fafchamps, M. and Gubert, F. (2007). The formation of risk sharing networks. *Journal of development Economics*, 83(2):326–350.

Field, C. A. and Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):369–390.

Fisher, R. A. (1935). *The design of experiments.* Oliver And Boyd; Edinburgh.

Freedman, D. and Lane, D. (1983a). A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–298.

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, 9(6):1218–1228.

Freedman, D. A. and Lane, D. (1983b). Significance testing in a nonstochastic setting. *A festschrift for Erich L. Lehmann*, pages 185–208.

Freedman, D. A. and Peters, S. C. (1984). Bootstrapping an econometric model: Some empirical results. *Journal of Business & Economic Statistics*, 2(2):150–158.

Gerber, A. S. and Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation.* WW Norton.

Godfrey, L. G. and Orme, C. D. (2001). Significance levels of heteroskedasticity-robust tests for specification and misspecification: some results on the use of wild bootstraps. *Unpublished Manuscript*.

Good, P. (2013). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media.

Hansen, B. (2018). The exact distribution of the t-ratio with robust and clustered standard errors.

Hansen, C. B. (2007). Asymptotic properties of a robust variance matrix estimator for panel data when t is large. *Journal of Econometrics*, 141(2):597–620.

Hartmanis, J. (1982). Computers and intractability: a guide to the theory of np-completeness (michael r. garey and david s. johnson). *Siam Review*, 24(1):90.

Higgins, J. J. (2004). *An introduction to modern nonparametric statistics*. Brooks/Cole Pacific Grove, CA.

Hinkelmann, K. and Kempthorne, O. (2007). *Design and analysis of experiments, volume 1: Introduction to experimental design*, volume 1. John Wiley & Sons.

Hoover, D. N. (1979). Relations on probability spaces and arrays of. *t, Institute for Advanced Study*.

Hornitzky, M. and Karlovskis, S. (1989). A culture technique for the detection of bacillus larvae in honeybees. *Journal of apicultural research*, 28(2):118–120.

Hubert, L. (1986). *Assignment methods in combinational data analysis*, volume 73. CRC Press.

Hubert, L. and Schultz, J. (1976). Quadratic assignment as a general data analysis strategy. *British journal of mathematical and statistical psychology*, 29(2):190–241.

Ibragimov, R. and Müller, U. K. (2010). t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics*, 28(4):453–468.

Ibragimov, R. and Müller, U. K. (2016). Inference with few heterogeneous clusters. *Review of Economics and Statistics*, 98(1):83–96.

Imbens, G. W. and Kolesar, M. (2016). Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics*, 98(4):701–712.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

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

Jones, R. H. (1981). Fitting a continuous time autoregression to discrete data. In *Applied time series analysis II*, pages 651–682. Elsevier.

Karp, R. M. (1972). Reducibility among combinatorial problems. In *Complexity of computer computations*, pages 85–103. Springer.

Kempthorne, O. (1992). Intervention experiments, randomization and inference. *Lecture Notes-Monograph Series*, pages 13–31.

Krackhardt, D. (1988). Predicting with networks: Nonparametric multiple regression analysis of dyadic data. *Social networks*, 10(4):359–381.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation.* Springer Science & Business Media.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses.* New York: Springer.

Lei, L. and Bickel, P. J. (2021). An assumption-free exact test for fixed-design linear models with exchangeable errors. *Biometrika*, 108(2):397–412.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, pages 673–687.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Liu, R. Y. et al. (1988). Bootstrap procedures under some non-iid models. *The Annals of Statistics*, 16(4):1696–1708.

MacKinnon, J. G., Nielsen, M. Ø., and Webb, M. D. (2021). Wild bootstrap and asymptotic inference with multiway clustering. *Journal of Business & Economic Statistics*, 39(2):505–519.

MacKinnon, J. G., Nielsen, M. Ø., and Webb, M. D. (2022). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics.*

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics*, 29(3):305–325.

Mammen, E. et al. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The annals of statistics*, 21(1):255–285.

Marshall, A. W., Olkin, I., and Arnold, B. C. (1979). *Inequalities: theory of majorization and its applications*, volume 143. Springer.

McCaffrey, D. F. and Bell, R. M. (2002). Bias reduction in standard errors for linear and generalized linear models with multi-stage samples. In *Proceedings of Statistics Canada Symposium*, pages 1–10.

McCullagh, P. (2000). Resampling and exchangeable arrays. *Bernoulli*, pages 285–301.

McCulloch, C. E. and Searle, S. R. (2004). *Generalized, linear, and mixed models*. John Wiley & Sons.

Menzel, K. (2021). Bootstrap with cluster-dependence in two or more dimensions. *Econometrica*, 89(5):2143–2188.

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of econometrics*, 32(3):385–397.

Neuhaus, G. (1993). Conditional rank tests for the two-sample problem under random censorship. *The Annals of Statistics*, pages 1760–1779.

Peters, S. and Freedman, D. (1984). Some notes on the bootstrap in regression problems. *Journal of Business & Economic Statistics*, 2(4):406–09.

Pitman, E. J. (1937). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130.

Rose, A. and Engel, C. (2002). Currency unions and international integration. *Journal of Money, Credit, and Banking*, 34:1067–1089.

Rosenbaum, P. R. (2002). Observational studies. In *Observational studies*, pages 1–17. Springer.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Shaikh, A. M. and Toulis, P. (2021). Randomization tests in observational studies with staggered adoption of treatment. *Journal of the American Statistical Association*, 116(536):1835–1848.

Snijders, T. A. (2011). Statistical models for social networks. *Annual review of sociology*, 37:131–153.

Tsybakov, A. (2009). *Introduction to nonparametric estimation*. Springer.

Volfovsky, A. and Hoff, P. D. (2015). Testing for nodal dependence in relational data matrices. *Journal of the American Statistical Association*, 110(511):1037–1046.

Wang, Y. S., Lee, S. K., Toulis, P., and Kolar, M. (2021). Robust inference for high-dimensional linear models via residual randomization. In *International Conference on Machine Learning*, pages 10805–10815. PMLR.

Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*, 38(3-4):330–336.

White, H. (1984). *Asymptotic theory for econometricians*. Academic press.

White, H. et al. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *econometrica*, 48(4):817–838.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Cengage learning.

Wu, C.-F. J. et al. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics*, 14(4):1261–1295.

Zeileis, A. (2004). Econometric computing with hc and hac covariance matrix estimators.

Zhao, A. and Ding, P. (2021). Covariate-adjusted fisher randomization tests for the average treatment effect. *Journal of Econometrics*, 225(2):278–294.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., Smith, G. M., et al. (2009). *Mixed effects models and extensions in ecology with R*, volume 574. Springer.