Covariate-adjusted Fisher randomization tests for the average treatment effect

Anqi Zhao and Peng Ding *

Abstract

Fisher's randomization test (FRT) delivers exact p-values under the strong null hypothesis of no treatment effect on any units whatsoever and allows for flexible covariate adjustment to improve the power. Of interest is whether the procedure could also be valid for testing the weak null hypothesis of zero average treatment effect. Towards this end, we evaluate two general strategies for FRT with covariate-adjusted test statistics: that based on the residuals from an outcome model with only the covariates, and that based on the output from an outcome model with both the treatment and the covariates. Based on theory and simulation, we recommend using the ordinary least squares (OLS) fit of the observed outcome on the treatment, centered covariates, and their interactions for covariate adjustment, and conducting FRT with the robust t-value of the treatment as the test statistic. The resulting FRT is finite-sample exact for the strong null hypothesis, asymptotically valid for the weak null hypothesis, and more powerful than the unadjusted analog under alternatives, all irrespective of whether the linear model is correctly specified or not. We develop the theory for complete randomization, cluster randomization, stratified randomization, and rerandomization, respectively, and give a recommendation for the test procedure and test statistic under each design. We first focus on the finite-population perspective and then extend the result to the super-population perspective, highlighting the difference in standard errors. Motivated by the similarity in procedure, we also evaluate the design-based properties of five existing permutation tests originally for linear models and show the superiority of the proposed FRT for testing the treatment effects.

Keywords: Finite-population inference; permutation test; randomization distribution; robust standard error; studentization; super-population inference

*Anqi Zhao, Department of Statistics and Applied Probability, National University of Singapore, 117546, Singapore (E-mail: staza@nus.edu.sg). Peng Ding, Department of Statistics, University of California, Berkeley, CA 94720 (E-mail: pengdingpku@berkeley.edu). Peng Ding was partially funded by the U.S. National Science Foundation (grant # 1945136). We thank Jason Wu, Cheng Gao, Kevin Guo, Thomas Richardson, Avi Feller, Xiaokang Luo, Xinran Li, Zhichao Jiang, Bin Yu, and Philip Stark for helpful comments.

1. Fisher's randomization test with covariate adjustment

Fisher (1935) viewed randomization as a reasoned basis for inference and proposed the randomization test as a universal way to generate finite-sample exact p-values without imposing modeling assumptions on the experimental outcomes. FRT becomes increasingly important with the popularity of field experiments in social sciences in addition to the traditional biomedical experiments. Proschan and Dodd (2019) reviewed the use of FRTs in randomized clinical trials and highlighted its strength in analyzing complex data. Based on a review of recent experimental papers in economics, Young (2019) showed the problems with model-based inference and advocated using FRT to obtain more credible p-values. There is an increasing interest in economics and related fields to use FRTs to analyze various types of empirical data (Freedman and Lane 1983; Kennedy 1995; Cattaneo et al. 2015; Canay et al. 2017; Ganong and Jäger 2018; Athey et al. 2018; Bugni et al. 2018; Young 2019; Heckman and Karapakula 2019; MacKinnon and Webb 2020).

The flexibility of FRT enables two natural strategies to incorporate covariate information to further improve the power. First, we can fit a statistical model of the outcome on the covariates and use the residuals as the pseudo outcomes to form the test statistic. Tukey (1993) used it with linear models, Gail et al. (1988) used it with generalized linear models, Raz (1990) used it with nonparametric regressions, Stephens et al. (2013) used it with the generalized estimating equation for clustered data, and Rosenbaum (2002) reviewed and extended it to not only randomized experiments but also matched observational studies. Second, we can directly fit an outcome model with both the treatment and covariates and use the model output, such as the coefficient of the treatment or the corresponding t-values, as the test statistic. The canonical choice is a linear model on the treatment and covariates, often known as the analysis of covariance (Fisher 1935; Freedman 2008; Lin 2013; Young 2019). Brillinger et al. (1978) gave an early application of this strategy with more complex statistical models.

The strong guarantees of FRT hold only under the strong null hypothesis of zero individual treatment effects, which is often criticized for being too restrictive for many practical applications. Adaptation to the weak null hypothesis of zero average treatment effect is one important direction for broadening its application. A natural class of test statistics for this purpose are the coefficients of the treatment from various outcome models under the above two strategies. They are consistent estimators of the average treatment effect and thus apply to both the strong and weak null hypotheses (Freedman 2008; Lin 2013). Of interest is the operating characteristics of the resulting FRTs when only the weak null hypothesis holds.

In general, not all test statistics can preserve the correct type one error rates under the weak null hypothesis even asymptotically (Romano 1990; Chung and Romano 2013; Wu and Ding 2020). We examine the asymptotic validity of the above two strategies and unify the discussion under the umbrella of OLS. Motivated by previous work on using studentized statistics for permutation tests (Janssen 1997; Chung and Romano 2013; Pauly et al. 2015; Wu and Ding 2020), we also examine the associated t-statistics studentized by the classic and robust standard errors, respectively (Eicker

1967; Huber 1967; White 1980). The robust t-statistic based on Lin (2013)'s estimator, as it turns out, guarantees both asymptotic validity and the highest power for testing the weak null hypothesis. His estimator equals the coefficient of the treatment in the OLS fit of the outcome on the treatment, centered covariates, and their interactions, but the aforementioned superior properties hold irrespective of whether the linear model is correctly specified or not. It is thus our final recommendation for testing the weak null hypothesis under complete randomization.

We first focus on complete randomization and then generalize the theory to other types of design. The extension to cluster randomization and stratified randomization is direct whereas that to rerandomization (Morgan and Rubin 2012) has some distinct features. In particular, covariate adjustment becomes more crucial since studentization alone does not ensure the appropriateness of FRT for the weak null hypothesis. In addition, it is common that the designer and analyzer do not communicate (Bruhn and McKenzie 2009; Heckman and Karapakula 2019), and if this happens, we recommend using FRT pretending that the experiment was completely randomized. In this non-ideal case, the proposed FRT is no longer finite-sample exact under the strong null hypothesis unless the original experiment is indeed completely randomized, but at least it preserves the correct type one error rates under the weak null hypothesis. Based on extensive theoretical investigations, we make final recommendations for FRT and the test statistic in each experimental design.

The above theory holds under the design-based framework conditioning on the finite population of units without any assumptions on the data generating process. With a slight modification of the standard error estimation, we extend the theory to the super-population framework and show the validity of the proposed procedures when the potential outcomes are independent draws from a super population. Further, the proposed procedures, though model-free in theory, make use of the OLS coefficients and t-statistics for easy implementation. It is thus curious to study their connections with existing permutation tests for coefficients in linear models (Freedman and Lane 1983; ter Braak 1992; Kennedy 1995; Manly 1997; DiCiccio and Romano 2017), as reviewed by Anderson and Legendre (1999), Anderson and Robinson (2001), and Lei and Bickel (2020). We evaluate the operating characteristics of these permutation tests for testing the treatment effects and demonstrate the superiority of FRT by various criteria. Among them, the recent proposal by DiCiccio and Romano (2017) is the closest to FRT and coincides in procedure with FRT based on Fisher (1935)'s estimator studentized by the robust standard error. Possible improvements include adding the treatment-covariates interactions into the linear model, such that it coincides with our final recommendation of FRT based on Lin (2013)'s estimator studentized by the robust standard error. DiCiccio and Romano (2017)'s original theory was developed under the linear model assumption for testing whether a coefficient is zero. Our extension provides an additional justification for it under the potential outcomes framework for testing the treatment effects.

We will use the following notation for permutations. Let Π be the set of all $N!$ random permutations of $\{1, \ldots, N\}$, indexed by π. For an $N \times 1$ vector $a = (a_1, \ldots, a_N)^{\mathrm{T}}$, let $a_\pi = (a_{\pi(1)}, \ldots, a_{\pi(N)})^{\mathrm{T}}$ be a permutation of its elements. If $b = b(a)$ is a function of a, let $b^\pi = b(a_\pi)$ be its value evaluated at a_π. Without introducing new notation, use π to also represent a random draw from Π, namely

$\pi \sim \text{Unif}(\Pi)$, with meaning clear from the context. With a slight abuse of notation, assume sets like $\{a_\pi : \pi \in \Pi\}$ to contain $|\Pi| = N!$ elements defined by $\pi \in \Pi$ throughout, such that a_π and $a_{\pi'}$ are two distinct elements so long as $\pi \neq \pi'$, even if $a_\pi = a_{\pi'}$.

2. Basic setup under complete randomization

2.1. Potential outcomes and Fisher's randomization test

Consider an intervention of two levels, $z = 0, 1$, and a finite population of N units, $i = 1, \ldots, N$. Let $Y_i(z)$ be the potential outcome of unit i under treatment z (Neyman 1923). The individual treatment effect is $\tau_i = Y_i(1) - Y_i(0)$, and the average treatment effect is $\tau = N^{-1} \sum_{i=1}^N \tau_i$. Let $x_i = (x_{i1}, \ldots, x_{iJ})^{\mathrm{T}}$ be the covariates for unit i, concatenated as an $N \times J$ matrix $X = (x_1, \ldots, x_N)^{\mathrm{T}}$. Center the covariates at $\bar{x} = N^{-1} \sum_{i=1}^N x_i = 0_J$ to simplify the presentation.

The designer assigns N_z units to receive level z with $N_1 + N_0 = N$ and $(p_1, p_0) = (N_1/N, N_0/N)$. Let Z_i denote the treatment level received by unit i, with $Z_i = 1$ for treatment and $Z_i = 0$ for control, vectorized as $Z = (Z_1, \ldots, Z_N)^{\mathrm{T}}$. Complete randomization samples Z uniformly from the set $\mathcal{Z}$ that contains all permutations of N_1 1's and N_0 0's. The observed outcome is $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$ for unit i, vectorized as $Y = (Y_1, \ldots, Y_N)^{\mathrm{T}}$. A test statistic is a function of the treatment vector, observed outcomes, and covariates, denoted by $T = T(Z, Y, X)$.

Write $Y = Y(Z)$ and $T = T(Z, Y(Z), X)$ to highlight the dependence of the observed outcomes and the test statistic on the treatment vector. Complete randomization induces a uniform distribution over $\{T(\boldsymbol{z}, Y(\boldsymbol{z}), X) : \boldsymbol{z} \in \mathcal{Z}\}$ as the *sampling distribution* of T. Fisher (1935) considered testing the strong null hypothesis

$$H_{0\mathrm{F}} : Y_i(1) = Y_i(0) \quad \text{for all } i = 1, \ldots, N$$

and proposed FRT to compute the p-value as

$$p_{\mathrm{FRT}} = |\Pi|^{-1} \sum_{\pi \in \Pi} 1\{T(Z_\pi, Y, X) \geq T(Z, Y, X)\}, \tag{1}$$

assuming a one-sided test. Each Z_π is a permutation of Z, and by symmetry, all possible values of Z_π over $\pi \in \Pi$ consist of $\mathcal{Z}$. FRT thus induces a uniform distribution over $\{T(\boldsymbol{z}, Y(Z), X) : \boldsymbol{z} \in \mathcal{Z}\}$ conditioning on the observed Z, known as the *randomization distribution* of T. Let $T^\pi = T(Z_\pi, Y(Z), X)$, where $\pi \sim \text{Unif}(\Pi)$, be a random variable from this distribution. The p_{FRT} in (1) as such gives the right-tail probability of the observed value of the test statistic with regard to its randomization distribution. Under $H_{0\mathrm{F}}$, the randomization distribution equals the sampling distribution which ensures the finite-sample exactness of FRT for arbitrary T.

In practice, we need to choose a test statistic that is sensitive to deviations from $H_{0\mathrm{F}}$. Computationally, FRT involves randomly permuting the treatment vector to generate Z_π. This justifies "permutation test" as its other name. If $|\Pi| = N!$ is too large, we can take a simple random sample

from Π to obtain a Monte Carlo approximation of p_{FRT}.

The nice properties of FRT under $H_{0\text{F}}$ inspires endeavors to extend it to other types of hypotheses. Consider the weak null hypothesis of zero average treatment effect (Neyman 1935):

$$H_{0\text{N}} : \tau = 0.$$

We can proceed with FRT by permuting the treatment vector Z and report p_{FRT} by (1) as if we were testing $H_{0\text{F}}$. Under $H_{0\text{N}}$, $Y(\boldsymbol{z})$ varies with $\boldsymbol{z} \in \mathcal{Z}$ such that the randomization distribution T^{π} no longer equals the sampling distribution T. Consequently, p_{FRT} loses its finite-sample exactness as the basis for controlling the type one error rates in general. Wu and Ding (2020) gave examples in which FRT yields invalid type one error rates under $H_{0\text{N}}$ even asymptotically.

Despite the possibly liberal type one error rates in general, sensible choices of the test statistic restore the validity of FRT for $H_{0\text{N}}$ at least asymptotically. Wu and Ding (2020) showed FRT preserves the correct type one error rates asymptotically with a class of appropriately studentized statistics. We extend their discussion to the setting with covariates and propose a general strategy for covariate-adjusted FRT that ensures both asymptotic validity and higher power for testing $H_{0\text{N}}$.

Assume the finite-population asymptotic framework that embeds $\mathcal{S} = \{Y_i(0), Y_i(1), x_i\}_{i=1}^{N}$ and $Z = (Z_i)_{i=1}^{N}$ into a sequence of finite populations and assignments for $N = 1, \ldots, \infty$. Technically, all quantities depend on N, but we omit the subscript N for simplicity.

Definition 1. A test statistic T is *proper* for testing $H_{0\text{N}}$ if under $H_{0\text{N}}$, $\lim_{N\to\infty} \text{pr}(p_{\text{FRT}} \leq \alpha) \leq \alpha$ for all $\alpha \in (0, 1)$ for almost all sequences of Z and any $\mathcal{S}$.

Assume a one-sided test and p-value as the right-tail probability as in (1). A statistic T is proper for testing $H_{0\text{N}}$ if under $H_{0\text{N}}$, the sampling distribution of T is stochastically dominated by its randomization distribution for almost all sequences of Z as $N \to \infty$.

2.2. Two strategies for covariate-adjusted FRT and twelve test statistics

We review two general strategies for covariate adjustment in FRT. We focus on test statistics based on estimators of τ to accommodate both $H_{0\text{F}}$ and $H_{0\text{N}}$, and unify them under the OLS formulation for easy implementation.

Let $\hat{Y}(z) = N_z^{-1} \sum_{i:Z_i=z} Y_i$ be the sample average of the outcomes under treatment z. The difference-in-means estimator $\hat{\tau}_{\text{N}} = \hat{Y}(1) - \hat{Y}(0)$ is unbiased for τ under complete randomization (Neyman 1923) and affords a natural statistic for testing both $H_{0\text{F}}$ and $H_{0\text{N}}$. Algebraically, it equals the coefficient of Z_i from the OLS fit of Y_i on $(1, Z_i)$. It is also common to use $\hat{\tau}_{\text{N}}/\hat{\text{se}}_{\text{N}}$ or $\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}$ as the test statistic, where $\hat{\text{se}}_{\text{N}}$ and $\tilde{\text{se}}_{\text{N}}$ are the classic and robust standard errors from the same OLS fit. Standard results suggest that $\tilde{\text{se}}_{\text{N}}^2$ is almost identical to the variance estimator proposed by Neyman (1923) whereas $\hat{\text{se}}_{\text{N}}^2$ is not, even asymptotically. Chung and Romano (2013) and Wu and Ding (2020) showed that randomization tests with appropriately studentized statistics are asymptotically valid for $H_{0\text{N}}$. We also consider studentization in covariate adjustment below.

Table 1: Twelve test statistics where $\hat{\text{se}}$ and $\tilde{\text{se}}$ denote the classic and robust standard errors.

	Neyman (1923)	Rosenbaum (2002)	Fisher (1935)	Lin (2013)
unstudentized	$\hat{\tau}_{\text{N}}$	$\hat{\tau}_{\text{R}}$	$\hat{\tau}_{\text{F}}$	$\hat{\tau}_{\text{L}}$
studentized by $\hat{\text{se}}$	$\hat{\tau}_{\text{N}}/\hat{\text{se}}_{\text{N}}$	$\hat{\tau}_{\text{R}}/\hat{\text{se}}_{\text{R}}$	$\hat{\tau}_{\text{F}}/\hat{\text{se}}_{\text{F}}$	$\hat{\tau}_{\text{L}}/\hat{\text{se}}_{\text{L}}$
studentized by $\tilde{\text{se}}$	$\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}$	$\hat{\tau}_{\text{R}}/\tilde{\text{se}}_{\text{R}}$	$\hat{\tau}_{\text{F}}/\tilde{\text{se}}_{\text{F}}$	$\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$

The first strategy for covariate adjustment is to fit an outcome model with covariates alone and use the residuals as the fixed, covariated-adjusted outcomes for conducting FRT. This appears to be the dominating approach advocated by Rosenbaum (2002); see also Gail et al. (1988), Raz (1990), Tukey (1993) and Ottoboni et al. (2018). Let $e = (e_1, \ldots, e_N)^{\text{T}}$ be the residuals from the OLS fit of Y_i on $(1, x_i)$, which can be viewed as pseudo outcomes unaffected by the treatment under $H_{0\text{F}}$. The difference in means of the residuals, $\hat{\tau}_{\text{R}} = \hat{e}(1) - \hat{e}(0)$, equals the coefficient of Z_i from the OLS fit of e_i on $(1, Z_i)$ and affords an intuitive estimator of τ after adjusting for the covariates. Similar to the discussion of $\hat{\tau}_{\text{N}}$, we can use $\hat{\tau}_{\text{R}}$, $\hat{\tau}_{\text{R}}/\hat{\text{se}}_{\text{R}}$, and $\hat{\tau}_{\text{R}}/\tilde{\text{se}}_{\text{R}}$ as the test statistics for testing $H_{0\text{F}}$ or $H_{0\text{N}}$ by FRT, where $\hat{\text{se}}_{\text{R}}$ and $\tilde{\text{se}}_{\text{R}}$ are the classic and robust standard errors from the OLS fit that yields $\hat{\tau}_{\text{R}}$. We regress Y_i on $(1, x_i)$ to form the residuals e whereas Rosenbaum (2002) regressed Y_i on x_i alone without the intercept. The difference does not affect $\hat{\tau}_{\text{R}}$ with centered covariates.

The second strategy for covariate adjustment is to directly fit an outcome model with both the treatment and covariates and use the coefficient or t-values of the treatment as the test statistic for FRT. Fisher (1935) suggested an estimator $\hat{\tau}_{\text{F}}$ for τ, which equals the coefficient of Z_i from the OLS fit of Y_i on $(1, Z_i, x_i)$. Lin (2013) recommended an improved estimator, $\hat{\tau}_{\text{L}}$, as the coefficient of Z_i from the OLS fit of Y_i on $\{1, Z_i, x_i - \bar{x}, Z_i(x_i - \bar{x})\}$ with centered covariates and treatment-covariates interactions. These two covariate-adjusted estimators, alongside their respective studentized variants, afford six additional test statistics, namely $\hat{\tau}_*$, $\hat{\tau}_*/\hat{\text{se}}_*$, and $\hat{\tau}_*/\tilde{\text{se}}_*$ $(* = \text{F}, \text{L})$, for testing $H_{0\text{F}}$ or $H_{0\text{N}}$ by FRT, where $\hat{\text{se}}_*$ and $\tilde{\text{se}}_*$ are the classic and robust standard errors from the respective OLS fits.

This gives us a total of twelve test statistics, three unadjusted and nine adjusted, for testing the treatment effects via FRT. Table 1 summarizes them, with the subscripts N, R, F, and L indicating Neyman (1923), Rosenbaum (2002), Fisher (1935), and Lin (2013), respectively. All twelve statistics are finite-sample exact for testing $H_{0\text{F}}$ irrespective of whether the models are correctly specified or not. Our goal is to evaluate their abilities for preserving the correct type one error rates under $H_{0\text{N}}$. Without loss of generality, we assume two-sided FRT for the rest of the text, or, equivalently, we use the absolute value of the test statistics in Table 1 to compute the p_{FRT} in (1).

The two strategies for covariate adjustment unify nicely under the OLS formulation yet differ materially with regard to the role of covariates under the permutations induced by the FRT procedure. The first strategy, on the one hand, adjusts for the covariates only once to form the pseudo outcomes e and proceeds with the permutations in a covariate-free fashion. The second strategy, on the other hand, adjusts for the covariates in each of the $N!$ permutations of Z.

Before giving the formal asymptotic results, we unify the above four estimators of τ. Let $S_x^2 = (N-1)^{-1}\sum_{i=1}^N x_i x_i^{\mathrm{T}}$ and $\hat{S}_{xY} = (N-1)^{-1}\sum_{i=1}^N x_i Y_i$ be the finite-population covariance matrices of the centered $(x_i)_{i=1}^N$ with itself and $(Y_i)_{i=1}^N$, respectively. Let $\hat{\tau}_x = \hat{x}(1) - \hat{x}(0)$ be the difference in means of the covariates under treatment and control, where $\hat{x}(z) = N_z^{-1}\sum_{i:Z_i=z} x_i$. Let $\hat{S}_{x(z)}^2 = (N_z-1)^{-1}\sum_{i:Z_i=z}\{x_i - \hat{x}(z)\}\{x_i - \hat{x}(z)\}^{\mathrm{T}}$ and $\hat{S}_{xY(z)} = (N_z-1)^{-1}\sum_{i:Z_i=z}\{x_i - \hat{x}(z)\}\{Y_i - \hat{Y}(z)\}$ be the sample covariance matrices of x_i with itself and Y_i under treatment z. Let $\hat{\gamma}_{\mathrm{R}}$ and $\hat{\gamma}_{\mathrm{F}}$ be the coefficients of x_i from the OLS fits of Y_i on $(1, x_i)$ and $(1, Z_i, x_i)$, respectively. Let $\hat{\gamma}_{\mathrm{L}} = p_0\hat{\gamma}_{\mathrm{L},1} + p_1\hat{\gamma}_{\mathrm{L},0}$, where $\hat{\gamma}_{\mathrm{L},z}$ is the coefficient of x_i from the OLS fit of Y_i on $(1, x_i)$ over the units under treatment z.

Proposition 1. For $\mathcal{D} = (Y_i, x_i, Z_i)_{i=1}^N$ from arbitrary data generating process,

$$\begin{aligned}
\hat{\tau}_* &= N_1^{-1}\sum_{i:Z_i=1}(Y_i - x_i^{\mathrm{T}}\hat{\gamma}_*) - N_0^{-1}\sum_{i:Z_i=0}(Y_i - x_i^{\mathrm{T}}\hat{\gamma}_*) = \hat{\tau}_{\mathrm{N}} - \hat{\tau}_x^{\mathrm{T}}\hat{\gamma}_*, \quad (* = \mathrm{R}, \mathrm{F}) \\
\hat{\tau}_{\mathrm{L}} &= N_1^{-1}\sum_{i:Z_i=1}(Y_i - x_i^{\mathrm{T}}\hat{\gamma}_{\mathrm{L},1}) - N_0^{-1}\sum_{i:Z_i=0}(Y_i - x_i^{\mathrm{T}}\hat{\gamma}_{\mathrm{L},0}) = \hat{\tau}_{\mathrm{N}} - \hat{\tau}_x^{\mathrm{T}}\hat{\gamma}_{\mathrm{L}},
\end{aligned}$$

where $\hat{\gamma}_{\mathrm{R}} = (S_x^2)^{-1}\hat{S}_{xY}$, $\hat{\gamma}_{\mathrm{F}} = \hat{\gamma}_{\mathrm{R}} - (1-1/N)^{-1}p_1p_0\hat{\tau}_{\mathrm{F}}(S_x^2)^{-1}\hat{\tau}_x$, and $\hat{\gamma}_{\mathrm{L},z} = (\hat{S}_{x(z)}^2)^{-1}\hat{S}_{xY(z)}$.

Proposition 1 entails only the algebraic properties of the OLS fits and holds under arbitrary outcome models. It unifies $\hat{\tau}_*$ $(* = \mathrm{R}, \mathrm{F}, \mathrm{L})$ as the difference-in-means estimators defined on the adjusted outcomes, or, equivalently, as $\hat{\tau}_{\mathrm{N}}$ with corrections based on the imbalance in means of the covariates.

Under $H_{0\mathrm{N}}$, FRT with $\hat{\tau}_{\mathrm{N}}$ does not preserve the correct type one error rates but FRT with $\hat{\tau}_{\mathrm{N}}/\tilde{\mathrm{se}}_{\mathrm{N}}$ does (Chung and Romano 2013; Ding and Dasgupta 2018). In the next section, we will extend the result to the nine covariate-adjusted test statistics in Table 1 and establish the properness of the four robustly-studentized t-statstics, namely $\hat{\tau}_*/\tilde{\mathrm{se}}_*$ $(* = \mathrm{N}, \mathrm{R}, \mathrm{F}, \mathrm{L})$, for testing $H_{0\mathrm{N}}$. Referred to them as the *robust t-statistics* hence. We will further show that among them, $\hat{\tau}_{\mathrm{L}}/\tilde{\mathrm{se}}_{\mathrm{L}}$ delivers the highest power under alternative hypotheses.

3. Asymptotic theory for FRTs for testing $\tau = 0$

3.1. Limiting distributions under complete randomization

We will develop in Theorems 1–4 the limiting distributions of the twelve test statistics in Table 1. The result elucidates both the asymptotic validity and power for testing $H_{0\mathrm{N}}$. Let $\bar{Y}(z) = N^{-1}\sum_{i=1}^N Y_i(z)$ and $S_z^2 = (N-1)^{-1}\sum_{i=1}^N\{Y_i(z) - \bar{Y}(z)\}^2$ be the finite-population mean and variance of $\{Y_i(z)\}_{i=1}^N$, respectively. Let $S_\tau^2 = (N-1)^{-1}\sum_{i=1}^N(\tau_i - \tau)^2$ be the finite-population variance of $(\tau_i)_{i=1}^N$. Let $S_{xY(z)} = (N-1)^{-1}\sum_{i=1}^N x_i Y_i(z)$ be the finite-population covariance matrix of $\{x_i, Y_i(z)\}_{i=1}^N$. Let $w_i(z) = (S_x^2)^{-1}x_i Y_i(z)$ with $\bar{w}(z) = N^{-1}\sum_{i=1}^N w_i(z) = (1-1/N)(S_x^2)^{-1}S_{xY(z)}$.

Condition 1. As $N \to \infty$, for $z = 0, 1$, (i) p_z has a limit in $(0, 1)$, (ii) the first two moments of $\{Y_i(1), Y_i(0), x_i\}_{i=1}^N$ have finite limits; S_x^2 and its limit are both positive definite; S_z^2 –

$S_{xY(z)}^{\mathrm{T}}(S_x^2)^{-1}S_{xY(z)}$ has a finite positive limit; the second moments of $\{w_i(1), w_i(0)\}_{i=1}^N$ have finite limits, and (iii) there exists a $c_0 < \infty$ independent of N such that $N^{-1}\sum_{i=1}^N Y_i^4(z) \leq c_0$, $N^{-1}\sum_{i=1}^N \|x_i\|_4^4 \leq c_0$, and $N^{-1}\sum_{i=1}^N \|w_i(z)\|_4^4 \leq c_0$.

Condition 1(ii) ensures S_τ^2 has a finite limit. We also use p_z, $\bar{Y}(z)$, τ, S_z^2, S_x^2, $S_{xY(z)}$, and S_τ^2 to denote their limiting values without introducing new symbols. The exact meaning should be clear from the context. Theorem 1 reviews the asymptotic distributions of the three unadjusted test statistics from Ding and Dasgupta (2018), and Theorems 2–4 extend them to the covariate-adjusted cases. Below, P_Z-a.s. indicates a statement that holds for almost all sequences of Z.

Theorem 1. Assume Condition 1 and complete randomization.

(a) $\sqrt{N}(\hat{\tau}_{\mathrm{N}} - \tau) \rightsquigarrow \mathcal{N}(0, v_{\mathrm{N}})$, and $\sqrt{N}\hat{\tau}_{\mathrm{N}}^{\pi} \rightsquigarrow \mathcal{N}(0, v_{\mathrm{N0}})$ P_Z-a.s., where $v_{\mathrm{N}} = p_1^{-1}S_1^2 + p_0^{-1}S_0^2 - S_\tau^2$ and $v_{\mathrm{N0}} = p_0^{-1}S_1^2 + p_1^{-1}S_0^2 + \tau^2$.

(b) $(\hat{\tau}_{\mathrm{N}} - \tau)/\hat{\mathrm{se}}_{\mathrm{N}} \rightsquigarrow \mathcal{N}(0, c'_{\mathrm{N}})$, and $(\hat{\tau}_{\mathrm{N}}/\hat{\mathrm{se}}_{\mathrm{N}})^{\pi} \rightsquigarrow \mathcal{N}(0,1)$ P_Z-a.s., where $c'_{\mathrm{N}} = v_{\mathrm{N}}/(v_{\mathrm{N0}} - \tau^2)$.

(c) $(\hat{\tau}_{\mathrm{N}} - \tau)/\tilde{\mathrm{se}}_{\mathrm{N}} \rightsquigarrow \mathcal{N}(0, c_{\mathrm{N}})$, and $(\hat{\tau}_{\mathrm{N}}/\tilde{\mathrm{se}}_{\mathrm{N}})^{\pi} \rightsquigarrow \mathcal{N}(0,1)$ P_Z-a.s., where $c_{\mathrm{N}} = v_{\mathrm{N}}/(v_{\mathrm{N}} + S_\tau^2) \leq 1$.

Building up intuitions for Theorem 1 helps to understand Theorems 2–4 for the covariate-adjusted cases below. The FRT procedure uses $\{(Y_i, Y_i)\}_{i=1}^N$ as the "pseudo potential outcomes" to generate the randomization distribution of $\hat{\tau}_{\mathrm{N}}$ represented by $\hat{\tau}_{\mathrm{N}}^{\pi}$, and ends up mixing $\{Y_i(1)\}_{i=1}^N$ and $\{Y_i(0)\}_{i=1}^N$ with proportions p_1 and p_0 in the absence of $H_{0\mathrm{F}}$. This results in the different forms of v_{N} and v_{N0}.

Computationally, $\hat{\tau}_{\mathrm{N}}^{\pi}$ is the coefficient of Z_π from the OLS fit of Y on $(1_N, Z_\pi)$, with $\tilde{\mathrm{se}}_{\mathrm{N}}^{\pi}$ and $\hat{\mathrm{se}}_{\mathrm{N}}^{\pi}$ as the classic and robust standard errors that satisfy $(\hat{\tau}_{\mathrm{N}}/\hat{\mathrm{se}}_{\mathrm{N}})^{\pi} = \hat{\tau}_{\mathrm{N}}^{\pi}/\hat{\mathrm{se}}_{\mathrm{N}}^{\pi}$ and $(\hat{\tau}_{\mathrm{N}}/\tilde{\mathrm{se}}_{\mathrm{N}})^{\pi} = \hat{\tau}_{\mathrm{N}}^{\pi}/\tilde{\mathrm{se}}_{\mathrm{N}}^{\pi}$. The "pseudo potential outcomes" rules out treatment effects and forces equal variances under treatment and control. This ensures the consistency of both $N(\hat{\mathrm{se}}_{\mathrm{N}}^{\pi})^2$ and $N(\tilde{\mathrm{se}}_{\mathrm{N}}^{\pi})^2$ for estimating the asymptotic variance of $\sqrt{N}\hat{\tau}_{\mathrm{N}}^{\pi}$ and thereby guarantees the convergence of $(\hat{\tau}_{\mathrm{N}}/\hat{\mathrm{se}}_{\mathrm{N}})^{\pi}$ and $(\hat{\tau}_{\mathrm{N}}/\tilde{\mathrm{se}}_{\mathrm{N}})^{\pi}$ to the standard normal irrespective of the true value of τ. The same intuition carries over to the covariate-adjusted statistics with the original potential outcomes replaced by the adjusted ones. We formalize the idea in Theorems 2–4.

Let $\gamma_z = (S_x^2)^{-1}S_{xY(z)}$ be the coefficient of x_i in the OLS fit of $Y_i(z)$ on $(1, x_i)$. Let $a_i(z) = Y_i(z) - \bar{Y}(z) - x_i^{\mathrm{T}}(p_1\gamma_1 + p_0\gamma_0)$ be the adjusted potential outcomes under treatment z, with finite-population mean zero and variance $S_{a(z)}^2$. Theorem 2 gives the asymptotic distributions of the three test statistics under the first strategy for covariate adjustment, with $\hat{\tau}_{\mathrm{R}}$, $\hat{\tau}_{\mathrm{R}}/\hat{\mathrm{se}}_{\mathrm{R}}$, and $\hat{\tau}_{\mathrm{R}}/\tilde{\mathrm{se}}_{\mathrm{R}}$ as the coefficient and t-values of Z_i from the OLS fit of e_i on $(1, Z_i)$, respectively.

Theorem 2. Assume Condition 1 and complete randomization.

(a) $\sqrt{N}(\hat{\tau}_{\mathrm{R}} - \tau) \rightsquigarrow \mathcal{N}(0, v_{\mathrm{R}})$, and $\sqrt{N}\hat{\tau}_{\mathrm{R}}^{\pi} \rightsquigarrow \mathcal{N}(0, v_{\mathrm{R0}})$ P_Z-a.s., where $v_{\mathrm{R}} = p_1^{-1}S_{a(1)}^2 + p_0^{-1}S_{a(0)}^2 - S_\tau^2$ and $v_{\mathrm{R0}} = p_0^{-1}S_{a(1)}^2 + p_1^{-1}S_{a(0)}^2 + \tau^2$.

(b) $(\hat{\tau}_{\mathrm{R}} - \tau)/\hat{\mathrm{se}}_{\mathrm{R}} \rightsquigarrow \mathcal{N}(0, c'_{\mathrm{R}})$, and $(\hat{\tau}_{\mathrm{R}}/\hat{\mathrm{se}}_{\mathrm{R}})^{\pi} \rightsquigarrow \mathcal{N}(0,1)$ P_Z-a.s., where $c'_{\mathrm{R}} = v_{\mathrm{R}}/(v_{\mathrm{R0}} - \tau^2)$.

(c) $(\hat{\tau}_{\text{R}} - \tau)/\tilde{\text{se}}_{\text{R}} \rightsquigarrow \mathcal{N}(0, c_{\text{R}})$, and $(\hat{\tau}_{\text{R}}/\tilde{\text{se}}_{\text{R}})^{\pi} \rightsquigarrow \mathcal{N}(0,1)$ P_Z-a.s., where $c_{\text{R}} = v_{\text{R}}/(v_{\text{R}} + S_\tau^2) \leq 1$.

Interestingly, Theorem 2 also holds for $\hat{\tau}_{\text{F}}$, $\hat{\tau}_{\text{F}}/\hat{\text{se}}_{\text{F}}$, and $\hat{\tau}_{\text{F}}/\tilde{\text{se}}_{\text{F}}$ under the second strategy, as the coefficient and t-values of Z_i from the OLS fit of Y_i on $(1, Z_i, x_i)$. This echos the numeric result from Proposition 1, which implies that the difference between $\hat{\gamma}_{\text{F}}$ and $\hat{\gamma}_{\text{R}}$ is of higher order under complete randomization.

Theorem 3. Theorem 2 holds if we replace all the subscripts R with F.

The asymptotic equivalence of $\hat{\tau}_{\text{R}}$ and $\hat{\tau}_{\text{F}}$ is perhaps no surprise after all, despite the distinction in procedure. Both statistics use a common coefficient, namely $\hat{\gamma}_{\text{R}}$ and $\hat{\gamma}_{\text{F}}$, to adjust the observed outcomes under both treatment and control and estimate this coefficient using the pooled data. Such practice, despite expeditious, can be problematic in experiments with unequal group sizes and heterogeneous treatment effects with respect to covariates (Freedman 2008).

Lin (2013)'s estimator, on the other hand, accommodates separate adjustments for outcomes under treatment and control evident from Proposition 1. Let $b_i(z) = Y_i(z) - \bar{Y}(z) - x_i^{\text{T}}\gamma_z$ be the adjusted potential outcomes under treatment-specific coefficient γ_z, with mean zero and finite-population variance $S^2_{b(z)}$. Let S^2_ξ be the finite-population variance of $\xi_i = b_i(1) - b_i(0)$. Theorem 4 gives the asymptotic distributions of $\hat{\tau}_{\text{L}}$, $\hat{\tau}_{\text{L}}/\hat{\text{se}}_{\text{L}}$, and $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$ under the second strategy, as the coefficient and t-values of Z_i from the OLS fit of Y_i on $(1, Z_i, x_i, Z_i x_i)$.

Theorem 4. Assume Condition 1 and complete randomization.

(a) $\sqrt{N}(\hat{\tau}_{\text{L}} - \tau) \rightsquigarrow \mathcal{N}(0, v_{\text{L}})$, and $\sqrt{N}\hat{\tau}^{\pi}_{\text{L}} \rightsquigarrow \mathcal{N}(0, v_{\text{L0}})$ P_Z-a.s., where $v_{\text{L}} = p_1^{-1}S^2_{b(1)} + p_0^{-1}S^2_{b(0)} - S^2_\xi$ and $v_{\text{L0}} = v_{\text{R0}} = v_{\text{F0}} = p_0^{-1}S^2_{a(1)} + p_1^{-1}S^2_{a(0)} + \tau^2$.

(b) $(\hat{\tau}_{\text{L}} - \tau)/\hat{\text{se}}_{\text{L}} \rightsquigarrow \mathcal{N}(0, c'_{\text{L}})$, and $(\hat{\tau}_{\text{L}}/\hat{\text{se}}_{\text{L}})^{\pi} \rightsquigarrow \mathcal{N}(0,1)$ P_Z-a.s., where $c'_{\text{L}} = v_{\text{L}}/(p_0^{-1}S^2_{b(1)} + p_1^{-1}S^2_{b(0)})$.

(c) $(\hat{\tau}_{\text{L}} - \tau)/\tilde{\text{se}}_{\text{L}} \rightsquigarrow \mathcal{N}(0, c_{\text{L}})$, and $(\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}})^{\pi} \rightsquigarrow \mathcal{N}(0,1)$ P_Z-a.s., where $c_{\text{L}} = v_{\text{L}}/(v_{\text{L}} + S^2_\xi) \leq 1$.

The asymptotic variance of $\hat{\tau}_{\text{L}}$ is less than or equal to $v_{\text{F}} = v_{\text{R}}$ (Lin 2013), but those of the randomization distributions are all equal, $v_{\text{L0}} = v_{\text{R0}} = v_{\text{F0}}$. Similar to the comments after Theorem 1, this is due to the mixing of the treated and control outcomes in the FRT procedure, which effectively results in covariate adjustment based on the pooled data even in constructing Lin (2013)'s estimator. In fact, Lemma S7 in the Supplementary Material gives a stronger result that $\hat{\tau}^{\pi}_*$ ($* = \text{R}, \text{F}, \text{L}$) are all asymptotically equivalent.

The asymptotic sampling distributions in Theorems 3 and 4 are not new (Freedman 2008; Lin 2013), but the randomization distributions are. Both the asymptotic sampling and randomization distributions of $\hat{\tau}_{\text{R}}$, $\hat{\tau}_{\text{R}}/\hat{\text{se}}_{\text{R}}$, and $\hat{\tau}_{\text{R}}/\tilde{\text{se}}_{\text{R}}$ in Theorem 2 are new. The analysis of randomization distributions builds upon the existing sampling distributions but requires additional technical tools, such as finite-population strong law of large numbers. We unify the existing and new results in the above four theorems to facilitate discussions on the asymptotic validity.

Technically, Condition 1 requires more moments than the usual asymptotic analysis of τ_* ($* = \text{N}, \text{F}, \text{L}$). This is due to the strong statement of the almost sure convergence of the randomization

distributions in Theorems 1–4. This is sufficient but unnecessary for showing that FRT controls the asymptotic type one error rate, which only requires that the quantiles of the asymptotic randomization distribution are greater than or equal to those of the asymptotic sampling distribution. We can use the "subsequence argument," a standard proving device for the bootstrap (van der vaart and Wellner 1996, Chapter 3.6), to relax the moment conditions. However, we keep the current version of Condition 1 to simplify the statements of the theorems and their proofs.

3.2. Asymptotic validity for testing $\tau = 0$

Theorems 1–4 establish the sampling and randomization distributions for all twelve test statistics in Table 1 as asymptotically normal. A statistic as such is proper under two-sided FRT if under $H_{0\text{N}}$, the asymptotic variance of its randomization distribution is greater than or equal to that of its sampling distribution for all $\mathcal{S}$. In general, v_*/v_{*0} and c'_* can be either greater or less than 1, suggesting the improperness of $\hat{\tau}_*$ and $\hat{\tau}_*/\hat{\text{se}}_*$ for $* = \text{N}, \text{R}, \text{F}, \text{L}$. On the other hand, $c_* \leq 1$, assuring the properness of $\hat{\tau}_*/\tilde{\text{se}}_*$ for $* = \text{N}, \text{R}, \text{F}, \text{L}$.

Corollary 1. Assume Condition 1 and complete randomization. The robust t-statistics $\hat{\tau}_*/\tilde{\text{se}}_*$ ($* = \text{N}, \text{R}, \text{F}, \text{L}$) are the only test statistics in Table 1 proper for testing $H_{0\text{N}}$ via FRT.

Corollary 1 highlights the necessity of robust studentization in constructing asymptotically valid FRT for testing $H_{0\text{N}}$. The other eight test statistics may also preserve the correct type one error rates asymptotically with additional conditions on (p_1, p_0) or $\mathcal{S}$. The former is within the control of the designer whereas the latter is not.

Corollary 2. Assume Condition 1 and complete randomization. As N goes to infinity,

(a) all twelve test statistics in Table 1 preserve the correct type one error rates if $p_1 = p_0 = 1/2$ or $\tau_i = \tau$ for all $i = 1, \ldots, N$;

(b) $\hat{\tau}_\text{N}$ and $\hat{\tau}_\text{N}/\hat{\text{se}}_\text{N}$ preserve the correct type one error rates if $S_1^2 = S_0^2$; $\hat{\tau}_\text{R}$, $\hat{\tau}_\text{R}/\hat{\text{se}}_\text{R}$, $\hat{\tau}_\text{F}$, and $\hat{\tau}_\text{F}/\hat{\text{se}}_\text{F}$ do so if $S_{a(1)}^2 = S_{a(0)}^2$; $\hat{\tau}_\text{L}$ and $\hat{\tau}_\text{L}/\hat{\text{se}}_\text{L}$ do so if $S_{b(1)}^2 = S_{b(0)}^2$.

3.3. Power under alternative hypotheses

A natural next question is the relative power of FRT with the four robust t-statistics under alternative hypotheses. Recall that Theorems 1–4 hold for arbitrary τ. A deviation from $H_{0\text{N}}$ shifts the center of $\hat{\tau}_*/\tilde{\text{se}}_*$ while leaving its asymptotic randomization distribution intact at $\mathcal{N}(0,1)$. The relative power thus depends on $|\hat{\tau}_*/\tilde{\text{se}}_*|$ under the alternative hypothesis. With $\hat{\tau}_*$ converging to τ in probability, the smaller the robust standard error, the higher the asymptotic relative power.

Corollary 3. Assume complete randomization and Condition 1. We have

$$\frac{\tilde{\text{se}}_*^2}{\tilde{\text{se}}_\text{N}^2} - \frac{p_1^{-1}S_{a(1)}^2 + p_0^{-1}S_{a(0)}^2}{p_1^{-1}S_1^2 + p_0^{-1}S_0^2} = o(1) \quad \text{for } * = \text{R and F}, \qquad \frac{\tilde{\text{se}}_\text{L}^2}{\tilde{\text{se}}_\text{N}^2} - \frac{p_1^{-1}S_{b(1)}^2 + p_0^{-1}S_{b(0)}^2}{p_1^{-1}S_1^2 + p_0^{-1}S_0^2} = o(1)$$

hold P_Z-a.s., with the limiting values of $\tilde{\text{se}}_{\text{L}}^2/\tilde{\text{se}}_*^2$ less than or equal to 1 for $* = \text{N}, \text{R}, \text{F}$.

Corollary 3 ensures $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$ to have the highest power asymptotically. The limiting values of $\tilde{\text{se}}_{\text{R}}^2/\tilde{\text{se}}_{\text{N}}^2$ and $\tilde{\text{se}}_{\text{F}}^2/\tilde{\text{se}}_{\text{N}}^2$, on the other hand, can be even greater than 1. This mirrors the asymptotic efficiency theory of point estimation and suggests $\hat{\tau}_{\text{R}}/\tilde{\text{se}}_{\text{R}}$ and $\hat{\tau}_{\text{F}}/\tilde{\text{se}}_{\text{F}}$ can be even less powerful than $\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}$ despite the extra use of covariates (Freedman 2008; Lin 2013). Lemma S6 in the Supplementary Material gives the technical details for the almost sure convergence in Corollary 3.

FRT with $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$, as a result, is finite-sample exact for testing $H_{0\text{F}}$, asymptotically valid for testing $H_{0\text{N}}$, and enjoys the highest power under alternatives, all irrespective of whether the linear model that generated it is correctly specified or not. It is thus our final recommendation for testing both $H_{0\text{F}}$ and $H_{0\text{N}}$ by FRT.

3.4. Confidence interval by inverting FRTs

We next extend the theory from testing hypotheses to constructing confidence intervals. This is conceptually straightforward given their duality. Consider using FRT to test $H_{0\text{N}}(c) : \tau = c$. We can pretend to be testing a strong null hypothesis of constant effect, $H_{0\text{F}}(c) : \tau_i = c$ for all $i = 1, \ldots, N$, and compute the p-value, denoted by $p_{\text{FRT}}(c)$, by using $Y - cZ$ as the fixed outcomes for FRT. Inverting a sequence of such FRTs yields

$$\text{CI}_{\text{FRT},\alpha} = \{c : p_{\text{FRT}}(c) \geq \alpha\}$$

as a tentative interval estimator for the average treatment effect τ. By duality, it is an asymptotic $1-\alpha$ confidence interval for τ if we use the robust t-statistics to perform the FRTs. Duality further suggests the one based on $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$ to have the smallest width asymptotically.

Alternatively, the robust Wald-type confidence intervals $(\hat{\tau}_* - q_{1-\alpha/2}\tilde{\text{se}}_*, \hat{\tau}_* + q_{1-\alpha/2}\tilde{\text{se}}_*)$ $(* = \text{N}, \text{F}, \text{R}, \text{L})$ cover τ with probability approaching $1 - \alpha$ as N goes to infinity, where $q_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal. These confidence intervals are asymptotically identical to $\text{CI}_{\text{FRT},\alpha}$ based on the robust t-statistics. They are convenient approximations for $\text{CI}_{\text{FRT},\alpha}$ which can be used as initial values in the grid search over c. We recommend using $\text{CI}_{\text{FRT},\alpha}$ based on $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$ because of its multiple guarantees: it has finite-sample exact coverage rate when $\tau_i = \tau$, has correct asymptotic coverage rate when the τ_i's vary, and has smaller width compared to the confidence interval without covariate adjustment.

4. Extensions to other experimental designs

4.1. Cluster randomization

Consider N units nested in M clusters of sizes n_i $(i = 1, \ldots, M; \sum_{i=1}^M n_i = N)$. The average cluster size is $\bar{n} = N/M$. Cluster randomization randomly assigns M_1 clusters to receive the treatment and the rest $M_0 = M - M_1$ clusters to receive the control. Let x_{ij} and $Y_{ij}(z)$ be the covariate

and potential outcomes for the jth unit in cluster i ($i = 1, \ldots, M;\ j = 1, \ldots, n_i$). The average treatment effect equals

$$\tau = N^{-1} \sum_{i=1}^{M} \sum_{j=1}^{n_i} \{Y_{ij}(1) - Y_{ij}(0)\} = M^{-1} \sum_{i=1}^{M} \{\tilde{Y}_{i\cdot}(1) - \tilde{Y}_{i\cdot}(0)\}, \tag{2}$$

where $\tilde{Y}_{i\cdot}(z) = \sum_{j=1}^{n_i} Y_{ij}(z)/\bar{n}$ defines the cluster total of potential outcomes scaled by $1/\bar{n}$.

Let Z_i be the treatment level received by cluster i and thus unit ij. The observed outcome for unit ij is $Y_{ij} = Z_i Y_{ij}(1) + (1 - Z_i) Y_{ij}(0)$. Let $\tilde{Y}_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}/\bar{n}$ and $\tilde{x}_{i\cdot} = \sum_{j=1}^{n_i} x_{ij}/\bar{n}$ be the scaled cluster totals of observed outcomes and covariates in cluster i. Then $\tilde{Y}_{i\cdot} = Z_i \tilde{Y}_{i\cdot}(1) + (1 - Z_i)\tilde{Y}_{i\cdot}(0)$ is the observed analog of $\tilde{Y}_{i\cdot}(z)$ under cluster randomization. This, together with the expression of τ from (2), suggests the equivalence of $(Z_i, \tilde{Y}_{i\cdot}, \tilde{x}_{i\cdot})_{i=1}^{M}$ to data from a complete randomization with potential outcomes $\{\tilde{Y}_{i\cdot}(1), \tilde{Y}_{i\cdot}(0)\}$ and average treatment effect τ (Middleton and Aronow 2015; Li and Ding 2017), and allows us to derive results in parallel with Theorems 1–4.

4.2. Stratified randomization

Consider N units in K strata of sizes $N_{[k]}$ ($k = 1, \ldots, K;\ \sum_{k=1}^{K} N_{[k]} = N$). Stratified randomization conducts an independent complete randomization in each stratum. Let $x_{[k]i}$, $Y_{[k]i}(z)$, and $Z_{[k]i}$ be the covariate, potential outcomes, and treatment indicator for the ith unit in stratum k ($k = 1, \ldots, K;\ i = 1, \ldots, N_{[k]}$). The average treatment effect equals

$$\tau = N^{-1} \sum_{k=1}^{K} \sum_{i=1}^{N_{[k]}} \{Y_{[k]i}(1) - Y_{[k]i}(0)\} = \sum_{k=1}^{K} \omega_{[k]} \tau_{[k]},$$

where $\omega_{[k]} = N_{[k]}/N$ and $\tau_{[k]} = N_{[k]}^{-1} \sum_{i=1}^{N_{[k]}} \{Y_{[k]i}(1) - Y_{[k]i}(0)\}$ define the proportion and the average treatment effect of the units in stratum k.

Assume $\hat{\tau}_{*[k]}$ and $\tilde{\text{se}}_{*[k]}$ as the basic estimator and robust standard error obtained from stratum k, where $*$ can be N, R, F, and L. The weighted average $\hat{\tau}_* = \sum_{k=1}^{K} \omega_{[k]} \hat{\tau}_{*[k]}$, with a slight abuse of notation, affords an intuitive estimator of τ with squared robust standard error $\tilde{\text{se}}_*^2 = \sum_{k=1}^{K} \omega_{[k]}^2 \tilde{\text{se}}_{*[k]}^2$. The abuse of notation causes little confusion because $\hat{\tau}_*$ and $\tilde{\text{se}}_*$ reduce to their definitions under complete randomization when $K = 1$. The properness of $\hat{\tau}_*/\tilde{\text{se}}_*$ for testing $H_{0\text{N}}$ is a direct application of Theorems 1–4.

Corollary 4. Assume stratified randomization and Condition 1 holds within all strata $k = 1, \ldots, K$. We have $(\hat{\tau}_* - \tau)/\tilde{\text{se}}_* \rightsquigarrow \mathcal{N}(0, c_*)$, and $(\hat{\tau}_*/\tilde{\text{se}}_*)^{\pi} \rightsquigarrow \mathcal{N}(0, 1)$ P_Z-a.s., where $c_* \leq 1$, for $* = \text{N}, \text{R}, \text{F}, \text{L}$.

Even if the original experiment is completely randomized, if a discrete covariate X is available, we can condition on the number of treated and control units landing in each stratum. The resulting assignment mechanism is identical to stratified randomization, such that we can permute the subvector of Z within each stratum of X as if the original experiment were stratified. This is

known as the *conditional randomization test*. Zheng and Zelen (2008) and Hennessy et al. (2016) perceived that they typically enhance the power if the covariates are predictive of the outcomes.

Among the four robust t-statistics, $\hat{\tau}_{\mathrm{N}}/\tilde{\mathrm{se}}_{\mathrm{N}}$ is the simplest and $\hat{\tau}_{\mathrm{L}}/\tilde{\mathrm{se}}_{\mathrm{L}}$ is the most powerful. Corollary 4 assumes $N_{[k]}$ goes to infinity for each k. With a large number of small strata, we need to modify the test statistic and the asymptotic scheme (Liu and Yang 2020). Since this involves different technical tools, we defer the technical details to future work.

4.3. Rerandomization

4.3.1. FRT with rerandomization

Morgan and Rubin (2012) proposed to use rerandomization to improve covariate balance in the design stage. We focus here on a special rerandomization that uses the Mahalanobis distance between covariate means as the balance criterion, known as ReM. The designer accepts a treatment vector Z if and only if

$$\mathcal{A}: \quad \hat{\tau}_x^{\mathrm{T}}\{\mathrm{cov}(\hat{\tau}_x)\}^{-1}\hat{\tau}_x < a \tag{3}$$

for a predetermined constant a. Let $\mathcal{Z}_a = \{\boldsymbol{z} : \ \boldsymbol{z} \in \mathcal{Z} \text{ satisfies (3)}\}$ be the set of acceptable assignments under threshold a. The sampling distribution of the test statistic T is uniform over $\{T(\boldsymbol{z}, Y(\boldsymbol{z}), X) : \boldsymbol{z} \in \mathcal{Z}_a\}$. FRT under ReM proceeds by permuting Z within $\mathcal{Z}_a$ and computes the p-value as

$$p_{\mathrm{FRT},\mathcal{A}} = |\mathcal{Z}_a|^{-1} \sum_{\pi: Z_\pi \in \mathcal{Z}_a} 1\{|T(Z_\pi, Y, X)| \geq |T(Z, Y, X)|\}. \tag{4}$$

It compares the observed value of T to its randomization distribution under ReM, denoted by $T^{\pi|\mathcal{A}}$. Under ReM in (3), $p_{\mathrm{FRT},\mathcal{A}}$ is finite-sample exact for $H_{0\mathrm{F}}$ for arbitrary T. Of interest is its large-sample validity for testing $H_{0\mathrm{N}}$, which relies on the asymptotic distributions of the test statistic. Theorem 5 summarizes the results based on the additional notation below.

Let $\epsilon \sim \mathcal{N}(0,1)$ and $\mathcal{L} \sim D_1 \mid (\|D\|_2^2 \leq a)$, where $D = (D_1, \ldots, D_J)^{\mathrm{T}} \sim \mathcal{N}(0_J, I_J)$, be independent standard and truncated normals, respectively, and let $r_{J,a} = P(\chi^2_{J+2} \leq a)/P(\chi^2_J \leq a) \in (0,1]$ be the variance of $\mathcal{L}$. Let $\mathcal{U}(\rho) = (1-\rho^2)^{1/2} \cdot \epsilon + \rho \cdot \mathcal{L}$ be a linear combination of ϵ and $\mathcal{L}$ for $\rho \in [0,1]$ with mean 0 and variance $v(\rho) = 1 - (1 - r_{J,a})\rho^2$. Recall v_* and v_{*0} in Theorems 1–4 as the asymptotic variances of $\sqrt{N}\hat{\tau}_*$ and $\sqrt{N}\hat{\tau}_*^{\pi}$ under complete randomization, with $v_{\mathrm{R}} = v_{\mathrm{F}}$ and $v_{\mathrm{R}0} = v_{\mathrm{F}0}$. Let $\rho_*^2 = 1 - v_{\mathrm{L}}/v_*$ and $\rho_{*0}^2 = 1 - v_{\mathrm{L}0}/v_{*0}$ for $* = \mathrm{N}, \mathrm{R}, \mathrm{F}$, with $\rho_{\mathrm{R}0} = \rho_{\mathrm{F}0} = 0$.

Theorem 5. Assume Condition 1, ReM in design, and $p_{\mathrm{FRT},\mathcal{A}}$ in (4) in analysis.

(a) $\sqrt{N}(\hat{\tau}_{\mathrm{N}} - \tau) \rightsquigarrow v_{\mathrm{N}}^{1/2} \cdot \mathcal{U}(\rho_{\mathrm{N}})$, $(\hat{\tau}_{\mathrm{N}} - \tau)/\hat{\mathrm{se}}_{\mathrm{N}} \rightsquigarrow (c'_{\mathrm{N}})^{1/2} \cdot \mathcal{U}(\rho_{\mathrm{N}})$, and $(\hat{\tau}_{\mathrm{N}} - \tau)/\tilde{\mathrm{se}}_{\mathrm{N}} \rightsquigarrow c_{\mathrm{N}}^{1/2} \cdot \mathcal{U}(\rho_{\mathrm{N}})$; $\sqrt{N}\hat{\tau}_{\mathrm{N}}^{\pi|\mathcal{A}} \rightsquigarrow v_{\mathrm{N}0}^{1/2} \cdot \mathcal{U}(\rho_{\mathrm{N}0})$, $(\hat{\tau}_{\mathrm{N}}/\hat{\mathrm{se}}_{\mathrm{N}})^{\pi|\mathcal{A}} \rightsquigarrow \mathcal{U}(\rho_{\mathrm{N}0})$, and $(\hat{\tau}_{\mathrm{N}}/\tilde{\mathrm{se}}_{\mathrm{N}})^{\pi|\mathcal{A}} \rightsquigarrow \mathcal{U}(\rho_{\mathrm{N}0})$ hold P_Z-a.s..

(b) $\sqrt{N}(\hat{\tau}_* - \tau) \rightsquigarrow v_*^{1/2} \cdot \mathcal{U}(\rho_*)$, $(\hat{\tau}_* - \tau)/\hat{\mathrm{se}}_* \rightsquigarrow (c'_*)^{1/2} \cdot \mathcal{U}(\rho_*)$, and $(\hat{\tau}_* - \tau)/\tilde{\mathrm{se}}_* \rightsquigarrow c_*^{1/2} \cdot \mathcal{U}(\rho_*)$;

$\sqrt{N}\hat{\tau}_*^{\pi|\mathcal{A}} \rightsquigarrow \mathcal{N}(0, v_{*0})$, $(\hat{\tau}_*/\hat{\text{se}}_*)^{\pi|\mathcal{A}} \rightsquigarrow \mathcal{N}(0,1)$, and $(\hat{\tau}_*/\tilde{\text{se}}_*)^{\pi|\mathcal{A}} \rightsquigarrow \mathcal{N}(0,1)$ hold P_Z-a.s. $(* = \text{R}, \text{F})$.

(c) $\hat{\tau}_\text{L}$, $\hat{\tau}_\text{L}/\hat{\text{se}}_\text{L}$, and $\hat{\tau}_\text{L}/\tilde{\text{se}}_\text{L}$ have identical sampling and randomization distributions as under complete randomization in Theorem 4.

Compare Theorem 5 under ReM with Theorems 1–4 under complete randomization. The asymptotic sampling and randomization distributions of $\hat{\tau}_\text{N}$, $\hat{\tau}_\text{N}/\hat{\text{se}}_\text{N}$, and $\hat{\tau}_\text{N}/\tilde{\text{se}}_\text{N}$ change to non-normal. The asymptotic sampling distributions of $\hat{\tau}_*$, $\hat{\tau}_*/\hat{\text{se}}_*$, and $\hat{\tau}_*/\tilde{\text{se}}_*$ $(* = \text{R}, \text{F})$ change to non-normal, whereas their asymptotic randomization distributions remain the same. ReM does not affect these two asymptotic randomization distributions because of the asymptotic independence between $\hat{\tau}_*^\pi$ $(* = \text{R}, \text{F})$ and $\hat{\tau}_x^\pi$. The asymptotic sampling and randomization distributions of $\hat{\tau}_\text{L}$, $\hat{\tau}_\text{L}/\hat{\text{se}}_\text{L}$, and $\hat{\tau}_\text{L}/\tilde{\text{se}}_\text{L}$ all remain unchanged. ReM does not affect them because of the asymptotic independence between $\hat{\tau}_\text{L}$ and $\hat{\tau}_x$ and that between $\hat{\tau}_\text{L}^\pi$ and $\hat{\tau}_x^\pi$.

In the case of symmetric yet non-normal limiting distributions as those of $\hat{\tau}_*$, $\hat{\tau}_*/\hat{\text{se}}_*$, and $\hat{\tau}_*/\tilde{\text{se}}_*$ for $* = \text{N}, \text{R}, \text{F}$, determination of properness entails comparisons of not only the variances but also all the central quantile ranges. A test statistic T is proper under a two-sided FRT if T has wider or equal central quantile ranges than $T^{\pi|\mathcal{A}}$ for all quantiles.

Corollary 5. Assume Condition 1, ReM in design, and $p_{\text{FRT},\mathcal{A}}$ in (4) in analysis. The covariate-adjusted robust t-statistics $\hat{\tau}_*/\tilde{\text{se}}_*$ $(* = \text{R}, \text{F}, \text{L})$ are the only test statistics in Table 1 proper for testing $H_{0\text{N}}$ via FRT.

Compare Corollary 5 with Corollary 1 to see that the unadjusted $\hat{\tau}_\text{N}/\tilde{\text{se}}_\text{N}$, whereas proper under complete randomization, is no longer proper under ReM due to the non-normal limiting distribution of $\hat{\tau}_\text{N}^\pi$. Cohen and Fogarty (2020) also noticed this phenomenon and gave a numeric example. They proposed a prepivoting approach to improve studentization. We do not pursue that direction given $\hat{\tau}_\text{N}/\tilde{\text{se}}_\text{N}$ is inferior to $\hat{\tau}_\text{L}/\tilde{\text{se}}_\text{L}$ even under complete randomization. The three covariate-adjusted robust t-statistics, namely $\hat{\tau}_\text{R}/\tilde{\text{se}}_\text{R}$, $\hat{\tau}_\text{F}/\tilde{\text{se}}_\text{F}$, and $\hat{\tau}_\text{L}/\tilde{\text{se}}_\text{L}$, are the only options in Table 1 proper for testing $H_{0\text{N}}$ under ReM. Covariate adjustment is thus essential for securing properness under ReM in addition to robust studentization. Similarly, FRT with $\hat{\tau}_\text{L}/\tilde{\text{se}}_\text{L}$ delivers the highest power among the three proper statistics. It is thus our recommendation for conducting FRT under ReM.

4.3.2. Rerandomization with tiers of covariates

When covariates have different levels of importance for the outcomes, Morgan and Rubin (2012) proposed using ReM with differing criteria for different tiers of covariates. The resulting FRT permutes the treatment vector Z within the subset of $\mathcal{Z}$'s that satisfy the tiered balance criteria. The sampling and randomization distributions of the twelve test statistics in Table 1 parallel those in Theorem 5, with $\hat{\tau}_\text{L}/\tilde{\text{se}}_\text{L}$ being the most powerful among the proper options. It is thus our recommendation for this extension as well. We omit the technical details due to its repetitiveness.

4.3.3. FRT in case of designer-analyzer information discrepancy

Discussion so far assumes the analyzer and the designer use the same covariates x_i and threshold a for doing ReM in the design and analysis stages, respectively. An interesting question, also a real concern in practice, is what if the designer and the analyzer do not communicate? Bruhn and McKenzie (2009) and Heckman and Karapakula (2019) gave examples arising in field experiments in economics; Li and Ding (2020) discussed optimal covariate adjustment based on estimation precision.

A relatively easy case is that the analyzer has access to additional covariates beyond those used in the design of ReM. Using FRT under this ReM with $\hat{\tau}_{\mathrm{L}}/\tilde{\mathrm{se}}_{\mathrm{L}}$ is again our final recommendation in this case. A more challenging case is that the analyzer is either unaware of the ReM in the design stage or does not have access to all covariates used in the ReM. In the absence of full information about the design, Heckman and Karapakula (2019) proposed to use the maximum p-value from the worst-case FRT over a set of designs consistent with the available information. Without completely specifying these designs, an alternative option is to use p_{FRT} in (1) such that the analysis coincides with that under complete randomization. Under $H_{0\mathrm{F}}$, the finite-sample exactness is lost unless the original experiment is indeed completely randomized. Of interest is how such information discrepancy further affects the test's properness for testing $H_{0\mathrm{N}}$.

Keep x_i as the covariates the analyzer uses in the analysis stage, and let d_i be the covariates the designer used for conducting ReM in the design stage, possibly different from x_i. The designer accepts an allocation if $\hat{\tau}_d^{\mathrm{T}}\{\mathrm{cov}(\hat{\tau}_d)\}^{-1}\hat{\tau}_d < a$ with $\hat{\tau}_d$ being the difference in means of d_i. The analyzer, on the other hand, proceeds with p_{FRT} in (1) using test statistics formed based on x_i, yielding randomization distributions identical to those under complete randomization.

Focus on the twelve test statistics in Table 1 for the rest of the discussion, with randomization distributions readily available from Theorems 1–4. The possible discrepancy between d_i and x_i causes the sampling distributions to deviate from those in Theorem 5. Let $S^2_{z|d}$, $S^2_{a(z)|d}$, $S^2_{b(z)|d}$, $S^2_{\tau|d}$, and $S^2_{\xi|d}$ be the finite-population variances of the linear projections of $Y_i(z)$, $a_i(z)$, $b_i(z)$, τ_i, and ξ_i onto d_i, respectively, for $z = 0, 1$. Let

$$\rho^2_{\mathrm{N}|d} = \frac{p_1^{-1}S^2_{1|d} + p_0^{-1}S^2_{0|d} - S^2_{\tau|d}}{p_1^{-1}S_1^2 + p_0^{-1}S_0^2 - S_\tau^2}, \qquad \rho^2_{\mathrm{R}|d} = \rho^2_{\mathrm{F}|d} = \frac{p_1^{-1}S^2_{a(1)|d} + p_0^{-1}S^2_{a(0)|d} - S^2_{\tau|d}}{p_1^{-1}S^2_{a(1)} + p_0^{-1}S^2_{a(0)} - S_\tau^2},$$
$$\rho^2_{\mathrm{L}|d} = \frac{p_1^{-1}S^2_{b(1)|d} + p_0^{-1}S^2_{b(0)|d} - S^2_{\xi|d}}{p_1^{-1}S^2_{b(1)} + p_0^{-1}S^2_{b(0)} - S_\xi^2}$$

be the squared multiple correlations between $\hat{\tau}_*$ ($* = \mathrm{N}, \mathrm{R}, \mathrm{F}, \mathrm{L}$) and $\hat{\tau}_d$.

Proposition 2. Assume Condition 1 holds for $\{Y_i(0), Y_i(1), x_i'\}_{i=1}^N$ with x_i' being the union of the x_i and d_i, and ReM using the d_i's. For $* = \mathrm{N}, \mathrm{R}, \mathrm{F}, \mathrm{L}$, we have

$$\sqrt{N}(\hat{\tau}_* - \tau) \rightsquigarrow v_*^{1/2} \cdot \mathcal{U}(\rho_{*|d}), \quad (\hat{\tau}_* - \tau)/\hat{\mathrm{se}}_* \rightsquigarrow (c_*')^{1/2} \cdot \mathcal{U}(\rho_{*|d}), \quad (\hat{\tau}_* - \tau)/\tilde{\mathrm{se}}_* \rightsquigarrow c_*^{1/2} \cdot \mathcal{U}(\rho_{*|d}).$$

Proposition 2 is a special case of Li and Ding (2020) and generalizes the sampling distribution results in Theorem 5 to allow for distinct covariates for the design and analysis stages, respectively. The resulting distributions are in general scaled $\mathcal{U}$ distributions as linear combinations of independent standard and truncated normals. In particular, $\rho_{\mathrm{L}|d} = 0$ if x_i can linearly represent d_i, rendering the limiting distributions of $\hat{\tau}_{\mathrm{L}}$, $\hat{\tau}_{\mathrm{L}}/\hat{\mathrm{se}}_{\mathrm{L}}$, and $\hat{\tau}_{\mathrm{L}}/\tilde{\mathrm{se}}_{\mathrm{L}}$ identical to those under complete randomization in Theorem 4. The following corollary holds by comparing Proposition 2 with Theorems 1–4.

Corollary 6. Assume Condition 1 holds for $\{Y_i(0), Y_i(1), x_i'\}_{i=1}^N$ with x_i' being the union of the x_i and d_i, ReM using the d_i's in design, and p_{FRT} in (1) in analysis. The robust t-statistics $\hat{\tau}_*/\tilde{\mathrm{se}}_*$ ($* = \mathrm{N}, \mathrm{R}, \mathrm{F}, \mathrm{L}$) are the only test statistics in Table 1 proper for testing $H_{0\mathrm{N}}$ via FRT.

Ironically, a less informed analysis recovers the properness of $\hat{\tau}_{\mathrm{N}}/\tilde{\mathrm{se}}_{\mathrm{N}}$ under ReM by restoring its asymptotic randomization distribution back to the standard normal. Nevertheless, this properness comes at the cost of being overly conservative. With the four robust t-statistics, p_{FRT} in (1) remains asymptotically valid under ReM even if the analyzer has only partial information on the covariates the designer used to form the balance criterion.

Further, the asymptotic randomization distributions of $\hat{\tau}_*/\tilde{\mathrm{se}}_*$ ($* = \mathrm{R}, \mathrm{F}, \mathrm{L}$) remain unchanged in computing p_{FRT} in (1) and $p_{\mathrm{FRT},\mathcal{A}}$ in (4). It might thus be tempting to ignore the rerandomization and conduct unrestricted FRT in the analysis stage whatsoever, even when exact information is available. We do not encourage such practice given its lack of finite-sample exactness under $H_{0\mathrm{F}}$ in the first place.

5. Connections

5.1. Connection with the super-population framework

It is also conventional to view the experimental units as random samples from a super population (Tsiatis et al. 2008; Berk et al. 2013; Bugni et al. 2018; Negi and Wooldridge 2020; Ye et al. 2020). We now extend the discussion to this framework and examine the operating characteristics of the proposed strategies for conducting FRT on random potential outcomes. Assume $\{Y_i(0), Y_i(1), Z_i, x_i\}_{i=1}^N$ are independent and identically distributed (IID) samples from a super population. Let $\tau = E\{Y_i(1) - Y_i(0)\}$ be the population average treatment effect with a slight abuse of notation. The goal is to test

$$H_0 : \tau = 0$$

as the analog of $H_{0\mathrm{N}}$ in the finite-population setting.

Recall $p_z = N_z/N$ as the treatment proportions for $z = 0, 1$. Without introducing new notation, let p_z also denote $P(Z_i = z)$. Let

$$\mu_z = E\{Y_i(z)\}, \quad \sigma_z^2 = \mathrm{var}\{Y_i(z)\}, \quad \mu_x = E(x_i), \quad \sigma_x^2 = \mathrm{cov}(x_i), \quad \sigma_{xY(z)} = \mathrm{cov}\{x_i, Y_i(z)\}$$

be the first two moments of the potential outcomes and covariates. We impose the following conditions under the super-population framework.

Condition 2. $\{Y_i(0), Y_i(1), Z_i, x_i\}_{i=1}^N$ are IID draws from the population with (i) $E(\|x_i\|_4^4) \leq \infty$, $E\{Y_i^4(z)\} \leq \infty$, $E\{\|x_i Y_i(z)\|_4^4\} \leq \infty$, $\sigma_z^2 - \sigma_{xY(z)}^{\mathrm{T}}(\sigma_x^2)^{-1}\sigma_{xY(z)} > 0$ for $z = 0, 1$, and (ii) $Z_i \perp\!\!\!\perp \{Y_i(0), Y_i(1), x_i\}$.

With a slight abuse of notation, let $\gamma_z = (\sigma_x^2)^{-1}\sigma_{xY(z)}$ be the coefficient of x_i in the population OLS fit of $Y_i(z)$ on $(1, x_i)$, and let $a_i(z) = Y_i(z) - \mu_z - (x_i - \mu_x)^{\mathrm{T}}(p_1\gamma_1 + p_0\gamma_0)$ and $b_i(z) = Y_i(z) - \mu_z - (x_i - \mu_x)^{\mathrm{T}}\gamma_z$ be the residuals. Let $\sigma_{a(z)}^2$ and $\sigma_{b(z)}^2$ be the variances of $a_i(z)$ and $b_i(z)$, respectively. Negi and Wooldridge (2020) reviewed the asymptotic distributions of the estimators: $\sqrt{N}(\hat{\tau}_* - \tau) \rightsquigarrow \mathcal{N}(0, v_*)$ $(* = \mathrm{N}, \mathrm{R}, \mathrm{F}, \mathrm{L})$, with

$$v_{\mathrm{N}} = p_1^{-1}\sigma_1^2 + p_0^{-1}\sigma_0^2, \qquad v_{\mathrm{R}} = v_{\mathrm{F}} = p_1^{-1}\sigma_{a(1)}^2 + p_0^{-1}\sigma_{a(0)}^2, \qquad v_{\mathrm{L}} = p_1^{-1}\sigma_{b(1)}^2 + p_0^{-1}\sigma_{b(0)}^2 + \Delta_{\bar{x}}, \quad (5)$$

where $\Delta_{\bar{x}} = (\gamma_1 - \gamma_0)^{\mathrm{T}}\sigma_x^2(\gamma_1 - \gamma_0)$. Technically, Negi and Wooldridge (2020) did not discuss $\hat{\tau}_{\mathrm{R}}$, but we can show that $\hat{\tau}_{\mathrm{R}}$ has the same asymptotic distribution as $\hat{\tau}_{\mathrm{F}}$. So we unify the results above. A key distinction from the finite-population case is that $\hat{\tau}_{\mathrm{L}}$ now has extra variability due to centering the covariates. In contrast, other estimators do not have this extra term $\Delta_{\bar{x}}$ because they remain unchanged whether or not we center the covariates.

The extra term due to centering goes away if we condition on all covariates. But if we stick to the IID assumption in Condition 2, we must modify the standard errors for Lin (2013)'s estimator to account for $\Delta_{\bar{x}}$ (Berk et al. 2013; Negi and Wooldridge 2020). In particular, define $\hat{\mathrm{se}}_{\mathrm{L}}^2$ and $\tilde{\mathrm{se}}_{\mathrm{L}}^2$ as the classic and robust standard errors squared plus $\hat{\Delta}_{\bar{x}}/N = \hat{\theta}^{\mathrm{T}} S_x^2 \hat{\theta}/N$, respectively, where S_x^2 is the sample covariance matrix of the x_i's and $\hat{\theta} = \hat{\gamma}_{\mathrm{L},1} - \hat{\gamma}_{\mathrm{L},0}$ is the coefficient of $Z_i(x_i - \bar{x})$ in the OLS fit of Y_i on $\{1, Z_i, x_i - \bar{x}, Z_i(x_i - \bar{x})\}$. We then use them to construct the studentized statistics and obtain in total twelve test statistics as in Table 1. Of interest is whether the resulting tests preserve the correct type one error rates under the super-population framework for testing H_0.

Theorem 6 gives the asymptotic sampling and randomization distributions of the twelve test statistics in Table 1. The same reasoning that underpins Corollaries 1 and 3 ensures the four robust t-statistics are the only options proper for testing H_0, with $\hat{\tau}_{\mathrm{L}}/\tilde{\mathrm{se}}_{\mathrm{L}}$ enjoying the highest power. Below, P_{S}-a.s. indicates a statement that holds for almost all sequences of $\{Y_i(0), Y_i(1), Z_i, x_i\}_{i=1}^N$.

Theorem 6. Assume Condition 2.

(a) $\sqrt{N}(\hat{\tau}_* - \tau) \rightsquigarrow \mathcal{N}(0, v_*)$, and $\sqrt{N}\hat{\tau}_*^{\pi} \rightsquigarrow \mathcal{N}(0, v_{*0})$ P_{S}-a.s., with v_* $(* = \mathrm{N}, \mathrm{R}, \mathrm{F}, \mathrm{L})$ defined in (5), $v_{\mathrm{N}0} = p_0^{-1}\sigma_1^2 + p_1^{-1}\sigma_0^2 + \tau^2$, and $v_{\mathrm{R}0} = v_{\mathrm{F}0} = v_{\mathrm{L}0} = p_0^{-1}\sigma_{a(1)}^2 + p_1^{-1}\sigma_{a(0)}^2 + \tau^2$.

(b) $(\hat{\tau}_* - \tau)/\hat{\mathrm{se}}_* \rightsquigarrow \mathcal{N}(0, c_*)$, and $(\hat{\tau}_*/\hat{\mathrm{se}}_*)^{\pi} \rightsquigarrow \mathcal{N}(0, 1)$ P_{S}-a.s., with $c_*' = v_*/(v_{*0} - \tau^2)$ for $* = \mathrm{N}, \mathrm{R}, \mathrm{F}$, and $c_{\mathrm{L}}' = v_{\mathrm{L}}/\{p_0^{-1}\sigma_{b(1)}^2 + p_1^{-1}\sigma_{b(0)}^2 + \Delta_{\bar{x}}\}$.

(c) $(\hat{\tau}_* - \tau)/\tilde{\mathrm{se}}_* \rightsquigarrow \mathcal{N}(0, 1)$, and $(\hat{\tau}_*/\tilde{\mathrm{se}}_*)^{\pi} \rightsquigarrow \mathcal{N}(0, 1)$ P_{S}-a.s. for $* = \mathrm{N}, \mathrm{R}, \mathrm{F}, \mathrm{L}$.

The sampling distributions, except for those indexed by R, are known, but the permutation distributions, especially those with corrections for $\Delta_{\bar{x}}$, are new. Intuitively, σ_z^2 mirrors the finite-population limit of S_z^2 in Condition 1 and affords the probability limit of S_z^2 under the super-population framework. The asymptotic variances of $\hat{\tau}_*$ $(* = \text{N}, \text{R}, \text{F})$ in Theorem 6 thus mirror their finite-population analogs in Theorem 1–3 without the conservativeness issue. In contrast, the asymptotic variance of $\hat{\tau}_\text{L}$ features the additional term $\Delta_{\bar{x}}$. The randomization distribution of $\hat{\Delta}_{\bar{x}}$ is of higher order such that the randomization distributions $(\hat{\tau}_\text{L}/\hat{\text{se}}_\text{L})^\pi$ and $(\hat{\tau}_\text{L}/\tilde{\text{se}}_\text{L})^\pi$ remain at $\mathcal{N}(0,1)$ even after this correction in the standard error.

Corollary 7. Assume Condition 2. The robustly-studentized $\hat{\tau}_*/\tilde{\text{se}}_*$ $(* = \text{N}, \text{R}, \text{F}, \text{L})$ are the only test statistics in Table 1 proper for testing H_0, with the limiting values of $\tilde{\text{se}}_\text{L}^2/\tilde{\text{se}}_*^2$ less than or equal to 1 for $* = \text{N}, \text{R}, \text{F}$ P_S-a.s..

Aronow et al. (2014) proposed an improvement on $\tilde{\text{se}}_\text{N}$ under the finite population framework, which can be extended to $\tilde{\text{se}}_*$ $(* = \text{R}, \text{F}, \text{L})$ as well. These alternatives, however, under estimate the asymptotic variances of $\hat{\tau}_*$ $(* = \text{N}, \text{R}, \text{F}, \text{L})$ under the super-population framework. We thus do not pursue them in our recommendation.

5.2. Connection with permutation tests for coefficients in linear models

We unified the twelve test statistics in Table 1 as outputs from OLS fits and permuted Z to compute the p-values via FRT. The procedures are similar in form to permutation tests based on linear models. Consider a linear model

$$Y = 1_N\alpha + Z\beta + X\gamma + \epsilon \tag{6}$$

that characterizes the treatment effect by the coefficient of Z. Testing the treatment effect reduces to testing $\beta = 0$. It is standard to use the t-test based on the classic or robust standard error. Permutation tests, on the other hand, afford compelling alternatives due to their accuracy in finite samples (Anderson and Legendre 1999; Anderson and Robinson 2001). We review here five existing permutation tests for testing $\beta = 0$ under model (6) and show their inferiority to FRT with $\hat{\tau}_\text{L}/\tilde{\text{se}}_\text{L}$ for testing the treatment effects.

5.2.1. A review of existing permutation tests for linear models

Without introducing new symbols, let $\hat{\tau}_\text{F}$ be the coefficient of Z from the OLS fit of model (6), with $\hat{\text{se}}_\text{F}$ and $\tilde{\text{se}}_\text{F}$ as the classic and robust standard errors. A recent proposal by DiCiccio and Romano (2017) used the $\hat{\tau}_\text{F}/\tilde{\text{se}}_\text{F}$ as the test statistic and constructed the reference distribution by permuting Z. It coincides with FRT with $\hat{\tau}_\text{F}/\tilde{\text{se}}_\text{F}$ in procedure and delivers asymptotically robust randomization test for $\beta = 0$ under the linear model framework. Our theory gives another justification of it for testing both $H_{0\text{F}}$ and $H_{0\text{N}}$ under the potential outcomes framework. Recall from Corollary 3 that FRT with $\hat{\tau}_\text{F}/\tilde{\text{se}}_\text{F}$ does not necessarily improve the power when testing the treatment effects.

A possible improvement is thus to add treatment-covariates interactions to model (6) such that the procedure coincides with the more powerful test based on $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$. Depending on whether we condition on the covariates or treat them as IID draws from a super population, we might need to modify the robust standard error as discussed in Section 5.1.

The rest four permutation tests, on the other hand, all used the classic $\hat{\tau}_{\text{F}}/\hat{\text{se}}_{\text{F}}$ as the test statistic yet employed distinct permutation schemes to generate the reference distributions. We review below their respective procedures to highlight the difference, and extend them to a unified theory with test statistics being the coefficients and the classic and robust t-statistics, respectively. For simplicity of presentation, we only give the p-value formulas based on their original proposals with $\hat{\tau}_{\text{F}}/\hat{\text{se}}_{\text{F}}$ as the test statistic. Those based on $\hat{\tau}_{\text{F}}/\tilde{\text{se}}_{\text{F}}$ and $\hat{\tau}_{\text{F}}$ can be derived similarly by replacing all occurrences of the classic standard errors with their robust counterparts or constant 1.

Freedman and Lane (1983) Let e be the residual vector from the OLS fit of Y on $(1_N, X)$ as the reduced model, with the fitted vector $Y - e$. Freedman and Lane (1983) proposed to permute e and obtain the p-value as follows:

FL-1: Permute $e = (e_1, \ldots, e_N)^{\text{T}}$ to obtain e_π, construct $Y^\pi = Y - e + e_\pi$ as the synthetic outcomes, and compute $(\hat{\beta}^\pi_{\text{FL}}, \hat{\text{se}}^\pi_{\text{FL}}, \tilde{\text{se}}^\pi_{\text{FL}})$ as the coefficient of Z and its classic and robust standard errors from the OLS fit of Y^π on $(1_N, Z, X)$.

FL-2: Compute $p_{\text{FL}} = |\Pi|^{-1} \sum_{\pi \in \Pi} 1(|\hat{\beta}^\pi_{\text{FL}}|/\hat{\text{se}}^\pi_{\text{FL}} \geq |\hat{\tau}_{\text{F}}|/\hat{\text{se}}_{\text{F}})$.

The pair (Z_i, x_i) remains intact under this procedure yet gets reshuffled under FRT. This exemplifies the difference between p_{FRT} and p_{FL}. Freedman and Lane (1983) first proposed the method without the intention of making it a formal test, but rather "an alternative interpretation of a reported significance level." It now becomes a standard permutation test for a coefficient in linear models (Anderson and Legendre 1999; Anderson and Robinson 2001).

Kennedy (1995) Kennedy (1995) also proposed to permute e with a slight modification of Freedman and Lane (1983). Let δ be the residual vector from the OLS fit of Z on $(1_N, X)$. The procedure proceeds as follows:

K-1: Permute e to obtain e_π, and compute $(\hat{\beta}^\pi_{\text{K}}, \hat{\text{se}}^\pi_{\text{K}}, \tilde{\text{se}}^\pi_{\text{K}})$ as the coefficient of δ and its classic and robust standard errors from the OLS fit of e_π on $(1_N, \delta)$.

K-2: Compute $p_{\text{K}} = |\Pi|^{-1} \sum_{\pi \in \Pi} 1(|\hat{\beta}^\pi_{\text{K}}|/\hat{\text{se}}^\pi_{\text{K}} \geq |\hat{\tau}_{\text{F}}|/\hat{\text{se}}_{\text{F}})$.

Kennedy (1995) and Anderson and Robinson (2001) pointed out that $\hat{\beta}^\pi_{\text{K}} = \hat{\beta}^\pi_{\text{FL}}$ due to the Frisch–Waugh–Lovell theorem whereas $\hat{\text{se}}^\pi_{\text{K}} \neq \hat{\text{se}}^\pi_{\text{FL}}$ and $\tilde{\text{se}}^\pi_{\text{K}} \neq \tilde{\text{se}}^\pi_{\text{FL}}$.

ter Braak (1992) ter Braak (1992) proposed to permute the residuals ϵ_{F} from the OLS fit of model (6), with $Y - \epsilon_{\text{F}}$ being the fitted vector. The procedure proceeds as follows:

Table 2: Five permutation tests for testing $\beta = 0$ in model (6).

procedure	T	model for computing T	T^π
DiCiccio and Romano (2017)	$\hat{\tau}_{\text{F}}/\tilde{\text{se}}_{\text{F}}$	$Y \sim 1_N + Z_\pi + X$	$\hat{\tau}^\pi_{\text{F}}/\tilde{\text{se}}^\pi_{\text{F}}$
Freedman and Lane (1983)	$\hat{\tau}_{\text{F}}/\hat{\text{se}}_{\text{F}}$	$Y - e + e_\pi \sim 1_N + Z + X$	$\hat{\beta}^\pi_{\text{FL}}/\hat{\text{se}}^\pi_{\text{FL}}$
Kennedy (1995)	same as above	$e_\pi \sim 1_N + \delta$	$\hat{\beta}^\pi_{\text{K}}/\hat{\text{se}}^\pi_{\text{K}}$
ter Braak (1992)	same as above	$Y - \epsilon_{\text{F}} + \epsilon_{\text{F},\pi} \sim 1_N + Z + X$	$(\hat{\beta}^\pi_{\text{TB}} - \hat{\tau}_{\text{F}})/\hat{\text{se}}^\pi_{\text{TB}}$
Manly (1997)	same as above	$Y_\pi \sim 1_N + Z + X$	$\hat{\beta}^\pi_{\text{M}}/\hat{\text{se}}^\pi_{\text{M}}$

TB-1 Permute ϵ_{F} to obtain $\epsilon_{\text{F},\pi}$, construct $Y^\pi = Y - \epsilon_{\text{F}} + \epsilon_{\text{F},\pi}$ as the synthetic outcomes, and compute $(\hat{\beta}^\pi_{\text{TB}}, \hat{\text{se}}^\pi_{\text{TB}}, \tilde{\text{se}}^\pi_{\text{TB}})$ as the coefficient of Z and its classic and robust standard errors from the OLS fit of Y^π on $(1_N, Z, X)$.

TB-2 Compute $p_{\text{TB}} = |\Pi|^{-1} \sum_{\pi\in\Pi} 1(|\hat{\beta}^\pi_{\text{TB}} - \hat{\tau}_{\text{F}}|/\hat{\text{se}}^\pi_{\text{TB}} \geq |\hat{\tau}_{\text{F}}|/\hat{\text{se}}_{\text{F}})$.

Compare TB-1 with FL-1 to see the difference in the models used for constructing the synthetic outcomes, with Freedman and Lane (1983) using the reduced model whereas ter Braak (1992) using the full model. Consequently, TB-2 differs from FL-2 in that $\hat{\beta}^\pi_{\text{TB}}$ must be centered by $\hat{\tau}_{\text{F}}$ under ter Braak (1992)'s procedure.

Manly (1997) Manly (1997) proposed to permute Y as follows:

M-1 Permute Y to obtain Y_π; compute $(\hat{\beta}^\pi_{\text{M}}, \hat{\text{se}}^\pi_{\text{M}}, \tilde{\text{se}}^\pi_{\text{M}})$ as the coefficient of Z and its classic and standard errors from the OLS fit of Y_π on $(1_N, Z, X)$.

M-2 Compute $p_{\text{M}} = |\Pi|^{-1} \sum_{\pi\in\Pi} 1(|\hat{\beta}^\pi_{\text{M}}|/\hat{\text{se}}^\pi_{\text{M}} \geq |\hat{\tau}_{\text{F}}|/\hat{\text{se}}_{\text{F}})$.

5.2.2. Finite-sample exactness for testing $H_{0\mathbf{F}}$ and properness for testing $H_{0\mathbf{N}}$

The procedures in Section 5.2.1 employed distinct permutation schemes to generate the reference distributions. DiCiccio and Romano (2017) permuted Z, Freedman and Lane (1983) and Kennedy (1995) permuted e, ter Braak (1992) permuted ϵ_{F}, and Manly (1997) permuted Y. Table 2 summarizes them.

An interesting question is how they compare with FRT and each other when applied to testing the treatment effects. Whereas the operating characteristics of DiCiccio and Romano (2017) follow directly from Theorem 3 due to its identicalness to FRT, those of the rest four tests are less straightforward and hinge on their respective reference distributions resulting from the distinct permutation schemes.

With a slight abuse of notation, we use the random variables under the reference distributions to represent the respective procedures for the rest of this section. The finite-sample exactness for testing $H_{0\text{F}}$, on the one hand, requires exact match of the reference distribution with the sampling distribution under $H_{0\text{F}}$. The difference in permutation schemes leaves the distributions of the unstudentized $\hat{\beta}^\pi_{\text{FL}}$, $\hat{\beta}^\pi_{\text{K}}$, $\hat{\beta}^\pi_{\text{TB}} - \hat{\tau}_{\text{F}}$, $\hat{\beta}^\pi_{\text{M}}$ and their studentized variants distinct from those of

$\hat{\tau}_{\text{F}}$, $\hat{\tau}_{\text{F}}/\hat{\text{se}}_{\text{F}}$, and $\hat{\tau}_{\text{F}}/\tilde{\text{se}}_{\text{F}}$ under $H_{0\text{F}}$, such that none of them is finite-sample exact for testing $H_{0\text{F}}$. Their properness and relative power for testing $H_{0\text{N}}$, on the other hand, depend on the asymptotic behaviors of the reference distributions, which are summarized below.

Theorem 7. Assume complete randomization and Condition 1. We have

(a) $\sqrt{N}\hat{\beta}^{\pi}_{*} \rightsquigarrow \mathcal{N}(0, v_{\text{F}0})$ for $* = \text{FL}, \text{K}$; $\quad \sqrt{N}(\hat{\beta}^{\pi}_{\text{TB}} - \hat{\tau}_{\text{F}}) \rightsquigarrow \mathcal{N}(0, v_{\text{F}0} - \tau^2)$; $\quad \sqrt{N}\hat{\beta}^{\pi}_{\text{M}} \rightsquigarrow \mathcal{N}(0, v_{\text{N}0})$;

(b) $\hat{\beta}^{\pi}_{*}/\hat{\text{se}}^{\pi}_{*} \rightsquigarrow \mathcal{N}(0,1)$ and $\hat{\beta}^{\pi}_{*}/\tilde{\text{se}}^{\pi}_{*} \rightsquigarrow \mathcal{N}(0,1)$ for $* = \text{FL}, \text{K}, \text{M}$;
$(\hat{\beta}^{\pi}_{\text{TB}} - \hat{\tau}_{\text{F}})/\hat{\text{se}}^{\pi}_{\text{TB}} \rightsquigarrow \mathcal{N}(0,1)$ and $(\hat{\beta}^{\pi}_{\text{TB}} - \hat{\tau}_{\text{F}})/\tilde{\text{se}}^{\pi}_{\text{TB}} \rightsquigarrow \mathcal{N}(0,1)$

hold P_Z-a.s., recalling $v_{\text{N}0}$ and $v_{\text{F}0}$ as the asymptotic variances of $\sqrt{N}\hat{\tau}^{\pi}_{\text{N}}$ and $\sqrt{N}\hat{\tau}^{\pi}_{\text{F}}$ defined in Theorems 1 and 3, respectively.

The asymptotic distribution of $\hat{\beta}^{\pi}_{\text{FL}}/\hat{\text{se}}^{\pi}_{\text{FL}}$ first appeared in Freedman and Lane (1983). Anderson and Robinson (2001) generalized it and gave a sketch of the proof for all four classic t-statistics in Table 2. We flesh out their proofs and furnish the new results on the unstudentized coefficients and their robustly-studentized versions.

Intuitively, the asymptotic variances of $\hat{\beta}^{\pi}_{\text{FL}} = \hat{\beta}^{\pi}_{\text{K}}$, $\hat{\beta}^{\pi}_{\text{TB}} - \hat{\tau}_{\text{F}}$, and $\hat{\beta}^{\pi}_{\text{M}}$ are proportional to the finite-population variances of e, ϵ_{F}, and Y, respectively, with increasing variability in the order of ϵ_{F}, e, and Y. In particular, the OLS fit of the reduced model adjusts for X such that e is less variable than Y; the OLS fit of the full model (6) further adjusts for Z such that ϵ_{F} is less variable than e. The four unstudentized test statistics elucidate the impact of different permutation schemes on the resulting reference distributions, with $\hat{\beta}^{\pi}_{\text{FL}}$ and $\hat{\beta}^{\pi}_{\text{K}}$ sharing the same limiting distribution with $\hat{\tau}^{\pi}_{\text{F}}$ and $\hat{\tau}^{\pi}_{\text{R}}$ irrespective of the true value of τ, whereas $\hat{\beta}^{\pi}_{\text{M}}$ sharing that with $\hat{\tau}^{\pi}_{\text{N}}$. The limiting distribution of $\hat{\beta}^{\pi}_{\text{TB}} - \hat{\tau}_{\text{F}}$, on the other hand, coincides with that of $\hat{\tau}^{\pi}_{\text{F}}$ or $\hat{\tau}^{\pi}_{\text{R}}$ if and only if $H_{0\text{N}}$ is true.

Despite the distinction between $\hat{\beta}^{\pi}_{*}$ for $* = \text{FL}, \text{K}, \text{TB}, \text{M}$, all eight studentized statistics are asymptotically standard normal and thus coincide with the limiting distributions of $(\hat{\tau}_{\text{F}}/\hat{\text{se}}_{\text{F}})^{\pi}$ and $(\hat{\tau}_{\text{F}}/\tilde{\text{se}}_{\text{F}})^{\pi}$ under FRT. Juxtapose Theorem 7 with Theorem 3 to see the four robustly-studentized variants as the only options proper for testing $H_{0\text{N}}$. We state the result in Corollary 8.

Corollary 8. Assume Condition 1 and complete randomization. The robustly studentized $\hat{\beta}^{\pi}_{\text{FL}}/\tilde{\text{se}}^{\pi}_{\text{FL}}$, $\hat{\beta}^{\pi}_{\text{K}}/\tilde{\text{se}}^{\pi}_{\text{K}}$, $(\hat{\beta}^{\pi}_{\text{TB}} - \hat{\tau}_{\text{F}})/\tilde{\text{se}}^{\pi}_{\text{TB}}$, and $\hat{\beta}^{\pi}_{\text{M}}/\tilde{\text{se}}^{\pi}_{\text{M}}$ are proper for testing $H_{0\text{N}}$ whereas the unstudentized and the classicly-studentized alternatives are not.

Despite the positive result in Corollary 8, we do not recommend these permutation tests for testing the treatment effects for several reasons. First, they are not finite-sample exact for $H_{0\text{F}}$ in the first place, such that properness under $H_{0\text{N}}$ can be a rather weak requirement. For instance, Manly (1997)'s permutation test with the unstudentized statistic yields a reference distribution far from the true distribution of $\hat{\tau}_{\text{F}}$ although the robust studentization repairs it asymptotically. Second, even within the scope of testing $H_{0\text{N}}$, the test statistic $\hat{\tau}_{\text{F}}/\tilde{\text{se}}_{\text{F}}$ is suboptimal compared to $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$ and may deliver even less power than $\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}$ despite the extra use of covariates. Further, Corollary 8

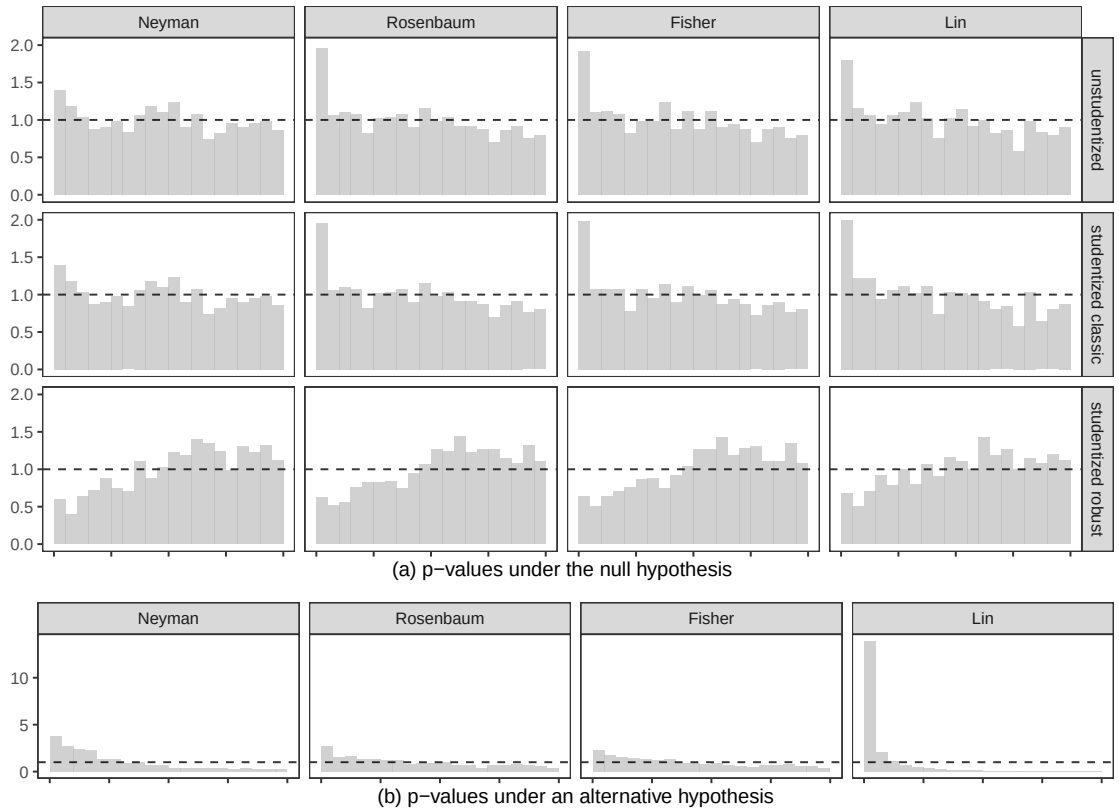

Figure 1: Empirical densities of the p_{FRT}'s under complete randomization with 20 bins in $(0, 1)$.

holds under complete randomization but may not extend to general designs even asymptotically. Overall, FRT is strictly superior for testing the treatment effects.

6. Simulation

6.1. Complete randomization

We first examine the large-sample validity of the twelve test statistics under $H_{0\text{N}}$. Consider a finite population of $N = 100$ units subjected to a complete randomization of size $(N_1, N_0) = (20, 80)$. For each i, we draw a univariate covariate x_i from $\text{Unif}(-1, 1)$ and generate potential outcomes as $Y_i(1) \sim \mathcal{N}(x_i^3, 1)$ and $Y_i(0) \sim \mathcal{N}(-x_i^3, 0.5^2)$. The $Y_i(1)$'s and $Y_i(0)$'s are centered to ensure $\tau = 0$. Fix $\{Y_i(1), Y_i(0), x_i\}_{i=1}^N$ in simulation. We draw a random permutation of N_1 1's and N_0 0's to obtain the observed outcomes and conduct FRTs. The procedure is repeated 1,000 times, with the p-values approximated by 500 independent permutations of the treatment vector in each replication. Figure 1(a) shows the p-values under $H_{0\text{N}}$. The four robust t-statistics, as shown in the last row, are the only ones that preserve the correct type one error rates. In fact, they are conservative, which is coherent with Theorems 1–4. All the other eight statistics yield type one error rates greater than the nominal levels and are thus not proper for testing $H_{0\text{N}}$.

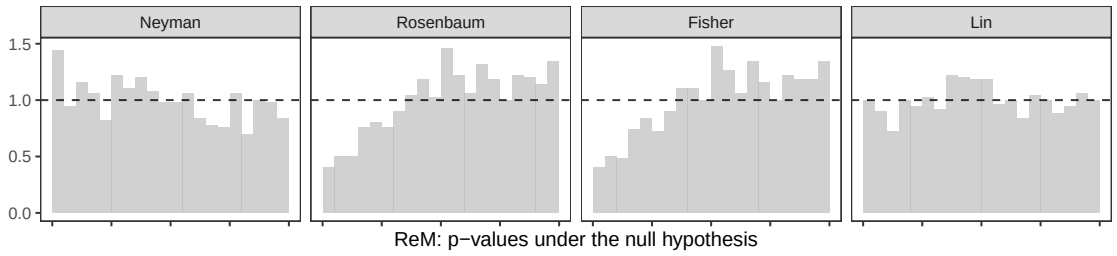

Figure 2: Empirical densities of the $p_{\text{FRT},\mathcal{A}}$'s based on the robust t-statistics with 20 bins in $(0, 1)$: ReM with a equaling the 0.05 quantile of χ^2_1.

We then evaluate the power of the four proper test statistics when $\tau \neq 0$. Take $Y_i(1) \sim \mathcal{N}(0.1 + x_i, 0.4^2)$ and $Y_i(0) \sim \mathcal{N}(-x_i, 0.1^2)$ for an alternative with τ close to 0.1 and inherit all the rest settings from the last paragraph. Figure 1(b) shows the p-values under the alternative. The theoretically most powerful $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$ indeed delivers the highest power among the four proper options. The tests based on $\hat{\tau}_{\text{F}}/\tilde{\text{se}}_{\text{F}}$ and $\hat{\tau}_{\text{R}}/\tilde{\text{se}}_{\text{R}}$, on the other hand, show even lower power than the unadjusted $\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}$. This is consistent with the theoretical results from Corollary 3 and concludes $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$ as our final recommendation for conducting FRT under complete randomization.

6.2. Rerandomization

We generate potential outcomes from $Y_i(1) = \epsilon_i$ and $Y_i(0) = x_i + \epsilon_i$ with ϵ_i and x_i as independent $\mathcal{N}(0, 1)$ and generate the treatment vector under ReM with a equaling the 0.05 quantile of χ^2_1. Inherit all the rest settings from the last subsection. Figure 2 shows the type one error rates of FRTs with the robust t-statistics. The unadjusted $\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}$ does not preserve the correct type one error rates whereas the other three adjusted statistics do. Compared with the results under complete randomization in Figure 1(a), FRTs with $\hat{\tau}_{\text{R}}/\tilde{\text{se}}_{\text{R}}$ and $\hat{\tau}_{\text{F}}/\tilde{\text{se}}_{\text{F}}$ become more conservative under ReM, due to their non-normal, narrower asymptotic sampling distributions. This is coherent with the theoretical results from Theorem 5.

7. Application

Chong et al. (2016) conducted a randomized experiment on 219 students of a rural secondary school in the Cajamarca district of Peru during the 2009 school year. They first provided the village with free iron supplements and trained the local staffs to distribute one free iron pill to any adolescent who requested one in person. They then randomly assigned the students to three arms with three different types of videos: in the first video, a popular soccer player was encouraging the use of iron supplements to maximize energy ("soccer" arm); in the second video, a physician was encouraging the use of iron supplements to improve overall health ("physician" arm); the third video did not mention iron and served as the control ("control" arm). The experiment was stratified by the class

Table 3: Re-analyzing the data from Chong et al. (2016). "N" corresponds to the unadjusted estimators and tests, and "L" corresponds to the covariate-adjusted estimators and tests.

(a) soccer versus control

	est	s.e.	p_{normal}	p_{frt}
class 1				
N	0.051	0.502	0.919	0.924
L	0.050	0.489	0.919	0.929
class 2				
N	−0.158	0.451	0.726	0.722
L	−0.176	0.452	0.698	0.700
class 3				
N	0.005	0.403	0.990	0.989
L	−0.096	0.385	0.803	0.806
class 4				
N	−0.492	0.447	0.271	0.288
L	−0.511	0.447	0.253	0.283
class 5				
N	0.390	0.369	0.291	0.314
L	0.443	0.318	0.164	0.186
all				
N	−0.051	0.204	0.802	0.800
L	−0.074	0.200	0.712	0.712

(b) physician versus control

	est	s.e.	p_{normal}	p_{frt}
class 1				
N	0.567	0.426	0.183	0.192
L	0.588	0.418	0.160	0.174
class 2				
N	0.193	0.438	0.659	0.666
L	0.265	0.409	0.517	0.523
class 3				
N	1.305	0.494	0.008	0.012
L	1.501	0.462	0.001	0.003
class 4				
N	−0.273	0.413	0.508	0.515
L	−0.313	0.417	0.454	0.462
class 5				
N	−0.050	0.379	0.895	0.912
L	−0.067	0.279	0.811	0.816
all				
N	0.406	0.202	0.045	0.047
L	0.463	0.190	0.015	0.017

level from 1 to 5. The three group sizes within classes are shown in the matrix below:

$$\begin{array}{l} \\ \text{soccer} \\ \text{physician} \\ \text{control} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{class 1} & \text{class 2} & \text{class 3} & \text{class 4} & \text{class 5} \end{array} \\ \begin{pmatrix} 16 & 19 & 15 & 10 & 10 \\ 17 & 20 & 15 & 11 & 10 \\ 15 & 19 & 16 & 12 & 10 \end{pmatrix} \end{array}.$$

One outcome of interest is the average grades in the third and fourth quarters of 2009, and an important background covariate is the anemia status at baseline. We make pairwise comparisons of the "soccer" arm versus the "control" arm and the "physician" arm versus the "control" arm. We also compare FRTs with and without adjusting for the covariate of baseline anemia status. We use their data set to illustrate FRTs under complete randomization and stratified randomization. The ten subgroup analyses within each class level use FRTs for complete randomization. The two overall analyses averaging over all class levels use FRTs for stratified randomization.

Table 3 shows the point estimators, the robust standard errors, the p-values based on large-sample approximations of the robust t-statistics, and the p-values based on FRTs. In most strata, covariate adjustment decreases the standard errors since the baseline anemia status is predictive of the outcome. Two exceptions are the pairwise comparison of the "soccer" arm versus the "control" arm within class 2 and the pairwise comparison of the "physician" arm versus the "control" arm within class 4, with differences both in the third digit after the decimal point. This is likely due

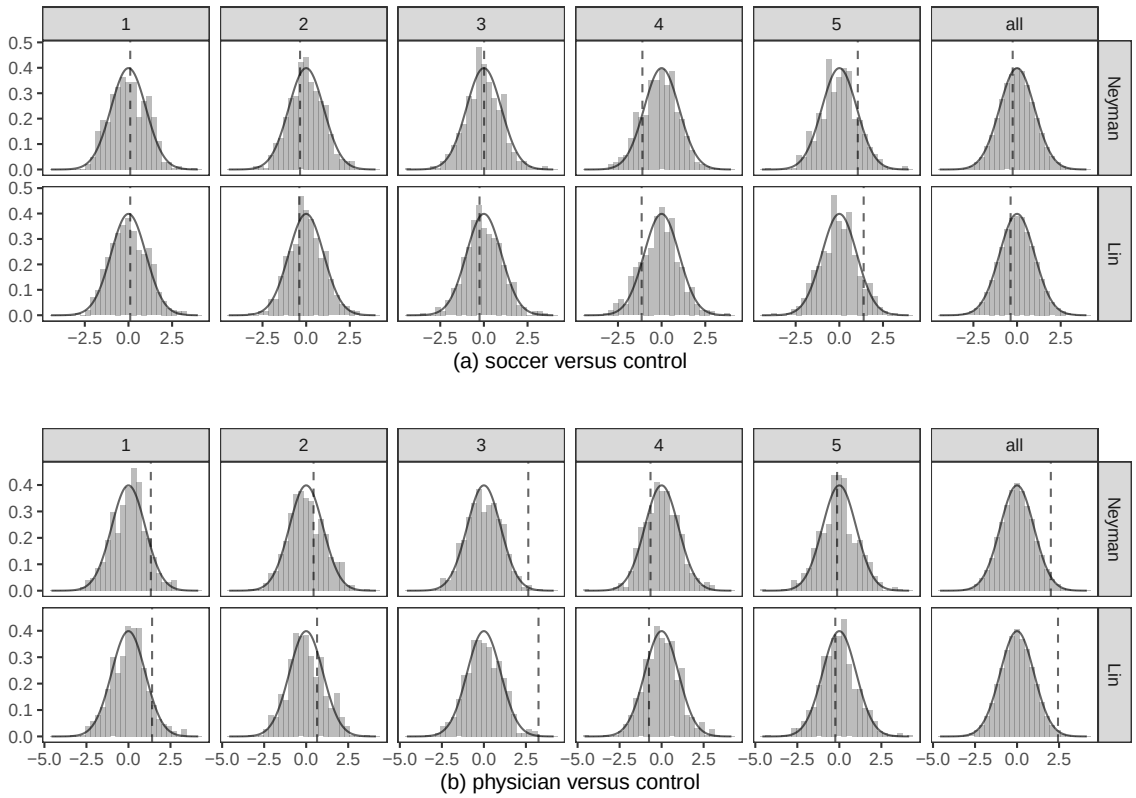

Figure 3: Randomization distributions based on 5×10^4 Monte Carlo simulations versus the asymptotic distributions $\mathcal{N}(0, 1)$.

to the small group sizes within these strata, leaving the asymptotic approximations dubious. The p-values from the large-sample approximations and FRTs are close with the latter being slightly larger in most cases. Based on the theory, the p-values based on FRTs should be trusted more given their additional guarantee of finite-sample exactness under the strong null hypothesis. This becomes important in this example given the relatively small group sizes within strata.

Bind and Rubin (2020) suggested reporting not only the p-values but also the randomization distributions of the test statistics when conducting FRT. Echoing their recommendation, we show in Figure 3 the histograms of the randomization distributions of the robust t-statistics alongside the asymptotic approximations. The discrepancy is quite clear in the subgroup analyses yet becomes unnoticeable after averaged over all class levels. Overall, the p-values based on large-sample approximations do not differ substantially from those based on FRTs in this application. The two approaches yield coherent conclusions: the video with a physician telling the benefits of iron supplements improved the academic performance and the effect was most significant among students in class 3; in contrast, the video with a popular soccer player telling the benefits of the iron supplements did not have any significant effect.

Table 4: Final recommendations for FRT and test statistic $\hat{\tau}_*/\tilde{\text{se}}_*$ in different experiments with $* = \text{L}$ if with covariates and $* = \text{N}$ if without.

design	no covariates	covariates	other comments
complete randomization	$* = \text{N}$	$* = \text{L}$	
cluster randomization	$* = \text{N}$	$* = \text{L}$	use cluster total outcomes
stratified randomization	$* = \text{N}$	$* = \text{L}$	weighted average over strata
ReM, complete design information		$* = \text{L}$	
ReM, incomplete design information	$* = \text{N}$	$* = \text{L}$	use p_{FRT} not $p_{\text{FRT},\mathcal{A}}$

8. Discussion

Echoing Fisher (1935), Proschan and Dodd (2019), Young (2019), and Bind and Rubin (2020), we believe FRT should be the default choice for analyzing experimental data given its flexibility to accommodate complex randomization schemes and arbitrary outcome generating processes. We established in this paper the theory for covariate adjustment in FRT under complete randomization, cluster randomization, stratified randomization, and rerandomization using the Mahalanobis distance, respectively, with final recommendations of the test statistics summarized in Table 4. Equipped with the finite-sample exactness under the strong null hypothesis, the recommended FRTs promise an additional guarantee under the weak null hypothesis and strictly dominate the analogs based on large-sample approximations.

We conjecture that the strategy of appropriately studentizing an efficient, covariate-adjusted estimator works for FRT in general experiments as well (e.g., Dasgupta et al. 2015; Lu 2016; Mukerjee et al. 2018; Middleton 2018; Fogarty 2018a,b). This strategy works for estimators with normal limiting distributions and may also work for estimators with non-normal limiting distributions as shown in the asymptotic theory of rerandomization. Cohen and Fogarty (2020)'s prepivoting approach may work more broadly but we leave the general theory to future research.

We focused on procedures based on OLS. It is of great interest to extend the theory to high dimensional settings (Bloniarz et al. 2016; Lei and Ding 2020), nonlinear models (Zhang et al. 2008; Moore and van der Laan 2009; Moore et al. 2011; Jiang et al. 2019; Guo and Basse 2020), and even estimators based on machine learning algorithms (Wager et al. 2016; Wu and Gagnon-Bartsch 2018; Farrell et al. 2020; Chen et al. 2020).

If the main parameter of interest is the average treatment effect, the asymptotic theory inevitably involves some moment conditions. Without these conditions, the inference becomes challenging (Bahadur and Savage 1956), and FRT may not control type one error rates even asymptotically with heavy-tailed outcomes. An alternative class of FRTs use rank statistics to gain robustness with respect to outliers (Lehmann 1975; Rosenbaum 2002). Although different rank statistics always work under the strong null hypothesis, they in general target parameters other than the average treatment effect (e.g., Rosenbaum 1999, 2003; Chung and Romano 2016). Chung and Romano (2016) proposed to studentize the Wilcoxon statistic in a permutation test, shedding light on the general theory of FRT with rank statistics.

References

M. J. Anderson and P. Legendre. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation*, 62:271–303, 1999.

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

P. Aronow, D. Green, and D. Lee. Sharp bounds on the variance in randomized experiments. *Annals of Statistics*, 42:850–871, 2014.

S. Athey, D. Eckles, and G. W. Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, 113:230–240, 2018.

R. R. Bahadur and L. J. Savage. The nonexistence of certain statistical procedures in nonparametric problems. *Annals of Mathematical Statistics*, 27:1115–1122, 1956.

R. Berk, E. Pitkin, L. Brown, A. Buja, E. George, and L. Zhao. Covariance adjustments for the analysis of randomized field experiments. *Evaluation Review*, 37:170–196, 2013.

M. A. C. Bind and D. B. Rubin. When possible, report a Fisher-exact P value and display its underlying null randomization distribution. *Proceedings of the National Academy of Sciences of the United States of America*, 117:19151–19158, 2020.

A. Bloniarz, H. Liu, C. Zhang, J. Sekhon, and B. Yu. Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 113:7383–7390, 2016.

D. R. Brillinger, L. V. Jones, and J. W. Tukey. The management of weather resources. Technical report, US Government Printing Office, Washington, DC, 1978.

M. Bruhn and D. McKenzie. In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, 1:200–232, 2009.

F. A. Bugni, I. A. Canay, and A. M. Shaikh. Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, 113:1784–1796, 2018.

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

M. D. Cattaneo, B. R. Frandsen, and R. Titiunik. Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate. *Journal of Causal Inference*, 3:1–24, 2015.

X. Chen, Y. Liu, S. Ma, and Z. Zhang. Efficient estimation of general treatment effects using neural networks with a diverging number of confounders. *arXiv preprint arXiv:2009.07055*, 2020.

A. Chong, I. Cohen, E. Field, E. Nakasone, and M. Torero. Iron deficiency and schooling attainment in Peru. *American Economic Journal: Applied Economics*, 8:222–55, 2016.

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

E. Chung and J. P. Romano. Asymptotically valid and exact permutation tests based on two-sample U-statistics. *Journal of Statistical Planning and Inference*, 168:97–105, 2016.

P. L. Cohen and C. B. Fogarty. Gaussian prepivoting for finite population causal inference. *https://arxiv.org/abs/2002.06654*, 2020.

T. Dasgupta, N. Pillai, and D. B. Rubin. Causal inference from 2^K factorial designs by using potential outcomes. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 77:727–753, 2015.

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

P. Ding. The Frisch–Waugh–Lovell theorem for standard errors. *Statistics and Probability Letters*, 168:108945, 2020.

P. Ding and T. Dasgupta. A randomization-based perspective of analysis of variance: a test statistic robust to treatment effect heterogeneity. *Biometrika*, 105:45–56, 2018.

F. Eicker. Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 59–82. Berkeley, CA: University of California Press, 1967.

M. H. Farrell, T. Liang, and S. Misra. Deep neural networks for estimation and inference. *Econometrica*, page in press, 2020.

R. A. Fisher. *The Design of Experiments*. Edinburgh, London: Oliver and Boyd, 1st edition, 1935.

C. B. Fogarty. On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80:1035–1056, 2018a.

C. B. Fogarty. Regression assisted inference for the average treatment effect in paired experiments. *Biometrika*, 105:994–1000, 2018b.

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

D. A. Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40:180–193, 2008.

W.A. Fuller. Some design properties of a rejective sampling procedure. *Biometrika*, 96:933–944, 2009.

M. H. Gail, W. Y. Tan, and S. Piantadosi. Tests for no treatment effect in randomized clinical trials. *Biometrika*, 75:57–64, 1988.

P. Ganong and S. Jäger. A permutation test for the regression kink design. *Journal of the American Statistical Association*, 113:494–504, 2018.

K. Guo and G. Basse. The generalized Oaxaca–Blinder estimator. *arXiv preprint arXiv:2004.11615*, 2020.

J. Hájek. Some extensions of the Wald–Wolfowitz–Noether theorem. *Annals of Mathematical Statistics*, 32:506–523, 1961.

J. Heckman and G. Karapakula. The Perry preschoolers at late midlife: A study in design-specific inference. *NBER Working Paper No. 25888*, 2019.

J. Hennessy, T. Dasgupta, L. Miratrix, C. Pattanayak, and P. Sarkar. A conditional randomization test to account for covariate imbalance in randomized experiments. *Journal of Causal Inference*, 4:61–80, 2016.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In Lucien M. Le Cam and Jerzy Neyman, editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233. Berkeley, California: University of California Press, 1967.

A. Janssen. Studentized permutation tests for non-iid hypotheses and the generalized Behrens–Fisher problem. *Statistics and Probability Letters*, 36:9–21, 1997.

F. Jiang, L. Tian, H. Fu, T. Hasegawa, and L. J. Wei. Robust alternatives to ANCOVA for estimating the treatment effect via a randomized comparative study. *Journal of the American Statistical Association*, 114:1854–1864, 2019.

P. E. Kennedy. Randomization tests in econometrics. *Journal of Business and Economic Statistics*, 13:85–94, 1995.

E. L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day, Inc., 1975.

L. Lei and P. J. Bickel. An assumption-free exact test for fixed-design linear models with exchangeable errors. *Biometrika*, page in press, 2020.

L. Lei and P. Ding. Regression adjustment in completely randomized experiments with a diverging number of covariates. *Biometrika*, page in press, 2020.

X. Li and P. Ding. General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112:1759–1169, 2017.

X. Li and P. Ding. Rerandomization and regression adjustment. *Journal of the Royal Statistical Society, Series B (Methodological)*, 82:241–268, 2020.

X. Li, P. Ding, and D. B. Rubin. Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 115:9157–9162, 2018.

W. Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Annals of Applied Statistics*, 7:295–318, 2013.

H. Liu and Y. Yang. Regression-adjusted average treatment effect estimates in stratified randomized experiments. *Biometrika*, page in press, 2020.

J. Lu. Covariate adjustment in randomization-based causal inference for 2^K factorial designs. *Statistics and Probability Letters*, 119:11–20, 2016.

J. G. MacKinnon and M. D. Webb. Randomization inference for difference-in-differences with few treated clusters. *Journal of Econometrics*, 218:435–450, 2020.

B. F. J. Manly. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall, 1997.

J. A. Middleton. A unified theory of regression adjustment for design-based inference. *arXiv preprint arXiv:1803.06011*, 2018.

J. A. Middleton and P. M. Aronow. Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Statistics, Politics and Policy*, 6:39–75, 2015.

K. L. Moore and M. J. van der Laan. Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Statistics in Medicine*, 28:39–64, 2009.

K. L. Moore, R. Neugebauer, T. Valappil, and M. J. van der Laan. Robust extraction of covariate information to improve estimation efficiency in randomized trials. *Statistics in Medicine*, 30: 2389–2408, 2011.

K. L. Morgan and D. B. Rubin. Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, 40:1263–1282, 2012.

R. Mukerjee, T. Dasgupta, and D. B. Rubin. Using standard tools from finite population sampling to improve causal inference for complex experiments. *Journal of the American Statistical Association*, 113:868–881, 2018.

A. Negi and J. M. Wooldridge. Revisiting regression adjustment in experiments with heterogeneous treatment effects. *Econometric Reviews*, page in press, 2020.

J. Neyman. On the application of probability theory to agricultural experiments (with discussion). *Statistical Science*, 5:465–472, 1923.

J. Neyman. Statistical problems in agricultural experimentation (with discussion). *Supplement to the Journal of the Royal Statistical Society*, 2:107–180, 1935.

K. Ottoboni, F. Lewis, and L. Salmaso. An empirical comparison of parametric and permutation tests for regression analysis of randomized experiments. *Statistics in Biopharmaceutical Research*, 10:264–273, 2018.

M. Pauly, E. Brunner, and F. Konietschke. Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 77:461–473, 2015.

M. A. Proschan and L. E. Dodd. Re-randomization tests in clinical trials. *Statistics in Medicine*, 38:2292–2302, 2019.

J. Raz. Testing for no effect when estimating a smooth function by nonparametric regression: a randomization approach. *Journal of the American Statistical Association*, 85:132–138, 1990.

J. P. Romano. On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85:686–692, 1990.

P. R. Rosenbaum. Reduced sensitivity to hidden bias at upper quantiles in observational studies with dilated treatment effects. *Biometrics*, 55:560–564, 1999.

P. R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17:286–327, 2002.

P. R. Rosenbaum. Exact confidence intervals for nonconstant effects by inverting the signed rank test. *American Statistician*, 57:132–138, 2003.

A. J. Stephens, E. J. Tchetgen Tchetgen, and V. De Gruttola. Flexible covariate-adjusted exact tests of randomized treatment effects with application to a trial of HIV education. *Annals of Applied Statistics*, 7:2106–2137, 2013.

C. J. F. ter Braak. Permutation versus bootstrap significance tests in multiple regression and ANOVA. In K.H. Jöckel, G. Rothe, and W. Sendler, editors, *Bootstrapping and Related Techniques*, pages 79–85. Berlin: Springer-Verlag, 1992.

A. A. Tsiatis, M. Davidian, M. Zhang, and X. Lu. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in Medicine*, 27:4658–4677, 2008.

J. W. Tukey. Tightening the clinical trial. *Controlled Clinical Trials*, 14:266–285, 1993.

A. W. van der vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer Verlag, 1996.

S. Wager, W. Du, J. Taylor, and R. J. Tibshirani. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 113:12673–12678, 2016.

H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48:817–838, 1980.

E. Wu and J. A. Gagnon-Bartsch. The LOOP estimator: Adjusting for covariates in randomized experiments. *Evaluation Review*, 42:458–488, 2018.

J. Wu and P. Ding. Randomization tests for weak null hypotheses in randomized experiments. *Journal of American Statistical Association*, 105:in press, 2020.

T. Ye, J. Shao, and Q. Zhao. Principles for covariate adjustment in analyzing randomized clinical trials. *arXiv preprint arXiv:2009.11828*, 2020.

A. Young. Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *Quarterly Journal of Economics*, 134:557–598, 2019.

M. Zhang, A. A. Tsiatis, and M. Davidian. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64:707–715, 2008.

L. Zheng and M. Zelen. Multi-center clinical trials: Randomization and ancillary statistics. *Annals of Applied Statistics*, 2:582–600, 2008.

Supplementary Material

Section S1 reviews the notation and some algebraic facts that hold for arbitrary data generating process. We omit the proofs because they are straightforward. It contains Lemma S1 on the univariate OLS fit, a basic yet powerful tool in later proofs when coupled with the Frisch–Waugh–Lovell theorems for both the regression coefficients and standard errors (Ding 2020). When referring to the Frisch–Waugh–Lovell theorems, we will simply say "by FWL."

Section S2 reviews the central limit theorems under complete randomization and random permutation, and gives a new finite population strong law of large numbers that works under not only simple random sampling and complete randomization but also rejective sampling and ReM (Fuller 2009; Morgan and Rubin 2012).

Section S3 gives the proofs of the main results under complete randomization.

Section S4 gives the proofs of the results under ReM.

Section S5 gives the proofs of the results related to the extensions to the super-population framework and permutation tests based on linear models.

S1. Notation and algebraic facts

Let 1_m and 0_m be the $m \times 1$ vectors of ones and zeros, respectively, and let I_m be the $m \times m$ identity matrix. We suppress the subscript m when the dimensions are clear from the context.

For $(u_i, v_i)_{i=1}^N$, where u_i and v_i are $m \times 1$ and $k \times 1$ vectors, respectively, let

$$\bar{u} = N^{-1} \sum_{i=1}^N u_i, \quad S_u^2 = (N-1)^{-1} \sum_{i=1}^N (u_i - \bar{u})(u_i - \bar{u})^{\mathrm{T}}, \quad S_{uv} = (N-1)^{-1} \sum_{i=1}^N (u_i - \bar{u})(v_i - \bar{v})^{\mathrm{T}}$$

be the finite-population mean and covariance matrices of $(u_i)_{i=1}^N$ with itself and $(v_i)_{i=1}^N$, respectively. The matrices S_u^2 and S_{uv} degenerate to the finite-population variance and covariance for $m = k = 1$. We suppress the "finite-population" when no confusion would arise. Let $\lambda = (N-1)/N$ be the scaling factor to accommodate the difference in normalizing by N or $N-1$.

S1.1. Potential outcomes and covariates

Assume standardized covariates with mean $\bar{x} = 0_J$ and covariance matrix $S_x^2 = I_J$ throughout the finite-population analysis to simplify the presentation. Let $X_j = (x_{1j}, \ldots, x_{Nj})^{\mathrm{T}}$ be the jth column of X with mean $N^{-1} \sum_{i=1}^N x_{ij} = 0$ and variance $(N-1)^{-1} \sum_{i=1}^N x_{ij}^2 = 1$. All procedures discussed in this paper are invariant to non-degenerate transformations of the covariates and thus unaffected by the standardization.

Recall $\gamma_z = (S_x^2)^{-1} S_{xY(z)} = (N-1)^{-1} \sum_{i=1}^N x_i Y_i(z)$ as the coefficient of x_i in the OLS fit of $Y_i(z)$ on $(1, x_i)$. Recall S_z^2, $S_{a(z)}^2$, $S_{b(z)}^2$, S_τ^2, and S_ξ^2 as the finite-population variances of $Y_i(z)$, $a_i(z) = Y_i(z) - \bar{Y}(z) - x_i^{\mathrm{T}}(p_1\gamma_1 + p_0\gamma_0)$, $b_i(z) = Y_i(z) - \bar{Y}(z) - x_i^{\mathrm{T}}\gamma_z$, $\tau_i = Y_i(1) - Y_i(0)$, and

$\xi_i = b_i(1) - b_i(0)$, respectively. Let

$$\gamma = p_1\gamma_1 + p_0\gamma_0, \quad S^2 = p_1S_1^2 + p_0S_0^2, \quad S_a^2 = p_1S_{a(1)}^2 + p_0S_{a(0)}^2, \quad S_b^2 = p_1S_{b(1)}^2 + p_0S_{b(0)}^2.$$

We have $a_i(z) = Y_i(z) - \bar{Y}(z) - x_i^{\mathrm{T}}\gamma$,

$$S_{a(z)}^2 = S_z^2 + \|\gamma\|_2^2 - 2\gamma_z^{\mathrm{T}}\gamma, \quad S_a^2 = S^2 - \|\gamma\|_2^2, \quad S_{b(z)}^2 = S_z^2 - \|\gamma_z\|_2^2, \quad S_\xi^2 = S_\tau^2 - \|\gamma_1 - \gamma_0\|_2^2, \quad \text{(S1)}$$

with $S_{a(z)}^2 - S_{b(z)}^2 = \|\gamma_z - \gamma\|_2^2 \geq 0$ for $z = 0, 1$.

S1.2. Observed outcomes and OLS fits

Let $\hat{Y}(z) = N_z^{-1}\sum_{i:Z_i=z} Y_i$ and $\hat{S}_z^2 = (N_z - 1)^{-1}\sum_{i:Z_i=z}\{Y_i - \hat{Y}(z)\}^2$ be the mean and variance of $\{Y_i : Z_i = z\}$, respectively. Let $\hat{\tau}_x = \hat{x}(1) - \hat{x}(0)$ be the difference in means of the covariates under treatment and control, where $\hat{x}(z) = N_z^{-1}\sum_{i:Z_i=z} x_i$. Centered covariates ensure

$$\hat{x}(1) = p_0\hat{\tau}_x, \quad \hat{x}(0) = -p_1\hat{\tau}_x. \quad \text{(S2)}$$

Let $\hat{S}_{x(z)}^2 = (N_z - 1)^{-1}\sum_{i:Z_i=z}\{x_i - \hat{x}(z)\}\{x_i - \hat{x}(z)\}^{\mathrm{T}}$ and $\hat{S}_{xY(z)} = (N_z - 1)^{-1}\sum_{i:Z_i=z}\{x_i - \hat{x}(z)\}\{Y_i - \hat{Y}(z)\}$ be the covariance matrices of $\{x_i : Z_i = z\}$ with itself and $\{Y_i : Z_i = z\}$.

Let $\hat{Y} = N^{-1}\sum_{i=1}^N Y_i$ and $\hat{S}^2 = (N-1)^{-1}\sum_{i=1}^N (Y_i - \hat{Y})^2$ be the mean and variance of $(Y_i)_{i=1}^N$, respectively. They satisfy

$$\hat{Y} = p_1\hat{Y}(1) + p_0\hat{Y}(0), \quad \hat{S}^2 = \frac{N_1 - 1}{N-1}\hat{S}_1^2 + \frac{N_0 - 1}{N-1}\hat{S}_0^2 + \frac{N}{N-1}p_0p_1\hat{\tau}_{\mathrm{N}}^2. \quad \text{(S3)}$$

Let $\hat{S}_{xY} = (N-1)^{-1}\sum_{i=1}^N x_iY_i$ be the covariance matrix of $\{(x_i, Y_i)\}_{i=1}^N$. The OLS fit of Y on $(1_N, X)$ has coefficients $(\hat{Y}, \hat{\gamma}_{\mathrm{R}})$ and residuals $e = (e_1, \ldots, e_N)^{\mathrm{T}}$ that satisfy

$$\begin{aligned}
&\hat{\gamma}_{\mathrm{R}} = (N-1)^{-1}\sum_{i=1}^N x_iY_i = \hat{S}_{xY}, \qquad e_i = Y_i - \hat{Y} - x_i^{\mathrm{T}}\hat{\gamma}_{\mathrm{R}} \quad (i = 1, \ldots, N), \\
&\hat{e} = N^{-1}\sum_{i=1}^N e_i = 0, \qquad \hat{S}_e^2 = (N-1)^{-1}\sum_{i=1}^N (e_i - \hat{e})^2 = (N-1)^{-1}\|e\|_2^2 = \hat{S}^2 - \|\hat{\gamma}_{\mathrm{R}}\|_2^2. \quad \text{(S4)}
\end{aligned}$$

Let $\hat{e}(z) = N_z^{-1}\sum_{i:Z_i=z} e_i$ and $\hat{S}_{e(z)}^2 = (N_z - 1)^{-1}\sum_{i:Z_i=z}\{e_i - \hat{e}(z)\}^2$ be the sample mean and variance of $\{e_i : Z_i = z\}$ for units in treatment group z.

Let $\epsilon_* = (\epsilon_{*,1}, \ldots, \epsilon_{*,N})^{\mathrm{T}}$ be the residuals from the OLS fit that generates $\hat{\tau}_*$, where $* = \mathrm{N}, \mathrm{R}, \mathrm{F}, \mathrm{L}$. Let $\hat{\epsilon}_*(z) = N_z^{-1}\sum_{i:Z_i=z}\epsilon_{*,i}$ and $\hat{S}_{*(z)}^2 = (N_z - 1)^{-1}\sum_{i:Z_i=z}\{\epsilon_{*,i} - \hat{\epsilon}_*(z)\}^2$ be the sample mean and variance of $\{\epsilon_{*,i} : Z_i = z\}$ under treatment z. They satisfy

$$\begin{aligned}
&\hat{S}_{\mathrm{N}(z)}^2 = \hat{S}_z^2, \quad \hat{S}_{\mathrm{R}(z)}^2 = \hat{S}_{e(z)}^2, \quad \hat{S}_{*(z)}^2 = \hat{S}_z^2 + \hat{\gamma}_*^{\mathrm{T}}\hat{S}_{x(z)}^2\hat{\gamma}_* - 2\hat{\gamma}_*^{\mathrm{T}}\hat{S}_{xY(z)} \quad \text{for } * = \mathrm{R}, \mathrm{F}, \\
&\hat{S}_{\mathrm{L}(z)}^2 = \hat{S}_z^2 + \hat{\gamma}_{\mathrm{L},z}^{\mathrm{T}}\hat{S}_{x(z)}^2\hat{\gamma}_{\mathrm{L},z} - 2\hat{\gamma}_{\mathrm{L},z}^{\mathrm{T}}\hat{S}_{xY(z)},
\end{aligned}$$

$$\hat{\mathrm{se}}_*^2 = \frac{N(N_1-1)}{(N-2)N_1N_0}\hat{S}^2_{*(1)} + \frac{N(N_0-1)}{(N-2)N_1N_0}\hat{S}^2_{*(0)} \quad \text{for } * = \text{N}, \text{R}, \text{F}, \text{L},$$
$$\tilde{\mathrm{se}}_*^2 = N_1^{-2}(N_1-1)\hat{S}^2_{*(1)} + N_0^{-1}(N_0-1)\hat{S}^2_{*(0)} \quad \text{for } * = \text{N}, \text{R}. \tag{S5}$$

S1.3. Projection matrices

Let H and $H_1 = N^{-1}1_N^{\mathrm{T}}1_N$ be the projection matrices onto the column spaces of $(1_N, X)$ and 1_N. Let $\delta = (I-H)Z = (\delta_1, \ldots, \delta_N)^{\mathrm{T}}$ be the residuals from the OLS fit of Z on $(1_N, X)$. They satisfy

$$\begin{aligned}
&e = (I-H)Y, \quad He = 0_N, \quad H_1 e = 0_N, \quad H1_N = 1_N, \quad HX = X,\\
&H = H_1 + X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} = N^{-1}1_N^{\mathrm{T}}1_N + (N-1)^{-1}XX^{\mathrm{T}},\\
&\delta_i = Z_i - p_1 - \lambda^{-1}p_1p_0x_i^{\mathrm{T}}\hat{\tau}_x, \quad \hat{\delta} = N^{-1}\sum_{i=1}^N \delta_i = 0, \quad H\delta = 0_N, \quad H_1\delta = 0_N,\\
&\|\delta\|_2^2 = Z^{\mathrm{T}}(I-H)Z = N(p_1p_0 - \lambda^{-1}p_1^2p_0^2\hat{\tau}_x^{\mathrm{T}}\hat{\tau}_x).
\end{aligned} \tag{S6}$$

S1.4. A lemma on the univariate OLS

Lemma S1. Let $u = (u_1, \ldots, u_N)^{\mathrm{T}}$ and $v = (v_1, \ldots, v_N)^{\mathrm{T}}$ be two $N \times 1$ vectors, and let $\hat{\tau}_0$ be the coefficient of v from the OLS fit of u on v, with the residual vector $\eta = u - v\hat{\tau}_0$. Let $\hat{\mathrm{se}}_0$ and $\tilde{\mathrm{se}}_0$ be the classic and robust standard errors, respectively. We have

$$\hat{\tau}_0 = \frac{v^{\mathrm{T}}u}{\|v\|_2^2}, \quad \hat{\mathrm{se}}_0^2 = \frac{1}{N-1}\frac{\|\eta\|_2^2}{\|v\|_2^2} = \frac{1}{N-1}\left(\frac{\|u\|_2^2}{\|v\|_2^2} - \hat{\tau}_0^2\right), \quad \tilde{\mathrm{se}}_0^2 = \frac{v^{\mathrm{T}}\,\mathrm{diag}(\eta_i^2)v}{(\|v\|_2^2)^2} = \frac{\eta^{\mathrm{T}}\,\mathrm{diag}(v_i^2)\eta}{(\|v\|_2^2)^2}.$$

S2. Probability measures and basic limiting theorems

S2.1. Probability measures

Recall $T^\pi = T(Z_\pi, Y(Z), X)$, where $\pi \sim \mathrm{Unif}(\Pi)$ and $Z_\pi \sim \mathrm{Unif}(\mathcal{Z})$, as a random variable following the randomization distribution of T conditioning on Z. Let $\hat{\mathrm{se}}_*^\pi$ and $\tilde{\mathrm{se}}_*^\pi$ be the classic and robust standard errors of $\hat{\tau}_*^\pi$ under Z_π for $* = \text{N}, \text{R}, \text{F}, \text{L}$. By definition, $\hat{\tau}_*^\pi$, $(\hat{\tau}_*/\hat{\mathrm{se}}_*)^\pi = \hat{\tau}_*^\pi/\hat{\mathrm{se}}_*^\pi$ and $(\hat{\tau}_*/\tilde{\mathrm{se}}_*)^\pi = \hat{\tau}_*^\pi/\tilde{\mathrm{se}}_*^\pi$ are the outputs from the OLS fits of Y on $(1_N, Z_\pi)$ for $* = \text{N}$; e on $(1_N, Z_\pi)$ for $* = \text{R}$; Y on $(1_N, Z_\pi, X)$ for $* = \text{F}$; and Y on $(1_N, Z_\pi, X, W^\pi)$ for $* = \text{L}$, respectively, where $W^\pi = (W_1^\pi, \ldots, W_N^\pi)^{\mathrm{T}}$ with $W_i^\pi = Z_{\pi(i)}x_i$. Let $(\hat{S}^2_{*(z)})^\pi$ be the analogs of $\hat{S}^2_{*(z)}$ based on the corresponding residuals for $* = \text{N}, \text{R}, \text{F}, \text{L}$.

Index by Z and S the probability measures induced by the treatment assignment and random sampling from the population under the finite and super-population frameworks, respectively. Write P_Z-a.s. if a result holds for almost all sequences of Z under the finite-population framework, write P_{S}-a.s. if a result holds for almost all sequences of $\{Y_i(1), Y_i(0), x_i, Z_i\}_{i=1}^N$ under the super-population framework, and write P_π-a.s. if a result holds for almost all sequences of $\pi \sim \mathrm{Unif}(\Pi)$ conditioning on a sequence of observed data $\mathcal{D} = (Y_i, x_i, Z_i)_{i=1}^N$. Write $P_{(Z,\pi)}$-a.s. if a result holds P_π-a.s. for almost all sequences of Z under the finite-population framework.

Let var_∞ and cov_∞ be the asymptotic variance and covariance, with the probability measure clear from the context. Write $A \dot{\sim} B$ for $\sqrt{N}(A - B) = o_{P,*}(1)$ for $* = Z, \pi, \text{S}$.

S2.2. Central limit theorems

To analyze complete randomization and ReM, we need a central limit theorem from Li and Ding (2017, Theorem 5): Under complete randomization and Condition 1,

$$\sqrt{N}\begin{pmatrix} \hat{\tau}_{\text{N}} - \tau \\ \hat{\tau}_x \end{pmatrix} \rightsquigarrow \mathcal{N}\left\{0_{J+1}, \begin{pmatrix} v_{\text{N}} & p_1^{-1}\gamma_1^{\text{T}} + p_0^{-1}\gamma_0^{\text{T}} \\ p_1^{-1}\gamma_1 + p_0^{-1}\gamma_0 & (p_1p_0)^{-1}I_J \end{pmatrix}\right\},$$

recalling $v_{\text{N}} = p_1^{-1}S_1^2 + p_0^{-1}S_0^2 - S_\tau^2$ from Theorem 1. When citing this result, we will simply say "by FPCLT."

To analyze random permutation, we need the following lemma due to Hájek (1961). To simplify the presentation, we give a version that involves slightly stronger moment conditions than Hájek (1961, Theorem 4.1).

Lemma S2. Let $u = (u_1, \ldots, u_N)^{\text{T}}$ and $v = (v_1, \ldots, v_N)^{\text{T}}$ be two $N \times 1$ vectors of real numbers, possibly depending on N. Let $\bar{u} = N^{-1}\sum_{i=1}^N u_i$, $S_u^2 = (N-1)^{-1}\sum_{i=1}^N (u_i - \bar{u})^2$, $\bar{v} = N^{-1}\sum_{i=1}^N v_i$, and $S_v^2 = (N-1)^{-1}\sum_{i=1}^N (v_i - \bar{v})^2$ be the means and variances, respectively. We have

(a) $E(N^{-1}u^{\text{T}}v_\pi) = \bar{u}\bar{v}$, $\quad \text{cov}(N^{-1}u^{\text{T}}v_\pi) = N^{-2}(N-1)S_u^2S_v^2$;

(b) $\sqrt{N}(N^{-1}u^{\text{T}}v_\pi - \bar{u}\bar{v}) \rightsquigarrow \mathcal{N}(0, S_u^2S_v^2)$ if (i) S_u^2 and S_v^2 have finite limits, and (ii) there exists an $\epsilon > 0$ such that $N^{-1}\sum_{i=1}^N (u_i - \bar{u})^{2+\epsilon} \le c_0$ and $N^{-1}\sum_{i=1}^N (v_i - \bar{v})^{2+\epsilon} \le c_0$ for some $c_0 < \infty$ independent of N.

S2.3. A finite-population strong law of large numbers

Based on Bloniarz et al. (2016, Lemma S1), Wu and Ding (2020, Lemma A3) proved a finite-population strong law of large numbers under simple random sampling. We further improve it to allow for rejective sampling in the sense of Fuller (2009), which includes simple random sampling as a special case with $a = \infty$ below. This new finite-population strong law of large numbers in Lemma S3 is useful for analyzing both complete randomization and ReM. Condition 1 ensures the sequences of $\{Y_i(z)\}_{i=1}^N$, $(x_{ij})_{i=1}^N$, and $\{x_{ij}Y_i(z)\}_{i=1}^N$ satisfy the condition required by Lemma S3 for all $z = 0, 1$ and $j = 1, \ldots, J$.

Lemma S3. Let $(W_i, x_i)_{i=1}^N$ be a sequence of finite populations with means $\bar{W} = N^{-1}\sum_{i=1}^N W_i$ and variances $S_W^2 = (N-1)^{-1}\sum_{i=1}^N (W_i - \bar{W})^2$ for $N = 1, \ldots, \infty$. Let $\mathcal{I} \subset \{1, \ldots, N\}$ be a random sample under rejective sampling, in the sense that we start with $\mathcal{I}$ as a simple random sample yet only accept it if $\hat{\tau}_x = |\mathcal{I}|^{-1}\sum_{i\in\mathcal{I}} x_i - (N - |\mathcal{I}|)^{-1}\sum_{i\notin\mathcal{I}} x_i$ satisfies $\hat{\tau}_x^{\text{T}}\{\text{cov}(\hat{\tau}_x)\}^{-1}\hat{\tau}_x < a$. Let $\hat{W}_\mathcal{I} = |\mathcal{I}|^{-1}\sum_{i\in\mathcal{I}} W_i$ and $\hat{S}_\mathcal{I}^2 = (|\mathcal{I}| - 1)^{-1}\sum_{i\in\mathcal{I}}(W_i - \hat{W}_\mathcal{I})^2$ be the sample mean and variance, respectively, and denote by $\mathcal{A}$ the event of $\hat{\tau}_x^{\text{T}}\{\text{cov}(\hat{\tau}_x)\}^{-1}\hat{\tau}_x < a$.

Assume as $N \to \infty$, (i) $\bar{W}$ and S_W^2 have finite limits, (ii) there exists a $c_0 < \infty$ independent of N such that $N^{-1}\sum_{i=1}^N W_i^4 \leq c_0$, (iii) $\lim_{N\to\infty} |\mathcal{I}|/N > 0$, and (iv) $\lim_{N\to\infty} P(\mathcal{A}) = r > 0$. We have $\hat{W}_\mathcal{I} - \bar{W} = o(1)$ and $\hat{S}_\mathcal{I}^2 - S_W^2 = o(1)$ for almost all sequences of $\mathcal{I}$.

Proof of Lemma S3. The probability measure under rejective sampling is equivalent to the probability measure under simple random sampling conditioning on $\mathcal{A}$. We take $\mathcal{I} \subset \{1, \ldots, N\}$ as a simple random sample of size $|\mathcal{I}|$ throughout the proof and reflect the rejective sampling via conditioning on $\mathcal{A}$. We proceed by verifying that there exists an n_0 such that for all $N > n_0$,

$$\max\left\{P(\hat{W}_\mathcal{I} - \bar{W} \geq t \mid \mathcal{A}),\ P(\hat{W}_\mathcal{I} - \bar{W} \leq -t \mid \mathcal{A})\right\} \leq 2r^{-1}\exp\left(-\frac{|\mathcal{I}|^2 t^2}{4NS_W^2}\right) \text{ for all } t \geq 0. \quad \text{(S7)}$$

The result then follows from the Borel–Cantelli lemma via identical reasoning as in the proof of Wu and Ding (2020, Lemma A3).

Let $p_\mathcal{A} = P(\mathcal{A})$ for notational simplicity, and let $\mathcal{A}^c$ be the complement of $\mathcal{A}$. The law of total probability ensures $P(\hat{W}_\mathcal{I} - \bar{W} \geq t) \geq p_\mathcal{A} \cdot P(\hat{W}_\mathcal{I} - \bar{W} \geq t \mid \mathcal{A})$ and $P(\hat{W}_\mathcal{I} - \bar{W} \leq t) \geq p_\mathcal{A} \cdot P(\hat{W}_\mathcal{I} - \bar{W} \leq t \mid \mathcal{A})$ for all $t \geq 0$, such that

$$\begin{aligned}
&\max\left\{P(\hat{W}_\mathcal{I} - \bar{W} \geq t \mid \mathcal{A}),\ P(\hat{W}_\mathcal{I} - \bar{W} \leq -t \mid \mathcal{A})\right\} \\
&\quad \leq p_\mathcal{A}^{-1}\max\left\{P(\hat{W}_\mathcal{I} - \bar{W} \geq t),\ P(\hat{W}_\mathcal{I} - \bar{W} \leq -t)\right\} \leq p_\mathcal{A}^{-1}\exp\left(-\frac{|\mathcal{I}|^2 t^2}{4NS_W^2}\right) \quad \text{for all } t \geq 0
\end{aligned}$$

by Wu and Ding (2020, Lemma A2). The sufficient condition in (S7) then follows from $\lim_{N\to\infty} p_\mathcal{A} = r$ such that there exists an n_0 with $p_\mathcal{A} \geq 2^{-1}r$ for all $N \geq n_0$.

□

S3. Finite-population inference under complete randomization

S3.1. Core lemmas

The following lemma gives some useful facts about Fisher (1935)'s analysis of covariance, i.e., the OLS fit of Y on $(1_N, Z, X)$.

Lemma S4. For $(Y_i, x_i, Z_i)_{i=1}^N$ from arbitrary data generating process with $\bar{x} = 0_J$, the coefficients from the OLS fit of Y on $(1_N, Z, X)$ are

$$\hat{\mu}_\text{F} = \hat{Y} - p_1\hat{\tau}_\text{F}, \quad \hat{\tau}_\text{F} = \frac{Z^\text{T}(I-H)Y}{Z^\text{T}(I-H)Z} = \hat{\tau}_\text{N} - \hat{\gamma}_\text{F}^\text{T}\hat{\tau}_x, \quad \hat{\gamma}_\text{F} = (X^\text{T}X)^{-1}X^\text{T}Y - Np_1p_0\hat{\tau}_\text{F}(X^\text{T}X)^{-1}\hat{\tau}_x,$$

respectively, with residuals $\epsilon_{\text{F},i} = Y_i - \hat{Y} - (Z_i - p_1)\hat{\tau}_\text{F} - x_i^\text{T}\hat{\gamma}_\text{F}$ for $i = 1, \ldots, N$ and standard errors

$$\hat{\text{se}}_\text{F}^2 = \frac{1}{N-2-J}\left\{\frac{Y^\text{T}(I-H)Y}{Z^\text{T}(I-H)Z} - \hat{\tau}_\text{F}^2\right\}, \quad \tilde{\text{se}}_\text{F}^2 = \frac{\eta^\text{T}\,\text{diag}(\delta_i^2)\eta}{\{Z^\text{T}(I-H)Z\}^2},$$

where $\eta = (I-H)Y - (I-H)Z\hat{\tau}_\text{F}$ and δ_i are the residuals from the OLS fit of Z on $(1_N, X)$.

If $\hat{\tau}_x = o(1)$, $N^{-1}\sum_{i=1}^N \|x_i\|_4^4 = O(1)$, and $N^{-1}\sum_{i=1}^N \epsilon_{\mathrm{F},i}^4 = O(1)$ as N goes to infinity, then

$$N\hat{\mathrm{se}}_{\mathrm{F}}^2 - \left(p_0^{-1}\hat{S}_{\mathrm{F}(1)}^2 + p_1^{-1}\hat{S}_{\mathrm{F}(0)}^2\right) = o(1), \qquad N\tilde{\mathrm{se}}_{\mathrm{F}}^2 - \left(p_1^{-1}\hat{S}_{\mathrm{F}(1)}^2 + p_0^{-1}\hat{S}_{\mathrm{F}(0)}^2\right) = o(1).$$

Proof of Lemma S4. First, let $u = e = (I-H)Y$ and $v = \delta = (I-H)Z$ in Lemma S1 to see

$$\hat{\tau}_0 = \frac{\delta^{\mathrm{T}} e}{\|\delta\|_2^2}, \quad \hat{\mathrm{se}}_0^2 = \frac{1}{N-1}\left(\frac{\|e\|_2^2}{\|\delta\|_2^2} - \hat{\tau}_{\mathrm{F}}^2\right), \quad \tilde{\mathrm{se}}_0^2 = \frac{\eta_0^{\mathrm{T}}\,\mathrm{diag}(\delta_i^2)\eta_0}{(\|\delta\|_2^2)^2},$$

where $\eta_0 = e - \delta\hat{\tau}_0$. The result for $\hat{\tau}_{\mathrm{F}}$, $\hat{\mathrm{se}}_{\mathrm{F}}$, and $\tilde{\mathrm{se}}_{\mathrm{F}}$ follows from $\hat{\tau}_{\mathrm{F}} = \hat{\tau}_0$, $(N-2-J)\hat{\mathrm{se}}_{\mathrm{F}}^2 = (N-1)\hat{\mathrm{se}}_0^2$, and $\tilde{\mathrm{se}}_{\mathrm{F}}^2 = \tilde{\mathrm{se}}_0^2$ by FWL.

Second, let $\chi = (1_N, Z, X)$ be the design matrix. That $N^{-1}X^{\mathrm{T}}Z = p_1p_0\hat{\tau}_x$ by (S2) ensures

$$N^{-1}\chi^{\mathrm{T}}\chi = \begin{pmatrix} 1 & p_1 & 0_J^{\mathrm{T}} \\ p_1 & p_1 & p_1p_0\hat{\tau}_x^{\mathrm{T}} \\ 0_J & p_1p_0\hat{\tau}_x & \lambda S_x^2 \end{pmatrix}, \quad \text{with} \quad \begin{pmatrix} 1 & p_1 \\ p_1 & p_1 \end{pmatrix}^{-1} = p_0^{-1}\begin{pmatrix} 1 & -1 \\ -1 & p_1^{-1} \end{pmatrix}, \tag{S8}$$

such that

$$N^{-1}\chi^{\mathrm{T}}\chi\begin{pmatrix} \hat{\mu}_{\mathrm{F}} \\ \hat{\tau}_{\mathrm{F}} \\ \hat{\gamma}_{\mathrm{F}} \end{pmatrix} = N^{-1}\chi^{\mathrm{T}}Y \quad \Longleftrightarrow \quad \begin{pmatrix} 1 & p_1 & 0_J^{\mathrm{T}} \\ p_1 & p_1 & p_1p_0\hat{\tau}_x^{\mathrm{T}} \\ 0_J & p_1p_0\hat{\tau}_x & \lambda S_x^2 \end{pmatrix}\begin{pmatrix} \hat{\mu}_{\mathrm{F}} \\ \hat{\tau}_{\mathrm{F}} \\ \hat{\gamma}_{\mathrm{F}} \end{pmatrix} = \begin{pmatrix} \hat{Y} \\ p_1\hat{Y}(1) \\ \lambda\hat{S}_{xY} \end{pmatrix}.$$

Directly comparing the rows verifies $\hat{\mu}_{\mathrm{F}} = \hat{Y} - p_1\hat{\tau}_{\mathrm{F}}$, $\hat{\tau}_{\mathrm{F}} = \hat{\tau}_{\mathrm{N}} - \hat{\gamma}_{\mathrm{F}}^{\mathrm{T}}\hat{\tau}_x$, and $\hat{\gamma}_{\mathrm{F}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y - \lambda^{-1}p_1p_0\hat{\tau}_{\mathrm{F}}(S_x^2)^{-1}\hat{\tau}_x$. The expression of $\epsilon_{\mathrm{F},i}$ then follows.

Third, $N^{-1}\sum_{i=1}^N \epsilon_{\mathrm{F},i}^4 = O(1)$ ensures $\hat{S}_{\mathrm{F}(z)}^2 = O(1)$ for $z = 0, 1$. The result for $\hat{\mathrm{se}}_{\mathrm{F}}^2$ follows from (S5). Further let $\Delta = \mathrm{diag}(\epsilon_{\mathrm{F},i}^2)$. The robust covariance estimator is $\tilde{V}_{\mathrm{F}} = (\chi^{\mathrm{T}}\chi)^{-1}(\chi^{\mathrm{T}}\Delta\chi)(\chi^{\mathrm{T}}\chi)^{-1}$ with $\tilde{\mathrm{se}}_{\mathrm{F}}^2$ as the $(2,2)$th element. That $N^{-1}\sum_{i=1}^N \|x_i\|_4^4 = O(1)$ and $N^{-1}\sum_{i=1}^N \epsilon_{\mathrm{F},i}^4 = O(1)$ ensures

$$N^{-1}\chi^{\mathrm{T}}\Delta\chi = N^{-1}\begin{pmatrix} 1_N^{\mathrm{T}}\Delta 1_N & 1_N^{\mathrm{T}}\Delta Z & 1_N^{\mathrm{T}}\Delta X \\ Z^{\mathrm{T}}\Delta 1_N & Z^{\mathrm{T}}\Delta Z & Z^{\mathrm{T}}\Delta X \\ X^{\mathrm{T}}\Delta 1_N & X^{\mathrm{T}}\Delta Z & X^{\mathrm{T}}\Delta X \end{pmatrix} = O(1)$$

with (i) $0 < 1_N^{\mathrm{T}}\Delta Z = Z^{\mathrm{T}}\Delta Z \le 1_N^{\mathrm{T}}\Delta 1_N \le (N\sum_{i=1}^N \epsilon_{\mathrm{F},i}^4)^{1/2} = O(N)$, (ii) $X^{\mathrm{T}}\Delta X = \sum_{i=1}^N \epsilon_{\mathrm{F},i}^2 x_i x_i^{\mathrm{T}} = O(N)$, and (iii) $1_N^{\mathrm{T}}\Delta X \le (1_N^{\mathrm{T}}\Delta 1_N)^{1/2}(X^{\mathrm{T}}\Delta X)^{1/2}$ and $Z^{\mathrm{T}}\Delta X \le (Z^{\mathrm{T}}\Delta Z)^{1/2}(X^{\mathrm{T}}\Delta X)^{1/2}$. This, together with $\hat{\tau}_x = o(1)$ in (S8), ensures

$$N\tilde{V}_{\mathrm{F}} = p_0^{-1}\begin{pmatrix} 1 & -1 & 0_J^{\mathrm{T}} \\ -1 & p_1^{-1} & 0_J^{\mathrm{T}} \\ 0_J & 0_J & p_0S_x^2 \end{pmatrix} N^{-1}\begin{pmatrix} 1_N^{\mathrm{T}}\Delta 1_N & 1_N^{\mathrm{T}}\Delta Z & 1_N^{\mathrm{T}}\Delta X \\ Z^{\mathrm{T}}\Delta 1_N & Z^{\mathrm{T}}\Delta Z & Z^{\mathrm{T}}\Delta X \\ X^{\mathrm{T}}\Delta 1_N & X^{\mathrm{T}}\Delta Z & X^{\mathrm{T}}\Delta X \end{pmatrix} p_0^{-1}\begin{pmatrix} 1 & -1 & 0_J^{\mathrm{T}} \\ -1 & p_1^{-1} & 0_J^{\mathrm{T}} \\ 0_J & 0_J & p_0S_x^2 \end{pmatrix} + o(1)$$

with the $(2,2)$th element as

$$\begin{aligned} N\tilde{\text{se}}_{\text{F}}^2 &= p_0^{-2}N^{-1}(-1,p_1^{-1})\begin{pmatrix} 1_N^{\text{T}}\Delta 1_N & 1_N^{\text{T}}\Delta Z \\ Z^{\text{T}}\Delta 1_N & Z^{\text{T}}\Delta Z \end{pmatrix}\begin{pmatrix} -1 \\ p_1^{-1} \end{pmatrix} + o(1) \\ &= p_0^{-2}N^{-1}\left(1_N^{\text{T}}\Delta 1_N - p_1^{-1}Z^{\text{T}}\Delta 1_N - p_1^{-1}1_N^{\text{T}}\Delta Z + p_1^{-2}Z^{\text{T}}\Delta Z\right) + o(1) \\ &= (p_0p_1)^{-2}N^{-1}(Z-p_1 1_N)^{\text{T}}\Delta(Z-p_1 1_N) + o(1) \\ &= p_1^{-1}\hat{S}_{\text{F}(1)}^2 + p_0^{-1}\hat{S}_{\text{F}(0)}^2 + o(1). \end{aligned}$$

□

Lemma S5. Assume Condition 1 and complete randomization. As $N \to \infty$, the following results hold P_Z-a.s.:

(a) $\hat{Y}(z) - \bar{Y}(z) = o(1)$, $\hat{S}_z^2 - S_z^2 = o(1)$, $\hat{x}(z) = o(1)$, $\hat{S}_{x(z)}^2 - S_x^2 = o(1)$, $\hat{S}_{xY(z)} - S_{xY(z)} = o(1)$ for $z = 0, 1$.

(b) The sequence of finite populations $(Y_i, Y_i, x_i)_{i=1}^N$ satisfies Condition 1 with $N^{-1}\sum_{i=1}^N Y_i^4 = O(1)$ and $\hat{Y} - \{p_1\bar{Y}(1) + p_0\bar{Y}(0)\} = o(1)$, $\hat{S}^2 - S^2 - p_1p_0\tau^2 = o(1)$, $\hat{\gamma}_{\text{R}} - \gamma = o(1)$, $\hat{S}_e^2 - S_a^2 - p_1p_0\tau^2 = o(1)$ giving the analogs of $\bar{Y}(z)$, S_z^2, γ_z, and both $S_{a(z)}^2$ and $S_{b(z)}^2$ for $z = 0, 1$, respectively.

(c) $\hat{e} = 0$, $\hat{S}_e^2 = S_a^2 + p_1p_0\tau^2 + o(1)$, $N^{-1}\sum_{i=1}^N e_i^4 = O(1)$.

(d) $\hat{\delta} = N^{-1}\sum_{i=1}^N \delta_i = 0$, $\hat{S}_\delta^2 = (N-1)^{-1}\sum_{i=1}^N(\delta_i - \hat{\delta})^2 = p_1p_0 + o(1)$, $N^{-1}\sum_{i=1}^N \delta_i^4 = O(1)$.

(e) $\hat{\mu}_{\text{F}} - \hat{Y} = o(1)$, $\hat{\tau}_{\text{F}} - \tau = o(1)$, $\hat{\gamma}_{\text{F}} - \gamma = o(1)$, with $\hat{\epsilon}_{\text{F}} = N^{-1}\sum_{i=1}^N \epsilon_{\text{F},i} = 0$, $\hat{S}_{\text{F}}^2 = (N-1)^{-1}\sum_{i=1}^N \epsilon_{\text{F},i}^2 = S_a^2 + o(1)$, $N^{-1}\sum_{i=1}^N \epsilon_{\text{F},i}^4 = O(1)$.

Lemma S5 ensures it suffices to focus on the sequences of Z that satisfy (a)–(e) when verifying results on almost sure convergence under P_Z.

Proof of Lemma S5. Recall from Condition 1 that $w_i(z) = (S_x^2)^{-1}x_iY_i(z) = x_iY_i(z)$ with mean and covariance matrix $\bar{w}(z) = N^{-1}\sum_{i=1}^N w_i(z) = \lambda S_{xY(z)}$ and $S_{w(z)}^2 = (N-1)^{-1}\sum_{i=1}^N\{w_i(z) - \bar{w}(z)\}^2$. Its observed analog $w_i = x_iY_i$ has sample means and covariance matrices $\hat{w}(z) = N_z^{-1}\sum_{i:Z_i=z} w_i$ and $\hat{S}_{w(z)}^2 = (N_z-1)^{-1}\sum_{i:Z_i=z}\{w_i - \hat{w}(z)\}\{w_i - \hat{w}(z)\}^{\text{T}}$ for $z = 0, 1$. Lemma S3 ensures

$$\hat{w}(z) - S_{xY(z)} = o(1)\ P_Z\text{-a.s.}, \quad \hat{S}_{w(z)}^2 - S_{w(z)}^2 = o(1)\ P_Z\text{-a.s.} \tag{S9}$$

under Condition 1. The result for $\hat{Y}(z)$, $\hat{S}_z^2$, $\hat{x}(z)$, and $\hat{S}_{x(z)}^2$ in statement (a) follows from Lemma S3 directly. The result for $\hat{S}_{xY(z)}$ then follows from

$$\hat{S}_{xY(z)} = (N_z-1)^{-1}\sum_{i:Z_i=z}\{x_i - \hat{x}(z)\}\{Y_i - \hat{Y}(z)\} = \frac{N_z}{N_z-1}\hat{w}(z) - \frac{N_z}{N_z-1}\hat{x}(z)\hat{Y}(z).$$

This verifies statement (a) and (S9) hold P_Z-a.s., such that it suffices to verify statements (b)–(d) hold for Z's that satisfy statement (a) and (S9). Fix one such sequence for the rest of proof.

For statement (b), the correspondence between the analogs follows from definitions with the analogs of $a_i(z)$ and $b_i(z)$ given by $Y_i - \hat{Y} - x_i^{\mathrm{T}}\hat{\gamma}_{\mathrm{R}} = e_i$ ensured by (S4). The limits of $\hat{Y}$ and $\hat{S}^2$ follow from statement (a) and (S3). With (w_i, w_i) as the analogs of $\{w_i(1), w_i(0)\}$ in the finite population $(Y_i, Y_i, x_i)_{i=1}^N$, the limits of $\hat{w} = N^{-1}\sum_{i=1}^N w_i$ and $\hat{S}_w^2 = (N-1)^{-1}\sum_{i=1}^N (w_i - \hat{w})(w_i - \hat{w})^{\mathrm{T}}$ follow from (S9) and applying (S3) entry-wise. This in turn ensures $\hat{S}_{xY}$, as the analog of $S_{xY(z)}$, satisfies $\hat{S}_{xY} = \hat{\gamma}_{\mathrm{R}} = \lambda^{-1}\hat{w} = \gamma + o(1)$ with $\hat{S}_e^2 = \hat{S}^2 - \|\hat{\gamma}_{\mathrm{R}}\|_2^2$ having finite positive limit $S^2 + p_1p_0\tau^2 - \|\gamma\|_2^2 = S_a^2 + p_1p_0\tau^2$ by (S1) and (S4). This verifies Condition 1(ii). Further, $N^{-1}\sum_{i=1}^N Y_i^4 \le N^{-1}\sum_{i=1}^N \{Y_i^4(1) + Y_i^4(0)\} \le 2c_0$; likewise for $N^{-1}\sum_{i=1}^N \|w_i\|_4^4 \le 2c_0$. This verifies Condition 1(iii) and hence statement (b).

For statement (c), that $\hat{e} = 0$ follows from (S4) and the variance follows from statement (b). Further, $\hat{\gamma}_{\mathrm{R}} = \gamma + o(1)$ from statement (b) ensures $\|\hat{\gamma}_{\mathrm{R}}\|_\infty = O(1)$ such that $(x_i^{\mathrm{T}}\hat{\gamma}_{\mathrm{R}})^4 \le c_1\|x_i\|_4^4$ for some c_1 independent of N. This, together with $e_i^4 = (e_i^2)^2 \le \{2(Y_i - \hat{Y})^2 + 2(x_i^{\mathrm{T}}\hat{\gamma}_{\mathrm{R}})^2\}^2 \le 8\{(Y_i - \hat{Y})^4 + (x_i^{\mathrm{T}}\hat{\gamma}_{\mathrm{R}})^4\}$, ensures $N^{-1}\sum_{i=1}^N e_i^4 = O(1)$ and hence statement (c).

For statement (d), the limit of $\hat{S}_\delta^2$ follows from (S6) and $\hat{\tau}_x = 0_J + o(1)$ by statement (a). With $\delta_i = Z_i - p_1 - \lambda^{-1}p_1p_0x_i^{\mathrm{T}}\hat{\tau}_x$ from (S6), we have $\delta_i^4 \le 8\{(Z_i - p_1)^4 + (\lambda^{-1}p_1p_0x_i^{\mathrm{T}}\hat{\tau}_x)^4\}$ and hence $N^{-1}\sum_{i=1}^N \delta_i^4 = O(1)$ by the same reasoning as that for $N^{-1}\sum_{i=1}^N e_i^4 = O(1)$ in statement (c).

Statement (e) follows from $N^{-1}Z^{\mathrm{T}}(I-H)Y = N^{-1}\sum_{i=1}^N \delta_i Y_i = p_1\hat{Y}(1) - p_1\hat{Y} - p_1p_0\hat{\gamma}_{\mathrm{R}}^{\mathrm{T}}\hat{\tau}_x = p_1p_0\hat{\tau}_{\mathrm{N}} - p_1p_0\hat{\gamma}_{\mathrm{R}}^{\mathrm{T}}\hat{\tau}_x$ and $N^{-1}Z(I-H)Z = p_1p_0 + o(1)$ by (S6). This ensures $\hat{\tau}_{\mathrm{F}} = \tau + o(1)$ by Lemma S4 and thus the result on $\hat{\mu}_{\mathrm{F}}$ and $\hat{\gamma}_{\mathrm{F}}$. Further, replace Y_i with $\epsilon_{\mathrm{F},i}$ in (S3) to see

$$\hat{S}_{\mathrm{F}}^2 = \frac{N_1 - 1}{N-1}\hat{S}_{\mathrm{F}(1)}^2 + \frac{N_0 - 1}{N-1}\hat{S}_{\mathrm{F}(0)}^2 + \frac{N}{N-1}p_0p_1\left\{\hat{\epsilon}_{\mathrm{F}}(1) - \hat{\epsilon}_{\mathrm{F}}(0)\right\}^2,$$

where $\hat{S}_{\mathrm{F}(z)}^2 = S_{a(z)}^2 + o(1)$ by (S1) and (S5), and $\hat{\epsilon}_{\mathrm{F}}(1) - \hat{\epsilon}_{\mathrm{F}}(0) = \hat{\tau}_{\mathrm{N}} - \hat{\tau}_{\mathrm{F}} - \hat{\tau}_x^{\mathrm{T}}\hat{\gamma}_{\mathrm{F}} = o(1)$ by $\epsilon_{\mathrm{F},i} = Y_i - \hat{Y} - (Z_i - p_1)\hat{\tau}_{\mathrm{F}} - x_i^{\mathrm{T}}\hat{\gamma}_{\mathrm{F}}$ from Lemma S4. We have $\hat{\epsilon}_{\mathrm{F}} = 0$ and $N^{-1}\sum_{i=1}^N \epsilon_{\mathrm{F},i}^4 \le 27\{N^{-1}\sum_{i=1}^N (Y_i - \hat{Y})^4 + N^{-1}\sum_{i=1}^N Z_i - p_1)^4\hat{\tau}_{\mathrm{F}}^4 + N^{-1}\sum_{i=1}^N (x_i^{\mathrm{T}}\hat{\gamma}_{\mathrm{F}})^4\} = O(1)$ by the Cauchy–Schwarz inequality. □

Technically, the first strategy by Rosenbaum (2002) takes e as the fixed input for conducting FRT and thus has no counterpart for $\hat{\gamma}_{\mathrm{R}}$ under Z_π. Nevertheless, the procedure is identical to one that takes $(Y_i, x_i)_{i=1}^N$ as the fixed input, regresses Y on $(1_N, X)$ to generate $e^\pi = e$ and $\hat{\gamma}_{\mathrm{R}}^\pi = \hat{\gamma}_{\mathrm{R}}$ independent of Z_π, and then regresses $e^\pi = e$ on $(1_N, Z_\pi)$ to generate $\hat{\tau}_{\mathrm{R}}^\pi$, $\hat{\mathrm{se}}_{\mathrm{R}}^\pi$, and $\tilde{\mathrm{se}}_{\mathrm{R}}^\pi$. This unifies the four procedures, $* = \mathrm{N}, \mathrm{R}, \mathrm{F}, \mathrm{L}$, as all taking $(Y_i, x_i)_{i=1}^N$ as the fixed input for conducting FRT. We take this perspective to simplify the presentation.

Lemma S6. Assume Condition 1 and complete randomization.

(a) $\hat{\gamma}_* - \gamma = o(1)$ for $* = \mathrm{R}, \mathrm{F}$, $\quad \hat{\gamma}_{\mathrm{L}} - (p_0\gamma_1 + p_1\gamma_0) = o(1)$,
$N\hat{\mathrm{se}}_{\mathrm{N}}^2 - (p_1p_0)^{-1}S^2 = o(1)$, $\quad N\tilde{\mathrm{se}}_{\mathrm{N}}^2 - \left(p_1^{-1}S_1^2 + p_0^{-1}S_0^2\right) = o(1)$,
$N\hat{\mathrm{se}}_*^2 - (p_1p_0)^{-1}S_a^2 = o(1)$, $\quad N\tilde{\mathrm{se}}_*^2 - (p_1^{-1}S_{a(1)}^2 + p_0^{-1}S_{a(0)}^2) = o(1)$ for $* = \mathrm{R}, \mathrm{F}$,
$N\hat{\mathrm{se}}_{\mathrm{L}}^2 - (p_1p_0)^{-1}S_b^2 = o(1)$, $\quad N\tilde{\mathrm{se}}_{\mathrm{L}}^2 - (p_1^{-1}S_{b(1)}^2 + p_0^{-1}S_{b(0)}^2) = o(1)$

hold P_Z-a.s., with $N\hat{\mathrm{se}}_*^2$ and $N\tilde{\mathrm{se}}_*^2$ all having positive finite limits.

(b) $\hat{\gamma}_*^\pi - \gamma = o(1)$ for $* = \text{R}, \text{F}, \text{L}$, with $\hat{\gamma}_\text{R}^\pi = \hat{\gamma}_\text{R}$,
$N(\hat{\text{se}}_\text{N}^2)^\pi - (p_1p_0)^{-1}S^2 - \tau^2 = o(1)$, $N(\tilde{\text{se}}_\text{N}^2)^\pi - (p_1p_0)^{-1}S^2 - \tau^2 = o(1)$,
$N(\hat{\text{se}}_*^2)^\pi - (p_1p_0)^{-1}S_a^2 - \tau^2 = o(1)$, $N(\tilde{\text{se}}_*^2)^\pi - (p_1p_0)^{-1}S_a^2 - \tau^2 = o(1)$ for $* = \text{R}, \text{F}, \text{L}$
hold $P_{(Z,\pi)}$-a.s., with $N(\hat{\text{se}}_*^2)^\pi$ and $N(\tilde{\text{se}}_*^2)^\pi$ all having positive finite limits.

Lin (2013, Theorem 2) showed $N\hat{\text{se}}_\text{F}^2 - (p_0^{-1}S_{a(1)}^2 + p_1^{-1}S_{a(0)}^2) = o_{P,Z}(1)$ and $N\tilde{\text{se}}_\text{F}^2 - (p_1^{-1}S_{a(1)}^2 + p_0^{-1}S_{a(0)}^2) = o_{P,Z}(1)$ as the probability limits. Lemma S6 strengthens the results to almost sure convergence.

Proof of Lemma S6. Lemma S5 ensures that it suffices to verify the result for sequences of Z that satisfy Lemma S5 statements (a)–(e). Fix one such sequence for the rest of the proof.

For statement (a) regarding the sampling distributions, the limits of $\hat{\gamma}_*$ follow from $\hat{\gamma}_\text{R} = (S_x^2)^{-1}\hat{S}_{xY}$, $\hat{\gamma}_\text{F} = \hat{\gamma}_\text{R} - \lambda^{-1}p_1p_0\hat{\tau}_\text{F}(S_x^2)^{-1}\hat{\tau}_x$, and $\hat{\gamma}_\text{L} = p_0\hat{\gamma}_{\text{L},1} + p_1\hat{\gamma}_{\text{L},0}$ with $\hat{\gamma}_{\text{L},z} = (\hat{S}_{x(z)}^2)^{-1}\hat{S}_{xY(z)}$ by Proposition 1, where $\hat{\gamma}_\text{R} = o(1)$, $\hat{\tau}_\text{F} - \tau = o(1)$, and $\hat{\gamma}_{\text{L},z} - \gamma_z = o(1)$ by Lemma S5. This ensures $\hat{S}_{\text{N}(z)}^2 - S_z^2 = o(1)$, $\hat{S}_{*(z)}^2 - S_{a(z)}^2 = o(1)$ for $* = \text{R}, \text{F}$, and $\hat{S}_{\text{L}(z)}^2 - S_{b(z)}^2 = o(1)$ by (S1) and (S5), and allows us to unify the standard errors as

$$N\hat{\text{se}}_*^2 - \left(p_0^{-1}\hat{S}_{*(1)}^2 + p_1^{-1}\hat{S}_{*(0)}^2\right) = o(1), \quad N\tilde{\text{se}}_*^2 - \left(p_1^{-1}\hat{S}_{*(1)}^2 + p_0^{-1}\hat{S}_{*(0)}^2\right) = o(1) \quad (* = \text{N}, \text{R}, \text{F}, \text{L}).$$

In particular, the result for $\hat{\text{se}}_*^2$ follows from (S5) provided $\hat{S}_{*(z)}^2$ all have finite limits. The result for $\tilde{\text{se}}_\text{N}^2$ and $\tilde{\text{se}}_\text{R}^2$ follows from (S5). The result for $\tilde{\text{se}}_\text{F}^2$ follows from Lemma S4 with the regularity condition ensured by Lemma S5(e). The result for $\tilde{\text{se}}_\text{L}^2$ follows from Li and Ding (2020, Theorem 8) given Condition 1 implies Li and Ding (2020, Condition 1) by Wu and Ding (2020, Proposition 1). Alternatively, almost identical algebra as in the proof of Lemma S4 for the limiting value of $N\tilde{\text{se}}_\text{F}^2$ attains the same end. Condition 1(ii) and (S1) ensure the limits are all positive.

For statement (b) regarding the randomization distributions, FRT takes $(Y_i, x_i)_{i=1}^N$ as the fixed input for permuting the treatment vector, and thereby induces a sequence of finite populations $(Y_i, Y_i, x_i)_{i=1}^N$ with Y_i as the "pseudo potential outcomes" under both treatment and control. The way we chose the fixed Z further ensures $(Y_i, Y_i, x_i)_{i=1}^N$ satisfies Condition 1. The result in statement (a) regarding the sampling distributions thus also holds for $\hat{\gamma}_*^\pi$ and $(\hat{S}_{*(z)}^2)^\pi$ P_π-a.s. if we replace γ with $\hat{\gamma}_\text{R}$, S_z^2 with $\hat{S}^2$, and both $S_{a(z)}^2$ and $S_{b(z)}^2$ with $\hat{S}_e^2$ by the correspondence result from Lemma S5(b). The result follows from $\hat{\gamma}_\text{R} - \gamma = o(1)$, $\hat{S}^2 - S^2 - p_1p_0\tau^2 = o(1)$, and $\hat{S}_e^2 - S_a^2 - p_1p_0\tau^2 = o(1)$. □

Lemma S7. Assume Condition 1 and complete randomization.

(a) $\hat{\tau}_\text{L} \dot{\sim} \hat{\tau}_\text{N} - (p_0\gamma_1 + p_1\gamma_0)^\text{T}\hat{\tau}_x$, $\hat{\tau}_\text{F} \dot{\sim} \hat{\tau}_\text{R} \dot{\sim} \hat{\tau}_\text{N} - \gamma^\text{T}\hat{\tau}_x$.

(b) $\hat{\tau}_\text{R}^\pi \dot{\sim} \hat{\tau}_\text{F}^\pi \dot{\sim} \hat{\tau}_\text{L}^\pi \dot{\sim} \hat{\tau}_\text{N}^\pi - \gamma^\text{T}\hat{\tau}_x^\pi$ holds P_Z-a.s..

Proof of Lemma S7. First, let $\gamma_\text{R} = \gamma_\text{F} = \gamma$ and $\gamma_\text{L} = p_0\gamma_1 + p_1\gamma_0$ to write $\hat{\gamma}_* - \gamma_* = o(1)$ P_Z-a.s. by Lemma S6. This, together with $\hat{\tau}_* = \hat{\tau}_\text{N} - \hat{\gamma}_*^\text{T}\hat{\tau}_x$ by Proposition 1 and $\sqrt{N}\hat{\tau}_x \rightsquigarrow \mathcal{N}\{0_J, (p_0p_1)^{-1}I_J\}$

by FPCLT, ensures $\sqrt{N}\{\hat{\tau}_* - (\hat{\tau}_{\mathrm{N}} - \gamma_*^{\mathrm{T}}\hat{\tau}_x)\} = (\hat{\gamma}_* - \gamma_*)^{\mathrm{T}}\sqrt{N}\hat{\tau}_x = o_{P,Z}(1)$ by Slutsky's theorem and hence $\hat{\tau}_* \dot{\sim} \hat{\tau}_{\mathrm{N}} - \gamma_*^{\mathrm{T}}\hat{\tau}_x$.

Second, Proposition 1 is algebraic and holds under the probability measure induced by π as well. This ensures $\hat{\tau}_*^{\pi} = \hat{\tau}_{\mathrm{N}}^{\pi} - \gamma^{\mathrm{T}}\hat{\tau}_x^{\pi} - (\hat{\gamma}_*^{\pi} - \gamma)^{\mathrm{T}}\hat{\tau}_x^{\pi}$ for $* = \mathrm{R}, \mathrm{F}, \mathrm{L}$, with $\hat{\gamma}_{\mathrm{R}}^{\pi} = \hat{\gamma}_{\mathrm{R}} = \gamma + o(1)$ P_Z-a.s. and $\hat{\gamma}_*^{\pi} - \gamma = o(1)$ P_π-a.s. for $* = \mathrm{F}, \mathrm{L}$ P_Z-a.s. by Lemma S6. Almost identical reasoning as that for $\hat{\tau}_*$ ensures $(\hat{\gamma}_*^{\pi} - \gamma)^{\mathrm{T}}\sqrt{N}\hat{\tau}_x^{\pi} = o_{P,\pi}(1)$ holds P_Z-a.s. for $* = \mathrm{R}, \mathrm{F}, \mathrm{L}$ and hence the result. □

S3.2. Proofs for main results under finite-population framework

Theorem 1 is a special case of Wu and Ding (2020). We give below a unified proof.

Proof of Theorems 1–4. First, Condition 1 ensures $S_{a(z)}^2$, S_a^2, $S_{b(z)}^2$, and S_b^2 have positive finite limits, and S_ξ^2 and S_τ^2 have finite limits by (S1).

The result for $\hat{\tau}_{\mathrm{N}}$ and $\hat{\tau}_{\mathrm{N}}^{\pi}$ follows from Ding and Dasgupta (2018), with $v_{\mathrm{N}} = p_1^{-1}S_1^2 + p_0^{-1}S_0^2 - S_\tau^2$ and $v_{\mathrm{N}0} = \lim_{N\to\infty}(p_1p_0)^{-1}\hat{S}^2 = p_0^{-1}S_1^2 + p_1^{-1}S_0^2 + \tau^2$. This ensures $\sqrt{N}(\hat{\tau}_{\mathrm{N}} - \tau)/v_{\mathrm{N}}^{1/2} \rightsquigarrow \mathcal{N}(0,1)$, and $\sqrt{N}\hat{\tau}_{\mathrm{N}}^{\pi}/v_{\mathrm{N}0}^{1/2} \rightsquigarrow \mathcal{N}(0,1)$ P_Z-a.s.. The result for the studentized variants follows from

$$\begin{aligned}
&\hat{\tau}_{\mathrm{N}}/\hat{\mathrm{se}}_{\mathrm{N}} = \sqrt{N}\hat{\tau}_{\mathrm{N}}/v_{\mathrm{N}}^{1/2}\cdot(v_{\mathrm{N}}/N\hat{\mathrm{se}}_{\mathrm{N}}^2)^{1/2}, && \hat{\tau}_{\mathrm{N}}/\tilde{\mathrm{se}}_{\mathrm{N}} = \sqrt{N}\hat{\tau}_{\mathrm{N}}/v_{\mathrm{N}}^{1/2}\cdot(v_{\mathrm{N}}/N\tilde{\mathrm{se}}_{\mathrm{N}}^2)^{1/2},\\
&\hat{\tau}_{\mathrm{N}}^{\pi}/\hat{\mathrm{se}}_{\mathrm{N}}^{\pi} = \sqrt{N}\hat{\tau}_{\mathrm{N}}^{\pi}/v_{\mathrm{N}0}^{1/2}\cdot\{v_{\mathrm{N}0}/N(\hat{\mathrm{se}}_{\mathrm{N}}^2)^{\pi}\}^{1/2}, && \hat{\tau}_{\mathrm{N}}^{\pi}/\tilde{\mathrm{se}}_{\mathrm{N}}^{\pi} = \sqrt{N}\hat{\tau}_{\mathrm{N}}^{\pi}/v_{\mathrm{N}0}^{1/2}\cdot\{v_{\mathrm{N}0}/N(\tilde{\mathrm{se}}_{\mathrm{N}}^2)^{\pi}\}^{1/2}
\end{aligned} \tag{S10}$$

by Slutsky's theorem with $(v_{\mathrm{N}}/N\hat{\mathrm{se}}_{\mathrm{N}}^2) = c'_{\mathrm{N}} + o(1)$ P_Z-a.s., $(v_{\mathrm{N}}/N\tilde{\mathrm{se}}_{\mathrm{N}}^2) = c_{\mathrm{N}} + o(1)$ P_Z-a.s., $v_{\mathrm{N}0}/N(\hat{\mathrm{se}}_{\mathrm{N}}^2)^{\pi} = 1 + o(1)$ $P_{(Z,\pi)}$-a.s., and $v_{\mathrm{N}0}/N(\tilde{\mathrm{se}}_{\mathrm{N}}^2)^{\pi} = 1 + o(1)$ $P_{(Z,\pi)}$-a.s. by Lemma S6.

For the nine covariate-adjusted variants with $* = \mathrm{R}, \mathrm{F}, \mathrm{L}$, the result for $\hat{\tau}_*$ and $\hat{\tau}_*^{\pi}$ follows from the result for $\hat{\tau}_{\mathrm{N}}$ and $\hat{\tau}_{\mathrm{N}}^{\pi}$, FPCLT, and Lemma S7, with

$$\begin{aligned}
v_{\mathrm{F}} = v_{\mathrm{R}} &= \mathrm{var}_\infty(\hat{\tau}_{\mathrm{R}}) = v_{\mathrm{N}} + (p_1p_0)^{-1}\|\gamma\|_2^2 - 2(p_1^{-1}\gamma_1 + p_0^{-1}\gamma_0)^{\mathrm{T}}\gamma = p_1^{-1}S_{a(1)}^2 + p_0^{-1}S_{a(0)}^2 - S_\tau^2,\\
v_{\mathrm{L}} &= \mathrm{var}_\infty(\hat{\tau}_{\mathrm{L}}) = v_{\mathrm{N}} - (p_1p_0)^{-1}\|p_0\gamma_1 + p_1\gamma_0\|_2^2 = p_1^{-1}S_{b(1)}^2 + p_0^{-1}S_{b(0)}^2 - S_\xi^2,\\
v_{*0} &= \mathrm{var}_\infty(\hat{\tau}_*^{\pi}) = v_{\mathrm{N}0} - (p_1p_0)^{-1}\|\gamma\|_2^2 = (p_1p_0)^{-1}S_a^2 + \tau^2 \quad \text{for } * = \mathrm{R}, \mathrm{F}, \mathrm{L}
\end{aligned}$$

by (S1). This ensures $\sqrt{N}\hat{\tau}_*/v_*^{1/2} \rightsquigarrow \mathcal{N}(0,1)$, and $\sqrt{N}\hat{\tau}_*^{\pi}/v_{*0}^{1/2} \rightsquigarrow \mathcal{N}(0,1)$ P_Z-a.s. for $* = \mathrm{R}, \mathrm{F}, \mathrm{L}$. The result for the studentized variants then follows from Slutsky's theorem and Lemma S6 via the same reasoning as in (S10). □

Proof of Proposition 1. That $\hat{\tau}_{\mathrm{R}} = \hat{\tau}_{\mathrm{N}} - \hat{\tau}_x^{\mathrm{T}}\hat{\gamma}_{\mathrm{R}}$ follows from (S4). That $\hat{\tau}_{\mathrm{F}} = \hat{\tau}_{\mathrm{N}} - \hat{\tau}_x^{\mathrm{T}}\hat{\gamma}_{\mathrm{F}}$ and the expression for $\hat{\gamma}_{\mathrm{F}}$ follows from Lemma S4. That $\hat{\tau}_{\mathrm{L}} = \hat{\tau}_{\mathrm{N}} - \{\hat{x}(1)\}^{\mathrm{T}}\hat{\gamma}_{\mathrm{L},1} + \{\hat{x}(0)\}^{\mathrm{T}}\hat{\gamma}_{\mathrm{L},0}$ for centered covariates follows from Lin (2013), which further simplifies to $\hat{\tau}_{\mathrm{L}} = \hat{\tau}_{\mathrm{N}} - \hat{\tau}_x^{\mathrm{T}}\hat{\gamma}_{\mathrm{L}}$ because of (S2). □

S4. FRT under ReM

Lemma S8. Assume Condition 1. Lemmas S5–S6 also hold under ReM.

Proof. Lemmas S5–S6 follow from the strong law of large numbers in Lemma S3 via the same reasoning as that under complete randomization, with the convergence in probability of the sample variances to their respective finite-population variances ensured by Li et al. (2018, Lemma A16). □

Proof of Theorem 5. The probability measure under ReM is equivalent to the probability measure under complete randomization conditioning on $\mathcal{A}$. We take the distribution under complete randomization as the default distribution throughout this proof, and use "$\mid \mathcal{A}$" to denote the condition of either $\hat{\tau}_x^{\mathrm{T}}\{\mathrm{cov}(\hat{\tau}_x)\}^{-1}\hat{\tau}_x < a$ under the sampling distribution or $(\hat{\tau}_x^{\pi})^{\mathrm{T}}\{\mathrm{cov}(\hat{\tau}_x^{\pi})\}^{-1}\hat{\tau}_x^{\pi} < a$ under the randomization distribution. As such, $T \mid \mathcal{A}$ gives the sampling distribution of T under ReM, and $T^{\pi} \mid \mathcal{A} \sim T^{\pi\mid\mathcal{A}}$ gives the randomization distribution of T under FRT with ReM.

The result for $\hat{\tau}_{\mathrm{L}}$, $\hat{\tau}_{\mathrm{L}}/\hat{\mathrm{se}}_{\mathrm{L}}$, and $\hat{\tau}_{\mathrm{L}}/\tilde{\mathrm{se}}_{\mathrm{L}}$ follows from the asymptotic independence between $\hat{\tau}_{\mathrm{L}}$ and $\hat{\tau}_x$ and between $\hat{\tau}_{\mathrm{L}}^{\pi}$ and $\hat{\tau}_x^{\pi}$ under complete randomization (Li and Ding 2020), such that ReM in either case does not affect the limiting distributions. We verify below the result for $* = \mathrm{N}, \mathrm{R}, \mathrm{F}$ together, starting from the unstudentized $\hat{\tau}_*$ and $\hat{\tau}_*^{\pi}$ for $* = \mathrm{N}, \mathrm{R}, \mathrm{F}$ and then moving onto the studentized variants.

For the distributions of the unstudentized $\hat{\tau}_*$ and $\hat{\tau}_*^{\pi}$, where $* = \mathrm{N}, \mathrm{R}, \mathrm{F}$, Lemma S7 ensures $\hat{\tau}_{\mathrm{F}}^{\pi} \dot{\sim} \hat{\tau}_{\mathrm{R}}^{\pi} \dot{\sim} \hat{\tau}_{\mathrm{L}}^{\pi}$ such that it suffices to verify the result for $\hat{\tau}_{\mathrm{N}}$, $\hat{\tau}_{\mathrm{R}}$, $\hat{\tau}_{\mathrm{F}}$, and $\hat{\tau}_{\mathrm{N}}^{\pi}$. Let $D \sim \mathcal{N}(0_J, I_J)$ and $\epsilon \sim \mathcal{N}(0,1)$ be two independent standard normals to represent

$$\sqrt{N}\hat{\tau}_x \rightsquigarrow v_x D, \quad \sqrt{N}\hat{\tau}_x^{\pi} \rightsquigarrow v_x D, \quad \sqrt{N}\hat{\tau}_{\mathrm{L}} \rightsquigarrow v_{\mathrm{L}}^{1/2}\epsilon, \quad \sqrt{N}\hat{\tau}_{\mathrm{L}}^{\pi} \rightsquigarrow v_{\mathrm{L}0}^{1/2}\epsilon,$$

where $v_x = (p_1 p_0)^{-1/2}$ by FPCLT. With

$$\hat{\tau}_{\mathrm{N}} \dot{\sim} \hat{\tau}_{\mathrm{L}} + (p_0\gamma_1 + p_1\gamma_0)^{\mathrm{T}}\hat{\tau}_x, \quad \hat{\tau}_{\mathrm{F}} \dot{\sim} \hat{\tau}_{\mathrm{R}} \dot{\sim} \hat{\tau}_{\mathrm{L}} - (p_1 - p_0)(\gamma_1 - \gamma_0)^{\mathrm{T}}\hat{\tau}_x, \quad \hat{\tau}_{\mathrm{N}}^{\pi} \dot{\sim} \hat{\tau}_{\mathrm{L}}^{\pi} + \gamma^{\mathrm{T}}\hat{\tau}_x \ P_Z\text{-a.s.}$$

also from Lemma S7, we have

$$\begin{aligned}
\sqrt{N}\hat{\tau}_{\mathrm{N}} \mid \mathcal{A} &\rightsquigarrow v_{\mathrm{L}}^{1/2}\epsilon + v_x(p_0\gamma_1 + p_1\gamma_0)^{\mathrm{T}}D \mid (\|D\|_2^2 < a), && \text{(S11)}\\
\sqrt{N}\hat{\tau}_{\mathrm{N}}^{\pi} \mid \mathcal{A} &\rightsquigarrow v_{\mathrm{L}0}^{1/2}\epsilon + v_x\gamma^{\mathrm{T}}D \mid (\|D\|_2^2 < a) \qquad P_Z\text{-a.s.}, && \text{(S12)}\\
\sqrt{N}\hat{\tau}_* \mid \mathcal{A} &\rightsquigarrow v_{\mathrm{L}}^{1/2}\epsilon - v_x(p_1 - p_0)(\gamma_1 - \gamma_0)^{\mathrm{T}}D \mid (\|D\|_2^2 < a) \quad \text{for } * = \mathrm{R}, \mathrm{F}, && \text{(S13)}
\end{aligned}$$

with

$$\begin{aligned}
v_x(p_0\gamma_1 + p_1\gamma_0)^{\mathrm{T}}D &\sim \mathcal{N}(0, v_{\mathrm{N}} - v_{\mathrm{L}}), \quad v_x\gamma^{\mathrm{T}}D \sim \mathcal{N}(0, v_{\mathrm{N}0} - v_{\mathrm{L}0}),\\
-v_x(p_1 - p_0)(\gamma_1 - \gamma_0)^{\mathrm{T}}D &\sim \mathcal{N}(0, v_{\mathrm{R}} - v_{\mathrm{L}}). && \text{(S14)}
\end{aligned}$$

Let $u = (v_{\mathrm{N}} - v_{\mathrm{L}})^{-1/2}v_x(p_0\gamma_1 + p_1\gamma_0)$ be a unit vector with $\|u\|_2^2 = 1$ and $D_{\mathrm{N}} = u^{\mathrm{T}}D \sim \mathcal{N}(0,1)$ by the first "$\sim$" in (S14). Complete u into an orthogonal matrix Γ_{N} with u^{T} as the first row and D_{N} as the first element of $\Gamma_{\mathrm{N}}D$. With $\Gamma_{\mathrm{N}}D \sim \mathcal{N}(0_J, I_J)$ and $\|\Gamma_{\mathrm{N}}D\|_2^2 = \|D\|_2^2$, it follows from (S14) that

$$v_x(p_0\gamma_1 + p_1\gamma_0)^{\mathrm{T}}D \mid (\|D\|_2^2 < a) \sim (v_{\mathrm{N}} - v_{\mathrm{L}})^{1/2}D_{\mathrm{N}} \mid (\|\Gamma_{\mathrm{N}}D\|_2^2 < a) \sim (v_{\mathrm{N}} - v_{\mathrm{L}})^{1/2}\mathcal{L}.$$

Plugging this back in (S11) proves

$$\sqrt{N}\hat{\tau}_{\text{N}} \mid \mathcal{A} \rightsquigarrow v_{\text{L}}^{1/2}\epsilon + (v_{\text{N}} - v_{\text{L}})^{1/2}\mathcal{L} = v_{\text{N}}^{1/2}\big\{(1-\rho_{\text{N}}^2)^{1/2}\cdot\epsilon + \rho_{\text{N}}\cdot\mathcal{L}\big\} = v_{\text{N}}^{1/2}\cdot\mathcal{U}(\rho_{\text{N}}).$$

The same reasoning verifies $\sqrt{N}\hat{\tau}_{\text{N}}^{\pi} \mid \mathcal{A} \rightsquigarrow v_{\text{N0}}^{1/2}\cdot\mathcal{U}(\rho_{\text{N0}})$ and $\sqrt{N}\hat{\tau}_{*} \mid \mathcal{A} \rightsquigarrow v_{*}^{1/2}\cdot\mathcal{U}(\rho_{*})$ for $* = \text{R}, \text{F}$ from (S12) and (S13) with $u = (v_{\text{N0}} - v_{\text{L0}})^{-1/2}v_x\gamma$ and $u = (v_{\text{R}} - v_{\text{L}})^{-1/2}v_x(p_1 - p_0)(\gamma_1 - \gamma_0)$, respectively. This verifies the result for $\hat{\tau}_*$ and $\hat{\tau}_*^{\pi}$, and ensures $\sqrt{N}\hat{\tau}_*/v_*^{1/2} \mid \mathcal{A} \rightsquigarrow \mathcal{U}(\rho_*)$ for $* = \text{N}, \text{R}, \text{F}$, and

$$\sqrt{N}\hat{\tau}_{\text{N}}^{\pi}/v_{\text{N0}}^{1/2} \mid \mathcal{A} \rightsquigarrow \mathcal{U}(\rho_{\text{N0}}) \quad P_{\mathcal{Z}}\text{-a.s.}, \quad \sqrt{N}\hat{\tau}_*^{\pi}/v_{*0}^{1/2} \mid \mathcal{A} \rightsquigarrow \mathcal{N}(0,1) \quad P_{\mathcal{Z}}\text{-a.s.} \quad \text{for } * = \text{R}, \text{F}.$$

It follows from Lemma S8 and Slutsky's Theorem that

$$\begin{aligned}
\sqrt{N}\hat{\tau}_*/\hat{\text{se}}_* \mid \mathcal{A} &= \sqrt{N}\hat{\tau}_*/v_*^{1/2}\cdot(v_*/N\hat{\text{se}}_*^2)^{1/2} \mid \mathcal{A} \rightsquigarrow (c_*')^{1/2}\cdot\mathcal{U}(\rho_*), \\
\sqrt{N}\hat{\tau}_*/\tilde{\text{se}}_* \mid \mathcal{A} &= \sqrt{N}\hat{\tau}_*/v_*^{1/2}\cdot(v_*/N\tilde{\text{se}}_*^2)^{1/2} \mid \mathcal{A} \rightsquigarrow c_*^{1/2}\cdot\mathcal{U}(\rho_*)
\end{aligned}$$

for $* = \text{N}, \text{R}, \text{F}$, and

$$\begin{aligned}
\sqrt{N}\hat{\tau}_{\text{N}}^{\pi}/\hat{\text{se}}_{\text{N}}^{\pi} \mid \mathcal{A} &= \sqrt{N}\hat{\tau}_{\text{N}}^{\pi}/v_{\text{N0}}^{1/2}\cdot(v_{\text{N0}}/N\hat{\text{se}}_{\text{N}}^2)^{1/2} \mid \mathcal{A} \rightsquigarrow \mathcal{U}(\rho_{\text{N0}}), \\
\sqrt{N}\hat{\tau}_{\text{N}}^{\pi}/\tilde{\text{se}}_{\text{N}}^{\pi} \mid \mathcal{A} &= \sqrt{N}\hat{\tau}_{\text{N}}^{\pi}/v_{\text{N0}}^{1/2}\cdot(v_{\text{N0}}/N\tilde{\text{se}}_{\text{N}}^2)^{1/2} \mid \mathcal{A} \rightsquigarrow \mathcal{U}(\rho_{\text{N0}}), \\
\sqrt{N}\hat{\tau}_*^{\pi}/\hat{\text{se}}_*^{\pi} \mid \mathcal{A} &\rightsquigarrow \mathcal{N}(0,1), \quad \sqrt{N}\hat{\tau}_*^{\pi}/\tilde{\text{se}}_*^{\pi} \mid \mathcal{A} \rightsquigarrow \mathcal{N}(0,1) \quad \text{for } * = \text{R}, \text{F}
\end{aligned}$$

holds $P_{\mathcal{Z}}$-a.s.. This completes the proof. □

Proof of Corollary 5. Wider or equal asymptotic central quantile ranges imply greater or equal asymptotic variance. A test statistic is thus not proper if the asymptotic variance of its randomization distribution is not greater than or equal to that of its sampling distribution for all $\mathcal{S}$.

For the unadjusted $\hat{\tau}_{\text{N}}$, $\hat{\tau}_{\text{N}}/\hat{\text{se}}_{\text{N}}$, and $\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}$, that $v_{\text{L}}/v_{\text{N}}$ can be either greater or less than $v_{\text{L0}}/v_{\text{N0}}$ suggests $\rho_{\text{N}}/\rho_{\text{N0}}$, and thus $v(\rho_{\text{N}})/v(\rho_{\text{N0}})$, can be either greater or less than 1. We have

$$\frac{\text{var}_\infty(\hat{\tau}_{\text{N}})}{\text{var}_\infty(\hat{\tau}_{\text{N}}^{\pi|\mathcal{A}})} \to c_{\text{N}}'\cdot\frac{v(\rho_{\text{N}})}{v(\rho_{\text{N0}})}, \quad \frac{\text{var}_\infty(\hat{\tau}_{\text{N}}/\hat{\text{se}}_{\text{N}})}{\text{var}_\infty\{(\hat{\tau}_{\text{N}}/\hat{\text{se}}_{\text{N}})^{\pi|\mathcal{A}}\}} \to c_{\text{N}}'\cdot\frac{v(\rho_{\text{N}})}{v(\rho_{\text{N0}})}, \quad \frac{\text{var}_\infty(\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}})}{\text{var}_\infty\{(\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}})^{\pi|\mathcal{A}}\}} \to c_{\text{N}}\cdot\frac{v(\rho_{\text{N}})}{v(\rho_{\text{N0}})},$$

have limiting values that can be either greater or less than 1. This shows none of $\hat{\tau}_{\text{N}}$, $\hat{\tau}_{\text{N}}/\hat{\text{se}}_{\text{N}}$, and $\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}$ are proper under ReM. Likewise for

$$\frac{\text{var}_\infty(\hat{\tau}_*)}{\text{var}_\infty(\hat{\tau}_*^{\pi|\mathcal{A}})} \to c_{\text{R}}'\cdot v(\rho_{\text{R}}), \quad \frac{\text{var}_\infty(\hat{\tau}_*/\hat{\text{se}}_*)}{\text{var}_\infty\{(\hat{\tau}_*/\hat{\text{se}}_*)^{\pi|\mathcal{A}}\}} \to c_{\text{R}}'\cdot v(\rho_{\text{R}}) \quad \text{for } * = \text{R}, \text{F}$$

to have limiting values that can be either greater or less than 1 with c_{R}' can be either greater or less than 1. This shows none of $\hat{\tau}_*$ and $\hat{\tau}_*/\hat{\text{se}}_*$ are proper under ReM for $* = \text{R}, \text{F}$.

Further, it follows from Li and Ding (2020, Lemma A4) that $q_{1-\alpha/2}(\rho) \leq q_{1-\alpha/2}(0) = q_{1-\alpha/2}$ for any $0 < \alpha < 1$, where $q_{1-\alpha/2}(\rho)$ is the $(1-\alpha/2)$th quantile of $\mathcal{U}(\rho)$. This ensures $|\mathcal{U}(\rho)|$ is

stochastically dominated by $|\epsilon|$ for arbitrary $0 \leq \rho \leq 1$, and thus the properness of $\hat{\tau}_*/\hat{\text{se}}_*$ under ReM for $* = \text{R}, \text{F}, \text{L}$. □

Proof of Proposition 2. The sampling distributions of $\hat{\tau}_*$ follow from Li and Ding (2020, Theorem 2). The sampling distributions of $\hat{\tau}_*/\hat{\text{se}}_*$ and $\hat{\tau}_*/\tilde{\text{se}}_*$ follow from Slutsky's theorem and Li and Ding (2020, Lemma A5), which ensures Lemma S6 holds under ReM even if $d_i \neq x_i$. □

Proof of Corollary 6. The result follows from comparing the asymptotic sampling distributions in Proposition 2 with the asymptotic randomization distributions under unrestricted FRT in Theorems 1–4. The reasoning is almost identical to that in the proof of Corollary 5, with $|\mathcal{U}(\rho)|$ being stochastically dominated by $|\epsilon|$ for arbitrary $0 \leq \rho \leq 1$. □

S5. Extensions and connections

S5.1. Connection with the super-population framework

Let $\gamma = p_1\gamma_1 + p_0\gamma_0$, $\sigma^2 = p_1\sigma_1^2 + p_0\sigma_0^2$, $\sigma_a^2 = p_1\sigma_{a(1)}^2 + p_0\sigma_{a(0)}^2$, and $\sigma_b^2 = p_1\sigma_{b(1)}^2 + p_0\sigma_{b(0)}^2$ analogous to γ, S^2, S_a^2, and S_b^2 in the finite-population setting, respectively. Let $(\hat{\text{se}}'_{\text{L}})^2$ and $(\tilde{\text{se}}'_{\text{L}})^2$ be the unmodified classic and robust standard errors of $\hat{\tau}_{\text{L}}$ under the super-population framework such that $\hat{\text{se}}_{\text{L}}^2 = (\hat{\text{se}}'_{\text{L}})^2 + \hat{\theta}^{\text{T}} S_x^2 \hat{\theta}/N$ and $\tilde{\text{se}}_{\text{L}}^2 = (\tilde{\text{se}}'_{\text{L}})^2 + \hat{\theta}^{\text{T}} S_x^2 \hat{\theta}/N$.

Lemma S9. Assume Condition 2.

(a) Lemma S5 statements (a)–(c) hold P_{S}-a.s..

(b) Lemma S6 statement (a) hold P_{S}-a.s. after changing $\hat{\text{se}}_{\text{L}}^2$ and $\tilde{\text{se}}_{\text{L}}^2$ to $(\hat{\text{se}}'_{\text{L}})^2$ and $(\tilde{\text{se}}'_{\text{L}})^2$, respectively, and changing S^2 and S_*^2 to σ^2 and σ_*^2 for $* = a, b, z, a(z), b(z)$, where $z = 0, 1$.

Proof. The proof is almost identical with that of Lemmas S5 and S6 under the finite-population framework. The classical strong law of large numbers ensures Lemma S5 statement (a) and (S9) hold P_{S}-a.s. under Condition 2. The rest of Lemma S5 then follows. Proposition 1 is algebraic and thus holds also under the super-population framework after replacing x_i with $x_i - \bar{x}$. Statement (b) then follows from statement (a) as Lemma S6 follows from Lemma S5. □

Proof of Theorem 6. The result for the sampling distributions of $\hat{\tau}_*$, where $* = \text{N}, \text{F}, \text{L}$, including $v_{\text{L}} \leq v_*$ for $* = \text{N}, \text{F}$, follows from Negi and Wooldridge (2020). The asymptotic equivalence of $\sqrt{N}\hat{\tau}_{\text{R}}$ and $\sqrt{N}\hat{\tau}_{\text{F}}$ follows from $\sqrt{N}(\hat{\tau}_{\text{F}} - \hat{\tau}_{\text{R}}) = \sqrt{N}\hat{\tau}_x^{\text{T}}(\hat{\gamma}_{\text{F}} - \hat{\gamma}_{\text{R}}) = o_{P,\text{S}}(1)$ by Proposition 1, Lemma S9, and Slutsky's theorem.

The result for $\hat{\tau}_*/\hat{\text{se}}_*$ and $\hat{\tau}_*/\tilde{\text{se}}_*$ follows from Lemma S9 (b) and Slutsky's theorem via the same reasoning as in the proof of Theorems 1–4. In particular, recall $\hat{\gamma}_{\text{L},z} = (\hat{S}_{x(z)}^2)^{-1}\hat{S}_{xY(z)}$ as the coefficient of x_i from the OLS fit of Y_i on $(1, x_i)$ with units in treatment group z. Algebraically, we have $\hat{\theta} = \hat{\gamma}_{\text{L},1} - \hat{\gamma}_{\text{L},0}$ such that $\hat{\theta} = \gamma_1 - \gamma_0 + o(1)$ P_{S}-a.s. under Condition 2 with $\hat{\gamma}_{\text{L},z} =$

$\gamma_z + o(1)$ P_S-a.s.. This, together with $S_x^2 = \sigma_x^2 + o(1)$ P_S-a.s., ensures $\hat{\Delta}_{\bar{x}} = \Delta_{\bar{x}} + o(1)$ P_S-a.s. such that $N\hat{\text{se}}_L^2 = p_0^{-1}\sigma_{b(1)}^2 + p_1^{-1}\sigma_{b(0)}^2 + \Delta_{\bar{x}} + o(1)$ P_S-a.s. and $N\tilde{\text{se}}_L^2 = v_L + o(1)$ P_S-a.s. by Lemma S9(b).

The result for the randomization distributions follows from Lemma S9 (a) and the sampling distribution results in Theorems 1–4 via almost identical reasoning as the proof of the randomization distribution results in Theorems 1–4.

In particular, Lemma S9 (a) ensures the sequence of $\{(Y_i, Y_i, x_i)\}_{i=1}^N$ that the FRT procedure takes as fixed input satisfies Condition 1 P_S-a.s. under Condition 2, with $\hat{\gamma}_R$, $\hat{S}^2$, and $\hat{S}_e^2$ giving the analogs of γ_z, S_z^2, and both $S_{a(z)}^2$ and $S_{b(z)}^2$, respectively. This ensures the analog of $\Delta_{\bar{x}}$ for the sequence of $\{(Y_i, Y_i, x_i)\}_{i=1}^N$ equals zero such that $\hat{\Delta}_{\bar{x}}^\pi = o(1)$ P_S-a.s., and thus $\sqrt{N}\hat{\text{se}}_L^\pi - \sqrt{N}(\hat{\text{se}}_L')^\pi = o(1)$ P_S-a.s., and $\sqrt{N}\tilde{\text{se}}_L^\pi - \sqrt{N}(\tilde{\text{se}}_L')^\pi = o(1)$ P_S-a.s..

The result follows from replacing the limiting values of γ_z, S_z^2, $S_{a(z)}^2$, and $S_{b(z)}^2$ in the sampling distributions in Theorems 1–4 with those of $\hat{\gamma}_R$, $\hat{S}^2$, $\hat{S}_e^2$, and $\hat{S}_e^2$, respectively, namely $\lim_{N\to\infty}\hat{\gamma}_R = \gamma$, $\lim_{N\to\infty}\hat{S}^2 = p_1\sigma_1^2 + p_0\sigma_0^2 + p_1p_0\tau^2$ and $\lim_{N\to\infty}\hat{S}_e^2 = p_1\sigma_{a(1)}^2 + p_0\sigma_{a(0)}^2 + p_1p_0\tau^2$.

The nuance here is that we are using the sampling distributions rather than the randomization distributions in Theorems 1–4 for the above reasoning. The change would be "replacing S^2 and S_*^2 with σ^2 and σ_*^2 for $* = a, b, z, a(z), b(z)$" if we use the randomization distributions in Theorems 1–4 instead; see the proof of Theorems 1–4 for the correspondence between the sampling and randomization distributions under the finite-population framework in the first place. □

S5.2. Permutation tests for linear models

Recall $\delta = (I - H)Z$ as the residual vector from the OLS fit of Z on $(1_N, X)$, with $\delta_i = Z_i - p_1 - \lambda^{-1}p_1p_0x_i^T\hat{\tau}_x$ by (S6). Let $C = (\|\delta\|_2^2)^{-1} = \{Z^T(I-H)Z\}^{-1}$. Lemma S5(d) ensures

$$NC = (p_1p_0)^{-1} + o(1) \ P_Z\text{-a.s..} \tag{S15}$$

Lemma S10. For $(Y_i, x_i, Z_i)_{i=1}^N$ from arbitrary data generating process with $\bar{x} = 0_J$, and $\pi \in \Pi$,

(a) the coefficients have explicit forms

$$\hat{\tau}_R^\pi = \frac{Z_\pi^T(I-H)e}{Z_\pi^T(I-H_1)Z_\pi}, \qquad \hat{\tau}_F^\pi = \frac{Z_\pi^T(I-H)e}{Z_\pi^T(I-H)Z_\pi},$$
$$\hat{\beta}_{FL}^\pi = \hat{\beta}_K^\pi = \frac{Z^T(I-H)e_\pi}{Z^T(I-H)Z}, \qquad \hat{\beta}_{TB}^\pi - \hat{\tau}_F = \frac{Z^T(I-H)\epsilon_{F,\pi}}{Z^T(I-H)Z}, \qquad \hat{\beta}_M^\pi = \frac{Z^T(I-H)Y_\pi}{Z^T(I-H)Z};$$

(b) the classic standard errors have explicit forms

$$(\hat{\text{se}}_{FL}^\pi)^2 = \frac{1}{N-2-J}\left\{C\cdot\|e\|_2^2 - C\cdot e_\pi^THe_\pi - (\hat{\beta}_{FL}^\pi)^2\right\},$$
$$(\hat{\text{se}}_K^\pi)^2 = \frac{1}{N-2}\left\{C\cdot\|e\|_2^2 - (\hat{\beta}_K^\pi)^2\right\},$$
$$(\hat{\text{se}}_{TB}^\pi)^2 = \frac{1}{N-2-J}\left\{C\cdot\|\epsilon_F\|_2^2 - C\cdot\epsilon_{F,\pi}^TH\epsilon_{F,\pi} - (\hat{\beta}_{TB}^\pi - \hat{\tau}_F)^2\right\},$$

$$(\hat{\text{se}}_{\text{M}}^{\pi})^2 = \frac{1}{N-2-J}\left\{C \cdot \|Y\|_2^2 - C \cdot Y_{\pi}^{\text{T}} H Y_{\pi} - (\hat{\beta}_{\text{M}}^{\pi})^2\right\},$$

and the robust standard errors have a unified form

$$(\tilde{\text{se}}_{*}^{\pi})^2 = C^2(\eta_{*}^{\pi})^{\text{T}} \operatorname{diag}(\delta_i^2)\eta_{*}^{\pi} \qquad (* = \text{FL}, \text{K}, \text{TB}, \text{M})$$

where $\eta_{\text{FL}}^{\pi} = (I-H)e_{\pi} - \delta^{\text{T}}\hat{\beta}_{\text{FL}}^{\pi}$, $\eta_{\text{K}}^{\pi} = e_{\pi} - \delta^{\text{T}}\hat{\beta}_{\text{K}}^{\pi}$, $\eta_{\text{TB}}^{\pi} = (I-H)\epsilon_{\text{F},\pi} - \delta^{\text{T}}(\hat{\beta}_{\text{TB}}^{\pi} - \hat{\tau}_{\text{F}})$, and $\eta_{\text{M}}^{\pi} = (I-H)Y_{\pi} - \delta^{\text{T}}\hat{\beta}_{\text{M}}^{\pi}$.

Proof of Lemma S10. For $\hat{\tau}_{\text{R}}$, let $(I-H_1)Z$ and $(I-H_1)e$ be the residual vectors from the OLS fits of Z on 1_N and e on 1_N, respectively. By FWL, $\hat{\tau}_{\text{R}}$ equals the coefficient of $(I-H_1)Z$ from the OLS fit of $(I-H_1)e$ on $(I-H_1)Z$ by Lemma S1. This ensures

$$\hat{\tau}_{\text{R}} = \frac{Z^{\text{T}}(I-H_1)e}{Z^{\text{T}}(I-H_1)Z} = \frac{Z^{\text{T}}(I-H)e}{Z^{\text{T}}(I-H_1)Z},$$

where the last identity follows from $He = H_1 e = 0_J$ by (S6).

The result for $\hat{\tau}_{\text{F}}$ follows from Lemma S4 and $(I-H)Y = e$.

The result for the Freedman–Lane procedure follows from replacing Y with $Y_{\text{FL}}^{\pi} = HY + e_{\pi}$ in Lemma S4, with $(I-H)Y_{\text{FL}}^{\pi} = (I-H)(HY + e_{\pi}) = (I-H)e_{\pi}$.

For the Kennedy procedure, let $\hat{\beta}'$, $(\hat{\text{se}}')^2$, and $(\tilde{\text{se}}')^2$ be the coefficient and standard errors from the OLS fit of e_{π} on δ. By (S6), $\delta = (I-H_1)\delta$ and $e_{\pi} = (I-H_1)e_{\pi}$, so δ and e_{π} can also be viewed as the residual vectors from the OLS fits of δ on 1_N and e_{π} on 1_N, respectively. With

$$\hat{\beta}_{\text{K}}^{\pi} = \hat{\beta}', \quad (\hat{\text{se}}_{\text{K}}^{\pi})^2 = \frac{N-1}{N-2}(\text{se}')^2, \quad (\tilde{\text{se}}_{\text{K}}^{\pi})^2 = (\tilde{\text{se}}')^2$$

by FWL, the result follows from $\hat{\beta}' = C \cdot \delta^{\text{T}} e_{\pi}$, $(\text{se}')^2 = (N-1)^{-1}\{C \cdot \|e\|_2^2 - (\hat{\beta}')^2\}$, and $(\tilde{\text{se}}')^2 = C^2(e_{\pi} - \delta\hat{\beta}')^{\text{T}} \operatorname{diag}(\delta_i^2)(e_{\pi} - \delta\hat{\beta}')$ by replacing u with e_{π} and v with δ in Lemma S1.

Recall $\hat{\mu}_{\text{F}}$ and $\hat{\gamma}_{\text{F}}$ as the coefficients of 1_N and X in the OLS fit of Y on $(1_N, Z, X)$. The ter Braak procedure uses $Y_{\text{TB}}^{\pi} = 1_N\hat{\mu}_{\text{F}} + Z\hat{\tau}_{\text{F}} + X\hat{\gamma}_{\text{F}} + \epsilon_{\text{F},\pi}$ as the synthetic outcome vector for estimating $\hat{\beta}_{\text{TB}}^{\pi}$, $\hat{\text{se}}_{\text{TB}}^{\pi}$, and $\tilde{\text{se}}_{\text{TB}}^{\pi}$, with

$$\begin{aligned}
(I-H)Y_{\text{TB}}^{\pi} &= (I-H)(1_N\hat{\mu}_{\text{F}} + Z\hat{\tau}_{\text{F}} + X\hat{\gamma}_{\text{F}} + \epsilon_{\text{F},\pi}) = (I-H)(Z\hat{\tau}_{\text{F}} + \epsilon_{\text{F},\pi}), \\
(Y_{\text{TB}}^{\pi})^{\text{T}}(I-H)Y_{\text{TB}}^{\pi} &= (Z\hat{\tau}_{\text{F}} + \epsilon_{\text{F},\pi})^{\text{T}}(I-H)(Z\hat{\tau}_{\text{F}} + \epsilon_{\text{F},\pi}) \\
&= C^{-1}\hat{\tau}_{\text{F}}^2 + 2\hat{\tau}_{\text{F}}Z^{\text{T}}(I-H)\epsilon_{\text{F},\pi} + \epsilon_{\text{F},\pi}^{\text{T}}(I-H)\epsilon_{\text{F},\pi}
\end{aligned}$$

by (S6). Replace Y with Y_{TB}^{π} in Lemma S4 to see

$$\begin{aligned}
\hat{\beta}_{\text{TB}}^{\pi} &= C \cdot Z^{\text{T}}(I-H)Y_{\text{TB}}^{\pi} = \hat{\tau}_{\text{F}} + C \cdot Z^{\text{T}}(I-H)\epsilon_{\text{F},\pi}, \\
(\hat{\text{se}}_{\text{TB}}^{\pi})^2 &= \frac{1}{N-2-J}\left\{C \cdot (Y_{\text{TB}}^{\pi})^{\text{T}}(I-H)Y_{\text{TB}}^{\pi} - (\hat{\beta}_{\text{TB}}^{\pi})^2\right\} \\
&= \frac{1}{N-2-J}\left\{C \cdot \epsilon_{\text{F},\pi}^{\text{T}}(I-H)\epsilon_{\text{F},\pi} + 2\hat{\tau}_{\text{F}}(\hat{\beta}_{\text{TB}}^{\pi} - \hat{\tau}_{\text{F}}) + \hat{\tau}_{\text{F}}^2 - (\hat{\beta}_{\text{TB}}^{\pi})^2\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{N-2-J}\left\{C\cdot\|\epsilon_{\mathrm{F}}\|_2^2 - C\cdot\epsilon_{\mathrm{F},\pi}^{\mathrm{T}} H\epsilon_{\mathrm{F},\pi} - (\hat{\beta}_{\mathrm{TB}}^{\pi}-\hat{\tau}_{\mathrm{F}})^2\right\},\\
(\tilde{\mathrm{se}}_{\mathrm{TB}}^{\pi})^2 &= C^2(\eta_{\mathrm{TB}}^{\pi})^{\mathrm{T}}\,\mathrm{diag}(\delta_i^2)\eta_{\mathrm{TB}}^{\pi}.
\end{aligned}
$$

The result for $\hat{\beta}_{\mathrm{M}}^{\pi}$, $\hat{\mathrm{se}}_{\mathrm{M}}^{\pi}$, and $\tilde{\mathrm{se}}_{\mathrm{M}}^{\pi}$ follows from replacing Y with Y_π in Lemma S4. □

Let $\Lambda = \mathrm{diag}(\delta_i^2)$ and $\langle u, v\rangle = u^{\mathrm{T}}\Lambda v$ for $u = (u_1,\ldots,u_N)^{\mathrm{T}}$ and $v = (v_1,\ldots,v_N)^{\mathrm{T}}$ be the corresponding inner product to simplify the presentation. The Cauchy–Schwarz inequality implies

$$
\langle u,u\rangle = \sum_{i=1}^{N} u_i^2\delta_i^2 \le \left(\|u\|_4^4\right)^{1/2}\left(\|\delta\|_4^4\right)^{1/2}, \quad \langle u,v\rangle \le \langle u,u\rangle^{1/2}\langle v,v\rangle^{1/2}. \tag{S16}
$$

Lemma S11. Assume Condition 1 and a sequence of Z that satisfies Lemma S5 statements (a)–(e).

(a) $N^{-1/2}\delta^{\mathrm{T}}e_\pi \rightsquigarrow \mathcal{N}\{0, p_1p_0(S_a^2+p_1p_0\tau^2)\}$, $\quad N^{-1/2}\delta^{\mathrm{T}}\epsilon_{\mathrm{F},\pi} \rightsquigarrow \mathcal{N}(0, p_1p_0S_a^2)$,
$N^{-1/2}\delta^{\mathrm{T}}Y_\pi \rightsquigarrow \mathcal{N}\{0, p_1p_0(S^2+p_1p_0\tau^2)\}$.

(b) $N^{-1}X^{\mathrm{T}}e_\pi = o_{P,\pi}(1)$, $\quad N^{-1}X^{\mathrm{T}}\epsilon_{\mathrm{F},\pi} = o_{P,\pi}(1)$, $\quad N^{-1}X^{\mathrm{T}}Y_\pi = o_{P,\pi}(1)$.

(c) $N^{-1}\langle e_\pi, e_\pi\rangle = p_1p_0(S_a^2+p_1p_0\tau^2)+o_{P,\pi}(1)$, $\quad N^{-1}\langle \epsilon_{\mathrm{F},\pi},\epsilon_{\mathrm{F},\pi}\rangle = p_1p_0S_a^2+o_{P,\pi}(1)$,
$N^{-1}\langle Y_\pi - 1_N\hat{Y}, Y_\pi - 1_N\hat{Y}\rangle = p_1p_0(S^2+p_1p_0\tau^2)+o_{P,\pi}(1)$.

Proof. Statement (a) follows from Lemma S2(b) by letting $u=\delta$ and $v = e, \epsilon_{\mathrm{F}}, Y$, respectively, with the respective means, variances, and bounded fourth moments following from Lemma S5.

For statement (b), $N^{-1}X_j^{\mathrm{T}}e_\pi = o_{P,\pi}(1)$ follows from Markov's inequality given $E(N^{-1}X_j^{\mathrm{T}}e_\pi) = 0$ and $\mathrm{var}(N^{-1}X_j^{\mathrm{T}}e_\pi) = N^{-1}\lambda\hat{S}_e^2 = o(1)$ by Lemmas S2(a) and S5(c). The proof for the rest of statement (b) is almost identical and thus omitted.

For statement (c), with a slight abuse of notation, let $\delta^2 = (\delta_1^2,\ldots,\delta_N^2)^{\mathrm{T}}$, $e^2 = (e_1^2,\ldots,e_N^2)^{\mathrm{T}}$, $\epsilon_{\mathrm{F}}^2 = (\epsilon_{\mathrm{F},1}^2,\ldots,\epsilon_{\mathrm{F},N}^2)^{\mathrm{T}}$, and $\nu = (\nu_1,\ldots,\nu_N)^{\mathrm{T}}$, where $\nu_i = (Y_i-\hat{Y})^2$, to write

$$
\begin{aligned}
&N^{-1}\langle e_\pi,e_\pi\rangle = N^{-1}\sum_{i=1}^{N}\delta_i^2e_{\pi(i)}^2 = N^{-1}(\delta^2)^{\mathrm{T}}(e^2)_\pi, \quad N^{-1}\langle\epsilon_{\mathrm{F},\pi},\epsilon_{\mathrm{F},\pi}\rangle = N^{-1}(\delta^2)^{\mathrm{T}}(\epsilon_{\mathrm{F}}^2)_\pi,\\
&N^{-1}\langle Y_\pi - 1_N\hat{Y}, Y_\pi - 1_N\hat{Y}\rangle = N^{-1}(\delta^2)^{\mathrm{T}}v_\pi.
\end{aligned}
$$

With $\hat{\delta^2} = N^{-1}\|\delta\|_2^2 = p_1p_0+o(1)$, $\hat{e^2} = N^{-1}\|e\|_2^2 = S_a^2+p_1p_0\tau^2+o(1)$ and $S_{\delta^2}^2 = (N-1)^{-1}\sum_{i=1}^N(\delta_i^2-\hat{\delta^2})^2 = O(1)$, $S_{e^2}^2 = (N-1)^{-1}\sum_{i=1}^N(e_i^2-\hat{e^2})^2 = O(1)$ by Lemma S5(c)–(d), Lemma S2(a) ensures $E\left\{N^{-1}(\delta^2)^{\mathrm{T}}(e^2)_\pi\right\} = \hat{\delta^2}\hat{e^2} = p_1(S_a^2+p_1p_0\tau^2)+o(1)$ and $\mathrm{var}\left\{N^{-1}(\delta^2)^{\mathrm{T}}(e^2)_\pi\right\} = N^{-1}\lambda S_{\delta^2}^2S_{e^2}^2 = o(1)$, which, coupled with Markov's inequality, imply

$$
N^{-1}\langle e_\pi,e_\pi\rangle = E\{N^{-1}(\delta^2)^{\mathrm{T}}(e^2)_\pi\}+o_{P,\pi}(1) = p_1p_0(S_a^2+p_1p_0\tau^2)+o_{P,\pi}(1).
$$

The proof for the rest of statement (c) is almost identical, with $\hat{\epsilon_{\mathrm{F}}^2} = N^{-1}\|\epsilon_{\mathrm{F}}\|_2^2 = S_a^2+o(1)$, $S_{\epsilon^2}^2 = (N-1)\sum_{i=1}^N(\epsilon_{\mathrm{F},i}^2-\hat{\epsilon_{\mathrm{F}}^2})^2 = O(1)$, $\hat{\nu} = \lambda\hat{S}^2 = S^2+p_1p_0\tau^2+o(1)$, and $S_\nu = (N-1)^{-1}\sum_{i=1}^N(v_i-\hat{\nu})^2 = O(1)$ by Lemma S5 statements (b) and (e). □

Freedman and Lane (1983) and Anderson and Robinson (2001) sketched proofs for the reference distributions of the statistics studentized by the classic standard errors. Below we give a unified proof for the statistics studentized by both the classic and robust standard errors.

Proof of Theorem 7. We first verify the asymptotic normality of $\sqrt{N}\hat{\beta}^\pi_*$ for $* = \text{FL}, \text{K}, \text{TB}, \text{M}$, and then prove the result on the studentized variants based on it.

Unstudentized coefficients Lemma S10(a) ensures $\sqrt{N}\hat{\beta}^\pi_{\text{FL}} = \sqrt{N}\hat{\beta}^\pi_{\text{K}} = (NC)\cdot N^{-1/2}\delta^{\text{T}}e_\pi$, $\sqrt{N}(\hat{\beta}^\pi_{\text{TB}} - \hat{\tau}_{\text{F}}) = (NC)\cdot N^{-1/2}\delta^{\text{T}}\epsilon_{\text{F},\pi}$ and $\sqrt{N}\hat{\beta}^\pi_{\text{M}} = (NC)\cdot N^{-1/2}\delta^{\text{T}}Y_\pi$, where $NC = (p_1p_0)^{-1} + o(1)$ P_Z-a.s. by (S15). The asymptotic normality of $\hat{\beta}^\pi_*$ follows from Lemma S11(a) by Slutsky's theorem. This ensures

$$\hat{\beta}^\pi_* = o_{P,\pi}(1) \quad P_Z\text{-a.s.} \quad \text{for } * = \text{FL}, \text{K}, \text{M}, \quad \hat{\beta}^\pi_{\text{TB}} - \hat{\tau}_{\text{F}} = o_{P,\pi}(1) \quad P_Z\text{-a.s.}. \tag{S17}$$

Studentized coefficients The result for $\hat{\beta}^\pi_*/\hat{\text{se}}^\pi_*$ and $\hat{\beta}^\pi_*/\tilde{\text{se}}^\pi_*$ follows from Slutsky's theorem and

$$\begin{aligned}
&N(\hat{\text{se}}^\pi_{\text{FL}})^2 = (p_1p_0)^{-1}S_a^2 + \tau^2 + o_{P,\pi}(1), && N(\tilde{\text{se}}^\pi_{\text{FL}})^2 = (p_1p_0)^{-1}S_a^2 + \tau^2 + o_{P,\pi}(1),\\
&N(\hat{\text{se}}^\pi_{\text{K}})^2 = (p_1p_0)^{-1}S_a^2 + \tau^2 + o_{P,\pi}(1), && N(\tilde{\text{se}}^\pi_{\text{K}})^2 = (p_1p_0)^{-1}S_a^2 + \tau^2 + o_{P,\pi}(1),\\
&N(\hat{\text{se}}^\pi_{\text{TB}})^2 = (p_1p_0)^{-1}S_a^2 + o_{P,\pi}(1), && N(\tilde{\text{se}}^\pi_{\text{TB}})^2 = (p_1p_0)^{-1}S_a^2 + o_{P,\pi}(1),\\
&N(\hat{\text{se}}^\pi_{\text{M}})^2 = (p_1p_0)^{-1}S^2 + \tau^2 + o_{P,\pi}(1), && N(\tilde{\text{se}}^\pi_{\text{M}})^2 = (p_1p_0)^{-1}S^2 + \tau^2 + o_{P,\pi}(1)
\end{aligned} \tag{S18}$$

hold P_Z-a.s..

We finish the proof by verifying (S18). Lemma S5 ensures it suffices to focus on sequences of Z that satisfy Lemma S5 statements (a)–(e) and (S17). Fix one such sequence with $Nc = (p_1p_0)^{-1} + o(1)$ by (S15).

For the classic standard errors $\hat{\text{se}}^\pi_*$, with $NC = (p_1p_0)^{-1} + o(1)$, $\hat{\beta}^\pi_* = o_{P,\pi}(1)$ by (S17), and

$$N^{-1}\|e\|_2^2 = S_a^2 + p_1p_0\tau^2 + o(1), \quad N^{-1}\|\epsilon_{\text{F}}\|_2^2 = S_a^2 + o(1), \quad N^{-1}\|Y\|_2^2 - \hat{Y}^2 = S^2 + p_1p_0\tau^2 + o(1)$$

by Lemma S5, a direct comparison of the expressions of $(\hat{\text{se}}^\pi_*)^2$ in Lemma S10(b) with (S18) suggests it suffices to verify

$$N^{-1}e_\pi^{\text{T}}He_\pi = o_{P,\pi}(1), \quad N^{-1}\epsilon_{\text{F},\pi}^{\text{T}}H\epsilon_{\text{F},\pi} = o_{P,\pi}(1), \quad N^{-1}Y_\pi^{\text{T}}HY_\pi - \hat{Y}^2 = o_{P,\pi}(1). \tag{S19}$$

Recall $H = N^{-1}1_N1_N^{\text{T}} + (N-1)^{-1}XX^{\text{T}}$ from (S6) to write

$$He_\pi = \lambda^{-1}N^{-1}XX^{\text{T}}e_\pi, \quad H\epsilon_{\text{F},\pi} = \lambda^{-1}N^{-1}XX^{\text{T}}\epsilon_{\text{F},\pi}, \quad HY_\pi = 1_N\hat{Y} + \lambda^{-1}N^{-1}XX^{\text{T}}Y_\pi. \tag{S20}$$

By Lemma S11(b), $N^{-1}e_\pi^{\text{T}}He_\pi = \lambda^{-1}(N^{-1}e_\pi^{\text{T}}X)(N^{-1}X^{\text{T}}e_\pi) = o_{P,\pi}(1)$ and likewise $N^{-1}\epsilon_{\text{F},\pi}^{\text{T}}H\epsilon_{\text{F},\pi} = o_{P,\pi}(1)$ and $N^{-1}Y_\pi^{\text{T}}HY_\pi = \hat{Y}^2 + o_{P,\pi}(1)$. This verifies (S19) and thus the result for $\hat{\text{se}}^\pi_*$.

For the robust standard errors, with $N(\tilde{\text{se}}^\pi_*)^2 = (NC)^2\cdot N^{-1}\langle\eta^\pi_*, \eta^\pi_*\rangle$ by Lemma S10(b), where

$NC = (p_1 p_0)^{-1} + o(1)$, it suffices to determine the probability limits of $N^{-1}\langle \eta_*^\pi, \eta_*^\pi \rangle$ for

$$\begin{aligned} &\eta_{\mathrm{FL}}^\pi = (I-H)e_\pi - \delta^{\mathrm{T}}\hat{\beta}_{\mathrm{FL}}^\pi, && \eta_{\mathrm{K}}^\pi = e_\pi - \delta^{\mathrm{T}}\hat{\beta}_{\mathrm{K}}^\pi, \\ &\eta_{\mathrm{TB}}^\pi = (I-H)\epsilon_{\mathrm{F},\pi} - \delta^{\mathrm{T}}(\hat{\beta}_{\mathrm{TB}}^\pi - \hat{\tau}_{\mathrm{F}}), && \eta_{\mathrm{M}}^\pi = (I-H)Y_\pi - \delta\hat{\beta}_{\mathrm{M}}^\pi = (I-H)(Y - 1_N\hat{Y})_\pi - \delta\hat{\beta}_{\mathrm{M}}^\pi. \end{aligned}$$

Lemma S5 ensures

$$\|Y\|_4^4 = O(N), \quad \|X_j\|_4^4 = O(N), \quad \|e\|_4^4 = O(N), \quad \|\delta\|_4^4 = O(N), \quad \|\epsilon_{\mathrm{F}}\|_4^4 = O(N). \tag{S21}$$

For $\eta_{\mathrm{K}}^\pi = e_\pi - \delta\hat{\beta}_{\mathrm{K}}^\pi$, (S16)–(S21) ensures $N^{-1}\langle \delta, \delta \rangle = O(1)$ and $N^{-1}\langle e_\pi, \delta \rangle = O(1)$ such that

$$\begin{aligned} N^{-1}\langle \eta_{\mathrm{K}}^\pi, \eta_{\mathrm{K}}^\pi \rangle &= N^{-1}\langle e_\pi, e_\pi \rangle - 2N^{-1}\langle e_\pi, \delta \rangle \cdot \hat{\beta}_{\mathrm{K}}^\pi + N^{-1}\langle \delta, \delta \rangle \cdot (\hat{\beta}_{\mathrm{K}}^\pi)^2 \\ &= N^{-1}\langle e_\pi, e_\pi \rangle + o(1) = p_1 p_0 (S_a^2 + p_1 p_0 \tau^2) + o_{P,\pi}(1) \end{aligned} \tag{S22}$$

by (S17) and $N^{-1}\langle e_\pi, e_\pi \rangle = p_1 p_0 (S_a^2 + p_1 p_0 \tau^2)$ from Lemma S11(c). This verifies the result for $\tilde{\mathrm{se}}_{\mathrm{K}}^\pi$.

For $\eta_{\mathrm{FL}}^\pi = e_\pi - \delta\hat{\beta}_{\mathrm{FL}}^\pi - He_\pi$, (S20) ensures

$$N^{-1}\langle He_\pi, He_\pi \rangle = \lambda^{-2}(N^{-1}X^{\mathrm{T}}e_\pi)^{\mathrm{T}}(N^{-1}X^{\mathrm{T}}\Lambda X)(N^{-1}X^{\mathrm{T}}e_\pi) = o_{P,\pi}(1),$$

where the last identity follows from $N^{-1}X^{\mathrm{T}}e_\pi = o_{P,\pi}(1)$ by Lemma S11(b) and $N^{-1}X^{\mathrm{T}}\Lambda X = O(1)$ by $N^{-1}\|\delta\|_4^4 = O(1)$ and $N^{-1}\|X_j\|_4^4 = O(1)$ from (S21). This, together with

$$N^{-1}\langle e_\pi - \delta\hat{\beta}_{\mathrm{FL}}^\pi, e_\pi - \delta\hat{\beta}_{\mathrm{FL}}^\pi \rangle = N^{-1}\langle e_\pi, e_\pi \rangle + o(1) = p_1 p_0 (S_a^2 + p_1 p_0 \tau^2) + o_{P,\pi}(1)$$

by the same reasoning as (S22), ensures

$$N^{-1}\langle e_\pi - \delta\hat{\beta}_{\mathrm{FL}}^\pi, He_\pi \rangle \leq \Big(N^{-1}\langle e_\pi - \delta\hat{\beta}_{\mathrm{FL}}^\pi, e_\pi - \delta\hat{\beta}_{\mathrm{FL}}^\pi \rangle\Big)^{1/2}\Big(N^{-1}\langle He_\pi, He_\pi \rangle\Big)^{1/2} = o_{P,\pi}(1)$$

by (S16) and thus

$$\begin{aligned} N^{-1}\langle \eta_{\mathrm{FL}}^\pi, \eta_{\mathrm{FL}}^\pi \rangle &= N^{-1}\langle e_\pi - \delta\hat{\beta}_{\mathrm{FL}}^\pi - He_\pi, e_\pi - \delta\hat{\beta}_{\mathrm{FL}}^\pi - He_\pi \rangle \\ &= N^{-1}\langle e_\pi - \delta\hat{\beta}_{\mathrm{FL}}^\pi, e_\pi - \delta\hat{\beta}_{\mathrm{FL}}^\pi \rangle + N^{-1}\langle He_\pi, He_\pi \rangle - 2N^{-1}\langle e_\pi - \delta\hat{\beta}_{\mathrm{FL}}^\pi, He_\pi \rangle \\ &= p_1 p_0 (S_a^2 + p_1 p_0 \tau^2) + o_{P,\pi}(1). \end{aligned}$$

This verifies the result for $\tilde{\mathrm{se}}_{\mathrm{FL}}^\pi$.

The result for $\tilde{\mathrm{se}}_{\mathrm{TB}}^\pi$ and $\tilde{\mathrm{se}}_{\mathrm{M}}^\pi$ follows from almost identical reasoning after replacing e_π with $\epsilon_{\mathrm{F},\pi}$ and $Y_\pi - 1_N\hat{Y}$ in the proof of $\tilde{\mathrm{se}}_{\mathrm{FL}}^\pi$. □