Unifying Design-based Inference: On Bounding and Estimating the Variance of any Linear Estimator in any Experimental Design

WORKING PAPER 1 OF 4

Joel A. Middleton[*]

April 4, 2021

[*] Charles and Louise Travers Department of Political Science, *University of California, Berkeley.* email: joel.middleton@gmail.com

1 Introduction

This paper provides a design-based framework for variance (bound) estimation in experimental analysis. Results are applicable to virtually any combination of experimental design, linear estimator (e.g., difference-in-means, OLS, WLS) and variance bound, allowing for unified treatment and a basis for systematic study and compairison of designs using matrix spectral analysis. A proposed variance estimator reproduces Eicker-Huber-White (aka. "robust", "heteroskedastic consistent", "sandwich", "White", "Huber-White", "HC", etc.) standard errors and "cluster-robust" standard errors as special cases. While past work has shown algebraic equivalences between design-based and the so-called "robust" standard errors under some designs, this paper motivates them for a wide array of design-estimator-bound triplets. In so doing, it provides a clearer and more general motivation for "robust" variance estimators.

1.1 The Neyman Causal Model

Consider a randomized experiment with k treatment arms. The Neyman causal model (NCM) assumes that the units in the experimental study represent a finite population of size n. For a given outcome measure, call it y, each unit, i, responds with one of k possible values in $\{y_{1i}, y_{2i}, ..., y_{ki}\}$, depending on their treatment assignment. The possible responses are referred to as the *potential outcomes.* In the NCM these values are considered (nonrandom) constants, which stands in contrast to other, more common, formulations where potential outcomes are assumed to be sampled from some (possibly nonparametric) distribution.

The only random element in the NCM is the treatment assignment indicators $\{R_{1i}, R_{2i}, ..., R_{ki}\}$, and they determine which potential outcome will be observed by the researcher. Since a unit can only be assigned to one arm of the experiment, only one of the indicators will realize a value of one, and the rest will be zero, such that $R_{1i} + R_{2i} + ... + R_{ki} = 1$ for all i.

A standard representation of the *observed* outcome for the i^{th} unit under the NCM would be,

$$Y_i^{obs} = y_{1i}R_{1i} + y_{2i}R_{2i} + ... + y_{ki}R_{ki},$$

which is itself random, due to the assignment indicators. For each unit, the observed data can then be represented as $\{Y_i^{obs}, R_{1i}, R_{2i}, ..., R_{ki}, x_i\}_{\forall i}$, where x_i is an additional vector of l covariates. Like the potential outcomes, x_i is nonrandom, but unlike the potential outcomes the same value is observed irrespective of the assignment.

Ideally, we would like to know, for a given individual, i, the difference between responses under various arms, called a *treatment effect.* It is clear from the definition of Y_i^{obs}, however, that individual treatment effects are not observable since only one of the potential outcomes can be observed for an individual, a problem known as *fundamental problem of causal inference* (Holland, 1986). As a result, researchers often try to estimate *averages* of across the units in study.

Example (*Treatment/Control Experiment*) : In an experiment with a control group (arm 0) and a treatment group (arm 1) the individual-level treatment effect, $y_{1i} - y_{0i}$, but this is not identified, so a researcher might try to estimate the average treatment effect $n^{-1}\sum_i (y_{1i} - y_{0i})$. △

Example (2×2 *Factorial Experiment*) : Consider a 2×2 factorial design with treatments A and B. Units in arm 1 are controls (no treatments), units in arm 2 are given treatment A only, units in arm 3 are given B only, and units in arm 4 are given both A and B. Similar to the treatment/control example, one could contrast the mean of an arm with a single treatment against the control mean, e.g., the average effect of A compared to no treatments, $n^{-1}\sum_i (y_{2i} - y_{1i})$. Another quantity of interest might be an *average marginal causal effect* (AMCE), e.g., the effect of A marginalizing over the levels of B, $n^{-1}\sum_i \frac{1}{2}(y_{2i} - y_{1i} + y_{4i} - y_{3i})$. Another example might be an omnibus test based on the contrast $n^{-1}\sum_i (y_{2i}/3 + y_{3i}/3 + y_{4i}/3 - y_{1i})$. △

Target quantities such as *local average treatment effects* or *conditional average treatment effects* might also be considered in this framework, but the primary focus of this paper is *variance* estimation for linear estimators for virtually any design.

Suffice to say that developing variance estimators before considering point estimation is appealing, if somewhat counter-intuitive, for two reasons. On the one hand, asymptotic analysis for point estimators can be made easier by having first established general variance expressions (for all linear estimators and virtually any design). On the other hand, a general framework for variance (bound) estimation can be developed even while a particular estimation target has yet to be defined, and even if an "estimator" does not estimate anything of interest, it's variance can still be studied.

1.2 Notation

To simplify notation, let y_1, y_2,...,y_k represent length n vectors of potential outcomes associated with each of the arms, with the i^{th} element of each corresponding to the i^{th} unit. Next, stack these vectors to create

$$y := (y_1' \ \ y_2' \ \ \ldots \ \ y_k')',$$

which is a column vector and has length kn containing all k potential outcomes for all n units.

Next, if we let 1_n be a n-length vector of ones, then a $kn \times k$ *intercept matrix* can be defined as,

$$\mathbb{1} := \begin{bmatrix} 1_n & & & \\ & 1_n & & \\ & & \ddots & \\ & & & 1_n \end{bmatrix},$$

which, for example, allows us to express a k-length vector of the means of each arm as $\frac{1}{n}\mathbb{1}'y$, or, equivalently, $(\mathbb{1}'\mathbb{1})^{-1}\mathbb{1}'y$. Next, define c as the *contrast vector*, of length k, such that $c'(\mathbb{1}'\mathbb{1})^{-1}\mathbb{1}'y$ gives contrasts between potential outcome means for the various arms.

Example (*Treatment/Control Experiment, continued*) : With two arms, control (arm 1) and treatment (arm 2), define $c = (-1 \ \ 1)'$. Then the *average treatment effect* is simply $c'(\mathbb{1}'\mathbb{1})^{-1}\mathbb{1}'y = n^{-1}\sum_i (y_{2i} - y_{1i})$. △

Example (2×2 *Factorial Experiment, continued*) : In a four-arm experiment, if $c = (-1 \ \ 1 \ \ 0 \ \ 0)'$ then $c'(\mathbb{1}'\mathbb{1})^{-1}\mathbb{1}'y = n^{-1}\sum_i (y_{2i} - y_{1i})$ is the avearge difference between the first two arms. Alternatively, if the researcher chooses $c = (-\frac{1}{2} \ \ \frac{1}{2} \ \ -\frac{1}{2} \ \ \frac{1}{2})'$ then $c'(\mathbb{1}'\mathbb{1})^{-1}\mathbb{1}'y = n^{-1}\sum_i \frac{1}{2}(y_{2i} - y_{1i} + y_{4i} - y_{3i})$. △

Next define an $n \times n$ diagonal matrix that has all n assignment indicators for treatment arm 1 on the diagonal,

$$\mathbf{R}_1 := \begin{bmatrix} R_{11} & & & & & \\ & R_{12} & & & & \\ & & \ddots & & & \\ & & & R_{1i} & & \\ & & & & \ddots & \\ & & & & & R_{1n} \end{bmatrix},$$

and define $\mathbf{R}_2$, $\mathbf{R}_3$, $\ldots$, $\mathbf{R}_k$ analogously. Arrange these matrices to create the diagonal $kn \times kn$ matrix

$$\mathbf{R} := \begin{bmatrix} \mathbf{R}_1 & & & \\ & \mathbf{R}_2 & & \\ & & \ddots & \\ & & & \mathbf{R}_k \end{bmatrix}$$

and note the a $kn \times kn$ diagonal matrix of assignment probabilities can be written as $\boldsymbol{\pi} := \mathrm{E}[\mathbf{R}]$, with the first n diagonal elements representing probabilities of assignment to arm 1, then the next n diagonal elements are probabilities of assignment to arm 2 and so on.

In this alternative notation the researcher can be said to observe the assignment, $\mathbf{R}$, the observed vector of outcomes, $\mathbf{R}y$, and also a matrix of l pre-treatment covariates, $\mathbf{x}$, which has size $n \times l$. In a randomized experiment $\boldsymbol{\pi}$ is also observed (known) in many cases. When intractable analytically, however, it might be estimated to arbitrary precision by repeating the original randomization until a target level of precision is achieved.

For covariate adjusted estimators, it is convenient to define the $kn \times (k + l)$ matrix,

$$\mathbb{x} := \begin{bmatrix} 1_n & & & \mathbf{x} \\ & 1_n & & \mathbf{x} \\ & & \ddots & \vdots \\ & & 1_n & \mathbf{x} \end{bmatrix}$$

which augments the intercept vector, $\mathbb{1}$, with covariates.

Remark 1. *For some cases, such adjusting for covariates separately by arm, it might be useful to define $\mathbb{x}$ with $\mathbf{x}$ matrices arranged along a block-diagonal. In that case, it is useful to stipulate that $\mathbf{x}$ have columns that sum to zero to avoid problems of coefficient interpretation (cf. Lin, 2013; Middleton, 2018). This will be discussed further in paper 3 of 4.*

2 Linear estimators

This paper focuses on the variance, bounding and variance bound estimation of the class of estimators that are linear in the observed outcome, y. This class includes everything from the difference-of-means, to the Horvitz-Thomposon estimator, to regression.

Note, however, that beyond presenting a general approach to variance bound estimation for the class of linear estimators, point estimation itself will be the focus of the third and fourth papers in the series. Questions such as consistency will be and causal identification will be considered then. For now, suffice it to be said that an estimator need not be consistent for any quantity of interest at all (causal or otherwise) in order to derive variance expressions for it.

2.1 Definition

Definition 2.1 (Linear Estimators)**.** *Linear estimators are defined as having the form,*

$$\widehat{\delta}_c := c'\mathbf{W}\mathbf{R}y, \tag{1}$$

where $\mathbf{W}$ *a matrix with* kn *columns and* k *rows if it is an unadjusted estimator and* $k+l$ *rows if it is a covariate adjusted estimator. The length of the contrast vector,* c*, is equal to the number of rows in* $\mathbf{W}$*. The first* k *entries of* c *are the contrast values, followed by* l *zeros in the case of covariate adjusted estimators.*

Also, for convenience, define $\mathbf{w}$ to be $\mathbf{W}$ evaluated at $\mathbf{R} = \boldsymbol{\pi}$, i.e.,

$$\mathbf{w} := \{\left.\mathbf{W}\right|_{\mathbf{R}=\boldsymbol{\pi}}\}. \tag{2}$$

Definition 2.2 (Horvitz-Thompson estimator)**.** *The Horvitz-Thompson estimator written as in Definition 2.1 with,*

$$\mathbf{W} = \mathbf{W}^{\text{HT}} := \left(\mathbb{1}'\mathbb{1}\right)^{-1}\mathbb{1}'\boldsymbol{\pi}^{-1},$$

noting that $\mathbf{W}^{\text{HT}} = \mathbf{w}^{\text{HT}}$ *since* $\mathbf{W}^{\text{HT}}$ *is nonrandom.*

Definition 2.3 (Contrast-of-means)**.** *Contrast-of-means (e.g., difference-of-means) can be written as in Definition 2.1 with,*

$$\mathbf{W} = \mathbf{W}^{\text{CM}} := \left(\mathbb{1}'\mathbf{R}\mathbb{1}\right)^{-1}\mathbb{1}'.$$

Definition 2.4 (Hajek estimator)**.** *The Hajek estimator can be written as Definition (2.1) with,*

$$\mathbf{W} = \mathbf{W}^{\text{HJ}} := \left(\mathbb{1}'\boldsymbol{\pi}^{-1}\mathbf{R}\mathbb{1}\right)^{-1}\mathbb{1}'\boldsymbol{\pi}^{-1}.$$

Definition 2.5 (OLS estimator)**.** *The OLS estimator can be written as Definition (2.1) with,*

$$\mathbf{W} = \mathbf{W}^{\text{OLS}} := \left(\mathbb{1}'\mathbf{R}\mathbb{1}\right)^{-1}\mathbb{1}'.$$

Definition 2.6 (WLS estimators)**.** *WLS estimators can be written as in Definition 2.1 with,*

$$\mathbf{W} = \mathbf{W}^{\text{WLS}} := \left(\mathbb{x}'\mathbf{m}\mathbf{R}\mathbb{x}\right)^{-1}\mathbb{x}'\mathbf{m}.$$

Remark 2. *WLS is a class that includes OLS, Hajek and contrast-of-means (e.g., difference-of-means) as special cases. It is equivalent to OLS when* $\mathbf{m} = \mathbf{i}_{kn}$ *(*$\mathbf{i}_{kn}$ *is the identity matrix). If* $\mathbf{m} = \mathbf{i}_{kn}$ *and, in addition,* $\mathbb{x} = \mathbb{1}$*, WLS is OLS without covariates, which is equivalent to the contrast-of-means (e.g., in the two-arm case, we call this the difference-of-means), underscoring Theorem 1 in Freedman (2008a). If* $\mathbf{m} = \boldsymbol{\pi}^{-1}$ *and* $\mathbb{x} = \mathbb{1}$*, then it is the Hajek estimator. The covariate adjusted WLS with* $\mathbf{m} = \boldsymbol{\pi}^{-1}$ *will be discussed further in paper 3 of 4, because it is algebraically equivalent to the generalized regression estimator introduced there.*

2.2 First-order Taylor approximation

In this section, a general approach to obtaining asymptotically valid variance expressions for linear estimators is given using a first-order approximation of a Tyalor series. The method is often used when an exact, closed-form variance expression is not tractable, as may be the case with any number of linear estimators. Examination of the W vectors defined above shows that, with the exception of Horvitz-Thompson, the estimators all had random denominators (i.e., inverted random matrices), making closed form variance expressions difficult.

The original estimator and its Taylor approximation are asymptotically equivalent (cite Pashley). As such, the original estimator "borrows" the closed-form variance expression given for the Taylor approximation, again justified given the asymptotic equivalence.

Lemma 2.7 (First-order Taylor approximation for linear estimators)**.** *First, assume a linear estimator as defined in Definition 2.1. Then, let* $\left\{\ .\ |_{\mathbf{R}=\boldsymbol{\pi}}\right\}$ *represent a function that evaluates the argument to the left of the vertical line at* $\mathbf{R}=\boldsymbol{\pi}$*. Similarly, let* $\left\{\ .\ |_{\mathbf{R}=\boldsymbol{\pi}}(\mathbf{R}-\boldsymbol{\pi})\right\}$ *evaluate its argument at* $\mathbf{R}=\boldsymbol{\pi}$ *and then multiply by* $(\mathbf{R}-\boldsymbol{\pi})$*. Then from Taylor's theorem and the product rule, we have the first-order Taylor approximation,* $\widehat{\delta}\approx\widehat{\delta}^{\text{T}}$*, with*

$$\begin{aligned}\widehat{\delta}_c^{\text{T}} :=& \left\{ c'\mathbf{W}\mathbf{R}y\ |_{\mathbf{R}=\boldsymbol{\pi}} \right\} + \left\{ c'\mathbf{W}|_{\mathbf{R}=\boldsymbol{\pi}} \right\}\left\{ \frac{d}{d\mathbf{R}}\mathbf{R}\bigg|_{\mathbf{R}=\boldsymbol{\pi}} (\mathbf{R}-\boldsymbol{\pi}) \right\}y \\ &+ \left\{ \frac{d}{d\mathbf{R}}c'\mathbf{W}\bigg|_{\mathbf{R}=\boldsymbol{\pi}} (\mathbf{R}-\boldsymbol{\pi}) \right\}\left\{ \mathbf{R}|_{\mathbf{R}=\boldsymbol{\pi}} \right\}y \\ =&\ a_c + c'\mathbf{w}\mathbf{R}y + \left\{ \frac{d}{d\mathbf{R}}c'\mathbf{W}\bigg|_{\mathbf{R}=\boldsymbol{\pi}} \mathbf{R} \right\}\boldsymbol{\pi}y \end{aligned} \tag{3}$$

where

$$a_c = -\left\{ \frac{d}{d\mathbf{R}}c'\mathbf{W}\bigg|_{\mathbf{R}=\boldsymbol{\pi}} \boldsymbol{\pi} \right\}\boldsymbol{\pi}y$$

is a constant.

Remark 3. *An expression for* a_c *is given but it is not important for the purposes of variance approximations because the term is a constant. Recall that the purpose of deriving a first-order Taylor approximation,* $\widehat{\delta}_c^{\text{T}}$*, is to identify a closed-form variance expression that might then be "borrowed" by the original linear estimator given in Definition 2.1.*

Theorem 2.8. *For a constant,* a_c*, and vector of constants,* z_c*, first-order Taylor approximations for linear estimators may be written as,*

$$\widehat{\delta}_c^{\text{T}} = a_c + n1_k'\mathbf{w}^{\text{HT}}\mathbf{R}z_c,$$

where z_c *has the form* $z_c = \boldsymbol{\pi}\text{diag}\,(\mathbf{l}y)\,\mathbf{t}'c$ *and where* $(k\times kn)$ *matrix* $\mathbf{t}$ *and* $(kn\times kn)$ *matrix* $\mathbf{l}$ *depend on the estimator. Hence, a first-order approximation of a Tyalor series using Taylor's theorem variance approximations will be expressed as the variance of a Horvitz-Thompson estimator of the ATE of* z_c *with contrast vector* $n1_k$*.*

Proof. With Equation (3), it is easy to see that the Taylor linearized approximation has the form

$$\widehat{\delta}_c^{\text{T}} = a_c + c'\mathbf{t}\mathbf{R}\mathbf{l}y$$

where matrices $\mathbf{t}$ and $\mathbf{l}$ are $(k \times kn)$ and $(kn \times kn)$, respectively, and will depend on the estimator. Noting that $c'\mathbf{t}$ is a $(1 \times kn)$ vector, write

$$\begin{aligned}\widehat{\delta}_c^{\mathrm{T}} =& a_c + 1_{kn}' \mathrm{diag}\left(c'\mathbf{t}\right) \mathbf{R}\mathbf{l}y \\ =& a_c + 1_{kn}' \mathbf{R} \mathrm{diag}\left(c'\mathbf{t}\right) \mathbf{l}y \\ =& a_c + 1_{kn}' \boldsymbol{\pi}^{-1} \mathbf{R} \boldsymbol{\pi} \mathrm{diag}\left(c'\mathbf{t}\right) \mathbf{l}y \\ =& a_c + n 1_k' (\mathbb{1}'\mathbb{1})^{-1} \mathbb{1}' \boldsymbol{\pi}^{-1} \mathbf{R} \boldsymbol{\pi} \mathrm{diag}\left(c'\mathbf{t}\right) \mathbf{l}y \\ =& a_c + n 1_k' \mathbf{w}^{\mathrm{HT}} \mathbf{R} z_c\end{aligned}$$

where $z_c := \boldsymbol{\pi} \mathrm{diag}\left(\mathbf{l}y\right) \mathbf{t}'c$. □

Remark 4. *The result shows that first order Taylor approximations are Horvitz-Thompson estimators. This highlights the importance of studying Horvitz-Thompson variance in order to develop asymptotic variance expressions for linear estimators in general.*

Remark 5. *The constant vector z_c is not directly observed. The next section will show that the plug-in principle provides a basis for asymptotically valid variance expressions.*

Table 1: Examples of linear estimators. $\mathbf{W}$ is as defined in Definition 2.1, z_c is as defined in Theorem 2.8.

Estimator	$\mathbf{W}$	z_c
Horvitz-Thompson	$\left(\mathbb{1}'\mathbb{1}\right)^{-1} \mathbb{1}'\boldsymbol{\pi}^{-1}$	$\mathrm{diag}(y)\mathbb{1}\left(\mathbb{1}'\mathbb{1}\right)^{-1}c$
Contrast-of-means	$\left(\mathbb{1}'\mathbf{R}\mathbb{1}\right)^{-1} \mathbb{1}'$	$\boldsymbol{\pi}\mathrm{diag}(y - \mathbb{1}\left(\mathbb{1}'\boldsymbol{\pi}\mathbb{1}\right)^{-1} \mathbb{1}'\boldsymbol{\pi}y)\mathbb{1}\left(\mathbb{1}'\boldsymbol{\pi}\mathbb{1}\right)^{-1}c$
Hajek	$\left(\mathbb{1}'\boldsymbol{\pi}^{-1}\mathbf{R}\mathbb{1}\right)^{-1} \mathbb{1}'\boldsymbol{\pi}^{-1}$	$\mathrm{diag}(y - \mathbb{1}\left(\mathbb{1}'\mathbb{1}\right)^{-1} \mathbb{1}'y)\mathbb{1}\left(\mathbb{1}'\mathbb{1}\right)^{-1}c$
OLS	$\left(\mathbb{x}'\mathbf{R}\mathbb{x}\right)^{-1} \mathbb{x}'$	$\boldsymbol{\pi}\mathrm{diag}\left(y - \mathbb{x}b^{\mathrm{OLS}}\right)\mathbb{x}\left(\mathbb{x}'\boldsymbol{\pi}\mathbb{x}\right)^{-1}c$
WLS	$\left(\mathbb{x}'\mathbf{m}\mathbf{R}\mathbb{x}\right)^{-1} \mathbb{x}'\mathbf{m}$	$\boldsymbol{\pi}\mathrm{diag}\left(y - \mathbb{x}b^{\mathrm{WLS}}\right)\mathbf{m}\mathbb{x}\left(\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x}\right)^{-1}c$
Generalized reg. $(b = b^{\mathrm{WLS}})$	$\mathbf{w}^{\mathrm{HT}}\left(\mathbf{i}_{kn} - \left(\mathbf{R} - \boldsymbol{\pi}\right)\mathbb{x}\mathbf{W}^{\mathrm{WLS}}\right)$	$\mathrm{diag}\left(y - \mathbb{x}b^{\mathrm{WLS}}\right)\mathbb{1}\left(\mathbb{1}'\mathbb{1}\right)^{-1}c$
IV	$\left(\widehat{\tilde{\mathbb{x}}}'\mathbf{R}\widehat{\tilde{\mathbb{x}}}\right)^{-1}\widehat{\tilde{\mathbb{x}}}'$ with: $\widehat{\tilde{\mathbb{x}}} := \mathbb{z}\left(\mathbb{z}'\mathbf{R}\mathbb{z}\right)^{-1}\mathbb{z}'\mathbf{R}\mathbb{x}$	$\boldsymbol{\pi}\mathrm{diag}\left(y - \mathbb{x}b^{\mathrm{IV}}\right)\tilde{\mathbb{x}}\left(\tilde{\mathbb{x}}'\boldsymbol{\pi}\tilde{\mathbb{x}}\right)^{-1}c$ with: $\tilde{\mathbb{x}} := \mathbb{z}\left(\mathbb{z}'\boldsymbol{\pi}\mathbb{z}\right)^{-1}\mathbb{z}'\boldsymbol{\pi}\mathbb{x}$, $b^{\mathrm{IV}} := \left(\tilde{\mathbb{x}}'\boldsymbol{\pi}\tilde{\mathbb{x}}\right)^{-1}\tilde{\mathbb{x}}'\boldsymbol{\pi}y$

Example (*Weighted least squares*) : Weighted least squares is a class that includes OLS $(\mathbf{m} = \mathbf{i}_{kn})$, contrast-of-means (e.g., difference of means, with $\mathbf{m} = \mathbf{i}_{kn}$ and $\mathbb{x} = \mathbb{1}$) and the Hajek estimator ($\mathbf{m} = \boldsymbol{\pi}^{-1}$ and $\mathbb{x} = \mathbb{1}$). To derive its Taylor approximation, first let $\mathbf{w}^{\mathrm{WLS}} = \mathbf{W}^{\mathrm{WLS}}|_{\mathbf{R}=\boldsymbol{\pi}} = \left(\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x}\right)^{-1}\mathbb{x}'\mathbf{m}$, and note that by the rules of matrix differentiation the third term in Equation (3) is

$$\begin{aligned}\left\{ \frac{\mathrm{d}}{\mathrm{d}\mathbf{R}} c' \left(\mathbb{x}'\mathbf{m}\mathbf{R}\mathbb{x}\right)^{-1} \mathbb{x}'\mathbf{m}\Big|_{\mathbf{R}=\boldsymbol{\pi}} \mathbf{R}\right\}\boldsymbol{\pi} y & \\ &= -c' \left(\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x}\right)^{-1} \left\{ \frac{\mathrm{d}}{\mathrm{d}\mathbf{R}} \left(\mathbb{x}'\mathbf{m}\mathbf{R}\mathbb{x}\right)\Big|_{\mathbf{R}=\boldsymbol{\pi}} \mathbf{R}\right\} \left(\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x}\right)^{-1} \mathbb{x}'\mathbf{m}\boldsymbol{\pi} y \\ &= -c' \left(\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x}\right)^{-1} \mathbb{x}'\mathbf{m}\left\{ \frac{\mathrm{d}}{\mathrm{d}\mathbf{R}}\mathbf{R}\Big|_{\mathbf{R}=\boldsymbol{\pi}} \mathbf{R}\right\}\mathbb{x} b^{\text{WLS}} \\ &= -c'\mathbf{w}^{\text{WLS}}\mathbf{R}\mathbb{x} b^{\text{WLS}}.\end{aligned}$$

Therefore, Equation (3) made specific to WLS is

$$\begin{aligned}\widehat{\delta}^{\text{T}(\text{WLS})} =& a_c^{\text{WLS}} + c'\mathbf{w}^{\text{WLS}}\mathbf{R}y - \mathbf{w}^{\text{WLS}}\mathbf{R}\mathbb{x} b^{\text{WLS}} \\ =& a_c^{\text{WLS}} + c'\mathbf{w}^{\text{WLS}}\mathbf{R}\left(y - \mathbb{x} b^{\text{WLS}}\right) \\ =& a_c^{\text{WLS}} + 1_{kn}'\text{diag}\left(c'\mathbf{w}^{\text{WLS}}\right)\mathbf{R}\text{diag}\left(y - \mathbb{x} b^{\text{WLS}}\right)1_{kn} \\ =& a_c^{\text{WLS}} + 1_{kn}'\mathbf{R}\text{diag}\left(y - \mathbb{x} b^{\text{WLS}}\right)\mathbf{w}^{\text{WLS}\prime}c \\ =& a_c^{\text{WLS}} + n1_k'\mathbf{w}^{\text{HT}}\mathbf{R}z_c^{\text{WLS}}\end{aligned}$$

where $z_c^{\text{WLS}} = \boldsymbol{\pi}\text{diag}\left(y - \mathbb{x} b^{\text{WLS}}\right)\mathbf{w}^{\text{WLS}\prime}c$ is recognizable in the form given in Theorem 2.8 with $\mathbf{t}^{\text{WLS}} = \mathbf{w}^{\text{WLS}}$ and $\mathbf{l}^{\text{WLS}} = \mathbf{i}_{kn} - \mathbb{x}\left(\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x}\right)^{-1}\mathbb{x}'\mathbf{m}\boldsymbol{\pi}$ is a "residual maker" matrix. △

3 Variance

Now that the importance of Horvitz-Thompson estimators for asymptotic variance expressions for the entire class of linear estimators (which includes, for example, OLS, WLS, Hajek, and difference-of-means) has been established, this section will give the variance of HT estimators and first-order approximates of linear estimators.

Throughout, we will make use of the $kn \times kn$ "first order design matrix", which will allow for easy comparison of designs using spectral analysis.

Definition 3.1. *The "first-order design matrix" is a variance-covariance matrix of inverse-probability weighted treatment assignments, written,*

$$\begin{aligned}\mathbf{d} :=& \mathrm{V}\left(1_{kn}'\boldsymbol{\pi}^{-1}\mathbf{R}\right) \\ =& \left(\mathrm{E}\left[\mathbf{R}1_{kn}1_{kn}'\mathbf{R}\right] - \boldsymbol{\pi}1_{kn}1_{kn}'\boldsymbol{\pi}\right) / \left(\boldsymbol{\pi}1_{kn}1_{kn}'\boldsymbol{\pi}\right),\end{aligned} \tag{4}$$

where "/" represents elementwise division.

Theorem 3.2 (Horvitz-Thompson Variance)**.** *An exact expression for the variance of Horvitz-Thompson estimators is given by*

$$\mathrm{V}\left(\widehat{\delta}_c^{\text{HT}}\right) = z_c^{\text{HT}\prime}\mathbf{d}z_c^{\text{HT}},$$

where $z_c^{\text{HT}\prime} = c'\left(\mathbb{1}'\mathbb{1}\right)^{-1}\mathbb{1}'\text{diag}\left(y\right)$.

Proof. Using the identity $y = \text{diag}\left(y\right)1_{kn}$, the Horvitz-Thompson estimator can be written

$$\begin{aligned}\widehat{\delta}^{\text{HT}} =& c'\left(\mathbb{1}'\mathbb{1}\right)^{-1}\mathbb{1}'\boldsymbol{\pi}^{-1}\mathbf{R}\text{diag}\left(y\right)1_{kn} \\ =& c'\left(\mathbb{1}'\mathbb{1}\right)^{-1}\mathbb{1}'\text{diag}\left(y\right)\boldsymbol{\pi}^{-1}\mathbf{R}1_{kn}. \\ =& z_c^{\text{HT}\prime}\boldsymbol{\pi}^{-1}\mathbf{R}1_{kn}.\end{aligned}$$

where $z_c^{\text{HT}\prime} = c' \left(\mathbb{1}'\mathbb{1}\right)^{-1} \mathbb{1}'\text{diag}\left(y\right)$. So the variance can be written,

$$\begin{aligned} \mathrm{V}\left(\widehat{\delta}_c^{\text{T}}\right) &= z_c^{\text{HT}\prime}\mathrm{V}\left(1'_{kn}\boldsymbol{\pi}^{-1}\mathbf{R}\right) z_c^{\text{HT}} \\ &= z_c^{\text{HT}\prime}\mathbf{d}z_c^{\text{HT}} \end{aligned}$$

□

Theorem 3.3 (Variance of first-order Taylor approximations)**.** *The variance of first-order Taylor approximations of linear estimators can be written as,*

$$\mathrm{V}\left(\widehat{\delta}_c^{\text{T}}\right) = z_c^{\text{T}\prime}\mathbf{d}z_c^{\text{T}}, \tag{5}$$

with examples of z_c^{T} given in Table 1.

Proof. By Theorem 2.8, linear approximations are Horvitz-Thompson estimators of a vector z_c and contrast vector $n1_k$. Now, $n1'_k \left(\mathbb{1}'\mathbb{1}\right)^{-1} \mathbb{1}'\text{diag}\left(z_c^{\text{T}}\right) = 1'_{kn}\text{diag}\left(z_c^{\text{T}}\right) = z_c^{\text{T}}$. Therefore, using Theorem 3.2,

$$\begin{aligned} \mathrm{V}\left(\widehat{\delta}_c^{\text{T}}\right) &= n1'_k \left(\mathbb{1}'\mathbb{1}\right)^{-1} \mathbb{1}'\text{diag}\left(z_c^{\text{T}}\right)\mathbf{d}\text{diag}\left(z_c^{\text{T}}\right)\mathbb{1}\left(\mathbb{1}'\mathbb{1}\right)^{-1} 1_k n \\ &= z_c^{\text{T}\prime}\mathbf{d}z_c^{\text{T}} \end{aligned}$$

□

Remark 6. *Equation (5) is an exact variance expression, however, they are not identified. The next subsection introduces the necessary concept of variance bounding.*

Remark 7. *The design matrix,* **d**, *will provide useful device in the study the best designs for outcomes with different characteristics as the next example will show.*

Example (*Comparing complete randomization and paired randomization*) : Consider a treatment/control (two-arm) experiment that is pair-randomized. A pair-randomized design is a special case of a block-randomized (i.e., stratified) design where blocks have size 2. In each pair/block, one unit is assigned to treatment and the other in control with equal (.5) probability. Across blocks, assignments are independent.

When $n = 4$ (and assuming w.l.o.g. that the data are sorted by pair), the design matrix is

$$\mathbf{d}^{pr} = \begin{bmatrix} 1 & -1 & & & -1 & 1 & & \\ -1 & 1 & & & 1 & -1 & & \\ & & 1 & -1 & & & -1 & 1 \\ & & -1 & 1 & & & 1 & -1 \\ -1 & 1 & & & 1 & -1 & & \\ 1 & -1 & & & -1 & 1 & & \\ & & -1 & 1 & & & 1 & -1 \\ & & 1 & -1 & & & -1 & 1 \end{bmatrix},$$

and note that empty cells are 0. The design matrix for complete randomization (where 2 of 4 are randomly assigned to treatment) is

$$\mathbf{d}^{cr} = \begin{bmatrix} 1 & -{}^1/_3 & -{}^1/_3 & -{}^1/_3 & -1 & {}^1/_3 & {}^1/_3 & {}^1/_3 \\ -{}^1/_3 & 1 & -{}^1/_3 & -{}^1/_3 & {}^1/_3 & -1 & {}^1/_3 & {}^1/_3 \\ -{}^1/_3 & -{}^1/_3 & 1 & -{}^1/_3 & {}^1/_3 & {}^1/_3 & -1 & {}^1/_3 \\ -{}^1/_3 & -{}^1/_3 & -{}^1/_3 & 1 & {}^1/_3 & {}^1/_3 & {}^1/_3 & -1 \\ -1 & {}^1/_3 & {}^1/_3 & {}^1/_3 & 1 & -{}^1/_3 & -{}^1/_3 & -{}^1/_3 \\ {}^1/_3 & -1 & {}^1/_3 & {}^1/_3 & -{}^1/_3 & 1 & -{}^1/_3 & -{}^1/_3 \\ {}^1/_3 & {}^1/_3 & -1 & {}^1/_3 & -{}^1/_3 & -{}^1/_3 & 1 & -{}^1/_3 \\ {}^1/_3 & {}^1/_3 & {}^1/_3 & -1 & -{}^1/_3 & -{}^1/_3 & -{}^1/_3 & 1 \end{bmatrix}.$$

Eigendecomposition of $\mathbf{d}^{cr} - \mathbf{d}^{pr}$ gives eigenvalues $2.67, 0, 0, 0, 0, 0, -1.33$, and -1.33 corresponding eigenvectors in Table 2. The eigenvectors associated with nonzero eigenvalues provide insight into the subspace in $\mathbb{R}^{2n}$ where one design may be preferable to another, for example, when the estimator is difference-in-means (which is equivalent to both Horvitz-Thopson and Hajek for these designs).

Table 2: Eigenvectors of $\mathbf{d}^{cr} - \mathbf{d}^{pr}$

e1	e2	e3	e4	e5	e6	e7	e8
-0.354	0.791	0.000	0.000	0.000	0.000	0.500	0.000
-0.354	0.158	-0.573	-0.178	-0.250	-0.421	-0.500	0.000
0.354	0.158	-0.180	-0.450	0.585	-0.149	0.000	0.500
0.354	0.158	0.319	-0.600	-0.260	-0.264	0.000	-0.500
0.354	0.474	0.282	0.524	0.100	-0.190	-0.500	0.000
0.354	-0.158	-0.291	0.346	-0.150	-0.611	0.500	0.000
-0.354	-0.158	0.102	0.073	0.685	-0.339	-0.000	-0.500
-0.354	-0.158	0.602	-0.077	-0.161	-0.454	-0.000	0.500

In this example, assuming the contrast matrix is $c = (-1, 1)'$, and examining the first eigenvector with eigenvalue 2.67, one can conclude that if the outcomes for the four units given in Table 3, then the difference-of-means would be much less precise under the completely randomized design. So, the eigenvector in a sense represents a "best-case" (normed) potential outcome vector for paired randomization. Inspection of the outcomes themselves confirms the intuition that pair randomization is better than complete randomization when units are homogenous within pairs.

Table 3: Pair randomization better than complete randomization

unit id	pair id	y_0	y_1
1	1	.3536	.3536
2	1	.3536	.3536
3	2	-.3536	-.3536
4	2	-.3536	-.3536

Next, considering the two eigenvectors associated with the eigenvalue -1.33, we see the implied potential outcomes in Table 4 give potential outcomes for which complete randomization is preferable. Note that either of the two sets is a "worst-case" scenario for paired randomization, as is any set of potential outcomes that can be generated by linear combinations of the two eigenvectors. Inspection of these outcomes is consistent with the observation that complete randomization can be better than paired randomization when paired units are maximally heterogeneous.

Table 4: Complete randomization better than pair randomization

unit id	pair id	y_0	y_1	y_0	y_1
1	1	-.5	-.5	0	0
2	1	.5	.5	0	0
3	2	0	0	-.5	-.5
4	2	0	0	.5	.5

△

Remark 8. *The example illustrates a relatively effortless method of identifying key insights about arbitrary designs through spectral analyses of first-order design matrices. In the example, the observation that pair randomization can hurt precision when units are not homogeneous within pairs is not new. However, this approach to comparing designs is perfectly general and can be applied to virtually any designs.*

4 Variance bounds

In spite of an exact expression for first-order Taylor approximations in Equation (8), the quantity is never identified because not all terms in the quadratic can be observed. Even if the elements of $\mathbf{R}z_c$ were observed directly (which is the case for $\mathbf{R}z_c^{\mathrm{HT}}$ but none of the other examples in Table 1), some pairs of potential outcomes can never be jointly observed. For example, for a given unit, only one of two (or more) potential outcomes can be observed, a problem is referred to as the "fundamental problem of causal inference" (Holland, 1986). Other design features, such as clustering or pair randomization, can also render various combinations of potential outcomes unobservable.

Starting with Neyman (1923) one proposed solution to unidentified variance has been to estimate a *variance bound*, i.e., a quantity that is provably greater than the variance, but which is identified. It should be understood that while the term *variance estimation* is often used as a shorthand in the literature, it is not, in general, an accurate phrase. *Variance bound estimation* is a more precise so it will be used here.

Definition 4.1 (Variance bound matrix). *Let $\tilde{\mathbf{d}}$ be an arbitrary $kn \times kn$ matrix and let be z an arbitrary vector with length kn. Then $\tilde{\mathbf{d}}$ is a* variance bound matrix *(or* bounding matrix*) for $\mathbf{d}$ if, for all $z \in \mathbb{R}^{kn}$, $z'\mathbf{d}z \leq z'\tilde{\mathbf{d}}z$.*

Lemma 4.2. *$\tilde{\mathbf{d}}$ is a bounding matrix $\mathbf{d}$ if and only if matrix $\tilde{\mathbf{d}} - \mathbf{d}$ is positive semi-definite.*

Proof. By the definition of a bound, $z'\tilde{\mathbf{d}}z - z'\mathbf{d}z \geq 0$ for all $z \in \mathbb{R}^{kn}$. This implies that $z'(\tilde{\mathbf{d}} - \mathbf{d})z \geq 0$, i.e., that $\tilde{\mathbf{d}} - \mathbf{d}$ is positive semi-definite. □

Definition 4.3 (Identified variance bound). *Let $\tilde{\mathbf{d}}$ be bounding matrix for $\mathbf{d}$. It gives an* identified variance bound *if*

$$\mathrm{I}(\mathbf{d} = -1) \circ \mathrm{I}(\tilde{\mathbf{d}} = 0) = \mathrm{I}(\mathbf{d} = -1)$$

where $\circ$ is element-wise multiplication, $\mathrm{I}\,(\mathbf{d} = -1)$ is an indicator function returning an $kn \times kn$ matrix of ones and zeros indicating whether each element of $\mathbf{d}$ is equal to -1 (an indication that the associated term in the variance quadratic is impossible to observe), and $\mathrm{I}(\tilde{\mathbf{d}} = 0)$ is, similarly, an indicator function returning an $kn \times kn$ matrix of ones and zeros indicating the location of zeros in $\tilde{\mathbf{d}}$.

4.1 Generalizing Neyman's variance bound

This section proposes a generalization of Neyman's (1923) variance bound. Let matrix $\mathbf{d}$ be partitioned into k^2 partitions of size $n \times n$. Then for $r, s \in \{1, 2, ..., k\}$, let the $\mathbf{d}_{rs}$ be the $(r, s)^{th}$ partition, having dimension $n \times n$. Also, let c_r be the r^{th} element of the length-k contrast vector, c. Then the following bounding method produces an identified bound for experiments when partitions $\mathrm{I}(\tilde{\mathbf{d}}^{\mathrm{N}}_{rr} == -1) = 0_{n\times n}$, i.e., there are no -1 values in the diagonal blocks, and $\tilde{\mathbf{d}}^{\mathrm{N}}_{rs} = \tilde{\mathbf{d}}^{\mathrm{N}}_{tu}$ for $r \neq s, t \neq u \in 1, 2, ..., k$ and $\sum_i c_i = 0$. Designs that meet this condition include complete randomization, cluster-randomization and block-randomization.

Definition 4.4 (Generalized Neyman variance bound)**.** *The "Generalized Neyman bound" is the is the bound corresponding to the block-diagonal bounding matrix, $\tilde{\mathbf{d}}^{\mathrm{N}}$, with block (r, r) given by,*

$$\tilde{\mathbf{d}}^{\mathrm{N}}_{rr} := \sum_{s=1}^{k} \frac{c_r}{c_s} \mathbf{d}_{rs}$$

where c_r and c_s are, respectively, elements r and s from from the contrast vector, c.

Theorem 4.5. *The generalized Neyman bound, with $\tilde{\mathbf{d}}^{\mathrm{N}}$ given in Definition 4.1 is an identified variance bound when partitions $\mathrm{I}(\tilde{\mathbf{d}}^{\mathrm{N}}_{rr} == -1) = 0_{n\times n}$, i.e., there are no -1 values in the diagonal blocks, $\tilde{\mathbf{d}}^{\mathrm{N}}_{rs} = \tilde{\mathbf{d}}^{\mathrm{N}}_{tu}$ for $r \neq s, t \neq u \in 1, 2, ..., k$, and $\sum_i c_i = 0$.*

Proof. First, with $z_c^{\mathrm{HT}} = \mathrm{diag}(y)\mathbb{1}(\mathbb{1}'\mathbb{1})^{-1}$ and letting y_r be the length-n vector of potential outcomes for the r^{th} treatment arm and $\tilde{\mathbf{d}}^{\mathrm{N}}_{rs}$ be the r, s partition of $\mathbf{d}$, we have

$$\begin{aligned}
n^2 z^{\mathrm{HT}\prime}\mathbf{d}z^{\mathrm{HT}} =& y'\mathrm{diag}(c'\mathbb{1})\mathbf{d}\mathrm{diag}(c'\mathbb{1})y \\
=& \sum_{r=1}^{k} c_r^2 y_r' \mathbf{d}_{rr} y_r + \sum_{r=1}^{k-1}\sum_{s=1}^{k} c_r c_s \left(y_r' \mathbf{d}_{rs} y_s + y_s' \mathbf{d}_{sr} y_r\right)
\end{aligned}$$

Next, define the r, s treatment effect as $\tau_{rs} := y_r - y_s$ and note that $\mathbf{d}_{12} = \mathbf{d}_{rs}$ for $r \neq s$. Then, by the definition of $\tilde{\mathbf{d}}^{\mathrm{N}}$,

$$\begin{aligned}
n^2 z^{\mathrm{HT}\prime}\tilde{\mathbf{d}}^{\mathrm{N}} z^{\mathrm{HT}} =& y'\mathrm{diag}(c'\mathbb{1})\tilde{\mathbf{d}}^{\mathrm{N}}\mathrm{diag}(c'\mathbb{1})y \\
=& \sum_{r=1}^{k} c_r^2 y_r' \mathbf{d}_{rr} y_r + \sum_{r=1}^{k-1}\sum_{s=r+1}^{k} c_r c_s \left(y_r' \mathbf{d}_{rs} y_r + y_s' \mathbf{d}_{sr} y_s\right) \\
=& \sum_{r=1}^{k} c_r^2 y_r' \mathbf{d}_{rr} y_r + \sum_{r=1}^{k-1}\sum_{s=r+1}^{k} c_r c_s \left(y_r' \mathbf{d}_{12} y_r + y_s' \mathbf{d}_{12} y_s\right) \\
=& \sum_{r=1}^{k} c_r^2 y_r' \mathbf{d}_{rr} y_r + \sum_{r=1}^{k-1}\sum_{s=r+1}^{k} c_r c_s \left(y_r' \mathbf{d}_{12} (y_s + \tau_{rs}) + y_s' \mathbf{d}_{12} (y_r - \tau_{rs})\right) \\
=& n^2 z^{\mathrm{HT}\prime}\mathbf{d}z^{\mathrm{HT}} + \sum_{r=1}^{k-1}\sum_{s=r+1}^{k} c_r c_s \left(y_r' \mathbf{d}_{12} \tau_{rs} - y_s' \mathbf{d}_{12} \tau_{rs}\right) \\
=& n^2 z^{\mathrm{HT}\prime}\mathbf{d}z^{\mathrm{HT}} + \sum_{r=1}^{k-1}\sum_{s=r+1}^{k} c_r c_s \left(y_r' \mathbf{d}_{12} \tau_{rs} - (y_r - \tau_{rs})' \mathbf{d}_{12} \tau_{rs}\right) \\
=& n^2 z^{\mathrm{HT}\prime}\mathbf{d}z^{\mathrm{HT}} + \sum_{r=1}^{k-1}\sum_{s=r+1}^{k} c_r c_s \tau_{rs}' \mathbf{d}_{12} \tau_{rs}.
\end{aligned}$$

Next, to show that the second term is non-negative, note that $\tau_{rs} = \tau_{rk} - \tau_{sk}$, and write

$$\begin{aligned}
\sum_{r=1}^{k-1} \sum_{s=r+1}^{k} c_r c_s \tau_{rs}' \mathbf{d}_{12} \tau_{rs} =& \frac{1}{2} \sum_{r=1}^{k} \sum_{s=1}^{k} c_r c_s \tau_{rs}' \mathbf{d}_{12} \tau_{rs} \\
=& \frac{1}{2} \sum_{r=1}^{k} \sum_{s=1}^{k} c_r c_s \left(\tau_{rk} - \tau_{sk}\right)' \mathbf{d}_{12} \left(\tau_{rk} - \tau_{sk}\right) \\
=& \frac{1}{2} \sum_{r=1}^{k} \sum_{s=1}^{k} c_r c_s \left(\tau_{rk}' \mathbf{d}_{12} \tau_{rk} + \tau_{sk}' \mathbf{d}_{12} \tau_{sk} - 2\tau_{sk}' \mathbf{d}_{12} \tau_{rk}\right) \\
=& \sum_{r=1}^{k} \sum_{s=1}^{k} c_r c_s \tau_{rk}' \mathbf{d}_{12} \tau_{rk} - \sum_{r=1}^{k} \sum_{s=1}^{k} c_r c_s \tau_{sk}' \mathbf{d}_{12} \tau_{rk} \\
=& \sum_{r=1}^{k} c_r \tau_{rk}' \mathbf{d}_{12} \tau_{rk} \left(\sum_{s=1}^{k} c_s\right) - \left(\sum_{s=1}^{k} c_s \tau_{sk}\right)' \mathbf{d}_{12} \left(\sum_{r=1}^{k} c_r \tau_{rk}\right) \\
=& 0 - \tau^{*\prime} \mathbf{d}_{12} \tau^{*} \\
\geq& 0
\end{aligned}$$

where the second to last line uses $\sum_{s=1}^{k} c_s = 0$ and the definition $\tau^* := \sum_{s=1}^{k} c_s \tau_{sk}$. The last line follows because $\mathbf{d}_{12}$ is negative semidefinite. □

4.2 A novel proof of the Aronow-Samii bound

Consider an identified bound proposed by Aronow and Samii (2017) that has the a unusual virtue of being perfectly general, i.e., applicable to arbitrary (identified) designs.

Definition 4.6 (Aronow-Samii variance bound)**.** *The "Aronow-Samii variance bound" is the bound corresponding to the bounding matrix,*

$$\tilde{\mathbf{d}}^{\text{AS}} := \mathbf{d} + \mathrm{I}\left(\mathbf{d} = -1\right) + \text{diag}(\mathrm{I}\left(\mathbf{d} = -1\right) 1_{kn})$$

where the indicator function, $\mathrm{I}(\mathbf{d} = -1)$*, returns a matrix of with ones indicating the location of -1 entries in* $\mathbf{d}$ *and zeros elsewhere, and* diag(.) *creates a diagonal matrix from a vector.*

Theorem 4.7. *The Aronow-Samii variance bound,* $n^{-2} y' \tilde{\mathbf{d}}^{\text{AS}} y$*, is an identified bound for* $n^{-2} y' \mathbf{d} y$*.*

Proof. By definition of $\tilde{\mathbf{d}}^{\text{AS}}$,

$$\tilde{\mathbf{d}}^{\text{AS}} - \mathbf{d} = \mathrm{I}\left(\mathbf{d} = -1\right) + \text{diag}\left(\mathrm{I}\left(\mathbf{d} = -1\right) 1_{kn}\right).$$

Note that by construction $(\tilde{\mathbf{d}}^{\text{AS}} - \mathbf{d})$ has diagonal elements set equal to the sum of the off-diagonal elements in its row (which by construction are either 0 or 1). The Gershgorin circle theorem implies that a real matrix is positive semi-definite if, for all i, the i^{th} diagonal element is greater or equal to the sum of the absolute values of the other elements in the i^{th} row. So, by the Gershgorin circle theorem $\tilde{\mathbf{d}}^{\text{AS}} - \mathbf{d}$ is positive semidefinite. Therefore, by Lemma (4.2), $n^{-2} y' \tilde{\mathbf{d}}^{\text{AS}} y$ is a variance bound. Moreover, as long as the design is an identified design (i.e., $0 < \pi_{1i} < 1$ for all i), it is an identified bound because $\mathrm{I}\left(\mathbf{d} = -1\right)$ ensures that the elements of $\mathbf{d}$ equal to -1 correspond to 0's in $\tilde{\mathbf{d}}^{\text{AS}}$. □

Remark 9. *Aronow and Samii (2017) derive their bound using Young's inequality. The above-theorem and proof using the Gershgorin circle theorem tie their insight to the current framework.*

4.3 Proposed algorithm for variance bounds for any design

The following is an algorithm which that can obtain an identified variance bound. Like the AS bound it has the virtue of being applicable to virtually any design. The algorithm is a proof of concept, demonstrating the utility of the notation scheme which allows for the application of matrix theory for the creation of alternative bounds. The subject of comparing bounds will be considered further in Section 4.4.

Algorithm 4.8.

1. Initialize $kn \times kn$ matrix $\mathbf{t}$. Examples could be $\mathrm{I}(\mathbf{d} = -1)$ or, if the conditions for the Neyman bound not be applicable, start with $\tilde{\mathbf{d}}^{\mathrm{N}} - \mathbf{d}$ which may approximate a bound
2. Obtain the eigen decomposition of matrix $\mathbf{t}$. If all eigenvalues are non-negative (within tolerance), goto Step 6, otherwise continue
3. Update $\mathbf{t} = \mathbf{v}(\mathbf{e} \circ \mathrm{I}(\mathbf{e} > 0))\mathbf{v}'$ where $\mathbf{v}$ is the matrix of eigenvectors and $\mathbf{e}$ is a diagonal matrix of eigenvalues
4. Update $\mathbf{t} = \mathrm{I}(\mathbf{d} = -1) + \mathrm{I}(\mathbf{d} \neq -1) \circ \mathbf{t}$
5. Return to Step 2
6. Set $\tilde{\mathbf{d}}^{\mathrm{M}} = \mathbf{d} + \mathbf{t}$

As above, $\circ$ is elementwise multiplication and, for example, $\mathrm{I}(\mathbf{e} > 0)$ is an indicator function returning a matrix of ones and zeros indicating which elements of $\mathbf{e}$ are greater than zero.

Conceptually, the goal of the algorithm is to create a matrix $\mathbf{t}$ that can be added to $\mathbf{d}$ yielding a $\tilde{\mathbf{d}}$ matrix that corresponds to an identified variance bound. By Lemma 4.2 and Definition 4.3, there are two requirements for $\mathbf{t}$. First it must be positive semi-definite, and, second, elements corresponding to -1's in the matrix $\mathbf{d}$ must equal one. In step 1, $\mathbf{t}$ meets the second criterion, but not the first. In step 3, the algorithm creates an approximation to the initial $\mathbf{t}$ matrix by way of the eigen decomposition that ensures positive semi-definiteness, thus meeting the first criterion. However, due to the approximation, $\mathbf{t}$ no longer meets the second criterion. Therefore, in step 4 the algorithm forces $\mathbf{t}$ to have 1's wherever $\mathbf{d}$ has -1's in order to again meet the second criteria. But doing so means that $\mathbf{t}$ will no longer meet the first criteria. So, the algorithm iterates through steps 2-4 until convergence is achieved (i.e., until all eigenvalues are non-negative in step 2) at which point $\mathbf{t}$ meets both criteria and, thus, $\tilde{\mathbf{d}}^{\mathrm{M}}$ corresponds to an identified bound.

4.4 Comparing bounds

Definition 4.9 (Tighter bound)**.** *Let $\tilde{\mathbf{d}}^a$ and $\tilde{\mathbf{d}}^b$ correspond to two identified bounds. Matrix $\tilde{\mathbf{d}}^a$ is corresponds to a tighter bound than $\tilde{\mathbf{d}}^b$ if $\tilde{\mathbf{d}}^b - \tilde{\mathbf{d}}^a$ is positive semidefinite.*

Definition 4.10 (Invariant bounding matrix)**.** *A matrix $\tilde{\mathbf{d}}$ is an invariant bounding matrix if it is an bounding matrix and if all $n \times n$ partitions, $\tilde{\mathbf{d}}_{ij} 1_{kn} = 0_{kn}$, i.e., all rows of the partition (or, equivalently, all columns) sum to zero.*

Example (*Paired randomization*) : Consider a pair-randomized design, whereby units are "blocked" (i.e., stratified) into groups of two, and then, in each block, one of the two units is randomly assigned to treatment while the other is assigned to control. Assignments across blocks are independent.

When $n = 4$ (and assuming w.l.o.g. that the data are sorted by pair), the matrix $\mathbf{d}$ is

$$\mathbf{d} = \begin{bmatrix} 1 & -1 & & & -1 & 1 & & \\ -1 & 1 & & & 1 & -1 & & \\ & & 1 & -1 & & & -1 & 1 \\ & & -1 & 1 & & & 1 & -1 \\ -1 & 1 & & & 1 & -1 & & \\ 1 & -1 & & & -1 & 1 & & \\ & & -1 & 1 & & & 1 & -1 \\ & & 1 & -1 & & & -1 & 1 \end{bmatrix},$$

noting that empty cells represent 0.

For the pair-randomized design, the Neyman bound cannot be applied because $\mathbf{d}_{00}$ and $\mathbf{d}_{11}$ have negative entries. The Aronow-Samii bound and Algorithm 4.8 have bounding matrices

$$\tilde{\mathbf{d}}^{\text{AS}} = \begin{bmatrix} 3 & & & & & 1 & & \\ & 3 & & & 1 & & & \\ & & 3 & & & & & 1 \\ & & & 3 & & & 1 & \\ & 1 & & & 3 & & & \\ 1 & & & & & 3 & & \\ & & & 1 & & & 3 & \\ & & 1 & & & & & 3 \end{bmatrix}, \quad \text{and} \quad \tilde{\mathbf{d}}^{\text{M}} = \begin{bmatrix} 2 & & & & & 2 & & \\ & 2 & & & 2 & & & \\ & & 2 & & & & & 2 \\ & & & 2 & & & 2 & \\ & 2 & & & 2 & & & \\ 2 & & & & & 2 & & \\ & & & 2 & & & 2 & \\ & & 2 & & & & & 2 \end{bmatrix},$$

respectively. By the Gershgorian circle theorem the difference,

$$\tilde{\mathbf{d}}^{\text{AS}} - \tilde{\mathbf{d}}^{\text{M}} = \begin{bmatrix} 1 & & & & & -1 & & \\ & 1 & & & -1 & & & \\ & & 1 & & & & & -1 \\ & & & 1 & & & -1 & \\ & -1 & & & 1 & & & \\ -1 & & & & & 1 & & \\ & & & -1 & & & 1 & \\ & & -1 & & & & & 1 \end{bmatrix},$$

is positive semi-definite, proving that $\tilde{\mathbf{d}}^{\text{M}}$ corresponds to a tighter variance bound. Confirmation also comes from eigendecomposition of the difference, $\tilde{\mathbf{d}}^{\text{AS}} - \tilde{\mathbf{d}}^{\text{M}}$, which yields all non-negative eigenvalues: 2, 2, 2, 2, 0, 0, 0, and 0.

One might alternatively choose the invariant bounding matrix,

$$\tilde{\mathbf{d}}^{\text{INVAR}} = \begin{bmatrix} 2 & & -1 & -1 & & 2 & -1 & -1 \\ & 2 & -1 & -1 & 2 & & -1 & -1 \\ -1 & -1 & 2 & & -1 & -1 & & 2 \\ -1 & -1 & & 2 & -1 & -1 & 2 & \\ & 2 & -1 & -1 & 2 & & -1 & -1 \\ 2 & & -1 & -1 & & 2 & -1 & -1 \\ -1 & -1 & & 2 & -1 & -1 & 2 & \\ -1 & -1 & 2 & & -1 & -1 & & 2 \end{bmatrix}.$$

The bound can be verified because the eigenvalues of $\tilde{\mathbf{d}}^{\text{INVAR}} - \mathbf{d}$ are 8, 0, 0, 0, 0, 0, 0, and 0. However, eigendecomposition of $\tilde{\mathbf{d}}^{\text{INVAR}} - \tilde{\mathbf{d}}^{\text{M}}$ gives eigenvalues 4, 0, 0, 0, 0, 0, 0, and -4, indicating that the better bound may depend on outcome vector, y, and perhaps the estimator as well. △

5 Variance bound estimation

With an identified variance bounds defined and several methods of obtaining matrices, $\tilde{\mathbf{d}}$, this section turns to the subject of variance bound *estimation.*

First define the $kn \times kn$ matrix of probabilities and joint probabilities of assignment,

$$\mathbf{p} := \mathrm{E}\left[\mathbf{R}1_{kn}1_{kn}'\mathbf{R}\right].$$

Next define an inverse probability weighted version of bounding matrix, $\tilde{\mathbf{d}}$, as

$$\tilde{\mathbf{d}}_{/\mathbf{p}} := \tilde{\mathbf{d}}/\mathbf{p} \tag{6}$$

with / denoting element-wise division defined such that division by zero equals zero. Then an unbiased estimator of a variance bound for the Horvitz-Thompson estimator can be written,

$$\widehat{\tilde{\mathrm{V}}}\left(\widehat{\delta}^{\mathrm{HT}}\right) := z_c^{\mathrm{HT}\prime}\mathbf{R}\tilde{\mathbf{d}}_{/\mathbf{p}}\mathbf{R}z_c^{\mathrm{HT}}, \tag{7}$$

with $z_c^{\mathrm{HT}} := \mathrm{diag}(y)\mathbb{1}\left(\mathbb{1}'\mathbb{1}\right)^{-1}c$. It is unbiased for the variance bound $z_c^{\mathrm{HT}\prime}\tilde{\mathbf{d}}z_c^{\mathrm{HT}}$ because $\mathrm{E}\left[\mathbf{R}\tilde{\mathbf{d}}_{/\mathbf{p}}\mathbf{R}\right] = \tilde{\mathbf{d}}$ by construction. Being inverse-probability weighted, the variance bound estimator in (7) is, itself, a Horvitz-Thompson estimator.

For other linear estimators, examples of which are given in Table 1, the bound $z_c'\tilde{\mathbf{d}}z_c$ cannot be estimated unbiasedly because the definition of z_c will often include quantities that, themselves, must be estimated. However, an appeal to the plug-in principle suggests the use of

$$\widehat{\tilde{\mathrm{V}}}\left(\widehat{\delta}_c^{\mathrm{T}}\right) := \widehat{z}_c'\mathbf{R}\tilde{\mathbf{d}}_{/\mathbf{p}}\mathbf{R}\widehat{z}_c \tag{8}$$

with $\widehat{z}_c$ having the same form as z_c but with sample analogues replacing some components.

Example (*The special case of Eicker-Huber-White (a.k.a. "heteroskedastic consistent", "sandwich", and "robust") standard errors*) : For the OLS estimator, z_c^{OLS} is defined in Table (1). The plug-in principle motivates the use of

$$\mathbf{R}\widehat{z}_c^{\mathrm{OLS}} = \boldsymbol{\pi}\mathrm{diag}\left(\mathbf{R}\widehat{u}\right)\mathbb{x}\left(\mathbb{x}'\mathbf{R}\mathbb{x}\right)^{-1}c,$$

where $\mathbf{R}\widehat{u} := \mathbf{R}(y - \mathbb{x}\widehat{b}^{\mathrm{OLS}})$ and $\widehat{b}^{\mathrm{OLS}} := \left(\mathbb{x}'\mathbf{R}\mathbb{x}\right)^{-1}\mathbb{x}'\mathbf{R}y$ is the OLS coefficient. Then from equation (8) we have,

$$\widehat{\tilde{\mathrm{V}}}\left(\widehat{\delta}_c^{\mathrm{T(OLS)}}\right) = c'\left(\mathbb{x}'\mathbf{R}\mathbb{x}\right)^{-1}\mathbb{x}'\mathrm{diag}\left(\mathbf{R}\widehat{u}\right)\boldsymbol{\pi}\tilde{\mathbf{d}}_{/\mathbf{p}}\boldsymbol{\pi}\mathrm{diag}\left(\mathbf{R}\widehat{u}\right)\mathbb{x}\left(\mathbb{x}'\mathbf{R}\mathbb{x}\right)^{-1}c.$$

This is a the variance bound estimator in (8) made specific to OLS. So far it is applicable to virtually any design and any variance bound.

Next, specify a Bernoulli design, in which units are assigned independently to treatment. (Probabilities of assignment may be equal across units, but they need not be in this example.) In this design, the diagonal elements of $\mathbf{d}$ are equal to the diagonal of $\boldsymbol{\pi}^{-1} - \mathbf{i}_{kn}$, where $\mathbf{i}_{kn}$ is an identity matrix. Further, any of the above bounding methods yields $\tilde{\mathbf{d}} = \boldsymbol{\pi}^{-1} - \mathbf{i} + \mathbf{i} = \boldsymbol{\pi}^{-1}$. Thus $\tilde{\mathbf{d}}_{/\mathbf{p}} = \boldsymbol{\pi}^{-2}$ so that $\boldsymbol{\pi}\tilde{\mathbf{d}}_{/\mathbf{p}}\boldsymbol{\pi} = \mathbf{i}_{kn}$ is the identity matrix. So the OLS variance bound estimator for Bernoulli designs simplifies to,

$$\widehat{\tilde{\mathrm{V}}}^{\mathrm{B}}\left(\widehat{\delta}_c^{\mathrm{T(OLS)}}\right) = c'\left(\mathbb{x}'\mathbf{R}\mathbb{x}\right)^{-1}\mathbb{x}'\mathrm{diag}\left(\mathbf{R}\widehat{u}^2\right)\mathbb{x}\left(\mathbb{x}'\mathbf{R}\mathbb{x}\right)^{-1}c.$$

This is White's (1980) canonical "sandwich" variance estimator, sometimes referred to as HC0. △

Remark 10. *The example shows that Eicker-Huber-White standard errors are a special case of (8) for OLS in a Bernoulli design. Note, however, that (8) is much more general. It applies to any linear estimator, virtually any design and any (identified) variance bound.*

Remark 11. *Adjustments for degrees of freedom (e.g., HC1) or leverage (e.g., HC2, HC3, etc.) can be applied as well.*

Example (*The special case of "cluster robust" standard errors*) : Also consider this variance bound estimator for OLS in designs in which clusters are assigned independently to treatment. Then, if we choose the Neyman bound $\tilde{\mathbf{d}}_{/\mathbf{p}}^{\mathrm{N}}$ (or $\tilde{\mathbf{d}}_{/\mathbf{p}}^{\mathrm{M}}$, which is equivalent in the case of for Bernoulli assignment of clusters), and assuming w.l.o.g. that units are sorted by cluster, then $\boldsymbol{\pi}\tilde{\mathbf{d}}_{/\mathbf{p}}^{\mathrm{N}}\boldsymbol{\pi}$ resolves to a block diagonal matrix of 1's with the blocks corresponding to clusters. Hence, (8) also reproduces the "cluster-robust" standard errors sometimes referred to as CR0 as a special case. $\triangle$

6 Asymptotics

6.1 Conditions for Convergence of Horvitz-Thompson Estimators

First establishing the unbiasedness of Horvitz-Thompson estimators will allow for straightforward proofs of consistency.

Lemma 6.1. *The Horvitz-Thompson estimator for an outcome vector, y, and given contrast, c, is unbiased for δ_c.*

Proof.

$$\begin{aligned}\mathrm{E}\left[c'\left(\mathbb{1}'\mathbb{1}\right)^{-1}\mathbb{1}'\mathbf{R}y\right] =& c'\left(\mathbb{1}'\mathbb{1}\right)^{-1}\mathbb{1}'\boldsymbol{\pi}^{-1}\mathrm{E}\left[\mathbf{R}\right]y\\ =& c'\left(\mathbb{1}'\mathbb{1}\right)^{-1}\mathbb{1}'\boldsymbol{\pi}^{-1}\boldsymbol{\pi}y\\ =& c'\left(\mathbb{1}'\mathbb{1}\right)^{-1}\mathbb{1}'y\\ =& \delta_c,\end{aligned}$$

□

Condition 1 (Bounded contrast). *The chosen contrast vector is finite, i.e., there exists a finite value u_c such that* $\max(|c|) < u_c$.

Condition 2 (Bounded outcomes). *There exists a finite value, u_y, such that* $\max(|y|) < u_y$ *for all n.*

Condition 3 (Design constraint for consistent Horvitz-Thompson estimators). *There exists a finite value, $u_{\mathbf{d}}$, such that $n^{-1}||\mathbf{d}||_{1,1} < u_{\mathbf{d}}$ for all n, where $||.||_{1,1}$ is the matrix norm that sums the absolute values of the matrix entries.*

Theorem 6.2 (Root-n consistency of HT estimators). *By Lemma 6.1 and Conditions 1-3 the Horvitz-Thompson estimator is root-n consistent.*

Proof. Given Lemma 6.1, it is sufficient to show that the variance converges at the parametric rate, i.e., that $n\mathrm{V}(\delta_c^{\mathrm{HT}})$ is bounded, in order to prove consistency. By Holder's Inequality,

$$\begin{aligned}n\mathrm{V}(\delta_c^{\mathrm{HT}}) &\leq n\max(|c'\left(\mathbb{1}'\mathbb{1}\right)^{-1}\mathbb{1}'y|)^2||\mathbf{d}||_{1,1}\\ &\leq \max(|c|)^2\max(|y|)^2 n^{-1}||\mathbf{d}||_{1,1}\\ &\leq u_c^2 u_y^2 u_{\mathbf{d}},\end{aligned}$$

with the last line using Conditions 1-3. □

partition	ij pattern	count	$\{\mathbf{d}_{ab}\}_{ij}$	count$\times\frac{1}{n}\{\mathbf{d}_{ab}\}_{ij}$
$\mathbf{d}_{11}$	$i{=}j$	n	$\frac{n_t}{n_c}$	$\frac{n_t}{n_c}{=}O(1)$
	$i{\neq}j$	$n(n{-}1)$	$-\frac{n_t}{n_c(n-1)}$	$-\frac{n_t}{n_c}{=}O(1)$
$\mathbf{d}_{12}$ or $\mathbf{d}_{21}$	$i{=}j$	$2n$	-1	$-2{=}O(1)$
	$i{\neq}j$	$2n(n{-}1)$	$\frac{1}{(n-1)}$	$2{=}O(1)$
$\mathbf{d}_{22}$	$i{=}j$	n	$\frac{n_c}{n_t}$	$\frac{n_c}{n_t}{=}O(1)$
	$i{\neq}j$	$n(n{-}1)$	$-\frac{n_c}{n_t(n-1)}$	$-\frac{n_c}{n_t}{=}O(1)$

Table 5: Analysis of Condition 3 for Complete Randomization

Example (*Checking consistency of HT estimators for completely randomized experiments*) : Consider a completely randomized experiment with n units where a fixed number of units, n_c, are randomly assigned to control, and the remainder, $n_t = n - n_c$, are assigned to treatment. Assume an asymptotic sequence of designs is such that there exists a constant value, π, such that $\frac{n_t}{n} \to \pi$ as $n \to \infty$ with $0 < \pi_t < 1$. Partition $\mathbf{d}$ into four $(n \times n)$ matrices and let $\mathbf{d}_{ab}$ represent the $a, b \in \{1, 2\}$ partition. Each partition has elements which take on two possible values, one on the diagonal and another on the off-diagonal. Because $\mathbf{d}_{12} = \mathbf{d}_{21}$, entries of matrix $\mathbf{d}$ take on one of six possible values. In Table 5, analysis of the these six values and their corresponding frequencies shows that a completely randomized design yields $\frac{1}{n}||\mathbf{d}||_{1,1} = 2\left(\frac{n_t}{n_c} + \frac{n_c}{n_t} + 2\right) = O(1)$. Thus, Condition 3 is satisfied. Therefore, by Theorem 6.2, Horvitz-Thompson estimators are consistent for completely randomized experiments for bounded contrast, c, and outcome vector, y. △

6.2 Conditions for Convergence of WLS estimator class

Condition 4 (Bounded covariates)**.** *There exists a finite value* $u_{\mathbf{x}}$ *that bounds the covariate values, i.e.,* $\max(|\mathbf{x}|) < u_{\mathbf{x}}$*, for all* n*.*

Condition 5 (Bounded $\boldsymbol{\pi}\mathbf{m}$)**.** *There exists a finite value* $u_{\boldsymbol{\pi}\mathbf{m}}$ *that bounds* $\boldsymbol{\pi}$ *times the WLS "weighting" matrix* $\mathbf{m}$*, i.e.,* $\max(|\boldsymbol{\pi}\mathbf{m}|) < u_{\boldsymbol{\pi}\mathbf{m}}$*, for all* n*.*

Lemma 6.3 (Root-n consistency of WLS)**.** *By conditions 2-5 and Theorem 6.2, the WLS "numerator" vector,* $\frac{1}{n}\mathbb{x}'\mathbf{R}y$*, is root-n consistent for* $\frac{1}{n}\mathbb{x}'\mathbf{m}\boldsymbol{\pi}y$*. Likewise, by conditions 3-5 and Theorem 6.2, the WLS "denominator" matrix,* $\frac{1}{n}\mathbb{x}'\mathbf{mR}\mathbb{x}$*, is root-n consistent for* $\frac{1}{n}\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x}$*. Further, by the continuous mapping theorem* $\widehat{b}^{\text{WLS}} \to b^{\text{WLS}}$*.*

Proof. Let $\mathbb{x}_i$ be the column vector created from the i^{th} column of $\mathbb{x}$. Then the i^{the} element of $\frac{1}{n}\mathbb{x}\mathbf{mR}y$ can be written,

$$\begin{aligned}\{\frac{1}{n}\mathbb{x}\mathbf{mR}y\}_i =& \frac{1}{n}\mathbb{x}_i'\mathbf{mR}y \\ =& \frac{1}{n}1_{kn}'\text{diag}\left(\mathbb{x}_i\right)\mathbf{mR}y \\ =& \frac{1}{n}1_{kn}'\mathbf{Rm}\text{diag}\left(\mathbb{x}_i\right)y \\ =& 1_k'\mathbf{w}^{\text{HT}}\mathbf{R}\boldsymbol{\pi}\mathbf{m}\text{diag}\left(\mathbb{x}_i\right)y \\ =& 1_k'\mathbf{w}^{\text{HT}}\mathbf{R}q\end{aligned}$$

with $q = \boldsymbol{\pi}\mathbf{m}\text{diag}\left(\mathbb{x}_i\right)y$, showing that the elements of the denominator matrix are Horvitz-Thompson estimators with outcome vector q and contrast vector 1_k. Now by Theorem 6.1, this is consistent because q is bounded, i.e., $\max(|q|) \leq u_{\boldsymbol{\pi}\mathbf{m}}u_{\mathbf{x}}u_y$.

Similarly, the i, j element of the WLS "denominator" matrix, can be written

$$\{\frac{1}{n}\mathbb{x}\mathbf{mR}\mathbb{x}\}_{ij} = 1_{k}'\mathbf{w}^{\text{HT}}\mathbf{R}r$$

with $r = \boldsymbol{\pi}\mathbf{m}\text{diag}\,(\mathbb{x}_i)\,\mathbb{x}_j$, showing that the elements of the denominator matrix are Horvitz-Thompson estimators with outcome vector r and contrast vector 1_k. Now by Theorem 6.1, this is consistent because r is bounded, i.e., $\max(|r|) \leq u_{\boldsymbol{\pi}\mathbf{m}}u_{\mathbf{x}}^2$. □

Condition 6 (Stability of WLS "denominator" estimand)**.** *The denominator of the "true" WLS coefficient, $\frac{1}{n}\mathbb{x}'\mathbf{m}\pi\mathbb{x}$, is invertable for all n and converges in probability to a matrix, $\mathbf{v}$, with finite entries.*

Theorem 6.4 (Consistency of the Taylor approximation)**.** *By Conditions 2-6, the Taylor approximate coefficient, $\widehat{b}^{\text{T(WLS)}} := b^{\text{WLS}} + (\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x})^{-1}\,\mathbb{x}'\mathbf{mR}(y - \mathbb{x}b^{\text{WLS}})$ is root-n consistent for $b^{\text{WLS}} := (\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x})^{-1}\,\mathbb{x}'\mathbf{m}\boldsymbol{\pi}y$, i.e., $\widehat{b}^{\text{T(WLS)}} - b^{\text{WLS}} = O_p(1/\sqrt{n})$.*

Proof. We have

$$\begin{aligned}
\widehat{b}^{\text{T(WLS)}} - b^{\text{WLS}} &:= (\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x})^{-1}\,\mathbb{x}'\mathbf{mR}(y - \mathbb{x}b^{\text{WLS}}) \\
&= (\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x})^{-1}\left((\mathbb{x}'\mathbf{mR}y - \mathbb{x}'\mathbf{m}\boldsymbol{\pi}y) - (\mathbb{x}'\mathbf{mR}\mathbb{x} - \mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x})\,b^{\text{WLS}}\right) \\
&= \left(\frac{1}{n}\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x}\right)^{-1}\left(\left(\frac{1}{n}\mathbb{x}'\mathbf{mR}y - \frac{1}{n}\mathbb{x}'\mathbf{m}\boldsymbol{\pi}y\right) - \left(\frac{1}{n}\mathbb{x}'\mathbf{mR}\mathbb{x} - \frac{1}{n}\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x}\right)b^{\text{WLS}}\right) \\
&= O_p(1)\left(O_p(1/\sqrt{n}) + O_p(1/\sqrt{n})\right), \\
&= O_p(1/\sqrt{n})
\end{aligned}$$

where the second to last line uses Lemma 6.3 and Condition 6. □

Theorem 6.5 (Asymptotic Equivalence of WLS and its Taylor Approximation)**.** *By Lemma 6.3 and Condition 6, WLS is asymptotically equivalent to the Taylor linear approximation for WLS.*

Proof. Let the "true" WLS coefficient be $b^{\text{WLS}} = (\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x})^{-1}\,\mathbb{x}'\mathbf{m}\boldsymbol{\pi}y$, then write the WLS estimator as,

$$\begin{aligned}
c'\mathbf{W}^{\text{WLS}}\mathbf{R}y =& c'\,(\mathbb{x}'\mathbf{mR}\mathbb{x})^{-1}\,\mathbb{x}'\mathbf{mR}y \\
=& c'b^{\text{WLS}} + c'\,(\mathbb{x}'\mathbf{mR}\mathbb{x})^{-1}\,\mathbb{x}'\mathbf{mR}\,(y - \mathbb{x}b^{\text{WLS}}) \\
=& c'b^{\text{WLS}} + c'\,(\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x})^{-1}\,\mathbb{x}'\mathbf{mR}\,(y - \mathbb{x}b^{\text{WLS}}) \\
&\quad + c'\left((\mathbb{x}'\mathbf{mR}\mathbb{x})^{-1} - (\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x})^{-1}\right)\mathbb{x}'\mathbf{mR}\,(y - \mathbb{x}b^{\text{WLS}}) \\
=& c'b^{\text{WLS}} + c'\,(\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x})^{-1}\,\mathbb{x}'\mathbf{mR}\,(y - \mathbb{x}b^{\text{WLS}}) \\
&\quad + c'\left((\mathbb{x}'\mathbf{mR}\mathbb{x})^{-1} - (\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x})^{-1}\right)\mathbb{x}'\mathbf{mR}\left(\mathbb{x}\widehat{b}^{\text{WLS}} - \mathbb{x}b^{\text{WLS}}\right) \\
&\quad + c'\left((\mathbb{x}'\mathbf{mR}\mathbb{x})^{-1} - (\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x})^{-1}\right)\mathbb{x}'\mathbf{mR}\left(y - \mathbb{x}\widehat{b}^{\text{WLS}}\right) \\
=& \widehat{\delta}_c^{\text{T(WLS)}} + c'\left(\left(\frac{1}{n}\mathbb{x}'\mathbf{mR}\mathbb{x}\right)^{-1} - \left(\frac{1}{n}\mathbb{x}'\mathbf{m}\boldsymbol{\pi}\mathbb{x}\right)^{-1}\right)\left(\frac{1}{n}\mathbb{x}'\mathbf{mR}\mathbb{x}\right)\left(\widehat{b}^{\text{WLS}} - b^{\text{WLS}}\right) \\
=& \widehat{\delta}_c^{\text{T(WLS)}} + O_p(1/\sqrt{n})O_p(1)O_p(1/\sqrt{n}) \\
=& \widehat{\delta}_c^{\text{T(WLS)}} + O_p(1/n)
\end{aligned}$$

□

6.3 Conditions for consistent variance estimation

Definition 6.6 (Second-order design matrix)**.** *The "second-order design matrix" is a forth order tensor* ($kn \times kn \times kn \times kn$) *of variances and covariances of inverse-probability weighted pairwise-joint inclusion indicators, written,*

$$\mathbb{d} := \Big(\mathrm{E}\left[(\mathbf{R}1_{n}1'_{kn}\mathbf{R}) \otimes (\mathbf{R}1_{kn}1'_{kn}\mathbf{R})\right] - \mathbf{p} \otimes \mathbf{p}\Big) / \left(\mathbf{p} \otimes \mathbf{p}\right),$$

where $\mathbf{p} := \mathrm{E}\left[\mathbf{R}1_{n}1'_{kn}\mathbf{R}\right]$ *is a matrix of with inclusion probabilities on the diagonal and pairwise joint inclusion probabilities off the diagonal, "$\otimes$" is the tensor outer product and "/" is elementwise division with division by zero resolving to zero.*

Condition 7 (Second order design constraint for consistent variance estimation)**.** *There exists a finite constant* $u_{\mathbb{d}}$ *such that* $\frac{1}{n}\left|\left|\left(\tilde{\mathbf{d}} \otimes \tilde{\mathbf{d}}\right) \circ \mathbb{d}\right|\right|_{1,1,1,1} < u_{\mathbb{d}}$ *for all* n*, where "$\otimes$" is tensor outer product, "$\circ$" is elementwise multiplication,* $||.||_{1,1,1,1}$ *gives the sum of the absolute values of the tensor entries.*

Theorem 6.7 (Consistency of the Horvitz-Thompson variance estimator)**.** *By Conditions 1-3 and 7 the variance estimator for the Horvitz-Thompson point estimator is consistent.*

Proof. The variance of n times the Horvitz-Thompson variance estimator (times n) is,

$$\begin{aligned} n\mathrm{V}\left(n\widehat{\tilde{\mathrm{V}}}\left(\widehat{\delta}^{\mathrm{HT}}\right)\right) =& n\mathrm{E}\left[\left(nz^{\mathrm{HT}\prime}\mathbf{R}\tilde{\mathbf{d}}_{/\mathbf{p}}\mathbf{R}z^{\mathrm{HT}} - nz^{\mathrm{HT}\prime}\tilde{\mathbf{d}}z^{\mathrm{HT}}\right)^{2}\right] \\ &\leq \max\left(|c|\right)^{4}\max\left(|y|\right)^{4}\frac{1}{n}\left|\left|\left(\tilde{\mathbf{d}} \otimes \tilde{\mathbf{d}}\right) \circ \mathbb{d}\right|\right|_{1,1,1,1} \\ &\leq u_{c}^{4}u_{y}^{4}u_{\mathbb{d}}, \end{aligned}$$

where the second line uses Holder's inequality and the last line uses Conditions 1, 2 and 7. □

Example (*Checking consistency of HT variance (bound) estimator for completely randomized experiments*) : Again consider a completely randomized experiment with n units where a fixed number of units, n_c, are randomly assigned to control, and the remainder, $n_t = n - n_c$, are assigned to treatment. Assume an asymptotic sequence of designs is such that there exists a constant value, π, such that $\frac{n_t}{n} \to \pi$ as $n \to \infty$ with $0 < \pi < 1$. Let the variance bound be the Neyman bound, i.e., $\tilde{\mathbf{d}} = \tilde{\mathbf{d}}^{\mathrm{N}}$ from Definition 4.1. Table 6.3 enumerates the unique values that appear in $\frac{1}{n}\left|\left|\left(\tilde{\mathbf{d}}^{\mathrm{N}} \otimes \tilde{\mathbf{d}}^{\mathrm{N}}\right) \circ \mathbb{d}\right|\right|_{1,1,1,1}$ along with their relative frequencies and shows that Condition 7 is satisfied for completely randomized experiments. Hence, the Horvitz-Thompson variance (bound) estimator given in Equation 7 is consistent for the Neyman bound. △

Treatment (T) or Control (C)	$ijkl$ pattern	count	$\{\mathrm{E}[(\mathbf{R}1_n1'_{kn}\mathbf{R}) \otimes(\mathbf{R}1_n1'_{kn}\mathbf{R})]\}_{ijkl}$	$\{\mathbf{p}\otimes\mathbf{p}\}_{ijkl}$	$\{\tilde{\mathbf{d}}^{\mathrm{N}}\otimes\tilde{\mathbf{d}}^{\mathrm{N}}\}_{ijkl}$	count $\times\frac{1}{n}\{(\tilde{\mathbf{d}}\otimes\tilde{\mathbf{d}})\circ\mathbf{d}\}_{ijkl}$
$i,j,k,l \in$C	$i=j=k=l$	n	$\frac{n_c}{n}$	$\frac{n_c^2}{n^2}$	$\frac{n^2}{n_c^2}$	$\frac{n^2n_t}{n_c^3}=O(1)$
	$i=j=k,l$ or $i=j=l,k$ or $i,j=k=l$ or $i=k=l,j$	$4n(n-1)$	$\frac{n_c(n_c-1)}{n(n-1)}$	$\frac{n_c^2(n_c-1)}{n^2(n-1)}$	$-\frac{n^2}{n_c^2(n-1)}$	$-\frac{4n^2n_t}{n_c^3}=O(1)$
	$i=j,k=l$	$n(n-1)$	$\frac{n_c(n_c-1)}{n(n-1)}$	$\frac{n_c^2}{n^2}$	$\frac{n^2}{n_c^2}$	$\frac{n_tn^2}{n_c^3}=O(1)$
	$i=j,k,l$ or $i,j,k=l$	$2n(n-1)(n-2)$	$\frac{n_c(n_c-1)(n_c-2)}{n(n-1)(n-2)}$	$\frac{n_c^2(n_c-1)}{n^2(n-1)}$	$-\frac{n^2}{n_c^2(n-1)}$	$\frac{4n_tn^2}{n_c^3}=O(1)$
	$i=k,j=l$ or $i=l,j=k$	$2n(n-1)$	$\frac{n_c(n_c-1)}{n(n-1)}$	$\frac{n_c^2(n_c-1)^2}{n^2(n-1)^2}$	$-\frac{n^2}{n_c^2(n-1)^2}$	$-\frac{2n^2}{n_c^2}=O(1)$
	$i=k,j,l$ or $i=l,j,k$ or $i,j=k,l$ or $i,j=l,k$	$4n(n-1)(n-2)$	$\frac{n_c(n_c-1)(n_c-2)}{n(n-1)(n-2)}$	$\frac{n_c^2(n_c-1)^2}{n^2(n-1)^2}$	$\frac{n^2}{n_c^2(n-1)^2}$	$\frac{4n^2}{n_c^2}\left(\frac{n(n_c-2)}{n_c(n_c-1)}-\frac{(n-2)}{(n-1)}\right)=O(1)$
	i,j,k,l	$n(n-1)(n-2)(n-3)$	$\frac{n_c(n_c-1)(n_c-2)(n_c-3)}{n(n-1)(n-2)(n-3)}$	$\frac{n_c^2(n_c-1)^2}{n^2(n-1)^2}$	$\frac{n^2}{n_c^2(n-1)^2}$	$\frac{-4n_tn^3}{n_c^2(n-1)(n_c-1)}-\frac{6n^2}{n_c^2(n-1)}+\frac{6n^3}{n_c^3(n_c-1)}=O(1)$
$i,j \in$T $k,l \in$C	$i=j=k=l$	$2n$	0	$\frac{n_cn_t}{n^2}$	$\frac{n^2}{n_tn_c}$	$-\frac{n^2}{n_tn_c}=O(1)$
or $i,j \in$C $k,l \in$T	$i=j=k,l$ or $i=j=l,k$ or $i,j=k=l$ or $i=k=l,j$	$8n(n-1)$	0	$\frac{n_cn_t(n_t-1)}{n^2(n-1)}$	$-\frac{n^2}{n_tn_c(n-1)}$	$-\frac{8n^2}{n_tn_c}=O(1)$
	$i=j,k=l$	$2n(n-1)$	$\frac{n_tn_c}{n^2}$	$\frac{n_cn_t}{n^2}$	$\frac{n^2}{n_tn_c}$	0
	$i=j,k,l$	$4n(n-1)(n-2)$	$\frac{n_tn_c(n_t-1)}{n^2(n-1)}$	$\frac{n_cn_t(n_t-1)}{n^2(n-1)}$	$\frac{n^2}{n_tn_c(n-1)}$	0
	$i=k,j=l$ or $i=l,j=k$	$4n(n-1)$	0	$\frac{n_cn_t(n_t-1)(n_c-1)}{n^2(n-1)^2}$	$\frac{n^2}{n_tn_c(n-1)^2}$	$\frac{4n^2}{n_tn_c(n-1)}=O(1/n)$
	$i=k=l,j$ or $i,j=k=l$	$4n(n-1)$	0	$\frac{n_cn_t(n_c-1)}{n^2(n-1)}$	$-\frac{n^2}{n_tn_c(n-1)}$	$-\frac{4n^2}{n_tn_c}=O(1)$
	$i=k,j,l$ or $i=l,j,k$ or $i,j=k,l$ or $i,j=l,k$	$8n(n-1)(n-2)$	0	$\frac{n_cn_t(n_c-1)(n_t-1)}{n^2(n-1)^2}$	$\frac{n^2}{n_tn_c(n-1)^2}$	$\frac{8n^2(n-2)}{n_tn_c(n-1)}=O(1)$
	i,j,k,l	$2n(n-1)(n-2)(n-3)$	$\frac{n_t(n_t-1)n_c(n_c-1)}{n(n-1)(n-2)(n-3)}$	$\frac{n_cn_t(n_c-1)(n_t-1)}{n^2(n-1)^2}$	$\frac{n^2}{n_tn_c(n-1)^2}$	$\frac{4n^2(n-3)}{n_tn_c(n-1)}=O(1)$
$i,j,k,l \in$T	$i=j=k=l$	n	$\frac{n_t}{n}$	$\frac{n_t^2}{n^2}$	$\frac{n^2}{n_t^2}$	$\frac{n^2n_c}{n_t^3}=O(1)$
	$i=j=k,l$ or $i=j=l,k$ or $i,j=k=l$ or $i=k=l,j$	$4n(n-1)$	$\frac{n_t(n_t-1)}{n(n-1)}$	$\frac{n_t^2(n_t-1)}{n^2(n-1)}$	$-\frac{n^2}{n_t^2(n-1)}$	$-\frac{4n^2n_c}{n_t^3}=O(1)$
	$i=j,k=l$	$n(n-1)$	$\frac{n_t(n_t-1)}{n(n-1)}$	$\frac{n_t^2}{n^2}$	$\frac{n^2}{n_t^2}$	$\frac{n_cn^2}{n_t^3}=O(1)$
	$i=j,k,l$ or $i,j,k=l$	$2n(n-1)(n-2)$	$\frac{n_t(n_t-1)(n_t-2)}{n(n-1)(n-2)}$	$\frac{n_t^2(n_t-1)}{n^2(n-1)}$	$-\frac{n^2}{n_t^2(n-1)}$	$\frac{4n_cn^2}{n_t^3}=O(1)$
	$i=k,j=l$ or $i=l,j=k$	$2n(n-1)$	$\frac{n_t(n_t-1)}{n(n-1)}$	$\frac{n_t^2(n_t-1)^2}{n^2(n-1)^2}$	$-\frac{n^2}{n_t^2(n-1)^2}$	$-\frac{2n^2}{n_t^2}=O(1)$
	$i=k,j,l$ or $i=l,j,k$ or $i,j=k,l$ or $i,j=l,k$	$4n(n-1)(n-2)$	$\frac{n_t(n_t-1)(n_t-2)}{n(n-1)(n-2)}$	$\frac{n_t^2(n_t-1)^2}{n^2(n-1)^2}$	$\frac{n^2}{n_t^2(n-1)^2}$	$\frac{4n^2}{n_t^2}\left(\frac{n(n_t-2)}{n_t(n_t-1)}-\frac{(n-2)}{(n-1)}\right)=O(1)$
	i,j,k,l	$n(n-1)(n-2)(n-3)$	$\frac{n_t(n_t-1)(n_t-2)(n_t-3)}{n(n-1)(n-2)(n-3)}$	$\frac{n_t^2(n_t-1)^2}{n^2(n-1)^2}$	$\frac{n^2}{n_t^2(n-1)^2}$	$\frac{-4n_cn^3}{n_t^2(n-1)(n_t-1)}-\frac{6n^2}{n_t^2(n-1)}+\frac{6n^3}{n_t^3(n_t-1)}=O(1)$

Table 6: Analysis of Condition 7 for Complete Randomization

References

Athey, S., and Imbens, G.W. 2017. The Econometrics of Randomized Experiments. *Handbook of Economic Field Experiments*, **1**: 73-140.

Arceneaux, Kevin, and David Nickerson. 2009. Modeling uncertainty with clustered data: A comparison of methods, Political Analysis, **17**: 177–90.

Aronow, Peter M. and Cyrus Samii. 2012. Conservative variance estimation for sampling designs with zero pairwise inclusion probabilities. *Survey Methodology* **39**(1): 231-241.

Aronow, Peter M. and Cyrus Samii. 2017. Estimating average causal effects under general interference, with application to a social network experiment. Forthcoming at *The Annals of Applied Statistics* .

Aronow, Peter M. and Joel A. Middleton. 2013. A class of unbiased estimators of average treatment effect in randomized experiments. *Journal of Causal Inference* **1**(1): 135-154.

Basse, G. and A. Feller. 2017. Analyzing two-stage experiments in the presence of interference, *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2017.1323641

Bloniarz, A., Liu, H., Zhang, C.H., Sekhon, J.S., and Yu, B. 2016. Lasso adjustments of treatment effect estimates in randomized experiments, *Proceedings of the National Academy of Sciences*, **113**(27): 7383-90.

Campbell, Stephen L., and Carl D. Meyer. 2009. Generalized Inverses of Linear Transformations. *https://doi.org/10.1137/1.9780898719048*

Fuller, W.A. 2009. *Sampling Statistics.* New Jersey: Wiley.

Fuller, W.A. and C.T. Isaki. 1981. Survey Design Under Superpopulation Models In: *Current Topics in Survey Sampling* Eds: Krewski, D. , J.N.K. Rao, R. Platek. New York, Academic Press.

Freedman, D.A. 2008a. On regression adjustments to experimental data. *Adv. in Appl. Math.* **40** 180–193.

Freedman, D.A. 2008b. On regression adjustments in experiments with several treatments. *Ann. Appl. Stat.* **2** 176–196.

Hansen, B. and Bowers, J. (2009). Attributing effects to a cluster-randomized get-out-the-vote campaign. *J. Amer. Statist. Assoc.* **104** 873–885.

Holland, P.W. 1986. Statistics and Causal Inference, *Journal of the American Statistical Association*, vol. 81, no. 396: 945-968.

Horvitz, D.G. and Thompson, D.J. 1952. A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47**: 663-684.

Isaki, C.T., and W.A. Fuller. 1982. Survey Design Under the Regression Superpopulation Model. *Journal of the American Statistical Association* **77**(377): 89-96

Li, Xinran and Ding, Peng. 2017. General Forms of Finite Population Central Limit Theorems with Applications to Causal Inference. *Journal of the American Statistical Association* **112**(520): 1759-1769

Li, Xinran, Peng Ding, and Donald B. Rubin. 2017. Asymptotic Theory of Rerandomization in Treatment-Control Experiments. *arXiv:1604.00698*

Lin, Winston. 2013. Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique. *Annals of Applied Statistics* **7**(1): 295-318

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

Middleton, J.A. 2008. Bias of the regression estimator for experiments using clustered random assignment. *Stat. Probability Lett.* **78** 2654–2659.

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

Middleton, Joel A. and Peter M. Aronow. 2015. Unbiased Estimation of the Average Treatment Effect in Cluster-Randomized Experiments. *Statistics, Politics and Policy* **1**:

Neyman, Jerzy Splawa, D. M. Dabrowska, and T. P. Speed. [1923.] 1990. On the application of probability theory to agricultural experiments: Essay on principles, section 9. *Statistical Science* **5**: 465–480.

Raj, D. 1965. On a method of using multi-auxiliary information in sample surveys. *J. Amer. Statist. Assoc.* **60** 270–277.

Rohde, Charles. 1965. Generalized Inverses of Partitioned Matrices. *Journal of the Society for Industrial and Applied Mathematics* **13**(4): 1033-1035.

Rubin, Donald. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**: 688–701.

Sarndal, C.-E., B. Swensson, and J. Wretman. 1992. Model Assisted Survey Sampling. New York: Springer.

Samii, Cyrus and Peter M. Aronow. 2012. On Equivalencies Between Design-Based and Regression-Based Variance Estimators for Randomized Experiments. *Statistics and Probability Letters.* **82**: 365–370.

Schochet, Peter Z. 2010. Is regression adjustment supported by the Neyman model for causal inference? *Journal of Statistical Planning and Inference* **140**: 246-259.

Sinclair, B., McConnell, M. and Green, D.P. 2012. Detecting Spillover Effects: Design and Analysis of Multilevel Experiments. *American Journal of Political Science* **56**(4): 1055-1069.

Wood, John. 2008. On the Covariance Between Related Horvitz-Thompson Estimators. *Journal of Official Statistics.* **24** 53–78.

Zhao, A., Ding, P., Mukerjee, R., and Dasgupta, T. 2017+. Randomization-Based Causal Inference From Unbalanced 2^2 Split-Plot Designs. *Annals of Statistics*, in press.

A Notation Index

n	Number of units in the finite population in the experiment
1_{kn}	Length-kn column vector of 1's. In matrix notation, serves as a replacement for the more common summation symbol, Σ
y_{0i}, y_{1i}	The control and treatment potential outcomes for the i^{th} unit, respectively
y_0, y_1	Length-n vectors of control and treatment potential outcomes, respetively
y	Length-kn vector of all potential outcomes. The first n elements are control potential outcomes multiplied by -1, followed by the treatment potential outcomes. Multiplication of control potential outcomes by -1 allows for the compact representation of the ATE as the sum of the elements of this vector divided by n
δ	Average treatment effect (ATE), the parameter of interest
R_{0i}, R_{1i}	Random indicators of the i^{th} unit's assignment to control and treatment, respectively
R_0, R_1	Length-n vectors of assignment indicators for control and treatment, respectively
$\mathbf{R}$	$kn \times kn$ diagonal matrix of assignment indicators. The first n diagonal elements represent the control indicators, followed by n treatment indicators
π_{0i}, π_{1i}	For the i^{th} unit, the probability of assignment to control and treatment, respectively
π_0, π_1	Length-n vectors of probabilities of assignment to control and treatment, respectively
$\boldsymbol{\pi}$	$kn \times kn$ diagonal matrix of assignment probabilities. The first n diagonal elements give the control probabilities, followed by the treatment probabilities
π_{0i0j}, π_{0i1j}, π_{1i0j}, π_{1i1j}	Joint assignment probabilities for units i and j. For example, π_{1i0j} is the probability that i is in treatment and j is in control
$\mathbf{d}$	$kn \times kn$ "design" matrix that gives the variance-covariance matrix of the vector $1'_{kn}\boldsymbol{\pi}^{-1}\mathbf{R}$. Allows for compact representation of variance of HT estimators as a quadratic in matrix form
$\mathbf{d}_{00}$, $\mathbf{d}_{01}$, $\mathbf{d}_{10}$, $\mathbf{d}_{11}$	The four $n \times n$ partitions of the matrix $\mathbf{d}$. For example, the top-right partition, $\mathbf{d}_{01}$, has i, j element $\frac{\pi_{0i1j} - \pi_{0i}\pi_{1j}}{\pi_{0i}\pi_{1j}}$
$\tilde{\mathbf{d}}$	A modified version of $\mathbf{d}$ that allows for compact representation of a variance *bound* for HT estimators as a quadratic in matrix form. While the variance of the HT estimator is not identified, a variance bound may be

$\mathbf{p}$	$kn \times kn$ "probability" matrix that gives the joint assignment probabilities
$\mathbf{p}_{00}$, $\mathbf{p}_{01}$, $\mathbf{p}_{10}$, $\mathbf{p}_{11}$	The four $n \times n$ quadrants of the matrix $\mathbf{p}$. For example, $\mathbf{p}_{01}$ has ij element π_{0i1j}
$\tilde{\mathbf{p}}$	A modified version of $\mathbf{p}$ that replaces zeros with ones. Allows for division by $\tilde{\mathbf{p}}$ without division-by-zero error
x_i	Length-k vector of covariates associated with the i^{th} unit
$\mathbf{x}$	An $n \times k$ matrix of covariates
$\tilde{\mathbf{x}}$	An $n \times (k + 1)$ matrix representing the concatenation of an intercept vector, 1_n, and $\mathbf{x}$
$\mathbb{x}$	A $kn \times l$ matrix of covariates. The first n rows are multiplied by -1 to mirror the vector y. Represents an arbitrary specification
$\mathbb{x}_{\text{I}}$	A $kn \times (k + 2)$ matrix of covariates. The "common slopes" specification. Elements in the first n rows are multiplied by -1 to mirror the vector y
$\mathbb{x}_{\text{II}}$	A $kn \times (2k + 2)$ matrix of covariates. The "separate slopes" specification. Elements in the first n rows are multiplied by -1 to mirror the vector y

B Supplementary Proofs