A Unifying Design-based Theory of Regression Adjustment

WORKING PAPER

Joel A. Middleton*

July 9, 2019

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

1 Introduction

1.1 The Neyman Causal Model for treatment-control experiments

Consider the Neyman causal model (NCM) and imagine a two-arm experiment. For convenience, refer to it as a treatment-control experiment, with one arm being referred to as the treatment and the other control. For the i^{th} unit of the (finite) study population of n units, there are two (fixed) potential outcomes: y_{0i} and y_{1i}. Which of i's outcomes is reveled is determined only by i's treatment assignment indicator, R_{1i}, which is random, the only random component in the NCM. The researcher observes for each unit the outcome $(R_{1i},\ R_{1i}y_{1i},\ (1-R_{1i})y_{0i},\ x_i)$, where x_i is an additional vector of k covariates which is observed for every unit irrespective of assignment.

The "fundamental problem of causal inference" (Holland, 1986) is that only one of the two potential outcomes can be observed for each unit such that a treatment effect can never be observed for a given individual. So, very commonly, the inferential target is the average treatment effect (ATE), which can be written,

$$\delta := n^{-1}\sum_i y_{1i} - n^{-1}\sum_i y_{0i}. \tag{1}$$

As will become evident later in the development, it will simplify exposition to re-represent the problem in a nonstandard way starting with the potential outcomes. First, represent the entire schedule of potential outcomes as a single column vector,

$$y := \begin{pmatrix} -y_{01} & -y_{02} & \dots & -y_{0n} & y_{11} & y_{12} & \dots & y_{1n} \end{pmatrix}', \tag{2}$$

and note that y has length $2n$. The first n elements are control potential outcomes multiplied by -1. The next n entries are treatment potential outcomes. With the control potential outcomes multiplied by -1, the ATE can be represented as the sum of this vector divided by n,

$$\delta = n^{-1}1_{2n}'y, \tag{3}$$

where 1_{2n} is a column vector of $2n$ ones. Of course, equations (1) and (3) are algebraically equivalent.

Given the definition of R_{1i} from above as the (random) treatment assignment indicator for the i^{th} individual, we can also define the control assignment indicator as $R_{0i} := 1 - R_{1i}$, then the $2n \times 2n$ matrix

$$\mathbf{R} := \begin{bmatrix} R_{01} & & & & & & & \\ & R_{02} & & & & & & \\ & & \ddots & & & & & \\ & & & R_{0n} & & & & \\ & & & & R_{11} & & & \\ & & & & & R_{12} & & \\ & & & & & & \ddots & \\ & & & & & & & R_{1n} \end{bmatrix},$$

gives these random indicators on its diagonal. In this notation the researcher can be said to observe $\mathbf{R}$ and $\mathbf{R}y$, and also an $n \times k$ matrix of pre-treatment covariates, $\mathbf{x}$.

A $2n \times 2n$ diagonal matrix of assignment probabilities can be defined as

$$\boldsymbol{\pi} := \mathrm{E}[\mathbf{R}] \tag{4}$$

with the first n diagonal elements representing probabilities of assignment to control and, the next n, probabilities of assignment to treatment. In a randomized experiment $\boldsymbol{\pi}$ is likely known, however, it could be estimated by simulation of the randomization procedure if necessary.

1.2 The Horvitz-Thompson Estimator

Next, the Horvitz-Thompson (HT) estimator of the ATE can be defined as

$$\widehat{\delta}^{HT} := n^{-1} 1_{2n}' \boldsymbol{\pi}^{-1} \mathbf{R} y. \tag{5}$$

where 1_{2n} is, as above, a column vector of $2n$ ones. Note that this formulation modifies equation (3) by substituting $\boldsymbol{\pi}^{-1}\mathbf{R}y$ in place of y. The diagonal matrix $\boldsymbol{\pi}^{-1}$ does the work of weighting the potential outcomes inversely proportional to the probability of being observed. Note that the HT estimator has the virtue of being unbiased for any identified design, i.e., designs in which $0 < \pi_{1i} < 1$ for all i, because, also using the definition in equation (4), $\mathrm{E}[\boldsymbol{\pi}^{-1}\mathbf{R}] = \mathbf{i}$, where $\mathbf{i}$ is a $(2n \times 2n)$ identity matrix.

The HT estimator has the same form as the inverse propensity of treatment weighted (IPTW) estimator, but the assignment probability is known in this setting by way of control over the design of the experiment. Moreover, variance expressions under the IPTW framework assume away joint inclusion probabilities, a limitation that is not shared by the development of HT estimators as will be seen in the next subsection.

1.3 Variance of the HT estimator

To express the variance of a HT estimator of the average treatment effect, first note that in equation (5), the elements of y can be seen as coefficients associated with the random vector $1_{2n}'\boldsymbol{\pi}^{-1}\mathbf{R}$. Thus, if one defines the $2n \times 2n$ "design" matrix

$$\mathbf{d} := \mathrm{V}\left(1_{2n}'\boldsymbol{\pi}^{-1}\mathbf{R}\right) \tag{6}$$

where $\mathrm{V}(.)$ represents variance-covariance, then the variance of the HT estimator of the average treatment effect can be written compactly as

$$\mathrm{V}\left(\widehat{\delta}^{HT}\right) = n^{-2} y' \mathbf{d} y. \tag{7}$$

As with the HT estimator given above, this variance expression is general, and applies to any design. Equivalent expressions are given by Aronow and Middleton (2015) and Aronow and Samii (2017), but (7) is considerably more concise.

It can be useful to examine the structure of $\mathbf{d}$. To do so, first define the joint probability that units i and j are both included in the treatment arm $\pi_{1i1j} := \mathrm{E}[R_{1i}R_{1j}]$, and note that the probability of assignment to treatment for unit i could be written π_{1i1i} or π_{1i}. Similarly, the joint probability of inclusion in the control group is $\pi_{0i0j} := \mathrm{E}[R_{0i}R_{0j}]$. Moreover, $\pi_{1i0j} := \mathrm{E}[R_{1i}R_{0j}]$ is the probability that i is in treatment and j is in control, and $\pi_{0i1j} := \mathrm{E}[R_{0i}R_{1j}]$ is the probability that i is in control and j is in treatment.

Next, note that $\mathbf{d}$ can be partitioned into four $n \times n$ matrices. Write

$$\mathbf{d} = \begin{bmatrix} \mathbf{d}_{00} & \mathbf{d}_{01} \\ \mathbf{d}_{10} & \mathbf{d}_{11} \end{bmatrix} \tag{8}$$

where, for example, the matrix $\mathbf{d}_{11}$ has ij element $\frac{\pi_{1i1j}}{\pi_{1i}\pi_{1j}} - 1$, and on the diagonal, with $i = j$, note the equivalence $\frac{\pi_{1i1i}}{\pi_{1i}\pi_{1i}} - 1 = \frac{1}{\pi_{1i}} - 1$. Similarly, the matrix $\mathbf{d}_{10}$ has ij element $\frac{\pi_{1i0j}}{\pi_{1i}\pi_{0j}} - 1$ with diagonal elements $\frac{\pi_{1i0i}}{\pi_{1i}\pi_{0i}} - 1 = -1$ because $\pi_{1i0i} = 0$ due to the *fundamental problem of causal inference*. This also leads to the more general observation, which will be important below, that any cells of the matrix $\mathbf{d}$ that have values of -1 indicate pairs of potential outcomes that can never be observed together, and, hence, correspond to components of the variance that are not identified. Sub-matrices $\mathbf{d}_{00}$ and $\mathbf{d}_{01}$ are defined analogously.

Additionally, if we define y_0 and y_1 to be length-n vectors of control and treatment potential outcomes, respectively, then the variance in equation (7) can also be written as the sum of four terms, $n^{-2}y_0'\mathbf{d}_{00}y_0$, $n^{-2}y_1'\mathbf{d}_{11}y_1$, $-n^{-2}y_1'\mathbf{d}_{10}y_0$, and $-n^{-2}y_0'\mathbf{d}_{01}y_1$, which give variance and covariance terms that are seen in other developments.

Since not all pairs of potential outcomes can be observed together, so-called *variance estimators* are actually *variance bound estimators*. Several contributions to variance bounding and bound estimation are made later in Section 3 after regression has been introduced.

1.4 Consistency of the HT estimator

To obtain root-n consistency of the HT estimator some regularity conditions are required on the sequence of outcome vectors, y_n, and sequence of designs with corresponding $\mathbf{d}_n$ as $n \to \infty$. Subscripts n are suppressed in the following theorem.

Theorem 1.1 (Sufficient conditions for root-n consistency of HT estimators)**.** *Let $|.|$ take the absolute value of each element of the matrix or vector, and assume there exists an $0 < l_y < \infty$ such that $\max(|y|) \leq l_y$ for all n. Moreover, assume there exists an $0 < l_{\mathbf{d}} < \infty$ such that $n^{-1}1'_{2n}|\mathbf{d}|1_{2n} < l_{\mathbf{d}}$ for all n. Then $\widehat{\delta}^{HT} - \delta$ is $O_p(n^{-0.5})$.*

Proof. Since HT estimators are unbiased, we simply need n times the variance to be $O(1)$. We have

$$\begin{aligned} n\mathrm{V}\left(\widehat{\delta}^{HT}\right) &\leq \max(|y|)^2 \left(n^{-1}1'_{2n}|\mathbf{d}|1_{2n}\right) \\ &\leq l_y^2\, l_{\mathbf{d}}. \end{aligned}$$

□

Remark 1. *The proof follows from the special case of Hölder's inequality where norm pair is the infinity norm and 1-norm. A more general statement is possible using norms r and s that satisfy $r^{-1} + s^{-1} = 1$.*

Remark 2. *Note that this statement separates the outcomes, y, from design elements that are collected in $\mathbf{d}$. By contrast, other treatments have tended to state requirements in terms of $\boldsymbol{\pi}^{-1}|y|$ and $\left(n^{-1}1'_{2n}|\boldsymbol{\pi}\mathbf{d}\boldsymbol{\pi}|1_{2n}\right)$. The term $\boldsymbol{\pi}^{-1}|y|$ incorporates both outcomes, which are a feature of the population, and assignment probabilities, which are an aspect of experimental design.*

Remark 3. *That $n^{-1}1'_{2n}|\mathbf{d}|1_{2n}$ converges to a constant as $n \to \infty$ is easily proven for complete randomization with constant proportion assigned to treatment, for example. It is likewise verifiable for cluster randomization with additional restrictions on the distribution of cluster sizes. For an arbitrary or complex design that can be nonetheless embedded in a hypothetical sequence, lack of growth of $n^{-1}1'_{2n}|\mathbf{d}|1_{2n}$ could be examined numerically.*

While finite sample central limit theorems for HT estimators exist for a number of designs, a CLT for arbitrary designs is beyond the scope of this paper. Isaki and Fuller (1982) and Fuller (2009) provide a general consistency results for HT estimators but derive asymptotic normality under a super-population model. Fuller and Isaki (1981) prove that a finite-population CLT holds for a particular unequal probability design, but this is not general. Aronow and Samii (2017) present sufficient conditions for certain complex designs, but these conditions do not include common designs such as complete randomization. Li and Ding (2017) summarize the literature on finite-population central limit theorems, which has mostly focused on the complete randomization case. The results of Li et al. (2017) underscore the fact that the asymptotic normality should not be taken for granted for some designs even though straightforward estimators, such as difference-of-means or HT, may be consistent and asymptotically unbiased.

1.5 Muli-Arm Designs, Other Estimands and Other Estimators

To show the flexibility of this framework, it is worth noting that contrasts between potential outcomes could be represented explicitly, rather than embedding $-1 \times y_0$ into the vector, y.

For example, consider a four-arm experiment, where y is a vector of length $4n$ with the first n elements of representing outcomes under arm 1, the next n representing outcomes associated with arm 2 and so on. Now define a *contrast matrix*,

$$\mathbf{c} = \begin{bmatrix} a_1\mathbf{i} & & & \\ & a_2\mathbf{i} & & \\ & & a_3\mathbf{i} & \\ & & & a_4\mathbf{i} \end{bmatrix},$$

where $\mathbf{i}$ is an $n \times n$ identity matrix and a_1, a_2, a_3, and a_4 are any constants that satisfy restrictions $a_1 + a_2 + a_3 + a_4 = 0$ and $|a_1| + |a_2| + |a_3| + |a_4| = 2$. Now in all previous equations, simply replace y with $\mathbf{c}y$. This allows the representation of various contrasts such as pairwise comparisons between arms, average marginal causal effects or interaction effects.

In designs where units are first sampled for inclusion in the experiment it may be necessary or useful to account for the uncertainty due to the sampling of units. In that case, "non-sampled" could be thought of as an additional treatment arm for purposes of mathematical analysis, with the understanding that the contrast weight for the associated outcomes would necessarily be 0.

Any weighted mean of potential outcomes can be represented with contrast matrices as well. Common motivations for re-weighting could include sub-group analysis, re-weightings to alternative demographic mixtures (i.e., estimating "population average treatment effects") or other quantities of interest such as attributable effects.

Contrast matrices also allow the representation of estimators other than HT in this framework. For example, suppose it is interesting to compare the variance and bias of the difference-of-means in a block randomized design with probabilities that differ across blocks. In that case, difference-of-means is just the HT estimator of the ATE for outcome $\mathbf{c}y$ with

$$\mathbf{c} = \begin{bmatrix} -n\, n_0^{-1} \boldsymbol{\pi}_0 & \\ & n\, n_1^{-1} \boldsymbol{\pi}_1 \end{bmatrix}.$$

The exact variance has the same form as (7) with the only difference being the introduction of the contrasted outcome $\mathbf{c}y$ in place of y.

The difference-of-means can be represented in this way for other designs as well, but n_1 and n_0 may be random variables, for example, in the case of cluster randomization with varying cluster sizes. This makes exact variance for the difference-of-means intractable. However, Taylor linearization provides approximate (asymptotic) variance expressions in terms of HT estimators of a location-shifted y.

A related result is that a number of estimators will take the form

$$n^{-1}\mathbf{m}\boldsymbol{\pi}^{-1}\mathbf{R}y$$

with $\mathbf{m}$ representing some matrix with $2n$ columns. Note that because $\mathbf{R}$ and $\boldsymbol{\pi}$ are diagonal, this can also be written $n^{-1}\mathbf{m}\text{diag}(y)\boldsymbol{\pi}^{-1}\mathbf{R}1_{2n}$ and taking the transpose reveals that this is equivalent to a vector of HT estimators, one for each of the columns of the matrix $\text{diag}\,(y)\,\mathbf{m}'$. Its exact variance-covariance expression is as in equation (7), but with $\text{diag}\,(y)\,\mathbf{m}'$ in place of y. Moreover, even if the matrix is random, a function of $\mathbf{R}$, but it converges at a sufficient rate to $\mathbf{m}$, then it is often possible express asymptotic or approximate variance in this way. This recognition provides some intuition for the asymptotic variance expressions of regression coefficients, given below.

The discussion about contrast matrices in this section notwithstanding, throughout the remainder of the paper the focus will be on two-arm experiments and the contrast between treatment and control arms will be embedded in y as given in equation (2).

2 Regression

Now that the basic framework has been established for estimation under the NCM, this section turns to the main subject of the paper, regression adjustment. In this section, a Generalized Regression estimator is proposed which is distinct from what will be referred to as *common regression practice*. Common regression practice directly interprets values contained in the coefficients to arrive at the ATE. Throughout I will refer to this practice as *interpreting the treatment coefficient*, noting that the phrase is shorthand and also refers to various equivalent practices, for example, in when taking the *difference* between two values in the coefficient vector as may be required depending on the specification.

In contrast to the common regression practice, the Generalized Regression estimator takes a a regression coefficient and uses it to make an *adjustment* to the HT estimator. For every conceivable common regression

coefficient, a corresponding GR estimator that utilizes that coefficient can be defined. Another way to put it, estimators in the class of GR are distinguished from one another by the particular regression coefficient that is used in making the adjustment. I refer a GR estimator and its respective coefficient as *conjugates*.

After introducing notation for covariate specifications in the new framework, the class of GR estimators is introduced, and the general parameters for consistency are provided. The first important result shows that $O_p(n^{-0.5})$ convergence of the GR estimator requires that the HT estimators converge at that rate under the design, but the conjugate coefficient is allowed to converge at any rate, which can be arbitrarily slow. What is more, the coefficient can converge to *any* finite value and, thus, the treatment coefficient itself need not be interpretable as the ATE in order for the GR estimator to obtain $O_p(n^{-0.5})$ consistency.

An important implication of the result is the GR estimator is consistent for the ATE even when its associated coefficient is not. One way to look at it then, is that using GR guards against errors that can lead to a coefficient that is not consistent.

Another important aspect of the result is that GR allows for coefficients that would otherwise not be conceived. Later, in Section 4, I propose two such coefficients, applicable to any design, that allow the conjugate GR estimator to obtain the asymptotic minimum variance. I refer to this as optimal covariate adjustment for arbitrary designs.

Additionally, I give the conditions that determine whether a regression/design combination will allow for interpretation of the treatment coefficient. This is a useful result in the sense of providing a reference for analysts engaging in common regression practice. One implication of the result is that the WLS coefficient that uses $\boldsymbol{\pi}^{-1}$-weights is generally consistent regardless of covariate specification and design.

2.1 Covariate specifications

To discuss covariate specifications, it helps to start with a well known example of the OLS coefficient estimator. First, let $\mathbf{x}$ be the zero-centered matrix of covariates with n rows and k columns. The covariates themselves are taken as constants that are not affected by treatment. Now consider the analysis practice of regressing observed outcomes on separate intercepts for each treatment arm and on $\mathbf{x}$ using OLS. In the present notation the OLS coefficient estimator can be written

$$\widehat{b}_{\mathrm{I}}^{ols} := \left(\mathbb{x}_{\mathrm{I}}'\mathbf{R}\mathbb{x}_{\mathrm{I}}\right)^{-1}\mathbb{x}_{\mathrm{I}}'\mathbf{R}y \tag{9}$$

where

$$\mathbb{x}_{\mathrm{I}} := \begin{bmatrix} -1_n & 0_n & -\mathbf{x} \\ 0_n & 1_n & \mathbf{x} \end{bmatrix}$$

is a $2n \times (k+2)$ matrix. Note that (9) is algebraically equivalent to the canonical OLS formulation that typically writes the estimator in terms of an $n \times (k+2)$ covariate matrix and a vector of observed outcomes that has length n. By contrast, the present formulation has the advantage that it separately represents the source of randomness ($\mathbf{R}$) and the fixed quantities ($\mathbb{x}_{\mathrm{I}}$ and y). Note that for convenience in later derivations, the leading column of $\mathbb{x}_{\mathrm{I}}$ is an intercept (constant) associated with the control group and the second column is an intercept associated with the treatment group. With only a slight change in interpretation of the coefficients, this could have instead been specified as a constant and a treatment indicator. Also, note that elements in the first n rows of $\mathbb{x}_{\mathrm{I}}$ are multiplied by -1, to mirror the definition of the vector y and thus ensuring that the elements of $\widehat{b}_{\mathrm{I}}^{ols}$ have the expected signs. The subscript on matrix $\mathbb{x}_{\mathrm{I}}$ is given to distinguish it from an alternative specification given below, and, in later derivations, in the absence of such a subscript, $\mathbb{x}$ will be taken to mean that it could be any specification. See also that $\widehat{b}_{\mathrm{I}}^{ols}$ shares the subscript indicating the particular specification of $\mathbb{x}$. The given specification will be referred to "specification I" or alternatively the "common slopes" specification.

Inspired by Lin (2013), "specification II" (or the "separate slopes" specification) is given by,

$$\mathbb{x}_{\mathrm{II}} := \begin{bmatrix} -1_n & -\mathbf{x} & 0_n & 0_{n\times k} \\ 0_n & 0_{n\times k} & 1_n & \mathbf{x} \end{bmatrix}$$

which is equivalent to including interactions between treatment and each covariate in $\mathbf{x}$. Lin (2013) proposes this specification as a remedy to Freedman's (2008a,b) critique that for completely randomized designs OLS with specification I can in some cases hurt asymptotic precision. Note that, as in specification I above, there is an intercept for each treatment arm, rather than a specifying a common intercept and a treatment indicator. Again, this convention simplifies some exposition below. It does not affect the properties of estimators discussed.

It has yet to be said just what coefficient estimators, such as $\widehat{b}_{\text{I}}^{ols}$ in (9), estimate. For the time being suffice it to say that researchers will often interpret the difference between intercept coefficients in $\widehat{b}_{\text{I}}^{ols}$ as an estimate of the ATE, i.e., $[-1 \;\; 1 \;\; 0_k' \,]\, \widehat{b}_{\text{I}}^{ols}$ is often taken to be the ATE estimator. However, a generalized regression estimator can be defined which broadens the class of regression estimators to include those with coefficients that may not be directly interpretable in this fashion. Expanding the range of possible regression coefficient estimators which are not directly interpretable will allow for the derivation of certain optimal estimators of the ATE that are not otherwise obvious (see Section 4).

2.2 Defining a class of generalized regression estimators

Three equivalent forms for the proposed class of generalized regression (GR) estimators are given in the definition below. Sampling theorists have the longest history with GR and the constructions that follow.[1] Doubly robust (point) estimators have the same form, though DR variance estimation assumes away joint probabilities of inclusion.[2] The literature on control functions has apparently reinvented the GR estimator as well.

Definition 2.1 (Generalized regression (GR) estimators)**.** *Three equivalent forms for "generalized regression estimators" of the average treatment effect (ATE) are given by*

$$\widehat{\delta}^{GR}(\widehat{b}) := n^{-1}1_{2n}'(\boldsymbol{\pi}^{-1}\mathbf{R}y - \boldsymbol{\pi}^{-1}\mathbf{R}\mathbb{x}\widehat{b} + \mathbb{x}\widehat{b}) \tag{10a}$$

$$= \widehat{\delta}^{HT} - \widehat{\delta}_{\mathbb{x}}^{HT}\widehat{b} \tag{10b}$$

$$= \widehat{\delta}_{\widehat{u}}^{HT} + n^{-1}1_{2n}'\mathbb{x}\widehat{b} \tag{10c}$$

where $\widehat{\delta}^{HT}$ is the HT estimator of the ATE, $\widehat{\delta}_{\mathbb{x}}^{HT} := n^{-1}1_{2n}'\left(\boldsymbol{\pi}^{-1}\mathbf{R} - \mathbf{i}_{2n}\right)\mathbb{x}$ (with identity matrix $\mathbf{i}_{2n}$) is a $(2n \times 2n)$ is a zero-centered vector of HT estimators of the column sums of $\mathbb{x}$ divided by n, and $\widehat{\delta}_{\widehat{u}}^{HT} := n^{-1}1_{2n}'\boldsymbol{\pi}^{-1}\mathbf{R}\widehat{u}$ is an (approximate) HT estimator of mean difference of residuals, $\widehat{u} := y - \mathbb{x}\widehat{b}$. Vector $\widehat{b}$ is an arbitrary coefficient estimator. Specific estimators in this class are distinguished by the particular $\widehat{b}$.

Each of the three forms of the GR estimator in (10) is useful at different times in subsequent sections. Form (10a) is the disaggregated form, explicitly showing three terms inside the parentheses. Collapsing the second and third term in (10a) leads to (10b), showing that the GR estimator can be written as a covariate adjusted HT estimator where the adjustment term is, itself, the product of another (zero-centered)

[1] The approach herein differs from the "GREG" estimator in the sampling literature in some key ways, however. First, their results were derived for the sampling setting rather than the causal inference context. Second, that literature has tended to focus on obtaining $\widehat{b}$ coefficients that are optimal under a model. This paper is fully design-based, so asymptotic optimality is considered from the design-based perspective. Third, the $\widehat{b}$ coefficients considered in the GREG literature has been typically limited to the class $\widehat{b}^{greg} = \left(\tilde{\mathbf{x}}\mathbf{m}\mathbf{R}\tilde{\mathbf{x}}\right)^{-1}\tilde{\mathbf{x}}\mathbf{m}\mathbf{R}y$ where $\mathbf{m}$ is a diagonal matrix with the i,i entry involving π_i (similar to WLS with $\boldsymbol{\pi}^{-1}$ weights) and often an estimate of (model) error variance. By contrast, this paper will propose estimators that have a somewhat different form in order to achieve asymptotic optimality in the design-based framework.

[2] There are three reasons not to refer to the GR estimator as "doubly robust", even though the latter term may be better known. First, the latter term was preceded by the term "generalized regression estimator", first coined by the sampling theorists some years before. Moreover, unlike doubly robust estimation which were fashioned for observational studies, in the current framework $\boldsymbol{\pi}$ is given by the design. Hence, the estimator is not "doubly robust" conditional on getting one or another set of modeling assumptions is correct; on the contrary, one could say that it is simply "robust" because the treatment assignment probabilities are given and thus correct by design. Moreover, variance expressions in the doubly robust literature do not account for joint assignment probabilities, and hence, are not useful in the current framework. By contrast, variance expressions derived here lead to asymptotically optimal estimators that would not be conceived of in a tradition that assumes away the essential role that joint assignment probabilities play in variance.

HT estimator multiplied by the regression coefficient. Collapsing the first and second term in (10a) leads to (10c), which gives the intuition that we might consider the estimator an average of predictions from the regression plus an (approximate) HT estimator for residuals.

Writing $\widehat{\delta}^{GR}(\widehat{b})$ is an abuse of notation, and, for example, it might be more exact to write $g(\widehat{\delta}^{HT}, \widehat{\delta}^{HT}_{\mathbb{x}}, \widehat{b})$, with function $g(a_1, a_2, a_3) := a_1 - a_2 a_3$. However, the notation $\widehat{\delta}^{GR}(\widehat{b})$ emphasizes two points. First, the appearance of $\widehat{\delta}$ emphasizes that this is an estimator of the ATE. Second, the coefficient estimator in parentheses emphasizes that particular members of the GR estimator class are specified by corresponding definitions of $\widehat{b}$.

Throughout, a superscript will be added to $\widehat{b}$ to signify a particular estimation method and a subscript will indicate covariate specifications. For example, $\widehat{b}^{ols}_{\text{I}}$ would be the "common slopes" OLS regression as given in (9) and the corresponding ATE estimator would be denoted $\widehat{\delta}^{GR}(\widehat{b}^{ols}_{\text{I}})$. Pairs such as $\widehat{b}^{ols}_{\text{I}}$, $\widehat{\delta}^{GR}(\widehat{b}^{ols}_{\text{I}})$ are referred to as "conjugates". Each unique $\widehat{b}$ corresponds to a conjugate ATE estimator in the class of GR estimators.

While the common use of regression involves interpreting the difference in intercept coefficients in $\widehat{b}$ as the ATE, the class of GR estimators is broader. As such, one way to look at the utility of the GR estimator is that it prescribes a general method of obtaining valid estimates of the ATE from a broader array of design-specification-coefficient combinations. For example, consider the OLS coefficient, $\widehat{b}^{ols}_{\text{I}}$. The difference of its intercept terms will not be consistent for the ATE in the case unequal probabilities of treatment assignment. By contrast, the conjugate GR estimator associated with the OLS coefficient, $\widehat{\delta}^{GR}(\widehat{b}^{ols}_{\text{I}})$, is consistent for arbitrary designs (under regularity conditions). Additionally, the GR class will allow for the derivation of coefficients with asymptotically optimal conjugates that would not otherwise be obtained (see Section 4).

In the next subsection, a general condition whereby the difference in intercept coefficients in $\widehat{b}$ will be algebraically equivalent to its conjugate ATE estimator, and hence directly interpretable, is given. This will show that the class of GR estimators subsumes common regression practice of interpreting the difference in intercept coefficients as the ATE. Moreover, the result will lead to a few insights that may be familiar, but which all follow nicely from the one theorem.

2.3 Variance of the GR estimator when $\widehat{b}$ is fixed

An exact expression for the variance of the GR estimator is straightforward when $\widehat{b}$ is a fixed vector of constants, call it b^f.[3] In theory, a researcher might obtain this fixed vector through examination of an auxiliary data set, or by way of conjecture, insight or divination. In practice, researchers will likely estimate coefficient values from the data at hand. Nonetheless, the variance of $\widehat{\delta}^{GR}(b^f)$ (the conjugate of the fixed coefficient b^f) is useful to consider for the following reasons. First, the variance expression for $\widehat{\delta}^{GR}(b^f)$ will help to establish the asymptotic variance expression for GR estimators (see section 2.4). Second, a value of b^f that is finite sample optimal is a quantity that a coefficient estimator might target to obtain *asymptotic* optimality (see section 4).

Definition 2.2 (Fixed-coefficient GR estimators)**.** *"Fixed-coefficient GR estimators" are in a subclass of GR estimators defined in* (10) *where* $\widehat{b} = b^f$ *and* $b^f \in \mathbb{R}^l$ *with* $l = k + 2$ *or* $l = 2k + 2$ *for specifications I and II, respectively.*

Lemma 2.3. *The finite sample variance of the fixed-coefficient GR estimator,* $\widehat{\delta}^{GR}(b^f)$*, with conjugate* b^f *being a fixed constant, is*

$$\mathrm{V}\left(\widehat{\delta}^{GR}(b^f)\right) = n^{-2} u' \mathbf{d} u \tag{11}$$

where $u := y - \mathbb{x} b^f$*.*

Proof. To see the result, start with the third form of the GR estimator given in (10c). Note that when $\widehat{b} = b^f$, a constant vector, the second term in (10c) is a constant. The first term is recognizable as a HT

[3] In sampling theory, when the b coefficients are fixed constants the corresponding estimator is called a "difference estimator".

estimator for the mean of vector $u := y - \mathbb{x}b^f$ (i.e., the residual vector), and note that u is fixed, not random, because b^f is not a function of the sample. Hence, the exact variance is constructed as in equation (7) but with u in place of y. □

2.4 HT estimators are central to the asymptotic behavior of GR

In this section conditions for GR estimators to be asymptotically unbiased, consistent, and asymptotically normal are given. The conditions given are somewhat high-level because greater specificity is difficult without first limiting asymptotic analysis to a particular design (e.g., complete randomization, cluster randomization, block randomization, etc.) and perhaps being more specific about the class of coefficient estimators (e.g., OLS, WLS, etc.).

That said, an important conclusion in this subsection is that, asymptotically speaking, a key consideration is whether a sequence of designs and finite populations are such that HT estimators converge at the asymptomatic rate and are asymptotically normal. If they are, then the coefficient, $\widehat{b}$, need only converge at some arbitrary rate.

Conditions:

1. (Convergence of HT estimators) Positive l_l, l_u exist such that, for all n, $l_l \leq nc'\mathrm{V}(\widehat{\delta}_{\mathbb{z}}^{HT})c \leq l_u$ where $\mathbb{z} = [y \;\; \mathbb{x}]$ and $|c| = 1$
2. (Convergence of $\widehat{b}$) $\widehat{b} - b = O_p(n^{-r})$ for some $r > 0$
3. (Multivariate normal HT estimators) $\left[\mathrm{V}(\widehat{\delta}_{\mathbb{z}}^{HT})\right]^{-0.5}(\widehat{\delta}_{\mathbb{z}}^{HT} - \delta_{\mathbb{z}})' \xrightarrow{d} N(0, \mathbf{i})$ where $\mathbb{z} = [y \;\; \mathbb{x}]$

Theorem 2.4. *Under Conditions 1-2, $\sqrt{n}(\widehat{\delta}^{GR}(\widehat{b}) - \delta)$ has limiting variance*

$$\lim_{n\to\infty} n^{-1}u'\mathbf{d}u \tag{12}$$

where $u := y - \mathbb{x}b$. Moreover, with the addition of Assumption 3,

$$n(u'\mathbf{d}u)^{-0.5}(\widehat{\delta}^{GR}(\widehat{b}) - \delta) \xrightarrow{d} N(0,1). \tag{13}$$

Proof. Starting with the form for the GR estimator given in (10b) and using Assumptions 1 and 2

$$\begin{aligned}\widehat{\delta}^{GR}(\widehat{b}) =& \widehat{\delta}^{HT} - \widehat{\delta}_{\mathbb{x}}^{HT}\widehat{b} \\ =& \widehat{\delta}^{HT} - \widehat{\delta}_{\mathbb{x}}^{HT}b - \widehat{\delta}_{\mathbb{x}}^{HT}(\widehat{b} - b) \\ =& \widehat{\delta}^{HT} - \widehat{\delta}_{\mathbb{x}}^{HT}b + O_p(n^{-0.5-r})\end{aligned}$$

Moreover, b is a fixed (limit) value so that, by Lemma 2.3, $\mathrm{V}\left(\widehat{\delta}^{HT} - \widehat{\delta}_{\mathbb{x}}^{HT}b\right) = n^{-2}u'\mathbf{d}u$ for all n. Expression (12) follows. Next, it follows from Condition 3 that $\sqrt{n}(\widehat{\delta}^{HT} - \widehat{\delta}_{\mathbb{x}}^{HT}b)$ has limiting normal distribution so that, by also using the variance expression in (12), (13) follows. □

2.5 A condition for interpreting the treatment coefficient as the ATE

The following theorem shows that, for commonly used design-specification-coefficient combinations, the difference in intercept coefficients is algebraically equivalent to the GR estimator, thus explaining in one sense equation (10) "generalizes" regression.

The main point of the following theorem is to establish conditions under which the first term in (10c) will be equal to zero for all possible randomizations.

Theorem 2.5. *Let* $\mathbf{m}$ *be any symmetric, positive definite* $2n\times 2n$ *matrix and* $\widehat{b}^{\mathbf{m}} = \left(\mathbb{x}'\mathbf{R}'\mathbf{m}^{-1}\mathbf{R}\mathbb{x}\right)^{-1}\mathbb{x}'\mathbf{R}'\mathbf{m}^{-1}\mathbf{R}y$ *(a class which encompasses GLS, WLS and OLS), then the conjugate GR estimator,* $\widehat{\delta}^{GR}(\widehat{b}^{\mathbf{m}})$*, is algebraically equivalent to* $n^{-1}1'_{2n}\mathbb{x}\widehat{b}^{\mathbf{m}}$ *if* $\exists z$ *such that*

$$\mathbf{R}\mathbb{x}z = (\mathbf{R}\mathbf{m}^{-1}\mathbf{R})^{(-)}\boldsymbol{\pi}^{-1}\mathbf{R}1_{2n} \tag{14}$$

where z *is some vector of constants that combines the* x*'s, and* $(.)^{(-)}$ *is the Moore-Penrose generalized inverse.*

Proof. First note that to prove the theorem, we need to show that the above condition implies that the first term in (10c) equals zero, i.e., that

$$\left(y - \mathbb{x}\widehat{b}^{\mathbf{m}}\right)'\boldsymbol{\pi}^{-1}\mathbf{R}1_{2n} = 0. \tag{15}$$

To see when the equality in (15) will hold, first note that from the definition of $\widehat{b}^{\mathbf{m}}$ given in the theorem

$$\left(y - \mathbb{x}\widehat{b}^{\mathbf{m}}\right)'\mathbf{R}'\mathbf{m}^{-1}\mathbf{R}\mathbb{x} = 0.$$

Therefore, it must also be the case that

$$\left(y - \mathbb{x}\widehat{b}^{\mathbf{m}}\right)'\mathbf{R}'\mathbf{m}^{-1}\mathbf{R}\mathbb{x}z = 0 \tag{16}$$

for any vector $z \in \mathbb{R}^{k+2}$.

Comparing the condition in (15) to the equality in (16), we can see that the condition in (15) is satisfied if

$$\mathbf{R}\mathbf{m}^{-1}\mathbf{R}\mathbb{x}z = \boldsymbol{\pi}^{-1}\mathbf{R}1_{2n}$$

holds for some value of z. Equivalently this is satisfied when there exits a z such that

$$\mathbf{R}\mathbb{x}z = (\mathbf{R}\mathbf{m}^{-1}\mathbf{R})^{(-)}\boldsymbol{\pi}^{-1}\mathbf{R}1_{2n},$$

completing the proof. □

Remark 4 (Algebraic equivalences for OLS in equal-π designs)**.** *For any identified design with equal* π_{1i} *for all* i *(such as a completely randomized design) when using OLS (i.e.,* $\mathbf{m}^{-1}$ *is an identity matrix), the condition in Theorem 2.5 reduces to* $\mathbf{R}\mathbb{x}z = \boldsymbol{\pi}^{-1}\mathbf{R}1_{2n}$*. This is trivially satisfied in specifications with an intercept for each treatment arm (such as specification I and specification II) and for equivalent specifications (such as a common intercept with a treatment indicator). For specification* I*, this means that the GR estimator* $\widehat{\delta}^{GR}(\widehat{b}^{ols}_{\text{I}})$ *is algebraically equivalent to the difference in intercept terms its conjugate coefficient estimator,* $\widehat{b}^{ols}_{\text{I}}$*. For specification* II *this means that, if the columns of* $\mathbf{x}$ *have mean zero, then the GR estimator* $\widehat{\delta}^{GR}(\widehat{b}^{ols}_{\text{II}})$ *is algebraically equivalent to the difference of intercept terms in its conjugate coefficient estimator,* $\widehat{b}^{ols}_{\text{II}}$*.*

Remark 5 (Algebraic equivalences for WLS with $\boldsymbol{\pi}^{-1}$ weights)**.** *For any identified design when using WLS with* $\boldsymbol{\pi}^{-1}$ *weights (i.e., when* $\mathbf{m}^{-1} = \boldsymbol{\pi}^{-1}$*) the condition in Theorem 2.5 reduces to* $\mathbf{R}\mathbb{x}z = \mathbf{R}1_{2n}$*. This is trivially satisfied in specifications with a separate intercept for each treatment arm (such as specification* I *and specification* II*) and for equivalent specifications (such as a common intercept with a treatment indicator). For specification* I *this means that the GR estimator* $\widehat{\delta}^{GR}(\widehat{b}^{\pi wls}_{\text{I}})$ *is algebraically equivalent to the difference in intercept terms in the conjugate coefficient estimator,* $\widehat{b}^{\pi wls}_{\text{I}}$*. For specification* II *this means that, if the columns of* $\mathbf{x}$ *have mean zero, then the GR estimator* $\widehat{\delta}^{GR}(\widehat{b}^{\pi wls}_{\text{II}})$ *is algebraically equivalent to the difference in intercept terms its conjugate coefficient estimator,* $\widehat{b}^{\pi wls}_{\text{II}}$*.*

Corollary 2.5.1. *One can ensure that the condition in Theorem 2.5 holds by including a vector, v, in $\mathbb{X}$ that satisfies $\mathbf{R}v = (\mathbf{R}\mathbf{m}^{-1}\mathbf{R})^{(-)}\boldsymbol{\pi}^{-1}\mathbf{R}1$.*

Remark 6. *Corollary 2.5.1 shows that for any identified design the condition in Theorem 2.5 is satisfied for OLS when the reciprocal of probability of assignment, $\boldsymbol{\pi}^{-1}1_{2n}$, is included as a covariate in $\mathbb{X}$. Moreover, including a zero-centered version of $\boldsymbol{\pi}^{-1}1_{2n}$ in $\mathbb{X}$ would allow for interpretation of the difference in intercept terms. However, for designs where some assignment probabilities are near zero, this approach may result in high leverage for corresponding units. In such cases inference could be problematic.*

Remark 7. *Note that in spite of Corollary 2.5.1, $(\mathbf{R}\mathbf{m}^{-1}\mathbf{R})^{(-)}\boldsymbol{\pi}^{-1}\mathbf{R}1$ could effectively be a random variable if $\mathbf{m}$ is not a diagonal matrix. Thus adding it to the matrix $\mathbb{X}$ could have unexpected consequences for variance and bias of the ATE estimator.*

Remarks 4-6 show that in some cases it is possible to directly interpret the difference in intercept coefficients as an estimated ATE. These are special cases of the GR estimator in Definition 2.1, and these relationships reveal in what sense GR "generalizes" (i.e., subsumes) regression approaches in common use. The generalization, however, will allow for a wider variety of possible design-specification-coefficient combinations even when the coefficient itself is not readily interpretable. As mentioned previously, a novel possibility includes plugging the OLS coefficient into the GR estimator irrespective of the design and $\mathbb{X}$ specification. Likewise, the form of the GR estimator allows for the definition of coefficients whose conjugate ATE estimators are consistent and obtain asymptotic minimum variance (see Section 4). These estimators would not be derived in a framework that required coefficients themselves to be interpretable.

3 Variance bounds and their estimation

In general, the variance expressions of the form (7) and (11) are not identified. This is due to the fact that some pairs of elements in the vector y can never be jointly observed, and hence, for example, some terms in the quadratic $n^{-2}y'\mathbf{d}y$ are never observable. One reason is that a given unit's potential outcomes, y_{0i} and y_{1i}, can never be observed together. This problem is referred to as the "fundamental problem of causal inference" (Holland, 1986). But other design features, such as clustering or pair randomization, render various combinations of potential outcomes jointly unobservable as well.

Starting with Neyman (1923) one proposed solution to unidentified variance has been to estimate a *variance bound*, i.e., a quantity that is known to be greater than the variance, but which is identified. "Variance estimation" is thus actually "variance bound estimation". It should be understood that estimating variance *bounds* is not a unique practice, though the more precise phrase may be less familiar.

3.1 Variance bounds defined

Definition 3.1 (Variance bound)**.** *For an arbitrary $2n \times 2n$ matrix $\tilde{\mathbf{d}}$, let $n^{-2}y'\tilde{\mathbf{d}}y$ be a "bound" for the variance $n^{-2}y'\mathbf{d}y$ if, for all $y \in \mathbb{R}^{2n}$, $n^{-2}y'\mathbf{d}y \leq n^{-2}y'\tilde{\mathbf{d}}y$.*

Lemma 3.2. *$n^{-2}y'\tilde{\mathbf{d}}y$ is a bound for the variance $n^{-2}y'\mathbf{d}y$ if and only if matrix $\tilde{\mathbf{d}}-\mathbf{d}$ is positive semi-definite.*

Proof. By the definition of a bound, $n^{-2}y'\tilde{\mathbf{d}}y - n^{-2}y'\mathbf{d}y \geq 0$ for all $y \in \mathbb{R}^n$. This implies that $n^{-2}y'(\tilde{\mathbf{d}}-\mathbf{d})y \geq 0$, i.e., that $\tilde{\mathbf{d}} - \mathbf{d}$ is positive semi-definite. □

Definition 3.3 (Identified variance bound)**.** *For an arbitrarily defined $2n \times 2n$ matrix $\tilde{\mathbf{d}}$, let $n^{-2}y'\tilde{\mathbf{d}}y$ be an "identified variance bound" for $n^{-2}y'\mathbf{d}y$ if it is a variance bound and 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 $2n \times 2n$ 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 $\mathrm{I}(\mathbf{d} = 0)$ is, likewise, an indicator function returning an $2n \times 2n$ matrix of ones and zeros indicating the location of zeros in $\tilde{\mathbf{d}}$.

The definition says that for a variance bound $n^{-2}y'\tilde{\mathbf{d}}y$ to be an identified bound the elements of matrix $\mathbf{d}$ equal to -1 must correspond to elements of $\tilde{\mathbf{d}}$ that equal 0.

3.2 Variance estimation for HT and GR estimators

Next, define

$$\mathbf{p} := \mathrm{E}\left[\mathbf{R}1_{2n}1'_{2n}\mathbf{R}\right],$$

which is the matrix of joint probabilities of assignment with the diagonal representing probabilities of assignment, i.e., $\mathrm{diag}\left(\boldsymbol{\pi}\right) = \mathrm{diag}\left(\mathbf{p}\right)$, and define

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

with / denoting element-wise division defined such that division by zero equals zero. Then

$$\widehat{\tilde{\mathrm{V}}}\left(\widehat{\delta}^{HT}\right) := n^{-2}y'\mathbf{R}\tilde{\mathbf{d}}_{/\mathbf{p}}\mathbf{R}y \tag{17}$$

is an unbiased estimator of the variance bound since $\mathrm{E}\left[\mathbf{R}\tilde{\mathbf{d}}_{/\mathbf{p}}\mathbf{R}\right] = \tilde{\mathbf{d}}$ by construction.

For GR, by analogy to equation (17) one can motivate the variance bound estimator as

$$\widehat{\tilde{\mathrm{V}}}\left(\widehat{\delta}^{GR}(\widehat{b})\right) :=n^{-2}\widehat{u}'\mathbf{R}\tilde{\mathbf{d}}_{/\mathbf{p}}\mathbf{R}\widehat{u} \tag{18}$$

where $\widehat{u} = y - \widehat{\mathbf{x}b}$, a result that is easily obtained using Taylor's theorem, and which relies on asymptotic arguments analogous to those invoked in the case of variance estimation in common regression. Refinements might involve adjustments motivated by concerns with small sample bias, such as leverage-type adjustments.

3.3 Generalizing the Neyman bound

This section proposes a generalization of Neyman's (1923) variance bound that can be defined for experiments with $\mathrm{I}(\mathbf{d}_{00} = -1) = \mathrm{I}(\mathbf{d}_{11} = -1) = 0_{n\times n}$ and $\mathbf{d}_{01} = \mathbf{d}_{10}$, covering a variety of designs.

Definition 3.4 (Generalized Neyman variance bound)**.** *The "Generalized Neyman variance bound" is*

$$\tilde{\mathrm{V}}^{N}\left(\widehat{\delta}^{HT}\right) := n^{-2}y'\tilde{\mathbf{d}}^{N}y \tag{19}$$

where

$$\tilde{\mathbf{d}}^{N} :=\mathbf{d} - \begin{bmatrix} \mathbf{d}_{01} & \mathbf{d}_{01} \\ \mathbf{d}_{01} & \mathbf{d}_{01} \end{bmatrix}.$$

[Equivalence proof]
[Proof of bound]

3.4 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 3.5 (Aronow-Samii variance bound)**.** *The "Aronow-Samii variance bound" is*

$$\tilde{\mathrm{V}}^{AS}\left(\widehat{\delta}^{HT}\right) := n^{-2}y'\tilde{\mathbf{d}}^{AS}y \tag{20}$$

where

$$\tilde{\mathbf{d}}^{AS} := \mathbf{d} + \mathrm{I}\left(\mathbf{d} = -1\right) + \mathrm{diag}\left(\mathrm{I}\left(\mathbf{d} = -1\right)1_{2n}\right)$$

and diag(.) *creates a diagonal matrix from a vector.*

Theorem 3.6. *The Aronow-Samii variance bound, $n^{-2}y'\tilde{\mathbf{d}}^{AS}y$, is an identified bound for $n^{-2}y'\mathbf{d}y$.*

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

$$\tilde{\mathbf{d}}^{AS} - \mathbf{d} = \mathrm{I}\left(\mathbf{d} = -1\right) + \mathrm{diag}\left(\mathrm{I}\left(\mathbf{d} = -1\right) 1_{2n}\right).$$

Note that by construction $(\tilde{\mathbf{d}}^{AS} - \mathbf{d})$ 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 $\tilde{\mathbf{d}}^{AS} - \mathbf{d}$ is positive semidefinite. Therefore, by Lemma (3.2), $n^{-2}y'\tilde{\mathbf{d}}^{AS}y$ 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 $\tilde{\mathbf{d}}^{AS}$. □

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

3.5 A proposed algorithm for finding a variance bound 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 virtually universally. The drawback is the potential computationally difficulty.

Algorithm 3.7.

1. Initialize $\mathbf{t}$. Examples could be $\mathrm{I}(\mathbf{d} = -1)$ or, if the conditions for the Neyman bound not be applicable, start with $\tilde{\mathbf{d}}^{N} - \mathbf{d}$ 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 $\tilde{\mathbf{d}}^{M} = \mathbf{d} + \mathbf{t}$

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 3.2 and Definition 3.3, 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, $\tilde{\mathbf{d}}^{M}$ corresponds to an identified bound.

3.6 Generalizing the Eicker–Huber–White variance bound estimator for regression coefficients

In this section I generalize the Eicker-Huber-White variance bound estimator for regression coefficients. Applications of the HC and CR versions are somewhat limited to designs where units of randomization independently assigned (or approximately so), or, alternatively, when assignment is (approximately) independent across clusters but completely coincident among units within a cluster. The generalization proposed here permits their use for virtually any design and also generalizes to allows for a variety of different bounding methods, three examples of which are given above. The section will conclude by demonstrating an algebraic equivalences.

For clarity, first write a regression coefficient in the form $\widehat{b}^{\mathbf{C}} = \mathbf{C}\boldsymbol{\pi}^{-1}\mathbf{R}y$. For example, if $\mathbf{C} = \left(\mathbb{x}'\boldsymbol{\pi}^{-1}\mathbf{R}\mathbb{x}\right)^{-1}\mathbb{x}'$, then $\widehat{b}^{\mathbf{C}}$ is $\boldsymbol{\pi}^{-1}$-weighted WLS. Alternatively, if $\mathbf{C} = \left(\mathbb{x}'\mathbf{R}\mathbb{x}\right)^{-1}\mathbb{x}'\boldsymbol{\pi}$ then the regression is OLS.

Definition 3.8 (The Generalized Eicker-Huber-White Sandwich variance bound estimator for regression coefficients)**.** *The "Generalized Eicker-Huber-White Sandwich variance bound estimator" is*

$$\widehat{\tilde{\mathrm{V}}}^{EHW}\left(\widehat{b}^{\mathbf{C}}\right) := \mathbf{C}\mathrm{diag}\left(\widehat{u}\mathbf{R}\right)\tilde{\mathbf{d}}_{\mathbb{p}}\mathrm{diag}\left(\mathbf{R}\widehat{u}\right)\mathbf{C}' \tag{21}$$

with $\widehat{u} = y - \mathbb{x}\widehat{b}^{\mathbf{C}}$.

To see the equivalence with the original formulation, assume a Bernoulli design whereby units are assigned to treatment independently with probabilities that need not be equal across units. Next note that in a Bernoulli design the diagonal elements of $\mathbf{d}$ are equal to the diagonal of $\boldsymbol{\pi}^{-1} - \mathbf{i}$ and that any of the above bounding methods will yield $\tilde{\mathbf{d}} = \tilde{\mathbf{d}}^{N} = \tilde{\mathbf{d}}^{AS} = \tilde{\mathbf{d}}^{M} = \boldsymbol{\pi}^{-1} - \mathbf{i} + \mathbf{i} = \boldsymbol{\pi}^{-1}$. Thus $\tilde{\mathbf{d}}_{\mathbb{p}} = \boldsymbol{\pi}^{-2}$. Then with the OLS coefficient, which corresponds to $\mathbf{C} = \left(\mathbb{x}'\mathbf{R}\mathbb{x}\right)^{-1}\mathbb{x}'\boldsymbol{\pi}$, the generalized EHW variance estimator becomes

$$\begin{aligned}
\widehat{\tilde{\mathrm{V}}}^{EHW}\left(\widehat{b}^{ols}\right) &= \left(\mathbb{x}'\mathbf{R}\mathbb{x}\right)^{-1}\mathbb{x}'\boldsymbol{\pi}\mathrm{diag}\left(\widehat{u}\mathbf{R}\right)\tilde{\mathbf{d}}_{\mathbb{p}}\mathrm{diag}\left(\mathbf{R}\widehat{u}\right)\boldsymbol{\pi}\mathbb{x}\left(\mathbb{x}'\mathbf{R}\mathbb{x}\right)^{-1} \\
&= \left(\mathbb{x}'\mathbf{R}\mathbb{x}\right)^{-1}\mathbb{x}'\mathrm{diag}\left(\widehat{u}\mathbf{R}\right)\left(\boldsymbol{\pi}\tilde{\mathbf{d}}_{\mathbb{p}}\boldsymbol{\pi}\right)\mathrm{diag}\left(\mathbf{R}\widehat{u}\right)\mathbb{x}\left(\mathbb{x}'\mathbf{R}\mathbb{x}\right)^{-1} \\
&= \left(\mathbb{x}'\mathbf{R}\mathbb{x}\right)^{-1}\mathbb{x}'\mathrm{diag}\left(\widehat{u}^2\mathbf{R}\right)\mathbb{x}\left(\mathbb{x}'\mathbf{R}\mathbb{x}\right)^{-1},
\end{aligned}$$

where in the last line $\left(\boldsymbol{\pi}\tilde{\mathbf{d}}_{\mathbb{p}}\boldsymbol{\pi}\right)$ collapses to an identity matrix, given the Bernoulli design, and $\mathrm{diag}\left(\widehat{u}^2\mathbf{R}\right)$ now represents a diagonal matrix of squared residuals. The form is recognizable as the canonical HC0.

Note also, that no special adjustments are necessary to obtain an algebraic equivalence with CR0 for designs in which clusters are assigned independently to treatment. In that case if we choose the Neyman bound $\tilde{\mathbf{d}}_{\mathbb{p}}^{N}$ or, equivlently in this case, $\tilde{\mathbf{d}}_{\mathbb{p}}^{M}$, then for units sorted by cluster $\left(\boldsymbol{\pi}\tilde{\mathbf{d}}_{\mathbb{p}}^{N}\boldsymbol{\pi}\right)$ resolves to a block diagonal matrix of 1's with the blocks corresponding to clusters. Hence, it is also algebraically equivalent to CR0 in designs where clusters are assigned independently.

Also consider completely randomized designs with Neyman's $\tilde{\mathbf{d}}_{\mathbb{p}}^{N}$ bound. The resulting generalized variance estimator is only different from HC0 in that the squared residuals are multiplied by factors of $\frac{n_1}{n_1-1}$ and $\frac{n_0}{n_0-1}$ for treatment and control outcomes, respectively. With only an intercept and treatment indicator, this is algebraically equivalent to HC2 and also Neyman's (1923) variance bound estimator, an equivalence discussed by Aronow and Samii for the case of complete randomization.

Leverage-type adjustements on exhibit in HC1, HC2 etc, are applicable here as well. In sum, EHW standard errors are now applicable to virtually any design.

It is important to note that the ability to bound the variance of the regression coefficient and estimate the bound should not be taken to imply that a regression coefficient can be meaningfully interpreted or that it is desirable or necessary to apply inferential techniques to regression coefficients. Instead, one can use GR. The common regression practice of interpreting the coefficient itself is acceptable if the conditions of Theorem (2.5) hold. Of course, meeting the condition implies the treatment coefficient is, itself, a GR

estimator, algebraically speaking. Hence, variance estimators for coefficients are not strictly essential for the purposes of causal inference.

3.7 Notes on the convergence of the variance estimator of the HT estimator

Now consider the variance of (n times) the variance estimator:

$$\begin{aligned}
\mathrm{V}\left(n\widehat{\mathrm{V}}\left(\widehat{\delta}^{HT}\right)\right) =& \mathrm{E}\left[\left(n^{-1}y'\mathbf{R}\tilde{\mathbf{d}}_{/\mathbf{p}}\mathbf{R}y - n^{-1}y'\tilde{\mathbf{d}}y\right)^2\right] \\
\leq& n^{-2}\max\left(|y|\right)^4 1'_{4n^2}\left|\left(\tilde{\mathbf{d}}\otimes\tilde{\mathbf{d}}\right)\circ\left(\mathrm{E}\left[(\mathbf{R}1_n1'_n\mathbf{R})\otimes(\mathbf{R}1_n1'_n\mathbf{R})\right]-\mathbf{p}\otimes\mathbf{p}\right)/\left(\mathbf{p}\otimes\mathbf{p}\right)\right|1_{4n^2} \\
=& n^{-2}\max\left(|y|\right)^4\sum_i\sum_j\sum_k\sum_l\left|\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}+t_{ij}\right)\left(\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l}+t_{kl}\right)\left(\frac{\pi_{ijkl}-\pi_{ij}\pi_{kl}}{\pi_{ij}\pi_{kl}}\right)\right|
\end{aligned}$$

where i,j,k,l index from 1 to $2n$ such that, for example, for $i \leq n$ define $\pi_i := \pi_{0i}$ and for $i > n$ we define $\pi_i := \pi_{1(i-n)}$. And, for example, t_{ij} is the i,j element of the matrix $\mathbf{t} := \left(\tilde{\mathbf{d}} - \mathbf{d}\right)$. That this quantity converges can be checked numerically for a hypothetical sequence of designs and populations.

[NOTES]

4 Minimum Variance GR for Arbitrary Designs

Lemma 2.3 gives the finite sample variance of $\widehat{\delta}^{GR}(b^f)$, the conjugate of the fixed regression coefficient, b^f, first introduced in Section 2.3. One might next ask, what value of $b^f \in \mathbb{R}^l$ minimizes the finite sample variance of $\widehat{\delta}^{GR}(b^f)$? That question is answered in subsection 4.1. The answer allows the derivation of coefficient estimators that target optimal values of b^f in subsections 4.2 and 4.3. By arguments in section 2.4, as long as the proposed coefficient estimators converges to the finite sample optimal value and Assumption 1 holds, then its conjugate ATE estimator obtains the asymptotic minimum variance in the class of GR estimators.

4.1 Optimality when $\widehat{b}$ is fixed

In this section, finite-sample optimal values of b^f, the fixed-coefficient introduced in Section 2.3, are derived.

Theorem 4.1. *Letting $(.)^{(-)}$ represent the Moore-Penrose generalized inverse[4], a coefficient value that is finite sample optimal for the fixed-coefficient GR estimator is*

$$b^{opt} := (\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}\mathbb{x}'\mathbf{d}y. \tag{22}$$

Proof. Starting with the finite sample variance of $\widehat{\delta}^{GR}(b^f)$ given in Theorem 2.3, we have

$$\begin{aligned}
n^2\mathrm{V}\left(\widehat{\delta}^{GR}(b^f)\right) =& u'\mathbf{d}u \\
=& (y-\mathbb{x}b^f)'\mathbf{d}(y-\mathbb{x}b^f) \\
=& y'\mathbf{d}y - 2y'\mathbf{d}\mathbb{x}b^f + {b^f}'\mathbb{x}'\mathbf{d}\mathbb{x}b^f
\end{aligned}$$

To minimize, take the derivative with respect to b^f, set equal to zero, and then rearrange to obtain

$$(\mathbb{x}'\mathbf{d}\mathbb{x})b^f = \mathbb{x}'\mathbf{d}y. \tag{23}$$

[4] A generalized inverse of $\mathbf{a}$, $\mathbf{a}^{(g)}$, has the property that $\mathbf{a}\mathbf{a}^{(g)}\mathbf{a} = \mathbf{a}$. When the inverse of $\mathbf{a}$ exists, a generalized inverse corresponds to the usual inverse.

Premultiplying the equality by $(\mathbb{x}'\mathbf{d}\mathbb{x})(\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}$ we have

$$\begin{aligned}(\mathbb{x}'\mathbf{d}\mathbb{x})(\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}(\mathbb{x}'\mathbf{d}\mathbb{x})b^f &= (\mathbb{x}'\mathbf{d}\mathbb{x})(\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}\mathbb{x}'\mathbf{d}y \\ \implies (\mathbb{x}'\mathbf{d}\mathbb{x})b^f &= (\mathbb{x}'\mathbf{d}\mathbb{x})(\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}\mathbb{x}'\mathbf{d}y\end{aligned}$$

where the second line follows from the definition of a generalized inverse. This implies that

$$b^f = (\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}\mathbb{x}'\mathbf{d}y$$

is a solution to (23). □

Remark 8. *Defining in terms of a generalized inverse is not simply to account for a few rare cases where the usual inverse is not applicable. There are an infinite number of optimal b^f in common settings, for example, any equal-π_1 design (such as complete randomization) with specification II.*

Remark 9. *The choice of the Moore-Penrose generalized inverse, in particular, is arbitrary in a statistical sense. On the one hand, in the special case where $(\mathbb{x}'\mathbf{d}\mathbb{x})$ is invertible, all generalized inverses produce the true inverse; in that case, there is a unique b^f vector that minimizes the variance of $\widehat{\delta}^{GR}(b^f)$. On the other hand, when $(\mathbb{x}'\mathbf{d}\mathbb{x})$ is not invertible, different generalized inverses will lead to different coefficients, all of which are optimal in the sense of minimizing the variance of their respective conjugate ATE estimators. There are two key features recommending the Moore-Penrose generalized, however. First, it has the virtue of being commonly implemented in software. Second, in addition to the generalized inverse property ($\mathbf{a}\mathbf{a}^{(-)}\mathbf{a} = \mathbf{a}$), it has the reflexive property ($\mathbf{a}^{(-)}\mathbf{a}\mathbf{a}^{(-)} = \mathbf{a}^{(-)}$) which is useful below.*

Remark 10. *Just as $n^{-2}y'\mathbf{d}y$ gives the variance of HT estimator, so too is $n^{-2}\mathbb{x}'\mathbf{d}\mathbb{x}$ a variance-covariance matrix of HT estimators. An insight is that the optimal coefficient values are determined by the joint distribution of estimated means of x's and y's, rather than the joint distribution of x's and y's. This is a slightly different way of thinking about the job of regression adjustment compared to the intuition that one should attempt to approximate the conditional expectation of y_{1i} (or y_{0i}) given x_i. Instead, one should be more concerned with the conditional expectation of $\widehat{\delta}^{HT}$ given $\widehat{\delta}^{HT}_{\mathbb{x}}$. The former conditional expectation may be well estimated by the latest in machine learning techniques, but, depending on the design, it need not correspond to the latter.*

4.2 A HT estimator of b^{opt}, namely, 3HT!

In this section, a HT estimator of b^{opt}, given in equation (22), is introduced. It has the usual limitations of HT estimators, imprecision and a general lack of invariance to location shifts in y. However, the estimator serves as a conceptual starting point, and the refinement in the next subsection may prove more useful. The coefficient estimator, call it 3HT!, takes its name from the fact that its conjugate is a constellation of three HT estimators, as can be seen by examining form (10b).

Definition 4.2 (The 3HT! optimal coefficient estimator)**.** *The "3HT! optimal coefficient estimator" is*

$$\widehat{b}^{3HT!} := (\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}\,\mathbb{x}'\mathbf{d}\boldsymbol{\pi}^{-1}\mathbf{R}y \tag{24}$$

where $(.)^{(-)}$ is the Moore-Penrose generalized inverse.

Remark 11. *Note that the estimator differs from a GLS-type estimator in a number of ways. First, the "denominator" matrix $(\mathbb{x}'\mathbf{d}\mathbb{x})$ is not random. Likewise, the "numerator" $\mathbb{x}'\mathbf{d}\boldsymbol{\pi}^{-1}\mathbf{R}y$ utilizes the fact that $\mathbb{x}$ is completely observed. Moreover, with GLS the linear model is assumed and the analogue of the $\mathbf{d}$ matrix is designed to minimize the variance of the coefficient vector, which is consistent under the linear model. In the current framework, there is no linear model implied, there are no stochastic errors since potential outcomes are fixed and the $\mathbf{d}$ matrix serves to allow the construction of variance-covariance matrices for HT estimators. Precision of the coefficient itself is not guaranteed. Precision guarantees are asymptotic for the conjugate, $\widehat{\delta}^{GR}(\widehat{b}^{3HT!})$.*

Remark 12. *Again, the use of the Moore-Penrose generalized inverse is for convenience. Fortunately, regardless of the generalized inverse chosen in the construction of $\widehat{b}^{3HT!}$, the conjugate estimators of the ATE are algebraically equivalent.*

Remark 13. *Like any HT estimator, it is unbiased. To see this, simply take the expectation of (24) and recall that* $\mathrm{E}[\mathbf{R}] = \boldsymbol{\pi}$.

For root-n convergence of the 3HT! coefficient, $(\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}\,\mathbb{x}'\mathbf{d}\boldsymbol{\pi}^{-1}\mathbf{R}y$, Lemma **??** shows that we need only the additional restriction that $\max\left(|\mathbf{d}|\,1_{2n}\right)$ be bounded.

Theorem 4.3 (Consistency of the 3HT! coefficient)**.** *As before, assume positive limiting values l_y, $l_{\mathbf{x}}$ and $l_{\mathbf{d}}$ exist such that* $\max(|y|) < l_y < \infty$, $\max\left(|\mathbf{x}|\right) < l_{\mathbf{x}} < \infty$ *and* $n^{-1}1'_{2n}\,|\mathbf{d}|\,1_{2n} < l_{\mathbf{d}} < \infty$. *Additionally, assume that positive $l_{\mathbf{m}}$ exists such that* $\max\left(|\mathbf{d}|\,1_{2n}\right) < l_{\mathbf{m}} < \infty$. *Then by Lemma* **??**, $n^{-1}\mathbb{x}'\mathbf{d}\boldsymbol{\pi}^{-1}\mathbf{R}y \to n^{-1}\mathbb{x}'\mathbf{d}y$ *as* $n \to \infty$, *and* $\widehat{b}^{3HT!} - b^{opt} = O_p(n^{-0.5})$.

Remark 14. *Note that a bound on* $\max\left(|\mathbf{d}|\,1_{2n}\right)$ *immediately implies a bound on* $n^{-1}1'_{2n}\,|\mathbf{d}|\,1_{2n}$.

Remark 15. *The requirement that* $\max\left(|\mathbf{d}|\,1_{2n}\right)$ *be bounded is provably true for completely randomized experiments. In a cluster-randomized experiment, it implies that there must be a limit on the size of the largest cluster.*

4.3 A GR estimator of b^{opt}, namely, 2GR! (-or- Regression adjusted regression adjustment)

Recognizing $\widehat{b}^{3HT!}$ in equation (22) as a HT estimator of the column sums of $\mathbf{b}$ in Lemma **??**, suggests that an improved estimation strategy may be to recursively apply GR adjustment. No new principles are required.

The regression-adjusted regression coefficient will be called $\widehat{b}^{2GR!}$. It takes its name from the fact that its conjugate, $\widehat{\delta}^{GR}(\widehat{b}^{2GR!})$, involves two levels of regression adjustment.

Subsequent to its definition, the invariance of its conjugate, $\widehat{\delta}^{GR}(\widehat{b}^{2GR!})$, will be proven. Its invariance is notable because its constituent parts are not generally invariant.

Definition 4.4 (The 2GR! optimal coefficient estimator)**.** *The "2GR! optimal coefficient estimator" is given by*

$$\begin{aligned}\widehat{b}^{2GR!} :=& \left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(-)}\mathbb{x}'\mathbf{d}\left(\boldsymbol{\pi}^{-1}\mathbf{R}y - \boldsymbol{\pi}^{-1}\mathbf{R}\mathbb{x}\widehat{b}^{\pi wls} + \mathbb{x}\widehat{b}^{\pi wls}\right) \\ =& \widehat{b}^{3HT!} - \left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(-)}\left(\mathbb{x}'\mathbf{d}\boldsymbol{\pi}^{-1}\mathbf{R}\mathbb{x} - \mathbb{x}'\mathbf{d}\mathbb{x}\right)\widehat{b}^{\pi wls}.\end{aligned} \tag{25}$$

where $\widehat{b}^{\pi wls} := (\mathbb{x}'\boldsymbol{\pi}^{-1}\mathbf{R}\mathbb{x})^{-1}\mathbb{x}'\boldsymbol{\pi}^{-1}\mathbf{R}y$ *is WLS with* $\boldsymbol{\pi}^{-1}$ *weights.*

The first line of (25) can be compared to (10a) to make clear that this is a regression-adjusted regression coefficient. The second line will be at the crux of asymptotic arguments: as long as $\widehat{b}^{3HT!} \xrightarrow{p} b^{opt}$, $\widehat{b}^{\pi wls} \xrightarrow{p} b^{\pi wls}$ and $\mathbb{x}\mathbf{d}\boldsymbol{\pi}^{-1}\mathbf{R}\mathbb{x} - \mathbb{x}\mathbf{d}\mathbb{x} \xrightarrow{p} 0$ then $\widehat{b}^{2GR!} \xrightarrow{p} b^{opt}$.

Next, the invariance of the regression estimator will be demonstrated, with the help of the following two lemmas.

Lemma 4.5. *Let* $y^* = e + fy$ *where*

$$e = c\begin{bmatrix}-1_n \\ 1_n\end{bmatrix}$$

and c and f are arbitrary constants, then for any specification with a constant (e.g., specification I) or separate constants for treatment arms (e.g., specification II) the two-step optimal coefficient estimated using y^* instead of y is

$$\widehat{b}^{2GR!*} = f\widehat{b}^{2GR!} + c\left(\mathbb{x}\mathbf{d}\mathbb{x}\right)^{(-)}\mathbb{x}\mathbf{d}\begin{bmatrix}-1_n \\ 1_n\end{bmatrix}.$$

Proof. Provided in Appendix. □

Lemma 4.6. *Let $y_1 = y_0 = 1_n$, then the finite-sample optimal coefficient is $b^{opt} = (\mathbb{x}\mathbf{d}\mathbb{x})^{(-)}\,\mathbb{x}'\mathbf{d}\begin{bmatrix}-1_n\\1_n\end{bmatrix}$ and the conjugate of this fixed value has expectation zero and variance zero.*

Theorem 4.7. *The 2GR! estimator of the ATE, $\widehat{\delta}^{GR}(\widehat{b}^{2GR!})$, is invariant to scale changes in y.*

Proof. As above, let $y^* = e + fy$ where

$$e = c\begin{bmatrix}-1_n\\1_n\end{bmatrix}$$

and c and f are arbitrary constants then

$$\begin{aligned}
\widehat{\delta}^{GR}(\widehat{b}^{2GR!*}) =& n^{-1}1'_{2n}\boldsymbol{\pi}^{-1}\mathbf{R}(y^* - \mathbb{x}\widehat{b}^{2GR!*}) + n^{-1}1'_{2n}\mathbb{x}\widehat{b}^{2GR!*}\\
=& n^{-1}1'_{2n}\boldsymbol{\pi}^{-1}\mathbf{R}\left(f(y - \mathbb{x}\widehat{b}^{2GR!}) + e - \mathbb{x}\,(\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}\,\mathbb{x}'\mathbf{d}e\right) + n^{-1}1'_{2n}\left(f\mathbb{x}\widehat{b}^{2GR!} + \mathbb{x}\,(\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}\,\mathbb{x}'\mathbf{d}e\right)\\
=& f\widehat{\delta}^{GR}(\widehat{b}^{2GR!}) + n^{-1}1'_{2n}\boldsymbol{\pi}^{-1}\mathbf{R}\left(e - \mathbb{x}\,(\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}\,\mathbb{x}'\mathbf{d}e\right) + n^{-1}1'_{2n}\left(\mathbb{x}\,(\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}\,\mathbb{x}'\mathbf{d}e\right)\\
=& f\widehat{\delta}^{GR}(\widehat{b}^{2GR!}).
\end{aligned}$$

The last line follows from Lemma 4.6 □

Theorem 4.8. *The 2GR! estimator of the ATE, $\widehat{\delta}^{GR}(\widehat{b}^{2GR!})$, is invariant to scale changes in $\mathbb{x}$.*

Proof. Let $\mathbf{f}$ be a $(l \times l)$ transformation matrix such that $\mathbf{f}^{-1}$ exists and let $\mathbb{x}^* = \mathbb{x}\mathbf{f}$. Next, write the two-step optimal estimator of the ATE computed with $\mathbb{x}^*$ in place of $\mathbb{x}$ as

$$\widehat{\delta}^{GR}(\widehat{b}^{2GR!*}) = \widehat{\delta}^{HT} - n^{-1}1_{2n}\left(\boldsymbol{\pi}^{-1}\mathbf{R} - \mathbf{i}\right)\mathbb{x}^*\left(\mathbb{x}^{*\prime}\mathbf{d}\mathbb{x}^*\right)^{(-)}\mathbb{x}^{*\prime}\mathbf{d}\left(\boldsymbol{\pi}^{-1}\mathbf{R}y - \left(\boldsymbol{\pi}^{-1}\mathbf{R} - \mathbf{i}\right)\mathbb{x}^*\widehat{b}^{\pi wls*}\right)$$

and note that $\mathbb{x}^*\widehat{b}^{\pi wls*} = \mathbb{x}\widehat{b}^{\pi wls}$ by the invariance of WLS. Now note that

$$\begin{aligned}
\mathbb{x}^*\left(\mathbb{x}^{*\prime}\mathbf{d}\mathbb{x}^*\right)^{(-)}\mathbb{x}^{*\prime} =& \mathbb{x}\mathbf{f}\left(\mathbf{f}'\mathbb{x}'\mathbf{d}\mathbb{x}\mathbf{f}\right)^{(-)}\mathbf{f}'\mathbb{x}'\\
=& \mathbb{x}\mathbf{f}\;\mathbf{f}^{-1}\left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(-)}\mathbf{f}'^{-1}\mathbf{f}'\mathbb{x}'\\
=& \mathbb{x}\left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(-)}\mathbb{x}'
\end{aligned}$$

where the second line follows from the properties of generalized inverses (cf. Campbell and Meyer, 2009). Hence, $\widehat{\delta}^{GR}(\widehat{b}^{2GR!*}) = \widehat{\delta}^{GR}(\widehat{b}^{2GR!})$. □

Remark 16. *Given its definition, $\widehat{\delta}^{GR}(\widehat{b}^{2GR!}) := \widehat{\delta}^{HT} - \widehat{\delta}^{HT}_{\mathbb{x}}\widehat{b}^{2GR!}$, invariance to location shifts in y and $\mathbb{x}$ not immediately obvious because the constituent parts, ($\widehat{\delta}^{HT}, \widehat{\delta}^{HT}_{\mathbb{x}}$, and $\widehat{b}^{2GR!}$), are not generally invariant. By contrast, the optimal GR estimator, $\widehat{\delta}^{GR}(\widehat{b}^{3HT!})$, is only invariant to location shifts in special cases (e.g., complete randomization).*

4.4 An MSE (bound) estimator for GR

The variance estimator in (18) may be unsatisfactory, particularly in small samples. However, one can unbiasedly estimate a bound on the MSE of the GR estimator.

First, define an arbitrary regression coefficient of the form $\widehat{b}^{\mathbf{C}} = \mathbf{C}y$ where $\mathbf{C}$ is a random matrix. For example, for OLS, $\mathbf{C} = \left(\mathbb{x}'\mathbf{R}\mathbb{x}\right)^{-1}\mathbb{x}'\mathbf{R}$. Now the zero-centered GR estimator can be written

$$\widehat{\delta}^{GR}(\widehat{b}^{\mathbf{C}}) - \delta = n^{-1}1_{2n}(\boldsymbol{\pi}^{-1}\mathbf{R} - \mathbf{i})\,(\mathbf{i} - \mathbb{x}\mathbf{C})\,y.$$

It has MSE

$$\mathrm{M}\left(\widehat{\delta}^{GR}(\widehat{b}^{\mathbf{C}})\right) = n^{-2}y'\mathbf{m}y$$

where

$$\mathbf{m} := \mathrm{E}\left[(\mathbf{i} - \mathbb{x}\mathbf{C})'\left(\boldsymbol{\pi}^{-1}\mathbf{R} - \mathbf{i}\right)1'_{2n}1_{2n}\left(\boldsymbol{\pi}^{-1}\mathbf{R} - \mathbf{i}\right)(\mathbf{i} - \mathbb{x}\mathbf{C})\right].$$

The $2n \times 2n$ matrix $\mathbf{m}$ could be obtained exactly in small enough samples by computing the matrix inside the expectation brackets for every possible randomization and averaging. Alternatively, the matrix could be simulated to arbitrary precision by averaging over a suitably large number of randomizations.

Next, a bound can be obtained by modifying Algorithm 3.7 to obtain $\tilde{\mathbf{m}}$ such that the elements corresponding to -1 values in $\mathbf{d}$ are zero and $\tilde{\mathbf{m}} - \mathbf{m}$ is positive semi-definite. Hence, the proposed MSE bound estimator is

$$\widehat{\tilde{\mathrm{M}}}\left(\widehat{\delta}^{GR}(\widehat{b}^{\mathbf{C}})\right) = n^{-2}y'\mathbf{R}\left(\tilde{\mathbf{m}}/\tilde{\mathbf{p}}\right)\mathbf{R}y$$

which is unbiased for the MSE bound

$$\tilde{\mathrm{M}}\left(\widehat{\delta}^{GR}(\widehat{b}^{\mathbf{C}})\right) = n^{-2}y'\tilde{\mathbf{m}}y.$$

4.5 Borrowing a tighter variance bound for $\widehat{\delta}^{GR}(\widehat{b}^{2GR\prime})$

In many instances the gap between the variance of the $\widehat{\delta}^{GR}(\widehat{b}^{2GR\prime})$, defined in section 4.3, and the bound on its variance estimated by (18) is exceedingly large, leading to overly conservative inference. Theorem 4.9 introduces an approach to minimizing the variance *bound* of the fixed-coefficient GR estimator, $\widehat{\delta}^{GR}(b^f)$, with respect to b^f. Then Theorem 4.10 gives a justification for "borrowing" its variance bound estimator and pairing it with $\widehat{\delta}^{GR}(\widehat{b}^{2GR\prime})$ for the purpose of inference.

Theorem 4.9. *Let $\tilde{\mathbf{d}}$ correspond to a variance bound and let $u = y - \mathbb{x}b^f$, where b^f is a fixed constant vector. Then for the fixed-coefficient GR estimator, a value of b^f that minimizes variance bound $n^{-2}u'\tilde{\mathbf{d}}u$ is*

$$\tilde{b}^{opt} := (\mathbb{x}'\tilde{\mathbf{d}}\mathbb{x})^{(-)}\mathbb{x}'\tilde{\mathbf{d}}y. \tag{26}$$

Proof. The result follows the same logic as the optimal finite sample b^f in Theorem 4.1. Rather than minimizing $u'\mathbf{d}u$, however, simply minimize $u'\tilde{\mathbf{d}}u$ with respect to b^f where $u = y - \mathbb{x}b^f$. □

Theorem 4.10. *Let $\tilde{\mathbf{d}}$ correspond to a variance bound and define $\tilde{u} = y - \mathbb{x}\tilde{b}^{opt}$ and $u = y - \mathbb{x}b^{opt}$, then for all $y \in \mathbb{R}^{2n}$, $n^{-2}u'\mathbf{d}u \leq n^{-2}\tilde{u}'\tilde{\mathbf{d}}\tilde{u} \leq n^{-2}u'\tilde{\mathbf{d}}u$. Hence, $n^{-2}\tilde{u}'\tilde{\mathbf{d}}\tilde{u}$ is a tighter bound for the variance of the fixed-coefficient GR estimator with optimal coefficient b^{opt} than $n^{-2}u'\tilde{\mathbf{d}}u$.*

Proof. By Theorem 4.1, because b^{opt} minimizes the variance of the fixed-coefficient GR estimator, $n^{-2}u'\mathbf{d}u \leq n^{-2}\tilde{u}'\mathbf{d}\tilde{u}$. Moreover, because $\tilde{\mathbf{d}}$ corresponds to a variance bound, $n^{-2}\tilde{u}'\mathbf{d}\tilde{u} \leq n^{-2}\tilde{u}'\tilde{\mathbf{d}}\tilde{u}$. Finally, by Theorem 4.9, because $\tilde{b}^{opt}$ minimizes the variance bound of the fixed-coefficient GR estimator, $n^{-2}\tilde{u}'\tilde{\mathbf{d}}\tilde{u} \leq n^{-2}u'\tilde{\mathbf{d}}u$. The result follows. □

The result motivates the variance bound estimator for $\widehat{\delta}^{GR}(\widehat{b}^{2GR\prime})$,

$$\widehat{\tilde{\mathrm{V}}}\left(\widehat{\delta}^{GR}(\widehat{b}^{2GR\prime})\right) := n^{-2}\widehat{\tilde{u}}'\mathbf{R}\left(\tilde{\mathbf{d}}/\tilde{\mathbf{p}}\right)\mathbf{R}\widehat{\tilde{u}}, \tag{27}$$

where $\widehat{\tilde{u}} = y - \mathbb{x}\widehat{\tilde{b}}^{opt}$ and $\widehat{\tilde{b}}^{opt}$ is an estimator of (26) that, for example, could be defined using a regression adjustment procedure analogous to (25). Again, additional adjustments for degrees of freedom or leverage may be advisable.

4.6 Conclusions about the proposed optimal estimators, $\widehat{\delta}^{GR}(\widehat{b}^{3HT!})$ and $\widehat{\delta}^{GR}(\widehat{b}^{2GR!})$

Estimators $\widehat{\delta}^{GR}(\widehat{b}^{3HT!})$ and $\widehat{\delta}^{GR}(\widehat{b}^{2GR!})$ have the virtue of being asymptotically optimal for arbitrary designs. However, asymptotic optimality does not necessarily imply good finite sample performance, and $\widehat{\delta}^{GR}(\widehat{b}^{3HT!})$ is not recommended in practice because it is unnecessarily imprecise and not generally invariant to location shifts in y. $\widehat{\delta}^{GR}(\widehat{b}^{2GR!})$ may be useful in some cases.

Alternatives to $\widehat{\delta}^{GR}(\widehat{b}^{3HT!})$ and $\widehat{\delta}^{GR}(\widehat{b}^{2GR!})$ are available for specific designs. In Section 5, complete randomization is considered, followed by Section 6 on clustered randomization. The sections show how to derive optimal estimators specific to those designs from this framework. Some of the results are known, but the derivation helps connect this framework to prior work (e.g. Lin, 2013).

5 Optimal Regression for Complete Randomization

This section draws the connection to two asymptotically optimal estimators for completely randomized designs, namely, OLS with specification II and the "tyranny of the minority" estimator. Lin (2013) originally proposed these estimators and shows that they are asymptotically optimal in response to Freedman's (2008a,b) critique that regression can hurt asymptotic precision. However, the proofs presented here are novel and the demonstration connects the current framework to Lin's results. In this section it is also shown that Lin's fully-interacted specification leads to tighter bounds on the variance in (20). Finally, it is proven that, for specification II, the 2GR! estimator, $\widehat{\delta}^{GR}(\widehat{b}^{2GR!}_{\text{II}})$, is algebraically equivalent to the OLS estimator, $\widehat{\delta}^{GR}(\widehat{b}^{ols}_{\text{II}})$. In that sense, 2GR! can be thought of as a generalization of Lin's OLS with specification II for arbitrary designs.

5.1 OLS is optimal for completely randomized designs with specification II

In this subsection, it will be shown that the population OLS coefficient, call it b^{ols}_{II}, is an optimal coefficient for the fixed-coefficient GR estimator with completely randomized designs and specification II. It will follow that, under Assumptions 1 and 2 in Section 2.4, the OLS coefficient *estimator*, call it $\widehat{b}^{ols}_{\text{II}}$, has a conjugate ATE estimator, $\widehat{\delta}^{GR}(\widehat{b}^{ols}_{\text{II}})$, that obtains minimum asymptotic variance.

Definition 5.1 (OLS coefficient)**.** *The "OLS coefficient for specification* II*" is*

$$\begin{aligned} b^{ols}_{\text{II}} &= (\mathbb{x}_{\text{II}}'\mathbb{x}_{\text{II}})^{-1}\,\mathbb{x}_{\text{II}}'y \\ &= \begin{bmatrix} \mu_{y_0} \\ \text{Var}(\mathbf{x})^{-1}\text{Cov}(\mathbf{x}, y_0) \\ \mu_{y_1} \\ \text{Var}(\mathbf{x})^{-1}\text{Cov}(\mathbf{x}, y_1) \end{bmatrix}, \end{aligned}$$

where $\text{Var}(\tilde{\mathbf{x}}) := n^{-1}\sum_i(\tilde{\mathbf{x}}_i - \mu_{\tilde{\mathbf{x}}})(\tilde{\mathbf{x}}_i - \mu_{\tilde{\mathbf{x}}})'$ *and* $\text{Cov}(\tilde{\mathbf{x}}, y_1) := n^{-1}\sum_i(\tilde{\mathbf{x}}_i - \mu_{\tilde{\mathbf{x}}})(y_{1i} - \mu_{y_1})'$ *are finite population variance and covariance, respectively.*[5]

Definition 5.2 (OLS coefficient estimator)**.** *The "OLS coefficient estimator for specification* II*" is*

$$\widehat{b}^{ols}_{\text{II}} = (\mathbb{x}_{\text{II}}'\mathbf{R}\mathbb{x}_{\text{II}})^{-1}\,\mathbb{x}_{\text{II}}'\mathbf{R}y.$$

To show that in completely randomized experiments with specification II the coefficient in Definition 5.1 is optimal for the fixed-coefficient GR estimator, the entire set of optimal coefficients for an arbitrary design is first defined. Subsequently, it can be shown that, for complete randomization, b^{ols}_{II} is in that set. The potentially infinite set of optimal coefficients for an arbitrary design is given in the following Lemma.

[5]Note that $\text{Var}(\tilde{\mathbf{x}})$ and $\text{Cov}(\tilde{\mathbf{x}}, y_1)$ summarize features of the finite population. They should not be taken to imply randomness in $\mathbf{x}$ and y_1. By contrast, V(.) is used throughout to characterize the design variance of an estimator (or variance-covariance of a vector of estimators, depending on context).

Lemma 5.3. *First, for an arbitrary design, for any given generalized inverse, denoted* $(.)^{(g)}$*, and a given* $z \in \mathbb{R}^l$ *where* l *is the number of columns of* $\mathbb{x}$*, let*

$$b^{opt,gz} := (\mathbb{x}'\mathbf{d}\mathbb{x})^{(g)}\,\mathbb{x}'\mathbf{d}y + \left(\mathbf{i}_l - (\mathbb{x}'\mathbf{d}\mathbb{x})^{(g)}(\mathbb{x}'\mathbf{d}\mathbb{x})\right) z \tag{28}$$

where $\mathbf{i}_l$ *is an* $l \times l$ *identity matrix. Then the entire set of solutions to* (23) *can be defined as*

$$\{b^{opt,gz} \mid z \in \mathbb{R}^l\}. \tag{29}$$

Proof. Provided in Appendix. □

In equation (28), the generalized inverse, g, is considered fixed and the set is defined with regard to all possible z. That said, g could be any generalized inverse. The point is that the entire set can be defined with reference to only a single generalized inverse. In keeping with the prior use of the Moore-Penrose generalized inverse above, it might have been sensible to also use it in (28) instead of the more generic g. However, in order to prove that OLS is optimal, a subsequent proof will use the fact that g can be some other inverse.

Next, before it can be proven that b_{II}^{ols} is in the set given by Lemma 5.3, it must be shown that a "separable" solutions can be optimal. By separable, it is meant that the sub-vector of coefficients associated with treatment units does not involve the terms $\mathbf{d}_{00}$, $\mathbf{d}_{01}$, $\mathbf{d}_{10}$ or y_0 (the vector of control potential outcomes), and, likewise, the sub-vector of coefficients associated with control potential outcomes does not involve the terms $\mathbf{d}_{11}$, $\mathbf{d}_{01}$, $\mathbf{d}_{10}$, or y_1 (the vector of treatment potential outcomes).[6]

Separability is provable with a less restrictive assumption than complete random assignment, namely, under equal-π_1 designs, (i.e., designs where $\pi_{1i} = \pi_{1j}$ for all i, j). Equal-π_1 designs include complete randomization, Bernoulli designs, cluster-randomized designs (i.e., complete random assignment of clusters) and block randomized designs where an equal fraction is assigned to treatment in every block.

Lemma 5.4. *For designs where* $\pi_{1i} = \pi_{1j}$ *for all* i, j *(e.g., completely randomized designs), defining* $\mathbf{d}_{**}$ *to be the matrix with* ij *element* $\pi_{1i1j} - \pi_{1i}\pi_{1j}$*, the following equalities hold:*

$$\mathbf{d}_{**} := \pi_{0i}^2\mathbf{d}_{00} = \pi_{1i}^2\mathbf{d}_{11} = -\pi_{1i}\pi_{0i}\mathbf{d}_{10} = -\pi_{0i}\pi_{1i}\mathbf{d}_{01}.$$

Proof. The result follows from the $\pi_{1i} = \pi_{1j}$ for all i, j and the defintion of the four partitions of $\mathbf{d}$ given in (8). □

Lemma 5.5. *Let* $\tilde{\mathbf{x}} = [1_n\ \mathbf{x}]$ *be the matrix of coefficients with the addition of a leading constant. For designs where* $\pi_{1i} = \pi_{1j}$ *for all* i, j *(e.g., completely randomized designs), the fixed-coefficient GR estimator with the "separated" coefficient*

$$b_{\text{II}}^{sep} := \begin{bmatrix} (\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}})^{(-)}\tilde{\mathbf{x}}'\mathbf{d}_{00}y_0 \\ (\tilde{\mathbf{x}}'\mathbf{d}_{11}\tilde{\mathbf{x}})^{(-)}\tilde{\mathbf{x}}'\mathbf{d}_{11}y_1 \end{bmatrix} \tag{30}$$

has finite-sample minimum variance, i.e., b_{II}^{sep} *is in the set of optimal fixed-coefficients given by Lemma 5.3.*

Proof. From Rohde (1965), for a positive semi-definite symmetric matrix

$$\mathbf{m} = \begin{bmatrix} \mathbf{a} & \mathbf{c} \\ \mathbf{c}' & \mathbf{b} \end{bmatrix}$$

a generalized inverse, call it g, is given by

$$\mathbf{m}^{(g)} := \begin{bmatrix} \mathbf{a}^{(-)} + \mathbf{a}^{(-)}\mathbf{c}\mathbf{q}^{(-)}\mathbf{c}'\mathbf{a}^{(-)} & -\mathbf{a}^{(-)}\mathbf{c}\mathbf{q}^{(-)} \\ -\mathbf{q}^{(-)}\mathbf{c}'\mathbf{a}^{(-)} & \mathbf{q}^{(-)} \end{bmatrix} \tag{31}$$

[6] Being separable implies that one way to minimize the overall variance of $\widehat{\delta}$ is to separately minimize (with respect to b) the variance of the estimated mean of each experimental arm while ignoring the other arm.

where $\mathbf{q} = \mathbf{b} - \mathbf{c}'\mathbf{a}^{(-)}\mathbf{c}$ and $(.)^{(-)}$ is the Moore-Penrose generalized inverse as before. Now if $\mathbf{m} = \mathbb{x}_{\text{II}}'\mathbf{d}\mathbb{x}_{\text{II}}$ then $\mathbf{q} = \tilde{\mathbf{x}}'\mathbf{d}_{11}\tilde{\mathbf{x}} - \tilde{\mathbf{x}}'\mathbf{d}_{10}\tilde{\mathbf{x}}\left(\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}}\right)^{(-)}\tilde{\mathbf{x}}'\mathbf{d}_{01}\tilde{\mathbf{x}}$. But because of Lemma 5.4, and using the definition of generalized inverse, it follows that $\mathbf{q} = 0$. So the inverse reduces to

$$\begin{aligned}\mathbf{m}^{(g)} &= \begin{bmatrix}\mathbf{a}^{(-)} & 0\\ 0 & 0\end{bmatrix}\\ &= \begin{bmatrix}(\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}})^{(-)} & 0\\ 0 & 0\end{bmatrix}.\end{aligned}$$

Therefore,

$$\begin{aligned}\left(\mathbb{x}_{\text{II}}'\mathbf{d}\mathbb{x}_{\text{II}}\right)^{(g)}\mathbb{x}_{\text{II}}'\mathbf{d}y &= \begin{bmatrix}(\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}})^{(-)} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\tilde{\mathbf{x}}'\mathbf{d}_{00}y_0 - \tilde{\mathbf{x}}'\mathbf{d}_{01}y_1\\ -\tilde{\mathbf{x}}'\mathbf{d}_{10}y_0 + \tilde{\mathbf{x}}'\mathbf{d}_{11}y_1\end{bmatrix}\\ &= \begin{bmatrix}(\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}})^{(-)}\left(\tilde{\mathbf{x}}'\mathbf{d}_{00}y_0 - \tilde{\mathbf{x}}'\mathbf{d}_{01}y_1\right)\\ 0\end{bmatrix}. \qquad (32)\end{aligned}$$

Next, using (32) and Lemma 5.3,

$$\begin{aligned}b_{\text{II}}^{opt,gz} &= \left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(g)}\mathbb{x}'\mathbf{d}y + \left(\mathbf{i}_l - \begin{bmatrix}(\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}})^{(-)} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}} & \tilde{\mathbf{x}}'\mathbf{d}_{01}\tilde{\mathbf{x}}\\ \tilde{\mathbf{x}}'\mathbf{d}_{10}\tilde{\mathbf{x}} & \tilde{\mathbf{x}}'\mathbf{d}_{11}\tilde{\mathbf{x}}\end{bmatrix}\right)z\\ &= \begin{bmatrix}(\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}})^{(-)}\left(\tilde{\mathbf{x}}'\mathbf{d}_{00}y_0 - \tilde{\mathbf{x}}'\mathbf{d}_{01}y_1\right)\\ 0\end{bmatrix} + \begin{bmatrix}\mathbf{i}_{(k+1)} - (\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}})^{(-)}\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}} & (\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}})^{(-)}\tilde{\mathbf{x}}'\mathbf{d}_{01}\tilde{\mathbf{x}}\\ 0 & \mathbf{i}_{(k+1)}\end{bmatrix}z. \qquad (33)\end{aligned}$$

Finally letting $z = \begin{bmatrix}0\\ (\tilde{\mathbf{x}}'\mathbf{d}_{11}\tilde{\mathbf{x}})^{(-)}\tilde{\mathbf{x}}\mathbf{d}_{11}y_1\end{bmatrix}$ leads to the result, with the last steps requiring the use of the reflexive property of the Moore-Penrose generalized inverse (i.e., for a symmetric matrix $\mathbf{a}$, $\mathbf{a}^{(-)}\mathbf{a}\mathbf{a}^{(-)} = \mathbf{a}^{(-)}$) and Lemma 5.4. □

Remark 17. *It is not always the case that an optimal coefficient vector has a "separable" solution. It can be the case that, in order to be optimal, the subvector of the coefficient associated with treatment potential outcomes must take account of control potential outcomes and vice versa. Surprisingly, this can be true even under the sharp null.*

Next, two additional lemmas will be necessary before showing the optimality of the OLS coefficient. Lemma 5.6 will show that, for completely randomized experiments, multiplying $\mathbf{d}_{11}$, $\mathbf{d}_{00}$, $\mathbf{d}_{01}$, or $\mathbf{d}_{10}$ by a length-n column vector zero-centers the vector and rescales by a constant. Lemma 5.7 will show that matrices such as $\tilde{\mathbf{x}}'\mathbf{d}_{11}\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}'\mathbf{d}_{11}y_1$ represent finite-population covariance matrices rescaled by constants.

Lemma 5.6. *In a completely randomized experiment,* $\mathbf{d}_{11}\tilde{\mathbf{x}} = \frac{nn_0}{(n-1)n_1}(\tilde{\mathbf{x}} - 1_n\mu_{\tilde{\mathbf{x}}})$*, with* 1_n *as a* $(n \times 1)$ *vector of ones and* $\mu_{\tilde{\mathbf{x}}}$ *a* $k+1$ *rowvector giving the column means of* $\tilde{\mathbf{x}}$*. Likewise, in a completely randomized experiment,* $\mathbf{d}_{00}\tilde{\mathbf{x}} = \frac{nn_1}{(n-1)n_0}(\tilde{\mathbf{x}} - 1_n\mu_{\tilde{\mathbf{x}}})$*. And* $\mathbf{d}_{10}\tilde{\mathbf{x}} = \mathbf{d}_{01}\tilde{\mathbf{x}} = -\frac{n}{n-1}(\tilde{\mathbf{x}} - 1_n\mu_{\tilde{\mathbf{x}}})$*.*

Proof. Provided in Appendix. □

Corollary 5.6.1. *For any constant vector,* c_n*,* $\mathbf{d}_{00}c_n = \mathbf{d}_{11}c_n = \mathbf{d}_{01}c_n = \mathbf{d}_{10}c_n = 0_n$*.*

Lemma 5.7. *In a completely randomized design*

$$\begin{aligned}\tilde{\mathbf{x}}'\mathbf{d}_{11}\tilde{\mathbf{x}} &= c_{11}\text{Var}(\tilde{\mathbf{x}})\\ \tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}} &= c_{00}\text{Var}(\tilde{\mathbf{x}}),\\ \tilde{\mathbf{x}}'\mathbf{d}_{01}\tilde{\mathbf{x}} &= c_{01}\text{Var}(\tilde{\mathbf{x}}),\\ \textit{and } \tilde{\mathbf{x}}'\mathbf{d}_{10}\tilde{\mathbf{x}} &= c_{10}\text{Var}(\tilde{\mathbf{x}}),\end{aligned}$$

where $\text{Var}(\tilde{\mathbf{x}}) := n^{-1}\sum_i(\tilde{\mathbf{x}}_i - \mu_{\tilde{\mathbf{x}}})(\tilde{\mathbf{x}}_i - \mu_{\tilde{\mathbf{x}}})'$ *is the finite-population variance-covariance matrix for* $\tilde{\mathbf{x}}$, $c_{11} := \frac{n^2 n_0}{(n-1)n_1}$, $c_{00} := \frac{n^2 n_1}{(n-1)n_0}$, *and* $c_{01} = c_{10} := -\frac{n^2}{(n-1)}$. *Similarly,*

$$\begin{aligned}\tilde{\mathbf{x}}'\mathbf{d}_{11}y_1 =& c_{11}\text{Cov}(\tilde{\mathbf{x}}, y_1)\\ \tilde{\mathbf{x}}'\mathbf{d}_{00}y_0 =& c_{00}\text{Cov}(\tilde{\mathbf{x}}, y_0),\\ \tilde{\mathbf{x}}'\mathbf{d}_{01}y_1 =& c_{01}\text{Cov}(\tilde{\mathbf{x}}, y_1),\\ \textit{and } \tilde{\mathbf{x}}'\mathbf{d}_{10}y_0 =& c_{10}\text{Cov}(\tilde{\mathbf{x}}, y_0).\end{aligned}$$

where, for example, $\text{Cov}(\tilde{\mathbf{x}}, y_1) := n^{-1}\sum_i(\tilde{\mathbf{x}}_i - \mu_{\tilde{\mathbf{x}}})(y_{1i} - \mu_{y_1})'$ *is a vector of finite-population covariances between* y_1 *and x's.*

Proof. Results follow from Lemma 5.6 and the fact that $\sum_i(\tilde{\mathbf{x}}_i - \mu_{\tilde{\mathbf{x}}})\tilde{\mathbf{x}}_i' = \sum_i(\tilde{\mathbf{x}}_i - \mu_{\tilde{\mathbf{x}}})(\tilde{\mathbf{x}}_i - \mu_{\tilde{\mathbf{x}}})'$. □

Finally, the next two theorems present the main results of the section.

Theorem 5.8. *In a completely randomized design with specification* II*, the OLS coefficient given in Definition 5.1 minimizes the variance of the fixed-coefficient GR estimator, i.e.,* b_{II}^{ols} *is in the set of optimal fixed-coefficients defined in Lemma 5.3.*

Proof. Using Lemma 5.7 write

$$\begin{aligned}(\tilde{\mathbf{x}}'\mathbf{d}_{11}\tilde{\mathbf{x}})^{(-)}\tilde{\mathbf{x}}'\mathbf{d}_{11}y_1 &= \begin{bmatrix}0 & 0_k'\\ 0_k & \text{Var}(\mathbf{x})\end{bmatrix}^{(-)}\begin{bmatrix}0\\ \text{Cov}(\mathbf{x}, y_1)\end{bmatrix}\\ &= \begin{bmatrix}0\\ \text{Var}(\mathbf{x})^{(-)}\text{Cov}(\mathbf{x}, y_1)\end{bmatrix}\end{aligned}$$

And note that unless the columns of $\mathbf{x}$ are colinear, $(.)^{(-)}$ is equivalent to the usual inverse. Now use Lemma 5.3 and let

$$z = \begin{bmatrix}\mu_{y_1}\\ 0_k\end{bmatrix},$$

where μ_{y_1} is he mean of treatment potential outcomes, to arrive at an optimal sub-vector for treatment potential outcomes is

$$\begin{bmatrix}\mu_{y_1}\\ \text{Var}(\mathbf{x})^{-1}\text{Cov}(\mathbf{x}, y_1)\end{bmatrix}.$$

An analogous optimal sub-vector for control potential outcomes can be defined. The result follows. □

Theorem 5.9. *Under Assumptions 1 and 2, in a completely randomized design with specification* II*, the OLS coefficient estimator given in Definition 5.2 has a conjugate ATE estimator that obtains asymptotic minimum variance within the class of GR estimators.*

Proof. Provided in Appendix. □

5.2 Tyranny of the minority is optimal for completely randomized designs with specification I

In this section, it will be shown that an optimal coefficient for the fixed-coefficient GR estimator for completely randomized designs and specification I is the "tyranny of the minority" coefficient (Lin, 2013), call it b_{I}^{tyr}, and a WLS estimator of the coefficient will be defined. It is noteworthy that, by contrast, there is no OLS analogue that is generally asymptotically optimal for specification I for completely randomized experiments. The section will also show that tyranny of the minority can achieve asymptotic precision using specification I that is as good as optimal estimators that use specification II.

First define the tyranny of the minority coefficient for specification I and its estimator.

Definition 5.10 (Tyranny of the minority coefficient)**.** *The "tyranny of the minority" coefficient for specification* I *is given by*

$$\begin{aligned} b_{\text{I}}^{tyr} &:= \left(\mathbb{x}_{\text{I}}{}' \left(\mathbf{i}_{2n} - \boldsymbol{\pi}\right) \mathbb{x}_{\text{I}}\right)^{-1} \mathbb{x}_{\text{I}}{}' \left(\mathbf{i}_{2n} - \boldsymbol{\pi}\right) y \\ &= \begin{bmatrix} \mu_{y_0} \\ \mu_{y_1} \\ \frac{n_1}{n}\text{Var}(\mathbf{x})^{(-)}\text{Cov}(\mathbf{x}, y_0) + \frac{n_0}{n}\text{Var}(\mathbf{x})^{(-)}\text{Cov}(\mathbf{x}, y_1) \end{bmatrix} \end{aligned}$$

where μ_{y_0} *and* μ_{y_1} *are means of control and treatment potential outcomes, respectively, and* $\mathbf{i}_{2n}$ *is a* $2n \times 2n$ *identity matrix.*

Note in the that $\text{Var}(\mathbf{x})^{(-)}\text{Cov}(\mathbf{x}, y_0)$ is the population least squares coefficients when regressing y_0 on $\mathbf{x}$, and $\text{Var}(\mathbf{x})^{(-)}\text{Cov}(\mathbf{x}, y_1)$ is, likewise, the population least squares coefficients when regressing y_1 on $\mathbf{x}$. The weights for combining these two components, $\frac{n_1}{n}$ and $\frac{n_0}{n}$, respectively, are such that the coefficient for the arm with fewer units gets more weight. Hence, the name "tyranny of the minority".

Definition 5.11 (Tyranny of the minority coefficient estimator)**.** *The "tyranny of the minority coefficient estimator" for specification* I *is*

$$\widehat{b}_{\text{I}}^{tyr} = \left(\mathbb{x}_{\text{I}}{}'\mathbf{R}\left(\boldsymbol{\pi}^{-1} - \mathbf{i}_{2n}\right)\mathbb{x}_{\text{I}}\right)^{-1} \mathbb{x}_{\text{I}}{}'\mathbf{R}\left(\boldsymbol{\pi}^{-1} - \mathbf{i}_{2n}\right) y.$$

where $\mathbf{i}_{2n}$ *is a* $2n \times 2n$ *identity matrix.*

To prove that b_{I}^{tyr} in Definition 5.10 is an optimal choice of coefficient for the fixed-coefficient GR estimator, first define an equivalent coefficient for specification II.

Definition 5.12 (Tyranny of the minority coefficient for specification II)**.** *The "tyranny of the minority coefficient for specification* II*" is*

$$b_{\text{II}}^{tyr} = \begin{bmatrix} \mu_{y_0} \\ \frac{n_1}{n}\text{Var}(\mathbf{x})^{-1}\text{Cov}(\mathbf{x}, y_0) + \frac{n_0}{n}\text{Var}(\mathbf{x})^{-1}\text{Cov}(\mathbf{x}, y_1) \\ \mu_{y_1} \\ \frac{n_1}{n}\text{Var}(\mathbf{x})^{-1}\text{Cov}(\mathbf{x}, y_0) + \frac{n_0}{n}\text{Var}(\mathbf{x})^{-1}\text{Cov}(\mathbf{x}, y_1) \end{bmatrix}.$$

Comparing Definition 5.12 to Definition 5.10 reveals that the "slope" coefficients, given by $\frac{n_1}{n}\text{Var}(\mathbf{x})^{-1}\text{Cov}(\mathbf{x}, y_0) + \frac{n_0}{n}\text{Var}(\mathbf{x})^{-1}\text{Cov}(\mathbf{x}, y_1)$, are identical for the two specifications. The implication is that $\mathbb{x}_{\text{I}} b_{\text{I}}^{tyr} = \mathbb{x}_{\text{II}} b_{\text{II}}^{tyr}$ and hence, the conjugate ATE estimators are algebraically equivalent. Therefore, if b_{II}^{tyr} is in the set of optimal choices for a fixed-coefficient in specification II, then b_{I}^{tyr} must be among the optimal coefficients for the fixed-coefficient GR estimator for specification I.

Theorem 5.13. *For completely randomized experiments with specification* II*, the tyranny of the minority coefficient given in Definition 5.12 is an optimal coefficient for the fixed-coefficient GR estimator.*

Proof. Beginning with Lemma 5.3 and again arriving at equation (33), this time let

$$z = \begin{bmatrix} \mu_{y_0} \\ 0_k \\ \mu_{y_1} \\ \frac{n_1}{n}\text{Var}(\mathbf{x})^{-1}\text{Cov}(\mathbf{x}, y_0) + \frac{n_0}{n}\text{Var}(\mathbf{x})^{-1}\text{Cov}(\mathbf{x}, y_1) \end{bmatrix}.$$

The result follows. □

Corollary 5.13.1. *For completely randomized experiments with specification* I*, the tyranny of the minority coefficient given in Definition 5.10 is an optimal coefficient for the fixed-coefficient GR estimator.*

Theorem 5.14. *Under Assumptions 1 and 2, in a completely randomized design with specification* I*, the tyranny of the minority coefficient estimator given in Definition 5.11 has a conjugate ATE estimator that obtains asymptotic minimum variance within the class of GR estimators.*

Proof. Provided in Appendix. □

5.3 OLS coefficients minimize AS bound for completely randomized designs with specification II

Given that OLS with specification II (see Section 5.1) and the tyranny of the minorty estimator with specification I (see Section 5.2) can be equally precise, it is unclear which might be preferable. One way to evaluate this is to see which leads to a tighter variance bound. In this section, it is shown that in completely randomized designs and specification II, the coefficients that minimize the AS variance bound are given by b_{II}^{ols} in Definition 5.1. The result suggests that, when using the AS bound, OLS with specification II will tend to lead to smaller intervals than tyranny of the minority.

Theorem 5.15. *In the completely randomized design with specification* II*, if we let* $u = y - \mathbb{x}_{\text{II}} b_{\text{II}}^f$ *with* b_{II}^f *being a fixed-coefficient, then a value of* b_{II}^f *that minimizes the bound on the variance,* $n^{-2}u'\tilde{\mathbf{d}}u$*, is* b_{II}^{ols} *from Definition 5.1.*

Proof. Provided in Appendix. □

5.4 2GR! is algebraically equivalent to OLS for completely randomized designs with specification II

It has been shown that OLS is asymptotically optimal in completely randomized experiments when specification II is used. In this section, it is demonstrated that the two-step optimal estimator, $\widehat{\delta}^{GR}(\widehat{b}_{\text{II}}^{2GR!})$, is algebraically equivalent to the OLS estimator, $\widehat{\delta}^{GR}(\widehat{b}_{\text{II}}^{ols})$.

Theorem 5.16. *The vector of residuals,* $\mathbf{R}_1\boldsymbol{\pi}^{-1}\widehat{u}_1$*, is orthogonal to the weights* $\mathbf{d}_{11}\tilde{\mathbf{x}}(\tilde{\mathbf{x}}'\mathbf{d}_{11}\tilde{\mathbf{x}})^{(g)}$*.*

Proof. From the lemmas above we have

$$\begin{aligned}\tilde{\mathbf{x}}'\mathbf{d}_{11}\mathbf{R}_1\boldsymbol{\pi}^{-1}\widehat{u}_1 =& c\sum_i(\tilde{\mathbf{x}}_i - \mu_x)\widehat{u}_{1i}R_i \\ =& c \times 0\end{aligned}$$

where $c = \frac{n}{n_1}\frac{n^2}{n_1}\left(1 - \frac{n_1-1}{n-1}\right)$ (with the first $\frac{n}{n_1}$ coming from $\boldsymbol{\pi}^{-1}$). The last line follows from the fact that we know that for OLS that the column space of (the observed) $\tilde{\mathbf{x}}$'s is orthogonal to the residuals. □

The result shows that the two-step optimal will not make any adjustment to the OLS estimates in the completely randomized case. The estimators are algebraically equivalent.

6 Optimal Regression for Cluster-Randomization

This section reports on results for experiments with complete randomization of clusters. As above, it is assumed that we have an identified design and that every arm has at least two units of randomization (clusters) assigned to it.

It is also assumed that there is no second-stage selection from within clusters, which is to say that covariates and outcomes are available for all cluster members. Extensions that account for second-stage sampling are possible but beyond the scope of the paper.

When analyzing cluster randomized experiments, one approach to estimating the ATE is to regress the individual-level data on the treatment indicator and covariates using OLS, but it is not asymptotically optimal. By contrast, as the next subsections will show, regression using cluster totals is asymptotically optimal.[7]

[7]It should also be noted that an alternative approach is to first take cluster averages before running regression. This approach is biased and not generally consistent for the ATE (Middleton, 2008). However, if one were content to estimate the average of cluster-level average effects, this approach may be acceptable. There are benefits to doing this. For example, results from

6.1 OLS with cluster totals is optimal for cluster-randomized designs with specification II

In this subsection, it will be shown that regression with cluster totals is asymptotically optimal for cluster randomized experiments using specification II.

First, let m, m_0, and m_1 be the number of clusters, the number of clusters in treatment and the number of clusters in control, respectively. Meanwhile, let c_i give the cluster id number for the cluster that includes unit i, and let $\tilde{\mathbf{x}}_n^c$ represent an $n \times (k+1)$ matrix of cluster totals, i.e., with i^{th} row giving the sum of rows in $\mathbf{x}$ associated with units in cluster c_i. By contrast, let $\tilde{\mathbf{x}}^c$ represent an $m \times (k+1)$ matrix of cluster totals, with g^{th} row giving the sum of rows of $\tilde{\mathbf{x}}$ associated with units in cluster g.

Definition 6.1 (OLS with cluster totals)**.** *The "OLS with cluster totals" coefficient for specification* II *is*

$$b_{\text{II}}^{ols,c} = \begin{bmatrix} \text{Var}(\tilde{\mathbf{x}}^c)^{(-)}\text{Cov}(\tilde{\mathbf{x}}^c, y_0^c) \\ \text{Var}(\tilde{\mathbf{x}}^c)^{(-)}\text{Cov}(\tilde{\mathbf{x}}^c, y_1^c) \end{bmatrix}$$

where $\tilde{\mathbf{x}}^c$ is the $m \times (k+1)$ matrix with row g including cluster totals for the g^{th} cluster. Likewise, y_0^c and y_1^c are length m with entry g representing cluster totals for the g^{th} cluster's y_{0i} and y_{1i} values, respectively.

Next, to define the corresponding coefficient estimator, first let specification II^c be as follows

$$\mathbb{x}_{\text{II}}^c = \begin{bmatrix} -1_m & 0_m & -\tilde{\mathbf{x}}^c & 0_{m\times(k+1)} \\ 0_m & 1_m & 0_{m\times(k+1)} & \tilde{\mathbf{x}}^c \end{bmatrix}.$$

Note $\mathbb{x}_{\text{II}}^c$ has $2k+4$ columns where $\mathbb{x}_{\text{II}}$ has only $2k+2$.

Definition 6.2 (OLS with cluster totals coefficient estimator)**.** *The "OLS with cluster totals coefficient estimator" for specification* II *is*

$$\begin{bmatrix} \widehat{a}_0 \\ \widehat{a}_1 \\ \widehat{b}_{\text{II}}^{ols,c} \end{bmatrix} = \left(\mathbb{x}_{\text{II}}^{c\prime}\mathbf{R}^c\mathbb{x}_{\text{II}}^c\right)^{-1}\mathbb{x}_{\text{II}}^{c\prime}\mathbf{R}^c y^c$$

where $\widehat{a}_0$ and $\widehat{a}_1$ are scalars and $\widehat{b}_{\text{II}}^{ols,c}$ has length $2k+2$ and $\mathbf{R}^c$ (an analog to $\mathbf{R}$) is a $2m \times 2m$ diagonal matrix with cluster-level assignment indicators on the diagonal.

The two lemmas that follow will lead into the final result of the section. Lemma 6.3 shows that, for cluster-randomized experiments, multiplying $\mathbf{d}_{11}$, $\mathbf{d}_{00}$, $\mathbf{d}_{01}$, or $\mathbf{d}_{10}$ by a length-n column vector returns a length-n vector of cluster totals, zero-centered and multiplied by a constant. Lemma 6.4 will show that matrices such as $\tilde{\mathbf{x}}'\mathbf{d}_{11}\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}'\mathbf{d}_{11}y_1$ represent finite-population covariance matrices for cluster totals rescaled by constants.

Lemma 6.3. *In a cluster randomized experiment, $\mathbf{d}_{11}\tilde{\mathbf{x}} = \frac{mm_0}{(m-1)m_1}(\tilde{\mathbf{x}}_n^c - \frac{n}{m}1_n\mu_{\tilde{\mathbf{x}}})$ where $\frac{n}{m}1_n\mu_{\tilde{\mathbf{x}}}$ is a matrix that subtracts off the average of cluster totals. Likewise, in a cluster-randomized experiment, $\mathbf{d}_{00}\tilde{\mathbf{x}} = \frac{mm_1}{(m-1)m_0}(\tilde{\mathbf{x}}_n^c - \frac{n}{m}1_n\mu_{\tilde{\mathbf{x}}})$. And $\mathbf{d}_{10}\tilde{\mathbf{x}} = \mathbf{d}_{01}\tilde{\mathbf{x}} = -\frac{m}{m-1}(\tilde{\mathbf{x}}_n^c - \frac{n}{m}1_n\mu_{\tilde{\mathbf{x}}})$.*

Proof. Provided in Appendix. □

Lemma 6.4. *In a cluster randomized experiment, $\tilde{\mathbf{x}}'\mathbf{d}_{11}\tilde{\mathbf{x}} = \frac{m^2m_0}{(m-1)m_1}\text{Var}(\tilde{\mathbf{x}}^c)$. Likewise, the $\tilde{\mathbf{x}}'\mathbf{d}_{11}y_1 = \frac{m^2m_0}{(m-1)m_1}\text{Cov}(\tilde{\mathbf{x}}^c, y_1^c)$ where y_1^c is an length-m vector with the g^{th} element representing cluster totals for the g^{th} cluster's y_{1i} values.*

Section 5 can be applied directly. Moreover, compared to analyzing cluster totals, high leverage observations, which can foul normal-theory inference, are less likely. Moreover, in the presence of treatment effects, summing to create cluster-level totals is likely to induce a correlation between the leverage of an observation and its treatment effect (in this case the sum of treatment effects for units in the cluster). The first-order term in regression's bias is the correlation between leverage and treatment effect (Lin, 2013; Freedman, 2008a,b).

Proof. Provided in Appendix. □

Theorem 6.5. *For cluster randomized experiments, the OLS with cluster totals coefficient in Definition 6.1 is optimal for the fixed-coefficient GR estimator.*

Proof. Since π_{1i} is equal for all i in a cluster randomized experiment, then using Lemma 5.5, we have that the optimal solution includes the separated coefficients in equation (30). So, by Lemma 6.4, for cluster randomized experiments

$$\begin{aligned} b_{\text{II}}^{sep} &= \begin{bmatrix} (\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}})^{(-)}\tilde{\mathbf{x}}'\mathbf{d}_{00}y_0 \\ (\tilde{\mathbf{x}}'\mathbf{d}_{11}\tilde{\mathbf{x}})^{(-)}\tilde{\mathbf{x}}'\mathbf{d}_{11}y_1 \end{bmatrix} \\ &= \begin{bmatrix} \text{Var}(\tilde{\mathbf{x}}^c)^{(-)}\text{Cov}(\tilde{\mathbf{x}}^c, y_0^c) \\ \text{Var}(\tilde{\mathbf{x}}^c)^{(-)}\text{Cov}(\tilde{\mathbf{x}}^c, y_1^c) \end{bmatrix}. \end{aligned}$$

□

Remark 18. *Note that the "intercept" terms are no longer constants when $\tilde{\mathbf{x}}$ is collapsed to $\tilde{\mathbf{x}}^c$. In a sense, the intercept terms now "control" for cluster size in OLS with cluster totals.*

Theorem 6.6. *Under Assumptions 1 and 2, in a cluster-randomized design with specification* II*, the OLS coefficient with cluster totals estimator given in Definition 6.2 has a conjugate ATE estimator that obtains asymptotic minimum variance within the class of GR estimators.*

Proof. Provided in Appendix. □

6.2 Tyranny of the minority with cluster totals is optimal for cluster randomized experiments with specification I

In this section, it will be shown that an optimal coefficient for the fixed-coefficient GR estimator for cluster-randomized designs and specification I is the "tyranny of the minority with cluster totals" coefficient, call it $b_{\text{I}}^{tyr,c}$. A WLS estimator of the coefficient will be defined. The section will also show that tyranny of the minority with cluster totals can achieve asymptotic precision using specification I that is as good as optimal estimators that use specification II.

First define the tyranny of the minority coefficient for cluster totals for specification I and its estimator.

Definition 6.7 (Tyranny of the minority with cluster totals coefficient)**.** *The "tyranny of the minority" with cluster totals coefficient for specification* I *is given by*

$$b_{\text{I}}^{tyr,c} := \frac{m_1}{m}\text{Var}(\tilde{\mathbf{x}}^c)^{(-)}\text{Cov}(\tilde{\mathbf{x}}^c, y_0^c) + \frac{m_0}{m}\text{Var}(\tilde{\mathbf{x}}^c)^{(-)}\text{Cov}(\tilde{\mathbf{x}}^c, y_1^c).$$

Next, to define the corresponding coefficient estimator, first let specification I^c be as follows

$$\mathbb{x}_{\text{I}}^c := \begin{bmatrix} -1_m & 0_m & -\tilde{\mathbf{x}}^c \\ 0_m & 1_m & \tilde{\mathbf{x}}^c \end{bmatrix}.$$

Note $\mathbb{x}_{\text{I}}^c$ has $l+1$ columns where $\mathbb{x}_{\text{I}}$ has only l.

Definition 6.8 (Tyranny of the minority with cluster totals coefficient estimator)**.** *The "tyranny of the minority with cluster totals coefficient estimator" for specification* I *is*

$$\begin{bmatrix} \widehat{a}_0 \\ \widehat{a}_1 \\ \widehat{b}_{\text{I}}^{tyr,c} \end{bmatrix} := \left(\mathbb{x}_{\text{I}}^{c\prime}\mathbf{R}^c\left((\boldsymbol{\pi}^c)^{-1} - \mathbf{i}_{2m}\right)\mathbb{x}_{\text{I}}^c\right)^{-1}\mathbb{x}_{\text{I}}^{c\prime}\mathbf{R}^c\left((\boldsymbol{\pi}^c)^{-1} - \mathbf{i}_{2m}\right)y^c.$$

where $\boldsymbol{\pi}^c$ is a $2m \times 2m$ matrix giving probabilities of assignment along the diagonal.

To prove that $b_{\text{I}}^{tyr,c}$ is an optimal choice of coefficient for the fixed-coefficient GR estimator, first define an equivalent coefficient for specification II.

Definition 6.9 (Tyranny of the minority with cluster totals for specification II)**.** *The "tyranny of the minority with cluster totals coefficient" for specification* II *is*

$$b_{\text{II}}^{tyr,c} = \begin{bmatrix} \frac{m_1}{m}\text{Var}(\tilde{\mathbf{x}}^c)^{-1}\text{Cov}(\tilde{\mathbf{x}}^c, y_0^c) + \frac{m_0}{m}\text{Var}(\tilde{\mathbf{x}}^c)^{-1}\text{Cov}(\tilde{\mathbf{x}}^c, y_1^c) \\ \frac{m_1}{m}\text{Var}(\tilde{\mathbf{x}}^c)^{-1}\text{Cov}(\tilde{\mathbf{x}}^c, y_0^c) + \frac{m_0}{m}\text{Var}(\tilde{\mathbf{x}}^c)^{-1}\text{Cov}(\tilde{\mathbf{x}}^c, y_1^c) \end{bmatrix}.$$

Comparing Definition 6.9 to Definition 6.7 reveals that the "slope" coefficients are identical in the two specifications. The implication is that $\not{\mathbf{x}}_{\text{I}}^c b_{\text{I}}^{tyr,c} = \not{\mathbf{x}}_{\text{II}}^c b_{\text{II}}^{tyr,c}$ and hence, the conjugate ATE estimators are algebraically equivalent. Therefore, if $b_{\text{II}}^{tyr,c}$ is in the set of optimal choices for a fixed-coefficient in specification II, then $b_{\text{I}}^{tyr,c}$ must be among the optimal coefficients for the fixed-coefficient GR estimator for specification I.

Theorem 6.10. *For cluster-randomized experiments with specification* II*, the tyranny of the minority with cluster totals coefficient given in Definition 6.9 is an optimal coefficient for the fixed-coefficient GR estimator.*

Proof. Beginning with Lemma 5.3 and again arriving at equation (33), this time let

$$z = \begin{bmatrix} 0_{(k+1)} \\ \frac{m_1}{m}\text{Var}(\tilde{\mathbf{x}}^c)^{-1}\text{Cov}(\tilde{\mathbf{x}}^c, y_0^c) + \frac{m_0}{m}\text{Var}(\tilde{\mathbf{x}}^c)^{-1}\text{Cov}(\tilde{\mathbf{x}}^c, y_1^c) \end{bmatrix}.$$

The result follows. □

Corollary 6.10.1. *For cluster-randomized experiments with specification* I*, the tyranny of the minority coefficient given in Definition 6.7 is an optimal coefficient for the fixed-coefficient GR estimator.*

Theorem 6.11. *Under Assumptions 1 and 2, in a cluster-randomized design with specification* I*, the tyranny of the minority with cluster totals coefficient estimator given in Definition 6.8 has a conjugate ATE estimator that obtains asymptotic minimum variance within the class of GR estimators.*

Proof. Provided in Appendix. □

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. 2015. 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. 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_{2n}	Length-$2n$ 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-$2n$ 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
$\mathbf{R}$	$2n \times 2n$ 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
$\boldsymbol{\pi}$	$2n \times 2n$ 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
$\mathbf{d}$	$2n \times 2n$ "design" matrix that gives the variance-covariance matrix of the vector $1'_{2n}\boldsymbol{\pi}^{-1}\mathbf{R}$. 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

$\mathbf{p}$	$2n \times 2n$ "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 $2n \times l$ 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 $2n \times (k+2)$ 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 $2n \times (2k+2)$ 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

Proof of Lemma 4.5. From the first line of (25), the two-step optimal coefficient when inputting y^* in place of y can be written

$$\begin{aligned}
\widehat{b}^{2GR!*} =& \left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(-)}\mathbb{x}\mathbf{d}\left(\boldsymbol{\pi}^{-1}\mathbf{R}y^* - \left(\boldsymbol{\pi}^{-1}\mathbf{R}-\mathbf{i}\right)\mathbb{x}\widehat{b}^{\pi wls*}\right)\\
=& \left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(-)}\mathbb{x}\mathbf{d}\left(\boldsymbol{\pi}^{-1}\mathbf{R} - \left(\boldsymbol{\pi}^{-1}\mathbf{R}-\mathbf{i}\right)\mathbb{x}\left(\mathbb{x}'\boldsymbol{\pi}^{-1}\mathbf{R}\mathbb{x}\right)^{-1}\mathbb{x}'\boldsymbol{\pi}^{-1}\mathbf{R}\right)\left(fy + c\begin{bmatrix}-1_n\\1_n\end{bmatrix}\right)\\
=& f\widehat{b}^{2GR!} + c\left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(-)}\mathbb{x}\mathbf{d}\left(\boldsymbol{\pi}^{-1}\mathbf{R} - \left(\boldsymbol{\pi}^{-1}\mathbf{R}-\mathbf{i}\right)\mathbb{x}\left(\mathbb{x}'\boldsymbol{\pi}^{-1}\mathbf{R}\mathbb{x}\right)^{-1}\mathbb{x}'\boldsymbol{\pi}^{-1}\mathbf{R}\right)\begin{bmatrix}-1_n\\1_n\end{bmatrix}\\
=& f\widehat{b}^{2GR!} + c\left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(-)}\mathbb{x}\mathbf{d}\left(\boldsymbol{\pi}^{-1}\mathbf{R} - \left(\boldsymbol{\pi}^{-1}\mathbf{R}-\mathbf{i}\right)\mathbb{x}\left(\mathbb{x}'\boldsymbol{\pi}^{-1}\mathbf{R}\mathbb{x}\right)^{-1}\mathbb{x}'\boldsymbol{\pi}^{-1}\mathbf{R}\right)\begin{bmatrix}-1_n\\1_n\end{bmatrix}\\
=& f\widehat{b}^{2GR!} + c\left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(-)}\mathbb{x}\mathbf{d}\left(\boldsymbol{\pi}^{-1}\mathbf{R}\begin{bmatrix}-1_n\\1_n\end{bmatrix} - \left(\boldsymbol{\pi}^{-1}\mathbf{R}-\mathbf{i}\right)\begin{bmatrix}-1_n\\1_n\end{bmatrix}\right)\\
=& f\widehat{b}^{2GR!} + c\left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(-)}\mathbb{x}\mathbf{d}\begin{bmatrix}-1_n\\1_n\end{bmatrix}
\end{aligned}$$

□

Proof of Lemma 5.3. The proof consists of two parts: first proving that all members of the set in (29) are solutions and, second, showing that all solutions are in the set.

First note that the fact that (22) is a solution to (23) implies that $(\mathbb{x}'\mathbf{d}\mathbb{x})(\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}\mathbb{x}'\mathbf{d}y = \mathbb{x}'\mathbf{d}y$. Next, premultiplying (28) by $(\mathbb{x}'\mathbf{d}\mathbb{x})$ yields

$$\begin{aligned}
(\mathbb{x}'\mathbf{d}\mathbb{x})b^{opt,z} =& (\mathbb{x}'\mathbf{d}\mathbb{x})\left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(-)}\mathbb{x}'\mathbf{d}y + (\mathbb{x}\mathbf{d}\mathbb{x})\left(\mathbf{i}_l - (\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}(\mathbb{x}'\mathbf{d}\mathbb{x})\right)z\\
\implies (\mathbb{x}'\mathbf{d}\mathbb{x})b^{opt,z} =& \mathbb{x}'\mathbf{d}y + \left((\mathbb{x}'\mathbf{d}\mathbb{x}) - (\mathbb{x}'\mathbf{d}\mathbb{x})(\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}(\mathbb{x}'\mathbf{d}\mathbb{x})\right)z\\
\implies (\mathbb{x}'\mathbf{d}\mathbb{x})b^{opt,z} =& \mathbb{x}'\mathbf{d}y
\end{aligned}$$

Hence, $b^{opt,z}$ is a solution to (23). This proves that all members of the set given by (29) are solutions.

Next, to prove that all solutions are in the set given by (29), let $b^{opt,*}$ represent a an arbitrary solution to (23) and then set $z = b^{opt,*}$. Then

$$\begin{aligned}
b^{opt,z} =& \left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(-)}\mathbb{x}'\mathbf{d}y + \left(\mathbf{i}_l - (\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}(\mathbb{x}'\mathbf{d}\mathbb{x})\right)b^{opt,*}\\
=& \left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(-)}\mathbb{x}'\mathbf{d}y + \left(b^{opt,*} - (\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}(\mathbb{x}'\mathbf{d}\mathbb{x})b^{opt,*}\right).\\
=& \left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(-)}\mathbb{x}'\mathbf{d}y + \left(b^{opt,*} - (\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}\mathbb{x}'\mathbf{d}y\right)\\
=& b^{opt,*}.
\end{aligned}$$

Hence, all solutions are represented in the set given by (29).

□

Proof of Lemma 5.6. By the definition of $\mathbf{d}_{11}$ above, in a completely randomized design the diagonal elements of $\mathbf{d}_{11}$ are

$$\begin{aligned}
\frac{\pi_{1i} - \pi_{1i}\pi_{1i}}{\pi_{1i}\pi_{1i}} =& \frac{\frac{n_1}{n} - \frac{n_1}{n}\frac{n_1}{n}}{\frac{n_1}{n}\frac{n_1}{n}}\\
=& \frac{n - n_1}{n_1}\\
=& \frac{n_0}{n_1}
\end{aligned}$$

and off-diagonal elements

$$\begin{aligned}\frac{\pi_{1i1j} - \pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}} =& \frac{\frac{n_1}{n}\frac{n_1-1}{n-1} - \frac{n_1}{n}\frac{n_1}{n}}{\frac{n_1}{n}\frac{n_1}{n}} \\ =& -\frac{1}{n-1}\frac{n_0}{n_1}.\end{aligned}$$

So if we define

$$\mathbf{d}_{11}^* = \frac{n_1(n-1)}{n_0 n}\mathbf{d}_{11}$$

then $\mathbf{d}_{11}^*$ has diagonal elements $\frac{n-1}{n}$ and off-diagonals $-\frac{1}{n}$, so that we can see that $\mathbf{d}_{11}^*\tilde{\mathbf{x}} = \tilde{\mathbf{x}} - 1_n\mu_{\tilde{\mathbf{x}}}$ returns the de-meaned $\tilde{\mathbf{x}}$. Therefore, $\mathbf{d}_{11}$ is a matrix that, when post-multiplied by $\tilde{\mathbf{x}}$, returns a de-meaned $\tilde{\mathbf{x}}$ that has been multiplied by the constant $\frac{nn_0}{(n-1)n_1}$. The proofs for $\mathbf{d}_{00}\tilde{\mathbf{x}}$, $\mathbf{d}_{01}\tilde{\mathbf{x}}$ and $\mathbf{d}_{01}\tilde{\mathbf{x}}$ are analogous. □

Proof of Theorem 5.9. To see that $\widehat{b}_{\text{II}}^{ols}$ estimates b_{II}^{ols} note that

$$\begin{aligned}\mathrm{E}\left[\mathbb{x}_{\text{II}}'\mathbf{R}\mathbb{x}_{\text{II}}\right] &= \begin{bmatrix} -1_n' & 0_n' \\ -\mathbf{x}' & 0_{(\frac{k}{2}-1)\times n}' \\ 0_n' & 1_n' \\ 0_{(\frac{k}{2}-1)\times n}' & \mathbf{x}' \end{bmatrix}\begin{bmatrix} -\frac{n_0}{n}1_n & -\frac{n_0}{n}\mathbf{x} & 0_n & 0_{(\frac{k}{2}-1)\times n} \\ 0_n & 0_{(\frac{k}{2}-1)\times n} & \frac{n_1}{n}1_n & \frac{n_1}{n}\mathbf{x}\end{bmatrix} \\ &= \begin{bmatrix} n_0 & 0 & 0 & 0 \\ 0 & n_0\mathrm{Var}(\mathbf{x}) & 0 & 0 \\ 0 & 0 & n_1 & 0 \\ 0 & 0 & 0 & n_1\mathrm{Var}(\mathbf{x})\end{bmatrix}\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}\left[\mathbb{x}_{\text{II}}'\mathbf{R}y\right] &= \begin{bmatrix} -1_n' & 0_n' \\ -\mathbf{x}' & 0_{(\frac{k}{2}-1)\times n}' \\ 0_n' & 1_n' \\ 0_{(\frac{k}{2}-1)\times n}' & \mathbf{x}' \end{bmatrix}\begin{bmatrix} -\frac{n_0}{n}y_0 \\ \frac{n_1}{n}y_1\end{bmatrix} \\ &= \begin{bmatrix} n_0\mu_{y_0} \\ n_0\mathrm{Cov}(\mathbf{x}, y_0) \\ n_1\mu_{y_1} \\ n_1\mathrm{Cov}(\mathbf{x}, y_1)\end{bmatrix}\end{aligned}$$

so that

$$\begin{aligned}\mathrm{E}\left[\mathbb{x}_{\text{II}}'\mathbf{R}\mathbb{x}_{\text{II}}\right]^{-1}\mathrm{E}\left[\mathbb{x}_{\text{II}}'\mathbf{R}y\right] &= \begin{bmatrix} n_0 & 0 & 0 & 0 \\ 0 & n_0\mathrm{Var}(\mathbf{x}) & 0 & 0 \\ 0 & 0 & n_1 & 0 \\ 0 & 0 & 0 & n_1\mathrm{Var}(\mathbf{x})\end{bmatrix}^{-1}\begin{bmatrix} n_0\mu_{y_0} \\ n_0\mathrm{Cov}(\mathbf{x}, y_0) \\ n_1\mu_{y_1} \\ n_1\mathrm{Cov}(\mathbf{x}, y_1)\end{bmatrix} \\ &= \begin{bmatrix} n_0^{-1} & 0 & 0 & 0 \\ 0 & n_0^{-1}\mathrm{Var}(\mathbf{x})^{-1} & 0 & 0 \\ 0 & 0 & n_1^{-1} & 0 \\ 0 & 0 & 0 & n_1^{-1}\mathrm{Var}(\mathbf{x})^{-1}\end{bmatrix}^{-1}\begin{bmatrix} n_0\mu_{y_0} \\ n_0\mathrm{Cov}(\mathbf{x}, y_0) \\ n_1\mu_{y_1} \\ n_1\mathrm{Cov}(\mathbf{x}, y_1)\end{bmatrix} \\ &= \begin{bmatrix} \mu_{y_0} \\ \mathrm{Var}(\mathbf{x})^{-1}\mathrm{Cov}(\mathbf{x}, y_0) \\ \mu_{y_1} \\ \mathrm{Var}(\mathbf{x})^{-1}\mathrm{Cov}(\mathbf{x}, y_1)\end{bmatrix}\end{aligned}$$

Hence under suitable regularity conditions $\widehat{b}_{\text{II}}^{ols} \to b_{\text{II}}^{ols}$ so that $\widehat{\delta}^{GR}(\widehat{b}_{\text{II}}^{ols})$ is asymptotically optimal. □

Proof of Theorem 5.14. First,

$$\begin{aligned}\mathrm{E}\left[\mathbb{x}_{\mathrm{I}}'\mathbf{R}\left(\boldsymbol{\pi}^{-1}-\mathbf{i}_{2n}\right)\mathbb{x}_{\mathrm{I}}\right] =&\mathbb{x}_{\mathrm{I}}'\boldsymbol{\pi}\left(\boldsymbol{\pi}^{-1}-\mathbf{i}_{2n}\right)\mathbb{x}_{\mathrm{I}}\\ =&\mathbb{x}_{\mathrm{I}}'\left(\mathbf{i}_{2n}-\boldsymbol{\pi}\right)\mathbb{x}_{\mathrm{I}}\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}\left[\mathbb{x}_{\mathrm{I}}'\mathbf{R}\left(\boldsymbol{\pi}^{-1}-\mathbf{i}_{2n}\right)y\right] =&\mathbb{x}_{\mathrm{I}}'\boldsymbol{\pi}\left(\boldsymbol{\pi}^{-1}-\mathbf{i}_{2n}\right)y\\ =&\mathbb{x}_{\mathrm{I}}'\left(\mathbf{i}_{2n}-\boldsymbol{\pi}\right)y\end{aligned}$$

so that

$$\mathrm{E}\left[\mathbb{x}_{\mathrm{I}}'\mathbf{R}\left(\boldsymbol{\pi}^{-1}-\mathbf{i}_{2n}\right)\mathbb{x}_{\mathrm{I}}\right]^{-1}\mathrm{E}\left[\mathbb{x}_{\mathrm{I}}'\mathbf{R}\left(\boldsymbol{\pi}^{-1}-\mathbf{i}_{2n}\right)y\right]=\left(\mathbb{x}_{\mathrm{I}}'\left(\mathbf{i}_{2n}-\boldsymbol{\pi}\right)\mathbb{x}_{\mathrm{I}}\right)^{-1}\mathbb{x}_{\mathrm{I}}'\left(\mathbf{i}_{2n}-\boldsymbol{\pi}\right)y$$

which is just the coefficient given in Definition 5.10. Thus, under suitable regularity conditions $\widehat{b}_{\mathrm{I}}^{tyr}\to b_{\mathrm{I}}^{tyr}$ so that its conjugate ATE estimator is asymptotically optimal. □

Proof of Theorem 5.15. First, let $\mathbb{x}_{\mathrm{II}}^{*}$ be an equivalent specification to specification II defined as

$$\mathbb{x}_{\mathrm{II}}^{*}=\begin{bmatrix}-1_n & 0_n & -\mathbf{x} & 0_{(n\times k)}\\ 0_n & 1_n & 0_{(n\times k)} & \mathbf{x}\end{bmatrix}.$$

Then,

$$\begin{aligned}\mathbb{x}_{\mathrm{II}}^{*\prime}\tilde{\mathbf{d}}\mathbb{x}_{\mathrm{II}}^{*} =&\mathbb{x}_{\mathrm{II}}^{*\prime}\mathbf{d}\mathbb{x}_{\mathrm{II}}^{*}+\mathbb{x}_{\mathrm{II}}^{*\prime}\begin{bmatrix}\mathbf{i} & \mathbf{i}\\ \mathbf{i} & \mathbf{i}\end{bmatrix}\mathbb{x}_{\mathrm{II}}^{*}\\ =&\begin{bmatrix}n & -n & 0 & 0\\ -n & n & 0 & 0\\ 0 & 0 & \mathbf{x}'\tilde{\mathbf{d}}_{00}\mathbf{x} & -\mathbf{x}'\tilde{\mathbf{d}}_{01}\mathbf{x}\\ 0 & 0 & -\mathbf{x}'\tilde{\mathbf{d}}_{10}\mathbf{x} & \mathbf{x}'\tilde{\mathbf{d}}_{11}\mathbf{x}\end{bmatrix}\end{aligned}$$

Now recall that $\mathbf{x}$ is zero-centered and using Lemma 5.7 we have for a completely randomized design

$$\begin{aligned}\mathbf{x}'\tilde{\mathbf{d}}_{00}\mathbf{x} =&\mathbf{x}'\mathbf{d}_{00}\mathbf{x}+\mathbf{x}'\mathbf{x}\\ =&\frac{n^2n_1}{(n-1)n_0}\mathrm{Var}(\mathbf{x})+n\mathrm{Var}(\mathbf{x})\\ =&c_a\mathrm{Var}(\mathbf{x})\end{aligned}$$

where $c_a:=\frac{n^2n_1+n(n-1)n_0}{(n-1)n_0}$. Likewise,

$$\begin{aligned}\mathbf{x}'\tilde{\mathbf{d}}_{11}\mathbf{x} =&c_b\mathrm{Var}(\mathbf{x}),\\ -\mathbf{x}'\tilde{\mathbf{d}}_{01}\mathbf{x} =&c_c\mathrm{Var}(\mathbf{x}),\\ \text{and}\quad -\mathbf{x}'\tilde{\mathbf{d}}_{10}\mathbf{x} =&c_c\mathrm{Var}(\mathbf{x})\end{aligned}$$

with $c_b:=\frac{n^2n_0+n(n-1)n_1}{(n-1)n_1}$ and $c_c:=\frac{-n^2+n-1}{(n-1)}$. Next, letting $c_q:=c_b-c_c^2c_a^{-1}$ and given that a generalized inverse of a partitioned matrix is given in (31),

$$\left(\mathbb{x}_{\mathrm{II}}^{*\prime}\tilde{\mathbf{d}}\mathbb{x}_{\mathrm{II}}^{*}\right)^{(g)}=\mathrm{Bdiag}\left(\begin{bmatrix}n & -n\\ -n & n\end{bmatrix}^{(-)},\begin{bmatrix}c_a^{-1}+c_a^{-2}c_c^2c_q^{-1} & -c_a^{-1}c_cc_q^{-1}\\ -c_a^{-1}c_cc_q^{-1} & c_q^{-1}\end{bmatrix}\otimes\mathrm{Var}(\mathbf{x})^{(-)}\right)$$

where Bdiag $(\mathbf{a}, \mathbf{b})$ makes a block diagonal matrix out of matrices $\mathbf{a}$ and $\mathbf{b}$ and $\otimes$ is the Kronecker product. Similarly,

$$\begin{aligned}
\mathbf{x}'\tilde{\mathbf{d}}_{00}y_0 =& c_a \text{Cov}(\mathbf{x}, y_0) \\
\mathbf{x}'\tilde{\mathbf{d}}_{11}y_1 =& c_b \text{Cov}(\mathbf{x}, y_1), \\
-\mathbf{x}'\tilde{\mathbf{d}}_{01}y_1 =& c_c \text{Cov}(\mathbf{x}, y_1), \\
\text{and} \quad -\mathbf{x}'\tilde{\mathbf{d}}_{10}y_0 =& c_c \text{Cov}(\mathbf{x}, y_0),
\end{aligned}$$

so that

$$\mathbb{x}_{\text{II}}'\tilde{\mathbf{d}}y = \begin{bmatrix} -1_{2n}'y \\ 1_{2n}'y \\ \begin{bmatrix} c_a \\ c_c \end{bmatrix} \otimes \text{Cov}(\mathbf{x}, y_0) + \begin{bmatrix} c_c \\ c_b \end{bmatrix} \otimes \text{Cov}(\mathbf{x}, y_1) \end{bmatrix}.$$

Therefore,

$$\left(\mathbb{x}_{\text{II}}^{*\prime}\tilde{\mathbf{d}}\mathbb{x}_{\text{II}}^{*}\right)^{(g)} \mathbb{x}_{\text{II}}^{*\prime}\tilde{\mathbf{d}}y = \begin{bmatrix} \begin{bmatrix} n & -n \\ -n & n \end{bmatrix}^{(-)} \begin{bmatrix} -1_{2n}'y \\ 1_{2n}'y \end{bmatrix} \\ \left(\begin{bmatrix} c_a^{-1} + c_a^{-2}c_c^2c_q^{-1} & -c_a^{-1}c_cc_q^{-1} \\ -c_a^{-1}c_cc_q^{-1} & c_q^{-1} \end{bmatrix} \otimes \text{Var}(\mathbf{x})^{(-)} \right) \left(\begin{bmatrix} c_a \\ c_c \end{bmatrix} \otimes \text{Cov}(\mathbf{x}, y_0) + \begin{bmatrix} c_c \\ c_b \end{bmatrix} \otimes \text{Cov}(\mathbf{x}, y_1) \right) \end{bmatrix}.$$

Focusing on the last $2k$ coefficients we have,

$$\begin{aligned}
&\left(\begin{bmatrix} c_a^{-1} + c_a^{-2}c_c^2c_q^{-1} & -c_a^{-1}c_cc_q^{-1} \\ -c_a^{-1}c_cc_q^{-1} & c_q^{-1} \end{bmatrix} \otimes \text{Var}(\mathbf{x})^{(g)} \right) \left(\begin{bmatrix} c_a \\ c_c \end{bmatrix} \otimes \text{Cov}(\mathbf{x}, y_0) + \begin{bmatrix} c_c \\ c_b \end{bmatrix} \otimes \text{Cov}(\mathbf{x}, y_1) \right) \\
&\quad = \left(\begin{bmatrix} c_a^{-1} + c_a^{-2}c_c^2c_q^{-1} & -c_a^{-1}c_cc_q^{-1} \\ -c_a^{-1}c_cc_q^{-1} & c_q^{-1} \end{bmatrix} \begin{bmatrix} c_a \\ c_c \end{bmatrix} \right) \otimes \text{Var}(\mathbf{x})^{(-)}\text{Cov}(\mathbf{x}, y_0) \\
&\qquad + \left(\begin{bmatrix} c_a^{-1} + c_a^{-2}c_c^2c_q^{-1} & -c_a^{-1}c_cc_q^{-1} \\ -c_a^{-1}c_cc_q^{-1} & c_q^{-1} \end{bmatrix} \begin{bmatrix} c_c \\ c_b \end{bmatrix} \right) \otimes \text{Var}(\mathbf{x})^{(-)}\text{Cov}(\mathbf{x}, y_1) \\
&\quad = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \otimes \text{Var}(\mathbf{x})^{(-)}\text{Cov}(\mathbf{x}, y_0) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \otimes \text{Var}(\mathbf{x})^{(-)}\text{Cov}(\mathbf{x}, y_1) \\
&\quad = \begin{bmatrix} \text{Var}(\mathbf{x})^{(-)}\text{Cov}(\mathbf{x}, y_0) \\ \text{Var}(\mathbf{x})^{(-)}\text{Cov}(\mathbf{x}, y_1) \end{bmatrix}.
\end{aligned}$$

The first equality follows from the mixed-product property of Kronecker products. The following line applies algebra and the definition of c_q. As long as there is no perfect collinearity in $\mathbf{x}$, $\text{Var}(\mathbf{x})^{(-)}$ represents the usual inverse matrix. The intercept coefficients are

$$\begin{bmatrix} n & -n \\ -n & n \end{bmatrix}^{(-)} \begin{bmatrix} -1_{2n}'y \\ 1_{2n}'y \end{bmatrix} = \frac{1}{2} \begin{bmatrix} -\delta \\ \delta \end{bmatrix},$$

but recognizing that the choice of generalized inverse was arbitrary, it can be seen that the full range of optimal intercepts includes

$$\begin{bmatrix} \mu_{y_0} \\ \mu_{y_1} \end{bmatrix}.$$

□

Proof of Lemma 6.3. By the definition of $\mathbf{d}_{11}$ above, in a cluster randomized designs the ij element of $\mathbf{d}_{11}$ when units i and j are in the same cluster is

$$\begin{aligned}\frac{\pi_{1i1j}-\pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}} =& \frac{\frac{m_1}{m}-\frac{m_1}{m}\frac{m_1}{m}}{\frac{m_1}{m}\frac{m_1}{m}}\\ =& \frac{m-m_1}{m_1}\\ =& \frac{m_0}{m_1}\end{aligned}$$

and for i, j not in the same cluster

$$\begin{aligned}\frac{\pi_{1i1j}-\pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}} =& \frac{\frac{m_1}{m}\frac{m_1-1}{m-1}-\frac{m_1}{m}\frac{m_1}{m}}{\frac{m_1}{m}\frac{m_1}{m}}\\ =& -\frac{1}{m-1}\frac{m_0}{m_1}.\end{aligned}$$

Now define

$$\mathbf{d}_{11}^{*}=\frac{m_1(m-1)}{m_0 m}\mathbf{d}_{11}$$

then $\mathbf{d}_{11}^{*}$ has i,j element equal to $\frac{m-1}{m}$ if i and j are in the same cluster and equal to $-\frac{1}{m}$ otherwise. So, $\mathbf{d}_{11}^{*}\tilde{\mathbf{x}}$ returns a length n vector $(\tilde{\mathbf{x}}_n^c-\frac{n}{m}1_n\mu_{\tilde{\mathbf{x}}})$ with the i^{th} row of $\tilde{\mathbf{x}}_n^c$ equal to the sums of x's for cluster c_i and with $\frac{n}{m}1_n\mu_{\tilde{\mathbf{x}}}$ doing the work of subtracting off the average of cluster totals. Therefore, $\mathbf{d}_{11}\tilde{\mathbf{x}}=\frac{mm_0}{(m-1)m_1}(\tilde{\mathbf{x}}_n^c-\frac{n}{m}1_n\mu_{\tilde{\mathbf{x}}})$. The proofs for $\mathbf{d}_{00}\tilde{\mathbf{x}}$, $\mathbf{d}_{01}\tilde{\mathbf{x}}$ and $\mathbf{d}_{01}\tilde{\mathbf{x}}$ are analogous. □

Proof of Lemma 6.4. Write

$$\begin{aligned}\tilde{\mathbf{x}}'\mathbf{d}_{11}\tilde{\mathbf{x}} =& \frac{mm_0}{(m-1)m_1}\tilde{\mathbf{x}}'\left(\tilde{\mathbf{x}}_n^c-\frac{n}{m}1_n\mu_{\tilde{\mathbf{x}}}\right)\\ =& \frac{mm_0}{(m-1)m_1}\tilde{\mathbf{x}}_m^{c\prime}\left(\tilde{\mathbf{x}}_m^c-\frac{n}{m}1_m\mu_{\tilde{\mathbf{x}}}\right)\\ =& \frac{m^2m_0}{(m-1)m_1}\mathrm{Var}(\tilde{\mathbf{x}}_m^c)\end{aligned}$$

where $\tilde{\mathbf{x}}_m^c$ is an $m\times(k-1)$ vector (one row per cluster) with the g^{th} row representing cluster totals of the rows of $\tilde{\mathbf{x}}$ associated with members of the g^{th} cluster. □

Proof of Theorem 6.11.

$$\begin{aligned}\mathrm{E}\left[\mathbb{x}_{\mathrm{I}}'\mathbf{R}\left(\boldsymbol{\pi}^{-1}-\mathbf{i}_{2n}\right)\mathbb{x}_{\mathrm{I}}\right] =& \mathbb{x}_{\mathrm{I}}'\boldsymbol{\pi}\left(\boldsymbol{\pi}^{-1}-\mathbf{i}_{2n}\right)\mathbb{x}_{\mathrm{I}}\\ =& \mathbb{x}_{\mathrm{I}}'\left(\mathbf{i}_{2n}-\boldsymbol{\pi}\right)\mathbb{x}_{\mathrm{I}}\\ =& \begin{bmatrix}0' & 1'\\ -1' & 1'\\ -\mathbf{x}' & \mathbf{x}'\end{bmatrix}\left(\mathbf{i}_{2n}-\boldsymbol{\pi}\right)\begin{bmatrix}0 & -1 & -\mathbf{x}\\ 1 & 1 & \mathbf{x}\end{bmatrix}\\ =& \begin{bmatrix}0' & 1'\\ -1' & 1'\\ -\mathbf{x}' & \mathbf{x}'\end{bmatrix}\begin{bmatrix}0 & -\frac{n_1}{n}1 & -\frac{n_1}{n}\mathbf{x}\\ \frac{n_0}{n}1 & \frac{n_0}{n}1 & \frac{n_0}{n}\mathbf{x}\end{bmatrix}\\ =& \begin{bmatrix}n_0 & n_0 & 0\\ n_0 & n & 0\\ 0 & 0 & n\mathrm{Var}(\mathbf{x})\end{bmatrix}\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left[\mathbb{x}_{\mathrm{I}}'\mathbf{R}\left(\boldsymbol{\pi}^{-1}-\mathbf{i}_{2n}\right)y\right] =& \mathbb{x}_{\mathrm{I}}'\boldsymbol{\pi}\left(\boldsymbol{\pi}^{-1}-\mathbf{i}_{2n}\right)y \\
=& \mathbb{x}_{\mathrm{I}}'\left(\mathbf{i}_{2n}-\boldsymbol{\pi}\right)\mathbb{x}_{\mathrm{I}} \\
=& \begin{bmatrix} 0' & 1' \\ -1' & 1' \\ -\mathbf{x}' & \mathbf{x}' \end{bmatrix}\left(\mathbf{i}_{2n}-\boldsymbol{\pi}\right)\begin{bmatrix} -y_0 \\ y_1 \end{bmatrix} \\
=& \begin{bmatrix} 0' & 1' \\ -1' & 1' \\ -\mathbf{x}' & \mathbf{x}' \end{bmatrix}\begin{bmatrix} -\frac{n_1}{n}y_0 \\ \frac{n_0}{n}y_1 \end{bmatrix} \\
=& \begin{bmatrix} n_0\mu_{y_1} \\ n_1\mu_{y_0}+n_0\mu_{y_1} \\ n_1\mathrm{Cov}(\mathbf{x},y_0)+n_0\mathrm{Cov}(\mathbf{x},y_1) \end{bmatrix}
\end{aligned}$$

so that

$$\begin{aligned}
&\mathrm{E}\left[\mathbb{x}_{\mathrm{I}}'\mathbf{R}\left(\boldsymbol{\pi}^{-1}-\mathbf{i}_{2n}\right)\mathbb{x}_{\mathrm{I}}\right]^{-1}\mathrm{E}\left[\mathbb{x}_{\mathrm{I}}'\mathbf{R}\left(\boldsymbol{\pi}^{-1}-\mathbf{i}_{2n}\right)y\right] \\
&\qquad= \begin{bmatrix} n_0 & n_0 & 0 \\ n_0 & n & 0 \\ 0 & 0 & n\mathrm{Var}(\mathbf{x}) \end{bmatrix}^{-1}\begin{bmatrix} n_0\mu_{y_1} \\ n_1\mu_{y_0}+n_0\mu_{y_1} \\ n_1\mathrm{Cov}(\mathbf{x},y_0)+n_0\mathrm{Cov}(\mathbf{x},y_1) \end{bmatrix} \\
&\qquad= \begin{bmatrix} nn_1^{-1}n_0^{-1} & -n_1^{-1} & 0 \\ -n_1^{-1} & n_1^{-1} & 0 \\ 0 & 0 & n^{-1}\mathrm{Var}(\mathbf{x})^{-1} \end{bmatrix}\begin{bmatrix} n_0\mu_{y_1} \\ n_1\mu_{y_0}+n_0\mu_{y_1} \\ n_1\mathrm{Cov}(\mathbf{x},y_0)+n_0\mathrm{Cov}(\mathbf{x},y_1) \end{bmatrix} \\
&\qquad= \begin{bmatrix} \delta \\ \mu_{y_0} \\ \frac{n_1}{n}\mathrm{Var}(\mathbf{x})^{-1}\mathrm{Cov}(\mathbf{x},y_0)+\frac{n_0}{n}\mathrm{Var}(\mathbf{x})^{-1}\mathrm{Cov}(\mathbf{x},y_1) \end{bmatrix}.
\end{aligned}$$

Thus, under suitable regularity conditions $\widehat{b}_{\mathrm{I}}^{tyr} \to b_{\mathrm{I}}^{tyr}$ so that $\widehat{\delta}^{GR}(\widehat{b}_{\mathrm{I}}^{tyr})$ is asymptotically optimal. □