A Unified Theory of Regression Adjustment for Design-based Inference

WORKING PAPER

Joel A. Middleton*

July 3, 2018

Abstract

Under the Neyman causal model, it is well-known that OLS with treatment-by-covariate interactions cannot harm asymptotic precision of estimated treatment effects in completely randomized experiments. But do such guarantees extend to experiments with more complex designs? This paper proposes a general framework for addressing this question and defines a class of generalized regression estimators that are applicable to experiments of virtually any design. The class subsumes common estimators (e.g., OLS). Within that class, two novel estimators are proposed that are applicable to arbitrary designs and asymptotically optimal. The first is composed of three Horvitz-Thompson estimators. The second recursively applies the principle of generalized regression estimation to obtain regression-adjusted regression adjustment. Additionally, variance bounds are derived that are tighter than those existing in the literature for arbitrary designs. Finally, a simulation study illustrates the potential for MSE improvements. Applications could include cluster-randomized experiments and experiments in networks where assignment to an exposure condition depends on the network structure.

1 Introduction

In the analysis of randomized experiments, regression adjustment is common. Even though its classical assumptions are not justified under Neyman's (1923) causal model (see Freedman, 2008a,b), regression's finite population and asymptotic properties may be nonetheless defensible. For completely randomized experiments for example, Lin (2013) demonstrates that asymptotic precision can not be harmed by ordinary least squares (OLS) regression if treatment-by-covariate interactions are included in the specification. And inference with heteroskedastic-consistent standard errors is asymptotically conservative.

Meanwhile, regression-type adjustments have been recommended for a variety of designs other than complete randomization. Within the framework of the Neyman-Rubin causal model, authors have considered adjustments for cluster-randomized designs (Hansen and Bowers, 2009; Middleton and Aronow, 2015), block-randomized designs (Athey and Imbens, 2017), arbitrary designs (Aronow and Middleton, 2015), two-stage designs with interference (Basse and Feller, 2017; Sinclair et al., 2012), arbitrary/complex designs with interference (Aronow and Samii, 2017), and factorial designs (Lu, 2016). Athey and Imbens (2017) consider block-, cluster- and pair-randomized designs.

With the notable exception of results for completely randomized designs (Lin, 2013; Bloniarz et al., 2016), however, few recommendations for practice have relied on guarantees of precision gains or claims of optimality under the Neyman-Rubin model, asymptotic or otherwise. Instead, justifications have tended to be heuristic or have relied on "model assisted" arguments. And while heuristic arguments that regression should help

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

more often than it hurts may be compelling, some amount of analysis is warranted to address concerns about regression adjustments for arbitrary designs of the sort originally raised by Freedman (2008a,b).

As such, a general framework for analysis of regression's properties for any design may be in order. This paper makes several contributions in that direction. First, a proposed framework allows for any design so that expressions can be maximally general. The re-representation of notation will attempt to make derivations compact and straightforward where they would otherwise be difficult. The benefit will be realized especially when expressions for variance are presented and when they are manipulated for the purpose of deriving expressions for optimal regression adjustment. Moreover, a benefit is realized when variance bounding is defined, allowing bounds to be derived that can be tighter than those existing in the literature for arbitrary designs.

Second, the proposed class of generalized regression estimators, analogous to estimators in the sampling literature, are applicable to arbitrary designs. It will be demonstrated that the class subsumes common regression practice. Future research may take as a starting point variations on this class, but the proposed class is conceptually useful.

Third, the paper proposes two asymptotically optimal regression estimators that can be applied to arbitrary designs. To develop them, the two key estimation principles presented in the paper, namely, the Horvitz-Thompson principle (inverse propensity of treatment weighting) and the generalized regression principle, are applied recursively. This yields one asymptotically optimal estimator consisting of three Horvitz-Thompson estimators and another that is regression-adjusted regression adjustment. While the proposed estimators can have MSE higher than OLS in small samples, the latter could may be useful in larger experiments with complex designs.An application could be, for example, experiments in networks where "exposure" conditions are determined by a complex network structure (e.g., Aronow and Samii, 2017). Cluster-randomized experiments with many clusters may be another application.

Fourth, the paper illustrates how the proposed framework might be used for analysis of specific designs, in particular, completely randomized and cluster randomized designs. In the case of completely randomized designs, results are known (see Lin, 2013). However, the proofs are novel, and, as such, they serve as a bridge between existing results and the proposed framework. Moreover, results provided may be useful in the analysis of designs not considered here.

1.1 Plan of the paper

In Section 2 the overall framework is presented. It will establish the context and notation for the discussion of regression adjustment. In Section 3 a generalized regression estimator is introduced and its connection to regression as commonly used is clarified. In Section 4 two asymptotically optimal estimators that can be applied to any design are proposed. Section 5 develops variance estimation for HT and generalized regression estimators. In two subsequent sections, results for specific designs are derived as special cases within the established framework. The specific designs considered are complete randomization and cluster randomization.[1] Section 8 uses simulation to illustrate the potential for reductions in MSE using asymptotically optimal regression adjustment.

2 Framework

This section establishes the overall framework which will be necessary to develop generalized regression adjustment for arbitrary designs.[2] In the next subsection, average treatment effect (ATE), the causal quantity of interest, is defined in the context of the Neyman-Rubin causal model (NRCM). A two-arm experiment is assumed. While generalization to multiple arms is straightforward, the added notation distracts from

[1] A few comments on block randomization are given in the discussion.

[2] Throughout, it should be understood that "arbitrary designs," is shorthand for "pretty much any design within reasonable limits." Limitations include that the design must be identified, i.e., every unit must have some chance of being in treatment and some chance of being in control. Additional limitations are required to ensure asymptotic properties. For example, a design where the treatment group has a fixed number of units as $n \to \infty$ is outside these limits. Within those limits, the framework is as general as possible.

the core developments. Throughout the paper, footnotes will attempt to highlight insights about multi-arm extensions.

In the second subsection, the HT estimator is formally introduced. The HT estimator serves as foundational estimator to which regression adjustments will be made. The estimator has the virtue of being unbiased for the ATE for any identified design, a property which implies that, asymptotically speaking, consistency requires only that its variance goes to zero. By contrast, an alternative estimator such as the difference-of-means could be badly biased and lack consistency in designs where assignment probabilities are unequal. Moreover, limitations of the HT estimator, namely imprecision and a lack of location invariance, will be addressed by the regression adjustment.[3] Importantly, the variance of the HT estimator will be discussed. As it will be demonstrated later, this variance expression is directly relevant to variance expressions for the generalized regression estimator. The topic of variance *estimation* will be postponed until Section 5, when an overall strategy can be presented for both HT estimators and generalized regression estimators.

2.1 The Neyman Causal Model for treatment-control experiments

Consider the Neyman-Rubin causal model (NRCM) 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 parameter of interest is the average treatment effect (ATE), which can be written

$$\delta := n^{-1}\sum_i \left(y_{1i} - y_{0i}\right).$$

The "fundamental problem of causal inference" (Holland, 1986) is that only one of the two potential outcomes can be observed for each unit.

For ease of derivations, it makes sense to re-represent the problem in a nonstandard way starting with the potential outcomes. First, represent the entire schedule of potential outcomes as the vector

$$y := \begin{bmatrix} -y_{01} \\ -y_{02} \\ \vdots \\ -y_{0n} \\ y_{11} \\ y_{12} \\ \vdots \\ y_{1n} \end{bmatrix}$$

and note that y has length $2n$ and that the control potential outcomes are multiplied by -1. This allows the ATE to be represented equivalently as

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

[3]Another obvious alternative would be to take Hajek estimator as foundational, but the random denominator introduces an inelegance that impedes the illustration of principles. It is also provably true that (under regularity conditions) generalized regression estimators that are HT-based obtain the equal asymptotic variance to Hajek-based counterparts. That said, for smaller samples, the development of a Hajek-based generalized regression estimator might be a worthwhile refinement. However, it is beyond the scope of this paper.

where 1_{2n} is a column vector of $2n$ ones.[4] This "stacked" representation of the potential outcomes allows for mathematical manipulation of a single entity, where the canonical frameworks requires manipulation of two entities. The benefits become evident in variance expressions and as covariates are added to the picture.

If, as above, R_{1i} is the (random) treatment assignment indicator for the i^{th} individual and defining the control assignment indicator as $R_{0i} := 1 - R_{1i}$, then the $2n \times 2n$ matrix

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

encodes on its diagonal which of the elements of y is observed. In matrix form, the researcher can be said to observe $\mathbf{R}$ and $\mathbf{R}y$, as well as an $n \times k$ matrix of covariates, $\mathbf{x}$, which will be assumed to be zero-centered. Zero-centering will make straightforward the interpretation of intercept terms in regression coefficients, but will otherwise not affect the analysis.

2.2 The Horvitz-Thompson Estimator

As an experiment, the researcher controls the mechanism that determines the distribution of $\mathbf{R}$. As such, the researcher is also able to compute quantities such as the probability of treatment assignment, $\pi_{1i} := \mathrm{E}[R_{1i}]$, and the probability of assignment to control, $\pi_{0i} := \mathrm{E}[R_{0i}]$ (either analytically or numerically to arbitrary precision). A $2n \times 2n$ diagonal matrix of assignment probabilities can be written

$$\boldsymbol{\pi} := \begin{bmatrix} \pi_{01} & & & & & & & \\ & \pi_{02} & & & & & & \\ & & \ddots & & & & & \\ & & & \pi_{0n} & & & & \\ & & & & \pi_{11} & & & \\ & & & & & \pi_{12} & & \\ & & & & & & \ddots & \\ & & & & & & & \pi_{1n} \end{bmatrix}.$$

Note that by definition $\boldsymbol{\pi} = \mathrm{E}[\mathbf{R}]$.

Next, the HT estimator of the average treatment effect can be defined as

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

where 1_{2n} is, as above, a column vector of $2n$ ones.[5,6] The diagonal matrix $\boldsymbol{\pi}^{-1}$ does the work of weighting the potential outcomes inversely proportional to the probability of being observed. As such, the HT estimator

[4] Alternatively, one could represent a contrast matrix explicitly, but at a notational cost. For example, one could define $\mathbf{c} = \begin{bmatrix} -\mathbf{i} & 0 \\ 0 & \mathbf{i} \end{bmatrix}$, where $\mathbf{i}$ is an $n \times n$ identity matrix, and write $\mathbf{c}y$ throughout instead of using the more compact y as defined above. Explicitly representing contrast matrices is useful when generalizing these results to more than two treatment arms.

[5] Equivalently, one could have written

$$\widehat{\delta}^{HT} = n^{-1} 1'_n \boldsymbol{\pi}_1^{-1} \mathbf{R}_1 y_1 - n^{-1} 1'_n \boldsymbol{\pi}_0^{-1} \mathbf{R}_0 y_0,$$

an expression that has the pedagogical benefit of clarifying that the estimator is a difference between two HT estimators, one for each treatment arm. However, the more compact convention above streamlines.

[6] To the extent possible, an attempt was made to adhere to the following notational conventions: upper case signifies random

has the virtue of being unbiased for any identified design, i.e., designs in which $0 < \pi_{1i} < 1$ for all i, because $\mathrm{E}[\mathbf{R}] = \boldsymbol{\pi}$.

An HT estimator is similar to an inverse propensity of treatment weighted (IPTW) estimator, but the "propensity score" is known in this setting by way of knowledge of the design of the experiment. Also, the estimator can sometimes be equivalent to the difference-of-means, such as in completely randomized designs.

2.3 Variance of the Horvitz-Thompson estimator

To express the variance of a HT estimator of the average treatment effect, first note that in equation (2), 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

$$\mathbf{d} := \mathrm{V}\left(1'_{2n}\boldsymbol{\pi}^{-1}\mathbf{R}\right) \tag{3}$$

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

$$\mathrm{V}\left(\widehat{\delta}^{HT}\right) = n^{-2}y'\mathbf{d}y. \tag{4}$$

Equivalent, though much more cumbersome expressions are given by Aronow and Middleton (2015) and Aronow and Samii (2017).[7] The compact representation of variance given here is essential to deriving a general expression for asymptotically optimal regression adjustment and results for variance bounds, below.

Insight into how to construct $\mathbf{d}$ in practice may arise from exploring its elements further. 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

$$\mathbf{d} = \begin{bmatrix} \mathbf{d}_{00} & \mathbf{d}_{01} \\ \mathbf{d}_{10} & \mathbf{d}_{11} \end{bmatrix} \tag{5}$$

where, for example, the matrix $\mathbf{d}_{11}$ has ij element $\frac{\pi_{1i1j}-\pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}}$ and the matrix $\mathbf{d}_{10}$ has ij element $\frac{\pi_{1i0j}-\pi_{1i}\pi_{0j}}{\pi_{1i}\pi_{0j}}$. Sub-matrices $\mathbf{d}_{00}$ and $\mathbf{d}_{01}$ are defined analogously. The equivalence in (5) will be useful below.

Since not all pairs of potential outcomes can be observed together for various reasons, not all terms in the quadratic in (4) are observable. Beginning with Neyman (1923), this has led to the idea of bounding the variance in the design-based paradigm. A general approach to variance bounds and their estimation for both HT and regression estimators will be described in Section 5.

variables (e.g., $R_{1i}, \mathbf{R}$); bold signifies a matrix (e.g., $\mathbf{R}, \mathbf{d}$); non-subscripted letters are vectors (e.g., y) or sometimes scalars (e.g., n, k, c); a subscript of 0 or 1 on a letter typically means the subvector associated with control or treatment outcomes, respectively, (e.g., y_0, y_1); letters with subscripts of i, j, k or l typically mean scalar elements of a vector or matrix (e.g., y_{1i}, π_{0i}, π_{1i1j}); and Greek typically signifies a quantity of interest or descriptive summary of a population characteristic (e.g., δ). An exception to this last rule, for reasons of tradition, is the use of π for assignment probability which is not a quantity of interest but a parameter set by the design.

[7] For example, Aronow and Middleton (2015) write the variance of the HT estimator as

$$\begin{aligned}\mathrm{V}\left(\widehat{\delta}^{HT}\right) = &\sum_i \frac{\pi_{1i}(1-\pi_{1i})}{\pi_{1i}\pi_{1i}} y_{1i}^2 + \sum_i \sum_{j\neq i} \frac{\pi_{1i1j}-\pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}} y_{1i}y_{1j} \\ &+ \sum_i \frac{\pi_{0i}(1-\pi_{0i})}{\pi_{0i}\pi_{0i}} y_{0i}^2 + \sum_i \sum_{j\neq i} \frac{\pi_{0i0j}-\pi_{0i}\pi_{0j}}{\pi_{0i}\pi_{0j}} y_{0i}y_{0j} \\ &- 2\sum_i y_{1i}y_{0i} - 2\sum_i \sum_{j\neq i} \frac{\pi_{1i0j}-\pi_{1i}\pi_{0j}}{\pi_{1i}\pi_{0j}} y_{1i}y_{0j}.\end{aligned}$$

3 Regression

Now that the framework has been established for estimation and variance estimation under the NCM, this section turns to the main subject of the paper, regression adjustment.

3.1 Covariate specifications

To discuss covariate specifications, it helps to move inductively from a well known example. First, let $\mathbf{x}$ be the zero-centered matrix of covariates with n rows and k columns. The covariates themselves are taken as given (i.e., the question of how to arrive at the right covariates, how to transform them, etc., is left to another paper) and fixed (i.e., nonrandom and 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. Using the current notational framework, this OLS coefficient estimator can be written

$$\widehat{b}_{\text{I}}^{ols} := \left(\mathbb{x}_{\text{I}}'\mathbf{R}\mathbb{x}_{\text{I}}\right)^{-1}\mathbb{x}_{\text{I}}'\mathbf{R}y \tag{6}$$

where

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

is a $2n \times (k+2)$ matrix. Note that (6) 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, similar to similar to Zhao et al. (2017+), the present formulation has the advantage that it separately represents the source of randomness ($\mathbf{R}$) and the fixed quantities ($\mathbb{x}_{\text{I}}$ and y). Note that for convenience in later derivations, the leading column of $\mathbb{x}_{\text{I}}$ 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 $\mathbb{x}_{\text{I}}$ are multiplied by -1, to mirror the definition of the vector y and thus ensuring that the elements of $\widehat{b}_{\text{I}}^{ols}$ have the expected signs. The subscript on matrix $\mathbb{x}_{\text{I}}$ 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 represent any arbitrary covariate specification. See also that $\widehat{b}_{\text{I}}^{ols}$ 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.

By contrast, "specification II" (or the "separate slopes" specification) is given by,

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

which is equivalent to including interactions between treatment and each covariate in $\mathbf{x}$. Lin (2013), for example, recommends 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 (6), 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 necessarily be directly interpretable in this fashion. Freeing regression coefficients from the burden of interpretable elements allows for the derivation of certain optimal estimators of the ATE that are not otherwise obvious.

3.2 Defining a class of generalized regression estimators

Three equivalent forms for the proposed class of generalized regression estimators are given in the definition below. Sampling theorists have the longest history with the idea of generalized regression and the constructions that follow.[8] Doubly robust estimators also use the form.[9] The literature on control functions has apparently reinvented the generalized regression estimator as well.

Definition 3.1 (Generalized regression estimators)**.** *Three equivalent forms for "generalized regression estimators" of the average treatment effect (ATE) are given by*

$$\widehat{\delta}^{R} := n^{-1}1'_{2n}(\boldsymbol{\pi}^{-1}\mathbf{R}y - \boldsymbol{\pi}^{-1}\mathbf{R}\mathbb{x}\widehat{b} + \mathbb{x}\widehat{b}) \tag{7a}$$

$$= \widehat{\delta}^{HT} - \widehat{\delta}^{HT}_{\mathbb{x}}\widehat{b} \tag{7b}$$

$$= n^{-1}1'_{2n}\boldsymbol{\pi}^{-1}\mathbf{R}\widehat{u} + n^{-1}1'_{2n}\mathbb{x}\widehat{b} \tag{7c}$$

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

The three forms of the generalized regression estimator in (7) are useful at different times in subsequent derivations. Form (7a) is the most disaggregated form. By grouping the terms inside the parentheses in different ways, the latter two forms can be derived.

To arrive at (7b), first factor out $\mathbb{x}\widehat{b}$ from the second and third terms inside the parenthesis in (7a). Then define the zero-centered HT estimator associated with $\mathbb{x}$, $\widehat{\delta}^{HT}_{\mathbb{x}} := n^{-1}1'_{2n}\left(\boldsymbol{\pi}^{-1}\mathbf{R} - \mathbf{i}_{2n}\right)\mathbb{x}$. To see that $\widehat{\delta}^{HT}_{\mathbb{x}}$ is zero-centered, simply take its expectation, noting that $\boldsymbol{\pi}^{-1}\mathrm{E}[\mathbf{R}] = \mathbf{i}_{2n}$. This form makes it clear that the generalized regression estimator is just the HT estimator of the ATE minus an adjustment term. Its compactness will make it useful for deriving an asymptotic result below.

To arrive at (7c), factor $\boldsymbol{\pi}^{-1}\mathbf{R}$ out of the first two terms inside the parenthesis in (7a) and define $\widehat{u} := y - \mathbb{x}\widehat{b}$, essentially a vector of residuals. $\mathbb{x}\widehat{b}$ is analogous to a vector of predicted values. The equivalence shows that the generalized regression estimator can be thought of as an HT estimator of the mean of residuals plus an average of predicted values. In this form, which is also useful for certain variance derivations, some will see the connection to doubly robust estimators (but see footnote 9).

As previously mentioned, particular members of the generalized regression estimator class are determined by the 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, $\widehat{b}^{ols}_{\mathrm{I}}$ would be the "common slopes" OLS regression as given in (6). The same superscript and subscript can be added to the corresponding ATE estimator, $\widehat{\delta}^{R,ols}_{\mathrm{I}}$, as well. The pair $(\widehat{b}^{ols}_{\mathrm{I}}, \widehat{\delta}^{R,ols}_{\mathrm{I}})$ can be referred to as "conjugates". Every unique $\widehat{b}$ implies a conjugate ATE estimator.

While the common use of regression involves interpreting the difference in intercept coefficients in $\widehat{b}$ as the ATE, the class of estimators defined here is broader. As such, one way to look at the utility of the

[8] 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 $\widehat{b}^{greg} = (\tilde{\mathbf{x}}\mathbf{m}\mathbf{R}\tilde{\mathbf{x}})^{-1}\tilde{\mathbf{x}}\mathbf{m}\mathbf{R}y$ 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.

[9] There are three reasons not to refer to the generalized regression 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.

generalized regression estimator is that it prescribes a general method of obtaining an estimate of the ATE from a broader array of specification-estimator-design combinations.[10] This, for example, will allow us to define coefficient vectors with asymptotically optimal conjugates that would not otherwise be obvious (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 generalized regression estimators subsumes common regression practice of interpreting the difference in intercept coefficients as ATE estimates. Moreover, the result will lead to a few insights that may be familiar, but which all follow nicely from the one theorem.

3.3 Common uses of regression are subsumed by the class of generalized regression estimators

Since the most common regression practice is to interpret the difference in intercept terms in coefficient vectors as ATE estimates, it serves to connect that practice to the generalized framework. The following theorem shows that, in common practice, the difference in intercept coefficients is algebraically equivalent to the generalized regression estimator, thus explaining in what sense equation (7) "generalizes" regression.

The main thrust of the following theorem is to establish conditions under which the first term in (7c) will be equal to zero, algebraically speaking.

Theorem 3.2. *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 generalized regression estimator,* $\widehat{\delta}^{R,\mathbf{m}}$*, is algebraically equivalent to* $n^{-1}1'_{2n}\mathbb{x}\widehat{b}^{\mathbf{m}}$ *if* $\exists z$ *such that*

$$\mathbf{R}\mathbb{x}z = (\mathbf{R}\mathbf{m}^{-1}\mathbf{R})^{(-)}\boldsymbol{\pi}^{-1}\mathbf{R}1_{2n} \tag{8}$$

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 (7c) equals zero, i.e., that

$$\left(y - \mathbb{x}\widehat{b}^{\mathbf{m}}\right)'\boldsymbol{\pi}^{-1}\mathbf{R}1_{2n} = 0. \tag{9}$$

To see when the equality in (9) will hold, first note that from the definition of $\widehat{b}^{\mathbf{m}}$ given in the theorem we have

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

Therefore, it must also be the case that

$$\left(y - \mathbb{x}\widehat{b}^{\mathbf{m}}\right)'\mathbf{R}'\mathbf{m}^{-1}\mathbf{R}\mathbb{x}z = 0 \tag{10}$$

for any vector z that linearly combines the x's in some way. Comparing the condition given in (9) to the equality in (10) we can see that the only need

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

[10] For example, suppose a researcher had block randomized with unequal assignment probabilities across blocks. Given the design, the first element of the OLS coefficient in (6) is not generally consistent for the ATE, and, hence, it can not be interpreted directly as such. However, (7) gives a general formulation for producing a consistent estimator of the ATE using a broad array of coefficient estimators, even those that are not directly interpretable given the design-specification combination. Consistency is discussed further, below.

for some value of z. 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 1 (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 3.2 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 generalized regression estimator $\widehat{\delta}_{\text{I}}^{R,ols}$ is algebraically equivalent to the difference in intercept terms its conjugate coefficient estimator, $\widehat{b}_{\text{I}}^{ols}$. For specification* II *this means that, if the columns of $\mathbf{x}$ have mean zero, then the generalized regression estimator $\widehat{\delta}_{\text{II}}^{R,ols}$ is algebraically equivalent to the difference of intercept terms in its conjugate coefficient estimator, $\widehat{b}_{\text{II}}^{ols}$.*

Remark 2 (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 3.2 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 generalized regression estimator $\widehat{\delta}_{\text{I}}^{R,\pi wls}$ is algebraically equivalent to the difference in intercept terms in the conjugate coefficient estimator, $\widehat{b}_{\text{I}}^{\pi wls}$. For specification* II *this means that, if the columns of $\mathbf{x}$ have mean zero, then the generalized regression estimator $\widehat{\delta}_{\text{II}}^{R,\pi wls}$ is algebraically equivalent to the difference in intercept terms its conjugate coefficient estimator, $\widehat{b}_{\text{II}}^{\pi wls}$.*

Corollary 3.2.1. *One can ensure that the condition in Theorem 3.2 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 3. *Corollary 3.2.1 shows that for any identified design the condition 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.*

Remark 4. *Note that in spite of Corollary 3.2.1, $(\mathbf{R}\mathbf{m}^{-1}\mathbf{R})^{(-)}\boldsymbol{\pi}^{-1}\mathbf{R}1$ 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 1-3 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 generalized regression estimator in Definition 3.1, and these relationships reveal in what sense it "generalizes" (i.e., subsumes) regression approaches in common use. The generalization, however, will allow for the definition of ATE estimators that obtain asymptotic optimality but for which the difference in intercept terms in the conjugate may not be readily interpretable.

Even when using an estimator-design-specification combination where the condition in Theorem 3.2 holds, knowing the point estimate of the ATE is $n^{-1}1_{2n}'\mathbb{x}\widehat{b}^{\mathbf{m}}$ can be useful. For example, it leads to the well-known maxim that the difference in intercept terms can only be directly interpreted in specification II if covariates, $\mathbf{x}$, are transformed to have mean zero. However, should the researcher forgo zero-centering, it still algebraically defines the process for arriving at an ATE estimate.

3.4 Variance of the generalized regression estimator when $\widehat{b}$ is fixed (-and- A post-hoc test of improved precision)

An exact expression for the variance of the generalized regression estimator is straightforward when $\widehat{b}$ is a fixed vector of constants, call it b^f.[11] In theory, a researcher might obtain this fixed vector through

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

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 $\widehat{\delta}^{R,f}$ (the conjugate of the fixed coefficient b^f) is useful to consider for the following reasons. First, the variance expression for $\widehat{\delta}^{R,f}$ will help to establish the asymptotic variance expression for generalized regression estimators (see section 3.5). 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). This section also provides the basis for a test of the null hypothesis that adjustment does not help precision.

Definition 3.3 (Fixed-coefficient generalized regression estimators)**.** *"Fixed-coefficient generalized regression estimators" are in a subclass of generalized regression estimators defined in* (7) *where* $\widehat{b} = b^f$ *and* $b^f \in \mathbb{R}^l$ *is a vector of constants.*

Lemma 3.4. *The finite sample variance of the fixed-coefficient generalized regression estimator,* $\widehat{\delta}^{R,f}$*, with conjugate* b^f *being a fixed constant, is*

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

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

Proof. To see the result, start with the third form of the generalized regression estimator given in (7c). Note that when $\widehat{b} = b^f$, a constant vector, the second term in (7c) is a constant. The first term is recognizable as a HT 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, for a given b^f. Hence, the exact variance is constructed as in equation (4) but with u in place of y. □

One question is whether fixed-coefficient generalized regression estimator improves precision over the HT estimator, and how one might be confident of that in practice. The answer to this question will suggest the basis of a hypothesis test that can be further explored after Section 5 on variance estimation. To begin to develop the idea of the test, note that (n^2 times) the difference in variances can be written

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

where the vector $u := y - \mathbb{x} b^f$ is not observed for every unit, so that the quantity cannot be observed. However, an estimator can be proposed by defining length-$2n$ column vector $v = (2b^{f'}\mathbb{x}'\mathbf{d}\mathrm{diag}(u))'$ and testing whether $\widehat{\delta}_v^{HT} := n^{-1}1_{2n}\boldsymbol{\pi}^{-1}\mathbf{R}v$ is greater than $-n^{-1}b^{f'}\mathbb{x}'\mathbf{d}\mathbb{x}b^f$. Since $\widehat{\delta}_v^{HT}$ is just an HT estimator, the same machinery that will be developed in Section 5 for conservative variance estimation can be applied to it and conservative inference can follow.

Note that the proposed method tests for the difference-of-variances, which is identified, even though the variances are not themselves identified. By contrast, a direct comparison of variance estimators based on those proposed in Section 5 would actually be a comparison of variance bounds, and, hence, less relevant.

An analogous test for generalized regression estimators with coefficients estimated from the data could also be developed. These tests could help reassure analysts with concerns that regression adjustment can sometimes hurt asymptotic precision (Freedman, 2008a,b). Of course, estimation decisions should be set ahead of time in pre-analysis plans, and such a test should only be used in retrospect. But such a test could still be useful for decision making, say, in a review of past studies to help a researcher determine whether to use regression adjustment in a future study.

3.5 An asymptotic argument

In this section conditions for generalized regression 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 are root-n consistent and asymptotically normal. If they are, then the coefficient, $\widehat{b}$, need only converge in probability. On the one hand, for sufficiently large n this provides a certain amount of freedom to choose coefficient estimators whose asymptotic normality or rate of convergence is uncertain. On the other hand, the reliance on the properties of HT estimators should be somewhat reassuring because they are well-studied, and their asymptotic properties are worked out under a variety of designs.

Assumptions:

1. (Root-n 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 $\widehat{b}$) $\widehat{b} - b = O(n^{-r})$ 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 3.5. *Under Assumptions 1-2, $\sqrt{n}(\widehat{\delta}^R - \delta)$ 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,

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

Proof. Starting with the form for the generalized regression estimator given in (7b) and using Assumptions 1 and 2 and the fact that $\mathrm{E}[\widehat{\delta}_{\mathbb{x}}^{HT}] = 0$ for all n,

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

Moreover, b is a fixed (limit) value so that, by Lemma 3.4, $\mathrm{V}\left(\widehat{\delta}^{HT} - \widehat{\delta}_{\mathbb{x}}^{HT}b\right) = n^{-2}u'\mathbf{d}u$ for all n. Expression (12) follows. Next, it follows from Assumption 3 that $\sqrt{n}(\widehat{\delta}^{HT} - \widehat{\delta}_{\mathbb{x}}^{HT}b)$ has limiting normal distribution so that, by also invoking the variance expression in (12), (13) follows. □

In spite of the fact that the result provides a reassurance that asymptotic properties hinge on a well-studied estimator, it is worth noting that a finite sample CLT for HT estimators that perfectly circumscribes designs for which asymptotic normality can be expected likely cannot exist. Isaki and Fuller (1982) and Fuller (2009) provide a general consistency result 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 and it is also in the sampling context. Aronow and Samii (2017) present a very general criteria, but it nonetheless does not cover all designs. Li and Ding (2017) do a nice job of summarizing the literature on finite-population central limit theorems.

The accuracy of a normal approximation will depend on the particular design, so it is hard say how concerned one should be about non-normality in general. On the one hand, the results of Li et al. (2017)

underscore the fact that the multi-variate normality should not be taken for granted. On the other hand, in practice, it would seem to take a bit of effort to construct asymptotic regimes where HT estimators are consistent and yet non-normality negatively impacts inference based on the normal approximation to a great degree. Moreover, the conservativeness of variance (bound) estimates should act as a counter-weight against consequences of non-normality. In cases where extreme caution is warranted, a researcher could consider intervals constructed using Chebyshev's inequality.

4 Optimal Regression For Arbitrary Designs

Lemma 3.4 gives the finite sample variance of $\widehat{\delta}^{R,f}$, the conjugate of the fixed regression coefficient, b^f, first introduced in Section 3.4. One might next ask, what value of $b^f \in \mathbb{R}^l$ minimizes the finite sample variance of $\widehat{\delta}^{R,f}$? 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 3.5, 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 generalized regression 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 3.4, are derived.

Theorem 4.1. *Letting $(.)^{(-)}$ represent the Moore-Penrose generalized inverse*[12]*, a coefficient value that is finite sample optimal for the fixed-coefficent generalized regression estimator is*

$$b^{opt} := (\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}\mathbb{x}'\mathbf{d}y. \tag{14}$$

Proof. Starting with the finite sample variance of $\widehat{\delta}^{R,f}$ given in Theorem 3.4, we have

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

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

$$(\mathbb{x}'\mathbf{d}\mathbb{x})b^f = \mathbb{x}'\mathbf{d}y. \tag{15}$$

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. □

[12] 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.

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.

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 $\widehat{\delta}^{R,f}$. 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.

It may be helpful to discuss briefly what matrices like $\mathbb{x}'\mathbf{d}\mathbb{x}$ represent. 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 Horvitz-Thompson estimator of b^{opt}, namely, 3HT! (-or- Horvitz-Thompson, Horvitz-Thompson, Horvitz-Thompson, trifecta)

In this section, a HT estimator of b^{opt}, given in equation (14), is introduced. It has the usual limitations of HT estimator, 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 ATE estimator is a constellation of three HT estimators, as can be seen by examining form (7b).

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

$$\widehat{b}^{3HT!} := (\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}\,\mathbb{x}'\mathbf{d}\boldsymbol{\pi}^{-1}\mathbf{R}y \tag{16}$$

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

Remark 5. *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, $\widehat{\delta}^{R,3HT!}$.*

Remark 6. *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 7. *Like any HT estimator, it is unbiased. To see this, simply take the expectation of (16) and recall that $E[\mathbf{R}] = \boldsymbol{\pi}$.*

Lemma 4.3. *The estimator of b^{opt} defined in* (16) *is just an HT estimator of the column sums of the $2n \times k$ matrix*

$$\mathbf{b} := \left((\mathbb{x}'\mathbf{d}\mathbb{x})^{(-)}\,\mathbb{x}'\mathbf{d}\mathrm{diag}(y)\right)'.$$

Proof. The proof involves the recognition that

$$\mathbf{R}\boldsymbol{\pi}^{-1}y = \text{diag}(y)\mathbf{R}\boldsymbol{\pi}^{-1}1_{2n}.$$

□

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

Recognizing $\widehat{b}^{3HT!}$ in equation (14) as a HT estimator of the column sums of $\mathbf{b}$ in Lemma 4.3, suggests that an improved estimation strategy may be to recursively apply generalized regression adjustment. No new principles are required.

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

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

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

$$\begin{aligned}\widehat{b}^{2R!} :=& \left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(-)}\mathbb{x}'\mathbf{d}\left(\boldsymbol{\pi}^{-1}\mathbf{R}y - \boldsymbol{\pi}^{-1}\mathbf{R}\mathbb{x}\widehat{b}^{\pi wls} + \mathbb{x}\widehat{b}^{\pi wls}\right) \\ =& \widehat{b}^{3HT!} - \left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(-)}\left(\mathbb{x}'\mathbf{d}\boldsymbol{\pi}^{-1}\mathbf{R}\mathbb{x} - \mathbb{x}'\mathbf{d}\mathbb{x}\right)\widehat{b}^{\pi wls}.\end{aligned} \tag{17}$$

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

The first line of (17) can be compared to (7a) to make clear that this is regression adjusted regression adjustment. 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 $\widehat{b}^{2R!} \xrightarrow{p} b^{opt}$.

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*

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

Proof. Provided in Appendix. □

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

Theorem 4.7. *The 2R! estimator of the ATE,* $\widehat{\delta}^{R,2R!}$*, 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

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

The last line follows from Lemma 4.6 □

Theorem 4.8. *The 2R! estimator of the ATE, $\widehat{\delta}^{R,2R!}$, 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

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

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 (Campbell and Meyer, 2009). Hence, $\widehat{\delta}^{R,2R!*} = \widehat{\delta}^{R,2R!}$. □

Remark 8. *Given its definition, $\widehat{\delta}^{R,2R!} := \widehat{\delta}^{HT} - \widehat{\delta}^{HT}_{\mathbb{x}}\widehat{b}^{2R!}$, 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 $\widehat{b}^{2R!}$), are not generally invariant. By contrast, the optimal generalized regression estimator, $\widehat{\delta}^{R,3HT!}$, is only invariant to location shifts in special cases (e.g., complete randomization).*

4.4 Conclusions about the proposed optimal estimators, $\widehat{\delta}^{R,3HT!}$ and $\widehat{\delta}^{R,2R!}$

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

Alternatives to $\widehat{\delta}^{R,3HT!}$ and $\widehat{\delta}^{R,2R!}$ are available for specific designs. In Section 6, complete randomization is considered, followed by Section 7 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 the framework herein to prior work (e.g. Lin, 2013).

5 Variance Estimation (-or, more exactly- Variance Bound Estimation)

In general, the variance expressions of the form (4) 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. Conservative inference follows.

In Section 5.1, the terms "variance bound" and "identified variance bound" are defined in terms of the current framework. Framing the problem in matrix terms facilitates insight and leads to methods of comparing alternative bounds. In Section 5.2, an important variance bound that has the virtue of being identified in any identified design, the Aronow-Samii (AS) bound, is defined. The AS bound serves as a benchmark against which other potential bounds might be compared. After that, Section 5.3 proposes an algorithm for finding an alternative variance bound, which can improve upon the AS bound substantially in some cases. Finally, Section 5.4 addresses the subject of how to estimate a variance bound, both for HT estimators and generalized regression estimators. And finally, in Section 5.6 a tighter variance bound specifically for the 2R! estimator is proposed that can substantially narrow intervals for that estimator.

5.1 Bounding the variance

Definition 5.1 (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 $y \in \mathbb{R}^{2n}$, $n^{-2}y'\mathbf{d}y \leq n^{-2}y'\tilde{\mathbf{d}}y$.*

Definition 5.2 (Tighter variance bound)**.** *For two $2n \times 2n$ matrices, $\tilde{\mathbf{d}}^a$ and $\tilde{\mathbf{d}}^b$, that correspond to different variance bounds, $\tilde{\mathbf{d}}^a$ corresponds to a "tighter variance bound" if for all $y \in \mathbb{R}^{2n}$ $n^{-2}y'\tilde{\mathbf{d}}^a y \leq n^{-2}y'\tilde{\mathbf{d}}^b y$.*

Definition 5.3 (Tighter variance bound under the sharp null)**.** *First, let y_0 and y_1 represent length-n vectors of control and treatment potential outcomes, respectively, and recall that under the sharp null $y_1 = y_0$. For two matrices $2n \times 2n$ matrices, $\tilde{\mathbf{d}}^a$ and $\tilde{\mathbf{d}}^b$, that correspond to different variance bounds, $\tilde{\mathbf{d}}^a$ corresponds to a "tighter variance bound under the sharp null" if, for all $y_1 = y_0 \in \mathbb{R}^n$, $n^{-2}y'\tilde{\mathbf{d}}^a y \leq n^{-2}y'\tilde{\mathbf{d}}^b y$.*

Lemma 5.4. *$n^{-2}y'\tilde{\mathbf{d}}y$ 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. □

Corollary 5.4.1. *For two $2n \times 2n$ matrices that define variance bounds, $\tilde{\mathbf{d}}^a$ and $\tilde{\mathbf{d}}^b$, $\tilde{\mathbf{d}}^a$ corresponds to a "tighter variance bound" if and only if the matrix $\tilde{\mathbf{d}}^b - \tilde{\mathbf{d}}^a$ is positive semi-definite.*

Corollary 5.4.2. *First writing the four $n \times n$ partitions of a matrix $\tilde{\mathbf{d}}$ as $\tilde{\mathbf{d}}_{00}$, $\tilde{\mathbf{d}}_{01}$, $\tilde{\mathbf{d}}_{10}$, and $\tilde{\mathbf{d}}_{11}$, define an $n \times n$ matrix*

$$\tilde{\mathbf{d}}_+ := \tilde{\mathbf{d}}_{00} + \tilde{\mathbf{d}}_{11} - \tilde{\mathbf{d}}_{10} - \tilde{\mathbf{d}}_{01}.$$

Then for two matrices, $\tilde{\mathbf{d}}^a$ and $\tilde{\mathbf{d}}^b$, that correspond to different variance bounds, $\tilde{\mathbf{d}}^a$ corresponds to a "tighter variance bound under the sharp null" if and only if $\tilde{\mathbf{d}}^b_+ - \tilde{\mathbf{d}}^a_+$ is positive semi-definite.

Remark 9. *Lemma 5.4 implies that a test for whether a candidate $\tilde{\mathbf{d}}$ corresponds to a bound is to check the eigenvalues of $\tilde{\mathbf{d}} - \mathbf{d}$ for nonnegativeness. Similarly, Corollary 5.4.1 implies that a test for whether a candidate $\tilde{\mathbf{d}}^a$ matrix corresponds to a tighter variance bound than $\tilde{\mathbf{d}}^b$ is to check the eigenvalues of $\tilde{\mathbf{d}}^b - \tilde{\mathbf{d}}^a$ for non-negativeness. Failing this test, Corollary 5.4.2 says that an adjudication between bounds might still be made by testing $\tilde{\mathbf{d}}^b_+ - \tilde{\mathbf{d}}^a_+$ for non-negative eigenvalues. Alternatively, if some eigenvalues of $\tilde{\mathbf{d}}^b - \tilde{\mathbf{d}}^a$ are positive and some negative, heuristic methods of choosing the better bound could include comparing the maximum and minimum eigenvalues or determining the sign of the sum of eigenvalues. Biased inference due to "cherry picking" the narrower confidence interval can be avoided because such comparisons can be done before observing the outcome variable.*

Definition 5.5 (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 and, for example, $\mathrm{I}(\mathbf{d} = -1)$ 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).

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.

5.2 Defining 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 5.6 (Aronow-Samii variance bound)**.** *The "Aronow-Samii variance bound" is*

$$\tilde{\mathrm{V}}^{AS}\left(\widehat{\delta}^{HT}\right) := n^{-2}y'\tilde{\mathbf{d}}^{AS}y \tag{18}$$

where

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

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

Theorem 5.7. *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 $\tilde{\mathbf{d}}^{AS}$,

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

Note that $\mathrm{I}(\mathbf{d} = -1) + \mathrm{diag}\left(\mathrm{I}(\mathbf{d} = -1)\, 1_{2n}\right)$ sets the diagonal elements of $(\tilde{\mathbf{d}}^{AS} - \mathbf{d})$ equal to the sum of off-diagonal elements in row i (which by construction are all non-negative). The Gershgorin circle theorem implies that a real matrix is positive semi-definite if, for all i, diagonal element ii 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 (5.4), $n^{-2}y'\tilde{\mathbf{d}}^{AS}y$ 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 $\mathrm{I}(\mathbf{d} = -1)$ 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.[13] The above-theorem and proof using the Gershgorin circle theorem tie their insight to the current framework.

The AS variance bound is elegant in its universal applicability and serves as a benchmark against which proposed alternatives might be compared. In some cases it can behave quite reasonably. For example, in completely randomized experiments it is exact under the sharp null. That said, simulation examples suggest it can be dramatically over-conservative for some designs.

In the next subsection an algorithm is proposed that may obtain a tighter identified bound for arbitrary designs. Additionally, an analytically-defined bound for cluster-randomized experiments is proposed in Section 7 which is exact under the sharp null and provably tighter than the AS bound.

[13] In sum, since, for example, $-y_{0i}$ and y_{1i} are impossible to jointly observe for all i (since units can only be assigned to treatment or control), the unobservable quantity $2y_{1i}y_{0i}$ which appears in the quadratic in (4) is bounded by the addition of identified quantity $y_{1i}^2 + y_{0i}^2$ in equation (18). By Young's inequality $2y_{1i}y_{0i} \leq y_{1i}^2 + y_{0i}^2$, and hence $\mathrm{V}\left(\widehat{\delta}^{HT}\right) \leq \tilde{\mathrm{V}}^{AS}\left(\widehat{\delta}^{HT}\right)$.

5.3 A proposed algorithm for a tighter variance bound

The following is an algorithm which, if it converges, obtains an identified variance bound. Like the AS bound it has the virtue of being widely applicable. The drawback is the potential computationally difficulty.

In some cases the proposed bound is strictly tighter than the AS bound in the sense of Definition 5.4.1. However, even when not strictly tighter, it can be tighter under the sharp null. In practice, relative tightness can be evaluated on a case-by-case basis using Corollary 5.4.1 and Corollary 5.4.2 and Remark 9.

Algorithm 5.8.

1. Set $\mathbf{t} = \mathrm{I}(\mathbf{d} = -1)$
2. Obtain the eigen decomposition of matrix $\mathbf{t}$
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. Repeat steps 2-4 until convergence is achieved (i.e., until all eigenvalues are non-negative in 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 5.4 and Definition 5.5, 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 matrix $\mathrm{I}(\mathbf{d} = -1)$ 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.

5.4 Estimating a variance bound

For a $\tilde{\mathbf{d}}$ associated with an identified variance bound for an HT estimator, $\tilde{\mathrm{V}}\left(\widehat{\delta}^{HT}\right)$, an unbiased estimator of that bound can then be constructed as

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

where / is element-wise division,

$$\tilde{\mathbf{p}} := \begin{bmatrix} \mathbf{p}_{00} & \mathbf{p}_{01} \\ \mathbf{p}_{10} & \mathbf{p}_{11} \end{bmatrix} + \begin{bmatrix} \mathrm{I}(\mathbf{p}_{00} = 0) & \mathrm{I}(\mathbf{p}_{01} = 0) \\ \mathrm{I}(\mathbf{p}_{10} = 0) & \mathrm{I}(\mathbf{p}_{11} = 0) \end{bmatrix}, \tag{20}$$

$\mathbf{p}_{00}$ has ij element π_{0i0j}, and $\mathbf{p}_{01}$ has ij element π_{0i1j}. $\mathbf{p}_{10}$ and $\mathbf{p}_{11}$ are defined analogously. Conceptually, $\tilde{\mathbf{p}}$ weights each term in the quadratic in (19) inversely proportional to the probability of observing that term. The second term on the right hand side of equation (20) serves to replace zeros in $\mathbf{p}_{00}$, $\mathbf{p}_{01}$, $\mathbf{p}_{10}$ and $\mathbf{p}_{11}$ with ones to ensure that there is no division by zero in $\tilde{\mathbf{d}}/\tilde{\mathbf{p}}$. The choice of replacing the zeros with a value of one is arbitrary and does not affect the estimate because zero elements in the first term on the right-hand-side of (20) correspond to zeros in $\tilde{\mathbf{d}}$ (by the definition of an identified variance bound). Unbiasedness follows from the recognition that $\mathrm{E}\left[\mathbf{R}\left(\tilde{\mathbf{d}}/\tilde{\mathbf{p}}\right)\mathbf{R}\right] = \tilde{\mathbf{d}}$.

Next, by analogy to equation (19) and an appeal to the asymptotics, one can motivate the variance bound estimator for generalized regression estimators as

$$\widehat{\tilde{\mathrm{V}}}\left(\widehat{\delta}^{R}\right) := n^{-2}\widehat{u}'\mathbf{R}\left(\tilde{\mathbf{d}}/\tilde{\mathbf{p}}\right)\mathbf{R}\widehat{u} \tag{21}$$

where $\widehat{u} = y - \mathbb{x}\widehat{b}$.

Unless $\widehat{b}$ is a fixed constant vector, its introduction in (21) means that the bound estimator is not generally unbiased. Refinements made to White's HC0 variance estimator could be applied here, such as degrees of freedom adjustments. An alternative is considered in the next section.

5.5 An MSE (bound) estimator for generalized regression

The variance estimator in (21) may be unsatisfactory, particularly in small samples. However, one can ubiasedly estimate a bound on the MSE of the generalized regression estimator, if one is willing to put in the computational effort.

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 generalized regression estimator can be written

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

It has MSE

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

where

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

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 5.8 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

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

which is unbiased for the MSE bound

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

5.6 Borrowing a tighter variance bound for $\widehat{\delta}^{R,2R!}$

In many instances the gap between the variance of the $\widehat{\delta}^{R,2R!}$, defined in section 4.3, and the bound on its variance estimated by (21) is exceedingly large, leading to overly conservative inference. Theorem 5.9 introduces an approach to minimizing the variance *bound* of the fixed-coefficient generalized regression estimator, $\widehat{\delta}^{R,f}$, with respect to b^f. Then Theorem 5.10 gives a justification for "borrowing" its variance bound estimator and pairing it with $\widehat{\delta}^{R,2R!}$ for the purpose of inference.

Theorem 5.9. *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 generalized regression estimator, a value of b^f that minimizes variance bound $n^{-2}u'\tilde{\mathbf{d}}u$ is*

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

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 5.10. *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 generalized regression 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 generalized regression 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 5.9, because $\tilde{b}^{opt}$ minimizes the variance bound of the fixed-coefficient generalized regression 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 $\widehat{\delta}^{R,2R!}$,

$$\widehat{\tilde{\mathrm{V}}}\left(\widehat{\delta}^{R,2R!}\right) := n^{-2}\widehat{\tilde{u}}'\mathbf{R}\left(\tilde{\mathbf{d}}/\tilde{\mathbf{p}}\right)\mathbf{R}\widehat{\tilde{u}}, \tag{23}$$

where $\widehat{\tilde{u}} = y - \mathbb{x}\widehat{\tilde{b}}^{opt}$ and $\widehat{\tilde{b}}^{opt}$ is an estimator of (22) that, for example, could be defined using a regression adjustment procedure analogous to (17). Again, additional adjustments for degrees of freedom or leverage may be advisable.

6 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 (18). Finally, it is proven that, for specification II, the 2R! estimator, $\widehat{\delta}_{\text{II}}^{R,2R!}$, is algebraically equivalent to the OLS estimator, $\widehat{\delta}_{\text{II}}^{R,ols}$. In that sense, 2R! can be thought of as a generalization of Lin's OLS with specification II for arbitrary designs.

6.1 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_{II}^{ols}, is an optimal coefficient for the fixed-coefficient generalized regression estimator with completely randomized designs and specification II. It will follow that, under Assumptions 1 and 2 in Section 3.5, the OLS coefficient *estimator*, call it $\widehat{b}_{\text{II}}^{ols}$, has a conjugate ATE estimator, $\widehat{\delta}_{\text{II}}^{R,ols}$, that obtains minimum asymptotic variance.

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

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

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.*[14]

[14]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).

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

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

To show that in completely randomized experiments with specification II the coefficient in Definition 6.1 is optimal for the fixed-coefficient generalized regression estimator, the entire set of optimal coefficients for an arbitrary design is first defined. Subsequently, it can be shown that, for complete randomization, b_{II}^{ols} is in that set. The potentially infinite set of optimal coefficients for an arbitrary design is given in the following Lemma.

Lemma 6.3. *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*

$$b^{opt,gz} := \left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(g)}\mathbb{x}'\mathbf{d}y + \left(\mathbf{i}_l - \left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(g)}\left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)\right)z \tag{24}$$

where $\mathbf{i}_l$ *is an* $l \times l$ *identity matrix. Then the entire set of solutions to* (15) *can be defined as*

$$\{b^{opt,gz} \mid z \in \mathbb{R}^l\}. \tag{25}$$

Proof. Provided in Appendix. □

In equation (24), 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 (24) 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 6.3, 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).[15]

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 6.4. *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 (5). □

Lemma 6.5. *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 generalized regression estimator with the "separated" coefficient*

$$b_{\text{II}}^{sep} := \begin{bmatrix}(\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}})^{(-)}\tilde{\mathbf{x}}'\mathbf{d}_{00}y_0 \\ (\tilde{\mathbf{x}}'\mathbf{d}_{11}\tilde{\mathbf{x}})^{(-)}\tilde{\mathbf{x}}'\mathbf{d}_{11}y_1\end{bmatrix} \tag{26}$$

has finite-sample minimum variance, i.e., b_{II}^{sep} *is in the set of optimal fixed-coefficients given by Lemma 6.3.*

[15] 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.

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

$$\mathbf{m}^{(g)} := \begin{bmatrix} \mathbf{a}^{(-)} + \mathbf{a}^{(-)}\mathbf{c}\mathbf{q}^{(-)}\mathbf{c}'\mathbf{a}^{(-)} & -\mathbf{a}^{(-)}\mathbf{c}\mathbf{q}^{(-)} \\ -\mathbf{q}^{(-)}\mathbf{c}'\mathbf{a}^{(-)} & \mathbf{q}^{(-)} \end{bmatrix} \tag{27}$$

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 6.4, and using the definition of generalized inverse, it follows that $\mathbf{q} = 0$. So the inverse reduces to

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

Therefore,

$$\begin{aligned} \left(\mathbb{x}_{\text{II}}'\mathbf{d}\mathbb{x}_{\text{II}}\right)^{(g)}\mathbb{x}_{\text{II}}'\mathbf{d}y &= \begin{bmatrix} \left(\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}}\right)^{(-)} & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \tilde{\mathbf{x}}'\mathbf{d}_{00}y_0 - \tilde{\mathbf{x}}'\mathbf{d}_{01}y_1 \\ -\tilde{\mathbf{x}}'\mathbf{d}_{10}y_0 + \tilde{\mathbf{x}}'\mathbf{d}_{11}y_1 \end{bmatrix} \\ &= \begin{bmatrix} \left(\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}}\right)^{(-)}\left(\tilde{\mathbf{x}}'\mathbf{d}_{00}y_0 - \tilde{\mathbf{x}}'\mathbf{d}_{01}y_1\right) \\ 0 \end{bmatrix}. \end{aligned} \tag{28}$$

Next, using (28) and Lemma 6.3,

$$\begin{aligned} b_{\text{II}}^{opt,gz} &= \left(\mathbb{x}'\mathbf{d}\mathbb{x}\right)^{(g)}\mathbb{x}'\mathbf{d}y + \left(\mathbf{i}_l - \begin{bmatrix} \left(\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}}\right)^{(-)} & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}} & \tilde{\mathbf{x}}'\mathbf{d}_{01}\tilde{\mathbf{x}} \\ \tilde{\mathbf{x}}'\mathbf{d}_{10}\tilde{\mathbf{x}} & \tilde{\mathbf{x}}'\mathbf{d}_{11}\tilde{\mathbf{x}} \end{bmatrix}\right) z \\ &= \begin{bmatrix} \left(\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}}\right)^{(-)}\left(\tilde{\mathbf{x}}'\mathbf{d}_{00}y_0 - \tilde{\mathbf{x}}'\mathbf{d}_{01}y_1\right) \\ 0 \end{bmatrix} + \begin{bmatrix} \mathbf{i}_{(k+1)} - \left(\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}}\right)^{(-)}\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}} & \left(\tilde{\mathbf{x}}'\mathbf{d}_{00}\tilde{\mathbf{x}}\right)^{(-)}\tilde{\mathbf{x}}'\mathbf{d}_{01}\tilde{\mathbf{x}} \\ 0 & \mathbf{i}_{(k+1)} \end{bmatrix} z. \end{aligned} \tag{29}$$

Finally letting $z = \begin{bmatrix} 0 \\ \left(\tilde{\mathbf{x}}'\mathbf{d}_{11}\tilde{\mathbf{x}}\right)^{(-)}\tilde{\mathbf{x}}\mathbf{d}_{11}y_1 \end{bmatrix}$ 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 6.4. □

Remark 10. *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 6.6 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 6.7 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 6.6. *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 6.6.1. *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 6.7. *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 6.6 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 6.8. *In a completely randomized design with specification* II*, the OLS coefficient given in Definition 6.1 minimizes the variance of the fixed-coefficient generalized regression estimator, i.e.,* b_{II}^{ols} *is in the set of optimal fixed-coefficients defined in Lemma 6.3.*

Proof. Using Lemma 6.7 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 6.3 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 6.9. *Under Assumptions 1 and 2, in a completely randomized design with specification* II*, the OLS coefficient estimator given in Definition 6.2 has a conjugate ATE estimator that obtains asymptotic minimum variance within the class of generalized regression estimators.*

Proof. Provided in Appendix. □

6.2 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 generalized regression 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 6.10 (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 6.11 (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 6.10 is an optimal choice of coefficient for the fixed-coefficient generalized regression estimator, first define an equivalent coefficient for specification II.

Definition 6.12 (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 6.12 to Definition 6.10 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 generalized regression estimator for specification I.

Theorem 6.13. *For completely randomized experiments with specification* II*, the tyranny of the minority coefficient given in Definition 6.12 is an optimal coefficient for the fixed-coefficient generalized regression estimator.*

Proof. Beginning with Lemma 6.3 and again arriving at equation (29), 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 6.13.1. *For completely randomized experiments with specification* I*, the tyranny of the minority coefficient given in Definition 6.10 is an optimal coefficient for the fixed-coefficient generalized regression estimator.*

Theorem 6.14. *Under Assumptions 1 and 2, in a completely randomized design with specification* I*, the tyranny of the minority coefficient estimator given in Definition 6.11 has a conjugate ATE estimator that obtains asymptotic minimum variance within the class of generalized regression estimators.*

Proof. Provided in Appendix. □

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

Given that OLS with specification II (see Section 6.1) and the tyranny of the minorty estimator with specification I (see Section 6.2) 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 6.1. 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 6.15. *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 6.1.*

Proof. Provided in Appendix. □

6.4 2R! 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. In this section, it is demonstrated that the two-step optimal estimator, $\widehat{\delta}^{R,2R!}$, is algebraically equivalent to the OLS estimator, $\widehat{\delta}^{R,ols}$.

Theorem 6.16. *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.

7 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.[16]

7.1 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 $\tilde{\mathbf{x}}^c_n$ 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 7.1 (OLS with cluster totals)**.** *The "OLS with cluster totals" coefficient for specification* II *is*

$$b^{ols,c}_{\text{II}} = \begin{bmatrix} \text{Var}(\tilde{\mathbf{x}}^c)^{(-)}\text{Cov}(\tilde{\mathbf{x}}^c, y^c_0) \\ \text{Var}(\tilde{\mathbf{x}}^c)^{(-)}\text{Cov}(\tilde{\mathbf{x}}^c, y^c_1) \end{bmatrix}$$

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^c_0 and y^c_1 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

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

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

Definition 7.2 (OLS with cluster totals coefficient estimator)**.** *The "OLS with cluster totals coefficient estimator" for specification* II *is*

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

where $\widehat{a}_0$ and $\widehat{a}_1$ are scalars and $\widehat{b}^{ols,c}_{\text{II}}$ 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.

[16] 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 Section 6 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).

The two lemmas that follow will lead into the final result of the section. Lemma 7.3 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 7.4 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 7.3. *In a cluster randomized experiment,* $\mathbf{d}_{11}\tilde{\mathbf{x}} = \frac{mm_0}{(m-1)m_1}(\tilde{\mathbf{x}}^c_n - \frac{n}{m}1_n\mu_{\tilde{\mathbf{x}}})$ *where* $\frac{n}{m}1_n\mu_{\tilde{\mathbf{x}}}$ *is a matrix that subtracts off the average cluster totals. Likewise, in a cluster-randomized experiment,* $\mathbf{d}_{00}\tilde{\mathbf{x}} = \frac{mm_1}{(m-1)m_0}(\tilde{\mathbf{x}}^c_n - \frac{n}{m}1_n\mu_{\tilde{\mathbf{x}}})$. *And* $\mathbf{d}_{10}\tilde{\mathbf{x}} = \mathbf{d}_{01}\tilde{\mathbf{x}} = -\frac{m}{m-1}(\tilde{\mathbf{x}}^c_n - \frac{n}{m}1_n\mu_{\tilde{\mathbf{x}}})$.

Proof. Provided in Appendix. □

Lemma 7.4. *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.*

Proof. Provided in Appendix. □

Theorem 7.5. *For cluster randomized experiments, the OLS with cluster totals coefficient in Definition 7.1 is optimal for the fixed-coefficient generalized regression estimator.*

Proof. Since π_{1i} is equal for all i in a cluster randomized experiment, then using Lemma 6.5, we have that the optimal solution includes the separated coefficients in equation (26). So, by Lemma 7.4, 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 11. *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 7.6. *Under Assumptions 1 and 2, in a cluster-randomized design with specification* II, *the OLS coefficient with cluster totals estimator given in Definition 7.2 has a conjugate ATE estimator that obtains asymptotic minimum variance within the class of generalized regression estimators.*

Proof. Provided in Appendix. □

7.2 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 generalized regression 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 7.7 (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

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

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

Definition 7.8 (Tyranny of the minority with cluster totals coefficient estimator)**.** *The "tyranny of the minority with cluster totals coefficient estimator" for specification* I *is*

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

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

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

Definition 7.9 (Tyranny of the minority with cluster totals for specification II)**.** *The "tyranny of the minority with cluster totals coefficient" for specification* II *is*

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

Comparing Definition 7.9 to Definition 7.7 reveals that the "slope" coefficients are identical in the two specifications. The implication is that $\mathbb{x}_{\mathrm{I}}^c b_{\mathrm{I}}^{tyr,c} = \mathbb{x}_{\mathrm{II}}^c b_{\mathrm{II}}^{tyr,c}$ and hence, the conjugate ATE estimators are algebraically equivalent. Therefore, if $b_{\mathrm{II}}^{tyr,c}$ is in the set of optimal choices for a fixed-coefficient in specification II, then $b_{\mathrm{I}}^{tyr,c}$ must be among the optimal coefficients for the fixed-coefficient generalized regression estimator for specification I.

Theorem 7.10. *For cluster-randomized experiments with specification* II*, the tyranny of the minority with cluster totals coefficient given in Definition 7.9 is an optimal coefficient for the fixed-coefficient generalized regression estimator.*

Proof. Beginning with Lemma 6.3 and again arriving at equation (29), this time let

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

The result follows. □

Corollary 7.10.1. *For cluster-randomized experiments with specification* I*, the tyranny of the minority coefficient given in Definition 7.7 is an optimal coefficient for the fixed-coefficient generalized regression estimator.*

Theorem 7.11. *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 7.8 has a conjugate ATE estimator that obtains asymptotic minimum variance within the class of generalized regression estimators.*

Proof. Provided in Appendix. □

7.3 A tighter variance bound for cluster randomized experiments

There are tighter bounds than that implied by Aronow and Samii's $\tilde{\mathbf{d}}^{AS}$ for cluster randomized experiments. Suppose, for example, you had complete randomization of clusters and at least two clusters assigned to each arm. Then $\tilde{\mathbf{d}}$ associated with the variance bound implied by Middleton and Aronow (2015) could be written

$$\tilde{\mathbf{d}}^{c} = \mathbf{d} + \left[\begin{array}{cc} \mathrm{I}\left(\mathbf{d}_{01} = -1\right) & \mathrm{I}\left(\mathbf{d}_{01} = -1\right) \\ \mathrm{I}\left(\mathbf{d}_{01} = -1\right) & \mathrm{I}\left(\mathbf{d}_{01} = -1\right) \end{array} \right]$$

Theorem 7.12. *For cluster randomized experiments, $n^{-2}y'\tilde{\mathbf{d}}^{c}y$ represents an identified bound for $n^{-2}y'\mathbf{d}y$. In other words, for all $y \in \mathbb{R}^{2n}$, $n^{-2}y'\tilde{\mathbf{d}}y \leq n^{-2}y'\tilde{\mathbf{d}}^{c}y$ and $n^{-2}y'\tilde{\mathbf{d}}^{c}y$ is identified.*

Proof. The quantity we are trying to bound is $\phi = 2\sum_i \sum_j y_{1i}y_{0j}\mathrm{I}\left(c_i = c_j\right)$, where c_i gives the cluster id for the i^{th} unit. The additional terms in the bound implied by $\tilde{\mathbf{d}}^{c}$ are

$$\begin{aligned} \sum_i \sum_j \left(y_{0i}y_{0j} + y_{1i}y_{1j}\right)\mathrm{I}\left(c_i = c_j\right) =& \sum_i \sum_j \left(y_{1i}y_{0j} + y_{1i}y_{0j} + \tau_i\tau_j\right)\mathrm{I}\left(c_i = c_j\right) \\ =& \phi + \sum_i \sum_j \tau_i\tau_j \mathrm{I}\left(c_i = c_j\right). \end{aligned}$$

The sum $\sum_i \sum_j \tau_i\tau_j \mathrm{I}\left(c_i = c_j\right)$ must be positive for each cluster. □

Theorem 7.13. *For cluster randomized experiments, if the sharp null holds then $n^{-2}y'\mathbf{d}y = n^{-2}y'\tilde{\mathbf{d}}^{c}y$.*

Theorem 7.14. *$\tilde{\mathbf{d}}^{c}$ provides a tighter bound under complete randomization of clusters than $\tilde{\mathbf{d}}^{AS}$ of associated with the Aronow-Samii bound.*

Proof. The additional terms being added by the AS method to bound the variance are $\sum_i \left(y_{0i}^2 + y_{1i}^2\right)$. So the question is whether $\sum_i \sum_j \left(y_{0i}y_{0j} + y_{1i}y_{1j}\right)\mathrm{I}\left(c_i = c_j\right) < \sum_i \left(y_{0i}^2 + y_{1i}^2\right)$. To simplify, for a single cluster, g, consider whether $\sum_{i\in g} \sum_{j\in g} y_{0i}y_{0j} \leq \sum_{i\in g} y_{0i}^2$. This inequality holds by Jensen's inequality. □

8 Simulations

Simulations illustrate the potential for efficiency gain in a hypothetical cluster-randomized experiment.

8.1 Data generation

A simulated data set was created with 1000 units and 100 clusters, 40 assigned to treatment. To create a range of cluster sizes, cluster membership, c_i, was determined by the equation

$$c_i = \mathrm{trunc}(1 + 100 \times ((i - .5)/1000)^{1.2}).$$

for $i \in \{1, 2, ..., 1000\}$. A table of cluster sizes can be seen in Table 1 below. The distribution of cluster sizes is right skewed.

Data were generated as follows:

$$\begin{aligned} x_{ci} =& \alpha_c + \epsilon_{xi} \\ y_{1ci} = y_{0ci} =& -\alpha_c + x_{ci}^{*} + n_c - 0.025n_c^2 + \epsilon_i \end{aligned}$$

where x_{ci} is a covariate for individual i in cluster c, α_c is drawn from a standard normal at the cluster level, ϵ_{xi} drawn from a standard normal at the individual level, ϵ_i is drawn from $N(0, 5)$ at the individual level, n_c is number of units in cluster c, and y_{0ci} and y_{1ci} are the potential outcomes under control and treatment, respectively. Note that $y_{1ci} = y_{0ci}$, i.e., the sharp null holds. Random components were drawn independently

Table 1: Cluster Sizes

cluster size	number
8	13
9	41
10	21
11	10
12	6
13	3
14	3
16	2
22	1
total	100

from one another.[17] A single finite population was generated using the above DGP and maintained across all simulations.[18]

8.2 Competing estimators

Competing Estimators:

1. *WLS/OLS.* The benchmark estimator is a generalized regression estimator using the WLS with $\boldsymbol{\pi}^{-1}$ weights coefficient. This is equivalent to OLS in this case since π_{1i} are equal for all i and specification II is used.
2. *3HT!.* The generalized regression estimator with coefficient given in Definition 4.2.
3. *2R!.* The generalized regression estimator with coefficient given in Definition 4.4.
4. *OLS with cluster totals.* OLS with cluster totals as described in Definition 6.2.

All estimators used specification II. There were four $\mathbf{x}$ specifications:

1. x.
2. x, and $\overline{x}_c$
3. x, $\overline{x}_c$ and n_c
4. x, $\overline{x}_c$, n_c and n_c^2

8.3 Results

Figure 1 presents the results of the simulations. Clockwise from the top left, the subfigures present the MSE, squared SE, the percent reduction in MSE (relative to WLS/OLS) and bias squared. From left to right on the x-axis are the four specifications.

Results suggest that the 3HT! has relatively poor performance overall. In particular the MSE is very high in part due to a substantial bias. The WLS/OLS estimator performs relatively poorly in terms of MSE for the first two specifications. For these specifications the regression adjusted 2R! and OLS with cluster totals performs well, obtaining an MSE that is about 60% lower than the benchmark, WLS/OLS. For the third specification (x, $\overline{x}$ and n_c) the 2R! performs the best, with an MSE that is about 13% lower than the benchmark WLS/OLS. For the final specification (x, $\overline{x}$, n_c and n_c^2) the 2R! and WLS/OLS perform about equally well, though both show evidence of model-overfit, i.e., an increase in MSE over the third specification.

[17] To give an idea about the relative contribution of ϵ_i to the overall variability of y_{0ci} (y_{1ci}), regressing y_{0ci} (y_{1ci}) on x_{ci}, $\overline{x}_c$, n_c and n_c^2, yielded an R^2 of 0.173.

[18] Using alternative random number seeds does not meaningfully change the results.

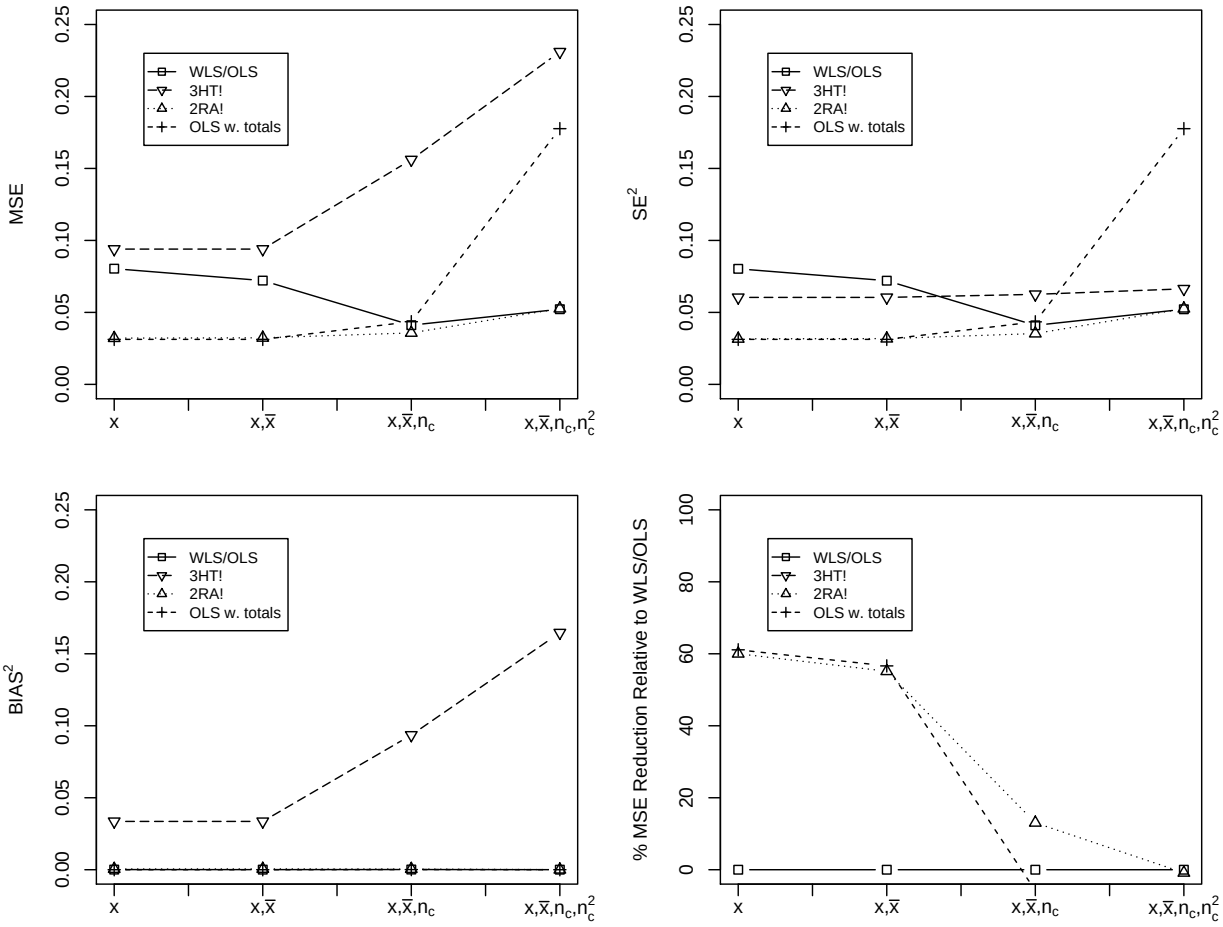

Figure 1: Results are from 5000 randomizations of a simulated cluster randomized experiment. Along the x-axis are four different covariate specifications, increasing in number of predictors from left to right. Each line depicts a coefficient estimation strategy. They are compared (clockwise from top left) in terms of MSE, SE^2, % MSE Reduction and Bias^2.

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**:1120.

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 (17), the two-step optimal coefficient when inputting y^* in place of y can be written

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

□

Proof of Lemma 6.3. The proof consists of two parts: first proving that all members of the set in (25) are solutions and, second, showing that all solutions are in the set.

First note that the fact that (14) is a solution to (15) 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 (24) 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 (15). This proves that all members of the set given by (25) are solutions.

Next, to prove that all solutions are in the set given by (25), let $b^{opt,*}$ represent a an arbitrary solution to (15) 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 (25).

□

Proof of Lemma 6.6. 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 6.9. 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 $\widehat{\delta}_{\text{II}}^{R,ols}$ is asymptotically optimal. □

Proof of Theorem 6.14. 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 6.10. 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 6.15. 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 6.7 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 (27),

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

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 7.3. 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 7.4. 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 7.11.

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

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

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

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