A Class of Unbiased Estimators of the Average Treatment Effect in Randomized Experiments

Peter M. Aronow* and Joel A. Middleton

Forthcoming at the *Journal of Causal Inference*

Abstract

We derive a class of design-based estimators for the average treatment effect that are unbiased whenever the treatment assignment process is known. We generalize these estimators to include unbiased covariate adjustment using any model for outcomes that the analyst chooses. We then provide expressions and conservative estimators for the variance of the proposed estimators.

1 Introduction

Randomized experiments provide the analyst with the opportunity to achieve unbiased estimation of causal effects. Unbiasedness is an important statistical property, entailing that the expected value of an estimator is equal to the true parameter of interest. Randomized experiments are often justified by the fact that they facilitate unbiased estimation of the average treatment effect (ATE). However, this unbiasedness is undermined when the analyst uses an inappropriate analytical tool.

Many statistical methods commonly used to estimate ATEs are biased and sometimes even inconsistent. Contrary to much conventional wisdom (Angrist and Lavy, 2002; Green and Vavreck, 2008; Arceneaux and Nickerson, 2009), even when all units have the same probability of entering treatment, the difference-in-means estimator is biased when clustering in treatment assignment occurs (Middleton, 2008; Middleton and Aronow, 2011). In fact, unless the number of clusters grows with N, the difference-in-means estimator is not generally consistent for the ATE. Similarly, in experiments with heterogeneous probabilities of treatment assignment, the inverse probability weighted (IPW)

*Peter M. Aronow is Doctoral Student, Department of Political Science, Yale University, P.O. Box 208301, New Haven, CT 06520, USA. Joel A. Middleton is Visiting Assistant Professor of Applied Statistics, New York University Steinhardt School of Culture, Education, and Human Development, New York, NY 10003, USA. Thanks to Don Green, Kosuke Imai and Allison Sovey for helpful comments. The authors express special gratitude to Cyrus Samii for invaluable discussions, comments and collaboration.

difference-in-means estimator is not generally unbiased. It is perhaps more well-known that covariate adjustment with ordinary least squares is biased for the analysis of randomized experiments under complete randomization (Freedman, 2008a,b; Schochet, 2010; Lin, in press). Ordinary least squares is, in fact, even inconsistent when fixed effects are used to control for heterogeneous probabilities of treatment assignment (Angrist, 1998; Humphreys, 2009). In addition, Rosenbaum (2002a)'s approach for testing and interval estimation relies on strong functional form assumptions (e.g., additive constant effects), which may lead to misleading inferences when such assumptions are violated (Samii and Aronow, 2012).

In this paper, we draw on classical sampling theory to develop and present an alternative approach that is always unbiased for the average treatment effect (both asymptotically and with finite N), regardless of the clustering structure of treatment assignment, probabilities of entering into treatment or functional form of treatment effects. This alternative also allows for covariate adjustment, also without risk of bias. We develop a generalized difference estimator, that will allow analysts to utilize any model for outcomes in order to reduce sampling variability. This difference estimator, which requires either prior information or statistical independence of some units' treatment assignment (including, e.g., blocked randomization, paired randomization or auxiliary studies), also confers other desirable statistical properties, including location invariance. We also develop estimators of the sampling variability of our estimators that are guaranteed to have a nonnegative bias whenever the difference estimator relies on prior information. These results extend those of Middleton and Aronow (2011), which provides unbiased estimators for experiments with complete randomization of clusters, including linear covariate adjustment.

Unbiasedness may not be the statistical property that analysts are most interested in. For example, analysts may choose an estimator with lower root mean squared error (RMSE) over one that is unbiased. However, in the realm of randomized experiments, where many small experiments may be performed over time, unbiasedness is particularly important. Results from unbiased but relatively inefficient estimators may be preferable when researchers seek to aggregate knowledge from many studies, as reported estimates may be systematically biased in one direction. Furthermore, clarifying the conditions under which unbiasedness will occur is an important enterprise. The class of estimators that is developed here is theoretically important, as it provides sufficient conditions for estimator unbiasedness.

This paper proceeds as follows. In Section 2, we provide a literature review of related work. In Section 3, we detail the Neyman-Rubin Causal Model and define the causal quantity of interest. In Section 4, we provide an unbiased estimator of the ATE and contrast it with other estimators in two common situations. In Section 5, we develop the generalized difference estimator of the ATE, which incorporates covariate adjustment. In Section 6, we define the sampling variance of our estimators and derive conservative estimators thereof. In Section 7, we provide a simple illustrative numerical example. In Section 8, we discuss practical implications of our findings.

2 Related Literature

Our work follows in the tradition of sampling-theoretic causal inference founded by Neyman (1923). In recent years, this framework has gained prominence, first with the popularization of a model of potential outcomes (Rubin, 1974, 1978), and then notably with Freedman (2008a,b)'s work on the bias of the regression estimator for the analysis of completely randomized experiments. The methods derived here relate to the design-based paradigm associated with two often disjoint literatures: that of survey sampling and that of causal inference. We discuss these two literatures in turn.

2.1 Design-based and Model-assisted Survey Sampling

Design-based survey sampling finds its roots in Neyman (1934), later formalized by Godambe (1955) and contemporaries (see Basu, 1971; Sarndal, 1978, for lucid discussions of the distinction between design-based and model-based survey sampling). The design-based survey sampling literature grounds results in the first principles of classical sampling theory, without making parametric assumptions about the response variable of interest (which is instead assumed to be fixed before randomization). All inference is predicated on the known randomization scheme. In this context, Horvitz and Thompson (1952) derive the workhorse, inverse-probability weighted estimator for design-based estimation upon which our results will be based. Refinements in the design-based tradition have largely focused on variance control; early important examples include Des Raj (1965)'s difference estimator and Hajek (1971)'s ratio estimator. Many textbooks (e.g., Thompson, 1997; Lohr, 1999) on survey sampling relate this classical treatment.

The model-assisted (Sarndal et al., 1992) mode of inference, combines features of a model-based and design-based approach. Here, modeling assumptions for the response variable are permitted, but estimator validity is judged by its performance from a design-based perspective. In this tradition, estimators are considered admissible if and only if they are consistent by design (Brewer, 1979; Isaki and Fuller, 1982). Model-assisted estimators include many variants of regression (see, e.g., Cochran, 1977, ch. 7) or weighting estimators (Holt and Smith, 1979), and, perhaps most characteristically, estimators that combine both approaches (e.g., the generalized regression estimators described in Sarndal et al., 1992). Our proposed estimators fall into the model-assisted mode of inference: unbiasedness is ensured by design but models may be used to improve efficiency.

2.2 Design-based Causal Inference

The design-based paradigm in causal inference may be traced to Neyman (1923), which considers finite-population-based sampling theoretic inference for randomized experiments. Neyman established a model of potential outcomes (detailed in Section 3), derived the sampling variability of the difference-in-means estimator for completely randomized experiments (defined in section 4.1.1) and proposed two conservative variance estimators.

Imbens and Rubin (2009, ch. 6) relates a modern treatment of Neyman's approach and Freedman, Pisani and Purves (1998, A32-A34) elegantly derives Neyman's conservative variance estimators.

Rubin (1974) repopularized a model of potential outcomes for statisticians and social scientists, though much associated work using potential outcomes falls into the model-based paradigm (i.e., in that it hypothesizes stochasticity beyond the experimental design). Although there exists a large body of research on causal inference in the model-based paradigm (i.e., sampling from a superpopulation) – textbook treatments can be found in, e.g., Morgan and Winship (2007), Angrist and Pischke (2009) and Hernan and Robins (in press) – we focus our discussion on research in the Neyman-style, design-based paradigm.[1]

Freedman (2008a,b,c) rekindled interest in the design-based analysis of randomized experiments. Freedman raises major issues posed by regression analysis as applied to completely randomized experiments, including efficiency, bias and variance estimation. Lin (in press) and Miratrix, Sekhon and Yu (in press) address these concerns by respectively proposing alternative regression-based and post stratification-based estimators that are both at least as asymptotically efficient as the unadjusted estimator (and, in fact, the post stratification-based estimator may be shown to be a special case of the regression-based estimator than Lin proposes). Turning to the issue of bias, Miratrix, Sekhon and Yu are also able to demonstrate that, for many experimental designs – including the completely randomized experiment – the post stratification-based estimator is conditionally unbiased. Our contribution is to propose a broad class of unbiased estimators that are applicable to any experimental design while still permitting covariate adjustment.

Variance identification and conservative variance estimation for completely-randomized and pair-randomized experiments are considered by Robins (1988) and Imai (2008) respectively, each showing how inferences may differ when a superpopulation is hypothesized. Samii and Aronow (2012) and Lin (in press) demonstrate that, in the case of completely randomized experiments, heteroskedasticity-robust variance estimates are conservative and Lin demonstrates that such estimates provide a basis for asymptotic inference under a normal approximation. Our paper extends this prior work by proposing a new Horvitz-Thompson-based variance estimator that is conservative for *any* experimental design, though additional regularity conditions would be required for use in constructing confidence intervals and hypothesis tests.

Finally, we note increased attention to the challenges of analysis of cluster-randomized experiments under the design-based paradigm, as evidenced in Middleton (2008), Hansen and Bowers (2009), Imai, King and Nall (2009) and Middleton and Aronow (2011). Middleton (2008) notes the bias of regression estimators for the analysis of cluster-randomized designs with complete randomization of clusters. As in this

[1] An alternative design-based tradition, typified by Rosenbaum (2002b), permits hypothesis testing, confidence interval construction, and Hodges-Lehmann point estimation via Fisher's exact test. Although links may be drawn between this Fisherian mode of inference and the Neyman paradigm (Samii and Aronow, 2012), the present work is not directly connected to the Fisherian mode of inference.

paper, Hansen and Bowers (2009) proposes innovative model-assisted estimators that allow for the regression fitting of outcomes, though conditions for unbiasedness are not established nor are the results generalized to alternative experimental designs. Imai, King and Nall (2009) propose that pair-matching is "essential" for cluster-randomized experiments at the design-stage and derive associated design-based estimators and conservative variance estimators. Middleton and Aronow (2011) propose Horvitz-Thompson-type unbiased estimators (including linear covariate adjustment), along with multiple variance estimators, for experiments with complete randomization of clusters. Our paper accommodates cluster-randomized designs, as well as any nonstandard design that might be imposed by the researcher.

3 Neyman-Rubin Causal Model

We begin by detailing the Neyman-Rubin nonparametric model of potential outcomes (Neyman, 1923; Rubin, 1978), which serves as the basis of our estimation approach. Define a binary treatment indicator T_i for units $i = 1, 2, ..., N$ such that $T_i = 1$ when unit i receives the treatment and $T_i = 0$ otherwise.[2] If the stable unit treatment value assumption (Rubin, 1978) holds, let Y_{1i} be the potential outcome if unit i is exposed to the treatment, and let Y_{0i} be the potential outcome if unit i is not exposed to the treatment. The observed outcome Y_i may be expressed as a function of the potential outcomes and the treatment:

$$Y_i = T_i Y_{1i} + (1 - T_i) Y_{0i}. \tag{1}$$

The causal effect of the treatment on unit i, τ_i, is defined as the difference between the two potential outcomes for unit i:

$$\tau_i \equiv Y_{1i} - Y_{0i}. \tag{2}$$

The ATE, denoted by Δ, is defined the average value of τ_i for all units i. In the Neyman-Rubin model, the only random component is the allocation of units to treatment and control groups.

Since $\tau_i \equiv Y_{1i} - Y_{0i}$, the ATE,

$$\Delta = \frac{\sum_{i=1}^{N} \tau_i}{N} = \frac{\sum_{i=1}^{N} (Y_{1i} - Y_{0i})}{N} = \frac{1}{N}\left[\sum_{i=1}^{N} Y_{1i} - \sum_{i=1}^{N} Y_{0i}\right] = \frac{1}{N}\left[Y_1^T - Y_0^T\right], \tag{3}$$

where Y_1^T is the sum of the potential outcomes if in the treatment condition and Y_0^T is the sum of potential outcomes if in the control condition. An estimator of Δ can be

[2]This assumption is made without loss of generality; multiple discrete treatments (or equivalently, some units not being sampled into either treatment or control) are easily accommodated in this framework. All instances of $(1 - T_i)$ in the text may be replaced by C_i, an indicator variable for whether or not unit i receives the control, with one exception to be noted in Section 6.

constructed using estimators of Y_0^T and Y_1^T:

$$\widehat{\Delta} = \frac{1}{N}\left[\widehat{Y_1^T} - \widehat{Y_0^T}\right], \tag{4}$$

where $\widehat{Y_1^T}$ is the estimated sum of potential outcomes under treatment and $\widehat{Y_0^T}$ is the estimated sum of potential outcomes under control.

Formally, the bias of an estimator is the difference between the expected value of the estimator and the true parameter of interest; an estimator is unbiased if this difference is equal to zero. If the estimators $\widehat{Y_0^T}$ and $\widehat{Y_1^T}$ are unbiased, the corresponding estimator of Δ is also unbiased since

$$\mathrm{E}[\widehat{\Delta}] = \frac{1}{N}\left[\mathrm{E}[\widehat{Y_1^T}] - \mathrm{E}[\widehat{Y_0^T}]\right] = \frac{1}{N}\left[Y_1^T - Y_0^T\right] = \Delta. \tag{5}$$

In the following sections, we demonstrate how to derive unbiased estimators of Y_0^T and Y_1^T and, in so doing, derive unbiased estimators of Δ.

4 Unbiased Estimation of Average Treatment Effects

Define N as the number of units in the study, π_{1i} as the probability that unit i is selected into treatment and π_{0i} as the probability that unit i is selected into control. We assume that, $\forall i$, $\pi_{1i} > 0$ and $\pi_{0i} > 0$, or that there is a nonzero probability that each unit will be selected into treatment and that there is a nonzero probability that each unit will be selected into control. (When all units are assigned to either treatment or control, $\pi_{0i} + \pi_{1i} = 1$.) To derive an unbiased estimator of the ATE, we first posit estimators of Y_0^T and Y_1^T. Define the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) of Y_1^T,

$$\widehat{Y_{1,HT}^T} = \sum_{i=1}^{N} \frac{1}{\pi_{1i}} T_i Y_{1i} = \sum_{i=1}^{N} \frac{1}{\pi_{1i}} T_i Y_i. \tag{6}$$

and, similarly, define the Horvitz-Thompson estimator of Y_0^T,

$$\widehat{Y_{0,HT}^T} = \sum_{i=1}^{N} \frac{1}{\pi_{0i}} (1 - T_i) Y_{0i} = \sum_{i=1}^{N} \frac{1}{\pi_{0i}} (1 - T_i) Y_i. \tag{7}$$

The estimators in (6) and (7) are unbiased for Y_1^T and Y_0^T, respectively, since $\mathrm{E}[T_i] = \pi_{1i}$ and $\mathrm{E}[1 - T_i] = 1 - \pi_{1i} = \pi_{0i}$.

From (4), it follows that we may construct an unbiased estimator of Δ:

$$\widehat{\Delta_{HT}} = \frac{1}{N}\left[\widehat{Y_{1,HT}^T} - \widehat{Y_{0,HT}^T}\right] = \frac{1}{N}\left[\sum_{i=1}^{N} \frac{1}{\pi_{1i}} Y_i T_i - \sum_{i=1}^{N} \frac{1}{\pi_{0i}} Y_i (1 - T_i)\right]. \tag{8}$$

We refer to this estimator of the ATE as the HT estimator. The HT estimator is subject to two major limitations. First, as proved in Appendix A, the estimator is not location invariant. By location invariance, we mean that, for a linear transformation of the data,

$$Y_i^* \equiv b_0 + b_1 \times Y_i, \tag{9}$$

where b_0 and b_1 are constants, the estimator based on the original data, $\widehat{\Delta}$, relates to the estimator computed based on the transformed data, $\widehat{\Delta}^*$, in the following way:

$$b_1\widehat{\Delta} = \widehat{\Delta}^*. \tag{10}$$

Failure of location invariance is an undesirable property because it implies that rescaling the data (e.g., recoding a binary outcome variable) can substantively alter the estimate that we compute based on the data. Second, $\widehat{\Delta_{HT}}$ does not account for covariate information, and so may be imprecise relative to estimators that incorporate additional information. We address both of these issues in Section 5.

4.1 Special Cases

The HT estimator is unbiased for all designs, but we will now demonstrate what the estimator reduces to under two common designs. The first, a fixed number of units (n out of N) is selected for inclusion in the treatment, each with equal probability $\left(\frac{n}{N}\right)$. In the second, we consider a case where units are selected as clusters into treatment.

4.1.1 Complete Random Assignment of Units into Treatment:

Consider a design where a fixed number of units (n out of N) is selected for inclusion in the treatment, each with equal probability $\left(\frac{n}{N}\right)$. The associated estimator is

$$\begin{aligned}\widehat{\Delta_{DM}} &= \frac{1}{N}\left[\sum_{i=1}^{N} T_i \frac{N}{n} Y_i - \sum_{i=1}^{N} (1 - T_i) \frac{N}{N-n} Y_i\right] \\ &= \frac{\sum\limits_{i \in I_1} Y_i}{n} - \frac{\sum\limits_{i \in I_0} Y_i}{N-n}.\end{aligned} \tag{11}$$

Equation (11) shows that, for the special case where n of N units are selected into treatment, the HT estimator reduces to the difference-in-means estimator: the average outcome among treatment units minus the average outcome among control units. While the difference-in-means estimator is not generally unbiased for all equal-probability designs, it is unbiased for the numerous experiments that use this particular design.

4.1.2 Complete Random Assignment of Clusters into Treatment:

Consider a design where a fixed number of *clusters* (m out of M clusters) is selected for inclusion in the treatment, each with equal probability $\left(\frac{m}{M}\right)$. The associated estimator is

$$\begin{aligned}\widehat{\Delta_C} &= \frac{1}{N}\left[\sum_{i=1}^{N} T_i \frac{M}{m} Y_i - \sum_{i=1}^{N} (1 - T_i) \frac{M}{M-m} Y_i\right] \\ &= \frac{M}{N}\left[\frac{\sum\limits_{i \in I_1} Y_i}{m} - \frac{\sum\limits_{i \in I_0} Y_i}{M-m}\right]. \end{aligned} \tag{12}$$

Contrast the estimator in (12) with the estimator in (11). A key insight is that (12) does not reduce to the difference-in-means estimator in (11). In fact, the difference-in-means estimator may be biased for cluster randomized experiments (Middleton and Aronow, 2011). Moreover, since the difference-in-means estimator is algebraically equivalent to simple linear regression, regression will likewise be biased for cluster randomized designs (Middleton, 2008).

5 Unbiased Covariate Adjustment

Regression for covariate adjustment is generally biased under the Neyman-Rubin Causal Model (Freedman, 2008a,b). In contrast, a generalized estimator can be constructed to obtain unbiased covariate adjustment. We draw upon the concept of difference estimation, which sampling theorists have been using since (at least) Des Raj (1965). The primary insight in constructing the estimators in (8) is that we need only construct unbiased estimators of totals under treatment and control conditions in order to construct an unbiased estimator of the ATE. Unlike Rosenbaum (2002a), we make no assumptions about the structure (e.g., additivity) of treatment effects.

Continuing with the above-defined notion of estimating totals, we can consider the following class of estimators,

$$\widehat{Y_1^{T*}} = \sum_{i=1}^{N}\left[T_i \frac{Y_{1i}}{\pi_{1i}} - T_i \frac{f(\mathbf{X_i}, \theta_\mathbf{i})}{\pi_{1i}} + f(\mathbf{X_i}, \theta_\mathbf{i})\right], \tag{13}$$

and

$$\widehat{Y_0^{T*}} = \sum_{i=1}^{N}\left[(1 - T_i) \frac{Y_{0i}}{\pi_{0i}} - (1 - T_i) \frac{f(\mathbf{X_i}, \theta_\mathbf{i})}{\pi_{0i}} + f(\mathbf{X_i}, \theta_\mathbf{i})\right], \tag{14}$$

where $f(.)$ is a predetermined real-valued function of pretreatment covariate vector $\mathbf{X_i}$

and of parameter vector $\theta_{\mathbf{i}}$.[3] In inspecting (13), one intuition is that if $f(\mathbf{X_i}, \theta_{\mathbf{i}})$ predicts the value of Y_{1i}, then across units in the study population $f(\mathbf{X_i}, \theta_{\mathbf{i}})$ and Y_{1i} will be correlated. By implication then, $\sum_{i=1}^{N} T_i \frac{Y_{1i}}{\pi_{1i}}$ and $\sum_{i=1}^{N} T_i \frac{f(\mathbf{X_i}, \theta_{\mathbf{i}})}{\pi_{1i}}$ will be correlated across randomizations, thus yielding $\widehat{Y_1^{T*}}$ such that $\text{Var}\left(\widehat{Y_1^{T*}}\right) < \text{Var}\left(\widehat{Y_1^{T}}\right)$. The same intuition holds for (14), so that both estimators will typically have precision gains.

There are many options for choosing $f(.)$. One option for $f(.)$ is a linear relationship between Y_i and X_i: $f(\mathbf{X_i}, \theta_{\mathbf{i}}) = \theta_{\mathbf{i}}'\mathbf{X_i}$. Similarly, if the relationship were thought to follow a logistic function (for binary Y_i), $f(\mathbf{X_i}, \theta_{\mathbf{i}}) = 1 - 1/(1 + \exp(\theta_{\mathbf{i}}'\mathbf{X_i}))$. While the choice of $f(.)$ may be relevant for efficiency, it has no bearing on the unbiasedness of the estimator, so long as the choice is determined prior to examining the data.

We may now define the generalized difference estimator:

$$\widehat{\Delta_G} = \frac{1}{N}\left(\widehat{Y_1^{T*}} - \widehat{Y_0^{T*}}\right). \tag{15}$$

The generalized difference estimator both confers location invariance (as demonstrated in Appendix C) and, very often, decreased sampling variability. $\widehat{\Delta_G}$ is equivalent to the Horvitz-Thompson estimator minus an adjustment term:

$$\widehat{\Delta_G} = \widehat{\Delta_{HT}} - \frac{1}{N}\left(\sum_{i=1}^{N} T_i \frac{f(\mathbf{X_i}, \theta_{\mathbf{i}})}{\pi_{1i}} - \sum_{i=1}^{N} (1 - T_i) \frac{f(\mathbf{X_i}, \theta_{\mathbf{i}})}{\pi_{0i}}\right). \tag{16}$$

The adjustment accounts for the fact that some samples will show imbalance on $f(\mathbf{X_i}, \theta_{\mathbf{i}})$. As we prove in Appendix B, a sufficient condition for $\widehat{\Delta_G}$ to be unbiased is that, for all i, $\text{Cov}(T_i, f(\mathbf{X_i}, \theta_{\mathbf{i}})) = 0$. The simplest way for this assumption to hold is for $\theta_{\mathbf{i}}$ to be derived from an auxillary or prior source, but we examine this selection process further in Section 5.1.[4]

5.1 Deriving $\theta_{\mathbf{i}}$ while Preserving Unbiasedness

If $\theta_{\mathbf{i}}$ is derived from the data, unbiasedness is not guaranteed because the value of $f(\mathbf{X_i}, \theta_{\mathbf{i}})$ can depend on the particular randomization, and thus $\text{Cov}(T_i, f(\mathbf{X_i}, \theta_{\mathbf{i}}))$ is not generally equal to zero. Formally, the estimator will generally be biased because

[3] An alternative is to also allow $f(.)$ to vary between the treatment and control groups, particularly if effect sizes are anticipated to be large. Many of our results will also hold under such a specification, although the conditions for unbiasedness (and conservative variance estimation) will be somewhat more restrictive.

[4] Interestingly (and perhaps unsurprisingly), $\widehat{\Delta_G}$ is quite similar to the double robust (DR) estimator proposed by Robins (1999) (and similar estimators, e.g., Rosenblum and van der Laan, 2010) the key differences between the DR estimator and the difference estimator follow. (a) The DR estimator utilizes estimated, rather than known, probabilities of entering treatment, and thus is subject to bias with finite N. (b) Even if known probabilities of entering treatment were used, θ in $f(\mathbf{X_i}, \theta_{\mathbf{i}})$ is chosen using a regression model, which typically fails to satisfy the restrictions necessary to yield unbiasedness established in Section 5.1. Thus, the DR estimator is subject to bias with finite N.

$$\mathrm{E}\left[\widehat{Y_1^{T*}}\right] = \mathrm{E}\left[\sum_{i=1}^{N} T_i \frac{Y_{1i}}{\pi_{1i}} - \sum_{i=1}^{N} T_i \frac{f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)}{\pi_{1i}} + \sum_{i=1}^{N} f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)\right]$$
$$= Y_1^T - \sum_{i=1}^{N} \mathrm{Cov}\left(T_i, \frac{f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)}{\pi_{1i}}\right), \tag{17}$$

and, likewise,

$$\mathrm{E}\left[\widehat{Y_0^{T*}}\right] = Y_0^T - \sum_{i=1}^{N} \mathrm{Cov}\left(\left(1 - T_i\right), \frac{f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)}{\pi_{0i}}\right). \tag{18}$$

In general, the bias of the estimator based on (13) and (14) will therefore be

$$\mathrm{E}\left[\widehat{\Delta_G}\right] - \Delta = \frac{1}{N}\left(\sum_{i=1}^{N} \mathrm{Cov}\left(\left(1 - T_i\right), \frac{f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)}{\pi_{0i}}\right) - \sum_{i=1}^{N} \mathrm{Cov}\left(T_i, \frac{f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)}{\pi_{1i}}\right)\right). \tag{19}$$

Consider two ways that one might derive $\theta_\mathbf{i}$ that satisfy $\mathrm{Cov}\left(T_i, f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)\right) = 0$. The simplest is to assign a fixed value to $\theta_\mathbf{i}$, so that $f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)$ has no variance, and thus no covariance with T_i. Assigning a fixed value to $\theta_\mathbf{i}$ requires using a prior insight or an auxiliary dataset. The choice of $\theta_\mathbf{i}$ may be suboptimal and, if chosen poorly, may increase the variance of the estimate, but, so long as the analyst does not use the data at hand in forming a judgment, there will be no consequence for bias. In fact, as we demonstrate in Section 6, this approach – where $\theta_\mathbf{i}$ is constant across randomizations – will provide benefits for variance estimation.

Following the basic logic of Williams (1961), a second option, only possible in some studies, is to exploit the fact that, for some units i, j, $T_i \perp\!\!\!\perp T_j$. Recall from (1) that $Y_i = T_i Y_{1i} + (1 - T_i) Y_{0i}$, where Y_{1i} and Y_{0i} are constants. Since the only stochastic component of Y_i is T_i, then $T_i \perp\!\!\!\perp T_j$, and $T_i \perp\!\!\!\perp Y_j$. If $\theta_\mathbf{i}$ is a function of any or all of the elements of the set $\{Y_j : T_i \perp\!\!\!\perp T_j\}$ and no other random variables, then $T_i \perp\!\!\!\perp \theta_\mathbf{i}$. It follows that $T_i \perp\!\!\!\perp f(\mathbf{X_i}, \theta_\mathbf{i})$ and therefore $\mathrm{Cov}\left(T_i, f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)\right) = 0$.[5] There are many studies where this option is available. Consider first an experiment where units are independently assigned to treatment. Without loss of generality, let us assume that the analyst chooses $f(.)$ to be linear. The analyst can then derive a parameter vector θ_i for each unit i in the following way: for each i, the analyst could perform an ordinary least squares regression of the outcome on covariates for all units except for unit i. Another example where the

[5] More formally and without loss of generality, let $f\left(\mathbf{X_i}, \theta_\mathbf{i}\right) = f\left(\mathbf{X_i}, g\left(\mathbf{Z}, Y_j\right)\right) = f\left(\mathbf{X_i}, g\left(\mathbf{Z}, h(Y_{0j}, Y_{1j}, T_j)\right)\right)$, where $\mathbf{Z}$ is a matrix of pretreatment covariates (that may or may not coincide with $\mathbf{X_j}$), $g(.)$ is an arbitrary function (e.g., the least squares fit), and $h(.)$ is the function implied by (1). Since only T_j is a random variable, the random variable $f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)$ equals some function $f'(T_j)$. $T_i \perp\!\!\!\perp T_j$ implies $T_i \perp\!\!\!\perp f'(T_j)$ (equivalently $T_i \perp\!\!\!\perp f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)$) which, in turn, implies $\mathrm{Cov}\left(T_i, f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)\right) = 0$.

second option could be used is a block-randomized experiment. In a block-randomized experiment, units are partitioned into multiple groups, with randomization only occurring within partitions. Since the treatment assignment processes in different partitions are independent, $T_i \perp\!\!\!\perp T_j$ for all i, j such that i and j are in separate blocks. To derive $\theta_{\mathbf{i}}$ for each i, the analyst could then use a regression of outcomes on covariates including all units not in unit i's block. Unmeasured block specific effects may cause efficiency loss, but would not lead to bias.

Note that there exists a special case where $\theta_{\mathbf{i}}$ may be derived from all Y_i without any consequence for bias. If there is no treatment effect whatsoever, then, for all units i, Y_i will be constant, and thus $f(\mathbf{X_i}, \theta_{\mathbf{i}})$ will have no variance (and thus no covariance with any random variables). This point will have greater importance in the following section, where we derive expressions for and develop estimators for the sampling variance of the proposed ATE estimators.

6 Sampling Variance

In the section, we will provide expressions for the sampling variance of the HT estimator and the generalized difference estimator. We then will derive conservative estimators of these sampling variances.

6.1 Sampling Variance of the HT Estimator

We begin by deriving the sampling variance of the HT estimator:

$$\begin{aligned}\text{Var}\left[\widehat{\Delta}_{HT}\right] &= \text{Var}\left[\frac{\widehat{Y_1^T} - \widehat{Y_0^T}}{N}\right] \\ &= \frac{1}{N^2}\left(\text{Var}\left[\widehat{Y_1^T}\right] + \text{Var}\left[\widehat{Y_0^T}\right] - 2\text{Cov}\left[\widehat{Y_1^T}, \widehat{Y_0^T}\right]\right) \qquad (20)\end{aligned}$$

By Horvitz and Thompson (1952), the variance of $\widehat{Y_1^T}$,

$$\begin{aligned}\text{Var}\left(\widehat{Y_1^T}\right) &= \sum_{i=1}^{N}\sum_{j=1}^{N}\text{Cov}\left(T_i, T_j\right)\frac{Y_{1i}}{\pi_{1i}}\frac{Y_{1j}}{\pi_{1j}} \\ &= \sum_{i=1}^{N}\text{Var}\left(T_i\right)\left(\frac{Y_{1i}}{\pi_{1i}}\right)^2 + \sum_{i=1}^{N}\sum_{j\neq i}\text{Cov}\left(T_i, T_j\right)\frac{Y_{1i}}{\pi_{1i}}\frac{Y_{1j}}{\pi_{1j}} \\ &= \sum_{i=1}^{N}\pi_{1i}(1-\pi_{1i})\left(\frac{Y_{1i}}{\pi_{1i}}\right)^2 + \sum_{i=1}^{N}\sum_{j\neq i}(\pi_{1i1j} - \pi_{1i}\pi_{1j})\frac{Y_{1i}}{\pi_{1i}}\frac{Y_{1j}}{\pi_{1j}}, \qquad (21)\end{aligned}$$

where π_{1i1j} is the probability that units i and j are jointly included in the treatment group. Similarly,

$$\text{Var}\left(\widehat{Y_0^T}\right) = \sum_{i=1}^{N} \pi_{0i}(1-\pi_{0i}) \left(\frac{Y_{0i}}{\pi_{0i}}\right)^2 + \sum_{i=1}^{N}\sum_{j\neq i}(\pi_{0i0j} - \pi_{0i}\pi_{0j})\frac{Y_{0i}}{\pi_{0i}}\frac{Y_{0j}}{\pi_{0j}}, \tag{22}$$

where π_{0i0j} is the probability that units i and j are jointly included in the control group. An expression for Cov $\left[\widehat{Y_1^T}, \widehat{Y_0^T}\right]$ may be found in Wood (2008),

$$\begin{aligned}\text{Cov}\left(\widehat{Y_1^T}, \widehat{Y_0^T}\right) &= \sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{1i0j} - \pi_{1i}\pi_{0j})\frac{Y_{1i}Y_{0j}}{\pi_{1i}\pi_{0j}} \\ &= \sum_{i=1}^{N}\sum_{\forall j:\pi_{1i0j}\neq 0}(\pi_{1i0j} - \pi_{1i}\pi_{0j})\frac{Y_{1i}Y_{0j}}{\pi_{1i}\pi_{0j}} - \sum_{i=1}^{N}\sum_{\forall j:\pi_{1i0j}=0} Y_{1i}Y_{0j}, \end{aligned} \tag{23}$$

where π_{1i0j} is the joint probability that unit i will be in the treatment group and unit j will be in the control group. Substituting (21), (22) and (23) into (20),

$$\begin{aligned}\text{Var}\left[\widehat{\Delta}_{HT}\right] = \frac{1}{N^2}\Bigg(&\sum_{i=1}^{N}\pi_{1i}(1-\pi_{1i})\left(\frac{Y_{1i}}{\pi_{1i}}\right)^2 + \sum_{i=1}^{N}\sum_{j\neq i}(\pi_{1i1j} - \pi_{1i}\pi_{1j})\frac{Y_{1i}}{\pi_{1i}}\frac{Y_{1j}}{\pi_{1j}} \\ &+ \sum_{i=1}^{N}\pi_{0i}(1-\pi_{0i})\left(\frac{Y_{0i}}{\pi_{0i}}\right)^2 + \sum_{i=1}^{N}\sum_{j\neq i}(\pi_{0i0j} - \pi_{0i}\pi_{0j})\frac{Y_{0i}}{\pi_{0i}}\frac{Y_{0j}}{\pi_{0j}} \\ &-2\sum_{i=1}^{N}\sum_{\forall j:\pi_{1i0j}\neq 0}(\pi_{1i0j} - \pi_{1i}\pi_{0j})\frac{Y_{1i}Y_{0j}}{\pi_{1i}\pi_{0j}} + 2\sum_{i=1}^{N}\sum_{\forall j:\pi_{1i0j}=0} Y_{1i}Y_{0j}\Bigg). \end{aligned} \tag{24}$$

6.2 Sampling Variance of the Generalized Difference Estimator

If $f(\mathbf{X_i}, \theta_\mathbf{i})$ are constants, this variance formula is also applicable to the generalized difference estimator. When $f(\mathbf{X_i}, \theta_\mathbf{i})$ is constant, we may simply redefine the outcome variable, $U_i = Y_i - f(\mathbf{X_i}, \theta_\mathbf{i})$. It follows that $U_{0i} = Y_{0i} - f(\mathbf{X_i}, \theta_\mathbf{i})$ and $U_{1i} = Y_{1i} - f(\mathbf{X_i}, \theta_\mathbf{i})$. If we rewrite (16), we can see that the generalized difference estimator is equivalent to the HT estimator applied to U_i:

$$\begin{aligned}\widehat{\Delta_G} &= \frac{1}{N}\left(\sum_{i=1}^{N} T_i\frac{Y_{1i}}{\pi_{1i}} - \sum_{i=1}^{N} T_i\frac{f(\mathbf{X_i}, \theta_\mathbf{i})}{\pi_{1i}} - \sum_{i=1}^{N}(1-T_i)\frac{Y_{0i}}{\pi_{0i}} + \sum_{i=1}^{N}(1-T_i)\frac{f(\mathbf{X_i}, \theta_\mathbf{i})}{\pi_{0i}}\right) \\ &= \frac{1}{N}\left(\sum_{i=1}^{N} T_i\frac{Y_{1i} - f(\mathbf{X_i}, \theta_\mathbf{i})}{\pi_{1i}} - \sum_{i=1}^{N}(1-T_i)\frac{Y_{0i} - f(\mathbf{X_i}, \theta_\mathbf{i})}{\pi_{0i}}\right) \\ &= \frac{1}{N}\left(\sum_{i=1}^{N}\frac{1}{\pi_{1i}}T_iU_i - \sum_{i=1}^{N}\frac{1}{\pi_{0i}}(1-T_i)U_i\right). \end{aligned} \tag{25}$$

Therefore, when $f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)$ is constant,

$$\begin{aligned}\text{Var}\left[\widehat{\Delta}_G\right] = \frac{1}{N^2}\Bigg(&\sum_{i=1}^{N}\pi_{1i}(1-\pi_{1i})\left(\frac{U_{1i}}{\pi_{1i}}\right)^2 + \sum_{i=1}^{N}\sum_{j\neq i}(\pi_{1i1j}-\pi_{1i}\pi_{1j})\frac{U_{1i}}{\pi_{1i}}\frac{U_{1j}}{\pi_{1j}} \\ &+ \sum_{i=1}^{N}\pi_{0i}(1-\pi_{0i})\left(\frac{U_{0i}}{\pi_{0i}}\right)^2 + \sum_{i=1}^{N}\sum_{j\neq i}(\pi_{0i0j}-\pi_{0i}\pi_{0j})\frac{U_{0i}}{\pi_{0i}}\frac{U_{0j}}{\pi_{0j}} \\ &-2\sum_{i=1}^{N}\sum_{\forall j:\pi_{1i0j}\neq 0}\frac{U_{1i}U_{0j}}{\pi_{1i}\pi_{0j}}(\pi_{1i0j}-\pi_{1i}\pi_{0j}) + 2\sum_{i=1}^{N}\sum_{\forall j:\pi_{1i0j}=0}U_{1i}U_{0j}\Bigg). \quad (26)\end{aligned}$$

Conversely, the HT estimator may be considered a special case of the generalized difference estimator where $f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)$ is zero for all units. As we proceed, for notational clarity, we use Y_i as the outcome measure, noting that the variances derived will also apply to the generalized difference estimator if U_i is substituted for Y_i (along with both associated potential outcomes) and $f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)$ is constant.

6.3 Accounting for Clustering in Treatment Assignment

We will rewrite the $\widehat{Y_1^T}$ and $\widehat{Y_0^T}$ estimators to account for clustering in treatment assignment. (Our reasons for doing so, while perhaps not obvious now, will become clearer when we derive variance estimators in Section 6.4 and Appendices D and E. While the treatment effect estimators are identical, such a notational switch will allow us to simplify and reduce the bias of our eventual variance estimators.) Note that if, for some units i, j, $\Pr(T_i \neq T_j) = 0$, then the total estimators may be equivalently rewritten. Define $K_k \in \mathbf{K}$ as the set of unit indices i that satisfy $T_i = T'_k$, where T'_k is indexed over all $|\mathbf{K}|$ unique random variables in $\{T_i : i = (1, 2, ..., N)\}$. Define π'_{1k} as the value of $\pi_{1i}, \forall i \in K_k$, π'_{0k} as the value of $\pi_{0i}, \forall i \in K_k$. Joint probabilities π'_{1k1l}, π'_{1k0l}, and π'_{0k0l} are defined analogously. Given these definitions, we can rewrite the HT estimator of the total of treatment potential outcomes as

$$\begin{aligned}\widehat{Y_{1,HT}^T} &= \sum_{i=1}^{N}\frac{1}{\pi_i}T_iY_i = \sum_{k=1}^{M}\sum_{i\in K_k}\frac{1}{\pi_{1i}}T_iY_i = \sum_{k=1}^{M}\sum_{i\in K_k}\frac{1}{\pi'_{1k}}T'_kY_i = \sum_{k=1}^{M}\frac{1}{\pi'_{1k}}T'_k\sum_{i\in K_k}Y_i \\ &= \sum_{k=1}^{M}\frac{1}{\pi'_{1k}}T'_kY'_k,\end{aligned}$$

where $Y'_k = \sum_{i\in K_k} Y_i$. And, similarly,

$$\widehat{Y_{0,HT}^T} = \sum_{k=1}^{M}\frac{1}{\pi'_{0k}}(1-T'_k)Y'_k.$$

In simple language, these estimators now operate over cluster totals as the units of observation. Since the units will always be observed together, they can be summed prior to estimation. The equivalency of these totaled and untotaled HT estimators serves as the basis for the estimation approach in Middleton and Aronow (2011). We may now derive variance expressions logically equivalent to those in (21), (22) and (23):

$$\text{Var}\left(\widehat{Y_1^T}\right) = \sum_{k=1}^{M} \pi'_{1k}(1-\pi'_{1k})\left(\frac{Y'_{1k}}{\pi'_{1k}}\right)^2 + \sum_{k=1}^{M}\sum_{l\neq k}(\pi'_{1k1l}-\pi'_{1k}\pi'_{1l})\frac{Y'_{1k}}{\pi'_{1k}}\frac{Y'_{1l}}{\pi'_{1l}}, \tag{27}$$

$$\text{Var}\left(\widehat{Y_0^T}\right) = \sum_{k=1}^{M} \pi'_{0k}(1-\pi'_{0k})\left(\frac{Y'_{0k}}{\pi'_{0k}}\right)^2 + \sum_{k=1}^{M}\sum_{l\neq k}(\pi'_{0k0l}-\pi'_{0k}\pi'_{0l})\frac{Y'_{0k}}{\pi'_{0k}}\frac{Y'_{0l}}{\pi'_{0l}} \tag{28}$$

and

$$\begin{aligned}\text{Cov}\left(\widehat{Y_1^T},\widehat{Y_0^T}\right) &= \sum_{k=1}^{M}\sum_{l=1}^{M}(\pi'_{1k0l}-\pi'_{1k}\pi'_{0l})\frac{Y'_{1k}Y'_{0l}}{\pi'_{1k}\pi'_{0l}} \\ &= \sum_{k=1}^{M}\sum_{\forall l:\pi'_{1k0l}\neq 0}(\pi'_{1k0l}-\pi'_{1k}\pi'_{0l})\frac{Y'_{1k}Y'_{0l}}{\pi'_{1k}\pi'_{0l}} - \sum_{k=1}^{M}\sum_{\forall l:\pi'_{1k0l}=0} Y'_{1k}Y'_{0l},\end{aligned} \tag{29}$$

where $Y'_{1k} = \sum_{i\in K_k} Y_{1i}$ and $Y'_{0k} = \sum_{i\in K_k} Y_{0i}$. The covariance expression in (29) may now be simplified, however. Since, for all pairs k, l such that $k \neq l$, $\Pr(T'_k \neq T'_l) > 0$, then $\pi'_{1k0l} > 0$ for all $l \neq k$.[6] Therefore, the covariance expression may be reduced further.

$$\text{Cov}\left(\widehat{Y_1^T},\widehat{Y_0^T}\right) = \sum_{k=1}^{M}\sum_{l\neq k}(\pi'_{1k0l}-\pi'_{1k}\pi'_{0l})\frac{Y'_{1k}Y'_{0l}}{\pi'_{1k}\pi'_{0l}} - \sum_{k=1}^{M} Y'_{1k}Y'_{0k}, \tag{30}$$

where the last term is the product of the potential outcomes for each cluster total.

6.4 Conservative Variance Estimation

Our goal is now to derive conservative variance estimators: although not unbiased, these estimators are guaranteed to have a nonnegative bias.[7] We can identify estimators of

[6] If there are multiple treatments, the following simplification cannot be used. Furthermore, the associated estimator in Section 6.4 must apply to (29), for which the derivation is trivially different.

[7] Although the variance estimator is nonnegatively biased, the associated standard errors may not be (due to Jensen's inequality) and any particular draw may be above or below the true value of the variance due to sampling variability.

$\text{Var}\left[\widehat{Y_1^T}\right]$, $\text{Var}\left[\widehat{Y_0^T}\right]$ and $\text{Cov}\left[\widehat{Y_1^T}, \widehat{Y_0^T}\right]$ that are (weakly) positively, positively and negatively biased respectively.[8] Recalling (20), the signs of these biases will ensure a non-negative bias for the overall variance estimator.

First, let us derive an unbiased estimator of $\text{Var}\left(\widehat{Y_1^T}\right)$ under the assumption that, for all pairs k, l, π'_{1k1l} and $\pi'_{0k0l} > 0$. This assumption is equivalent to assuming that all pairs of units have nonzero probability of being assigned to the same treatment condition. This assumption is violated in, e.g., pair-randomized studies, wherein the joint probability of two units in the same pair being jointly assigned to treatment is zero. We propose the Horvitz and Thompson (1952)-style estimator,

$$\widehat{\text{Var}}\left(\widehat{Y_1^T}\right) = \sum_{k=1}^{M} T'_k(1-\pi'_{1k})\left(\frac{Y'_k}{\pi'_{1k}}\right)^2 + \sum_{k=1}^{M}\sum_{l\neq k}\frac{T'_kT'_l}{\pi'_{1k1l}}(\pi'_{1k1l}-\pi'_{1k}\pi'_{1l})\frac{Y'_k}{\pi'_{1k}}\frac{Y'_l}{\pi'_{1l}}, \quad (31)$$

which is unbiased by $\text{E}\left[T'_k\right] = \pi'_{1k}$ and $\text{E}\left[T'_kT'_l\right] = \pi'_{1k1l}$.

What if, for some k, l, $\pi'_{1k1l} = 0$? Blinded authors (in press) prove that $\widehat{\text{Var}}\left(\widehat{Y_1^T}\right)$ will be conservative, or non negatively biased, if, for all k, $Y'_{1k} \geq 0$ (or, alternatively, all $Y'_{1k} \leq 0$). Blinded authors (in press) provide an alternative conservative estimator of the variance for the general case, where Y'_{1k} may be positive or negative:

$$\begin{aligned}\widehat{\text{Var}}_C\left(\widehat{Y_1^T}\right) = &\sum_{k=1}^{M} T'_k(1-\pi'_{1k})\left(\frac{Y'_k}{\pi'_{1k}}\right)^2 \\ &+ \sum_{k=1}^{M}\sum_{l\neq k:\pi'_{1k1l}>0}\frac{T'_kT'_l}{\pi'_{1k1l}}(\pi'_{1k1l}-\pi'_{1k}\pi'_{1l})\frac{Y'_k}{\pi'_{1k}}\frac{Y'_l}{\pi'_{1l}} \\ &+ \sum_{k=1}^{M}\sum_{\forall l:\pi'_{1k1l}=0}\left(T'_k\frac{(Y'_k)^2}{2\pi'_{1k}} + T'_l\frac{(Y'_l)^2}{2\pi'_{1l}}\right). \end{aligned} \quad (32)$$

By an application of Young's inequality, $\text{E}\left[\widehat{\text{Var}}_C\left(\widehat{Y_1^T}\right)\right] \geq \text{Var}\left(\widehat{Y_1^T}\right)$. (An abbreviated proof is presented in Appendix D.) Likewise, a generally conservative estimator of

[8] The variance estimators derived in this paper do not reduce to those proposed by Neyman (1923), Imai (2008) or Middleton and Aronow (2011), due to differences in how the covariance term is approximated.

$\text{Var}\,(\widehat{Y_0^T})$,

$$\widehat{\text{Var}}_C(\widehat{Y_0^T}) = \sum_{k=1}^{M}(1-T_k')(1-\pi_{0k}')\left(\frac{Y_k'}{\pi_{0k}'}\right)^2 + \sum_{k=1}^{M}\sum_{l\neq k:\pi_{0k0l}'>0}\frac{(1-T_k')(1-T_l')}{\pi_{0k0l}'}(\pi_{0k0l}'-\pi_{0k}'\pi_{0l}')\frac{Y_k'}{\pi_{0k}'}\frac{Y_l'}{\pi_{0l}'} + \sum_{k=1}^{M}\sum_{\forall l:\pi_{0k0l}'=0}\left((1-T_k')\frac{(Y_k')^2}{2\pi_{0k}'}+(1-T_l')\frac{(Y_l')^2}{2\pi_{0l}'}\right). \quad (33)$$

Unfortunately, it is impossible to develop a generally unbiased estimator of the covariance between $\widehat{Y_1^T}$ and $\widehat{Y_0^T}$ because Y_{1k}' and Y_{0k}' can never be jointly observed. However, again using Young's inequality, we can derive a generally conservative (which is, in this case, nonpositively biased) covariance estimator:

$$\widehat{\text{Cov}}_C\left(\widehat{Y_1^T},\widehat{Y_0^T}\right) = \sum_{k=1}^{M}\sum_{l\neq k}\frac{T_k'(1-T_l')}{\pi_{1k0l}'}(\pi_{1k0l}'-\pi_{1k}'\pi_{0l}')\frac{Y_k'Y_l'}{\pi_{1k}'\pi_{0l}'} - \sum_{k=1}^{M}T_k'\frac{(Y_k')^2}{2\pi_{1k}'} - \sum_{k=1}^{M}(1-T_k')\frac{(Y_k')^2}{2\pi_{0k}'}. \quad (34)$$

In Appendix E, we prove that $\text{E}\left[\widehat{\text{Cov}}_C\left(\widehat{Y_1^T},\widehat{Y_0^T}\right)\right] \leq \text{Cov}\left(\widehat{Y_1^T},\widehat{Y_0^T}\right)$. One important property of this estimator is that, under the sharp null hypothesis of no treatment effect whatsoever, this estimator is unbiased.

Combining and simplifying (32), (33) and (34), we can construct a conservative estimator of $\text{Var}\,(\widehat{\Delta}_{HT})$, $\widehat{\text{Var}}_C(\widehat{\Delta}_{HT}) =$

$$\frac{1}{N^2}\sum_{k=1}^{M}\left[T_k'\left(\frac{Y_k'}{\pi_{1k}'}\right)^2+(1-T_k')\left(\frac{Y_k'}{\pi_{0k}'}\right)^2+\sum_{l\neq k}\left(\frac{T_k'T_l'}{\pi_{1k1l}'+\epsilon_{1k1l}}(\pi_{1k1l}'-\pi_{1k}'\pi_{1l}')\frac{Y_k'}{\pi_{1k}'}\frac{Y_l'}{\pi_{1l}'}+\right.\right.$$
$$\left.\frac{(1-T_k')(1-T_l')}{\pi_{0k0l}'+\epsilon_{0k0l}}(\pi_{0k0l}'-\pi_{0k}'\pi_{0l}')\frac{Y_k'}{\pi_{0k}'}\frac{Y_l'}{\pi_{0l}'}-2\frac{T_k'(1-T_l')}{\pi_{1k0l}'+\epsilon_{1k0l}}(\pi_{1k0l}'-\pi_{1k}'\pi_{0l}')\frac{Y_k'Y_l'}{\pi_{1k}'\pi_{0l}'}\right)$$
$$\left.+\sum_{\forall l:\pi_{1k1l}'=0}\left(T_k'\frac{(Y_k')^2}{2\pi_{1k}'}+T_l'\frac{(Y_l')^2}{2\pi_{1l}'}\right)+\sum_{\forall l:\pi_{0k0l}'=0}\left((1-T_k')\frac{(Y_k')^2}{2\pi_{0k}'}+(1-T_l')\frac{(Y_l')^2}{2\pi_{0l}'}\right)\right], \quad (35)$$

where $\epsilon_{akbl} = 1[\pi_{akbl}' = 0]$. ($\epsilon_{akbl}$ is included to avoid division by zero.) The fact that $\widehat{\text{Var}}_C(\widehat{\Delta}_{HT})$ is conservative has been established. But also note that when, for all k, l,

$\pi'_{1k1l} > 0$ and $\pi'_{0k0l} > 0$, and there is no treatment effect whatsoever, the estimator is exactly unbiased. A proof of this statement trivially follows from the fact that when $Y_{1i} = Y_{0i}$, (1) reduces to $Y_i = Y_{1i} = Y_{0i}$, which is not a random variable.[9]

7 Illustrative Numerical Example

In this section, we present an illustrative numerical example to demonstrate the properties of our estimators. This example is designed to be representative of small experiments in the social sciences that may be subject to both clustering and blocking. Consider a hypothetical randomized experiment run on 16 individuals organized into 10 clusters across two blocks. A single prognostic covariate $\mathbf{X}$ is available, and two clusters in each block are randomized into treatment, with the others randomized into control. In Table 1, we detail the structure of the randomized experiment, including the potential outcomes for each unit. Note that we have assumed no treatment effect whatsoever.

Unit	Block	Cluster	Y_{1i}	Y_{0i}	$\mathbf{X_i}$	π_{1i}
1	1	1	1	1	4	2/4
2	1	1	1	1	0	2/4
3	1	2	1	1	4	2/4
4	1	2	1	1	1	2/4
5	1	3	1	1	4	2/4
6	1	4	0	0	2	2/4
7	2	5	1	1	4	2/6
8	2	5	1	1	1	2/6
9	2	5	0	0	2	2/6
10	2	6	1	1	5	2/6
11	2	6	1	1	4	2/6
12	2	7	1	1	1	2/6
13	2	7	1	1	4	2/6
14	2	8	0	0	2	2/6
15	2	9	0	0	2	2/6
16	2	10	0	0	3	2/6

Table 1: Details of numerical example.

We may now assess the performance of both (average) treatment effect estimators and

[9] $\widehat{\mathrm{Var}}_C(\widehat{\Delta}_G)$, the variance estimator as applied to U_i, is not generally guaranteed to be conservative. Specifically, when $f(\mathbf{X_i}, \theta_i)$ not constant, there is no guarantee that $\widehat{\mathrm{Var}}_C(\widehat{\Delta}_G)$ will be conservative, though an analogy to linearized estimators suggests that it should be approximately conservative with large N. Importantly, however, when, for all k, l, $\pi'_{1k1l} > 0$ and $\pi'_{0k0l} > 0$ and the sharp null hypothesis of no treatment effect holds, $\widehat{\mathrm{Var}}_C(\widehat{\Delta}_G)$ is unbiased for $\mathrm{Var}(\widehat{\Delta}_G)$.

associated variance estimators, by computing the estimated average treatment effect and variance over all 90 possible randomizations.

7.1 Estimators

Let us first consider four traditional, regression based, estimators. The simplest regression based estimator is the simple IPW difference-in-means estimator, logically equivalent to an IPW least squares regression of the outcome on the treatment indicator. The IPW difference-in-means estimator is a consistent estimator if the finite population grows in such a fashion that the WLLN holds, e.g., independent assignment of units or a growing number of clusters (see Middleton and Aronow, 2011, for a discussion of the consistency of the difference-in-means estimator in the equal-probability, clustered random assignment case). To estimate the variance of this estimator, we use the Huber-White "robust" clustered variance estimator from a IPW least squares regression.

We then examine an alternative regression strategy: ordinary least squares, holding fixed effects for randomization strata (the "FE" estimator). Under modest regularity conditions, Angrist (1998) demonstrates that the fixed effects estimator converges to a reweighted causal effect; in this case, the estimator would be consistent as the treatment effect is constant (zero) across all observations. Similarly, we also use the fixed effects estimator including the covariate $\mathbf{X}$ (the "FE (Cov.)" estimator) in the regression. For both estimators, the associated variance estimator is the Huber-White "robust" clustered variance estimator. Last among the regression estimators is the random effects estimator (the "RE (Cov.)" estimator), as implemented using the `lmer()` function in the `lme4` (Bates and Maechler, 2010) package in R. As with the general recommendation of Green and Vavreck (2008) for cluster randomized experiments, we assume a Gaussian random effect associated with each cluster, fixed effects for randomization strata and control for the covariate $\mathbf{X}$. Variance estimates are empirical Bayes estimates also produced by the `lmer()` function.

We now examine four cases of the Horvitz-Thompson based estimators proposed in this paper. Referring back to Eqs. (8) and (35), we first use $\widehat{\Delta_{HT}}$ and $\widehat{\mathrm{Var}}_C(\widehat{\Delta}_{HT})$ to estimate the ATE and variance. We then use three different forms of the generalized difference estimator. In all cases, we use the general formulations in $\widehat{\Delta_G}$ and $\widehat{\mathrm{Var}}_C(\widehat{\Delta}_G)$, but vary the form and parameters of $f(\mathbf{X_i}, \theta_\mathbf{i})$. In the "G (Prior)" specification, we set $f(\mathbf{X_i}, \theta_\mathbf{i}) = 0.5$ (a reasonable agnostic choice for a binary outcome), neither varying the fitting function according to observed data nor incorporating information on the covariate. In the "G (Linear)" specification, $f(\mathbf{X_i}, \theta_\mathbf{i}) = \beta_{0b} + \beta_{1b} X_i$, where b indicates the block of the unit. For units in block 1, we estimate β_{01} and β_{11} from an OLS regression of Y_i on the covariate (including an intercept) using only units in block 2. β_{02} and β_{12} are similarly estimated from an OLS regression using only units in block 1. As detailed in Section 5, this procedure preserves unbiasedness since, for all units, $\mathrm{Cov}\left(T_i, f(\mathbf{X_i}, \theta_\mathbf{i})\right) = 0$. In the "G (Logit)" specification, $f(\mathbf{X_i}, \theta_\mathbf{i}) = 1 - 1/(1 + \exp(\beta_{0b} + \beta_{1b} X_i))$. β_{01} and β_{11} are

	Regression Based				Horvitz-Thompson Based			
	IPW	FE	FE (Cov.)	RE (Cov.)	HT	G (Prior)	G (Linear)	G (Logit)
Δ	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\mathrm{E}\left[\widehat{\Delta}\right]$	-0.014	-0.016	-0.012	-0.010	0.000	0.000	0.000	0.000
Bias	-0.014	-0.016	-0.012	-0.010	0.000	0.000	0.000	0.000
SE	0.276	0.283	0.191	0.197	0.429	0.302	0.162	0.170
RMSE	0.277	0.283	0.191	0.197	0.429	0.302	0.162	0.170
Var	0.076	0.080	0.036	0.039	0.184	0.091	0.026	0.029
$\mathrm{E}\left[\widehat{\mathrm{Var}}\right]$	0.071	0.074	0.031	0.038	0.184	0.091	0.026	0.029
Bias	-0.005	-0.006	-0.005	-0.001	0.000	0.000	0.000	0.000
SE	0.017	0.019	0.011	0.006	0.037	0.020	0.007	0.007
RMSE	0.018	0.020	0.012	0.007	0.037	0.020	0.007	0.007

Table 2: ATE and variance estimator properties for numerical example.

now derived from a logistic regression using only the units in block 2 (and vice versa for β_{02} and β_{12}). The G (Logit) specification is also unbiased by $\mathrm{Cov}\left(T_i, f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)\right) = 0$. However, while the variance estimators for the HT-based estimators are not generally unbiased (only conservative), the variance estimators will be unbiased since the sharp null hypothesis of no treatment effect holds and for all clusters k, l, $\pi'_{1k1l} > 0$ and $\pi'_{0k0l} > 0$.

7.2 Results

In Table 2, we demonstrate that the only unbiased estimators are the Horvitz-Thompson based estimators: all of the regression-based estimators, including variance estimators, are negatively biased. Although the relative efficiency (e.g., RMSE) of each estimator depends on the particular characteristics of the data at hand, the example demonstrates a case wherein the Horvitz-Thompson based estimators that exploit covariate information have lower RMSE than do the regression-based estimators.

Furthermore, the proposed variance estimators have RMSE on par with the regression-based estimators. However, since the regression-based estimators are negatively biased, researchers may run the risk of systematically underestimating the variance of the estimated ATE when using standard regression estimators. In randomized experiments, where conservative variance estimators are typically preferred, the negative bias of traditional estimators may be particularly problematic.

8 Discussion

The estimators proposed here illustrate a principled and parsimonious approach for unbiased estimation of average treatment effects in randomized experiments of any design. Our method allows for covariate adjustment, wherein covariates can be allowed to have any relationship to the outcome imaginable. Conservative variance estimation also flows directly from the design of the study in our framework. Randomized experiments have been justified on the grounds that they create conditions for unbiased causal inference but the design of the experiment can not generally be ignored when choosing an estimator. Bias may be introduced by the method of estimation, and even consistency may not be guaranteed.

In this paper, we return to the sampling theoretic foundations of the Neyman (1923) model to derive unbiased, covariate adjusted estimators. Sampling theorists developed a sophisticated understanding about the relationship between unbiased estimation and design decades ago. As we demonstrate in this paper, applying sampling theoretic insights to the analysis of randomized experiments permits a broad class of intuitive and clear estimators that highlight the design of the experiment.

References

Angrist, J. D. 1998. Estimating the labor market impact of voluntary military service using Social Security data on military applicants. *Econometrica* **66**: 249–288.

Angrist, J. D. and Lavy, J. 2002. The Effect of High School Matriculation Awards: Evidence from Randomized Trials. NBER Working Paper 9389.

Angrist, J. D., and Pischke, J-S. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press.

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

Basu, D. 1971. An essay on the logical foundations of survey sampling, Part I. In V. Godambe and D. Sprott (Eds.), *Foundations of Statistical Inference*. Toronto: Holt, Rinehart, and Winston.

Bates, D. and Maechler, M. 2010. lme4: Linear mixed-effects models using S4 classes. R package, version 0.999375-37.

Blinded authors. In press. Conservative variance estimation for sampling designs with zero pairwise inclusion probabilities. Forthcoming at *Survey Methodology*.

Brewer, K.R.W. 1979. A class of robust sampling designs for large-scale surveys. *Journal of the American Statistical Association* **74**: 911–915.

Cochran, W. G. 1977. *Sampling Techniques*, 3rd ed. Wiley, New York

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

Fisher, R. A. 1935. *The Design of Experiments*. Edinburgh: Oliver & Boyd.

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.

Freedman, D.A. 2008c. Randomization does not justify logistic regression. *Statistical Science* **23**: 237–49.

Freedman, D.A., Pisani R. and Purves, R.A. 1998. *Statistics*, 3rd ed. New York: Norton.

Godambe, V. P. A Unified Theory of Sampling From Finite Populations. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 17, No. 2

Green, D. P. and Vavreck, L. 2008. Analysis of cluster-randomized experiments: A comparison of alternative estimation approaches. *Political Analysis* **16** 138–152.

Hajek, J. 1971. Comment on 'An essay on the logical foundations of survey sampling, Part I.' In V. Godambe and D. Sprott (Eds.), *Foundations of Statistical Inference*. Toronto: Holt, Rinehart, and Winston.

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

Hernan, M. and Robins, J. In press. *Causal Inference*. London: Chapman and Hall.

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.

Holt, D. and Smith, T. M. F. 1979. Post stratification. *J. Roy. Statist. Soc. Ser. A.* **142**: 33–46.

Humphreys, M. 2009. Bounds on least squares estimates of causal effects in the presence of heterogeneous assignment probabilities. Working paper. Available at: http://www.columbia.edu/˜mh2245/papers1/monotonicity4.pdf.

Imai, K. 2008. Variance identication and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in Medicine* **27**, 4857–4873.

Imai, K., King, G. and Nall, C. 2009. The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Statist. Sci.* **24** 29–53.

Imbens, G.W. and Rubin, D. 2009. Causal inference in statistics. Unpublished textbook.

Isaki, C. T. and Fuller, W. A. 1982. Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.* **77,** 89–96.

Lin, W. In press. Agnostic Notes on Regression Adjustments to Experimental Data: Re-examining Freedman's Critique. *Annals of Applied Statistics*.

Lohr, S. L. *Sampling: design and analysis*. Pacific Grove, CA: Duxbury Press, 1999.

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

Middleton, J.A. and Aronow, P.M. 2011. Unbiased Estimation of the Average Treatment Effect in Cluster-Randomized Experiments. Working paper. Yale University.

Miratrix, L., Sekhon, J. and Yu, B. In press. Adjusting Treatment Effect Estimates by Post-Stratication in Randomized Experiments. *Journal of the Royal Statistical Society. Series B (Methodological)*.

Morgan, S.L. and Winship, C. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge Univ Press.

Neyman, J. 1934. On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, Vol. 97, No. 4: 558–625

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

Robins, J. M. 1988. Confidence intervals for causal parameters. *Statist. Med.* **7,** 773–785.

Robins, J.M. 1999. Association, Causation, and Marginal Structural Models. *Synthese* **121**, 151–179.

Rosenbaum, P. R. 2002a. Covariance Adjustment in Randomized Experiments and Observational Studies. *Statistical Science*. **17**: 286–304.

Rosenbaum, P.R. 2002b. *Observational Studies*. 2nd ed. New York, NY: Springer.

Rosenblum, M. and van der Laan, M.J. 2010. Simple, Efficient Estimators of Treatment Effects in Randomized Trials Using Generalized Linear Models to Leverage Baseline Variables. *The International Journal of Biostatistics* **6**,1: Article 13.

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

Rubin, D. 1978. Bayesian inference for causal effects. *The Annals of Statistics* **6**: 34–58.

Sarndal, C-E. 1978. Design-Based and Model-Based Inference in Survey Sampling. *Scandinavian Journal of Statistics*. **5**, 1: 27–52.

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

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

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

Thompson, M. E. 1997. *Theory of Sample Surveys*. London: Chapman and Hall.

Williams, W. H. 1961. Generating unbiased ratio and regression estimators. *Biometrics* **17** 267–274.

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

A Non-invariance of the Horvitz-Thompson Estimator

This proof follows from Middleton and Aronow (2011). To show that the estimator in (8) is not invariant, let Y_{1i}^* be a linear transformation of the treatment outcome for the i^{th} person such that

$$Y_{1i}^* \equiv b_0 + b_1 \times Y_{1i} \tag{36}$$

and likewise, the control outcomes,

$$Y_{0i}^* \equiv b_0 + b_1 \times Y_{0i}. \tag{37}$$

We can demonstrate that the HT estimator is not location invariant because the estimate based on this transformed variable will be

$$\begin{aligned}
\widehat{\Delta}^{*}_{HT} &= \frac{1}{N}\left[\sum_{i=1}^{N}\frac{1}{\pi_{1i}}Y_i^{*}T_i - \sum_{i=1}^{N}\frac{1}{\pi_{0i}}Y_i^{*}(1-T_i)\right] \\
&= \frac{1}{N}\left[\sum_{i=1}^{N}\frac{1}{\pi_{1i}}\left(b_0+b_1Y_i\right)T_i - \sum_{i=1}^{N}\frac{1}{\pi_{0i}}\left(b_0+b_1Y_i\right)(1-T_i)\right] \\
&= \frac{1}{N}\left[\sum_{i=1}^{N}\frac{1}{\pi_{1i}}b_0T_i - \sum_{i=1}^{N}\frac{1}{\pi_{0i}}b_0(1-T_i)\right] + \frac{1}{N}\left[\sum_{i=1}^{N}\frac{1}{\pi_{1i}}b_1Y_iT_i - \sum_{i=1}^{N}\frac{1}{\pi_{0i}}b_1Y_i(1-T_i)\right] \\
&= \frac{b_0}{N}\left[\sum_{i=1}^{N}\frac{1}{\pi_{1i}}T_i - \sum_{i=1}^{N}\frac{1}{\pi_{0i}}(1-T_i)\right] + \frac{b_1}{N}\left[\sum_{i=1}^{N}\frac{1}{\pi_{1i}}Y_iT_i - \sum_{i=1}^{N}\frac{1}{\pi_{0i}}Y_i(1-T_i)\right] \\
&= \frac{b_0}{N}\left[\sum_{i=1}^{N}\frac{1}{\pi_{1i}}T_i - \sum_{i=1}^{N}\frac{1}{\pi_{0i}}(1-T_i)\right] + b_1\widehat{\Delta}_{HT}. \qquad (38)
\end{aligned}$$

Unless $b_0 = 0$, the term on the left does not generally reduce to zero but instead varies across treatment assignments, so (38) does not generally equal (10) for a given randomization. Therefore, the HT estimator is not generally location invariant. The equation also reveals that multiplicative scale changes where $b_0 = 0$ and $b_1 \neq 0$ (e.g. transforming from feet to inches) need not be of concern. However, a transformation that includes an additive component, such as reverse coding a binary indicator variable ($b_0 = 1$ and $b_1 = -1$), will lead to a violation of invariance. So, for any given randomization, transforming the data in this way can yield substantively different estimates.

B Unbiasedness of the Generalized Difference Estimator

Assume that $\mathrm{Cov}\left(T_i, f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)\right) = 0$, for all $i \in (1, 2, ..., N)$.

$$\begin{aligned}
\mathrm{E}\left[\widehat{Y_1^{T*}}\right] &= \mathrm{E}\left[\sum_{i=1}^{N}T_i\frac{Y_{1i}}{\pi_{1i}} - \sum_{i=1}^{N}T_i\frac{f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)}{\pi_{1i}} + \sum_{i=1}^{N}f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{N}T_i\frac{Y_{1i}}{\pi_{1i}}\right] - \mathrm{E}\left[\sum_{i=1}^{N}T_i\frac{f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)}{\pi_{1i}}\right] + \mathrm{E}\left[\sum_{i=1}^{N}f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)\right] \\
&= \sum_{i=1}^{N}\frac{\pi_{1i}Y_{1i}}{\pi_{1i}} - \sum_{i=1}^{N}\pi_{1i}\mathrm{E}\left[\frac{f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)}{\pi_{1i}}\right] + \sum_{i=1}^{N}\mathrm{E}\left[f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)\right] \\
&= Y_1^T - \sum_{i=1}^{N}\mathrm{E}\left[f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)\right] + \sum_{i=1}^{N}\mathrm{E}\left[f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)\right] \\
&= Y_1^T \qquad (39)
\end{aligned}$$

and, likewise,

$$\mathrm{E}\left[\widehat{Y_0^{T*}}\right] = Y_0^T. \tag{40}$$

The third line of (39) follows from $\mathrm{Cov}\left(T_i, f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)\right) = 0$. The key insight is that, if $\mathrm{E}\left[\sum_{i=1}^{N} T_i \frac{f(\mathbf{X_i}, \theta_\mathbf{i})}{\pi_{1i}}\right] = \mathrm{E}\left[\sum_{i=1}^{N} f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)\right]$, the two right-most terms in (13) and (14) cancel in expectation and, therefore, the terms do not lead to bias in the estimation of Y_1^T or Y_0^T.

C Location Invariance of the Generalized Difference Estimator

Unlike the HT estimator, the Generalized Difference Estimator is location invariant. If the outcome measure changes such that $Y_i^* = b_0 + b_1 Y_i$, we assume that the predictive function will also change by the identical transformation such that

$$f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)^* = b_0 + b_1 f\left(\mathbf{X_i}, \theta_\mathbf{i}\right). \tag{41}$$

If we conceptualize $f(.)$ as a function designed to predict the value of Y_i, then the intuition behind this transformation is clear; if we change the scaling of the outcome variable, it logically implies that the numerical prediction of the outcome will change accordingly. By (16) and (41),

$$\begin{aligned}
\widehat{\Delta_G}^* &= \widehat{\Delta_{HT}}^* - \frac{1}{N}\left(\sum_{i=1}^{N} T_i \frac{f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)^*}{\pi_{1i}} - \sum_{i=1}^{N}\left(1 - T_i\right)\frac{f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)^*}{\pi_{0i}}\right) \\
&= \widehat{\Delta_{HT}}^* - \frac{1}{N}\left(\sum_{i=1}^{N} T_i \frac{b_0 + b_1 f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)}{\pi_{1i}} - \sum_{i=1}^{N}\left(1 - T_i\right)\frac{b_0 + b_1 f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)}{\pi_{0i}}\right) \\
&= \widehat{\Delta_{HT}}^* - \frac{1}{N}\left(\sum_{i=1}^{N} T_i \frac{b_0}{\pi_{1i}} - \sum_{i=1}^{N}\left(1 - T_i\right)\frac{b_0}{\pi_{0i}}\right) \\
&\quad - \frac{1}{N}\left(\sum_{i=1}^{N} T_i \frac{b_1 f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)}{\pi_{1i}} - \sum_{i=1}^{N}\left(1 - T_i\right)\frac{b_1 f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)}{\pi_{0i}}\right) \\
&= b_1 \widehat{\Delta_{HT}} - \frac{1}{N}\left(\sum_{i=1}^{N} T_i \frac{b_1 f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)}{\pi_{1i}} - \sum_{i=1}^{N}\left(1 - T_i\right)\frac{b_1 f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)}{\pi_{0i}}\right) \\
&= b_1 \widehat{\Delta_{HT}} - \frac{b_1}{N}\left(\sum_{i=1}^{N} T_i \frac{f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)}{\pi_{1i}} - \sum_{i=1}^{N}\left(1 - T_i\right)\frac{f\left(\mathbf{X_i}, \theta_\mathbf{i}\right)}{\pi_{0i}}\right) \\
&= b_1 \widehat{\Delta_G}.
\end{aligned} \tag{42}$$

The fourth line in (42) follows from the substitution of (38) for $\widehat{\Delta_{HT}}^*$.

D Abbreviated proof of conservative variance estimator

We present an abbreviated proof from Blinded authors (in press). Without loss of generality, we prove that $\widehat{\mathrm{Var}}_C(\widehat{Y_1^T})$ will be positively biased for $\mathrm{Var}\,(\widehat{Y_1^T})$.

$$\mathrm{Var}\,(\widehat{Y_1^T}) = \sum_{k=1}^{M} \pi'_{1k}(1-\pi'_{1k}) \left(\frac{Y'_{1k}}{\pi'_{1k}}\right)^2 + \sum_{k=1}^{M}\sum_{l\neq k} (\pi'_{1k1l} - \pi'_{1k}\pi'_{1l}) \frac{Y'_{1k}}{\pi'_{1k}} \frac{Y'_{1l}}{\pi'_{1l}},$$

By Young's inequality,

$$\begin{aligned}\mathrm{Var}\,(\widehat{Y_1^T}) \leq \mathrm{Var}_C(\widehat{Y_1^T}) = &\sum_{k=1}^{M} \pi'_{1k}(1-\pi'_{1k}) \left(\frac{Y'_{1k}}{\pi'_{1k}}\right)^2 \\ &+ \sum_{k=1}^{M} \sum_{l\neq k:\pi'_{1k1l}>0} (\pi'_{1k1l} - \pi'_{1k}\pi'_{1l}) \frac{Y'_{1k}}{\pi'_{1k}} \frac{Y'_{1l}}{\pi'_{1l}} \\ &+ \sum_{k=1}^{M} \sum_{\forall l:\pi'_{1k1l}=0} \left(\frac{(Y'_{1k})^2}{2} + \frac{(Y'_{1l})^2}{2}\right).\end{aligned}$$

$\mathrm{Var}_C(\widehat{Y_1^T})$ may be estimated without bias:

$$\begin{aligned}\widehat{\mathrm{Var}}_C(\widehat{Y_1^T}) = &\sum_{k=1}^{M} T'_k(1-\pi'_{1k}) \left(\frac{Y'_k}{\pi'_{1k}}\right)^2 + \sum_{k=1}^{M} \sum_{l\neq k:\pi'_{1k1l}>0} \frac{T'_k T'_l}{\pi'_{1k1l}} (\pi'_{1k1l} - \pi'_{1k}\pi'_{1l}) \frac{Y'_k}{\pi'_{1k}} \frac{Y'_l}{\pi'_{1l}} \\ &+ \sum_{k=1}^{M} \sum_{\forall l:\pi'_{1k1l}=0} \left(T_k \frac{(Y'_k)^2}{2\pi'_{1k}} + T_l \frac{(Y'_l)^2}{2\pi'_{1l}}\right),\end{aligned}$$

by $\mathrm{E}\,[T'_k T'_l] = \pi'_{1k1l}$, $\mathrm{E}\,[T'_k] = \pi'_{1k}$ and (1). Since $\mathrm{E}\left[\widehat{\mathrm{Var}}_C(\widehat{Y_1^T})\right] = \mathrm{Var}_C(\widehat{Y_1^T})$, $\mathrm{E}\left[\widehat{\mathrm{Var}}_C(\widehat{Y_1^T})\right] \geq \mathrm{Var}\,(\widehat{Y_1^T})$. By inspection, $\widehat{\mathrm{Var}}_C(\widehat{Y_0^T})$ is also conservative.

Examining this estimator reveals why we have totaled clusters prior to estimation of variances. By combining totals, we apply Young's inequality to all pairs of cluster totals, instead of all cluster-crosswise pairs of individual units. The bounds need only apply to a single totaled quantity, rather than to each of the constituent components. This step therefore will typically reduce the bias of the estimator.

E Proof of conservative covariance estimator

$$\mathrm{Cov}\left(\widehat{Y_1^T}, \widehat{Y_0^T}\right) = \sum_{k=1}^{M}\sum_{l\neq k} (\pi'_{1k0l} - \pi'_{1k}\pi'_{0l}) \frac{Y'_{1k}Y'_{0l}}{\pi'_{1k}\pi'_{0l}} - \sum_{k=1}^{M} Y'_{1k}Y'_{0k}.$$

By Young's inequality,

$$\text{Cov}\left(\widehat{Y_1^T}, \widehat{Y_0^T}\right) \geq \text{Cov}_C\left(\widehat{Y_1^T}, \widehat{Y_0^T}\right) \qquad = \sum_{k=1}^{M}\sum_{l\neq k}(\pi'_{1k0l} - \pi'_{1k}\pi'_{0l})\frac{Y'_{1k}Y'_{0l}}{\pi'_{1k}\pi'_{0l}} - \sum_{k=1}^{M}\frac{(Y'_{1k})^2}{2} - \sum_{k=1}^{M}\frac{(Y'_{0k})^2}{2}. \tag{43}$$

$\text{Cov}_C\left(\widehat{Y_1^T}, \widehat{Y_0^T}\right)$ may be estimated without bias:

$$\widehat{\text{Cov}}_C\left(\widehat{Y_1^T}, \widehat{Y_0^T}\right) = \sum_{k=1}^{M}\sum_{l\neq k}\frac{T'_k(1-T'_l)}{\pi'_{1k0l}}(\pi'_{1k0l} - \pi'_{1k}\pi'_{0l})\frac{Y'_kY'_l}{\pi'_{1k}\pi'_{0l}} - \sum_{k=1}^{M}T'_k\frac{(Y'_k)^2}{2\pi'_{1k}} - \sum_{k=1}^{M}(1-T'_k)\frac{(Y'_k)^2}{2\pi'_{0k}},$$

by $\text{E}\left[T'_k(1-T'_l)\right] = \pi'_{1k0l}$, $\text{E}\left[T'_k\right] = \pi'_{1k}$, $\text{E}\left[1-T'_k\right] = \pi'_{0k}$ and (1). Since $\text{E}\left[\widehat{\text{Cov}}_C\left(\widehat{Y_1^T}, \widehat{Y_0^T}\right)\right] = \text{Cov}_C\left(\widehat{Y_1^T}, \widehat{Y_0^T}\right)$, $\text{E}\left[\widehat{\text{Cov}}_C\left(\widehat{Y_1^T}, \widehat{Y_0^T}\right)\right] \leq \text{Cov}\left(\widehat{Y_1^T}, \widehat{Y_0^T}\right)$. Unbiasedness under the sharp null hypothesis of no effect is ensured by (43), where if $Y'_{0k} = Y'_{1k}$, $\text{Cov}_C\left(\widehat{Y_1^T}, \widehat{Y_0^T}\right) = \text{Cov}\left(\widehat{Y_1^T}, \widehat{Y_0^T}\right)$. Much as in Appendix D, the bias of the estimator is reduced by totaling clusters prior to estimation. In fact, unbiasedness under the sharp null hypothesis of no effect only holds because we have totaled clusters. Otherwise, the bounds would have to operate over all units, and pairs of units, within each cluster.