The LOOP Estimator: Adjusting for Covariates in Randomized Experiments

James Wu* Johann Gagnon-Bartsch*

December 3, 2016

Abstract

When conducting a randomized controlled trial, it is common to specify in advance the statistical analyses that will be used to analyze the data. Typically these analyses will involve adjusting for small imbalances in baseline covariates. However, this poses a dilemma, since adjusting for too many covariates can hurt precision more than it helps, and it is often unclear which covariates are predictive of outcome prior to conducting the experiment. For example, both post-stratification and OLS regression adjustments can actually increase variance (relative to a simple difference in means) if too many covariates are used. OLS is also biased under the Neyman-Rubin model. In this paper, we introduce the LOOP ("Leave-One-Out Potential outcomes") estimator of the average treatment effect. We leave out each observation and then impute that observation's treatment and control potential outcomes using a prediction algorithm, such as a random forest. This estimator is unbiased under the Neyman-Rubin model, generally performs at least as well as the unadjusted estimator, and the experimental randomization largely justifies the statistical assumptions made. Most importantly, the LOOP estimator also enables us to take advantage of automatic variable selection when using random forests.

1 Introduction

In randomized controlled trials, it is common to specify in advance the statistical analyses to be performed. For example, various authors have advocated for the reporting of statistical methods in the trial protocol (*e.g.*, [2], [16]). Typically these analyses will involve adjusting for small imbalances in baseline covariates, which can improve the precision of the treatment effect estimate. However, in cases where the analysis methods are pre-specified, it can be unclear which covariates should be used and if covariate adjustment will even be helpful. An overly aggressive adjustment that adjusts for too many covariates can hurt precision more than it helps.

Covariate adjustment is commonly done through regression; for example, Young [19] cites

*Department of Statistics, University of Michigan, Ann Arbor, MI.

53 experimental papers from the economics literature between 2007 and 2014 in which regression adjustment is used. Although it is common, OLS does have disadvantages. One of the virtues of randomized experiments is that the physical act of randomization largely justifies the statistical assumptions of the Neyman-Rubin model,[1] a non-parametric model which was first introduced by Jerzy Neyman [17] and further developed by Donald Rubin [15]. However, the Neyman-Rubin model is quite different than the standard OLS model, and as noted by Freedman [7], randomization fails to justify the standard assumptions of OLS. Moreover, Freedman shows that the regression estimate is biased under the Neyman-Rubin model and can, in certain circumstances, be outperformed by a simple difference in means. In a response to Freedman's paper, Lin [12] argues that "without taking the regression model literally, we can still make use of properties of OLS that do not depend on the model assumptions." In other words, even when the regression model is incorrect, regression adjustment can be a useful tool. Both Freedman and Lin note that the OLS estimator performs well in large sample sizes. However, in cases where we have only a moderate sample size and a relatively large number of covariates, variable selection may be required.

While regression is a common method of covariate adjustment, there are others. For example, Bloniarz et al. [4] propose the use of lasso adjustments when the number of covariates is large, especially when the number of covariates exceeds the number of experimental units. Another covariate adjustment method is post-stratification [10]. Post-stratification is an adjustment made by stratifying on a pretreatment variable, estimating the treatment effect within each stratum, and taking the weighted average over all strata. Miratrix, Sekhon, and Yu [13] explore the properties of the post-stratified estimator under the Neyman-Rubin model. Rosenbaum [14] also discusses covariate adjustment in the context of randomization inference. Rosenbaum uses the covariates to estimate the control outcome for each unit, calculates residuals from these estimates, and permutes the residuals to test hypothesized values of the treatment effect. He then inverts the hypothesis tests to yield confidence intervals. Rosenbaum notes that one can obtain the residuals using any fitting algorithm and cites robust linear regression, rank linear regression, or a smoother as examples (in addition to OLS). While Rosenbaum's method relies only on randomization as the basis for inference, it assumes a fixed treatment effect for each unit.

Aronow and Middleton [1] introduce another estimator, which is related to the Horvitz-Thompson estimator [11]. This design-based estimator involves the estimation of a function of the covariates. So long as this function is independent of the treatment assignment, the resulting estimate will be unbiased. We propose a special case of this estimator, the LOOP ("Leave-One-Out Potential outcomes") estimator. We leave out each observation and then impute that observation's treatment and control potential outcomes using a prediction algorithm, such as a random forest [5]. Our work is similar to that of Wager, Du, Taylor, and Tibshirani [18], who also propose a set of estimators that build on the work of Aronow and Middleton, and use machine learning methods to impute potential outcomes. Wager et al.

[1] One important assumption that is not guaranteed by randomization is that one unit's outcome is not affected by another unit's treatment status. This assumption is sometimes referred to as the Stable Unit Treatment Value Assumption (SUTVA).

assume that the experimental units are drawn from a superpopulation, and focus primarily on the population average treatment effect.

In this paper, we analyze the LOOP estimator assuming that the potential outcomes and the covariates are fixed and that the only source of randomness is in the treatment assignment. We derive an estimate for the variance of the LOOP estimator. Aronow and Middleton also provide an estimate for the variance of their estimator, but assume that the function of the covariates is a constant. Note that our variance estimate also differs from that of Wager et al., as we work under a different model.

We discuss the imputation of each unit's potential outcomes using various methods such as decision trees. We show that using the LOOP estimator and imputing potential outcomes using a decision tree is equivalent to post-stratification. Because random forests are typically an improvement over individual decision trees, our hope is that we can use the LOOP estimator with random forests to improve upon post-stratification. Miratrix et al. note that post-stratification is nearly as efficient as blocking, and we therefore hope to obtain an estimate that works as well or better than if we had used a blocked design.

To summarize, the primary advantages of the LOOP estimator are: (1) it is design-based, meaning that the experimental randomization largely justifies the statistical assumptions; (2) it is exactly unbiased; (3) it generally performs no worse than the simple difference-in-means estimator, but can often substantially improve performance; and importantly (4) it allows for automatic variable selection, so we do not need to know which covariates to use ahead of time.

The paper is organized as follows. Section 2 provides a motivating example. In Section 3, we introduce notation and assumptions and discuss the simple difference and LOOP estimators. In Section 4, we discuss three different methods of imputing the potential outcomes and relate the LOOP estimator with imputation done by decision trees and random forests to post-stratification. In Section 5, we discuss how to modify the procedures to account for different experimental designs such as block designs. In Section 6, we provide an estimate of the variance. In Section 7, we apply the LOOP estimator to two examples: one using simulated data and one using real experimental data. Section 8 concludes.

2 Motivation

Our motivating example is a so-called "pay for success" program in the state of Illinois [9]. In brief, a pay for success program is one in which the government contracts an outside organization to provide needed services, but only pays the organization if the services are shown to be effective, typically in a randomized controlled experiment. In our example, the contracted organization is to provide special social services to at-risk youth, and one metric for success (among others) is a reduction in the number of days spent in juvenile detention. Success of the program will be evaluated according to the results of a six year experiment in which eligible youth are randomly selected to receive either the special services or ordi-

nary care. The evaluation will be conducted by researchers in the School of Social Work at the University of Michigan; author Gagnon-Bartsch of this paper assisted the evaluators in planning the design and analysis of the experiment. Unfortunately, the experiment has only recently begun so we do not yet have any data on which to apply the methods we develop in this paper, and our discussion of the pay for success program is therefore limited to this section (we explore an alternative dataset in Section 7). Nonetheless, the challenges presented by the pay for success program are instructive, and we outline them briefly.

Several hundred youth are expected to take part in the program. Eligible participants are randomized to treatment or control, each with probability 1/2. Treatment assignments are independent. More elaborate designs were considered, but were too logistically challenging. A key difficulty is the fact that the participants enter into the experiment continually over time, making designs such as blocking infeasible.

Several baseline covariates will be available, at least some of which (*e.g.*, age) are known to be highly predictive of outcome. It was agreed that some form of adjustment for these covariates was desirable, but initially there was no clear consensus on which adjustment procedure should be used or which covariates should be included. Given the need to specify the analysis protocol in advance, this led to considerable discussion. In the end, it was agreed to use a post-stratification estimator, partly on the grounds that it is unbiased under the Neyman-Rubin model, whereas other common estimators (*e.g.*, linear regression) are not. Unbiasedness is arguably more inherently desirable in this example than in many other applications because the state's payment rate for the services provided will be directly proportional to the estimated size of the treatment effect. Any bias in the estimator therefore effectively results in a bias in the payment.[2] Moreover, it was agreed to post-stratify on just two variables that are known to be highly predictive of outcome; other covariates will not be used due to the risk that an overly aggressive adjustment could end up hurting precision rather than improving it.

In this paper, we are motivated to produce a method that provides automatic variable selection in order to eliminate the guesswork in deciding which covariates to use, while remaining unbiased under the Neyman-Rubin model. An initial idea was to randomly split the data in half, use one half to empirically determine which covariates are predictive of outcome and construct a set of strata that are optimal in some sense (perhaps using a decision tree), and then use the other half of the data to compute a post-stratified estimate using the optimal strata. Since the data used to construct the strata would be independent of the data used in the estimation step, the estimator would remain unbiased. Only half of the data would be used in the estimation step, however, this procedure could then be repeated many times and the results averaged to produce an aggregate estimator that is also unbiased but effectively makes use of all of the data.

[2]Note that even if the estimator is unbiased, payment will still be biased for other reasons. In particular, no payment will be made at all unless the observed treatment effect achieves statistical significance, which results in a payment bias against the service provider. On the other hand, if the observed treatment effect turns out to be negative, the service provider will not receive a negative payment (*i.e.*, will not be required to pay the state), which results in a bias in favor of the service provider.

The method we develop in this paper, which is a special case of the estimator proposed by Aronow and Middleton [1], is very similar in spirit the procedure just described. It is in some sense a limiting case in which the data is split, not in half, but rather such that all of the observations except for one are used to determine the optimal strata, and (counter-intuitively) only one observation is left for estimation. Moreover, instead of relying on just one single set of optimal strata, we use many nearly optimal sets; these sets of strata are effectively determined by a random forest algorithm.

3 The LOOP Estimator

In this section, we introduce the LOOP ("Leave-One-Out Potential outcomes") estimator, which we can use to obtain an unbiased estimate of the average treatment effect while adjusting for covariates.

3.1 Model and Notation

Consider a randomized controlled experiment in which there are N participants, indexed by $i = 1, 2, ..., N$. Each participant is randomly assigned to either treatment or control, and we let X_i denote the i-th participant's treatment assignment, such that $X_i = 1$ if the i-th participant is assigned to treatment and $X_i = 0$ if the i-th participant is assigned to control. For each participant, we observe (in addition to the treatment assignment X_i) a response variable Y_i and a q-dimensional vector of baseline covariates Z_i.

We assume that the treatment assignments are independent of one another, *i.e.*

$$X_i \perp\!\!\!\perp X_j \tag{1}$$

for $i \neq j$. We let p_i denote the i-th participant's probability of being assigned to treatment, *i.e.*

$$p_i = P(X_i = 1) \tag{2}$$

and assume $0 < p_i < 1$. In some parts of this paper, we assume for simplicity (and without much loss of generality) that $p_i = \frac{1}{2}$ for all i, but for now we explicitly let p_i be arbitrary.

Associated with each of the N participants are two fixed (non-random) potential outcomes, t_i and c_i. We assume that we observe t_i if participant i is assigned to treatment and c_i if participant i is assigned to control. That is, the observed outcome Y_i for participant i is

$$Y_i = X_i t_i + (1 - X_i) c_i. \tag{3}$$

We define the individual treatment effect τ_i as

$$\tau_i = t_i - c_i \tag{4}$$

and the average treatment effect $\bar{\tau}$ as

$$\bar{\tau} = \frac{1}{N}\sum_{i=1}^{N} \tau_i \tag{5}$$

which is our primary parameter of interest.

Lastly, some additional notation. Let $T = \{i : X_i = 1\}$ and $C = \{i : X_i = 0\}$. Let n be the (random) number of participants assigned to treatment and $N - n$ be the number assigned to control. For each participant, we define the important quantity m_i as

$$m_i = (1 - p_i)t_i + p_i c_i. \tag{6}$$

Note that when $p_i = \frac{1}{2}$, this is simply the mean of t_i and c_i. We will use the notation $\hat{m}_i$ to denote an estimate of m_i. Finally, we define the random variable U_i as

$$U_i = \begin{cases} 1/p_i, & X_i = 1 \\ -1/(1 - p_i), & X_i = 0 \end{cases} \tag{7}$$

and note that U_i has expectation 0.

3.2 Average and Individual Treatment Effects

It is not possible to observe any single participant's treatment effect τ_i, because for each participant we are only able to observe the treatment response t_i or the control response c_i. However, it is well known that the average treatment effect τ can be estimated. We define the *simple difference estimator* $\hat{\tau}_{sd}$ to be the difference of the average of the observed treatment responses and the average of the observed control responses:

$$\hat{\tau}_{sd} = \frac{1}{n}\sum_{i \in T} Y_i - \frac{1}{N - n}\sum_{i \in C} Y_i. \tag{8}$$

This provides an unbiased estimate of the average treatment effect (conditional on $0 < n < N$).

Less well known is the fact that it is also possible to provide an unbiased estimate of an individual participant's treatment effect τ_i. For example, $Y_i U_i$ is one such estimator:

$$Y_i U_i = \begin{cases} t_i/p_i, & X_i = 1 \\ -c_i/(1 - p_i), & X_i = 0 \end{cases} \tag{9}$$

and thus

$$\mathbb{E}(Y_i U_i) = \frac{t_i}{p_i} P(X_i = 1) + \frac{-c_i}{1 - p_i} P(X_i = 0) \tag{10}$$

$$= t_i - c_i. \tag{11}$$

This estimator is essentially mathematical trickery. Suppose, for example, that $p_i = 1/2$. Then if participant i is assigned to treatment we would estimate his treatment effect as $2Y_i$, and if he was assigned to control we would estimate his treatment effect as $-2Y_i$. Although this does result in an unbiased estimator of τ_i, it is clearly useless for all practical purposes. A more sanguine way of putting this would be that the estimator, despite being unbiased, likely has very high variance.

As an alternative estimator of τ_i, consider

$$\hat{\tau}_i = (Y_i - \hat{m}_i)U_i. \tag{12}$$

If $\hat{m}_i$ is independent of U_i — that is, if $\hat{m}_i$ is independent of the i-th participant's treatment assignment — then $\hat{\tau}_i$ is an unbiased estimator of τ_i:

$$\begin{aligned} \mathbb{E}(\hat{\tau}_i) &= \mathbb{E}\left[(Y_i - \hat{m}_i)U_i\right] \\ &= \mathbb{E}(Y_i U_i) - \mathbb{E}(\hat{m}_i)\mathbb{E}(U_i) \\ &= \tau_i \end{aligned} \tag{13}$$

where in the last line we use the fact that $\mathbb{E}(U_i) = 0$. The advantage of this estimator is that it will have a low variance as long as $\hat{m}_i \approx m_i$. To see why, suppose that $\hat{m}_i = m_i$ exactly. Then

$$(Y_i - m_i)U_i = \begin{cases} (t_i - m_i)/p_i, & X_i = 1 \\ (-c_i + m_i)/(1 - p_i), & X_i = 0 \end{cases} \tag{14}$$

but both $(t_i - m_i)/p_i$ and $(-c_i + m_i)/(1 - p_i)$ work out to be τ_i, and thus $\hat{\tau}_i$ is not only unbiased but also has zero variance. When $\hat{m}_i$ only approximately equals m_i, then the variance of $\hat{\tau}_i$ is no longer zero but is small. More precisely, in Section 6 we show that

$$\mathrm{Var}(\hat{\tau}_i) = \frac{1}{p_i(1 - p_i)} \mathbb{E}\left[(\hat{m}_i - m_i)^2\right]. \tag{15}$$

To summarize then, $\hat{\tau}_i$ will be unbiased and have low variance as long as: (1) $\hat{m}_i$ is independent of X_i; and (2) $\hat{m}_i$ is a good estimator of m_i.

3.3 Leave-One-Out Imputation

We now define the LOOP estimator of the average treatment effect $\bar{\tau}$ as:

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} \hat{\tau}_i \tag{16}$$

where $\hat{\tau}_i$ is defined as in (12) and where $\hat{m}_i$ is obtained as follows. For each i, we drop observation i and use the remaining $N-1$ observations to impute t_i and c_i, using any method of our choosing (*e.g.*, linear regression, random forests, etc.). Having obtained estimates $\hat{t}_i$ and $\hat{c}_i$ we then set

$$\hat{m}_i = (1 - p_i)\hat{t}_i + p_i \hat{c}_i. \tag{17}$$

As an example, suppose we wish to estimate $\hat{m}_i$ using linear regression. For each i, we would drop observation i and then regress Y on X and Z using only the remaining $N-1$ observations. We would then calculate $\hat{t}_i$ and $\hat{c}_i$ using the fitted model, plugging in Z_i for the covariates, and then compute $\hat{m}_i$ as in (17).

Because we leave out the i-th observation when we compute $\hat{m}_i$, it follows that X_i and $\hat{m}_i$ are independent and thus that $\hat{\tau}_i$ is unbiased. It immediately follows that $\hat{\tau}$ is also unbiased. This will be true no matter how we estimate t_i and c_i, as long as we leave out observation i so that $\hat{t}_i$ and $\hat{c}_i$ are independent of X_i. Importantly, note that we impute both t_i and c_i, even though one of them is actually observed and therefore known. If we were to use the true observed value, then $\hat{m}_i$ would no longer be independent of X_i.

It is worth noting that although we use the individual treatment effect estimates $\hat{\tau}_i$ in this paper simply as an intermediate step in the estimation of the average treatment effect $\bar{\tau}$, these individual treatment effect estimates may be useful for other purposes as well, such as in estimating treatment effect heterogeneity. With this in mind, we summarize below three useful facts about the $\hat{\tau}_i$, the latter two of which we show in Section 6:

$$\mathbb{E}(\hat{\tau}_i) = \tau_i \tag{18}$$

$$\mathrm{Var}(\hat{\tau}_i) = \frac{1}{p_i(1-p_i)}\mathbb{E}\left[(\hat{m}_i - m_i)^2\right] \tag{19}$$

$$\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j) = \mathrm{Cov}(\hat{m}_i U_i, \hat{m}_j U_j) \tag{20}$$

The covariance term $\mathrm{Cov}(\hat{m}_i U_i, \hat{m}_j U_j)$ is usually negligible and can be ignored in most applications (note that U_i and U_j are independent).

4 Imputing the Potential Outcomes

In the subsequent sections, we propose several methods for imputing the potential outcomes in order to estimate m_i. First, we impute the potential outcomes without making use of covariates, simply taking the mean of the observed outcomes in each treatment group. When we do this, we see that the LOOP estimator is exactly equal to the simple difference estimator. We also impute the potential outcomes using decision trees and discuss the connection between post-stratification and the LOOP estimator. Finally, we propose the use of random forests, which provide an improvement on the post-stratification and allow us to take advantage of automatic variable selection.

4.1 Imputing Potential Outcomes Ignoring Covariates: LOOP equals the Simple Difference Estimator

In this section, we impute the potential outcomes without making use of covariates. We simply take the mean of the observed outcomes in the treatment group (excluding observation i) to estimate t_i and the mean of the observed outcomes in the control group (excluding observation i) to estimate c_i. If the assignment probabilities are all equal, *i.e.*, if $p_i = p$

for all i and for some fixed p, then the LOOP estimator is exactly equivalent to the simple difference estimator, as we show below:

$$
\begin{aligned}
\hat{\tau} &= \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{m}_i\right)U_i \\
&= \frac{1}{N}\left[\sum_{i=1}^{N}\frac{1}{p}\left(Y_i - \hat{m}_i\right)X_i + \sum_{i=1}^{N}\frac{1}{1-p}\left(\hat{m}_i - Y_i\right)\left(1 - X_i\right)\right] \\
&= \frac{1}{N}\left\{\sum_{i=1}^{N}\frac{1}{p}\left[Y_i - \left(\frac{\sum_{k\in T\setminus i}(1-p)Y_k}{n - X_i} + \frac{\sum_{k\in C\setminus i} pY_k}{(N-n) - (1-X_i)}\right)\right]X_i + \right. \\
&\qquad \left. \sum_{i=1}^{N}\frac{1}{1-p}\left[\left(\frac{\sum_{k\in T\setminus i}(1-p)Y_k}{n - X_i} + \frac{\sum_{k\in C\setminus i} pY_k}{(N-n) - (1-X_i)}\right) - Y_i\right]\left(1 - X_i\right)\right\} \\
&= \frac{1}{N}\left[\sum_{i\in T}\left(\frac{Y_i}{p} - \frac{1-p}{p}\frac{\sum_{k\in T\setminus i}Y_k}{n-1} - \frac{\sum_{k\in C}Y_k}{N-n}\right) + \sum_{i\in C}\left(\frac{\sum_{k\in T}Y_k}{n} + \frac{p}{1-p}\frac{\sum_{k\in C\setminus i}Y_k}{(N-n)-1} - \frac{Y_i}{1-p}\right)\right] \\
&= \frac{1}{N}\left[\sum_{i\in T}\frac{Y_i}{p} - \sum_{i\in C}\frac{Y_i}{1-p} - \frac{1-p}{p}\frac{(n-1)\sum_{k\in T}Y_k}{n-1} - \frac{n\sum_{k\in C}Y_k}{N-n} + \frac{(N-n)\sum_{k\in T}Y_k}{n}\right. \\
&\qquad \left. + \frac{p}{1-p}\frac{((N-n)-1)\sum_{k\in C}Y_k}{(N-n)-1}\right] \\
&= \frac{1}{N}\left[\sum_{i\in T}\frac{Y_i - (1-p)Y_i}{p} - \sum_{i\in C}\frac{Y_i - pY_i}{1-p} - \frac{n\sum_{k\in C}Y_k}{N-n} + \frac{(N-n)\sum_{k\in T}Y_k}{n}\right] \\
&= \frac{1}{N}\left[\sum_{i\in T}Y_i - \sum_{i\in C}Y_i - \frac{n\sum_{k\in C}Y_k}{N-n} + \frac{(N-n)\sum_{k\in T}Y_k}{n}\right] \\
&= \frac{1}{N}\left[\frac{((N-n)+n)\sum_{k\in T}Y_k}{n} - \frac{(n+(N-n))\sum_{k\in C}Y_k}{N-n}\right] \\
&= \frac{\sum_{k\in T}Y_k}{n} - \frac{\sum_{k\in C}Y_k}{N-n} \\
&= \hat{\tau}_{sd}.
\end{aligned}
\tag{21}
$$

As a result of this equivalence, we conclude that in practice the LOOP estimator will typically perform no worse than the simple difference estimator. That is, the LOOP estimator will outperform the simple difference estimator as long as we improve the imputation of the potential outcomes beyond this baseline approach. In particular, we find it reassuring that the leave-one-out procedure does not inherently introduce extra variance.

Technical note: One minor difference between the simple difference estimator and the LOOP estimator in this case is that the simple difference estimator is undefined whenever n is equal to 0 or N, whereas the LOOP estimator is undefined whenever n is equal to 0, 1, $N-1$, or N.

4.2 Imputing Potential Outcomes using Decision Trees: LOOP equals Post-stratification

In this section, we discuss the connection between the LOOP estimator and post-stratification. Post-stratification is a covariate adjustment method made by stratifying on pretreatment variables, estimating the treatment effect within each stratum by taking a simple difference in means, and then taking the weighted average over all strata [13]. We argue that when we impute potential outcomes using a decision tree, the LOOP estimator is equivalent to post-stratification.

Given a single decision tree (fixed in advance), we impute the potential outcomes as follows. First, we assign each observation i to a group; this is done by applying the decision tree to observation i's covariates. (This group may be viewed as a "leaf" or a "stratum.") For each i, we then impute t_i using the average observed outcome of the treated units within the same group (excluding observation i itself). We impute c_i similarly. Thus, using the same argument given above in Section 4.1, it is simple to show that the average of the $\hat{\tau}_i$ within a group is equal to the simple difference within that group. Thus, the average of all the τ_i is a weighted average of the within-group simple differences, *i.e.*, it is a post-stratification estimator.

4.3 Imputing Potential Outcomes using Random Forests

In their analysis of post-stratification, Miratrix et al. show that it is nearly as efficient as blocking. However, one disadvantage of post-stratification is that we must be parsimonious in the number of variables selected. If we include too many covariates, we end up partitioning our data too finely. We can overcome this limitation and also improve on the post-stratified estimate using the LOOP estimator. One advantage of the LOOP estimator is that estimation of m_i is very flexible. One can impute the potential outcomes using any method, so long as $\hat{m}_i$ and X_i are independent. In particular, we can use ensemble methods such as boosting or bagging to improve our estimates over a single decision tree.

One such method is the random forest algorithm, and random forests will be our method of choice for imputing the potential outcomes for the remainder of the paper. In order to impute the potential outcomes using random forests, we could first omit observation i, and then create a random forest using the remaining $N - 1$ observations, which we could use to impute t_i and c_i. However, this would be computationally demanding. Fortunately, it is also unnecessary. Random forests are naturally suited for the LOOP estimator. Although we describe a leave-one-out procedure, we can make use of the out-of-bag predictions in practice. We can therefore fit a single random forest. For each i, we predict c_i and t_i using the out-of-bag predictions, *i.e.,* using only the trees that do not include observation i. By contrast, when imputing the potential outcomes using many other methods, such as OLS, we do need to create a separate model for each i. As a result, imputing the potential outcomes with random forests can be relatively computationally efficient.

Because random forests are typically an improvement over individual decision trees, they

allow us to obtain a more precise estimate of the ATE. By using random forests to effectively improve upon post-stratification, we might even hope to obtain an estimate of the ATE that works as well as or better than if we had used a blocked experimental design. Moreover, random forests essentially provide automatic variable selection, making it unnecessary to decide in advance which covariates should be used. Biau [3] shows that the rate of convergence of the random forest algorithm depends on the number of important variables present, rather than how many noise variables there are.

5 Dependent Treatment Assignments

In the preceding sections, we assumed that the treatment assignments are independent of each other. It is common for researchers to randomly assign a fixed number n of participants to treatment and leave the remaining $N - n$ as controls. In such cases, treatment assignments are not independent. However, we can ensure the independence of X_i and $\hat{m}_i$ as follows: if the i-th observation is assigned to treatment, we randomly pick one of the control observations and drop that observation as well as observation i when fitting our prediction model. Conversely, if the i-th observation is control, we randomly drop one of the treatment observations. Thus, regardless of whether X_i is equal to 0 or 1, when we estimate $\hat{m}_i$, we use $N - 2$ of the remaining $N - 1$ observations. Of these $N - 2$ observations, $n - 1$ will be assigned to treatment, $N - n - 1$ will be assigned to control, and the specific allocation will be independent of X_i. As an example, suppose we have a trial with 10 participants where 7 are randomly assigned to treatment, and suppose that participant 1 is in the treatment group. When we calculate $\hat{m}_1$, we use the 6 remaining participants assigned to treatment, and a random selection of 2 of the participants assigned to control. Similarly, if participant 1 is assigned to control, we would use a random selection of 6 of the participants assigned to treatment, and the remaining 2 partipants assigned to control. Moreover, knowing X_1 doesn't tell us anything about *which* of the remaining 9 partipants will be the 6 in treatment and the 2 in control that we use when we estimate m_1. Thus, $\hat{m}_1$ is independent of X_1.

Since this procedure ensures that $\hat{m}_i$ and X_i are independent, $\hat{\tau}_i$ will remain unbiased. By dropping an extra observation we are losing some information. However, we could repeat this entire procedure many times, producing an unbiased estimate of $\hat{\tau}_i$ each time, which we could then average. In the aggregate, we would then make use of all remaining $N - 1$ observations. Note that in practice, the use of the random drop procedure would not change our estimates much. For example, if we use the random drop procedure with a decision tree, we would still obtain the post-stratified estimate. (See Appendix A for further discussion.)

Note that a similar procedure could be used in a block-randomized experiment, in which a fixed number of participants within each block are assigned to treatment, and the rest to control. In this case, when computing $\hat{m}_i$, we would need to drop an observation that is in the same block as i. This procedure could even be extended to paired designs. In a paired design, both observation i and observation i's pair would need to be dropped. However, all of the remaining observations from the experiment could still be used to produce an estimate of m_i.

6 Variance Estimation

Aronow and Middleton give a conservative estimate of the variance of the Horvitz-Thompson estimator. They also provide an estimate for the variance of their own estimator, but only when the function of the covariates (*i.e.*, our $\hat{m}_i$) is a constant fixed in advance, not computed from the data. In this section, we derive an estimate for the variance of the LOOP estimator. Given the leave-one-out method we use to impute potential outcomes, the jackknife would be an obvious choice for estimating the variance. As Efron and Stein show, the jackknife variance estimate tends to be conservative [6]. However, we found this estimate to be too conservative to be of practical use, especially in the presence of treatment effect heterogeneity. In this section, we provide a different estimate for the variance of our estimator.

6.1 Variance of $\hat{\tau}$

We show that

$$\mathrm{Var}(\hat{\tau}_i) = \frac{1}{p_i(1-p_i)}\mathrm{MSE}(\hat{m}_i) \tag{22}$$

and that

$$\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j) = \gamma_{ij} \tag{23}$$

where

$$\gamma_{ij} = \mathrm{Cov}(\hat{m}_i U_i, \hat{m}_j U_j). \tag{24}$$

We argue that γ_{ij} is negligible, and thus conclude

$$\mathrm{Var}(\hat{\tau}) \approx \frac{1}{N}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{p_i(1-p_i)}\mathrm{MSE}(\hat{m}_i)\right]. \tag{25}$$

First, we find the variance of a single $\hat{\tau}_i$:

$$\begin{aligned}
\mathrm{Var}(\hat{\tau}_i) &= \mathrm{Var}\left[\mathbb{E}(\hat{\tau}_i|\hat{m}_i)\right] + \mathbb{E}[\mathrm{Var}(\hat{\tau}_i|\hat{m}_i)] \\
&= \mathrm{Var}(\tau_i) + \mathbb{E}\left[\mathrm{Var}\left(\frac{1}{p_i}(Y_i-\hat{m}_i)X_i + \frac{1}{1-p_i}(\hat{m}_i - Y_i)(1-X_i)|\hat{m}_i\right)\right] \\
&= 0 + \mathbb{E}\left[\mathrm{Var}\left(\frac{1}{p_i}(t_i-\hat{m}_i)X_i + \frac{1}{1-p_i}(\hat{m}_i - c_i)(1-X_i)|\hat{m}_i\right)\right] \\
&= \frac{1}{p_i^2(1-p_i)^2}\mathbb{E}\left[\mathrm{Var}((1-p_i)(t_i-\hat{m}_i)X_i + p_i(\hat{m}_i-c_i)(1-X_i)|\hat{m}_i)\right] \\
&= \frac{1}{p_i^2(1-p_i)^2}\mathbb{E}[\mathrm{Var}[((1-p_i)t_i + p_i c_i - \hat{m}_i)X_i + p_i(\hat{m}_i - c_i)|\hat{m}_i]] \\
&= \frac{1}{p_i^2(1-p_i)^2}\mathbb{E}[\mathrm{Var}[(m_i-\hat{m}_i)X_i + p_i(\hat{m}_i-c_i)|\hat{m}_i]] \\
&= \frac{1}{p_i^2(1-p_i)^2}\mathbb{E}[(m_i-\hat{m}_i)^2\mathrm{Var}(X_i|\hat{m}_i)] \\
&= \frac{1}{p_i(1-p_i)}\mathbb{E}[(m_i-\hat{m}_i)^2] = \frac{1}{p_i(1-p_i)}\mathrm{MSE}(\hat{m}_i).
\end{aligned} \tag{26}$$

We now analyze the covariance term:

$$\begin{aligned}\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j) &= \mathrm{Cov}[(Y_i - \hat{m}_i)U_i, (Y_j - \hat{m}_j)U_j] \\ &= \mathrm{Cov}(Y_iU_i, Y_jU_j) - \mathrm{Cov}(Y_iU_i, \hat{m}_jU_j) - \mathrm{Cov}(\hat{m}_iU_i, Y_jU_j) \\ &\quad + \mathrm{Cov}(\hat{m}_iU_i, \hat{m}_jU_j). \end{aligned} \tag{27}$$

The first term is zero, as Y_iU_i and Y_jU_j are independent. The second and third terms are also zero:

$$\begin{aligned}\mathrm{Cov}(Y_iU_i, \hat{m}_jU_j) &= \mathbb{E}(Y_iU_i\hat{m}_jU_j) - \mathbb{E}(Y_iU_i)\mathbb{E}(\hat{m}_jU_j) \\ &= \mathbb{E}(Y_iU_i\hat{m}_j)\mathbb{E}(U_j) - \mathbb{E}(Y_iU_i)\mathbb{E}(\hat{m}_j)\mathbb{E}(U_j) \\ &= 0. \end{aligned} \tag{28}$$

Thus,

$$\mathrm{Cov}[(Y_i - \hat{m}_i)U_i, (Y_j - \hat{m}_j)U_j] = \mathrm{Cov}(\hat{m}_iU_i, \hat{m}_jU_j). \tag{29}$$

In most cases, this value is negligible, as U_i and U_j are independent, and $\hat{m}_i$ generally does not depend heavily on U_j, while $\hat{m}_j$ generally does not depend heavily on U_i. There are certain pathological cases where the correlation is large. For example, suppose that for all i $\hat{m}_i = \prod_{k\neq i} U_k$. Then $\hat{m}_iU_i = \prod_{k=1}^{N} U_k$ for all i, so the correlation between $\hat{m}_iU_i$ and $\hat{m}_jU_j$ is 1 for all i and j. Note that we provide a method for estimating these covariances in Appendix B. Nevertheless, the covariance term is generally small. We denote

$$\mathrm{Cov}(\hat{m}_iU_i, \hat{m}_jU_j) = \gamma_{ij} \tag{30}$$

and thus, our expression for the variance becomes

$$\begin{aligned}\mathrm{Var}(\hat{\tau}) &= \frac{1}{N^2}\left[\sum_{i=1}^{N} \frac{1}{p_i(1-p_i)}\mathrm{MSE}(\hat{m}_i) + \sum_{i\neq j} \mathrm{Cov}(\hat{m}_iU_i, \hat{m}_jU_j)\right] \\ &= \frac{1}{N^2}\left[\sum_{i=1}^{N} \frac{1}{p_i(1-p_i)}\mathrm{MSE}(\hat{m}_i) + \sum_{i\neq j} \gamma_{ij}\right] \\ &\approx \frac{1}{N}\left[\frac{1}{N}\sum_{i=1}^{N} \frac{1}{p_i(1-p_i)}\mathrm{MSE}(\hat{m}_i)\right]. \end{aligned} \tag{31}$$

6.2 Estimating the Variance using Cross Validation

In this section, we estimate the variance of the LOOP estimator using cross validation. We assume for the sake of simplicity that $p_i = 1/2$ for all i, so

$$\mathrm{Var}(\hat{\tau}) \approx \frac{4}{N}\left[\frac{1}{N}\sum_{i=1}^{N} \mathrm{MSE}(\hat{m}_i)\right]. \tag{32}$$

In order to estimate this quantity, we use cross validation to estimate the average mean squared errors of $\hat{t}_i$ and $\hat{c}_i$, which we then use to estimate the average MSE of $\hat{m}_i$ (the bracketed term above). We first express the MSE of $\hat{m}_i$ in terms of the MSEs of $\hat{t}_i$ and $\hat{c}_i$:

$$\begin{aligned}\text{MSE}(\hat{m}_i) &= [\mathbb{E}(\hat{m}_i - m_i)]^2 + \text{Var}(\hat{m}_i) \\ &= \frac{1}{4}[\text{MSE}(\hat{t}_i) + \text{MSE}(\hat{c}_i) + 2\text{Cov}(\hat{t}_i, \hat{c}_i) + 2\text{Bias}(\hat{t}_i)\text{Bias}(\hat{c}_i)] \\ &\leq \frac{\text{MSE}(\hat{t}_i) + \text{MSE}(\hat{c}_i) + 2\sqrt{\text{MSE}(\hat{t}_i)\text{MSE}(\hat{c}_i)}}{4}. \end{aligned} \tag{33}$$

See Appendix C for the derivation. We can therefore bound the variance of our estimator as follows:

$$\begin{aligned}\text{Var}(\hat{\tau}) &= \frac{4}{N}\left[\frac{1}{N}\sum_{i=1}^{N}\text{MSE}(\hat{m}_i)\right] \\ &\leq \frac{1}{N}\sum_{i=1}^{N}\text{MSE}(\hat{t}_i) + \frac{1}{N}\sum_{i=1}^{N}\text{MSE}(\hat{c}_i) + 2\frac{1}{N}\sum_{i=1}^{N}\sqrt{\text{MSE}(\hat{t}_i)\text{MSE}(\hat{c}_i)} \\ &\leq \frac{1}{N}\sum_{i=1}^{N}\text{MSE}(\hat{t}_i) + \frac{1}{N}\sum_{i=1}^{N}\text{MSE}(\hat{c}_i) + 2\sqrt{\frac{1}{N}\sum_{i=1}^{N}\text{MSE}(\hat{t}_i)\frac{1}{N}\sum_{i=1}^{N}\text{MSE}(\hat{c}_i)}. \end{aligned} \tag{34}$$

For each observation, we have already estimated $\hat{t}_i$ and $\hat{c}_i$ when we estimated $\hat{m}_i$. We can then calculate the mean squared error on the units that actually received treatment, which we use as our estimate of the average MSE of $\hat{t}_i$ (we do the same procedure on the control units to obtain the estimate for $\hat{c}_i$). We denote the average MSE estimates as:

$$\hat{M}_t = \frac{1}{n}\sum_{i\in T}(\hat{t}_i - t_i)^2 \tag{35}$$

$$\hat{M}_c = \frac{1}{N-n}\sum_{i\in C}(\hat{c}_i - c_i)^2. \tag{36}$$

Our final estimate for the variance is therefore

$$\widehat{\text{Var}}(\hat{\tau}) = \frac{1}{N}\left[\hat{M}_t + \hat{M}_c + 2\sqrt{\hat{M}_t\hat{M}_c}\right]. \tag{37}$$

7 Results

Below, we apply the LOOP estimator (with random forests) to both simulated and actual data. We first provide an illustrative example with simulated data to demonstrate the bias of the point estimate and standard error for the OLS estimator. We then apply the LOOP estimator to the experiment conducted by Gerber and Green on voter turnout [8].

7.1 Simulation

Consider a randomized experiment in which there are $N = 30$ subjects and there is a single covariate, Z, with three possible values: 0, 1, and 2. For each value of Z, there are 10 subjects and each subject has potential outcomes that are generated from a normal distribution with standard deviation 0.1. For $Z = 0$, the control and treatment outcomes have expectations 0 and 1, respectively; for $Z = 1$, the control and treatment outcomes both have expectation 1; and for $Z = 2$, the control and treatment outcomes have expectations 1 and 2.

After generating the treatment and control potential outcomes for the 30 subjects (which we do only once), we create 100,000 random assignment vectors (X) and the 100,000 corresponding vectors of observed outcomes (Y). For each of these, we estimate the average treatment effect and nominal standard error. Below, we compare the results using OLS, the LOOP estimator with random forests, and cross estimation [18] with random forests.[3] The bias is estimated as the mean point estimate minus the true ATE. We also show the mean nominal standard error and estimate the true standard error using the standard deviation of the 100,000 point estimates. The nominal standard errors for the LOOP estimator are calculated using the method of Section 6, while the nominal standard errors for cross estimation are calculated using the estimator provided by [18]. For OLS, the nominal standard errors are calculated using the usual formulas.

Table 1: Simulation Results: LOOP, Cross Estimation, and OLS

Method	**Bias Estimate**	**Mean Nominal SE**	**Estimate of True SE**
LOOP - RF	-0.0006	0.0411	0.0344
Cross Estimation - RF	0.0005	0.0947	0.0334
OLS	-0.0124	0.0949	0.0356

Note: The bias estimates for LOOP and Cross Estimation are not statistically significant from 0.

We can see that the OLS estimate is biased, while the LOOP and cross estimation estimators are both unbiased. Moreover, while the true standard errors of the three methods are similar, the nominal standard errors for OLS and cross estimation are both quite biased. The nominal standard error for LOOP is also biased, but less so.

7.2 Gerber and Green

7.2.1 Description

The Gerber and Green data set comes from an experiment in 1998 involving 29,380 individuals. The experiment sought to examine the effect of three different treatments (personal canvassing, telephone calls, and direct mailers) on voting. For our analysis, we will concentrate on the direct mail experiment. In the direct mail experiment, 14,719 of the subjects

[3]We use the code provided by [18], with two modifications: we increase the number of trees from 500 to 1,500 and we remove the specified node size parameter. These modifications improve performance in the context of this simulation.

were assigned to treatment and the remaining 14,661 were assigned to control. The treatment group was split evenly into groups who would receive one, two, or three mailers. Although the personal canvassing treatment was independent of the mail treatment, telephone calls were not. Because people who received mailers were also more likely to receive phone calls, the treatment of interest is effectively receiving at least one mailer and being more likely to receive phone calls.

7.2.2 Results

In this section, we apply the LOOP estimator to the Gerber and Green data set using the same covariates as used by Gerber and Green. For reasons we will explain shortly, it will be helpful to assume in our analyses that the treatment effect is zero. This assumption is largely supported by the data, although in truth there may be a very small effect. For example, using the simple difference estimator (and all 29,380 observations), we estimate the ATE to be 0.00461.

We compare the nominal standard errors for the simple difference estimator, linear regression, the LOOP estimator (using random forests), and cross estimation (using random forests). We also compare the estimator proposed by Lin [12], which is obtained from the OLS regression of Y on X, Z, and the interaction between X and Z. In our analyses, we consider both the entire set of 29,380 observations and subsets as small as 50. In this way, we can compare the performance of the different methods across a variety of sample sizes.

Nominal standard errors are calculated as in the previous simulation example. We also calculate "true" standard errors as follows. First we assume that there is no treatment effect, and specifically assume that $\tau_i = 0$ for all i. Under this assumption, we may regard all potential outcomes as observed, since $t_i = c_i$ for all i. We may therefore calculate true standard errors by considering permutations of the treatment assignment vector (X) as in the previous simulation example. However, here we consider only 100 permutations rather than 100,000, for computational feasibility.

For each sample size (50, 100, 200, 1000), we take 100 samples, and perform the entire analysis just describe for each sample (resulting in 100 nominal standard errors and 100 "true" standard errors). We then average the results over the 100 samples to get an average nominal SE and an average true SE. The results are given in Table 2.

Table 2: Comparison of Methods using all Covariates

N	Simple Difference	LOOP	Cross Estimation	OLS	OLS Interact
50	0.143	0.135	0.134	0.200	2.535
	0.144	*0.135*	*0.133*	*0.197*	*8.544*
100	0.100	0.096	0.093	0.108	0.155
	0.101	*0.094*	*0.094*	*0.109*	*0.160*
200	0.071	0.066	0.065	0.067	0.072
	0.071	*0.065*	*0.064*	*0.068*	*0.073*
1,000	0.032	0.028	0.028	0.028	0.028
	0.032	*0.028*	*0.028*	*0.028*	*0.028*
29,380	0.0058	0.0049	0.0050	0.0050	0.0050

For each sample size, the first row contains the average nominal standard error and the second row contains the average estimated true standard error in italics.

The LOOP estimate and cross estimation outperform linear regression at small sample sizes, while having comparable performance at large sample sizes. For example, we can see that LOOP and cross estimation outperform both OLS estimates at $N = 100$, while the performance of all four estimators are identical at $N = 1000$. The OLS estimates are outperformed by the simple difference estimator at small sample sizes. Cross estimation and the LOOP estimator with random forests outperform the simple difference estimator at all sample sizes.

Lin states that the OLS with interactions estimator is at least as efficient asymptotically as the simple difference estimator. However, we can see that issues arise in small sample sizes. For example, at $N = 100$, the standard error of the OLS with interactions estimator is much higher than that of the OLS estimator, while the estimates of the standard errors converge for larger sample sizes.

7.2.3 Variable Selection

One reason the OLS estimators perform poorly at smaller sample sizes is that there are a relatively large number of covariates. In this section, we use the same estimators as above, but drop the "Ward" variable, which is a categorical variable with 29 levels.

In this case, the LOOP estimator and the OLS estimators perform similarly, and both outperform the simple difference estimator at all sample sizes. However, the LOOP estimator is essentially unchanged whether we include the "Ward" variable or not, while the OLS estimators improve for small sample sizes. In other words, to improve performance of the OLS estimate, we would have needed to perform variable selection, but this is not necessary when using the LOOP estimator with random forests.

Table 3: Comparison of Methods without "Ward" Variable

N	Simple Difference	LOOP	Cross Estimation	OLS	OLS Interact
50	0.144	0.135	0.131	0.133	0.138
	0.144	*0.128*	*0.129*	*0.134*	*0.139*
100	0.101	0.095	0.091	0.091	0.091
	0.099	*0.091*	*0.091*	*0.089*	*0.090*
200	0.071	0.065	0.063	0.063	0.063
	0.071	*0.063*	*0.063*	*0.063*	*0.063*
1,000	0.032	0.028	0.028	0.028	0.028
	0.031	*0.028*	*0.028*	*0.028*	*0.028*
29,380	0.0058	0.0050	0.0051	0.0051	0.0051

For each sample size, the first row contains the nominal standard error and the second row contains the estimated true standard error in italics.

8 Discussion

One notable benefit of the simple difference estimator is that it is not the result of a specification search. Furthermore, it provides a baseline for comparison. While methods of covariate adjustment can improve the precision of the estimate of the average treatment effect, they often require the researchers to perform variable selection and can result in data snooping. For example, when using post-stratification, we must be careful not to use too many covariates otherwise we partition the data set too finely. Over-adjustment can result in poorer performance with linear regression as well. Freedman notes that the OLS estimator is biased under the Neyman-Rubin model and can be outperformed by the simple difference estimator. He further notes that these issues occur because randomization fails to justify the assumptions of OLS. As seen in the Gerber and Green example, OLS adjustment can hurt performance unless we select a subset of the covariates.

The LOOP estimator solves many of the issues with linear regression. It is an unbiased estimate of the average treatment effect and randomization justifies the assumptions made. One advantage of the LOOP estimator is that estimation of m_i is very flexible. One can impute the potential outcomes using any method, so long as $\hat{m}_i$ and X_i are independent. One baseline approach is to estimate m_i without making use of covariates, simply taking the mean of the observed outcomes in each treatment group. In this case, the LOOP estimator is exactly equal to the simple difference estimator. This suggests that the LOOP estimator will generally outperform the simple difference estimator, so long as we use a sensible method for imputing the potential outcomes. For example, one could estimate m_i using a decision tree (resulting in a post-stratified estimator) or k-Nearest Neighbors.

In this paper, we suggest the use of random forests to impute the potential outcomes, as they are computationally efficient relative to other methods, improve performance over the post-stratified estimate, and allow for automatic variable selection. Because of the auto-

matic variable selection, we can make adjust for covariates without knowing ahead of time which covariates we wish to use. We see that in the Gerber and Green example, our estimator outperforms the simple difference estimator regardless of whether we use the full set or the subset of covariates. Furthermore, researchers are often concerned with the validity of statistical inference after model selection. Because model selection occurs in a "black box" with our method, any post-selection inference is still valid. In particular, when imputing the potential outcomes using random forests, the researcher will not have to do any manual variable selection and can take advantage of the automatic variable selection.

References

[1] Aronow, P. M. and Middleton, J. A. (2013), A class of unbiased estimators of the average treatment effect in randomized experiments, *Journal of Causal Inference* 1(1), 135154.

[2] Begg, C., Cho, M., Eastwood, S., et al. (1996), Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *Journal of the American Medical Association* 276, 637-639.

[3] Biau, G, (2012), Analysis of a random forests model, *Journal of Machine Learning Research* 13, 1063-1095.

[4] Bloniarz, A., Liu, H., Zhang C., Sekhon, J., and Yu B. (2016), Lasso adjustments of treatment effect estimates in randomized experiments, *PNAS* 113(27), 7383-7090.

[5] Breiman, L. (2001), Random forests, *Machine Learning* 45(1), 5-32.

[6] Efron, B. and Stein, C. (1981), The jackknife estimate of variance, *The Annals of Statistics* 9(3), 586-596

[7] Freedman, D. A. (2008), On regression adjustments to experimental data, *Advances in Applied Mathematics* 40(2), 180-193.

[8] Gerber, A. S. and Green, D. P. (2000), The effects of canvassing, telephone calls, and direct mail on voter turnout: a field experiment, *American Political Science Review* 94(3), 653-663.

[9] Governor's Office, Illinois (2016), *Rauner administration moves to improve outcomes for dually-involved youth* [Press Release]. Retrieved from `http://www3.illinois.gov/PressReleases/ShowPressRelease.cfm?SubjectID=3&RecNum=13897`.

[10] Holt, D. and Smith, T.M.F. (1979), Post stratification, *Journal of the Royal Statistical Society, Series A* 142(1), 3346.

[11] Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association* 47, 663685.

[12] Lin, W. (2013), Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique, *The Annals of Applied Statistics* 7(1), 295-318.

[13] Miratrix, L. W., Sekhon, J. S., and Yu, B. (2012), Adjusting treatment effect estimates by post-stratification in randomized experiments, *Journal of the Royal Statistical Society, Series B* 75(2), 369-396.

[14] Rosenbaum, P. R. (2002), Covariance adjustment in randomized experiments and observational studies, *Statistical Science* 17(3), 286-327.

[15] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688-701.

[16] Schulz, K., Altman, D., and Moher, D. (2010), CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials, *BMC medicine* 8(1), 1.

[17] Splawa-Neyman, J., Dabrowska, D.M., and Speed, T.P. (1990), On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science* 5(4), 465-472.

[18] Wager, S., Du, W., Taylor, J., and Tibshirani, R. J (2016), High-dimensional regression adjustments in randomized experiments, *Proceedings of the National Academy of Sciences* 113(45), 12673-12678.

[19] Young, A. (2016), Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results, Working Paper.

A Expectation of the Random Drop Procedure

In this section, we examine how the random drop procedure affects the value of $\hat{\tau}$. Consider the case where we impute t_i as the average of the treated units and c_i as the average of the control units (omitting observation i each time). If unit i was in the control group, then each time we estimate $\hat{m}_i$, we would drop a random observation in the treatment group before taking the averages of the observed outcomes. While we could repeat this procedure many times and average the resulting estimates to get our final estimate of $\hat{m}_i$, we could instead take the expected value of the "random drop" estimate over all possible drops. In this case, the estimate of $\hat{m}_i$ is exactly equal to the estimate had we not dropped any observations in the first place. Without loss of generality, we assume that observation i is assigned to control. Let $\hat{m}_{i,-k}$ and $\hat{\tau}_{i,-k}$ denote the estimates where we randomly dropped the k-th observation

and let $\hat{m}_{i,\cdot}$ and $\hat{\tau}_{i,\cdot}$ denote their expected values over all possible drops.

$$\begin{aligned}
\mathbb{E}_k(\hat{m}_{i,-k}) &= \frac{1}{n}\sum_{k\in T}\hat{m}_{i,-k} \\
&= \frac{1}{n}\sum_{k\in T}\left[\frac{\sum_{j\in T\setminus\{i,k\}}Y_j}{n-1} + \frac{\sum_{j\in C\setminus\{i,k\}}Y_j}{N-n-1}\right] \\
&= \frac{1}{n}\sum_{k\in T}\left[\frac{\sum_{j\in T\setminus\{k\}}Y_j}{n-1}\right] + \frac{1}{n}\sum_{k\in T}\left[\frac{\sum_{j\in C\setminus\{i\}}Y_j}{N-n-1}\right] \\
&= \frac{1}{n}\left[\frac{(n-1)\sum_{j\in T}Y_j}{n-1}\right] + \frac{1}{n}\left[\frac{n\sum_{j\in C\setminus\{i\}}Y_j}{N-n-1}\right] \\
&= \frac{\sum_{j\in T}Y_j}{n} + \frac{\sum_{j\in C\setminus\{i\}}Y_j}{N-n-1}. \qquad (38)
\end{aligned}$$

This last line is equal to the value of $\hat{m}_i$ that we would have gotten had we not dropped any observations besides i. Our estimate for $\hat{\tau}$ would also be the same as if we had not used the random drop procedure (*i.e.*, $\mathbb{E}_k(\hat{\tau}_{i,-k}) = \hat{\tau}_i$). A similar argument can be used to show that if we were to use the random drop procedure when estimating $\hat{m}_i$ using a decision tree, the expected value of $\hat{\tau}$ would still be the post-stratified estimate.

B Estimating the Covariance of $\hat{m}_i U_i$ and $\hat{m}_j U_j$

In this section, we provide an estimate for the covariance of $\hat{m}_i U_i$ and $\hat{m}_j U_j$. First,

$$\begin{aligned}
\mathrm{Cov}(\hat{m}_i U_i, \hat{m}_j U_j) &= \mathrm{Cov}\left[[(1-p_i)\hat{t}_i + p_i\hat{c}_i]U_i, [(1-p_j)\hat{t}_j + p_j\hat{c}_j]U_j\right] \\
&= (1-p_i)(1-p_j)\mathrm{Cov}(\hat{t}_i U_i, \hat{t}_j U_j) + (1-p_i)p_j\mathrm{Cov}(\hat{t}_i U_i, \hat{c}_j U_j) \\
&\quad + p_i(1-p_j)\mathrm{Cov}(\hat{c}_i U_i, \hat{t}_j U_j) + p_i p_j\mathrm{Cov}(\hat{c}_i U_i, \hat{c}_j U_j). \qquad (39)
\end{aligned}$$

Now, we let $\hat{t}_i^{+j}$ denote the estimate of t_i including the j-th observation, where all the treatment assignments of the other $N-2$ observations are kept as is. Similarly, we let $\hat{t}_i^{-j}$ denote the estimate of t_i excluding the j-th observation. Then we have

$$\begin{aligned}
\mathrm{Cov}(\hat{t}_i U_i, \hat{t}_j U_j | U_{k\notin\{i,j\}}) &= \hat{t}_i^{+j}\hat{t}_j^{+i} - \hat{t}_i^{-j}\hat{t}_j^{+i} - \hat{t}_i^{+j}\hat{t}_j^{-i} + \hat{t}_i^{-j}\hat{t}_j^{-i} \\
&= (\hat{t}_i^{+j} - \hat{t}_i^{-j})(\hat{t}_j^{+i} - \hat{t}_j^{-i}) \\
\mathrm{Cov}(\hat{t}_i U_i, \hat{c}_j U_j | U_{k\notin\{i,j\}}) &= \hat{t}_i^{+j}\hat{c}_j^{-i} - \hat{t}_i^{-j}\hat{c}_j^{-i} - \hat{t}_i^{+j}\hat{c}_j^{+i} + \hat{t}_i^{-j}\hat{c}_j^{+i} \\
&= \left(\hat{t}_i^{+j} - \hat{t}_i^{-j}\right)\left(\hat{c}_j^{-i} - \hat{c}_j^{+i}\right) \\
\mathrm{Cov}(\hat{c}_i U_i, \hat{t}_j U_j | U_{k\notin\{i,j\}}) &= \hat{c}_i^{-j}\hat{t}_j^{+i} - \hat{c}_i^{+j}\hat{t}_j^{+i} - \hat{c}_i^{-j}\hat{t}_j^{-i} + \hat{c}_i^{+j}\hat{t}_j^{-i} \\
&= \left(\hat{c}_i^{-j} - \hat{c}_i^{+j}\right)\left(\hat{t}_j^{+i} - \hat{t}_j^{-i}\right) \\
\mathrm{Cov}(\hat{c}_i U_i, \hat{c}_j U_j | U_{k\notin\{i,j\}}) &= \hat{c}_i^{-j}\hat{c}_j^{-i} - \hat{c}_i^{+j}\hat{c}_j^{-i} - \hat{c}_i^{-j}\hat{c}_j^{+i} + \hat{c}_i^{+j}\hat{c}_j^{+i} \\
&= \left(\hat{c}_i^{-j} - \hat{c}_i^{+j}\right)\left(\hat{c}_j^{-i} - \hat{c}_j^{+i}\right). \qquad (40)
\end{aligned}$$

Note that $\hat{t}_i^{+j}$ is calculable when $X_j = 1$, but not when $X_j = 0$, as t_j is not observable when $X_j = 0$. Similarly, $\hat{c}_i^{+j}$ is calculable when $X_j = 0$, but not when $X_j = 1$. Thus, we use following estimate of the covariance (where all the terms are estimable):

$$\widehat{\mathrm{Cov}}(\hat{m}_i U_i, \hat{m}_j U_j) = \begin{cases} \frac{(1-p_i)(1-p_j)}{p_i p_j}(\hat{t}_i^{+j} - \hat{t}_i^{-j})(\hat{t}_j^{+i} - \hat{t}_j^{-i}), & X_i = X_j = 1 \\ (\hat{t}_i^{+j} - \hat{t}_i^{-j})(\hat{c}_j^{-i} - \hat{c}_j^{+i}), & X_i = 0, X_j = 1 \\ (\hat{c}_i^{-j} - \hat{c}_i^{+j})(\hat{t}_j^{+i} - \hat{t}_j^{-i}), & X_i = 1, X_j = 0 \\ \frac{p_i p_j}{(1-p_i)(1-p_j)}(\hat{c}_i^{-j} - \hat{c}_i^{+j})(\hat{c}_j^{-i} - \hat{c}_j^{+i}), & X_i = X_j = 0 \end{cases} \tag{41}$$

which is an unbiased estimate of the covariance:

$$\begin{aligned} &\mathbb{E}[\widehat{\mathrm{Cov}}(\hat{m}_i U_i, \hat{m}_j U_j) | U_{k \notin \{i,j\}}] \\ &= p_i p_j \frac{(1-p_i)(1-p_j)}{p_i p_j}(\hat{t}_i^{+j} - \hat{t}_i^{-j})(\hat{t}_j^{+i} - \hat{t}_j^{-i}) + (1-p_i)p_j(\hat{t}_i^{+j} - \hat{t}_i^{-j})(\hat{c}_j^{-i} - \hat{c}_j^{+i}) \\ &\quad + p_i(1-p_j)(\hat{c}_i^{-j} - \hat{c}_i^{+j})(\hat{t}_j^{+i} - \hat{t}_j^{-i}) + (1-p_i)(1-p_j)\frac{p_i p_j}{(1-p_i)(1-p_j)}(\hat{c}_i^{-j} - \hat{c}_i^{+j})(\hat{c}_j^{-i} - \hat{c}_j^{+i}) \\ &= (1-p_i)(1-p_j)\mathrm{Cov}(\hat{t}_i U_i, \hat{t}_j U_j | U_{k \notin \{i,j\}}) + (1-p_i)p_j \mathrm{Cov}(\hat{t}_i U_i, \hat{c}_j U_j | U_{k \notin \{i,j\}}) \\ &\quad + p_i(1-p_j)\mathrm{Cov}(\hat{c}_i U_i, \hat{t}_j U_j | U_{k \notin \{i,j\}}) + p_i p_j \mathrm{Cov}(\hat{c}_i U_i, \hat{c}_j U_j | U_{k \notin \{i,j\}}) \\ &= \mathrm{Cov}(\hat{m}_i U_i, \hat{m}_j U_j | U_{k \notin \{i,j\}}). \end{aligned} \tag{42}$$

We take the expectation across all randomizations to show $\widehat{\mathrm{Cov}}(\hat{m}_i U_i, \hat{m}_j U_j)$ is unbiased.

$$\begin{aligned} \mathbb{E}[\mathrm{Cov}(\hat{m}_i U_i, \hat{m}_j U_j | U_{k \notin \{i,j\}})] &= \mathbb{E}[\mathbb{E}(\hat{m}_i U_i \hat{m}_j U_j | U_{k \notin \{i,j\}}) - \mathbb{E}(\hat{m}_i U_i | U_{k \notin \{i,j\}})\mathbb{E}(\hat{m}_j U_j | U_{k \notin \{i,j\}})] \\ &= \mathbb{E}[\mathbb{E}(\hat{m}_i U_i \hat{m}_j U_j | U_{k \notin \{i,j\}})] \\ &= \mathbb{E}(\hat{m}_i U_i \hat{m}_j U_j)] \\ &= \mathrm{Cov}(\hat{m}_i U_i, \hat{m}_j U_j) \end{aligned} \tag{43}$$

Summing across all i, j pairs yields an unbiased estimate of $\sum_{i \neq j} \mathrm{Cov}(\hat{m}_i U_i, \hat{m}_j U_j)$.

C The Mean Squared Error of $\hat{m}_i$

Below, we express $\text{MSE}(\hat{m}_i)$ in terms of the MSEs of $\hat{t}_i$ and $\hat{c}_i$:

$$\begin{aligned}
\text{MSE}(\hat{m}_i) &= [\mathbb{E}(\hat{m}_i - m_i)]^2 + \text{Var}(\hat{m}_i) \\
&= \left[\mathbb{E}(\frac{1}{2}(\hat{t}_i + \hat{c}_i - t_i - c_i))\right]^2 + \text{Var}(\frac{\hat{t}_i + \hat{c}_i}{2}) \\
&= \frac{1}{4}[[\mathbb{E}(\hat{t}_i - t_i + \hat{c}_i - c_i)]^2 + \text{Var}(\hat{t}_i) + \text{Var}(\hat{c}_i) + 2\text{Cov}(\hat{t}_i, \hat{c}_i)] \\
&= \frac{1}{4}[[\text{Bias}(\hat{t}_i) + \text{Bias}(\hat{c}_i)]^2 + \text{Var}(\hat{t}_i) + \text{Var}(\hat{c}_i) + 2\text{Cov}(\hat{t}_i, \hat{c}_i)] \\
&= \frac{1}{4}[\text{Bias}^2(\hat{t}_i) + \text{Bias}^2(\hat{c}_i) + 2\text{Bias}(\hat{t}_i)\text{Bias}(\hat{c}_i) + \text{Var}(\hat{t}_i) + \text{Var}(\hat{c}_i) + 2\text{Cov}(\hat{t}_i, \hat{c}_i)] \\
&= \frac{1}{4}[\text{MSE}(\hat{t}_i) + \text{MSE}(\hat{c}_i) + 2\text{Cov}(\hat{t}_i, \hat{c}_i) + 2\text{Bias}(\hat{t}_i)\text{Bias}(\hat{c}_i)] \\
&\leq \frac{\text{MSE}(\hat{t}_i) + \text{MSE}(\hat{c}_i) + 2\sqrt{\text{MSE}(\hat{t}_i)\text{MSE}(\hat{c}_i)}}{4}. \qquad (44)
\end{aligned}$$

To show inequality (44), we prove that:

$$\text{Cov}(\hat{t}_i, \hat{c}_i) + \text{Bias}(\hat{t}_i)\text{Bias}(\hat{c}_i) \leq \sqrt{\text{MSE}(\hat{t}_i)\text{MSE}(\hat{c}_i)}. \qquad (45)$$

The proof is trivial, but included here from sake of completeness.

Proof. Let $\text{Cov}(\hat{t}_i, \hat{c}_i) = C$, $\text{Bias}(\hat{t}_i) = B_t$, $\text{Bias}(\hat{c}_i) = B_c$, $\text{Var}(\hat{t}_i) = V_t$, $\text{Var}(\hat{c}_i) = V_c$:

$$\begin{aligned}
C + B_t B_c &\leq \sqrt{\text{MSE}(\hat{t}_i)\text{MSE}(\hat{c}_i)} \\
(C + B_t B_c)^2 &\leq (B_t^2 + V_t)(B_c^2 + V_c) \\
C^2 + 2CB_t B_c + B_t^2 B_c^2 &\leq V_t V_c + V_t B_c^2 + V_c B_t^2 + B_t^2 B_c^2 \\
C^2 + 2CB_t B_c &\leq V_t V_c + V_t B_c^2 + V_c B_t^2. \qquad (46)
\end{aligned}$$

$\text{Cov}(\hat{t}_i, \hat{c}_i)$ is less than or equal to $\sqrt{\text{Var}(\hat{t}_i)\text{Var}(\hat{c}_i)}$ so it is sufficient to show:

$$\begin{aligned}
V_t V_c + 2\sqrt{V_t V_c} B_t B_c &\leq V_t V_c + V_t B_c^2 + V_c B_t^2 \\
2\sqrt{V_t V_c} B_t B_c &\leq V_t B_c^2 + V_c B_t^2 \\
0 &\leq V_t B_c^2 - 2\sqrt{V_t V_c} B_t B_c + V_c B_t^2 \\
0 &\leq (\sqrt{V_t} B_c - \sqrt{V_c} B_t)^2. \qquad (47)
\end{aligned}$$

□