

The LOOP Estimator: Adjusting for Covariates in Randomized Experiments

Evaluation Review
2018, Vol. 42(4) 458-488
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0193841X18808003
journals.sagepub.com/home/erx



Edward Wu¹  and Johann A. Gagnon-Bartsch¹

Abstract

Background: 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, as 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. **Objectives:** This article aims to produce a covariate adjustment method that allows for automatic variable selection, so that practitioners need not commit to any specific set of covariates prior to seeing the data. **Results:** In this article, we propose the “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. In addition to allowing for automatic variable selection, this estimator is unbiased under the

¹ Department of Statistics, University of Michigan, Ann Arbor, MI, USA

Corresponding Author:

Edward Wu, Department of Statistics, University of Michigan, 311 West Hall, 1085 University Ave., Ann Arbor, MI 48109, USA.

Email: jameswu@umich.edu

Neyman–Rubin model, generally performs at least as well as the unadjusted estimator, and the experimental randomization largely justifies the statistical assumptions made.

Keywords

causal inference, covariate adjustment, potential outcomes, randomized trials

It is common when analyzing randomized controlled trials to adjust for small imbalances in baseline covariates in order to improve the precision of the treatment effect estimate.¹ To avoid the possibility of data snooping, and to ensure the validity of statistical inference, several authors have advocated that the statistical methods be fully specified in advance and reported in the trial protocol (e.g., Begg et al., 1996; Schulz, Altman, & Moher, 2010).² However, in cases where the analysis methods must be prespecified, 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 (e.g., Freedman 2008; Miratrix, Sekhon, & Yu 2013).

A second concern when adjusting for baseline covariates is bias. Statisticians often allow for biased estimates in order to reduce the overall mean squared error, and many common methods for covariate adjustment do introduce a small amount of bias. However, in some cases, practitioners may find exact unbiasedness inherently desirable for various reasons. We discuss one such example in the section “Motivation.” Spiess (2018) presents another argument for unbiasedness when analyzing randomized experiments.

In this article, we propose a covariate adjustment method, the “leave-one-out potential outcomes” (LOOP) estimator, to simultaneously address both the concerns discussed above. The method is unbiased and model selection occurs in a “black box,” so any postselection inference remains valid. In particular, the method allows for automatic variable selection, so one need not know which covariates to use ahead of time. This method is also design based, meaning that the experimental randomization largely justifies the statistical assumptions, and it generally performs no worse than the simple difference-in-means estimator but can often substantially improve performance.

This article is organized as follows. The second section reviews the covariate adjustment literature and relates our method to other estimators.

The third section discusses the randomized trial that motivates our work; in this example, both model selection and bias were concerns. The fourth section introduces notation and assumptions and discusses the simple difference-in-means and LOOP estimators. The fifth section relates the LOOP estimator to poststratification and the simple difference-in-means estimator. In the sixth section, we provide an estimate of the variance. The seventh section discusses how to modify the procedures to account for different experimental designs such as block designs. In the eighth section, we apply the LOOP estimator to examples using simulated data and real experimental data. The ninth section concludes.

Relation to Prior Literature

One of the virtues of randomized experiments is that the physical act of randomization largely justifies the statistical assumptions of the Neyman–Rubin model, a nonparametric model which was first introduced by Neyman (Splawa-Neyman, Dabrowska, & Speed, 1990; translation of the original 1923 paper) and further developed by Rubin (1974). Covariate adjustment is often done through linear regression; however, the standard ordinary least squares (OLS) model is quite different from the Neyman–Rubin model and randomization fails to justify the standard assumptions of OLS. In fact, the OLS estimate is biased under the Neyman–Rubin model; see Freedman (2008) and Lin (2013) for further discussion on OLS adjustments. Other types of regression adjustments can be used: Berk et al. (2013a) build on the work of Freedman (2008) and Lin (2013), while Bloniarz, Liu, Zhang, Sekhon, and Yu (2016) 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. In addition, regression adjustments can be used to analyze randomized experiments besides treatment-control studies (e.g., Lu 2016).

Various other covariate adjustment methods have been proposed, including several that are explicitly design based. For example, poststratification (Holt & Smith, 1979) 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 (2013) explore the properties of the poststratified estimator under the Neyman–Rubin model. Koch, Amara, Davis, and Gillings (1982) and Koch, Tangen, Jung, and Amara (1998) propose a method that tests Fisher’s sharp null hypothesis (i.e., that all individual treatment effects are zero). They compute the covariance matrix of the treatment and covariates under the sharp null and

note that a quadratic form involving this covariance matrix has an approximate χ^2 distribution, which they use to obtain a p value. Rosenbaum (2002) introduces a similar covariate adjustment method that involves inverting hypothesis tests of the sharp null to obtain an estimate of the treatment effect. Rosenbaum's method is quite flexible and allows for automatic variable selection; however, it assumes a constant treatment effect across units. In this article, we propose the LOOP estimator, which is also design based and allows for automatic variable selection. Unlike Rosenbaum, we do not assume a constant treatment effect.

Aronow and Middleton (2013) introduce another design-based estimator, which is related to the Horvitz–Thompson (1952) estimator. This estimator involves the estimation of a function of the covariates such that the function is predictive of the outcome, resulting in a reduction in variance. In addition, so long as this function is independent of the treatment assignment, the resulting estimate of the average treatment effect will be unbiased. Following a result from Williams (1961), Aronow and Middleton (2013) suggest sample splitting to ensure independence when estimating the function of the covariates. However, many of their calculations assume that the function is a constant fixed in advance and not estimated using a sample splitting procedure. In this article, we propose a special case of Aronow and Middleton's estimator with a sample splitting approach. We successively leave out each observation and then impute that observation's treatment and control potential outcomes using a prediction algorithm such as a random forest (Breiman, 2001).

Our work is similar to that of Wager, Du, Taylor, and Tibshirani (2016), who also propose a set of estimators that build on the work of Aronow and Middleton. Wager et al. propose the use of sample splitting and machine learning methods to impute potential outcomes. They also provide a variance estimate but work under a model in which they assume that the experimental units are drawn from a superpopulation and focus primarily on the population average treatment effect. In this article, we assume that the potential outcomes and the covariates are fixed and that the only source of randomness is in the treatment assignment. While the point estimate for the average treatment effect need not change under this model, variance estimation is different, and we derive an estimate for the variance of the LOOP estimator under this framework. Note that we focus specifically on the case where the sample splitting is a leave-one-out procedure. As we will show later, this allows for direct comparison to traditional estimators such as a simple difference-in-means and poststratification.

Our method is also related to the augmented inverse probability weighted (“AIPW”) estimator, which was proposed and developed by Robins, Rotnitzky, and Zhao (1994), Robins (2000), and Scharfstein, Rotnitzky, and Robins (1999) to estimate treatment effects in observational studies with missing data. Like the estimator proposed by Aronow and Middleton (2013), AIPW can be considered an extension of the Horvitz–Thompson estimator. It involves a difference in means (inversely weighted by the propensity score) and a regression adjustment based on the expectation of the outcome conditional on the covariates and treatment assignment. See also Chernozhukov et al. (2018) for a related estimator, which employs both sample splitting and machine learning methods to estimate the treatment effect in a high-dimensional setting.

Several other methods use an AIPW-like estimator specifically in randomized experiments (e.g., Spiess, 2018; Rothe, 2018; Tsiatis, Davidian, Zhang, & Lu, 2008). Tsiatis, Davidian, Zhang, and Lu (2008) separate the modeling of covariate-outcome relationships and the evaluation of the treatment effect in order to ensure valid inference after variable selection. Other methods have been proposed to ensure valid postselection inferences. For example, Moore and van der Laan (2009) use targeted maximum likelihood estimation to make covariate adjustments when the outcome is binary. This method involves modeling the probability that the outcome will be 0 or 1 conditional upon the covariates and the treatment assignment. One can use any procedure to model these conditional probabilities, including methods with automatic variable selection. Steingrimsson, Hanley, and Rosenblum (2017) give recommendations for the use of targeted maximum likelihood estimation in practice.

Motivation

Our work is motivated by a so-called pay for success program in the state of Illinois. In brief, a pay for success program is one in which a 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 6-year experiment in which eligible youth are randomly

selected to receive either the special services or ordinary care. The evaluation will be conducted by researchers in the School of Social Work at the University of Michigan, and we assisted the evaluators in planning the design and analysis of the experiment.

Several hundred youth are expected to take part in the program. Eligible participants are independently randomized to treatment or control, each with probability $1/2$. 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. The interested parties (the state, the outside organization providing the services, and the evaluators) agreed that some form of adjustment for these covariates would be desirable. However, there was initially no clear consensus on which adjustment procedure to use.

One concern was bias. Unbiasedness was felt to be desirable, perhaps more so in this example than in many others, because the state's payment rate 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. Indeed, one high ranking state official was opposed to any amount of bias, even if it might reduce the mean squared error. To paraphrase, the magnitude of the error was not so much a concern, as long as it was a fair bet. Other officials were open to using a biased estimator, so long as the bias was negligible. Critically, however, it was felt that the bias should still be quantified, and in the case of biased estimators, it was unclear how to produce a concrete number for the bias. For this reason as well, an unbiased estimator was preferred. Ultimately, it was decided to use poststratification.

A second concern was which covariates to adjust for. It was required to fully specify the analysis protocol in advance. Many potential covariates were available; however, adjusting for too many covariates could result in overadjustment, leading to inflated variance. Poststratification is especially sensitive to overadjustment, and considerable discussion was required to come to a consensus on both the number of covariates and which specific covariates to be used.

The challenges outlined above motivate our work. We wish to produce a method that provides automatic variable selection in order to eliminate the guesswork in deciding which covariates to use, while remaining exactly unbiased under the Neyman–Rubin model.

The LOOP Estimator

In this section, we introduce the LOOP estimator, which we can use to obtain an unbiased estimate of the average treatment effect while adjusting for covariates.

Model and Notation

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

$$T_i \perp\!\!\!\perp T_j,$$

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

$$p_i = P(T_i = 1),$$

and assume $0 < p_i < 1$. In some parts of this article, we assume for simplicity (and without much loss of generality) that $p_i = p$ for all i and for some fixed constant p , but for now, we explicitly let p_i vary from subject to subject.

Associated with each of the N participants are two fixed (nonrandom) 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

$$Y_i = T_i t_i + (1 - T_i) c_i.$$

We define the individual treatment effect τ_i as

$$\tau_i = t_i - c_i,$$

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

$$\bar{\tau} = \frac{1}{N} \sum_{i=1}^N \tau_i, \quad (1)$$

which is our primary parameter of interest.

Lastly, some additional notation. Let $T = \{i : T_i = 1\}$ and $C = \{i : T_i = 0\}$. 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

$$m_i = (1 - p_i)t_i + p_i c_i.$$

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 (signed) inverse probability weights U_i as

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

and note that U_i has expectation 0.

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 $\bar{\tau}$ 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:

$$\hat{\tau}_{sd} = \frac{1}{n} \sum_{i \in T} Y_i - \frac{1}{N - n} \sum_{i \in C} Y_i. \quad (2)$$

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

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:

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

and thus

$$\begin{aligned} \mathbb{E}(Y_i U_i) &= \frac{t_i}{p_i} P(T_i = 1) + \frac{-c_i}{1 - p_i} P(T_i = 0), \\ &= t_i - c_i. \end{aligned}$$

Although this is an unbiased estimator of τ_i , it generally has very high variance and is therefore not useful for practical purposes. 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 $-2Y_i$.

As an alternative estimator of τ_i , consider

$$\hat{\tau}_i = (Y_i - \hat{m}_i)U_i. \quad (3)$$

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 :

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

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

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

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 the section “Variance Estimation,” we show that

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

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

Finally, note that \hat{m}_i in Equation 3 plays the same role as the “augmented” portion of the AIPW estimator as described by Lunceford and Davidian (2004) and the function of the covariates in the estimator of Aronow and Middleton (2013).

Leave-One-Out Imputation

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

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^N \hat{\tau}_i, \quad (4)$$

where $\hat{\tau}_i$ is defined as in Equation (3) 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 and random forests). Having obtained estimates \hat{t}_i and \hat{c}_i , we then set

$$\hat{m}_i = (1 - p_i)\hat{t}_i + p_i\hat{c}_i. \quad (5)$$

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 T 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 compute \hat{m}_i as in Equation 5.

Because we leave out the i th observation when we compute \hat{m}_i , it follows that T_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 T_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 T_i .

It is worth noting that although we use the individual treatment effect estimates $\hat{\tau}_i$ in this article 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. Athey and Imbens (2016) and Nie and Wager (2017) use similar formulations for estimating heterogeneous treatment effects. With this in mind, we summarize below three useful facts about $\hat{\tau}_i$, the latter two of which we show in the section “Variance Estimation”:

$$\begin{aligned} \mathbb{E}(\hat{\tau}_i) &= \tau_i, \\ \text{Var}(\hat{\tau}_i) &= \frac{1}{p_i(1 - p_i)} \mathbb{E}[(\hat{m}_i - m_i)^2], \\ \text{Cov}(\hat{\tau}_i, \hat{\tau}_j) &= \text{Cov}(\hat{m}_i U_i, \hat{m}_j U_j). \end{aligned} \quad (6)$$

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

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 poststratification and the LOOP estimator. Finally, we propose the use of random forests, which may provide an improvement over poststratification and allow us to take advantage of automatic variable selection.

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 . That is, we estimate t_i and c_i as:

$$\hat{t}_i = \frac{\sum_{k \in T \setminus i} Y_k}{n - T_i}, \quad (7)$$

$$\hat{c}_i = \frac{\sum_{k \in C \setminus i} Y_k}{(N - n) - (1 - T_i)}. \quad (8)$$

If the assignment probabilities are all equal, that is, 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 in Online Appendix A. As a result of this equivalence, we conclude that in practice the LOOP estimator will typically perform no worse, or at least not that much worse, than the simple difference estimator. More precisely, in the section “Variance Estimation,” we show that the variance of the LOOP estimator is directly related to the mean squared error of the \hat{m}_i terms. Thus, the LOOP estimator will outperform the simple difference estimator as long as we improve the imputation of the potential outcomes beyond this baseline approach (mean imputation). In addition, the equivalence between the LOOP estimator and the simple difference estimator provides us with some reassurance that the leave-one-out procedure does not inherently introduce extra variance.

Imputing Potential Outcomes Using Decision Trees: LOOP Equals Poststratification

In this section, we discuss the connection between the LOOP estimator and poststratification. Poststratification 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 (Holt & Smith, 1979). We argue that when we impute potential outcomes using a decision tree (see James, Witten, Hastie, & Tibshirani, 2013, for a summary of decision trees), the LOOP estimator is equivalent to poststratification.

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 the section "Imputing Potential Outcomes Ignoring Covariates: LOOP Equals the Simple Difference Estimator," it is simple to show that the average of the \hat{t}_i within a group is equal to the simple difference within that group. Thus, the average of all the \hat{t}_i is a weighted average of the within-group simple differences, that is, it is a poststratification estimator.

Imputing Potential Outcomes Using Random Forests

In their analysis of poststratification, Miratrix et al. (2013) show that poststratification is nearly as efficient as blocking. However, one disadvantage of poststratification 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 poststratified 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 T_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 this article. For a description of tree-based methods, including random forests, see James, Witten, Hastie, and Tibshirani (2013). 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, doing this for each i would be computationally demanding. Fortunately, it is also unnecessary. Because we are using a leave-one-out procedure, and because

out-of-bag predictions are essentially leave-one-out predictions, we can simply make use of the out-of-bag predictions. To clarify, random forests are an ensemble of many decision trees, each of which is constructed using a bootstrap sample. In fitting any given tree, some number of observations will be left out. The out-of-bag prediction for the i th observation is the prediction made using the trees that do not include observation i and is effectively a leave-one-out prediction. We can therefore fit just two random forests (one on the treatment units and one on the control units) and impute the potential outcomes using the out-of-bag predictions. 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 average treatment effect $\bar{\tau}$. By using random forests to effectively improve upon poststratification, we might even hope to obtain an estimate of $\bar{\tau}$ 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 (2012) 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. Given these properties and the computational efficiency of random forests, we see that random forests are naturally suited for the LOOP estimator.

Variance Estimation

Aronow and Middleton (2013) give a conservative estimate of the variance of the Horvitz–Thompson estimator. They also provide an estimate for the variance of their own estimator; however, this estimate is derived under the assumption that the function of the covariates (i.e., our \hat{m}_i) is a constant fixed in advance, not computed from the data. Wager et al. (2016) provide a variance estimate for their method but assume that the experimental units are drawn from a superpopulation. In this section, we derive an estimate for the variance of the LOOP estimator under the assumption that the treatment assignment is the only source of randomness. 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 (1981) show, the jackknife variance estimate tends to be conservative. However, we

found this estimate to be excessively conservative in the presence of treatment effect heterogeneity. Here we provide a different estimate for the variance of our estimator. In the section “Variance of $\hat{\tau}$,” we calculate the true variance of $\hat{\tau}$, and then in the section “Estimating the Variance,” we produce an estimate.

Variance of $\hat{\tau}$

In Online Appendix B.1, we show that:

$$\text{Var}(\hat{\tau}_i) = \frac{1}{p_i(1-p_i)} \text{MSE}(\hat{m}_i), \quad (9)$$

and that

$$\gamma_{ij} = \text{Cov}(\hat{\tau}_i, \hat{\tau}_j) = \rho_{ij} \sqrt{\frac{\text{Var}(\hat{m}_i) \text{Var}(\hat{m}_j)}{p_i p_j (1-p_i)(1-p_j)}}, \quad (10)$$

where

$$\rho_{ij} = \text{Corr}(\hat{m}_i U_i, \hat{m}_j U_j).$$

Combining Equations 9 and 10 yields:

$$\text{Var}(\hat{\tau}) = \frac{1}{N^2} \left[\sum_{i=1}^N \frac{1}{p_i(1-p_i)} \text{MSE}(\hat{m}_i) + \sum_{i \neq j} \gamma_{ij} \right]. \quad (11)$$

Limiting our attention to the special case that $p_i = p$ for all i ,

$$\begin{aligned} \text{Var}(\hat{\tau}) &= \frac{\overline{\text{MSE}}}{Np(1-p)} + \frac{\sum_{i \neq j} \gamma_{ij}}{N^2}, \\ &= \frac{\overline{\text{MSE}}}{Np(1-p)} + \frac{(N-1)\bar{\gamma}}{N}, \end{aligned} \quad (12)$$

where

$$\overline{\text{MSE}} = \frac{1}{N} \sum_{i=1}^N \text{MSE}(\hat{m}_i),$$

and

$$\bar{\gamma} = \frac{1}{N(N-1)} \sum_{i \neq j} \gamma_{ij}.$$

Estimating the Variance

In Online Appendix B.2, we show that when $p_i = p$ for all i ,

$$\frac{\overline{\text{MSE}}}{Np(1-p)} \leq \frac{1}{N} \left[\frac{1-p}{p} M_t + \frac{p}{1-p} M_c + 2\sqrt{M_t M_c} \right], \quad (13)$$

where

$$M_t = \frac{1}{N} \sum_{i=1}^N \text{MSE}(\hat{t}_i),$$

and

$$M_c = \frac{1}{N} \sum_{i=1}^N \text{MSE}(\hat{c}_i).$$

We estimate M_t and M_c by leave-one-out cross validation:

$$\hat{M}_t = \frac{1}{Np} \sum_{i \in T} (\hat{t}_i - t_i)^2, \quad (14)$$

$$\hat{M}_c = \frac{1}{N(1-p)} \sum_{i \in C} (\hat{c}_i - c_i)^2. \quad (15)$$

In Online Appendix B.3, we show that these estimates are unbiased. We plug Equations 14 and 15 into the bound (Equation 13) to obtain an estimate for the first term in Equation 12:

$$\frac{1}{N} \left[\frac{1-p}{p} \hat{M}_t + \frac{p}{1-p} \hat{M}_c + 2\sqrt{\hat{M}_t \hat{M}_c} \right]. \quad (16)$$

Next, we provide an unbiased estimator of γ_{ij} (and thus, $\bar{\gamma}$) in Online Appendix B.4. Specifically, we have:

$$\hat{\gamma}_{ij} = \begin{cases} \frac{(1-p)^2}{p^2} (\hat{t}_i^{+j} - \hat{t}_i^{-j}) (\hat{t}_j^{+i} - \hat{t}_j^{-i}), & T_i = T_j = 1 \\ \begin{cases} (\hat{t}_i^{+j} - \hat{t}_i^{-j}) (\hat{c}_j^{-i} - \hat{c}_j^{+i}), & T_i = 0, T_j = 1 \\ (\hat{c}_i^{-j} - \hat{c}_i^{+j}) (\hat{t}_j^{+i} - \hat{t}_j^{-i}), & T_i = 1, T_j = 0 \end{cases} \\ \frac{p^2}{(1-p)^2} (\hat{c}_i^{-j} - \hat{c}_i^{+j}) (\hat{c}_j^{-i} - \hat{c}_j^{+i}), & T_i = T_j = 0 \end{cases}, \quad (17)$$

where \hat{t}_i^{-j} is an estimate of t_i excluding the j th observation (in addition to the i th observation). We let \hat{t}_i^{+j} denote an estimate of t_i including the j th observation and assuming $T_j = 1$. Note that this is only calculable when $T_j = 1$, in which case $\hat{t}_i^{+j} = \hat{t}_i$. We define \hat{c}_i^{-j} and \hat{c}_i^{+j} similarly. We plug this estimate, $\hat{\gamma}_{ij}$, into the second term of Equation 12 and add to the plug-in estimator (Equation 16) for the bound (Equation 13) to obtain an estimate of the variance of $\hat{\tau}$:

$$\widehat{\text{Var}}(\hat{\tau}) = \frac{1}{N} \left[\frac{1-p}{p} \hat{M}_t + \frac{p}{1-p} \hat{M}_c + 2\sqrt{\hat{M}_t \hat{M}_c} \right] + \frac{1}{N^2} \sum_{i \neq j} \hat{\gamma}_{ij}. \quad (18)$$

Estimating the Variance in Practice

In practice, we recommend making two modifications when estimating the variance. First, we recommend estimating M_t and M_c as

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

and

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

particularly when N is small. Note that these approximations require that $0 < n < N$.

Second, we recommend omitting the second term in Equation 18 for computational efficiency. In many cases, γ_{ij} is negligible in the sense that γ_{ij} (and likewise $\bar{\gamma}$) goes to zero faster than $1/N$. For example, suppose that under suitable regularity conditions, $\text{Var}(\hat{m}_i)$ and $\text{Var}(\hat{m}_j)$ go to 0 at rate $1/N$. Then, if ρ_{ij} goes to zero (at any rate), γ_{ij} will go to zero faster than $1/N$. Online Appendix C gives a more formal argument. We also provide simulation results in the section “Simulation 3: Negligibility of $\bar{\gamma}$ ” to demonstrate empirically that $N\bar{\gamma}$ goes to 0 as N increases.

To see why we might expect ρ_{ij} (and likewise $\bar{\rho}$) to go to zero, recall that U_i and U_j are independent. Thus, even if \hat{m}_i and \hat{m}_j are correlated (which they typically will be), ρ_{ij} may still be negligible. Indeed, if \hat{m}_i and \hat{m}_j are perfectly correlated, then $\rho_{ij} = 0$. The only reason for $\hat{m}_i U_i$ and $\hat{m}_j U_j$ to be correlated would be through the dependence of \hat{m}_i on U_j , and of \hat{m}_j on U_i . These dependencies will typically decay as N grows. As an illustrative example, suppose that for all i , \hat{m}_i is a linear estimator, that is, for some constants $a_{i,k}$

$$\hat{m}_i = a_{i,0} + \sum_{k \neq i} a_{i,k} U_k.$$

In this case, it can be shown (see Online Appendix C.1) that \bar{p} goes to zero at rate $1/N$; more specifically, we show $\bar{p} \leq 1/(N-1)$. Indeed, we further show that if \hat{m}_i is a polynomial function of degree D for all i , then $\bar{p} \leq D/(N-1)$. Note that there do exist certain pathological cases where \bar{p} can be large. For example, suppose that for all i , $\hat{m}_i = \prod_{k \neq i} U_k$. Then, $\hat{m}_i U_i = \prod_{k=1}^N U_k$ for all i , so the correlation between $\hat{m}_i U_i$ and $\hat{m}_j U_j$ is exactly 1.

The two modifications discussed in this section yield the following estimate for the variance of $\hat{\tau}$:

$$\widetilde{\text{Var}}(\hat{\tau}) = \frac{1}{N} \left[\frac{1-p}{p} \tilde{M}_t + \frac{p}{1-p} \tilde{M}_c + 2\sqrt{\tilde{M}_t \tilde{M}_c} \right]. \quad (19)$$

If there is concern that in a particular application \bar{p} is not negligible—either due to concern that \bar{p} may not go to zero faster than $1/N$ or simply due to concern that N is not large enough—we can instead use Equation 18 to estimate the variance of $\hat{\tau}$.

Relationship Between $\widetilde{\text{Var}}(\hat{\tau})$ and the Sample Variance

We show in Online Appendix D that when we impute potential outcomes ignoring covariates (i.e., we calculate \hat{c}_i and \hat{t}_i as in Equations 7 and 8),

$$\tilde{M}_t = \frac{n}{n-1} s_t^2, \quad (20)$$

and

$$\tilde{M}_c = \frac{N-n}{N-n-1} s_c^2, \quad (21)$$

where s_t^2 and s_c^2 are the standard sample variances (of the treated and control units). We show in Online Appendix D that plugging Equations 20 and 21 into Equation 19 yields the following inequality:

$$\begin{aligned} \widetilde{\text{Var}}(\hat{\tau}) &\leq \left(\frac{n}{Np} \right) \frac{s_t^2}{n-1} + \left(\frac{N-n}{N(1-p)} \right) \frac{s_c^2}{N-n-1}, \\ &\approx \frac{s_t^2}{n-1} + \frac{s_c^2}{N-n-1}, \end{aligned} \quad (22)$$

with equality in Equation 22 when \tilde{M}_t and \tilde{M}_c are equal. Thus, our variance estimate provides a result roughly equal to or slightly better than if we had performed a t test. For a related discussion, see Aronow, Green, and Lee (2014).

Dependent Treatment Assignments

In the preceding sections, we assumed that the treatment assignments are independent of each other. In this section, we consider study designs in which the treatment assignments are not independent. For example, it is common for researchers to randomly assign a fixed number n of participants to treatment and leave the remaining $N - n$ as controls (i.e., complete randomization). In such cases, treatment assignments are not independent. However, we can ensure the independence of T_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 imputation model. Conversely, if the i th observation is control, we randomly drop one of the treatment observations. Thus, regardless of whether T_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 T_i . We give an example to illustrate this “random drop” procedure in Online Appendix E.1.

Because this procedure ensures that \hat{m}_i and T_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 poststratified estimate (see Online Appendix E.2 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 .

Table 1. Simulation 1: Potential Outcome Values.

Treatment Effects	Z Value	Mean c_i	Mean t_i
Heterogeneous	0	0	1
	1	1	1
	2	1	2
Homogeneous	0	0	1
	1	1	2
	2	1	2

Finally, we note that the independence of \hat{m}_i and T_i implies that Equation 9 continues to hold. However, Equation 10 is no longer valid, due to the dependence of U_i and U_j . Variance estimation in this context may therefore require a modified approach.

Results

Below, we apply the LOOP estimator (with random forests) to both simulated and actual data. In our first simulation, we compare methods when the treatment effects are either homogeneous or heterogeneous and also demonstrate the bias of the point estimate and standard error for the OLS estimator. Next, we consider a simulation in which we examine the performance of the LOOP estimator when many of the covariates are not predictive. In our third simulation, we empirically demonstrate that the covariance terms discussed in the section “Variance Estimation” are negligible. Finally, we apply the LOOP estimator to the experiment conducted by Barrera-Osorio, Bertrand, Linden, and Perez-Calle (2011) on the effects of various cash transfer programs on educational outcomes in Colombia.

Simulation 1: Heterogeneous and Homogeneous Treatment Effects

Consider a randomized experiment in which there are N subjects and there is a single covariate, Z , with three possible values: 0, 1, and 2. For each value of Z , there are $N/3$ subjects and each subject has potential outcomes that are generated from a normal distribution with means given in Table 1 and standard deviation .1. We consider two scenarios, one where the treatment effects are heterogeneous and the other with homogeneous treatment effects. We consider four cases: (a) $N = 30$ and heterogeneous treatment effects, (b) $N = 100$ and heterogeneous treatment effects, (c) $N = 30$ and homogeneous treatment effects, and (d) $N = 100$ and homogeneous treatment effects.

For each of the four cases, we do the following. We generate a single set of treatment and control potential outcomes for the N subjects. We then create 100,000 random assignment vectors (T), where the treatment assignments are independent Bernoulli random variables with probability 1/2. For each of these 100,000 treatment assignment vectors, we compute the observed outcomes (Y) and estimate the average treatment effect and nominal standard error.

We compare the results using OLS, the LOOP estimator with random forests, and cross estimation with random forests (Wager, Du, Taylor, & Tibshirani, 2016). Note that for cross estimation, we use the code provided on GitHub; however, we remove the specified node size parameter. This modification improves performance in the context of this simulation. The bias is estimated as the mean point estimate minus the true average treatment effect. 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 the section “Estimating the Variance in Practice,” while the nominal standard errors for cross estimation are calculated using the estimator provided by Wager et al. (2016). For OLS, the point estimate is obtained by regressing Y on T and Z (without any interaction terms), while the nominal standard errors are calculated using the usual formulas (not robust standard errors). Z is treated as a continuous variable in the regression (not as a factor). We also compute the coverage probabilities at a confidence level of 95%. We show the results in Table 2. Finally, note that in Table 2, the nominal standard error refers to the mean nominal standard error over the 100,000 trials, while the true standard error refers to the estimate for the true standard error described above. We continue this practice throughout the remainder of this article.

We can see that LOOP and cross estimation are both unbiased, while the OLS estimate is biased. This bias is smaller for homogeneous treatment effects and when N is larger. We can also see that the true standard errors are similar for LOOP and cross estimation. However, in the case of heterogeneous treatment effects, the nominal standard error of cross estimation is quite conservative, even when N increases. The nominal standard error for LOOP is also conservative, but less so. In the case of cross estimation, this conservative bias is partially because Wager et al. assume that the experimental units are drawn from a superpopulation and must account for this additional uncertainty. For LOOP, the conservative bias is related to the inequality (Equation 13). For a discussion on a related inequality for the simple difference estimator, see Aronow et al. (2014).

Table 2. Simulation I Results.

Method	Bias Est.	Nominal SE	True SE	Coverage (%)
(a) $N = 30$, Heterogeneous treatment effects				
LOOP	−.00001	.045	.039	98.56
Cross estimation	.00059	.106	.038	99.70
OLS	−.01396	.108	.044	99.90
(b) $N = 100$, Heterogeneous treatment effects				
LOOP	.00000	.021	.014	99.73
Cross estimation	.00000	.051	.014	100.00
OLS	−.00339	.051	.015	100.00
(c) $N = 30$, Homogeneous treatment effects				
LOOP	−.00003	.046	.040	98.53
Cross estimation	.00007	.045	.039	98.43
OLS	−.00152	.086	.083	95.72
(d) $N = 100$, Homogeneous treatment effects				
LOOP	.00000	.021	.014	99.71
Cross estimation	.00000	.021	.014	99.73
OLS	.00014	.052	.049	95.90

Note. LOOP = leave-one-out potential outcomes estimator; OLS = ordinary least squares.

Technical note. Cross estimation is slightly biased as implemented. This is due to the difference between the out-of-bag and the leave-one-out estimates of the potential outcomes. This issue can easily be fixed by reducing (by one) the size of the bootstrap sample used in the random forest when making out-of-bag predictions of the potential outcomes.

Simulation 2: Estimating the Treatment Effect for a Binary Response

In our second simulation, we consider a randomized experiment in which the response is either zero or one. Each of the N subjects has one of three sets of potential outcomes: (a) zero regardless of treatment assignment, (b) zero if control and one if treatment, and (c) one regardless of treatment assignment. Like in the previous simulation, the treatment assignments are independent Bernoulli random variables with probability 1/2. We also have one covariate (Z_1) that is predictive of the outcome. Higher values of this covariate indicate that the participant is more likely to be in groups (b) or (c) than group (a). Finally, we assume there are k noise covariates (Z_k).

We generate Z_1 from a standard normal distribution. For each subject i , the probabilities that the subject ends up in each group are

determined as follows: We calculate $w_{i1} = 1$, $w_{i2} = \exp(0.5c \times Z_{i1})$, and $w_{i3} = \exp(c \times Z_{i1})$, where c is a positive constant. The probability that observation i is assigned to group j is $p_{ij} = w_{ij} / (w_{i1} + w_{i2} + w_{i3})$. Thus, higher values of c indicate Z_1 is more predictive of outcome. In addition, observation i is most likely to be in the third group (and least likely to be in the first group) if Z_{i1} is positive.

Under this framework, we consider three sets of simulations. First, we assume that both the number of subjects ($N = 200$) and the predictive power of Z_1 ($c = 3$) are constant and vary the number of noise covariates (from $k = 5$ to $k = 100$ in increments of 5). Next, we fix the predictive power of Z_1 ($c = 3$) and the number of noise covariates ($k = 50$) and vary the number of subjects from 100 to 1,000 in increments of 50. Finally, we fix the number of subjects ($N = 200$) and noise covariates ($k = 50$) and vary the predictive power of Z_1 (from $c = 0$ to $c = 5.5$ in increments of 0.5). We run 10,000 trials for each simulation. For each set of simulations, we index the results to the true standard error for the simple difference estimator. We show the results in Figure 1.

We observe that while the performance of OLS declines as the number of noise covariates increases, the performance of LOOP remains constant relative to the simple difference estimator. Similarly, OLS performs worse than the simple difference estimator when the number of subjects is small, while the LOOP estimator outperforms the simple difference estimator for all sample sizes. Finally, it is important to note that covariate adjustment does not help when the covariates are not useful for predicting the outcomes. When Z_1 is predictive of the outcome, LOOP outperforms the simple difference estimator. However, we note that even when Z_1 is not predictive of outcome, the performance of the LOOP estimator is still comparable to that of the simple difference estimator. We discuss this further in the section “Cash Transfer Programs and Enrollment,” where we apply the LOOP estimator to actual experimental data.

Technical note. We slightly modify the procedure described in the section “Simulation 1: Heterogeneous and Homogeneous Treatment Effects.” This is because we compare different simulations in each chart with varying parameter values, and we wish to avoid the variability associated with using a single set of potential outcomes for each simulation. For each of the 10,000 trials, we generate new covariates and potential outcomes and obtain a point estimate and a nominal standard error. We then calculate the nominal standard error as the average of the 10,000 nominal standard errors and the true standard error by taking the standard deviation of the 10,000

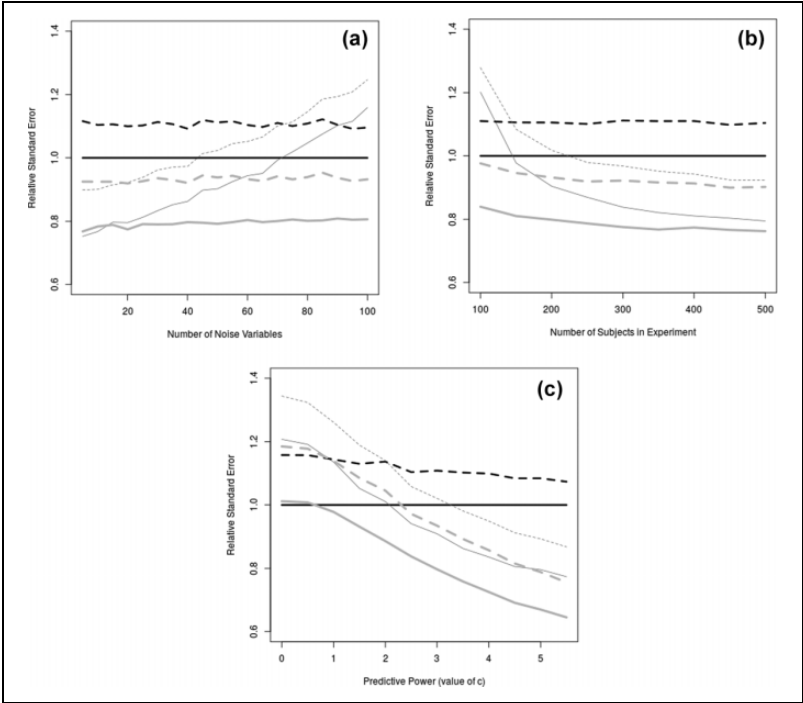


Figure 1. Comparison of standard errors for Simulation 2. All standard errors are relative. That is, each value has been divided by the standard error for the simple difference estimator. We use solid lines to denote the the true standard error and dotted lines to denote the nominal standard error. Method used is shown by the color and width of the lines: (1) simple difference estimator, black lines, (2) ordinary least squares, thin gray lines, and (3) leave-one-out potential outcomes estimator, bold light gray lines.

differences between each point estimate and the true $\bar{\tau}$ for that trial (i.e., the standard deviation of the 10,000 values for $\hat{\tau} - \bar{\tau}$).

Simulation 3: Negligibility of $\bar{\gamma}$

In the section “Estimating the Variance in Practice,” we argue that $\bar{\gamma}$ is typically negligible and can often be ignored when estimating the variance of $\hat{\tau}$. To support this argument, we show via simulation that $N\bar{\gamma}$ goes to zero as N increases. For this simulation, we generate a single set of

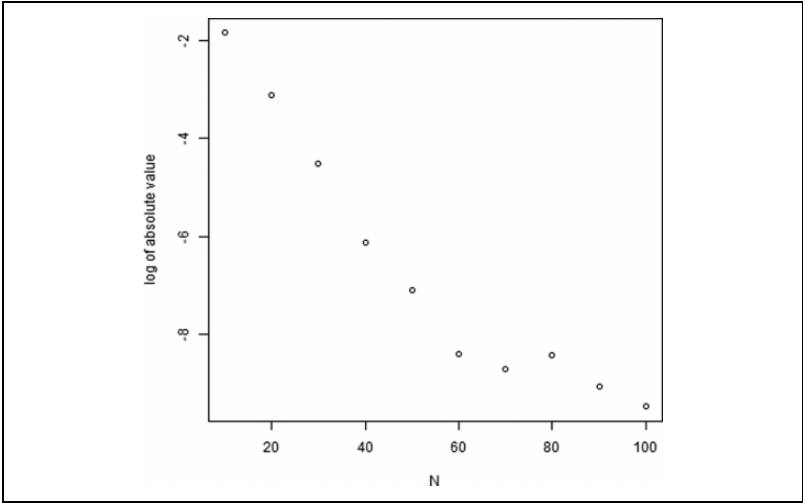


Figure 2. Estimate of $|N\tilde{\gamma}|$ for different values of N ; values are plotted on a log scale. Note that the estimates begin to taper at around $N = 70$. This is due to the standard error of our estimate $\tilde{\gamma}$ of $\bar{\gamma}$. See Online Appendix F.1 for more details.

$N = 100$ subjects using the setup of Simulation 1 (with heterogeneous treatment effects) for the covariates and potential outcomes. We then estimate

$$\bar{\gamma} = \frac{\sum_{i \neq j} \gamma_{ij}}{N(N-1)},$$

for each of the first $N = 10, 20, \dots, 100$ observations. For each N , we generate 100,000 treatment assignment vectors, calculate the $\hat{\tau}_i$ s for each treatment assignment vector, and use the results to obtain a simulation estimate $\tilde{\gamma}$ of $\bar{\gamma}$, along with a standard error for this estimate. In Figure 2, we plot $|N\tilde{\gamma}|$ on a log scale. We can see that the value of $|N\tilde{\gamma}|$ declines as N increases. For a table of the values and standard errors of $\tilde{\gamma}$, see Online Appendix F.1.

Cash Transfer Programs and Enrollment

In their experiment in 2005, Barrera-Orsorio et al. studied the effects of several conditional cash transfer programs on educational outcomes for students in Bogota, Colombia. They conducted experiments in two

localities of Bogota, San Cristobal and Suba. For our analysis, we focus on the San Cristobal experiment. The San Cristobal experiment involved 10,907 students from Grades 6 to 11. These students were selected by lottery to be assigned to one of the two treatments or to control: 3,427 students were assigned to the “basic” treatment, 3,424 to the “savings” treatment, and the remaining 4,056 were assigned to control. In the basic treatment, each student received a bimonthly payment of roughly US\$15 so long as the student attended school at least 80% of days that month. In the savings treatment, each student received a bimonthly payment of roughly US\$10 so long as they met the attendance threshold. The remaining third was held in a bank account and paid to the students’ families when it was time to reenroll for the subsequent year. Barrera-Osorio et al. use the following covariates. For each student, they use age, age squared, gender, grade, years behind (or ahead) relative to their grade, and indicator for whether the student is over age for their grade. They also record the marital status, age, and years of education for the head of household, as well as several household characteristics: Whether or not the residence is rented or owned, income, total number of people, number of children, and an indicator for single parent household. Finally, they include household values for indices that relate to access to utilities, possession of durable goods, the physical infrastructure of the house, and poverty.

In their experiment, Barrera-Osorio et al. collected reenrollment status from administrative records. However, they were unable to obtain reenrollment status for approximately 10% of the observations. In our analysis, we consider both reenrollment status itself and whether the reenrollment status is missing as outcome variables. For each outcome variable, we estimate the average treatment effect for the basic treatment compared to the savings treatment, the basic treatment compared to control, and the savings treatment compared to control. We use the same covariates and restrict our analysis to students in Grades 6–10 as in Barrera-Osorio et al. (2011). We compare the standard errors using LOOP (with random forests), the simple difference estimator, OLS, and cross estimation (with random forests) in Table 3. See Online Appendix F.2 for the full results, including additional methods (LOOP with OLS and OLS with interaction terms) and the point estimates for the treatment effect.

As we can see, OLS, cross estimation, and LOOP provide improvement over the simple difference estimator when missing status is the outcome variable of interest. We can also see that even in this traditional setting (i.e., a large sample size with relatively few covariates), LOOP performs at least as well as OLS. Finally, covariate adjustment does not help when

Table 3. Comparison of Standard Errors With Missing and Reenrollment Status as Outcomes.

Treatments	Method	Missing Status ($\times 10^{-3}$)	Reenrollment Status ($\times 10^{-3}$)
Basic vs. savings	LOOP	6.0	11.8
	Simple difference	7.4	11.8
	OLS	6.3	11.6
	Cross estimation	6.0	11.6
Basic vs. control	LOOP	5.8	11.6
	Simple difference	7.1	11.6
	OLS	6.1	11.5
	Cross estimation	5.7	11.5
Saving vs. control	LOOP	5.7	11.3
	Simple difference	7.0	11.4
	OLS	6.1	11.2
	Cross estimation	5.7	11.2

Note. LOOP = leave-one-out potential outcomes estimator; OLS = ordinary least squares.

reenrollment status is the outcome variable, as the covariates are less predictive of outcome.

Discussion

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. For example, when using poststratification, we must be careful not to use too many covariates, otherwise we partition the data set too finely. Overadjustment can result in poorer performance with linear regression as well: OLS performs poorly when the sample size is small relative to the number of covariates or as the number of noise covariates increases.

The LOOP estimator 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 T_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. It is possible to harm precision in certain cases: We might have a small number of observations and an overly flexible imputation method, which could result in overfitting. However, if we were to use a sufficiently regularized imputation method, we would generally expect that the LOOP estimator would perform at least as well as, or at least not much worse than, the simple difference estimator. For example, we might use an ensemble method that includes mean imputation within the ensemble. While we have not explored such an imputation method in this article, we expect that it would likely help guard against overfitting.

In this article, we suggest the use of random forests to impute the potential outcomes, as they are computationally efficient relative to other methods, likely improve performance over a poststratified estimate, and allow for automatic variable selection. Because of the automatic variable selection, we can adjust for covariates without knowing ahead of time which covariates we wish to use, and any postselection inference is still valid. Finally, as with any covariate adjustment method, the LOOP estimator only improves precision over the unadjusted estimator if the covariates are predictive of outcome. However, we see that even when the covariates are not predictive of outcome, the LOOP estimator generally performs as well as the simple difference estimator.

Implementation in R

The LOOP estimator is implemented in R as the “loop.estimator” package (version 1.0.0.0) and is available on GitHub at <https://github.com/wuje/LOOP>.

Acknowledgments

We would like to thank Yotam Shem-Tov, Luke Miratrix, and Ben Hansen for helpful comments and suggestions.


Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This material is based upon work supported by the National Science Foundation under Grant No. 1646108.

ORCID iD

Edward Wu  <https://orcid.org/0000-0001-8647-2567>

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Depending on the statistical model being used, these adjustments may also be viewed as adjusting for conditional bias (i.e., bias due to the realized allocation of treatment and resulting covariate imbalance). However, in the model we will consider, the treatment assignment vector T is the only source of randomness. The experimental units, their covariates, and their potential outcomes are all modeled as fixed. Conditioning on T therefore removes all randomness and fixes the treatment effect estimate. For this reason, although the covariate adjustment method we present may be viewed in spirit as adjusting for conditional bias, our discussion will be in terms of improved precision.
2. Other authors have suggested different ways to ensure valid postselection inferences; for example, Berk, Brown, Buja, Zhang, and Zhao (2013b) introduce a method for valid postselection confidence intervals and Lee, Sun, Sun, and Taylor (2016) propose a general framework for valid inference after model selection. For a further discussion on data snooping when analyzing experimental data, see Mutz, Pemantle, and Pham (2018).

References

- Aronow, P. M., Green, D. P., & Lee, D. K. (2014). Sharp bounds on the variance in randomized experiments. *The Annals of Statistics*, 42, 850–871.
- Aronow, P. M., & Middleton, J. A. (2013). A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference*, 1, 135–154.
- Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113, 7353–7360.
- Barrera-Osorio, F., Bertrand, M., Linden, L. L., & Perez-Calle, F. (2011). Improving the design of conditional transfer programs: Evidence from a randomized education experiment in Colombia. *American Economic Journal: Applied Economics*, 3, 167–195.
- Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., . . . Stroup, D. F. (1996). Improving the quality of reporting of randomized controlled trials: The CONSORT statement. *The Journal of the American Medical Association*, 276, 637–639.

- Berk, R., Pitkin, E., Brown, L., Buja, A., George, E., & Zhao, L. (2013a). Covariance adjustments for the analysis of randomized field experiments. *Evaluation Review*, 37, 170–196.
- Berk, R., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2013b). Valid post-selection inference. *The Annals of Statistics*, 41, 802–837.
- Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13, 1063–1095.
- Bloniarz, A., Liu, H., Zhang, C., Sekhon, J. S., & Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113, 7383–7390.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21, C1–C68.
- Efron, B., & Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics*, 9, 586–596.
- Freedman, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40, 180–193.
- Holt, D., & Smith, T. M. F. (1979). Post stratification. *Journal of the Royal Statistical Society, Series A*, 142, 33–46.
- Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663–685.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, Chap. 8). New York, NY: Springer.
- Koch, G. G., Amara, I. A., Davis, G. W., & Gillings, D. B. (1982). A review of some statistical methods for covariance analysis of categorical data. *Biometrics*, 38, 563–595.
- Koch, G. G., Tangen, C. M., Jung, J. W., & Amara, I. A. (1998). Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statistics in Medicine*, 17, 1863–1892.
- Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44, 907–927.
- Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7, 295–318.
- Lu, J. (2016). Covariate adjustment in randomization-based causal inference for 2^k factorial designs. *Statistics & Probability Letters*, 119, 11–20.

- Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in medicine*, 23, 2937–2960.
- Miratrix, L. W., Sekhon, J. S., & Yu, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society, Series B*, 75, 369–396.
- Moore, K. L., & van der Laan, M. J. (2009). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in Medicine*, 28, 39–64.
- Mutz, D. C., Pemantle, R., & Pham, P. (2018). The perils of balance testing in experimental design: Messy analyses of clean data. *The American Statistician*, 1–11. Retrieved from <https://www.tandfonline.com/doi/abs/10.1080/00031305.2017.1322143>
- Nie, X., & Wager, S. (2017). Learning objectives for treatment effect estimation. *arXiv preprint arXiv:1712.04912*
- Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*, 1999, 6–10.
- Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846–866.
- Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17, 286–327.
- Rothe, C. (2018). Flexible covariate adjustments in randomized experiments. *Working Paper*. Retrieved from http://www.christophrothe.net/papers/fca_apr2018.pdf
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701.
- Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Rejoinder. *Journal of the American Statistical Association*, 94, 1135–1146.
- Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomized trials. *Annals of Internal Medicine*, 152, 726–732.
- Spiess, J. (2018). Optimal estimation when researcher and social preferences are misaligned. Tech. rep. *Job Market Paper*. Retrieved from <https://scholar.harvard.edu/files/spiess/files/alignedestimation.pdf>
- Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. P. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5, 465–472.

- Steingrimsdottir, J. A., Hanley, D. F., & Rosenblum, M. (2017). Improving precision by adjusting for prognostic baseline variables in randomized trials with binary outcomes, without regression model assumptions. *Contemporary Clinical Trials*, 54, 18–24.
- Tsiatis, A. A., Davidian, M., Zhang, M., & Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*, 27, 4658–4677.
- Wager, S., Du, W., Taylor, J., & Tibshirani, R. J. (2016). High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113, 12673–12678.
- Williams, W. (1961). Generating unbiased ratio and regression estimators. *Biometrics*, 17, 267–274.

Author Biographies

Edward Wu is a graduate student in the Department of Statistics at the University of Michigan.

Johann A. Gagnon-Bartsch is an assistant professor in the Department of Statistics at the University of Michigan.