Rebar+LOOP=Awesome

December 14, 2018

1 Methodological Background

1.1 Causal Inference from Experiments

Consider a randomized experiment to estimate the average effect of a binary treatment T on an outcome Y. Following Neyman [1923] and Rubin [1974], for subject $i = 1, \dots, N$, let potential outcomes y_{Ti} and y_{Ci} represent the outcome value Y_i that i would have exhibited if he or she had (perhaps counterfactually) been assigned to treatment, $T_i = 1$ or control, $T_i = 0$. Then define the treatment effect for i as $\tau_i = y_{Ti} - y_{Ci}$; our goal will be to estimate the average treatment effect (ATE), $\bar{\tau} \equiv \sum_i \tau_i / N$.

If both y_{Ci} and y_{Ti} were known for each subject i, statistical modeling would be unnecessary—researchers could calculate $\bar{\tau}$ exactly, without error, by simply averaging observed τ. In practice, we never observe both y_{Ci} and y_{Ti}. Instead, we rely on the experimental setup to estimate and infer causation. Since the treatment group is a random sample of the N participants, survey sampling literature provides design-based unbiased estimators of the mean of y_T based on observed Y in the treatment group and the known distribution of T. These estimators, and their associated inference, depend only on the experimental design, and not on modeling assumptions. Likewise, the survey sampling literature suggests analogous unbiased, design-based estimators for the mean of y_C based on observed Y values in the control group, which is itself a random sample. The survey sample structure of randomized experiments allows us to infer counterfactual potential outcomes (at least on average) and estimate $\bar{\tau}$ as if τ_i were available for each i, albeit with sampling error.

We will use this framework to analyze the 22 TestBed experiments. These are examples of "Bernoulli experiments," in which each T_i is an independent Bernoulli trial: $Pr(T_i = 1) \equiv p_i$, with $0 < p_i < 1$, and $T_i \perp\!\!\!\perp T_j$ if $i \neq j$. In the TestBed experiments, $p_i = 1/2$ for all i. Observed outcomes are a function of treatment assignment and potential outcomes:

$$Y_i = T_i y_{Ti} + (1 - T_i) y_{Ci} = y_{Ci} + \tau_i T_i.$$

In this model, Y_i is only random due to its dependence on T_i; Y_i has a discrete distribution, with $Pr(Y_i = y_{Ci}) = 1 - p_i$, $Pr(Y_i = y_{Ti}) = p_i$, and $Pr(Y_i = y') = 0$ for any $y' \notin \{y_{Ci}, y_{Ti}\}$. Since either y_{Ti} or y_{Ci} is unobserved, Y_i's distribution is never known. Along the same lines, let $M_i = T_i y_{Ci} + (1 - T_i) y_{Ti}$, i's unobserved counterfactual outcome—when i is treated, $M_i = y_{Ci}$ and when i is in the control condition $M_i = y_{Ti}$. Then i's treatment effect may be expressed as $\tau_i = Y_i - M_i$ if i is in the treatment group or $\tau_i = M_i - Y_i$ if i is in the control group. Although M_i is, by definition, unobserved, it plays a central role in causal inference, as does its expectation,

$$m_i \equiv p_i y_{Ci} + (1 - p_i) y_{Ti}$$

which will play a prominent role in the method we are proposing.

Under this model, estimation and inference about $\bar{\tau}$ is based on the observed values of Y and T. Let

$$U_i = \begin{cases} \frac{1}{p_i} & T_i = 1 \\ -\frac{1}{1-p_i} & T_i = 0 \end{cases}$$

be subject i's signed inverse probability weights. Note that $\mathbb{E}U_i = 0$, and $\mathbb{E}U_i Y_i = \tau_i$. (To see this, note that when $T = 1$, with probability p_i, $Y_i = y_{Ti}$ and $U_i Y_i = y_{Ti}/p_i$; when $T = 0$, with probability $1 - p_i$, $U_i Y_i = -yci/(1 - p_i)$). Then $U_i Y_i$ may be thought of as an unbiased estimate of τ_i, and $\hat{\tau}^{IPW} = \sum_i U_i Y_i / N$ is an unbiased estimate of $\bar{\tau}$. In fact, $\hat{\tau}^{IPW}$ is identical to the "Horvitz-Thompson" estimator of, e.g., Aronow and Middleton [2013]

$$\hat{\tau}^{IPW} = \frac{1}{N} \sum_{i \in \mathcal{T}} \frac{Y_i}{p_i} - \frac{1}{N} \sum_{i \in \mathcal{C}} \frac{Y_i}{1 - p_i}$$

where $\mathcal{C} = \{i | T_i = 0\}$ is the control group and $\mathcal{T} = \{i | T_i = 1\}$ is the treatment group. This, in turn, is the difference between the Horvitz-Thomson estimates of $\bar{y_T}$ and $\bar{y_C}$ [Horvitz and Thompson, 1952].

The sampling variance of $\hat{\tau}^{IPW}$ proceeds from the same principals: the variance of $U_i Y_i$ is

$$V(U_i Y_i) = \left(y_{Ti} \sqrt{\frac{1 - p_i}{p_i}} + y_{Ci} \sqrt{\frac{p_i}{1 - p_i}} \right)^2 = \frac{m_i^2}{p_i(1 - p_i)} \tag{1}$$

and, since subjects' treatment assignments are mutually independent, $V(\hat{\tau}^{IPW}) = 1/N^2 \sum_i m_i^2 / \{p_i(1 - p_i)\}$. Since y_{Ti} and y_{Ci} are never simultaneously observed, $V(\hat{\tau}^{IPW})$ is not identified; however, it may be bounded in expectation, as $\hat{V}(\hat{\tau}^{IPW}) = \sum_i U_i^2 Y_i^2 / N^2$: $\mathbb{E}\hat{V}(\hat{\tau}^{IPW}) \leq V(\hat{\tau}^{IPW})$. (See Aronow and Middleton 2013 for equivalent expressions for more general experimental designs.)

Classical survey sampling theory implies that $\hat{\tau}^{IPW}$ is asymptotically normal, with asymptotic variance of at most $\hat{V}(\hat{\tau}^{IPW})$, so Wald-type confidence intervals of the form $\hat{\tau}^{IPW} \pm z_{\alpha/2} \hat{V}(\hat{\tau}^{IPW})^{1/2}$ achieve at least nominal coverage in large samples. These guarantees hold regardless of the distribution of $\{y_C, y_T\}$—they depend only on the experimental design.

1.2 Design-Based Covariate Adjustment

The reason for error in estimating $\hat{\tau}$, is our inability to observe counterfactual potential outcomes M. As we've seen, randomized trials, coupled with design-based estimators like

$\hat{\tau}^{IPW}$, use comparison groups and survey sampling theory to fill in this missing information. Baseline covariates—a vector $\boldsymbol{x}_i$ of data for subject i gathered prior to treatment randomization—may provide an alternative strategy. To see how, say a researcher had constructed algorithms $\hat{y}_C(\boldsymbol{x}_i)$ and $\hat{y}_T(\boldsymbol{x}_i)$ designed to predict y_C and y_T, respectively, from $\boldsymbol{x}$. Then, if $\hat{M}_i$ is an estimate of i's missing counterfactual, either $\hat{y}_C(\boldsymbol{x}_i)$ or $\hat{y}_T(\boldsymbol{x}_i)$, then $Y_i - \hat{M}_i$ (if $T_i = 1$) or $\hat{M}_i - Y_i$ (if $T_i = 0$) may be considered estimates for τ_i. In general, the bias of algorithms such as $\hat{y}_C(\cdot)$ and $\hat{y}_T(\cdot)$, will be unknown, so these effect estimates may be inadvisable. On the other hand, imperfect or potentially biased predictions of potential outcomes can, *when combined with randomization*, yield substantial benefits.

The approach we will take to combining covariate adjustment with randomization follows Wu and Gagnon-Bartsch [2017], as will its presentation here. It has antecedents in Rosenbaum [2002], Aronow and Middleton [2013], Wager et al. [2016], and Patikorn et al. [2017]. In a Bernoulli experiment, note that

$$\begin{aligned}
&U_i(Y_i - m_i)\\
&\qquad = \begin{cases} \frac{1}{p_i}(y_{Ti} - p_i y_{Ci} - (1-p_i)y_{Ti}) & T_i = 1\\ -\frac{1}{1-p_i}(y_{Ci} - p_i y_{Ci} - (1-p_i)y_{Ti}) & T_i = 0 \end{cases}\\
&\qquad = \begin{cases} \frac{p_i(y_{Ti}-y_{Ci})}{p_i} & T_i = 1\\ \frac{(1-p_i)(y_{Ti}-y_{Ci})}{1-p_i} & T_i = 0 \end{cases}\\
&\qquad = \tau_i
\end{aligned}$$

. This suggests using predictions $\hat{y}_C(\boldsymbol{x}_i)$ and $\hat{y}_T(\boldsymbol{x}_i)$, to estimate m_i as $\hat{m}_i$, and estimating τ_i as

$$\hat{\tau}_i^m \equiv U_i(Y_i - \hat{m}_i)$$

It turns out that $\hat{\tau}_i^m$ is unbiased if prediction algorithms $\hat{y}_C(\cdot)$ and $\hat{y}_T(\cdot)$ are constructed in such a way that

$$\{\hat{y}_C(\boldsymbol{x}_i), \hat{y}_T(\boldsymbol{x}_i)\} \perp\!\!\!\perp T_i. \tag{2}$$

Since $\boldsymbol{x}_i \perp\!\!\!\perp T_i$ by design, (2) is tantamount to requiring that T_i, and variables such as Y_i that depend on it, play no role in constructing prediction algorithms $\hat{y}_C(\cdot)$ and $\hat{y}_T(\cdot)$.

Under (2), $\hat{\tau}_i^m$ is indeed unbiased:

$$\mathbb{E}\hat{\tau}_i^m = \mathbb{E}U_iY_i + \mathbb{E}U_i\hat{m}_i = \mathbb{E}U_iY_i + \mathbb{E}U_i\mathbb{E}\hat{m}_i = \mathbb{E}U_iY_i = \tau_i \tag{3}$$

where we use the facts that $\mathbb{E}U_i = 0$ and $\mathbb{E}U_iY_i = \tau_i$. Finally, define ATE estimate

$$\hat{\tau}^m = \frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_i^m = \frac{1}{N}\sum_{i=1}^{N}\frac{T_i(Y_i - \hat{m}_i)}{p_i} - \frac{(1-T_i)(Y_i - \hat{m}_i)}{1-p_i} \tag{4}$$

The unbiasedness of $\hat{\tau}^m$ for $\bar{\tau}$ follows from the unbiasedness of each of its summands, $\hat{\tau}_i^m$ for τ_i.

Crucially, this unbiasedness holds even if predictions $\hat{y}_C(\boldsymbol{x}_i)$ and $\hat{y}_T(\boldsymbol{x}_i)$ are biased; prediction algorithms $\hat{y}_C(\cdot)$ and $\hat{y}_T(\cdot)$ need not be unbiased, consistent, or correct in any sense.

As long as $\hat{y}_C(\boldsymbol{x}_i)$ and $\hat{y}_T(\boldsymbol{x}_i)$ are constructed to be independent of T_i, $\hat{\tau}_i^m$ will be unbiased. The same cannot be said for regression-based covariance adjustment, the common technique of regressing Y on T and $\boldsymbol{x}$ [Freedman, 2008].

The goal of the covariate adjustment in $\hat{\tau}_i^m$ is to estimate average effects with greater precision; its success in this regard depends on the predictive accuracy of $\hat{y}_C(\boldsymbol{x}_i)$ and $\hat{y}_T(\boldsymbol{x}_i)$. Wu and Gagnon-Bartsch [2017] show that

$$V(\hat{\tau}_i^m|\hat{m}_i) = \frac{(\hat{m}_i - m_i)^2}{p_i(1-p_i)} \tag{5}$$

Accurate predictions of y_{Ci} and y_{Ti}, and hence of $\hat{m}_i$, yield precise estimation of τ_i. On the other hand, inaccurate predictions (such that $(\hat{m}_i - m_i)^2 > m_i^2$) will decrease precision—though, again, without causing bias. The sampling variance of $\hat{\tau}^m$ depends on the dependence structure of $\hat{m}$, and will be discussed in the following two sections.

Under this framework, successful covariate adjustment requires predictions $\hat{y}_C(\boldsymbol{x}_i)$ and $\hat{y}_T(\boldsymbol{x}_i)$ that are accurate and independent of T. To satisfy the independence condition, i's outcome Y_i, which is a function of T, cannot play a role in the construction of the algorithms $\hat{y}_C(\cdot)$ and $\hat{y}_T(\cdot)$; they must be fit using other data. Recent literature proposes two solutions to this problem. One approach [Sales et al., 2018] suggests estimating prediction algorithm $\hat{y}_C(\cdot)$ for all participants in the experiment using an entirely separate dataset: covariate and outcome data from subjects that were not part of the randomized experiment. A second approach [Wu and Gagnon-Bartsch, 2017] fits a separate algorithm for each experimental participant i, using data from experimental units other than i. The following two subsections will review these two approaches in some depth. The remainder of the paper will discuss their combination.

1.3 Auxiliary Data: the Remnant from an Experiment

Modern field trials are often conducted within a very data-rich context, in which high-dimensional and rich covariate data is automatically, or already, collected for all experiment participants. For instance, in the TestBed experiments, system administrators have access to log data for every problem and skill builder each participating student worked before the onset of the experiment. In other contexts, such as healthcare or education, rich administrative data is available. In fact, these covariates are available for a much wider population than just the experimental participants—in the TestBed case, there is log data for all ASSISTments users. In healthcare or education examples, administrative data is available for every student or patient in the system, not just for those who were randomized to a treatment or control condition. Often, as in the TestBed case, the outcome variable Y is also drawn from administrative or log data. We refer to subjects within the same data system in which the experiment took place—i.e. for whom covariate and outcome data are available—but who were not part of the experiment, as the "remnant" from the experiment. The remnant from a TestBed experiment consists of all ASSISTments users for whom log data is available but who did not participate in the experiment.

Clearly, pooling data from the remnant with data from the experiment undermines the benefits and justification for randomization. On the other hand, Sales et al. [2018] argues that data from the remnant can play a role in covariate adjustment. When an RCT contrasts an experimental condition in $\mathcal{T}$ with business as usual in $\mathcal{C}$, then only the control condition will be present in the remnant. Then, an analyst may fit a model $\hat{y}_C^{rem}(\cdot)$ to data from the remnant, and use the fitted model, in conjunction with experimental participants' own covariates $\boldsymbol{x}$, to predict their control potential outcomes as $\tilde{x}_i \equiv \hat{y}_C^{rem}(bmx_i)$. Finally, estimate m_i as $\hat{m}_i^{rebar} \equiv \tilde{x}_i$, and use $\hat{m}_i^{rebar}$ to construct effect estimators $\hat{\tau}_i^m$ and $\hat{\tau}^m$. Sales et al. [2018] calls this method "remnant-based residualization," or "rebar."

Since the model $\hat{y}_C^{rem}(\cdot)$ is fit using data from a separate sample from the experiment, and $\boldsymbol{x} \perp\!\!\!\perp T$ by design, predictions $\tilde{x}_i$ satisfy the independence criterion (2). In fact, $\tilde{x}_i$ may be treated formally just like any other covariate—a fact which is reflected in its notation. Furthermore, $\hat{y}_C^{rem}(\cdot)$ may be fit and assessed in any way, as long as only remnant data is used. This process can be iterative, so that an analyst may fit a candidate model, assess its performance (perhaps with $k-$fold cross-validation), modify the model, and repeat until suitable performance is achieved. This follows from the fact that inference proceeds from the randomization of T, and models fit in the remnant are invariant to T. Any realization of the assignment vector $\boldsymbol{T}$ would have given rise to precisely the same predictions $\tilde{x}_i$. The frequent problem of post-selection inference, which is exacerbated when the dimension of $\boldsymbol{x}$ is large, does not apply here.

Sales et al. [2018] used rebar to analyze the 22 TestBed experiments, and in 16 of the experiments, rebar reduced standard errors by 25-45% relative to estimates without covariate adjustment. However, in three experiments, the rebar estimates had higher standad errors than their unadjusted counterparts (in the remaining three experiments, the improvement was moderate). In general, when $\tilde{x}_i$ is a poor prediction of y_{Ci}, so that $(m_i - \hat{m}_i^{rebar})^2 > m_i^2$, covariance adjustment increases sampling variance. This will be the case if a predictive model fit in the remnant extrapolates poorly to the experimental sample—for instance, if the distribution of $\boldsymbol{x}$, or the distribution of y_C conditional on $\boldsymbol{x}$, differs subsantially between the two samples. To make matters worse, the performance of $\hat{y}_C^{rem}(\cdot)$ in the experimental set—where it counts—may not be checked directly. Once a researcher uses observed experimental outcomes Y to select $\hat{y}_C^{rem}(\cdot)$, the resulting predictions $\tilde{x}_i$ will no longer be independent of T, violating (2).

An additional problem with rebar is that, since typically only the control condition is present in the remnant, $\hat{y}_C^{rem}(\cdot)$ is used to predict both potential outcomes as $\tilde{x}_i$. This may further increase squared prediction error $(m_i - \hat{m}_i^{rebar})^2$. Of course, a preliminary estimate of $\bar{\tau}$ is available from the experimental data, but as before, incorporating experimental outcomes into $\hat{m}$ induces a dependence between $\hat{m}$ and T.

The remnant from an experiment is often much larger than the experimental sample, and may provide fertile ground for predicting potential outcomes, especially in the presence of rich high-dimensional covariates. However, absent methods to use experimental data to assess predictive accuracy and account for possible treatment effects, covariance adjustment using the remnant is risky.

1.4 LOOP

An additional risk of covariate adjustment, even when data from the remnant is not used, is the possibility of overfitting. This is particularly a concern when there is a large number of covariates with little predictive power. Overfitting may result in adjustments that harm rather than improve precision [Freedman, 2008, **?**]. This is a concern of Wu and Gagnon-Bartsch [2017], and their LOOP estimator allows for automatic variable selection to avoid overfitting. Importantly, Wu and Gagnon-Bartsch [2017] argue that LOOP will typically not harm precision and perform no worse than the simple difference estimator.

The LOOP estimator is a leave-one-out method that proceeds as follows. For each i, we first drop observation i, and then use the remaining $N-1$ observations to construct prediction models for the control and treatment potential outcomes, denoted $\hat{y}_C^{(-i)}(\mathbf{x})$ and $\hat{y}_T^{(-i)}(\mathbf{x})$, respectively. These models may be fit by any method, for example linear regression or random forests. In particular, methods that allow for automatic variable selection, or other forms of dimensionality reduction or regularization to prevent overfitting may be used.

Next, following equation (**??**), we set

$$\hat{m}_i = p\hat{y}_C^{(-i)}(\mathbf{x}_i) + (1-p)\hat{y}_T^{(-i)}(\mathbf{x}_i) \tag{6}$$

and the LOOP estimator is then given by $\hat{\tau}^m$ in (4). Note that here $\hat{m}_i \perp\!\!\!\perp T_i$ due to the fact that $\hat{m}_i$ is computed without observation i. It follows that $\hat{\tau}^m$ is unbiased.

Wu and Gagnon-Bartsch [2017] provide an estimate for the variance of the LOOP estimator. Let

$$\hat{E}_C = \frac{1}{n}\sum_{i\in\mathcal{C}}(\hat{y}_{Ci} - y_{Ci})^2 \tag{7}$$

and define $\hat{E}_T$ similarly. Note $\hat{E}_C$ and $\hat{E}_T$ are leave-one-out cross validation mean squared errors. The estimated variance is then given by

$$\widehat{\mathrm{Var}}(\hat{\tau}) = \frac{1}{N}\left[\frac{p}{1-p}\hat{E}_C + \frac{1-p}{p}\hat{E}_T + 2\sqrt{\hat{E}_C\hat{E}_T}\right]. \tag{8}$$

Wu and Gagnon-Bartsch [2017] note that (8) will typically be somewhat conservative. This is due to the fact that $\mathrm{Var}(\hat{\tau})$ is unidentifiable, which itself derives from the fact that we only ever observe one potential outcome for each unit, and thus the correlation of the potential outcomes cannot be estimated. Instead, a bound must be used. This difficulty is not unique to the LOOP estimator; similar comments apply to the simple difference estimator [Aronow et al., 2014]. Note also that

$$\begin{aligned}\frac{1}{N}\left[\frac{p}{1-p}\hat{E}_C + \frac{1-p}{p}\hat{E}_T + 2\sqrt{\hat{E}_C\hat{E}_T}\right] &\leq \frac{\hat{E}_C}{N(1-p)} + \frac{\hat{E}_T}{Np} && (9)\\ &\approx \frac{\hat{E}_C}{N-n} + \frac{\hat{E}_T}{n} && (10)\end{aligned}$$

which is similar in form to the variance estimate typically used in a two-sample t-test, namely $\frac{s_C^2}{N-n} + \frac{s_T^2}{n}$, where s_C^2 and s_T^2 denote the control group and treatment group sample variances. In (10), s_C^2 and s_T^2 are replaced by $\hat{E}_C$ and $\hat{E}_T$. In other words, the variances are replaced by the estimated mean squared errors of the imputations.

A special case of the LOOP estimator occurs when the potential outcomes are imputed by simply taking the mean of the observed outcomes (after dropping observation i). That is, we set

$$\hat{y}_C^{(-i)}(\mathbf{x}) = \bar{y}_{Ci}^{(-i)} \equiv \frac{1}{|\mathcal{C} \setminus i|} \sum_{j \in \mathcal{C} \setminus i} y_{Cj} \tag{11}$$

and similarly for $\hat{y}_T^{(-i)}(\mathbf{x})$. Note that in this case the covariates are ignored. It can be shown that when the potential outcomes are mean-imputed in this manner the LOOP estimator $\hat{\tau}^m$ is exactly equal to the simple difference estimator

$$\hat{\tau}^{SD} = \frac{1}{n} \sum_{i \in \mathcal{T}} Y_i - \frac{1}{N-n} \sum_{i \in \mathcal{C}} Y_i \tag{12}$$

[Wu and Gagnon-Bartsch, 2017]. Moreover, $\hat{E}_C = \frac{N-n}{N-n-1} s_C^2$ and $\hat{E}_T = \frac{n}{n-1} s_C^2$ and thus the variance estimate given by (10) is nearly identical to the ordinary t-test variance estimate.

In short, when using mean imputation for the potential outcomes, LOOP essentially simplifies to an ordinary t-test. The effect estimate is identical, and the variance estimate is nearly identical. This is highly reassuring. Any imputation strategy that improves upon mean-imputation in terms of mean squared error will reduce the variance of the LOOP estimator relative to the simple difference estimator. Most modern machine learning methods employ some form of regularization to guard against overfitting, and thus typically perform no worse, or at least not substantially worse, than mean-imputation. Thus in practice there is relatively little risk that LOOP will hurt precision.

2 Method

Our goal is to construct a method that, like rebar, is able to exploit data in the remnant but, like LOOP, poses little risk of harming precision.

Recall that $\tilde{x}_i$ are the imputed control potential outcomes for the participants in the experiment, using a predictive model $\hat{y}_C^{rem}(\cdot)$ fit in the remnant, and that $\tilde{x}_i$ may also be thought of simply as an additional covariate, along with those in $\mathbf{x}_i$. we may therefore include $\tilde{x}_i$ as a covariate in LOOP. We now discuss three options for doing so.

Strategy 1: Define

$$\tilde{\mathbf{x}}_i \equiv (\tilde{x}_i, x_{i1}, x_{i2}, ..., x_{ip}) \tag{13}$$

or in other words, $\tilde{\mathbf{x}}_i$ is $\mathbf{x}_i$ augmented with $\tilde{x}_i$.

The most straight-forward option would be to run LOOP on the experimental data, but using the augmented set of covariates $\tilde{\mathbf{x}}$ instead of $\mathbf{x}$. The hope is that by including $\tilde{x}_i$ we can implicitly exploit information in the remnant in much the same way that rebar does. Moreover, by using LOOP, we also reduce the risk of accidentally hurting precision.

A few comments: (1) For this option we would use random forests as the imputation strategy within LOOP, as suggested by Wu and Gagnon-Bartsch [2017]. We refer to LOOP with random forests as LOOP-RF. (2) $\tilde{x}_i$ is a function of the other covariates and thus, in at least some sense, does not contain any additional information. However, the function $\hat{y}_C^{rem}(\mathbf{x})$ is fitted on the remnant, which may be much larger than the experimental sample, and thus $\hat{y}_C^{rem}(\mathbf{x})$ may be a more accurate imputation function than what we would be able to obtain using the experimental data alone. In this sense, $\tilde{x}_i$ does contain additional information, which LOOP-RF can exploit by heavily weighting $\tilde{x}_i$ over the other covariates. (3) If the imputations $\tilde{x}_i$ are poor, then LOOP-RF may simply downweight or effectively ignore them. In particular, poor imputations from the remnant should not harm precision. (4) Biased imputations can still be helpful. Importantly, because the $\tilde{x}_i$ are used as a covariate within LOOP-RF, they do not necessarily need to accurately impute the potential outcomes in the experimental sample; rather, it suffices that they are merely predictive of the potential outcomes. If the experimental sample is systematically different from the remnant, e.g., the potential outcomes in the experimental sample are on average higher than those in the remnant, the $\tilde{x}_i$ will still be useful as long as they are correlated with the experimental potential outcomes. (5) We have implicitly assumed here that all units in the remnant are untreated, and the $\tilde{x}_i$ are imputed control potential outcomes. In light of the previous comment (4), the $\tilde{x}_i$ may still be used as a covariate within LOOP-RF for predicting treatment potential outcomes, even if the y_T differ from the y_C due to a treatment effect, as long as $\tilde{x}$ is predictive of y_T. (6) If the $\tilde{x}_i$ are highly accurate, using them as a covariate within a nonparametric method like a random forest may be statistically inefficient. The random forest may effectively just add noise or bias. It is this concern that motivates the next strategy.

Strategy 2: A second option would be to run LOOP on the experimental data, but using only $\tilde{x}_i$ as a predictor, and using linear regression instead of a random forest. That is, for each i

$$\hat{y}_{Ci} = \hat{\beta}_{0C}^{(-i)} + \hat{\beta}_{1C}^{(-i)} \tilde{x}_i \tag{14}$$

where $\hat{\beta}_{0C}^{(-i)}$ and $\hat{\beta}_{1C}^{(-i)}$ are the coefficients from a regression of the observed y_C on $\tilde{x}$, omitting observation i. The expression for $\hat{y}_{Ti}$ would be analogous.

Strategy 2 may be preferable to strategy 1 when the $\tilde{x}_i$ are highly accurate imputations of y_C. In that case, $\hat{\beta}_{1C}^{(-i)} \approx 1$ and $\hat{y}_{Ci} \approx \tilde{x}_i$. In other words, the imputations from the remnant "pass through" the LOOP procedure largely unmodified, resulting in a rebar-like adjustment. However, in contrast to rebar, poor imputations $\tilde{x}$ will not necessarily harm precision. Consider the extreme case in which $\tilde{x}$ is pure noise. We would then expect $\hat{\beta}_{1C}^{(-i)} \approx 0$ and $\hat{\beta}_{0C}^{(-i)} \approx \bar{y}_{Ci}^{(-i)}$ so that $\hat{y}_{Ci} \approx \bar{y}_{Ci}^{(-i)}$. That is, we revert approximately to mean-imputation, and the final estimator is therefore approximately equal to the simple difference estimator.

Strategy 3: In practice it may not always be clear whether strategy 1 or 2 will perform better; it depends on the quality of the imputations $\tilde{x}$ as well as the predictive power of the covariates in the experimental sample. Thus, a final option would be to combine strategies 1 and 2 by taking a weighted average. Let $\hat{y}_{Ci}^{(-i,S1)}(\tilde{\mathbf{x}}_{\mathbf{i}})$ and $\hat{y}_{Ci}^{(-i,S2)}(\tilde{x}_i)$ denote imputations from strategies 1 and 2, respectively. We then let

$$\hat{y}_{Ci}^{(-i,S3)}(\tilde{\mathbf{x}}_{\mathbf{i}}) = \alpha_i \hat{y}_{Ci}^{(-i,S1)}(\tilde{\mathbf{x}}_{\mathbf{i}}) + (1-\alpha_i)\hat{y}_{Ci}^{(-i,S2)}(\tilde{x}_i) \tag{15}$$

where α_i is given by

$$\alpha_i = \operatorname*{arg\,min}_{\alpha \in [0,1]} \sum_{j \in \mathcal{C} \setminus i} \left[Y_j - \left(\alpha \hat{y}_{Cj}^{(-i,S1)}(\tilde{\mathbf{x}}_{\mathbf{j}}) + (1-\alpha)\hat{y}_{Cj}^{(-i,S2)}(\tilde{x}_j) \right) \right]^2 \tag{16}$$

References

Peter M Aronow and Joel A Middleton. A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference*, 1(1):135–154, 2013.

Peter M. Aronow, Donald P. Green, and Donald K. K. Lee. Sharp bounds on the variance in randomized experiments. *Ann. Statist.*, 42(3):850–871, 06 2014. doi: 10.1214/13-AOS1200. URL `https://doi.org/10.1214/13-AOS1200`.

David A. Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193, 2008.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260): 663–685, 1952.

J. Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5:463–480, 1923. 1990; transl. by D.M. Dabrowska and T.P. Speed.

Thanaporn Patikorn, Douglas Selent, Neil T Heffernan, Joseph E Beck, and Jian Zou. Using a single model trained across multiple experiments to improve the detection of treatment effects. In *Proceedings of the 10th International Conference of Educational Data Mining*, 2017.

P.R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3), 2002.

D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology; Journal of Educational Psychology*, 66(5):688, 1974.

Adam C Sales, Anthony Botelho, Thanaporn M Patikorn, and Neil T Heffernan. Using big data to sharpen design-based inference in a/b tests. In *Proceedings of the 11th International Conference on Educational Data Mining. International Educational Data Mining Society*, pages 479–486, 2018.

Stefan Wager, Wenfei Du, Jonathan Taylor, and Robert J Tibshirani. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(45):12673–12678, 2016.

Edward Wu and Johann Gagnon-Bartsch. The loop estimator: Adjusting for covariates in randomized experiments. *arXiv preprint arXiv:1708.01229*, 2017.