Leveraging covariate adjustments at scale in online A/B testing

Lorenzo Masoero   masoerl@amazon.com
Amazon.com

Doug Hains   dhains@amazon.com
Amazon.com

James McQueen   jmcq@amazon.com
Amazon.com

Editor:

Abstract

Companies offering web services routinely run randomized online experiments to estimate the "causal impact" associated with the adoption of new features and policies on key performance metrics of interest. These experiments are used to estimate a variety of effects: the increase in click rate due to the repositioning of a banner, the impact on subscription rate as a consequence of a discount or special offer, etc. In these settings, even effects whose sizes are very small can have large downstream impacts. The simple difference in means estimator (Splawa-Neyman et al., 1923/1990) is still the standard estimator of choice for many online A/B testing platforms due to its simplicity. This method, however, can fail to detect small effects, even when the experiment contains thousands or millions of observational units. As a byproduct of these experiments, however, large amounts of additional data (covariates) are collected. In this paper, we discuss benefits, costs and risks of allowing experimenters to leverage more complicated estimators that make use of covariates when estimating causal effects of interest. We adapt a recently proposed general-purpose algorithm for the estimation of causal effects with covariates to the setting of online A/B testing. Through this paradigm, we implement several covariate-adjusted causal estimators. We thoroughly evaluate their performance at scale, highlighting benefits and shortcomings of different methods. We show on real experiments how "covariate-adjusted" estimators can (i) lead to more precise quantification of the causal effects of interest and (ii) fix issues related to imbalance across treatment arms — a practical concern often overlooked in the literature. In turn, (iii) these more precise estimates can reduce experimentation time, cutting cost and helping to streamline decision-making processes, allowing for faster adoption of beneficial interventions.

Keywords: Online A/B testing, Randomized experiments, Covariate adjustments

1 Introduction

The continued growth and product improvement for online companies relies on efficiently finding new opportunities and accurately measuring the impact of decisions on customers. To estimate the causal impact of a change to a product or feature, online companies heavily rely on A/B tests (randomized controlled trials/online randomized experiments). Under

minimal assumption, A/B tests are indeed guaranteed to produce unbiased estimates of the causal impact of the interventions that are being tested (Splawa-Neyman et al., 1923/1990).

In what follows, we will assume that the experimental units of interest are "customers". In practice, different experiments might be tracking different units (sellers, streamers, shopping missions, etc.). A/B tests work by randomly assigning some customers (usually half, the "treatment group") to see the new experience, while the other customers (the "control group") see the old, *status quo* experience. Over a fixed experimental period, different relevant metrics of interest of these customers are measured and recorded. A simple way for experimenters to quantify how the change in the experience impacts customers with an A/B test is to compare the average value of a metric of interest or "key performance indicator" [KPI] across customers in the treatment group to the average value of the same metric in the control group. This "difference in means" [DIM] approach produces a simple estimate of the average causal effect of the change in the experience on the KPI in question.

The simplicity and low marginal cost of running A/B tests with millions of customers has led them to be ubiquitous in the industry. A/B testing is used to evaluate front-end and back-end changes to search engines (Google, Bing, Yandex), online retailers (Amazon, eBay, Etsy), streaming media services (Netflix, Twitch, YouTube), social networks (Facebook, LinkedIn, Twitter), travel services (Lyft, Uber, Airbnb, Booking.com), etc. See Gupta et al. (2019) for a thorough discussion of the role and use of A/B tests in the industry.

Due to the opportunity cost of experimentation time, small treatment effect sizes, and large heterogeneity amongst customers, the difference in means approach can often fail to detect effects of the intervention when they are present, even when the experiment contains thousands or millions of customers. However, large amounts of data often unrelated with the A/B test (the "covariates") are collected before and throughout the experiment about the experimental units. This abundance of data gives experimenters the potential to adopt more complex "covariate-adjusted" methods to form their estimates. In particular, any feature that is independent of the intervention (such as any measures taken prior to the experiment) can be leveraged to estimate the causal effect of interest. Resulting covariate-adjusted estimators can lead to improved, less variable, estimates of the "causal effects" of interest. It has been observed empirically that simple covariate adjusted approaches, such as the popular "CUPED" (Deng et al., 2013), can lead to significant variance reduction relative to the difference in means estimator. Furthermore, covariate adjustment has been used defensively to guard against an unlucky randomization, where the intervention may appear artificially better or worse due to luck (see Tukey (1991)).

The literature related to covariate adjustment methods is continuously growing (see, e.g. Guo et al. (2021); Jin and Ba (2021) for recent contributions). In this paper, we explore benefits and costs of expanding the toolkit of experimenters by using larger sets of covariates and more complex estimators for the estimation of causal effects in the context of online A/B tests. We show in our experiments that covariate-adjusted methods can lead to non-trivial gains in terms of estimation accuracy and variance reduction.

The rest of this paper is organized as follows: we introduce notation for the problem of interest in Section 2. Next, we describe in Section 3 the class of Generalized Oaxaca-Blinder Estimators [GOBEs] — a flexible, general purpose method to produce estimators of the causal effects leveraging any arbitrary number of additional covariates. We discuss the

potential use of these estimators for experimentation in Section 4, and present experimental results in Section 5. We conclude with a discussion and next steps in Section 6.

2 Potential outcomes, randomized experiments and causal effects

The field of causal inference is a collection of theoretically sound tools, methodologies and procedures which can help practitioners answer questions about the impact of interventions they may want to implement. While making rigorous causal claims about interventions is appealing and desirable, this ability comes at the cost of collecting data through carefully designed experiments. In order for causal claims to be valid, experimenters have to make sure that the data is collected in such a way that no bias or flaw is introduced in the analysis. The standard approach to ensure that the data we are collecting will allow us to formulate causal claims, is to perform a *randomized controlled trial* (RCT), or A/B test. In its simplest form, an A/B test is implemented by exposing each experimental unit to either the control (A) or treatment (B) experience *at random*. Randomization is the key technical device which allows experimenters to draw causal conclusions from the experiment.

To make our discussion precise, we here adopt the causal model of potential outcomes (Splawa-Neyman et al., 1923/1990; Rubin, 1977). In a nutshell, we assume that in an experiment in which we observe N units, every individual unit $n \in [N] := \{1, \ldots, N\}$ is exposed to one of $T \geq 2$ different "treatments" or "policies". For example, the N units might be different customers in the experiment. For each *potential* allocation of unit $n \in [N]$ to policy t, we assume that there exists a "potential" outcome $Y_n(t)$. That is, each unit in the experiment is associated with a (latent) vector of potential outcomes, $[Y_n(0), \ldots, Y_n(T-1)]^\top$, of which only one coordinate is observed in an experiment. These outcomes are assumed to be fixed conditionally on the assignment. For simplicity in what follows we will consider the case of $T = 2$ alternative treatments, that we will simply call "control" ($t = 0$) or "treatment" ($t = 1$), even though our discussion naturally extends to any $T > 2$. Given these definitions, we can formally define what we mean by a "causal effect". The most important effect (or estimand) of interest, and the one we will focus on, is the average treatment effect [ATE]. The ATE is the (average) *causal effect* in the population of exposing a unit to the treatment ($t = 1$) instead of the alternative control ($t = 0$). Often, for decision-making, we think of the treatment as an alternative policy to a standard baseline (potentially more expensive or riskier). The ATE quantifies the impact on the outcome of interest of adopting this alternative strategy. Formally,

$$\bar{Y}_k := \sum_{n=1}^{N} \frac{Y_n(k)}{N}, \text{ and ATE} := \bar{Y}_1 - \bar{Y}_0 = \sum_{n=1}^{N} \frac{[Y_n(1) - Y_n(0)]}{N}.$$

The ATE can not be directly computed or observed in practice, because units are either exposed to treatment or control, but never to both. RCTs or A/B tests are used to estimate the ATE. The fundamental mechanism underlying an A/B test is its random assignment mechanism (or triggering logic), which determines the experience to which each unit will be exposed. In the simplest case, each unit n is endowed with a binary random variable with mean $\pi \in (0, 1)$:

$$J_n \sim \text{Bernoulli}(\pi). \tag{1}$$

3

If $J_n = 0$, then unit n is exposed to the control. Otherwise, if $J_n = 1$, the treatment experience is rendered. The ATE is estimated by comparing the observed outcomes for the units in control and treatment. Denote with $\mathcal{I}_t := \{n \in [N] : J_n = t\}$ for the units in group $t \in \{0, 1\}$. The "difference-in-means" [DIM] estimator is simply the difference between the average outcome in each treatment group:

$$\widehat{\text{ATE}}_{\text{DIM}} := \left[\frac{1}{|\mathcal{I}_1|} \sum_{n \in \mathcal{I}_1} Y_n(1) \right] - \left[\frac{1}{|\mathcal{I}_0|} \sum_{n \in \mathcal{I}_0} Y_n(0) \right]. \tag{2}$$

The theoretical properties of this extremely simple estimator are well understood (Splawa-Neyman et al., 1923/1990): it is unbiased, and under mild conditions it obeys a central limit theorem in large samples. See Li and Ding (2017) and the references therein for a thorough overview and discussion.

3 Leveraging covariates: Generalized Oaxaca-Blinder estimators

Often, when collecting data from our experiment, we have access to additional covariates measured at the unit level, hereafter denoted as $\boldsymbol{z}_n := [z_{n,1}, \ldots, z_{n,K}]^\top \in \mathbb{R}^K$, for $n \in [N]$ and some fixed $K \in \mathbb{N}$. For example, when running an experiment on the engagement of customers subscribing to a video streaming service, we might have access to previous measurements of the customer activity, the longevity of the customer's account, whether they have subscribed to for pay-per-view channels, etc. If these covariates are:

(C1) independent of the assignment variable J_n

(C2) correlated with the outcome variable of interest Y_n

they can bee leveraged to form covariate-adjusted estimators. See Imbens and Rubin (2015, Chapter 7) for a detailed discussion on the validity of regression adjustments in randomized experiments.

In what follows, we describe a general recipe to build "covariate-adjusted" estimators. The key intuition underlying this approach is to view the estimation of the ATE as a "missing data" or "imputation" problem. For each treatment t, we can fit a regression model using the observed data within the group $\mathcal{I}_t$, and use the regression to impute the "missing" values of units assigned to the other treatment group(s) — $n \in \mathcal{I}_t^C$. That is, we fit for $t \in \{0, 1\}$ a regression model $Y_n \sim f_t(\boldsymbol{z}_n; \boldsymbol{\theta}_t)$ using covariates and outcomes in the corresponding treatment group, $\mathcal{D}_t := \{(Y_n, \boldsymbol{z}_n)\}_{n \in \mathcal{I}_t}$. Here $\boldsymbol{\theta}_t$ is a finite dimensional parameter that characterizes the regression model (e.g., the slope and intercept of a linear regression model). Given $\mathcal{D}_t$, we estimate $\boldsymbol{\theta}$ by minimizing a loss function $\mathcal{L}$ computed on $\mathcal{D}_t$ and parametrized by $\boldsymbol{\theta}_t$:

$$\hat{\boldsymbol{\theta}}_t \in \arg \min_{\theta} \mathcal{L}\{\mathcal{D}_t; \boldsymbol{\theta}\}. \tag{3}$$

This gives us the imputation operator:

$$\hat{f}_t(Y_n, \boldsymbol{z}_n, J_n; \hat{\boldsymbol{\theta}}_t) = \begin{cases} Y_n & \text{if } J_n = t, \\ f(\boldsymbol{z}_n; \hat{\boldsymbol{\theta}}_t) & \text{otherwise.} \end{cases}$$

This approach induces the large class of "Generalized Oaxaca-Blinder Estimators" [GOBEs, Guo and Basse (2021)] of the type

$$\widehat{\text{ATE}}_{\mathcal{M}} = \frac{1}{N} \sum_{n=1}^{N} \left\{ \hat{Y}_n(1) - \hat{Y}_n(0) \right\}, \tag{4}$$

where $\hat{Y}_n(t) = \hat{f}_t(Y_n, \boldsymbol{z}_n, J_n; \hat{\boldsymbol{\theta}}_t)$ and $\mathcal{M} := \{f_0, f_1\}$ is used to emphasize the dependency of the estimator on the regression functions used. We summarize this procedure in Algorithm 1.

Algorithm 1 Generalized Oaxaca-Blinder Estimators

Require: Data $\{(Y_n, J_n, \boldsymbol{z}_n)\}_{n=1}^{N}$, regression model $\mathcal{M} = \{f_0, f_1\}$, where f_0 is the regression model for T_0 and f_1 for T_1.

For $t \in \{0, 1\}$, and $\mathcal{D}_t := \{(Y_n, \boldsymbol{z}_n)\}_{n \in \mathcal{I}_t}$, fit regression model f_t:

$$\hat{\boldsymbol{\theta}}_t \in \arg\min_{\boldsymbol{\theta}} \mathcal{L}\left\{\mathcal{D}_t; \boldsymbol{\theta}\right\}.$$

For $n = 1, \ldots, N$, impute values

$$\hat{Y}_n(t) = \hat{f}_t(Y_n, \boldsymbol{z}_n, J_n; \hat{\boldsymbol{\theta}}_t) = \begin{cases} Y_n & \text{if } t = J_n \\ \hat{f}_t(Y_n, \boldsymbol{z}_n, J_n; \hat{\boldsymbol{\theta}}_t) & \text{if } t = 1 - J_n. \end{cases}$$

For $t = 0, 1$ estimate the the mean-squared error of the model:

$$\widehat{\text{MSE}}_{\mathcal{M}, t} = \frac{1}{|\mathcal{I}_t| - 1} \sum_{n \in \mathcal{I}_t} \left\{ Y_n - \hat{Y}_n(t) \right\}^2,$$

and

$$\widehat{\text{Var}}_{\mathcal{M}} := \frac{\widehat{\text{MSE}}_{\mathcal{M}, 1}}{|\mathcal{I}_1|} + \frac{\widehat{\text{MSE}}_{\mathcal{M}, 0}}{|\mathcal{I}_0|}.$$

return Estimate and corresponding confidence intervals $\widehat{\text{ATE}}_{\mathcal{M}}$ as per Equation (4) and

$$\widehat{\text{CI}}_{\mathcal{M}}(\alpha) := \widehat{\text{ATE}}_{\mathcal{M}} \pm z_{1 - \frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}_{\mathcal{M}}},$$

where z_α is the $100 \times \alpha\%$ quantile of the cumulative density function [CDF] of the standard normal distribution.

The Gaussian assumption on the confidence intervals returned by Algorithm 1 is asymptotically justified under mild conditions for large classes of models (see Guo and Basse (2021, Theorem 4)).

Difference-in-means as GOBE Notice that the difference-in-mean estimator introduced in Equation (2) can be viewed as a generalized Oaxaca-Blinder estimator. Indeed $\widehat{\text{ATE}}_{\text{DIM}}$ satisfies Equation (4) for the choice $\hat{f}_t(Y_n, \boldsymbol{z}_n, J_n; \hat{\boldsymbol{\theta}}_t) = \frac{1}{|\mathcal{I}_t|} \sum_{n' \in \mathcal{I}_t} Y_{n'}$, where we impute the missing values with the group mean, irrespective of the value of the covariates.

Linear regression as GOBE The standard linear-regression adjusted estimator is a GOBE. This estimator is obtained by fitting via ordinary least squares [OLS] the following regression:

$$Y_n \sim \beta_0 + \beta_1 J_n + \boldsymbol{\gamma}^\top \boldsymbol{z}_n + \boldsymbol{\delta}^\top J_n(\boldsymbol{z}_n - \bar{\boldsymbol{z}}), \tag{5}$$

and using the estimate $\widehat{\text{ATE}}_{\text{LR}} := \hat{\beta}_1$. Here $\bar{\boldsymbol{z}} \in \mathbb{R}^K$ is the average for each of the K components, computed across the N units. Lin (2013) shows that asymptotically $\widehat{\text{ATE}}_{\text{LR}}$ is unbiased and has smaller variance than $\widehat{\text{ATE}}_{\text{DIM}}$. To see $\widehat{\text{ATE}}_{\text{LR}}$ as a "GOBE", notice that it can be equivalently obtained by fitting via OLS in two separate linear regressions: for $n \in \mathcal{I}_t$, $t \in \{0,1\}$, fit $Y_n \sim \boldsymbol{\theta}_t^\top(\boldsymbol{z}_n - \bar{\boldsymbol{z}})$. We recover the GOBE formulation of Equation (4) by letting

$$\hat{f}_t(Y_n, \boldsymbol{z}_n, J_n; \hat{\boldsymbol{\theta}}_t) = \begin{cases} Y_n & \text{if } t = J_n \\ \hat{\boldsymbol{\theta}}_t^\top \boldsymbol{z}_n & \text{if } t \neq J_n. \end{cases} \tag{6}$$

See, e.g. Lin (2013, Lemma 3) for a proof of why Equation (6) and Equation (5) lead to the same estimator.

One can also employ general non-linear regression models to perform adjustments. Guo and Basse (2021) provide conditions under which regression models produce unbiased and asymptotically normal estimates, justifying the Gaussian approximation in Algorithm 1. Building on these results, Cohen and Fogarty (2020) propose a two-step GOBE. First a GOBE is fitted, and then a second GOBE with a linear regression model using imputed values $x_n := \hat{f}_{1-J_n}(Y_n, \boldsymbol{z}_n, J_n; \hat{\boldsymbol{\theta}}_{1-J_n})$ as the only covariate for each outcome Y_n is used to produce the final estimate. This "two-step" GOBE is asymptotically unbiased, normally distributed, and more efficient than the difference-in-means estimator. See Guo et al. (2021); Jin and Ba (2021) for other recent approaches to develop flexible models in online A/B testing.

4 Implementing GOBEs at scale

As already discussed in Section 1, drawing conclusions from online A/B tests can be challenging: experiments often consist of small changes related to details of the user experience. Consequently, associated effect sizes can be very small and hard to detect even in large samples. Despite being small, these can lead to large downstream impacts. Exactly because of this reason, employing models that allow for precise estimates of the causal effects is important: more precise estimates of the effects of the interventions can allow practitioners to detect smaller effect sizes and crucially shorten the experimentation time needed in order to obtain a conclusive answer about the effectiveness of a treatment.

Ultimately, it would be desirable to have an end-to-end automated inference engine which produces, for each experiment, the "best" possible estimate for the causal effect under study, without requiring experimenters to specify which model and covariates should be employed for this task. In practice, assessing which estimator is best is far from being trivial. Indeed, while on the one hand the idea of developing ad-hoc large models with curated covariates for an individual experiment of interest seems appealing for variance reduction, on the other hand large-scale causal inference engines have to rely on estimators that perform well on average across all experiments. That is, the methods used need to be:

- Scalable: companies typically run a very large number of experiments every year, and their computational resources are limited. It is undesirable for practitioners to have to wait for their results due to long analysis run times (e.g., to solve the minimization problem in Algorithm 1).

- Reliable: the team maintaining the infrastructure is often small relative to the customer base it serves. The methods implemented need to rely on algorithmically sound routines that produce stable estimates of the causal effects of interest.

- Interpretable: the results of the experiments are used by practitioners for policy-making. It is therefore imperative that the estimates produced are transparent, easy to interpret, and do not require specialized knowledge.

Because of these reasons, in our experiments presented in Section 5 we only employ linear models, their regularized counterparts (LASSO, ElasticNet, Ridge and principal components regression), and one simple instance of a generalized linear model. Extending our analysis to more complicated models, and assessing their feasibility in a production setting is part of ongoing investigations.

We here describe in detail the regression models we fit to experimental data to benchmark the performances of different GOBE estimators. We have already discussed the difference in means [DIM] and simple linear regression [LR] estimators, and their characterization as GOBEs in Section 3. Ridge regression, LASSO and elastic net are extremely popular "regularized" counterparts of simple linear regression model, in which the weight vector $\boldsymbol{\theta}$ is "shrunk" using a penalty. Formally, given a regularization parameter $\gamma > 0$, we minimize with respect to $\boldsymbol{\theta}$ the loss function

$$\mathcal{L}(\mathcal{D}_t; \boldsymbol{\theta}) := \sum_{n \in \mathcal{I}_t} \left(y_n - \boldsymbol{\theta}^\top \boldsymbol{z}_n \right)^2 + \gamma \|\boldsymbol{\theta}\|_\ell^2,$$

where $\ell = 1$ for LASSO and $\ell = 2$ for Ridge regression (i.e., regularize using the ℓ-1 or ℓ-2 norm). Elastic net regression is obtained by combining the ℓ_1 and ℓ_2 penalties on the regression coefficients. Formally, in this case, we minimize the loss function

$$\mathcal{L}(\mathcal{D}_t; \boldsymbol{\theta}) := \sum_{n \in \mathcal{I}_t} \left(y_n - \boldsymbol{\theta}^\top \boldsymbol{z}_n \right)^2 + \gamma \lambda \sum_{k=1}^{K} |\theta_{t,k}| + \frac{\gamma(1-\lambda)}{2} \sum_{k=1}^{K} \theta_{t,k}^2.$$

Here λ trades off the importance of the ℓ_1 and ℓ_2 penalties. Differently from linear regression and the simple difference in means, these regularized models crucially depend on the tuning of some regularization hyperparameter. Towards the goal of having a streamlined, automated procedure to fit these models, we adopt a standard cross-validation approach. For each model, we repeatedly minimize the objective function across a predetermined number of different values of the regularization parameters. For each of these values, we split the data in 5 random folds, and fit the model 5 times by iteratively leaving out one fold of the data. For each fold, we compute the coefficient of determination on the left-out-data using the fitted coefficient and choose the optimal regularization level by picking the value that achieved the maximum average coefficient of determination across the folds, and re-fit the model using the full dataset. We also consider principal component regression [PCR]
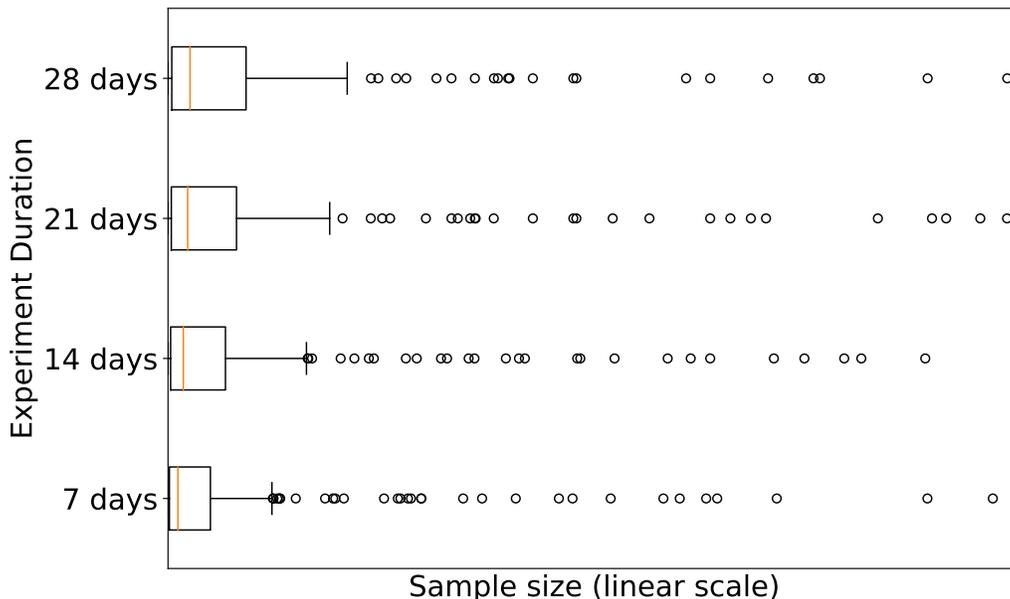
7

Figure 1: Empirical distribution of the sample sizes of the experiments considered in the meta-analysis.

— where we first reduce the dimensionality of the regressors using their projections onto principal components, and then use these as covariates in a linear regression, as well as an instance of a generalized linear model using a Tweedie distribution kernel.

5 Experiments

5.1 Data description

For our experiments, we consider a representative set of $W = 100$ A/B tests. These have been running in production over the course of the last two years, at different times of the year. Each experiment corresponds to a different intervention. For simplicity, in our analysis we only consider one pairwise comparison per experiment (T_1 versus T_0) — even though some experiments might have more than two treatment arms. For each experiment, we run our data analysis pipeline and compute estimates of the causal effects after collecting data for a total time of $D \in \{7, 14, 21, 28\}$ days. For any analysis duration, the sizes of the experiments (total number of customers in the T_1 and T_0 arms) varies considerably (Figure 1).

For all these experiments, we track the same KPI of interest. Since the scale of this KPI varies across experiments, in our illustrations and analysis we focus on the percent ATE (or lift), which is defined as $\text{LIFT}(:) = \text{ATE}/|\bar{Y}_0|$. We use as covariates less than 10 pre-exposure values of customer metrics correlated with the KPI.

5.2 Variance reduction

Intuitively, a more precise estimator (with estimated lower variance) leads to better estimates, and directly translate in faster and better decision making (as we further discuss in Section 5.4). We estimate the performance of model $\mathcal{M}$ in terms of precision by computing their estimated percentage variance reduction $\widehat{\mathrm{VR}}_{\mathcal{M}}$ with respect to the baseline DIM. Recalling that $\widehat{\mathrm{Var}}_{\mathcal{M}}$ is the estimated variance of the ATE under model $\mathcal{M}$ (as per Algorithm 1), we define:

$$\widehat{\mathrm{VR}}_{\mathcal{M}} := 100 \times \left\{ 1 - \frac{\widehat{\mathrm{Var}}_{\mathcal{M}}}{\widehat{\mathrm{Var}}_{\mathrm{DIM}}} \right\}.$$

We plot in Figure 2 the variance reduction as a function of the duration of the analysis across all the experiments in the meta-analysis. Three main findings emerge:

- As expected (see, e.g. Guo and Basse (2021, Theorem 4)), covariate adjusted estimates generally have smaller variances.

- Larger variance reduction is observed in longer analyses, which are characterized by more stable customer behavior.

- The performance observed across different covariate adjusted methods is similar. We analyze in Section 5.5 the performance and computation cost of these methods in relation with the number of (potentially noisy) regressors.

Next, we further try to understand the relationship between analysis duration, sample size and model precision. For a given duration of the analyses (e.g., $D = 7$ days), let $F_{N,D} : \mathbb{N} \to [0,1]$ be the empirical cumulative density function [CDF] of the sample size N of the experiments after D days, and let $F_{N,D}^{-1} : [0,1] \to \mathbb{N}$ be its inverse. E.g., $F_{N,7}^{-1}(0.6)$ is the sample size of the 60%-largest experiment amongst the 7-day analyses. We consider the variance reduction gains within the first quartile (experiments with sample size $N \in [F_{N,D}^{-1}(0), F_{N,D}^{-1}(0.25))$ and in the last quartile (experiments with sample size $N \in [F_{N,D}^{-1}(0.75), F_{N,D}^{-1}(1))$. We observe different behaviors at duration $D = 7$ and $D = 28$. Specifically, for the shorter analysis time ($D = 7$ days) variance reduction is particularly evident in *larger* experiments. However, for $D = 28$ smaller experiments seem to be benefitting the most from covariate adjustments. More broadly, we expect the variance reduction induced by covariate adjustments can vary with the experiment size and duration, and the choice of covariates and model used. We advise practitioners to extensively analyze their data, before the experiment is run, prior to adopting a regression model.

5.3 Robustness to chance imbalance

In an A/B test, treatment arms should be *ex-ante* comparable. That is, by virtue of the randomization, the distribution of the covariates for the units in the treatment and control group should coincide. This is not only a property of a correctly constructed A/B test, but also a fundamental requirement that such an experiment should satisfy to yield valid inferences. Consider an experiment in which the value of a given covariate x is predictive of the outcome y (e.g., units with higher x tend to have higher y). If the triggering logic systematically allocates with higher (or lower) probability units with higher value of x to the
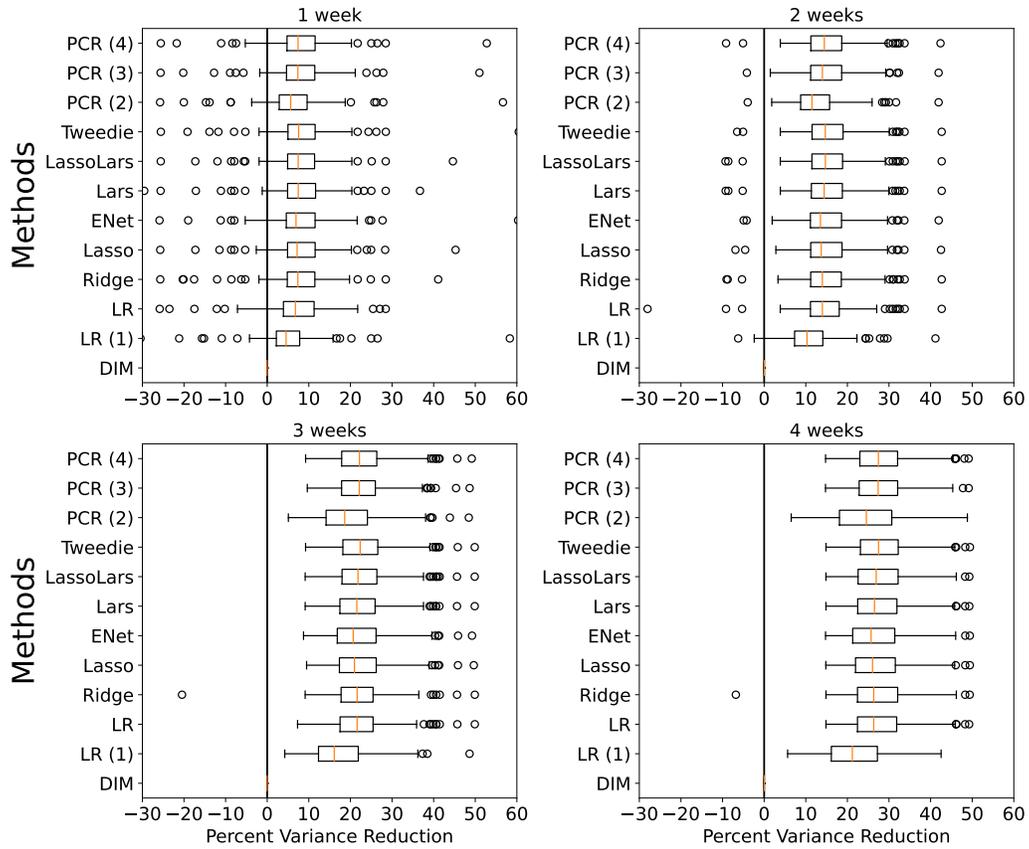
Figure 2: Boxplots of the estimated variance reduction across the experiments considered in the analysis. Each subplot in the figure corresponds to a different analysis duration time, and each row in the boxplot refers to a different model $\mathcal{M}$.
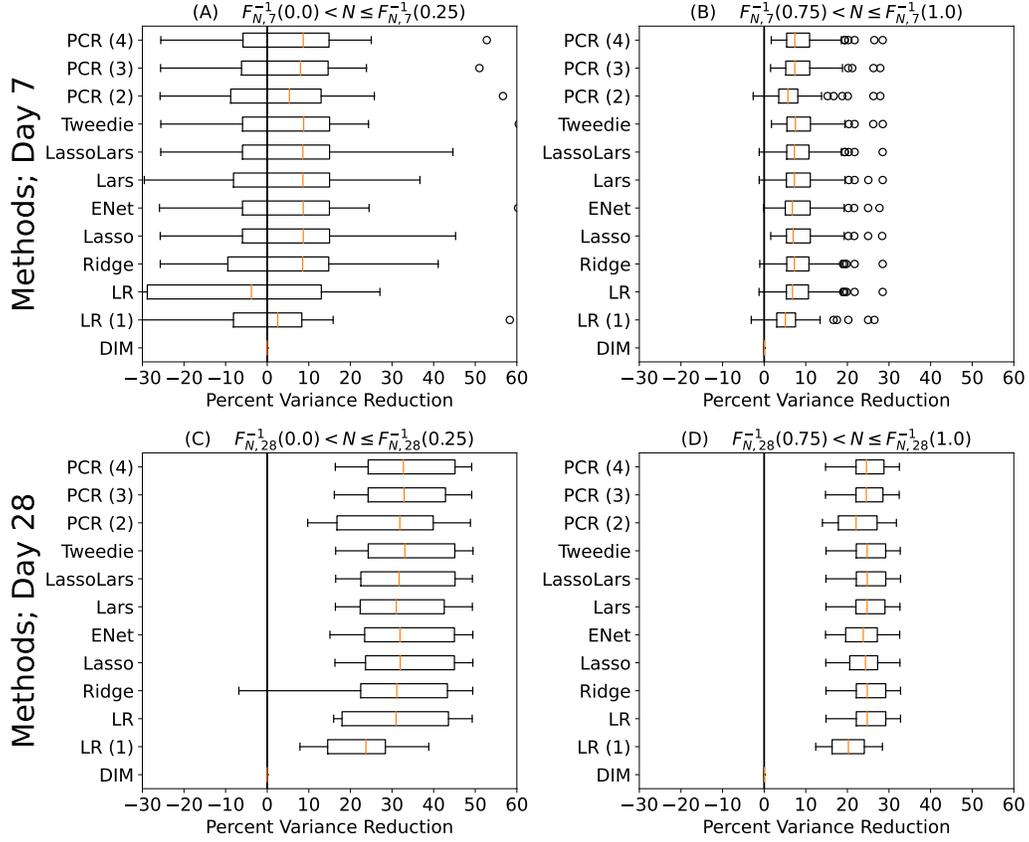
Figure 3: Boxplots of the estimated variance reduction. First row: 7-day analyses. Second row: 28-day analyses. Left subplots (A, C): $\widehat{\mathrm{VR}}_{\mathcal{M}}$ in the smallest 25 experiments. Right subplots (B, D) $\widehat{\mathrm{VR}}_{\mathcal{M}}$ for the largest 25 experiments. Each row in each boxplot corresponds to a different model $\mathcal{M}$.

treatment, condition (C1) in Section 3 is violated. As a consequence, inferences obtained from the A/B test are going to be invalid.

Even for experiments in which the triggering logic determining treatment assignments is correctly specified, however, it can be the case that in practice an experiment leads to imbalanced treatment arms. E.g., an experiment in which the triggering logic follows Equation (1) can result in an "unlucky" split of the data, in which the covariate values in treatment arms are not comparable. Concretely, assignment variables $J_{1:N} := \{J_1, \ldots, J_N\}$ could define a control group $\mathcal{I}_0$ containing units that have on average much higher values of the KPI of interest in the pre-experimental period with respect to $\mathcal{I}_1$ (or *vice versa*). In the presence of high pre-experimental covariate imbalance, practitioners worry whether they can trust their findings. We here empirically show the following:

- Under high imbalance, the difference in means estimator can systematically lead to wrong conclusions. In other terms: $\widehat{\mathrm{ATE}}_{\mathrm{DIM}}$ is unbiased unconditionally on the covariate imbalance, but it can be conditionally biased. See Figure 4.

- Covariate adjusted methods alleviate this concern, and are robust to pre-experimental chance imbalance. See Figure 5.

Notice: in this section, we focus on simple linear regression — qualitative findings for other methods are similar and omitted. To get us started, we need an operational definition of imbalance to quantify the pre-experimental comparability of the control and treatment arms. Let x_n denote the pre-experimental value of the KPI of interest Y_n for unit n, and define the imbalance parameter ζ:

$$\zeta := \zeta(x_{1:N}, J_{1:N}) = \bar{x}(1) - \bar{x}(0), \tag{7}$$

where $\bar{x}(t) := \sum_{n=1}^{N} x_n 1(J_n = t)/\{\sum_{n=1}^{N} 1(J_n = t)\}$ is the average value of the KPI in the pre-experimental period for units later exposed to treatment arm t. Intuitively, when $|\zeta|$ is large, the two groups (control and treatment) are not ex-ante comparable. When the imbalance is sufficiently severe, experimenters worry that the estimates might not be trustworthy. In turn, this typically leads to the necessity of re-randomizing the experiment. This causes inefficiency in the experimentation pipeline: re-randomizing is expensive as it requires using additional computational resources, and postponing launch decisions. Methods whose inferences are less sensitive to randomization bias are therefore preferable.

We now show on our real data that using a covariate adjusted estimator can lead to substantially better results than the unadjusted estimator, even in the presence of severe imbalance. In turn, this reduces the need to re-randomize and experimentation cost.

5.3.1 A/A analysis

To study robustness of covariate adjustments to chance imbalance, we here adopt the following "A/A" approach. Given an experiment of interest, we restrict our attention to a single arm in the experiment (e.g., control). Namely, we only consider the subset of the units $n \in \mathcal{I}_t$ exposed to policy t, for a fixed t, and discard all the other units, together with their covariates. For notation simplicity, in what follows we denote $\mathcal{I} := \mathcal{I}_t$ the set of units in treatment arm t. After this pre-processing, we can treat the units in $\mathcal{I}$ as if they were
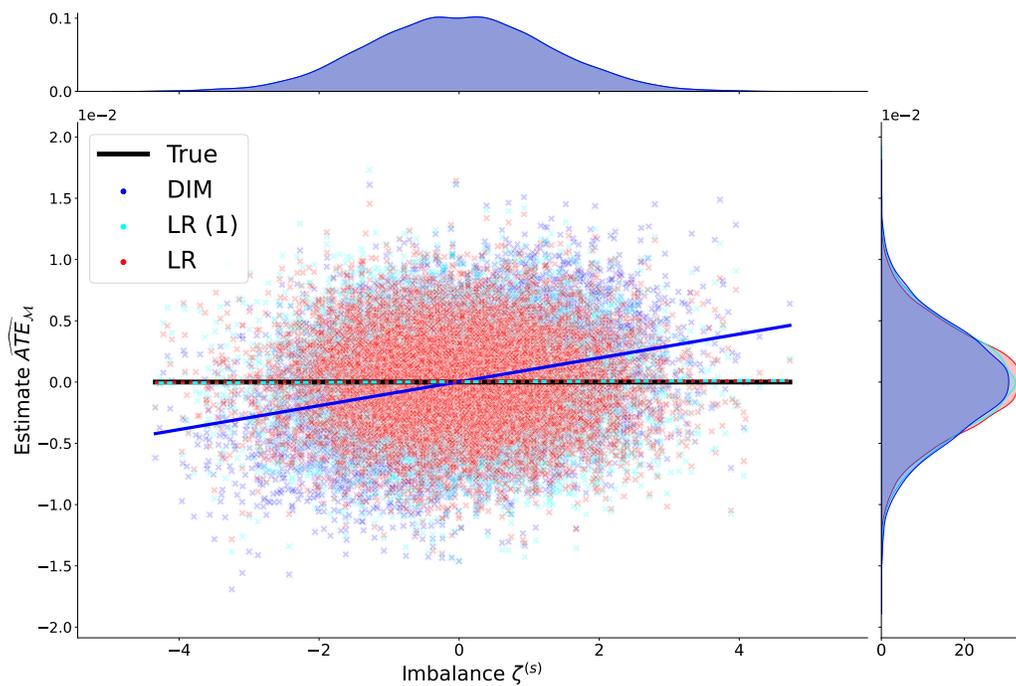
Figure 4: Comparing DIM and LR estimates for a single experiment for $S = 10,000$ re-randomizations. Each dot corresponds to a different $(\zeta^{(s)}, \widehat{\text{ATE}}_{\mathcal{M}})$ combination.

obtained from an A/A test. That is, if we were to randomly split them into two "fake" treatment arms, we would have by construction that the ground truth average causal effect is known and equal to 0. A similar experimental setup is adopted e.g. in Guo and Basse (2021, Section 4).

5.3.2 MONTE CARLO SIMULATION

To assess robustness to imbalance, we adopt a Monte Carlo approach. We fix a large integer S and for each $s = 1, \ldots, S$, we randomly split $\mathcal{I}$ into two groups, creating A/A re-randomization groups $\mathcal{I}_0^{(s)}, \mathcal{I}_1^{(s)}$ such that $\mathcal{I}_0^{(s)} \cup \mathcal{I}_1^{(s)} = \mathcal{I}$ and $\mathcal{I}_0^{(s)} \cap \mathcal{I}_1^{(s)} = \varnothing$. We then define the A/A arm indicator $J_n^{(s)} := 1(n \in \mathcal{I}_1^{(s)})$, and use Algorithm 1 to fit $\widehat{\mathrm{ATE}}_{\mathcal{M}}^{(s)}$ using data $\mathcal{D}_t^{(s)} = \{Y_{\mathcal{I}}, J_{\mathcal{I}}^{(s)}, X_{\mathcal{I}}\}$ for all the estimators $\mathcal{M}$ under consideration. Here $Y_{\mathcal{I}} = \{Y_n : J_n \in \mathcal{I}\}$. That is, we fit $\widehat{\mathrm{ATE}}$ assuming that the units in control are those with index in $\mathcal{I}_0^{(s)}$, and the unit in treatment are indexed by $\mathcal{I}_1^{(s)}$. Importantly, notice that for every re-randomization s we induce a split-specific level of imbalance $\zeta^{(s)}$ as per Equation (7). Moreover, by construction the ATE is 0, since we here let the outcome Y_n be fixed, regardless of the value of $J_n^{(s)}$. We summarize this procedure in Algorithm 2.

Algorithm 2 A/A test

Require: Data $\mathcal{D} := \{(y_n, z_n)\}_{n \in \mathcal{I}}$, set of regression models $\mathfrak{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_W\}$, treatment arm t.
    Let $\mathcal{I} := \{n \in [N] : J_n = t\}$.
    for $s = 1, \ldots, S$ **do**
        Split $\mathcal{I}$ into $\mathcal{I}_0^{(s)}, \mathcal{I}_1^{(s)}$ at random such that $\mathcal{I}_0^{(s)} \cup \mathcal{I}_1^{(s)} = \mathcal{I}$ and $\mathcal{I}_0^{(s)} \cap \mathcal{I}_1^{(s)} = \varnothing$. Let $J_n^{(s)} := 1(n \in \mathcal{I}_1^{(s)})$.
        Let $\bar{x}(\mathcal{I}_t^{(s)}) := \sum_{n \in \mathcal{I}_t^{(s)}} x_n / |\mathcal{I}_t^{(s)}|$ and compute

$$\zeta^{(s)} = \bar{x}(\mathcal{I}_1^{(s)}) - \bar{x}(\mathcal{I}_0^{(s)}).$$

        for $\mathcal{M} \in \mathfrak{M}$ **do**
            With $\mathcal{D}^{(s)} := \{Y_{\mathcal{I}}, J_{\mathcal{I}}^{(s)}, X_{\mathcal{I}}\}$, estimate $\widehat{\mathrm{ATE}}_{\mathcal{M}}^{(s)}, \widehat{\mathrm{CI}}_{\mathcal{M}}^{(s)}(\alpha)$ using Algorithm 1.
        end for
    end for

5.3.3 VALIDATION

Once we have performed the Monte Carlo procedure described above, we have access to an empirical bivariate distribution of the estimator $\widehat{\mathrm{ATE}}_{\mathcal{M}}$ as a function of the imbalance level ζ. We provide a visualization of how different estimators perform in Figure 4, where Algorithm 2 has been run on the control arm of a single experiment in the meta-analysis for $S = 10{,}000$. We make a scatterplot of the imbalance level ζ (horizontal axis) against $\widehat{\mathrm{ATE}}_{\mathcal{M}}$ for $\mathcal{M} \in \{\mathrm{DIM}, \mathrm{LR1}, \mathrm{LR}\}$ (vertical axis). Here LR1 signifies that we regress the KPI only against its pre-experimental value, while LR uses all the available covariates. We plot

on the top the marginal distribution of the imbalance level ζ (which is as expected centered around 0), and on the right the marginal density of each estimator $\widehat{\text{ATE}}_{\mathcal{M}}$. We also plot a solid color line that is the linear fit of $\widehat{\text{ATE}}_{\mathcal{M}}$ against ζ. It is evident from Figure 4 that $\widehat{\text{ATE}}_{\text{DIM}}$ is unconditionally unbiased (its expectation over re-randomization $s = 1, \ldots, S$ coincides with the true ATE 0), but it is not unbiased conditionally on imbalance.

Because the true value of the true underlying average causal effect is known, this information directly translates in a joint distribution for the accuracy of the estimator as a function of the imbalance. Estimators that are less sensitive to the imbalance allow experimenters to be confident about the results obtained even when such imbalance is present. *Vice versa*, an estimator that is sensitive to the imbalance level will make experimenters doubt their findings when the pre-experimental covariates are imbalanced. In turn, stable estimators will result in more efficient experimentation pipelines, in which data from "unlucky splits" are still useful to draw conclusions about the causal effect of interest.

We now introduce a number of metrics that allow us to translate this intuition into a quantitative assessment of the quality of the estimator as a function of the imbalance level. Fix a value $\kappa \in \mathbb{N}$, and create index sets $\mathcal{S}_1, \ldots, \mathcal{S}_\kappa$, where for each $j = 1, \ldots, \kappa$, the index set $\mathcal{S}_j \subset \{1, \ldots, S\}$ contains the indices associated with the values of $\zeta^{(s)}$ within the $100 \times \frac{j-1}{\kappa}\%$ to the $100 \times \frac{j}{\kappa}\%$ quantile of the empirical distribution of $\zeta^{(s)}$. Then, we partition the values $(\widehat{\text{ATE}}_{\mathcal{M}}^{(s)}, \zeta^{(s)})$ into κ splits of equal size (κ-iles), according to the (sorted) value of $\zeta^{(s)}$. For each $j = 1, \ldots, \kappa$, let $G_{\mathcal{M},j} : \mathbb{R} \to [0, 1]$ be the empirical cumulative density function of the estimates $\widehat{\text{ATE}}_{\mathcal{M}}^{(s)}$ falling into the j-th bucket, and let $G_{\mathcal{M},j}^{-1} : [0, 1] \to \mathbb{R}$ be its inverse (e.g., for $\kappa = 10$, we obtain the median value of the estimated ATEs which fall between the 30th and 40th percentile via $G_{3,\mathcal{M}}^{-1}(0.5)$). Within each of the κ buckets, we compute

- the estimated MSE in the j-th bucket:

$$\widehat{\text{MSE}}_{\mathcal{M},j} = |\mathcal{S}_j|^{-1} \sum_{s \in \mathcal{S}_j} \left(\widehat{\text{ATE}}_{\mathcal{M}}^{(s)} - \text{ATE} \right)^2. \tag{8}$$

- the square distance of the median value of $\widehat{\text{ATE}}_{\mathcal{M}}^{(s)}$ in the j-th bucket to the true value ATE:

$$\widehat{\text{Mediandist}}_{\mathcal{M},j} = (G_{\mathcal{M},j}^{-1}(1/2) - \text{ATE})^2. \tag{9}$$

- the excess fraction of $\widehat{\text{ATE}}_{\mathcal{M}}^{(s)}$ in the j-th bucket which underestimate or overestimate the true effect ATE:

$$\widehat{\text{Excessfrac}}_{\mathcal{M},j} = \frac{\max\{\hat{q}_{\mathcal{M},j}^+, 1 - \hat{q}_{\mathcal{M},j}^-\} - 1/2}{1/2}, \tag{10}$$

where $\hat{q}_{\mathcal{M},j}^+ = \arg\min_\alpha \{G_{\mathcal{M},j}^{-1}(\alpha) \geq \text{ATE}\}$ is the quantile associated with the smallest value $\widehat{\text{ATE}}_{\mathcal{M}}^{(s)}$ in the j-th bucket to be above ATE and $\hat{q}_{\mathcal{M},j}^- = \arg\max_\alpha \{G_{\mathcal{M},j}^{-1}(\alpha) \leq \text{ATE}\})$ is the quantile associated with the largest value $\widehat{\text{ATE}}_{\mathcal{M}}^{(s)}$ in the j-th bucket to
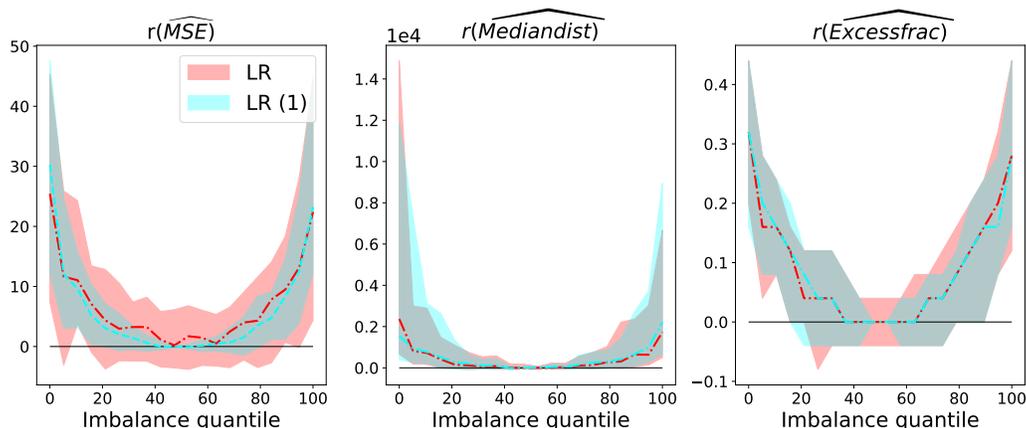
15

Figure 5: Summaries of the relative counterparts (Equation (11)) of the robustness metrics introduced in Equations (8) to (10) across the M experiments at day 7 of the analysis. Higher values correspond to higher sensitivity of DIM relative to LR, LR1. Solid line: median across experiments; Shaded regions : $(25\% - 75\%)$ percentiles.

be below the true ATE. Notice: when the distribution of the estimates is centered around the true value, Equation (10) is close to 0. In the presence of large conditional bias, it approaches 1.

Since these three metrics can all be regarded as notions of loss (the lower the value, the better), and since we're interested in assessing whether a covariate adjusted method $\mathcal{M}$ achieves lower loss than the default unadjusted method DIM, we also define for a method $\mathcal{M}$ their "relative" (to DIM) counterpart as

$$r(\widehat{\mathrm{Metric}}_{\mathcal{M},j}) := \frac{\widehat{\mathrm{Metric}}_{\mathrm{DIM},j} - \widehat{\mathrm{Metric}}_{\mathcal{M},j}}{\widehat{\mathrm{Metric}}_{\mathrm{DIM},j}}, \tag{11}$$

for $\widehat{\mathrm{Metric}} \in \{\widehat{\mathrm{MSE}}, \widehat{\mathrm{Mediandist}}, \widehat{\mathrm{Excessfrac}}\}$. For these relative metrics, larger values indicate higher sensitivity of the difference in means estimator $\widehat{\mathrm{ATE}}_{\mathrm{DIM}}$ with respect to an alternative covariate adjusted estimator $\widehat{\mathrm{ATE}}_{\mathcal{M}}$ to the imbalance level $\zeta^{(s)}$. We report the value attained by these relative metrics across all the $W = 100$ experiments on the control arm at analysis day $D = 7$. Specifically, for each experiment we re-run Algorithm 2 for $S = 10,000$ re-randomizations and compute the relative metrics across $\kappa = 20$ buckets (Figure 5). We retain for each experiment and for each bucket the median value of the relative metric, and plot a solid line connecting the median (of these medians) across the W experiments, for all imbalance quantiles $\mathcal{S}_j$, for $j = 1, \ldots, \kappa$, as well as a 50% centered empirical intervals through the shaded region. We find a similar behavior across the metrics: they are close to zero when the imbalance $|\zeta^{(s)}|$ is small (i.e., around its median value across re-randomizations). As the absolute value of the imbalance level $|\zeta^{(s)}|$ increases, however, the value of the relative metrics also sharply increases, indicating larger sensitivity to the
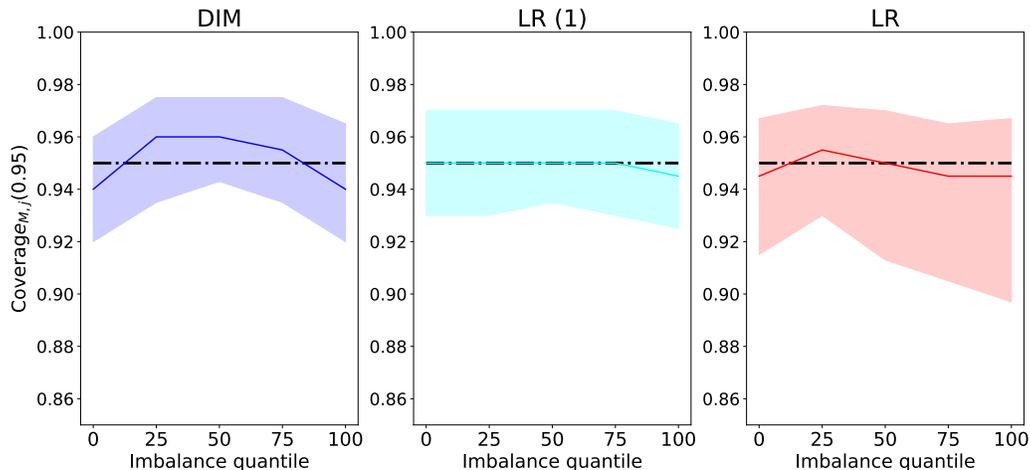
Figure 6: Summaries of $\widehat{\text{Coverage}}_{\mathcal{M},j}(\alpha)$ for $\mathcal{M} \in \{\text{DIM}, \text{LR1}, \text{LR}\}$ (vertical axis, left to right), at $\alpha = 95\%$ across $\kappa = 5$ buckets (horizontal axis). The solid line tracks the median coverage across the M experiments, and the shaded regions cover 10% – 90% percentiles across these experiments. Results are relative to the analysis at day $D = 7$. The solid black line is the target nominal value 95%.

imbalance of the difference in means estimator with respect to the linear adjusted estimators (either using one or many covariates). We conclude by checking estimators' calibration (Figure 6). Given width $\alpha = 0.95$, we compute the fraction of times that the estimated confidence interval spans the true value for each bucket $\mathcal{S}_j$:

$$\widehat{\text{Coverage}}_{\mathcal{M},j}(\alpha) = \frac{\sum_{s \in \mathcal{S}_j} \mathbb{1}\left(\text{ATE} \in \widehat{\text{CI}}_\alpha^{(s)}\right)}{|\mathcal{S}_j|}. \tag{12}$$

Attaining the nominal coverage α means that the confidence intervals are well calibrated. For this robustness metric the performance of the difference in means estimator is less less sensitive to pre-experimental imbalance than for the metrics considered in Figure 5.

5.4 Impact on experimentation time

We next illustrate how smaller estimated variances can lead to shorter experimentation time. We consider a hypothesis testing framework, where H_0 is a null hypothesis of no effect of the treatment — $H_0 : \{\text{LIFT} = 0\}$ — and H_1 is a fixed alternative the treatment has a fixed percent effect of size δ, $H_1 : \{\text{LIFT} = \delta\}$. Based on the data collected so far (e.g., the first $D = 7$ days of the analysis), we form a prediction on the number of future units that are going to trigger in the experiment as it progresses (adapting recent sample-size prediction methods, see Masoero et al. (2022); Richardson et al. (2022); Camerlenghi et al. (2022)). Given the hypotheses, the predictions, and the estimated variances at $D = 7$, we compute the first future day D' at which we expect to be able to reject the null hypothesis of no effect with at least $\pi = 80\%$ power under the alternative H_1 with a given significance
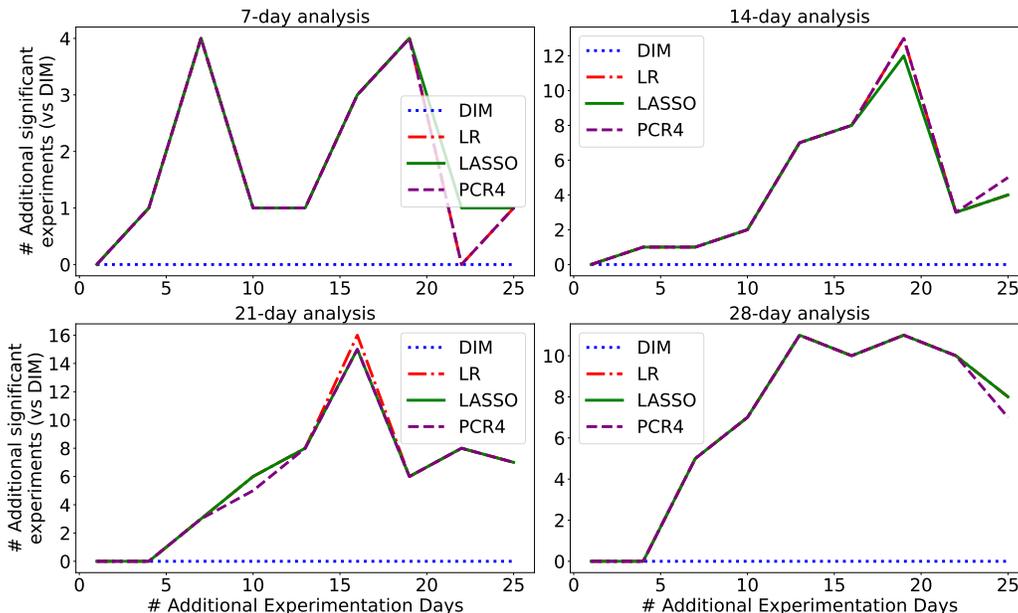
Figure 7: For a given number of additional experimentation days (horizontal axis), we plot the additional number of experiments (vertical axis) for which H_0 can be rejected at significance $\alpha = 0.05$ with 80% power given H_1 when the estimates are obtained using a model $\mathcal{M}$ as opposed to the default DIM. Different subplots refer to different experimental durations.

$\alpha = 0.05$. As seen in Figure 1, higher precision (smaller variance) directly translates in shorter experimentation time. We see, again, very similar performance across the different covariate adjusted methods considered.

We emphasize that the predictions in Figure 7 depend on a number of factors: from the properties of the experiment (e.g., the observed means and variances of the KPI), to the choice of hypothesized fixed effect value δ. However, the trend displayed in Figure 7 — whereby smaller estimated variances translate in higher power and hence shorter duration of the experiments — is expected: smaller estimated variances directly translate into shorter experiments.

5.5 Making tradeoffs at scale: computation, robustness, interpretability

As already discussed in Section 4, large scale inference engines should be designed keeping in mind the constraints imposed by the scale at which they operate. We have already discussed how linear models and regularizations thereof are robust (e.g., to chance imbalance) and interpretable. We here analyze how the computation cost and the estimation accuracy scales with the number of (noisy) additional covariates. Specifically, we test how computation and accuracy are affected by augmenting the K covariates with additional spurious covariates. To do so, we compute for each covariate k the (empirical) first moment $\hat{\mu}_k$ and second moment $\hat{\sigma}_k^2$, and draw for each $n = 1, \ldots, N$, a set of spurious
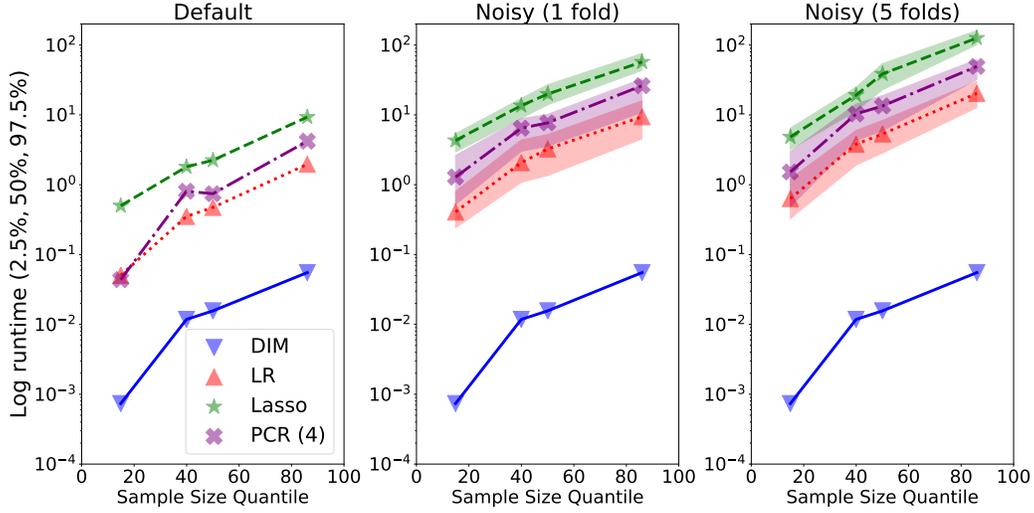
18

Figure 8: Log computation time (vertical axis) on four representative datasets having sample size $N = F_{N,7}^{-1}(\alpha)$ for $\alpha \in \{0.15, 0.4, .06, 0.85\}$ (horizontal axis). In the left plot we use the default set of K covariates. In the center plot we add one fold of K spurious covariates, in the right plot we add five folds of spurious covariates as per Equation (13). In both these last cases, we repeat the fit $N_{MC} = 100$ times on 100 different randomly drawn noisy sets of covariates.

covariates $\tilde{z}_{n,k} \sim \mathcal{N}(\hat{\mu}_k, \hat{\sigma}_k)$ i.i.d., for $k = 1, \ldots, K$. This produces an augmented set of covariates, $\tilde{\boldsymbol{z}}_n := [z_{n,1}, \ldots, z_{n,K}, \tilde{z}_{n,1}, \ldots, \tilde{z}_{n,K}]^\top$. In our experiments, we also consider even larger sets of covariates, obtained by drawing L times from each kernel $\mathcal{N}(\hat{\mu}_k, \hat{\sigma}_k)$ for every $n = 1, \ldots, N$. In the general case where we draw L spurious folds, the covariates used are:

$$
\tilde{\boldsymbol{z}}_n := [\overbrace{z_{n,1}, \ldots, z_{n,K}}^{\text{Real Covariates}}, \overbrace{\tilde{z}_{n,1}, \ldots, \tilde{z}_{n,K}}^{\text{First Spurious Fold}}, \overbrace{\tilde{z}_{n,K+1}, \ldots, \tilde{z}_{n,2K}}^{\text{Second Spurious Fold}}, \\
\ldots, \underbrace{\tilde{z}_{n,(L-1)K+1}, \ldots, \tilde{z}_{n,LK}}_{L-\text{th Spurious Fold}}]^\top \in \mathbb{R}^{(L+1) \times K}. \tag{13}
$$

Equation (13) simulates a setting in which we might be using a large set of not curated covariates, some of which are noisy and uncorrelated with the outcomes (violating condition (C2) in Section 3). We analyze in Figure 8 the computation cost of running different covariate adjusted methods, relative to the baseline DIM, as a function of the size of the experiment and the number of covariate used. In our experiments, even for larger ones, the computation cost of covariate adjusted methods is moderate. We run experiments using the popular `scipy` python library (Virtanen et al., 2020) on a 16-core Intel(R) Xeon(R) CPU E5-2686 v4 @ 2.30GHz. Computation cost increases at faster rate with the sample size than with the covariates' dimensionality.

We next assess the robustness of different regression-adjusted methods to the presence of noisy covariates. Because the true causal effect is unknown, we treat the estimate obtained using the linear regression model on the full data as the "ground truth" — i.e., ATE :=
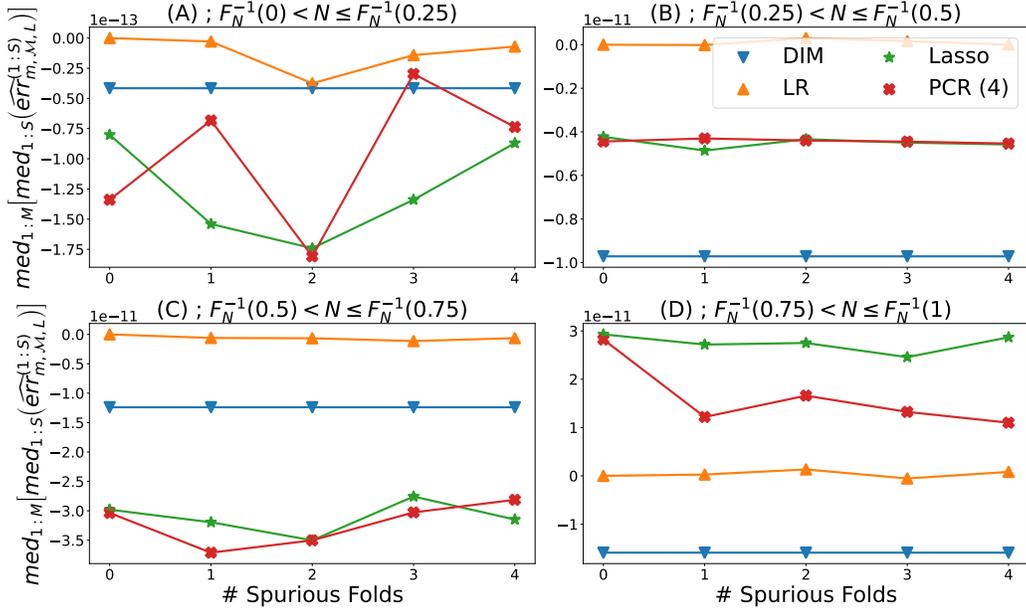
19

Figure 9: Median of median errors (vertical axis) as a function of the number of spurious folds (horizontal axis) at day $D = 7$ of the analysis. Each subplot refers to a different quartile experiments according to their sample size as per $F_{N,7}$.

$\widehat{\text{ATE}}_{\text{LR}}$ with data $\{y_{1:N}, J_{1:N}, \boldsymbol{z}_{1:N}\}$. For a fixed $L \geq 1$ and a large number of Monte Carlo random draws S, we (i) draw spurious covariates $\tilde{\boldsymbol{z}}_{1:N}$ and (ii) compute the empirical distribution of the percentage absolute difference (or error) in the estimate in the presence of spurious covariates with respect to the ground truth:

$$\widehat{\text{err}}^{(s)}_{w,\mathcal{M},L} := \frac{|\widehat{\text{ATE}}^{(s)}_{w,\mathcal{M},L} - \text{ATE}_w|}{|\text{ATE}_w|}.$$

For experiment w, $\widehat{\text{ATE}}^{(s)}_{w,\mathcal{M},L}$ is the estimate of the true ATE_w using model $\mathcal{M}$ on the s-th random re-draw of covariates with L spurious folds. Results are displayed in Figure 9, where we break down the distribution of the error by clustering experiments in the meta-analysis according to their sample size. Specifically, we divide experiments into four quartiles according to $F_{N,7}$ ((A)–(D)). Within each quartile, we compute for each experiment w and for each method $\mathcal{M}$ and fold L the median error across $s = 1, \dots, S$ ($\text{med}_{1:S}(\widehat{\text{err}}^{(1:S)}_{w,\mathcal{M},L})$). For each quartile of the sample size distribution, this procedure gives us a list of 25 values for each $\mathcal{M}, L$. We plot in Figure 9 the median across these 25 median errors (vertical axis) as a function of the number of spurious folds (horizontal axis) across different methods. The estimators considered are extremely robust to noise in terms of point estimates, even in the presence of several noisy covariates. Adding spurious noisy covariates has also limited impact on variance reduction gains, as showed in Figure 10. This shows us that the performance of covariate adjusted methods is reliable in the presence of noisy covariates, and the
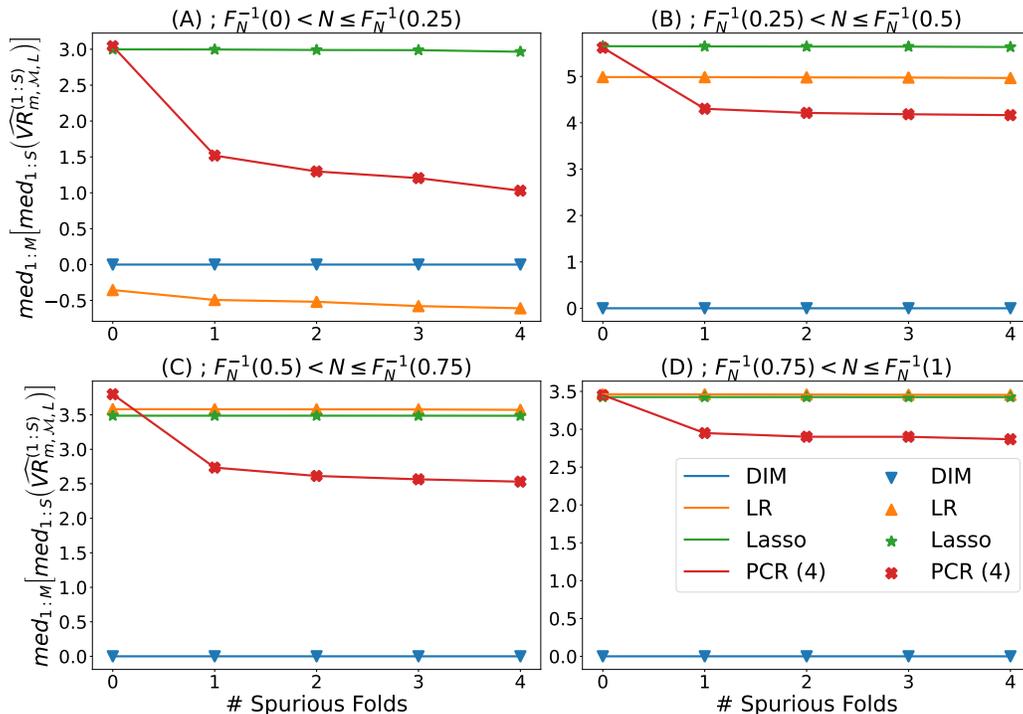
20

Figure 10: Median of median variance reduction $\widehat{\mathrm{VR}}_{\mathcal{M}}$ (vertical axis) as a function of the number of spurious folds (horizontal axis) at day $D = 7$ of the analysis. Each subplot refers to a different quartile experiments according to their sample size as per $F_{N,7}$.

computation cost — for a well optimized library — is not prohibitive, even in the presence of large sample sizes.

6 Discussion

In this paper, we have discussed the value and potential of adopting a large class of covariate adjusted models (Generalized Oaxaca-Blinder Estimators) for the estimation of causal effects in online A/B testing. GOBEs rely on a simple but very general procedure, discussed in Algorithm 1. By leveraging additional covariates and adopting linear and non-linear regression models, we showed in Section 5 on extensive experiments on real data that these estimators lead to precise (Section 5.2) and robust (Section 5.3) estimates of the causal effects of interest, which vastly outperform the simple difference in means estimator. Adoption of these estimators can help practitioners understand the effectiveness of the intervention being tested in shorter periods of time, cutting experimentation cost and streamlining the adoption of beneficial innovations (Section 5.4). The upfront cost to be paid in order to obtain these more precise estimates is both computational and statistical. We discuss these drawbacks in Section 5.5, in which we analyse the performance and computation cost incurred by linear models, and regularized versions thereof on our meta-analysis.

In particular, we focus on how such performance scales with the sizes of experiments and dimensionality of covariates. We find that for the models considered, computation cost is not prohibitive even for larger experiments, and inferences are reliable even in the presence of several spurious covariates.

In light of the practical concerns and desiderata outlined in Section 4, we choose to only consider interpretable and simple linear models, and their regularized versions. We emphasize, however, that the generalized Oaxaca-Blinder framework for the estimation of causal effects introduced in Section 3 can be straightforwardly applied to complicated, non-linear regression functions (e.g., neural networks). Fitting flexible, nonlinear regression models typically involves solving a complicated, non-convex optimization problem (like the minimization problem of Equation (3)), and might require to employ cross-fitting approaches like the ones discussed in Section 4 in order to tune regularization parameters. The design of paradigms to automate these procedure, and related cost-benefit analyses is an active research area. In settings different from the one we considered, flexible non-linear methods have the potential to vastly outperform the simple linear methods here considered. See, e.g., the discussion in Guo et al. (2021).

We envision a number of exciting avenues for future research. On the methodological side, Guo and Basse (2021) laid the foundations of a framework to provide provable guarantees for a large class of regression models for the estimation of the causal effects. Enlarging the class of models for which these guarantees hold is an exciting avenue for future work (Cohen and Fogarty, 2020; List et al., 2022). Additionally, simplifying the conditions necessary for these guarantees to hold, and strengthening their characterization could further increase the popularity of these approaches. On the applied side, the development of pipelines to automate the identification of the optimal regression functions in the presence of large and heterogeneous datasets is a crucial step towards the adoption of these methods at scale. Towards this goal, thorough investigation of the benefits, costs and risks of adopting large, flexible covariate adjusted regression methods are exciting challenges for practitioners in the upcoming years.

References

Federico Camerlenghi, Stefano Favaro, Lorenzo Masoero, and Tamara Broderick. Scaled process priors for bayesian nonparametric estimation of the unseen genetic variation. *Journal of the American Statistical Association*, pages 1–12, 2022.

Peter L Cohen and Colin B Fogarty. No-harm calibration for generalized Oaxaca-Blinder estimators. *arXiv preprint arXiv:2012.09246*, 2020.

Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 123–132, 2013.

Kevin Guo and Guillaume Basse. The generalized Oaxaca-Blinder estimator. *Journal of the American Statistical Association*, 2021.

Yongyi Guo, Dominic Coey, Mikael Konutgan, Wenting Li, Chris Schoener, and Matt Goldman. Machine learning for variance reduction in online experiments. *NeurIPS 2021*, 2021.

Somit Gupta, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, et al. Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter*, 21(1):20–35, 2019.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Ying Jin and Shan Ba. Towards optimal variance reduction in online controlled experiments. *arXiv preprint arXiv:2110.13406*, 2021.

Xinran Li and Peng Ding. General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112 (520):1759–1769, 2017.

Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

John A List, Ian Muir, and Gregory K Sun. Using machine learning for efficient flexible regression adjustment in economic experiments. Technical report, National Bureau of Economic Research, 2022.

Lorenzo Masoero, Federico Camerlenghi, Stefano Favaro, and Tamara Broderick. More for less: predicting and maximizing genomic variant discovery via bayesian nonparametrics. *Biometrika*, 109(1):17–32, 2022.

Thomas S Richardson, Yu Liu, James McQueen, and Doug Hains. A bayesian model for online activity sample sizes. In *International Conference on Artificial Intelligence and Statistics*, pages 1775–1785. PMLR, 2022.

Donald B Rubin. Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, 2(1):1–26, 1977.

Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1923/1990.

John W Tukey. Use of many covariates in clinical trials. *International Statistical Review/Revue Internationale de Statistique*, pages 123–137, 1991.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.