Meta-Analysis of Randomized Experiments with Applications to Heavy-Tailed Response Data

Nilesh Tripuraneni[*]
University of California, Berkeley

Dominique Perrault-Joncas [†]
Amazon, Seattle

Dhruv Madeka
Amazon, NYC

Dean Foster
Amazon, NYC

Michael I. Jordan
University of California, Berkeley, Amazon

Abstract

A central obstacle in the objective assessment of treatment effect (TE) estimators in randomized control trials (RCTs) is the lack of ground truth (or validation set) to test their performance. In this paper, we propose a novel cross-validation-like methodology to address this challenge. The key insight of our procedure is that the noisy (but unbiased) difference-of-means estimate can be used as a ground truth "label" on a portion of the RCT, to test the performance of an estimator trained on the other portion. We combine this insight with an aggregation scheme, which borrows statistical strength across a large collection of RCTs, to present an end-to-end methodology for judging an estimator's ability to recover the underlying treatment effect. We evaluate our methodology across 699 RCTs implemented in the Amazon supply chain. In this heavy-tailed setting, our methodology suggests that procedures that aggressively downweight or truncate large values, while introducing bias, lower the variance enough to ensure that the treatment effect is more accurately estimated.

1 Introduction

Causal inference is widely used across numerous disciplines such as medicine, technology, and economics to inform important downstream decisions Hernan and Robins [2020]. Inferring causal relationships between an intervention and outcome requires estimating the treatment effect (TE): the difference between what happened given an intervention and what would have happened in its absence. A central difficulty is that these two events are never jointly observed Rubin [2005]. TE estimation leverages randomized controlled trials (RCTs)—which randomly assign the products of interest into either the treatment or control groups—to counter selection biases and allow causal effects to be estimated via a simple differences-in-means estimate.

Indeed, the simplest "model-free" unbiased estimator of a treatment effect is the difference-in-means (DM) estimate Rubin [2005]. Such an estimator may, however, suffer from high variance in real-world scenarios which often involve heterogeneous, high-dimensional and heavy-tailed data[3]. A plethora of additional information is thus often used to improve TE estimates relative to this simple baseline. For example, pretreatment regression adjustments can significantly reduce the variance of a treatment effect estimate while adding little additional bias Angrist and Pischke [2008], Imbens and Rubin [2015]. Similarly, a host of other regularization and robustness modifications can be used to trade off bias and variance.

[*] Work done while at Amazon.

[†] Correspondence to joncas [at] amazon dot com.

[3] Such heavy-tailed data is commonplace in the large-scale RCTs which motivate our study.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

As the complexity of such estimators increases, so do the assumptions (and work) needed to establish their statistical validity. One particular setting in which this becomes easier, and which we argue arises in many practical applications,[4] is when large RCTs can be run on the same population. This setting provides an opportunity to get at the fundamental attributes of interest—the mean-squared error (MSE) of a given treatment effect estimator. Our simple insight is that the DM estimator can function as a noisy, but unbiased “label" for the treatment effect. Noisy estimates for a TE estimator performance can then be computed by comparing this estimator to the (unbiased) difference-in-means estimator via a simple, held-out validation estimate (see (4)). Our goal in this work is to judge the performance of TE estimators by pooling noisy (but unbiased) estimates of their performance *across many RCTs*. Such a procedure is desirable because it targets the actual quantity of interest, the estimator MSE, in an assumption/estimator-agnostic fashion. The primary contributions of this work are as follows:

- We process a corpus of 699 genuine RCTs implemented at Amazon across several years and we highlight the heavy-tailed nature of the response and covariate variables. The unique challenges associated with heavy-tailed estimation require careful navigation of the bias-variance tradeoff which motivates the development of an objective selection procedure for TE estimation.
- We present a selection scheme which borrows statistical strength across the corpus of RCTs in order to judge the relative performance of several commonly used TE estimators.
- We use this framework to argue that in the presence of heavy-tailed data—that often arise in large-scale technology and logistics applications—aggressive downweighting and truncation procedures are needed to control variance.

1.1 Related Work

The literature on causal inference and treatment effect estimation is vast and a comprehensive review is beyond the scope of this paper. Hernan and Robins [2020], Imbens and Rubin [2015], Angrist and Pischke [2008], Hadad [2020] and Wager [2020] provide modern perspectives on both the theory and practice of treatment effect estimation. Cross-validation (CV) also has been (and remains) a major subject of statistical inquiry as it is amongst the most widely used tools to assess the quality of an estimator and perform model selection Bayle et al. [2020], Lei [2020], Stone [1974], Geisser [1975].

Relatively little work has been done in the intersection of these two domains. Part of the difficulty stems from the fact that the standard procedure of CV breaks down for treatment effect estimation since the true treatment effect is never observed in data. Athey and Imbens [2016] and Powers et al. [2018] do provide model-specific selection methods in the context of treatment effect estimation. However, these works do not apply to arbitrary TE estimators. Closest to our work is that of Schuler et al. [2018], who use a data-splitting methodology to evaluate several risk functions to assess *heterogeneous* treatment effect estimators. This differs from our work in two principal ways. First, our framework is targets the problem of *average* treatment effect estimation—in many scenarios that we are interested in, treatments cannot be individualized and must be applied in an all-or-nothing fashion to the entire population. Our statistical scheme also differs since we provide a provably *unbiased* estimate[5] of the mean-squared error of a TE estimator, and we introduce an aggregation scheme to borrow statistical strength across different RCTs to compare estimators. Additionally, our work uses a large corpus of 699 *actual* randomized RCTs conducted at Amazon over the course of several years as our test-bed for estimator selection in contrast to synthetic data simulations.

One of our main motivations is to highlight the unique challenges associated with heavy-tailed data often present in applications arising at large-scale technology and logistics companies. Semiparametric TE estimators for heavy-tailed datasets inspired by similar applications have been explored Fithian and Wager [2014] and Taddy et al. [2016]. However, these works do not address the problem of model selection which is our central focus. Specifically, we focus on methods to select among simple estimators (with few to no tuning parameters) that are widely used in practice.

[4] Including AB testing of forecasting model improvements, website changes, supply-chain modifications, or a number of other interventions.

[5] Leveraging the unbiased nature of the DM estimator.

1.2 Preliminaries

We work within the Rubin potential outcomes model Rubin [2005] where we imagine we are given a domain of objects $\mathcal{Y}$ and a target variable of interest $Y(\cdot)$ given a possible intervention. For a fixed intervention I, our goal is to estimate the population average treatment effect (ATE):

$$\Delta = \mathbb{E}[Y(1) - Y(0)], \tag{1}$$

where $Y(1)$ corresponds to the value of an experimental unit—in our case a product in the supply chain—given the treatment and $Y(0)$ its unobserved counterfactual control (and vice versa). In general, we also allow the existence of other covariates in our model $\mathbf{X} \in \mathcal{X}$. In a given RCT, we first randomly sample an equal number of products into a treatment group, $\mathcal{T}$, and a control group $\mathcal{C}$. We further let the $(\mathbf{X}_i, T_i, Y_i)$ be the covariates, treatment dummy, and value of the ith product. By a standard argument, using the assumption of randomization (independence of $\{Y_i(1), Y_i(0)\}$ and T_i), the differences-in-means estimator,

$$\hat{\Delta}_{DM} = \frac{1}{|\mathcal{T}|}\sum_{i \in \mathcal{T}} Y_i(1) - \frac{1}{|\mathcal{C}|}\sum_{i \in \mathcal{C}} Y_i(0), \tag{2}$$

provides an unbiased estimate of Δ [Rubin, 2005]. A primary benefit of the DM estimator is that it is “model-free." That is, it makes no explicit assumptions on the data-generation process for Y_i as a function of the other covariates.

1.3 Dataset Description

We use 699 RCTs that were run at Amazon since 2017 on a population of products. The interventions in each RCT consist of various modifications and (potential) improvements to the way in which products are processed through the supply chain. The RCTs are most often constructed with 50% of products in an RCT randomly placed in the treatment group and 50% in the control group, though some are not evenly balanced. The RCTs vary in size from tens of thousands of products to those with several millions. Each RCT is run over the course of approximately 27 weeks with the intervention instituted at a trigger date at 10 weeks in the treatment group.

At each week in an RCT, the response variable generated from each product is computed. Each RCT was preprocessed to contain the averaged pretreatment response (denoted X), a strictly nonnegative averaged pretreatment auxiliary covariate (denoted D), averaged posttreatment response (denoted Y), and binary treatment indicator (denoted T) for each product. Auxiliary covariates (such as D) often arise in naturally occurring applications where it is feasible to forecast a related quantity to Y (such as the number of expected products needed in a time period to satisfy user demand).

2 Heavy Tails and Hard Estimation Case Study

The difficulties associated with treatment effect estimation of an intervention in large-scale commerce RCT datasets are many fold. The most salient difficulty for our consideration is that the response distribution over the range of products has a *heavy tail*. Similar heavy-tailed distributions are known to exist in user revenue distributions as well as user engagement metrics at large-scale technology companies [Fithian and Wager, 2014, Taddy et al., 2016]. Estimation in this setting is difficult and requires balancing several considerations when considering the pros and cons of various estimation techniques. Our exploration of these issues serves a dual purpose: (1) to highlight the ubiquitous occurrence of such heavy tails in naturally occurring data, and (2) to motivate the need for a model selection procedure to navigate the bias-variance tradeoff.

Let us investigate the data inside a single RCT to assist in further making this point. The RCT under consideration consists of millions of distinct products. This RCT (a representative choice) displays significant heavy-tail behavior, as shown in Fig. 2.

We implement the Hill estimator to obtain an estimate of the power-law behavior η in the right tail distribution of $\sim y^{-\eta}$ across all the RCTs under consideration. The Hill cutoff hyperparameter is chosen to discard points near the center of the distribution (i.e., near zero) and allows the formulation of a bias-variance tradeoff [Drees et al., 2000]. We avoid a more sophisticated data-driven choice of this cutoff since the precise Hill value is not of particular interest in our setting.[6]. Rather, it is

[6] Indeed we have tens of thousands of points in all RCTs, so small-sample difficulties associated with “Hill horror plots" seem not to arise.

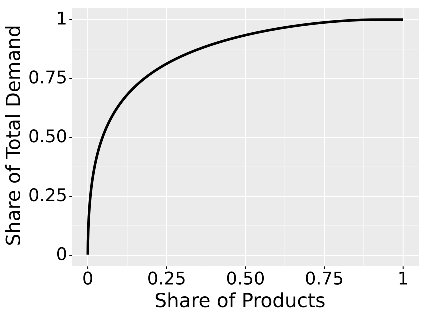

Figure 1: Gini plot of a single RCT showing the cumulative share of demand vs. product population share ordered by descending popularity. Demand is heavy-tailed with the top 20% most popular products accounting for nearly 80% of the demand share.

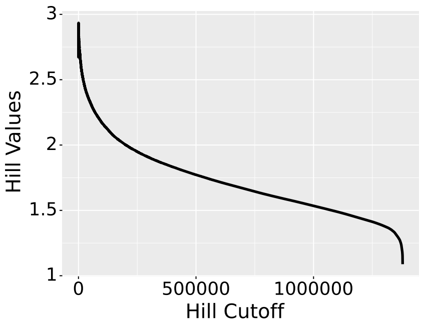

Figure 2: Hill plot of the right tail of the response variable distribution in a single RCT versus the Hill cutoff hyperparameter. The Hill values are an estimate of the power η in the asymptotic tail behavior of the response distribution variable, Y, $p(y) \sim y^{-\eta}$.

apparent the power η can be conservatively judged to be between $1 - 3$ in Fig. 2. Analyzing the response distribution across the entire corpus of 699 RCTs and choosing the Hill cutoff parameter at the 5th percentile shows that the average decay exponent is ≈ 2.32 with a standard deviation of 0.79, and median of 2.15.

The difficulties seen in this case study reinforce the conclusion that handling the heavy tails inherent in our data likely requires more sophisticated (regularized) estimators than the DM estimator. Ultimately this boils down to balancing the tradeoff between bias and variance in estimation. Navigating this bias-variance tradeoff is one of the primary motivations for our aggregation methodology for TE estimator selection.

3 Validation Procedure for Treatment Effect Estimators

In this section, we present the key idea behind the validation procedure we use to assess the quality of an arbitrary treatment effect estimator, $\hat{\Delta}_E(\cdot, \cdot)$, in the RCT denoted j. Let Δ denote the population ATE shown in (1). Given the groups $\mathcal{T}$ and $\mathcal{C}$, we first randomly partition them into disjoint groups $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{C}_1$, $\mathcal{C}_2$. Now, consider the (potentially complicated) treatment effect estimator $\hat{\Delta}_E(\mathcal{T}_1, \mathcal{C}_1)$ trained on the first fold of data. We can obtain an estimate of its performance by how well it targets the difference-of-means estimator computed on the hold-out set $\hat{\Delta}_{DM}(\mathcal{T}_2, \mathcal{C}_2)$:

$$\widehat{\text{MSE}}_{E,j}((\mathcal{T}_1, \mathcal{C}_1), (\mathcal{T}_2, \mathcal{C}_2)) = (\hat{\Delta}_E(\mathcal{T}_1, \mathcal{C}_1) - \hat{\Delta}_{DM}(\mathcal{T}_2, \mathcal{C}_2))^2. \tag{3}$$

A simple argument shows that this quantity can be used to compare the relative MSE of two different estimators. Given two different treatment effect estimators A and B in the aforementioned setting, we have:

$$\begin{aligned} &\mathbb{E}[(\hat{\Delta}_A(\mathcal{T}_1, \mathcal{C}_1) - \hat{\Delta}_{DM}(\mathcal{T}_2, \mathcal{C}_2))^2] \leq \mathbb{E}[(\hat{\Delta}_B(\mathcal{T}_1, \mathcal{C}_1) - \hat{\Delta}_{DM}(\mathcal{T}_2, \mathcal{C}_2))^2] \implies \\ &\mathbb{E}[(\hat{\Delta}_A(\mathcal{T}_1, \mathcal{C}_1) - \Delta)^2] \leq \mathbb{E}[(\hat{\Delta}_B(\mathcal{T}_1, \mathcal{C}_1) - \Delta)^2]. \end{aligned} \tag{4}$$

See Appendix A for a proof. This result motivates using the held-out sample error as a metric to assess the relative merit of two estimators $\hat{\Delta}_A$ and $\hat{\Delta}_B$. However, simply using this estimator on a single RCT provides a (potentially very) noisy estimate of the population error, not the population error itself. Indeed, if the estimator $\hat{\Delta}_{DM}(\mathcal{T}_2, \mathcal{C}_2)$ is sufficiently good to estimate Δ, why even bother to use another estimator? Said another way, the error estimate in (3) will always suffer at least the variance of the unbiased estimate (2). In practice we use a cross-validated version of (3) to reduce the subsampling variance due to the random train/test splits (see Appendix D). This procedure will not decrease the variance of the DM estimator arising from the underlying heavy-tailed data however.

Our proposal for resolving this conundrum is to note that in many situations we have access to *multiple* RCTs from the same underlying population or process given different interventions. Thus, aggregating the set of error estimates

$$\hat{\mathbf{A}} = \{\widehat{\text{MSE}}_{A,1}((\mathcal{T}_1, \mathcal{C}_1), (\mathcal{T}_2, \mathcal{C}_2)), \ldots, ...\widehat{\text{MSE}}_{A,J}((\mathcal{T}_1, \mathcal{C}_1), (\mathcal{T}_2, \mathcal{C}_2))\} \tag{5}$$

and comparing to

$$\hat{\mathbf{B}} = \{\widehat{\text{MSE}}_{B,1}((\mathcal{T}_1, \mathcal{C}_1), (\mathcal{T}_2, \mathcal{C}_2)), \ldots, ..., \widehat{\text{MSE}}_{B,J}((\mathcal{T}_1, \mathcal{C}_1), (\mathcal{T}_2, \mathcal{C}_2))\}, \tag{6}$$

for various interventions $\mathcal{J} = \{1, \ldots, J\}$, can allow us to pool information across RCTs. We sidestep the methodological complexities of performing this aggregation and instead turn to an investigation of simple, practically-motivated schemes.

3.1 An Aggregation Scheme

Aggregating the mean-squared errors requires handling a practical consideration. Since the RCTs and interventions across RCTs themselves may be different, the overall scales of the MSEs between different RCTs may be different. As an example, consider a corpus of two RCTs on which estimator A obtain errors $\{1, 10\}$ and estimator B obtains errors $\{2, 9\}$. Simply averaging the errors or doing a rank-based test of performance would indicate both estimators are equivalent. However, intuitively we believe a relative improvement of estimator B from 10 to 9 on the second RCT does not outweigh the degradation from 1 to 2 on the first RCT.

This observation motivates the definition of a normalized score to compare the estimators A vs B, as a function of the vectors of their noisy errors.[7] For each intervention $j \in \{1, ..., J\}$ we define the normalized score:

$$S_j(\hat{A}_j, \hat{B}_j) = \frac{\hat{B}_j - \hat{A}_j}{\hat{B}_j + \hat{A}_j}, \tag{7}$$

for $\hat{A}_j \in \hat{\mathbf{A}}$ and $\hat{B}_j \in \hat{\mathbf{B}}$. Where $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ are defined according to (5) and (6) respectively.

This normalized score vector (which we denote by $\hat{\mathbf{S}}(\hat{\mathbf{A}}, \hat{\mathbf{B}})$) to bound them in the range $[-1, 1]$. Each element of this vector is a noisy score of estimator A's performance relative to B *on one RCT in the corpus*.[8] If the estimator has many elements that are positive, it suggests that estimator B has larger errors than estimator A. In this case, we would expect estimator A to be better than estimator B.

To formalize this intuition we use a two-sided one-sample t-test applied to this normalized score vector to test the null that the "population mean" of the $\hat{\mathbf{S}}$ "distribution" is 0, i.e., that the performance of estimator A is indistinguishable from the performance of estimator B. Overall, this procedure interpolates between two extremes. A purely rank-based test of performance might only count the number of RCTs for which A is better than B irrespective of how much better one is in a particular RCT. Meanwhile, a procedure which only looks at the raw (unnormalized) RCT errors has the property that RCTs with large MSE values for both estimators would drown out signal from RCTs with small MSE values. We stress that the t-test heuristic provides a simple way of converting the information contained in $\hat{\mathbf{S}}(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ to a single number, but we recommend looking at the score histograms for a more complete picture.

4 Results

In this section, we present results obtained from a corpus of 699 RCTs performed at Amazon over several years as described in Section 1.3. We compare commonly used estimators for TE estimation by their out-of-sample MSE computed via the cross-validation procedure described in Section 3. See Appendix Appendix B for more details on the specific estimators.

We begin by studying several of the normalized score histograms to facilitate the comparison of our estimators; additional results are provided in Appendix C. In judging two estimators A, B via their score distribution $\hat{\mathbf{S}}(\hat{\mathbf{A}}, \hat{\mathbf{B}})$, we note that a left-skewed score distribution indicates B is a better estimator (in terms of its MSE) than A.

[7] As noted earlier, in practice each error estimate is averaged over several resampled train/test splits, but we suppress this extra notation for clarity.

[8] Our notion of a normalized score vector is element-wise transitive. That is, $\frac{b-a}{a+b} > 0$ and $\frac{c-b}{b+c} > 0$ imply $\frac{c-a}{a+c} > 0$.

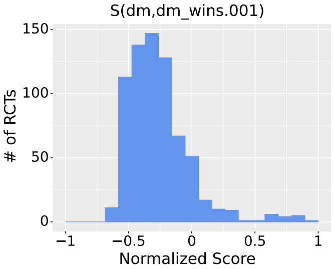

Figure 3: Histogram of the score distribution for dm vs Winsorized (at 0.001) dm estimator.

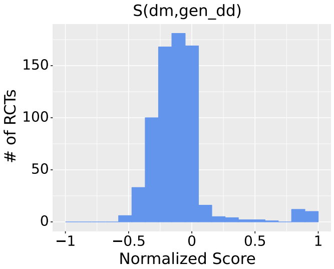

Figure 4: Histogram of the score distribution for dm vs gen_dd estimator.

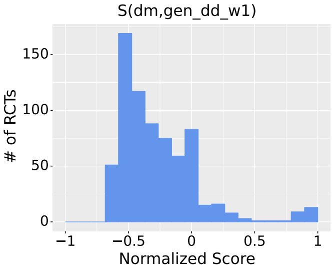

Figure 5: Histogram of the score distribution for dm vs gen_dd_w1 estimator.

In Table 1, we use the t-test heuristic from Section 3.1 to summarize each score histogram. For the sake of brevity, we do not display all the methods tested in the table. Overall, we see several phenomena that accord with our expectations. First, adjusting for the pretreatment covariate reduces variance (i.e., gen_dd is better then dm). Second, downweighting large values of Y provides significant value: inverse weighting by D and Winsorization performs generically the best under our metric (gen_dd_w1 and all Winsorized estimators perform well). We also see that the dm estimator is dominated by every other method in Table 1; such as the median of median-of-means estimator (mom1000), whose robustness underlies its improved performance.

We summarize this table by converting it into a table of pairwise comparisons of wins/losses/ties using a p-value to determine the significance of the win or loss. The question of extracting an ordered ranking from the table of wins/losses is a classic problem. The natural procedure of simply summing up the number of row-wise wins is commonly referred to as the Copeland/Borda counting method (see [Saari and Merlin, 1996] and references within).

Table 1: Comparison of Estimators via one-sample t-test applied to their normalized score vector. Easiest to read row-wise. The index (A, B) of the table computes the pair of the (t-statistic, p-value) associated with the score $\hat{\mathbf{S}}(\hat{\mathbf{A}}, \hat{\mathbf{B}})$. A large positive t-statistic at index (A, B) indicates estimator A is better then estimator B and vice-versa.

Method	dm	mom1000	gen_dd	gen_dd_w1	dm_wins.001	gen_dd_wins.001	gen_dd_w1_wins.001
dm	x	(-3.58, 0.000363)	(-12.68, 2.38e-33)	(-22.36, 3.6e-84)	(-28.19, 7.99e-118)	(-25.33, 2.96e-101)	(-24.96, 4.11e-99)
mom1000	(3.58, 0.000363)	x	(-2.12, 0.0342)	(-11.89, 7.32e-30)	(-13.51, 3.78e-37)	(-14.61, 1.94e-42)	(-15.72, 5.33e-48)
gen_dd	(12.68, 2.38e-33)	(2.12, 0.0342)	x	(-21.1, 4.73e-77)	(-19.01, 2e-65)	(-25.15, 3.11e-100)	(-23.49, 1.14e-90)
gen_dd_w1	(22.36, 3.6e-84)	(11.89, 7.32e-30)	(21.1, 4.73e-77)	x	(-0.26, 0.794)	(-5.12, 3.87e-07)	(-9.56, 1.87e-20)
dm_wins.001	(28.19, 7.99e-118)	(13.51, 3.78e-37)	(19.01, 2e-65)	(0.26, 0.794)	x	(-4.17, 3.41e-05)	(-5.39, 9.62e-08)
gen_dd_wins.001	(25.33, 2.96e-101)	(14.61, 1.94e-42)	(25.15, 3.11e-100)	(5.12, 3.87e-07)	(4.17, 3.41e-05)	x	(-4.12, 4.2e-05)
gen_dd_w1_wins.001	(24.96, 4.11e-99)	(15.72, 5.33e-48)	(23.49, 1.14e-90)	(9.56, 1.87e-20)	(5.39, 9.62e-08)	(4.12, 4.2e-05)	x

Applying such a method by inspection returns the following rankings:

gen_dd_w1_wins.001 $>$ **gen_dd_wins.001** $>$ **dm_wins.001** $\approx$ **gen_dd_w1** $>$ **gen_dd** $>$ **mom1000** $>$ **dm**

Overall, these results suggest that aggressively Winsorizing and/or downweighting heavy tails can profitably trade variance for some additional bias.

5 Conclusion

In this work, we develop a simple methodology for treatment effect model/estimator selection which pools the performance of estimators across RCTs. The methodology allows us to compare estimators on a held-out data fold in an unbiased way. The results align with a priori intuitions of estimator performance for our data corpus. One insight is that we should be trading off variance for more bias to reduce the MSE of treatment effect estimation in problems with heavy tails. Further investigation into better estimators (as judged by their held-out MSE) and their coverage is warranted.

While our corpus consists of RCTs at Amazon run over several years, we hope our primary methodological contribution – to propose a cross-validation-like methodology to evaluate TE estimators – can be used to objectively evaluate causal inference techniques in settings where large corpora of RCTs are available.

A Proofs of Estimator Validation Lemmas

We present below the proof of (4).

[Proof of (4)] We simplify the MSE of a treatment effect estimator E by centering the DM estimator around its mean and expanding the square:

$$\mathbb{E}[(\hat{\Delta}_A(\mathcal{T}_1, \mathcal{C}_1) - \hat{\Delta}_{DM}(\mathcal{T}_2, \mathcal{C}_2))^2] = \mathbb{E}[(\hat{\Delta}_A(\mathcal{T}_1, \mathcal{C}_1) - \Delta + \Delta - \hat{\Delta}_{DM}(\mathcal{T}_2, \mathcal{C}_2))^2] =$$

$$\mathbb{E}[(\hat{\Delta}_A(\mathcal{T}_1, \mathcal{C}_1) - \Delta)^2] + \mathbb{E}[(\Delta - \hat{\Delta}_{DM}(\mathcal{T}_2, \mathcal{C}_2))^2] + 2\mathbb{E}[(\hat{\Delta}_A(\mathcal{T}_1, \mathcal{C}_1) - \Delta)]\cancelto{0}{\mathbb{E}[(\Delta - \hat{\Delta}_{DM}(\mathcal{T}_2, \mathcal{C}_2))]} \implies$$

$$\mathbb{E}[(\hat{\Delta}_A(\mathcal{T}_1, \mathcal{C}_1) - \hat{\Delta}_{DM}(\mathcal{T}_2, \mathcal{C}_2))^2] = \mathbb{E}[(\hat{\Delta}_A(\mathcal{T}_1, \mathcal{C}_1) - \Delta)^2] + \mathbb{E}[(\Delta - \hat{\Delta}_{DM}(\mathcal{T}_2, \mathcal{C}_2))^2], \tag{8}$$

where the cancellation uses the independence of the first/second folds of data to factor the expectation over the two terms, and the unbiased estimation property of the DM estimator over the second fold [Rubin, 2005][9]. We then obtain the following variances for two estimators A and B:

$$\mathbb{E}[(\hat{\Delta}_A(\mathcal{T}_1, \mathcal{C}_1) - \hat{\Delta}_{DM}(\mathcal{T}_2, \mathcal{C}_2))^2] - \mathbb{E}[(\hat{\Delta}_B(\mathcal{T}_1, \mathcal{C}_1) - \hat{\Delta}_{DM}(\mathcal{T}_2, \mathcal{C}_2))^2] = \tag{9}$$

$$\mathbb{E}[(\hat{\Delta}_A(\mathcal{T}_1, \mathcal{C}_1) - \Delta)^2] - \mathbb{E}[(\hat{\Delta}_B(\mathcal{T}_1, \mathcal{C}_1) - \Delta)^2], \tag{10}$$

from which the claim follows.

B Estimators

In this Appendix we formally define the estimators used in the paper.

For the following estimators, we note that each admits a "Winsorization" which can be used to trade off bias and variance. To do this, we can simply Winsorize the covariates and targets, X, D, Y, *in only the training fold*, to reduce variance. The test folds are always left untrimmed/Winsorized so (4) remains valid. Explicitly we define Winsorization at level 0.001 to Winsorize the X, Y distributions at $P0.1$, $P99.9$ and the (positive) auxiliary D distribution at $P99.9$.

The simple **difference-of-means estimator**,

$$\hat{\Delta}_{DM} = \frac{1}{|\mathcal{T}|}\sum_{i \in \mathcal{T}} Y_i(1) - \frac{1}{|\mathcal{C}|}\sum_{i \in \mathcal{C}} Y_i(0), \tag{11}$$

as defined before is the first estimator we consider. We also consider the **Difference-of-Median-of-Means (mom) estimator**

$$\hat{\Delta}_{DMoM} = \text{MoM}(\{Y_i(1)\}_{i=1}^{|\mathcal{T}|}, B) - \text{MoM}(\{Y_i(0)\}_{i=1}^{|\mathcal{C}|}, B). \tag{12}$$

Where $\text{MoM}(\{Y_i(1)\}_{i=1}^{|\mathcal{T}|}, B)$ indicates we bucket the data into B blocks, compute the mean in each block, and the median across all the blocked means. We use mom1000 in our experiments to denote the median-of-means estimator chosen with 1000 total blocks. Next we also consider what we refer to as the **Generalized Difference-in-Differences (gen_dd) estimator** which assumes access to a pretreatment product-specific covariate X_i corresponding to the response value Y_i. So, assuming the model,

$$Y = \alpha + T \cdot \Delta + X \cdot \beta + \epsilon, \tag{13}$$

we can estimate the ATE for a binary treatment by (least-squares) regressing Y_i onto $(1, T_i, X_i)$, where ϵ_i represents a general conditionally mean-zero noise term (which may depend on X_i). If the covariates X_i are strongly correlated with the response value Y_i, incorporating them into the regression can significantly reduce the variance.

Finally we consider a reweighted version of the previous estimator we refer to as the **Weighted Generalized LR (and Generalized Difference-in-Differences) (gen_dd_w1)** estimator. That is, we can consider estimation objectives of the form:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{(1+D_i)^\gamma}(Y_i - \alpha - \Delta T_i - \beta_i X_i)^2. \tag{14}$$

[9] Throughout we also implicitly use the fact the subfolds are (uniformly) randomly sampled from the treatment and control groups—so the expectation over the subfold is equivalent to the expectations over the entire treatment/control groups.

to estimate α β, and most importantly the TE Δ. In practice, the covariate D is taken as an auxiliary covariate, which serves as positive surrogate capturing the shape of the distribution of Y. In this case the weighting has the effect of downweighting large values of Y which can be useful to regularize heavy-tailed distributions.

C Additional Results

First we present several additional estimator histograms.

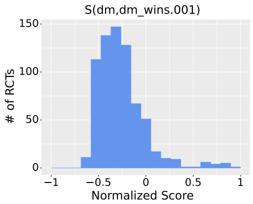

Figure 6: Histogram of the score distribution for dm vs Winsorized (at 0.001) dm estimator.

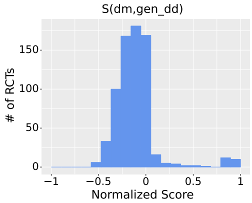

Figure 7: Histogram of the score distribution for dm vs gen_dd estimator.

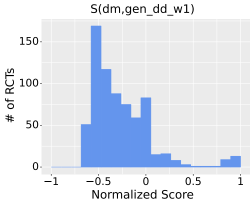

Figure 8: Histogram of the score distribution for dm vs gen_dd_w1 estimator.

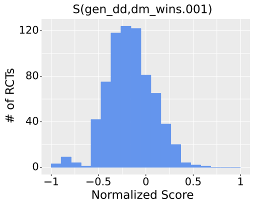

Figure 9: Histogram of the score distribution for gen_dd vs Winsorized (at 0.001) dm estimator.

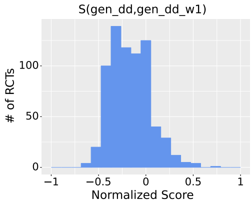

Figure 10: Histogram of the score distribution for gen_dd vs gen_dd_w1 estimator.

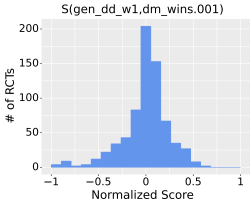

Figure 11: Histogram of the score distribution for gen_dd_w1 vs Winsorized (at 0.001) dm estimator.

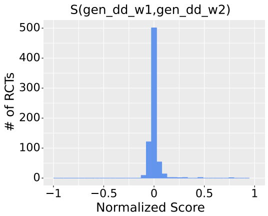

Figure 12: Histogram of the score distribution for gen_dd_w1 vs gen_dd_w2 estimator.

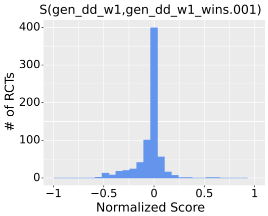

Figure 13: Histogram of the score distribution for gen_dd_w1 vs gen_dd_w1_wins.001 estimator.

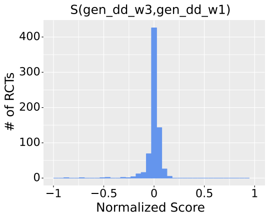

Figure 14: Histogram of the score distribution for gen_dd_w3 vs gen_dd_w1 estimator.

In this section we present additional results from our aggregation methodology to explore their stability under using different bootstrapped train/test splits to compute the normalized score vectors $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$. Tables 2 and 3 show consistent results.

D Cross-Validation Methodology

The cross-validation methodologies described in Section 3 are for the most part intuitive; nonetheless, it is worthwhile to present all the details of how we partition $\mathcal{T}$ and $\mathcal{C}$ as well as how we repeat the

Method	dm	mom1000	gen_dd	gen_dd_w1	gen_dd_w_norm	dm_wins.001	gen_dd_wins.001	gen_dd_w1_wins.001
dm	x	(-3.58, 0.000369)	(-12.49, 1.68e-32)	(-21.95, 6.88e-82)	(-17.57, 1.27e-57)	(-27.53, 5.04e-114)	(-24.74, 6.91e-98)	(-24.47, 2.51e-96)
mom1000	(3.58, 0.000369)	x	(-2.03, 0.043)	(-11.7, 5.02e-29)	(-9.15, 5.86e-19)	(-13.23, 7.35e-36)	(-14.3, 6.26e-41)	(-15.43, 1.68e-46)
gen_dd	(12.49, 1.68e-32)	(2.03, 0.043)	x	(-20.42, 3.05e-73)	(-13.23, 7.11e-36)	(-18.44, 2.49e-62)	(-24.2, 9.27e-95)	(-22.75, 1.9e-86)
gen_dd_w1	(21.95, 6.88e-82)	(11.7, 5.02e-29)	(20.42, 3.05e-73)	x	(6.83, 1.8e-11)	(-0.22, 0.828)	(-4.82, 1.78e-06)	(-9.39, 7.72e-20)
gen_dd_w_norm	(17.57, 1.27e-57)	(9.15, 5.86e-19)	(13.23, 7.11e-36)	(-6.83, 1.8e-11)	x	(-4.38, 1.37e-05)	(-8.76, 1.46e-17)	(-11.22, 5.47e-27)
dm_wins.001	(27.53, 5.04e-114)	(13.23, 7.35e-36)	(18.44, 2.49e-62)	(0.22, 0.828)	(4.38, 1.37e-05)	x	(-4.03, 6.21e-05)	(-5.27, 1.79e-07)
gen_dd_wins.001	(24.74, 6.91e-98)	(14.3, 6.26e-41)	(24.2, 9.27e-95)	(4.82, 1.78e-06)	(8.76, 1.46e-17)	(4.03, 6.21e-05)	x	(-4.11, 4.44e-05)
gen_dd_w1_wins.001	(24.47, 2.51e-96)	(15.43, 1.68e-46)	(22.75, 1.9e-86)	(9.39, 7.72e-20)	(11.22, 5.47e-27)	(5.27, 1.79e-07)	(4.11, 4.44e-05)	x

Table 2: Comparison of Estimators via one-sample t-test applied to their normalized score vector. This table was computed using error vectors from only 50 resampled train/test splits to feed into $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$. Easiest to read row-wise. The index (A, B) of the table computes the pair of the (t-statistic, p-value) associated with the score $\hat{\mathbf{S}}(\hat{\mathbf{A}}, \hat{\mathbf{B}})$. A large positive t-statistic at index (A, B) indicates estimator A is better then estimator B and vice versa.

Method	dm	mom1000	gen_dd	gen_dd_w1	gen_dd_w_norm	dm_wins.001	gen_dd_wins.001	gen_dd_w1_wins.001
dm	x	(-3.44, 0.000613)	(-12.49, 1.73e-32)	(-22.14, 5.73e-83)	(-17.7, 2.55e-58)	(-27.85, 7.64e-116)	(-25.18, 2.11e-100)	(-24.73, 8.29e-98)
mom1000	(3.44, 0.000613)	x	(-2.24, 0.0252)	(-11.93, 5.02e-30)	(-9.39, 8.29e-20)	(-13.64, 8.54e-38)	(-14.76, 3.62e-43)	(-15.8, 2.07e-48)
gen_dd	(12.49, 1.73e-32)	(2.24, 0.0252)	x	(-20.86, 1.08e-75)	(-13.48, 4.75e-37)	(-18.83, 2e-64)	(-24.98, 2.95e-99)	(-23.22, 4.23e-89)
gen_dd_w1	(22.14, 5.73e-83)	(11.93, 5.02e-30)	(20.86, 1.08e-75)	x	(6.72, 3.64e-11)	(-0.37, 0.714)	(-5.31, 1.47e-07)	(-9.42, 6.27e-20)
gen_dd_w_norm	(17.7, 2.55e-58)	(9.39, 8.29e-20)	(13.48, 4.75e-37)	(-6.72, 3.64e-11)	x	(-4.52, 7.21e-06)	(-9.1, 8.98e-19)	(-11.23, 5e-27)
dm_wins.001	(27.85, 7.64e-116)	(13.64, 8.54e-38)	(18.83, 2e-64)	(0.37, 0.714)	(4.52, 7.21e-06)	x	(-4.2, 3.05e-05)	(-5.32, 1.37e-07)
gen_dd_wins.001	(25.18, 2.11e-100)	(14.76, 3.62e-43)	(24.98, 2.95e-99)	(5.31, 1.47e-07)	(9.1, 8.98e-19)	(4.2, 3.05e-05)	x	(-3.87, 0.000119)
gen_dd_w1_wins.001	(24.73, 8.29e-98)	(15.8, 2.07e-48)	(23.22, 4.23e-89)	(9.42, 6.27e-20)	(11.23, 5e-27)	(5.32, 1.37e-07)	(3.87, 0.000119)	x

Table 3: Comparison of Estimators via one-sample t-test applied to their normalized score vector. This table was computed using error vectors from only 50 resampled train/test splits to feed into $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ distinct from those in previous table. Easiest to read row-wise. The index (A, B) of the table computes the pair of the (t-statistic, p-value) associated with the score $\hat{\mathbf{S}}(\hat{\mathbf{A}}, \hat{\mathbf{B}})$. A large positive t-statistic at index (A, B) indicates estimator A is better then estimator B and vice versa.

procedure to cross-validate (3). We start with a formal definition of treatment $\mathcal{T}$ and control $\mathcal{C}$ groups. Let j be some lab in $\mathcal{J}$; then, the treatment group for this lab is the set of outcomes $Y_{i,j}$ and features $X_{i,j}$ for each product i under the in the treatment arm $T_{i,j} = 1$, that is $\mathcal{T}_j = \{(Y_{i,j}, X_{i,j}) | T_{i,j} = 1\}$ where $|\mathcal{T}_j| = K_j$ is the number of products that were assigned to the treatment arm $T_{i,j} = 1$. Similarly, the control group is given by $\mathcal{C} = \{(Y_{i,j}, X_{i,j}) | T_{i,j} = 0\}_{i=K_j+1}^{M_j}$ where $|\mathcal{C}_j| = M_j - K_j$ is the number of products assigned to the control arm $T_{i,j} = 0$ and M_j is the total number of products in the lab.

The goal of our methodology is to find an optimal estimator $\hat{\Delta}_j$ for the ATE Δ_j or an optimal roll out policy D_j under some objective function L. This means finding a function of $\mathcal{T}$ and $\mathcal{C}$ such that $f(\mathcal{T}, \mathcal{C}) \in \mathbb{R}$ or $f(\mathcal{T}, \mathcal{C}) \in \{0, 1\}$ respectively, and that optimizes the expected objective:

$$\mathbb{E}[L(f(\mathcal{T}, \mathcal{C}), \Delta)] \tag{15}$$

for f in some functional space $\mathcal{F}$. As discussed in Section 3, to do this in the context of an RCT where Δ is unknown, we rely on the fact that the difference-in-means estimator $\hat{\Delta}(\mathcal{T}, \mathcal{C})$ is unbiased for the ATE Δ. Specifically, for any lab i, we randomly split the treatment and control group using two random subsets of product indices $S_j = \{1, \ldots, K_j\}$ and $R_j = \{K_j + 1, \ldots, M_j\}$ so that we end up with the four following sets:

- $\mathcal{T}_{j,1} = \{(Y_{i,j}, X_{i,j}) | T_{i,j} = 1 \text{ and } i \in S_j\}$
- $\mathcal{C}_{j,1} = \{(Y_{i,j}, X_{i,j}) | T_{i,j} = 0 \text{ and } i \in R_j\}$
- $\mathcal{T}_{j,2} = \{(Y_{i,j}, X_{i,j}) | T_{i,j} = 1 \text{ and } i \notin S_j\}$
- $\mathcal{C}_{j,2} = \{(Y_{i,j}, X_{i,j}) | T_{i,j} = 0 \text{ and } i \notin R_j\}$.

We also pick the size of S_j and R_j so that the split proportion p is constant across treatment, control, and labs:

$$\frac{|S_j|}{K_j} = \frac{|R_j|}{M_j - K_j} = p.$$

With this splitting methodology, we can now replace (15) with the empirical mean of the objective over all the labs in $\mathcal{J}$:

$$\frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} L(f(\mathcal{T}_{j,1}, \mathcal{C}_{j,1}), \hat{\Delta}_{DM}(\mathcal{T}_{j,2}, \mathcal{C}_{j,2})).$$

We can now optimize empirical objective for f similarly to empirical risk minimization for supervised learning. We can also "cross-validate" the empirical mean of the objective to reduce the subsampling variance and to get confidence intervals. To do this we simply repeat the splitting procedure multiple times so that every random index set S_j and R_j is now also indexed by a split $b \in \{1, \ldots, B\}$. Putting all of this together, we now have:

$$\frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \frac{1}{B} \sum_{b=1}^{B} L(f(\mathcal{T}_{j,b,1}, \mathcal{C}_{j,b,1}), \hat{\Delta}_{DM}(\mathcal{T}_{j,b,2}, \mathcal{C}_{j,b,2})). \tag{16}$$

This is how we estimated (3) in the paper, using $p = 0.5$ and $B = 100$. It is worth noting that in the case of (3), we ended up replacing the outer sum of (16) with the aggregation methodology of Section 3.1 to deal with the heavy-tailed nature of M_j, i.e. to ensure that the largest labs did not dominate the value of (16).

References

Miguel Hernan and James Robins. *Causal Inference: What If.* Boca Raton: Chapman and Hall/CRC, 2020.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics*. Princeton University Press, 2008.

Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Vitor Hadad. Ml-based causal inference tutorial. https://bookdown.org/stanfordgsbsilab/tutorial/, 2020.

Stefan Wager. Stats 361: Causal inference. *None*, 2020.

Pierre Bayle, Alexandre Bayle, Lucas Janson, and Lester Mackey. Cross-validation confidence intervals for test error. *arXiv preprint arXiv:2007.12671*, 2020.

Jing Lei. Cross-validation with confidence. *Journal of the American Statistical Association*, 115 (532):1978–1997, 2020.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

Seymour Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328, 1975.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H Shah, Trevor Hastie, and Robert Tibshirani. Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, 37(11):1767–1787, 2018.

Alejandro Schuler, Michael Baiocchi, Robert Tibshirani, and Nigam Shah. A comparison of methods for model selection when estimating individual treatment effects. *arXiv preprint arXiv:1804.05146*, 2018.

William Fithian and Stefan Wager. Semiparametric exponential families for heavy-tailed data. *arXiv preprint arXiv:1307.7830*, 2014.

Matt Taddy, Hedibert Freitas Lopes, and Matt Gardner. Scalable semiparametric inference for the means of heavy-tailed distributions. *arXiv preprint arXiv:1602.08066*, 2016.

Holger Drees, Sidney Resnick, and Laurens de Haan. How to make a Hill plot. *The Annals of Statistics*, 28(1):254–274, 2000.

Donald G Saari and Vincent R Merlin. The Copeland method. *Economic Theory*, 8(1):51–76, 1996.