When Should We Use Biased Estimators of the Average Treatment Effect?*

Jake Bowers† Gustavo Diaz‡ Christopher Grady§

November 7, 2022

Link to most recent version

Abstract

Researchers routinely use biased estimators of the average treatment effect when analyzing experimental data. Under desirable conditions in terms of sample size, treatment assignment probabilities, and the distribution of outcomes, the cost of a small increase in bias can translate into large gains in precision. When conditions are not desirable, however, the bias outweighs the precision enhancement. How can applied researchers learn which situation their particular research design leaves them in? We discuss this bias-precision trade-off in the context of common and uncommon approaches to the estimation of average treatment effects in experiments: randomization within fixed strata or blocks and M-estimation for outcomes with large outliers or skew. Using simulations, we dramatize situations where the increase in precision warrants the use of a biased estimator, as well as situations where one should not choose a biased estimator. We illustrate how to apply these ideas in a realistic setting by following the research design of a block-randomized messaging field experiment with rare and skewed outcomes conducted by the Office of Evaluation Sciences. Our main contribution is to provide guidelines to address the bias-precision trade-off at the pre-analysis planning stage so that researchers can make informed decisions about the choice of estimator for the average treatment effect in randomized experiments.

*__Work in progress, please ask before citing.__ We thank Cyrus Samii for valuable feedback.

†Associate Professor. Department of Political Science and Statistics. University of Illinois at Urbana-Champaign. E-mail: jwbowers@illinois.edu

‡Postdoctoral Fellow. Department of Political Science. McMaster University. E-mail: diazg2@mcmaster.ca

§Senior Metrics Advisor. United States Agency for International Development. E-mail: cgrady@usaid.gov

Contents

1 Introduction **3**

2 Block-randomized experiments and the bias in the fixed-effects approach to estimation **4**

2.1 Setting 5

2.2 Simulation 7

2.3 Summary of section 9

3 Estimating the mean with something other than the mean: M-estimators **10**

3.1 Setting 13

3.2 Simulation 15

3.3 Summary of section 17

4 Application to the GSA Auctions Experiment **17**

4.1 Simulations 18

4.2 Summary of section 18

5 Discussion and Conclusion **18**

5.1 A note on testing versus estimation in randomized experiments 18

5.2 Conclusion 19

References **19**

1 Introduction

Most researchers estimate the average treatment effect of their randomized experimental treatment using biased estimators. For example, consider the expression $Y_i = \beta_0 + \beta_1 Z_i + \beta_2 x_1$. If Z is a binary treatment, Y a continuous outcome, and x_1 a covariate, then estimating the ATE using OLS to find $\hat{\beta}_2$ in that expression is to use a biased estimator (Freedman 2008; Lin 2013). Yet, for the vast majority of experiments (when sample sizes are large, treatment assignment probabilities not extreme, and outcomes not extremely skewed or rare), this bias is small and the gains to precision may be large (Green 2009; Schochet 2010). As a result, using this biased estimator is a widespread practice. Other biased estimators are equally common — the use of fixed effects to weight block-randomized experiments (which is a practice closely related to covariance adjustment) — or should be used more frequently — M-estimators to estimate the ATE in the presence of very skewed outcomes or outcomes with severe outliers.[1]

The big questions to answer when selecting a biased vs an unbiased estimator are "How much bias?" and "How much more precision?" In this paper we present simulations to dramatize situations where (1) the bias is small and the gains in precision are large and (2) the bias is large, overwhelms the precision gains, and threatens the integrity of the p-values and confidence intervals used for statistical inference. The point of presenting these simulations is to show how one might answer the question "Should I use a biased estimator in this given situation?" They also suggest some general scenarios where it may be especially worthwhile to launch the kinds of simulation studies that we demonstrate here.[2]

Our advice is to use biased estimators when you can convince yourself that the bias will be small. Often it will be. And the act of such convincing — using simulations — protects you from the cases when the bias is large.

Our plan in this paper is to introduce simulated data scenarios for one common practice (es-

[1] In simple experiments the unbiased estimator of the ATE is simply a difference of means whatever the distribution of the outcome.

[2] We provide commented code for these simulations so that others can use our approach — which in turn relies heavily on the Blair et al. (2019) package. See TODO INSERT GITHUB LINK.

timating average treatment effects in block-randomized studies using "fixed effects") and one uncommon practice (estimating average treatment effects using extreme point or skewed distribution robust M-estimators). We present two designs per estimation approach: one where the bias is small and precision improvement is large and one where the bias is large and problematic.

After we analyze simulated data that exhibits the bias-variance trade-off, we apply these approaches to a messaging field experiment run by the OES: the GSA Auctions Experiment. We are not allowed to directly share the data from that experiment, but we simulated treatment assignment, blocking variables, and outcomes that match the overall pattern in the Auctions Experiment.[3]. In that experiment, it was clear that the robust estimator of the ATE was better than the simple ATE and that the biases from fixed effects were small compared to the gains in precision given the zero-inflated outcome.

Throughout we use the DeclareDesign approach to simulation. We have found that DeclareDesign makes learning about the properties of estimators much easier than past bespoke simulation approaches. All code and data available in http:://github.com/BLAH.

2 Block-randomized experiments and the bias in the fixed-effects approach to estimation

Researchers use block-randomization to increase the precision of ATE estimates (Imai, King, and Stuart 2008).[4] In a block-randomized experiment, researchers group observations in strata, conduct independent experiments in each, and combine results across blocks with a weighted ATE. Recent work shows that blocking on prognostic covariates, relative to complete randomization, tends to improve bias and precision (Pashley and Miratrix 2021). Still, researchers must choose how to weight the ATEs from each block to calculate an overall ATE. The two primary approaches are block-size weights and precision weights (Gerber and Green 2012). Precision weights are equivalent to fixed effects in linear regression.

[3]The pattern of assignments and blocking exactly matches the actual experiment, the outcomes are simulated to match the observed outcome distribution

[4]Another reason to use block-randomization is to facilitate sub-group comparisons (Bowers 2011). We do not discuss this here.

Previous work favors block-size weights since they are an unbiased estimator of the ATE in a block-randomized experiment (e.g. Gibbons, Serrato, and Urbancic 2018; Humphreys 2009).[5] However, precision weights can yield biased yet more precise estimates, especially when the proportion of treated units varies by block (Gerber and Green 2012, chap. 4).[6] Since experiments are costly, some researchers may prefer to trade unbiasedness for precision up to some point. The expression for the bias created by precision-weights is well known: TODO insert it here and discuss it. However, our discussion of that expression highlights the uncertainty than an individual researcher may face when making decisions about bias versus precision in a given case. In that case, we suggest that the researcher should try to use simulation to represent the finite-sample details of their own study and to thus reason about this trade-off in the very context in which the research is being conducted.

We illustrate the trade-offs between alternative weighting schemes with simulations that vary the proportion of treated units and the distribution of treatment effects across blocks. As implied by the expression for bias above, we find that the choice of weighting scheme depends on the correlation between treatment assignment probabilities and treatment effects across blocks. The choice is irrelevant when the proportion of treated units is constant across blocks. However, the choice of weighting scheme matters when the proportion of treated units varies across blocks. In that case, precision weights are as unbiased and more precise than block-size weights when treatment effects across blocks are small, but they introduce bias when treatment effects are large.

2.1 Setting

Consider an experiment on a sample of n units grouped in B blocks, with n_b units per block. In each block, m_b units are assigned to treatment and $(n_b - m_b)$ to control. Let the individual causal effect be $\tau_i = Y_i(1) - Y_i(0)$, with $Y_i(1)$ and $Y_i(0)$ the potential outcomes under treatment and control, respectively. We are interested in the overall ATE $\tau \equiv (1/n) \sum_{i=1}^{n} \tau_i$, which we cannot observe. In an experiment using complete or simple randomization we would estimate τ using

[5] See also https://declaredesign.org/blog/biased-fixed-effects.html

[6] See also Kalton (1968) and Hansen and Bowers (2008) for more arguments showing that precision weights produce the most precise statistical tests.

the difference in mean observed outcomes, Y_i, $\hat{\tau} = \frac{\sum_{i \in t} Y_i}{m} - \frac{\sum_{i \in c} Y_i}{(n-m)}$ but that estimator is usually biased under block-randomization unless every block is the same size and the proportion assigned to treatment in each block is the same size. (TODO show how the bias expression changes in this case). Instead, we estimate the block-level ATE for each block b, $\tau_b \equiv (1/n_b) \sum_{i=1}^{n_b} \tau_i$

using

$$\hat{\tau_b} = \frac{\sum_{i \in t} Y_{ib}}{m_b} - \frac{\sum_{i \in c} Y_{ib}}{(n_b - m_b)} \tag{1}$$

where the summation subscripts denote units in the treatment (t) and control (c) groups. Given a set of $\hat{\tau_b}$ for $b = 1 \dots B$, how shall we combine them to estimate τ? If blocks differ in size an obvious choice is to create a weighted average of each block level estimate

$$\hat{\tau}_{wt} = \frac{1}{B} \sum_{b=1}^{B} w_b \hat{\tau_b} \tag{2}$$

and to choose $w_b = \frac{n_b}{n}$ such that we have the block-size weights estimator

$$\hat{\tau}_{nbwt} = \frac{1}{B} \sum_{b=1}^{B} \frac{n_b}{n} \hat{\tau_b}. \tag{3}$$

This estimator makes intuitive sense in that each individual in the study receives equal weight (in the unweighted estimator, individuals in small blocks would receive too much weight and individuals in large blocks, too little). (TODO show how this is an unbiased estimator).

Although we can see that this estimator is unbiased, is it best? An unbiased estimator produces results that are not systematically different from the truth but it need not produce results that are close to the truth (or closest to the truth). If we want to choose weights such that $Var(\hat{\tau}_{wt})$ is small (or even such that the RMSE of $\hat{\tau}_{wt}$) is small), it turns out that weighting by the variance of the estimator of each block specific $\hat{\tau}$ is best. We call this the precision or harmonic weights

estimator or "fixed effects" or "least squares dummy variable" estimator:

$$\hat{\tau}_{hbwt} = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{h_b} \hat{\tau}_b \tag{4}$$

where $h_b = n_b p_b (1 - p_b)$, with p_b as the proportion treated units in block b. (TODO show the math connecting this expression to the OLS expression. Maybe in the appendix.)

The block-size weights estimator is an unbiased estimator of the ATE in a block-randomized experiment, even in cases with a wild distribution of block sizes and the proportion of treated units within each (Pashley and Miratrix 2020). However, as we illustrate below, in some circumstances the distribution of estimates from hypothetical realizations of the same experiment has high variance, which means the estimator has low precision. As its name suggests, the precision weights estimator yields a more precise distribution of estimates under similar conditions. Therefore, researchers interested in detecting non-zero effects may choose to sacrifice unbiasedness for increased precision when analyzing data from block-randomized experiments. The next sub-section presents dramatized scenarios to illustrate how researchers might use simulation to navigate the bias versus precision trade-off.

2.2 Simulation

We simulate stylized block-randomized experiments with three blocks. We assume potential outcomes under control $\boldsymbol{Y(0)} \sim N(0, 1)$ and $Y_i(1) = Y_i(0) + \tau_b + \epsilon_i$, with $\boldsymbol{\epsilon} \sim N(0, 1)$. This represents varying treatment effects within blocks (ϵ_i) and across blocks (τ_b). To facilitate illustration, we fix block sizes so that $\boldsymbol{n_b} = (100, 100, 100)$. Across experiments, we vary the proportion of treated units, considering homogeneous, $\boldsymbol{p_b} = (.5, .5, .5)$, and varied, $\boldsymbol{p_b} = (.05, .4, .5)$, blocks. We also consider small and large treatment effects (measured as standard deviations of $Y_i(0)$, $\boldsymbol{\tau_b} = (.25, .2, 0)$ and $\boldsymbol{\tau_b} = (4, 2, 0)$. We discuss alternative parameter values in section TODO of the supplementary materials.

For each parameter combination, we simulate $1,000$ experiments. Figure 1 reports the distribu-

tion of ATE estimates using block-size and precision weights, which we compare against the true ATE under the corresponding research design. An estimator is less biased as the center of the distribution approaches the true ATE. An estimator is more precise if the distribution is narrower.

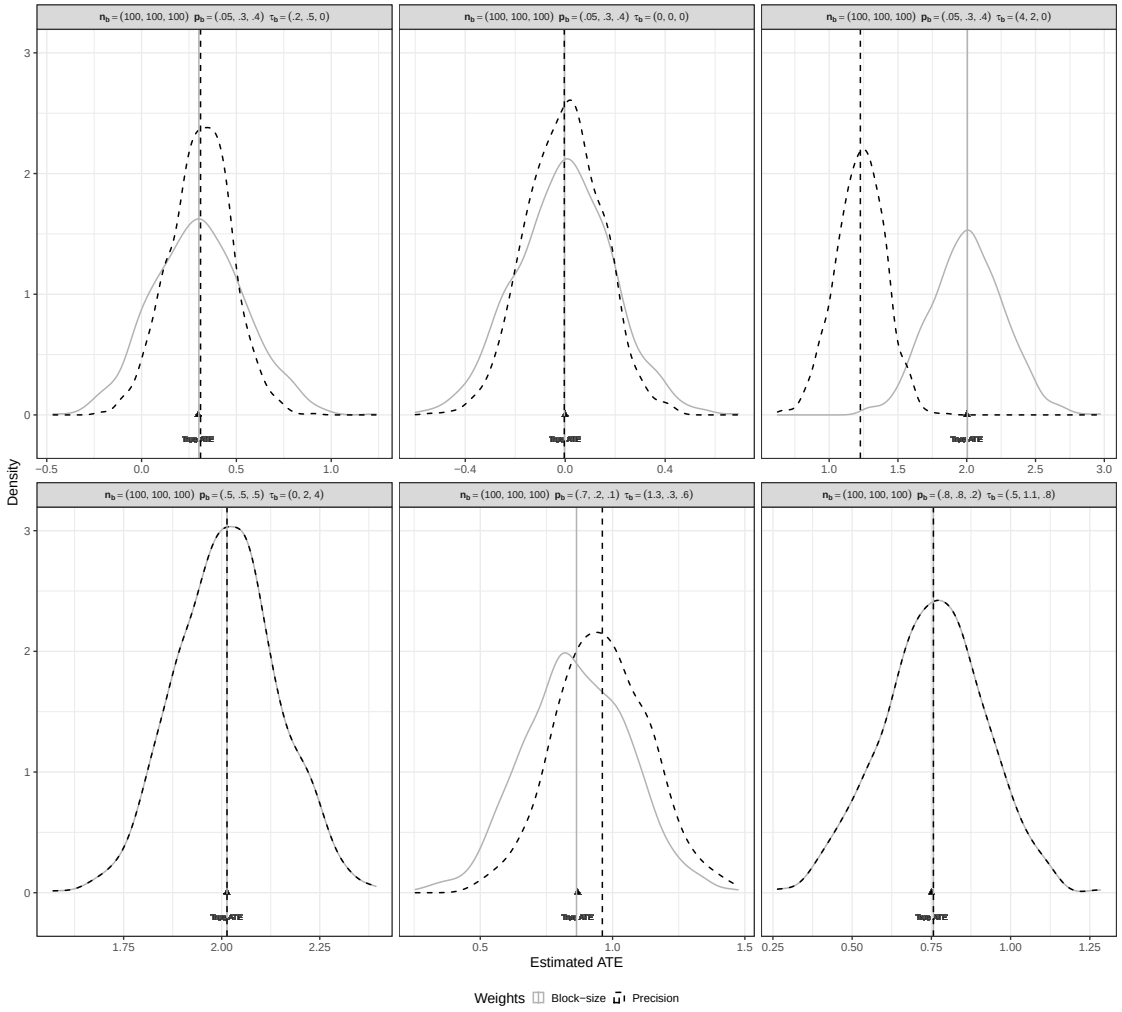

Figure 1: Estimates from simulated block-randomized experiments using alternative weighting schemes. Each panel is based on $1,000$ simulations.

Figure 1 shows what the mathematical expressions for bias and precision suggested in general form, but makes the trade-off more vivid in the cases that we chose to explore. First, we see that both weighting schemes perform the same when $\boldsymbol{p_b}$ is (nearly) the same in each block, regardless of $\boldsymbol{\tau_b}$ (shown in only one panel here, with $p_b = (.5, .5, .5)$) and with $p_b = (.8, .8, .2)$. When p_b varies but there are constant treatment effects (or no effects at all), such that $\tau_b = (0, 0, 0)$, then

the precision weighted estimator is more precise with little bias. And the same pattern holds for $p_b = (.05, .3, .4)$ and $\tau_b = (.2, .5, 0)$.

We see both weighting schemes perform the same when the correlation between probabilities of assignment and treatment effect is zero (which means that the precision weights have no correlation with treatment effects): both the case with p_b constant at .5 (implying no correlation) and $p_b = (.8, .8, .2)$ with $\tau_b = (.5, 1.1, .8)$ have zero correlation between treatment effects and precision weights. Second, the choice of weighting scheme matters when $\boldsymbol{p_b}$ varies and is also sensitive to the pattern of true underlying treatment effects as they vary across the blocks. If the set of $\boldsymbol{\tau_b}$ is small, precision weights have very little bias, but they are more precise (we see this in the panel labeled TODO). This means researchers interested in detecting non-zero effects may prefer precision weights in this case. Third, we were able to create a situation with large bias.[7] (In other materials we can switch the sign of the bias by changing the sign of the correlation between τ_b and p_b vectors. This bolsters the case for block-size weights. We can see here that the direction of the bias depends on whether there is a correlation between treatment effect heterogeneity across blocks and the pattern of weights: in this case all blocks have the same block-size weight but only differ in precision-weights and the blocks with probabilities of treatment (p_b) closer to .5 have the largest precision-weight. This is because in a two-arm experiment with the same variance in treated and control potential outcomes, an experiment assigning half of the units to treatment and half to control has the most precision (see Chapter 3, Gerber and Green 2012). We see that the bias is positive in this case when the block with the largest treatment effect has the most precision weight and the block with the smallest treatment effect has the lowest precision weight (and the bias is negative in the opposite case.

2.3 Summary of section

In a study of villages of different sizes but with a fixed budget for an experimental treatment to be randomly assigned within village (or school or other site), a researcher might end up a research design with unequal probabilities of treatment assignment across such strata or blocks. In such

[7] Hat tip to https://declaredesign.org/blog/biased-fixed-effects.html for the inspiration.

a case, the common approach, to estimate average treatment effects using OLS with fixed effects for blocks (which we showed above is identical to the precision-weighting approach of combing block-specific estimates), will produce a biased estimate of the overall average treatment effect. The amount of bias in this estimate will depend on the correlation between the probabilities of treatment and treatment effect sizes across the blocks. If the probabilities of treatment vary across blocks in a way that is uncorrelated with (or only slightly correlated with) treatment effects, then the bias will be small and researchers might prefer the bias vs precision trade-off. If the probabilities of treatment do not vary (i.e. the researchers can assign the same *proportion* of people to treatment in each village), then either approach will yield the same estimates — the two weights are identical.

How would a researcher know which situation they are in? After the experiment is complete, it is not difficult to calculate the block-specific estimates. And probabilities of treatment assignment within each block are controlled by and known by the researcher as well. So an investigator could calculate this correlation. Here we engineered a correlation of either roughly -0.97 (in the case with $p_b = \{.05, .3, .4\}$ and $\tau_b = \{4, 2, 0\}$ which produced high bias) and -0.17 (in the case with $p_b = \{.05, .3, .4\}$ and $\tau_b = \{.2, .5, 0\}$ which produced little bias). If the researcher wondered whether this pattern might allow an increase in precision at the cost of little bias, the researcher could simulate as we have done — but using the values calculated from the observed experiment.

3 Estimating the mean with something other than the mean: M-estimators

Under standard experimental assumptions, design-based estimators provide an unbiased estimate of the average treatment effect. However, the mean, the difference in means, and their OLS analogues are sensitive to unusual data with skewed outcomes (TODO canonical cites to the robust statistics work). As an illustration, consider the schedule of potential outcomes in Table 1. In this toy experiment with four observations, unit 4 has unusually high potential outcomes compared to the others. Still, the true ATE is

$$\frac{1}{N}\sum_{i=1}^{N} Y_i(1) - Y_i(0) = 4/4 = 1 \tag{5}$$

Table 1: Toy example to illustrate the sensitivity of unbiased estimators to skewed outcomes

ID	$Y_i(0)$	$Y_i(1)$	$Y_i(1) - Y_i(0)$	Z_i	Z_i'
1	1	2	1	0	1
2	2	3	1	1	0
3	3	4	1	0	1
4	10	11	1	1	0

What happens when we estimate the true ATE ($\tau = 1$) with the unbiased estimator of the difference in means? Consider the two alternative treatment assignments listed in Table 1, Z_i and Z_i'. Under Z_i, the estimate is $(3+11)/2-(1+3)/2 = 5$. Under Z_i', it is $(2+4)/2-(2+10)/2 = -3$: in this small example estimates can vary considerably depending on whether we assign the high outlier to treatment or control.

In the presence of skewed outcomes or outcomes with a few extreme values, unbiased estimators can exhibit high variance in the sampling distribution of estimates they produce. A measure is robust if changes in a distribution have small effects on their value (Wilcox 2021). A mean has long been known to not be robust. Thus, the difference in means estimator is not a robust estimator of the ATE even if it is unbiased. The lack of robustness in the difference in means estimator can translate to a lack of precision — after all, as we just showed above, the values from the estimator can vary widely in the presence of an outlier.

To bring estimators closer to the truth, robust estimators use alternative measures to compute the ATE. These estimators increase precision by reducing dependency on extreme observations at the cost of unbiasedness. For example, one could still use the difference in means estimators

but after dropping the largest value of the outcome. Or one could replace that largest value with the next largest value. Neither of these estimators of the ATE are unbiased, but they are simple and might reduce the chance that a single estimator (the one reported by an analyst) might be quite far from the truth. For regression-based estimators, one could replace OLS with quantile regression (Koenker and Hallock 2001); estimating the difference in mean potential outcomes using differences in medians, for example. Again, this would be a biased estimator.

Here we explore the class of estimators known as M-estimators because they aim to directly estimate means (or differences of means) but allow the analyst to choose a function that downweights extreme points, and because there is a well-developed theory about how such functions perform (in large samples). The disadvantage of M-estimators is that one still needs to make decisions on the most appropriate objective function to penalize observations with high residuals. For example, a recent review identifies 48 different robust M-estimators for regression alone (Menezes et al. 2021). We choose to focus on M-estimation because they give researchers flexibility to specify the choice of robust estimator before observing the data. Moreover, M-estimators are directly motivated by cases in which means and OLS regression, both common approaches to estimate ATEs, yield biased yet imprecise estimates. Researchers can translate the ideas and simulation code that generates the results in this section to any other robust estimator that accomplishes the same goal.[8]

The theory of M-estimation arises from the theory of maximum-likelihood estimation (in regards their optimality properties). In this paper we use those ideas but do not specify likelihood functions. We do not need to have a correct likelihood-function (or any likelihood function) here: we are still working within the design-based statistical inference framework common in the analyis of randomized experiments and we will use simulation to discover whether these estimators improve our ability to learn from an experiment with large outliers and/or very skewed outcomes.

[8]Other robust estimators include replacing the mean with winsorized means, trimmed means, or the sample median (Wilcox 2021). For regression-based estimators, one could replace OLS with quantile regression (Koenker and Hallock 2001).

3.1 Setting

We focus on the estimation of a single mean to facilitate exposition although we will show estimates of differences of means below. We normally understand the mean as the expected value of a random variable $E[X]$, but more generally we can think of the expected value as single-number summary of the data. A good summary should be as close as possible to every observation in X.

Let $E[X-c]^2$ represent the expected square distance between each observation in X with respect to a measure of location c. We want the value of c that minimizes this expression. We can find this value by differentiating this function with respect to c and setting the result to zero, which leads to

$$E[X - c] = 0 \tag{6}$$

By definition, the unbiased estimator of the true mean, μ, such that $E[X] = \mu$, satisfies this condition by minimizing the sum of squared residuals. Any other measure would lead to a biased estimator. Yet, as illustrated above, the mean may be too sensitive to outliers. The logic behind M-estimation is to use a standard measure of location as a point of departure and then improve upon to it to satisfy some desirable properties. In this case, we want a measure that is as close as possible to the mean that is more robust to extreme observations. In other words, we want a (slightly) biased yet more precise estimator (and, even more than "precise" as in "low variance", we want "close to the truth").

Let $\xi(X - \mu_m)$ be a function that measures the expected distance from a measure of location μ_m. This is the objective function of the M-estimator. For example, for the case of the mean, $\xi(X - \mu_m) = (X - \mu_m)^2$. Which leads to the maximum likelihood estimator:

$$\Psi(X - \mu_m) = -2(X - \mu_m) \tag{7}$$

for which we find the optimal values of μ_m by setting $E[\Psi(X - \mu_m)] = 0$. The key is to find functions ξ and Ψ that lead to more robust estimates. This implies assigning different weights to observations depending on their distance to a measure of location. For example, the objective function of Huber's (1964) canonical M-estimator is:

$$\xi_H(x) = \begin{cases} \frac{1}{2}x^2 \text{ for } |x| \leq k \\ k|x| - \frac{1}{2}x^2 \text{ for } |x| > k \end{cases} \tag{8}$$

where k is a tuning constant. A rule of thumb is $k = 1.345\sigma$, where σ is the standard deviation of the distribution of initial squared distances. Other M-estimators have different default parameter values.

Beyond rules of thumb, choosing and tuning M-estimators involves striking a balance between breakdown and efficiency. The breakdown point is the proportion of unusual observations a sample can have before rendering estimates uninformative. The highest possible breakdown point is 0.5, this is because no more than half of the observations can be considered outliers. Efficiency relates to the relative asymptotic performance of M-estimators under ideal circumstances. The target is usually 95% asymptotic efficiency compared to a normal distribution.

Most M-estimators sacrifice one performance criterion at the expense of the other. For example, Huber's estimator has high efficiency but can easily break down with more than one outlier. Yohai (1987) proposes MM-estimation as a solution to achieve a high breakdown point and high efficiency. This procedure involves two steps. The first step finds initial values using a high breakdown point estimator, an S-estimator. The second step refines these values with a high efficiency M-estimator.

Since most M-estimators do not have an analytic solution, estimation requires an iterative reweighting procedure. The specifics of choosing the right objective functions are beyond the scope of this paper, but statistical software and textbooks often reduce this decision to a few choices

(Wilcox 2021).

3.2 Simulation

We simulate complete randomized experiments with 1,000 observations. The potential outcomes under control $\boldsymbol{Y(0)}$ are drawn from a contaminated normal distribution, so that individual i's potential outcome follows $N(0,1)$ with probability $(1-\alpha)$ or $N(0,\sigma^2)$ with probability α. We fix $\alpha = 0.1$ to reflect a setting with a small but non-negligible number of outliers. We consider values of $\sigma = (1, 5, 10, 20)$. Figure 2 shows individual draws from contaminated normal distributions with different σ. Larger values do not change the expect number of contaminated observations, but they increase the chances of drawing an observation with extreme values relative to the standard normal. We fix the true ATE $\tau = 0.5$.

For each experiment, we estimate treatment effects using the difference in means and two robust estimators estimated via iteratively re-weighted least squares. The first is Huber's (1964) canonical M-estimator outlined in the previous section. We also consider an MM-estimator using an S-estimator with a tuning constant $k = 1.548\sigma$ in the initial step to achieve a high breakdown point and a biweight estimator in the final step to enhance precision (Yohai 1987; Wilcox 2021).[9]

We again simulate 1,000 experiments for each parameter combination. Figure 3 reports the distribution of estimates across estimators against the true ATE. As in the previous section, the figure reflects the theoretical motivation for preferring M-estimators in a general case. When we consider a constant additive treatment effect, both the difference in means and robust estimators yield unbiased estimates, as they distributions center around the true ATE. Then, as σ increases, we see a wider distribution for the difference in means estimator, whereas both the M and MM-estimators preserve the same level of precision across values of σ.

This is only possible because in our simulations the probability of contamination is independent of treatment assignment and potential outcomes. In practice, one would expect extreme observations to also react differently to treatment. To capture this, TODO simulations with contamination

[9] These are the default parameters for the `rlm` function in the `MASS` R package.

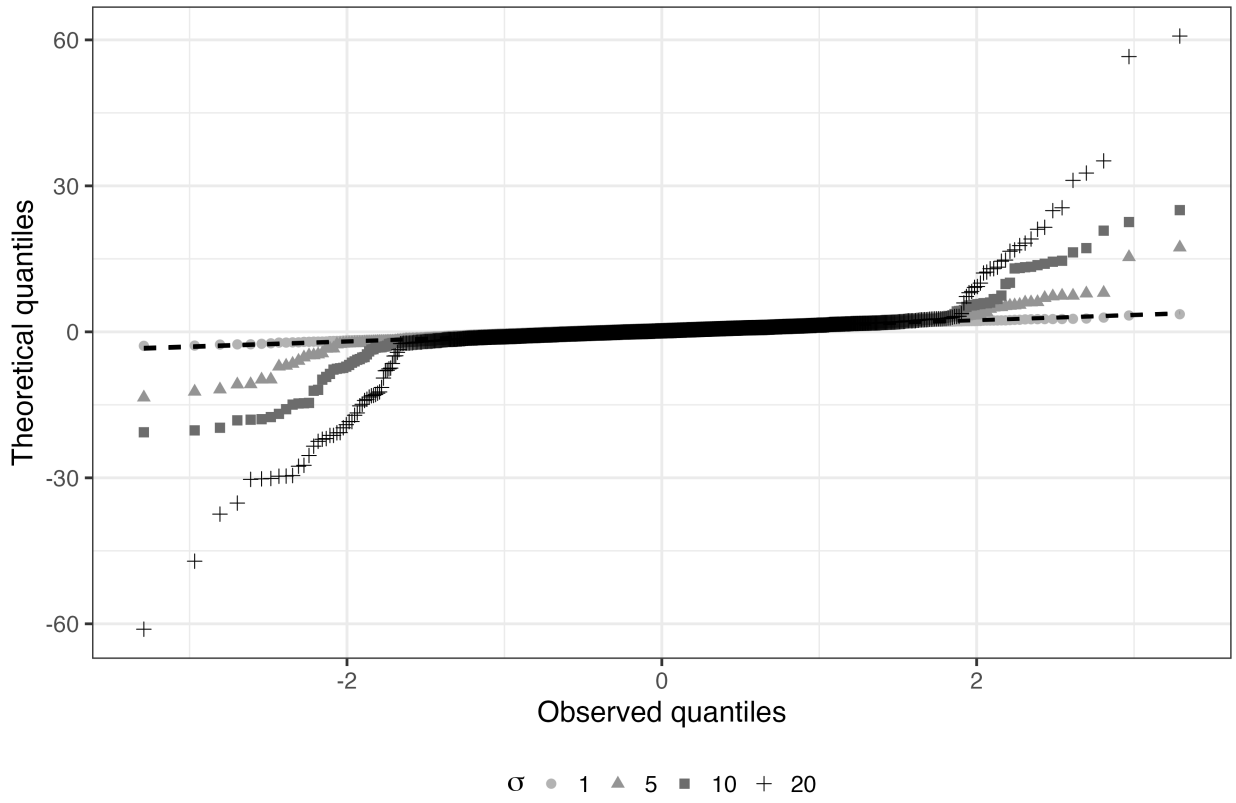

Figure 2: Individual draws from contaminated normal distributions.

that correlates with potential outcomes.

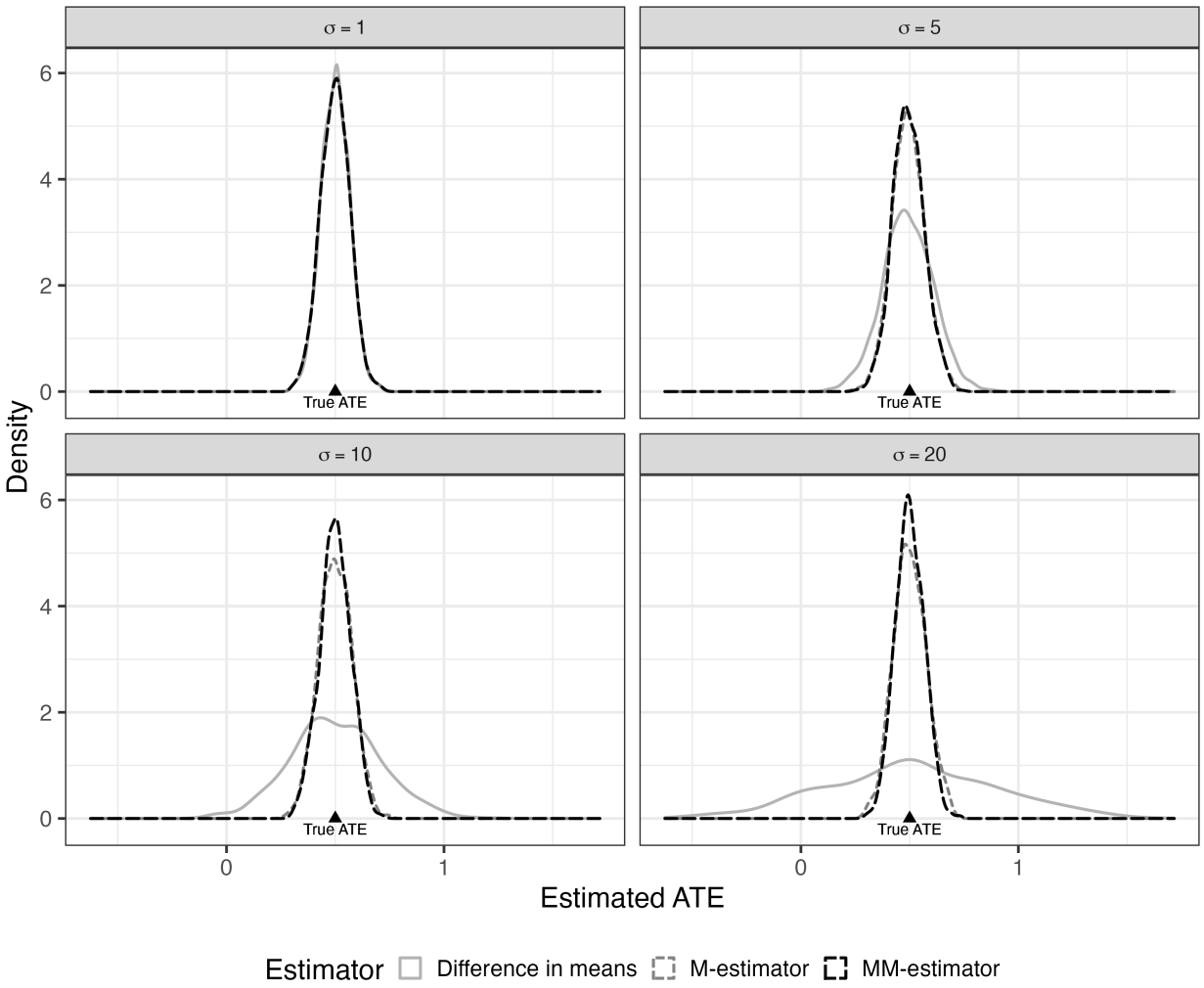

Figure 3: Estimates from simulated experiments with extreme observations in the outcome. Each panel is based on 1,000 simulations.

3.3 Summary of section

TODO after finding a case where M-estimators introduce bias that offsets precision gains. Right now the M-estimators just look uniformly and always better than the simple difference-of-means estimators.

4 Application to the GSA Auctions Experiment

In 2015 the General Services Administration designed and executed an experiment aiming to decrease the amount of unsold property that the U.S. government would have to manage and dispose of. The design involved assigning roughly 800,000 email messages to either a treatment

condition (a reminder that a person had bid on an item similar to the one for sale) or a control condition (no reminder) within about 7000 blocks or strata defined by the type of item. [10] We have simulated data to look like the outcome of this experiment and use the same configuration of block-sizes but we do not have access to the raw outcomes or individual level data from this experiment. In the actual experiment some of the blocks were very large (TODO) and some were very small (TODO), roughly half were assigned to treatment in every block, and outcomes ranged from $0 bid (in more than half of all blocks) to over $2,000,000 bid (for a helicopter), with a median bid of around $10. In the report on that experiment the OES team notes that a confidence interval for the ATE built using conventional methods of analyzing block-randomized experiments (the fixed-effects approach) yielded $p = .3$ for the test of the weak null hypothesis of no average effects using the difference-of-means as a test statistic with a 95% CI of [-$17,$67]. Here we compare the block-size weighting, precision-weighting, simple differencs of means, and M-estimated means, in regards their bias and mse in a version of this experiment simulated to look like the one here (with a true ATE of $20).

TODO

4.1 Simulations

4.2 Summary of section

5 Discussion and Conclusion

5.1 A note on testing versus estimation in randomized experiments

We have shown situations where the bias in "fixed effects" aka precision-weighted estimators and M-estimators is so large that any researcher would prefer to forgo any precision gains associated with those approaches. We mention testing here because many researchers prefer precise estimators to imprecise estimators in order to construct p-values for tests of the weak null of no average effect of the treatment (or equivalently, to create confidence intervals to be interpreted only in regards whether the confidence interval includes 0). In this case, where precision is desired for

[10] Details of the experiment can be found here. In fact those 800,000 messages were sent to 37,000 users across 7100 auction lots. We ignore the dependence across lots by user here since are using simulated data for confidentiality reasons.

the purposes of testing (or creating confidence intervals), researchers may prefer a hybrid approach rather than using a biased estimator to *both* estimate the ATE *and* act as a test statistic for a test of the weak null of no average effects. In fact, randomization-based tests of the sharp null hypothesis of no effects are valid under a way array of test-statistics.[11] Thus, researchers do not really have to make a choice between bias and statistical power. A researcher can report an unbiased estimate of the ATE and then a p-value from a test with high statistical power and correct false positive rate, each arising from a different calculation: the estimator of the ATE using, say, block-size weights and the test of a sharp hypothesis of no effects using precision-weights and a randomization justified reference distribution.

We included tests of this sort in all of the simulations done above. We here show that those tests using precision weighted test statistics have more statistical power than block-size weighted tests of the weak null even when the bias in the precision weighted estimators is high. Thus, while we have written throughout this paper about the bias versus precision trade-off from the perspective of reporting a point estimate that is close to the truth, we note here that such a trade off may not in fact be necessary if a unbiased-but-possibly-far-from-the-truth estimate is required and if a randomization-based test of the sharp null can be substituted for the randomization-based test of the weak null.

5.2 Conclusion

TODO

References

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Declaring and Diagnosing Research Designs." *American Political Science Review* 113 (3): 838–59.

Bowers, Jake. 2011. "Making Effects Manifest in Randomized Experiments." In *Cambridge Handbook of Experimental Political Science*, edited by James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia. New York, NY: Cambridge University Press.

[11] Bowers and Leavitt (2020) show how randomization-based tests of the sharp null have this property (while randomization-based tests of the weak null of no average effects using the difference-in-means test statistics requires that test statistic arise from an unbiased estimator of the ATE).

Bowers, Jake, and Thomas Leavitt. 2020. "Causality & Design-Based Inference." In *The SAGE Handbook of Research Methods in Political Science and International Relations*, edited by Luigi Curini and Robert Franzese. Sage Publications Ltd.

Freedman, David A. 2008. "On regression adjustments to experimental data." *Advances in Applied Mathematics* 40 (2): 180–93.

Gerber, Alan S, and Donald P Green. 2012. *Field experiments: Design, analysis, and interpretation*. WW Norton.

Gibbons, Charles E., Juan Carlos Suárez Serrato, and Michael B. Urbancic. 2018. "Broken or Fixed Effects?" *Journal of Econometric Methods* 8 (1). https://doi.org/10.1515/jem-2017-0002.

Green, Donald P. 2009. "Regression Adjustments to Experimental Data: Do David Freedman's Concerns Apply to Political Science." In *26th Annual Meeting of the Society for Political Methodology, Yale University, July*, 23–25. Citeseer.

Hansen, B. B., and J. Bowers. 2008. "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." *Statistical Science* 23: 219.

Huber, Peter J. 1964. "Robust Estimation of a Location Parameter." *The Annals of Mathematical Statistics* 35 (1): 73–101. https://doi.org/10.1214/aoms/1177703732.

Humphreys, Macartan. 2009. "Bounds on Least Squares Estimates of Causal Effects in the Presence of Heterogeneous Assignment Probabilities." *Manuscript, Columbia University*.

Imai, Kosuke, Gary King, and Elizabeth A. Stuart. 2008. "Misunderstandings Between Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171 (2): 481–502. https://doi.org/10.1111/j.1467-985x.2007.00527.x.

Kalton, G. 1968. "Standardization: A Technique to Control for Extraneous Variables." *Applied Statistics* 17: 118–36.

Koenker, Roger, and Kevin F Hallock. 2001. "Quantile Regression." *Journal of Economic Perspectives* 15 (4): 143–56. https://doi.org/10.1257/jep.15.4.143.

Lin, Winston. 2013. "Agnostic notes on regression adjustments to experimental data: Reexamin-

ing Freedman's critique." *The Annals of Applied Statistics* 7 (1): 295–318.

Menezes, D. Q. F. de, D. M. Prata, A. R. Secchi, and J. C. Pinto. 2021. "A Review on Robust m-Estimators for Regression Analysis." *Computers & Chemical Engineering* 147 (April): 107254. https://doi.org/10.1016/j.compchemeng.2021.107254.

Pashley, Nicole E., and Luke W. Miratrix. 2020. "Insights on Variance Estimation for Blocked and Matched Pairs Designs." *Journal of Educational and Behavioral Statistics* 46 (3): 271–96. https://doi.org/10.3102/1076998620946272.

———. 2021. "Block What You Can, Except When You Shouldn't." *Journal of Educational and Behavioral Statistics*, July, 107699862110272. https://doi.org/10.3102/10769986211027240.

Schochet, Peter Z. 2010. "Is Regression Adjustment Supported by the Neyman Model for Causal Inference?" *Journal of Statistical Planning and Inference* 140 (1): 246–59.

Wilcox, Rand R. 2021. *Introduction to Robust Estimation and Hypothesis Testing*. Elsevier. https://doi.org/10.1016/c2019-0-01225-3.

Yohai, Victor J. 1987. "High Breakdown-Point and High Efficiency Robust Estimates for Regression." *The Annals of Statistics* 15 (2). https://doi.org/10.1214/aos/1176350366.