A Comparison of Causal Inference Methods for Estimating Sales Lift

Mike Hankin, Google
David Chan, Google
Michael Perry, Google

September 29, 2020

Abstract

In this paper, we compare a variety of methods for causal inference through simulation, examining their sensitivity to and asymptotic behavior in the presence of correlation between (heterogeneous) treatment effect size and propensity to be treated, as well as their robustness to model mis-specification. We limit our focus to well-established methods relevant to the estimation of sales lift, which initially motivated this paper and serves as an illustrative example throughout. We demonstrate that popular matching methods often fail to adequately debias lift estimates, and that even doubly robust estimators, when naively implemented, fail to deliver statistically valid confidence intervals. The culprit is inadequate standard error estimators, which often yield insufficient confidence interval coverage because they fail to take into account uncertainty at early stages of the causal model. As an alternative, we discuss a more reliable approach: the use of a doubly robust point estimator with a sandwich standard error estimator.

1 Introduction

Causal inference methods are tools for measuring the effect of an intervention based on observational or experimental data. Their purpose is more complicated than that of descriptive statistics: they are intended to infer information not about the existing, realized state of the world, but an alternative, unobserved state. Decisionmakers across a wide variety of fields use the insights generated by these methods because they measure the incremental impact of an intervention, making them more useful than associative measurements.

1.1 Initial Motivation: Estimating Sales Lift

The motivation for this paper is the complex problem of the measurement of offline sales lift due to online advertising. In particular, estimating the increase in a household's offline purchasing due to exposure of a household member to online advertising for the brand.

If all people were influenced by all ads equally, ad targeting would not be necessary. Of course, this is not true, and advertisers generally target people who they believe would be most influenced by an ad. For example, a maternity clothes retailer may target women of child-bearing age over other groups of people because it anticipates that these women are more likely to be influenced by an ad for maternity clothes. This rational behavior by advertisers means that those who are most likely to be influenced by the ad are more likely to be exposed to it.

The sales lift due to advertising can't be obtained by simply comparing those who were exposed, to those who weren't, as these two groups are inherently different. But this view of ad targeting also suggests a more subtle concept: the correlation between the magnitude of the treatment effect on an individual, and that individual's propensity to be exposed to the treatment. This connection between the targeting and the individual's change in behavior induced by ad exposure will feature prominently in our simulations and analysis.

The methods used to extract causal effect measurements from observational data vary by field. In this paper, we limit our examination to the methods of causal inference used to estimate the causal impact of ad exposure on a household's tendency to purchase a product. We also omit discussion of methods like Media Mix Models that measure effectiveness of advertising spend on an aggregate level, as the challenges associated with using such models to extract causal estimates runs much deeper than statistical validity [5].

1.2 Overview

This paper takes the following form: in Section 2 we introduce the fundamentals of causal inference along with a motivating toy example; in Section 3, we introduce the methods of causal inference examined in our simulations; in Section 4 we describe the data generation scenarios employed in our simulations, and consider their implications and relationship to real world problems; in Section 5 we present the results of our simulations; and in Section 6 we present our analysis of those results, and summarize our comparative findings.

2 Causal Inference

Here we introduce a toy example of the problem at hand, set the notation required to formally discuss causal inference, articulate the fundamental problem, and present the assumptions necessary for inference to be possible.

2.1 Toy Example

To illustrate our problem, consider a toy example where the causal graph in Figure 1, as defined by Pearl [18, 22], represents the true causal process. Here, we consider a situation in which an advertising campaign for "Sugar Inferno" candy is primarily targeted at male YouTube users because their manufacturer erroneously believes that male users will be more responsive to the ad campaign. Clearly, a user's chance of being exposed to the ad increases the more time they spend on the platform. Additionally, their familiarity with the platform and the regularity with which they're exposed to online video advertising will likely affect the magnitude of the ad's effect on them, inducing a heterogeneous treatment effect [7]. Finally, their pre-existing preference for sweet treats will directly affect their outcomes, possibly in a heterogeneous manner.

In our simplistic example, the ad serving system does not have any way of estimating their predilection for sweets, so that preference cannot directly affect their probability of being exposed to the ad. We also assume that gender has no direct impact on sales, except through ad exposure, or on their taste for candy. Our population is all YouTube users, and we assume no network effects, which is to say that each user's ad exposure status and potential purchase outcomes are independent of those of all other users.

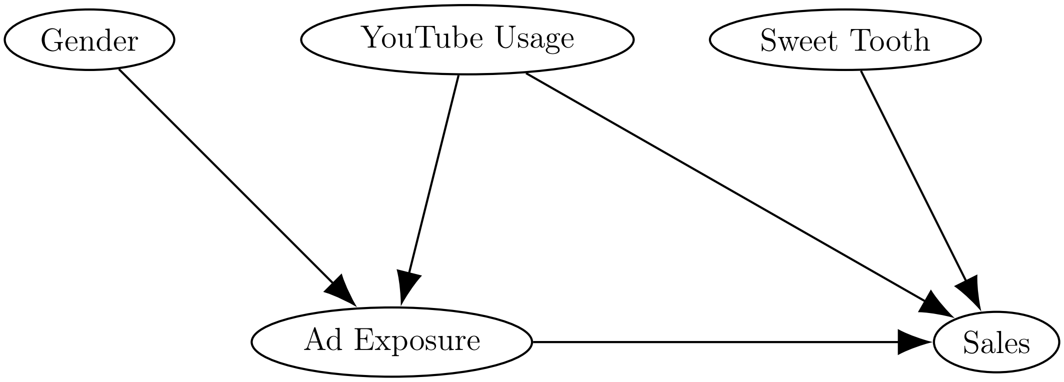


Figure 1: Sugar Inferno toy example generating directed acyclic graph (DAG).

2.2 Notation

Taking treatment to mean exposure to the ad campaign, first define X to be a vector of pre-treatment covariates (X_i the covariates for the ith unit), T to be a binary treatment indicator (T_i the treatment indicator for the ith unit), $Y(0)$ and $Y(1)$ to be the untreated and treated potential outcomes ($Y_i(0)$ and $Y_i(1)$) and Y with no argument to be the observed outcome, equivalent to $Y(T)$ ($Y_i = Y_i(T_i)$).

In the toy example, X_i would contain the user's gender, a measurement of their pre-treatment YouTube usage, and (in a perfect world for the measurer) a quantification of their fondness for sweets. T_i is a binary representation of their ad exposure status[1]. $Y_i(0)$ is their post-campaign period spend had they not been exposed to the ad, while $Y_i(1)$ is their post-campaign period spend in the case where they had been exposed; Y_i without a treatment argument represents the amount of money they actually spent on Sugar Inferno in the post-campaign period and can be computed from the treatment indicator and potential outcomes as $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$.

We will use $b(X) = P(T = 1 \mid X)$ to denote a user's propensity to be treated, conditional upon their covariates, and reference some modeled approximation of this value as $\hat{b}(X) = b\left(X; \hat{\beta}\right)$ which we refer to as the propensity score.

2.3 Basic Problem Statement

Causal inference is the process of estimating or understanding the causal impact of an action on an outcome. In our toy example, the goal is to assess the average change in purchasing behavior induced by exposure to a Sugar Inferno advertisement.

In the case of a binary (yes/no) action, a naive approach would be to take the difference of the average outcomes of the treated and untreated populations. However doing so yields a biased estimate, as demonstrated by the fundamental identity of causal inference [24]:

[1] For our toy example, we ignore multiple exposures. A user who has been exposed to one or more ads is counted as treated.

$$
\begin{aligned}
\mathbf{E}\left[Y\left(1\right) \mid T=1\right]-\mathbf{E}\left[Y\left(0\right) \mid T=0\right] &= \left(\mathbf{E}\left[Y\left(1\right) \mid T=1\right]-\mathbf{E}\left[Y\left(0\right) \mid T=1\right]\right) \\
&\quad + \left(\mathbf{E}\left[Y\left(0\right) \mid T=1\right]-\mathbf{E}\left[Y\left(0\right) \mid T=0\right]\right) \\
&= \tau_{\text{ATT}} + \text{bias}_{\text{Selection}}
\end{aligned}
$$

where τ_{ATT} is the treatment effect on the treated (see Section 3.1) and $\text{bias}_{\text{Selection}}$ captures the bias due to differences in potential outcomes between the treated and untreated population. The goal of causal methods is to remove that bias term from the impact estimate, and to quantify the uncertainty of the resulting estimate. This is generally accomplished by balancing the treated and control groups based on some pre-treatment measurements, thereby comparing the treated users to a group of "look-a-like" untreated users.

2.4 Assumptions

We require the standard assumptions of causal inference [14] hold. Without these assumptions, causal inference becomes either impossible or so complex (requiring convoluted dependence assumptions) as to be essentially impossible. They are:

- Probabilistic treatment assignment: every experimental unit has a non-zero chance of receiving (or not receiving) the treatment.
  - $0 < b\left(X\right) < 1$ almost surely (i.e. $P\left(0 < b\left(X\right) < 1\right) = 1$)
  - In the toy example, while male users are more likely to be exposed, the targeting scheme also serves ads to female users, albeit at a lower rate.
- Stable unit treatment value assumption (SUTVA): for each pair of experimental units, the potential outcomes of one are independent of the treatment assignment of the other.
  - $\forall i \neq j \quad Y_i\left(0\right), Y_i\left(1\right) \perp T_j$
  - In the toy example, user i's purchasing rate is independent of user j's ad exposure status. Our assumption of the absence of network effects (one user's exposure doesn't affect other users' purchasing) grants us SUTVA.
- Individualistic Assignment: The probability of unit i being assigned to the treatment group depends only on its own covariates and potential outcomes, not that of any other unit.
  - $\exists q : \mathbb{X} \times \mathbb{Y} \times \mathbb{Y} \to [0,1]$ such that $P\left(T_i = 1 \mid X_j, Y_j\left(0\right), Y_j\left(1\right) \, \forall j\right) = q\left(X_i, Y_i\left(0\right), Y_i\left(1\right)\right)$
  - In the toy example, user i's probability of being exposed to the ad depends only on their gender and YouTube usage; it is independent of any information (pre- or post-treatment) related to any other user.
- Unconfoundedness: The probability of unit i being assigned to the treatment group does not depend on any unit's potential outcomes, including its own.
  - $P\left(T_i = 1 \mid X_j, Y_j\left(0\right), Y_j\left(1\right) \, \forall j\right) = P\left(T_i = 1 \mid X_j \, \forall j\right)$

  - Combining this with individualistic assignment reduces to a dependence of a unit's probability of receiving treatment depending on its covariates alone, i.e.

$$\exists q : \mathbb{X} \to [0,1] \text{ such that } P\left(T_i = 1 \mid X_j, Y_j\left(0\right), Y_j\left(1\right) \forall j\right) = q\left(X_i\right)$$

  - In the toy example, the targeting scheme doesn't have any information about a user's post-campaign period potential purchasing rates that isn't captured in X.

- Conditional Ignorability: given the pre-treatment covariates X, the treatment assignment is independent of the potential outcomes.
  - $Y\left(1\right), Y\left(0\right) \perp T \mid X$
  - In the toy example, given their gender, YouTube usage, and preference for sweets, the quantity of Sugar Inferno a given user would purchase in the measured period if they had been treated (or if they hadn't) is independent of whether or not they actually were treated. In other words, once we've conditioned on the covariates, knowing the treatment status tells us nothing about the user's treated or untreated purchasing behavior. Violations of this assumption arise when there are unobserved variables that affect both the user's purchase behavior and their likelihood to see an ad. For example we are unlikely to be able to observe "predilection for sweets" but ad-serving might have proxies for that in its targeting (such as previous browsing behavior) and it probably affects purchase behavior. A model without a measure of "predilection for sweets" violates this assumption.

The final assumption, Conditional Ignorability, is commonly adopted but rarely actually satisfied. It is a demanding assumption, requiring that all relevant pre-treatment data is available to the analyst. Omission of any information relevant to both the treatment assignment mechanism and the potential outcomes in an estimation procedure will almost certainly incur bias. In our toy example, it is unlikely that the investigator performing the measurement would know about the users' preferences for sweets. We do not examine the effects of violating this assumption in this paper although, it is a significant risk.

When the assumptions above are satisfied, the practitioner has a wide array of methodologies at their disposal.

3 Estimation Methods

The goal of the methods introduced here is to accurately estimate the treatment effect given some observational data of the form $\left(\left(X_1, T_1, Y_1\right), ..., \left(X_n, T_n, Y_n\right)\right)$, and to quantify the uncertainty of that estimate. In this section, we clarify the different estimands that are the targets of these causal methods. We then describe a variety of relevant families of such methods, and finally we introduce the sandwich estimator, a scheme for robust uncertainty quantification.

3.1 Estimands

In the following discussion, we focus on the estimand Average Treatment Effect on the Treated

$$\tau = E[Y\left(1\right) - Y\left(0\right) \mid T = 1] \quad \text{(ATT)} \tag{1}$$

as opposed to Average Treatment Effect

$$\mathbf{E}\left[Y\left(1\right)-Y\left(0\right)\right] \quad \text{(ATE)} \tag{2}$$

except where the latter makes for a more comprehensible introduction. These two quantities may differ when the treatment effect is heterogeneous, and that heterogeneity is correlated with the treatment assignment mechanism. In the toy example, ATT is the effect that ad exposure had on those YouTube users who were actually exposed to the ad, whereas ATE is the average effect ad exposure would have had on the entire population of YouTube users.

In addition to point estimates, decision-makers want some quantification of the uncertainty of the estimate. This often takes the form of a confidence interval (CI), a range of values that is likely to contain the true effect size. CIs will be discussed in greater depth in Section 5.2, but for now suffice it to say that they are usually calculated using both the point estimate and the estimate for the standard deviation of the distribution of the estimator. The latter of these quantities is commonly referred to as the standard error (SE), and it will feature prominently in the rest of this paper.

3.2 Overview of Methods and Estimators

In this section, we introduce a number of the standard methods of causal inference. First, we describe matching methods in which treated units are paired with similar untreated units (generally discarding unpaired units). This matching decorrelates treatment assignment from the covariates, allowing effect estimation by application of simple methods to the matched data. After introducing this class of methods and surveying two relevant examples, we discuss their computational characteristics and uncertainty quantification.

Next, we introduce doubly-robust augmented inverse propensity weighting (AIPW) by first describing its simpler inspiration, inverse propensity weighting (IPW). In IPW, observations are reweighted based on an estimated propensity score, which approximates the assignment mechanism as a function of the covariates. AIPW extends this by including an outcome model, and is the first example of a wider class of "doubly-robust" (DR) methods that employ estimates of both the treatment assignment mechanism and the outcome model. These methods are consistent when either the outcome or propensity submodel is correctly specified. We briefly survey the wider class of DR estimators, and consider the possibility of DR matching.

Throughout this subsection, we discuss standard errors (SE) for the various estimators. We conclude with a deeper examination of the issue of SE estimators for DR models, and their tendency to squander the double robustness of the point estimator. To remedy this, we describe the sandwich estimator for robust SEs.

3.2.1 Matched ANCOVA

Matched ANCOVA (Analysis of Covariance) methods are two-stage methods in which the first stage (matching) seeks to decorrelate the treatment assignment from the covariates, thereby removing selection bias, and the second stage (ANCOVA) performs the actual treatment effect estimation in addition to any further debiasing required due to residual treatment-covariate correlation. ANCOVA estimates differences between groups while controlling for covariates using a linear model. The model is easy to use but requires strong assumptions: that the outcome is a linear function of the covariates, that the treatment effect is homogeneous and additive, and that the error terms are normally distributed, identically distributed, and pairwise independent. In practice these assumptions are unlikely to be satisfied. For a detailed discussion of Matched ANCOVA, see Schafer and Kang [19].

Matching Methods All matching methods are designed to establish covariate balance across the treated and control populations, decorrelating the treatment from the covariates. The most common matching methods, on which we will focus, do so by establishing a pseudo-metric[2] used to find similar units. For every treated unit, the matching scheme will pair the M distinct, valid control units nearest to them according to the pseudo-metric for some positive integer M. Once applied to a match, a control unit may be taken out of consideration for future matches (known as matching without replacement), or left available to future matches to augment its weight in the post-matched dataset (matching with replacement). An upper limit on the maximum pseudo-metric distance between matched units may be used. How to handle treatment units without enough matches is a choice for the modeler, but often they are discarded which affects the legitimacy of the estimate.

Propensity Score Matching (PSM) Propensity distance, $|\hat{b}\left(X_i\right) - \hat{b}\left(X_j\right)|$, is a canonical pseudo-metric for matching. It is motivated by the conditional independence of the covariates and treatment assignment given the propensity score: $X \perp T \mid b\left(X\right)$, as well as the conditional independence of the potential outcomes and the treatment assignment given the propensity score: $Y\left(1\right), Y\left(0\right) \perp T \mid b\left(X\right)$ [14]. By reducing the matching to a single dimension, PSM avoids the failed matches that commonly afflict matching methods for high dimensional data, at the cost of specifying a form for the propensity model.

PSM increases covariate balance, particularly in severely unbalanced datasets. However, see King [15] for some of the potential pitfalls of PSM on datasets that are already fairly well-balanced across assignment groups.

Coarsened Exact Matching (CEM) In CEM, individual continuous covariates are first transformed into categorical variables through coarsening. With all covariates now of finite cardinality, exact matches may be found for treated units. Usually, the coarsening scheme is iterative: a coarsening function is applied to each covariate, matching is attempted, and if it fails to match a sufficient portion of the incoming units, further coarsening may be applied; see Imbens [14] for details. Because the implicit pseudo-metric is discrete, it may become difficult to find an adequate number of matching units as the dimensionality of the data increases.

CEM is a commonly used in a variety of fields, including sales lift measurement. For example, CEM was used in a 2019 study cited in Newsweek [10] examining the effect of a vegetarian diet on risk of cardiovascular events [25], an economic comparison of organic and conventional dairy farms [9], and meta-analysis of the effect of marriage on dementia risk [21].

Unmatched If one believes the groups are already sufficiently balanced or that the treatment assignment mechanism is fully random, then no matching is required. Depending on the matching scheme, avoiding any matching may yield a larger dataset for estimation, trading increased bias for a reduction in variance.

Effect Estimation After matching, an ANCOVA model of the form specified in Equation 3 is used to model the observed outcome y as a function of covariates z and treatment indicator t. The set of covariates z of dimension p doesn't necessarily have to be the same set of covariates on which matching is performed.

[2] A binary operator satisfying all the requirements of a metric except that for pseudo-metrics $d\left(x, y\right) = 0 \not\Rightarrow x = y$

$$y_i = \alpha + z_i'\beta + t_i\tau + \epsilon_i \tag{3}$$

$$\epsilon_i \sim N\left(0, \sigma_{\text{err}}^2\right) \quad \text{iid} \quad \forall i = 1, ..., n$$

The treatment effect τ is given as the coefficient of the treatment indicator and the parameters of the model are estimated via standard regression techniques.

$$\frac{\hat{\tau} - \tau}{\hat{\sigma}_\tau / \sqrt{n - p - 1}} \sim \text{Student's t}\left(n - p - 1\right) \tag{4}$$

Equation 4 provides a formula for defining a confidence interval for τ, where $\hat{}$s imply estimates, and $\hat{\sigma}_\tau$ is the standard error of the treatment effect estimator, which can be obtained using simple robust SE estimators such as the Huber-White heteroskedasticity robust estimator [27], itself a special case of the sandwich estimator. For reasonably large n, the Student's t distribution will converge to a standard normal.

There exist a variety of approaches to estimating SEs for ANCOVA when modeling assumptions are violated. However, while such SE estimators may increase the accuracy of the estimate for the data provided to the ANCOVA model, they will fail to take into account the uncertainty in the matching and almost certainly under-estimate the overall SE. There is no general agreement on how to handle SEs for matching methods [12]. There have been attempts made to account for uncertainty introduced in the matching process itself when the effect estimation is performed via ANCOVA (see [2]), but these SE estimators are not easily generalized to more complex effect estimation schemes and fail to account for uncertainty in the steps preceding the matching. For instance, training of the propensity model in the case of propensity matching.

The covariate balancing approach used prior to ANCOVA estimation determines the interpretation of that estimated coefficient. When all treated units are included, with a fixed number of control matches for each of them, the estimand is ATT.

Estimation Within Strata In a variant of matched ANCOVA, practitioners use the matching scheme to build blocks or strata of units in such a manner as to resemble a block randomized experiment. For each stratum, they perform an in-strata treatment effect estimation, often via ANCOVA, because matching does not fully eliminate covariate imbalance but the units are close enough that the outcome should be locally well predicted by a linear function of the covariates and treatment indicator. They then compute the overall treatment effect estimate by computing a weighted average of the strata treatment effects. This allows for a less biased estimate of the effect size when the outcome function is smooth but non-linear, however it makes accurate standard errors very difficult to estimate. See King [16] for further details.

This technique is particularly applicable to coarsened exact matching. In the toy example, we would calculate that treated user's Individual Treatment Effect (ITE) by taking the difference of their outcome from the mean or median of their matched controls' outcomes; ATT would be calculated by averaging the ITEs over all of the treated users.

Computational Burden Matching algorithms require either a nearest-neighbor search, which is computationally expensive, or an approximate nearest-neighbor search which adds unmeasured uncertainty to the procedure. Propensity score matching reduces the dimensionality on which matching is performed to 1, greatly lessening the burden, though it does require an extra model fitting stage.

3.2.2 Doubly Robust Methods

Doubly robust methods estimate both a propensity model ($\hat{b}(x)$ an estimate for $\mathbf{E}[T \mid X = x]$) and an outcome model ($\hat{m}(x, t)$ an estimate for $\mathbf{E}[Y(t) \mid X = x]$), and are so named because they remain consistent (see Equation 5)

$$\forall \epsilon > 0, \quad \lim_{n \to \infty} \mathrm{P}\left(\mid \hat{\theta}(X_1, ..., X_n) - \theta_0 \mid > \epsilon\right) = 0 \tag{5}$$

if either one of the models is correctly specified [20]. It is important to note that, while the treatment effect estimates may be DR, their basic standard error estimators are not, and tend to result in lower confidence interval coverage when one of the models is misspecified.

Inverse Propensity Weighting Though not a DR estimator, we describe the IPW estimator here to introduce the reader to AIPW.

IPW is itself the difference of two estimators, the average treated potential outcome and the average control potential outcome. We elaborate on this by first introducing these estimators for the simpler ATE variant, then modifying them to produce the ATT variant. The most common average potential outcome estimator employed in IPW is the Horvitz-Thompson (HT) estimator [13]. The HT estimator for the treated potential outcome takes the form

$$\hat{\mu}_{\text{Treated}} = \frac{1}{N} \sum_{i=1}^{N} \frac{T_i \cdot Y_i}{\hat{b}(X_i)}$$

and arises from the fact that the individual terms have expectation equal to the expected treated outcome

$$\mathbf{E}\left[\frac{T \cdot Y}{b(X)}\right] = \mathbf{E}[Y(1)].$$

Estimation of the average control potential outcome is handled similarly. Combining the two, we can thus estimate ATE by

$$\hat{\tau}_{\text{ATE}} = \frac{1}{N} \sum_{i=1}^{N} \frac{T_i \cdot Y_i}{\hat{b}(X_i)} - \frac{1}{N} \sum_{i=1}^{N} \frac{(1 - T_i) \cdot Y_i}{1 - \hat{b}(X_i)}.$$

When estimating ATT, the averages must be taken over the population of treated units. For the treated potential outcome, this simplifies the estimator:

$$\hat{\mu}_{\text{Treated}|T=1} = \frac{\sum_{i=1}^{N} T_i \cdot Y_i}{\sum T_i}.$$

However, it complicates the average untreated potential outcome estimator, as it amount to a reweighting:

$$O_{b,i} = \frac{\hat{b}(X_i)}{1 - \hat{b}(X_i)}$$

$$\hat{\mu}_{\text{Control}|T=1} = \left(\frac{1}{N} \sum (1 - T_i) \cdot Y_i \cdot O_{b,i}\right) / \left(\frac{\sum \hat{b}(X_i)}{N}\right) = \frac{\sum (1 - T_i) \cdot Y_i \cdot O_{b,i}}{\sum \hat{b}(X_i)}.$$

From the two average potential outcome HT estimators over the treated units, we obtain the IPW ATT estimate:

$$\hat{\tau}_{\text{ATT}} = \frac{\sum T_i \cdot Y_i}{\sum T_i} - \frac{\sum (1 - T_i) \cdot Y_i \cdot O_{b,i}}{\sum \hat{b}(X_i)}.$$

The propensity score must be correctly specified for either estimand to be unbiased because in the expectation calculation we treat is as the true propensity function.

Doubly-Robust Augmented Inverse Propensity Weighting AIPW [11] is a weighted sum of terms of the form

$$\begin{aligned}\hat{\tau}_i &= \frac{T_i \cdot Y_i - \hat{m}(X_i, 1) \cdot \left(T_i - \hat{b}(X_i)\right)}{\hat{b}(X_i)} - \frac{(1 - T_i) \cdot Y_i + \hat{m}(X_i, 0) \cdot \left(T_i - \hat{b}(X_i)\right)}{1 - \hat{b}(X_i)} \\ &= \frac{T_i \cdot (Y_i - \hat{m}(X_i, 1))}{\hat{b}(X_i)} - \frac{(1 - T_i) \cdot (Y_i - \hat{m}(X_i, 0))}{1 - \hat{b}(X_i)} + \hat{m}(X_i, 1) - \hat{m}(X_i, 0).\end{aligned}$$

taking $\hat{m}$ to approximate the conditional expected outcome function $m(x, t) = \mathbf{E}[Y \mid X = x, T = t]$. This can be interpreted as adjusting the estimate produced by the basic outcome model by an IPW estimate of the outcome model's errors.

The general form of the estimator is

$$\hat{\tau} = \sum_{i}^{N} \hat{w}_i \hat{\tau}_i.$$

where $\hat{w}_i = 1/N$ for ATE and $\hat{w}_i = \frac{\hat{b}(X_i)}{\sum_{j=1}^{N} \hat{b}(X_j)}$ for ATT.

Alternative Doubly-Robust Estimators AIPW is just one example of the larger class of DR estimators. We focus on it in this paper because it is the easiest to understand and implement. For interested readers, we suggest investigating Targeted Maximum Likelihood Estimation [26], Entropy Balancing [28], and Double Machine Learning [6] as other examples of DR estimators.

It is important to note that any blackbox classifier and regressor, respectively, can be used for the propensity and outcome models in all of these DR methods.

Is PSM ANCOVA Doubly Robust? For the ANCOVA model to be correctly specified, we require that

1. the true, generating outcome model is linear in the covariates,
2. the treatment effect is homogeneous and additive,
3. and the error noise is homoskedastic, uncorrelated with the covariates or treatment assignment, and Gaussian.

In the unlikely event that all of those assumptions are satisfied, then regardless of any matching, its estimate should be consistent as long as the matching scheme leaves enough data points for estimation. Alternatively, if the propensity model is perfectly specified and has a consistent estimator, the covariates in the matched data should be asymptotically independent of the treatment assignment. However, if the outcome model is misspecified, it is unclear what conditions are necessary to still achieve a consistent estimator.

If PSM ANCOVA is DR, it has yet to be proven. Even if verifiable conditions for double robustness could be found, there is no known robust SE estimator that would account for the stochasticity in propensity model training and the matching scheme, yielding invalid confidence intervals. Abadie and Imbens discuss conditions for consistency and the difficulty of extracting accurate standard errors [1]. Antonelli et al introduce a modified propensity-and-outcome matched estimator that exhibits DR properties [4], but it does not present a clear path to robust standard errors and is computationally less attractive.

3.3 Sandwich Estimator

The sandwich estimator is a framework for developing SE estimators with desirable robustness properties [23, 8, 3]. Namely, consistency of the SE, even if the underlying model is misspecified. In our application, if we employ a DR estimator with at least one of the submodels correctly specified then the sandwich estimator will asymptotically yield a confidence interval with correct coverage for our estimand. This extension of double-robustness to the area of confidence intervals significantly enhances the usefulness of these estimators. Without such a robust SE estimator, the double robustness property would be squandered in that any time one of the submodels was misspecified, the SE could no longer be trusted.

The sandwich estimator cannot be used with a discrete matching process because it requires differentiability with respect to all of its parameters. Similarly, it could not be applied to a causal method with a non-smooth submodel (ex: AIPW with a random forest propensity model). For details on the sandwich estimator in general see Stefanski [23]. See Section A.1 for a write up of the derivation of the sandwich estimator for an AIPW estimator.

4 Data Generation Scenarios

We use simulated data to highlight the areas where each estimator does well or performs poorly. We begin by describing the basic data generating process, then explain the two scenarios examined in this paper. The first varies the sample size, holding the data generating process constant, and examines asymptotic behavior. The second holds the sample size constant, but varies a parameter of the data generating process that controls how much the treatment effect size is correlated with the propensity to be treated, allowing for comparisons of robustness. This correlation of propensity to be exposed and treatment effect size is central to our exploration, in that it emulates the effect of ad targeting.

4.1 Data Generating Scheme

Our data points take the form of tuples of covariates, treatment assignments, and observed outcomes: (X, T, Y). For each sample, we first generate the covariates X, then the treatment assignment T dependent on X, followed by the expected potential outcomes conditioned on X, and finally sample the observed outcome given the expected potential outcomes and T.

The covariates are drawn from a 5 dimensional Gaussian distribution with a low (but non-zero) degree of correlation between them, and homogeneous variances:

$$X \sim \mathrm{N}\left(\mathbf{0}, \Sigma\right)$$

$$\Sigma_{i,j} = \sigma_X^2 \rho_X^{|i-j|}$$

The probability of a unit receiving treatment is generated using a logistic model designed such that, given the covariate data generating process (DGP), the unconditional probability of receiving treatment is approximately 10%. That is:

$$b\left(X\right) \coloneqq P\left(T = 1 \mid X\right) = \mathrm{sigmoid}\left(\alpha_T + X'\beta_T\right)$$
$$\mathbf{E}\left[T\right] = 0.1$$

where

$$\mathrm{sigmoid}\left(x\right) = \frac{1}{1+e^{-x}}$$
$$\mathrm{logit}\left(x\right) = \mathrm{sigmoid}^{-1}\left(x\right) = \log\left(\frac{x}{1-x}\right)$$

and we have

$$\forall x \in \mathbb{R} \quad 0 < \mathrm{sigmoid}\left(x\right) < 1, \quad \frac{\partial \mathrm{sigmoid}\left(x\right)}{\partial x} > 0.$$

The conditional expected untreated outcome is a quadratic function of the covariates, with interaction terms set to 0 and second order coefficients small with respect to the linear coefficients. Taking α.s to be scalar, β.s to be p-dimensional vectors, and σ.s to be positive scalars, define:

$$\mathbf{E}\left[Y(0) \mid X\right] \coloneqq \mu_0\left(X\right) = \alpha_0 + X'\beta_0 + X'\mathrm{diag}\left(\gamma_0\right)X.$$

where γ_0 is a vector of dimension p and the diag($\cdot$) function takes a dimension p vector and returns a $p \times p$ matrix with the vector populating the diagonal, and 0's elsewhere.

The conditional expected treated outcome is the sum of the conditional expected untreated outcome and a conditional expected treatment term, which is linear in the covariates:

$$\mathbf{E}\left[Y(1) \mid X\right] \coloneqq \mu_1\left(X\right) = \mu_0\left(X\right) + \tau\left(X\right)$$

$$\tau\left(X\right) = \alpha_{\text{eff}} + X'\beta_{\text{eff}}.$$

Finally, the observed outcome is generated from a normal distribution using the following model:

$$\forall t \in \{0,1\} \quad Y\left(t\right) \mid X \sim N\left(\mu_t\left(X\right), \sigma_{\text{noise}}^2\right).$$

We wish to generate the treatment effect such that it is a combination of a component highly correlated with the propensity score, and a random component orthogonal to propensity score, where we control the balance of those two components. We do so by first constructing a vector $\beta_\perp$ such that $\beta_\perp \perp \beta_T$ and define

$$
\begin{aligned}
\operatorname{logit}\left(b\left(X\right)\right) &= \alpha_T + X'\beta_T \\
\mu_X^{\mathrm{lo}} &= \mathbf{E}\left[\operatorname{logit}\left(b\left(X\right)\right)\right] \\
\sigma_X^{\mathrm{lo}} &= \operatorname{StdDev}\left(\operatorname{logit}\left(b\left(X\right)\right)\right) \\
Z_T &= \frac{\operatorname{logit}\left(b\left(X\right)\right) - \mu_X^{\mathrm{lo}}}{\sigma_X^{\mathrm{lo}}} \\
Z_\perp &= \frac{X'\beta_\perp - \mathbf{E}\left[X'\beta_\perp\right]}{\operatorname{StdDev}\left[X'\beta_\perp\right]}
\end{aligned}
$$

where the Z's are standard normal. Using those symbols, we can decompose the treatment effect as

$$
\begin{aligned}
\tau\left(X\right) &= \alpha_{\text{eff}} + X'\beta_{\text{eff}} \\
&= \alpha_{\text{eff}} + \sigma_{\text{eff}}\rho_{\text{eff}}Z_T + \sigma_{\text{eff}}\sqrt{1-\rho_{\text{eff}}^2}Z_\perp
\end{aligned}
$$

so that $\rho_{\text{eff}} \in (-1, 1)$ controls the correlation between the log odds of the propensity and the treatment effect for all units, treated and control. Heterogeneity and magnitude are adjusted through σ_{eff} and α_{eff}, respectively. For details on parameter settings see Appendix B.

For all of our models, we assume linearity in both the outcome and in the log odds of the propensity. Given the DGP described here, the propensity models are correctly specified and the outcome models are misspecified. We employ a version of AIPW that essentially uses an ANCOVA outcome model, so that it does not end up with extra degrees of freedom with which it might better fit the outcome model.

4.2 Sample Size Scenario

Although the consistency properties of DR estimators and the sandwich estimator for their SEs may seem to imply that they should outperform other methods in the presence of mis-specification, consistency is an asymptotic property. It's not clear at what sample sizes a given DR estimator will produce estimates close to the true parameter. At small sample sizes, simpler methods that involve fewer nuisance parameter estimates may even deliver better performance when both bias and variance are taken into account. For these reasons, we compare the estimators at a range of sample sizes, with 10,000 Monte Carlo repetitions at each sample size, otherwise holding the DGP fixed. The number of units ranges from 200 to 25,600 in a grid that mixes linear and log scale to achieve higher granularity in multiple regions of interest.

4.3 Propensity-Effect Size Correlation Scenario

In any generating process involving treatment targeting it is likely that treatment effect size will be heterogeneous, and that it will be correlated with the propensity to be treated. Were this not the case, targeting would be totally unnecessary. That is, any candidate for exposure would have the same expected increase in outcome measurement, hence removing the need for targeting. This correlation is particularly problematic for causal inference, as discussed in detail in Appendix A.3. To summarize its conclusions, when they are uncorrelated, all methods examined here should eliminate bias. When they are highly correlated, all methods will likely suffer increased bias, with those more beholden to the assumption of homogeneous treatment effects suffering the worst effects.

In this scenario, we generate 5,000 units from the same model described in Section 4.1, varying ρ_{eff}. It should be noted that this directly correlates the Conditional Average Treatment Effect (CATE: $\mathbf{E}[Y(1) - Y(0) \mid X]$) and the propensity to be treated, but cannot induce complete correlation between the CATE and the treatment assignment because treatment assignment is stochastic and probabilistic. We vary α_{eff} along with ρ_{eff} in order to maintain a uniform ATT.

Again, we generate 10,000 Monte Carlo repetitions for each correlation value.

5 Simulation Results

In this section, we list the specific configurations of the estimators tested in our simulations, discuss the metrics that are used for comparison, and review the results from the two scenarios.

5.1 Estimator Implementations

For our simulations, we assess several of the estimators discussed in Section 3.2, with a closer focus on the ANCOVA models due to their prevalence in applications within our area of interest. These estimators are DR augmented inverse propensity weighting and three ANCOVA based methods. We further outline the specific configuration chosen for each estimator.

Our AIPW estimator uses a logistic regression for its propensity model, and linear regression with a treatment indicator for its outcome model (essentially ANCOVA). Generally AIPW would be implemented with separate treated and control outcome models, but we wanted to put it on equal footing with true ANCOVA models. We examine an unmatched ANCOVA and two flavors of matched ANCOVA models: PSM and CEM. No regularization is applied to the ANCOVA regression in any of these models, as doing so would bias the resulting estimate because it is extracted directly from the coefficient of the treatment indicator. For both matched ANCOVAs, we choose control units with replacement so that another treated unit may reuse them (by increasing their weighting in the ANCOVA) instead of being forced to clip a treated unit in the case where all of its valid matches have already been used in the matching process.

For PSM, we used an a elastic net penalized logistic regression for the propensity scores, with regularization strength chosen via cross-validation. We then paired three distinct control units to every treated unit, as our pool of control units was significantly larger than our pool of treated units, and we wanted to include as much data as possible in the ANCOVA stage while still preserving balance. Again, we do allow the same control unit to be paired to multiple treated units. Higher and lower ratios of treated-to-control matches were examined in experiments not included in this paper, but in the settings examined here none outperformed 3:1.

CEM's configuration is more complex, so we include three variants to demonstrate different performance regimes. The first coarsens each covariate to two levels, splitting on the median value of that covariate across all units, and matches one control unit to every treated unit with replacement. The second coarsens to three levels based on the quantiles of the covariates over the treated units, and also pairs one control to every treated unit with replacement. The third and final variant coarsens to two levels over the treated units' covariate distributions, and matches three distinct controls to every treated unit, again with replacement.

Estimator	Matching Type	Control/Treated Match Ratio	Notes
PSM ANCOVA	PSM	3/1	Logistic model for propensity with L^2 regularization
CEM ANCOVA 2A1-1	Quantile on all units (treated and control); CEM	1/1	2 coarsening levels per covariate; split at median of population values
CEM ANCOVA 3T1-1	Quantile on treated units; CEM	1/1	3 coarsening levels per covariate; split at 33% and 66% of treated population values
CEM ANCOVA 2T3-1	Quantile on treated units; CEM	3/1	2 coarsening levels per covariate; split at median of treated population values
ANCOVA	N/A	N/A	
AIPW	N/A	N/A	Logistic model for propensity; linear model with treatment indicator for outcome

Table 1: Table of Estimators

5.2 Evaluation Metrics

In comparing the performance of the estimators described in Section 3.2, we focus on three metrics: bias, root mean squared error (RMSE), and confidence interval coverage. The bias of an estimator with respect to its estimand is the difference between the expected value of the estimator and the true value of its estimand. Bias elimination is the raison d'etre for causal methods, and is the primary differentiator between causal methods and non-causal methods like simple difference of means. We must therefore confirm that any viable estimator does indeed yield unbiased (technically, consistent) results under adverse conditions, when presented with enough data.

Confidence interval coverage is the probability that the confidence interval includes the true treatment effect,

$$P_\theta\left(L\left(X\right) \leq \theta \leq U\left(X\right)\right),$$

where θ is the parameter for which the confidence interval is to be calculated, X denotes the observed data, and $L(\cdot)$ and $U(\cdot)$ are the upper and lower bound functions that define the confidence interval. Methods with actual coverage less than nominal coverage provide unrealistically precise estimates. A valid confidence interval methodology should hold regardless of the actual values involved:

$$\min_{\theta\in\Theta} P_\theta\left(L\left(X\right) \leq \theta \leq U\left(X\right)\right) \geq 1-\alpha.$$

That is to say, we require that the worst-case coverage level still exceeds $1-\alpha$. Subject to the valid coverage requirement, we want the narrowest interval possible. In our analyses, we use the normal approximation to calculate the confidence interval. That is to say, we estimate the point estimate $\hat{\theta}(X)$ and the standard error estimate $\hat{s}\left(X\right)$, yielding the CI

$$\begin{aligned} L\left(X\right) &= \hat{\theta}\left(X\right) + z_{\alpha/2}\hat{s}\left(X\right) \\ U\left(X\right) &= \hat{\theta}\left(X\right) + z_{1-\alpha/2}\hat{s}\left(X\right). \end{aligned}$$

However, confidence interval coverage does not tell the full story. An unbiased estimator with accurate standard errors might still be inefficient, with standard errors much larger than an alternative estimation scheme. To fully evaluate an estimator we examine the root mean squared error (RMSE) of the model, as that simultaneously measures both bias and variance. The RMSE is defined (and decomposed) as follows:

$$\begin{aligned} \text{RMSE} &= \sqrt{\frac{1}{n_{\text{MC}}}\sum_{i=1}^{n_{\text{MC}}}\left(\hat{\theta}_{[i]}-\theta_0\right)^2} \\ &= \sqrt{\left(\hat{\theta}-\theta_0\right)^2 + \frac{1}{n_{\text{MC}}}\sum_{i=1}^{n_{\text{MC}}}\left(\hat{\theta}_{[i]}-\hat{\theta}\right)^2} \\ &= \sqrt{\text{Bias}^2 + \text{Variance}} \end{aligned}$$

where $\hat{\theta}_{[i]}$ is the parameter estimate from the i^{th} Monte Carlo iteration, and θ_0 is the true parameter value.

In all scenarios, for each configuration, we generate $n_{\text{MC}} = 10,000$ datasets and average these metrics over them. We calculate ground truth by averaging the ATT (calculated from the pre-noise potential outcomes) over all of the datasets.

5.3 Results with Varying Sample Size

Convergence of an estimator is usually compared to the rate of $\sqrt{n}$, so here we present the bias (not scaled by ground truth) and RMSE multiplied by that factor in our visualizations, in addition to relative bias without the $\sqrt{n}$ multiplier.

In Figures 2 and 3, we observe that PSM ANCOVA and AIPW approach unbiasedness at a rate exceeding the expected $\sqrt{n}$. This is particularly significant because the treatment assignment probability is correlated with not only the control potential outcome, but also with the heterogeneous CATE, making for a very difficult inference problem. AIPW and PSM ANCOVA both succeed in essentially eliminating selection bias when provided with a sufficiently large sample.

On the other hand, Unmatched ANCOVA and the CEM ANCOVAs immediately demonstrate their inability to handle the difficulties of the scenario. The crude matching and lack thereof, fail to eliminate the imbalance even though the number of covariates is relatively low. In Figure 2 we observe that they converge to biased estimates ranging from 5% to 25% for CEM and approximately 15% for unmatched. Note also the tradeoff between CEM configurations of sample size required for stabilization and the extent to which the bias can be reduced. The "2A1-1" variant quantiles on the entire population, which appears to be an inefficient to achieving balance for an ATT estimate, so performs even worse than ANCOVA without any matching at all due to the reduction in the number of data points available without any significant improvement to balance. This is likely due to the vast majority of treated units falling into the same few coarsening "buckets," resulting in control units whose covariates may be in the same range as those of the treated units, but likely have very different distributions. The "2T3-1" quantiles on the target population's covariate distributions and thus does a much better job of balancing than the first configuration, as evidenced by the improved bias reduction. The "3T1-1" configuration performs best of all of the CEM variants in bias reduction because the finer grained matching allows for significantly better balancing. For the latter two configurations, we note that even their unsatisfactory level of bias reduction requires a sizeable sample to be achieved. This is due to the difficulty in assembling a sufficiently sized control group to match the treated group.

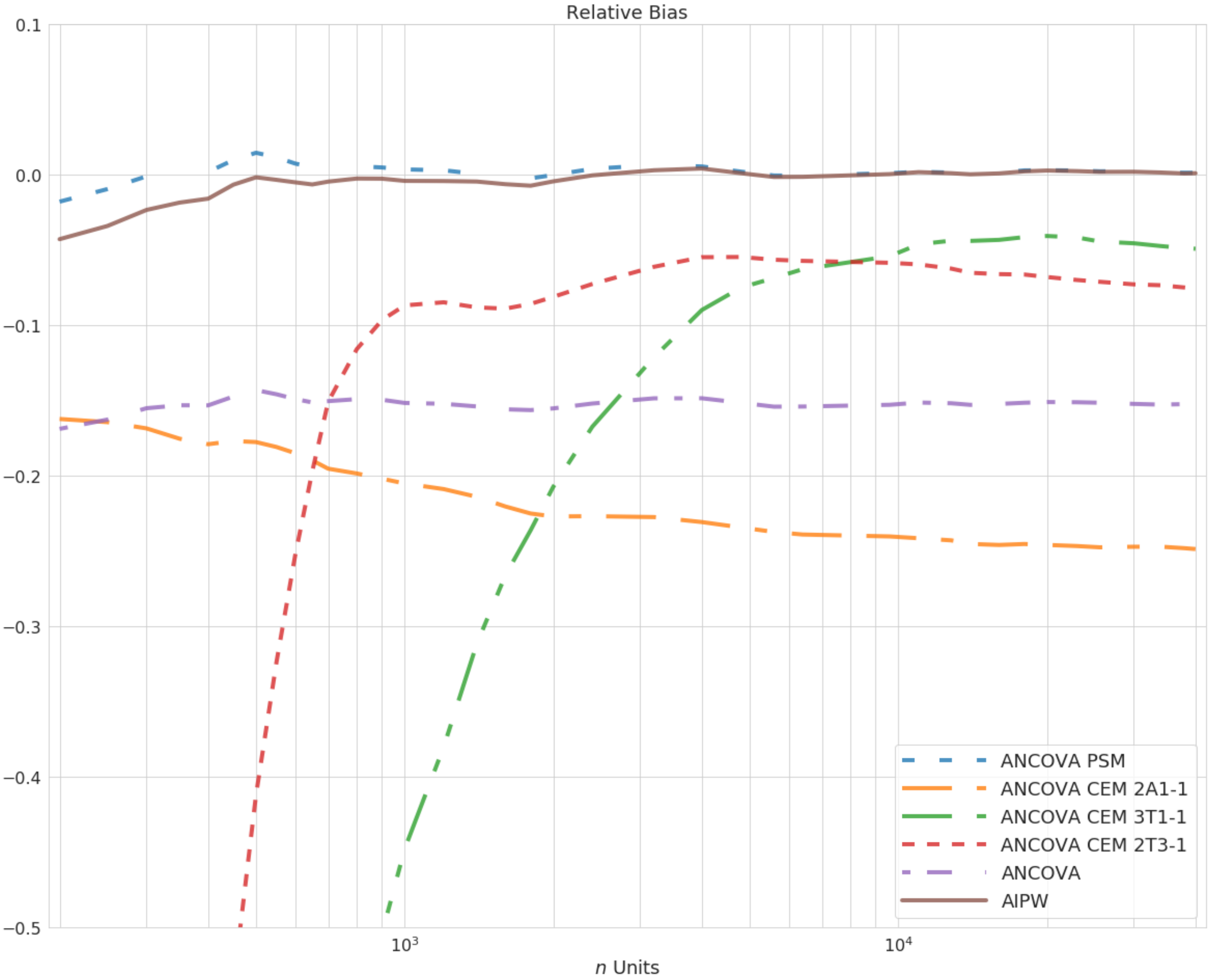


Figure 2: Estimator relative bias ($\frac{\text{bias}}{\text{ground truth ATT}}$) over sample size. PSM and AIPW are able to remove essentially all bias. CEM and unmatched converge to estimates off by $\sim 45\%$ and $\sim 30\%$, respectively.

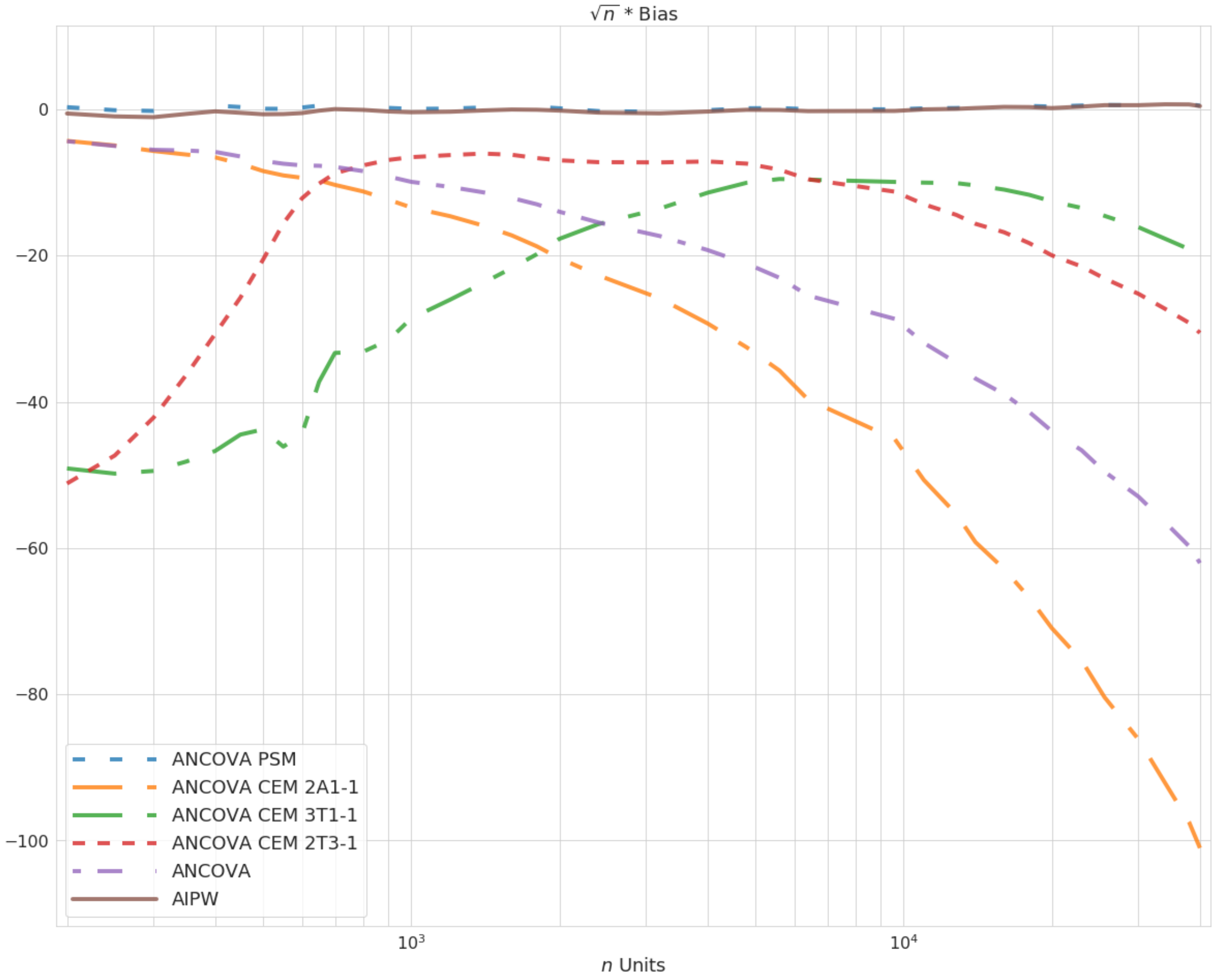


Figure 3: Estimator bias (not scaled by the ground truth effect) times the square root of sample size in the estimation of ATT. CEM and unmatched ANCOVA's scaled bias diverges.

Figure 4 shows coverage rates for 80% confidence intervals for each method, so good coverage levels should approach 0.8 as the sample size increases. We note that coverage for unmatched ANCOVA and the CEM ANCOVAs plummet as the sample size grows. This is due to the fact that as the sample size increases, the confidence intervals shrinks around the still biased point estimates, resulting in poorer coverage, as shown in Figure 3.

The difference in asymptotic coverage between AIPW methods and PSM ANCOVA is smaller, but still clearly apparent. AIPW maintains coverage levels above the nominal level, due to the sandwich SE estimator capturing all of its uncertainty. It achieves statistical validity at very low sample sizes in these simulations, and uniformly dominates in that metric. The matched PSM ANCOVA falls below the nominal coverage rates, in spite of the use of robust standard errors for the ANCOVA models. This is because they fail to take into account the uncertainty involved in the matching process, and therefore underestimate their standard errors, resulting in undercoverage. Without heteroskedasticity robust SEs, the ANCOVA estimators would suffer even greater undercoverage.

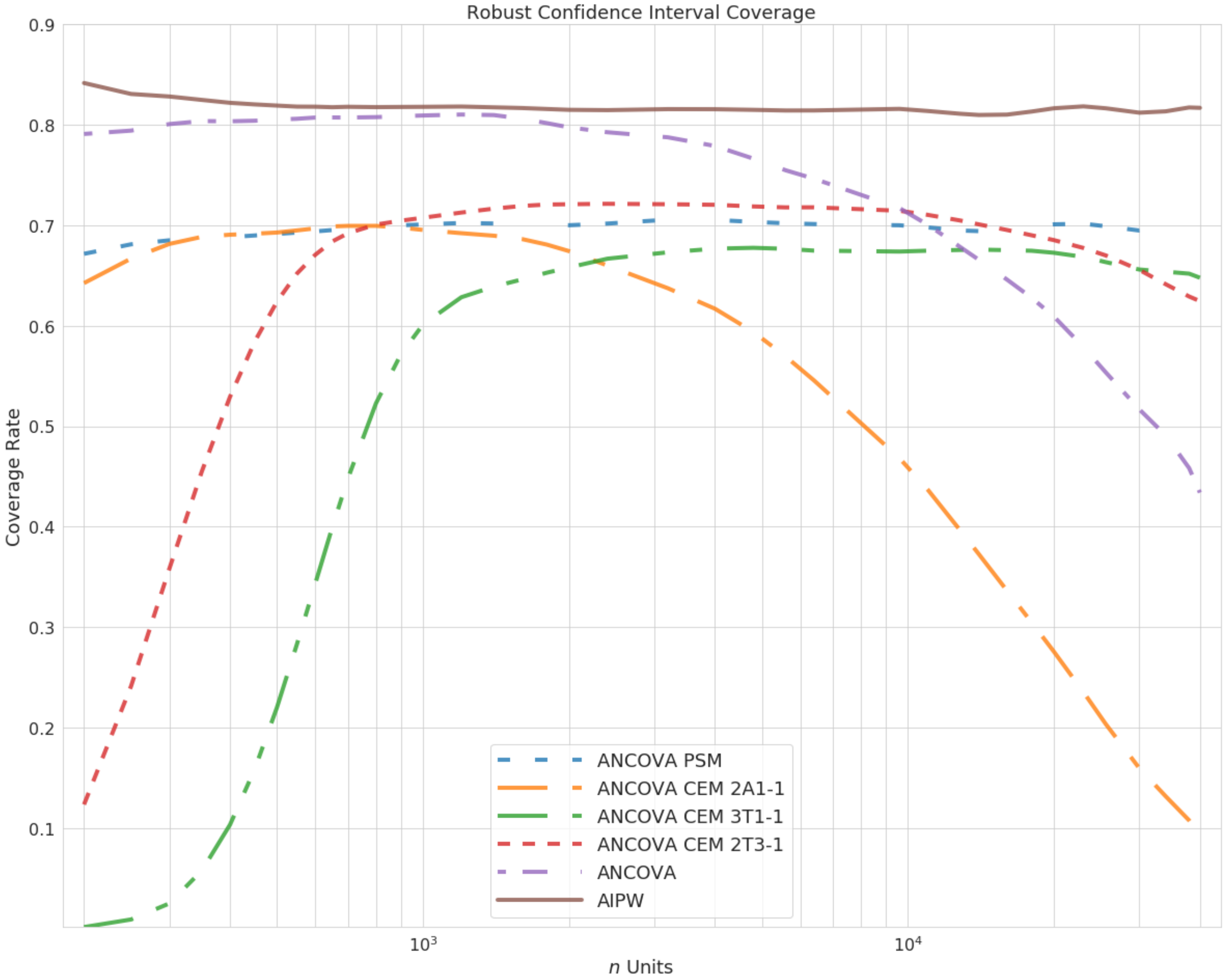


Figure 4: Confidence interval coverage using robust, model standard errors over sample size. AIPW maintains nominal coverage at all samples sizes. PSM fails to do so, though it does not plunge in the way CEM and unmatched ANCOVA do. Target: 80%.

The RMSE of the estimators (Figure 5) further reinforces the takeaways from the first three plots: unmatched and CEM ANCOVA fail to adequately debias their estimates. The exact matching of CEM is very sensitive to configuration and requires enormous amounts of data before achieving its maximal debiasing. Additionally, it often underperforms even unmatched ANCOVA because it eliminates too much of the data (increasing variance) while still failing to adequately balance (preserving or even enhancing bias). The initial decrease in the scaled RMSE of the CEM ANCOVA variants is due to a decrease in variance as sample size grows, but soon that narrowing leads to a concentration of the estimates near the biased value, and a blow up for scaled RMSE.

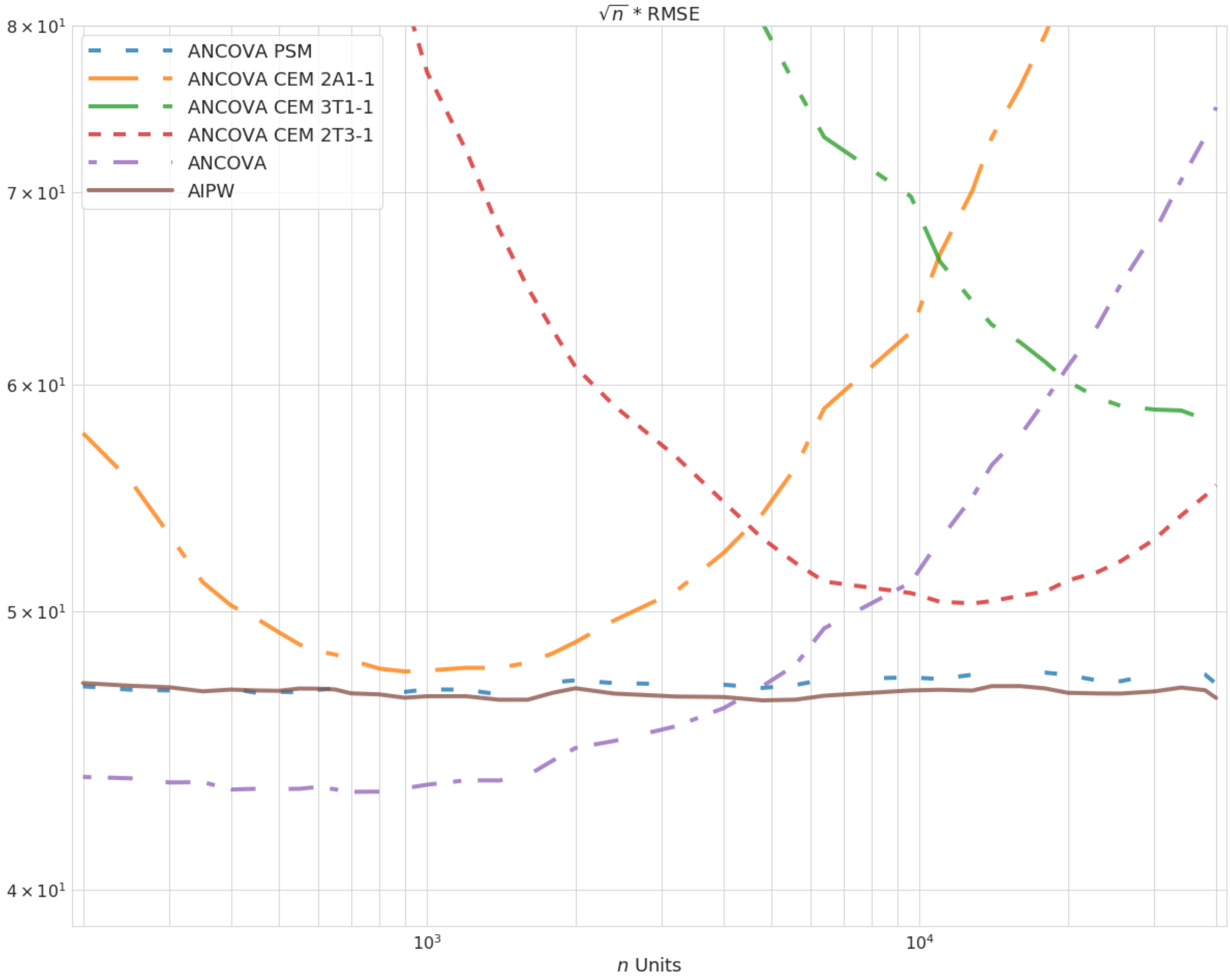


Figure 5: RMSE times the square root of sample size over sample size. Scaled RMSE of CEM and unmatched ANCOVA diverges.

5.4 Results with Varying Propensity-Effect Size Correlation

In a real world setting, CATE is often heterogeneous and correlated with the propensity to be treated. For instance, an individual's online behavior will affect both their likelihood of being exposed to an advertisement, and the effect that such an ad exposure has on their behavior. A first time YouTube user is much more likely to remember the first video ad to which they were exposed than a long-time user is their ten-thousandth ad exposure. To show the effect of this correlation on the methods examined here, we vary it across a wide range of values, as described in Section 4.3, holding sample size fixed at 5,000. We include only the 2A1-1 CEM variant in this section because the other two are still dominated by their variance at sample sizes of 5,000, as can be seen in Figure 5. It is, however, a poor configuration relative to the other two presented here, with its relative bias never dropping below 15% as sample size increases.

In Figure 6, we observe that as expected, all methods result in essentially unbiased results when the treatment assignment propensity and the CATE are uncorrelated. Bias tends to grow with the magnitude of the correlation, and in its opposing direction. As in the sample size simulations, CEM ANCOVA and unmatched ANCOVA are uniformly dominated by the other methods in terms of bias reduction. AIPW demonstrates slightly inferior performance to PSM ANCOVA.

In the simulations presented here, CEM underperforms unmatched ANCOVA, but that is a function of the CEM configuration and the low sample size. At higher sample sizes with better configurations, (for instance, the other two presented in the previous simulation scenario) CEM will tend to dominate unmatched ANCOVA, however the sample size necessary for that to occur will explode as the number of dimensions, granularity of the coarsening, and separation of the treated and control groups increases. Here we focus only on CEMs performance as the correlation parameter varies, recognizing that its relative performance with respect to unmatched ANCOVA is an artifact. The main relevant takeaway is that the bias reduction of both of these methods is highly dependent on the correlation. They both succeed at eliminating bias in their treatment effect estimation when the correlation is 0, and fail elsewhere.

As discussed in Section 4, the generating propensity models are linear in the log odds space meaning that the propensity models for PSM ANCOVA and AIPW are correctly specified, so it stands to reason that they would perform well. While there are a number of simple modifications to the generating outcome models that could spoil the success of the PSM ANCOVA model, the balance induced by effective propensity matching goes a long way towards eliminating bias. Likewise, AIPW's correctly specified propensity model ensures significant bias reduction.

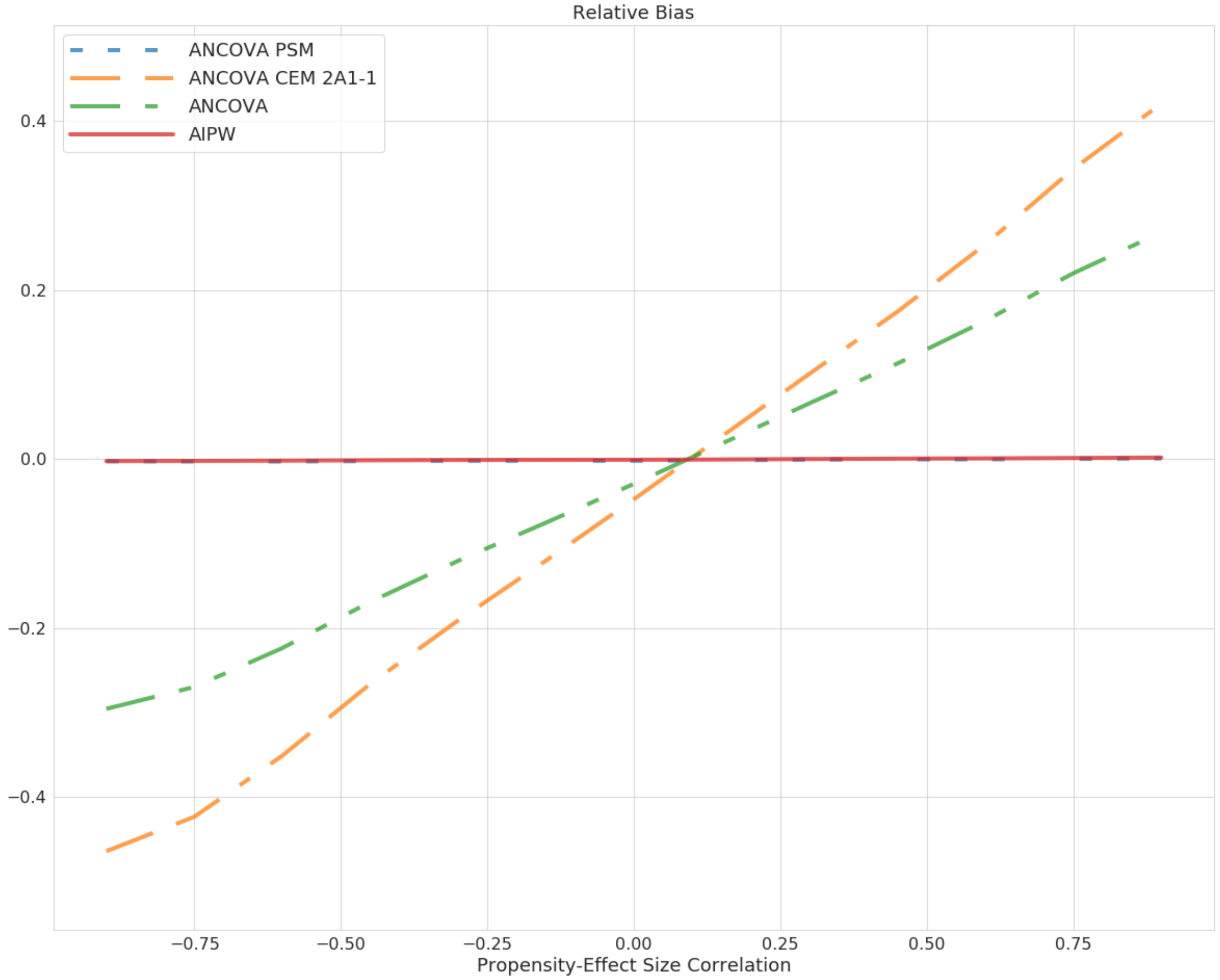


Figure 6: Estimator relative bias ($\frac{\text{bias}}{\text{ground truth ATT}}$) by propensity-effect size correlation. CEM and unmatched ANCOVA demonstrate high sensitivity to this correlation.

Although PSM ANCOVA succeeds in debiasing its estimates, it still fails to reach nominal coverage levels, whereas AIPW achieve full coverage, regardless of propensity-effect size correlation, as demonstrated in Figure 7. The misspecification in this generating scenario was chosen because between propensity misspecification and outcome misspecification, outcome misspecification is more favorable for PSM ANCOVA. That is, propensity matching should still eliminate the bias, and robust standard errors can be applied to the misspecified outcome model. However, it still fails to achieve the desired performance, thereby emphasizing the necessity of including the matching uncertainty in the overall standard error estimate in order to provide statistically valid confidence intervals.

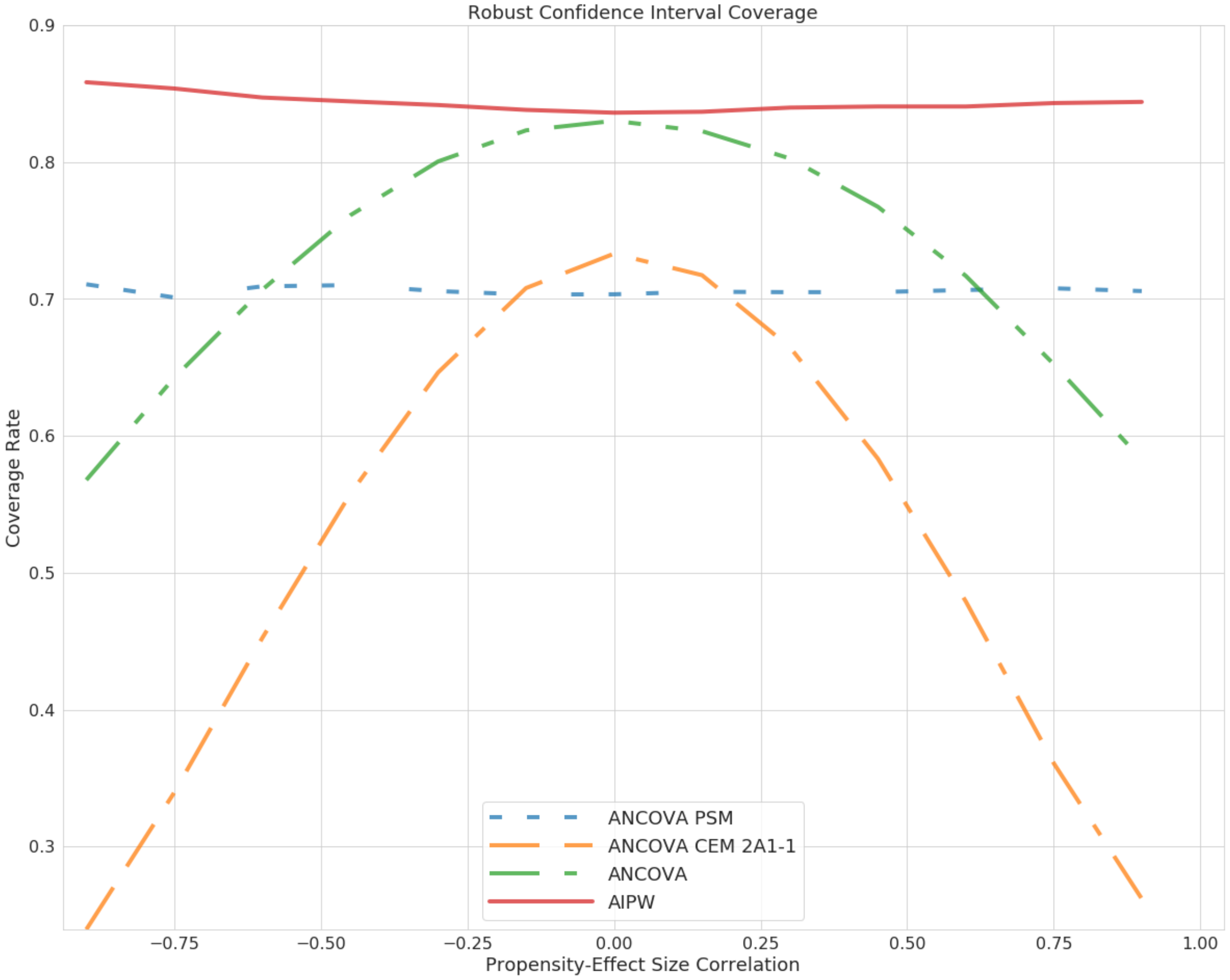


Figure 7: Confidence interval coverage using robust model standard errors by propensity-effect size correlation. Target: 80%.

Again, CEM and unmatched ANCOVA fail to consistently achieve valid coverage, largely due to the bias of their point estimates. When propensity and effect size are uncorrelated, unmatched ANCOVA actually achieves the nominal coverage level. The fact that the unscaled RMSE of CEM and unmatched ANCOVA dips below that of AIPW and PSM ANCOVA when correlation is near 0 in Figure 8, when all four methods have bias near 0, indicates that the variance of the former two methods is lower than that of the latter. This drives home the point that ANCOVA models will very quickly concentrate around their expected value, regardless of whether they're correctly specified or not.

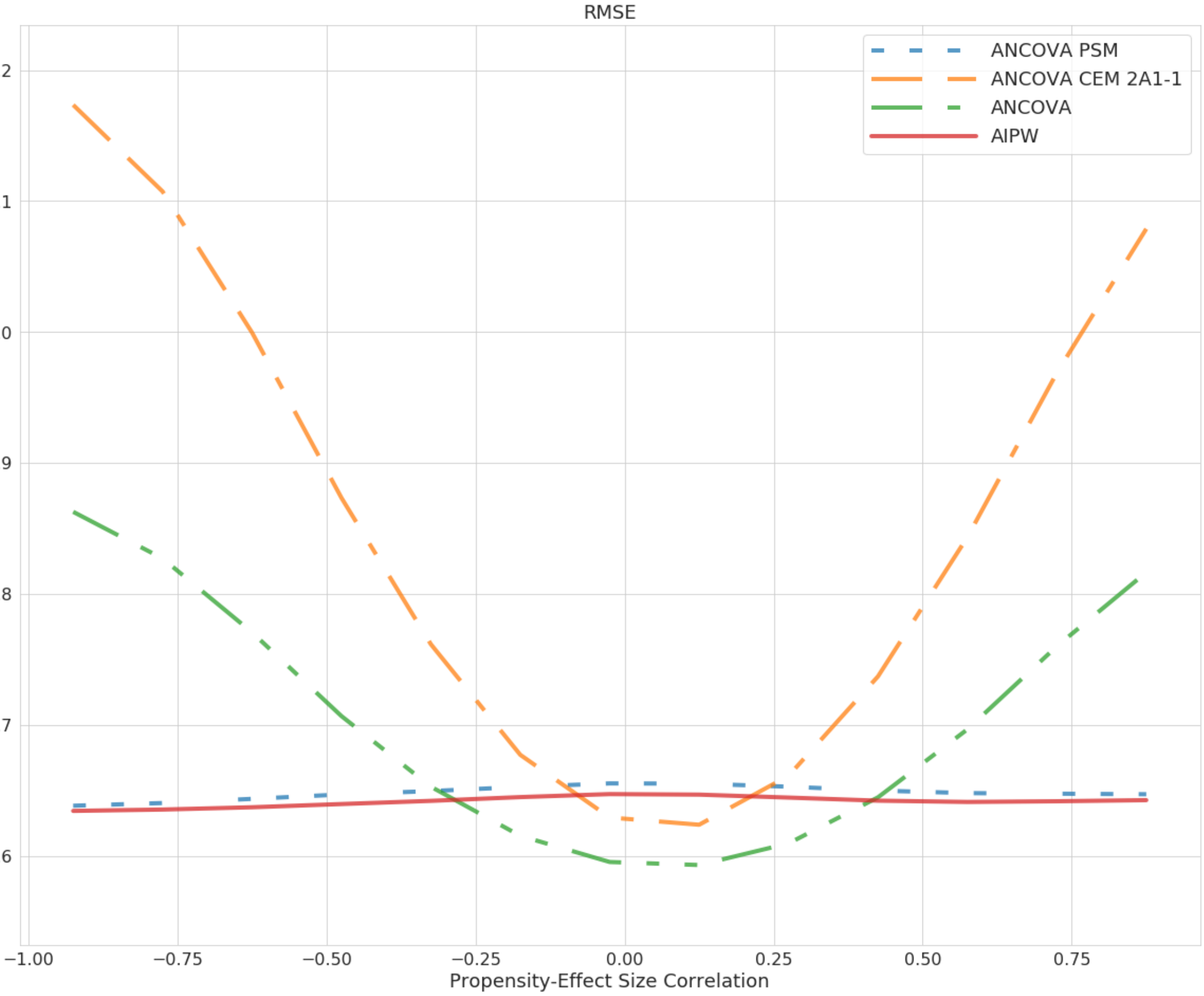


Figure 8: RMSE by propensity-effect size correlation.

6 Conclusion

In this paper, we compared the performance of a set of commonly used causal estimators in two illustrative simulation scenarios. We examined the robustness of these estimators in both point estimation and uncertainty quantification. Robust uncertainty quantification is particularly important when the point estimates can be small relative to the noise, making distinguishing statistically significant results from statistically insignificant results challenging.

We introduce DR estimators which remain consistent if either their propensity model or outcome model is correctly specified. However, this advantage is limited when overly restrictive parametric models are employed, as they will all tend to have some degree of misspecification in most real-life problems. The promise of a consistent estimate can be achieved by using more flexible semi-parametric or non-parametric propensity and outcome models.

While adequately specified DR point estimators are consistent, naive implementations of their standard error estimators are not robust to model misspecification. This can lead to erroneous conclusions about the validity of the result. A main contribution of this paper is the introduction of a robust standard estimator, implemented for the AIPW method described in this paper and applicable to a wide range of parametric and semi-parametric models. Such a robust standard estimator is not available for matching methods, although robust standard errors can be obtained for the ANCOVA outcome models.

We first note that for DR methods to achieve un-biased point estimates and correct CI coverage, large sample sizes are generally required. For the simulation scenarios used in this paper, with a moderate number of covariates and signal to noise ratio, adequate sample size requirements were observed to be in the thousands of observations. For many real-life problems where the noise-to-signal ratio could be high and the effect size-propensity correlation severe, many observations may be required before the estimates stabilize. Unfortunately, it is not possible for us to estimate the necessary sample size here because of the myriad other factors that affect sales lift estimation (eg. What the publisher considers to be an ad exposure and data vendor's sources and coverage). However, based on our experience with running real-life sales lift studies, we have required millions of observations before we start seeing consistency of results between the propensity based method and results from an experiment. See the discussion of the effect of sample size and volatility on confidence in randomized control trials in [17] for details.

In all scenarios examined in this paper, CEM and unmatched ANCOVA demonstrated poor bias reduction and statistical validity in comparison to the other methods, concentrating around biased estimates. This bias is more pronounced as the propensity effect size correlation increases. PSM ANCOVA did a better job of achieving unbiased point estimates in the presence of such correlation, but fails to obtain adequate CI coverage, due to the unavailability of a robust standard error estimator. While robust standard errors for ANCOVA can be obtained, unmatched ANCOVA suffers from bias, due to correlation between propensity and effect size.

We performed, but did not include in this paper, experiments using the ESTIMATION WITHIN STRATA approach. The results were interesting and somewhat distinct from the CEM ANCOVA approach, but still failed to outperform the other methods (PSM ANCOVA and AIPW) in bias reduction or statistical validity. In a scenario in which both propensity and outcome are highly non-linear, with absurdly large sample sizes it's conceivable that the stratification approach (with ANOVA in the strata) might outperform a grossly misspecified AIPW estimator, but as long as the AIPW used sufficiently flexible submodels it would require a pathological data generating process.

7 Acknowledgements

We would like to thank many of our colleagues at Google for their comments and suggestions. In particular we acknowledge Jim Koehler and Yijia Feng for their crucial feedback and innumerable discussions.

References

[1] Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

[2] Alberto Abadie and Guido W Imbens. Matching on the estimated propensity score. *Econometrica*, 84(2):781–807, 2016.

[3] Takeshi Amemiya. *Advanced econometrics*. Harvard university press, 1985.

[4] Joseph Antonelli, Matthew Cefalu, Nathan Palmer, and Denis Agniel. Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics*, 74(4):1171–1179, 2018.

[5] David Chan and Mike Perry. Challenges and opportunities in media mix modeling. Technical report, 2017.

[6] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[7] Xavier Drèze and François-Xavier Hussherr. Internet advertising: Is anybody watching? *Journal of Interactive Marketing*, 17(4):8–23, 2003.

[8] Ludwig Fahrmeir and Heinz Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, pages 342–368, 1985.

[9] Moritz Flubacher, George Sheldon, Adrian Müller, et al. Comparison of the economic performance between organic and conventional dairy farms in the swiss mountain region using matching and stochastic frontier analysis. *J. Socio-Econ. Agric*, 8:76–84, 2015.

[10] Kashmira Gander. Vegetarian, vegan diets linked to lower heart disease risk, but higher risk of stroke. *Newsweek*, 2019.

[11] Adam N Glynn and Kevin M Quinn. An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, 18(1):36–56, 2010.

[12] Jennifer Hill and Jerome P Reiter. Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine*, 25(13):2230–2256, 2006.

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

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

[15] Gary King and Richard Nielsen. Why propensity scores should not be used for matching. *Political Analysis*, 2019 2019.

[16] Gary King, Richard Nielsen, Carter Coberley, James E Pope, and Aaron Wells. Comparative effectiveness of matching methods for causal inference. *Unpublished manuscript*, 15, 2011.

[17] Randall A Lewis and Justin M Rao. The unfavorable economics of measuring the returns to advertising. *The Quarterly Journal of Economics*, 130(4):1941–1973, 2015.

[18] Judea Pearl. *Causality*. Cambridge university press, 2009.

[19] Joseph L Schafer and Joseph Kang. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods*, 13(4):279, 2008.

[20] Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable dropout using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

[21] Andrew Sommerlad, Joshua Ruegger, Archana Singh-Manoux, Glyn Lewis, and Gill Livingston. Marriage and risk of dementia: systematic review and meta-analysis of observational studies. *J Neurol Neurosurg Psychiatry*, 89(3):231–238, 2018.

[22] Peter Spirtes. Introduction to causal inference. *Journal of Machine Learning Research*, 11(May):1643–1662, 2010.

[23] Leonard A Stefanski and Dennis D Boos. The calculus of m-estimation. *The American Statistician*, 56(1):29–38, 2002.

[24] Morgan Stephen and Winship Christopher. Counterfactuals and causal inference: Methods and principles for social research, 2007.

[25] Tammy YN Tong, Paul N Appleby, Kathryn E Bradbury, Aurora Perez-Cornago, Ruth C Travis, Robert Clarke, and Timothy J Key. Risks of ischaemic heart disease and stroke in meat eaters, fish eaters, and vegetarians over 18 years of follow-up: results from the prospective epic-oxford study. *bmj*, 366:l4897, 2019.

[26] Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data.* Springer Science & Business Media, 2011.

[27] Halbert White et al. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

[28] Qingyuan Zhao and Daniel Percival. Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1), 2017.

A Derivations

A.1 Sandwich Estimator for AIPW

A.1.1 M-Estimators and a Toy Example

Use of the sandwich estimator for SE estimation requires only that the estimator can be represented as the solution to a system of equations of the form

$$\sum_i \Psi\left(X_i, \alpha, \beta\right) = 0$$

where Ψ is a possibly vector valued function of the individual observations X_i, the parameters being estimated α, and possibly a set of auxiliary parameters β, and that Ψ is differentiable with respect to α and β. In the canonical case, take Ψ to be the gradient of a log likelihood function with respect to the distribution parameters; then the MLE is an M-estimator. In such a case, if the likelihood was misspecified the sandwich estimator would give an accurate estimate of the SE for the parameters about their *true* mean, however that true mean might no longer correspond to the real estimand of interest. For example, if we are interested in estimating the mean of some data we believe to be normally distributed, we would set

$$\alpha = \mu$$

$$\beta = \sigma^2$$

$$\Psi\left(X, \mu, \sigma^2\right) = \begin{pmatrix} X - \mu \\ (X-\mu)^2 - \sigma^2 \end{pmatrix}.$$

In the context of causal inference, this takes the form

$$\sum_i \Psi\left(X_i, T_i, Y_i, \beta, \tau\right) = 0$$

which allows for computation of the estimated covariance matrix for all of the parameter estimators, and therefore the SE of the treatment effect estimator.

A.1.2 For AIPW

Here we briefly demonstrate the framing of the AIPW estimator as an M-estimator, so that a robust sandwich estimator estimate for its SE can be derived. We begin demonstrating the framing of AIPW as an M-estimator, then give a cursory walkthrough of the application of the sandwich estimator.

First define the following loss functions

1. ℓ_C is a (possibly penalized) classification loss function
   (a) ex: $T \log\left(1 + e^{-X\beta}\right) + (1 - T)\log\left(1 + e^{X\beta}\right) + \lambda \|\beta\|_2^2$
2. ℓ_O is a (possibly penalized) outcome loss function
   (a) ex for a continuous outcome: $(Y - X\beta)^2 + \lambda\|\beta\|_2^2$

Then our system of M equations takes the form

$$\psi_t\left(X, T; \beta_C\right) = \nabla_{\beta_t} \ell_C\left(X, T; \beta_t\right)$$

$$\psi_1\left(X, T, Y; \beta_1\right) = T \nabla_{\beta_1} \ell_O\left(X, Y; \beta_1\right)$$

$$\psi_0\left(X, T, Y; \beta_1\right) = (1 - T)\nabla_{\beta_0} \ell_O\left(X, Y; \beta_0\right)$$

$$\psi_\tau\left(X, T, Y; \beta_t, \beta_0, \beta_1, \tau\right) = \frac{T}{b\left(X; \beta_t\right)}\left(Y - m\left(X; \beta_1\right)\right) - \frac{1-T}{1 - b\left(X; \beta_t\right)}\left(Y - m\left(X; \beta_0\right)\right) + m\left(X; \beta_1\right) - m\left(X; \beta_0\right) - \tau$$

The final Ψ is the concatenation of ψ_t, ψ_0, ψ_1, and ψ_τ (the propensity, control outcome, treated outcome, and effect size loss function gradients), and Γ is the concatenation of β_t, β_0, β_1, and τ (the parameters of the propensity, control outcome, and treated outcome functions and the effect size). Using the notation of [23], define

$$\hat{\Gamma}: \quad \sum_i \Psi\left(X_i, T_i, Y_i; \hat{\Gamma}\right) = 0$$

$$A_n = -\frac{1}{N}\sum_i \nabla_\Gamma \Psi\left(X_i, T_i, Y_i; \hat{\Gamma}\right)$$

$$B_n = \frac{1}{N}\sum_i \Psi\left(X_i, T_i, Y_i; \hat{\Gamma}\right)\Psi\left(X_i, T_i, Y_i; \hat{\Gamma}\right)^T$$

$$\Sigma_n = A_n^{-1} B_n A_n^{-1}$$

Then the bottom right element of Σ_n is the sandwich variance estimate for $\hat{\tau}$.

A.2 IPW Formal Derivations

For ATE estimates, both the treated and control means take the same form. We present the derivation for the treated side:

$$\begin{aligned}
\mathbf{E}\left[\frac{T \cdot Y}{b(X)}\right] &= \mathbf{E}\left[\mathbf{E}\left[\frac{T \cdot Y}{b(X)} \mid X\right]\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[\frac{T \cdot Y(1)}{b(X)} \mid X\right]\right] \\
&= \mathbf{E}\left[\frac{1}{b(X)} \cdot \mathbf{E}[T \mid X] \cdot \mathbf{E}[Y(1) \mid X]\right] \\
&= \mathbf{E}\left[\frac{1}{b(X)} \cdot b(X) \cdot \mathbf{E}[Y(1) \mid X]\right] \\
&= \mathbf{E}[\mathbf{E}[Y(1) \mid X]] = \mathbf{E}[Y(1)]
\end{aligned}$$

Substituting $1 - T$ for T and $1 - b(X)$ for $b(X)$, we get

$$\mathbf{E}\left[\frac{(1-T) \cdot Y}{1 - b(X)}\right] = \mathbf{E}[Y(0)]$$

via the same calculations. We can thus estimate ATE by

$$\hat{\tau}_{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N} \frac{T_i \cdot Y_i}{\hat{b}(X_i)} - \frac{1}{N}\sum_{i=1}^{N} \frac{(1-T_i) \cdot Y_i}{1 - \hat{b}(X_i)}$$

Estimation of ATT is similar, however the groups means are not treated as symmetrically. First, consider the conditional treated mean estimator

$$\hat{\mu}_{\text{Treated,ATT}} = \frac{\sum T_i \cdot Y_i}{\sum T_i}$$

$$\begin{aligned}
\mathbf{E}[T \cdot Y] &= \mathbf{E}[T \cdot Y(1)] \\
&= \mathbf{E}[T \cdot Y(1) \mid T = 1] \cdot P(T = 1) \\
&+ \mathbf{E}[T \cdot Y(1) \mid T = 0] \cdot P(T = 0) \\
&= \mathbf{E}[T \cdot Y(1) \mid T = 1] \cdot P(T = 1) + \mathbf{E}[0 \mid T = 0] \\
&= \mathbf{E}[T \cdot Y(0) \mid T = 1] \cdot P(T = 1) + \mathbf{E}[T \cdot Y(0) \mid T = 0] \cdot P(T = 0) \\
&= \mathbf{E}[Y(0) \mid T = 1] \cdot P(T = 1) + 0 \\
\frac{\mathbf{E}[T \cdot Y]}{\mathbf{E}[T]} &= \mathbf{E}[Y(0) \mid T = 1]
\end{aligned}$$

The conditional control mean estimator

$$O_{b,i} = \frac{\hat{b}\left(X_i\right)}{1-\hat{b}\left(X_i\right)}$$

$$\hat{\mu}_{\text{Control,ATT,HT}} = \left(\frac{1}{N}\sum\left(1-T_i\right)\cdot Y_i\cdot O_{b,i}\right)/\left(\frac{\sum\hat{b}\left(X_i\right)}{N}\right) = \frac{\sum\left(1-T_i\right)\cdot Y_i\cdot O_{b,i}}{\sum\hat{b}\left(X_i\right)}$$

arises from the following chain equalities:

$$\begin{aligned}\mathbf{E}\left[\left(1-T\right)\cdot Y\cdot\frac{b\left(X\right)}{1-b\left(X\right)}\right] &= \mathbf{E}\left[\mathbf{E}\left[Y\left(0\right)\mid X\right]\cdot b\left(X\right)\right]\\ &= \mathbf{E}\left[\mathbf{E}\left[Y\left(0\right)\mid X\right]\cdot\mathbf{E}\left[T\mid X\right]\right]\\ &= \mathbf{E}\left[\mathbf{E}\left[T\cdot Y\left(0\right)\mid X\right]\right]\\ &= \mathbf{E}\left[T\cdot Y\left(0\right)\right]\\ &= \mathbf{E}\left[T\cdot Y\left(0\right)\mid T=1\right]\cdot P\left(T=1\right)+\mathbf{E}\left[T\cdot Y\left(0\right)\mid T=0\right]\cdot P\left(T=0\right)\\ &= \mathbf{E}\left[Y\left(0\right)\mid T=1\right]\cdot P\left(T=1\right)+0\end{aligned}$$

A.3 Propensity-Effect Size Correlation Rationale

Consider a linear baseline outcome with heterogeneous additive treatment effect

$$y = \alpha + X'\beta + T\tau\left(X\right) + \epsilon.$$

A sufficiently flexible estimator that correctly models the outcome, allowing for treatment effect heterogeneity, should be able to recover the ATE or ATT. If the treatment effect size is independent of the propensity, conditional upon the covariates, then even ANCOVA can recover the ATE (equivalent to the ATT), as demonstrated here:

$$\begin{aligned} y &= \alpha + X'\beta + T\tau\left(X\right) + \epsilon\\ &= \alpha + X'\beta + T\tau_0 + T\left(\tau\left(X\right)-\tau_0\right) + \epsilon\end{aligned}$$

where $\tau_0 = \mathbf{E}\left[\tau\left(X\right)\right]$ over the population in the dataset. Taking the full error term to be

$$\eta = T\left(\tau\left(X\right)-\tau_0\right) + \epsilon$$

we demonstrate that the full error term is uncorrelated with the treatment assignment, using the following identity, derived in Section A.3.2 :

$$\text{Cov}\left(T,\eta\right) = \mathbf{E}\left[b\left(X\right)\left(\tau\left(X\right)-\tau_0\right)\right]\left(1-\mathbf{E}\left[b\left(X\right)\right]\right)$$

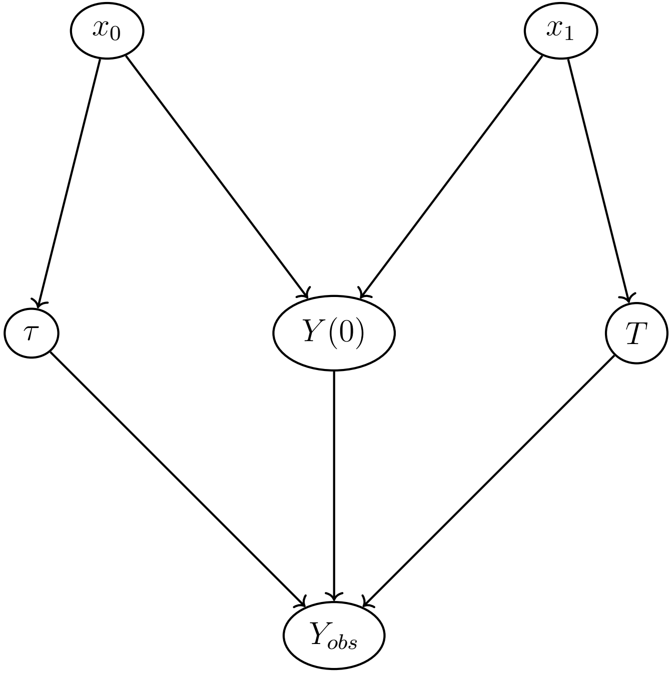
Now if $b\left(X\right)\perp\tau\left(X\right)$ then

$$\mathbf{E}\left[b\left(X\right)\left(\tau\left(X\right)-\tau_0\right)\right] = \mathbf{E}\left[b\left(X\right)\right]\mathbf{E}\left[\tau\left(X\right)-\tau_0\right] = 0$$

hence,

$$\mathrm{Cov}\left(T,\eta\right)=0$$

A covariate's coefficient in a linear model is only biased when that covariate is correlated with the error term. Therefore, while we cannot trust the coefficients estimated for the covariates (returning to our standard use of the terminology), the treatment effect coefficient should be unbiased. To see this, take $T\left(\tau\left(X\right)-\tau_0\right)$ to be the missing covariate (its coefficient being 1). If we consider a toy example in which there are only two covariates, both of which affect the control outcome, one of which affects the effect size, and the other of which affects the treatment propensity, we can draw the following causal DAG:



From this DAG we can use the backdoor criterion [18, ch. 3]to show that $P\left(Y_{obs}\mid do\left(T\right)\right)$ is identifiable, so the ATE is recoverable even if the CATE is not. For a further illustration of this unbiasedness from an omitted variable perspective, see Section A.3.1.

A.3.1 Omitted Variable Bias with Partial Correlations

Take $u,\beta_u\in\mathbb{R}^{p_u}$ to be the set of unobserved covariates and their coefficients in our model, $x_0,\beta_0\in\mathbb{R}^{p_0}$ to be another set of covariates (and coefficients) in our model, and $x_1,\beta_1\in\mathbb{R}^{p_1}$ to be a final such set where the final set x_1 is uncorrelated with the unobserved set u. Now assume the relationship between x_0 and u is of the form

$$u=A^Tx_0+\delta$$

where

$$\delta\sim N\left(0_u,\sigma_\delta^2I_u\right)$$

such that

$$\mathrm{Cov}\left(x_1,u\right)=0$$

Finally, define the outcome model

$$y=x_0\beta_0+x_1\beta_1+u\beta_u+\epsilon$$

with

$$\epsilon \sim N\left(0, \sigma_\epsilon^2\right)$$

Forcing everything into matrix form we have

$$\begin{aligned} Y &= X_0\beta_0 + X_1\beta_1 + U\beta_u + \epsilon \\ &= X_0\left(\beta_0 + A\beta_u\right) + X_1\beta_1 + \left(\delta\beta_u + \epsilon\right) \end{aligned}$$

Thus a regression of y on (x_0, x_1) should yield an unbiased estimate of β_1 (though a biased estimate of β_0).

A.3.2 Treatment assignment and error term are uncorrelated in our DGP

$$\begin{aligned} \operatorname{Cov}(T,\eta) &= \mathbf{E}\left[\mathbf{E}\left[T\eta \mid X\right]\right] - \mathbf{E}\left[\mathbf{E}\left[T \mid X\right]\right]\mathbf{E}\left[\mathbf{E}\left[\eta \mid X\right]\right] \\ &= \mathbf{E}\left[\mathbf{E}\left[T^2\left(\tau(X) - \tau_0\right) \mid X\right]\right] - \mathbf{E}\left[b(X)\right]\mathbf{E}\left[\mathbf{E}\left[T\left(\tau(X) - \tau_0\right) \mid X\right]\right] \\ &= \mathbf{E}\left[\mathbf{E}\left[T\left(\tau(X) - \tau_0\right) \mid X\right]\right] - \mathbf{E}\left[b(X)\right]\mathbf{E}\left[\mathbf{E}\left[T \mid X\right]\left(\tau(X) - \tau_0\right)\right] \\ &= \mathbf{E}\left[b(X)\left(\tau(X) - \tau_0\right)\right]\left(1 - \mathbf{E}\left[b(X)\right]\right) \end{aligned}$$

$$b(X) \perp \tau(X) \Rightarrow \mathbf{E}\left[b(X)\left(\tau(X) - \tau_0\right)\right] = 0$$

B Parameter Values

1. Covariates
   (a) $\rho_X = 0.25$
   (b) $\sigma_X^2 = 0.25$
2. Treatment
   (a) $\alpha_T = -2.33$
   (b) $\beta_T = (0.0, 0.7, 0.55, 0.2, 0.0)$
3. Outcome
   (a) $\alpha_0 = 15.78$
   (b) $\beta_0 = (0.468, 0.223, 0.362, -0.516, -0.144)$
   (c) $\gamma_0 = (-0.1, .2, 0.24, -0.3, 0.15)$
   (d) $\sigma_{\text{noise}} = 2.0$
   (e) Treatment Effect
      i. $\alpha_{\text{eff}} = -2.0$
         A. Varies in propensity-effect size correlation grid to maintain constant ATT.
      ii. $\beta_\perp = (1.48, -0.58, 1.28, -1.5, -1.08)$
      iii. $\rho_{\text{eff}} = 0.4$

A. Varies in propensity-effect size correlation grid.

iv. $\sigma_{\text{eff}} = 15.0$

4. $n = 5,000$

   (a) Varies in sample size grid.