Two Chances are Better than One: Comparing the Performance of Doubly Robust Methods

December 18, 2025

Abstract

To estimate Average Treatment Effects (ATE) in causal inference studies, Doubly Robust Estimators have been widely used. Their advantage lies that they combine two models—a propensity score model and an outcome model—such that the estimator remains consistent if only one of them is correctly specified. The researcher has two opportunities to correctly specify their model, making doubly robust estimators statistically preferred over other estimators. Thus, we ran simulations of different doubly robust estimators (Augmented Inverse Probability Weighting or AIPW, Double Score Matching, and Targeted Maximum Likelihood Estimation) to assess their robustness under model misspecification scenarios, comparing their performance to non-doubly robust estimators (Inverse Probability Weighting or IPW, Linear Regression). Our results demonstrate the overall superiority of doubly robust methods when one model is misspecified. However, we found slight bias with Targeted Maximum Likelihood Estimation and Double Score Matching when the outcome model is misspecified.

Keywords: simulation, causal inference, AIPW, TMLE, DSM

1 Introduction

The Average Treatment Effect, or ATE measures the average causal impact of a treatment. It is a highly desired value to obtain in causal inference studies, especially within the medical field as it quantifies the general impact of health treatments (Wang et al. 2023), medical drugs (Michoel & Zhang 2023), and social determinants of health (Alsan & Wanamaker 2018). Many estimators have been proposed for estimating the ATE, including a variety of propensity score methods and outcome models.

The most desirable estimators, however, are able to combine propensity score models and outcome models. Such estimators have the quality of double robustness. They are consistent for the ATE if only one of the two components—the propensity score model or the outcome model—is correctly specified (Tan et al. 2025). However, the performance of different types of doubly robust estimators across different model specifications, and their overall performance against propensity score and outcome model methods is still poorly understood.

We used a simulation study comparing three doubly robust methods—Augmented Inverse Probability Weighting, Targeted Maximum Likelihood Estimation, and double score matching—as well as one propensity score-based method and one outcome-based method. The propensity score-based method is Inverse Probability Weighting. The outcome-based method is linear regression. We evaluated the estimates for these five methods across four scenarios: (1) both models are correctly specified, (2) only the propensity score model is correctly specified, (3) only the outcome model is correctly specified, and (4) both models are incorrectly specified. In the first scenario, we expect all methods to be unbiased. In the second scenario, we expect all methods except linear regression to be unbiased. In the

third scenario, we expect all methods except Inverse Probability Weighting to be unbiased. The fourth scenario has led to different results across various studies. Kang & Schafer (2007) and Cao et al. (2009) found that if both models are unspecified, bias is magnified. However, Tan (2008) states that in this scenario, doubly robust methods perform similarly to other methods. Thus, we seek to answer the following questions using our simulation:

- How do doubly robust estimators compare to non-doubly robust estimators when both of its modeling components are incorrectly specified?
- How do doubly robust methods compare to the propensity score-based (outcome-based) method when the propensity score (outcome) model is correctly specified? What about when it is incorrectly specified?

2 Causal Inference Overview

A sleep-deprived college student takes an exam. Later in the semester, they learn that they received a failing grade on this exam. Did their sleep deprivation cause them to fail the exam? How would the average student's grade be affected by sleep deprivation? These questions, and many other questions, are at the core of causal inference, a statistical field concerned with finding and characterizing causal relationships.

While many machine learning methods aim to maximize prediction, they cannot disentangle association from causation. For example, a linear regression model could likely predict exam scores with reasonable accuracy from the students' GPA. However, GPA does not cause exam scores—rather, exam scores are a component of GPA (Cunningham n.d.).

Causation of a treatment X on outcome Y requires statistical association, but the converse

is not necessarily true. Association can occur in the above example, where Y causes X. It can also occur when there are common causes, or another variable L which causes both X and Y. For example, perhaps mental illness both causes the student to be sleep deprived and also reduces their testing performance. Causation contains additional complexity which cannot be found from statistical association alone (Miguel A. Hernan 2025).

Suppose X is a treatment with two levels: 0 indicating no treatment or the control, 1 indicating received treatment. $Y^{X=0}$ and $Y^{X=1}$ are potential outcomes, where $Y^{X=0}$ indicates the outcome for if the individual did not receive the treatment $Y^{X=1}$ indicates the outcome for if the individual received the treatment. The fundamental problem of causal inference is, however, that only one of the potential outcomes can be observed. We never see the counterfactual, the hidden outcome, so the individual causal impact is unobserved. This is why causal inference procedures most often measure the Average Treatment Effect, or ATE τ. It is defined as $E[Y^{X=1}] - E[Y^{X=0}]$ (Miguel A. Hernan 2025).

2.1 Observational Studies

Ideally, we would have experimental data to establish causality, but running experiments can be time-consuming, require funding, or unethical. Most data in the world is observational. The usage of observational data to establish causality has been controversial, but only including randomized data does a disservice to the relationships which can be drawn from careful statistical analysis to mitigate selection bias (Rubin 1974). Observational data must meet three criteria before being used for causal inference (Miguel A. Hernan 2025).

1. Positivity - To establish causality of X on Y, individuals must have a positive, non-zero probability of receiving the treatment and the control. This allows for a compar-

ison to be made between the groups. If all students were equally sleep deprived, we could not find the isolated impact of sleep on exam grades(Miguel A. Hernan 2025).

2. Consistency (also known as Stable Unit Treatment Value Assumption or SUTVA) - The treatment and control groups must be well defined so that each group receives the same amount of treatment—or in the control group's case, no treatment (Miguel A. Hernan 2025). How are sleep deprived and not sleep deprived people categorized? Is a student who sleeps 6 hours a day considered sleep-deprived?

3. Exchangeability - The observed values of the control group would be equal in expectation to the potential outcomes of the treatment group if they had not received the treatment. This is often the hardest assumption to meet in the case of observational data, as covariates often are not equal between the treatment and control groups (Miguel A. Hernan 2025).

Usually, these assumptions are violated to some degree in observational studies, and it is up to the researcher to address them (Miguel A. Hernan 2025).

Conversely, Randomized Controlled Trials, also known as RCTs, involve randomly assigning sample participants to the treatment group $X = 1$ or the control group $X = 0$. In the "real world", confounders often affect the relationship between X and Y. However, the random assignment aspect allows the experimenter complete control over the assignment mechanism. If designed properly, a randomized controlled trial should balance the confounders between the treatment group $X = 1$ and the control group $X = 0$, isolating the effects of the treatment. They are the gold standard for assessing causality, automatically fulfilling the three assumptions: positivity, consistency, and exchangeability. (Miguel A. Hernan 2025).

3 Causal Inference Methods for Observational Studies

3.1 Propensity Score Modeling

For RCT's, we know the assignment of treatment X is entirely under the researcher's control. However, for observational studies, the assignment of treatment X happens without manipulation, and so individuals with certain characteristics may be more likely to receive treatment X than others. Thus, the distribution of covariates L are usually unevenly distributed between $X = 1$ and $X = 0$. To adjust for the imbalance, propensity score models and outcome models can be used (Tan et al. 2025).

A propensity score $e(X)$ is the conditional probability of receiving the treatment X, given covariates L. In a RCT where half of the individuals are assigned the treatment $X = 1$ and the other half the control $X = 0$, $e(X) = 0.5$ for all individuals—in observational studies, this usually is not the case. We must create propensity score models to estimate the propensity scores for observational studies (Miguel A. Hernan 2025). Even when the models are misspecified, propensity score methods can reduce bias (Zhang et al. 2022). To estimate the propensity scores, logistic regression modeling popular (Miguel A. Hernan 2025). We could use a logistic regression model to predict the probability a student is sleep-deprived given their age, sex, workload, and presence of mental illness, for instance, and this probability would be the student's propensity score.

Propensity matching is a popular propensity score method which involves pairing observations based on similar propensity scores and comparing their outcomes, one from the treatment group $X = 1$ and one from the control group $X = 0$. Since propensity scores are continuous, an exact match is unlikely so we must define closeness carefully. There is a bias-variance tradeoff—matching values too exactly will result in high variance, but match-

ing too loosely will result in a different $e(X)$ distribution and exchangeability will not hold (Miguel A. Hernan 2025). However, Figure 1 demonstrates a problem with propensity score matching: the imbalance around $e(X) = 0$ and $e(X) = 1$. That is, observations which have received the treatment are unlikely to be around $e(X) = 0$ and observations which have not received the treatment are unlikely to be around $e(X) = 1$. Many observations are not matched, wasting data. The unmatched observations are discarded, created further generalizability issues if the remaining sample is unrepresentative of the target population (Zhang et al. 2022).

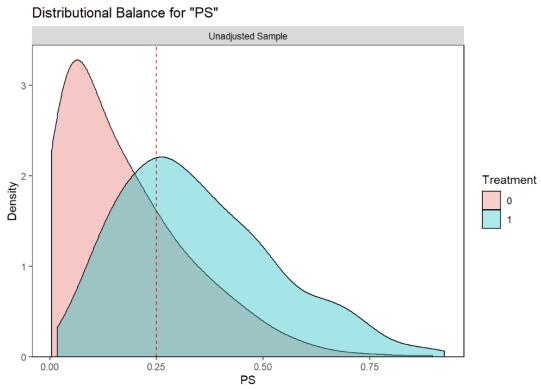

Figure 1: Propensity Scores by Treatment Level for NHANES 2017-2018 (Karim 2023)

3.1.1 Inverse Probability Weighting (IPW)

Instead of wasting data, we can use the IPW estimator to estimate ATE τ, which relies on building a propensity score model, and then using the predicted probability as a weight in subsequent analyses. By re-weighting the observations, it creates a pseudo-population where the treatment assignment is able to be independent of measured confounders. From the exchangeability assumption, the potential outcomes are unbiased (Kurz 2022).

We use $E(\frac{XY}{\hat{e}(X)})$ to estimate $E(Y(1))$. XY is the outcome for the treatment group $X = 1$. This outcome is reweighted by propensity scores $e(X)$ in the denominator so that

observations with lower propensity scores are weighted more. Since observations in the treatment group are unlikely to have small propensity scores, the observations that do must represent more observations. The converse holds for $E(\frac{(1-X)Y}{1-\hat{e}(X)})$, which estimates $E(Y(0))$ (Kurz 2022). The ATE estimate for IPW is characterized by

$$\hat{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}[\frac{X_i Y_i}{\hat{e}(X_i)} - \frac{(1-X_i)Y_i}{1-\hat{e}(X_i)}].$$

The IPW estimator is unbiased when the propensity scores are known. It is consistent when $e(x)$ is the true propensity score. The IPW estimator is a singly robust estimator, because it is consistent when the treatment model is correctly specified (Tan et al. 2025).

There are issues with very unbalanced sample sizes when the propensity scores get close to 0 or 1 leading to potential violations of positivity. For example, at a propensity score of 0, subjects are very unlikely to have received the treatment, so there are very few subjects who actually have received the treatment and they have an outsized impact on the ATE estimate $\hat{\tau}$. Thus, the IPW estimator may be very unstable (Kurz 2022).

3.2 Outcome Modeling

Outcome models predict the potential outcomes Y (Tan et al. 2025). To de-bias the prediction of outcome Y, the outcome model accounts for other explanations (Cunningham n.d.). These alternative explanations for outcome Y may include the same explanations for propensity scores $e(X)$. For the sleep deprivation example, an outcome model to predict test scores might account for amount of time studied, general intelligence, access to academic resources, and presence of mental illness.

Linear regression models are a popular choice for outcome models. Looking at the equation of a multiple linear regression model where $\beta_0, \beta_1, ...\beta_p$ are the coefficients and $x_1, x_2, ..., x_p$

are the predictors:

$y_i = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon_i$.

It is already in the structure we want for causal inference. x_1 can represent the treatment, and $x_2, ..., x_p$ as alternative explanations for the outcome. We obtain an estimate for the ATE τ as β_1. (Miguel A. Hernan 2025). Linear regression is singly robust in estimating ATE τ, provided the model is correctly specified (Tan et al. 2022).

However, if there are substantial differences in the distribution of covariates between the treatment group and the control group, regression models struggle in establishing the treatment effect due to the high multicollinearity. Linear regression also requires parametric assumptions on the underlying distribution of the data (Tan et al. 2025).

4 Doubly Robust Estimation

Both propensity score modeling and outcome modeling are consistent for predicting ATE τ when their models are correctly specified, but in real-world settings, this is difficult to ascertain. Given such uncertainty, we prefer two chances to correctly specify the model instead of one (Xu & Zhao 2024).

Doubly robust estimation combines a propensity score model and an outcome model, allowing for a consistent estimate of ATE τ if only one of them is correctly specified. If both are correctly specified, then the estimator is semiparametrically efficient (Xu & Zhao 2024).

Doubly robust estimation was originally developed for missing data models, where it continues to be widely used. In missing data modeling, doubly robust estimators are consistent

if either the model for the missingness mechanism or the model for the distribution of complete data is correct (Bang & Robins 2005).

There are many different types of doubly robust estimators, which have different ways of combining the propensity score model and the outcome model. We discuss three—Augmented Inverse Probability Weighting, Targeted Maximum Likelihood Estimation, and Double Score Matching.

4.1 Augmented Inverse Probability Weighting

The AIPW estimator was developed by Robins et al. (1994) to improve upon the IPW estimator by reducing variability and increasing efficiency. While the IPW estimator only leverages a propensity score model, the AIPW estimator also uses outcome predictions (Kurz 2022).

The AIPW estimator has the following estimate for ATE:

$$\hat{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_1 X_i - \hat{\mu}_0 X_i + \frac{X_i}{\hat{e}(X_i)}(Y_i - \hat{\mu}_1 X_i) - \frac{1 - X_i}{1 - \hat{e}(X_i)}(Y_i - \hat{\mu}_0 X_i)].$$

This is a hefty equation, but can be understood by its two components:

1. $\frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_1 X_i - \hat{\mu}_0 X_i]$ is the direct regression adjustment. It is a weighted average of the two outcome model estimates, for the treatment group $X = 1$ and for the control group $X = 0$ (Kurz 2022).
2. $\frac{1}{n}\sum_{i=1}^{n}[\frac{W_i}{\hat{e}(X_i)}(Y_i - \hat{\mu}_1 X_i) - \frac{1-W_i}{1-\hat{e}(X_i)}(Y_i - \hat{\mu}_0 X_i)]$ is very similar to the IPW estimator. However, it is applied to the residuals, not the original outcomes (Kurz 2022).

Since the AIPW is a doubly robust estimator, it is consistent for ATE if either the propensity score model or the outcome model is correctly specified. The AIPW also has good

large-sample properties in that it is asymptotically normally distributed. Standard errors can be estimated through bootstrap (Kurz 2022).

However, AIPW has the same issues as IPW with data sparsity in the extreme ends creating potential violation of positivity (Kurz 2022).

4.2 Targeted Maximum Likelihood Estimation (TMLE)

Maximum likelihood estimation (MLE) fits a model to the data by minimizing a global measure, such as mean squared error. We may perform MLE for a particular parameter of interest, considered the target parameter; the other parameters are nuisance parameters. We prefer estimates with small bias and variance for the target parameter, even at the expense of increasing the bias and variance for nuisance parameters (Susan Gruber 2009).

Targeted MLE reduces bias in the estimate for the target parameter, the ATE τ in this case. Sometimes, this bias reduction procedure can increase the variance of the estimate, but often in finite samples the variance is reduced as well (Susan Gruber 2009). The "targeted" step involves computing the initial outcome estimates and then updating it (Tan et al. 2025).

The process is as follows:

First, initial outcome estimates of $Y^{X=1} = E[Y^{X=1}|L]$ and $Y^{X=0} = E[Y^{X=0}|L]$ and propensity scores $e(X_i)$ are calculated from outcome and propensity score models. Then, the inverse propensity—$H_1 = [e(X_i)]^{-1}$ for observations who have received and $H_0 = -[1-e(X_i)]^{-1}$ for observations which have not received the treatment—must be calculated for each observation. H_1 and H_0 are known as clever covariates. A logistic transformation can be applied to the ATE τ, which becomes $logit[E(Y^{X=1}|L)] = logit[\hat{E}(Y^{X=1}|L)] + \epsilon_1 H_1$ and

$logit[E(Y^{X=0}|L)] = logit[\hat{E}(Y^{X=0}|L)] + \epsilon_0 H_0$. These equations have very similar structure to logistic regression models. $\hat{E}(Y^{X=0}|L)$ and $\hat{E}(Y^{X=1}|L)$ can be seen as the intercepts, and the clever covariates H_1 and H_0 as the covariates for a logistic regression. ϵ_1 and ϵ_0 are the fluctuation parameters because they provides information about how much to change, or fluctuate, our initial outcome estimates, and they are estimated through fitting the logistic regression model. The clever covariate and fluctuation parameter allows us to update the the initial outcome estimate to become $logit[\hat{E}^*(Y^{X=1}|L)] = logit[\hat{E}(Y^{X=1}|L)] + \hat{\epsilon}_1 H_1$ and $logit[\hat{E}^*(Y^{X=0}|L)] = logit[\hat{E}(Y^{X=0}|L)] + \hat{\epsilon}_0 H_0$ (Hoffman 2020).

The final estimate for the ATE is given by

$$\hat{\tau}_{TMLE} = \frac{1}{n}\sum_{i=1}^{n}[\hat{E}^*(Y^{X=1}|L) - \hat{E}^*(Y^{X=0}|L)].$$

H_X is the vector of inverse propensity scores which correspond to the treatments the subjects actually received. $\hat{E}[Y^X|L]$ is the initial ATE estimate expected outcomes of all observations. $\hat{E}^*[Y^X|L] = \hat{E}[Y^X|L] + \hat{\epsilon} H_X$ is the updated expected outcome, given the treatments the subjects actually received. This term is used to estimate the standard error, stated below (Hoffman 2020).

The standard error for the TMLE is computed from an Influence Function, which calculates how much the final estimate for the ATE τ is influenced by each observation; it is an empirical estimate. The Influence Function is:

$$\hat{IF} = Y - \hat{E}^*(Y^X|L)H_X + \hat{E}^*(Y^{X=1}|L) - \hat{E}^*(Y^{X=0}|L) - \hat{\tau}_{TMLE}.$$

After computing the Influence Function, the standard error is (Hoffman 2020):

$$\hat{SE} = \sqrt{\frac{\hat{IF}}{n}}.$$

4.3 Double Score Matching (DSM)

Double score matching is a matching method which leverages propensity scores and prognostic scores (Tan et al. 2025).

Prognostic scores $\phi(X)$ are the outcomes under the condition of no treatment. They are estimated through an outcome model which is fit only on the $X = 0$ group. It is assumed that the predicted outcomes from this model represent the baseline risk, and generalize to all observations (Stuart et al. 2013). Prognostic scores $\phi(X)$ are sufficient for $Y^{X=0}$ in that $Y^{X=0} \perp\!\!\!\perp X|\phi(X)$ (Hansen 2008).

Both prognostic score matching and propensity score matching are popular causal inference methods. However, prognostic scores tend to demonstrate less extreme distributional differences between the treatment group $X = 1$ and the control group $X = 0$, allowing less observations to be wasted (Zhang et al. 2022).

Double score matching involves matching on both prognostic scores and propensity scores using Mahalanobis distance. It is doubly robust in that the estimator is consistent if either the propensity score model or the prognostic score model are correctly specified (Tan et al. 2025). The double score matching estimator is:

$$\hat{\tau}_{DSM} = \frac{1}{n}\sum_{i=1}^{n}(2X_i - 1)(1 + M^{-1}K_{S,i})Y_i.$$

The augmented score $S = [e(X)\phi(X)^T]^T$ is the matching variable which combines the propensity score and prognosis score. $J_{S,i}$ is the index set for the matched observations. $K_{S,i} = \sum_{i=1}^{n} I(i \in J_{S,i})$ is the number of times i is used in a match. M is the fixed number of matches (Yang & Zhang 2020).

Standard errors are typically estimated through bootstrap (Yang & Zhang 2020).

Monte Carlo Simulation Study

Simulation allows us to create data whilst directly fixing the ATE. This means that we know the true ATE, and we can test whether the various doubly robust estimators are able to capture this value. It is difficult to evaluate the performance of an estimator without knowing the true ATE, as doubly robust estimators do not have measures of performance. We also know the propensity score model and the outcome model, so we know when the doubly robust estimators have the correct model components. The performance of the doubly robust estimators for predicting ATE under different scenarios is compared in Table 1 and Table 2.

The following doubly robust estimators will be used and their corresponding R packages:

- Augmented Inverse Probability Weighting (`Rcal`)
- Targeted Maximum Likelihood Estimation (`tmle`)
- Double Score Matching (`dsmatch`)

Other singly-robust estimators will also be used to compare efficacy against

- Inverse Probability Weighting (`Rcal`) – propensity score-based
- Linear Regression – outcome-based

Simulations were performed using Rv4.5.1.

5 Data Generation

Six covariates $L_1, L_2, L_3, L_4, L_5, L_6$ were drawn from a standard normal distribution. Four more covariates L_7, L_8, L_9, L_{10} were drawn from an exponential distribution with rate

parameter 1. X was assigned 0 or 1 from a Bernoulli distribution where the probability of $X = 1$, or the propensity score was defined by the logistic regression equation $log(odds) = L_1 + L_2 + L_3$. The outcome Y was calculated from the linear equation $Y = X + L_3 + L_4 + L_5 + L_6 + \epsilon$, where ϵ represents random error drawn from a normal distribution with mean 0 and standard deviation 3. With this outcome model, the coefficient of X is 1, so the true ATE is $\tau = 1$. This process was inspired by Kurz (2022). We make 100 datasets with 1000 observations each, for each scenario. Conditions were verified for the propensity score model and the outcome model.

We look at all four scenarios of the propensity score model and the outcome model—both correctly specified, one is correctly specified but not the other, both misspecified.

6 Scenarios

6.1 Scenario 1: Both Propensity Score Model and Outcome Model Correct

Ideally, both models would be correctly specified. We used a logistic regression with covariates L_1, L_2, L_3 to estimate the propensity scores $e(X)$, and a linear regression with covariates X, L_3, L_4, L_5, L_6. Figure 2 shows that all estimators are centered around the true ATE, $\tau = 1$. IPW demonstrates the most variation. AIPW, Double score matching, Targeted MLE, and linear regression demonstrate small variation around $\hat{\tau} = 1$.

```
Warning: The `update` argument of `modify_header()` is deprecated as of gtsummary 2.0.0
i Use `modify_header(...)` input instead. Dynamic dots allow for syntax like
  `modify_header(!!!list(...))`.
```

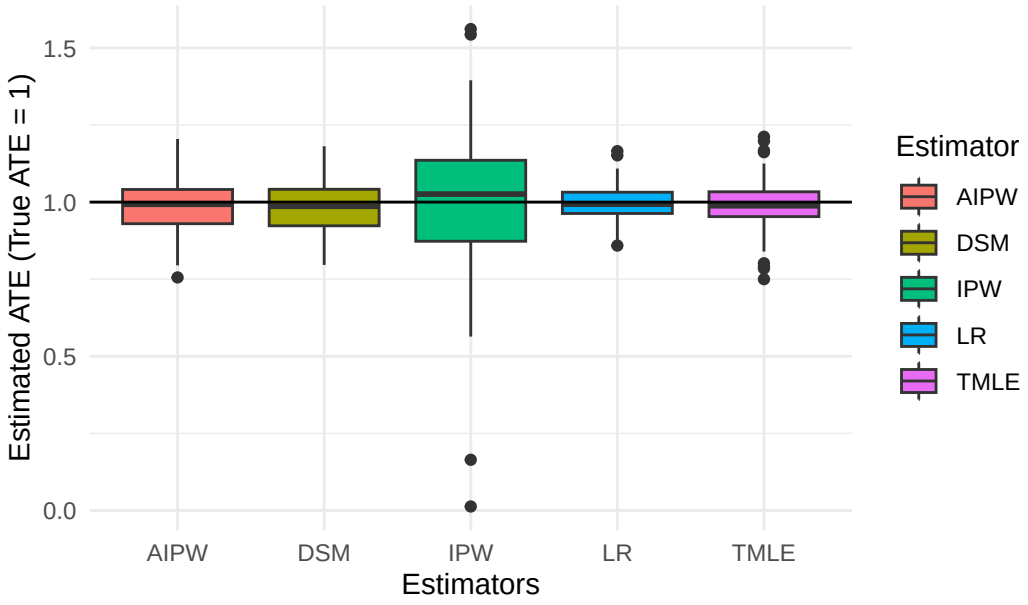

Figure 2: Scenario 1 Boxplots of Estimators

```
i The deprecated feature was likely used in the gtsummary package.

  Please report the issue at <https://github.com/ddsjoberg/gtsummary/issues>.
```

6.2 Scenario 2: Only Propensity Score Model Correct

In this scenario, the propensity score model is correctly specified and the outcome model is misspecified. We changed the outcome model to only include X, L_7, L_8, L_9, despite the true outcomes being generated from $X + L_3 + L_4 + L_5 + L_6 + \epsilon$. Outcome model misspecification is demonstrated with the inclusion of L_7, L_8, L_9, which were generated from an exponential distribution with rate parameter 1, but L_3, L_4, L_5, L_6 were generated from a standard normal. Linear regression was verified to pass conditions despite the non-normally distributed predictors. Figure 3 demonstrates that linear regression is biased, with estimates of ATE τ ranging from 1.30 to 1.94, completely failing to capture the true ATE of 1. Targeted MLE and double score matching became slightly biased, with a mean

estimate of 1.23 and 1.07 respectively. AIPW and IPW were unbiased. The variation in the estimates for Targeted MLE, double score matching, and linear regression increased compared to the first scenario.

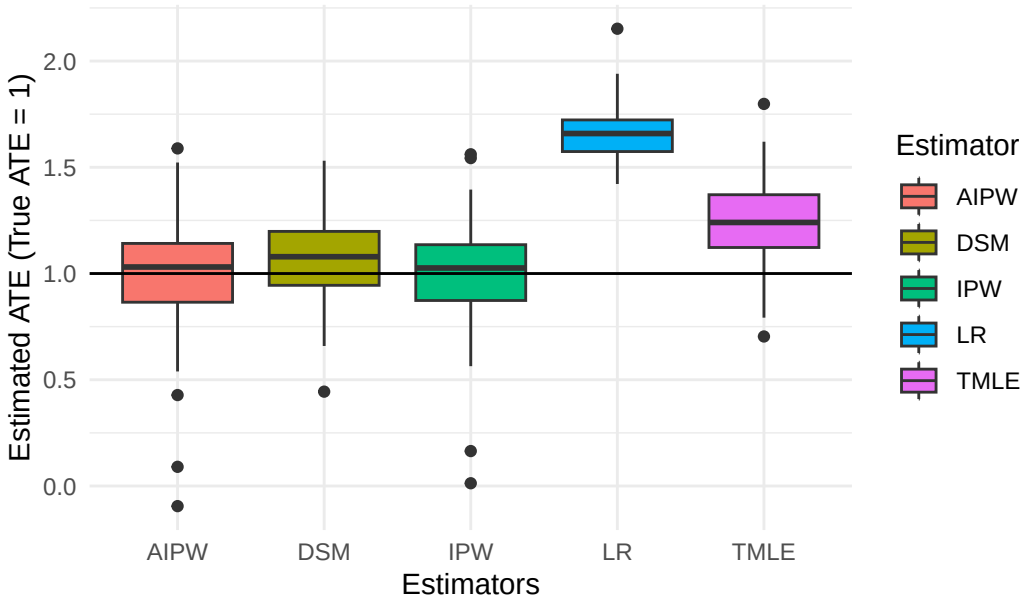

Figure 3: Scenario 2 Boxplots of Estimators

6.3 Scenario 3: Only Outcome Model Correct

In this scenario, the outcome model is correctly specified and the propensity score model is misspecified. We change the propensity score model to include only L_{10} as a covariate, despite the true outcomes being generated from a logistic regression with L_1, L_2, L_3. L_{10} was generated from an exponential distribution with rate parameter 1 whereas L_1, L_2, L_3 were generated from standard normal distributions. The conditions for the misspecified logistic regression model were verified. Figure 4 shows that only IPW is biased in this scenario. Its estimates range from 1.42 to 2.14, completely failing to capture ATE $\tau = 1$. The other four estimators perform very similarly, with similar variance and are all unbiased.

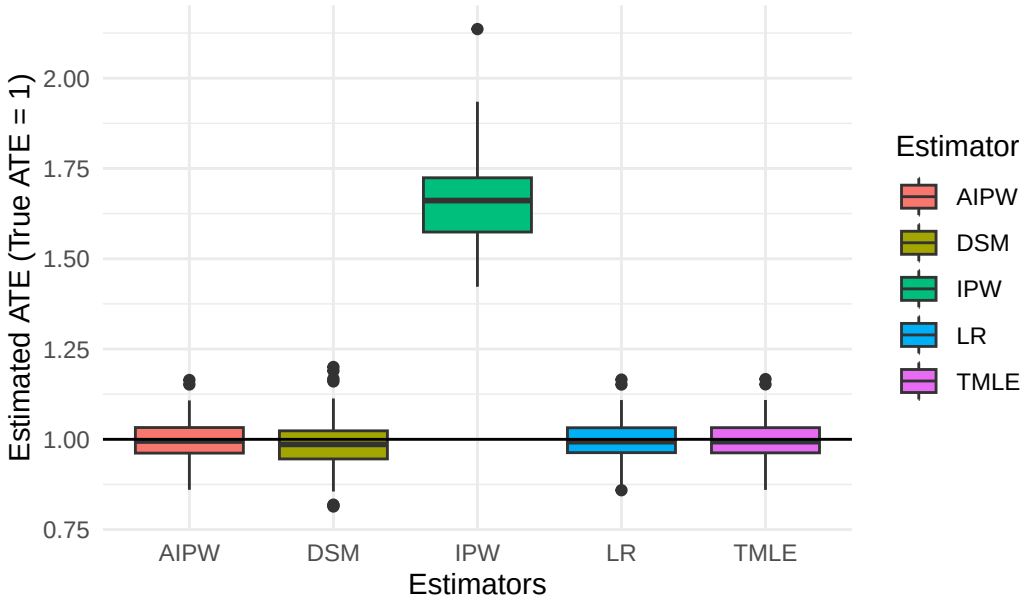

Figure 4: Scenario 3 Boxplots of Estimators

6.4 Scenario 4: Both Propensity Score Model and Outcome Model Incorrect

This is the most common scenario in predicting the ATE from observational studies, where both the propensity score model and the outcome model are misspecified (Alves n.d.). The incorrect specifications follow Scenarios 2 and 3. In Figure 5, all estimators are similarly biased. The variation is also similar. All of them failed to capture the true ATE $\tau = 1$.

7 Discussion

A few interesting results were demonstrated by our simulation.

Our results for the first three scenarios, where at least one model was correctly specified, were somewhat as expected. That is, the IPW estimator was biased when the propensity

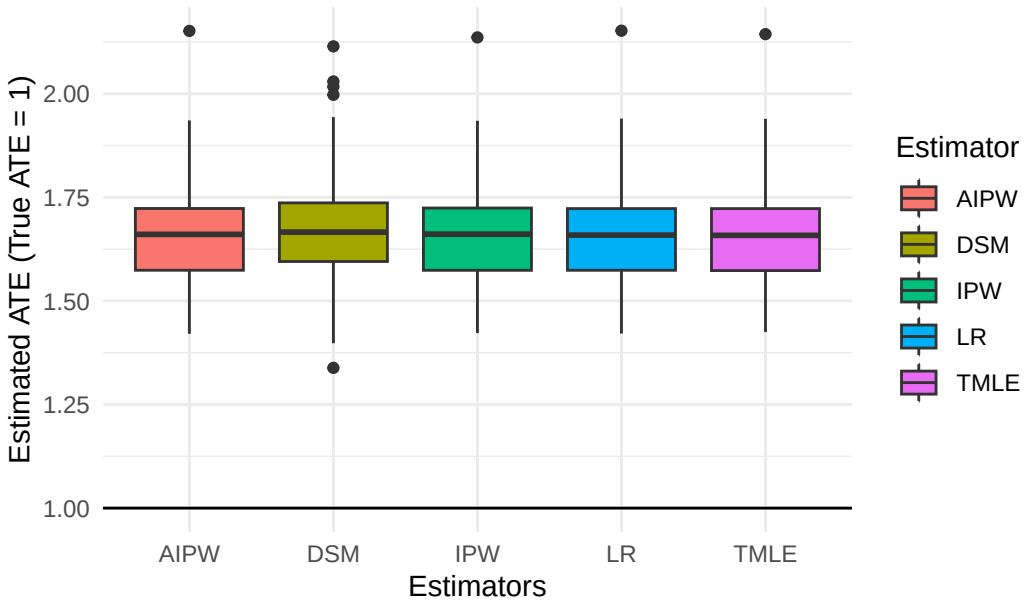

Figure 5: Scenario 4 Boxplots of Estimators

Table 1: Results of Estimators in Scenarios 1 and 2

Estimator	Scenario 1: Both Models Correct[1]	Scenario 2: Propensity Score Model Correc
AIPW	0.98 (0.08)	1.00 (0.25)
TMLE	0.99 (0.08)	1.24 (0.19)
DSM	0.99 (0.09)	1.06 (0.20)
IPW	1.00 (0.23)	1.00 (0.23)
LR	0.99 (0.06)	1.66 (0.12)

[1]Mean (SD)

Table 2: Results of Estimators in Scenarios 3 and 4

Estimator	Scenario 3: Outcome Model Correct[1]	Scenario 4: Both Models Incorrect[1]
AIPW	1.00 (0.06)	1.66 (0.12)
TMLE	1.00 (0.06)	1.66 (0.12)
DSM	0.99 (0.07)	1.67 (0.14)
IPW	1.66 (0.12)	1.66 (0.12)
LR	0.99 (0.06)	1.66 (0.12)

[1]Mean (SD)

score model was incorrectly specified. Linear regression was biased when the outcome model was incorrectly specified. The doubly robust estimators were unbiased for all three of these scenarios—with a caveat. Targeted MLE and double score matching were slightly biased when the outcome model was misspecified, with mean estimates $\hat{\tau}_{TMLE} = 1.24$ and $\hat{\tau}_{DSM} = 1.06$ respectively, suggesting that these methods may be more sensitive to outcome model misspecification than AIPW, which remained unbiased.

The standard deviations for all the doubly robust estimators increased substantially from scenario 1 to scenario 2, demonstrating less predictive accuracy when the outcome model is misspecified. Conversely, the standard deviations for the doubly robust estimators in scenario 3 are comparable, even slightly smaller than the standard deviations for scenario 1. The standard deviation for the IPW estimator also declined. The misspecified propensity score model may have contained less variation than the correctly specified model. While smaller variation allows for more precise estimation, it may fail to capture the true estimate.

In the fourth scenario, where both models were incorrectly specified, all estimators performed similarly. All five estimators had mean estimates around $\hat{\tau} = 1.66$, standard deviations of around 0.12, and failed to capture true ATE $\tau = 1$ in their range. While the estimators performed similarly when both the propensity score model and the outcome model were misspecified, their superiority in achieving unbiased or only slightly biased estimates when one of the models is wrong makes them a desirable choice for a causal inference study.

The propensity scores and outcomes for our simulation study were simulated with linear equations, but real-world data often does not follow a linear structure. Doubly robust machine learning methods, also known as double machine learning, which leverage nonparametric machine learning models such as SuperLearner have been exciting sites of analysis. Tan et al. (2025) demonstrated that nonparametric machine learning methods such as SuperLearner paired with doubly robust estimators improve their performance. Future work might extend our simulation study to evaluate double machine learning methods with nonlinear data.

References

Alsan, M. & Wanamaker, M. (2018), 'Tuskegee and the health of black men', *The quarterly journal of economics* **133**, 407–455.

Alves, M. F. (n.d.).

Bang, H. & Robins, J. M. (2005), 'Doubly robust estimation in missing data and causal inference models', *Biometrics* **61**, 962–973.

Cao, W., Tsiatis, A. A. & Davidian, M. (2009), 'Improving efficiency and robustness of

the doubly robust estimator for a population mean with incomplete data', *Biometrika* **96**(3), 723–734.

Cunningham, S. (n.d.), 'Causal inference: The mixtape'.

URL: *https://mixtape.scunning.com/*

Hansen, B. B. (2008), 'The prognostic analogue of the propensity score', *Biometrika* **95**(2), 481–488.

URL: *https://doi.org/10.1093/biomet/asn004*

Hoffman, K. (2020), 'KHstats - An Illustrated Guide to TMLE, Part II: The Algorithm — khstats.com', https://www.khstats.com/blog/tmle/tutorial-pt2#part-ii-the-algorithm. [Accessed 14-12-2025].

Kang, J. D. Y. & Schafer, J. L. (2007), 'Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data', *Statistical Science* **22**(4), 523–539.

URL: *http://www.jstor.org/stable/27645858*

Kurz, C. F. (2022), 'Augmented inverse probability weighting and the double robustness property', *Medical Decision Making* **42**(2), 156–167. PMID: 34225519.

URL: *https://doi.org/10.1177/0272989X211027181*

Michoel, T. & Zhang, J. D. (2023), 'Causal inference in drug discovery and development', *Drug discovery today* **28**.

Miguel A. Hernan, J. M. R. (2025), *Causal Inference: What if?*, CRC.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994), 'Estimation of regression coefficients when some regressors are not always observed', *J. Am. Stat. Assoc.* **89**(427), 846.

Rubin, D. B. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies', *J. Educ. Psychol.* **66**(5), 688–701.

Stuart, E. A., Lee, B. K. & Leacy, F. P. (2013), 'Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research', *J. Clin. Epidemiol.* **66**(8 Suppl), S84–S90.e1.

Susan Gruber, M. J. v. d. L. (2009), *Targeted Maximum Likelihood Estimation: A Gentle Introduction*.

Tan, X., Yang, S., Ye, W., Faries, D. E., Lipkovich, I. & Kadziola, Z. (2022), *Combining Doubly Robust Methods and Machine Learning for Estimating Average Treatment Effects for Observational Real-world Data*.

Tan, X., Yang, S., Ye, W., Faries, D. E., Lipkovich, I. & Kadziola, Z. (2025), 'Double machine learning methods for estimating average treatment effects: a comparative study', *J. Biopharm. Stat.* **35**(6), 1176–1195.

Tan, Z. (2008), 'Comment: Understanding OR, PS and DR'.

Wang, X., Panickan, V. A., Cai, T., Xiong, X., Cho, K., Cai, T. & Bourgeois, F. T. (2023), 'Endovascular aneurysm repair devices as a use case for postmarketing surveillance of medical devices', *JAMA Internal Medicine* **183**(10), 1090–1097.

Xu, T. & Zhao, J. (2024), 'Relaxed doubly robust estimation in causal inference', *Stat. Theory Relat. Fields* **8**(1), 69–79.

Yang, S. & Zhang, Y. (2020), 'Multiply robust matching estimators of average and quantile treatment effects'.

URL: *https://arxiv.org/abs/2001.06049*

Zhang, Y., Yang, S., Ye, W., Faries, D. E., Lipkovich, I. & Kadziola, Z. (2022), 'Practical recommendations on double score matching for estimating causal effects', *Stat. Med.* **41**(8), 1421–1445.