Demystifying Double Robust, Flexible Adjustment Methods for Causal Inference

Nathan I. Hoffmann

August 24, 2023

Abstract

Double robust methods for flexible covariate adjustment in causal inference have proliferated in recent years. Despite their apparent advantages, these methods remain underutilized by social scientists. It is also unclear whether these methods actually outperform more traditional methods in finite samples. This paper has two aims: It is a guide to some of the latest methods in double robust, flexible covariate adjustment for causal inference, and it compares these methods to more traditional statistical methods. It does this by using both simulated data where the treatment effect estimate is known, and then using comparisons of experimental and observational data from the National Supported Work Demonstration. Methods covered include Augmented Inverse Propensity Weighting, Targeted Maximum Likelihood Estimation, and Double/Debiased Machine Learning. Results suggest that these methods do not necessarily outperform OLS regression or matching on propensity score estimated by logistic regression, even in cases where the data generating process is not linear.

Introduction

Statistical methods for flexible covariate adjustment in causal inference have proliferated in recent years. These methods have a number of strengths over traditional regression methods: They make few functional form assumptions, can accommodate large numbers of covariates, and produce easily interpretable treatment effect estimates. Many of these methods also have a "double robust" property: They estimate one model for the treatment exposure and another for the outcome, and as long as at least one is correctly specified, then the treatment effect will be estimated consistently. Despite their apparent advantages, these methods remain underutilized by social scientists. Part of the barrier has been lack of familiarity with these methods. It has also been unclear how these methods compare, or whether such methods actually perform better than traditional methods in finite samples.

This paper makes advances on these fronts. First, it is a guide to some of the latest methods in double robust, flexible covariate adjustment for causal inference, explaining the methods to a social scientist audience. Second, it compares these methods to more traditional statistical methods using both simulations (Dorie et al., 2019) and the National Support for Work Demonstration (NSW) originally analyzed by LaLonde (1986).

Methods covered include Targeted Maximum Likelihood Estimation (TMLE, van der Laan & Rubin, 2006), Double or Debiased Machine Learning (DML, Chernozhukov et al., 2018), and Augmented Inverse Propensity Weighting (AIPW, Glynn & Quinn, 2010). This paper reviews the theory behind these methods as well as simple R implementations of them on simulations and real data. These methods are compared to two methods commonly used by social scientists: ordinary least squares (OLS) regression, and matching on propensity scores estimated from logistic regression (PSM).

Conceptual Overview

Double robust methods estimate two models:

- an outcome model

$$\mu_d(X_i) = E(Y_i \mid D_i = d, X_i)$$

- and an exposure model (or treatment model or propensity score):

$$\pi(X_i) = E(D_i \mid X_i)$$

where $\mu_d(\cdot)$ is the model of control or treatment $D_i = d = \{0, 1\}$, X_i is a vector of covariates for unit $i = 1, \ldots, N$ for treatment (1) and control (0), Y_i is the outcome, and $\pi(\cdot)$ is the exposure model. The covariates included in X_i can be different for the two models.

An estimator is called "double robust" if it achieves consistent estimation of the ATE (or whatever estimand the researcher is interested in) as long as at least one of these two models is consistently estimated. This means that the outcome model can be completely misspecified, but as long as the exposure model is correct, our estimation of the ATE will be consistent. This also means that the exposure model can be completely wrong, as along as the outcome model is correct.

Origins of Doubly Robust Methods

According to Bang & Robins (2005), double robust methods have their origins in missing data models. Robins et al. (1994) and Rotnitzky et al. (1998) developed augmented orthogonal inverse probability-weighted (AIPW) estimators in missing data models, and Scharfstein et al. (1999) showed that AIPW was double robust and extended to causal inference.

But Kang & Schafer (2007) argue that double robust methods are older. They cite work by Cassel et al. (1976), who proposed "generalized regression estimators" for population means from surveys where sampling weights must be estimated. Arguably, double robust methods go back even further than this. The form of double robust methods is similar to residual-on-residual regression, which dates back to Frisch & Waugh (1933) famous FWL theorem:

$$\beta_D = \frac{\text{Cov}(\tilde{Y}_i, \tilde{D}_i)}{\text{Var}(\tilde{D}_i)}$$

where $\tilde{D}_i$ is the residual part of D_i after regressing it on X_i, and $\tilde{Y}_i$ is the residual part of Y_i after regressing it on X_i. This formulation writes the regression coefficient as composed of an outcome model ($\tilde{Y}_i$) and exposure model ($\tilde{D}_i$), the two models used in double robust estimators.

There are also links between double robust methods and matching with regression adjustment. This work goes back to at least Rubin (1973), who suggested that regression adjustment in matched data produces less biased estimates that either matching (exposure adjustment) or regression (outcome adjustment) do by themselves.

Assumptions

Most double robust methods require almost all of the standard assumptions necessary formost methods that depend on selection on observables. Although some double robust methods relax one or two of these, the six standard assumptions are:

1. Consistency
2. Positivity/overlap
3. One version of treatment
4. No interference
5. IID observations
6. Conditional ignorability: $\{Y_{i0}, Y_{i1}\} \perp\!\!\!\perp D_i \mid X_i$

Special attention should be paid to Assumption 6: double robust methods will not work if we do not measure an important confounder that affects both treatment and exposure. But notably, the double robust methods covered in this tutorial make no functional form assumptions. These methods are designed to incorporate flexible machine learning algorithms to estimate both the outcome and exposure models, with regularization (often through cross-fitting) to avoid overfitting.

Overview of Techniques

Each of the methods reviewed in this paper can be thought of as a collection of estimation techniques. Each involves a model for the outcome and another for the treatment exposure, but the ways these relate and are combined varies from method to method. Choice of estimation technique for these two models is left to the discretion of the user; often ensemble learning is recommended, but in practice simpler methods can also work well.

Augmented Inverse Propensity Weighting (AIPW)

The oldest of these modern methods, AIPW arose in the context of missing data imputation. The method simply combines estimates from a model for the treatment exposure, $\pi(X)$, and a model for the outcome, $\mu(X)$. The name comes from the close similarity to inverse propensity weights (IPW), but whereas IPW only weights for propensity of treatment, AIPW "augments" these weights with an estimate of the response surface as well.

Formally, the model can be written as the difference between an estimated outcome for treated units and the estimated outcome for untreated units (see the demonstration below):

$$\widehat{ATE}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i(Y_i - \hat{\mu}_1(\mathrm{X}_i))}{\hat{\pi}(\mathrm{X}_i)} + \hat{\mu}_1(\mathrm{X}_i)\right) - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(1-D_i)(Y_i - \hat{\mu}_0(\mathrm{X}_i))}{1-\hat{\pi}(\mathrm{X}_i)} + \hat{\mu}_0(\mathrm{X}_i)\right)$$

Glynn & Quinn (2010) provide an alternate formula, where the basic inverse probability weight (IPW) estimator (which incorporates only the exposure model $\hat{\pi}$) is corrected using a weighted average of two outcome regression estimates:

$$\widehat{ATE}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left[\frac{D_i Y_i}{\hat{\pi}(\mathrm{X}_i)} - \frac{(1-D_i)Y_i}{1-\hat{\pi}(\mathrm{X}_i)}\right] - \frac{D_i - \hat{\pi}(\mathrm{X}_i)}{\hat{\pi}(\mathrm{X}_i)(1-\hat{p}i(\mathrm{X}_i))}[(1-\hat{\pi}(\mathrm{X}_i))\hat{\mu}_1(\mathrm{X}_i) + \hat{\pi}(\mathrm{X}_i)\hat{\mu}_0(\mathrm{X}_i)]\right\}$$

Targeted Maximum Likelihood Estimation (TMLE)

TMLE begins by estimating the relevant part of the data-generating distribution $P(Y)$, i.e. the conditional density $Q = P(Y \mid X)$. It next estimates the exposure model. Although any estimation method can be used for these steps, the originators of the method suggest using a "super learner," i.e. ensemble learning with cross-validation. Next, the exposure model is used to calculate a "clever covariate," which is similar to an IPW. The coefficient for this clever covariate is estimated using maximum likelihood – whence the "MLE" in "TMLE." Finally, the estimate of Q is updated in a function involving the clever covariate. This process can be iterated, but usually one iteration is enough. The estimate of the distribution Q can be used to calculate the estimand of interest.

Formulaly, first generate estimates of $\mu_d(\mathrm{X}_i) = E(Y \mid D = d, \mathrm{X}_i)$ and $\pi(\mathrm{X}_i) = P(D = 1 \mid \mathrm{X}_i)$. Then create variable for targeting step:

$$H_{di} = \frac{I(D_i = 1)}{\hat{\pi}(\mathrm{X}_i)} - \frac{I(D_i = 0)}{1-\hat{\pi}(\mathrm{X}_i)}$$

Next, calculate the clever covariates for each individual in the data. These quantities are similar to inverse probability weights:

$$H_{1i}(D = 1, \mathrm{X}_i) = \frac{d_i}{\hat{\pi}(\mathrm{X}_i)}, \quad H_{0i}(D = 0, \mathrm{X}_i) = \frac{1-d_i}{1-\hat{\pi}(\mathrm{X}_i)}.$$

Then estimate fluctuation parameters $\epsilon = (\epsilon_0, \epsilon_1)$ through maximum likelihood of the following logistic regression with fixed intercept $\mathrm{logit}(\mu_{di})$:

$$\mathrm{logit}[E(Y = 1 \mid D, \mathrm{X})] = \mathrm{logit}(\hat{\mu}_{di}) + \epsilon_0 H_{0i} + \epsilon_1 H_{1i}$$

Then generate updated ("targeted") estimates of potential outcomes:

$$\hat{\mu}_1^*(\mathrm{X}_i) = \text{expit}[\text{logit}(\hat{\mu}_1(\mathrm{X}_i)) + \hat{\epsilon}H_{1i}]$$
$$\hat{\mu}_0^*(\mathrm{X}_i) = \text{expit}[\text{logit}(\hat{\mu}_0(\mathrm{X}_i)) + \hat{\epsilon}H_{0i}]$$

Finally, estimate targeted parameter (ATE):

$$\widehat{ATE}_{TMLE} = \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_1^*(\mathrm{X}_i) - \hat{\mu}_1^*(\mathrm{X}_i)]$$

Double or Debiased Machine Learning* (DML)

The most recent of the methods reviewed here, DML is motivated by the need to handle problems with high-dimensional nuisance parameters, i.e. a large number of measured confounders. Flexible machine learning is appropriate for this task, but such methods suffer from regularization bias. DML removes this bias in a two-step procedure. First, it solves the auxiliary problem of estimating the treatment exposure model $E(D|X) = \pi(X)$. It then uses this model to remove bias: Neyman orthogonalization allows the creation of an orthogonalized regressor, essentially partialing out the effect of covariates X from treatment D. The debiased D is then used to estimate the conditional mean of the outcome $E(Y \mid X) = \mu(X)$, which can be used to calculate the estimand of interest.

We can motivate this as follows. Suppose we want to estimate δ in the following framework:

$$y_i = \delta d_i + g_0(\mathrm{x}_i) + u_i,$$

$$d_i = m_0(\mathrm{x}_i) + v_i.$$

The idea is to estimate g_0 and m_0 separately, then use residual-on-residual regression to obtain an estimate of δ. However this leaves a term in the asymptotic distribution of $\hat{\delta}$ that biases the estimate. To avoid this, DML uses sample splitting.

We randomly split the sample of n observations into two sets, I and I^c, each of size $n/2$. We then estimate the response and treatment models using only set I^c:

1) Estimate $d_i = \hat{m}_0(\mathrm{x}_i) + \hat{v}_i, i \in I^c$.

2) Estimate $y_i = \hat{g}_0(\mathrm{x}_i) + \hat{u}_i, i \in I^c$, without using d_i.

Next, we use the estimated models to perform residual-on-residual regression on the left out set I to obtain an estimate of δ:

$$\hat{\delta}(I^c, I) = \left(\sum_{i\in I}\hat{v}_i d_i\right)^{-1}\sum_{i\in I}\hat{v}_i(y_i - \hat{g}_0(\mathrm{x}_i)).$$

Using half the sample results in efficiency loss. To rectify this, we repeat the above procedure, switching the split sets. We then have $\delta(I^c, I)$ and $\delta(I, I^c)$. The cross-fitting DML estimator is:

$$\widehat{ATE}_{DML} = \frac{\hat{\delta}(I^c, I) + \hat{\delta}(I, I^c)}{2}.$$

A simple demonstration using AIPW

To demonstrate double robustness, this section presents one of the simpler double robust estimators, AIPW (Glynn & Quinn, 2010). As shown above, we can write this estimator as follows:

$$\widehat{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} + \hat{\mu}_1(X_i)\right) - \frac{1}{N}\sum_{i=1}^{N}\left(\frac{(1-D_i)(Y_i - \hat{\mu}_0(X_i))}{1-\hat{\pi}(X_i)} + \hat{\mu}_0(X_i)\right)$$

For each individual in the sample, this estimator calculates two quantities:

- The treated potential outcome

$$\hat{Y}_{1i} = \frac{D_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} + \hat{\mu}_1(X_i)$$

- The control potential outcome

$$\hat{Y}_{0i} = \frac{(1 - D_i)(Y_i - \hat{\mu}_0(X_i))}{1 - \hat{\pi}(X_i)} + \hat{\mu}_0(X_i)$$

Let's focus on the treated model:

$$\hat{Y}_{1i} = \frac{D_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} + \hat{\mu}_1(X_i)$$

First, assume that the outcome model $\mu_1(X_i)$ is correctly specified and the exposure model $\pi(X_i)$ is incorreclty specified. Let's also assume (for now) that we're dealing with a treated unit, i.e. $D_i = 1$. Then

$$\hat{\mu}_1(X_i) = Y_i$$

and hence

$$\hat{Y}_{1i} = \frac{D_i(0)}{\hat{\pi}(X_i)} + \hat{\mu}_1(X_i) = \hat{\mu}_1(X_i).$$

So the model relies only on the outcome model. The incorrectly specified exposure model completely disappears from the equation. If we're dealing with a control unit ($D_i = 0$), we get the same result:

$$\hat{Y}_{1i} = \frac{0(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} + \hat{\mu}_1(X_i) = \hat{\mu}_1(X_i).$$

Now, what if the exposure model $\pi(X_i)$ is correctly specified and the outcome model $\mu_1(X)$ is incorrect? First, we rewrite the estimator for the treated outcome:

$$\begin{aligned}
\hat{Y}_{1i} &= \frac{D_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} + \hat{\mu}_1(X_i) \\
&= \frac{D_i Y_i}{\hat{\pi}(X_i)} - \frac{D_i \hat{\mu}_1(X_i)}{\hat{\pi}(X_i)} + \frac{\hat{\pi}(X_i)\hat{\mu}_1(X_i)}{\hat{\pi}(X_i)} \\
&= \frac{D_i Y_i}{\hat{\pi}(X_i)} - \left(\frac{D_i - \hat{\pi}(X_i)}{\hat{\pi}(X_i)} \right) \hat{\mu}_1(X_i). \qquad (*)
\end{aligned}$$

Since the exposure model is correclty specified, we have $D_i = \hat{\pi}(X_i)$ on average, so

$$E[D_i - \hat{\pi}(X_i)] = 0.$$

This means that the second term in equation $(*)$ is 0, so

$$E[\hat{Y}_{1i}] = E\left[\frac{D_i Y_i}{\hat{\pi}(X_i)} \right].$$

This shows that when the exposure model is correct, then the estimator depends only on the exposure model. We can make similar arguments for the control model $\hat{Y}_{0i}$.

This demonstration shows that this estimator achieves double robustness: the estimator is robust to misspecification of either the exposure or the outcome model (but not both).

Methods

These methods have many similarities. How do the results they give compare? This section tests the performance of each in practice using two methods. First, results are compared using simulated data from a causal inference competition (Dorie et al. (2019)). The true treatment effect is known, so these simulations allow assessment of accuracy. Second, these methods are applied to data from LaLonde's (1986) study of the National Supported Work Demonstration (NSW). The NSW randomly provided training to disadvantaged workers, allowing an experimental estimate of the effect of the intervention, and data assembled by Dehejia & Wahba (1999) allows these experimental estimates to be compared to observational ones.

The three double robust methods are compared to two traditional estimation methods used as benchmarks: linear regression estimated using ordinary least squares regression ("OLS") and propensity-score matching with scores estimated from logistic regression using the `MatchIt` package ("PSM").

Double robust can use a variety of techniques to estimate the underlying treatment and outcome models. The simulation results below compare three estimation techniques. First is using a logistic regression for the exposure model and an OLS regression for the outcome model. Second is generalized random forests (GRF, Athey et al., 2019) using the `grf` R package, with separate models for exposure and outcome. The final technique is the SuperLearner (as promoted by the makers of TMLE) using the `SuperLearner` package, again with separate models for exposure and outcome. GLM, glmnet (lasso), and XGBoost models are considered for the SuperLearner.

For the simulations, I have coded original functions to to implement the three double robust methods. This allows the use of the same exposure and outcome models' predictions in implementing these methods, permitting maximum comparability. However, the evaluation of the simulations only considers point estimates and not standard errors. For evaluation of the experimental LaLonde data, I use availabel R packages that calculate standard errors, using generalized random forests as the estimation technique. The augmented inverse propensity weighted estimator and targeted maximum likelihood estimation use the `grf` package, and double/debiased machine learning uses `DoubleML` and `mlr3` packages.

Results

Simluations with Dorie et al. (2019) Data

In 2016, the Atlantic Causal Inference Conference hosted a competition for causal inference methods that adjust on observables. Dorie et al. (2019) published the results of this competition, along with the data used in the competition. Below, I test double robust methods on the 20 data sets used for the "do-it-yourself" part of the competition. The data represent a hypothetical twins study investigating the impact of birth weight on IQ. The data have 4802 observations and 52 covariates. The authors of the study specify a different data generating process for the potential outcomes in each data set. In all cases, ignorability holds, but the authors vary the following:

- degree of nonlinearity
- percentage of treated
- overlap for the treatment group
- alignment (correspondence between the assignment mechanism and the response surface)
- treatment effect heterogeneity
- overall magnitude of the treatment effect

The 20 data sets used here cover a range of these attributes; see the supplemental material from Dorie et al. (2019) for details. I use 10 simulations of each data set, resulting in 200 data sets. I also use 100 simulations of each of the two datasets with linear data generating processes (numbers 1 and 3) to show how these methods perform when the underlying models are linear. I then calculate bias, percent bias (the estimator's bias as a percentage of its standard error), root mean squared error (rmse), and median absolute error (mae). I also present the number of datasets for which the method fails and the average computation time for each data set, in seconds.

Results for the full range of simulations are shown in Table 1 and Figure 1. Using AIPW as written above results in some wildly biased estimates for the ols_logit estimator, due to dividing by some very small

Table 1: Results of Monte Carlo simulations using the first 20 datasets from Dorie et al. (2019), 10 replications each. Percent bias is calculated as the estimator's bias as a percentage of its standard error, rmse is root mean squared error, mae is median absolute error, and comp_time is average computation time measured in seconds for each dataset.

method	estimator	bias	percent_bias	rmse	mae	comp_time
ols	NA	0.25	0.157	0.74	0.41	0.055
psm	NA	0.20	0.131	0.86	0.53	0.672
aipw	ols_logit	-8.24	-1.547	10.06	7.63	0.542
aipw	grf	0.06	0.039	0.45	0.22	33.092
aipw	superlearner	-0.50	-0.170	2.87	0.72	114.177
tmle	ols_logit	-1.58	-0.676	2.27	1.47	0.569
tmle	grf	0.35	0.230	0.58	0.33	33.110
tmle	superlearner	0.60	0.038	15.80	3.90	114.189
dml	ols_logit	1.37	0.168	8.18	1.04	0.697
dml	grf	1.39	0.171	8.12	1.15	36.458
dml	superlearner	1.40	0.172	8.16	1.25	122.503

propensity scores. Hence I present estimates from a trimmed AIPW estimator, where predicted exposure model values are set to 0.01 if they are less than 0.01 and to 0.99 if they are greater than 0.99.

The lowest bias is achieved by AIPW with GRF, followed by propensity score matching and then OLS. The lowest root mean squared error is obtained by AIPW with GRF, followed by TMLE with GRF, and next by OLS regression. Results are similar if we consider percent bias or mean absolute error. DML does not perform particularly well and also has the highest computation times due to sample splitting.

If we consider only the linear datasets (Table 2 and Figure 2), OLS obtains the lowest bias, followed by AIPW with GRF and then propensity score matching. The lowest root mean squared error is achieved by OLS, then AIPW with GRF, and then TMLE with GRF. Percent bias and mean absolute error show similar results.

Overall, traditional methods perform surprisingly well in comparison with the double robust methods. Even in the full range of datasets – which include highly nonlinear exposure and outcome data-generating processes – OLS and propensity score matching obtain some of the smallest bias. Generalized random forests appear to be the best estimators for the double robust methods, though a SuperLearner than incorporates GRF has the potential to outperform GRF alone, at the expense of computation time. Notably, the method with the longest computation time – DML with a SuperLearner – takes over 2,000 times as along as OLS.

Comparisons using LaLonde NSW Data

As another test, I use data from LaLonde's (1986) study of the National Supported Work Demonstration (NSW), as provided by Dehejia & Wahba (1999). Between March 1975 and July 1977, the NSW randomly provided training to disadvantaged workers. LaLonde used earnings in 1978 as the outcome of interest; comparing earnings in this year for treated and untreated workers allows an experimental estimate of the effect of the intervention. Restricting the sample to men, this study had 297 treated and 425 control participants. Covariates include age, education in years of schooling, earnings in 1975, and dichotomous variables for Black and Hispanic race, married, and not having a high school degree. Following Dehejia & Wahba (1999), I add a variable indicating whether each respondent's earnings in 1975 was \$0 – i.e., they were unemployed.

LaLonde compared these experimental estimates to control samples drawn from the Panel Study of Income Dynamics (PSID) and Westat's Matched Current Population Survey-Social Security Administration File (CPS). The PSID-1 sample (n = 2,490) contains all male household heads under 55 who did not classify themselves as retired in 1975, and the PSID-3 sample (n = 128) further restricts this to men who were not working in spring of 1976 or 1975. The CPS-1 sample (n = 15,992) includes all CPS males under 55, and CPS-3 (n = 429) restricts this two those who were not working in March 1976 whose earnings in 1975 was

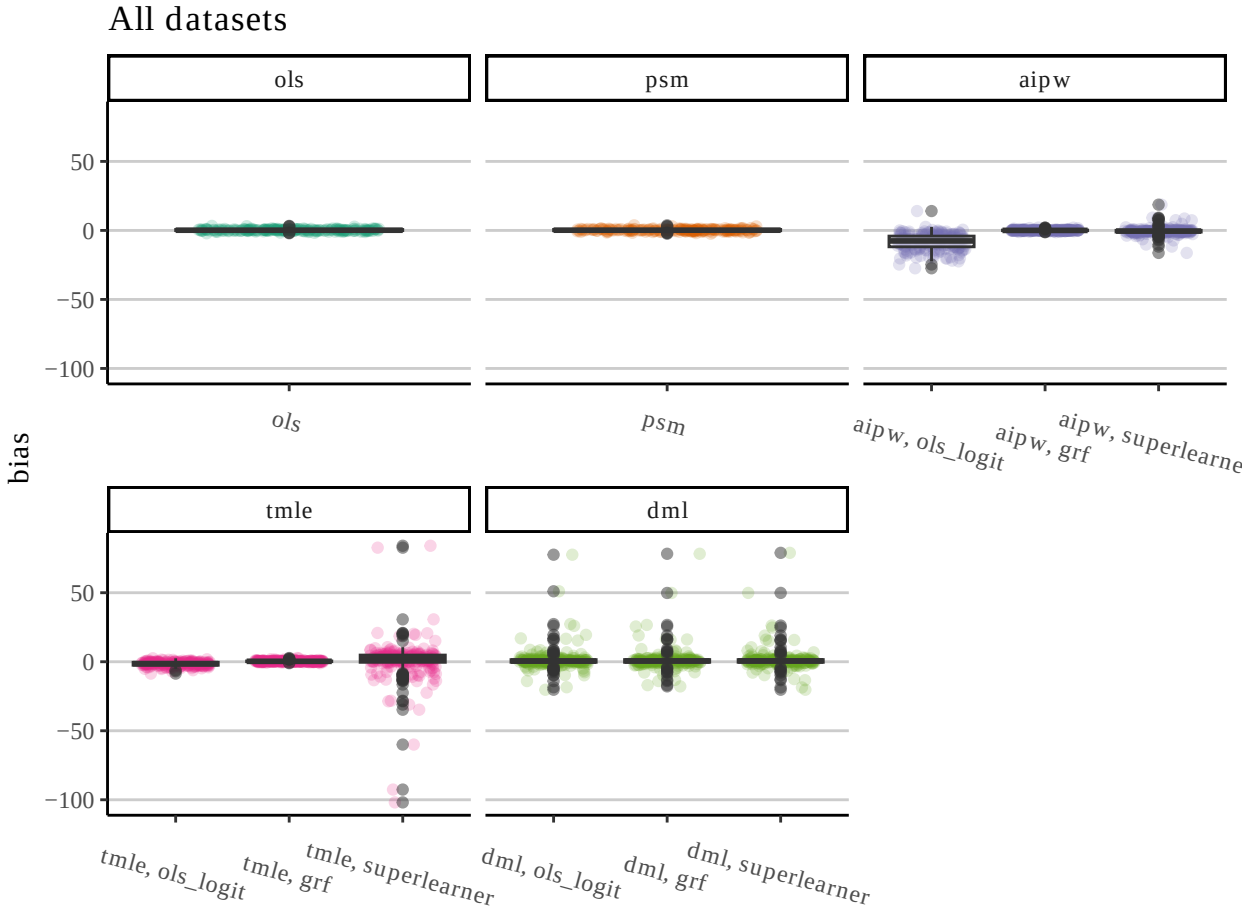

Figure 1: Results of Monte Carlo simulations using the first 20 datasets from Dorie et al. (2019), 10 replications each.

Table 2: Results of Monte Carlo simulations using the two datasets from Dorie et al. (2019), with linear data generating processes, 100 replications each ("linear"). Percent bias is calculated as the estimator's bias as a percentage of its standard error, rmse is root mean squared error, mae is median absolute error, and comp_time is average computation time measured in seconds for each dataset.

method	estimator	bias	percent_bias	rmse	mae	comp_time
ols	NA	-0.024	-0.017	0.391	0.12	0.055
psm	NA	-0.053	-0.039	0.522	0.12	0.672
aipw	ols_logit	-3.435	-1.012	4.755	2.06	0.610
aipw	grf	-0.035	-0.024	0.402	0.23	30.774
aipw	superlearner	-0.637	-0.535	0.934	0.57	114.216
tmle	ols_logit	-0.536	-0.339	0.905	0.33	0.632
tmle	grf	0.305	0.220	0.483	0.33	30.790
tmle	superlearner	0.421	0.098	4.507	2.41	114.237
dml	ols_logit	1.177	0.597	1.777	1.27	0.707
dml	grf	1.187	0.647	1.611	1.25	31.794
dml	superlearner	1.191	0.653	1.611	1.25	126.777

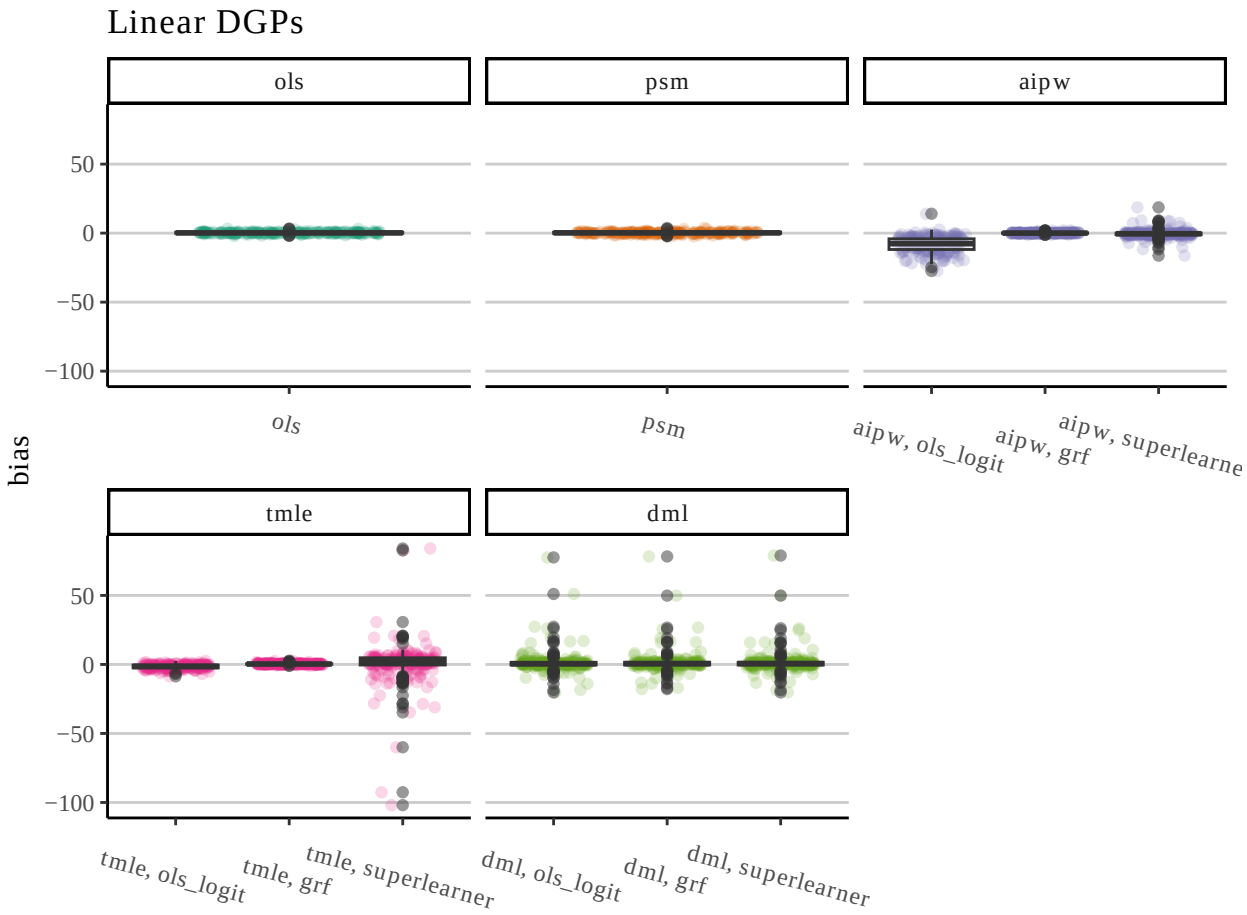

Figure 2: Results of Monte Carlo simulations using the two datasets from Dorie et al. (2019), with linear data generating processes, 100 replications each ("linear").

Table 3: ATE estimates for Lalonde NSW data as provided by Dehejia and Wahba (1999), with CPS and PSID comparison groups. Standard errors shown in parentheses. Covariates include age, education in years of schooling, earnings in 1975, and dichotomous variables for Black and Hispanic race, married, not having a high school degree, and having no earnings in 1975. The "With 1974 earnings" estimates additionally include earnings in 1974 as a covariate, along with an indicator for having no earnigns in 1974.

sample	method	experimental	PSID-1	PSID-3	CPS-1	CPS-3
Original LaLonde	ols	807 (468)	-1458 (802)	518 (1011)	-993 (452)	-993 (452)
Original LaLonde	psm	805 (520)	-2438 (824)	-1411 (1676)	-758 (574)	-758 (574)
Original LaLonde	aipw	791 (501)	-1799 (859)	-1415 (1281)	-301 (514)	-301 (514)
Original LaLonde	tmle	772 (490)	-764 (1334)	-2002 (2530)	335 (562)	335 (562)
Original LaLonde	dml	702 (495)	-3307 (793)	-516 (1061)	-1058 (486)	-1058 (486)
With 1974 earnings	ols	1676 (639)	752 (915)	1833 (1160)	699 (548)	699 (548)
With 1974 earnings	psm	1875 (671)	1084 (824)	2079 (1522)	1679 (720)	1679 (720)
With 1974 earnings	aipw	1844 (684)	1223 (777)	1263 (938)	1530 (701)	1530 (701)
With 1974 earnings	tmle	1803 (672)	1922 (1418)	2072 (2697)	1954 (734)	1954 (734)
With 1974 earnings	dml	1757 (661)	-657 (959)	455 (1091)	916 (687)	916 (687)

below the poverty level. Restricting these observational samples gets closer to the group eligible for the NSW.

Following Dehejia & Wahba (1999), I present results for the original samples analyzed by LaLonde (1986), but I also include results using a subsample of the experimental group that has 1974 earnings data available (185 treated and 260 control participants) and include this additional covariate, along with an indicator variable for no earnings in 1974.

I again compare the three double robust methods to a linear model fitted by OLS ("ols") and propensity score matching using logistic regression ("psm"). Results are presented in Table 3.

The "experimental" column provides a baseline for the comparison, suggesting that the program resulted in an earnings gain of about $700 to $800. If selection on observables holds, then we should be able to recover these estimates from the non-experimental control groups. Most of the methods do not perform very well; the only ones to estimate a positive treatment effects are OLS for the PSID-3 sample and TMLE for the CPS samples, and these are still smaller than the experimental baselines.

Including 1974 earnings data results in much better estimates with the observational control groups. For PSID-1, TMLE provides estimates closest to the experimental ones, while DML performs much worse than even OLS. For PSID-3, OLS and TMLE both perform well. For CPS-1 and CPS-3, TMLE performs the best.

These results highlight the importance of selection on observables holding. Without including 1974 earnings as a covariate, it appears that selection on observables does not hold, as most methods provide highly inaccurate estimates with the wrong sign. Once 1974 earnings are included, all of the methods provide estimates much closer to the experimental values. PSM, AIPW, and TMLE all do fairly well, while OLS and DML are more unstable across samples.

Conclusion

This paper aimed to provide an introduction to and evaluation of double robust methods for covariate adjustment in causal inference. By comparing AIPW, TMLE, and DML to the more traditional statistical methods of OLS and PSM, it allows evaluation of whether these methods are worth the effort and (computational) time for social scientists to adopt them.

Results are mixed. On the one hand, in the full range of simulated data, AIPW with GRF estimators achieves the lowest bias and root mean squared error. This is the simplest (and most computationally efficient) of the double robust methods, although GRF is computationally expensive. TMLE with GRF also does fairly well. But OLS and PSM have strong showings, in the top three smallest bias or root mean

squared error. In the datasets with linear data-generating processes, OLS achieves the lowest bias and root mean squared error.

In the experimental LaLonde data, OLS does not perform as strongly. Although it is one of the few methods to provide experimental estimates with the correct sign in the original LaLonde set of covariates, its estimates are highly unstable across samples. When 1974 earnings are used as a covariate, PSM, AIPW, and TMLE all perform fairly well, though TMLE comes out on top.

Overall, these double robust methods do not show a clear advantage over simpler, more computationally efficient methods such as OLS and PSM. While they are useful for social scientists to understand, in most applications, OLS or PSM provide similar results. However, in certain circumstances these double robust methods may be a better choice. Especially when paired with a highly flexible estimator like GRF, these methods can be useful when the number of covariates is high or even exceeds the number of observations. If the researcher believes the data-generating process is highly nonlinear, they may also be a good a choice, at least as a sensitivity check.

References

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. The Annals of Statistics, 47(2), 1148–1178. https://doi.org/10.1214/18-AOS1709

Bang, H., & Robins, J. M. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. Biometrics, 61(4), 962–973. https://doi.org/10.1111/j.1541-0420.2005.00377.x

Cassel, C. M., Särndal, C. E., & Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika, 63(3), 615–620. https://doi.org/10.1093/biomet/63.3.615

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1), C1–C68. https://doi.org/10.1111/ectj.12097

Dehejia, R. H., & Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. Journal of the American Statistical Association, 94(448), 1053–1062. https://doi.org/10.2307/2669919

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition. Statistical Science, 34(1), 43–68. https://doi.org/10.1214/18-STS667

Frisch, R., & Waugh, F. V. (1933). Partial Time Regressions as Compared with Individual Trends. Econometrica, 1(4), 387–401. https://doi.org/10.2307/1907330

Glynn, A. N., & Quinn, K. M. (2010). An Introduction to the Augmented Inverse Propensity Weighted Estimator. Political Analysis, 18(1), 36–56. https://doi.org/10.1093/pan/mpp036

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. https://doi.org/10.1214/07-STS227

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. The American Economic Review, 76(4), 604–620. https://www.jstor.org/stable/1806062

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. Journal of the American Statistical Association, 89(427), 846–866. https://doi.org/10.1080/01621459.1994.10476818

Rotnitzky, A., Robins, J. M., & Scharfstein, D. O. (1998). Semiparametric Regression for Repeated Outcomes with Nonignorable Nonresponse. Journal of the American Statistical Association, 93(444), 1321–1339. https://doi.org/10.2307/2670049

Rubin, D. B. (1973). The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. Biometrics, 29(1), 185–203. https://doi.org/10.2307/2529685

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models. Journal of the American Statistical Association, 94(448), 1096–1120. https://doi.org/10.1080/01621459.1999.10473862

van der Laan, M. J., & Rubin, D. (2006). Targeted Maximum Likelihood Learning. The International Journal of Biostatistics, 2(1). https://doi.org/10.2202/1557-4679.1043