Demystifying Double Robust, Flexible Adjustment Methods for Causal Inference

Nathan I. Hoffmann

February 20, 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 doubly 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 complex survey data from the Program for International Student Assessment (PISA). Methods covered include Targeted Maximum Likelihood Estimation, Double/Debiased Machine Learning, and Augmented Inverse Propensity Weighting. Although these methods are robust to either the exposure or response model being misspecified, preliminary results show that, when both models are incorrect, bias grows rapidly in small samples.

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 doubly 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 a type of data that social scientists frequently encounter: cross-national survey data. It does this by using both simulated data where the treatment effect estimate is known, and then using complex survey data from the Program for International Student Assessment (PISA).

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

Doubly 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 "doubly 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), doubly 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 doubly robust and extended to causal inference.

But Kang & Schafer (2007) argue that doubly 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, doubly robust methods go back even further than this. The form of doubly 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 doubly robust estimators.

There are also links between doubly 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 doubly robust methods require almost all of the standard assumptions necessary formost methods that depend on selection on observables. Although some doubly 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: doubly robust methods will not work if we do not measure an important confounder that affects both treatment and exposure. But notably, the doubly robust methods covered in this tutorial make no functional form assumptions. Most use flexible machine learning algorithms to estimate both the outcome and exposure models, with regularization (often through cross-fitting) to avoid overfitting.

A simple demonstration

To demonstrate double robustness, this section presents one of the simpler doubly robust estimators: Augmented Inverse Propensity Weighting (AIPW) (Glynn & Quinn, 2010). 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).

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. As shown in the demonstration above, 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.

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.

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.

These methods have many similarities. How do the results they give compare? The next section tests the performance of each in practice.

Methods

For the comparisons in the next section, estimation methods include ordinary least squares regression ("OLS"), propensity-score matching with scores estimated from logistic regression using the `MatchIt` package ("PSM"), the augmented inverse propensity weighted estimator with generalized additive models using the `CausalGAM` package (AIPW), targeted maximum likelihood estimation using the `tmle` package (TMLE) with the default Superlearner that relies on the `SL.glm`,`SL.step`, and `GL.glm.interaction` functions, and double/debiased machine learning with random forests with the `DoubleML` and `mlr3` packages.

Table 1: Results of Monte Carlo simulations using the first 20 datasets from Dorie et al. (2019), 5 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.

label	bias	percent_bias	rmse	mae	fail_count	comp_time	n
lm	0.250	0.157	0.250	0.41	0	0.06	200
psm	0.203	0.130	0.203	0.53	6	0.77	200
aipw	0.134	0.093	0.134	0.38	0	8.74	200
tmle	0.018	0.013	0.018	0.36	0	7.85	200
dml	0.374	0.243	0.374	0.42	0	6.25	200

Table 2: Results of Monte Carlo simulations using the linear DGP for dataset 3 from Dorie et al. (2019), 5 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.

label	bias	percent_bias	rmse	mae	fail_count	comp_time	n
lm	0.023	0.014	0.023	0.050	0	0.059	200
psm	0.021	0.013	0.021	0.049	12	0.762	200
aipw	0.165	0.105	0.165	0.157	0	8.856	200
tmle	0.035	0.023	0.035	0.093	0	7.862	200
dml	0.490	0.306	0.490	0.450	0	6.367	200

Results

Simluation 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. For these preliminary results, I used 5 simulations of each data set, resulting in 100 data sets. I then calculated 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 include the number of data sets for which the method failed and the average computation time for each data set, in seconds.

Results are shown in Table 1. The methods all have similar levels of bias and error, but AIPW performs better than the other methods on all measures. AIPW did fail, however, on the most data sets; the package `CausalGAM` package does not do well with lack of overlap of factor variables between treated and control groups.

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.

sample	method	experimental	PSID-1	PSID-3	CPS-1	CPS-3
Original LaLonde	lm	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	lm	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)

Comparisons using LaLonde NSW Data

As another benchmark test, I use data from LaLonde's (1986) study of the National Supported Work Demonstration (NWS), as provided by Dehejia & Wahba (1999). The NWS randomly provided training to disadvantaged workers, allowing an experimental estimate of the effect of the intervention. This study had 297 treated and 425 control participants

LaLonde compared these experimental estimates to estimates comparing the treated group to 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 income in 1975 was below the poverty level. Restricting these observational samples is meant to get closer to the population eligible for the NWS.

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).

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

For the original NSW data, none of the methods perform especially well. For the PSID samples, none of the estimates have a positive sign, as in the experimental estimates. For CPS, TMLE is the only method that achieves a positive sign.

Including 1974 earnings data allows much better estimates with the observational control groups. (Note that the restricted experimental sample has a greater experimental effect estimate than the full experimental sample.) For PSID-1, TMLE performs the best, 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.

References

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.

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