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 $i = 1, \dots, N$ 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 we're 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.

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; Rotnitzky et al., 1998). 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:

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

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

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

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" part of "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 practice.

Preliminary Findings

Monte Carlo Simulations

The structure of these simulations is based on Kang & Schafer (2007). For each unit $i = 1, \ldots, n$, let $Z = (z_{i1}, z_{i2}, z_{i3}, z_{i4})^\top$ be distributed independently as $N(0, I)$, where I is the 4×4 identity matrix. Furthermore, let $d_i \in \{0, 1\}$ be an indicator for treatment status with a Bernoulli distribution with probability π_i of receiving treatment status. Values of π_i are generated as:

$$\pi_i = \text{expit}(z_{i1} + 0.5z_{i2} - 0.33z_{i3} - 0.2z_{i4}),$$

where $\text{expit}(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$. These are then used as the probability of treatment assignment in a series of Bernoulli draws for d_i. Outcomes y_i are generated as

$$y_i = 500 + 50d_i + 30z_{i1} - 35z_{i2} - 60z_{i3} + 50z_{i4} + \epsilon_i,$$

where $\epsilon_i \sim N(0, 10)$. Now assume that the researcher cannot measure the z_{ij}'s. Instead we observe $X = (x_{i1}, x_{i2}, x_{i3}, x_{i4})^\top$:

$$\begin{aligned} x_{i1} &= \exp(z_{i1}/2) \\ x_{i2} &= z_2/(1 + \exp(z_{i1})) + 10 \\ x_{i3} &= (z_{i1}z_{i3}/25 + 0.6)^3 \\ x_{i4} &= (z_2 + z_4 + 20)^2 \end{aligned}$$

For these preliminary results, datasets of 200 observations are generated 100 times. For each estimation method, three specifications are used:

- "correct": the model is fit to the true data-generating variables, Z
- "incorrect": the model is fit to the transformed variables X
- "ovb": the model is fit to the $(x_2, x_3, x_4)^\top$ (x_1 is omitted).

Results are shown in Table 1. 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

Table 1: Results of Monte Carlo simulations over 100 replications. Percent bias is calculated as the estimator's bias as a percentage of its standard error, rmse is root mean squared error, and mae is median absolute error.

label	n	model	bias	percent_bias	rmse	mae
OLS	200	correct	0.10	6.7	1.6	1.1
OLS	200	incorrect	7.89	61.2	15.1	10.7
OLS	200	ovb	27.38	231.1	29.8	26.1
PSM	200	correct	0.09	5.8	1.6	1.1
PSM	200	incorrect	9.95	79.8	15.9	11.8
PSM	200	ovb	29.70	240.2	32.2	30.0
AIPW	200	correct	0.04	2.0	1.8	1.2
AIPW	200	incorrect	5.45	30.0	18.9	13.2
AIPW	200	ovb	23.73	172.2	27.4	23.0
TMLE	200	correct	0.04	2.4	1.6	1.0
TMLE	200	incorrect	6.49	45.2	15.7	11.9
TMLE	200	ovb	27.81	236.5	30.2	27.7
DML	200	correct	1.06	25.0	4.3	3.1
DML	200	incorrect	1.39	14.4	9.7	6.3
DML	200	ovb	25.38	229.4	27.7	23.4

likelihood estimation using the `tmle` package (TMLE), and double/debiased machine learning with random forests with the `DoubleML` and `mlr3` packages.

Surprisingly, the double robust methods do not necessarily perform better. All methods perform well when the true data-generating variables are provided. When transformed variables are provided instead, bias, RMSE, and MAE rise for all of the estimation methods, even those that use flexible methods to model the response and treatment assignment surfaces. When a variable is omitted, bias also rises for all of the methods.

Next Steps

The Monte Carlo simulations presented here use small sample sizes and few variables, while double robust methods are perhaps most advantageous for large, high-dimensional samples. In the final paper, the simulations here will be supplemented samples of 1,000 and 10,000, and all will be replicated 1,000 times. Data sets with many covariates will also be simulated. The final paper will also test these methods on cross-national PISA test score data and provide R code to implement these methods.

References

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

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

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

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