Doubly Robust Estimation and Causal Inference Model

Weitong Yao Yuwei Zheng

Version: 0.01

Last update: June 23, 2019

Abstract

Doubly robust estimation (DR) is a combination of outcome regression (OR) and inverse propensity weighted score model (IPW). If one of these models is correctly specified, the ATE of doubly robust is the unbiased, consistent estimator of average treatment effect estimator under the assumption of no unmeasured confounders. This paper provides a conceptual overview of IPW and DR, compares different MSE and ATE from simulations under different circumstances, and discusses the possible improvements and limitations of DR.

1 Introduction

Doubly Robust is a robust protected model to study causal inferences. The term doubly robustness is constructed based on the Augmented Inverse Probability Weighted Estimators (AIPW), proposed by Robins, Rotnitzky and Zhao (1994). They provided a method to study regression coefficients with missing regressors, which improves the efficiency of estimation. This algorithm is further extended and issues about this methodology are studied by many researchers (Robins, J. M., Van der Laan, M. J., Brookhart, M. A., etc.).

In the introduction, we mainly review some related concepts and algorithms for DR estimations. This article is organized as follows: in Section 2, we begin to compare different models, including pure linear model, IPW, traditional DR, constrained weighting ensemble models, and unconstrained weighting ensemble models. In Section 3, we obtain conclusions and results from our simulation studies. The discussion of DR is illustrated in Section 4.

Conceptual Review

We define $X = 1$ if it's treated and $X = 0$ if it's under control, Y as observed outcome and Z as observed, measurable confounders, and the full dataset is $(Y_i, X_i, Z_i) \quad i = 1, ..., n$ for individuals. What we are interested in is to find the average treatment effect $\tau = E(Y|X = 1, Z) - E(Y|X = 0, Z)$. It is a counterfactual issue because we can't see all potential outcomes (Y_1, Y_0) for all individuals, instead we can only observe $Y = Y_1X + Y_0(1 - X)$, which is critical.

Regression is one approach to estimate τ, and it requires the postulated regression model is true and identical and no unmeasurable confounders. In this situation, the ATE of outcome regression is calculated as follows, that's the marginal effect of treatment (maximum likelihood estimate)is exactly the ATE for regression method.

$$\begin{aligned} E(Y|X,Z) =& \beta_0 + \beta_x X + \beta_z Z \\ \tau_{reg} =& E(Y|X=1,Z) - E(Y|X=0,Z) \\ =& (\hat{\beta}_0 + \hat{\beta}_x \times 1 + \hat{\beta}_z Z) - (\hat{\beta}_0 + \hat{\beta}_x \times 0 + \hat{\beta}_z Z) \\ =& \hat{\beta}_x \end{aligned} \tag{1}$$

Another way to estimate τ is inverse propensity weighted score model(IPW), the reason for "inverse propensity" is to provide even weighting for treatment group and control group:

$$\tau_{ipw} = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{X_i Y_i}{P(X=1|Z)} - \frac{(1-X_i)\,Y_i}{1-P(X=1|Z)}\Big] \tag{2}$$

Under assumptions that at least one of two models is correctly specified and no unmeasurable confounders, the ATE of doubly robust is defined as follows, and it's obviously a linear combination of outcome regression and inverse propensity weighted model with weighs equal $\frac{P(X|Z)-X_i}{P(X|Z)}$ and 1 respectively.

$$\begin{aligned} \tau_{DR} =& \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{P(X=1|Z)-X_i}{P(X=1|Z)}\widehat{Y_1} + \frac{X_i Y_i}{P(X=1|Z)}\Big] - \\ & \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{(1-P(X=1|Z))-(1-X_i)}{1-P(X=1|Z)}\widehat{Y_0} + \frac{(1-X_i)\,Y_i}{1-P(X=1|Z)}\Big] \\ =& \tau_{1,DR} - \tau_{0,DR} \end{aligned} \tag{3}$$

$$\begin{aligned} where,\ \ \tau_{1,DR} =& E(Y_1) + E\Big[\frac{P(X=1|Z)-X_i}{P(X=1|Z)}(\widehat{Y_1} - Y_1)\Big] \\ and,\ \ \tau_{0,DR} =& E(Y_0) + E\Big[\frac{X_i - P(X=1|Z)}{1-P(X=1|Z)}\widehat{Y_0}\Big] \end{aligned} \tag{4}$$

$\tau_{1,DR}$ is the function of observed outcome Y_1 and predicted outcome $\widehat{Y_1}$, $\tau_{0,DR}$ depends on $\widehat{Y_0}$ only as we can't observe Y_0, and $P(X=1|Z)$ is derived from logistic regression. If the OR is true, $E(Y-\hat{Y})=0$; if IPW is true, we have $E(P(X=1|Z)-X)=0$. In either scenario, ATE of DR is always $E(Y_1)-E(Y_0)$, which is an unbiased, consistent estimator for ATE.

2 Stacking Models and Comparison

What if the weights are more flexible, will the stacking model of OR and IPW perform better than traditional doubly robust model? In this section, we apply more flexible weights with each model and compare its ATE to different traditional models with different simulations and number of observations. All simulations are conducted by using cross-validation with five folders.

2.1 Constrained Weighted Stacking Model

The advantage of constrained weighting is that it provides better interpretation. If the weight of OR is defined as w_1, the weight of IPW is $w_2 = 1 - w_1$. The logic to find optimal weighting is to minimize mean squared error. Details about simulation are:

$$\begin{aligned} Z \sim & N(0,2),\ X \sim binomial(1, sigmoid(Z)),\ Y = X + Z + \varepsilon \\ OR = & Y \sim X + Z \qquad IPW = \frac{X_i Y_i}{P(X=1|Z)} + \frac{(1-X_i)\,Y_i}{1-P(X=1|Z)} \end{aligned} \tag{5}$$

And the Constrained Weighted Stacking Model is defined as:

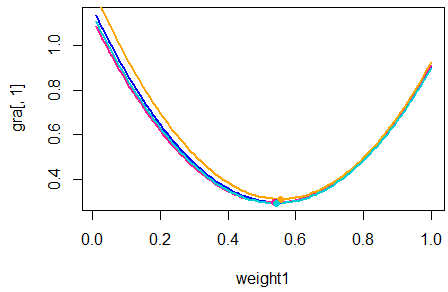
$$\begin{aligned} Stack_1 = & w_1 \times \widehat{Y_1} + w_2 \times \frac{X_i Y_i}{P(X=1|Z)} \\ Stack_0 = & w_1 \times \widehat{Y_0} + w_2 \times \frac{(1-X_i)\,Y_i}{1-P(X=1|Z)} \\ & w_1 + w_2 = 1 \end{aligned} \tag{6}$$

The simulation contains three kinds of $P(X = 1|Z)$: gaussian, logit and probit. The mean optimal weight of OR under different number of observations and $P(X = 1|Z)$ is

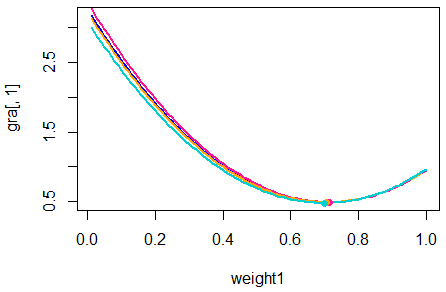
Table 1: Average Optimal Weight of OR

	Gaussian	Logit	Probit
n = 500	0.5443383	0.6085503	0.5982036
n = 1000	0.7070444	0.8213404	0.8078634
n = 2000	0.8346533	0.9117676	0.9143612
n = 5000	0.9505096	0.951561	0.9718606

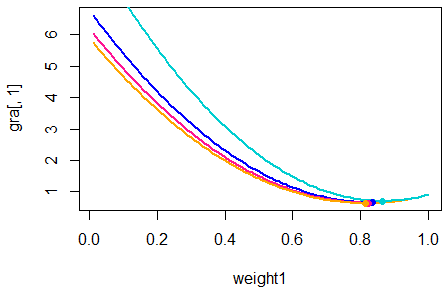
The results make sense: (1) As the number of observations increase, the simulation is more convincible and the weight of OR increases largely, according with $Y = X + Z + \varepsilon$. (2) If the independent variable is binary, there is little difference between logit regression and probit regression. (3) The speed of convergence(the fitted model converged to OR) is faster under Logit and Probit, compared with Gaussian. The vertical axis of graphs represent MSE for testing dataset. (4) Further discovery: the optimal weight of OR is larger than that of IPW until $\alpha >= 4$ in $Y = Z^{\alpha} + X + \varepsilon$, it seems OR is a better fitted model (however, the MSE is so large under $\alpha > 1$, which means the discovery may be unreasonable.)



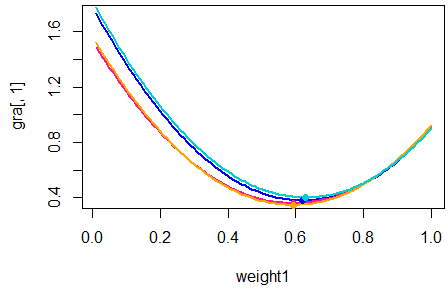
(a) gaussian with n = 500



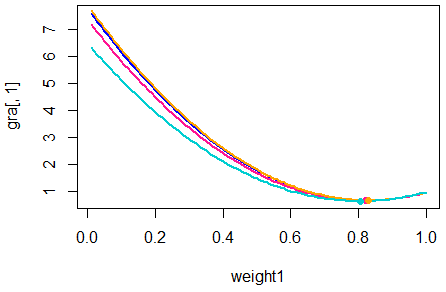
(b) gaussian with n = 1000



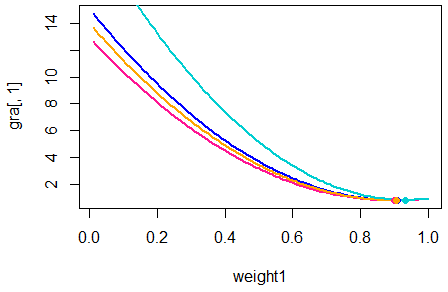
(c) gaussian with n = 2000



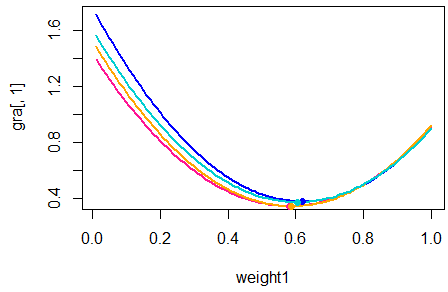
(d) logit with n = 500



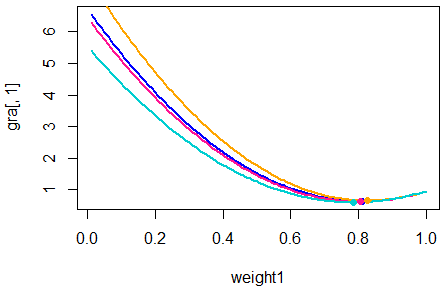
(e) logit with n = 1000



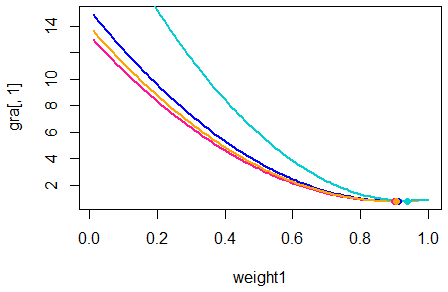
(f) logit with n = 2000



(g) probit with n = 500



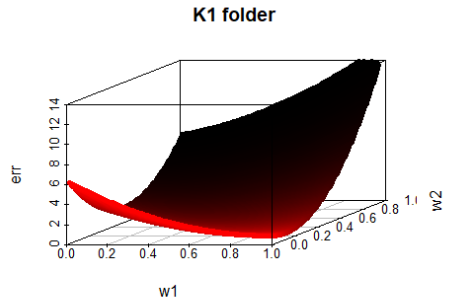
(h) probit with n = 1000



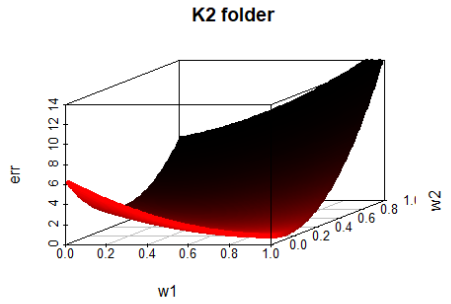
(i) probit with n = 2000

2.2 Unconstrained Weighted Stacking Model

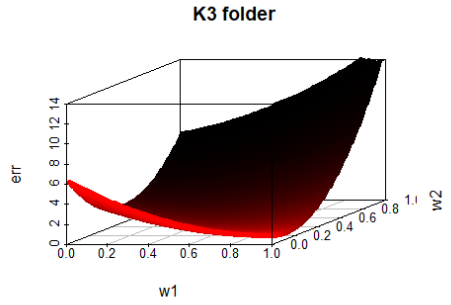
The unconstrained weighted stacking model may provide better fitted outcome at the cost of better interpretation. Suppose w_1, w_2 are weights for OR and IPW respectively and keep the settings same as section 2.1. By using "brute traversal" weights between 0 and 1, we obtain 3D plots for each folder :



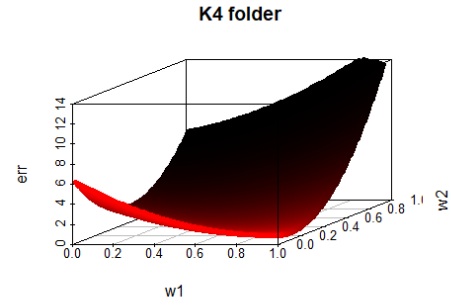
(j) 1st training fold with n = 1000



(k) 2nd training fold with n = 1000



(l) 3rd training fold with n = 1000



(m) 4th training fold with n = 1000

The optimal weights are stable at around 0.91 and 0.05 with increasing n.

2.3 Result Comparison

2.3.1 Model Definition

We study and compare MSE, ATE and variance of outcome regression, IPW, Doubly Robust, OLS ensemble(lsen), constrained weighted stacking model(stack) and unconstrained weighted stacking model(Unstack). As defined earlier,

$$\begin{aligned} f_1 &= w_1 \times \widehat{Y_1} + w_2 \times \frac{X_i Y_i}{P(X=1|Z)} \\ f_0 &= w_1 \times \widehat{Y_0} + w_2 \times \frac{(1-X_i)\,Y_i}{1-P(X=1|Z)} \\ \tau_{stack} &= E(f_1) - E(f_0) \quad with \quad w_1 + w_2 = 1 \\ \tau_{Unstack} &= E(f_1) - E(f_0) \quad with \quad w_1, w_2 \in [0,1] \end{aligned} \tag{7}$$

The OLS ensemble model is obtain using the following procedure:

- Using training data set in cv, regress $\widehat{Y_1}$ and $\frac{X_i Y_i}{P(X=1|Z)}$ on Y | X=1(observed Y_1) then we get $Y_1 = \hat{\beta}_1 \widehat{Y_1} + \hat{\beta}_2 \frac{X_i Y_i}{P(X=1|Z)}$
- Using training data set in cv, regress $\widehat{Y_0}$ and $\frac{(1-X_i)Y_i}{1-P(X=1|Z)}$ on Y | X=0(observed Y_0) then we get $Y_0 = \hat{\beta}_3 \widehat{Y_0} + \hat{\beta}_4 \frac{(1-X_i)Y_i}{1-P(X=1|Z)}$
- Apply two linear regression model on testing data and get $\mathcal{Y}_1$ and $\mathcal{Y}_0$
- $\tau_{lsen} = E(\mathcal{Y}_1) - E(\mathcal{Y}_0)$

2.3.2 Model Comparison

We compare the efficiency of six causal inference model by changing different initial settings and number of observations by spliting the dataset into 10 folders and applying cross-validation, the comparisons are shown in the following table.

Table 2: Comparing Causal Inference Model with n = 500

Settings	Indicator	Reg	IPW	DR	LSEN	Stack	Unstack
Y = linear(X) + linear(Z)	ATE	0.951430	1.011689	1.009087	3.444393	1.017201	0.978468
X = binomial(1,linear(Z))	VAR	0.023993	1.112388	0.232103	0.033396	0.054101	0.058459
	MSE	0.934337	85.261919	3.847946	6.811522	4.129348	3.402911
Y = linear(X) + linear(Z)	ATE	0.956788	1.189190	1.016797	1.340817	1.021133	0.918498
X = binomial(1,nonlinear(Z))	VAR	0.019725	0.148412	0.087840	0.042385	0.029081	0.026488
	MSE	1.053720	3.820766	0.466968	5.616494	3.008335	1.738237
Y = linear(X) + nonlinear(Z)	ATE	1.061157	-2.275825	-1.665877	1.583260	0.908970	0.796402
X = binomial(1,linear(Z))	VAR	0.177363	52.932163	27.393468	0.124896	0.374979	0.331102
	MSE	39.34201	5026.19116	81.44477	60.19720	44.22266	37.23538
Y = linear(X) + nonlinear(Z)	ATE	1.210787	1.219049	1.184054	1.262922	1.227717	1.071796
X = binomial(1,nonlinear(Z))	VAR	0.009586	0.106280	0.105656	0.0164993	0.021225	0.027807
	MSE	1.341292	3.283949	0.650278	1.682166	0.511069	0.466270

Table 3: Comparing Causal Inference Model with n = 1000

Settings	Indicator	Reg	IPW	DR	LSEN	Stack	Unstack
Y = linear(X) + linear(Z)	ATE	1.007846	1.236580	0.952120	3.424797	1.073456	1.039096
X = binomial(1,linear(Z))	VAR	0.006284	0.184614	0.024730	0.017164	0.026608	0.022017
	MSE	1.025680	40.461838	2.806959	6.732188	3.618579	3.227080
Y = linear(X) + linear(Z)	ATE	1.004839	1.254511	0.995902	1.399795	1.089641	1.000316
X = binomial(1,nonlinear(Z))	VAR	0.004731	0.010886	0.016957	0.003589	0.006722	0.007132
	MSE	0.975406	4.700372	0.467718	5.184206	3.158589	2.020850
Y = linear(X) + nonlinear(Z)	ATE	1.468347	2.108269	1.547489	1.587310	1.509558	1.364751
X = binomial(1,linear(Z))	VAR	0.072504	1.945301	1.245016	0.091852	0.303723	0.312551
	MSE	33.50649	624.16712	219.46862	52.80868	47.38703	40.32269
Y = linear(X) + nonlinear(Z)	ATE	1.251615	1.194133	1.173154	1.289442	1.238894	1.098957
X = binomial(1,nonlinear(Z))	VAR	0.008945	0.013843	0.013961	0.004397	0.004731	0.005538
	MSE	1.4309120	3.7667123	0.6473534	1.8486960	0.5469495	0.5199483

Table 4: Comparing Causal Inference Model with n = 2000

Settings	Indicator	Reg	IPW	DR	LSEN	Stack	Unstack
Y = linear(X) + linear(Z)	ATE	0.995271	0.940532	0.930769	3.424313	1.034398	1.007690
X = binomial(1,linear(Z))	VAR	0.002309	1.156691	0.041513	0.006603	0.005387	0.005614
	MSE	0.988404	358.591610	4.104773	7.000557	4.035834	3.637824
Y = linear(X) + linear(Z)	ATE	1.009137	1.361061	1.060845	1.429234	1.107688	1.027161
X = binomial(1,nonlinear(Z))	VAR	0.003655	0.020703	0.005703	0.008695	0.006180	0.005427
	MSE	0.965065	5.280279	0.517221	5.552783	3.218770	2.166259
Y = linear(X) + nonlinear(Z)	ATE	1.363651	-0.938856	-0.354882	1.600783	1.207503	1.109078
X = binomial(1,linear(Z))	VAR	0.071956	15.757928	10.003057	0.116070	0.068612	0.082408
	MSE	31.64986	6225.74181	118.98011	52.89056	45.81513	41.19371
Y = linear(X) + nonlinear(Z)	ATE	1.265785	1.275776	1.238021	1.318620	1.265688	1.120080
X = binomial(1,nonlinear(Z))	VAR	0.004881	0.061766	0.057562	0.005951	0.011241	0.012562
	MSE	1.414781	4.014497	0.679947	1.746575	0.565457	0.539640

3 Results and Conclusions

We study ATE and MSE of different causal inference models from Table2 to Table4 in Section 2.3 under different number of observations and settings. By comparison, we can summarize the results as follows:

(i) By comparing ATE, since the initial setting is $Y \sim X + nonlinear/linear(Z)$, it's straightforward that we can assume the average treatment effect of X is around 1. In either scenario, no matter what n is, the

best three estimations which can derive ATE about 1 are Reg, Stack and Unstack, because there's no large deviation and all ATEs are larger tha n 0.

(ii) As sample size increases, all ATEs converge to 1. For n = 2000, the estimation can be ordered as: Unstack, Stack and Reg. For each sample size, Unstack provides best performance in nonlinear initial settings due to its flexible stacking.

(iii) According to MSE, Reg preforms the best. DR, Stack, and Unstack also perform well, and it can be interpreted as it depends on the weighting on outcome regression model, which is also decided by the settings of simulations.

(iv) Under each circumstance, the risk of fixed weighting in DR is presented, compared with Stack and Unstack. If the Reg is correctly defined while IPW is not, (for example, the third simulation with n=2000, in which Reg has ATE = 1.36 and IPW has IPW = -0.94) DR suffers a lot from fixed weighting and Stack and Unstack are affected far less.

These results illustrate that the risk of poor estimation of DR does come from fixed weighting, and more flexible weightings provide better average treatment effect estimators, especially in a nonlinear situation, or said, the assumption that at least one of two models is correctly specified is violated. However, the outcome regression is more stable and also has good predictive power in linearity.

4 Discussion

The basic idea of doubly robust is to ensemble two basic model: OR and IPW, assuming that either $E(Y-\hat{Y}) = 0$ or $E(P(X = 1|Z) - X) = 0$, or both of them are correct. Although the robustness works well in many aspects, it has a weakness of fixed weighting, which illustrates DR largly, directly depends on OR and IPW and lacks the ability of self-adjustment. Thus, we consider flexible weighting for DR.

The constrained weighted stacking model and unconstrained weighted stacking model defined in the article show better performances compared with traditional DR when the assumption does not hold any more. And also, it presents that outcome regression is a simple but efficient method to estimate the marginal effect of treatment. Except for OLS ensemble, it's reasonable to apply machine learning to boost them one by one. For example, using random forest to fit each basic model alternately, and derive residuals to train the ensemble model, which can obtain better weights to mimic the complicated real model.

5 Bibliography

[1] Jiaming Mao, "Foundations of Causal Inference"

[2] Marie Davidian, "Doubly Robustness in Estimation of Causal Treatment Effects", Department of Statistics, North Carolina State University

[3] Romain Neugebauer, Mark van der Laan, "Why prefer double robust estimators in causal inference?", *Journal of Statistical Planning and Inference*, 129(2005) 405-426

[4] Trevor Hastie, Robert Tibshirani, Jerome Friedman , “The Elements of Statistical Learning”, Springer ISBN: 9780387848570, 2009-10-01, 288-290

[5] Michele J.F., Daniel Westreich, Chris Wiesen, Til Sturmer, M. Alan Brookhart, Marie Davidian, “Doubly Robust Estimation of Causal Effects”, *American Journal of Epidemiology*, Vol.173 No.7, DOI: 10.1093/aje/kwq439, 2011