

Doubly Robust Estimators

Ye Wang

University of North Carolina at Chapel Hill

Linear Methods in Causal Inference

POLI784

Review

- ▶ We have learned four methods to deal with confounders under strong ignorability.
- ▶ Matching, weighting, regression, and balancing.
- ▶ NN matching requires no extra restrictions but introduces a bias term and is inefficient.
- ▶ PS matching and IPW require accurate estimates of the propensity scores.
- ▶ They are sensitive to the violation of positivity.
- ▶ Regression is built upon the correct specification of the response surface.
- ▶ Balancing is valid when either the propensity score or the response surface satisfies a certain form.

Combine estimators

- ▶ We can actually combine previously mentioned estimators for a better performance.
- ▶ The combined estimator is usually more efficient.
- ▶ It may also possess the property we call “double robustness” (Robins, Rotnitzky, and Zhao 1994).
- ▶ Remember that different methods impose different structural restrictions, which may not hold in practice.
- ▶ The doubly robust estimators produce credible results when structural restrictions hold for either method.

The AIPW estimator

- ▶ A classic example of doubly robust estimators is the augmented IPW (AIPW) estimator:

$$\hat{\tau}_{AIPW} = \frac{1}{N} \sum_{i=1}^N \left[\frac{D_i(Y_i - \hat{m}_1(\mathbf{X}_i))}{\hat{g}(\mathbf{X}_i)} - \frac{(1 - D_i)(Y_i - \hat{m}_0(\mathbf{X}_i))}{1 - \hat{g}(\mathbf{X}_i)} \right] + \frac{1}{N} \sum_{i=1}^N [\hat{m}_1(\mathbf{X}_i) - \hat{m}_0(\mathbf{X}_i)],$$

where $\hat{g}(\mathbf{X}_i)$ is the estimated propensity score, $\hat{m}_1(\mathbf{X}_i)$ is an estimate for $E[Y_i|D_i = 1, \mathbf{X}_i]$ and $\hat{m}_0(\mathbf{X}_i)$ is an estimate for $E[Y_i|D_i = 0, \mathbf{X}_i]$.

- ▶ For example, we can assume that $g(\mathbf{X}_i) = \frac{e^{\mathbf{X}_i' \beta}}{1 + e^{\mathbf{X}_i' \beta}}$,
 $m_0(\mathbf{X}_i) = \mathbf{X}_i' \beta_0$, and $m_1(\mathbf{X}_i) = \mathbf{X}_i' \beta_1$.
- ▶ Each model has its own structural restrictions.

The AIPW estimator

- ▶ Suppose the regression models are correctly specified and the propensity score model is not, then

$$E \left[\frac{1}{N} \sum_{i=1}^N [\hat{m}_1(\mathbf{X}_i) - \hat{m}_0(\mathbf{X}_i)] \right] = \tau.$$

- ▶ Moreover, $\hat{\varepsilon}_i = Y_i - \hat{m}_{D_i}(\mathbf{X}_i)$ is a random noise such that $E[\hat{\varepsilon}_i | \mathbf{X}_i] \rightarrow 0$.

- ▶ Now,

$$\begin{aligned} E[\hat{\tau}_{AIPW}] &= \frac{1}{N} \sum_{i=1}^N E \left[\frac{D_i \hat{\varepsilon}_i}{\hat{g}(\mathbf{X}_i)} - \frac{(1 - D_i) \hat{\varepsilon}_i}{1 - \hat{g}(\mathbf{X}_i)} \right] \\ &\quad + \frac{1}{N} \sum_{i=1}^N E [\hat{m}_1(\mathbf{X}_i) - \hat{m}_0(\mathbf{X}_i)] \\ &\rightarrow 0 + \tau = \tau \end{aligned}$$

- ▶ This is true even when $\hat{g}(\mathbf{X}_i) \not\rightarrow g(\mathbf{X}_i)$.

The AIPW estimator

- ▶ Suppose it is the other way around, we can see that the estimator is equivalent to

$$\hat{\tau}_{AIPW} = \frac{1}{N} \sum_{i=1}^N \left[\frac{D_i Y_i}{\hat{g}(\mathbf{X}_i)} - \frac{(1 - D_i) Y_i}{1 - \hat{g}(\mathbf{X}_i)} \right] - \frac{1}{N} \sum_{i=1}^N \left[\frac{(D_i - \hat{g}(\mathbf{X}_i)) \hat{m}_1(\mathbf{X}_i)}{\hat{g}(\mathbf{X}_i)} - \frac{(D_i - \hat{g}(\mathbf{X}_i)) \hat{m}_0(\mathbf{X}_i)}{1 - \hat{g}(\mathbf{X}_i)} \right]$$

- ▶ The first part is just the IPW estimator thus consistent.
- ▶ Since $\hat{g}(\mathbf{X}_i) \rightarrow g(\mathbf{X}_i)$ and $E[\hat{\nu}_i | \mathbf{X}_i] = E[D_i - g(\mathbf{X}_i) | \mathbf{X}_i] = 0$,

$$\frac{1}{N} \sum_{i=1}^N E \left[\frac{\hat{\nu}_i \hat{m}_1(\mathbf{X}_i)}{\hat{g}(\mathbf{X}_i)} - \frac{\hat{\nu}_i \hat{m}_0(\mathbf{X}_i)}{1 - \hat{g}(\mathbf{X}_i)} \right] \rightarrow 0.$$

- ▶ This is true even when $\hat{m}_{D_i}(\mathbf{X}_i) \not\rightarrow m_{D_i}(\mathbf{X}_i)$.

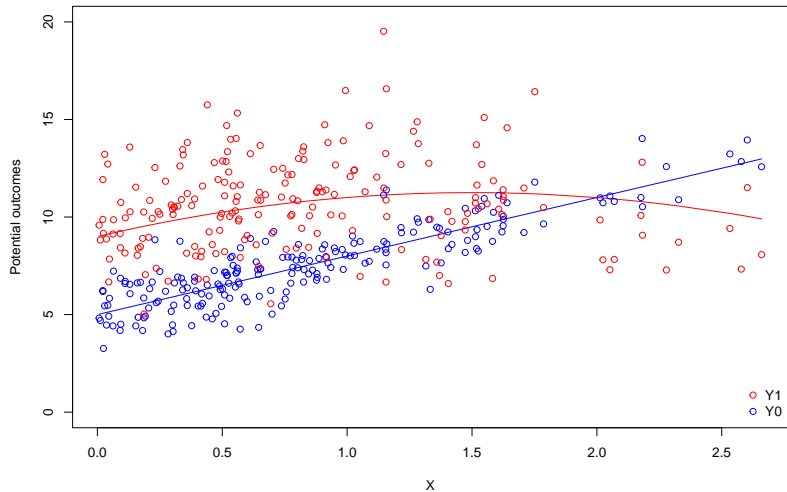
The AIPW estimator

- ▶ When either model is correctly specified, $\hat{\tau}_{AIPW}$ is consistent for τ .
- ▶ When both are correctly specified, $\hat{\tau}_{AIPW}$ reaches the efficiency bound proved by Hahn (1998).
- ▶ To estimate the variance of $\hat{\tau}_{AIPW}$, note that $\hat{\tau}_{AIPW} = \frac{1}{N} \sum_{i=1}^N l_i$, where

$$l_i = \frac{D_i(Y_i - \hat{m}_1(\mathbf{X}_i))}{\hat{g}(\mathbf{X}_i)} - \frac{(1 - D_i)(Y_i - \hat{m}_0(\mathbf{X}_i))}{1 - \hat{g}(\mathbf{X}_i)} - [\hat{m}_1(\mathbf{X}_i) - \hat{m}_0(\mathbf{X}_i)]$$

- ▶ Therefore, $\widehat{Var}[\hat{\tau}_{AIPW}] = \frac{1}{N^2} \sum_{i=1}^N (l_i - \hat{\tau}_{AIPW})^2$.
- ▶ l_i represents the (efficient) influence function for $\hat{\tau}_{AIPW}$.
- ▶ The variance can be obtained by regression l_i on 1.
- ▶ It does not account for the uncertainties from estimating the nuisance parameters either.

The AIPW estimator: simulation



The SATE is 3.1335

The AIPW estimator: simulation

```
## The SATE is 3.133
## Estimate from the right regression model is 3.143
## Estimate from the right ipw estimator is 3.111
## Estimate from the wrong regression model is 3.792
## Estimate from the wrong ipw estimator is 4.169
## Estimate from the doubly robust estimator
## with wrong regression model is 3.142
## Estimate from the doubly robust estimator
## with wrong pscore model is 3.143
## Estimate from the doubly robust estimator
## with correct models is 3.122
```

Bias correction in matching

- ▶ Lin, Ding, and Han (2023) proved that the bias correction estimator proposed by Abadie and Imbens (2011) is also doubly robust when M grows with N .
- ▶ We first estimate $\hat{m}_1(\mathbf{X}_i)$ and $\hat{m}_0(\mathbf{X}_i)$.
- ▶ Then, for each treated observation i , we have

$$\hat{Y}_i^{bc}(1) = \begin{cases} Y_i & D_i = 1 \\ \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} (Y_j + \hat{m}_0(\mathbf{X}_i) - \hat{m}_0(\mathbf{X}_j)) & D_i = 0, \end{cases}$$
$$\hat{Y}_i^{bc}(0) = \begin{cases} \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} (Y_j + \hat{m}_1(\mathbf{X}_i) - \hat{m}_1(\mathbf{X}_j)) & D_i = 1 \\ Y_i & D_i = 0, \end{cases}$$

- ▶ Similarly, the ATE estimate is

$$\hat{\tau}_M^{bc} = \frac{1}{N} \sum_{i=1}^N (\hat{Y}_i^{bc}(1) - \hat{Y}_i^{bc}(0)).$$

Bias correction in matching

- ▶ They show that

$$\begin{aligned}\hat{\tau}_M^{bc} &= \frac{1}{N} \sum_{i=1}^N (\hat{m}_1(\mathbf{X}_i) - \hat{m}_0(\mathbf{X}_i)) \\ &\quad + \frac{1}{N} \sum_{i=1}^N (2D_i - 1) \left(1 + \frac{K_M(i)}{M} \right) \hat{\varepsilon}_i\end{aligned}$$

- ▶ As $M \rightarrow \infty$,

$$1 + \frac{K_M(i)}{M} \rightarrow \begin{cases} \frac{1}{g(\mathbf{X}_i)} & D_i = 1 \\ \frac{1}{1-g(\mathbf{X}_i)} & D_i = 0, \end{cases}$$

- ▶ It approximates the AIPW estimator.
- ▶ They suggest that we should choose $M = N^{2/(2+\kappa)}$ based on simulation evidence.

Bias correction in matching: application

```
##
## Estimate...    2295
## AI SE.....   1321.4
## T-stat.....   1.7368
## p.val.....    0.082416
##
## Original number of observations..... 2675
## Original number of treated obs..... 185
## Matched number of observations..... 185
## Matched number of observations (unweighted). 932
##
## Estimate...   1468.7
## AI SE.....   1385.5
## T-stat.....   1.06
## p.val.....    0.28914
##
## Original number of observations..... 2675
## Original number of treated obs..... 185
```

Double robustness in regression

- ▶ Double robustness is built upon a simple idea from regression analysis.
- ▶ Consider the regression model

$$Y_i = \tau D_i + \mathbf{X}'_i \beta + \varepsilon_i.$$

- ▶ The OLS estimate $\hat{\tau}_{OLS}$ is consistent when ε_i is uncorrelated with either Y_i or D_i .
- ▶ Recall that the FWL theorem suggests

$$\hat{\tau}_{OLS} = (\hat{\nu}' \hat{\nu})^{-1} (\hat{\nu}' \hat{\varepsilon}),$$

where $\hat{\nu} = \mathbf{QD}$, $\hat{\varepsilon} = \mathbf{QY}$, and $\mathbf{Q} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Double robustness in regression

- ▶ Regressing D_i on \mathbf{X}_i implies the following regression model:

$$D_i = \mathbf{X}_i' \delta + \nu_i.$$

- ▶ We can show that

$$\hat{\tau}_{OLS} = \tau + (\nu' \mathbf{Q} \nu)^{-1} (\nu' \mathbf{Q} \varepsilon),$$

- ▶ $\hat{\tau}_{OLS} \rightarrow \tau$ when either $\nu' \mathbf{Q} \rightarrow 0$ or $\mathbf{Q} \varepsilon \rightarrow 0$.
- ▶ The former holds when $E[\nu_i | \mathbf{X}_i] = 0$ and the latter holds when $E[\varepsilon_i | \mathbf{X}_i] = 0$.
- ▶ The correct specification of either model can ensure the consistency of $\hat{\tau}_{OLS}$.
- ▶ The idea is generalized to nonparametric regression by Robinson (1988) (Robinson's transformation).

Summary

- ▶ Doubly robust estimators are robust to model misspecification rather than the violation of identification assumptions.
- ▶ Without strong ignorability, doubly robust estimators will be inconsistent.
- ▶ Both parts of the estimator should be functions of the same set of covariates.
- ▶ You cannot use the response surface to control for \mathbf{X}_i and the propensity score model to control for \mathbf{Z}_i .
- ▶ Nor can we expect the estimator to have negligible bias when both models are slightly biased.
- ▶ Positivity is essential for these estimators to work.
- ▶ But how can we ensure that strong ignorability holds?
- ▶ Unconfoundedness seems more plausible when we condition on more variables.
- ▶ The opposite could be true for positivity! (D'Amour et al. 2021)

Machine learning and causal inference

- ▶ Estimating nuisance parameters becomes challenging when we have a lot of potential confounders.
- ▶ How do you fit a logistical model when $P \gg N$?
- ▶ This is where machine learning (ML) can be useful.
- ▶ ML algorithms help us determine which variables really matter for prediction.
- ▶ They produce accurate predictions for nuisance parameters even in the high-dimensional setting.

Machine learning and causal inference

- ▶ But ML is designed to minimize the MSE rather than to estimate causal parameters.
- ▶ Simply plugging into nuisance parameters estimated by ML leads to large biases.
- ▶ This is known as regularization bias (Chernozhukov, Hansen, and Spindler 2015).
- ▶ Fortunately, this problem is less severe for doubly robust estimators.
- ▶ When we combine them with cross-fitting, biases caused by regularization can be negligible.
- ▶ We estimate nuisance parameters on the training set, and estimate causal parameters using AIPW on the test set.
- ▶ This is known as “double machine learning” proposed by Chernozhukov et al. (2017).

References I

- Abadie, Alberto, and Guido W Imbens. 2011. "Bias-Corrected Matching Estimators for Average Treatment Effects." *Journal of Business & Economic Statistics* 29 (1): 1–11.
- Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2017. "Double/Debiased/Neyman Machine Learning of Treatment Effects." *American Economic Review* 107 (5): 261–65.
- Chernozhukov, Victor, Christian Hansen, and Martin Spindler. 2015. "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach." *Annu. Rev. Econ.* 7 (1): 649–88.
- D'Amour, Alexander, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. 2021. "Overlap in Observational Studies with High-Dimensional Covariates." *Journal of Econometrics* 221 (2): 644–54.

References II

- Hahn, Jinyong. 1998. “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects.” *Econometrica*, 315–31.
- Lin, Zhexiao, Peng Ding, and Fang Han. 2023. “Estimation Based on Nearest Neighbor Matching: From Density Ratio to Average Treatment Effect.” *Econometrica* 91 (6): 2187–217.
- Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao. 1994. “Estimation of Regression Coefficients When Some Regressors Are Not Always Observed.” *Journal of the American Statistical Association* 89 (427): 846–66.
- Robinson, Peter M. 1988. “Root-n-Consistent Semiparametric Regression.” *Econometrica: Journal of the Econometric Society*, 931–54.