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:

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

  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 $E\left[\frac{1}{N}\sum_{i=1}^{N}\left[\hat{m}_1(\mathbf{X}_i) - \hat{m}_0(\mathbf{X}_i)\right]\right] = \tau$.
- Moreover, $\hat{\varepsilon}_i = Y_i - \hat{m}_{D_i}(\mathbf{X}_i)$ is a random noise such that $E[\hat{\varepsilon}_i|\mathbf{X}_i] \to 0$.
- Now,

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

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

The AIPW estimator

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

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

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

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

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

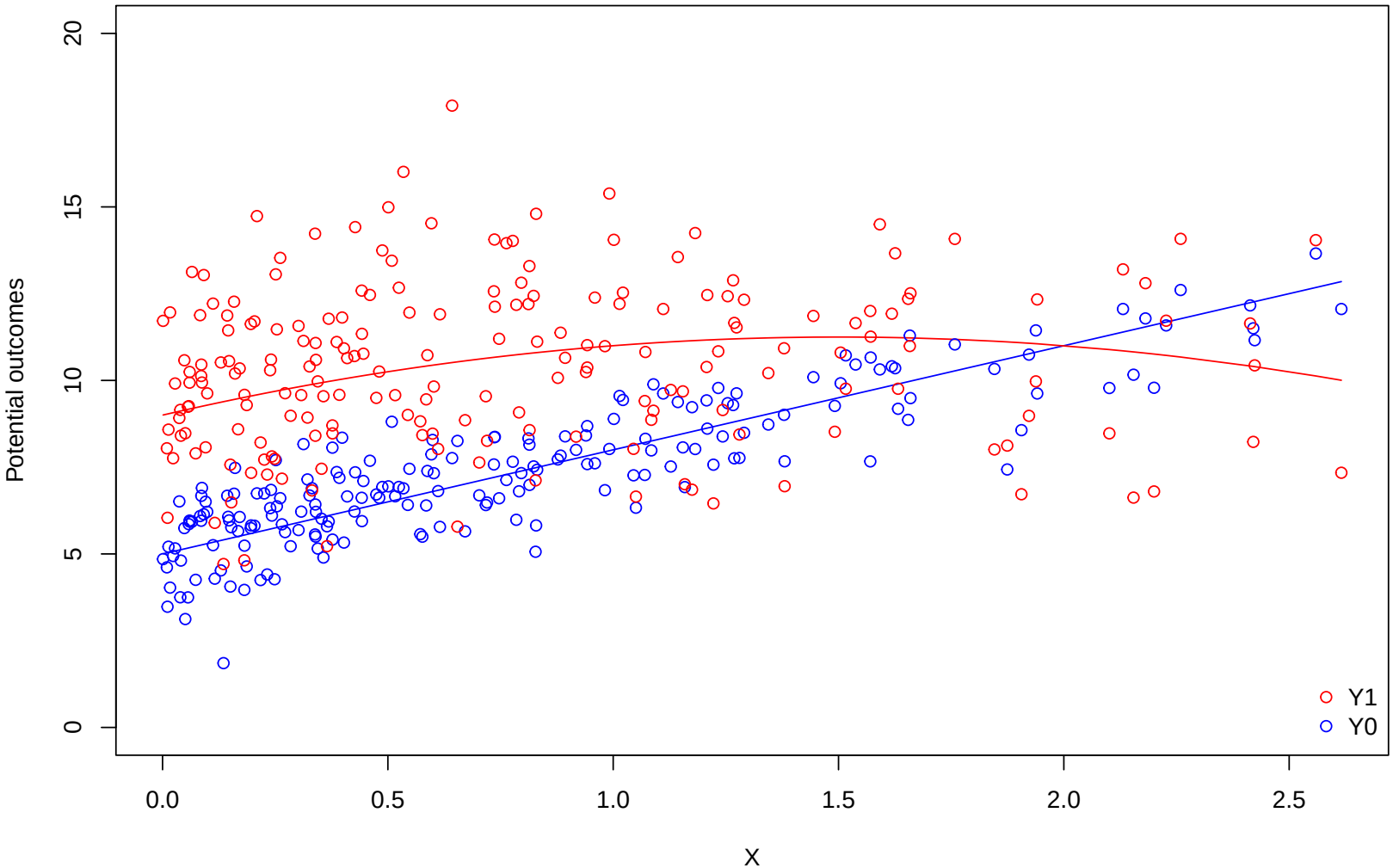
The AIPW estimator

- 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 $\hat{\tau}_{AIPW} = \frac{1}{N}\sum_{i=1}^{N} I_i$, where

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

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

The AIPW estimator: simulation




```
## The SATE is 3.09747
```

The AIPW estimator: simulation

```
## The SATE is 3.097

## Estimate from the right regression model is 3.162

## Estimate from the right ipw estimator is 3.115

## Estimate from the wrong regression model is 3.602

## Estimate from the wrong ipw estimator is 4.213

## Estimate from the doubly robust estimator
##  with wrong regression model is 3.111

## Estimate from the doubly robust estimator
##  with wrong pscore model is 3.162

## Estimate from the doubly robust estimator
##  with correct models is 3.107
```

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

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

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

- Similarly, the ATE estimate is

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

Bias correction in matching

- They show that

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

- As $M \to \infty$, $\frac{N_0}{N_1}\frac{K_M(i)}{M} \to \frac{f_{\mathbf{X}|D=1}(\mathbf{X}_i)}{f_{\mathbf{X}|D=0}(\mathbf{X}_i)}$, the density ratio at $\mathbf{X}_i$.
- Consequently,

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

- 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

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

- $\hat{\tau}_{OLS} \to \tau$ when either $\nu'\mathbf{Q} \to 0$ or $\mathbf{Q}\varepsilon \to 0$.
- 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 $\hat{\tau}_{OLS}$.

Summary

- Doubly robust estimators are robust to model misspecification not to the violation of identification assumptions.
- If strong ignorability is not satisfied, doubly robust estimators will be inconsistent.
- Therefore, 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)

Why machine learning?

- Now, let's assume that strong ignorability holds conditional on a large set of confounders.
- The dimensionality of the confounders can be even larger than the sample size, $P \gg N$.
- E.g., high-order terms and all the interaction terms of some covariates.
- If we know the values of the nuisance parameters, all the methods can still be applied.
- But estimating the nuisance parameters becomes really challenging.
- How do you run regression on 200 covariates when $N = 100$?
- This is where machine learning (ML) can be useful.

Basic ideas of machine learning

- We are interested in the relationship between Y_i and $\mathbf{X}_i$, where the dimensionality of $\mathbf{X}_i$ could be high:

$$Y_i = f(\mathbf{X}_i) + \varepsilon_i.$$

- The goal is to find an estimate $\hat{f}(\cdot)$ such that $E\left[\hat{f}(\mathbf{X}_i) - f(\mathbf{X}_i)\right]^2$ is minimized.
- Compared with conventional approaches, machine learning algorithms have two unique features: penalization and cross-validation.
- They allow us to select variables that have a strong prediction power of Y_i.

Basic ideas of machine learning

- In regression, we try to minimize the SSR:

$$\hat{f} = \arg\min_f \sum_{i=1}^{N} [Y_i - f(\mathbf{X}_i)]^2.$$

- In machine learning, we augment the objective function by adding a penalty term:

$$\hat{f} = \arg\min_f \sum_{i=1}^{N} [Y_i - f(\mathbf{X}_i)]^2 + \phi_\lambda(f),$$

where $\phi_\lambda(f)$ measures the complexity of our estimate.

Basic ideas of machine learning

- For example, the famous least absolute shrinkage and selection operator (LASSO) can be estimated from

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (Y_i - \mathbf{X}_i'\beta)^2 + \lambda \sum_{p=2}^{P} |\beta_p|.$$

- The first part is the familiar SSR, but we penalize models in which many variables have a non-zero coefficient.
- Consequently, the coefficient of many variables equals 0 in $\hat{\beta}$.
- The magnitude of λ decides the severity of penalty.
- If $\lambda = 0$, LASSO becomes linear regression.
- If $\lambda = \infty$, all the coefficients are zero and we predict Y_i with $\bar{Y}$.
- It controls the bias-variance trade-off and should be selected via cross-validation.

Basic ideas of machine learning

- Similar to bandwidth selection, we select a sequence of possible values for λ.
- Then, we randomly split the sample into the training set and the test set.
- For each λ, we solve $\hat{\beta}_\lambda$ on the training set.
- We test the performance of the linear model on the test set with:

$$\psi(\lambda) = \frac{1}{|i \in S_{test}|} \sum_{i \in S_{test}} (Y_i - \mathbf{X}_i' \hat{\beta}_\lambda)^2.$$

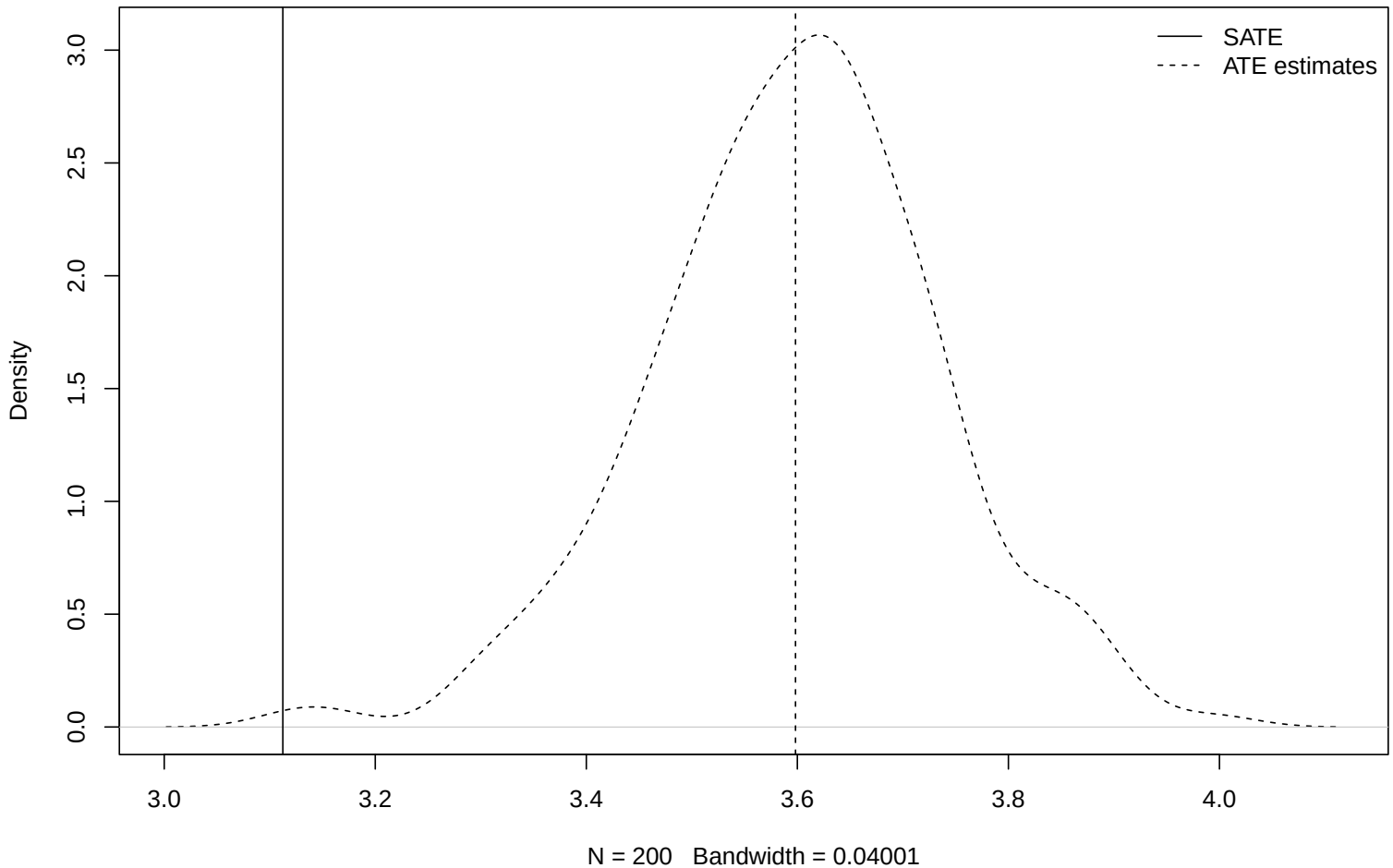
- The optimal choice, λ^*, minimizes $\psi(\lambda)$.
- Finally, we solve $\hat{\beta}^*$ using the entire sample and λ^*.
- LASSO works when we do not know which variables among a large set of candidates actually affect Y_i.

Basic ideas of machine learning

- Machine learning algorithms are designed to maximize our ability to make predictions.
- Should we just let the algorithms predict the relationship between Y_i and $(D_i, \mathbf{X}_i)$ for us and consider problem as solved?
- Not that simple!
- Even the most advanced AI cannot do causal inference, just like androids do not dream of electric sheep.
- They are selecting predictors rather than confounders.
- Variables that are strongly correlated with D_i but weakly correlated with Y_i may not be selected by LASSO (Chernozhukov, Hansen, and Spindler 2015).
- We should use ML algorithms to approximate the nuisance parameters.
- They are very useful when we don't know which variables from a large set of candidates are actually confounders.
- Yet some modifications of our basic methods are still necessary.

ML in causal inference

- A byproduct of penalization is slow convergence rate, which causes biases in estimation.



ML in causal inference

- Fortunately, this problem is less severe for doubly robust estimators.
- When we combine them with cross-fitting, biases caused by regularization can be negligible.
- This is an idea known as "double machine learning" proposed by Chernozhukov et al. (2017).
- For the AIPW estimator, we randomly split the sample into K folds: $\{I_k\}_{k=1}^K$.
- For $i \in I_k$, we apply ML algorithms to estimate all the nuisance parameters, $(g(\cdot), m_0(\cdot), m_1(\cdot))$, using units in $\cup_{l \neq k} I_l$.
- Then, we predict the values of the nuisance parameters for i, $(\hat{g}(\mathbf{X}_i), \hat{m}_0(\mathbf{X}_i), \hat{m}_1(\mathbf{X}_i))$, and plug them into the AIPW estimator.
- We still use all the units hence do not lose any efficiency.

ML in causal inference

```
## The SATE is 3.112

## Estimate from the naive ML estimator is 3.608

## Estimate from the DML estimator (no CF) is 3.347

## Estimate from the DML estimator is 3.354
```

ML in causal inference

- Cross-fitting is similar to cross-validation we saw before.
- It ensures that the irreducible error in I_k is independent to that from $\cup_{l \neq k} I_l$ (no "double dipping").
- $(\hat{g}(\cdot), \hat{m}_0(\cdot), \hat{m}_1(\cdot))$ behave as if they are known functions for units in I_k.
- The AIPW estimator satisfies a property we call "Neyman orthogonality."
- It means that the ATE estimate is insensitive/orthogonal to bias from estimating the nuisance parameters.
- For the AIPW estimator under cross-fitting, the regularization bias is bounded by

$$||\hat{g} - g|| * ||\hat{m}_D - m_D||$$

- If the convergence rate for both estimators is higher than $N^{1/4}$ (true for most ML algorithms), the bias will be negligible in large samples.

Summary

- ML algorithms are designed for prediction rather than causal inference.
- They rely on penalization and cross-validation to find models with the highest prediction power.
- We can use them to estimate the nuisance parameters when the number of potential confounders is large.
- To eliminate bias from regularization, we need 1) estimators that satisfy Neyman orthogonality, and 2) cross-fitting.
- The AIPW estimator with nuisance parameters estimated by ML algorithms is root-N consistent and asymptotically normal.
- We can estimate its variance using the influence function.
- There are many other approaches to incorporate ML into causal inference.
- A fast-growing field in causal inference.

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.