University of Amsterdam
Amsterdam School of Economics

Machine Learning for Econometrics

Week 6: Causal Inference and Machine Learning

Yi He

December 6, 2022

Plan for Today

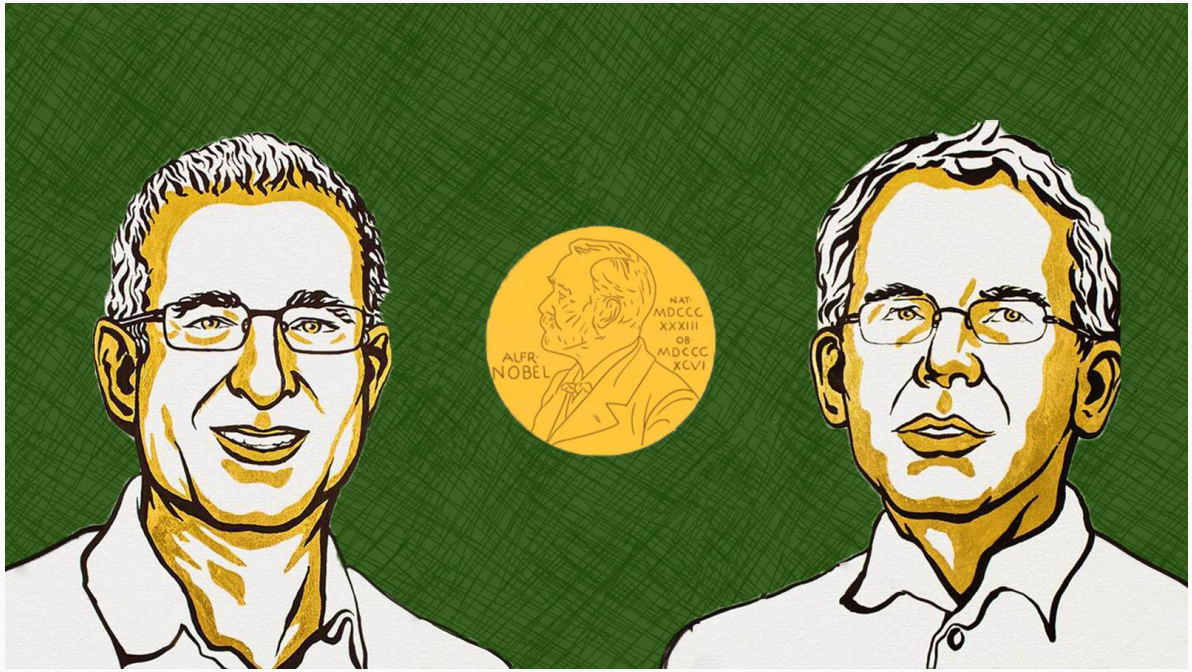
1. Introduction

2. Confounder Selection

3. Debiased Machine Learning

Introduction

Nobel Prize Economics 2021



Joshua Angrist and Guido Imbens [Dutch] share half the award

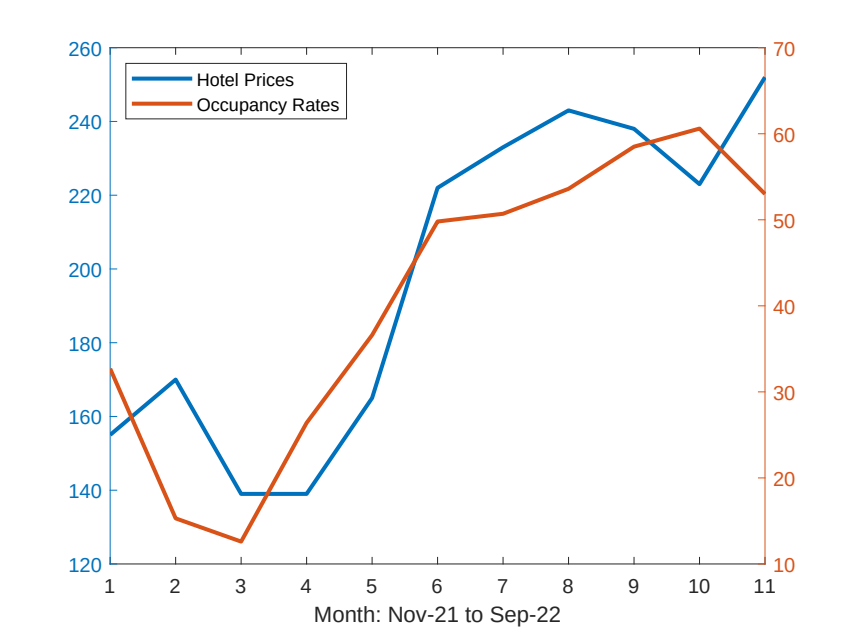
Achievement:

methodological contributions to the analysis of causal relationships

A Motivating Example

Imagine a hotel data set containing data about

- Occupancy rates
- Prices



Blue: the average rate per night for a standard double room in Amsterdam;

Red: hotel occupancy rate (in percentage) in the Netherlands

- Prices are easy to obtain through price comparison sites
- Occupancy rates are typically not made public by hotels
- Estimating the occupancy rates of *competitors*, based on publicly available prices, is a prediction problem:

  $$\text{Prices} \uparrow \text{ suggests occupancy rates} \uparrow$$

- Estimating how occupancy would change if the hotel raised prices across the board (e.g., $+5\%$ in prices in every state of the world) is a question of **causal inference**:

  $$\text{Prices} \uparrow \text{ causes occupancy rates} \uparrow ?$$

- Even though prices and occupancy are positively correlated in a typical dataset, we would not conclude that raising prices would increase occupancy

Potential Outcome Model

- Wish to find the causal effect of a treatment/policy variable D on the outcome variable Y
- Treatment ($D = 1$) or control ($D = 0$) groups
- $Y_i^{(d)}$ is the outcome for individual i when given treatment $d \in \{1, 0\}$:

$$Y_i = \begin{cases} Y_i^{(1)} & \text{if } D_i = 1, \\ Y_i^{(0)} & \text{if } D_i = 0, \end{cases} \quad \text{or } Y_i = D_i Y_i^{(1)} + (1 - D_i) Y_i^{(0)}.$$

- Binary treatment effect for individual i is given by

$$\delta_i = Y_i^{(1)} - Y_i^{(0)}$$

- *Counterfactual*: only observe $Y_i^{(D_i)} \in \{Y_i^{(1)}, Y_i^{(0)}\}$ but not both

The Fundamental Problem of Causal Inference

Table 2.1 in Morgan and Winship (2014)

Group	$Y^{(1)}$	$Y^{(0)}$
Treatment group ($D = 1$)	Observable as Y	Counterfactual
Control group ($D = 0$)	Counterfactual	Observable as Y

- Impossible to calculate individual-level treatment effects
- Estimate the *average treatment effect* (ATE)

$$\tau = \frac{1}{n}\sum_{i=1}^{n}\tau_i,\ \tau_i = \mathbb{E}\delta_i = \mathbb{E}\left[Y_i^{(1)} - Y_i^{(0)}\right] = \mathbb{E}\,Y_i^{(1)} - \mathbb{E}\,Y_i^{(0)}$$

where n denotes the total number of individuals

A Naive Estimator of the ATE

The difference in mean outcomes between

- the treatment group $\{i : D_i = 1\}$ with $n_1 = \sum_{i=1}^{n} D_i$ participants; and
- the control group $\{i : D_i = 0\}$ with $n_0 = n - n_1$ participants:

$$\begin{aligned}\widehat{\tau} =& \frac{1}{n_1} \sum_{D_i=1} Y_i - \frac{1}{n_0} \sum_{D_i=0} Y_i \\ =& \frac{1}{n_1} \sum_{D_i=1} Y_i^{(1)} - \frac{1}{n_0} \sum_{D_i=0} Y_i^{(0)}\end{aligned}$$

Is this a consistent estimator?

Even for the favorable case:

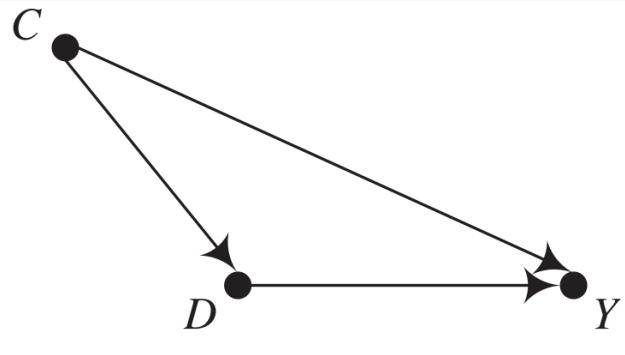
- $(Y_i^{(1)}, Y_i^{(0)}, D_i) \sim (Y^{(1)}, Y^{(0)}, D)$ are i.i.d.
- The treatment effect $\tau \equiv \tau_i \equiv \delta_i \equiv \delta$ is constant

$$
\begin{aligned}
\widehat{\tau} \xrightarrow{P} & \mathbb{E}[Y^{(1)} \mid D = 1] - \mathbb{E}[Y^{(0)} \mid D = 0] \\
= & \mathbb{E}[\delta + Y^{(0)} | D = 1] - \mathbb{E}[Y^{(0)} | D = 0] \\
= & \delta + \underbrace{\Big\{ \mathbb{E}[Y^{(0)} | D = 1] - \mathbb{E}[Y^{(0)} | D = 0] \Big\}}_{\text{bias}}
\end{aligned}
$$

- For $(Y^{(1)}, Y^{(0)})$ independent of D: $(Y^{(1)}, Y^{(0)}) \perp\!\!\!\perp D$
- ... the bias vanishes: $\widehat{\tau} \xrightarrow{P} \delta$

Unconfoundedness Assumption

- Confounding factors C influencing both D and Y:



- Example: college education and income both affected by parents' income & education, etc.
- Confounding implies that treatment assignment is not random but (self-)selected based on characteristics.
- Ideally confounders can be represented by features $X \in \mathbb{R}^d$.
- Unconfoundedness Assumption:

$$(Y^{(1)}, Y^{(0)}) \perp\!\!\!\perp D \mid X = x, \quad x \in \mathcal{X}$$

Confounder Selection

Structural Equations

If the treatment effect $\delta = Y_i^{(1)} - Y_i^{(0)}$ is fixed

$+$ linear model $Y_i^{(0)} = X_i^T \beta + \varepsilon_i = \sum_{j=1}^{d} X_{i,j}\beta_j + \varepsilon_i$

implies that

$$Y_i = \delta D_i + \sum_{j=1}^{d} X_{i,j}\beta_j + \varepsilon_i, \quad \underbrace{\mathbb{E}[\varepsilon_i | X_i, D_i] = 0}_{\text{Unconfoundedness}},$$

where we assume a zero intercept for simplicity.

Combining both information on policy Y and treatment D

$$\begin{cases} Y_i = \delta D_i + X_i^T \beta + \varepsilon_i, & \mathbb{E}[\varepsilon_i | X_i, v_i] = 0 \\ D_i = p(X_i) + v_i, & \mathbb{E}[v_i | X_i] = 0, \end{cases}$$

where $p(x)$ is called the *propensity score* function given by

$$p(x) = \mathbb{E}[D|X = x] = \mathbb{P}(D = 1|X = x)$$

Linear Model For Treatment Equation

For simplicity, consider an approximate linear probability model

$$p(X_i) = \sum_{j=1}^{d} X_{i,j}\gamma_j + r_i = X_i^T\gamma + r_i$$

where r_i should be negligible (stochastically).

Rewrite the model into a reduced form given by

$$\begin{cases} Y_i = X_i^T\bar{\beta} + \bar{\varepsilon}_i \\ D_i = X_i^T\gamma + \bar{v}_i, \end{cases}$$

where $\bar{\beta} = \delta\gamma + \beta$, $\bar{\varepsilon}_i = \delta\bar{v}_i + \varepsilon_i$ and $\bar{v}_i = r_i + v_i$.

Post-Selection Estimation

- When the dimension of confounders d is high relative to the sample size (with a large d/n), selecting a small set of the most useful confounders could be wise for variance reduction purpose.
- Run the least-squares regression on only the variables selected
- Post-selection estimator is not stable in general, but it can provide solid inference under sparsity assumption.

Sparsity Assumption

- If d is small, δ may be estimated by least-squares regression using the outcome equation:

$$Y_i = \delta D_i + X_i^T \beta + \varepsilon_i$$

- When d is large, assume that the true support set is however sparse:

$$\mathcal{J} = \{1 \leq j \leq d : \beta_j \neq 0\} \text{ has size } |\mathcal{J}| \ll n$$

- Only a few confounders are relevant to the population models, but we do not know which.

Single Selection Using The Outcome Equation

- Use the LASSO regression minimizing

$$\frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \delta D_i - \sum_{j=1}^{d} X_{i,j}\beta_j\right)^2 + \lambda\sum_{j=1}^{d}|\beta_j|$$

  where treatment is *not* penalized
- The confounders should be normalized.
- LASSO performs model selection: for non-trivial $\lambda > 0$

$$\widehat{\mathcal{J}} = \{1 \leq j \leq d : \widehat{\beta}_j \neq 0\} \ll n$$

Why Post-Selection Estimation?

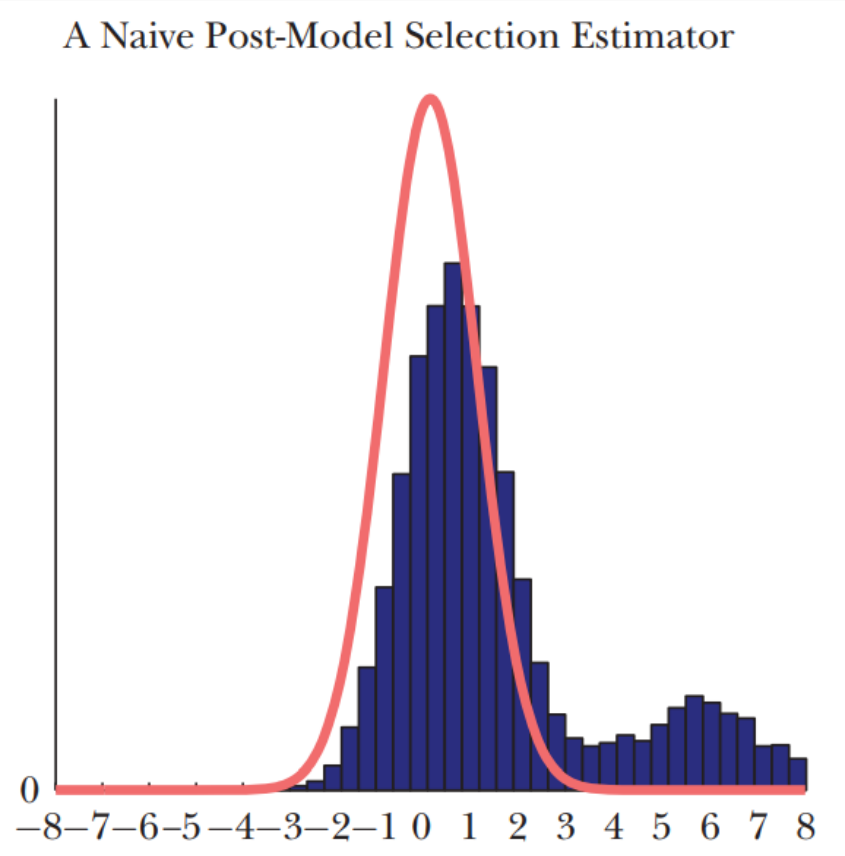
- Under more regularity conditions and an appropriate choice of penalty coefficient λ, Belloni and Chernozhukov (2013) show that with high probability and some constant C

$$\mathcal{J} \subset \widehat{\mathcal{J}}, \text{ and } \left|\widehat{\mathcal{J}}\right| \leq C|\mathcal{J}| \ll n$$

- The former implies a correct model selection, and the latter says our selected model is small (enough)
- LASSO estimator $\widehat{\delta}_{\text{LASSO}}$ often has a non-trivial bias: see, e.g., Belloni et al. (2013) for iterative algorithms for choosing λ
- Post-Lasso estimation tends to reduce bias: least-squares estimator after selection.

Naive Post Selection Estimator

- Single selection could miss the variables with true coefficients "moderately close to zero" with a non-trivial chance
- Select more variables: how?



Double Selection: Belloni et al. (2014)

Recall the reduced form

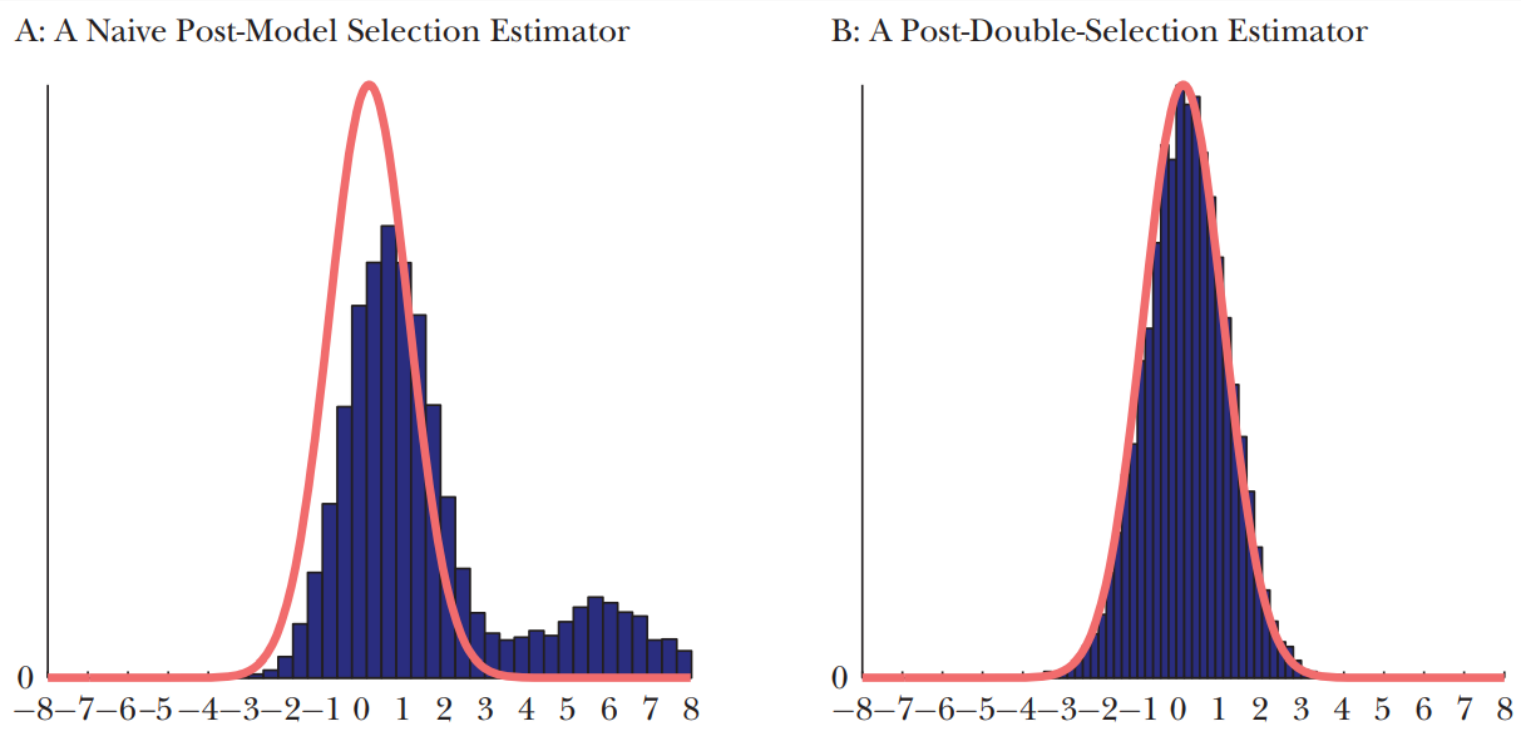
$$\begin{cases} Y_i = X_i^T \bar{\beta} + \bar{\varepsilon}_i \\ D_i = X_i^T \gamma + \bar{v}_i, \end{cases}$$

- Sparse index sets $\widehat{\mathcal{J}}_1, \widehat{\mathcal{J}}_2 \subset \{1, \ldots, d\}$ from **each** equation (using LASSO regression)
- Take a (small) set $\widehat{\mathcal{J}}_3$ that you think is important for ensuring robustness, if necessary
- The final index set is given by

$$\widehat{\mathcal{J}} = \widehat{\mathcal{J}}_1 \cup \widehat{\mathcal{J}}_2 \cup \widehat{\mathcal{J}}_3$$

- Least-squares regression after double selection.

Simulation Example: Belloni et al. (2014)



The post-double-selection estimator $\widehat{\delta}$ reduces bias, and its distribution gets closer to normal for sparse models.

Debiased Machine Learning

Auxiliary ML Estimator

Consider the partially linear model

$$Y_i = \delta D_i + \mu^{(0)}(X_i) + \varepsilon_i$$

where $\mu^{(0)}(x) = \mathbb{E}[Y^{(0)}|X = x]$ may be non-linear.

Suppose we have a machine learning estimator $\widehat{\mu}^{(0)}$ from another auxiliary sample independent of our training set with a learning bias

$$b(x) = \mathbb{E}\left[\widehat{\mu}^{(0)}(x)\right] - \mu^{(0)}(x) \not\equiv 0,$$

which is not identically zero.

The learning bias can emerge due to the bias-variance tradeoff in ML estimator.

Rewrite the model as

$$Y_i - \widehat{\mu}^{(0)}(X_i) = \delta D_i + \widetilde{\varepsilon}_i,\ \widetilde{\varepsilon}_i = \varepsilon_i + \mu^{(0)}(X_i) - \widehat{\mu}^{(0)}(X_i).$$

Regressing $Y_i - \widehat{\mu}^{(0)}(X_i)$ on D_i gives the least-squares estimator

$$\begin{aligned}
\widehat{\delta} &= \left(\sum D_i^2\right)^{-1}\left(\sum D_i(Y_i - \widehat{\mu}^{(0)}(X_i))\right) \\
&= \left(\sum D_i^2\right)^{-1}\left(\sum D_i(Y_i - \mu^{(0)}(X_i))\right) \\
&\quad + \left(\sum D_i^2\right)^{-1}\left(\sum D_i(\underbrace{\mu^{(0)}(X_i) - \widehat{\mu}^{(0)}(X_i)}_{\text{'prediction error' } e(X_i)})\right) \\
&= \delta + \underbrace{\left(\sum D_i^2\right)^{-1}\left(\sum D_i\varepsilon_i\right)}_{\text{variance component}} + \underbrace{\left(\sum D_i^2\right)^{-1}\left(\sum D_i e(X_i)\right)}_{\text{bias component}}
\end{aligned}$$

because

$$\mathbb{E}[D_i e(X_i)] = \mathbb{E}[D_i \cdot \mathbb{E}[e(X_i) \mid S]] = \mathbb{E}[\underbrace{D_i}_{p(X_i)+v_i} \cdot b(X_i)] \neq 0$$

Endogeneity and IV Regression

$$Y_i - \widehat{\mu}^{(0)}(X_i) = \delta D_i + \widetilde{\varepsilon}_i,\ \widetilde{\varepsilon}_i = \varepsilon_i + \mu^{(0)}(X_i) - \widehat{\mu}^{(0)}(X_i).$$

The learning bias issue is the endogeneity issue :

$$\mathbb{E}[D_i \widetilde{\varepsilon}_i] = \mathbb{E}[D_i b(X_i)] \neq 0.$$

because both D_i and $b(X_i)$ depend on X_i.

- Using the decomposition

$$D_i = p(X_i) + v_i, \quad \mathbb{E}[v_i | X_i] = 0$$

- If v_i were observable, it is a valid instrument for D_i as

$$\begin{cases} \mathbb{E}[v_i D_i] \neq 0 \\ \mathbb{E}[v_i \widetilde{\varepsilon}_i] = 0 \end{cases} \Rightarrow \delta = \frac{\mathbb{E}[v_i \left(Y_i - \widehat{\mu}^{(0)}(X_i)\right)]}{\mathbb{E}[v_i D_i]}$$

$$= \frac{\mathbb{E}[v_i \left(Y_i - \widehat{\mu}^{(0)}(X_i)\right)]}{\mathbb{E}[v_i^2]}$$

Debiased/Double ML: Chernozhukov et al. (2018)

One may estimate δ using the ratio of the sample averages

- Split into independent folds with index sets $\mathcal{I}$ and $\mathcal{I}^{\complement}$
- Learn both the regression function $\mu^{(0)}(\cdot)$ and propensity score functions $p(\cdot)$ on the auxiliary sample $\mathcal{I}^{\complement}$.
- Instrumental regression in sample $\mathcal{I}$:

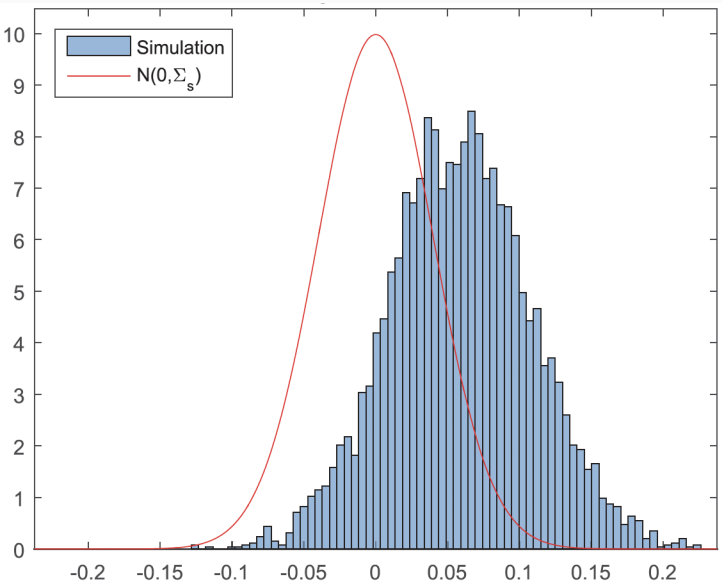
$$\widehat{\delta} = \left(\sum_{i \in \mathcal{I}} \widehat{v}_i D_i \right)^{-1} \sum_{i \in \mathcal{I}} \widehat{v}_i (Y_i - \widehat{\mu}^{(0)}(X_i)), \text{ where } \widehat{v}_i = D_i - \widehat{p}(X_i).$$

or

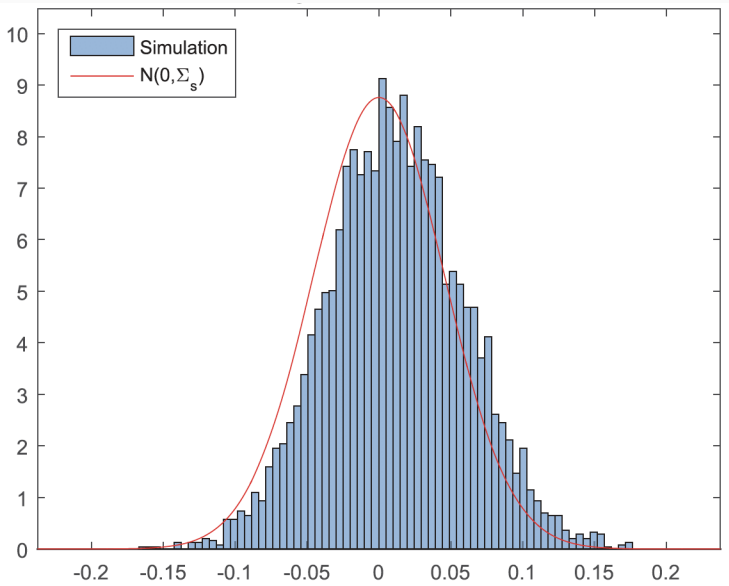
$$\widehat{\delta} = \left(\sum_{i \in \mathcal{I}} \widehat{v}_i^2 \right)^{-1} \sum_{i \in \mathcal{I}} \widehat{v}_i (Y_i - \widehat{\mu}^{(0)}(X_i))$$

- Optional: swap $\mathcal{I}$ and $\mathcal{I}^{\complement}$ and average the estimates

Simulation Example: Chernozhukov et al. (2018)



Conventional ML



Debiased ML

- $\mu^{(0)}(x)$ and $p(x)$ are learned by using random forests
- DML estimator has a smaller bias

Remark

- Due to the time constraint, today we cannot cover all machine learning methods in causal inference
- . . . such as the causal forest proposed Wager and Athey (2018) and the neural network approach in Farrell et al. (2021) based on influence functions
- You can find more explanations and details in our online reader.
- You will also learn more during the tutorial session.

References i

Belloni, A. and Chernozhukov, V., 2013. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19, 521–547.

Belloni, A., Chernozhukov, V. and Hansen, C., 2013. Inference for high-dimensional sparse econometric models. In *Advances in Economics and Econometrics: Tenth World Congress*. Vol. 3, edited by Daron Acemoglu, Manuel Arellano, and Eddie Dekel, 245–295. Cambridge, UK: Cambridge University Press.

Belloni, A., Chernozhukov, V. and Hansen, C., 2014. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81, 608–650.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21, C1–C68.

References ii

Farrell, M.H., Liang, T. and Misra, S., 2021. Deep neural networks for estimation and inference. *Econometrica*, 89, 181–213.

Wager, S. and Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113, 1228–1242.

Morgan, S.L. and Winship, C., 2014. Counterfactuals and Causal Inference (2nd Ed.). Cambridge University Press.