

# Causal Machine Learning

## Post-Double-Selection Procedure

Anthony Strittmatter

# Reference

Belloni, Chernozhukov, and Hansen (2014): "High-Dimensional Methods and Inference on Structural and Treatment Effects", Journal of Economic Perspectives, 28 (2), pp. 29-50, [download](#).

# Outline

## Selection Bias

### Selection-on-Observables Identification Strategy

### Power of Conditioning

### Multivariate Regression

### Post-Double-Selection Procedure

# Impact Evaluation

- ▶ Impact evaluation is a fascinating field of econometrics.
- ▶ It allows to make policy recommendations and business decisions.
- ▶ It enables to answer questions like:  
What is the causal impact of variable  $D$  on variable  $Y$ ?
- ▶ Examples include:
  - ▶ What is the causal effect of one additional year of education on wages?
  - ▶ Do micro-finance programs reduce poverty in developing countries?
  - ▶ What is the causal effect of value added taxes on customer purchases?
  - ▶ How large is the incumbency advantage in elections?
  - ▶ What is the causal effect of a marketing campaign on revenues?

⇒ The ability to conduct and/or interpret an impact evaluation study is useful in (almost) every field of economics and management!

# Causal Effect

- ▶ Lets say we want to analyse the causal effect of participation in a job search assistance course on earnings
- ▶  $D$  is a dummy indicating the participation in the job search assistance course
  - ▶  $D = 1$  under participation
  - ▶  $D = 0$  under non-participation
  - We often call this the treatment dummy
- ▶ Potential outcomes:
  - ▶  $Y^1$  denotes the potential earnings under participation in the job search assistance course
  - ▶  $Y^0$  denotes the potential earnings under non-participation in the job search assistance course
- ▶ The expected causal effect for a randomly selected individual from the population (Average Treatment Effect, ATE) is

$$ATE = E[Y^1] - E[Y^0]$$

# Stable Unit Treatment Value Assumption (SUTVA)

$$Y = D \cdot Y^1 + (1 - D) \cdot Y^0 \quad (1)$$

1. This assumption states that there are only two types of treatment (participation and non-participation),
  - ▶ The job search assistance course is always the same (no heterogeneous treatments, e.g., the duration of the course should not vary)
  - ▶ There are no alternative treatments (e.g., there should be no substitute for job search assistance for non-participants)
2. It excludes general equilibrium effects of the job search assistance course
  - ▶ Spillover effects could occur if course participants would inform non-participants about the course contents
  - ▶ Crowding-out effects could occur when course participants get jobs that would be devoted to non-participants in the absence of the course
  - ▶ Large courses could have externalities on the business cycle

⇒ We refer to (1) often as the “observational rule” (OR)

# Selection Bias

- ▶ We assume that  $0 < \Pr(D = 1) < 1$
- ▶ Naive estimation strategy:

$$\begin{aligned} E[Y|D = 1] - E[Y|D = 0] &\stackrel{OR}{=} E[Y^1|D = 1] - E[Y^0|D = 0] \\ &= \underbrace{(E[Y^1|D = 1] - E[Y^0|D = 1])}_{\text{ATET}} \\ &\quad + \underbrace{(E[Y^0|D = 1] - E[Y^0|D = 0])}_{\text{Selection Bias ATET}} \\ &= \underbrace{(E[Y^1|D = 0] - E[Y^0|D = 0])}_{\text{ATENT}} \\ &\quad + \underbrace{(E[Y^1|D = 1] - E[Y^1|D = 0])}_{\text{Selection Bias ATENT}} \end{aligned}$$

- ▶ ATET: Average Treatment Effect on the Treated
- ▶ ATENT: Average Treatment Effect on the Non-Treated

# Law of Iterative Expectations (LIE)

- ▶ Law of Iterative Expectations:

$$E[Y] = E[E[Y|X]] = E_X[E[Y|X]] = \int E[Y|X]f_X(x)dx$$

- ▶ Special case for dummy variables:

$$E[Y] = Pr(D = 1) \cdot E[Y|D = 1] + Pr(D = 0) \cdot E[Y|D = 0]$$

## Average Treatment Effect (ATE):

$$ATE \stackrel{LIE}{=} Pr(D = 1) \cdot ATET + Pr(D = 0) \cdot ATENT$$



# Randomised Experiments

- ▶ Randomised experiments are often called the “gold standard” of impact evaluation
- ▶ Under random assignment (RA), the potential outcomes are independent of the treatment, such that  $(Y^1, Y^0) \perp\!\!\!\perp D$  is satisfied

- ▶ ATET:

$$E[Y^1|D = 1] - E[Y^0|D = 1] = E[Y^1] - E[Y^0] = ATE$$

- ▶ ATENT:

$$E[Y^1|D = 0] - E[Y^0|D = 0] = E[Y^1] - E[Y^0] = ATE$$

- ▶ Selection Bias ATET:

$$E[Y^0|D = 1] - E[Y^0|D = 0] = E[Y^0] - E[Y^0] = 0$$

- ▶ Selection Bias ATENT:

$$E[Y^1|D = 1] - E[Y^1|D = 0] = E[Y^1] - E[Y^1] = 0$$

# Some Disadvantages Experiments

- ▶ Minimal social acceptance:
  - ▶ Would you agree to randomize the years of schooling for your children?
  - ▶ Would you agree to randomize police interventions to combat domestic violence?
- ▶ Randomisation technically impossible or impractical:
  - ▶ We cannot randomize climate change, gender, and incumbency.
  - ▶ Randomizing the Fed rate or value added taxes on the unit level is impractical (or even impossible).
- ▶ Costly and time consuming:
  - ▶ Poverty programs can be randomized, but the randomization can cause welfare losses during the experimental period.

# Some Disadvantages Experiments

- ▶ External validity:
  - ▶ Are experiments carried-out with a small group of economic students externally valid?
- ▶ Imperfect compliance
  - ▶ We can randomize the offer to participate in training programs, but not everybody participates.
  - ▶ We can randomize phone calls of get-out-the-vote (GOTV) campaigns, but not everybody answers the phone.

⇒ There is need for alternative empirical strategies!

# Outline

Selection Bias

Selection-on-Observables Identification Strategy

Power of Conditioning

Multivariate Regression

Post-Double-Selection Procedure

# Notation

- ▶  $D$ : Binary treatment dummy which can have values  $d \in \{0, 1\}$
- ▶  $Y^1, Y^0$ : Potential outcomes under treatment and non-treatment
- ▶  $Y = D \cdot Y^1 + (1 - D) \cdot Y^0$ : Observed outcome with support  $\mathcal{Y} \subseteq \mathbb{R}$  (we assume SUTVA throughout)
- ▶  $X$ :  $K$ -dimensional vector of exogenous pre-treatment control variables which can have values  $x \in \mathcal{X}$  (with  $\mathcal{X} \subseteq \mathbb{R}^K$  being the support of  $X$ ). The first element of  $X$  is a constant term.

# Notation

- ▶  $\mu_d(x) = E[Y^d|X = x]$ : Conditional expectation of the potential outcome  $Y^d$  (for  $d \in \{0, 1\}$ ) when control variables have values  $x$
- ▶  $p(x) = Pr(D = 1|X = x)$ : Condition probability that  $D = 1$  when control variables have values  $x$  (propensity score)
- ▶ We assume to observe i.i.d. (independent and identically distributed) data on the triple  $(Y, D, X)$  throughout

# Individual Causal Effects

$$\delta_i = Y_i^1 - Y_i^0$$

for observation units  $i = 1, \dots, N$  (e.g., individuals)

- ▶ Most of the time we omit the subscript  $i$  for ease of notation. We only use it when needed for clarity.
- ▶ Here the subscript makes clear that we allow for heterogeneous effects of each observation units.
- ▶ However, individual causal effects can only be identified under unrealistic assumptions

# Parameters of Interest

- ▶ Average Treatment Effects (ATE):

$$\delta = E[Y^1 - Y^0] = E[\delta_i]$$

- ▶ Average Treatment Effects on the Treated (ATET):

$$\theta = E[Y^1 - Y^0 | D = 1] = E[\delta_i | D = 1]$$

- ▶ Average Treatment Effects on the Non-Treated (ATENT):

$$\rho = E[Y^1 - Y^0 | D = 0] = E[\delta_i | D = 0]$$

- ▶ Conditional Average Treatment Effects (CATE):

$$\delta(x) = E[Y^1 - Y^0 | X = x] = E[\delta_i | X = x] = \mu_1(x) - \mu_0(x)$$



# Identifying Assumptions

## Assumptions for non-parametric models:

1. SUTVA (or observational rule, OR)
2. Conditional Independence Assumption (CIA):

$$(Y^1, Y^0) \perp\!\!\!\perp D | X = x \text{ for all } x \in \mathcal{X}$$

3. Common Support (CS) Assumption:

$$0 < p(x) = Pr(D = 1 | X = x) < 1 \text{ for all } x \in \mathcal{X}$$

# Interpretation of Assumptions

## Conditional Independence Assumption (CIA):

- ▶ Potential outcomes  $Y^1$  and  $Y^0$  are independent of the treatment  $D$  conditional on the covariates  $X$ .
- ▶ Implies that we have to control for all covariates that have a joint impact on the treatment and the potential outcomes.
- ▶ All covariates  $X$  have to be exogenous (typically determined pre-treatment).
- ▶ The CIA is an untestable assumption. We have the use application specific economic arguments to justify this assumptions.

## Common Support (CS) Assumption:

- ▶ Requires that we observe for each treated observation unit a comparable (in terms of covariates  $X$ ) non-treated observation unit.
- ▶ The CS assumption can be tested.

# Identification of ATEs

Under Assumption 1-3, we can identify  $\delta$  from observable data  $(Y, D, X)$ :

$$\begin{aligned}\delta &= E[Y^1 - Y^0] = E[Y^1] - E[Y^0] \\ &\stackrel{LIE}{=} \int (E[Y^1|X=x] - E[Y^0|X=x])f_X(x)dx \\ &\stackrel{CS, CIA}{=} \int (E[Y^1|D=1, X=x] - E[Y^0|D=0, X=x])f_X(x)dx \\ &\stackrel{OR}{=} \int (E[Y|D=1, X=x] - E[Y|D=0, X=x])f_X(x)dx \\ &= E_X[E[Y|D=1, X=x] - E[Y|D=0, X=x]] \quad \square\end{aligned}$$

# Outline

Selection Bias

Selection-on-Observables Identification Strategy

Power of Conditioning

Multivariate Regression

Post-Double-Selection Procedure

# Power of Conditioning

- ▶  $Y$ : Earnings (in Euro).
- ▶  $D$ : Dummy for participation in a job search assistant program ( $D = 1$  under participation,  $D = 0$  under non-participation).
- ▶  $X$ : Gender dummy ( $X = 1$  for women,  $X = 0$  for men).
- ▶ We observe a sample  $(Y, D, X)$  with  $N = 100$ .
- ▶ Observations:

		Participants	Non-participants
		$D = 1$	$D = 0$
Women	$X = 1$	$N = 10$	$N = 30$
Men	$X = 0$	$N = 40$	$N = 20$

# Power of Conditioning

- Observable expected earnings:

	$E[Y^1 D = 1, X = x]$ $= E[Y D = 1, X = x]$	$E[Y^0 D = 0, X = x]$ $= E[Y D = 0, X = x]$
Women ( $X = 1$ )	4000	3000
Men ( $X = 0$ )	5000	5000

- Counterfactual expected earnings (unobservables are in **red**):

	$E[Y^0 D = 1, X = x]$	$E[Y^1 D = 0, X = x]$
Women ( $X = 1$ )	3000	4000
Men ( $X = 0$ )	5000	5000

# True Causal Effects

- Average Treatment Effect on the Treated (ATET):

$$\begin{aligned} \text{ATET} &= \Pr(X = 1|D = 1) \cdot (E[Y^1|D = 1, X = 1] - E[Y^0|D = 1, X = 1]) \\ &\quad + \Pr(X = 0|D = 1) \cdot (E[Y^1|D = 1, X = 0] - E[Y^0|D = 1, X = 0]) \\ &= \frac{10}{50} \cdot (4000 - 3000) + \frac{40}{50} \cdot (5000 - 5000) = 200 \end{aligned}$$

- Average Treatment Effect on the Non-Treated (ATENT):

$$\begin{aligned} \text{ATENT} &= \Pr(X = 1|D = 0) \cdot (E[Y^1|D = 0, X = 1] - E[Y^0|D = 0, X = 1]) \\ &\quad + \Pr(X = 0|D = 0) \cdot (E[Y^1|D = 0, X = 0] - E[Y^0|D = 0, X = 0]) \\ &= \frac{30}{50} \cdot (4000 - 3000) + \frac{20}{50} \cdot (5000 - 5000) = 600 \end{aligned}$$

- Average Treatment Effect (ATE):

$$\begin{aligned} \text{ATE} &= \Pr(D = 1) \cdot \text{ATET} + \Pr(D = 0) \cdot \text{ATENT} \\ &= \frac{50}{100} \cdot 200 + \frac{50}{100} \cdot 600 = 400 \end{aligned}$$

# Naive Estimator

- ▶ Expected earnings of participants:

$$\begin{aligned} E[Y|D=1] &= Pr(X=1|D=1) \cdot E[Y|D=1, X=1] \\ &\quad + Pr(X=0|D=1) \cdot E[Y|D=1, X=0] \\ &= \frac{10}{50} \cdot 4000 + \frac{40}{50} \cdot 5000 = 4800 \end{aligned}$$

- ▶ Expected earnings of non-participants:

$$\begin{aligned} E[Y|D=0] &= Pr(X=1|D=0) \cdot E[Y|D=0, X=1] \\ &\quad + Pr(X=0|D=0) \cdot E[Y|D=0, X=0] \\ &= \frac{30}{50} \cdot 3000 + \frac{20}{50} \cdot 5000 = 3800 \end{aligned}$$

- ▶ Naive estimator:

$$E[Y|D=1] - E[Y|D=0] = 4800 - 3800 = 1000$$



# Is the CIA Valid?

- Conditional mean independence assumption for potential outcome under treatment ( $Y^1$ ):

$$E[Y^1|D = 1, X = x] = E[Y^1|D = 0, X = x]$$

$$4000 = E[Y^1|D = 1, X = 1] = E[Y^1|D = 0, X = 1] = 4000$$

$$5000 = E[Y^1|D = 1, X = 0] = E[Y^1|D = 0, X = 0] = 5000$$

- Conditional mean independence assumption for potential outcome under non treatment ( $Y^0$ ):

$$E[Y^0|D = 0, X = x] = E[Y^0|D = 1, X = x]$$

$$3000 = E[Y^0|D = 0, X = 1] = E[Y^0|D = 1, X = 1] = 3000$$

$$5000 = E[Y^0|D = 0, X = 0] = E[Y^0|D = 1, X = 0] = 5000$$

# Average Treatment Effect on the Treated (ATET)

Under Assumptions 1-3,

$$\begin{aligned}E[Y^1 - Y^0|D = 1] &= E[Y^1|D = 1] - E[Y^0|D = 1] \\&= E[Y^1|D = 1] - Pr(X = 1|D = 1) \cdot E[Y^0|D = 1, X = 1] \\&\quad - Pr(X = 0|D = 1) \cdot E[Y^0|D = 1, X = 0] \\&= E[Y^1|D = 1] - Pr(X = 1|D = 1) \cdot E[Y^0|D = 0, X = 1] \\&\quad - Pr(X = 0|D = 1) \cdot E[Y^0|D = 0, X = 0] \\&= E[Y|D = 1] - Pr(X = 1|D = 1) \cdot E[Y|D = 0, X = 1] \\&\quad - Pr(X = 0|D = 1) \cdot E[Y|D = 0, X = 0] \\&= 4800 - \frac{10}{50} \cdot 3000 - \frac{40}{50} \cdot 5000 = 200\end{aligned}$$

→ Positive bias (= 1000 - 200 = 800)!

# Average Treatment Effect on the Non-Treated (ATENT)

Under Assumptions 1-3,

$$\begin{aligned}E[Y^1 - Y^0|D = 0] &= E[Y^1|D = 0] - E[Y^0|D = 0] \\&= Pr(X = 1|D = 0) \cdot E[Y^1|D = 0, X = 1] \\&\quad + Pr(X = 0|D = 0) \cdot E[Y^1|D = 0, X = 0] - E[Y^0|D = 0] \\&= Pr(X = 1|D = 0) \cdot E[Y^1|D = 1, X = 1] \\&\quad + Pr(X = 0|D = 0) \cdot E[Y^1|D = 1, X = 0] - E[Y^0|D = 0] \\&= Pr(X = 1|D = 0) \cdot E[Y|D = 1, X = 1] \\&\quad + Pr(X = 0|D = 0) \cdot E[Y|D = 1, X = 0] - E[Y|D = 0] \\&= \frac{30}{50} \cdot 4000 + \frac{20}{50} \cdot 5000 - 3800 = 600\end{aligned}$$

→ Positive bias (= 1000 - 600 = 400)!

# Average Treatment Effect (ATE)

► ATE:

$$\begin{aligned} E[Y^1 - Y^0] &= Pr(D = 1) \cdot E[Y^1 - Y^0 | D = 1] \\ &\quad + Pr(D = 0) \cdot E[Y^1 - Y^0 | D = 0] \\ &= \frac{50}{100} \cdot 200 + \frac{50}{100} \cdot 600 = 400 \end{aligned}$$

→ The average effect of participation in job search assistance on earnings is 400 Euro.

# Simpson's Paradox

- ▶ Suppose we investigate the gender wage gap.
- ▶ We observe the following average wages of 100 women and 100 men in management and non-management positions:

	Women	Men
Non-management	1581.65 Euro ( $N = 87$ )	1507.59 Euro ( $N = 5$ )
Management	2796.22 Euro ( $N = 13$ )	2659.91 Euro ( $N = 4$ )

- ▶ In the sample, 13 women and 43 men have a management position.
- ▶ How large is the gender wage gap?

# Simpson's Paradox

- ▶ On average women earn less in this example:

$$\underbrace{\left( \frac{13}{100} \cdot 2796.22 + \frac{87}{100} \cdot 1581.65 \right)}_{\text{Average Wage Women}} - \underbrace{\left( \frac{41}{100} \cdot 2659.91 + \frac{59}{100} \cdot 1507.59 \right)}_{\text{Average Wage Men}} = -240.50$$

- ▶ Without conditioning on management position, women earn on average 240.50 Euro less than men.

# Simpson's Paradox

- ▶ But in each sub-category women earn more than men:
  - ▶ Management:  $2796.22 - 2659.91 = 136.31$
  - ▶ Non-management:  $1581.65 - 1507.59 = 74.06$
- ▶ The gender wage gap after conditioning on management position is:

$$\frac{13 + 41}{200} \cdot 136.31 + \frac{87 + 59}{200} \cdot 74.06 = 90.87$$

- ▶ After conditioning on management position, women earn on average 90.87 Euro more than men.

# Simpson's Paradox

- ⇒ **What is the correct gender wage gap?**
- ⇒ **Do we need to control for management position or not?**
- ▶ The seemingly contradicting results of the conditional and unconditional estimator are called Simpson's Paradox.
- ▶ The correct answers depends on the (typically untestable) assumptions we impose.
- ⇒ **Let us elaborate more on this on the next couple of slides using the DAG framework.**

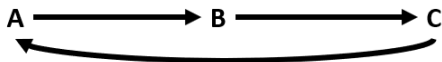


# Directed Acyclic Graphs (DAGs)

- ▶ Undirected graphs:



- ▶ Directed cyclic graphs:

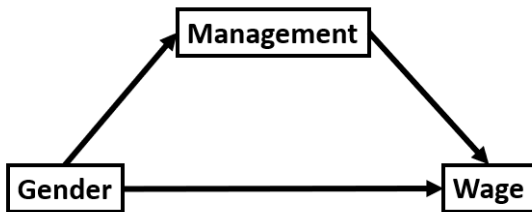


- ▶ Directed acyclic graphs:

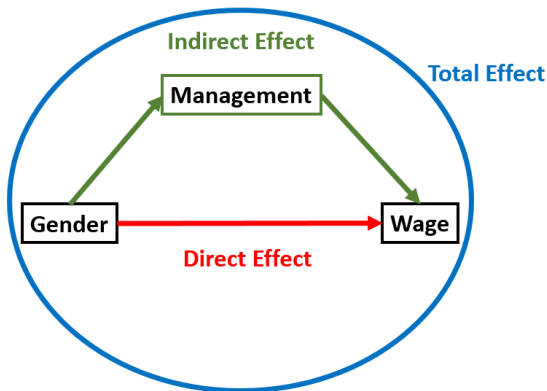


# Gender Wage Gap

→ We are interested in the effect of gender (left variable) on wages (right variable). → gender wage gap



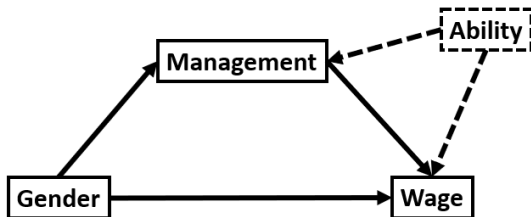
# Gender Wage Gap



- ▶ Total effect is identified when we do not condition on management.
- ▶ Direct effect is identified when we do condition on management.

# Gender Wage Gap

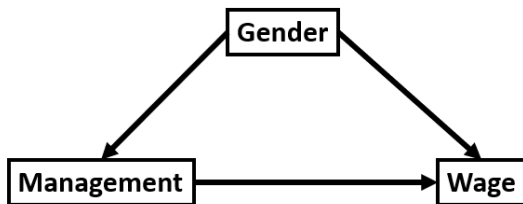
→ Now we consider additionally the unobserved variable ability (dashed means unobserved).



- ▶ We can only identify the total effect.
- ▶ Controlling for management would introduce an omitted variable bias.

# Manager Wage Premium

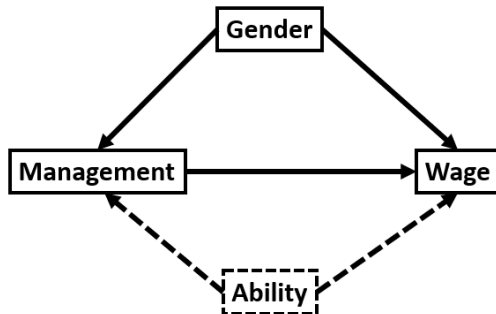
- We are interested in the effect of management position (left variable) on wages (right variable) → manager wage premium.



- We must control for gender, otherwise we have an omitted variable bias.

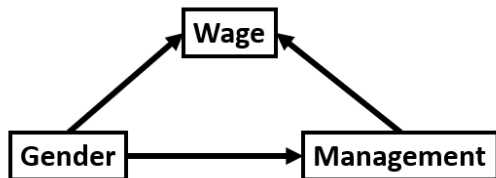
# Manager Wage Premium

- ▶ But unobserved ability could still cause an omitted variable bias, even after controlling for gender!
- ▶ Then the manager wage premium is not identified.



# Glass Ceiling in Career Ladders

- We are interested in the effect of gender (left variable) on management position (right variable) → glass ceiling in career ladders



- ▶ Do not control for wage!
- ▶ It is an outcome variable and accordingly endogenous (even if we ignore the unobserved ability).

# Overview

Selection Bias

Selection-on-Observables Identification Strategy

Power of Conditioning

Multivariate Regression

Post-Double-Selection Procedure



# Conditional Expectations of Potential Outcomes

- We saw on Slide 19, that

$$\delta = \int (\underbrace{E[Y^1|X=x]}_{=\mu_1(x)} - \underbrace{E[Y^0|X=x]}_{=\mu_0(x)}) f_X(x) dx$$

- We can identify  $\mu_1(x)$  and  $\mu_0(x)$  from observable data

$$\mu_1(x) = E[Y^1|X=x] \stackrel{CS, CIA}{=} E[Y^1|D=1, X=x] \stackrel{OR}{=} E[Y|D=1, X=x]$$

$$\mu_0(x) = E[Y^0|X=x] \stackrel{CS, CIA}{=} E[Y^0|D=0, X=x] \stackrel{OR}{=} E[Y|D=0, X=x]$$

- Using the sample analogy principle, an estimator for ATE is

$$\hat{\delta} = \frac{1}{N} \sum_{i=1}^N (\tilde{\mu}_1(X_i) - \tilde{\mu}_0(X_i)) \quad (2)$$

where  $\tilde{\mu}_1(X_i)$  and  $\tilde{\mu}_0(X_i)$  are the estimated conditional expectation of the potential outcome for observation units with characteristics  $X_i$

# Regression Model

- ▶ There are many possible ways how we can estimate  $\tilde{\mu}_1(X_i)$  and  $\tilde{\mu}_0(X_i)$
- ▶ A very simple way is to use OLS regressions
- ▶ We can estimate  $\tilde{\mu}_1(\cdot)$  and  $\tilde{\mu}_0(\cdot)$  in two separate empirical models

$\tilde{\mu}_1(X_i) = X_i \tilde{\beta}^1$  in the sample of participants with  $D = 1$

$\tilde{\mu}_0(X_i) = X_i \tilde{\beta}^0$  in the sample of non-participants with  $D = 0$

- ▶ After we have estimated the coefficients  $\tilde{\beta}^1$  and  $\tilde{\beta}^0$ , we can calculate  $\tilde{\mu}_1(X_i)$  and  $\tilde{\mu}_0(X_i)$  for the entire sample (since  $X_i$  is observed for all units  $i = 1, \dots, N$ )
- ▶ Accordingly, we have all ingredients to estimate (2)

# Additional Assumptions

- ▶ For the regression model we have to make additional parametric assumptions:
  1. **Linearity:**

We have to assume that the linear functional form is correct for  $\tilde{\mu}_1(X_i) = X_i\tilde{\beta}^1$  and  $\tilde{\mu}_0(X_i) = X_i\tilde{\beta}^0$
  2. **No Perfect Multicollinearity:**

We have to assume that the design matrix has full rank, otherwise the objective function of the OLS estimator has multiple solutions
- ▶ However, both additional assumptions can be relaxed:
  - ▶ We can add many non-linear and interaction terms in  $X$  to allow for more flexible functional forms
  - ▶ We can use linear machine learning estimators (e.g., Lasso, Ridge, Elastic Net) instead of OLS, which make it easier to handle very flexible models and can even deal with perfect multicollinearity

# Alternative Representation

- The empirical model interacted with the treatment dummy is an alternative representation for the conditional expectations of the potential outcomes

$$\tilde{\mu}_d(x) = \tilde{E}[Y^d | D = d, X = x] = x \cdot \tilde{\beta}^0 + d \cdot x \cdot \underbrace{(\tilde{\beta}^1 - \tilde{\beta}^0)}_{=\tilde{\gamma}} \quad (3)$$

where  $\tilde{\gamma}$  is a K-dimensional vector of coefficients

- We can rewrite the T-Learner as

$$\hat{\delta} = \frac{1}{N} \sum_{i=1}^N X_i \tilde{\gamma}$$

# Effect Homogeneity

- ▶ We assume additionally that the treatment effects do not vary with regard to the characteristics  $X$ , such that  $X\beta^1 = X\beta^0 + \alpha$ , where  $\alpha$  is a scalar
- ▶ Under effect homogeneity, the empirical model (3) simplifies to

$$\tilde{\mu}_d(x) = \tilde{E}[Y^d|D = d, X = x] = x \cdot \tilde{\beta}^0 + d \cdot \tilde{\alpha} \quad (4)$$

and the T-Learner simplifies to  $\hat{\delta} = \tilde{\alpha}$

- ▶ **Note that the canonical model in (4) is used very often to estimate ATEs, even though this model makes unnecessarily strong assumptions about linearity and effect homogeneity**

# Exclusion Restriction and Common Support

## ► Exclusion Restriction:

- In the undergraduate studies you learned that the exclusion restriction  $E[u|D, X] = 0$  is an important assumption to identify models like in (4)

$$Y = X\beta^0 + D\alpha + u$$

- The exclusion restriction is stronger than the CIA, but it would be sufficient to assume  $E[u|D, X] = E[u|X]$  if we are only interested in consistent estimates for  $\alpha$  and do not care so much about the estimates of  $\beta^0$

## ► Common Support:

- If the functional forms  $X\tilde{\beta}^1$  and  $X\tilde{\beta}^0$  are correct, we can relax the common support assumption, because we can extrapolate out of support.
- But too much extrapolation might lead to overfitting. Accordingly, we should be careful about common support violations even in OLS regressions!

# Overview

Selection Bias

Selection-on-Observables Identification Strategy

Power of Conditioning

Multivariate Regression

Post-Double-Selection Procedure



# Estimation Target

- ▶ Multivariate Linear Regression Model:

$$Y_i = D_i\delta + X_i\beta_g + U_i \quad (\text{structural model})$$

$$D_i = X_i\beta_m + V_i \quad (\text{selection model})$$

- ▶ Parameter of interest:  $\delta$
- ▶ Nuisance parameters:  $\beta_g$  and  $\beta_m$
- ▶  $X_i$  contains  $p \gg N$  covariates.
- ▶ We assume controlling for  $K \ll N$  covariates is sufficient to identify  $\delta$ .
- ▶ Controlling for too many irrelevant covariates may reduce the efficiency of OLS.

# Types of Covariates

Relation between covariates and outcome (for some  $s_g > 0$ ):

- ▶  $|\beta_{gj}| > s_g$ : covariate  $X_j$  has a **strong association** with  $Y_i$
- ▶  $0 < |\beta_{gj}| \leq s_g$ : covariate  $X_j$  has a **weak association** with  $Y_i$
- ▶  $\beta_{gj} = 0$ : covariate  $X_j$  has a **no association** with  $Y_i$

Relation between covariates and treatment (for some  $s_m > 0$ ):

- ▶  $|\beta_{mj}| > s_m$ : covariate  $X_j$  has a **strong association** with  $D_i$
- ▶  $0 < |\beta_{mj}| \leq s_m$ : covariate  $X_j$  has a **weak association** with  $D_i$
- ▶  $\beta_{mj} = 0$ : covariate  $X_j$  has a **no association** with  $D_i$

→ All covariates are standardised

## Types of Covariates (cont.)

	$\beta_{gj} = 0$	$0 <  \beta_{gj}  \leq s_g$	$ \beta_{gj}  > s_g$
$\beta_{mj} = 0$	Irrelevant	Irrelevant	Irrelevant
$0 <  \beta_{mj}  \leq s_m$	Irrelevant	Unclear?	Weak Confounder
$ \beta_{mj}  > s_m$	Irrelevant	Weak Confounder	Strong Confounder

- ▶  $|\beta_{gj}| > s_g$  and  $0 < |\beta_{mj}| \leq s_m$ : "Weak Outcome Confounder"
- ▶  $|\beta_{mj}| > s_m$  and  $0 < |\beta_{gj}| \leq s_g$ : "Weak Treatment Confounder"

# Naive Approach I: Structural Model

Apply Lasso to the structural model

$$\min_{\beta_g} \{E[(Y_i - D_i\delta - X_i\beta_g)^2] + \lambda\|\beta_g\|_1\}$$

without a penalty on  $\delta$  and estimate a Post-Lasso model using all covariates with non-zero  $\beta_g$  coefficients.

Covariates that are weakly associated with  $Y_i$  could be dropped.

→ Potentially we drop “weak treatment confounders”

Covariates that are strongly associated with  $D_i$  could be dropped.

→ Potentially we drop “strong confounders”

## Naive Approach II: Selection Model

Apply Lasso to the selection model

$$\min_{\beta_m} \{E[(D_i - X_i\beta_m)^2] + \lambda\|\beta_m\|_1\}$$

and estimate a Post-Lasso structural model using all covariates with non-zero  $\beta_m$  coefficients.

Covariates that are weakly associated with  $D_i$  could be dropped.

→ Potentially we drop “weak outcome confounders”

# Double Selection Procedure

1. Apply Lasso to the reduced form models

$$\min_{\tilde{\beta}_g} \{E[(Y_i - X_i \tilde{\beta}_g)^2] + \lambda \|\tilde{\beta}_g\|_1\}, \quad (5)$$

$$\min_{\beta_m} \{E[(D_i - X_i \beta_m)^2] + \lambda \|\beta_m\|_1\}, \quad (6)$$

with  $\tilde{\beta}_g = \delta \beta_m + \beta_g$ .

2. Take the union of all covariates  $\tilde{X}_i$  with either non-zero  $\beta_m$  or  $\tilde{\beta}_g$  coefficients and estimate the Post-Lasso structural model

$$Y_i = D_i \delta + \tilde{X}_i \beta_g^* + u_i.$$

# Double Selection Procedure (cont.)

Potentially (6) omits “weak outcome confounders”

$\tilde{\beta}_{gj} \approx \beta_g$  when  $0 < |\beta_{mj}| \leq s_m$ , such that the missing “weak outcome confounders” are likely selected in (5).

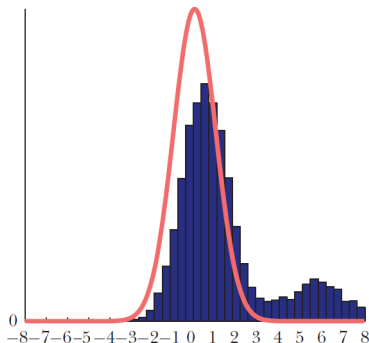
Disadvantages:

- Potentially we omit “very weak” confounders with  $0 < |\beta_{gj}| \leq s_g$  and  $0 < |\beta_{mj}| \leq s_g$ .
- All procedures potentially include irrelevant variables.

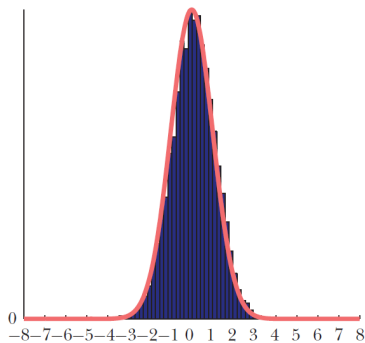
# Simulation Exercise

## Distribution of Estimators

Naive Single-Post-Selection  
on Structural Model



Double-Post-Selection



Source: [Belloni, Chernozhukov, and Hansen \(2014\)](#)



# Asymptotic Results

- Consistency and asymptotic normality

$$\sqrt{N}(\hat{\delta} - \delta) \xrightarrow{d} N(0, \sigma^2).$$

- Model selection step is asymptotically negligible for building confidence intervals.
- Optimal penalty parameter  $\lambda^* = 2c \cdot \Phi^{-1}(1 - \gamma/2p)/\sqrt{N}$  (e.g.,  $c = 1.1$  and  $\gamma \leq 0.05$ ) for "Feasible LASSO"

$$\min_{\beta} E[(Y_i - X_i\beta)^2] + \lambda^* \|\beta\|_1.$$

Reference: [Belloni, Chernozhukov, and Hansen \(2014\)](#)