Introduction to Causal Machine Learning

Accounting for Confounders with Double ML

Anthony Strittmatter

Reference

Huber (2023): "Causal Analysis: Impact Evaluation and Causal Machine Learning with Applications in R", Chapters 5.1-5.2, online version.

Chernozhukov, Hansen, Kallus, Spindler, Syrgkanis (2024): "The Causal ML Book", Chapter 4, download.

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.

Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey (2017): "Double/Debiased/Neyman Machine Learning of Treatment Effects", American Economic Review, 107 (5), pp. 261-265, download.

Outline

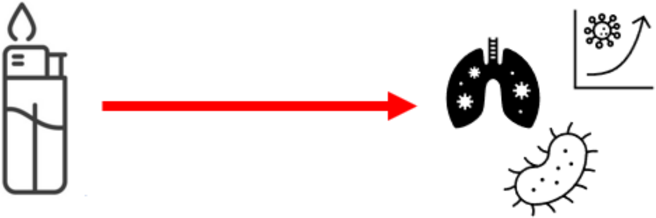
Impact Evaluation

- 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?

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

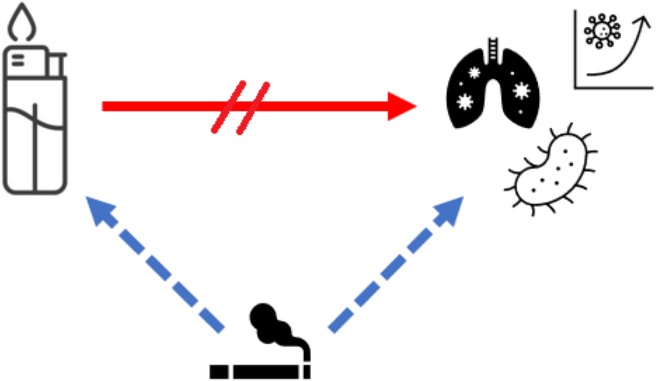
Causal Pitfalls

Impact of having a lighter in your pocket on the likelihood of lung cancer?



Causal Pitfalls

Impact of having a lighter in your pocket on the likelihood of lung cancer?



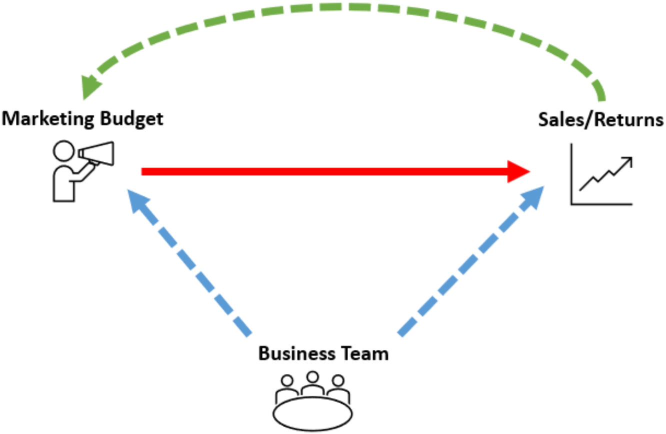
Philip Morris Advertisement

Causal Pitfalls

Impact of marketing budget on sales/returns?

Causal Pitfalls

Impact of marketing budget on sales/returns?



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

  $\rightarrow$ 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

$$\text{ATE} = E[Y^1] - E[Y^0]$$

Stable Unit Treatment Value Assumption (SUTVA)

$$Y = D \cdot Y^1 + (1 - D) \cdot Y^0 \tag{1}$$

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:

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

  - 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):

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

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:
  $$E[Y^1|D=1] - E[Y^0|D=1] = E[Y^1] - E[Y^0] = \text{ATE}$$
  - ATENT:
  $$E[Y^1|D=0] - E[Y^0|D=0] = E[Y^1] - E[Y^0] = \text{ATE}$$
  - 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 of Randomised 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 of Randomised 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.

$\Rightarrow$ There is need for alternative empirical strategies!

Outline

Notation

- We assume to observe i.i.d. (independent and identically distributed) data on the triple (Y, D, X)
- 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
- $\mu_d(x) = E[Y^d|X = x]$: 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)

Individual Causal Effects

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

for observation units $i = 1, ..., N$ (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 assumptions that are unplausible in most applications

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 <u>joint</u> impact on the treatment and the potential outcomes.
- All covariates X have to be exogeneous (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 δ from observable data (Y, D, X):

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

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

Power of Conditioning

- Observable expected earnings:

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

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

	$E[Y^0 \vert D=1, X=x]$	$E[Y^1 \vert D=0, X=x]$
Women ($X=1$)	3500	3500
Men ($X=0$)	4875	5750

True Causal Effects

- Average Treatment Effect on the Treated (ATET):

$$\begin{aligned} ATET =& Pr(X=1|D=1)\cdot(E[Y^1|D=1,X=1]-E[Y^0|D=1,X=1]) \\ &+Pr(X=0|D=1)\cdot(E[Y^1|D=1,X=0]-E[Y^0|D=1,X=0]) \\ =& \frac{10}{50}\cdot(4000-3500)+\frac{40}{50}\cdot(5000-4875)=200 \end{aligned}$$

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

$$\begin{aligned} ATENT =& Pr(X=1|D=0)\cdot(E[Y^1|D=0,X=1]-E[Y^0|D=0,X=1]) \\ &+Pr(X=0|D=0)\cdot(E[Y^1|D=0,X=0]-E[Y^0|D=0,X=0]) \\ =& \frac{30}{50}\cdot(3500-3000)+\frac{20}{50}\cdot(5750-5000)=600 \end{aligned}$$

- Average Treatment Effect (ATE):

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

Naive Estimator

- Expected earnings of participants:

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

- Expected earnings of non-participants:

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

- Naive estimator:

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

Average Treatment Effect on the Treated (ATET)

Under Assumptions 1-3,

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

Selection bias for ATET:

- Share of women lower among participants than non-participants (and *vice versa* for men) $(-)$
- Effects of participation are lower for treated women than treated men $(-)$

$\rightarrow$ Positive bias $(= 1000 - 200 = 800)$!

Average Treatment Effect on the Non-Treated (ATENT)

Under Assumptions 1-3,

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

Average Treatment Effect (ATE)

- ATE:

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

$\rightarrow$ 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 ($N = 59$)
Management	2796.22 Euro ($N = 13$)	2659.91 Euro ($N = 41$)

- 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

$\Rightarrow$ **What is the correct gender wage gap?**

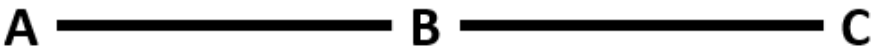
$\Rightarrow$ **Do we need to control for management position or not?**

$\Rightarrow$ The seemingly contradicting results of the conditional and unconditional estimator are called Simpson's Paradox.

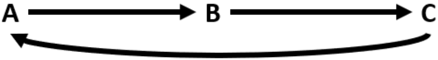
$\Rightarrow$ The correct answers depends on the (typically untestable) assumptions we impose.

Directed Acyclic Graphs (DAGs)

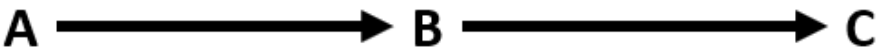
- Undirected graphs:



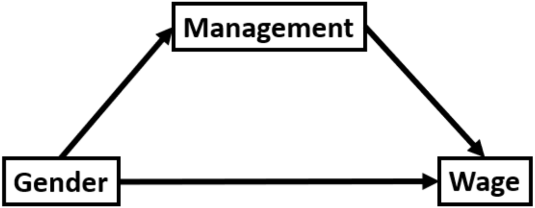
- Directed cyclic graphs:



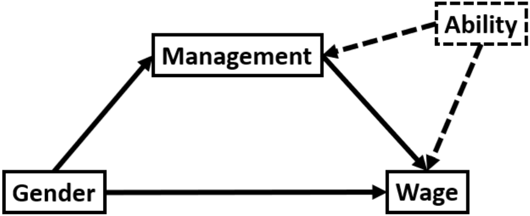
- Directed acyclic graphs:



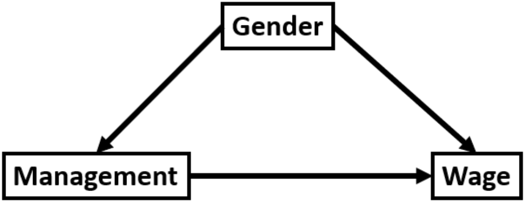
Gender Wage Gap



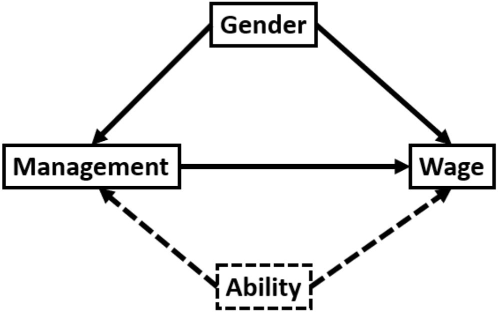
Gender Wage Gap



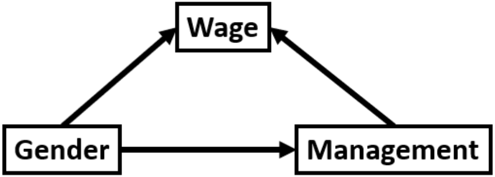
Manager Wage Premium



Manager Wage Premium



Glass Ceiling Effect



Overview

Conditional Expectations of Potential Outcomes

- We saw on Slide 22, that

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

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

$$\mu_1(x) = E[Y|D=1, X=x] \overset{OR}{=} E[Y^1|D=1, X=x] \overset{CS,CIA}{=} E[Y^1|X=x]$$
$$\mu_0(x) = E[Y|D=0, X=x] \overset{OR}{=} E[Y^0|D=0, X=x] \overset{CS,CIA}{=} E[Y^0|X=x]$$

T-Learner

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

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

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

$$\tilde{\mu}_1(X_i) = X_i\tilde{\beta}^1 \text{ in the sample of participants with } D = 1$$
$$\tilde{\mu}_0(X_i) = X_i\tilde{\beta}^0 \text{ in the sample of non-participants with } D = 0$$

- 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 $i = 1, ..., N$)
- Accordingly, we have all ingredients to estimate (2)

Alternative Representation

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

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

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 α is a scalar
- Under effect homogeneity, the empirical model (3) simplifies to

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

  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**

Overview

Estimation Target

- Multivariate Linear Regression Model:

$$
\begin{aligned}
Y_i &= D_i\delta + X_i\beta_g + U_i && \text{(structural model)} \\
D_i &= X_i\beta_m + V_i && \text{(selection model)}
\end{aligned}
$$

- Parameter of interest: δ
- Nuisance parameters: β_g and β_m

- X_i contains $p \gg N$ covariates.
- We assume controlling for $K \ll N$ covariates is sufficient to identify δ.
- 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

$\rightarrow$ 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 δ and estimate a Post-Lasso model using all covariates with non-zero β_g coefficients.

Covariates that are weakly associated with Y_i could be dropped.

$\rightarrow$ Potentially we drop "weak treatment confounders"

Covariates that are strongly associated with D_i could be dropped.

$\rightarrow$ Potentially we drop "strong confounders"

Illustration Naive Approach I

- Effect of assignment to a training programme on earnings
- Stratified experiment $\rightarrow$ randomisation within gender groups
- Women are more likely to be assigned to training programme

$\Rightarrow$ The only confounder is gender

	OLS Unbiased	OLS Biased	Lasso	Post-Lasso
Assignment	18.969	-52.473	-49.872	-47.306
Female	-87.451			
High School			16.675	44.458
White			7.916	23.954
African-American			-18.539	-37.835
Work Experience			28.373	41.210
Employed			5.568	24.790
Employed Last Year			21.552	31.755
Previous Earnings			14.720	52.168
Intercept	236.186	231.316	208.011	188.298

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 β_m coefficients.

Covariates that are weakly associated with D_i could be dropped.

$\rightarrow$ 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 β_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"

$\tilde{\beta}_g \approx \beta_g$ when $0 < |\beta_m| \leq s_m$, such that the missing "weak outcome confounders" are likely selected in (5).

Disadvantages:

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

Excursus: Omitted Variable Bias

- Suppose the true cause-and-effect relationship is:

$$Y = D\delta + X\beta_g + U$$

- The omitted variable in (5) is:

$$D = X\beta_m + V$$

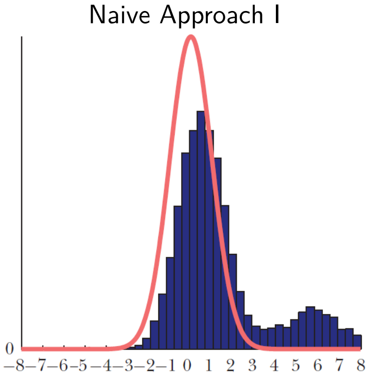
- Merging the two equations gives:

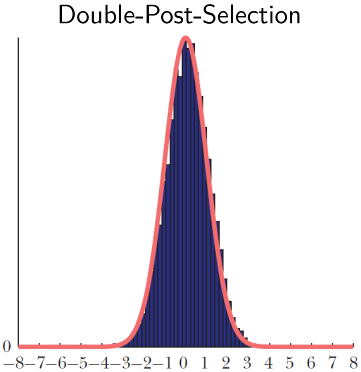
$$\begin{aligned} Y &= (X\beta_m + V)\delta + X\beta_g + U \\ &= X\beta_m\delta + V\delta + X\beta_g + U \\ &= X\underbrace{(\beta_m\delta + \beta_g)}_{\tilde{\beta}_g} + (V\delta + U) \end{aligned}$$

- The omitted variable bias is: $\beta_m\delta$
- When $0 < |\beta_m| \leq s_m$, the omitted variable bias is ≈ 0 and $\tilde{\beta}_g \approx \beta_g$

Simulation Exercise

Distribution of Estimators





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)

Outline

Partial Regression

Frisch-Waugh-Lovell (FWL) Theorem

- Suppose we want to estimate the coefficient δ in the model:

$$Y = D\delta + X\beta_g + U$$

- Applying the FWL Theorem, we can retrieve the estimated coefficient $\hat{\delta}$ from

$$\tilde{Y} = \tilde{D}\hat{\delta} + U$$

after partialling-out

$$\tilde{Y} = Y - X\hat{\beta}_g$$
$$\tilde{D} = D - X\hat{\beta}_m$$

$\Rightarrow$ YouTube Video explaining FWL Theorem

Double Lasso Procedure

1. Apply Lasso to the reduced form models

$$\min_{\tilde{\beta}_g} \{E[(Y_i - X_i\hat{\beta}_g)^2] + \lambda \|\hat{\beta}_g\|_1\},$$
$$\min_{\beta_m} \{E[(D_i - X_i\hat{\beta}_m)^2] + \lambda \|\hat{\beta}_m\|_1\},$$

   and obtain the resulting residuals:

$$\tilde{Y}_i = Y_i - X_i\hat{\beta}_g$$
$$\tilde{D}_i = D_i - X_i\hat{\beta}_m$$

2. We run the least squares regression of $\tilde{Y}_i$ on $\tilde{D}_i$ to obtain the estimate $\hat{\delta}$. We can use standard results from this regression, ignoring that the input variables were previously estimated, to perform inference about $\hat{\delta}$.

Partialling Out Procedure

Main Advantages:

- Generic approach, can be combined with any supervised ML estimator
- Sparsity assumptions can be avoided by appropriate choice of estimators

Main Disadvantages:

- Still assumption of linearity for the main effect
- Does not incorporate effect heterogeneity

Outline

T-Learner for ATE

Identification:

$$\begin{aligned}\delta =& E[Y_i(1)] - E[_iY(0)] \\ =& \int \underbrace{E[Y_i|D_i = 1, X_i = x]}_{=\mu_1(x)} - \underbrace{E[Y_i|D_i = 0, X_i = x]}_{=\mu_0(x)} f_X(x)dx\end{aligned}$$

Estimator:

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

with $\hat{\mu}_1(x) = \hat{E}[Y_i|D_i = 1, X_i = x]$ and $\hat{\mu}_0(x) = \hat{E}[Y_i(0)|D_i = 0, X_i = x]$ being the estimated conditional expectations of the potential outcomes.

T-Learner

Main Advantages:

- Generic approach
- Sparsity assumptions can be avoided by appropriate choice of estimator for propensity score
- Heterogeneous treatment effects

Main Disadvantages:

- Potentially omitting "weak selection confounders"
- Not $\sqrt{N}$-consistent in high-dimensional setting

Modified Outcome Method for ATE

Inverse Probability Weighting:

$$Y^*_{i,IPW} = \frac{D_i}{p(X_i)} Y_i - \frac{1 - D_i}{1 - p(X_i)} Y_i = \frac{D_i - p(X_i)}{p(X_i)(1 - p(X_i))} Y_i$$

with the propensity score $p(x) = Pr(D_i = 1 | X_i = x)$.

ATE: $\delta = E[Y^*_{i,IPW}]$ and $\hat{\delta} = \frac{1}{N} \sum_{i=1}^{N} \hat{Y}^*_{i,IPW}$

Proof of Identification

$$
\begin{aligned}
\delta = E\left[Y_i(1)\right] - E\left[Y_i(0)\right] &\overset{LIE}{=} \int E\left[Y_i(1)|X_i = x\right] - E\left[Y_i(0)|X_i = x\right] f_X(x)dx \\
&\overset{CIA}{=} \int E\left[Y_i(1)|D_i = 1, X_i = x\right] - E\left[Y_i(0)|D_i = 0, X_i = x\right] f_X(x)dx \\
&= \int E\left[Y_i|D_i = 1, X_i = x\right] - E\left[Y_i|D_i = 0, X_i = x\right] f_X(x)dx \\
&= \int E\left[D_i Y_i|D_i = 1, X_i = x\right] - E\left[(1 - D_i)Y_i|D_i = 0, X_i = x\right] f_X(x)dx \\
&\overset{LIE}{=} \int E\left[\left.\frac{D_i Y_i}{p(X_i)}\right| X_i = x\right] - E\left[\left.\frac{(1 - D_i)Y_i}{1 - p(X_i)}\right| X_i = x\right] f_X(x)dx \\
&= \int E\left[\left.\frac{D_i Y_i}{p(X_i)} - \frac{(1 - D_i)Y_i}{1 - p(X_i)}\right| X_i = x\right] f_X(x)dx \\
&= \int E\left[\left.\frac{D_i - p(X_i)}{p(X_i)(1 - p(X_i))} Y_i\right| X_i = x\right] f_X(x)dx \overset{LIE}{=} E\left[\frac{D_i - p(X_i)}{p(X_i)(1 - p(X_i))} Y_i\right]
\end{aligned}
$$

Reference: Horvitz and Thompson (1952)

Modified Outcome Method with IPW

Main Advantages:

- Generic approach
- Sparsity assumptions can be avoided by appropriate choice of estimator for propensity score
- Heterogeneous treatment effects

Main Disadvantages:

- Potentially omitting "weak outcome confounders"
- Shows weak performance in simulations (Knaus, Lechner, and Strittmatter, 2018)
- Not $\sqrt{N}$-consistent in high-dimensional setting

Double/Debiased Machine Learning (DML)

Efficient Score:

$$
\begin{aligned}
Y^*_{i,DML} &= \mu_1(X_i) - \mu_0(X_i) + \frac{D_i - p(X_i)}{p(X_i)(1-p(X_i))} Y_i - \frac{D_i}{p(X_i)}\mu_1(X_i) + \frac{1-D_i}{1-p(X_i)}\mu_0(X_i) \\
&= \mu_1(X_i) - \mu_0(X_i) + \frac{D_i(Y_i - \mu_1(X_i))}{p(X_i)} - \frac{(1-D_i)(Y_i - \mu_0(X_i))}{1-p(X_i)}
\end{aligned}
$$

ATE: $\delta = E[Y^*_{i,DML}]$ and $\hat{\delta} = \dfrac{1}{N}\sum_{i=1}^{N} \hat{Y}^*_{i,DML}$

We can use standard ML methods to estimate $\hat{\mu}_1(x)$, $\hat{\mu}_0(x)$, and $\hat{p}(x)$.

Reference: Chernozhukov et al., 2017

Proof of Identification

$$\delta = E\left[\mu_1(X_i) - \mu_0(X_i) + \frac{D_i(Y_i - \mu_1(X_i))}{p(X_i)} - \frac{(1-D_i)(Y_i - \mu_0(X_i))}{1-p(X_i)}\right]$$

$$= \int \mu_1(x) - \mu_0(x) + \frac{E\left[D_i Y_i \middle| X_i = x\right] - E\left[D_i \mu_1(x) \middle| X_i = x\right]}{p(x)}$$

$$- \frac{E\left[(1-D_i)Y_i \middle| X_i = x\right] - E\left[(1-D_i)\mu_0(x) \middle| X_i = x\right]}{1-p(x)} f_X(x)dx$$

$$= \int \mu_1(x) - \mu_0(x) + \frac{p(x)(E\left[Y_i \middle| D_i = 1, X_i = x\right] - \mu_1(x))}{p(x)}$$

$$- \frac{(1-p(x))(E\left[Y_i \middle| D_i = 0, X_i = x\right] - \mu_0(x))}{1-p(x)} f_X(x)dx$$

$$= \int \mu_1(x) - \mu_0(x) + (\underbrace{E\left[Y_i^1 \middle| X_i = x\right]}_{=\mu_1(x)} - \mu_1(x)) - (\underbrace{E\left[Y_i^0 \middle| X_i = x\right]}_{=\mu_0(x)} - \mu_0(x)) f_X(x)dx$$

$$= \int \mu_1(x) - \mu_0(x) + \underbrace{(\mu_1(x) - \mu_1(x))}_{=0} - \underbrace{(\mu_0(x) - \mu_0(x))}_{=0} f_X(x)dx$$

$$= \int \mu_1(x) - \mu_0(x) f_X(x)dx$$

DML Cross-Fitting Algorithm

1. Partition the data randomly in samples S^A and S^B
2. Estimate the nuisance parameters $\hat{\mu}_1^A(x)$,$\hat{\mu}_0^A(x)$, and $\hat{p}^A(x)$ in S^A; and $\hat{\mu}_1^B(x)$,$\hat{\mu}_0^B(x)$, and $\hat{p}^B(x)$ in S^B with ML
3. Calculate the efficient scores in samples S^A and S^B, respectively:

$$\hat{Y}_{i,DML}^{A*} = \hat{\mu}_1^B(X_i^A) - \hat{\mu}_0^B(X_i^A) + \frac{D_i^A(Y_i^A - \hat{\mu}_1^B(X_i^A))}{\hat{p}^B(X_i^A)} - \frac{(1-D_i^A)(Y_i^A - \hat{\mu}_0^B(X_i^A))}{1-\hat{p}^B(X_i^A)}$$

$$\hat{Y}_{i,DML}^{B*} = \hat{\mu}_1^A(X_i^B) - \hat{\mu}_0^A(X_i^B) + \frac{D_i^B(Y_i^B - \hat{\mu}_1^A(X_i^B))}{\hat{p}^A(X_i^B)} - \frac{(1-D_i^B)(Y_i^B - \hat{\mu}_0^A(X_i^B))}{1-\hat{p}^A(X_i^B)}$$

4. Calculate ATE,

$$\hat{\delta} = \frac{1}{2}(\underbrace{\hat{E}[\hat{Y}_{i,DML}^{A*}|S^A]}_{=\hat{\delta}_A} + \underbrace{\hat{E}[\hat{Y}_{i,DML}^{B*}|S^B]}_{=\hat{\delta}_B}),$$

Asymptotic Results for ATE

- Main Regularity Condition: Convergence rate of nuisance parameters is at least $\sqrt[4]{N}$.
- ATE can be estimated $\sqrt{N}$-consistently

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

with $\sigma^2 = Var(Y^*_{i,DML})$ and $Var(\hat{\delta}) = \sigma^2/N$

- Split sample estimator of σ^2

$$\hat{\sigma}^2 = \frac{1}{2}\left(\hat{\sigma}^2_A + (\hat{\delta}_A - \hat{\delta})^2\right) + \frac{1}{2}\left(\hat{\sigma}^2_B + (\hat{\delta}_B - \hat{\delta})^2\right)$$

for $\hat{\delta} = 1/2(\hat{\delta}_A + \hat{\delta}_B)$

Advantages of DML

Advantages compared to IPW and T-Learner:

- Treatment and outcome equations are modelled explicitly
- Double robustness property
- $\sqrt{N}$-consistent and asymptotically normal even under high-dimensional confounding

Efficient Score for ATET

$$Y^*_{i,ATET} = \frac{D_i(Y_i - \mu_0(X_i))}{p} - \frac{p(X_i)(1 - D_i)(Y_i - \mu_0(X_i))}{p(1 - p(X_i))}$$

with $p = Pr(D_i = 1)$.

ATET: $\rho = E[Y^*_{i,ATET}]$ and $\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N} \hat{Y}^*_{i,ATET}$

Estimator of Asymptotic Variance:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{Y}^*_{i,ATET} - \hat{\rho}\right)^2$$

and $\hat{Var}(\hat{\rho}) = \hat{\sigma}^2/N$

References: Chernozhukov et al., 2017, Farrell, 2015

Proof of Identification for ATET

$$
\begin{aligned}
\rho =& E\left[\frac{D_i(Y_i-\mu_0(X_i))}{p}-\frac{p(X_i)(1-D_i)(Y_i-\mu_0(X_i))}{p(1-p(X_i))}\right] \\
=& \int E\left[\frac{D_iY_i}{p}-\frac{p(X_i)(1-D_i)Y_i}{p(1-p(X_i))}-\frac{(D_i-p(X_i))\mu_0(X_i)}{p(1-p(X_i))}\middle| X_i=x\right] f_X(x)dx \\
=& \int \left(\frac{E[D_iY_i|X_i=x]}{p}-\frac{p(x)E[(1-D_i)Y_i|X_i=x]}{p(1-p(x))}\right. \\
& \left.-\frac{E[D_i-p(X_i)|X_i=x]}{p(1-p(x))}\mu_0(x)\right) f_X(x)dx \\
=& \int \left(\frac{E[D_iY_i|X_i=x]}{p}-\frac{p(x)E[(1-D_i)Y_i|X_i=x]}{p(1-p(x))}\right) f_X(x)dx \\
=& \int \frac{p(x)}{p}\left(E[D_iY_i|D_i=1,X_i=x]-E[(1-D_i)Y_i|D_i=0,X_i=x]\right) f_X(x)dx \\
=& \int \left(E[Y_i(1)|D_i=1,X_i=x]-E[Y_i(0)|D_i=0,X_i=x]\right) f_{X|D=1}(x)dx \\
=& \int \left(E[Y_i(1)|D_i=1,X_i=x]-E[Y_i(0)|D_i=1,X_i=x]\right) f_{X|D=1}(x)dx \\
=& E[Y_i(1)-Y_i(0)|D_i=1]
\end{aligned}
$$