

Introduction to causal inference

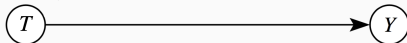
Julie Josse (Inria) & Bernard Sebastien (Sanofi R&D)

Charlotte VOINOT

March 5, 2024

What is Causal Inference ?

⇒ Effect of a policy/intervention/treatment T on an outcome Y



Causal Inference : example of questions

⇒ **Effect of a policy/intervention/treatment T on an outcome Y**

- What is the impact of an oncology medicine on long term mortality ?
- What impact do social networks have on the mental health of adolescents and young adults ?

Causal Inference : example of questions

⇒ **Effect of a policy/intervention/treatment T on an outcome Y**

- What is the impact of an oncology medicine on long term mortality ?
- What impact do social networks have on the mental health of adolescents and young adults ?

In your related topic :



What is the effect of using a specific organic fertilizer on a specific crop yields ?

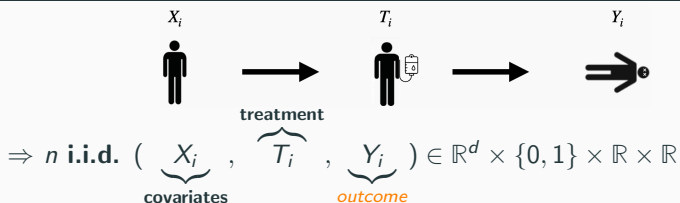


How do water management techniques affect crop growth and yield ?



What is the impact of specific genetic variations on the expression of genes involved in a given metabolic pathway ?

Potential outcomes



Let's say that in our example $X_1 = \text{age}$ and $X_2 = \text{sex}$.

Covariates		Treatment	Outcome	Potential outcomes ¹	
X_1	X_2	T	Y	$Y(0)$	$Y(1)$
20	F	1	67	?	67
45	F	0	83	83	?
...
52	M	0	100	100	?

Our goal is to compute the individual causal effect of the treatment:

$$\Delta_i = Y_i(1) - Y_i(0)$$

¹Donald B Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies, 1974

Identification of Average Treatment Effect

Individual causal effect of the treatment:

$$\Delta_i = Y_i(1) - Y_i(0)$$

However, the two potential outcomes cannot be observed : **fundamental problem of causal inference**.

In order to fix the problem, we need to define the Average Treatment Effect:

Average Treatment Effect (ATE)

$$\tau = \mathbb{E}[\Delta] = \mathbb{E}[Y(1) - Y(0)]$$

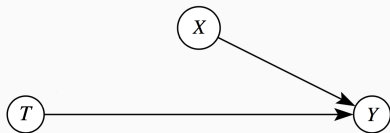
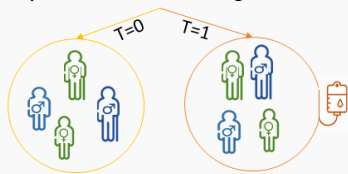
The ATE is the difference of the average outcome had everyone gotten treated and the average outcome had nobody gotten the treatment.

Causal Inference in RCT

Randomized clinical trial (RCT)



1) Random treatment assignment



Corresponding assumptions

1. $T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1), X_i\}$ (random treatment assignment)
2. $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ (STUVA)

Randomized Controlled Trial

Identifiability assumptions

- $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$
(STUVA : Consistency & No interference)
- $T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1), X_i\}$ (random treatment assignment)
Flip a coin to assign the treatment

$$\begin{aligned}\text{We now have } \tau &= \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)] \\ &= \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \\ &= \mathbb{E}[Y_i(1) | T_i = 1] - \mathbb{E}[Y_i(0) | T_i = 0] \\ &= \mathbb{E}[Y_i | T_i = 1] - \mathbb{E}[Y_i | T_i = 0]\end{aligned}$$

We say that τ is **identifiable** if it can be computed using a infinite number of observations from it.

Randomized Controlled Trial

Identifiability assumptions

- $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ (STUVA)
- $T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1), X_i\}$ (random treatment assignment)

Flip a coin to assign the treatment

$$\begin{aligned}\text{We now have } \tau &= \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \\ &= \mathbb{E}[Y_i | T_i = 1] - \mathbb{E}[Y_i | T_i = 0]\end{aligned}$$

Covariates		Treatment	Outcome	Potential outcomes	
X_1	X_2	T	Y	Y(0)	Y(1)
20	F	1	67	?	67
45	F	0	83	83	?

52	M	0	100	100	?

$$\hat{\tau}_{DM} = \frac{1}{n_1} \sum_{T_i=1} Y_i - \frac{1}{n_0} \sum_{T_i=0} Y_i; \quad \tau = \text{mean(blue)} - \text{mean(red)}$$

Randomized Controlled Trial

Identifiability assumptions

- $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ (consistency)
- $T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1), X_i\}$ (random treatment assignment)

Flip a coin to assign the treatment

Difference-in-means estimator


$$\hat{\tau}_{DM} = \frac{1}{n_1} \sum_{i=1}^n T_i Y_i - \frac{1}{n_0} \sum_{i=1}^n (1 - T_i) Y_i$$

where $n_1 = \sum_{i=1}^n T_i$ and $n_0 = \sum_{i=1}^n 1 - T_i$

$\hat{\tau}_{DM}$ unbiased and \sqrt{n} -consistent $\sqrt{n}(\hat{\tau}_{DM} - \tau) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, V_{DM})$

with $V_{DM} = \frac{\text{Var}(Y_i(0))}{\mathbb{P}(T_i=0)} + \frac{\text{Var}(Y_i(1))}{\mathbb{P}(T_i=1)}$.

Randomized Controlled Trial (RCT)

- **gold standard** (allocation )
- same covariate distributions of treated and control groups
⇒ **High internal validity**

Randomized Controlled Trial (RCT)

- **gold standard** (allocation )
- same covariate distributions of treated and control groups
⇒ **High internal validity**
- expensive, long, ethical limitations
- small sample size: restrictive inclusion criteria
⇒ No personalized medicine
- **trial sample different from the population eligible for treatment**
⇒ **Low external validity**

Data sources & evidences to estimate the treatment effect

Randomized Controlled Trial (RCT)

- **gold standard** (allocation )
- same covariate distributions of treated and control groups
⇒ **High internal validity**
- expensive, long, ethical limitations
- small sample size: restrictive inclusion criteria
⇒ No personalized medicine
- **trial sample different from the population eligible for treatment**
⇒ **Low external validity**

Observational data

- low cost con
- large amounts of data (registries, biobanks, EHR, claims)
⇒ patient's heterogeneity
- **representative of the target populations**
⇒ **High external validity**

Data sources & evidences to estimate the treatment effect

Randomized Controlled Trial (RCT)

- **gold standard** (allocation )
- same covariate distributions of treated and control groups
⇒ **High internal validity**
- expensive, long, ethical limitations
- small sample size: restrictive inclusion criteria
⇒ No personalized medicine
- **trial sample different from the population eligible for treatment**
⇒ **Low external validity**

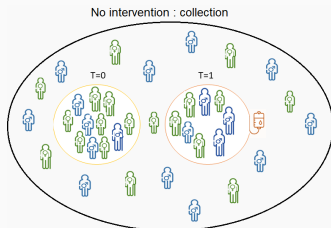
Observational data

- “big data”: low quality
- lack of a controlled design opens the door to **confounding bias**
⇒ **Low internal validity**
- low cost con
- large amounts of data (registries, biobanks, EHR, claims)
⇒ patient's heterogeneity
- **representative of the target populations**
⇒ **High external validity**

Observational Trial

The population is observed without any intervention by the investigator :
non experimental study so non random assignment.

Let's say that we focus on the same treatment in an observational study :



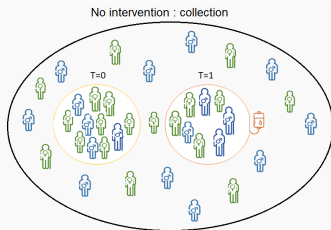
We obtain surprising results :

	Survived	Deceased	$P(\text{Survived} \mid \text{Treatment})$	$P(\text{Deceased} \mid \text{Treatment})$
No treated	205	45	0,82	0,18
Treated	27	23	0,54	0,46

Observational Trial

The population is observed without any intervention by the investigator :
non experimental study so non random assignment.

Let's say that we focus on the same treatment in an observational study :

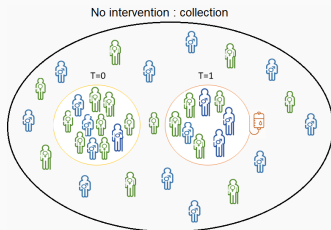


We obtain surprising results :

	Survived	Deceased	$P(\text{Survived} \mid \text{Treatment})$	$P(\text{Deceased} \mid \text{Treatment})$
No treated	205	45	0,82	0,18
Treated	27	23	0,54	0,46

- Is the treatment killing people ?

What could be the problem ?



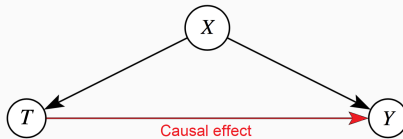
If we focus on the adjustment of covariates, we can see that the covariates are **unadjusted** between the groups of treatment

Covariates	T=0	T=1
Severity (from grade 1 to 3)	1,3	2,5
Age	60	75

Severe patients and older patients (with a higher risk of death) are more likely to be treated \Rightarrow **Confounding bias**

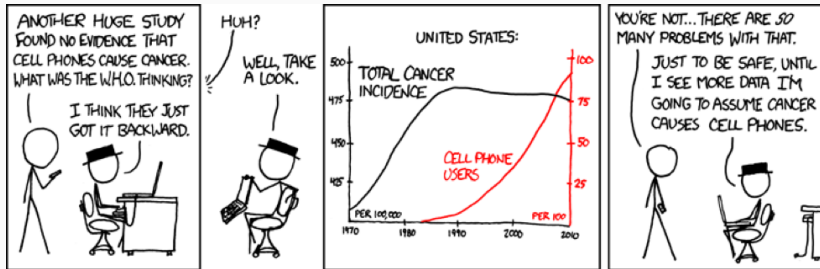
Confounding bias ?

⇒ Effect of a policy/intervention/treatment T on an outcome Y



- Let T be the treatment of interest
- Y the outcome
- X the confounding variables

We want to predict what would happen if we change the system



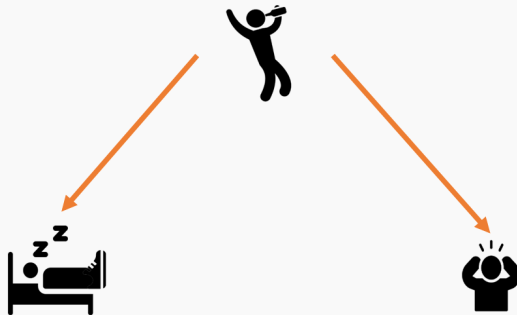
Key point : Correlation does not imply causation

Causal Inference : Correlation does not imply causation



Sleeping with shoes on is strongly correlated with waking up with a headache

Causal Inference: Correlation does not imply causation



Sleeping with shoes on is strongly correlated with waking up with a headache

Common cause : drinking the night before

Observational Trial : How to adjust covariates ?

Unconfoundedness

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$$

Measure all possible confounders

Unobserved confounders make it impossible to separate correlation and causality when correlated to both the outcome and the treatment.

Assumption for ATE identifiability in observational data

Overlap

Propensity score: probability of treatment given observed covariates.

$$e(x) \triangleq \mathbb{P}(T_i = 1 | X_i = x) \quad \forall x \in \mathcal{X}.$$

We assume overlap, i.e. $\eta < e(x) < 1 - \eta$, $\forall x \in \mathcal{X}$ and some $\eta > 0$

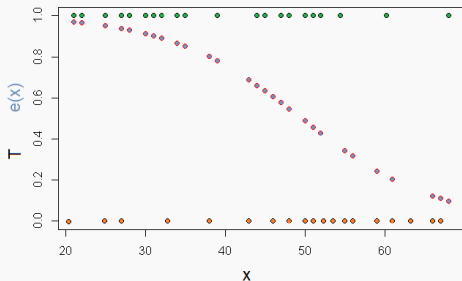


Figure 1: Example of propensity score estimation in one dimensional case : logistic regression

G-formula estimator

Average treatment effect (ATE): $\tau = \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$

Identifiability assumptions in observational data

- $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$ **(Unconfoundedness)**
- $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ **(Consistency)**
- $\eta < e(x) < 1 - \eta, \quad \forall x \in \mathcal{X}$ and some $\eta > 0$ **(Positivity)**

Using the law of total expectation,

$$\begin{aligned}\tau &= \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \\ &= \mathbb{E}[\mathbb{E}[Y_i(1)|X]] - \mathbb{E}[\mathbb{E}[Y_i(1)|X]] && \text{Law of total probability} \\ &= \mathbb{E}[\mathbb{E}[Y_i(1)|T_i = 1, X]] - \mathbb{E}[\mathbb{E}[Y_i(0)|T_i = 0, X]] && \text{Unconfoundedness \& Positivity} \\ &= \mathbb{E}[\mathbb{E}[Y_i|T_i = 1, X]] - \mathbb{E}[\mathbb{E}[Y_i|T_i = 0, X]] && \text{Consistency}\end{aligned}$$

G-formula estimator

Average treatment effect (ATE): $\tau = \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$

Identifiability assumptions in observational data

- $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$ (Unconfoundedness)
- $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ (Consistency)
- $\eta < e(x) < 1 - \eta, \quad \forall x \in \mathcal{X}$ and some $\eta > 0$ (Positivity)

Using the law of total expectation,

$$\begin{aligned}\tau &= \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \\ &= \mathbb{E}[\mathbb{E}[Y_i(1) \mid X]] - \mathbb{E}[\mathbb{E}[Y_i(1) \mid X]] \quad \text{Law of total probability} \\ &= \mathbb{E}[\mathbb{E}[Y_i(1) \mid T_i = 1, X]] - \mathbb{E}[\mathbb{E}[Y_i(0) \mid T_i = 0, X]] \quad \text{Unconfoundedness \& Positivity} \\ &= \mathbb{E}[\mathbb{E}[Y_i \mid T_i = 1, X]] - \mathbb{E}[\mathbb{E}[Y_i \mid T_i = 0, X]] \quad \text{Consistency}\end{aligned}$$

G-formula estimator

$$\hat{\tau}_G = \frac{1}{n} \sum_{i=1}^n \hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i)$$

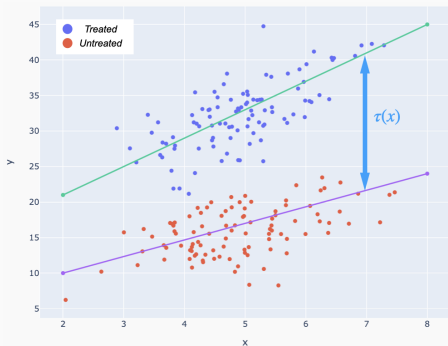
where $\mu_{(t)}(X) = \mathbb{E}[Y \mid T = t, X]$

G-formula estimator

G-formula estimator

$$\hat{\tau}_G = \frac{1}{n} \sum_{i=1}^n \hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i)$$

where $\mu_{(t)}(X) = \mathbb{E}[Y|T = t, X]$



In assuming that assumption of **Unconfoundedness**, **Consistency** and **Positivity** are satisfied and for $t \in \{0, 1\}$ we have:

$$\mathbb{E}[\hat{\mu}_{t,n}(X)] \xrightarrow{P} \mathbb{E}[\mu_t(X)]$$

then **T-learner estimator** is an **unbiased estimator of the ATE**:

$$\mathbb{E}[\hat{\tau}_G] = \tau$$

Inverse-propensity weighting estimator

Average treatment effect (ATE): $\tau = \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$

Identifiability assumptions in observational data

- $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$ **(Unconfoundedness)**
- $\eta < e(x) < 1 - \eta, \quad \forall x \in \mathcal{X}$ and some $\eta > 0$ **(Overlap)**
- $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ **(Consistency)**

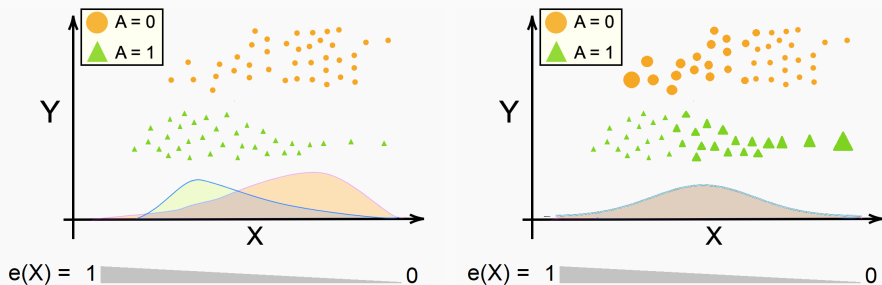
Propensity score : $e(x) = \mathbb{P}(T_i = 1 \mid X_i = x)$

$$\begin{aligned}\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\ &= \mathbb{E}[\mathbb{E}[Y_i(1) \mid X_i] - \mathbb{E}[Y_i(0) \mid X_i]] \\ &= \mathbb{E}\left[\frac{\mathbb{E}[T_i \mid X_i] \mathbb{E}[Y_i(1) \mid X_i]}{e(X_i)} - \frac{\mathbb{E}[1 - T_i \mid X_i] \mathbb{E}[Y_i(0) \mid X_i]}{1 - e(X_i)}\right] \text{ def. of } e(X) \\ &= \mathbb{E}\left[\frac{\mathbb{E}[T_i Y_i(1) \mid X_i]}{e(X_i)} - \frac{\mathbb{E}[(1 - T_i) Y_i(0) \mid X_i]}{1 - e(X_i)}\right] \text{ unconfoundedness} \\ &= \mathbb{E}\left[\frac{T_i Y_i}{e(X_i)} - \frac{(1 - T_i) Y_i}{1 - e(X_i)}\right]\end{aligned}$$

Inverse-propensity weighting estimator

IPW estimator

$$\hat{\tau}_{IPW} = \frac{1}{n} \sum_{i=1}^n \left(\frac{T_i Y_i}{\hat{e}(X_i)} - \frac{(1 - T_i) Y_i}{1 - \hat{e}(X_i)} \right)$$



⇒ Balance the differences between the two groups.

$\hat{\tau}_{IPW}$ unbiased and \sqrt{n} -consistent $\sqrt{n}(\hat{\tau}_{IPW} - \tau) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, V_{IPW})$

with $V_{IPW} = \mathbb{E} \left[\frac{(Y^{(0)})^2}{1 - e(X)} + \frac{(Y^{(1)})^2}{e(X)} \right] - \tau^2$ when $\hat{e}(\cdot)$ is consistent

Augmented Inverse-propensity weighting estimator

Average treatment effect (ATE): $\tau = \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$

Identifiability assumptions in observational data

- $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$ **(Unconfoundedness)**
- $\eta < e(x) < 1 - \eta, \quad \forall x \in \mathcal{X}$ and some $\eta > 0$ **(Overlap)**
- $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ **(Consistency)**

Model Treatment on Covariates $e(x) = \mathbb{P}(T_i = 1 \mid X_i = x)$

Model Outcome on Covariates $\mu_{(w)}(x) = \mathbb{E}[Y_i(w) \mid X_i = x]$

Augmented Inverse-propensity weighting estimator

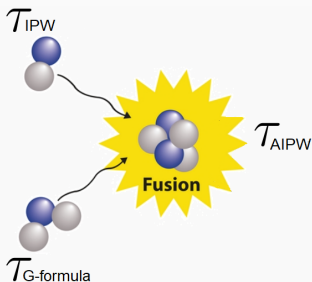
Average treatment effect (ATE): $\tau = \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$

Identifiability assumptions in observational data

- $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$ (Unconfoundedness)
- $\eta < e(x) < 1 - \eta$, $\forall x \in \mathcal{X}$ and some $\eta > 0$ (Overlap)
- $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ (Consistency)

Model Treatment on Covariates $e(x) = \mathbb{P}(T_i = 1 \mid X_i = x)$

Model Outcome on Covariates $\mu_{(w)}(x) = \mathbb{E}[Y_i(w) \mid X_i = x]$



Augmented Inverse-propensity weighting estimator

Average treatment effect (ATE): $\tau = \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$

Identifiability assumptions in observational data

- $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$ (Unconfoundedness)
- $\eta < e(x) < 1 - \eta$, $\forall x \in \mathcal{X}$ and some $\eta > 0$ (Overlap)
- $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ (Consistency)

Model Treatment on Covariates $e(x) = \mathbb{P}(T_i = 1 \mid X_i = x)$

Model Outcome on Covariates $\mu_{(w)}(x) = \mathbb{E}[Y_i(w) \mid X_i = x]$

AIPW estimator

$$\hat{\tau}_{AIPW} = \frac{1}{n} \sum_{i=1}^n \left(\mu_{(1)}(X_i) - \mu_{(0)}(X_i) + \frac{T_i \cdot (Y_i - \mu_{(1)}(X_i))}{e(X_i)} - \frac{(1 - T_i)(Y_i - \mu_{(0)}(X_i))}{1 - e(X_i)} \right)$$

$\hat{\tau}_{AIPW}$ unbiased and \sqrt{n} -consistent if either the $\hat{\mu}_{(w)}(x)$ are consistent or $\hat{e}(x)$ is consistent ² \Rightarrow **Doubly Robust** estimator

² Chernozhukov, Double/Debiased Machine Learning for Treatment and Causal Parameters, 2017

Augmented Inverse-propensity weighting estimator

Model Treatment on Covariates $e(x) = \mathbb{P}(T_i = 1 | X_i = x)$

Model Outcome on Covariates $\mu_{(w)}(x) = \mathbb{E}[Y_i(w) | X_i = x]$

AIPW estimator

$$\hat{\tau}_{AIPW} = \frac{1}{n} \sum_{i=1}^n \left(\mu_{(1)}(X_i) - \mu_{(0)}(X_i) + \frac{T_i \cdot (Y_i - \mu_{(1)}(X_i))}{e(X_i)} - \frac{(1 - T_i)(Y_i - \mu_{(0)}(X_i))}{1 - e(X_i)} \right)$$

Doubly Robust estimator ³ \Rightarrow If we have:

$$\mathbb{E} \left[(\hat{\mu}_w(X) - \mu_w(X))^2 \right] \mathbb{E} \left[(\hat{e}(X) - e(X))^2 \right] = o \left(\frac{1}{n} \right)$$

then $\hat{\tau}_{AIPW}$ is a consistent and asymptotically normal estimator of the τ :

$$\sqrt{n} (\hat{\tau}_{AIPW} - \tau) \Rightarrow \mathcal{N}(0, V^*)$$

$$V^* = \text{Var} [\tau(X_i)] + \mathbb{E} \left[\frac{\sigma_0^2(X_i)}{1 - e(X_i)} \right] + \mathbb{E} \left[\frac{\sigma_1^2(X_i)}{e(X_i)} \right]$$

³ Chernozhukov, Double/Debiased Machine Learning for Treatment and Causal Parameters, 2017

Conclusion

When measuring a causal effect, removing all confounding bias can be done two different ways:

$$\tau_{RD} = \mathbb{E} [Y^{(1)}] - \mathbb{E} [Y^{(0)}]$$

Randomized Controlled Trial (RCT)

Randomized clinical trial (RCT)



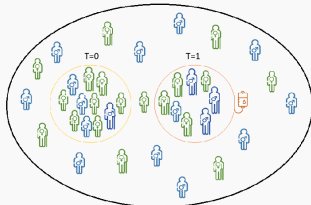
1) Random treatment assignment



$$\hat{\tau}_{DM} = \frac{1}{n_1} \sum_{i=1}^n T_i Y_i - \frac{1}{n_0} \sum_{i=1}^n (1 - T_i) Y_i$$

Observational data

No intervention : data collection



$$\tau_{IPW} \quad \tau_{AIPW} \quad \tau_G$$

- What if a covariate is missing (break the unconfoundedness assumption) ?
- Importance in variable selections (Should I add only confounding variables in the observational estimators ?)
- Possibilities to take into account the heterogeneity in the treatment effect : $\tau(x) = \mathbb{E}[\Delta_i | X_i = x] = \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x] \Rightarrow$
Personalized medicine (Causal tree, Causal Forest)

Thank you for your attention

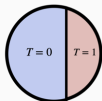
Appendix

$Y(t)$ Vs $Y|T = t$

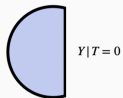
Population



Subpopulations



Conditioning



Intervening



AIPW : 2 ways

$$\begin{aligned}\hat{\tau}_{AIPW_1} &= \underbrace{\frac{1}{n} \sum_{i=1}^n \left(\frac{T_i Y_i}{\hat{e}(X_i)} - \frac{(1 - T_i) Y_i}{1 - \hat{e}(X_i)} \right)}_{\text{the IPW estimator}} \\ &+ \underbrace{\frac{1}{n} \sum_{i=1}^n \left(\hat{\mu}_{(1)}(X_i) \left(1 - \frac{T_i}{\hat{e}(X_i)} \right) - \hat{\mu}_{(0)}(X_i) \left(1 - \frac{1 - T_i}{1 - \hat{e}(X_i)} \right) \right)}_{\approx \text{mean-zero noise}}, \\ \hat{\tau}_{AIPW_2} &= \underbrace{\frac{1}{n} \sum_{i=1}^n (\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i))}_{\text{a consistent treatment effect estimator}} \\ &+ \underbrace{\frac{1}{n} \sum_{i=1}^n \left(\frac{T_i}{\hat{e}(X_i)} (Y_i - \hat{\mu}_{(1)}(X_i)) - \frac{1 - T_i}{1 - \hat{e}(X_i)} (Y_i - \hat{\mu}_{(0)}(X_i)) \right)}_{\approx \text{mean-zero noise}},\end{aligned}$$

It makes group more similar before doing the extrapolation (linear model extrapolate far away, changing a bit slope will change a lot the results (credit Susan Athey)).