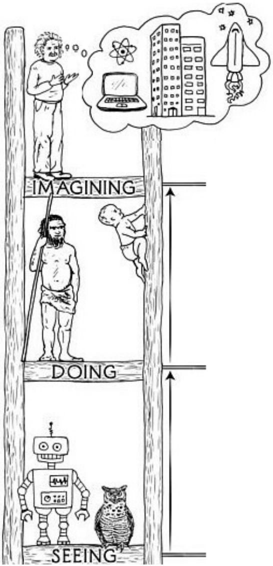
Counterfactual reasoning

Charles K. Assaad, Emilie Devijver

emilie.devijver@univ-grenoble-alpes.fr

Table of contents

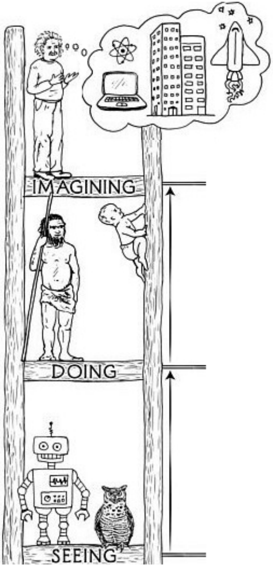
- Causality's ladder
- Counterfactuals
    - Example and definition
    - Deterministic counterfactuals
    - Probabilistic counterfactuals
    - Linear models
    - Graphical representation
- Estimation
    - Potential Outcomes Framework
    - ITE and ATE
    - First assumptions and RCTs
    - Conditional assumptions and methods
    - Hidden confounders: some solutions



Counterfactuals

Interventions

Associations



Counterfactuals
I took an aspirin, and my headache is gone: would I have had a headache had I not taken that aspirin?

Interventions
It I take an aspirin now, will I wake up with a headache?
P(*headache*|do(*aspirin*))

Associations
I took an aspirin after dinner, will I wake up with a headache?

A first example

I took an aspirin, and my headache is gone: would I have had a headache had I not taken that aspirin?

- T: observed treatment (aspirin)
- Y: observed outcome (headache)
- i: used in subscript to denote a specific individual (me)
- $Y_i(1)$: potential outcome under treatment for individual i
- $Y_i(0)$: potential outcome under no treatment for individual i

$$do(T = 1) \rightarrow Y_i(1) = 1$$
$$do(T = 0) \rightarrow Y_i(0) =?$$

A first example

I took an aspirin, and my headache is gone: would I have had a headache had I not taken that aspirin?

- T: observed treatment (aspirin)
- Y: observed outcome (headache)
- i: used in subscript to denote a specific individual (me)
- $Y_i(1)$: potential outcome under treatment for individual i
- $Y_i(0)$: potential outcome under no treatment for individual i

$$\begin{array}{rl} \text{factual} & do(T=1) \rightarrow Y_i(1) = 1 \\ \text{counterfactual} & do(T=0) \rightarrow Y_i(0) = ? \end{array}$$

Definition

$$Y(t) |\ T = t', Y = y'$$

where t is the hypothetical condition, and $T = t', Y = y'$ is the observation.

Interest in an individual level

From an experimentalist perspective, there is a profound gap between population and individual levels of analysis: the do(x)-operator captures the behavior of a population under intervention, whereas $Y_x(u)$ describes the behavior of a specific individual under such interventions.

Definition

$$Y(t) \mid T = t', Y = y'$$

where t is the hypothetical condition, and $T = t', Y = y'$ is the observation.

Interest in an individual level
From an experimentalist perspective, there is a profound gap between population and individual levels of analysis: the do(x)-operator captures the behavior of a population under intervention, whereas $Y_x(u)$ describes the behavior of a specific individual under such interventions.

Fundamental law of counterfactuals

T, Y be two variables, not necessarily connected by a single equation, described in a structural model M.
Let M_t stand for the modified version of M, with the equation of T replaced by $T = t$.
Formal definition of $Y_t(u)$: $Y_t(u) = Y_{M_t}(u)$

SUTVA: Stable Unit Treatment Value Assumption: if $T = t$, then $Y_t = Y$.

Example with binary treatment

If T is binary, then the consistency rule takes the convenient form:

$$Y = TY_1 + (1 - T)Y_0$$

For example,

- Y: being happy or unhappy (1 or 0)
- T: get a dog or don't (1 or 0)
- U: unobserved variable describing the individual (1 if dog-person or 0 if anti-dog person)

then $Y_1 = U$ and $Y_0 = 1 - U$.

Observations: $T = 0$ and $Y = 0$
$U = 1$ and $Y_u(1) = 1$

Example with binary treatment

If T is binary, then the consistency rule takes the convenient form:

$$Y = TY_1 + (1 - T)Y_0$$

For example,

- Y: being happy or unhappy (1 or 0)
- T: get a dog or don't (1 or 0)
- U: unobserved variable describing the individual (1 if dog-person or 0 if anti-dog person)

then $Y_1 = U$ and $Y_0 = 1 - U$.

Observations: $T = 0$ and $Y = 0$
$U = 1$ and $Y_u(1) = 1$

Example with binary treatment

If T is binary, then the consistency rule takes the convenient form:

$$Y = TY_1 + (1 - T)Y_0$$

For example,

- Y: being happy or unhappy (1 or 0)
- T: get a dog or don't (1 or 0)
- U: unobserved variable describing the individual (1 if dog-person or 0 if anti-dog person)

then $Y_1 = U$ and $Y_0 = 1 - U$.

Observations: $T = 0$ and $Y = 0$

$U = 1$ and $Y_u(1) = 1$

Example with binary treatment

If T is binary, then the consistency rule takes the convenient form:

$$Y = TY_1 + (1 - T)Y_0$$

For example,

- Y: being happy or unhappy (1 or 0)
- T: get a dog or don't (1 or 0)
- U: unobserved variable describing the individual (1 if dog-person or 0 if anti-dog person)

then $Y_1 = U$ and $Y_0 = 1 - U$.

Observations: $T = 0$ and $Y = 0$
$U = 1$ and $Y_u(1) = 1$

Example with binary treatment

If T is binary, then the consistency rule takes the convenient form:

$$Y = TY_1 + (1 - T)Y_0$$

For example,

- Y: being happy or unhappy (1 or 0)
- T: get a dog or don't (1 or 0)
- U: unobserved variable describing the individual (1 if dog-person or 0 if anti-dog person)

then $Y_1 = U$ and $Y_0 = 1 - U$.

Observations: $T = 0$ and $Y = 0$
$U = 1$ and $Y_u(1) = 1$

General steps for computing deterministic counterfactuals

1. **Abduction:** use the observations to determine the value of U
2. **Action:** modify the model M by removing the structural equations for the variables in T and replacing them with the appropriate functions $T = t$, to obtain the modified model M_t
3. **Prediction:** use the modified model M_t and the value of U to compute the value of $Y(t)$, the consequence of the counterfactual

Example with binary treatment, cont'd

What if we can't solve for U?

$$Y = \begin{cases} 1 & \text{if individual always happy} \\ 0 & \text{if individual never happy} \\ T & \text{if individual dog-needer} \\ 1 - T & \text{if individual dog-hater} \end{cases}$$

For example,

- Y: being happy or unhappy (1 or 0)
- T: get a dog or don't (1 or 0)
- U: unobserved variable describing the individual (1 if dog-person or 0 if anti-dog person)

Observations: $T = 1$ and $Y = 0$: $Y_u(1) = 0$. What is $Y_u(0)$? We don't know if the individual is never happy or a dog-hater.

Example with binary treatment, cont’d

What if we can’t solve for U?

$$Y = \begin{cases} 1 & \text{if individual always happy} \\ 0 & \text{if individual never happy} \\ T & \text{if individual dog-needer} \\ 1 - T & \text{if individual dog-hater} \end{cases}$$

For example,

- Y: being happy or unhappy (1 or 0)
- T: get a dog or don’t (1 or 0)
- U: unobserved variable describing the individual (1 if dog-person or 0 if anti-dog person)

Observations: $T = 1$ and $Y = 0$: $Y_u(1) = 0$. What is $Y_u(0)$? We don’t know if the individual is never happy or a dog-hater.

Example with binary treatment, cont'd

What if we can't solve for U?

$$Y = \begin{cases} 1 & \text{if individual always happy} \\ 0 & \text{if individual never happy} \\ T & \text{if individual dog-needer} \\ 1 - T & \text{if individual dog-hater} \end{cases}$$

For example,

- Y: being happy or unhappy (1 or 0)
- T: get a dog or don't (1 or 0)
- U: unobserved variable describing the individual (1 if dog-person or 0 if anti-dog person)

Observations: $T = 1$ and $Y = 0$: $Y_u(1) = 0$. What is $Y_u(0)$?
We don't know if the individual is never happy or a dog-hater.

Example with binary treatment, cont'd

What if we can't solve for U?

$$Y = \begin{cases} 1 & \text{if individual always happy} \\ 0 & \text{if individual never happy} \\ T & \text{if individual dog-needer} \\ 1 - T & \text{if individual dog-hater} \end{cases}$$

For example,

- Y: being happy or unhappy (1 or 0)
- T: get a dog or don't (1 or 0)
- U: unobserved variable describing the individual (1 if dog-person or 0 if anti-dog person)

Observations: $T = 1$ and $Y = 0$: $Y_u(1) = 0$. What is $Y_u(0)$? We don't know if the individual is never happy or a dog-hater.

Example with binary treatment, cont'd

We add a probability distribution over U:

$$P(U \text{ always happy }) = 0.3 \qquad P(U \text{ never happy }) = 0.2$$
$$P(U \text{ dog-needer}) = 0.4 \qquad P(U \text{ dog-hater}) = 0.1.$$

Because we know that $Y_u(1) = 0$, we are only interested into $P(U \text{ never happy} | T = 1, Y = 0)$ and $P(U \text{ dog-hater} | T = 1, Y = 0)$ (this individual cannot be always happy because $Y_u(1) = 0$).

$$\begin{aligned} P(U \text{ never happy} | T = 1, Y = 0) &= \frac{P(U \text{ never happy}, T = 1, Y = 0)}{P(T = 1, Y = 0)} \\ &= 0.2/(0.2 + 0.1) = 2/3 \\ P(U \text{ dog-hater} | T = 1, Y = 0) &= 0.1/(0.2 + 0.1) = 1/3 \end{aligned}$$

Then $Y_u(0) = 0$ if this individual is never happy: with proba 2/3, and $Y_u(0) = 1$ with probability 1/3.

Example with binary treatment, cont'd

We add a probability distribution over U:

$$P(U \text{ always happy }) = 0.3 \qquad P(U \text{ never happy }) = 0.2$$
$$P(U \text{ dog-needer}) = 0.4 \qquad P(U \text{ dog-hater}) = 0.1.$$

Because we know that $Y_u(1) = 0$, we are only interested into $P(U \text{ never happy}|T = 1, Y = 0)$ and $P(U \text{ dog-hater}|T = 1, Y = 0)$ (this individual cannot be always happy because $Y_u(1) = 0$).

$$\begin{aligned} P(U \text{ never happy}|T = 1, Y = 0) &= \frac{P(U \text{ never happy}, T = 1, Y = 0)}{P(T = 1, Y = 0)} \\ &= 0.2/(0.2 + 0.1) = 2/3 \\ P(U \text{ dog-hater}|T = 1, Y = 0) &= 0.1/(0.2 + 0.1) = 1/3 \end{aligned}$$

Then $Y_u(0) = 0$ if this individual is never happy: with proba 2/3, and $Y_u(0) = 1$ with probability 1/3.

Example with binary treatment, cont'd

We add a probability distribution over U:

$$P(U \text{ always happy }) = 0.3 \qquad P(U \text{ never happy }) = 0.2$$
$$P(U \text{ dog-needer}) = 0.4 \qquad P(U \text{ dog-hater}) = 0.1.$$

Because we know that $Y_u(1) = 0$, we are only interested into $P(U \text{ never happy}|T = 1, Y = 0)$ and $P(U \text{ dog-hater}|T = 1, Y = 0)$ (this individual cannot be always happy because $Y_u(1) = 0$).

$$\begin{aligned} P(U \text{ never happy}|T = 1, Y = 0) &= \frac{P(U \text{ never happy}, T = 1, Y = 0)}{P(T = 1, Y = 0)} \\ &= 0.2/(0.2 + 0.1) = 2/3 \\ P(U \text{ dog-hater}|T = 1, Y = 0) &= 0.1/(0.2 + 0.1) = 1/3 \end{aligned}$$

Then $Y_u(0) = 0$ if this individual is never happy: with proba 2/3, and $Y_u(0) = 1$ with probability 1/3.

Example with binary treatment, cont'd

We add a probability distribution over U:

$$P(U \text{ always happy }) = 0.3 \qquad P(U \text{ never happy }) = 0.2$$
$$P(U \text{ dog-needer}) = 0.4 \qquad P(U \text{ dog-hater}) = 0.1.$$

Because we know that $Y_u(1) = 0$, we are only interested into $P(U \text{ never happy}|T = 1, Y = 0)$ and $P(U \text{ dog-hater}|T = 1, Y = 0)$ (this individual cannot be always happy because $Y_u(1) = 0$).

$$\begin{aligned} P(U \text{ never happy}|T = 1, Y = 0) &= \frac{P(U \text{ never happy}, T = 1, Y = 0)}{P(T = 1, Y = 0)} \\ &= 0.2/(0.2 + 0.1) = 2/3 \\ P(U \text{ dog-hater}|T = 1, Y = 0) &= 0.1/(0.2 + 0.1) = 1/3 \end{aligned}$$

Then $Y_u(0) = 0$ if this individual is never happy: with proba 2/3, and $Y_u(0) = 1$ with probability 1/3.

General steps for computing probabilisttic counterfactuals

1. **Abduction:** use the observations to update the distribution of U
2. **Action:** modify the model M by removing the structural equations for the variables in T and replacing them with the appropriate functions $T = t$, to obtain the modified model M_t
3. **Prediction:** use the modified model M_t and the updated distribution of U to compute the value of $Y(t)$, the consequence of the counterfactual

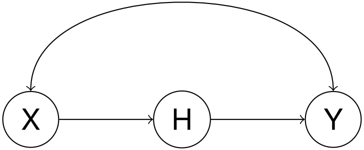
Another example: fully specified linear model M

$$
\begin{aligned}
X &= U_X \\
H &= 0.5X + U_H \\
Y &= 0.7X + 0.4H + U_Y \\
\sigma_{U_iU_j} &= 0 \text{ for all } i, j \in \{X, H, Y\}
\end{aligned}
$$

Encouragement

Homework

Exam score



Observation: a student named Joe, $X = 0.5$, $H = 1$, $Y = 1.5$

Another example: fully specified linear model M

$$
\begin{aligned}
X &= U_X && \text{Encouragement} \\
H &= 0.5X + U_H && \text{Homework} \\
Y &= 0.7X + 0.4H + U_Y && \text{Exam score} \\
\sigma_{U_i U_j} &= 0 \text{ for all } i, j \in \{X, H, Y\}
\end{aligned}
$$

Observation: a student named Joe, $X = 0.5, H = 1, Y = 1.5$

Another example: fully specified linear model M

$$
\begin{aligned}
X &= U_X && \text{Encouragement}\\
H &= 0.5X + U_H && \text{Homework}\\
Y &= 0.7X + 0.4H + U_Y && \text{Exam score}\\
\sigma_{U_i U_j} &= 0 \text{ for all } i, j \in \{X, H, Y\}
\end{aligned}
$$

Observation: a student named Joe, $X = 0.5$, $H = 1$, $Y = 1.5$
What would Joe's score have been had he doubled his study time?

$U_X = 0.5, U_H = 0.75, U_Y = 0.75$

$Y_{H=2}(U_X = 0.5, U_H = 0.75, U_Y = 0.75) = 1.90$

Another example: fully specified linear model *M*

$X = U_X$	Encouragement
$H = 0.5X + U_H$	Homework
$Y = 0.7X + 0.4H + U_Y$	Exam score
$\sigma_{U_i U_j} = 0$ for all $i, j \in \{X, H, Y\}$	

Observation: a student named Joe, $X = 0.5$, $H = 1$, $Y = 1.5$
What would Joe's score have been had he doubled his study time?

$U_X = 0.5$, $U_H = 0.75$, $U_Y = 0.75$

$Y_{H=2}(U_X = 0.5, U_H = 0.75, U_Y = 0.75) = 1.90$

Another example: fully specified linear model M

$$
\begin{aligned}
X &= U_X && \text{Encouragement} \\
H &= 0.5X + U_H && \text{Homework} \\
Y &= 0.7X + 0.4H + U_Y && \text{Exam score} \\
\sigma_{U_i U_j} &= 0 \text{ for all } i, j \in \{X, H, Y\}
\end{aligned}
$$

Observation: a student named Joe, $X = 0.5$, $H = 1$, $Y = 1.5$
What would Joe's score have been had he doubled his study time?
$U_X = 0.5, U_H = 0.75, U_Y = 0.75$
$Y_{H=2}(U_X = 0.5, U_H = 0.75, U_Y = 0.75) = 1.90$

Another example: fully specified linear model M

$$\begin{aligned} X &= U_X & &\text{Encouragement} \\ H &= 0.5X + U_H & &\text{Homework} \\ Y &= 0.7X + 0.4H + U_Y & &\text{Exam score} \\ \sigma_{U_i U_j} &= 0 \text{ for all } i, j \in \{X, H, Y\} \end{aligned}$$

Observation: a student named Joe, $X = 0.5$, $H = 1$, $Y = 1.5$
What would Joe's study time have been had he doubled his score?

Another example: fully specified linear model *M*

$$
\begin{aligned}
X &= U_X && \text{Encouragement}\\
H &= 0.5X + U_H && \text{Homework}\\
Y &= 0.7X + 0.4H + U_Y && \text{Exam score}\\
\sigma_{U_i U_j} &= 0 \text{ for all } i, j \in \{X, H, Y\}
\end{aligned}
$$

Observation: a student named Joe, $X = 0.5, H = 1, Y = 1.5$
What would Joe's study time have been had he doubled his score?
$U_X = 0.5, U_H = 0.75, U_Y = 0.75$

Another example: fully specified linear model M

$$
\begin{aligned}
X &= U_X && \text{Encouragement} \\
H &= 0.5X + U_H && \text{Homework} \\
Y &= 0.7X + 0.4H + U_Y && \text{Exam score} \\
\sigma_{U_i U_j} &= 0 \text{ for all } i, j \in \{X, H, Y\}
\end{aligned}
$$

Observation: a student named Joe, $X = 0.5$, $H = 1$, $Y = 1.5$
What would Joe's study time have been had he doubled his score?
$U_X = 0.5, U_H = 0.75, U_Y = 0.75$
$H_{Y=2}(U_X = 0.5, U_H = 0.75, U_Y = 0.75) = 1$

Another example: fully specified linear model M

$X = U_X$	Encouragement
$H = 0.5X + U_H$	Homework
$Y = 0.7X + 0.4H + U_Y$	Exam score
$\sigma_{U_i U_j} = 0 \text{ for all } i, j \in \{X, H, Y\}$	

Observation: a student named Joe, $X = 0.5$, $H = 1$, $Y = 1.5$
What would Joe's study time have been had he doubled his score?

$U_X = 0.5, U_H = 0.75, U_Y = 0.75$

$H_{Y=2}(U_X = 0.5, U_H = 0.75, U_Y = 0.75) = 1$

Counterfactual conditions are on the future, not on the past!

Another example: fully specified linear model M, cont'd

Again, in that case, some questions can't be explicitly determined.

- Suppose Joe had a scored $Y = y$ in the exam. What is the probability that Joe's score would be $Y = y'$ had he had five more hours of encouragement training?
- What would his expected score be in such hypothetical world?

We do not have information on X, H: we cannot therefore determine uniquely the value u that pertains to Joe.

Another example: fully specified linear model M, cont'd

Again, in that case, some questions can't be explicitly determined.

- Suppose Joe had a scored $Y = y$ in the exam. What is the probability that Joe's score would be $Y = y'$ had he had five more hours of encouragement training?
- What would his expected score be in such hypothetical world?

We do not have information on X, H: we cannot therefore determine uniquely the value u that pertains to Joe.

Counterfactuals in linear models

Theorem

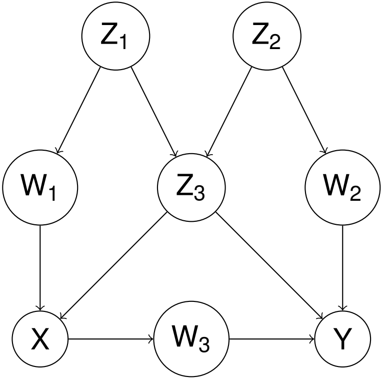
Let τ be the slope of the total effect of X on Y,

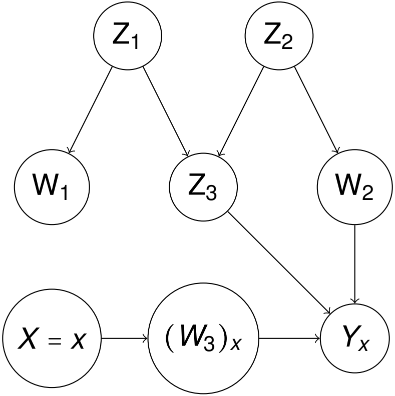
$$\tau = E(Y|do(x+1)) - E(Y|do(x))$$

then, for any observation $Z = e$, we have

$$E(Y_{X=x}|\, Z = e) = E(Y|Z = e) + \tau(x - E(X|Z = e))$$

Graphical representations of counterfactuals





Backdoor criterion

Theorem If a set Z of variables satisfies the backdoor condition relative to (X, Y), then, for all x, the counterfactual Y_x is conditionally independent of X given Z:

$$P(Y_x|X, Z) = P(Y_x|Z)$$

It helps when estimating the probabilities of counterfactuals from observational studies.

$$\begin{aligned} P(Y_x = y) &= \sum_z P(Y_x = y|Z = z)P(Z = z) \\ &= \sum_z P(Y_x = y|Z = z, X = x)P(Z = z) \\ &= \sum_z P(Y = y|Z = z, X = x)P(Z = z). \end{aligned}$$

Backdoor criterion

Theorem If a set Z of variables satisfies the backdoor condition relative to (X, Y), then, for all x, the counterfactual Y_x is conditionally independent of X given Z:

$$P(Y_x|X,Z) = P(Y_x|Z)$$

It helps when estimating the probabilities of counterfactuals from observational studies.

$$\begin{aligned} P(Y_x = y) &= \sum_z P(Y_x = y|Z = z)P(Z = z) \\ &= \sum_z P(Y_x = y|Z = z, X = x)P(Z = z) \\ &= \sum_z P(Y = y|Z = z, X = x)P(Z = z). \end{aligned}$$

Actual causation / potential causation: a motivation for counterfactual world

We assume we know the causal graph $\mathcal{G}$.

We don't know which one, among all the potential causes, are the actual causes for a specific individual.

- Example for smoking / cancer
- Example of recommander system

Actual causation / potential causation: a motivation for counterfactual world

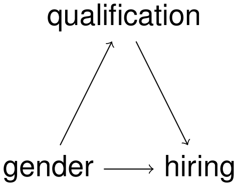
We assume we know the causal graph $\mathcal{G}$.

We don't know which one, among all the potential causes, are the actual causes for a specific individual.

- Example for smoking / cancer
- Example of recommander system

Mediation

We want to know whether and to what degree a company discriminates by gender in its hiring practices. However, gender also affects hiring practices in other ways, often, for instance, women are more or less likely to go into a particular field than men, or to have achieved advanced degrees in that field: mediating variable of qualifications.



qualification

gender

hiring

Mediation

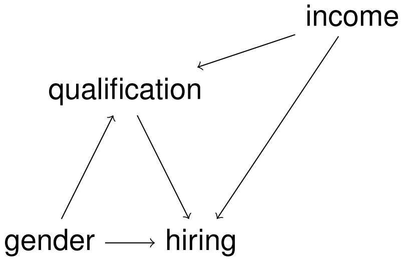
To find the direct effect of gender on hiring, we need to somehow hold qualifications steady and measure the remaining relationship between ender and hiring:

$$P(\text{hired}|\text{Female, highly qualified})$$
$$P(\text{hired}|\text{Male, highly qualified})$$

This is true if the previous graph is true.

Mediation

What if there are confounders of the mediating variable and the outcome variable?



→ needs of intervention!

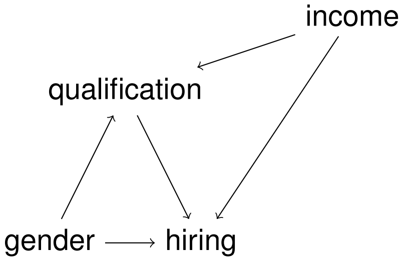
Controlled direct effect:

$$P(Y = y|do(X = x, Z = z)) - P(Y = y|do(X = x', Z = z))$$

do-calculus allows to compute this (using backdoor criterion)

Mediation

What if there are confounders of the mediating variable and the outcome variable?



income

qualification

gender

hiring

→ needs of intervention!

Controlled direct effect:

$$P(Y = y|do(X = x, Z = z)) - P(Y = y|do(X = x', Z = z))$$

do-calculus allows to compute this (using backdoor criterion)

Mediation

A policy maker wishes to access the extent to which gender disparity in hiring can be reduced by making hiring decisions gender-blind, rather than eliminating gender inequality in education or job training.

→ direct effect or indirect effet (mediated via job qualification)?

- The hiring status (Y) of a female applicant with qualification $Q = q$ given that the employer treats her as though she is a male is captured by the counterfactual $Y_{X=1,Q=q}$ (where $X = 1$ refers to being a male)
- Since the value q would vary among applicants, we need to average:

$$\sum_q E(Y_{X=1,Q=q})P(Q = q|X = 0)$$

- Male applicants:

$$\sum_q E(Y_{X=1,Q=q})P(Q = q|X = 1)$$

Mediation

$$\sum_q E(Y_{X=1,Q=q})(P(Q=q|X=0)-P(Q=q|X=1))$$

is the indirect effect of gender on hiring, mediated by qualification: **natural indirect effect**, because we allow the qualification Q to vary naturally from applicant to applicant, as opposed to the controlled direct effect

Mediation

$$\begin{aligned} t &= f_T(u_T) \\ m &= f_M(t, u_M) \\ y &= f_Y(t, m, u_Y) \end{aligned}$$

where T, M and Y are discrete or continuous random variables, f_T, f_M, f_Y are arbitrary functions, and U_T, U_M, U_Y are omitted factors that influence T, M and Y.

- Total effect: $E(Y|do(T=1)) - E(Y|do(T=0))$
- Controlled direct effect: $E(Y|do(T=1, M=m)) - E(Y|do(T=0, M=m))$
- Natural direct effect: $E(Y_{1,M_0} - Y_{0,M_0})$
- Natural indirect effect: $E(Y_{1,M_1} - Y_{1,M_0})$

Mediation

$$
\begin{aligned}
t &= f_T(u_T) \\
m &= f_M(t, u_M) \\
y &= f_Y(t, m, u_Y)
\end{aligned}
$$

where T, M and Y are discrete or continuous random variables, f_T, f_M, f_Y are arbitrary functions, and U_T, U_M, U_Y are omitted factors that influence T, M and Y.

- Total effect: $E(Y|do(T = 1)) - E(Y|do(T = 0))$
- Controlled direct effect: $E(Y|do(T = 1, M = m)) - E(Y|do(T = 0, M = m))$
- Natural direct effect: $E(Y_{1,M_0} - Y_{0,M_0})$
- Natural indirect effect: $E(Y_{1,M_1} - Y_{1,M_0})$

Mediation

$$
\begin{aligned}
t &= f_T(u_T) \\
m &= f_M(t, u_M) \\
y &= f_Y(t, m, u_Y)
\end{aligned}
$$

where T, M and Y are discrete or continuous random variables, f_T, f_M, f_Y are arbitrary functions, and U_T, U_M, U_Y are omitted factors that influence T, M and Y.

- Total effect: $E(Y|do(T = 1)) - E(Y|do(T = 0))$ do calculus
- Controlled direct effect: $E(Y|do(T = 1, M = m)) - E(Y|do(T = 0, M = m))$ do calculus
- Natural direct effect: $E(Y_{1,M_0} - Y_{0,M_0})$ counterfactual
- Natural indirect effect: $E(Y_{1,M_1} - Y_{1,M_0})$counterfactual

Mediation

We need conditions to have identification of natural effect (direct or indirect)!

Exercise Consider the structural model:

$$\begin{aligned} y &= \beta_1 m + \beta_2 t + u_y \\ m &= \gamma_1 t + u_m \end{aligned}$$

Determine the total effect, the natural direct effect and the natural indirect effect.

Identifiability in counterfactuals

Same work done for interventions can be done for counterfactuals

ID* algorithm [1]

[1] Shpitser and Pearl, 2008, JMLR

Potential Outcomes (PO)

- Set of n units indexed by i (individuals)
- T_i be the value of a treatment assigned to individual i

Definition (Potential outcomes)

The potential outcome under treatment level t, denoted by $Y_i(t)$, is the value that the outcome would have taken were T_i set to t, possibly contrary to the fact.

- For binary T_i, $Y_i(0)$ is the potential outcome if the unit i does not receive the treatment (control), and $Y_i(1)$ is the potential outcome if the unit i does receive the treatment (treated).
  - Treatment Group: The group of subjects or units that receive the specific intervention or treatment being studied.
  - Control Group: The group of subjects or units that do not receive the treatment, serving as a comparison to assess the treatment's effectiveness.

Potential Outcomes (PO)

- Set of n units indexed by i (individuals)
- T_i be the value of a treatment assigned to individual i

Definition (Potential outcomes)

The potential outcome under treatment level t, denoted by $Y_i(t)$, is the value that the outcome would have taken were T_i set to t, possibly contrary to the fact.

- For binary T_i, $Y_i(0)$ is the potential outcome if the unit i does not receive the treatment (control), and $Y_i(1)$ is the potential outcome if the unit i does receive the treatment (treated).
  - Treatment Group: The group of subjects or units that receive the specific intervention or treatment being studied.
  - Control Group: The group of subjects or units that do not receive the treatment, serving as a comparison to assess the treatment's effectiveness.

Potential Outcomes (Contd.)

Key Points about Potential Outcomes:

- Fundamental for understanding causal effects and comparing the effects of interventions.
- Crucial for estimating the impact of treatments or interventions in causal inference studies.
- Potential outcomes enable us to translate causal questions into the estimation of a causal estimand.

History of the concept

- Started from Neyman (1923) and Fisher's (1935) work on understanding experiments
- Formalized by Rubin in a series of papers (from 1974)
- Potential outcomes has evolved into an entire framework for causal inquiry.

Can be seen as an alternative way to express counterfactuals by the do operator (Pearl, 2000).

Potential Outcomes (Contd.)

Key Points about Potential Outcomes:

- Fundamental for understanding causal effects and comparing the effects of interventions.
- Crucial for estimating the impact of treatments or interventions in causal inference studies.
- Potential outcomes enable us to translate causal questions into the estimation of a causal estimand.

History of the concept

- Started from Neyman (1923) and Fisher's (1935) work on understanding experiments
- Formalized by Rubin in a series of papers (from 1974)
- Potential outcomes has evolved into an entire framework for causal inquiry.

Can be seen as an alternative way to express counterfactuals by the do operator (Pearl, 2000).

Example of Treatment and Control Groups

Illustrative Example:

- Consider a clinical trial evaluating the effectiveness of a new drug for a specific medical condition.
- The patients receiving the actual drug constitute the treatment group, while those receiving a placebo or standard treatment form the control group.
- By comparing the outcomes between the two groups, researchers can assess the causal impact of the new drug on the patients' health outcomes.

Counterfactuals for Estimating Causal Effects

Using Counterfactuals to Estimate Causal Effects:

- Counterfactuals provide a hypothetical comparison of what would have happened under different treatment conditions.
- Used to estimate the causal effect of an intervention by comparing the observed outcome with the hypothetical outcome that would have occurred without the intervention.

Application of Counterfactuals in Causal Inference:

- Essential for determining the causal impact of treatments, policies, or interventions in observational and experimental studies.
- Enable researchers to evaluate the effectiveness of interventions by comparing the actual outcomes with the hypothetical outcomes in the absence of the intervention.

Counterfactuals for Estimating Causal Effects

Using Counterfactuals to Estimate Causal Effects:

- Counterfactuals provide a hypothetical comparison of what would have happened under different treatment conditions.
- Used to estimate the causal effect of an intervention by comparing the observed outcome with the hypothetical outcome that would have occurred without the intervention.

Application of Counterfactuals in Causal Inference:

- Essential for determining the causal impact of treatments, policies, or interventions in observational and experimental studies.
- Enable researchers to evaluate the effectiveness of interventions by comparing the actual outcomes with the hypothetical outcomes in the absence of the intervention.

Illustrative Example: Counterfactuals in a Study

Example Scenario:

- Consider a study evaluating the impact of a new teaching method on student performance in a particular subject.
- The counterfactual comparison involves assessing the performance of students who received the new teaching method with the hypothetical performance they would have had if they had not received the new method.
- By comparing the actual performance with the hypothetical performance, researchers can estimate the causal effect of the new teaching method on student achievement.

Individual Treatment Effect (ITE)

Represents the causal effect of a treatment or intervention on an individual unit within a study.

Definition (ITE)

For each individual i,

$$ITE_i = Y_{1i} - Y_{0i}$$

where:

- ITE_i is the Individual Treatment Effect for the ith unit,
- Y_{1i} is the PO for the ith unit under the treatment,
- Y_{0i} is the PO for the ith unit under the control.

We can consider other quantities (ratio, percentage increase...) but always some contrast measure between two POs.

Individual Treatment Effect (ITE)

Represents the causal effect of a treatment or intervention on an individual unit within a study.

Definition (ITE)

For each individual i,

$$ITE_i = Y_{1i} - Y_{0i}$$

where:

- ITE_i is the Individual Treatment Effect for the ith unit,
- Y_{1i} is the PO for the ith unit under the treatment,
- Y_{0i} is the PO for the ith unit under the control.

We can consider other quantities (ratio, percentage increase...) but always some contrast measure between two POs.

The fundamental Problem of Causal Inference

- Can we do something to estimate the ITE?
- **The fundamental Problem of Causal Inference** (Holland, 1986)
  It is impossible to observe the value of $Y_i(1)$ and $Y_i(0)$ for the same unit, therefore it is impossible to observe the ITE.

Average Treatment Effect (ATE)

- Represents the average causal effect of a treatment or intervention on the outcome variable within a population.
- Provides an overall assessment of the treatment's impact on the entire population under study.

Definition (ATE)

$$ATE = E[Y_1 - Y_0]$$

where:

- Y_1 is the potential outcome under the treatment,
- Y_0 is the potential outcome under the control,
- $E[\cdot]$ denotes the expectation or average over the entire population.

Differences between ATE and ITE

Distinguishing ATE and ITE:

- ATE provides the average treatment effect for the entire study population, while ITE focuses on the specific effects for individual units.
- ATE assesses the overall impact of a treatment at a population level, while ITE emphasizes individual-level variations in treatment effects.
- ATE is used for evaluating the general effectiveness of interventions, whereas ITE is crucial for understanding personalized treatment effects.

Example of Average Treatment Effect (ATE)

Real-World Scenario:

- A study assessing the impact of a new educational program on student performance.
- ATE is calculated by comparing the average test scores of students who participated in the program with those who did not. We need assumptions here!
- The difference in average scores provides an estimate of the average effect of the educational program on the overall student population.

Example of Individual Treatment Effect (ITE)

Real-World Scenario:

- A clinical trial investigating the efficacy of a new drug for a specific medical condition in a diverse patient population.
- ITE is computed by analyzing the individual response to the drug compared to the response they would have had without the treatment.
- The variation in treatment effects among different patient subgroups helps in identifying specific patient characteristics that influence the drug's effectiveness.

Real-World Examples of Treatment Effects (Contd.)

Key Takeaways from Real-World Examples:

- ATE provides insights into the overall impact of interventions on a study population, guiding policy and program decisions.
- ITE helps in understanding the heterogeneous responses to treatments among individuals, enabling personalized treatment strategies and interventions.

How to estimate ATE?

- **Hypothetical world**: we observe every potential outcome for every individual.
- **In reality**: we observe one (at most) for each individual.

Can we average the observations from control and treatment?
Yes but under very stringent assumptions!

How to estimate ATE?

- **Hypothetical world**: we observe every potential outcome for every individual.
- **In reality**: we observe one (at most) for each individual.

Can we average the observations from control and treatment?
Yes but under very stringent assumptions!

Assumptions

SUTVA

Definition (SUTVA: Stable Unit Treatment Value Assumption)

Observed outcome = potential outcome of the observed treatment: $Y_i(t) = Y_i$ if $T_i = t$.

For a binary treatment, this writes

$$Y_i = Y_i(1) \times T_i + Y_i(0) \times (1 - T_i)$$

- No interference: manipulating another unit's tratment does not affect a unit's PO
- Consistency: for each unit, no different form or version of each treatment level, which lead to different PO

Assumptions
SUTVA

Definition (SUTVA: Stable Unit Treatment Value Assumption)

Observed outcome = potential outcome of the observed treatment: $Y_i(t) = Y_i$ if $T_i = t$.

For a binary treatment, this writes

$$Y_i = Y_i(1) \times T_i + Y_i(0) \times (1 - T_i)$$

- No interference: manipulating another unit's tratment does not affect a unit's PO
- Consistency: for each unit, no different form or version of each treatment level, which lead to different PO

Assumptions

Positivity and ignorability

Definition

Positivity: we assume that, for all units *i* and treatment levels *t*,

$$P(T_i = t) > 0$$

Ignorability: we assume that, for all treatment levels, *t*,

$$Y_i(t) \perp\!\!\!\perp T_i$$

This means that the average outcome in the treated group is representative of what we would see on average if everyone got treated (same for the controls).

$$E(Y_i(1)) = E(Y_i(1)|T_i = 1) = E(Y_i(1)|T_i = 0)$$

First estimator of ATE

Under SUTVA, positivity, ignorability,

$$\begin{aligned} E(Y_i|T_i = 1) - E(Y_i|T_i = 0) &= E(Y_i(1)|\ T_i = 1) - E(Y_i(0)|T_i = 0) \\ &= E(Y_i(1)) - E(Y_i(0)) \\ &= ATE \end{aligned}$$

Sample means estimator

$$\widehat{\mathrm{ATE}} = \frac{1}{n_t}\sum_{i=1}^{n} Y_i T_i - \frac{1}{n_c}\sum_{i=1}^{n} Y_i(1 - T_i)$$

Estimator unbiased, consistent and asymptotically gaussian.

First estimator of ATE

Under SUTVA, positivity, ignorability,

$$\begin{aligned} E(Y_i|T_i=1) - E(Y_i|T_i=0) &= E(Y_i(1)|\ T_i=1) - E(Y_i(0)|T_i=0) \\ &= E(Y_i(1)) - E(Y_i(0)) \\ &= ATE \end{aligned}$$

Sample means estimator

$$\widehat{\mathrm{ATE}} = \frac{1}{n_t}\sum_{i=1}^{n} Y_i T_i - \frac{1}{n_c}\sum_{i=1}^{n} Y_i(1-T_i)$$

Estimator unbiased, consistent and asymptotically gaussian.

Randomized Controlled Trials (RCTs)

The easiest way to collect data that satisfies those assumptions is to perform a randomized experiment.

Definition (Randomized Experiment)

An experiment is a study in which the probability of treatment assignment $P(T_i = t)$ is directly under the control of a researcher.

Definition and Application:

- RCTs are experimental studies where participants are randomly assigned to either the treatment or control group.
- They are considered the gold standard for estimating causal effects as randomization helps control for both observed and unobserved confounding variables.

Challenges: non-compliance, arm switches, ...

Randomized Controlled Trials (RCTs)

Example:

- Clinical trials for testing the efficacy of a new drug in a controlled setting.
- Analyzing the impact of a policy change on employment using survey data and statistical controls.

Drawbacks: not always possible (unethical, need for a large sample size, bias in population selection, lack of follow-up...)

Challenges and Solutions:

- Observational studies use data from naturally occurring settings and are prone to confounding and bias.
- Techniques such as multivariate regression, stratification, and sensitivity analysis help control for confounding factors and improve causal inference from observational data.

Randomized Controlled Trials (RCTs)

Example:

- Clinical trials for testing the efficacy of a new drug in a controlled setting.
- Analyzing the impact of a policy change on employment using survey data and statistical controls.

Drawbacks: not always possible (unethical, need for a large sample size, bias in population selection, lack of follow-up...)

Challenges and Solutions:

- Observational studies use data from naturally occurring settings and are prone to confounding and bias.
- Techniques such as multivariate regression, stratification, and sensitivity analysis help control for confounding factors and improve causal inference from observational data.

Randomized Controlled Trials (RCTs)

Example:

- Clinical trials for testing the efficacy of a new drug in a controlled setting.
- Analyzing the impact of a policy change on employment using survey data and statistical controls.

Drawbacks: not always possible (unethical, need for a large sample size, bias in population selection, lack of follow-up...)

Challenges and Solutions:

- Observational studies use data from naturally occurring settings and are prone to confounding and bias.
- Techniques such as multivariate regression, stratification, and sensitivity analysis help control for confounding factors and improve causal inference from observational data.

A tool: the propensity score

Definition (Propensity score)

The propensity score is the probability of receiving the treatment:

$$e(x) = P(T = 1 | X = x)$$

Estimating the Propensity Score Logistic Regression Formula:

$$\text{logit}(e) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Where:

- e represents the estimated propensity score,
- $X_1, X_2, \ldots, X_p$ represent the observed covariates or confounding variables,
- $\beta_0, \beta_1, \ldots, \beta_p$ are the coefficients of the logistic regression model.

More recently, many more ML methods: boosting, NN, RF ...

Another set of assumptions

Definition

Conditional positivity: we assume that, for all units *i* and treatment levels *t*,

$$P(T_i = t | X_i = x) > 0$$

for all *x* in the domain.
Conditional ignorability: we assume that $Y_i(1), Y_i(0) \perp\!\!\!\perp T_i | X_i = x$ for all *x* and *t*.

- Knowing a unit's covariate values will never determine what treatment that unit gets with certainty
- The covariate tell the whole story of the treatment assignment process, and within levels of X_i, treatment is assigned as-if-random.

How to pick the good set of covariates?

Another set of assumptions

Definition

Conditional positivity: we assume that, for all units *i* and treatment levels *t*,

$$P(T_i = t | X_i = x) > 0$$

for all *x* in the domain.

Conditional ignorability: we assume that $Y_i(1), Y_i(0) \perp\!\!\!\perp T_i | X_i = x$ for all *x* and *t*.

- Knowing a unit's covariate values will never determine what treatment that unit gets with certainty
- The covariate tell the whole story of the treatment assignment process, and within levels of X_i, treatment is assigned as-if-random.

How to pick the good set of covariates?

Another set of assumptions

Definition

Conditional positivity: we assume that, for all units i and treatment levels t,

$$P(T_i = t | X_i = x) > 0$$

for all x in the domain.
Conditional ignorability: we assume that $Y_i(1), Y_i(0) \perp\!\!\!\perp T_i | X_i = x$ for all x and t.

- Knowing a unit's covariate values will never determine what treatment that unit gets with certainty
- The covariate tell the whole story of the treatment assignment process, and within levels of X_i, treatment is assigned as-if-random.

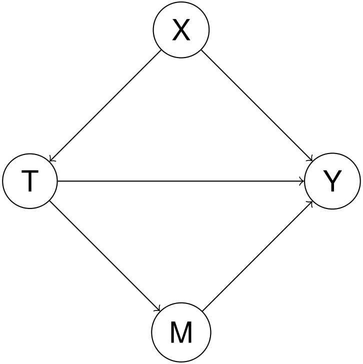
How to pick the good set of covariates?

How to pick the good set of covariates?

It comes back to the causal graph!

We look for variables that

- block all non-causal paths from T to Y
- let all causal paths from T to Y open.



Setting on observational dataset

- Dataset $(T_i, Y_i, X_i)_{1 \leq i \leq n}$
- Assumptions: SUTVA, conditional ignorability wrt X, conditional positivity wrt X
- Methods:
  - Post-stratification
  - IPTW
  - Matching
  - Regression-based (S-learner, T-learner)
  - Double Robustness
  - Double Machine Learning
  - Causal Forest

Post-stratification

- Within strate, we can identify the ATE by the difference-in-means
- $\mathrm{CATE}(x) = E(Y(1) - Y(0)|X = x)$
- Come back to ATE:

$$\begin{aligned} & E(Y(1) - Y(0)) \\ = & E(E(Y(1) - Y(0)|X)) \\ = & \sum_x (E(Y|T = 1, X = x) - E(Y|T = 0, X = x))P(X = x) \end{aligned}$$

estimated by

$$\widehat{\mathrm{ATE}} = \sum_x \widehat{\mathrm{CATE}}(x)\frac{n_x}{n}$$

Drawbacks:

- If too many strates, too few units in each strate
- continuous covariates...

Post-stratification

- Within strate, we can identify the ATE by the difference-in-means
- $\text{CATE}(x) = E(Y(1) - Y(0)|X = x)$
- Come back to ATE:

$$\begin{aligned} & E(Y(1) - Y(0)) \\ = & E(E(Y(1) - Y(0)|X)) \\ = & \sum_x (E(Y|T = 1, X = x) - E(Y|T = 0, X = x))P(X = x) \end{aligned}$$

estimated by

$$\widehat{\text{ATE}} = \sum_x \widehat{\text{CATE}}(x) \frac{n_x}{n}$$

Drawbacks:

- If too many strates, too few units in each strate
- continuous covariates...

Inverse Probability of Treatment Weighting (IPTW)

Fact: in a randomized experiment, covariate distribution are balanced across treatment groups, but not in observational studies.

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} w_i(Y_i T_i - Y_i(1 - T_i))$$

Which weight?propensity score!

IPTW (or Horowitz-Thompson estimator): weighted estimator with $w_i = (e(X_i))^{-1}$ for treated units and $w_i = (1 - e(X_i))^{-1}$ for control units

Inverse Probability of Treatment Weighting (IPTW)

Fact: in a randomized experiment, covariate distribution are balanced across treatment groups, but not in observational studies.

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} w_i(Y_i T_i - Y_i(1 - T_i))$$

Which weight?propensity score!

IPTW (or Horowitz-Thompson estimator): weighted estimator with $w_i = (e(X_i))^{-1}$ for treated units and $w_i = (1 - e(X_i))^{-1}$ for control units

Matching

Definition and Application: how to fill in the missing outcome for each unit?

- Propensity score matching is used to estimate the causal effect of a treatment or intervention by balancing the distribution of observed covariates between the treatment and control groups.
- It involves matching treated and untreated units

Definition (Matching)

For each unit *i*, find the unit *j* with opposite treatment and most similar covariate values and use their outcome as the missing one for *i*.

Which similarity? Euclidian, propensity score-based, ...

Example in Social Sciences

Research Question: Does participation in a mentoring program improve academic performance in at-risk students?
Propensity Score Matching Process:

1. Collect demographic data, socioeconomic background, and previous academic performance of at-risk students.
2. Estimate the propensity scores using logistic regression, considering relevant covariates.
3. Match treated students who participated in the mentoring program with similar untreated students who did not participate, based on their propensity scores.
4. Compare the academic performance of the matched groups to evaluate the impact of the mentoring program on the students' academic outcomes.

Regression-based methods

We blocked the non-causal open path from T to Y by adjusting on X.
How to model the relationship between $Y_i(t)$ and X_i?
Regression model:

$$E(Y_i|T_i, X_i) = \alpha + \beta T_i + X_i \gamma$$

Do $\hat{\beta}$ is an estimate of the treatment effect?
Only if (the model is correctly specified and) the treatment effect is constant across units → **homogeneous treatment effect**

Heterogeneous treatment effect

$$E(Y_i|T_i, X_i) = \alpha + \beta T_i + X_i \gamma + T_i X_i \lambda$$

Then ATE $= \beta + E(X_i)\lambda$.

Regression-based methods

We blocked the non-causal open path from T to Y by adjusting on X.
How to model the relationship between $Y_i(t)$ and X_i?
Regression model:

$$E(Y_i|T_i, X_i) = \alpha + \beta T_i + X_i\gamma$$

Do $\hat{\beta}$ is an estimate of the treatment effect?
Only if (the model is correctly specified and) the treatment effect is constant across units → **homogeneous treatment effect**

Heterogeneous treatment effect

$$E(Y_i|T_i, X_i) = \alpha + \beta T_i + X_i\gamma + T_iX_i\lambda$$

Then ATE $= \beta + E(X_i)\lambda$.

Regression-based methods

We blocked the non-causal open path from T to Y by adjusting on X.
How to model the relationship between $Y_i(t)$ and X_i?
Regression model:

$$E(Y_i|T_i, X_i) = \alpha + \beta T_i + X_i\gamma$$

Do $\hat{\beta}$ is an estimate of the treatment effect?
Only if (the model is correctly specified and) the treatment effect is constant across units → **homogeneous treatment effect**

Heterogeneous treatment effect

$$E(Y_i|T_i, X_i) = \alpha + \beta T_i + X_i\gamma + T_i X_i \lambda$$

Then ATE $= \beta + E(X_i)\lambda$.

Regression-based methods

We blocked the non-causal open path from T to Y by adjusting on X.
How to model the relationship between $Y_i(t)$ and X_i?
Regression model:

$$E(Y_i|T_i, X_i) = \alpha + \beta T_i + X_i\gamma$$

Do $\hat{\beta}$ is an estimate of the treatment effect?
Only if (the model is correctly specified and) the treatment effect is constant across units → **homogeneous treatment effect**

Heterogeneous treatment effect

$$E(Y_i|T_i, X_i) = \alpha + \beta T_i + X_i\gamma + T_iX_i\lambda$$

Then ATE $= \beta + E(X_i)\lambda$.

Regression-based methods

We blocked the non-causal open path from T to Y by adjusting on X.

How to model the relationship between $Y_i(t)$ and X_i?

Regression model:

$$E(Y_i|T_i, X_i) = \alpha + \beta T_i + X_i\gamma$$

Do $\hat{\beta}$ is an estimate of the treatment effect?

Only if (the model is correctly specified and) the treatment effect is constant across units → **homogeneous treatment effect**

Heterogeneous treatment effect

$$E(Y_i|T_i, X_i) = \alpha + \beta T_i + X_i\gamma + T_iX_i\lambda$$

Then ATE $= \beta + E(X_i)\lambda$.

Regression-based methods

$$\text{ATE} = \beta + E(X_i)\lambda$$

One would like $E(X_i) = 0$: de-meaning covariates
Then in very specific cases with very strong assumptions, you can express the ATE with the regression coefficients!

Other solution:

S-learner	$\mu(t, x) = E(Y \mid T = t, X = x)$
T-learner	$\mu(1, x) = E(Y \mid T = 1, X = x)$
	$\mu(0, x) = E(Y \mid T = 0, X = x)$

Regression-based methods

$$\text{ATE} = \beta + E(X_i)\lambda$$

One would like $E(X_i) = 0$: de-meaning covariates
Then in very specific cases with very strong assumptions, you can express the ATE with the regression coefficients!

Other solution:

$$\begin{aligned} &\text{S-learner} && \mu(t,x) = E(Y|T=t, X=x) \\ &\text{T-learner} && \mu(1,x) = E(Y|T=1, X=x) \\ & && \mu(0,x) = E(Y|T=0, X=x) \end{aligned}$$

Double robustness

- Most estimator (IPTW, S-learner, T-learner) are sensitive to model misspecification
- Doubly robust estimator: combine them! For example, augmented IPW:

$$\widehat{\mathrm{ATE}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) \\ + T_i \frac{Y_i - \hat{\mu}_{(1)}(X_i)}{\hat{e}(X_i)} - (1 - T_i)\frac{Y_i - \hat{\mu}_{(0)}(X_i)}{1 - \hat{e}(X_i)}$$

- Even with double robustness, needs of correct specification

Double robustness

- Most estimator (IPTW, S-learner, T-learner) are sensitive to model misspecification
- Doubly robust estimator: combine them! For example, augmented IPW:

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) \\ + T_i \frac{Y_i - \hat{\mu}_{(1)}(X_i)}{\hat{e}(X_i)} - (1 - T_i)\frac{Y_i - \hat{\mu}_{(0)}(X_i)}{1 - \hat{e}(X_i)}$$

- Even with double robustness, needs of correct specification

Double Machine Learning

$$Y = g_0(T, X) + U$$
$$ATE = E(g_0(1, X) - g_0(0, X))$$

g_0 is learnt by ML

- flexibility,
- heterogenous treatment effects,
- high dimensional X

1st method

1. $\hat{g}_0$ using ML
2. plug in predictions to estimate ATE

Caution! ML methods address the variance-bias trade-off, which leads to a bias of the causal estimate

Double Machine Learning

$$Y = g_0(T, X) + U$$

$$ATE = E(g_0(1, X) - g_0(0, X))$$

g_0 is learnt by ML

- flexibility,
- heterogenous treatment effects,
- high dimensional X

1st method

1. $\hat{g}_0$ using ML
2. plug in predictions to estimate ATE

Caution! ML methods address the variance-bias trade-off, which leads to a bias of the causal estimate

Double Machine Learning

Definition (Neyman Orthogonality)

The error terms that arise due to regularization do not affect the causal estimate. For ψ a score function, $\mathcal{D}$ a dataset, η the nuisance part,

$$E(\psi(\mathcal{D}; \text{ATE}, \eta)) = 0.$$

Sample splitting: split the sample into two parts: one for the ML estimation, one for the causal estimation ATE.

Theorem

Central limit theorem for the double ML estimator under regularity conditions with known covariance matrix.

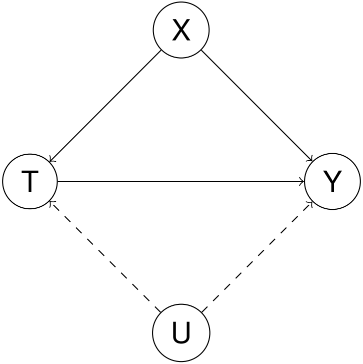
Causal forest

- A causal tree is constructing leaves such that the individuals *i* come from a randomized experiment. Then, sum over the leaves.
- Causal forest: ensemble of causal trees
- This is a consistent estimator of CATE
- Variable importance deduced from causal RFs

Unmeasured confounding

If we suspect some unmeasured confounding,

- Conditional ignorability does not hold with respect to X
- We assume that there is an unmeasured variable U such that conditional ignorability hold with respect to X and U



Instrumental Variable Analysis

Definition and Application:

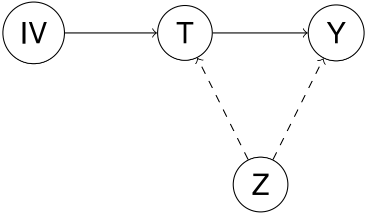
- Instrumental variables IV: correlated with the treatment but not directly associated with the outcome, allowing for the estimation of causal effects in the presence of unobserved biases.
- Key aspect of an IV: should be strongly correlated with the endogenous variable (X) but uncorrelated with the error term and any unobserved confounders (Z). This helps address potential issues of endogeneity and omitted variable bias, improving the validity of your causal inference.
- In other words: we want to split the variation in X_i that is uncorrelated with the noise, to estimate the causal effect.

Instrumental Variable Analysis

Definition (Instrumental Variable)

An instrumental variable IV must satisfy three conditions:

1. **Relevance** IV has a causal effect on T
2. **Exclusion restriction** the causal effect of IV on Y is fully mediated by Y
3. **Instrumental unconfoundedness** The relationship between IV and Y is unconfounded of confounded only by variables we measure and can adjust on.



Instrumental Variable Analysis

Estimation through IV

Under the linear model,

$$\begin{aligned} Z &= \varepsilon_Z \\ IV &= \varepsilon_{IV} \\ T &= \beta_{z,t} Z + \beta_{iv,t} IV + \varepsilon_T \\ Y &= \beta_{z,y} Z + \beta_{t,y} T + \varepsilon_Y. \end{aligned}$$

Two-stage least squares estimator (2SLS estimator)

Regressing T on IV, we get $\hat{T}$ as a function of only IV.

Then, regressing Y on $\hat{T}$ we get an estimator for $\beta_{t,y}$.

Caution! this can have large variance if the value $\beta_{IV,T}$ is near zero. In such cases, IV is called a weak instrument.

Instrumental Variable Analysis

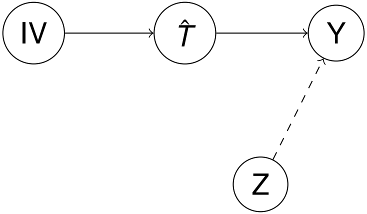
Estimation through IV

Under the linear model,

$$\begin{aligned} Z &= \varepsilon_Z \\ IV &= \varepsilon_{IV} \\ T &= \beta_{z,t} Z + \beta_{iv,t} IV + \varepsilon_T \\ Y &= \beta_{z,y} Z + \beta_{t,y} T + \varepsilon_Y . \end{aligned}$$

Two-stage least squares estimator (2SLS estimator)

Regressing T on IV, we get $\hat{T}$ as a function of only IV.



Then, regressing Y on $\hat{T}$ we get an estimator for $\beta_{t,y}$.

Instrumental Variable Analysis

Estimation through IV
Under the linear model,

$$\begin{aligned} Z &= \varepsilon_Z \\ IV &= \varepsilon_{IV} \\ T &= \beta_{z,t} Z + \beta_{iv,t} IV + \varepsilon_T \\ Y &= \beta_{z,y} Z + \beta_{t,y} T + \varepsilon_Y. \end{aligned}$$

Two-stage least squares estimator (2SLS estimator)
Regressing T on IV, we get $\hat{T}$ as a function of only IV.
Then, regressing Y on $\hat{T}$ we get an estimator for $\beta_{t,y}$.
Caution! this can have large variance if the value $\beta_{IV,T}$ is near zero. In such cases, IV is called a weak instrument.

Example in Social Sciences

Research Question: Does increased spending on education lead to improved long-term economic outcomes for individuals?
Instrumental Variable Analysis Process:

1. Identify an instrumental variable, such as a policy change affecting education spending at the regional level.
2. Verify that the instrumental variable is correlated with education spending but not directly associated with individual economic outcomes.
3. Use the instrumental variable to estimate the causal effect of education spending on long-term economic outcomes, addressing the endogeneity issue.

Sum up

- Causal sufficiency
  - Under SUTVA, positivity, ignorability: sample means estimator

$$\widehat{\text{ATE}} = \frac{1}{n_t}\sum_{i=1}^{n} Y_i T_i - \frac{1}{n_c}\sum_{i=1}^{n} Y_i(1 - T_i)$$

  - When those assumptions hold? ... only RCTs ...
  - Lightning assumptions: SUTVA, conditional positivity, conditional ignorability, wrt backdoor/frontdoor set
    - Post-stratification:
    - IPTW: propensity score as weights,
    - Matching
    - Regression-based methods - double ML
- Hidden confounders
  - Instrumental variable analysis

Sum up

- Causal sufficiency
  - Under SUTVA, positivity, ignorability: sample means estimator

$$\widehat{\text{ATE}} = \frac{1}{n_t}\sum_{i=1}^{n} Y_i T_i - \frac{1}{n_c}\sum_{i=1}^{n} Y_i(1 - T_i)$$

  - When those assumptions hold? ... only RCTs ...
  - Lightning assumptions: SUTVA, conditional positivity, conditional ignorability, wrt backdoor/frontdoor set
    - Post-stratification:

$$\widehat{\text{ATE}} = \sum_{x} \widehat{\text{CATE}}(x)\frac{n_x}{n}$$

    - IPTW: propensity score as weights,

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} w_i(Y_i T_i - Y_i(1 - T_i))$$

    - Matching
    - Regression-based methods - double ML
- Hidden confounders
  - Instrumental variable analysis

Sum up

- Causal sufficiency
  - Under SUTVA, positivity, ignorability: sample means estimator

$$\widehat{\mathrm{ATE}} = \frac{1}{n_t}\sum_{i=1}^{n} Y_i T_i - \frac{1}{n_c}\sum_{i=1}^{n} Y_i(1 - T_i)$$

  - When those assumptions hold? ... only RCTs ...
  - Lightning assumptions: SUTVA, conditional positivity, conditional ignorability, wrt backdoor/frontdoor set
    - Post-stratification:
    - IPTW: propensity score as weights,
    - Matching
    - Regression-based methods - double ML
- Hidden confounders
  - Instrumental variable analysis

References

- *Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies*, Donald Rubin, Journal of Educational Psychology, 1974
- *Causal Inference: The Mixtape*, Scott Cunningham, 2021
- *Causal inference in statistics, social, and biomedical sciences*, Imbens, G. W., & Rubin, D. B., Cambridge University Press, 2015
- *Double/debiased machine learning for treatment and structural parameters*, Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, James Robins, The Econometrics Journal, 2018
- *Estimation and Inference of Heterogeneous Treatment Effects using Random Forests*, Stefan Wager & Susan Athey, Journal of the American Statistical Association, 2018