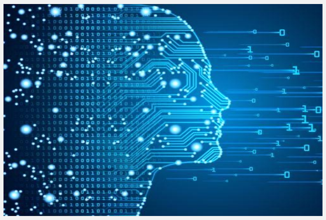
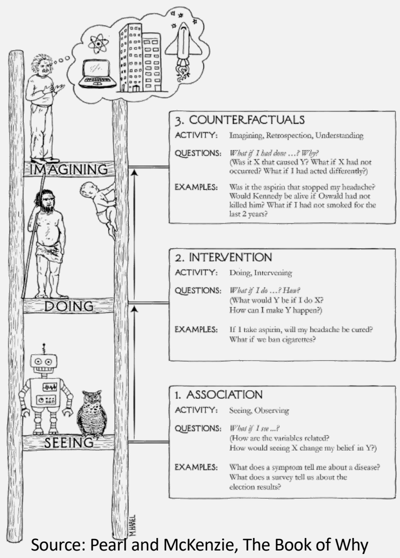


Principles of Causal Inference

Vasant G. Honavar

Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics, Public Health Sciences and Neuroscience
Center for Artificial Intelligence Foundations and Scientific Applications
Institute for Computational and Data Sciences
Huck Institutes of the Life Sciences
Clinical and Translational Sciences Institute
Northeast Big Data Hub
Pennsylvania State University
vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu

Ladder of Causation



Source: Pearl and McKenzie, The Book of Why

- Seeing: Most animals, learning machines populate the first rung. They learn from association.
- Doing: Tool users, including early humanoids, and perhaps some animals, populate the second rung. They can reason about and learn from interventions.
- Imagining: Humans populate the top rung. They can imagine worlds that do not exist and reason about, and learn from, counterfactuals.

What exactly are counterfactuals?

- Would US have elected a Democrat as President had Biden not withdrawn from the race in July 2024?
- Would Joe have bought On sneakers had he not been targeted by an ad?
- Would Jen have recovered from cancer had she not received chemotherapy?
- The unrealized antecedent is called the counterfactual

What exactly are counterfactuals?

- Suppose I am driving from my home in Centre Hill to the University Park airport on a football weekend
  - Option 1: Take East Branch Road to Lemont and take 322 West – $\mathrm{do}(X = 1)$
  - Option 2: Take East Branch Road to Atherton Street and take University Ave – $do(X = 0)$
- I choose Option 2 (shorter route that I am used to)
  - It takes me an hour to get to the airport and I miss my flight
  - I say to myself – I should have taken 322 West instead
  - What doe this mean?
  - If I had taken 322 West, I would have reached the airport sooner

Motivation for Counterfactuals

- After reaching the airport, I tell myself "I should have taken 322 West"
- I am thinking had I taken 322 West, I would have reached the airport sooner (and managed to catch my flight)
- My thinking is informed by my experience – that it took me 1 hour to reach the airport via University Ave
- When I decided to take University Ave, had I anticipated that it would take me 1 hour to get to the airport via University Ave, I would have taken 322 West instead if I thought that doing so would get me to the airport in less than 1 hour!
- What information did I need to make a rational choice?

How can we express counterfactuals?

- When I decided to take University Ave, had I anticipated that it would take me 1 hour to get to the airport via Univ. Ave, I would have taken 322 West instead if I thought that doing so would get me to the airport in less than 1 hour!
- What information did I need to make a rational choice?
  - The expected time to reach the airport via 322 West conditioned on the observation that it took me an hour to reach the airport via Univ. Ave

How can we express counterfactuals?

- How can I get the expected time to reach the airport via 322 West conditioned on the observation that it took me an hour to reach the airport via Univ. Ave?
- Should we compare $\mathbb{E}(t \mid do(X = 1), t = 1\ hour)$ with the time it actually took me to reach the airport via University Ave?
- What is $\mathbb{E}(t \mid do(X{=}1), t = 1\ hour)$?
- $\mathbb{E}(t \mid do(X{=}1), t = 1\ hour) =$ 1 hour!
- If that was the case, taking 322 West should make no difference
- What is wrong with my logic?
- The t whose expectation we are taking and the t we are conditioning are are not the same t
- How can I express the quantity I want to express?

Note on notation

- We use $Y^{a=1}$, $Y_{a=1}$ or Y_1 to denote the value of Y under the intervention $do(A = 1)$

Motivating the counterfactuals

- The *do* operator lets us distinguish between $P(t|do(X=0)$ and $P(t|do(X=1)$
- But the *do* operator is too crude to distinguish between the hypothetical driving time to the airport on 322 West conditioned on the actual driving time on University Ave
- We need a notation to distinguish between
  - Driving time to airport via 322 West: $Y_{X=1}$ or Y_1
  - Actual (observed) driving time Y to airport via Univ Ave
- We need to estimate $\mathbb{E}(Y_{X=1}\,|X=0, Y=1)$
- The expression contains a hypothetical event $Y_{X=1}$ predicated on the event $do(X=1)$, conditioned on a conflicting events $X=0$ and $Y=1$ that actually occurred (and hence observed)!
- That is, $Y=Y_{X=1}$ and $X=0$ (and $Y=Y_{X=0}=1$) occur in different worlds!

Do expressions are not enough to express counterfactuals

- We need a notation to distinguish between
  - Driving time to airport via 322 West: $Y_{X=1}$ or Y_1
  - Actual (observed) driving time Y to airport via Univ Ave
- We need to estimate $\mathbb{E}(Y_{X=1} \,|X = 0, Y = 1)$
- $Y = Y_{X=1}$ and $X = 0$ (and $Y = Y_{X=0} = 1$) occur in different worlds!
- $\mathbb{E}(Y_{X=1} \,|X{=}0, Y{=} Y_0 = 1)$ is very different from $\mathbb{E}(Y|do(X = 0))$
  - The first is about expectation of Y in the counterfactual world conditioned on observations in the factual world.
  - The second is about expectation of Y in a world conditioned on intervention in the same world.
- We can't reduce the first expression to a do expression
- We can't estimate it from an intervention experiment

Motivating the counterfactuals

- We can't reduce $E(Y_{X=1} \mid X=0, Y = Y_0 = 1)$ to a do expression
- Hence, we cannot apply do-calculus!
  - You can only
    - *do* an intervention on everyone in the population (or everyone with the same covariates X)
  - However, as the preceding example shows, there are interesting causal questions having to do with individual level counterfactuals that cannot be operationalized using the *do*-operator
- What does it say about the completeness of do-calculus?
- Nothing!
- Why? *do*-calculus is about causal effects in populations, NOT individuals!

Motivating the counterfactuals

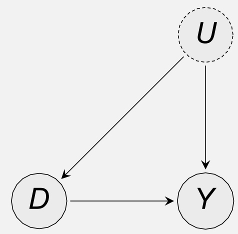
- Can we use an RCT to get at $E(Y_{X=1} \mid X=0, Y=Y_0=1)$?
  - An RCT will get us $E(Y \mid do(X=0))$ and $E(Y \mid do(X=1))$
  - An RCT will NOT get us $E(Y_{X=1} \mid X=0, Y=Y_0=1)$!
  - Why not?
  - Because X cannot simultaneously be both 1 and 0!
- If we cannot estimate $E(Y_{X=1} \mid X=0, Y=Y_0=1)$ from an RCT, there is no hope of estimating it from observational data!
- What if we estimate the freeway driving time for another driver or at another time of the day as a surrogate for your driving time from SC to NYC had you taken the freeway?
  - That would be an approximation
  - The quality of the approximation depends on many factors

Structural Causal Models Recap

A structural causal model $M = (V, U, F, \mathrm{P}(u))$ where:

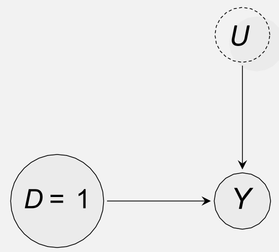
- V is a set of endogenous (observed) variables.
- U is a set of exogenous (unobserved) variables.
- F is a set of functions $f : D \rightarrow V_i$ where $D \subseteq V \cup U$ and $V_i \in V$.
- $P(u)$ is a probability distribution on U.

Recap: Causal Effects as Interventions



- This model corresponds to the following structural equations
- $D = f_D(U)$
- $Y = f_Y(D, U)$
- What do the graph and the equations look like when we intervene and "do" $D = 1$?

Recap: Causal Effects as Interventions



- This model corresponds to
- $D = 1$
- $Y = f_Y(1, U)$

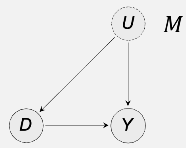
Recap: Causal Effects as Interventions

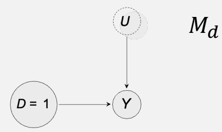
- If we $do(D = 1)$, then $D = 1$, and $Y = f_Y(1, U)$
- This Y under $do(D = 1)$ is a function of U and hence differs across individuals
- The mean of Y under the intervention $do(D = 1)$ is:

$$E[Y \mid do(D = 1)] = \sum_u f_Y(1, u) P(U = u)$$

- $f_Y(1, u)$ is Y if D is set to 1 for a unit with infinitely many features u
- This value $f_Y(1, u)$ is in fact a (individual-level) counterfactual
- "What would Y be if D were set to 1 in an individual with covariates u"?

Structural Interpretation of Counterfactuals





- If we $do(D = d)$ in an SCM M,
  - We get the SCM M_d where $D = f_D(u)$ is replaced by $D = d$ The counterfactual value of Y in unit u in model M when D is set to d is $Y_{M_d}(u)$, or $Y_d(u)$ or just Y_d
- The variable Y is passively observed, and the variable Y_d denotes the result of an intervention $D = d$
- This definition of counterfactuals, because it refers to interventional outcome, relies on a causal model

Fundamental law of counterfactuals

Counterfactuals obey:

- Consistency: if $D = d$ then $Y_d = Y$
- If D is binary, $Y = D\ Y_1 + (1 - D)\ Y_0$

- Y_1 is the observed value of Y when X is set to 1

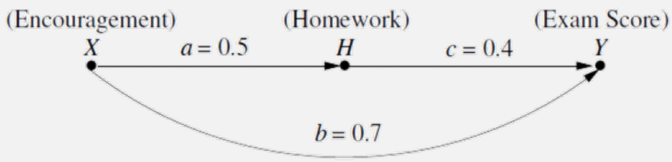
Causal Effects using Counterfactuals

- What is then the average causal effect of binary D on Y using not the *do*-operator, but counterfactuals?
  - $\mathbb{E}[Y_1] - \mathbb{E}[Y_0]$
- In the literature (a la Rubin) that uses only counterfactuals but no graphs, this is often called the average treatment effect (of D on Y)
- In the language of causal models:
  - $\mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \mathbb{E}[Y|do(D = 1)] - \mathbb{E}[Y|do(D = 0)]$
- But counterfactuals allow us to also think about causal effects for individuals: $Y_1(u) - Y_0(u)$
- This individual treatment effect will vary across individuals as a function of u

Interpreting counterfactuals

- Suppose M is a structural causal model $(V, U, F\ )$, exogenous variables U (latent) with known domains
- $U = u$ implies an individual in the population (e.g., a person, a situation in Nature)
- $X(u)$ denotes the characteristics of an individual with $U = u$
- Law of counterfactuals (LoC)
  - $Y_d(u) = Y_{M_d}(u)$
    - We can think of LoC as the solution for]Y in the surgically modified version of M, namely, M_d
    - LoC provides answer to questions such as what would Y have been had D been set to d?

From population data to individual behavior

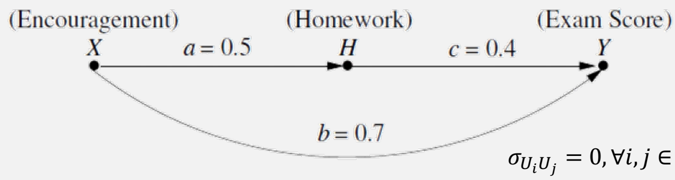


$$X = U_X$$
$$H = a\,X + U_H$$
$$Y = b\,X + c\,H + U_Y$$

$$0.5 = U_X$$
$$1 = (0.5)(0.5) + U_H$$
$$1.5 = (0.7)\,(0.5) + (0.4)\,(1) + U_Y$$

- Suppose Joe has $X = 0.5$, $H = 1$, and $Y = 1.5$
  - We find that
    - $U_X = 0.5$
    - $U_H = 0.75$ and
    - $U_Y = 0.75$

From population data to individual behavior



$\sigma_{U_i U_j} = 0, \forall i, j \in \{X, H, Y\}$

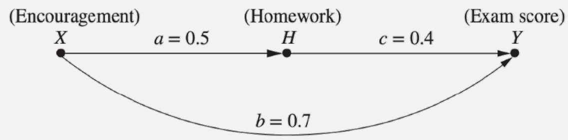
Each variable has mean= 0 and variance = 1

$$X = U_X$$
$$H = a\,X + U_H$$
$$Y = b\,X + c\,H + U_Y$$

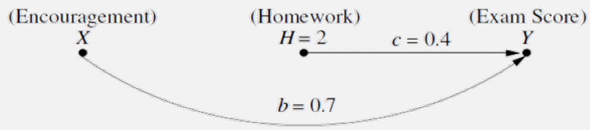
$$0.5 = U_X$$
$$1 = (0.5)(0.5) + U_H$$
$$1.5 = (0.7)\,(0.5) + (0.4)\,(1) + U_Y$$

- Suppose Joe has $X = 0.5$, $H = 1$, and $Y = 1.5$
  - We find that
    - $U_X = 0.5$
    - $U_H = 0.75$ and
    - $U_Y = 0.75$

From population data to individual behavior



$$X = U_X$$
$$H = a\,X + U_H$$
$$Y = b\,X + c\,H + U_Y$$



$$X = U_X$$
$$H = 2$$
$$Y = b\,X + c\,H + U_Y$$

- What happens to Joe's score when we double the homework?
- $H = 2, U_X = 0.5, U_H = 0.75$, and $U_Y = 0.75$

$$Y_{H=2} = (0.7)(0.5) + (2)(0.4) + 0.75$$
$$= 0.35 + 0.8 + 0.75$$
$$= 1.90$$

Counterfactuals in Linear Systems

- Structural model $Y = \alpha + \beta D + E$
- This model claims that for every unit u, $Y_d(u) = \alpha + \beta d + E$ so that for every u, $Y_1(u) - Y_0(u) = \beta$
- β is one structural coefficient (identifiable from observational data under certain conditions)
- Given the causal assumptions embodied in this structural causal model, β, the causal effect of D on Y the same for every individual.
- This is almost always wrong
  - If motherhood M affects wages W differently among women
  - We couldn't possibly assert that $W = \alpha + \beta M + E$
- Structural models are not regressions, but the structural coefficients, under certain conditions (which we went over in previous lectures), can be identified from observational data

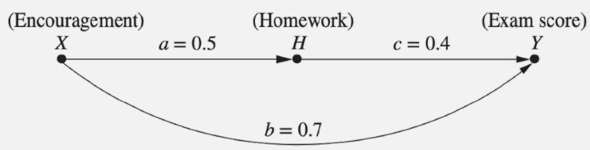
Computing Deterministic Counterfactuals

- Abduction[1]: Use evidence $E = e$ to determine the value of (past) U
- Action: Modify the model M, by removing the structural equations for the variables in D by replacing them with $D = d$, to get the modified model M_d
- Prediction: Use the modified Model M_d and the (past) value of U to compute the value of Y_d, the consequence of the counterfactual based on our understanding of the past and the imagined intervention $D = d$
- Counterfactuals, which are taken as primitives in Rubin's potential outcomes framework, are derived properties of structural equation models

- [1] Peirce, C. S. *Collected Papers of Charles Sanders Peirce*, C. Hartshorne, P. Weiss, and A. Burks (ed), 1931–1958, Cambridge MA: Harvard University Press.

Probabilistic counterfactuals

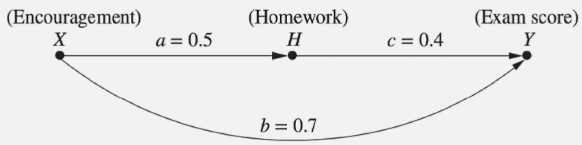
- What if counterfactuals pertain to a subset of individuals in a population?
- It is unlikely that their counterfactual outcomes are identical
  - Example: Effect of motherhood on income of women
- Suppose we wanted to know what would have happened if all students with $Y < 2$ had their homework H doubled?



$$X = U_X$$
$$H = a\,X + U_H$$
$$Y = b\,X + c\,H + U_Y$$

Probabilistic counterfactuals

- Suppose we wanted to know what would have happened if all students with $Y < 2$ had their homework H doubled?

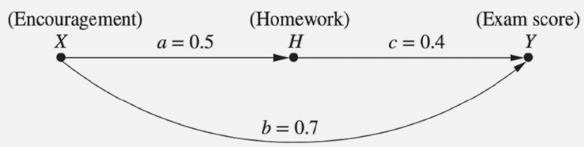


$$X = U_X$$
$$H = a\,X + U_H$$
$$Y = b\,X + c\,H + U_Y$$

- Can we use do expressions to express such counterfactuals?
- No. Do expressions cannot restrict the set of individuals intervened on in the manner specified

Probabilistic counterfactuals

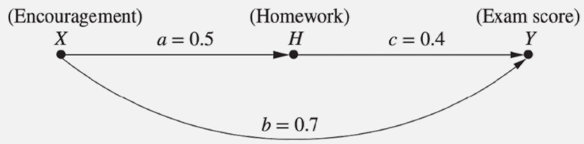
- Suppose we wanted to know what would have happened if all students with $Y < 2$ had their homework H doubled?



$$X = U_X$$
$$H = a\,X + U_H$$
$$Y = b\,X + c\,H + U_Y$$

- Suppose Joe's score was less than 2. We want to know what his score would have been had his homework been doubled?
  - Unlike in the deterministic case, we don't know everything (X, Y, H) about Joe.
  - All we know is that he is in the group with $Y < 2$

Probabilistic counterfactuals



$$X = U_X$$
$$H = a\,X + U_H$$
$$Y = b\,X + c\,H + U_Y$$

- Suppose Joe's score was less than 2. We want to know what his score would have been had his homework been doubled?
- Unlike in the deterministic case, we don't know everything (X, Y, H) about Joe. All we know is that he is in the group with $Y < 2$
- We cannot determine the precise value of $U = \{U_X, U_H, U_Y\}$ for Joe
- $P(U)$ induces a distribution over the observables $\{X, Y, H\}$
- This presents us with the problem of answering probabilistic counterfactual queries

Probabilistic Counterfactual Given a Causal Model

- Given that we observe the feature $E = e$ for a given individual, what is the expected outcome Y for that individual had D been d?
  - That is, we want to know: $\mathbb{E}[Y_{D=d} | E = e]$
- Computing the probabilistic counterfactual given a causal model M involves 3 steps:
  - Abduction: Use evidence $E = e$ to update $P(U)$ to $P(U|E = e)$
  - Action: Modify the model M, by removing the structural equations by setting $D = d$, to get the modified model M_d
  - Prediction: Use the modified Model M_d and $P(U|E = e)$ to compute the expectation of Y, the consequence of the counterfactual
- Counterfactuals, which are taken as primitives in Rubin's potential outcomes framework, are derived properties of structural equation models

Example

$X = aU$

$Y = bX + U$

Suppose $a = b = 1$

$U = \{1,2,3\}$

$P(u=1) = \frac{1}{2}$, $P(u=2) = \frac{1}{3}$ and $P(u=3) = \frac{1}{6}$

u	$X(u)$	$Y(u)$	$Y_1(u)$	$Y_2(u)$	$Y_3(u)$	$X_1(u)$	$X_2(u)$	$X_3(u)$
1	1	2	2	3	4	1	1	1
2	2	4	3	4	5	2	2	2
3	3	6	4	5	6	3	3	3

$X(1) = (1)(1) = 1.$

$Y(1) = (1)X(1) + 1 = (1)(1) + 1 = 2$

How do we compute $Y_1(2)$

$Y_1(2)$ is the result of intervention setting $X = 1$ on Y with $U = 2$

- Drop the first Structural equation and set $X = 1$.
- Use second structural equation to calculate $Y_1(2) = (1)(1) + 2 = 3$

Example

$X = aU$

$Y = bX + U$

Suppose $a = b = 1$

$U = \{1,2,3\}$

$P(u = 1) = \frac{1}{2}$, $P(u = 2) = \frac{1}{3}$ and $P(u = 3) = \frac{1}{6}$

u	$X(u)$	$Y(u)$	$Y_1(u)$	$Y_2(u)$	$Y_3(u)$	$X_1(u)$	$X_2(u)$	$X_3(u)$
1	1	2	2	3	4	1	1	1
2	2	4	3	4	5	2	2	2
3	3	6	4	5	6	3	3	3

- We can compute the probability that Y would be 3 had X been 2
  - $P(Y_2 = 3)$
  - $Y_2(u) = 3$ occurs only in the first row, when $U = 1$ which occurs with probability $P(1) = 1/2$

Example

$X = aU$

$Y = bX + U$

Suppose $a = b = 1$

$U = \{1,2,3\}$

$P(u = 1) = \frac{1}{2}$, $P(u = 2) = \frac{1}{3}$ and $P(u = 3) = \frac{1}{6}$

u	$X(u)$	$Y(u)$	$Y_1(u)$	$Y_2(u)$	$Y_3(u)$	$X_1(u)$	$X_2(u)$	$X_3(u)$
1	1	2	2	3	4	1	1	1
2	2	4	3	4	5	2	2	2
3	3	6	4	5	6	3	3	3

- We can compute any counterfactual probability
  - $P(Y_2 = 4) = P(U = 2) = 1/3$
- We can compute any joint probability
  - $P(Y_1 < 4, Y_2 > 3) = 1/3$
  - Note that this is a cross-world event spanning $X = 1$ and $X = 2$ which intersect at $U = 2$

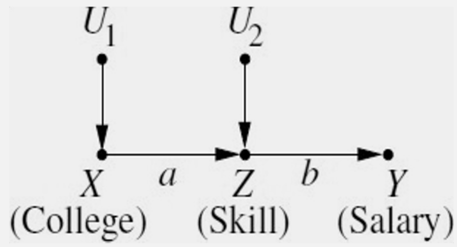
The power of probabilistic counterfactuals

- Given an SCM, we can compute any counterfactual probability
- Given an SCM, we can compute any joint probability over combinations of counterfactuals
  - E.g. $P(Y_1 = y_1 \quad , Y_2 = y_2)$
- This allows us to compute conditional probabilities over counterfactuals and define independence among counterfactuals just as we did over observables
- This is something we cannot do using the $do(X = x)$ notation

Counterfactuals and do-expressions reexamined

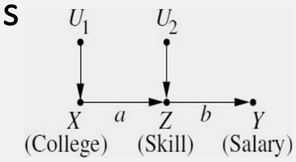
Example model:

- $X = 1$ denotes college educated
- U_2 =1 denotes having work experience
- Z denotes skill level
- Y denotes salary



$$X = U_1 \quad Z = aX + U_2, \quad Y = bZ$$

Limitation of the do-expressions



$X = U_1 \quad Z = aX + U_2, \quad Y = bZ$

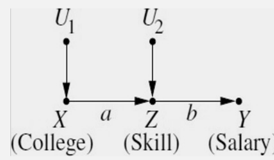
Suppose we want to compute $\mathbb{E}[Y_{X=1} | Z = 1]$ the expected salary of individuals with skill level $Z = 1$, had they received a college education

Can we use $\mathbb{E}[Y \,| do(X = 1), Z = 1]$?

- The do-expression stands for the expected salary of individuals who all finished college and have since acquired skill level $Z = 1$.
- The salaries of these individuals, as the SCM shows, depend only on their skill, and are not affected by whether they obtained the skill through college or work experience.
- Conditioning on $Z = 1$, in this case, cuts off the effect of the intervention that we're interested in.

Limitation of the do-expressions

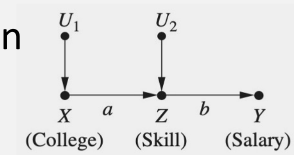
Suppose we want to compute $E[Y_{X=1}|Z = 1]$ the expected salary of individuals with skill level $Z = 1$, had they received a college education



$$X = U_1 \quad Z = aX + U_2, \quad Y = bZ$$

- The individuals that are relevant for computing $\mathbb{E}[Y_{X=1}|Z = 1]$ are excluded by the do-expression $\mathbb{E}[Y \,|do(X = 1), Z = 1]$
- In general,
  - $\mathbb{E}[Y \,|do(X = 1), Z = 1] = \mathbb{E}[Y \,|do(X = 0), Z = 1]$
    but
  - $\mathbb{E}[Y_{X=1}|Z = 1] \neq \mathbb{E}[Y_{X=0}|Z = 1]$
  - Why?

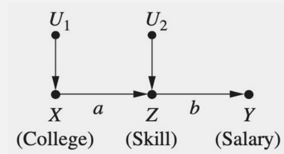
Counterfactual versus do-expression



$$X = U_1 \quad Z = aX + U_2, \quad Y = bZ$$

- $\mathbb{E}[Y \mid do(X = 1), Z = 1] = \mathbb{E}[Y \mid do(X = 0), Z = 1]$
  - Y only depends on Z – Conditioning on Z d-separates X from Y
  - $Z = 1$ refers to current skills; intervention $do(X = 1)$ is an imagined intervention on education in an unrealized past, given current skills
- $\mathbb{E}[Y_{X=1} \mid Z = 1] \neq \mathbb{E}[Y_{X=0} \mid Z = 1]$
  - $Z = 1$ selects a subset of the population in which we examine the effect of intervening on X
  - $Z = 1$ and $X = 1$ refer to different worlds (pre- and post-intervention)

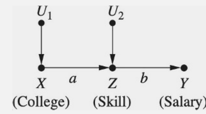
Can counterfactual encode a do-expression?



$X = U_1 \quad Z = aX + U_2, \quad Y = bZ$

- Yes. $\mathbb{E}[Y \mid do(X = 1), Z = 1] = \mathbb{E}[Y_{X=1} \mid Z_{X=1} = 1]$
- That is, we condition on the post-intervention value of Z
- $P[Y = y \mid do(X = 1), Z = z] = \frac{P(Y=y, Z=z \mid do(X=1))}{P(Z=z \mid do(X=1))}$

Counterfactual and do Calculations



$Y_0(u) = Y_{X=0}(u)$ $Z_0(u) = Z_{X=0}(u)$
$Y_1(u) = Y_{X=1}(u)$ $Z_1(u) = Z_{X=1}(u)$

$X = u_1 \quad Z = aX + u_2 \quad Y = bZ$

u_1	u_2	$X(u)$	$Z(u)$	$Y(u)$	$Y_0(u)$	$Y_1(u)$	$Z_0(u)$	$Z_1(u)$
0	0	0	0	0	0	ab	0	a
0	1	0	1	b	b	$(a+1)b$	1	$a+1$
1	0	1	a	ab	0	ab	0	a
1	1	1	$a+1$	$(a+1)b$	b	$(a+1)b$	1	$a+1$

Suppose $a \neq 1, a \neq 0, ab \neq 0$

$\mathbb{E}[Y_1 | Z = 1] = (a+1)b$

$\mathbb{E}[Y_0 | Z = 1] = b$

$\mathbb{E}[Y | do(X = 1), Z = 1] = b$

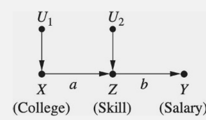
$\mathbb{E}[Y | do(X = 0), Z = 1] = b$

$\mathbb{E}[Y_1 - Y_0 \mid Z = 1] = ab$

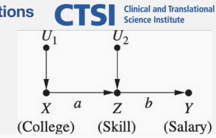
- Even though Z d-separates X from Y, X has a causal effect on Y among those with $Z = 1$
- While the salary of those at skill level $Z = 1$ depends only on their skill and not on education

Counterfactual and do Calculations

- Even though Z d-separates X from Y, X has a causal effect on Y among those with $Z = 1$
- While the salary of those at skill level $Z = 1$ depends only on their skill and not on education X, the salary of individuals currently at skill level $Z = 1$ could have been different had they had a different past
- Dependencies of this sort needed for retrospective reasoning about an unrealized past are not represented in standard structural causal models and cannot be expressed using do expressions
- Performing such reasoning requires augmenting causal graphs with counterfactual variables



Counterfactual and do Calculations



$X = u_1 \quad Z = aX + u_2 \quad Y = bZ$

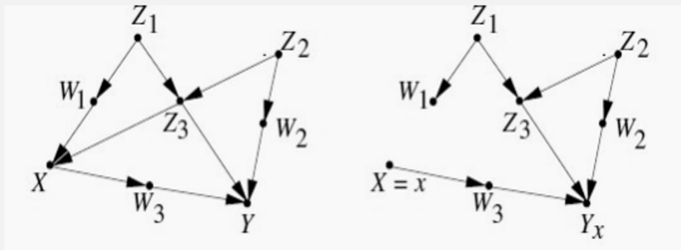
u_1	u_2	$X(u)$	$Z(u)$	$Y(u)$	$Y_0(u)$	$Y_1(u)$	$Z_0(u)$	$Z_1(u)$
0	0	0	0	0	0	ab	0	a
0	1	0	1	b	b	$(a+1)b$	1	$a+1$
1	0	1	a	ab	0	ab	0	a
1	1	1	$a+1$	$(a+1)b$	b	$(a+1)b$	1	$a+1$

- With $a \neq 0, a \neq 1$, $P(U_1)$ and $P(U_2)$ do not appear in the calculations because the condition $Z = 1$ occurs only for $u_1 = 0$ and $u_2 = 1$ forcing Y, Y_1 and Y_2 to take a definite value.
- But with $a = 1$, $Z = 1$ occurs when $u_1 = 0$ and $u_2 = 1$ as well as when $u_1 = 1$ and $u_2 = 0$
- $\mathbb{E}[Y_{X=1}|Z = 1] = b\left(1 + \frac{P(u_1=0)P(u_2=1)}{P(u_1=0)P(u_2=1)+P(u_1=1)P(u_2=0)}\right)$
- $\mathbb{E}[Y_{X=0}|Z = 1] = b\left(\frac{P(u_1=0)P(u_2=0)}{P(u_1=0)P(u_2=1)+P(u_1=1)P(u_2=0)}\right)$

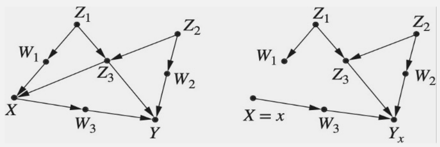
Graphical Representation of Counterfactuals

Can we see counterfactual in our causal model's graph?

Yes. Based on the fundamental law of counterfactuals $Y_x(u) = Y_{M_x}(u)$

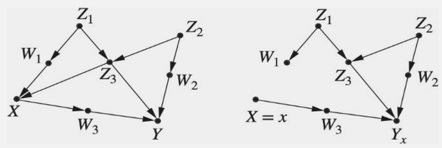


Graphical representation of counterfactuals



- How can we visualize counterfactual Y_x?
  - Remove arrows going into X yielding M_x in which Y is now Y_x
- Y_x denotes the value of Y when X is held constant at $X = x$

The Graphical Representation of Counterfactuals



Unobserved variables are not shown in the figure

- What variables would cause Y_x to vary?
  - All exogenous variables capable of influencing Y in M_x
  - Observed and unobserved parents of Y, and parents of nodes on the pathways between X and Y
    - Z_3, W_2, U_{w_3} and U_y
- If we can identify a set of variables Z in M_x that satisfy the back door criterion with respect to (X, Y_x), we render X independent of Y_x given Z

Counterfactual Interpretation of 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

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

- How can we calculate $P(y_x)$ from data?

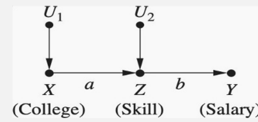
$$P(y_x) = \sum_z P(y_x|Z = z)\ P(Z = z) \qquad \text{LoT}$$

$$= \sum_z P(y_x|x, Z = z)\ P(Z = z) \qquad \text{BDC}$$

$$= \sum_z P(y|x, Z = z)\ P(Z = z) \qquad \text{Consistency}$$

This is just backdoor adjustment in the counterfactual setting!

Counterfactual Independence

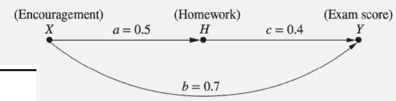


- Does the effect of education on salary (Y_x) depend on education (X), given skill Z?

  $Y_x \perp\!\!\!\perp X \mid Z$? or $\mathbb{E}[Y_x|X,Z] = \mathbb{E}[Y_x|Z]$?
- We know $\mathbb{E}[Y|X,Z] = \mathbb{E}[Y|Z] \because Z$ blocks all paths from X to Y
- Is the situation different for Y_x?
  - Yes!
  - Remove arrows into X to get M_x in which Y is Y_x
  - Which variables cause Y_x to vary when conditioned on Z?
  - U_2 - Why? Because U_2 and X become d-connected when conditioned on Z
  - Hence, $\mathbb{E}[Y_x|X,Z] \neq \mathbb{E}[Y_x|Z]$
  - In this case, Education matters in estimating the causal effect of Skill (Z) on Salary (Y)!

Counterfactual in Experimental Settings

- We saw that counterfactual queries can be answered from a fully specified structural model
- Consider data for 10 students

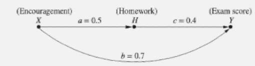


Participant	U_X	U_H	U_Y	X	Y	H	Y_0	Y_1	H_0	H_1	$Y_{00}\ldots$
	Participant characteristics			Observed behavior			Predicted potential outcomes				
1	0.5	0.75	0.75	0.5	1.50	1.0	1.05	1.95	0.75	1.25	0.75
2	0.3	0.1	0.4	0.3	0.71	0.25	0.44	1.34	0.1	0.6	0.4
3	0.5	0.9	0.2	0.5	1.01	1.15	0.56	1.46	0.9	1.4	0.2
4	0.6	0.5	0.3	0.6	1.04	0.8	0.50	1.40	0.5	1.0	0.3
5	0.5	0.8	0.9	0.5	1.67	1.05	1.22	2.12	0.8	1.3	0.9
6	0.7	0.9	0.3	0.7	1.29	1.25	0.66	1.56	0.9	1.4	0.3
7	0.2	0.3	0.8	0.2	1.10	0.4	0.92	1.82	0.3	0.8	0.8
8	0.4	0.6	0.2	0.4	0.80	0.8	0.44	1.34	0.6	1.1	0.2
9	0.6	0.4	0.3	0.6	1.00	0.7	0.46	1.36	0.4	0.9	0.3
10	0.3	0.8	0.3	0.3	0.89	0.95	0.62	1.52	0.8	1.3	0.3

$$X = U_X$$
$$H = a\,X + U_H$$
$$Y = b\,X + c\,H + U_Y$$
$$\sigma_{u_H u_Y} = 0$$

- We used the model to predict the potential outcomes
- In reality, we never can get such data (why?)
- Nevertheless, we can use the model to compute $\mathbb{E}[Y_{X=1} - Y_{X=0}]$

Counterfactual in Experimental Settings



- Suppose we do not have the model
- But we have data from an experiment in which X is assigned at random to members of the population
- The observed data correspond to the last two columns

Participant	Predicted potential outcomes Y_0	Predicted potential outcomes Y_1	Observed outcomes Y_0	Observed outcomes Y_1
1	1.05	1.95	1.05	■
2	0.44	1.34	■	1.34
3	0.56	1.46	■	1.46
4	0.50	1.40	■	1.40
5	1.22	2.12	1.22	■
6	0.66	1.56	0.66	■
7	0.92	1.82	■	1.82
8	0.44	1.34	0.44	■
9	0.46	1.36	■	1.36
10	0.62	1.52	0.62	■

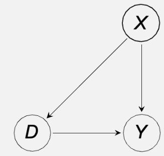
True average treatment effect: 0.90   Study average treatment effect: 0.68

- Now, because X is randomly assigned, the backdoor adjustment formula applies in the counterfactual setting with $Z = \{\ \}$
- $\mathbb{E}[Y_x]$=$\mathbb{E}[Y\ \ |X = x]$

- Because $\mathbb{E}[Y_x]$=$\mathbb{E}[Y\ \ |X = x]$, we can estimate $\mathbb{E}[Y_{X=1} - Y_{X=0}]$ $= \mathbb{E}[Y_{X=1}]$ $-\mathbb{E}[Y_{X=0}]$ from the observed data!
- Note that the quality of the estimate depends on sample size etc.

Applications of Counterfactuals

Freedom of choice and causal effects



- In most modern nations, some may attend university but are not forced to
- For those who attend university, does it pay off in terms of lifetime earnings?
- Y lifetime earnings, $D = 1$ if you go to university
- We can model the above using a very simple SCM :
- $Y = f_Y(D, U_Y) \qquad D = f_D(X, U_D)$
- U_Y and U_D may be correlated
- Why?
  - You love learning (X) so you attend university
  - But your love of learning can impact your earnings regardless of whether you attend university

Freedom of choice and causal effects

- How can you get at the question as to whether for those who attend university, does it pay off in terms of lifetime earnings?
- We need average treatment effect on the treated (ATT) which is expressed in terms of counterfactuals

$$\mathbb{E}[Y_1 - Y_0 | D = 1] = \mathbb{E}[Y_1 - Y_0 | D = 1]$$
$$= \mathbb{E}[Y_1 | D = 1] - \mathbb{E}[Y_0 | D = 1]$$

Estimation of ATT

Suppose X satisfies BDC with re (D, Y)

$$ATT = \mathbb{E}[Y_1 - Y_0|D = 1] =$$

$$\sum_x (\mathbb{E}[Y|D = 1, X = x] - E[Y|D = 0, X = x]) \cdot P(X = x|D = 1)$$

Estimation strategy:

- In each stratum of X, estimate X –specific effect
- If X is continuous, discretize it into strata
- Estimate $\mathbb{E}[Y|D = 1, X = x] - \mathbb{E}[Y|D = 0, X = x]$ via regression (regress D on Y and X
- Estimate $P(X = x|D = 1)$ nonparametrically

$$P(X = x|D = 1) = \frac{P(X = x, D = 1)}{P(D = 1)} \quad \text{(Bayes' rule)}$$

Relevance of ATT

- In general, if ATT = ATE, individuals do not make choices depending on their potential outcomes/no confounding
- If ATT > ATE, individuals (or some other force) optimizes the choice of treatment with respect to the outcome Y
- If ATT < ATE, people (or someone else) chooses the treatment as to "hurt" people (with respect to Y)!
- Identification of ATT and ATE gives some insight into decision-making processes

Effectiveness of a Job Training Program

- The government funds a job training program aimed at getting jobless people back into the workforce
- A pilot randomized experiment shows that the program is effective
  - A higher percentage of people were hired among those who finished the the program than among those who did not enroll in the program
- The program is approved, and the training is offered to any unemployed person who wants to enroll
- The hiring rate among those who complete the program turns out to be even higher than in the randomized pilot study

Effectiveness of a Job Training Program

- The hiring rate among those who complete the program turns out even higher than that in the randomized pilot study
- Critic:
  - Those who self-enroll, may be more intelligent, more resourceful, and more socially connected than those who were eligible but did not enroll and hence were more likely to have found a job regardless of the training
- What we really need to estimate is the differential benefit of the program on those enrolled
  - The extent to which hiring rate has increased among the enrolled, compared to what it would have been had they not enrolled

Effectiveness of a Job Training Program

- What we really need to estimate is the differential benefit of the program on those enrolled: the extent to which hiring rate has increased among the enrolled, compared to what it would have been had they not enrolled.
- ATT to the rescue
  - Let $X = 1$ represent training and $Y = 1$ represent hiring
  - The effect of training on the trained is $\mathbb{E}[Y_{X=1} - Y_{X=0} \mid X = 1]$
  - While there are situations in which ATT may not be identifiable, in many cases, as we have already seen, it is possible to identify ATT using backdoor or other adjustments

Personal Decision Making

- Cancer patients must decide between two treatments:
  - lumpectomy alone, or
  - lumpectomy plus irradiation
- Ms. Jones, in consultation with her oncologist, decides on the second option
  - Ten years later, Ms. Jones is alive, and the tumor has not recurred
  - She wonders: Do I owe my life to irradiation?
- Mrs. Smith, on the other hand, chooses the first option and her tumor recurred after a year
  - She wonders: Should I have undergone irradiation?

Can these speculations be substantiated using data?

- In general, no. Yes, under some assumptions.

Discriminatory hiring practices

- Mary files a lawsuit against Omega Inc, alleging discriminatory hiring practices
- She claims she applied for a job with Omega Inc. and despite being well-qualified job, she was not interviewed, allegedly because Omega, Inc. realized that she is female
- She claims that the hiring record of Omega Inc shows consistent preferences for male employees
- Does Mary have a case? Can hiring records prove whether Omega Inc. was discriminating when declining her job application?

Discriminatory hiring practices

- Does Mary have a case? Can hiring records prove whether Omega Inc. was discriminating when declining her job application?
  - Y = Mary being invited to interview (1 if invited, 0 if not)
  - X = Omega Inc.'s perception of Mary's gender (1 if male, 0 if female)
- What should we estimate? $P(Y_1 = 1|X = 1, Y = 0)$
- Probability that Mary would have been invited to interview conditional on Omega inc. believed her to be male and she was not invited to interview
- If there is no gender-based discrimination, this probability should be close to 0
- If this probability is greater than some threshold, that could be used as evidence for discrimination

Reducing gender disparity in hiring

- Two potential policy choices
  - Making hiring decisions gender-blind (easy)
    - Making hiring decisions gender blind helps us get at the direct effect of gender on hiring – the indirect effect of gender through other variables, e.g., qualification, etc. are not eliminated by gender-blind hiring
  - Additionally, eliminating gender inequality in education or job training (hard)
    - Eliminating gender disparity in hiring requires eliminating the indirect effects or the effect of gender mediated by other variables, e.g., qualification

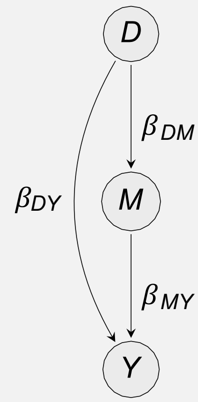
Direct and indirect effects
Path disabling interventions

History of (in)direct effects

- Being clear about the theoretical causal mechanism is a precondition for a good theory
- Very often, disagreement is not about direction of some causal effect, but about the mechanism
- The methodological literature on how to learn about causal mechanisms from data only started around 2000!
- While this is clearly of interest to **all** sciences
  - Pearl in 1st ed. of "Causality" (2000): "Indirect effects lack intrinsic operational meaning"
  - Rubin (2004): Indirect effects are "ill-defined" and "more deceptive than helpful"
  - Pearl changed his opinion in 2001 and gave general definitions, identification results, and policy implications of indirect effects

Direct and indirect effects in linear models

- What is the direct, what is the indirect effect of *D* on *Y* in this model?
- Direct: β_{DY} , indirect: $\beta_{DM}\beta_{MY}$
- Can you isolate the indirect effect using *do*?
- No! not possible to $do(D)$ so as to isolate $\beta_{DM}\beta_{MY}$
- But as we have seen, linear causal models permit causal estimation using series of linear regressions



From linear to general case

- In linear structural equation models, it is clear what direct and indirect effects are
  - Just look at path coefficients
- However, generalizing this notion to nonlinear models, or unknown functions *f* is more complicated
  - Pearl in 1st ed. of "Causality" 2000)
    - Indirect effects lack intrinsic operational meaning
    - It is not possible to express them using *do*-operator
    - It is not clear to which action they correspond
  - Pearl 2001, also 2nd ed. Of "Causality (2009)
    - An indirect effect is the effect of a variable when its direct effect is disabled
- Counterfactuals offer us a way out

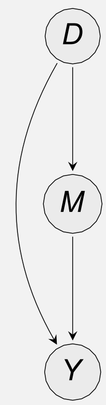
Gender-blind auditioning of musicians

- There is evidence that major orchestras have historically discriminated against female instrumental musicians
- Women $G = 1$ acquire musical skills M.
  - They audition for an orchestra, and then are hired $Y = 1$ with some probability p_1
  - They audition for an orchestra behind a curtain, and then are hired $Y = 1$ with a different probability p_2
- Having the musicians play behind a curtain makes sure the committee does not know gender, thus disabling the direct effect of gender on hiring
- Goldin and Rouse (2000)* found that introduction of gender-blind audition in various professional US orchestras substantially increased the representation of women in orchestra

* Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American economic review*, *90*(4), 715-741.

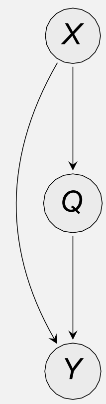
Path-disabling interventions

- A drug D helps to cure some illness Y
  - D results in headache which leads patients to take aspirin M.
  - Aspirin possibly also affects Y
- Suppose the drug company manipulates D so that it does not cause headache, while maintaining its direct effect of D on Y
  - Disable path $D \rightarrow M$
  - Some people may still naturally take aspirin for other reasons, which may moderate the direct effect of D on Y
- If we can identify the direct effect of D not going through M from observational data, we can get at the direct causal effect of the drug D on Y



Path disabling interventions

- Because we seek to disable the influence of one variable through only one (and not all) paths, we cannot express the effect of such interventions using the do operator
- However, we can express it using the language of counterfactuals
- Suppose we want to assess hiring disparity after putting in place gender-blind hiring procedures
  - We require that all female $(X = 0)$ applicants be treated like males $(X = 1)$ with comparable qualifications $(Q = q)$ and proceed to estimate the hiring rate under this counterfactual condition $Y_{X=1,Q=q}$
  - But because q varies among applicants, we need to average this according to distribution of q



Path disabling interventions

- Suppose we want to assess hiring disparity after putting in place gender-blind hiring procedures
  - We require that all female ($X = 0$) applicants be treated like males ($X = 1$) with comparable qualifications ($Q = q$) and proceed to estimate the hiring rate under this counterfactual condition $Y_{X=1,Q=q}$
  - But because q varies among applicants, we need to average $Y_{X=1,Q=q}$ according to distribution of q among females

$$\sum_q \mathbb{E}\left[Y_{X=1,Q=q}\right] P(Q = q|X = 0)$$

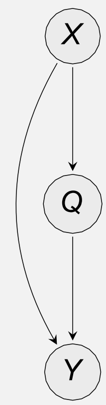
Path disabling interventions

- Similarly, the hiring rate among males is obtained by averaging $Y_{X=1,Q=q}$ with respect to distribution of q among males

$$\sum_q \mathbb{E}\left[Y_{X=1,Q=q}\right] P(Q=q|X=1)$$

- The indirect effect of gender on hiring as mediated by qualification is given by

$$\sum_q \mathbb{E}\left[Y_{X=1,Q=q}\right] [P(Q=q|X=0) - P(Q=q|X=1)]$$



Path disabling interventions

- The indirect effect of gender on hiring mediated by qualification is given by

$$\sum_q \mathbb{E}\left[Y_{X=1,Q=q}\right]\left[P(Q=q|X=0)-P(Q=q|X=1)\right]$$

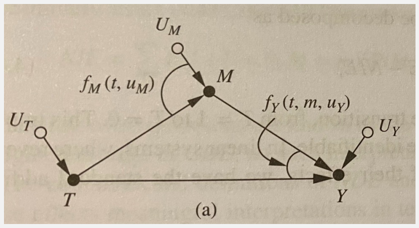
- This is the Natural indirect effect (NIE) of X on Y, mediated by q
- Can we estimate NIE from observational data?
- In the absence of confounding, we can show that

$$\sum_q \mathbb{E}\left[Y_{X=1,Q=q}\right]\left[P(Q=q|X=0)-P(Q=q|X=1)\right]$$

$$=\sum_q \mathbb{E}[Y|X=1,Q=q]\left[P(Q=q|X=0)-P(Q=q|X=1)\right]$$

We call this the mediation formula

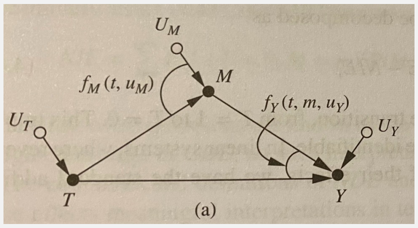
A toolkit for mediation



$$t = f_T(u_T)$$
$$m = f_M(t, u_M)$$
$$y = f_Y(t, m, u_Y)$$

- Total effect $TE = \mathbb{E}[Y_1 - Y_0]$
  $= \mathbb{E}[Y|do(T = 1) - \mathbb{E}(Y|do(T = 0)]$
- TE measures the expected change in Y as the treatment changes from $T = 0$ to $T = 1$, while the mediator is allowed to track the change in T naturally as dictated by f_M

A toolkit for mediation



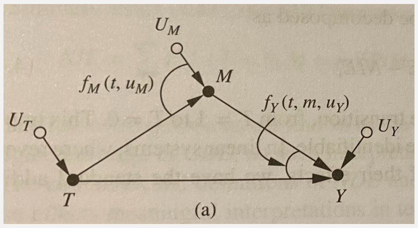
$t = f_T(u_T)$
$m = f_M(t, u_M)$
$y = f_Y(t, m, u_Y)$

- Controlled direct effect

$$CDE(m) = \mathbb{E}[Y_{1,m} - Y_{0,m}]$$
$$= \mathbb{E}[Y|do(T = 1, M = m) - \mathbb{E}(Y|do(T = 0, M = m)]$$

- CDE measures the expected change in Y as the treatment changes from $T = 0$ to $T = 1$, while the mediator is set to m uniformly for the entire population

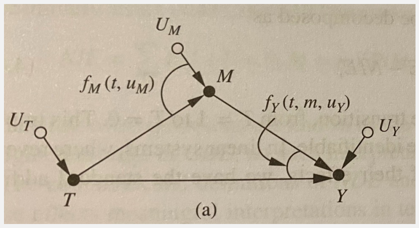
A toolkit for mediation



$$t = f_T(u_T)$$
$$m = f_M(t, u_M)$$
$$y = f_Y(t, m, u_Y)$$

- Natural direct effect $NDE = \mathbb{E}[Y_{1,M_0} - Y_{0,M_0}]$
- NDE measures the expected change in Y as the treatment changes from $T = 0$ to $T = 1$, while the mediator is set to whatever value it would have attained, for each individual, prior to the change, i.e., under $T = 0$

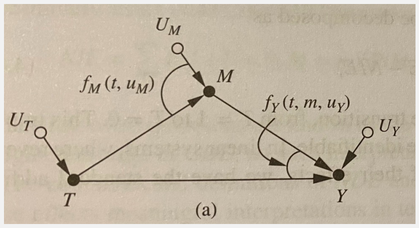
A toolkit for mediation



$$t = f_T(u_T)$$
$$m = f_M(t, u_M)$$
$$y = f_Y(t, m, u_Y)$$

- Natural indirect effect $NIE = \mathbb{E}[Y_{0,M_1} - Y_{0,M_0}]$
- NIE measures the expected change in Y when the treatment is held constant at $T = 0$ while the mediator M changes to whatever value it would have attained, for each individual, under $T = 1$.
- NIE captures, the portion of the effect that can be explained by mediation alone, while disabling the capacity of Y to respond to T

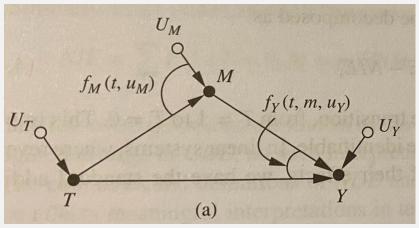
A toolkit for mediation



$t = f_T(u_T)$
$m = f_M(t, u_M)$
$y = f_Y(t, m, u_Y)$

- In general, $TE = NDE - NIE_{\ r}$
- Where $NIE_{\ r}$ denotes the NIE under the reverse change, i.e., $T = 1$ to $T = 0$ $NIE_r = \mathbb{E}[Y_{0,M_0} - Y_{0,M_1}]$
- In linear systems, $TE = NDE + NIE$
- Why?
- Because reversal of change flips the sign of the coefficient

Response fraction due to mediation



$$t = f_T(u_T)$$
$$m = f_M(t, u_M)$$
$$y = f_Y(t, m, u_Y)$$

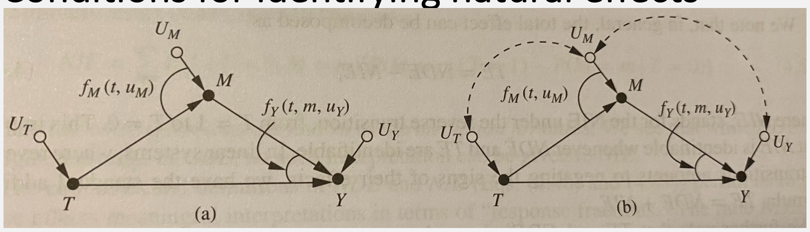
- NDE/TE measures the fraction of the response that is transmitted directly, with M frozen
- NIE/TE measures the fraction of the response that is transmitted through M, with Y blinded to T
- $(TE - NDE)/TE$ measures the fraction of the response that is necessarily due to M

Identifiability

- TE and CDE(m) are do-effects
  - They can be expressed using do-operator
  - They can be obtained from experiments

    Or
  - They can be obtained from observations, whenever possible, using do-calculus
- What about NDE and NIE?
  - They are not do-effects
  - They are expressed using counterfactuals
  - We need new identifiability conditions

Conditions for identifying natural effects



We can identify NDE and NIE provided there exists a set W of measured covariates such that

1. No member of W is a descendent of T
2. W blocks all backdoor paths from M to Y (after removing $T \rightarrow M$ and $T \rightarrow Y$)
3. The W-specific effect of T on M is identifiable (from experiments or observations)
4. The W-specific joint effect of $\{T, M\}$ on Y is identifiable (from experiments or observations)

Identification of NDE

When the first two conditions hold,

$NDE = \sum_m \sum_w [\mathbb{E}[Y|do(T=1), M=m, W=w] - \mathbb{E}[Y|do(T=0), M=m, W=w]]. P(M=m|do(T=0), W=w)P(W=w)$

In addition, if W de-confounds the relationships in 3 and 4, we can replace interventional probabilities by their observational counterparts

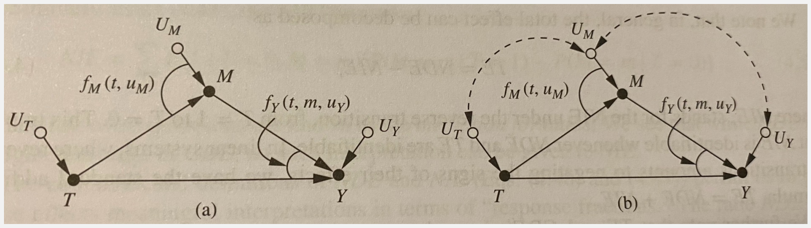
$NDE = \sum_m \sum_w [\mathbb{E}[Y|T=1, M=m, W=w] - \mathbb{E}[Y|T=0, M=m, W=w]].\ P(M=m|T=0, W=w)P(W=w)$

In the non-confounded case, this simplifies to

$NDE = \sum_m \sum_w [\mathbb{E}[Y|T=1, M=m] - \mathbb{E}[Y|T=0, M=m]]P(M=m|T=0)$

$$NIE = \sum_m \mathbb{E}\,[Y|T=0, M=m][P(M=m|T=1) - P(M=m|T=0)]$$

Response fraction due to mediation



- NDE is a weighted average of CDE
- NDE/TE measures the fraction of the response that is transmitted directly, with M frozen
- NIE/TE measures the fraction of the response that is transmitted through M, with Y blinded to T
- $(TE - NDE)/TE$ measures the fraction of the response that is necessarily due to M

Summary: Counterfactuals, path disabling interventions, Mediation

- Mediation is easy to analyze in the case of linear causal models
  - Need to only worry about unobserved confounding
- But in general, causal models may not be linear
- Definition of direct/indirect effects relies on path-disabling interventions instead of variable-setting or do interventions
  - This leads to
    - Nested counterfactuals $\mathbb{E}[Y_{D=1,M_{D=0}}]$ and definition of natural direct (as distinct from controlled direct) and indirect effects
    - Identification via blocking back-door paths

Summary: Counterfactuals, path disabling interventions, Mediation

- Definition of direct/indirect effects relies on path-disabling interventions instead of variable-setting or do interventions
- Leads to
  - Nested counterfactuals $\mathbb{E}[Y_{D=1,M_{D=0}}]$ and definition of natural direct (as distinct from controlled direct) and indirect effects.
  - Identification via blocking back-door paths
- Estimand is very similar to ATT/ATE estimand:
  - For all values m of mediator M compute mean differences in Y for different values d of treatment;
  - Weight each difference with distribution of M under baseline value of D
- Nonparametric identification of NDE/NIE impossible in the presence of post-treatment confounders