Counterfactuals and Mediation

Brady Neal

causalcourse.com

Counterfactuals Basics

Important Application: Mediation

Counterfactuals Basics

Important Application: Mediation

Fundamental Problem of Causal Inference

$\mathrm{do}(T = 1)$

$Y_i(1) = 1$

$\mathrm{do}(T = 0)$

$Y_i(0) = 0$

T	: observed treatment
Y	: observed outcome
i	: used in subscript to denote a specific unit/individual
$Y_i(1)$	: potential outcome under treatment
$Y_i(0)$	: potential outcome under no treatment

Causal effect

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

Fundamental Problem of Causal Inference

Counterfactual

$\mathrm{do}(T=1)$

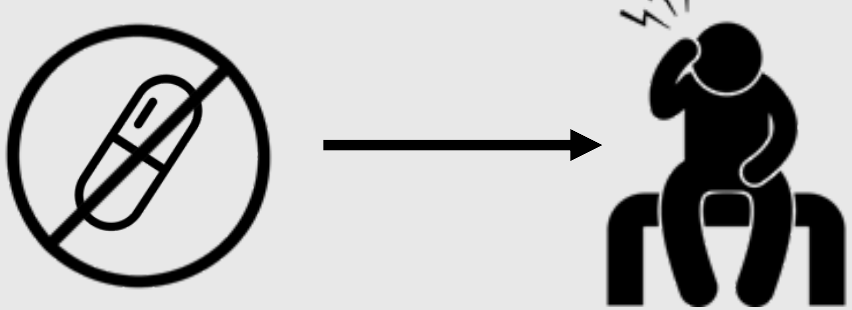
$Y_i(1) = 1$

T	: observed treatment
Y	: observed outcome
i	: used in subscript to denote a specific unit/individual
$Y_i(1)$	: potential outcome under treatment
$Y_i(0)$	: potential outcome under no treatment

Factual

$\mathrm{do}(T=0)$

$Y_i(0) = 0$



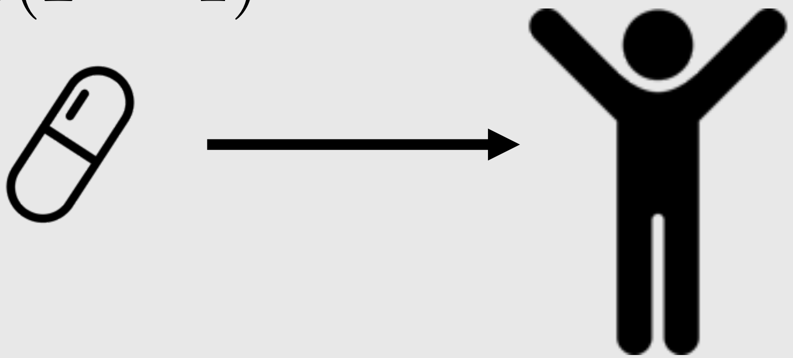
Causal effect

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

Fundamental Problem of Causal Inference

$\mathrm{do}(T = 1)$ $\qquad$ $Y_i(1) = 1$

Factual



T	: observed treatment
Y	: observed outcome
i	: used in subscript to denote a specific unit/individual
$Y_i(1)$	: potential outcome under treatment
$Y_i(0)$	: potential outcome under no treatment

Counterfactual

$\mathrm{do}(T = 0)$ $\qquad$ $Y_i(0) = 0$

Causal effect

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

We can compute counterfactuals using a parametric SCM.

Counterfactuals

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

Counterfactuals

Counterfactual: $P(Y(t) \mid \overbrace{T = t', Y = y'}^{\text{observation}})$

Counterfactuals

Counterfactual: $P(\underset{\substack{\uparrow \\ \text{hypothetical condition}}}{Y(t)} \mid \overbrace{T = t', Y = y'}^{\text{observation}})$

Counterfactuals

$$\text{Counterfactual: } P(\underset{\substack{\uparrow \\ \text{hypothetical condition}}}{Y(t)} \mid \overbrace{T = t', Y = y'}^{\text{observation}})$$

Different from CATE: $\mathbb{E}[Y(t) \mid X = x]$

Counterfactuals

Counterfactual: $P(\underbrace{Y(t)}_{\text{hypothetical condition}} \mid \overbrace{T = t', Y = y'}^{\text{observation}})$

Different from CATE: $\mathbb{E}[Y(t) \mid X = x] = \mathbb{E}[Y \mid do(t), X = x]$

Counterfactuals

Counterfactual: $P(\underbrace{Y(t)}_{\text{hypothetical condition}} \mid \overbrace{T = t', Y = y'}^{\text{observation}})$

Different from CATE: $\mathbb{E}[Y(t) \mid X = x] = \mathbb{E}[Y \mid do(t), X = x]$

Cannot express counterfactuals using do-notation

Roadmap for Computing Counterfactuals

Roadmap for Computing Counterfactuals

Given: Observation of (T, Y) (observation of potential outcome $Y(t)$ where t is the observed value of T)

Roadmap for Computing Counterfactuals

Given: Observation of (T, Y) (observation of potential outcome Y(t) where t is the observed value of T)

Main ingredient necessary: correct parametric model for the structural equation for Y

Roadmap for Computing Counterfactuals

Given: Observation of (T, Y) (observation of potential outcome Y(t) where t is the observed value of T)

Main ingredient necessary: correct parametric model for the structural equation for Y

Result: access to counterfactuals Y(t') at the unit-level

Computing Counterfactuals: Simple Example

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

Computing Counterfactuals: Simple Example

Y: happy or unhappy (1 or 0)

T: get a dog or don't (1 or 0)

Computing Counterfactuals: Simple Example

Y: 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; 0 if anti-dog person)

Computing Counterfactuals: Simple Example

Y: 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; 0 if anti-dog person)

SCM:
$$T := \dots$$
$$Y := UT + (1 - U)(1 - T)$$

Computing Counterfactuals: Simple Example

Y: 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; 0 if anti-dog person)

SCM: $T := \dots$

$Y := UT + (1 - U)(1 - T)$

Observation: T = 0 and Y = 0

Computing Counterfactuals: Simple Example

Y: 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; 0 if anti-dog person)

SCM: $T := \ldots$

$Y := UT + (1 - U)(1 - T)$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

Computing Counterfactuals: Simple Example

Y: 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; 0 if anti-dog person)

SCM: $T := \dots$

$Y := UT + (1 - U)(1 - T)$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$Y_u(1)?$

Computing Counterfactuals: Simple Example

Y: 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; 0 if anti-dog person)

SCM: $T := \dots$

$$Y := UT + (1 - U)(1 - T)$$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$Y_u(1)$?

Computing Counterfactuals: Simple Example

Y: 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; 0 if anti-dog person)

SCM:

$$T := \dots$$

$$\underline{Y := UT + (1 - U)(1 - T)}$$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$Y_u(1)$?

Step 1: Solve for U

Computing Counterfactuals: Simple Example

Y: 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; 0 if anti-dog person)

SCM: $T := \dots$

$\underline{Y := UT + (1-U)(1-T)}$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$Y = UT + (1-U)(1-T) \quad Y_u(1)?$

Computing Counterfactuals: Simple Example

Y: 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; 0 if anti-dog person)

SCM:
$$T := \dots$$
$$\underline{Y := UT + (1 - U)(1 - T)}$$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$$Y = UT + (1 - U)(1 - T)$$
$$0 = U(0) + (1 - U)(1 - 0)$$

$Y_u(1)$?

Computing Counterfactuals: Simple Example

Y: 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; 0 if anti-dog person)

SCM:
$$T := \dots$$
$$\underline{Y := UT + (1-U)(1-T)}$$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$$Y = UT + (1-U)(1-T)$$
$$0 = U(0) + (1-U)(1-0)$$
$$0 = 1 - U$$

$Y_u(1)$?

Computing Counterfactuals: Simple Example

Y: 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; 0 if anti-dog person)

SCM:
$$T := \dots$$
$$\underline{Y := UT + (1-U)(1-T)}$$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$$Y = UT + (1-U)(1-T)$$
$$0 = U(0) + (1-U)(1-0)$$
$$0 = 1 - U$$
$$U = 1$$

$$Y_u(1)?$$

Computing Counterfactuals: Simple Example

Y: 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; 0 if anti-dog person)

SCM:
$$T := \dots$$
$$\underline{Y := UT + (1-U)(1-T)}$$

Step 2: Individualized SCM

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$$Y = UT + (1-U)(1-T)$$
$$0 = U(0) + (1-U)(1-0)$$
$$0 = 1 - U$$
$$U = 1$$

$Y_u(1)$?

Computing Counterfactuals: Simple Example

Y: 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; 0 if anti-dog person)

SCM:
$$T := \dots$$
$$\underline{Y := UT + (1-U)(1-T)}$$

Step 2: Individualized SCM
$$T := \dots$$
$$Y := (1)T + (1-1)(1-T)$$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$$Y = UT + (1-U)(1-T)$$
$$0 = U(0) + (1-U)(1-0)$$
$$0 = 1 - U$$
$$U = 1$$

$Y_u(1)$?

Computing Counterfactuals: Simple Example

Y: 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; 0 if anti-dog person)

SCM:
$$T := \ldots$$
$$\underline{Y := UT + (1 - U)(1 - T)}$$

Step 2: Individualized SCM
$$T := \ldots$$
$$Y := T$$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$$Y = UT + (1 - U)(1 - T)$$
$$0 = U(0) + (1 - U)(1 - 0)$$
$$0 = 1 - U$$
$$U = 1$$

$Y_u(1)$?

Computing Counterfactuals: Simple Example

Y: 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; 0 if anti-dog person)

SCM:
$$T := \dots$$
$$\underline{Y := UT + (1-U)(1-T)}$$

Step 2: Individualized SCM
$$T := 1$$
$$Y := T$$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$$Y = UT + (1-U)(1-T)$$
$$0 = U(0) + (1-U)(1-0)$$
$$0 = 1 - U$$
$$U = 1$$

$Y_u(1)$?

Computing Counterfactuals: Simple Example

Y: 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; 0 if anti-dog person)

SCM: $T := \ldots$

$$\underline{Y := UT + (1 - U)(1 - T)}$$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$$Y = UT + (1 - U)(1 - T)$$

$$0 = U(0) + (1 - U)(1 - 0)$$

$$0 = 1 - U$$

$$U = 1$$

$$Y_u(1)?$$

Step 2: Individualized SCM

$$T := 1$$

$$Y := T$$

$$Y_u(1) = 1$$

Computing Counterfactuals: Simple Example

Y: 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; 0 if anti-dog person)

SCM: $T := \dots$

$$\underline{Y := UT + (1 - U)(1 - T)}$$

Observation: T = 0 and Y = 0 $\quad (Y_u(0) = 0)$

$$Y = UT + (1 - U)(1 - T)$$

$$0 = U(0) + (1 - U)(1 - 0)$$

$$0 = 1 - U$$

$$U = 1$$

$Y_u(1)$?

Step 2: Individualized SCM

$$T := 1$$

$$Y := T$$

$$Y_u(1) = 1$$

ITE: $Y_u(1) - Y_u(0) = 1 - 0 = 1$

General Steps for Deterministic Counterfactuals

General Steps for Deterministic Counterfactuals

From Chapter 4 of Pearl et al. (2016)'s Primer:

General Steps for Deterministic Counterfactuals

From Chapter 4 of Pearl et al. (2016)'s Primer:

1. Abduction: Use an observation to determine the value of U

General Steps for Deterministic Counterfactuals

From Chapter 4 of Pearl et al. (2016)'s Primer:

1. Abduction: Use an observation to determine the value of U
2. Action: Modify the SCM, by replacing the structural equation for T with T := t

General Steps for Deterministic Counterfactuals

From Chapter 4 of Pearl et al. (2016)'s Primer:

1. Abduction: Use an observation to determine the value of U
2. Action: Modify the SCM, by replacing the structural equation for T with T := t
3. Prediction: Use the value of U from step 1 and the modified SCM from step 2 to compute the value of Y(t)

Question:

Given the observation T = 1 and Y = 0, compute Y(0) for this individual given the following SCM:

$$
\begin{aligned}
T &:= \dots \\
Y &:= UT + (1 - U)(1 - T)
\end{aligned}
$$

Can’t Always Determine Counterfactual

Even when we have the structural equation for Y, we can’t always determine counterfactuals with probability 1

Can't Always Determine Counterfactual

Even when we have the structural equation for Y, we can't always determine counterfactuals with probability 1

What if we can't solve for U (function that maps U to Y for a fixed value of T isn't invertible)?

Can't Always Determine Counterfactual

Even when we have the structural equation for Y, we can't always determine counterfactuals with probability 1

What if we can't solve for U (function that maps U to Y for a fixed value of T isn't invertible)?

Example:

Structural equation for Y:

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

Can't Always Determine Counterfactual

Even when we have the structural equation for Y, we can't always determine counterfactuals with probability 1

What if we can't solve for U (function that maps U to Y for a fixed value of T isn't invertible)?

Example:

Observation: T = 1 and Y = 0

Structural equation for Y:

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

Can't Always Determine Counterfactual

Even when we have the structural equation for Y, we can't always determine counterfactuals with probability 1

What if we can't solve for U (function that maps U to Y for a fixed value of T isn't invertible)?

Example:

Observation: T = 1 and Y = 0

Structural equation for Y:

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

Non-Invertible Example

Structural equation for Y:

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

Observation: T = 1 and Y = 0

Non-Invertible Example

Structural equation for Y:

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

Observation: T = 1 and Y = 0

$$(Y_u(1) = 0)$$

Non-Invertible Example

Structural equation for Y:

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

Observation: T = 1 and Y = 0

$$(Y_u(1) = 0)$$

$$Y_u(0) =?$$

Non-Invertible Example

Structural equation for Y:

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

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

Observation: T = 1 and Y = 0

$$(Y_u(1) = 0)$$

$$Y_u(0) = ?$$

Non-Invertible Example

Structural equation for Y:

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

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

Observation: T = 1 and Y = 0

$$(Y_u(1) = 0)$$

$$Y_u(0) = ?$$

Non-Invertible Example

Structural equation for Y:

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

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

Observation: T = 1 and Y = 0

$$(Y_u(1) = 0)$$

$$P(U = \text{never happy} \mid T = 1, Y = 0) = \frac{0.2}{0.2 + 0.1} = \frac{2}{3}$$
$$P(U = \text{dog-hater} \mid T = 1, Y = 0) = \frac{0.1}{0.2 + 0.1} = \frac{1}{3}$$

$$Y_u(0) = ?$$

Non-Invertible Example

Structural equation for Y:

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

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

Observation: T = 1 and Y = 0

$$(Y_u(1) = 0)$$

$$Y_u(0) = ?$$

$$P(U = \text{never happy} \mid T = 1, Y = 0) = \frac{0.2}{0.2 + 0.1} = \frac{2}{3}$$
$$P(U = \text{dog-hater} \mid T = 1, Y = 0) = \frac{0.1}{0.2 + 0.1} = \frac{1}{3}$$

$$P(Y_u(0) = 1) = \frac{1}{3}$$

General Steps for Probabilistic Counterfactuals

General Steps for Probabilistic Counterfactuals

From Chapter 4 of Pearl et al. (2016)'s Primer:

General Steps for Probabilistic Counterfactuals

From Chapter 4 of Pearl et al. (2016)'s Primer:

1. Abduction: Use an observation Z to update the distribution of U: P(U | Z)

General Steps for Probabilistic Counterfactuals

From Chapter 4 of Pearl et al. (2016)'s Primer:

1. Abduction: Use an observation Z to update the distribution of U: P(U | Z)
2. Action: Modify the SCM, by replacing the structural equation for T with T := t

General Steps for Probabilistic Counterfactuals

From Chapter 4 of Pearl et al. (2016)'s Primer:

1. Abduction: Use an observation Z to update the distribution of U: P(U | Z)
2. Action: Modify the SCM, by replacing the structural equation for T with T := t
3. Prediction: Use the the updated distribution of U step 1 and the modified SCM from step 2 to compute the distribution of Y(t)

No Unit-Level Counterfactuals without Parametric Model

Main ingredient necessary for computing counterfactuals: parametric model for the structural equation for Y

No Unit-Level Counterfactuals without Parametric Model

Main ingredient necessary for computing counterfactuals: parametric model for the structural equation for Y

Strong assumption

No Unit-Level Counterfactuals without Parametric Model

Main ingredient necessary for computing counterfactuals: parametric model for the structural equation for Y

Strong assumption

Without it, we are stuck with the fundamental problem of causal inference.

Question:

Given the observation T = 1 and Y = 1, compute Y(0) for this individual given the following SCM and prior:

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

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