# Causal Networks

## Counterfactuals

Fabio Stella

Department of Informatics, Systems and Communication,

University of Milan-Bicocca

Viale Sarca 336, 2016 Milan, ITALY

e-mail: fabio.stella@unimib.it

Twitter: FaSt@FabioAStella

In observational studies, there could also be some unobserved confounder, thus we'd like to know how robust our estimates are to unobserved confounding(s).

Thus, in this lecture we will introduce:

- Observational counterfactual decomposition

- No assumptions bound

- Nonnegative/Nonpositive monotone treatment response lower bound

- Monotone treatment selection upper bound

- Optimal treatment selection

- Sensitivity Analysis

# Part I

# Introduction to Counterfactuals

While driving home yesterday, I came to a fork in the road,

I took surface street ($X = 0$), only to find out that the traffic was touch and go.





where I had to make a choice, take

- the **freeway** ($X = 1$) or
- the **surface street** ($X = 0$).

As I arrived home, an hour later, I said to myself:

"*Gee, I should have taken the freeway.*"

What does it mean to say, "I should have taken the freeway"?

- Colloquially, it means, "If I had taken the freeway, I would have gotten home earlier."

- Scientifically, it means that my mental estimate of the expected driving time on the freeway, on that same day, under the identical circumstances, and governed by the same idiosyncratic driving habits that I have, would have been lower than my actual driving time.

While driving home yesterday, I came to a fork in the road,



(X = 0) SURFACE    FREEWAY (X = 1)

**HYPOTHETICAL CONDITION** or **ANTECEDENT**

The statement "If had taken the freeway, I would have gotten home earlier." where the "if" statement is untrue or unrealized—is known as a **COUNTERFACTUAL**.

Counterfactuals are used to compare two **OUTCOMES** (e.g., driving times) under the exact same conditions, differing only in **ANTECEDENT.**

I took surface street ($X = 0$), only to find out that the traffic was touch and go.



I took freeway ($X = 1$).



driving time = 1h

COMPARE
OUTCOMES

driving time = ?h

The fact that we know the outcome of our actual decision is important, because my estimated driving time on the freeway after seeing the consequences of my actual decision (to take surface street) may be totally different from my estimate prior to seeing the consequence.

Take surface street ($X = 0$) $\rightarrow$ driving time = 1h

⇓

Provides valuable evidence for the assessment, for example, that the traffic was particularly heavy on that day, and that it might have been due to a brush fire.

⇓

My statement "I should have taken the freeway" conveys the judgment that whatever mechanisms impeded my speed on the surface street would not have affected the speed on the freeway to the same extent.

⇓

My retrospective estimate is that a freeway drive would have taken less than 1 hour, and this estimate is clearly different than my prospective estimate was, when I made the decision prior to seeing the consequences—otherwise, I would have taken the freeway to begin with.

I took surface street ($X = 0$), only to find out that the traffic was touch and go.



I took freeway ($X = 1$).

If we try to express this estimate using do-expressions, we come to an impasse. Writing

$$\mathbb{E}[driving\ time | do(freeway), driving\ time = 1\ hour]$$

leads to a clash between the driving time we wish to estimate and the actual driving time observed.

We must distinguish symbolically between the following variables:

- Actual driving time
- Hypothetical driving time under freeway conditions when actual surface driving time is known to be 1 hour.

Unfortunately, the *do*-operator is too crude to make this distinction. The do-operator allows us to distinguish between:

$P(\text{driving time} | do(\text{freeway}))$ and $P(\text{driving time} | do(\text{surface street}))$

it does not offer us the means of distinguishing between the two variables themselves, one standing for the time on surface street, the other for the hypothetical time on the freeway.

We need this distinction in order to let the actual driving time (on surface street) inform our assessment of the hypothetical driving time.

I took surface street ($X = 0$), only to find out that the traffic was touch and go.



I took freeway ($X = 1$).

Fortunately, making this distinction is easy; we simply use different subscripts to label the two outcomes.

$Y(X = 1) = Y(1)$      freeway driving time

$Y(X = 0) = Y(0)$      surface street driving time

In our case $Y = Y(0) = 1\ hour$ is actually observed, then the quantity we wish to estimate is

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

HYPOTHETICAL CONDITION or ANTECEDENT     OBSERVED VARIABLES

$Y(1) = y$ and $X = 0$ are—and must be—events occurring under different conditions, sometimes referred to as "different worlds."

A randomized controlled experiment on the two decision options will never get us the estimate we want, i.e., we can get

$$\mathbb{E}[Y(1)] = \mathbb{E}[Y|do(freeway)] \qquad \mathbb{E}[Y(0)] = \mathbb{E}[Y|do(surface\ street)]$$

But the fact that we cannot take both the freeway and surface street simultaneously prohibits us from estimating the quantity we wish to estimate $\mathbb{E}[Y(1)|X = 0, Y = 1]$.

I took surface street ($X = 0$), only to find out that the traffic was touch and go.



I took freeway ($X = 1$).

Such approximations may be appropriate for estimating the target quantity under some circumstances, but they are not appropriate for defining it.

⇧

In either case, the driving time we would be measuring under such surrogates will only be an approximation of the one we set out to estimate, $Y(1)$, and the degree of approximation would vary with the assumptions we can make on how similar those surrogate conditions are to my own driving time had I taken the freeway.

⇧

One might be tempted to circumvent this difficulty by measuring the freeway time at a later time, or of another driver, but then conditions may change with time, and the other driver may have different driving habits than I.

⇧

But the fact that we cannot take both the freeway and surface street simultaneously prohibits us from estimating the quantity we wish to estimate $\mathbb{E}[Y(1)|X=0, Y=1]$.

I took surface street ($X = 0$), only to find out that the traffic was touch and go.



I took freeway ($X = 1$).

# Part II

# Defining and Computing Counterfactuals

Consider a fully specified model $M = \langle U, V, F \rangle$, for which we know both the functions $F$ and the values of all exogenous variables $U$.

In such a deterministic model,

- every assignment $U = u$ to the exogenous variables corresponds to a single member of, or "unit" in a population, or to a "situation" in nature.

The reason for this correspondence is as follows:

- each assignment $U = u$ uniquely determines the values of all variables in $V$,

- the characteristics (salary, address, education, propensity to engage in musical activity, …) of each individual "unit" in a population have unique values, depending on that individual's identity.

**STRUCTURAL CAUSAL MODEL (SCM)**

A structural causal model is a tuple $M = \langle U, V, F \rangle$ of the following sets:

- $U$; a set of exogenous variables,

- $V$; a set of endogenous variables,

- $F$; a set of functions, one to generate each endogenous variable as a function of other variables.

If

- $U = u$ stands for the defining characteristics of an individual named Joe, and

- $X$ stands for a variable named "salary," then

- $X(u)$ stands for Joe's salary.

Consider now the counterfactual sentence,

"$Y$ would be $y$ had $X$ been $x$, in situation $U = u$"
$(Y_x(u) = y)$

⬇

to be interpreted as an instruction to make a minimal modification in the current model so as to establish the antecedent condition $X = x$, which is likely to conflict with the observed value of $X$, $X(u)$.

⬇

Such a minimal modification amounts to replacing the equation for $X$ with a constant $x$, which may be thought of as an external intervention $do(X = x)$, not necessarily by a human experimenter.

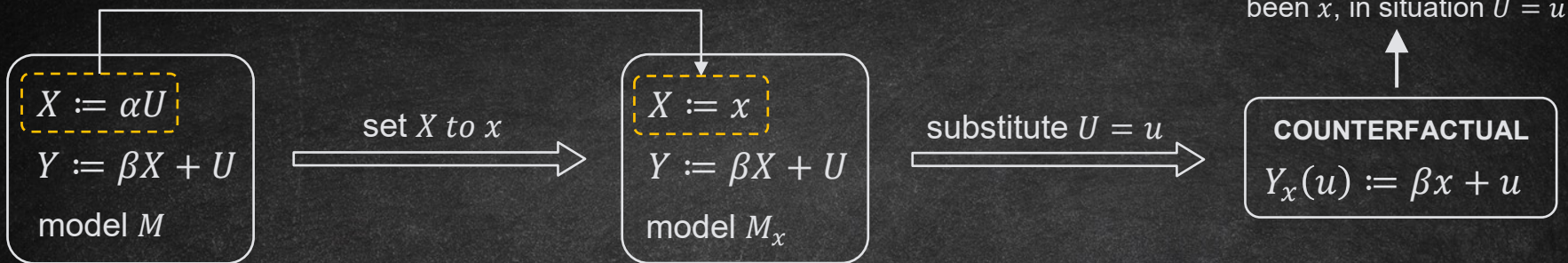⬄

**STRUCTURAL CAUSAL MODEL (SCM)**

A structural causal model is a tuple $M = \langle U, V, F \rangle$ of the following sets:

- $U$; a set of exogenous variables,

- $V$; a set of endogenous variables,

- $F$; a set of functions, one to generate each endogenous variable as a function of other variables.
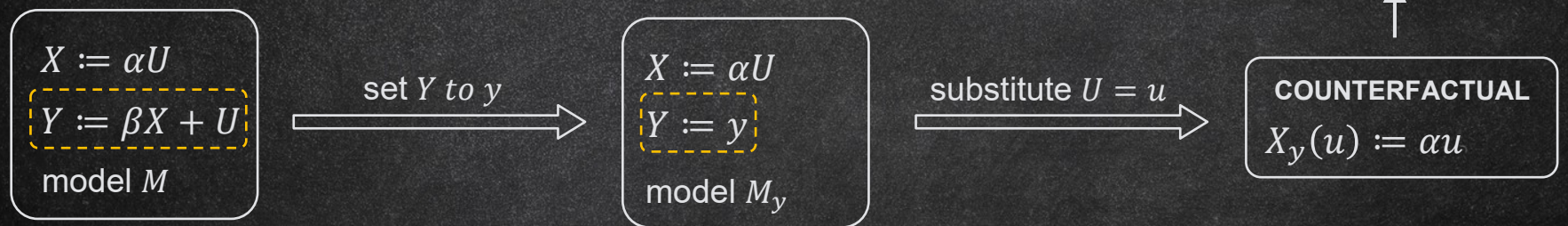
This replacement permits the constant $x$ to differ from the actual value of $X$ (namely, $X(u)$) without rendering the system of equations inconsistent, and in this way, it allows all variables, exogenous as well as endogenous, to serve as antecedents to other variables.

We demonstrate the previous definition on the following simple causal model

what $Y$ would be had $X$
been $x$, in situation $U = u$

$$X := \alpha U$$
$$Y := \beta X + U$$
model $M$

set $X$ to $x$

$$X := x$$
$$Y := \beta X + U$$
model $M_x$

substitute $U = u$

**COUNTERFACTUAL**
$$Y_x(u) := \beta x + u$$

Let us examine the counterfactual $X_y(u)$, that is, what $X$ would be
had $Y$ been $y$ in situation $U = u$.

what $X$ would be had $Y$
been $y$, in situation $U = u$

$$X := \alpha U$$
$$Y := \beta X + U$$
model $M$

set $Y$ to $y$

$$X := \alpha U$$
$$Y := y$$
model $M_y$

substitute $U = u$

**COUNTERFACTUAL**
$$X_y(u) := \alpha u$$

$X$ remains unaltered by the hypothetical condition "had $Y$ been $y$".

Indeed, $X$ happens first than $Y$, thus setting the value of $Y$ (by an intervention) to any given value $y$ can not
change the value of $X$ (because it happened in the past).

Each SCM encodes within it many counterfactuals, corresponding to the various values that its variables can take.

To illustrate additional counterfactuals generated by the model

$$X := \alpha U$$
$$Y := \beta X + U$$

let us assume that $U$ can take on three values, 1, 2, and 3, and let $\alpha = \beta = 1$.

| $u$ | $X(u)$ | $Y(u)$ | $Y_1(u)$ | $Y_2(u)$ | $Y_3(u)$ | $X_1(u)$ | $X_2(u)$ | $X_3(u)$ |
|---|---|---|---|---|---|---|---|---|
| 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 |

Table 14.1

For example, to compute $Y_2(u)$, for $u = 2$,

Every structural equation model assigns a definitive value to every conceivable counterfactual.

$$X := U$$
$$Y := X + U$$

$$X = X(u) = X(2) = 2$$
$$U = u = 2$$

$$X := X(2) = 2$$
$$Y := X + U$$

For each situation $U = u$, we obtained a definite number, $Y_x(u)$, which stands for that hypothetical value of $Y$ in that situation.

individual level

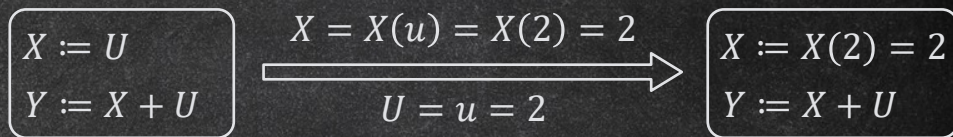Counterfactuals are different than ordinary interventions, captured by the *do*-operator.

The *do*-operator is only defined on probability distributions and always delivers probabilistic results such as $\mathbb{E}[Y|do(x)]$.

population level

$$X(2) := 2$$
$$Y_2(2) := 4$$

We are now ready to generalize the concept of counterfactuals to any structural model, $M$.
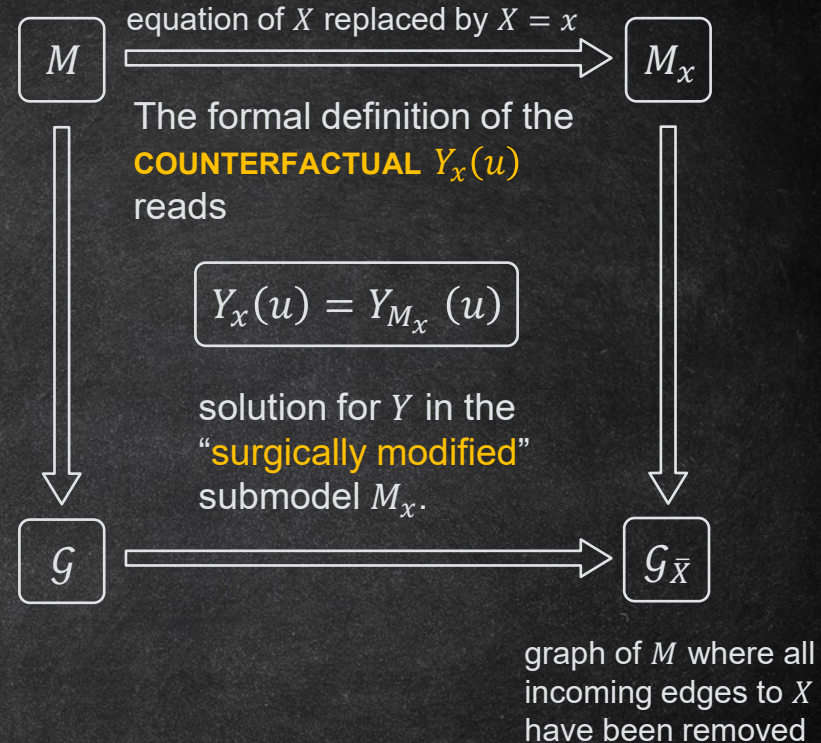
Consider any arbitrary two variables $X$ and $Y$, not necessarily connected by a single equation.

Counterfactuals allow us to take our scientific conception of reality, $M$, and use it to generate answers to an enormous number of hypothetical questions of the type "What would $Y$ be had $X$ been $x$?"

The same definition is applicable when $X$ and $Y$ are sets of variables, if by $M_x$ we mean a model where the equations of all members of $X$ are replaced by constants.

How can a simple model, consisting of just a few equations, assign values to so many counterfactuals?

The answer is that the values that these counterfactuals receive are not totally arbitrary, but must cohere with each other to be consistent with an underlying model.

$M$ — equation of $X$ replaced by $X = x$ → $M_x$

The formal definition of the **COUNTERFACTUAL** $Y_x(u)$ reads

$$Y_x(u) = Y_{M_x}(u)$$

solution for $Y$ in the "surgically modified" submodel $M_x$.

$\mathcal{G}$ → $\mathcal{G}_{\bar{X}}$

graph of $M$ where all incoming edges to $X$ have been removed

If we observe $X(u) = 1$ and $Y(u) = 0$, then

$$Y_{X=1}(u) = 0$$

because setting $X$ to a value it already has, $X(u)$, should produce no change in the world.

Hence, $Y$ should stay at its current value of $Y(u) = 0$.

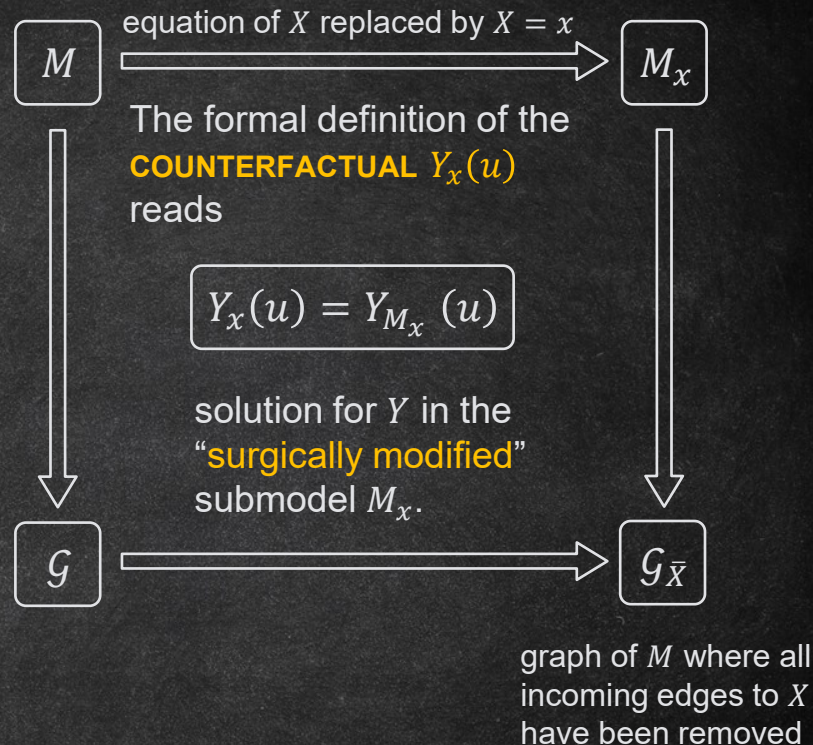In general, counterfactuals obey the following consistency rule:

**COUNTERFACTUAL CONSISTENCY RULE**

If $X = x$   then   $Y_X = Y$

If $X$ is binary then we have the convenient form

$Y = XY_1 + (1 - X)Y_0$

- $Y_1$ is equal to the observed value of $Y$ whenever $X$ takes the value 1.

- $Y_0$ is equal to the observed value of $Y$ whenever $X$ is 0.

$M$  →  equation of $X$ replaced by $X = x$  →  $M_x$

The formal definition of the **COUNTERFACTUAL** $Y_x(u)$ reads

$$Y_x(u) = Y_{M_x}(u)$$

solution for $Y$ in the "surgically modified" submodel $M_x$.

$\mathcal{G}$  →  $\mathcal{G}_{\bar{X}}$

graph of $M$ where all incoming edges to $X$ have been removed

**ENCOURAGEMENT DESIGN** (Figure 14.1).

- $X$; amount of time a student spends in an after-school remedial program,

- $H$; the amount of homework a student does, and

- $Y$; a student's score on the exam.

For example, if $Y = 1$, then the student scored 1 standard deviation above the mean on his or her exam.

This model represents a randomized pilot program, in which students are assigned to the remedial sessions by the luck of the draw.
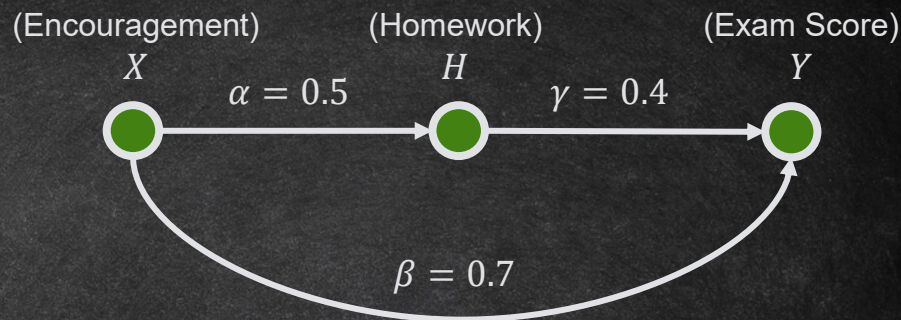


Figure 14.1

$$X := U_X$$

$$H := \alpha X + U_H$$

$$Y := \beta X + \gamma H + U_Y$$

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

$$\alpha = 0.5$$
$$\beta = 0.7$$
$$\gamma = 0.4$$

given or recovered from population data

**ENCOURAGEMENT DESIGN** (Figure 14.1).

- $X$; amount of time a student spends in an after-school remedial program,

- $H$; the amount of homework a student does, and

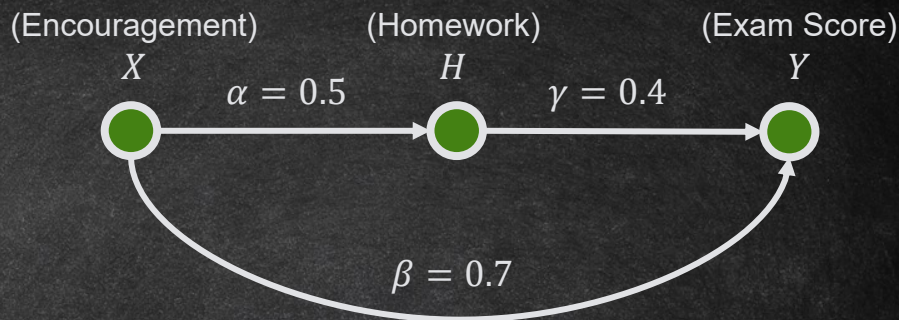- $Y$; a student's score on the exam.

(Encouragement) $X$    $\alpha = 0.5$    (Homework) $H$    $\gamma = 0.4$    (Exam Score) $Y$

$\beta = 0.7$

Figure 14.1

student Joe

$X = 0.5$
$H = 1.0$
$Y = 1.5$

What would Joe's score have been had he doubled his study time?

$U_X = 0.5$
$U_H = H - 0.5X = 0.75$
$U_Y = Y - 0.7X - 0.4H = 0.75$

$$X := U_X$$
$$H := \alpha X + U_H$$
$$Y := \beta X + \gamma H + U_Y$$
$$\sigma_{U_i U_j} = 0, \forall i, j \in \{X, H, Y\}$$

$\alpha = 0.5$
$\beta = 0.7$
$\gamma = 0.4$

given or recovered from population data

Then, we simulate the action of doubling Joe's study time by replacing the structural equation for $H$ with the constant $H = 2$.

**ENCOURAGEMENT DESIGN** (Figure 14.1).

- $X$; amount of time a student spends in an after-school remedial program,

- $H$; the amount of homework a student does, and

- $Y$; a student's score on the exam.

$X = 0.5$
$H = 1.0$
$Y = 1.5$

**What would Joe's score have been had he doubled his study time?**

student Joe

(Encouragement)
$X$

(Homework)
$H$

(Exam Score)
$Y$

$\gamma = 0.4$

$\beta = 0.7$

Figure 14.2

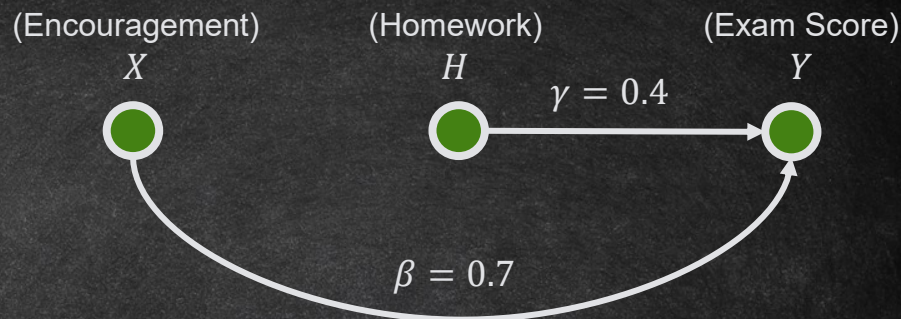Then, we simulate the action of doubling Joe's study time by replacing the structural equation for $H$ with the constant $H = 2$.

$$X := U_X$$

$$H := 2$$

$$Y := \beta X + \gamma H + U_Y$$

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

$\alpha = 0.5$
$\beta = 0.7$
$\gamma = 0.4$

given or recovered from population data

**ENCOURAGEMENT DESIGN** (Figure 14.1).

- $X$; amount of time a student spends in an after-school remedial program,

- $H$; the amount of homework a student does, and

- $Y$; a student's score on the exam.



$X = 0.5$
$H = 1.0$
$Y = 1.5$

What would Joe's score have been had he doubled his study time?

student Joe

Figure 14.2

Finally, we compute the value of $Y$ in our modified model (Figure 14.2) using the updated $U$ values, giving

$$X := U_X$$
$$H := 2$$
$$Y := \beta X + \gamma H + U_Y$$
$$\sigma_{U_i U_j} = 0, \forall i, j \in \{X, H, Y\}$$

$$\alpha = 0.5$$
$$\beta = 0.7$$
$$\gamma = 0.4$$

given or recovered from population data

$$Y_{H=2}(U_X = 0.5, U_X = 0.75, U_X = 0.75) := 0.5\ 0.7 + 0.4\ 2 + 0.75$$

$$:= 1.9$$

## ENCOURAGEMENT DESIGN (Figure 14.1).

- $X$; amount of time a student spends in an after-school remedial program,

- $H$; the amount of homework a student does, and

- $Y$; a student's score on the exam.



(Encouragement) $X$     (Homework) $H$     (Exam Score) $Y$

$\gamma = 0.4$

$\beta = 0.7$

Figure 14.2

$X = 0.5$
$H = 1.0$
$Y = 1.5$

student
Joe

What would Joe's score have been had he doubled his study time?

Joe's score, had he doubled his homework

$$H = 1 \rightarrow H = 2,$$

would have been 1.9 instead of 1.5. (An increase to 1.9 stdv above the mean, instead of the current 1.5)
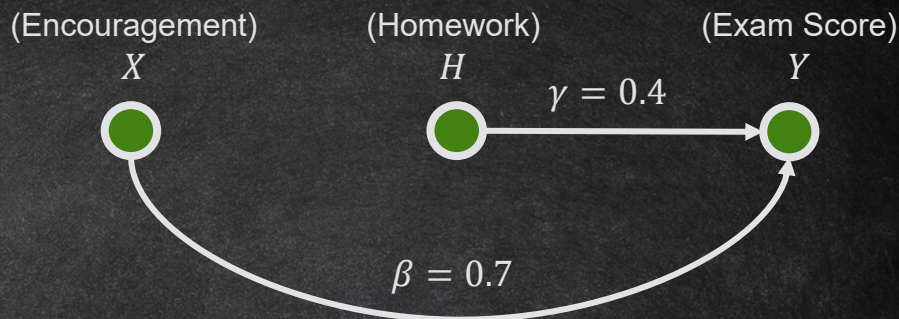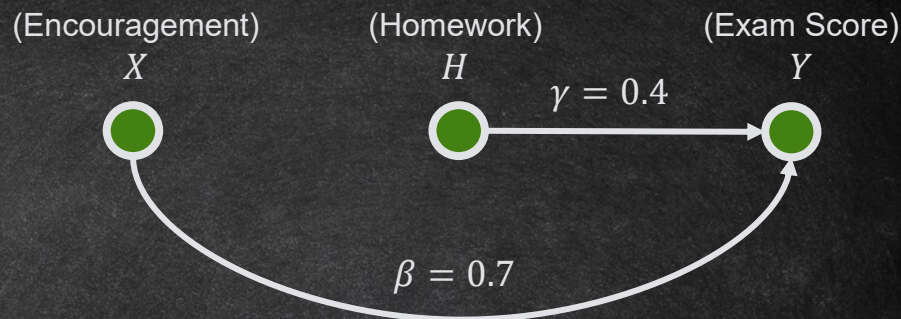
$$X := U_X$$
$$H := 2$$
$$Y := \beta X + \gamma H + U_Y$$
$$\sigma_{U_i U_j} = 0, \forall i, j \in \{X, H, Y\}$$

$\alpha = 0.5$
$\beta = 0.7$
$\gamma = 0.4$

given or recovered from population data

$$Y_{H=2}(U_X = 0.5, U_X = 0.75, U_X = 0.75) := 0.5 \, 0.7 + 0.4 \, 2 + 0.75$$

$$:= 1.9$$

Any deterministic counterfactual is computed by the following steps:       **Temporal Metaphors**

- **ABDUCTION:** use evidence $E = e$ to determine the value of $U$.    ⟹    explains the past $(U)$ in light of the current evidence $e$

- **ACTION:** modify the model, $M$, by removing the structural equations for the variables in $X$ and replacing them with the appropriate functions $X = x$, to obtain the modified model, $M_x$.    ⟹    bends the course of history (minimally) to comply with the hypothetical antecedent $X = x$

- **PREDICTION:** use the modified model, $M_x$, and the value of $U$ to compute the value of $Y$, the consequence of the counterfactual.    ⟹    predicts the future $(Y)$ based on our new understanding of the past and our newly established condition, $X = x$

The three steps will solve any **DETERMINISTIC COUNTERFACTUAL**, that is, counterfactuals pertaining to a single unit of the population in which we know the value of every relevant variable.
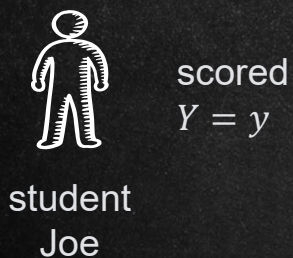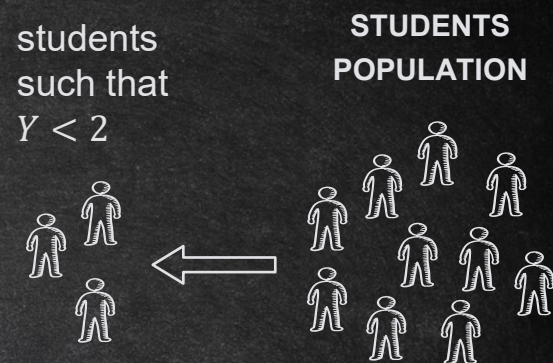
Structural equation models are able to answer counterfactual queries of this nature because each equation represents the mechanism by which a variable obtains its values.

If we know these mechanisms, we should also be able to predict what values would be obtained had some of these mechanisms been altered, given the alterations.

Counterfactuals can also be probabilistic, pertaining to a class of units within the population;

▪ we might want to know what would have happened if all students for whom $Y < 2$ had doubled their homework time ($H = 1 \rightarrow H = 2$).

These probabilistic counterfactuals differ from *do*-operator interventions because, like their deterministic counterparts, they restrict the set of individuals intervened upon, which *do*-expressions cannot do.

students such that $Y < 2$

**STUDENTS POPULATION**

| deterministic counterfactual | move to → | probabilistic counterfactual | allows to → | ask questions about probabilities and expectations of counterfactuals |

**student Joe**

scored $Y = y$

▪ What is the probability that Joe's score would be $Y = y'$ had he had five more hours of encouragement training ($H = 1 \rightarrow H = 6$)?

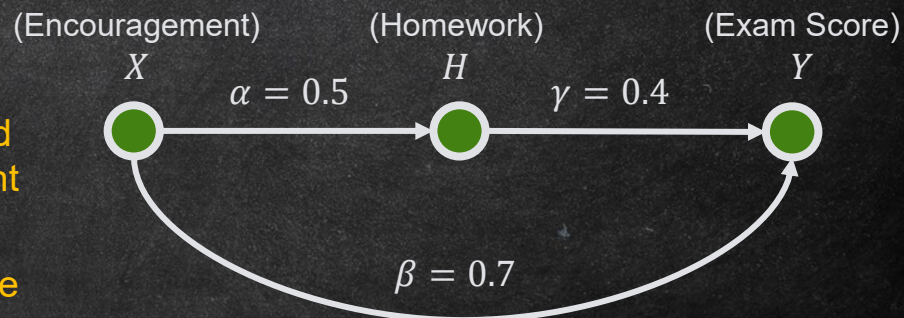▪ What would his expected score be in such hypothetical world?

(Encouragement) $X$    $\alpha = 0.5$    (Homework) $H$    $\gamma = 0.4$    (Exam Score) $Y$

$\beta = 0.7$

Figure 14.1

$$X := U_X$$
$$H := 2$$
$$Y := \beta X + \gamma H + U_Y$$
$$\sigma_{U_i U_j} = 0, \forall i, j \in \{X, H, Y\}$$

Unlike in the example of the previous model

(Encouragement)   (Homework)   (Exam Score)
$X$   $H$   $Y$

$\alpha = 0.5$   $\gamma = 0.4$

$\beta = 0.7$

Figure 14.1

we do not have information on all three variables, $\{X, H, Y\}$, and we cannot therefore determine uniquely the value $u$ that pertains to Joe.
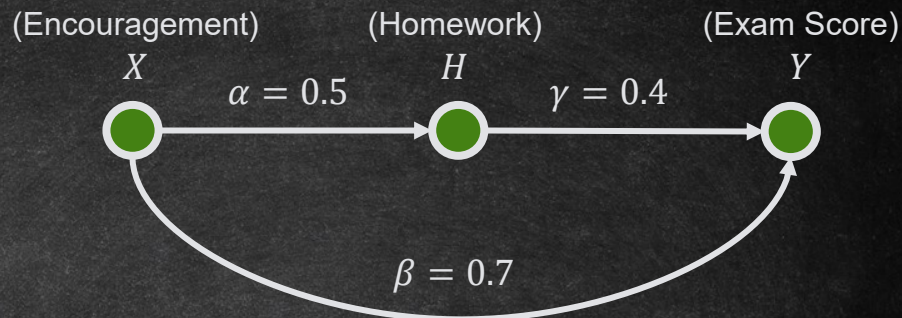
**STUDENTS POPULATION**

class of units compatible with the evidence available

student Joe

Nondeterminism enters causal models by assigning probabilities $P(U = u)$ over the exogenous variables $U$.

These represent our uncertainty as to the identity of the subject under consideration or, when the subject is known, what other characteristics that subject has that might have bearing on our problem.

Instead, Joe may belong to a large class of units compatible with the evidence available, each having a different value of $u$.

exogenous probability $P(U = u)$

induces a unique probability distribution on the endogenous variables $V$

$P(V = v)$

**STRUCTURAL CAUSAL MODEL (SCM)**

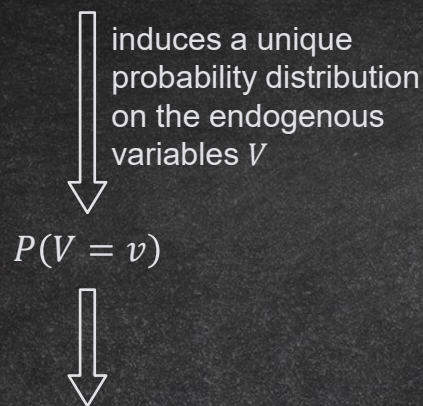A structural causal model is a tuple $M = \langle U, V, F \rangle$ of the following sets:

- $U$; a set of exogenous variables,
- $V$; a set of endogenous variables,
- $F$; a set of functions, one to generate each endogenous variable as a function of other variables.

It allows to define and compute not only the probability of any single counterfactual, $Y_X = y$, but also the joint distributions of all combinations of observed and counterfactual variables

typical query

"Given that we observe feature $E = e$ for a given individual, what would we expect the value of $Y$ for that individual to be if $X$ had been $x$?"

$E = e$ after the conditioning bar represents all information (or evidence) we might have about the individual, potentially including the values of $X$, $Y$, or any other variable.

The subscript $X = x$ in $Y_{X=x}$ represents the antecedent specified by the counterfactual sentence.

This expectation is denoted

$$\mathbb{E}[Y_{X=x} | E = e],$$

where we allow $E = e$ to conflict with the antecedent $X = x$.

Given an arbitrary counterfactuals of the form,

$$\mathbb{E}[Y_{X=x}|E=e]$$

the three-step process for **PROBABILISTIC COUNTERFACTUAL** reads:

- **ABDUCTION:** Update $P(U)$ by the evidence to obtain $P(U|E=e)$.

- **ACTION:** modify the model, $M$, by removing the structural equations for the variables in $X$ and replacing them with the appropriate functions $X=x$, to obtain the modified model, $M_x$.

- **PREDICTION:** use the modified model, $M_x$, and the updated probabilities over the $U$ variables, $P(U|E=e)$, to compute the expectation of $Y$, the consequence of the counterfactual.

Suppose we wish to estimate, using Figure 14.1, the effect on test score ($Y$) provided by a school policy that sends students who are lazy on their homework ($H \leq H_0$) to attend the after-school program for $X=1$.

We can't simply intervene on $X$ to set it equal to 1 in cases where $H$ is low, because in our model, $X$ is one of the causes of $H$.



(Encouragement)
$X$

(Homework)
$H$

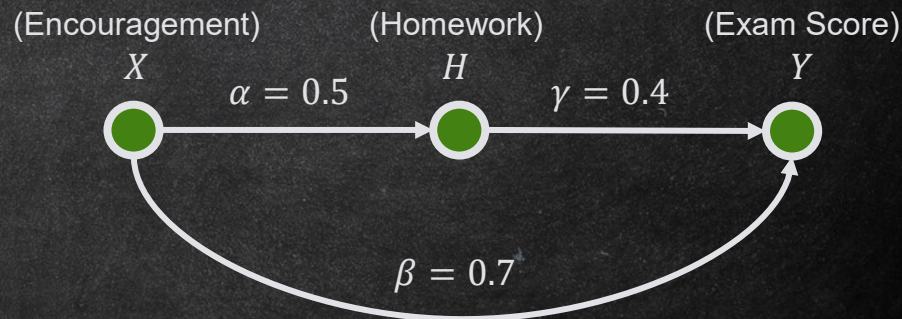(Exam Score)
$Y$

$\alpha = 0.5$

$\gamma = 0.4$

$\beta = 0.7$

Figure 14.1

Given an arbitrary counterfactuals of the form,

$$\mathbb{E}[Y_{X=x}|E=e]$$

the three-step process for **PROBABILISTIC COUNTERFACTUAL** reads:

- **ABDUCTION:** Update $P(U)$ by the evidence to obtain $P(U|E=e)$.

- **ACTION:** modify the model, $M$, by removing the structural equations for the variables in $X$ and replacing them with the appropriate functions $X = x$, to obtain the modified model, $M_x$.

- **PREDICTION:** use the modified model, $M_x$, and the updated probabilities over the $U$ variables, $P(U|E=e)$, to compute the expectation of $Y$, the consequence of the counterfactual.

Instead, we express the expected value of this quantity in counterfactual notation as

$$\mathbb{E}[Y_{X=1}|H \leq H_0],$$

which can, in principle, be computed using the above steps.

Counterfactual reasoning and the above procedure are necessary for estimating the effect of actions and policies on subsets of the population characterized by features that, in themselves, are affected by the policy (e.g., $H \leq H_0$).
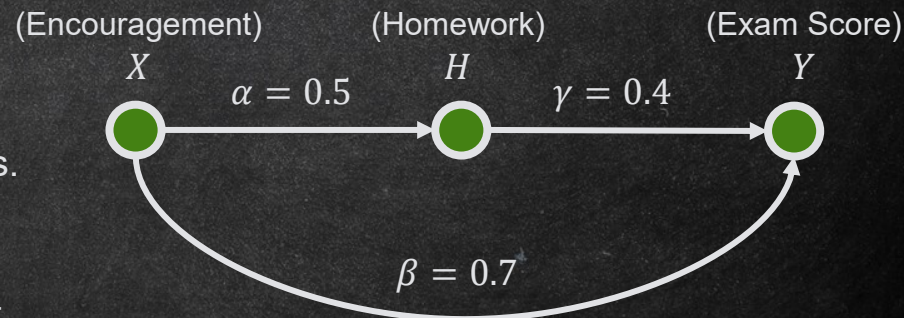


(Encouragement) $X$    $\alpha = 0.5$    (Homework) $H$    $\gamma = 0.4$    (Exam Score) $Y$

$\beta = 0.7$

Figure 14.1

# Part II

# Nondeterministic Counterfactuals

To examine how nondeterminism is reflected in the calculation of counterfactuals, let us assign probabilities to the values of $U$ in the model of to the right.

Imagine that $U = \{1,2,3\}$ represents three types of individuals in a population, occurring with probabilities

$$X := \alpha U$$
$$Y := \beta X + U$$

$$P(U = 1) = \frac{1}{2} \qquad P(U = 2) = \frac{1}{3} \qquad P(U = 3) = \frac{1}{6}$$

| $u$ | $X(u)$ | $Y(u)$ | $Y_1(u)$ | $Y_2(u)$ | $Y_3(u)$ | $X_1(u)$ | $X_2(u)$ | $X_3(u)$ |
|---|---|---|---|---|---|---|---|---|
| 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 |

Table 14.1

All individuals within a population type have the same values of the counterfactuals, as specified by the corresponding rows in Table 14.1.

We can compute the proportion of units for which $Y$ would be 3 had $X$ been 2, or $Y_2(u) = 3$. $\Longrightarrow$ This condition occurs only in the first row of Table 14.1 and, since it is a property of $U = 1$, we conclude that it will occur with probability

$$P(Y_2 = 3) = P(U = 1) = \frac{1}{2}$$

However, we can also compute joint probabilities of every combination of counterfactual and observable events. For example,

$$P(Y_2 > 3, Y_1 < 4)$$

$Y_2 > 3$ in the $X = 2$ world

$Y_1 < 4$ in the $X = 1$ world

$\Longrightarrow u = 2, P(Y_2 > 3, Y_1 < 4) = \frac{1}{3}$

Cross-world probabilities are as simple to derive as intra-world ones:

$$X := \alpha U$$
$$Y := \beta X + U$$

- we simply identify the rows in which the specified combination is true and sum up the probabilities assigned to those rows.

- This allows us to compute conditional probabilities among counterfactuals and defining notions such as dependence and conditional independence among counterfactuals.

| $u$ | $X(u)$ | $Y(u)$ | $Y_1(u)$ | $Y_2(u)$ | $Y_3(u)$ | $X_1(u)$ | $X_2(u)$ | $X_3(u)$ |
|---|---|---|---|---|---|---|---|---|
| 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 |

Table 14.1

Joint probabilities over multiple-world counterfactuals

$$P(Y_1 = y_1, Y_2 = y_2)$$

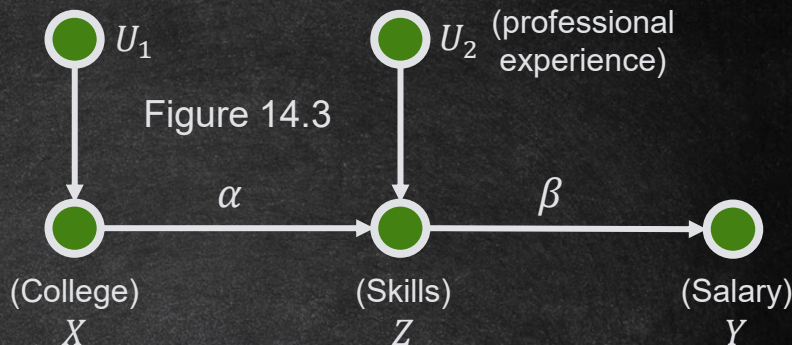can be computed from any structural model as we did in Table 14.1.

They cannot however be expressed using the $do(x)$ notation, because the latter delivers just one probability for each intervention $X = x$.

Consider the model to the right.

Let

- $X = 1$, having a college education,
- $U_2 = 1$ having professional experience,
- $Z$; level of skill needed for a given job,
- $Y$; salary.

$$X := U_1$$
$$Z := \alpha X + U_2$$
$$Y := \beta Z$$

$U_1$

$U_2$ (professional experience)

Figure 14.3

$\alpha$                    $\beta$

(College)                (Skills)                (Salary)
$X$                      $Z$                      $Y$

Which is expected salary of individuals with skill level $Z = 1$, had they received a college education $X = 1$ ? $\Longrightarrow$ $\mathbb{E}[Y_{X=1}|Z = 1]$ $\Longleftarrow$

Not possible to use the *do*-expression to compute it, because

The *do*-expression attempt to capture this hypothetical salary

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

would not reveal the desired information.

The *do*-expression stands for the expected salary of individuals who all finished college and have since acquired skill level $Z = 1$.

- condition $Z = 1$ represents current skills
- antecedent $X = 1$ represents a hypothetical education in an unrealized past
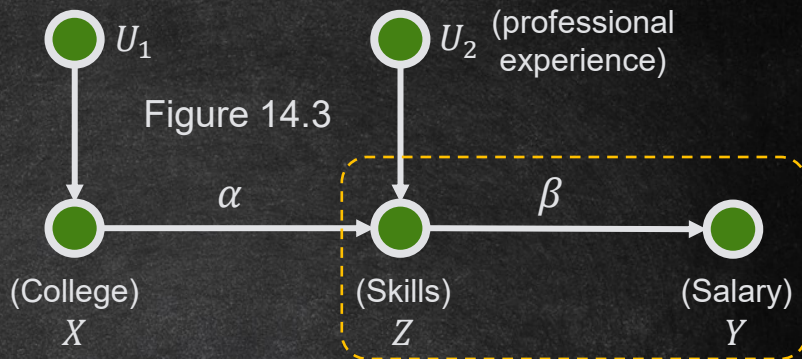
Indeed, $Z = 1$ and $X = 1$ refer to two different worlds.

Consider the model to the right.

Let

- $X = 1$, having a college education,
- $U_1 = 1$ having professional experience,
- $Z$; level of skill needed for a given job,
- $Y$; salary.

$$X := U_1$$
$$Z := \alpha X + U_2$$
$$Y := \beta Z$$



Figure 14.3

**Which is expected salary of individuals with skill level $Z = 1$, had they received a college education $X = 1$ ?** $\Longrightarrow$ $\mathbb{E}[Y_{X=1}|Z = 1]$

The *do*-expression attempt to capture this hypothetical salary

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

would not reveal the desired information.

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 graph shows, depend only on their skill, and are not affected by whether they obtained the skill through college or through work experience.

Conditioning on $Z = 1$, in this case, cuts off the effect of the intervention that we're interested in.

In contrast, some of those who currently have $Z = 1$ might not have gone to college and would have attained higher skill (and salary) had they gotten college education.

Consider the model to the right.

Let

- $X = 1$, having a college education,
- $U_1 = 1$ having professional experience,
- $Z$; level of skill needed for a given job,
- $Y$; salary.

$$X := U_1$$
$$Z := \alpha X + U_2$$
$$Y := \beta Z$$



Figure 14.3

Which is expected salary of individuals with skill level $Z = 1$, had they received a college education $X = 1$ ? $\Longrightarrow$ $\mathbb{E}[Y_{X=1}|Z=1]$

Their salaries are of great interest to us, but they are not included in the *do*-expression.

Thus, in general, the *do*-expression will not capture our counterfactual question:

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

The salaries of these individuals, as the graph shows, depend only on their skill, and are not affected by whether they obtained the skill through college or through work experience.

Conditioning on $Z = 1$, in this case, cuts off the effect of the intervention that we're interested in.
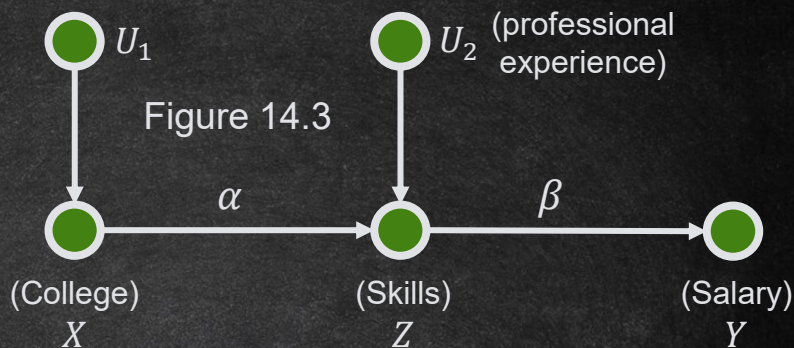
In contrast, some of those who currently have $Z = 1$ might not have gone to college and would have attained higher skill (and salary) had they gotten college education.

Consider the model to the right.

Let

- $X = 1$, having a college education,
- $U_1 = 1$ having professional experience,
- $Z$; level of skill needed for a given job,
- $Y$; salary.

$$X := U_1$$
$$Z := \alpha X + U_2$$
$$Y := \beta Z$$



Figure 14.3

$U_1$

$U_2$ (professional experience)

$\alpha$

$\beta$

(College) $X$

(Skills) $Z$

(Salary) $Y$

**Which is expected salary of individuals with skill level $Z = 1$, had they received a college education $X = 1$ ?** $\Longrightarrow$ $\mathbb{E}[Y_{X=1}|Z = 1]$

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

treat $Z = 1$ as a **POSTINTERVENTION** condition that prevails for two different sets of units under the two antecedents

$$\mathbb{E}[Y_{X=1}|Z = 1] \neq \mathbb{E}[Y_{X=0}|Z = 1] \Longrightarrow$$

treat $Z = 1$ as defining one set of units in the current world that would react differently under the two antecedents

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

Consider the model to the right.

Let

- $X = 1$, having a college education,
- $U_1 = 1$ having professional experience,
- $Z$; level of skill needed for a given job,
- $Y$; salary.

$$X := U_1$$
$$Z := \alpha X + U_2$$
$$Y := \beta Z$$



Figure 14.3

$U_2$ (professional experience)

(College) $X$ — $\alpha$ → (Skills) $Z$ — $\beta$ → (Salary) $Y$
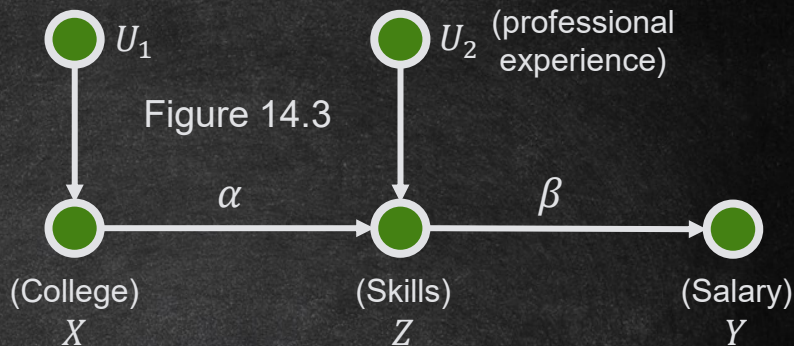
Which is expected salary of individuals with skill level $Z = 1$, had they received a college education $X = 1$ ? $\implies$ $\mathbb{E}[Y_{X=1}|Z=1]$

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

$$\mathbb{E}[Y_{X=1}|Z=1] \neq \mathbb{E}[Y_{X=0}|Z=1] \Longleftarrow$$

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

The expression $\mathbb{E}[Y|do(X=1), Z=1]$ on the other hand, invokes only postintervention events, and that is why it is expressible in $do(x)$ notation. $\Uparrow$

$do(x)$ cannot capture this difference, because

- $X = 1$ refers to preintervention world,
- $Z = 1$ refers to postintervention world.

Consider the model to the right.

Let

- $X = 1$, having a college education,
- $U_1 = 1$ having professional experience,
- $Z$; level of skill needed for a given job,
- $Y$; salary.

$$X := U_1$$
$$Z := \alpha X + U_2$$
$$Y := \beta Z$$



Figure 14.3

(College)        (Skills)        (Salary)
$X$              $Z$             $Y$

Which is expected salary of individuals with skill level $Z = 1$, had they received a college education $X = 1$ ? $\implies$ $\mathbb{E}[Y_{X=1}|Z = 1]$

Does the counterfactual notation capture the postintervention, single-world expression $\mathbb{E}[Y|do(X = 1), Z = 1]$?

$\Downarrow$ yes
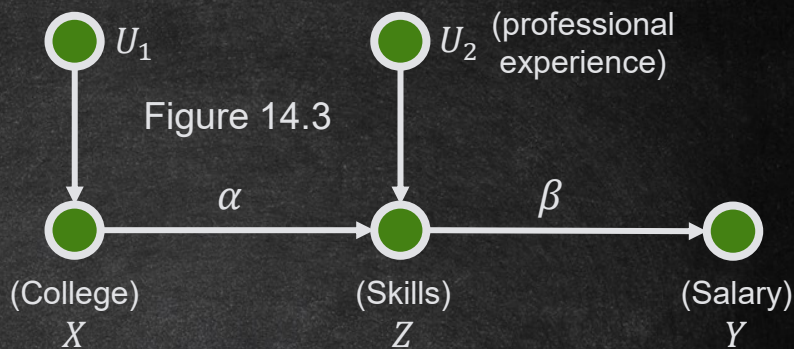
$$\mathbb{E}[Y|do(X = 1), Z = 1] = \mathbb{E}[Y_{X=1}|Z_{X=1} = 1]$$

This shows explicitly how the dependence of $Z$ on $X$ should be treated.

- $Z_{X=1}$; value that $Z$ would attain had $X$ been 1, and this is precisely what we mean when we put $Z = z$ in a *do*-expression by Bayes' rule.

$$P(Y = y|do(X = 1), Z = z) = \frac{P(Y = y, Z = z|do(X = 1))}{P(Z = z|do(X = 1))}$$

Table 14.2 depicts the counterfactuals associated with the model to the right, with all subscripts denoting the state of $X$.

$$X := U_1$$
$$Z := \alpha X + U_2$$
$$Y := \beta Z$$

| $U_1$ | $U_2$ | $X(u)$ | $Z(u)$ | $Y(u)$ | $Y_0(u)$ | $Y_1(u)$ | $Z_0(u)$ | $Z_1(u)$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | $\alpha\beta$ | 0 | $\alpha$ |
| 0 | 1 | 0 | 1 | $\beta$ | $\beta$ | $(\alpha+1)\beta$ | 1 | $\alpha+1$ |
| 1 | 0 | 1 | $\alpha$ | $\alpha\beta$ | 0 | $\alpha\beta$ | 0 | $\alpha$ |
| 1 | 1 | 1 | $\alpha+1$ | $(\alpha+1)\beta$ | $\beta$ | $(\alpha+1)\beta$ | 1 | $\alpha+1$ |

Table 14.2

Table 14.2 is obtained by the same method we used in constructing Table 14.1: replacing the equation $X = u$ with the appropriate constant (0 or 1) and solving for $Y$ and $Z$.

$U_1 = 0 \implies$ $\begin{aligned} X &:= 0 \\ Z &:= U_2 \end{aligned}$ $\quad U_2 = 0 \implies Y := 0$

$\quad U_2 = 1 \implies Y := \beta$

$U_1 = 1 \implies$ $\begin{aligned} X &:= 1 \\ Z &:= \alpha + U_2 \end{aligned}$ $\quad U_2 = 0 \implies Y := \alpha\beta$

$\quad U_2 = 1 \implies Y := (\alpha+1)\beta$

Table 14.2 depicts the counterfactuals associated with the model to the right, with all subscripts denoting the state of $X$.

$$X := U_1$$
$$Z := \alpha X + U_2$$
$$Y := \beta Z$$

| $U_1$ | $U_2$ | $X(u)$ | $Z(u)$ | $Y(u)$ | $Y_0(u)$ | $Y_1(u)$ | $Z_0(u)$ | $Z_1(u)$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | $\alpha\beta$ | 0 | $\alpha$ |
| 0 | 1 | 0 | 1 | $\beta$ | $\beta$ | $(\alpha+1)\beta$ | 1 | $\alpha+1$ |
| 1 | 0 | 1 | $\alpha$ | $\alpha\beta$ | 0 | $\alpha\beta$ | 0 | $\alpha$ |
| 1 | 1 | 1 | $\alpha+1$ | $(\alpha+1)\beta$ | $\beta$ | $(\alpha+1)\beta$ | 1 | $\alpha+1$ |

Table 14.2

Table 14.2 is obtained by the same method we used in constructing Table 14.1: replacing the equation $X = u$ with the appropriate constant (0 or 1) and solving for $Y$ and $Z$.

Using Table 14.2, we can verify that

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

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

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



Figure 14.3

Table 14.2 depicts the counterfactuals associated with the model to the right, with all subscripts denoting the state of $X$.

$$X := U_1$$
$$Z := \alpha X + U_2$$
$$Y := \beta Z$$

| $U_1$ | $U_2$ | $X(u)$ | $Z(u)$ | $Y(u)$ | $Y_0(u)$ | $Y_1(u)$ | $Z_0(u)$ | $Z_1(u)$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | $\alpha\beta$ | 0 | $\alpha$ |
| 0 | 1 | 0 | 1 | $\beta$ | $\beta$ | $(\alpha+1)\beta$ | 1 | $\alpha+1$ |
| 1 | 0 | 1 | $\alpha$ | $\alpha\beta$ | 0 | $\alpha\beta$ | 0 | $\alpha$ |
| 1 | 1 | 1 | $\alpha+1$ | $(\alpha+1)\beta$ | $\beta$ | $(\alpha+1)\beta$ | 1 | $\alpha+1$ |

Table 14.2

Table 14.2 is obtained by the same method we used in constructing Table 14.1: replacing the equation $X = u$ with the appropriate constant (0 or 1) and solving for $Y$ and $Z$.

Using Table 14.2, we can verify that

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

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

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

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



Figure 14.3

$do(X=0) \rightarrow$

$Z = 1$

$\alpha$     $\beta$

$Y = \beta$

(College)    (Skills)    (Salary)
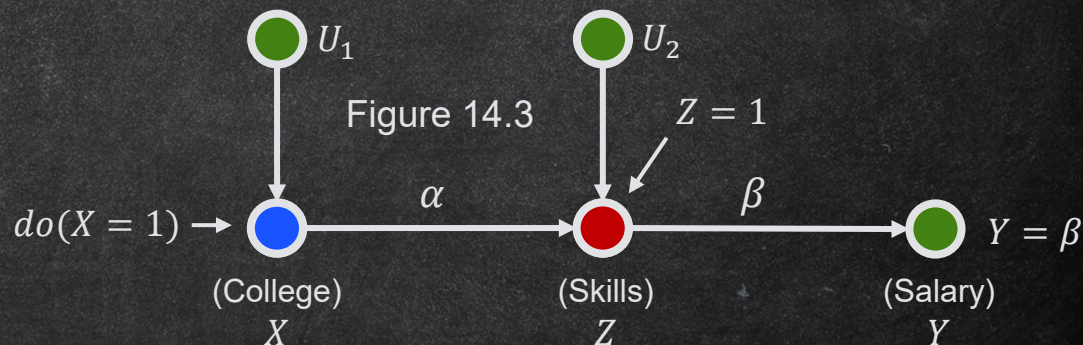$X$      $Z$      $Y$

Table 14.2 depicts the counterfactuals associated with the model to the right, with all subscripts denoting the state of $X$.

$$X := U_1$$
$$Z := \alpha X + U_2$$
$$Y := \beta Z$$

| $U_1$ | $U_2$ | $X(u)$ | $Z(u)$ | $Y(u)$ | $Y_0(u)$ | $Y_1(u)$ | $Z_0(u)$ | $Z_1(u)$ |
|-------|-------|--------|--------|--------|----------|----------|----------|----------|
| 0 | 0 | 0 | 0 | 0 | 0 | $\alpha\beta$ | 0 | $\alpha$ |
| 0 | 1 | 0 | 1 | $\beta$ | $\beta$ | $(\alpha+1)\beta$ | 1 | $\alpha+1$ |
| 1 | 0 | 1 | $\alpha$ | $\alpha\beta$ | 0 | $\alpha\beta$ | 0 | $\alpha$ |
| 1 | 1 | 1 | $\alpha+1$ | $(\alpha+1)\beta$ | $\beta$ | $(\alpha+1)\beta$ | 1 | $\alpha+1$ |

Table 14.2

Table 14.2 is obtained by the same method we used in constructing Table 14.1: replacing the equation $X = u$ with the appropriate constant (0 or 1) and solving for $Y$ and $Z$.

Using Table 14.2, we can verify that

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

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

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

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

These equations provide numerical confirmation of the inequality to the right

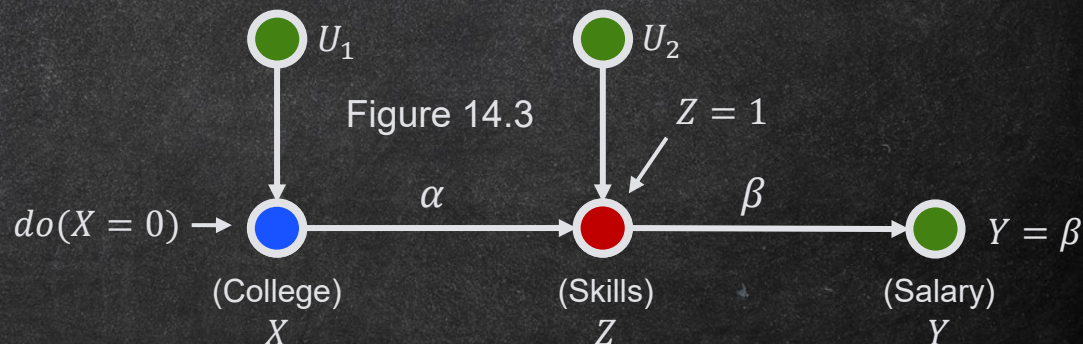$$\mathbb{E}[Y|do(X=1), Z=1] \neq \mathbb{E}[Y_{X=1}|Z=1]$$

Table 14.2 depicts the counterfactuals associated with the model to the right, with all subscripts denoting the state of $X$.

$$X := U_1$$
$$Z := \alpha X + U_2$$
$$Y := \beta Z$$

| $U_1$ | $U_2$ | $X(u)$ | $Z(u)$ | $Y(u)$ | $Y_0(u)$ | $Y_1(u)$ | $Z_0(u)$ | $Z_1(u)$ |
|-------|-------|--------|--------|--------|----------|----------|----------|----------|
| 0 | 0 | 0 | 0 | 0 | 0 | $\alpha\beta$ | 0 | $\alpha$ |
| 0 | 1 | 0 | 1 | $\beta$ | $\beta$ | $(\alpha+1)\beta$ | 1 | $\alpha+1$ |
| 1 | 0 | 1 | $\alpha$ | $\alpha\beta$ | 0 | $\alpha\beta$ | 0 | $\alpha$ |
| 1 | 1 | 1 | $\alpha+1$ | $(\alpha+1)\beta$ | $\beta$ | $(\alpha+1)\beta$ | 1 | $\alpha+1$ |

Table 14.2

Table 14.2 is obtained by the same method we used in constructing Table 14.1: replacing the equation $X = u$ with the appropriate constant (0 or 1) and solving for $Y$ and $Z$.

Using Table 14.2, we can verify that

$X$ affects $Y$

$$\mathbb{E}[Y_1|Z = 1] = (\alpha + 1)\beta$$
$$\mathbb{E}[Y_0|Z = 1] = \beta$$

$$\mathbb{E}[Y_1 - Y_0|Z = 1] = (\alpha + 1)\beta - \beta = \alpha\beta \neq 0$$

The equations also demonstrate a peculiar property of counterfactual conditioning.

Despite the fact that $Z$ separates $X$ from $Y$ in the graph of Figure 14.3, we find that $X$ has an effect on $Y$ for those units falling under $Z = 1$.

Figure 14.3

separates $X$ from $Y$

$U_1$   $U_2$

$\alpha$   $\beta$

(College)   (Skills)   (Salary)
$X$   $Z$   $Y$

$$X := U_1$$
$$Z := \alpha X + U_2$$
$$Y := \beta Z$$

| $U_1$ | $U_2$ | $X(u)$ | $Z(u)$ | $Y(u)$ | $Y_0(u)$ | $Y_1(u)$ | $Z_0(u)$ | $Z_1(u)$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | $\alpha\beta$ | 0 | $\alpha$ |
| 0 | 1 | 0 | 1 | $\beta$ | $\beta$ | $(\alpha+1)\beta$ | 1 | $\alpha+1$ |
| 1 | 0 | 1 | $\alpha$ | $\alpha\beta$ | 0 | $\alpha\beta$ | 0 | $\alpha$ |
| 1 | 1 | 1 | $\alpha+1$ | $(\alpha+1)\beta$ | $\beta$ | $(\alpha+1)\beta$ | 1 | $\alpha+1$ |

Table 14.2

The equations also demonstrate a peculiar property of counterfactual conditioning.

The reason for this behavior is best explained in the context of our salary ⟸ example.

- the salary of those who have acquired skill level $Z = 1$ depends only on their skill, not on $X$,

- the salary of those who are currently at $Z = 1$ would have been different had they had a different past.

Despite the fact that $Z$ separates $X$ from $Y$ in the graph of Figure 14.3, we find that $X$ has an effect on $Y$ for those units falling under $Z = 1$.

Figure 14.3



separates $X$ from $Y$

(College)   (Skills)   (Salary)
$X$         $Z$        $Y$

| $U_1$ | $U_2$ | $X(u)$ | $Z(u)$ | $Y(u)$ | $Y_0(u)$ | $Y_1(u)$ | $Z_0(u)$ | $Z_1(u)$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | $\alpha\beta$ | 0 | $\alpha$ |
| 0 | 1 | 0 | 1 | $\beta$ | $\beta$ | $(\alpha+1)\beta$ | 1 | $\alpha+1$ |
| 1 | 0 | 1 | $\alpha$ | $\alpha\beta$ | 0 | $\alpha\beta$ | 0 | $\alpha$ |
| 1 | 1 | 1 | $\alpha+1$ | $(\alpha+1)\beta$ | $\beta$ | $(\alpha+1)\beta$ | 1 | $\alpha+1$ |

$$X := U_1$$
$$Z := \alpha X + U_2$$
$$Y := \beta Z$$

Table 14.2

- Retrospective reasoning of this sort, concerning dependence on the unrealized past, is not shown explicitly in the graph of Figure 14.3.

- To facilitate such reasoning, we need to devise means of representing counterfactual variables directly in the graph.

The equations also demonstrate a peculiar property of counterfactual conditioning.

Despite the fact that $Z$ separates $X$ from $Y$ in the graph of Figure 14.3, we find that $X$ has an effect on $Y$ for those units falling under $Z = 1$.



Figure 14.3

separates $X$ from $Y$

(College) $X$    (Skills) $Z$    (Salary) $Y$
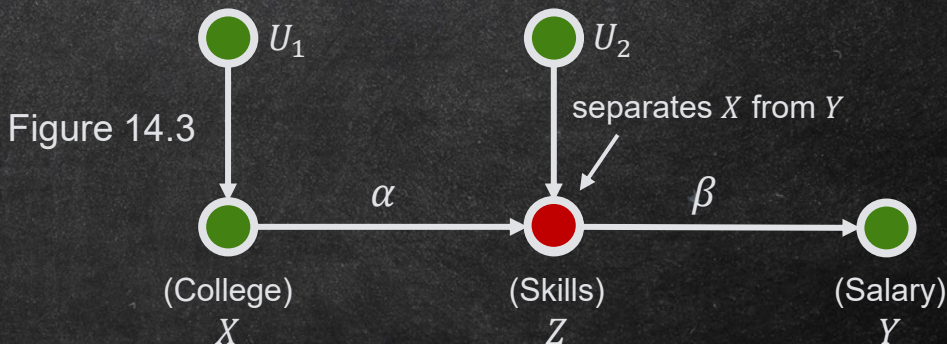
Thus far, the relative magnitudes of the probabilities of $P(U_1)$ and $P(U_1)$ have not entered into the calculations, because the condition $Z = 1$ occurs only for $U_1 = 0$ and $U_2 = 1$ (assuming that $\alpha \neq 0$ and $\alpha \neq 1$), and under these conditions, each of $Y$, $Y_1$, and $Y_0$ has a definite value.

These probabilities play a role, however, if we assume $\alpha = 1$ in the model,

since $Z = 1$ can now occur under two conditions: $(U_1 = 0, U_2 = 1)$ and $(U_1 = 1, U_2 = 0)$.

| $U_1$ | $U_2$ | $X(u)$ | $Z(u)$ | $Y(u)$ | $Y_0(u)$ | $Y_1(u)$ | $Z_0(u)$ | $Z_1(u)$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | $\alpha\beta$ | 0 | $\alpha$ |
| 0 | 1 | 0 | 1 | $\beta$ | $\beta$ | $(\alpha+1)\beta$ | 1 | $\alpha+1$ |
| 1 | 0 | 1 | $\alpha$ | $\alpha\beta$ | 0 | $\alpha\beta$ | 0 | $\alpha$ |
| 1 | 1 | 1 | $\alpha+1$ | $(\alpha+1)\beta$ | $\beta$ | $(\alpha+1)\beta$ | 1 | $\alpha+1$ |

Table 14.2

$\alpha = 1$

| $U_1$ | $U_2$ | $X(u)$ | $Z(u)$ | $Y(u)$ | $Y_0(u)$ | $Y_1(u)$ | $Z_0(u)$ | $Z_1(u)$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | $\beta$ | 0 | 1 |
| 0 | 1 | 0 | 1 | $\beta$ | $\beta$ | $2\beta$ | 1 | 2 |
| 1 | 0 | 1 | 1 | $\beta$ | 0 | $\beta$ | 0 | 1 |
| 1 | 1 | 1 | 2 | $2\beta$ | $\beta$ | $2\beta$ | 1 | 2 |

$$\mathbb{E}[Y_1|Z=1] = (\alpha+1)\beta \qquad \xrightarrow{\ \alpha=1\ } \qquad \mathbb{E}[Y_1|Z=1] = 2\beta \quad \Longleftarrow \quad P(U_1=0)P(U_2=1)$$

$$\mathbb{E}[Y_0|Z=1] = \beta \qquad\qquad\qquad\qquad\qquad \mathbb{E}[Y_0|Z=1] = \beta \quad \Longleftarrow \quad P(U_1=1)P(U_2=0)$$

Thus far, the relative magnitudes of the probabilities of $P(U_1)$ and $P(U_1)$ have not entered into the calculations, because the condition $Z = 1$ occurs only for $U_1 = 0$ and $U_2 = 1$ (assuming that $\alpha \neq 0$ and $\alpha \neq 1$), and under these conditions, each of $Y$, $Y_1$, and $Y_0$ has a definite value.

| $U_1$ | $U_2$ | $X(u)$ | $Z(u)$ | $Y(u)$ | $Y_0(u)$ | $Y_1(u)$ | $Z_0(u)$ | $Z_1(u)$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | $\beta$ | 0 | 1 |
| 0 | 1 | 0 | 1 | $\beta$ | $\beta$ | $2\beta$ | 1 | 2 |
| 1 | 0 | 1 | 1 | $\beta$ | 0 | $\beta$ | 0 | 1 |
| 1 | 1 | 1 | 2 | $2\beta$ | $\beta$ | $2\beta$ | 1 | 2 |

$$\mathbb{E}[Y_0|Z = 1] = \beta \left( \frac{P(U_1 = 0)\, P(U_2 = 1)}{P(U_1 = 0)\, P(U_2 = 1) + P(U_1 = 1)\, P(U_2 = 0)} \right) + 0 \left( \frac{P(U_1 = 1)\, P(U_2 = 0)}{P(U_1 = 0)\, P(U_2 = 1) + P(U_1 = 1)\, P(U_2 = 0)} \right)$$

$$= \beta \left( \frac{P(U_1 = 0)\, P(U_2 = 1)}{P(U_1 = 0)\, P(U_2 = 1) + P(U_1 = 1)\, P(U_2 = 0)} \right)$$

$$\mathbb{E}[Y_1|Z = 1] = 2\beta \left( \frac{P(U_1 = 0)\, P(U_2 = 1)}{P(U_1 = 0)\, P(U_2 = 1) + P(U_1 = 1)\, P(U_2 = 0)} \right) + \beta \left( \frac{P(U_1 = 1)\, P(U_2 = 0)}{P(U_1 = 0)\, P(U_2 = 1) + P(U_1 = 1)\, P(U_2 = 0)} \right)$$

$$= \beta \left( 1 + \frac{P(U_1 = 0)\, P(U_2 = 1)}{P(U_1 = 0)\, P(U_2 = 1) + P(U_1 = 1)\, P(U_2 = 0)} \right)$$

$$\mathbb{E}[Y_1|Z = 1] = 2\beta \qquad \Longleftarrow \qquad P(U_1 = 0)P(U_2 = 1)$$

$$\mathbb{E}[Y_0|Z = 1] = \beta \qquad \Longleftarrow \qquad P(U_1 = 1)P(U_2 = 0)$$

| $U_1$ | $U_2$ | $X(u)$ | $Z(u)$ | $Y(u)$ | $Y_0(u)$ | $Y_1(u)$ | $Z_0(u)$ | $Z_1(u)$ |
|-------|-------|--------|--------|--------|----------|----------|----------|----------|
| 0 | 0 | 0 | 0 | 0 | 0 | $\beta$ | 0 | 1 |
| 0 | 1 | 0 | 1 | $\beta$ | $\beta$ | $2\beta$ | 1 | 2 |
| 1 | 0 | 1 | 1 | $\beta$ | 0 | $\beta$ | 0 | 1 |
| 1 | 1 | 1 | 2 | $2\beta$ | $\beta$ | $2\beta$ | 1 | 2 |

$$\beta \left( 1 + \frac{P(U_1 = 0)\, P(U_2 = 1)}{P(U_1 = 0)\, P(U_2 = 1) + P(U_1 = 1)\, P(U_2 = 0)} \right) > \beta \left( \frac{P(U_1 = 0)\, P(U_2 = 1)}{P(U_1 = 0)\, P(U_2 = 1) + P(U_1 = 1)\, P(U_2 = 0)} \right)$$

$$\mathbb{E}[Y_1 | Z = 1] > \mathbb{E}[Y_0 | Z = 1]$$

The skill-specific causal effect of education on salary is nonzero, despite the fact that salaries are determined by skill only, not by education.

This is to be expected, since a nonzero fraction of the workers at skill level $Z = 1$ did not receive college education, and, had they been given college education, their skill would have increased to $Z_1 = 2$, and their salaries to $2\beta$.

Counterfactuals are byproducts of structural equation models, thus we can see them in the causal graphs associated with those models.

Indeed, the **FUNDAMENTAL LAW OF COUNTERFACTUALS** tells us that, if we modify model $M$ to obtain the submodel $M_x$, then the outcome variable $Y$ in the modified model is the counterfactual $Y_x$ of the original model.

$$Y_x(u) = Y_{M_x}(u)$$

Since modification calls for removing all arrows entering the variable $X$, as illustrated in Figure 14.4, we conclude that the node associated with the $Y$ variable serves as a surrogate for $Y_x$, with the understanding that the substitution is valid only under the modification.
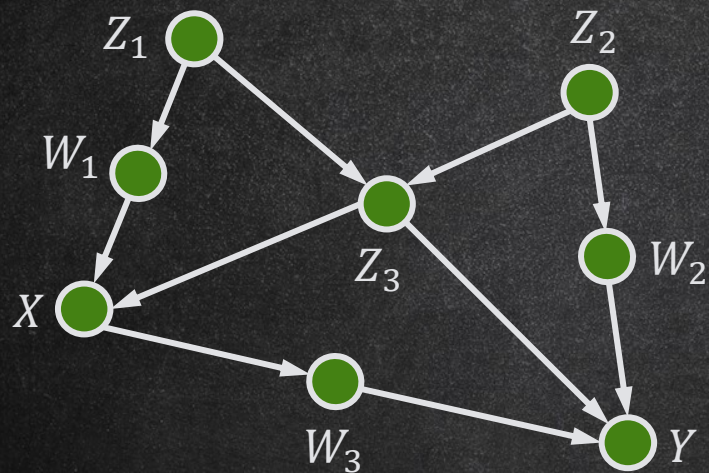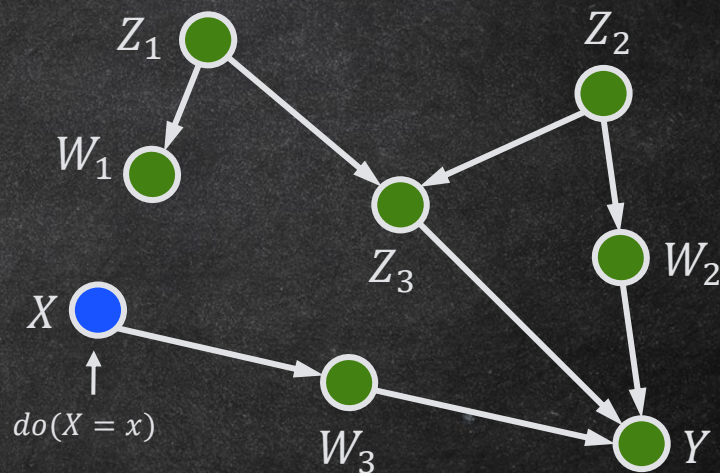


Figure 14.4 (a)  $M$

Figure 14.4 (b)  $M_x$

This temporary visualization of counterfactuals is sufficient to answer some fundamental questions about the statistical properties of $Y_x$ and how those properties depend on other variables in the model, specifically when those other variables are conditioned on.

When we ask about the statistical properties of $Y_x$, we need to examine what would cause $Y_x$ to vary. According to its structural definition, $Y_x$ represents the value of $Y$ under a condition where $X$ is held constant at $X = x$.

Statistical variations of $Y_x$ are therefore governed by all exogenous variables capable of influencing $Y$ when $X$ is held constant, that is, when the arrows entering $X$ are removed, as in Figure 14.4(b).
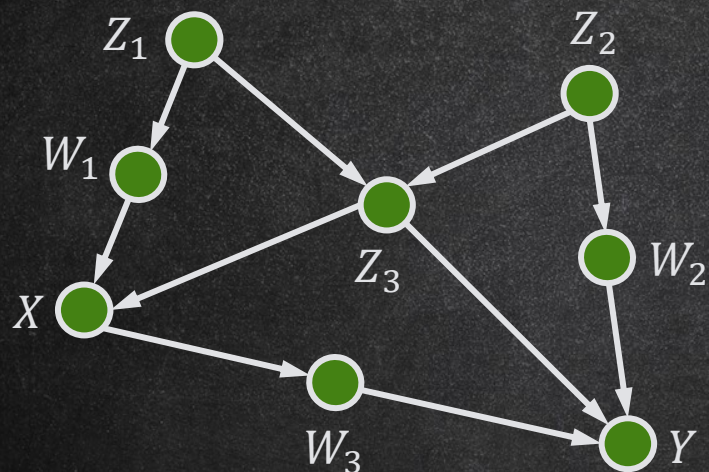


Figure 14.4 (a) $M$
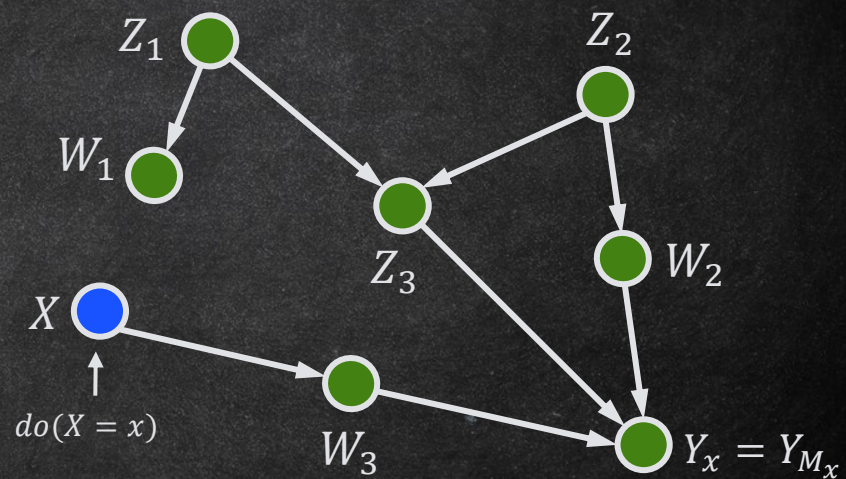
Figure 14.4 (b) $M_x$

Under such conditions, the set of variables capable of transmitting variations to $Y$ are the parents of $Y$, (observed and unobserved) as well as parents of nodes on the pathways between $X$ and $Y$.

In Figure 14.4(b), for example, these nodes are $\{Z_3, W_2, U_Y, U_3\}$, where $U_Y$ and $U_3$, the error terms of $Y$ and $W_3$.

Any set of variables that blocks a path to these parents also blocks that path to $Y_x$, and will result in, therefore, a conditional independence for $Y_x$.

In particular, if we have a set $\mathbf{Z}$ of covariate that satisfies the backdoor criterion in $M$, that set also blocks all paths between $X$ and those parents, and consequently, it renders $X$ and $Y_x$ independent in every stratum $\mathbf{Z} = \mathbf{z}$.

**THE BACKDOOR CRITERION**

Given an ordered pair of variables $(X, Y)$ in a DAG $\mathcal{G}$, a set of variables $\mathbf{S}$ satisfies the backdoor criterion relative to $(X, Y)$ if no node in $\mathbf{S}$ is a descendant of $X$, and $\mathbf{S}$ blocks every path between $X$ and $Y$ that contains an arrow into $X$.
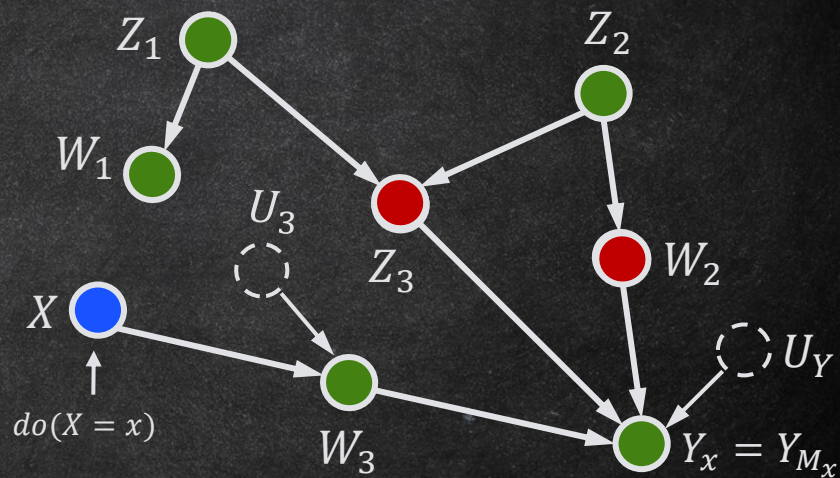


Figure 14.4 (b)  $M_x$

**COUNTERFACTUAL INTERPRETATION OF BACKDOOR**

If a set $\mathbf{Z}$ of variables satisfies the backdoor condition relative to $(X, Y)$, then, for all $x$, the counterfactual $Y_x$ is conditionally independent of $X$ given $\mathbf{Z}$

$$P(Y_x | X, \mathbf{Z}) = P(Y_x | \mathbf{Z})$$

The theorem to the left has far-reaching consequences when it comes to estimating the probabilities of counterfactuals from observational studies.

In particular, it implies that $P(Y_x = y)$ is identifiable by the adjustment formula.

**ADJUSTMENT FORMULA**

$$P\big(Y = y | do(X = x)\big) = \sum_z P(Y = y | X = x, \mathbf{Z} = \mathbf{z})\, P(\mathbf{Z} = \mathbf{z})$$

**COUNTERFACTUAL CONSISTENCY RULE**

If $X = x$  then  $Y_X = Y$

If $X$ is binary then we have the convenient form

$$Y = XY_1 + (1 - X)Y_0$$

$$P(Y_x = y) = \sum_z P(Y_x = y | \mathbf{Z} = \mathbf{z})\, P(\mathbf{Z} = \mathbf{z})$$

$$= \sum_z P(Y_x = y | \mathbf{Z} = \mathbf{z}, X = x)\, P(\mathbf{Z} = \mathbf{z})$$
(by the counterfactual interpretation of backdoor)

$$= \sum_z P(Y = y | \mathbf{Z} = \mathbf{z}, X = x)\, P(\mathbf{Z} = \mathbf{z})$$
(by the counterfactual consistency rule)

$$Y_x(u) = Y_{M_x}(u)$$

sometimes called
"**CONDITIONAL IGNORABILITY**"

converts a graphical
model into algebraic
notation, and allows
us to derive

**COUNTERFACTUAL INTERPRETATION OF BACKDOOR**

If a set $\mathbf{Z}$ of variables satisfies the backdoor condition relative to $(X, Y)$, then, for all $x$, the counterfactual $Y_x$ is conditionally independent of $X$ given $\mathbf{Z}$

$$P(Y_x|X, \mathbf{Z}) = P(Y_x|\mathbf{Z})$$

Gives the notion of conditional ignorability a scientific interpretation and permits us to test whether it holds in any given model.

$$P(Y_x = y) = \sum_z P(Y_x = y|\mathbf{Z} = z)\,P(\mathbf{Z} = z)$$

$$= \sum_z P(Y_x = y|\mathbf{Z} = z, X = x)\,P(\mathbf{Z} = z)$$

(by the counterfactual interpretation of backdoor)

$$= \sum_z P(Y = y|\mathbf{Z} = z, X = x)\,P(\mathbf{Z} = z)$$

(by the counterfactual consistency rule)

The derivation to the left, invokes only algebraic steps; it makes no reference to the model once we ensure that $\mathbf{Z}$ satisfies the backdoor criterion.

Having a graphical representation for counterfactuals, we can resolve the dilemma in Figure 14.3, and explain graphically why a stronger education ($X$) would have had an effect on the salary ($Y$) of people who are currently at skill level $Z = z$, despite the fact that, according to the model, salary is determined by skill only.

Formally, to determine if the effect of education on salary ($Y_x$) is statistically independent of the level of education, we need to locate $Y_x$ in the graph and see if it is d-separated from $X$ given $Z$.

Referring to Figure 14.3, we see that $Y_x$ can be identified with $U_2$, the only parent of nodes on the causal path from $X$ to $Y$ (and therefore, the only variable that produces variations in $Y_x$ while $X$ is held constant).

Inspecting Figure 14.3 tells us that $Z$ acts as a collider between $X$ and $U_2$, and, therefore, $X$ and $U_2$ (and similarly $X$ and $Y_x$) are not d-separated given $Z$.



Figure 14.3

node on the
causal path
from $X$ to $Y$
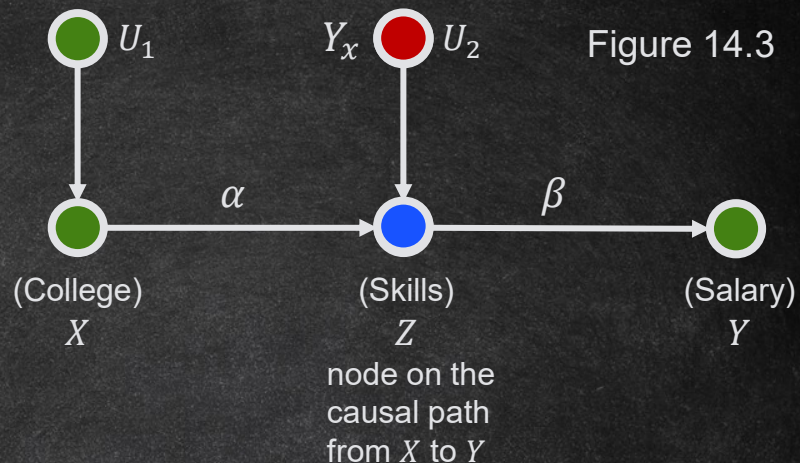
$$\mathbb{E}[Y_x | X, Z] \neq \mathbb{E}[Y_x | Z]$$

despite the fact that

$$\mathbb{E}[Y | X, Z] = \mathbb{E}[Y | Z]$$

We now know that every counterfactual question can be answered from a fully specified structural model.
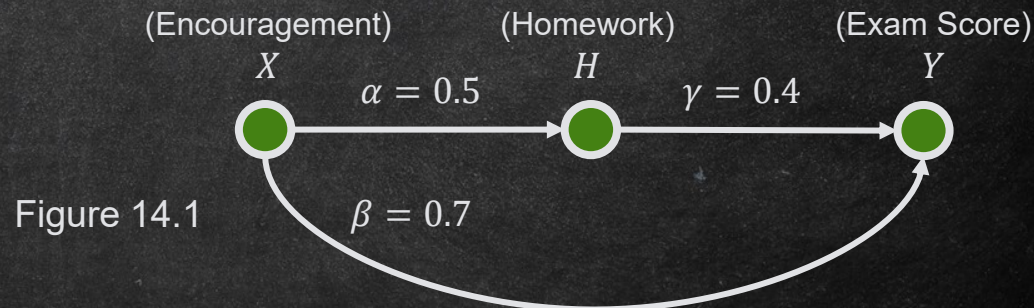
However, what happens in an experimental setting, where a model is not available, and the experimenter must answer interventional questions on the basis of a finite sample of observed individuals?

Let us refer back to the "encouragement design" model of Figure 14.1, in which we analyzed the behavior of an individual named Joe, and assume that the experimenter observes a set of 10 individuals, with Joe being participant 1.

Each individual is characterized by a distinct vector $U_i = (U_X, U_H, U_Y)$, as shown in the first three columns of Table 14.3.

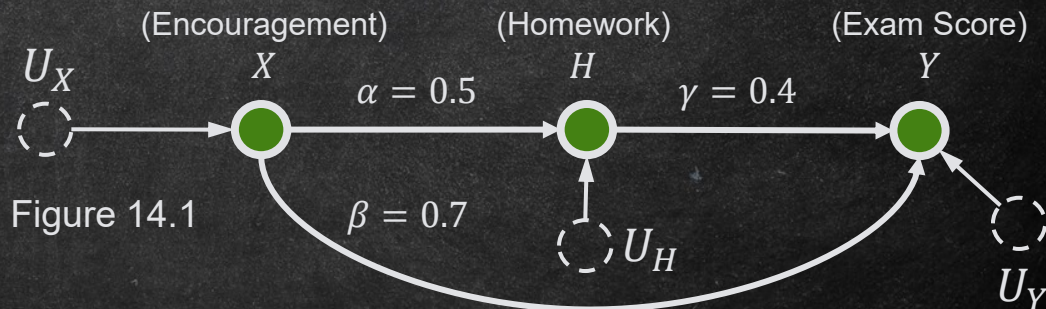| Participant | Participant Characteristic | | | Observed Behaviour | | | Predicted Potential Outcomes | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $U_X$ | $U_H$ | $U_Y$ | $X$ | $H$ | $Y$ | $Y_0$ | $Y_1$ | $H_0$ | $H_1$ | $Y_{00}$ |
| 1 | 0.5 | 0.75 | 0.75 | 0.5 | 1.00 | 1.50 | 1.05 | 1.95 | 0.75 | 1.25 | 0.75 |
| 2 | 0.3 | 0.1 | 0.4 | 0.3 | 0.25 | 0.71 | 0.44 | 1.34 | 0.1 | 0.6 | 0.4 |
| 3 | 0.5 | 0.9 | 0.2 | 0.5 | 1.15 | 1.01 | 0.56 | 1.46 | 0.9 | 1.4 | 0.2 |
| 4 | 0.6 | 0.5 | 0.3 | 0.6 | 0.80 | 1.04 | 0.50 | 1.40 | 0.5 | 1.0 | 0.3 |
| 5 | 0.5 | 0.8 | 0.9 | 0.5 | 1.05 | 1.67 | 1.22 | 2.12 | 0.8 | 1.3 | 0.9 |
| 6 | 0.7 | 0.9 | 0.3 | 0.7 | 1.25 | 1.29 | 0.66 | 1.56 | 0.9 | 1.4 | 0.3 |
| 7 | 0.2 | 0.3 | 0.8 | 0.2 | 0.24 | 1.10 | 0.92 | 1.82 | 0.3 | 0.8 | 0.8 |
| 8 | 0.4 | 0.6 | 0.2 | 0.4 | 0.80 | 0.80 | 0.44 | 1.34 | 0.6 | 1.1 | 0.2 |
| 9 | 0.6 | 0.4 | 0.3 | 0.6 | 0.70 | 1.00 | 0.46 | 1.36 | 0.4 | 0.9 | 0.3 |
| 10 | 0.3 | 0.8 | 0.3 | 0.3 | 0.95 | 0.89 | 0.62 | 1.52 | 0.8 | 1.3 | 0.3 |

Table 14.3



Figure 14.1

From the first six columns of Table 14.3, we can create a full data set that complies with the model.

For each triplet $(U_X, U_H, U_Y)$, the model of Figure 14.1 enables us to complete a full row of the table, including:

- $Y_0$; potential outcomes under treatment ($X = 1$),

- $Y_1$, potential outcomes under control ($X = 0$).

- Columns $X$, $H$, $Y$ predict the results of observational studies.

- Columns $Y_0$, $Y_1$, $H_0$ and $H_1$ predict hypothetical outcome under two treatment regimes, $X = 0$ and $X = 1$.

| | Participant Characteristic | | | Observed Behaviour | | | Predicted Potential Outcomes | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Participant | $U_X$ | $U_H$ | $U_Y$ | $X$ | $H$ | $Y$ | $Y_0$ | $Y_1$ | $H_0$ | $H_1$ | $Y_{00}$ |
| 1 | 0.5 | 0.75 | 0.75 | 0.5 | 1.00 | 1.50 | 1.05 | 1.95 | 0.75 | 1.25 | 0.75 |
| 2 | 0.3 | 0.1 | 0.4 | 0.3 | 0.25 | 0.71 | 0.44 | 1.34 | 0.1 | 0.6 | 0.4 |
| 3 | 0.5 | 0.9 | 0.2 | 0.5 | 1.15 | 1.01 | 0.56 | 1.46 | 0.9 | 1.4 | 0.2 |
| 4 | 0.6 | 0.5 | 0.3 | 0.6 | 0.80 | 1.04 | 0.50 | 1.40 | 0.5 | 1.0 | 0.3 |
| 5 | 0.5 | 0.8 | 0.9 | 0.5 | 1.05 | 1.67 | 1.22 | 2.12 | 0.8 | 1.3 | 0.9 |
| 6 | 0.7 | 0.9 | 0.3 | 0.7 | 1.25 | 1.29 | 0.66 | 1.56 | 0.9 | 1.4 | 0.3 |
| 7 | 0.2 | 0.3 | 0.8 | 0.2 | 0.24 | 1.10 | 0.92 | 1.82 | 0.3 | 0.8 | 0.8 |
| 8 | 0.4 | 0.6 | 0.2 | 0.4 | 0.80 | 0.80 | 0.44 | 1.34 | 0.6 | 1.1 | 0.2 |
| 9 | 0.6 | 0.4 | 0.3 | 0.6 | 0.70 | 1.00 | 0.46 | 1.36 | 0.4 | 0.9 | 0.3 |
| 10 | 0.3 | 0.8 | 0.3 | 0.3 | 0.95 | 0.89 | 0.62 | 1.52 | 0.8 | 1.3 | 0.3 |

Table 14.3



Figure 14.1

From the first six columns of Table 14.3, we can create a full data set that complies with the model.

Many more, in fact infinite, potential outcomes may be predicted;

- for example, $Y_{X=0.5,H=2.0}$ as computed for Joe (Figure 14.2), as well as all combinations of subscripted variables.

From this synthetic population, one can estimate the probability of every counterfactual query on variables $X, H, Y$ assuming, of course, that we are in possession of all entries of the table.

The estimation would require us to simply count the proportion of individuals that satisfy the specified query as previously demonstrated.

| | Participant Characteristic | | | Observed Behaviour | | | Predicted Potential Outcomes | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Participant | $U_X$ | $U_H$ | $U_Y$ | $X$ | $H$ | $Y$ | $Y_0$ | $Y_1$ | $H_0$ | $H_1$ | $Y_{00}$ |
| 1 | 0.5 | 0.75 | 0.75 | 0.5 | 1.00 | 1.50 | 1.05 | 1.95 | 0.75 | 1.25 | 0.75 |
| 2 | 0.3 | 0.1 | 0.4 | 0.3 | 0.25 | 0.71 | 0.44 | 1.34 | 0.1 | 0.6 | 0.4 |
| 3 | 0.5 | 0.9 | 0.2 | 0.5 | 1.15 | 1.01 | 0.56 | 1.46 | 0.9 | 1.4 | 0.2 |
| 4 | 0.6 | 0.5 | 0.3 | 0.6 | 0.80 | 1.04 | 0.50 | 1.40 | 0.5 | 1.0 | 0.3 |
| 5 | 0.5 | 0.8 | 0.9 | 0.5 | 1.05 | 1.67 | 1.22 | 2.12 | 0.8 | 1.3 | 0.9 |
| 6 | 0.7 | 0.9 | 0.3 | 0.7 | 1.25 | 1.29 | 0.66 | 1.56 | 0.9 | 1.4 | 0.3 |
| 7 | 0.2 | 0.3 | 0.8 | 0.2 | 0.24 | 1.10 | 0.92 | 1.82 | 0.3 | 0.8 | 0.8 |
| 8 | 0.4 | 0.6 | 0.2 | 0.4 | 0.80 | 0.80 | 0.44 | 1.34 | 0.6 | 1.1 | 0.2 |
| 9 | 0.6 | 0.4 | 0.3 | 0.6 | 0.70 | 1.00 | 0.46 | 1.36 | 0.4 | 0.9 | 0.3 |
| 10 | 0.3 | 0.8 | 0.3 | 0.3 | 0.95 | 0.89 | 0.62 | 1.52 | 0.8 | 1.3 | 0.3 |

Table 14.3



(Encouragement) $X$ — (Homework) $H$ — $\gamma = 0.4$ — (Exam Score) $Y$
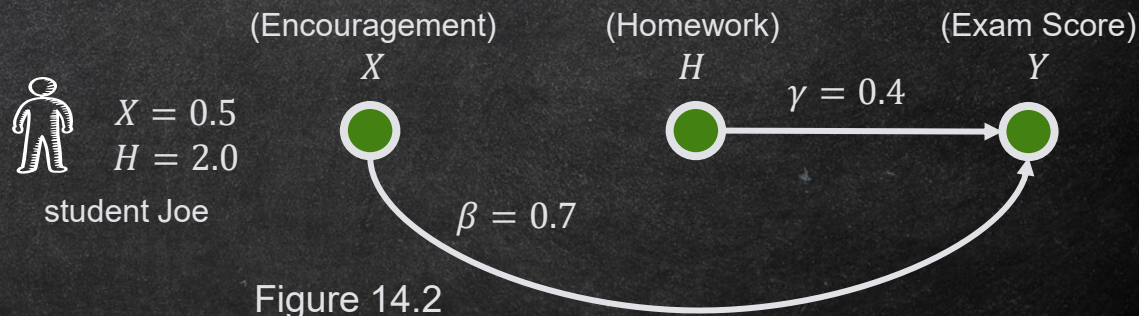
$X = 0.5$
$H = 2.0$

student Joe

$\beta = 0.7$

Figure 14.2

The information conveyed by Table 14.3 is not available to us in either observational or experimental studies.

This information was deduced from a parametric model such as the one in Figure 14.2, from which we could infer the defining characteristics $\{U_X, U_H, U_Y\}$ of each participant, given the observations $\{X, H, Y\}$.

| | Participant Characteristic | | | Observed Behaviour | | | Predicted Potential Outcomes | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Participant | $U_X$ | $U_H$ | $U_Y$ | $X$ | $H$ | $Y$ | $Y_0$ | $Y_1$ | $H_0$ | $H_1$ | $Y_{00}$ |
| 1 | 0.5 | 0.75 | 0.75 | 0.5 | 1.00 | 1.50 | 1.05 | 1.95 | 0.75 | 1.25 | 0.75 |
| 2 | 0.3 | 0.1 | 0.4 | 0.3 | 0.25 | 0.71 | 0.44 | 1.34 | 0.1 | 0.6 | 0.4 |
| 3 | 0.5 | 0.9 | 0.2 | 0.5 | 1.15 | 1.01 | 0.56 | 1.46 | 0.9 | 1.4 | 0.2 |
| 4 | 0.6 | 0.5 | 0.3 | 0.6 | 0.80 | 1.04 | 0.50 | 1.40 | 0.5 | 1.0 | 0.3 |
| 5 | 0.5 | 0.8 | 0.9 | 0.5 | 1.05 | 1.67 | 1.22 | 2.12 | 0.8 | 1.3 | 0.9 |
| 6 | 0.7 | 0.9 | 0.3 | 0.7 | 1.25 | 1.29 | 0.66 | 1.56 | 0.9 | 1.4 | 0.3 |
| 7 | 0.2 | 0.3 | 0.8 | 0.2 | 0.24 | 1.10 | 0.92 | 1.82 | 0.3 | 0.8 | 0.8 |
| 8 | 0.4 | 0.6 | 0.2 | 0.4 | 0.80 | 0.80 | 0.44 | 1.34 | 0.6 | 1.1 | 0.2 |
| 9 | 0.6 | 0.4 | 0.3 | 0.6 | 0.70 | 1.00 | 0.46 | 1.36 | 0.4 | 0.9 | 0.3 |
| 10 | 0.3 | 0.8 | 0.3 | 0.3 | 0.95 | 0.89 | 0.62 | 1.52 | 0.8 | 1.3 | 0.3 |

Table 14.3

In general, in the absence of a parametric model, there is very little we learn about the potential outcomes $Y_1$ and $Y_0$ of individual participants, when all we have is their observed behavior $\{X, H, Y\}$.
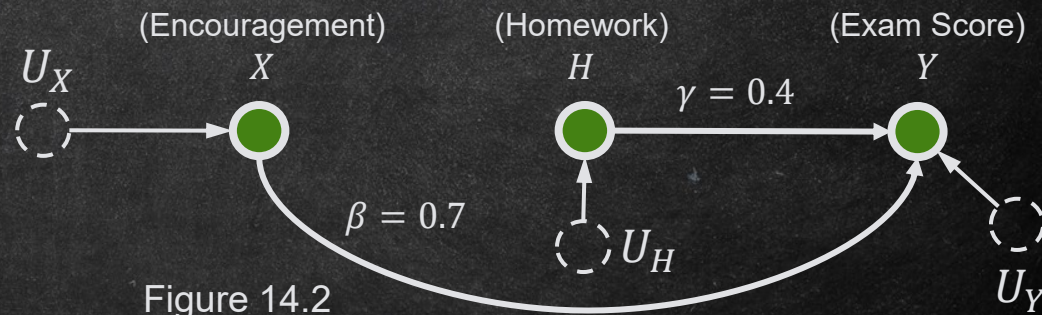


Figure 14.2

Theoretically, the only connection we have between the counterfactuals $\{Y_1, Y_0\}$ and the observables $\{X, H, Y\}$ is the consistency rule

**COUNTERFACTUAL CONSISTENCY RULE**

If $X = x$ then $Y_X = Y$

If $X$ is binary then we have the convenient form

$Y = XY_1 + (1 - X)Y_0$

| Participant | Participant Characteristic | | | Observed Behaviour | | | Predicted Potential Outcomes | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $U_X$ | $U_H$ | $U_Y$ | $X$ | $H$ | $Y$ | $Y_0$ | $Y_1$ | $H_0$ | $H_1$ | $Y_{00}$ |
| 1 | 0.5 | 0.75 | 0.75 | 0.5 | 1.00 | 1.50 | 1.05 | 1.95 | 0.75 | 1.25 | 0.75 |
| 2 | 0.3 | 0.1 | 0.4 | 0.3 | 0.25 | 0.71 | 0.44 | 1.34 | 0.1 | 0.6 | 0.4 |
| 3 | 0.5 | 0.9 | 0.2 | 0.5 | 1.15 | 1.01 | 0.56 | 1.46 | 0.9 | 1.4 | 0.2 |
| 4 | 0.6 | 0.5 | 0.3 | 0.6 | 0.80 | 1.04 | 0.50 | 1.40 | 0.5 | 1.0 | 0.3 |
| 5 | 0.5 | 0.8 | 0.9 | 0.5 | 1.05 | 1.67 | 1.22 | 2.12 | 0.8 | 1.3 | 0.9 |
| 6 | 0.7 | 0.9 | 0.3 | 0.7 | 1.25 | 1.29 | 0.66 | 1.56 | 0.9 | 1.4 | 0.3 |
| 7 | 0.2 | 0.3 | 0.8 | 0.2 | 0.24 | 1.10 | 0.92 | 1.82 | 0.3 | 0.8 | 0.8 |
| 8 | 0.4 | 0.6 | 0.2 | 0.4 | 0.80 | 0.80 | 0.44 | 1.34 | 0.6 | 1.1 | 0.2 |
| 9 | 0.6 | 0.4 | 0.3 | 0.6 | 0.70 | 1.00 | 0.46 | 1.36 | 0.4 | 0.9 | 0.3 |
| 10 | 0.3 | 0.8 | 0.3 | 0.3 | 0.95 | 0.89 | 0.62 | 1.52 | 0.8 | 1.3 | 0.3 |

Table 14.3

which informs us that, $Y_1$ must be equal to $Y$ in case $X = 1$, and $Y_0$ must be equal to $Y$ in case $X = 0$.

But aside from this tenuous connection, most of the counterfactuals associated with the individual participants will remain unobserved.
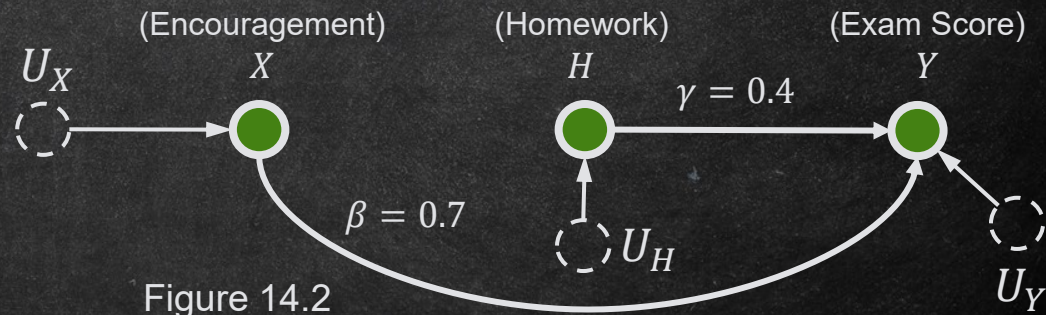


Figure 14.2

Fortunately, there is much we can learn about those counterfactuals at the population level, such as estimating their probabilities or expectation.

This we have witnessed already through the adjustment formula,

$$P(Y_x = y) = \sum_{\mathbf{z}} P(Y = y | \mathbf{Z} = \mathbf{z}, X = x) \, P(\mathbf{Z} = \mathbf{z})$$

where we were able to compute

$$\mathbb{E}[Y_1 - Y_0]$$

using the graph alone, instead of a complete model.

Much more can be obtained from experimental studies, where even the graph becomes dispensable.

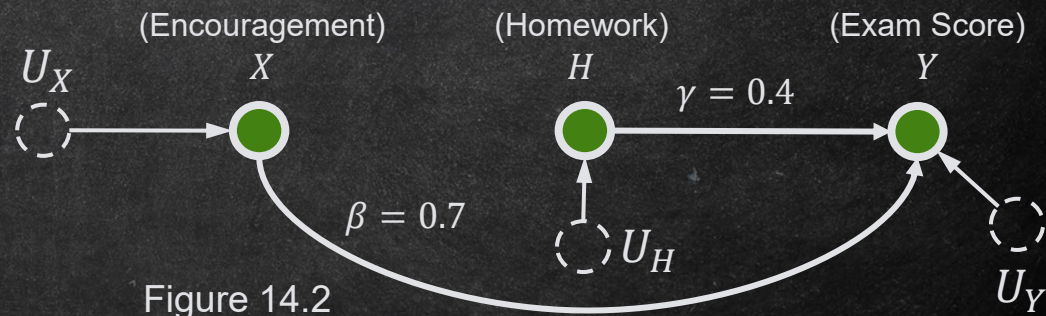| | Participant Characteristic | | | Observed Behaviour | | | Predicted Potential Outcomes | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Participant | $U_X$ | $U_H$ | $U_Y$ | $X$ | $H$ | $Y$ | $Y_0$ | $Y_1$ | $H_0$ | $H_1$ | $Y_{00}$ |
| 1 | 0.5 | 0.75 | 0.75 | 0.5 | 1.00 | 1.50 | 1.05 | 1.95 | 0.75 | 1.25 | 0.75 |
| 2 | 0.3 | 0.1 | 0.4 | 0.3 | 0.25 | 0.71 | 0.44 | 1.34 | 0.1 | 0.6 | 0.4 |
| 3 | 0.5 | 0.9 | 0.2 | 0.5 | 1.15 | 1.01 | 0.56 | 1.46 | 0.9 | 1.4 | 0.2 |
| 4 | 0.6 | 0.5 | 0.3 | 0.6 | 0.80 | 1.04 | 0.50 | 1.40 | 0.5 | 1.0 | 0.3 |
| 5 | 0.5 | 0.8 | 0.9 | 0.5 | 1.05 | 1.67 | 1.22 | 2.12 | 0.8 | 1.3 | 0.9 |
| 6 | 0.7 | 0.9 | 0.3 | 0.7 | 1.25 | 1.29 | 0.66 | 1.56 | 0.9 | 1.4 | 0.3 |
| 7 | 0.2 | 0.3 | 0.8 | 0.2 | 0.24 | 1.10 | 0.92 | 1.82 | 0.3 | 0.8 | 0.8 |
| 8 | 0.4 | 0.6 | 0.2 | 0.4 | 0.80 | 0.80 | 0.44 | 1.34 | 0.6 | 1.1 | 0.2 |
| 9 | 0.6 | 0.4 | 0.3 | 0.6 | 0.70 | 1.00 | 0.46 | 1.36 | 0.4 | 0.9 | 0.3 |
| 10 | 0.3 | 0.8 | 0.3 | 0.3 | 0.95 | 0.89 | 0.62 | 1.52 | 0.8 | 1.3 | 0.3 |

Table 14.3



Figure 14.2

Assume that we have no information whatsoever about the underlying model.

All we have are measurements on $Y$ taken in an experimental study in which $X$ is randomized over two levels, $X = 0$ and $X = 1$.

Table 14.4 describes the responses of the same 10 participants (Joe being participant 1) under such experimental conditions, with participants

student
Joe

| Participant | Predicted Potential Outcomes | | Observed Outcomes | |
|---|---|---|---|---|
| | $Y_0$ | $Y_1$ | $Y_0$ | $Y_1$ |
| 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 | □ |

Table 14.4

Assume that we have no information whatsoever about the underlying model.

All we have are measurements on $Y$ taken in an experimental study in which $X$ is randomized over two levels, $X = 0$ and $X = 1$.

Table 14.4 describes the responses of the same 10 participants (Joe being participant 1) under such experimental conditions, with participants

- 1, 5, 6, 8, and 10 assigned to $X = 0$,
- ~~2, ..., ... assigned to $X = 1$~~.

student Joe

Control group $X = 0$

| Participant | Predicted Potential Outcomes | | Observed Outcomes | |
|---|---|---|---|---|
| | $Y_0$ | $Y_1$ | $Y_0$ | $Y_1$ |
| 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 | □ |

Table 14.4

Assume that we have no information whatsoever about the underlying model.

All we have are measurements on $Y$ taken in an experimental study in which $X$ is randomized over two levels, $X = 0$ and $X = 1$.

Table 14.4 describes the responses of the same 10 participants (Joe being participant 1) under such experimental conditions, with participants

- 1, 5, 6, 8, and 10 assigned to $X = 0$,

- 2, 3, 4, 7, and 9 assigned to $X = 1$.

student Joe

Treatment group $X = 1$

| Participant | Predicted Potential Outcomes | | Observed Outcomes | |
| --- | --- | --- | --- | --- |
| | $Y_0$ | $Y_1$ | $Y_0$ | $Y_1$ |
| 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 | ☐ |

Table 14.4

The first two columns give the true potential outcomes (taken from Table 4.3), while the last two columns describe the information available to the experimenter, where a square indicates that the response was not observed.

student Joe

- $Y_0$ is observed only for participants assigned to $X = 0$
- $Y_1$ is observed only for participants assigned to $X = 1$.

RANDOMIZATION assures us that, although half of the potential outcomes are not observed, the difference between the observed means in the treatment and control groups will converge to the difference of the population averages, $\mathbb{E}[Y_1 - Y_0] = 0.9$.

This is because randomization distributes the black squares at random along the two rightmost columns of Table 14.4, independent of the actual values of $Y_0$ and $Y_1$, so as the number of sample increases, the sample means converge to the population means.

| Participant | Predicted Potential Outcomes | | Observed Outcomes | |
|---|---|---|---|---|
| | $Y_0$ | $Y_1$ | $Y_0$ | $Y_1$ |
| 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

Table 14.4

This unique and important property of randomized experiments is not new to us, since randomization, like interventions, renders $X$ independent of any variable that may affect $Y$ (as in Figure 14.4(b)).
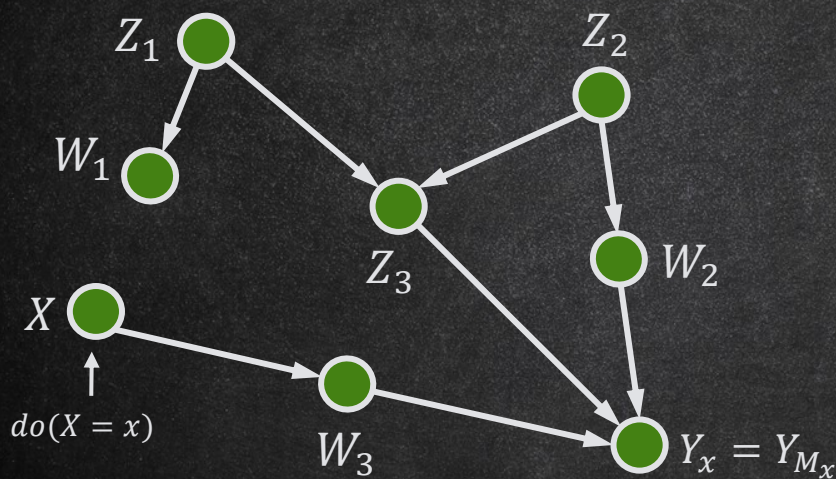
student Joe

| | Predicted Potential Outcomes | | Observed Outcomes | |
|---|---|---|---|---|
| Participant | $Y_0$ | $Y_1$ | $Y_0$ | $Y_1$ |
| 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

Table 14.4



Figure 14.4 (b) $M_x$

$Z_1$

$Z_2$

$W_1$

$Z_3$

$W_2$

$X$

$do(X = x)$

$W_3$

$Y_x = Y_{M_x}$

Under such conditions, the adjustment formula

$$P(Y_x = y) = \sum_z P(Y = y | \mathbf{Z} = \mathbf{z}, X = x)\, P(\mathbf{Z} = \mathbf{z})$$

is applicable with $\mathbf{Z} = \{\emptyset\}$, yielding

$$\mathbb{E}[Y_x] = \mathbb{E}[Y | X = x],$$

where $X = 1$ represents treated units and $X = 0$ untreated.

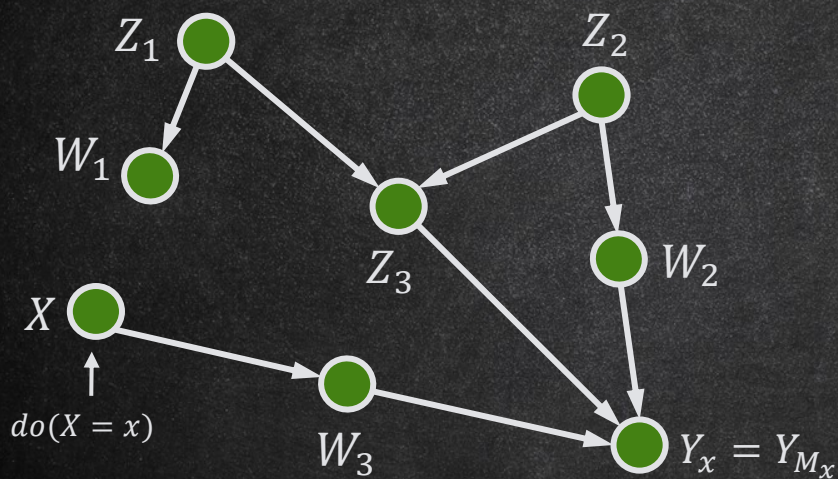student Joe

| Participant | Predicted Potential Outcomes | | Observed Outcomes | |
|---|---|---|---|---|
| | $Y_0$ | $Y_1$ | $Y_0$ | $Y_1$ |
| 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

Table 14.4



Figure 14.4 (b) $M_x$

$do(X = x)$

$Y_x = Y_{M_x}$

Table 14.4 helps us understand what is actually computed when we take sample averages in experimental settings and how those averages are related to the underlying counterfactuals, $Y_1$ and $Y_0$.

student Joe

| Participant | Predicted Potential Outcomes | | Observed Outcomes | |
|---|---|---|---|---|
| | $Y_0$ | $Y_1$ | $Y_0$ | $Y_1$ |
| 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

Table 14.4

$Z_1$

$W_1$

$Z_2$

$Z_3$

$W_2$
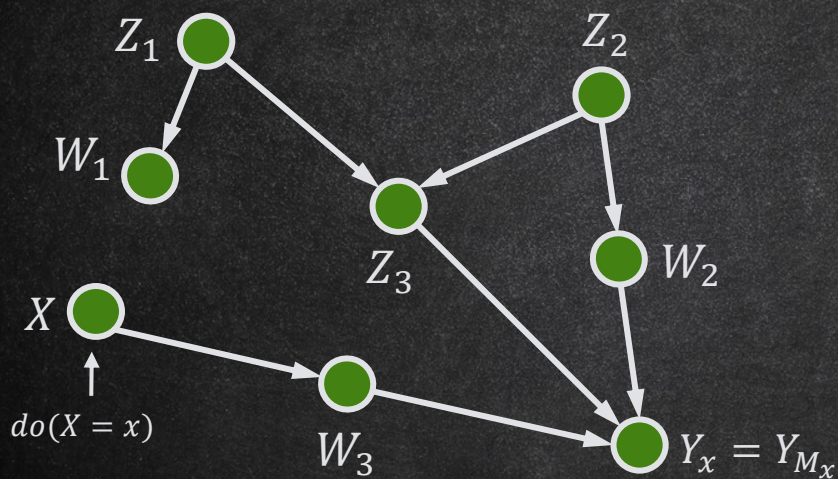
$X$

$do(X = x)$

$W_3$

$Y_x = Y_{M_x}$

Figure 14.4 (b)  $M_x$

In nonparametric models, counterfactual quantities of the form

$$\mathbb{E}[Y_{X=x}|\mathbf{Z} = \mathbf{z}]$$

may not be identifiable, even if we have the luxury of running experiments.

In fully linear models, however, things are much easier; any counterfactual quantity is identifiable whenever the model parameters are identified.

This is because the parameters fully define the model's functions, and as we have seen earlier, once the functions are given, counterfactuals are computable using

$$Y_x(u) = Y_{M_x}(u)$$

Since in linear models every model parameter is identifiable from interventional studies using the interventional definition of direct effects, we conclude that in linear models, every counterfactual is experimentally identifiable.

**LINEAR COUNTERFACTUAL IDENTIFIABILITY**

Can counterfactuals be identified in observational studies, when some of the model parameters are not identified?

Any counterfactual of the form

$$\mathbb{E}[Y_{X=x}|\mathbf{Z} = \mathbf{e}]$$

with $e$ an arbitrary set of evidence, is identified whenever

$$\mathbb{E}[Y|do(X = x)]$$

is identified.

Intuitive interpretation of counterfactuals in linear models:

$$\mathbb{E}[Y_{X=x}|\mathbf{Z} = e]$$

can be computed by:

1) calculating the best estimate of $Y$ conditioned on the evidence $e$, $\mathbb{E}[Y|\mathbf{Z} = e]$,

2) adding to it whatever change is expected in $Y$ when $X$ is shifted from its current best estimate, $\mathbb{E}[Y|\mathbf{Z} = e]$, to its hypothetical value, $x$.

Can counterfactuals be identified in observational studies, when some of the model parameters are not identified?

**RELATIONSHIP BETWEEN** $\mathbb{E}[Y_{X=x}|\mathbf{Z} = e]$ and $\mathbb{E}[Y|do(X = x)]$

Let $\tau$ be the slope of the total effect of $X$ on $Y$,

$$\tau = \mathbb{E}[Y|do(x + 1)] - \mathbb{E}[Y|do(x + 1)]$$

then, for any evidence $\mathbf{Z} = e$, we have

$$\mathbb{E}[Y_{X=x}|\mathbf{Z} = e] = \mathbb{E}[Y|\mathbf{Z} = e] + \tau(x - \mathbb{E}[X|\mathbf{Z} = e])$$

**LINEAR COUNTERFACTUAL IDENTIFIABILITY**

Any counterfactual of the form

$$\mathbb{E}[Y_{X=x}|\mathbf{Z} = e]$$

with $e$ an arbitrary set of evidence, is identified whenever

$$\mathbb{E}[Y|do(X = x)]$$

is identified.

Methodologically, the importance of the theorem to the right lies in enabling researchers to answer hypothetical questions about individuals (or sets of individuals) from population data.

In the situation illustrated by Figure 14.1, we computed the counterfactual $Y_{H=2}$ under the evidence
$$e = \{X = 0.5, H = 1, Y = 1\}$$

**RELATIONSHIP BETWEEN** $\mathbb{E}[Y_{X=x}|\mathbf{Z} = e]$ and $\mathbb{E}[Y|do(X = x)]$

Let $\tau$ be the slope of the total effect of $X$ on $Y$,
$$\tau = \mathbb{E}[Y|do(x + 1)] - \mathbb{E}[Y|do(x + 1)]$$

then, for any evidence $\mathbf{Z} = e$, we have
$$\mathbb{E}[Y_{X=x}|\mathbf{Z} = e] = \mathbb{E}[Y|\mathbf{Z} = e] + \tau(x - \mathbb{E}[X|\mathbf{Z} = e])$$

We now demonstrate how the Theorem to the left can be applied to this model in computing the
**EFFECT OF TREATMENT ON THE TREATED**

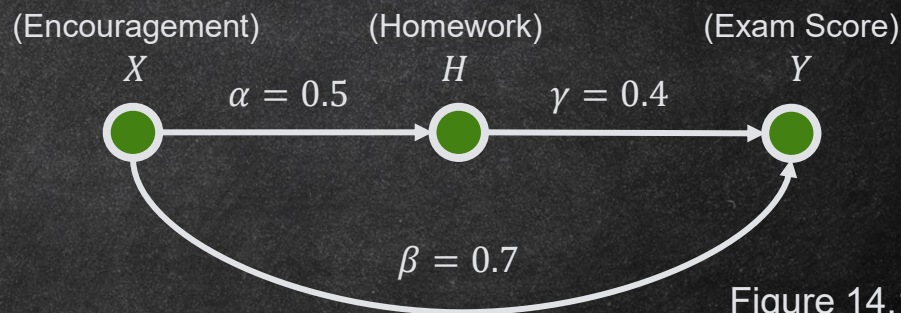$$\boldsymbol{ETT} = \mathbb{E}[Y_1 - Y_0|X = 1]$$



Figure 14.1

Methodologically, the importance of the theorem to the right lies in enabling researchers to answer hypothetical questions about individuals (or sets of individuals) from population data.

In the situation illustrated by Figure 14.1, we computed the counterfactual $Y_{H=2}$ under the evidence
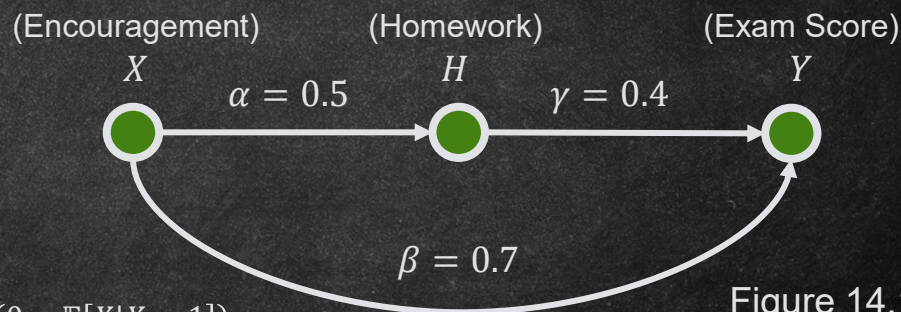$$e = \{X = 0.5, H = 1, Y = 1\}$$

**RELATIONSHIP BETWEEN** $\mathbb{E}[Y_{X=x}|\mathbf{Z} = e]$ and $\mathbb{E}[Y|do(X = x)]$

Let $\tau$ be the slope of the total effect of $X$ on $Y$,
$$\tau = \mathbb{E}[Y|do(x + 1)] - \mathbb{E}[Y|do(x + 1)]$$
then, for any evidence $\mathbf{Z} = e$, we have
$$\mathbb{E}[Y_{X=x}|\mathbf{Z} = e] = \mathbb{E}[Y|\mathbf{Z} = e] + \tau(x - \mathbb{E}[X|\mathbf{Z} = e]) \quad (*)$$

We now demonstrate how the Theorem to the left can be applied to this model in computing the
**EFFECT OF TREATMENT ON THE TREATED**

$$\boldsymbol{ETT} = \mathbb{E}[Y_1 - Y_0|X = 1]$$

Substituting the evidence $e = \{X = 0.5\}$ to (*)

$$\boldsymbol{ETT} = \mathbb{E}[Y_1|X = 1] - \mathbb{E}[Y_0|X = 1]$$

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

$$= \tau$$

$$= \beta + \alpha\gamma = 0.9$$



(Encouragement) $X$ — $\alpha = 0.5$ → (Homework) $H$ — $\gamma = 0.4$ → (Exam Score) $Y$

$\beta = 0.7$

Figure 14.1

The effect of treatment on the treated is equal to the effect of treatment on the entire population.

A general result in linear systems that can be seen directly from (*)

$$\mathbb{E}[Y_{x+1} - Y_x | \mathbf{Z} = e] = \tau$$

independent on the evidence of $e$.

**RELATIONSHIP BETWEEN** $\mathbb{E}[Y_{X=x} | \mathbf{Z} = e]$ and $\mathbb{E}[Y | do(X = x)]$

Let $\tau$ be the slope of the total effect of $X$ on $Y$,

$$\tau = \mathbb{E}[Y | do(x + 1)] - \mathbb{E}[Y | do(x + 1)]$$

then, for any evidence $\mathbf{Z} = e$, we have

$$\mathbb{E}[Y_{X=x} | \mathbf{Z} = e] = \mathbb{E}[Y | \mathbf{Z} = e] + \tau(x - \mathbb{E}[X | \mathbf{Z} = e]) \quad (*)$$

Things are different when a multiplicative (i.e., nonlinear) interaction term is added to the output equation.

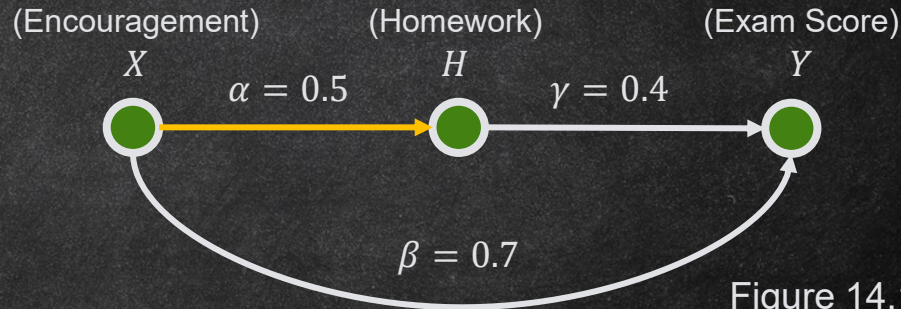For example, if the arrow $X \to H$ were reversed in Figure 14.1, and the equation for $Y$ read

(Encouragement)    (Homework)    (Exam Score)

$X$    $H$    $Y$

$\alpha = 0.5$    $\gamma = 0.4$

$\beta = 0.7$

Figure 14.1

**RELATIONSHIP BETWEEN** $\mathbb{E}[Y_{X=x}|\mathbf{Z} = \boldsymbol{e}]$ and $\mathbb{E}[Y|do(X = x)]$

A general result in linear systems that can be seen directly from (*)

$$\mathbb{E}[Y_{x+1} - Y_x|\mathbf{Z} = \boldsymbol{e}] = \tau$$

independent on the evidence of $\boldsymbol{e}$.

Let $\tau$ be the slope of the total effect of $X$ on $Y$,

$$\tau = \mathbb{E}[Y|do(x + 1)] - \mathbb{E}[Y|do(x + 1)]$$

then, for any evidence $\mathbf{Z} = \boldsymbol{e}$, we have

$$\mathbb{E}[Y_{X=x}|\mathbf{Z} = \boldsymbol{e}] = \mathbb{E}[Y|\mathbf{Z} = \boldsymbol{e}] + \tau(x - \mathbb{E}[X|\mathbf{Z} = \boldsymbol{e}]) \quad (*)$$

Things are different when a multiplicative (i.e., nonlinear) interaction term is added to the output equation.

For example, if the arrow $X \to H$ were reversed in Figure 14.1, and the equation for $Y$ read

$$Y := \beta X + \gamma H + \delta XH + U_Y \implies \tau \neq ETT$$

nonlinear interaction term



(Encouragement) $X$    $\alpha = 0.5$    (Homework) $H$    $\gamma = 0.4$    (Exam Score) $Y$

$\beta = 0.7$

Figure 14.5

# PART IV

# MATHEMATICAL TOOL KITS FOR ATTRIBUTION AND MEDIATION

Assuming binary events, let

- $X = x$ and $Y = y$ represent treatment and outcome, respectively, and

- $X = x'$, $Y = y'$ their negations.

Our target quantity is defined by the English sentence:

"Find the probability that if $X$ had been $x'$, $Y$ would be $y'$, given that, in reality, $X$ is $x$ and $Y$ is $y$."

mathematically

**PROBABILITY OF NECESSITY** (PN)

Captures the legal criterion of "but for," according to which judgment in favor of a plaintiff should be made if and only if it is "more probable than not" that the damage would not have occurred but for the defendant's action.

$$PN(x, y) = P(Y_{x'} = y' | X = x, Y = y)$$

What assumptions permit us to identify PN from empirical studies, be they observational, experimental, or a combination thereof?

If $Y$ is monotonic relative to $X$, that is, $Y_1(u) \geq Y_0(u)$ for all $u$, then $PN$ is identifiable whenever the causal effect $P(y|do(x))$ is identifiable, and

$$PN = \frac{P(y) - P(y|do(x'))}{P(x, y)} \quad \text{or, substituting} \quad P(y) = P(y|x)P(x) + P(y|x')(1 - P(x))$$

$$PN = \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y|do(x'))}{P(x, y)}$$

**EXCESS RISK RATIO (ERR)**

often used in court cases in the absence of experimental data

**CONFOUNDING FACTOR (CF)**

represents a correction needed to account for confounding bias, that is, $P(y|do(x')) \neq P(y|x')$

Confounding occurs when the proportion of population for whom $Y = y$, when $X$ is set to $x'$ for everyone is not the same as the proportion of the population for whom $Y = y$ among those acquiring $X = x'$ by choice.

$$PN = \underbrace{\frac{P(y|x) - P(y|x')}{P(y|x)}}_{} + \underbrace{\frac{P(y|x') - P(y|do(x'))}{P(x, y)}}_{}$$

Tells us how much more likely people are to die in crashes when driving one of the manufacturer's cars.

Corrects for the bias that people who buy the manufacturer's cars are more likely to drive fast (leading to deadlier crashes) than the general populations.

For instance, suppose there is a case brought against a car manufacturer, claiming that its car's faulty design led to a man's death in a car crash.

If it turns out that people who buy the manufacturer's cars are more likely to drive fast (leading to deadlier crashes) than the general populations, the second term will correct for that bias.

**EXCESS RISK RATIO**
**(ERR)**

**CONFOUNDING FACTOR**
**(CF)**

often used in court
cases in the absence
of experimental data

represents a correction
needed to account for
confounding bias, that is,
$P(y|do(x') \neq P(y|x')$

$$PN = \overbrace{\frac{P(y|x) - P(y|x')}{P(y|x)}} + \overbrace{\frac{P(y|x') - P(y|do(x'))}{P(x,y)}}$$

The formula provides an estimable measure of **NECESSARY CAUSATION**, which can be used for monotonic $Y_x(u)$ whenever the causal effect $P(y|do(x))$ can be estimated, be it from randomized trials or from graph-assisted observational studies (e.g., through the backdoor criterion).

It has also been shown that the
expression

$$PN = \frac{P(y) - P(y|do(x'))}{P(x,y)}$$

provides a lower bound for PN in
the general nonmonotonic case.

In particular, the upper and lower bounds on PN are given by

$$max\left\{0, \frac{P(y) - P(y|do(x'))}{P(x,y)}\right\} \leq PN \leq min\left\{1, \frac{P(y'|do(x')) - P(x',y')}{P(x,y)}\right\}$$

**EXCESS RISK RATIO
(ERR)**

**CONFOUNDING FACTOR
(CF)**

often used in court
cases in the absence
of experimental data

represents a correction
needed to account for
confounding bias, that is,
$P(y|do(x')) \neq P(y|x')$

$$PN = \overbrace{\frac{P(y|x) - P(y|x')}{P(y|x)}} + \overbrace{\frac{P(y|x') - P(y|do(x'))}{P(x,y)}}$$

In drug-related litigation, it is not uncommon to obtain data from both experimental and observational studies. The former is usually available from the manufacturer or the agency that approved the drug for distribution (e.g., FDA, WHO), whereas the latter is often available from surveys of the population.

$$CF \triangleq \frac{P(y|x') - p(Y_{x'})}{P(x,y)}$$

A few algebraic steps allow us to express the lower bound (LB) and upper bound (UB) as

$$ERR \triangleq 1 - \frac{1}{RR} = 1 - \frac{P(y|x')}{P(y|x)}$$

$$LB = ERR + CF$$

$$UB = ERR + CF + q$$

$$q \triangleq \frac{P(y'|x)}{P(y|x)}$$

$$max\left\{0, \overbrace{\frac{P(y) - P(y|do(x'))}{P(x,y)}}\right\} \leq PN \leq min\left\{1, \overbrace{\frac{P(y'|do(x')) - P(x',y')}{P(x,y)}}\right\}$$

- $CF$ represents the normalized degree of confounding among the unexposed ($X = x'$),

- $ERR$ is the "excess risk ratio" and,

- $q$ is the ratio of negative to positive outcomes among the exposed.

**EXCESS RISK RATIO (ERR)**

often used in court cases in the absence of experimental data

**CONFOUNDING FACTOR (CF)**

represents a correction needed to account for confounding bias, that is, $P(y|do(x') \neq P(y|x')$

$$PN = \overbrace{\frac{P(y|x) - P(y|x')}{P(y|x)}} + \overbrace{\frac{P(y|x') - P(y|do(x'))}{P(x,y)}}$$

$$CF \triangleq \frac{P(y|x') - p(Y_{x'})}{P(x,y)}$$

$$ERR \triangleq 1 - \frac{1}{RR} = 1 - \frac{P(y|x')}{P(y|x)}$$

$$q \triangleq \frac{P(y'|x)}{P(y|x)}$$

A few algebraic steps allow us to express the lower bound (LB) and upper bound (UB) as
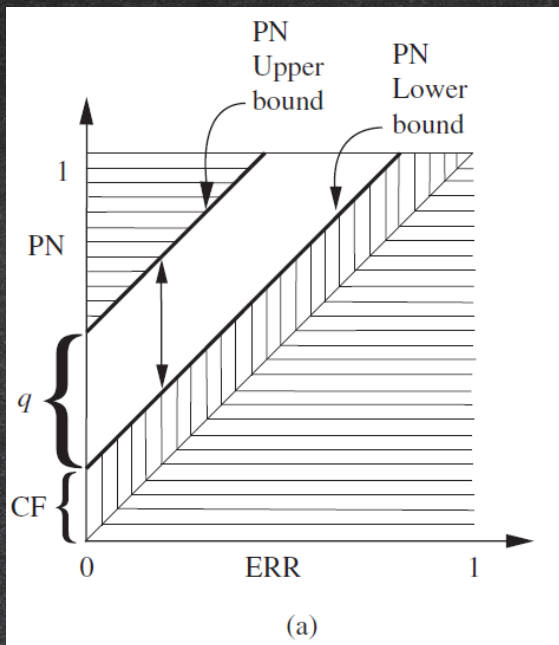
$$LB = ERR + CF$$

$$UB = ERR + CF + q$$

$$max\left\{0, \overbrace{\frac{P(y) - P(y|do(x'))}{P(x,y)}}\right\} \leq PN \leq min\left\{1, \overbrace{\frac{P(y'|do(x')) - P(x',y')}{P(x,y)}}\right\}$$

$$PN = \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y|do(x'))}{P(x,y)}$$

The figure to the right shows bounds as a function of $ERR$.

- regardless of confounding, the interval $UB - LB$ remains constant and depends on only one observable parameter, $\frac{P(y'|x)}{P(y|x)}$.

- $CF$ may raise the lower bound to meet the criterion of "more probable than not," $PN > \frac{1}{2}$, when the $ERR$ alone would not suffice.

- the amount of "rise" to both bounds is given by $CF$, which is the only estimate needed from the experimental data; the causal effect $P(y_x) - P(y_{x'})$ is not needed.



(a)

$$ERR \triangleq 1 - \frac{1}{RR} = 1 - \frac{P(y|x')}{P(y|x)}$$

$$q \triangleq \frac{P(y'|x)}{P(y|x)} \qquad CF \triangleq \frac{P(y|x') - p(Y_{x'})}{P(x,y)}$$

$$LB = ERR + CF \qquad\qquad UB = ERR + CF + q$$
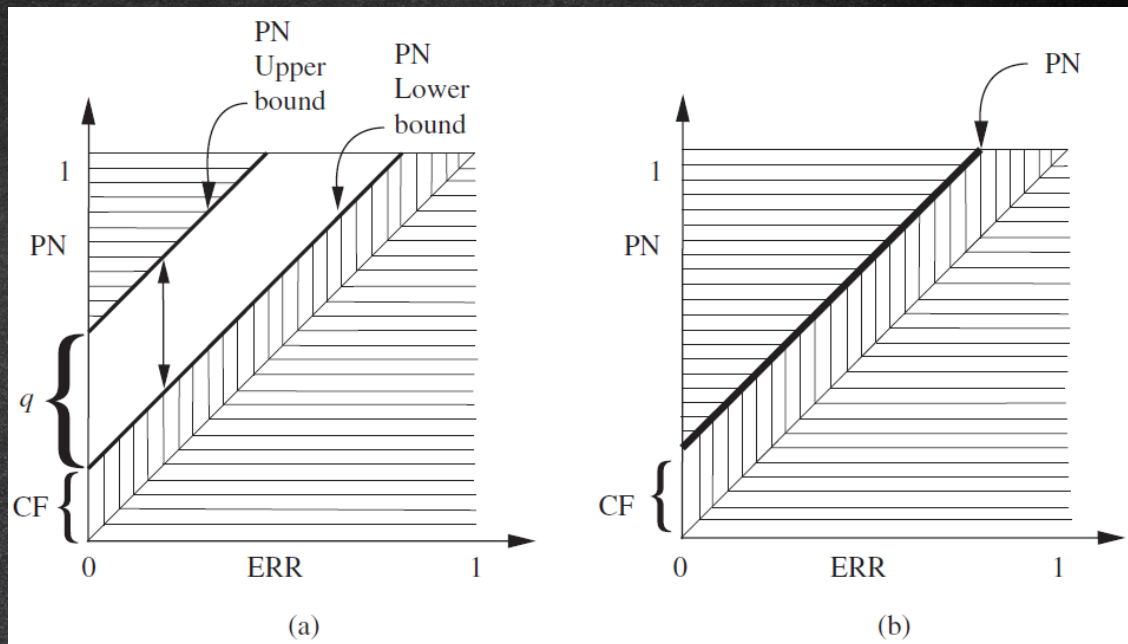
$$max\left\{0, \frac{P(y) - P(y|do(x'))}{P(x,y)}\right\} \leq PN \leq min\left\{1, \frac{P(y'|do(x')) - P(x',y')}{P(x,y)}\right\}$$

If monotonicity can be assumed, the upper and lower bounds coincide, and the gap collapses entirely, as shown in the figure(b) to the right.

This collapse does not reflect $q = 0$, but a shift from the bounds of bottom right to the identified condition of

$$PN = \frac{P(y) - P(y|do(x'))}{P(x,y)}$$

If it is the case that the experimental and survey data have been drawn at random from the same population, then the experimental data can be used to estimate the counterfactuals of interest, for example, $P(Y_x = y)$, for the observational as well as experimental sampled populations.



(a)                                               (b)

$$LB = ERR + CF \qquad\qquad UB = ERR + CF + q$$

$$max\left\{0, \frac{P(y) - P(y|do(x'))}{P(x,y)}\right\} \leq PN \leq min\left\{1, \frac{P(y'|do(x')) - P(x',y')}{P(x,y)}\right\}$$

## ATTRIBUTION IN LEGAL SETTING



- A lawsuit is filed against the manufacturer of drug $X$, charging that the drug is likely to have caused the death of Mr A, who took it to relieve back pains.

- The manufacturer claims that experimental data on patients with back pains show conclusively that drug $X$ has only minor effects on death rates.



- However, the plaintiff argues that the experimental study is of little relevance to this case because it represents average effects on patients in the study, not on patients like Mr A who did not participate in the study.
- In particular, argues the plaintiff, Mr A is unique in that he used the drug of his own volition, unlike subjects in the experimental study, who took the drug to comply with experimental protocols.

- To support this argument, the plaintiff furnishes nonexperimental data on patients who, like Mr A, chose drug $X$ to relieve back pains but were not part of any experiment, and who experienced lower death rates than those who didn't take the drug.

- The court must now decide, based on both the experimental and nonexperimental studies, whether it is "more probable than not" that drug $X$ was in fact the cause of Mr A's death.

**ATTRIBUTION IN LEGAL SETTING**

$$LB = ERR + CF \qquad UB = ERR + CF + q$$

To illustrate the usefulness of the bounds

$$max\left\{0, \frac{P(y) - P(y|do(x'))}{P(x,y)}\right\} \leq PN \leq min\left\{1, \frac{P(y'|do(x')) - P(x',y')}{P(x,y)}\right\}$$

Consider (hypothetical) data associated with the two studies shown in Table 14.5. (In the analyses below, we ignore sampling variability.)

The experimental data provide the estimates

| | Experimental | | Nonexperimental | |
|---|---|---|---|---|
| | $do(x)$ | $do(x')$ | $x$ | $x'$ |
| Deaths $(y)$ | 16 | 14 | 2 | 28 |
| Survivals $(y')$ | 984 | 986 | 998 | 972 |

Table 14.5

$$P(y|do(x)) = \frac{16}{1,000} = 0.016$$

$$P(y|do(x')) = \frac{14}{1,000} = 0.014$$

$$P(y) = \frac{30}{2,000} = 0.015 \qquad P(x,y) = \frac{2}{2,000} = 0.001$$

whereas the nonexperimental data provide the estimates

$$P(y|x) = \frac{2}{1,000} = 0.002 \qquad P(y|x') = \frac{28}{1,000} = 0.028$$

## ATTRIBUTION IN LEGAL SETTING

Assuming that drug $X$ can only cause (but never prevent) death, monotonicity holds, thus we can apply the equation to the right to obtain:

$$PN = \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y|do(x'))}{P(x, y)}$$

$$= \frac{0.002 - 0.028}{0.002} + \frac{0.028 - 0.014}{0.001}$$

The experimental data provide the estimates

$$P(y|do(x)) = \frac{16}{1,000} = 0.016$$

$$P(y|do(x')) = \frac{14}{1,000} = 0.014$$

$$P(y) = \frac{30}{2,000} = 0.015 \qquad P(x, y) = \frac{2}{2,000} = 0.001$$

whereas the nonexperimental data provide the estimates

$$P(y|x) = \frac{2}{1,000} = 0.002 \qquad P(y|x') = \frac{28}{1,000} = 0.028$$

## ATTRIBUTION IN LEGAL SETTING

Assuming that drug $X$ can only cause (but never prevent) death, monotonicity holds, thus we can apply the equation to the right to obtain:

$$PN = \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y|do(x'))}{P(x,y)}$$

$$= \underbrace{\frac{0.002 - 0.028}{0.002}}_{\substack{\textbf{EXCESS RISK RATIO} \\ \textbf{(ERR)} \\ \text{often used in court} \\ \text{cases in the} \\ \text{absence of} \\ \text{experimental data}}} + \underbrace{\frac{0.028 - 0.014}{0.001}}_{\substack{\textbf{CONFOUNDING FACTOR} \\ \textbf{(CF)} \\ \text{a correction needed to} \\ \text{account for confounding} \\ \text{bias, that is,} \\ P(y|do(x')) \neq P(y|x')}}$$

$$\underbrace{-13}_{\substack{\text{gives the} \\ \text{impression that} \\ \text{the drug } X \text{ is} \\ \text{actually} \\ \text{preventing deaths}}} \qquad \underbrace{14}_{\substack{\text{rectifies this} \\ \text{impression and} \\ \text{sets the probability} \\ \text{of necessity to} \\ PN = 1.}}$$

## ATTRIBUTION IN LEGAL SETTING

Moreover, since the lower bound of to the right becomes 1, we conclude that $PN = 1$ even without assuming monotonicity.

Thus, the plaintiff was correct; barring sampling errors, the data provide us with **100% assurance that drug $X$ was in fact responsible for the death of Mr A.**

$$LB = ERR + CF \qquad UB = ERR + CF + q$$

$$max\left\{0, \frac{P(y) - P(y|do(x'))}{P(x,y)}\right\} \leq PN \leq min\left\{1, \frac{P(y'|do(x')) - P(x',y')}{P(x,y)}\right\}$$

$$= \frac{0.002 - 0.028}{0.002} + \frac{0.028 - 0.014}{0.001}$$

**EXCESS RISK RATIO (ERR)**

often used in court cases in the absence of experimental data

$-13$

gives the impression that the drug $X$ is actually preventing deaths

**CONFOUNDING FACTOR (CF)**

a correction needed to account for confounding bias, that is, $P(y|do(x') \neq P(y|x')$

$14$

rectifies this impression and sets the probability of necessity to $PN = 1$.

The canonical model for a typical **MEDIATION PROBLEM** takes the form:

In Figure 14.6(a), the omitted factors are assumed to be arbitrarily distributed but mutually independent.
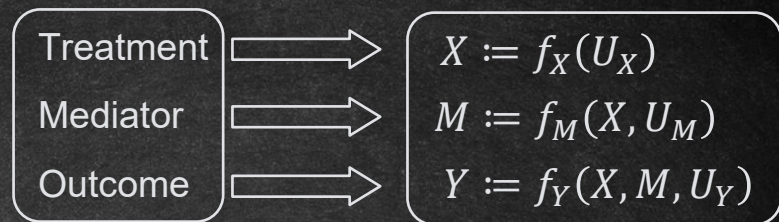
In Figure 14.6(b), the dashed arcs connecting $U_X$ and $U_M$ (as well as $U_M$ and $U_X$) encode the understanding that the factors in question may be dependent.
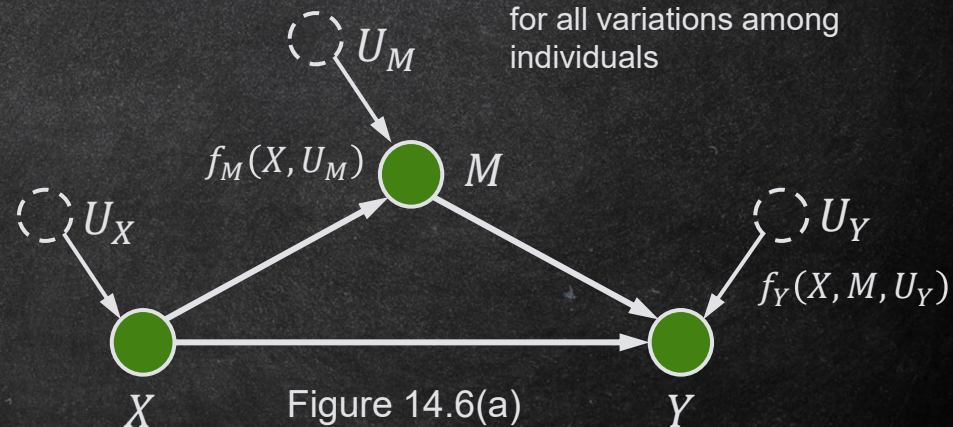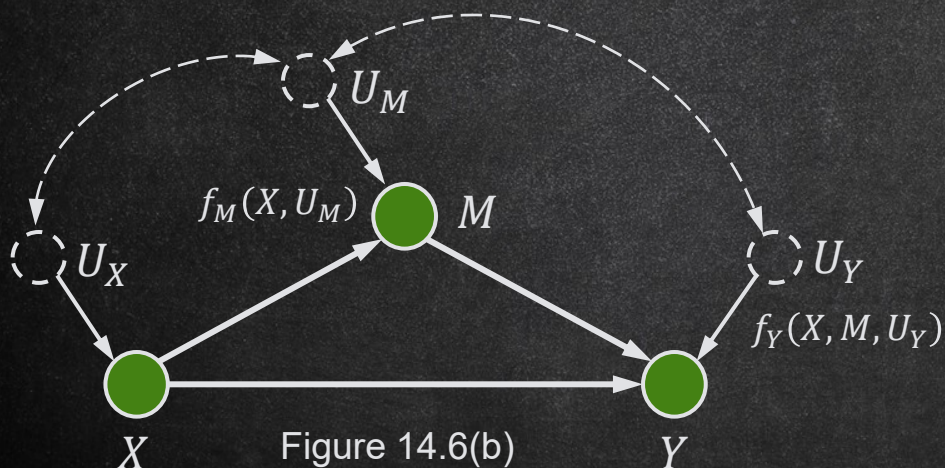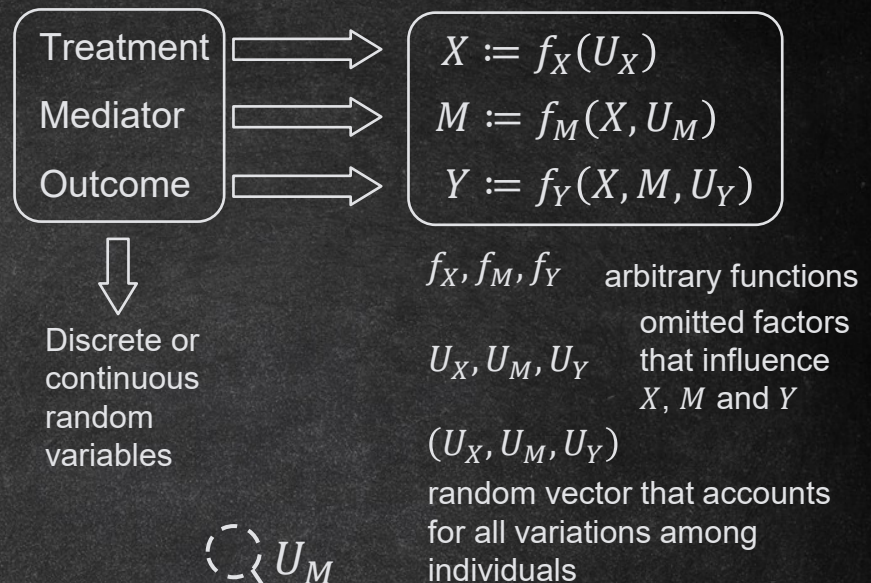
Treatment $\Longrightarrow$ $X := f_X(U_X)$

Mediator $\Longrightarrow$ $M := f_M(X, U_M)$

Outcome $\Longrightarrow$ $Y := f_Y(X, M, U_Y)$

$\Downarrow$

Discrete or continuous random variables

$f_X, f_M, f_Y$    arbitrary functions

$U_X, U_M, U_Y$    omitted factors that influence $X$, $M$ and $Y$

$(U_X, U_M, U_Y)$

random vector that accounts for all variations among individuals



$f_M(X, U_M)$   $M$

$U_M$

$U_X$

$U_Y$

$f_Y(X, M, U_Y)$

$X$    Figure 14.6(b)    $Y$

$f_M(X, U_M)$   $M$

$U_M$

$U_X$

$U_Y$

$f_Y(X, M, U_Y)$

$X$    Figure 14.6(a)    $Y$

## COUNTERFACTUAL DEFINITION OF DIRECT AND INDIRECT EFFECTS

Using the structural model to the right and the counterfactual notation, four types of effects can be defined for the transition from $X = 0$ to $X = 1$.

Generalizations to arbitrary reference points, say from $X = x$ to $X = x'$, are straightforward:

Treatment $\Longrightarrow$ $X := f_X(U_X)$

Mediator $\Longrightarrow$ $M := f_M(X, U_M)$

Outcome $\Longrightarrow$ $Y := f_Y(X, M, U_Y)$

$\Downarrow$

Discrete or continuous random variables

$f_X, f_M, f_Y$    arbitrary functions

$U_X, U_M, U_Y$    omitted factors that influence $X$, $M$ and $Y$

$(U_X, U_M, U_Y)$

random vector that accounts for all variations among individuals

### TOTAL EFFECT

$$TE = \mathbb{E}[Y_1 - Y_0] = \mathbb{E}[Y|do(X = 1)] - \mathbb{E}[Y|do(X = 0)]$$

$TE$ measures the expected increase in $Y$ as the treatment changes from $X = 0$ to $X = 1$, while the mediator is allowed to track the change in $X$ naturally, as dictated by the function $f_M$.
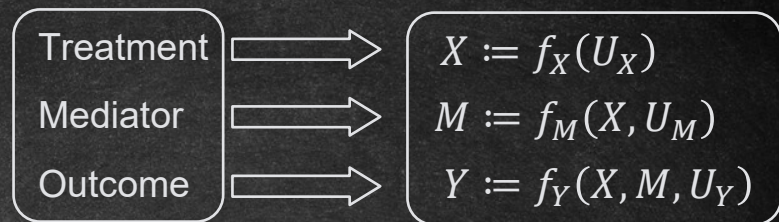
### CONTROLLED EFFECT

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

$CDE$ measures the expected increase in $Y$ as the treatment changes from $X = 0$ to $X = 1$, while the mediator is set to a specified level $M = m$ uniformly over the entire population.

## COUNTERFACTUAL DEFINITION OF DIRECT AND INDIRECT EFFECTS

Using the structural model to the right and the counterfactual notation, four types of effects can be defined for the transition from $X = 0$ to $X = 1$.

Generalizations to arbitrary reference points, say from $X = x$ to $X = x'$, are straightforward:

| Treatment | $X := f_X(U_X)$ |
| Mediator | $M := f_M(X, U_M)$ |
| Outcome | $Y := f_Y(X, M, U_Y)$ |

Discrete or continuous random variables

$f_X, f_M, f_Y$    arbitrary functions

$U_X, U_M, U_Y$    omitted factors that influence $X$, $M$ and $Y$

$(U_X, U_M, U_Y)$

random vector that accounts for all variations among individuals

### NATURAL DIRECT EFFECT

$$NDE = \mathbb{E}\big[Y_{1,M_0} - Y_{0,M_0}\big]$$

$NDE$ measures the expected increase in $Y$ as the treatment changes from $X = 0$ to $X = 1$, while the mediator is set to whatever value it would have attained (for each individual) prior to the change, that is, under $X = 0$.

### NATURAL INDIRECT EFFECT

$$NIE = \mathbb{E}\big[Y_{0,M_1} - Y_{0,M_0}\big]$$

$NIE$ measures the expected increase in $Y$ when the treatment is held constant, at $X = 0$, and $M$ changes to whatever value it would have attained (for each individual) under $X = 1$. It captures, therefore, the portion of the effect that can be explained by mediation alone, while disabling the capacity of $Y$ to respond to $X$.

We note that, in general, the **TOTAL EFFECT** can be decomposed as

$$TE = \mathbb{E}[Y_1 - Y_0] = \mathbb{E}[Y|do(X = 1)] - \mathbb{E}[Y|do(X = 0)]$$

This implies that $NIE$ is identifiable whenever $NDE$ and $TE$ are identifiable.

$$TE = NDE - NIE_r \qquad NIE_r = \mathbb{E}[Y_{0,M_0} - Y_{0,M_1}]$$

In linear systems, where reversal of transitions amounts to negating the signs of their effects, we have the standard additive formula to the right.

$$TE = NDE + NIE$$

We further note that $TE$ and $CDE(m)$ are do-expressions and can, therefore, be estimated from experimental data or in observational studies using the backdoor or front-door adjustments.

**CONDITIONS FOR IDENTIFYING NATURAL EFFECTS**

The following set of conditions, are sufficient for identifying both direct and indirect natural effects.

Not so for the $NDE$ and $NIE$; a new set of assumptions is needed for their identification.

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

i.   No member of **W** is a descendant of $X$.

ii.  **W** blocks all backdoor paths from $M$ to $Y$ (after removing $X \rightarrow M$ and $X \rightarrow Y$).

iii. The **W**-specific effect of $X$ on $M$ is identifiable (possibly using experiments or adjustments).

iv.  The **W**- specific joint effect of $\{X, M\}$ on $Y$ is identifiable (possibly using experiments or adjustments).

## IDENTIFICATION OF THE NDE

When conditions $i$) and $ii$) hold, the natural direct effect is experimentally identifiable and is given by

$$NDE = \sum_m \sum_w \big[ \mathbb{E}[Y|do(X=1, M=m), \mathbf{W}=\boldsymbol{w}] - \mathbb{E}[Y|do(X=0, M=m), \mathbf{W}=\boldsymbol{w}] \big] \times P(M=m|do(X=0), \mathbf{W}=\boldsymbol{w}) \, P(\mathbf{W}=\boldsymbol{w})$$

The identifiability of the do-expressions in the above equation is guaranteed by conditions $iii$) and $iv$) and can be determined using the backdoor or front-door criteria.
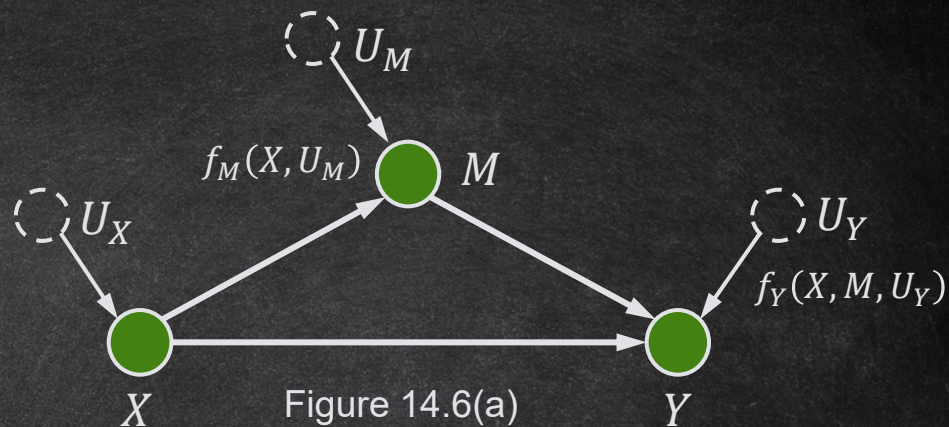
## IDENTIFICATION OF THE NDE

If conditions $i$) and $ii$) are satisfied by a set $\mathbf{W}$ that also deconfounds the relationships in $iii$) and $iv$), then the do-expressions in the above equation are reducible to conditional expectations, and the natural direct effect becomes

$$NDE = \sum_m \sum_w \big[ \mathbb{E}[Y|X=1, M=m, \mathbf{W}=\boldsymbol{w}] - \mathbb{E}[Y|X=0, M=m, \mathbf{W}=\boldsymbol{w}] \big] \times P(M=m|X=0, \mathbf{W}=\boldsymbol{w}) \, P(\mathbf{W}=\boldsymbol{w})$$

In the nonconfounding case (Figure 14.6(a)), $NDE$ reduces to

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

Figure 14.6(a)

**IDENTIFICATION OF THE NDE**

If conditions *i*) and *ii*) are satisfied by a set **W** that also deconfounds the relationships in *iii*) and *iv*), then the do-expressions in the above equation are reducible to conditional expectations, and the natural direct effect becomes

$$NDE = \sum_m \sum_{\boldsymbol{w}} \big[ \mathbb{E}[Y|X = 1, M = m, \mathbf{W} = \boldsymbol{w}] - \mathbb{E}[Y|X = 0, M = m, \mathbf{W} = \boldsymbol{w}] \big] \times P(M = m|X = 0, \mathbf{W} = \boldsymbol{w}) \, P(\mathbf{W} = \boldsymbol{w})$$

In the nonconfounding case (Figure 14.6(a)), $NDE$ reduces to

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

Similarly, using $\quad TE = NDE - NIE_r \implies$

and $\quad TE = \mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0]$

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

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

**MEDIATION FORMULAS**

**IDENTIFICATION OF THE NDE**

If conditions *i*) and *ii*) are satisfied by a set $\mathbf{W}$ that also deconfounds the relationships in *iii*) and *iv*), then the do-expressions in the above equation are reducible to conditional expectations, and the natural direct effect becomes

$$NDE = \sum_m \sum_{\mathbf{w}} \big[\mathbb{E}[Y|X=1, M=m, \mathbf{W}=\mathbf{w}] - \mathbb{E}[Y|X=0, M=m, \mathbf{W}=\mathbf{w}]\big] \times P(M=m|X=0, \mathbf{W}=\mathbf{w}) \, P(\mathbf{W}=\mathbf{w})$$

In the nonconfounding case (Figure 14.6(a)), $NDE$ reduces to

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

We see that while $NDE$ is a weighted average of $CDE$, no such interpretation can be given to $NIE$.

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

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

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

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

The counterfactual definitions of $NDE$ and $NIE$

$$NDE = \mathbb{E}\big[Y_{1,M_0} - Y_{0,M_0}\big]$$

$$NIE = \mathbb{E}\big[Y_{0,M_1} - Y_{0,M_0}\big]$$

permit us to give these effects meaningful interpretations in terms of "**RESPONSE FRACTIONS**."

- $\dfrac{NDE}{TE}$ measures the fraction of the response that is transmitted directly, with $M$ "frozen."

- $\dfrac{NIE}{TE}$ measures the fraction of the response that may be transmitted through $M$, with $Y$ blinded to $X$.

Consequently,

- $\dfrac{TE - NDE}{TE}$ measures the fraction of the response that is necessarily due to $M$.

**ENCOURAGEMENT DESIGN** (Figure 14.1).

- $X$; amount of time a student spends in an after-school remedial program,

- $H$; the amount of homework a student does, and

- $Y$; a student's score on the exam.

For example, if $Y = 1$, then the student scored 1 standard deviation above the mean on his or her exam.

This model represents a randomized pilot program, in which students are assigned to the remedial sessions by the luck of the draw.
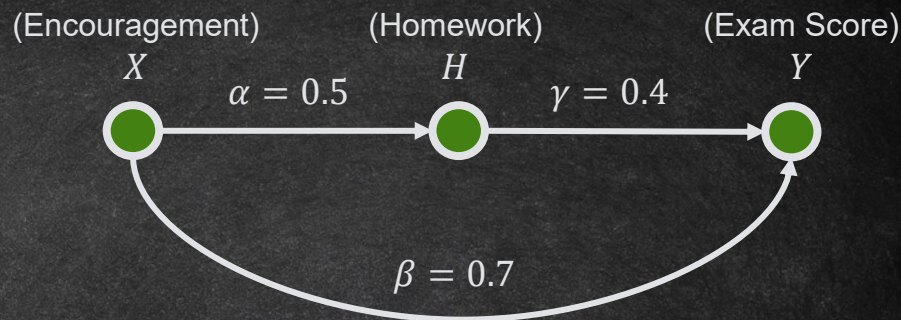


Figure 14.1

$$X := U_X$$
$$H := \alpha X + U_H$$
$$Y := \beta X + \gamma H + U_Y$$
$$\sigma_{U_i U_j} = 0, \forall i, j \in \{X, H, Y\}$$

$$\alpha = 0.5$$
$$\beta = 0.7$$
$$\gamma = 0.4$$

given or recovered from population data

## NUMERICAL EXAMPLE: MEDIATION WITH BINARY VARIABLES

To anchor these mediation formulas in a concrete example, we return to the encouragement design example and assume that

- $X = 1$ stands for participation in an enhanced training program,
- $Y = 1$ for passing the exam, and
- $M = 1$ for a student spending more than 3 hours per week on homework.
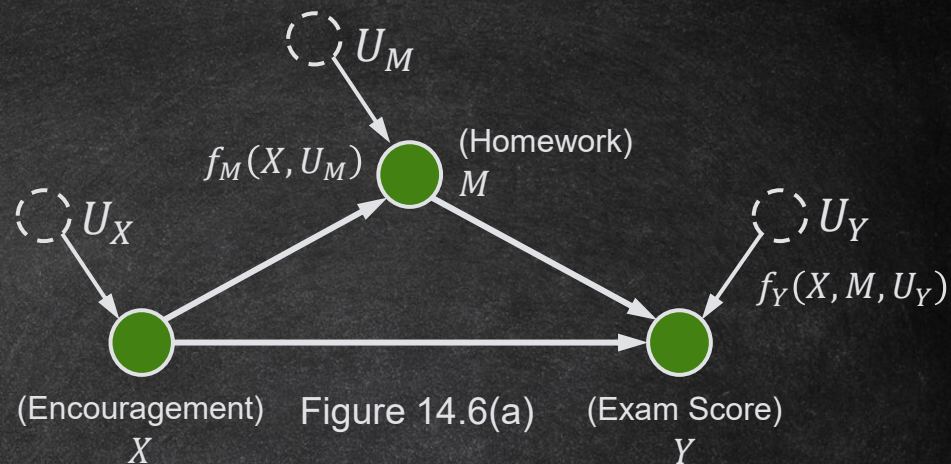
Figure 14.6(a)

Assume further that the data described in Tables 14.6 and 14.7 were obtained in a randomized trial with no mediator-to-outcome confounding (Figure 14.6(a)).

| Treatment $X$ | Homework $M$ | Success Rate $\mathbb{E}[Y \mid X = x, M = m]$ |
|---|---|---|
| 1 | 1 | 0.8 |
| 1 | 0 | 0.4 |
| 0 | 1 | 0.3 |
| 0 | 0 | 0.2 |

Table 14.6

| Treatment $X$ | Homework $\mathbb{E}[M \mid X = x]$ |
|---|---|
| 0 | 0.40 |
| 1 | 0.75 |

Table 14.7

Our research question asks for the extent to which students' homework contributes to their increased success rates regardless of the training program.

The policy implications of such questions lie in evaluating policy options that either curtail or enhance homework efforts, for example, by counting homework effort in the final grade or by providing students with adequate work environments at home.

- An extreme explanation of the data, with significant impact on educational policy, might argue that the program does not contribute substantively to students' success, save for encouraging students to spend more time on homework, an encouragement that could be obtained through less expensive means.

- Opposing this theory, we may have teachers who argue that the program's success is substantive, achieved mainly due to the unique features of the curriculum covered, whereas the increase in homework efforts cannot alone account for the success observed.

| Treatment | Homework | Success Rate |
|-----------|----------|--------------|
| $X$ | $M$ | $\mathbb{E}[Y \mid X = x, M = m]$ |
| 1 | 1 | 0.8 |
| 1 | 0 | 0.4 |
| 0 | 1 | 0.3 |
| 0 | 0 | 0.2 |

Table 14.6

| Treatment | Homework |
|-----------|----------|
| $X$ | $\mathbb{E}[M \mid X = x]$ |
| 0 | 0.40 |
| 1 | 0.75 |

Table 14.7

Our research question asks for the extent to which students' homework contributes to their increased success rates regardless of the training program.

Substituting the data into ⟹

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

$$NDE = \sum_m \left[ \mathbb{E}[Y|X = 1, M = m ] - \mathbb{E}[Y|X = 0, M = m ]\right] \times P(M = m|X = 0)$$

$NDE = (0.40 - 0.20)(1 - 0.40) + (0.80 - 0.30)\ 0.40 = 0.32$

$NIE = (0.75 - 0.40)(0.30 - 0.20) = 0.035$

■ $\dfrac{NDE}{TE}$ measures the fraction of the response that is transmitted directly, with $M$ "frozen."

$TE = 0.80 \times 0.75 + 0.40 \times 0.25 - (0.30 \times 0.40 + 0.20 \times 0.10) = 0.46$ ⟹ the program as a whole has increased the success rate by 46%

$\dfrac{NIE}{TE} = 0.07,$ $\dfrac{NDE}{TE} = 0.696,$ $1 - \dfrac{NDE}{TE} = 0.304$ ⟹ a significant portion, 30.4%, of this increase is due to the capacity of the program to stimulate improved homework effort.

| Treatment $X$ | Homework $M$ | Success Rate $\mathbb{E}[Y|X = x, M = m ]$ |
|---|---|---|
| 1 | 1 | 0.8 |
| 1 | 0 | 0.4 |
| 0 | 1 | 0.3 |
| 0 | 0 | 0.2 |

Table 14.6

| Treatment $X$ | Homework $\mathbb{E}[M|X = x]$ |
|---|---|
| 0 | 0.40 |
| 1 | 0.75 |

Table 14.7

Our research question asks for the extent to which students' homework contributes to their increased success rates regardless of the training program.

Substituting the data into ⟹

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

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

At the same time, only 7% of the increase (46%) can be explained by stimulated homework alone without the benefit of the program itself

- $\dfrac{NIE}{TE}$ measures the fraction of the response that may be transmitted through $M$, with $Y$ blinded to $X$.

$$\frac{NIE}{TE} = 0.07, \qquad \frac{NDE}{TE} = 0.696, \qquad 1 - \frac{NDE}{TE} = 0.304 \Longrightarrow$$

a significant portion, 30.4%, of this increase is due to the capacity of the program to stimulate improved homework effort.

| Treatment | Homework | Success Rate |
|---|---|---|
| $X$ | $M$ | $\mathbb{E}[Y|X = x, M = m]$ |
| 1 | 1 | 0.8 |
| 1 | 0 | 0.4 |
| 0 | 1 | 0.3 |
| 0 | 0 | 0.2 |

Table 14.6

| Treatment | Homework |
|---|---|
| $X$ | $\mathbb{E}[M|X = x]$ |
| 0 | 0.40 |
| 1 | 0.75 |

Table 14.7

Our research question asks for the extent to which students' homework contributes to their increased success rates regardless of the training program.