Scenario

Example

"I took bus 12 home yesterday, but it was very crowded and I had to wait a long time. If I had taken another bus line, I would have been home earlier."

Question: Can we express this using $\mathrm{do}(\cdot)$-notation?

Counterfactuals: definition

counterfactual: a statement involving a condition (the *antecedent*) that is contrary to fact

E.g. my travel time yesterday if taking bus x', when really I took bus x

Refers to outcome in a ‘different world’: same as the real world in all other ways (similar to perfect intervention)

Counterfactuals vs. interventions: distributions vs. values

Interventions allow us to talk about changes in the *distribution* of a population

Counterfactuals let us talk about changes in *values* for individuals

Notation

The value of Y, had X been 1

Notation:

$$Y_{X=1} \qquad \text{(often abbreviated to } Y_1 \text{ if clear from context)}$$

Here:

- $X = 1$ is the antecedent (can refer to endo-/exogenous variables)
- Y is an endogenous variable

Counterfactuals vs. interventions: generalization

Counterfactuals generalize interventions: any interventional quantity can be written as a counterfactual

A counterfactual simplifies to an interventional quantity if it does not include contradictory conditions:

- $P(Y_{X=x} = y \mid Z_{X=x} = z)$ simplifies to $P(Y = y \mid Z = z, \mathrm{do}(X = x))$
- $P(Y_{X=x} = y \mid Z = z)$ does not simplify in general: Z and Y_x exist in different 'worlds'

Consistency rule

Consistency rule:

$$\text{if } X = x \text{ then } Y_x = Y$$

Expresses that if the 'counter'factual is actually in agreement with fact, then it reduces to the original random variable

Level 3

Counterfactuals form a third level in the following hierarchy:

Associations	$P(Y \mid X = x)$	What does a symptom tell me about a disease?
Interventions	$P(Y \mid \mathrm{do}(X = x))$	What if I take aspirin, will my headache be cured?
Counterfactuals	$P(Y_{X=x})$	Was it aspirin that cured my headache?

Warning

Interventional quantities can be measured directly in experiments (at least in theory)

Counterfactuals can never be measured directly!

We may be able to measure a surrogate, e.g.

- in a similar situation at a different time, or
- in a different but similar individual

... but without additional assumptions these might turn out ot be quite different from the counterfactual itself

Counterfactuals in SCMs

As with interventions, structural causal models allow us to talk about counterfactuals

In SCM terms, “the same” unit/individual/situation translates to: the same values of all exogenous variables ($U = u$; u is a vector)

If all the exogenous values are known, so are the values of all endogenous variables, so we may write e.g. $X(u)$

Computing counterfactuals (individual case): example 1/2

Original model:

$$\begin{aligned} X &= aU \\ Y &= bX + U \end{aligned}$$

To compute $Y_{X=x}(u)$: look at modified model

$$\begin{aligned} X &= x \\ Y &= bX + U \end{aligned}$$

Fill in x and u:

$$Y_{X=x}(u) = bx + u$$

Computing counterfactuals (individual case): example 2/2

Now compute $X_{Y=y}(u)$ instead of $Y_{X=x}(u)$. Original model:

$$X = aU$$
$$Y = bX + U$$

To compute $X_{Y=y}(u)$: look at modified model

$$X = aU$$
$$Y = y$$

Fill in (y and) u:

$$X_{Y=y}(u) = au$$

Computing counterfactuals (individual case): steps

Steps:

1. Find u (using the original model)
2. Do an intervention (modifying the model)
3. Compute the value of the counterfactual quantity

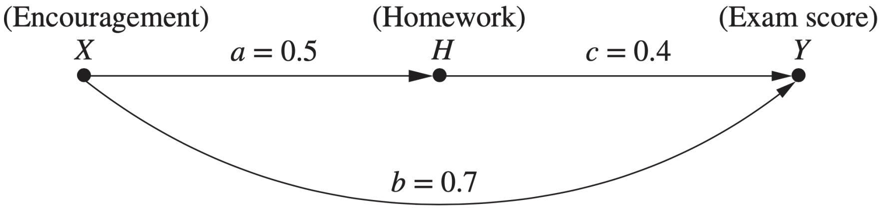


Figure 4.1 A model depicting the effect of Encouragement (X) on student's score

$$X = U_X$$
$$H = 0.5 \cdot X + U_H$$
$$Y = 0.7 \cdot X + 0.4 \cdot H + U_Y$$

Consider Joe, for whom we measure:

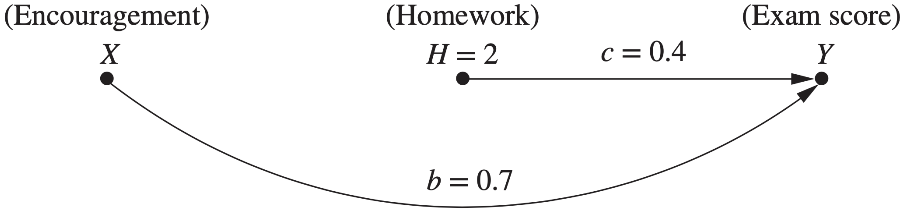
$$X = 0.5 \qquad H = 1 \qquad Y = 1.5$$

Question: would would Joe's exam score have been if he had doubled his study time ($H = 2$)?

First, find the values of U:

$$U_X = 0.5$$
$$U_H = 1 - 0.5 \cdot 0.5 = 0.75$$
$$U_Y = 1.5 - 0.7 \cdot 0.5 - 0.5 \cdot 1 = 0.75$$

Next, modify the model:



Finally, compute

$$Y_{H=2}(u) = 0.5 \cdot 0.7 + 2.0 \cdot 0.4 + 0.75 = 1.9$$

Counterfactuals: probabilistic case

In practice, we may not know enough to identify u exactly:

- we might be interested in an individual, but not all variables are observed
- we might be interested in a subpopulation, e.g. all individuals with $Y < 2$

The evidence $E = e$ we have leaves a probability distribution on U in the first step of computing a counterfactual: $P(U \mid E = e)$

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

u	P(u)
1	1/2
2	1/3
3	1/6