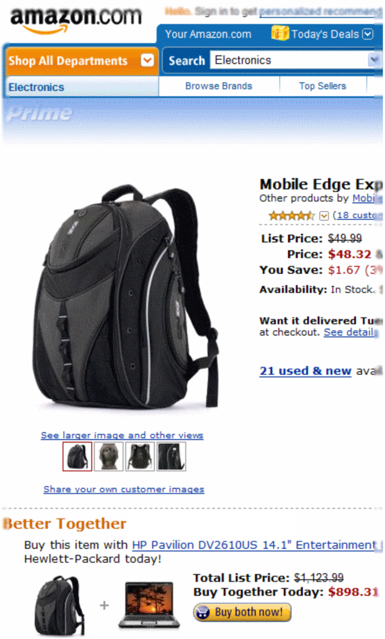
StatML.io CDT: Causality Module

Robin J. Evans

Imperial College London and University of Oxford

February 2024

Example



Mobile Edge Exp
Other products by Mobi
(18 custo

List Price: ~~$49.99~~
Price: $48.32
You Save: $1.67 (3
Availability: In Stock.

Want it delivered Tue
at checkout. See detail

21 used & new ava

Better Together

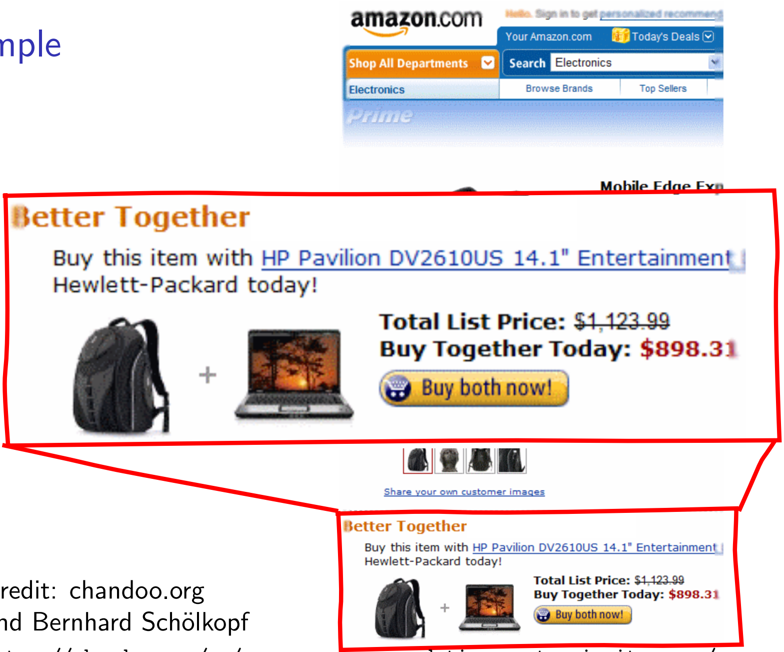
Buy this item with HP Pavilion DV2610US 14.1" Entertainment
Hewlett-Packard today!

Total List Price: ~~$1,123.99~~
Buy Together Today: $898.31

Credit: chandoo.org
and Bernhard Schölkopf

`https://chandoo.org/wp/amazons-recommendation-system-is-it-crazy/`

Example



Better Together

Buy this item with HP Pavilion DV2610US 14.1" Entertainment
Hewlett-Packard today!

Total List Price: ~~$1,123.99~~
Buy Together Today: $898.31

Buy both now!

Better Together

Buy this item with HP Pavilion DV2610US 14.1" Entertainment
Hewlett-Packard today!

Total List Price: ~~$1,123.99~~
Buy Together Today: $898.31

Buy both now!

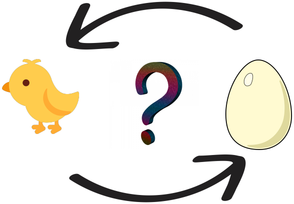
Credit: chandoo.org
and Bernhard Schölkopf

`https://chandoo.org/wp/amazons-recommendation-system-is-it-crazy/`

Causality

'Causality' is a very large (by no means entirely statistical) topic, but encompasses two important subfields:

- **causal discovery:** determining causal structure from observational data;



- **causal inference:** estimating causal effects from data, given the structure.

Learning objectives

The plan for these two weeks is to introduce basic concepts of causal learning (reasoning, modelling, and inference) to enable you to read more advanced ‘causal’ papers.

We will focus on:

- formulating causal (research) questions;
- understanding sources of (avoidable and unavoidable) bias;
- some basic inference methods, such as adjustment and inverse weighting;
- a couple of ML methods (e.g. double machine learning).

Please **ASK** if you have question or comments.

Outline

1. Introduction

2. Single-World Intervention Graphs

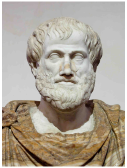
3. Estimation Methods

4. Machine Learning Methods

History of Causal Inference

Causal inference is a topic with a history almost as long as history itself!

Aristotle is generally credited as the earliest to consider this question **philosophically**.



- There are distinct traditions in different disciplines. Notably, medicine, epidemiology, econometrics, and psychiatry.
- These approaches have different terminology, accepted assumptions and sources of data.
- Only in the past few years has some convergence emerged across fields.
- Causality is **fundamental** to many research questions in different scientific fields.

Statistical Causality

(Associational) statistics asks ‘what?’

Causality asks ‘why?’ and ‘how?’ and ‘what if?’

Causation / causality: philosophical, moral and other usages of the term—not what we are concerned with here.

This module takes particular (narrow) view of causality most relevant for scientific enquiries: causality we can **implement**.

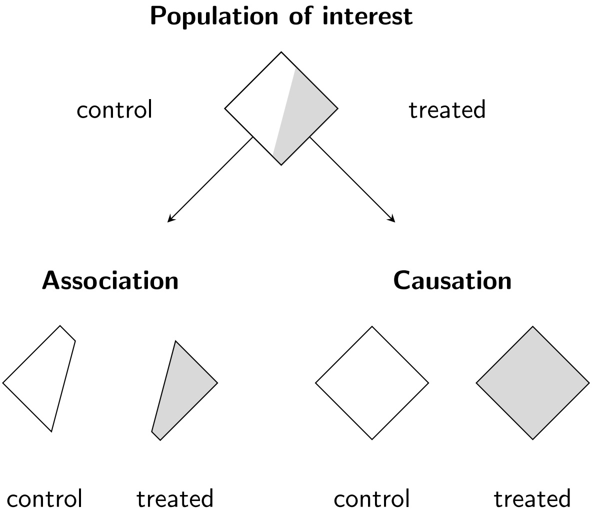
We are interested in a ‘causal effect’; that is, a difference in outcomes, or their distribution, between (hypothetical) experiments we might do.

Basic Causal Concepts

1. Introduction

- Basic Causal Concepts
- Motivation
- Conditional Independence
- Directed Acyclic Graphs
- Confounding and Adjustment
- Selection Bias
- Potential Outcomes
- Other Causal Effects

Causation vs Association



Population of interest

control

treated

Association

Causation

control

treated

control

treated

(From Hernán and Robins, 2025)

Causal Questions

Descriptive or predictive questions

(A) "Is this patient suitable for surgery?"

(B) "Is this patient at high risk of developing complications during surgery?"

Causal questions

(C) "Which type of anaesthetic should this patient receive to minimise the risk of complications during surgery?"

(C') "How does the amount of anaesthetic affect the risk of complications during surgery?"

(D) "What can be done to reduce the risk of complications during surgery for an average / a particular type of patient?"

Causal Questions

Much of the art of causal inference lies in:

1. formulating a **causal** question;
2. turning that causal question into something that can be answered with data (i.e. a **statistical** question.)

The general strategy is:

1. determine a **causal** quantity that will answer the scientific question of interest;
2. check that the quantity is **identifiable** from the data you have and assumptions you are willing to make;
3. choose a **statistical** estimator based on your data and assumptions.

Target Trials

Formulate the **ideal randomized clinical trial** that you would use to answer your question. Use the usual trial design criteria (e.g. Hernán and Robins, 2016):

PICOT

Population: identify which subjects should be included/excluded.

Intervention: what is the treatment you wish to investigate?

Comparison: what will you use as a baseline?

Outcome: what outcome measure will you consider?

Time: the horizon over which comparisons should be made.

Your (likely observational) data can be manipulated to fit with the answers you give. See Hernán et al. (2008) for an example applied to hormone replacement therapy.

Chris will talk about this more next week.

References

Hernán, M. A., ... & Robins, J. M. (2008). Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*, 19(6), 766-779.

Hernán, M. A. and Robins, J. M. *What if.* CRC Press, 2025(?)

Hernán, M. A. and Robins, J. M. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology*, 183 (8): 758–764, 2016.

Motivation

1. Introduction

- Basic Causal Concepts
- Motivation
- Conditional Independence
- Directed Acyclic Graphs
- Confounding and Adjustment
- Selection Bias
- Potential Outcomes
- Other Causal Effects

A Causal Story

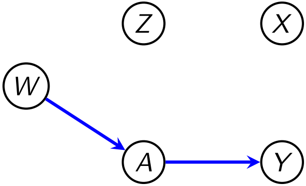
Consider the following situation.

> *An obstetrician is interested in whether giving a vitamin A supplement (A) to new mothers may help to reduce the risk of post-natal depression (Y). She implements an encouragement W to take such supplements, by offering it to a randomly selected subset of half the new mothers in her ward. This is assumed not to have any effect other than increasing the change of mothers taking vitamin A.*
>
> *She suspects that age (Z) is also a determinant of how likely a mother is to take the supplement, and that this also affects the baby's health (X). This in turn affects the likelihood of post-natal depression, but not the probability of taking the supplement.*
>
> *There is assumed to be no direct effect of age on post-natal depression.*

How should we represent the information contained in this paragraph?

Directed Graphs

Use a graph!



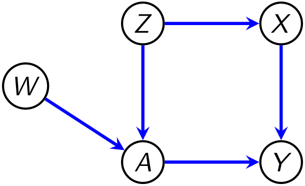
- Z age;
- W encourage;
- X infant health;
- A vitamin A;
- Y post-natal depression;

$[Y]$ is assumed to be directly affected by the treatment (A).

W is **randomly** assigned, and affects only A.

Directed Graphs

Use a graph!



- Z age;
- W encourage;
- X infant health;
- A vitamin A;
- Y post-natal depression;

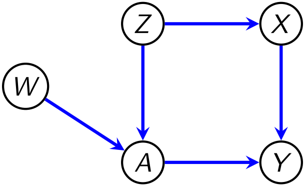
Z may determine A and X, but not Y.

X is predictive of Y, but does not affect A.

Interpreting a Graph

Now that we've drawn a graph, we get a nice representation of the (possible) causal structure underlying our data.

We can immediately see which paths are **causal** (they're directed!) and which are not.



This is useful **even** if we do not intend to use the graph for inference!

Conditional Independence

1. Introduction

- Basic Causal Concepts
- Motivation
- Conditional Independence
- Directed Acyclic Graphs
- Confounding and Adjustment
- Selection Bias
- Potential Outcomes
- Other Causal Effects

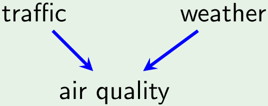
Independence

Definition

Given two random variables X and Y defined on a Cartesian product space, we say that X is **independent** of Y under p (denoted $X \perp\!\!\!\perp Y\,[p]$) if

$$p(x \,|\, y) = p(x).$$

Example. Air quality is adversely affected by both traffic and local weather conditions, but these two factors may not be related.



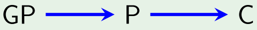
Conditional Independence

Definition

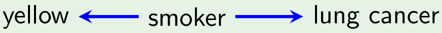
Given two random variables X and Y defined on a Cartesian product space, and a third variable Z, we say that X is **conditionally independent** of Y given Z under p (denoted $X \perp\!\!\!\perp Y \mid Z\,[p]$) if

$$p(x \mid y, z) = p(x \mid z).$$

Example. People's genes are conditionally independent of their grandparents's genes, given their parents's genes.



Example. Lung cancer is conditionally independent of having yellow fingers, given one's smoking status.



Alternative Characterizations

Theorem

Let X, Y, Z be random variables, with joint density p. Then we can write

$$p(x, y, z) = f(x, z) \cdot g(y, z)$$

if and only if $X \perp\!\!\!\perp Y \mid Z\,[p]$.

This can be very useful if we only know the density up to a constant of proportionality.

Simpson's Paradox

Conditional independence is sometimes quite unintuitive.

Below is the margin of an infamous dataset on death penalty convictions in Florida between 1976 and 1987.

Death Penalty?	Defendant's Race	
	White	Black
Yes	53	15
No	430	176

White defendants are slightly more likely than black defendants to face the death penalty.

Simpson's Paradox

Here is the full dataset.

Victim's Race	Death Penalty?	Defendant's Race	
		White	Black
White	Yes	53	11
	No	414	37
Black	Yes	0	4
	No	16	139

Now we can see that if we condition on the victim's race, the dependence of the penalty applied conditional on the defendant's race is completely reversed!

Morals

Let:

- D be an indicator that the death penalty was imposed;
- V be an indicator for the race of the victim;
- R be an indicator for the race of the defendant.

By changing the numbers only very slightly, it is easy to obtain either:

$$
\begin{array}{llcl}
 & D \perp\!\!\!\perp R & \text{and} & D \not\perp\!\!\!\perp R \mid V, \\[1em]
\text{or} & D \not\perp\!\!\!\perp R & \text{and} & D \perp\!\!\!\perp R \mid V.
\end{array}
$$

Graphoids

Conditional independences obey several rules called **semi-graphoid axioms** (though in this context they are not really axioms!) These are:

1. Symmetry: $X \perp\!\!\!\perp Y \mid Z \implies Y \perp\!\!\!\perp X \mid Z$;
2. Decomposition: $X \perp\!\!\!\perp Y, W \mid Z \implies X \perp\!\!\!\perp Y \mid Z$;
3. Weak union: $X \perp\!\!\!\perp Y, W \mid Z \implies X \perp\!\!\!\perp W \mid Y, Z$;
4. Contraction: $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid Y, Z \implies X \perp\!\!\!\perp Y, W \mid Z$;

We can summarize axioms 2–4 as a 'chain rule':

$$X \perp\!\!\!\perp Y \mid Z \quad \text{and} \quad X \perp\!\!\!\perp W \mid Y, Z \quad \iff \quad X \perp\!\!\!\perp W, Y \mid Z.$$

In addition, if $p > 0$ then we have:

5. Intersection: $X \perp\!\!\!\perp Y \mid W, Z$ and
   $X \perp\!\!\!\perp W \mid Y, Z \implies X \perp\!\!\!\perp Y, W \mid Z$.

All five rules are called the **graphoid axioms**.

Directed Acyclic Graphs

1. Introduction

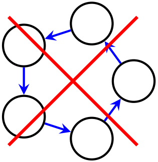
- Basic Causal Concepts
- Motivation
- Conditional Independence
- Directed Acyclic Graphs
- Confounding and Adjustment
- Selection Bias
- Potential Outcomes
- Other Causal Effects

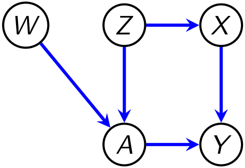
Directed Acyclic Graphs

vertices $V \in \boldsymbol{V}$

edges →

no directed cycles





directed acyclic graph (DAG), $\mathcal{G}$

We will associate the vertices/nodes with **random variables**, and the edges will denote **causal** dependence.

Terminology and Notation

For a DAG $\mathcal{G}$ with vertices $\boldsymbol{V}$...

If ...	we say...	and we write...
$X \to Y$	X is a **parent** of Y Y is a **child** of X	$X \in \mathrm{pa}_{\mathcal{G}}(Y)$ $Y \in \mathrm{ch}_{\mathcal{G}}(X)$
$X \to \cdots \to Y$ or $X = Y$	X is an **ancestor** of Y Y is a **descendant** of X	$X \in \mathrm{an}_{\mathcal{G}}(Y)$ $Y \in \mathrm{de}_{\mathcal{G}}(X)$

A **path** is a sequence of adjacent edges, without repeating a vertex.

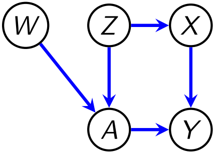
A **directed path** is a path where all the edges are oriented pointing towards the final vertex.

A **directed cycle** is a directed path from X to Y and an edge $Y \to X$.

Given a **topological ordering** (parents precede their children) of the variables $V_1, \dots, V_p$ we write $\mathrm{pre}_<(i) = \{1, \dots, i-1\}$ for each i.

DAG Models (aka Bayesian Networks)

graph $\mathcal{G}$



$\iff$

model $\mathcal{M}(\mathcal{G})$

$$p(\boldsymbol{v}) = \prod_{v \in V} p(v \mid \mathrm{pa}_{\mathcal{G}}(v)).$$

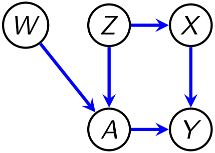
(factorization)

So in example above:

$$p(\boldsymbol{v}) = p(w) \cdot p(z) \cdot p(x \mid z) \cdot p(a \mid w, z) \cdot p(y \mid a, x).$$

DAG Models (aka Bayesian Networks)

Can also define model as a list of conditional independences:



pick a topological ordering $<$ of the graph: e.g. W, Z, X, A, Y.

Can **always** factorize a joint distribution as:

$$p(w, z, a, x, y) = p(w) \cdot p(z \mid w) \cdot p(a \mid w, z) \cdot p(x \mid w, z, a) \cdot p(y \mid w, z, a, x).$$

The model is the same as setting (e.g.)

$$p(y \mid w, z, a, x) = p(y \mid a, x) = p(y \mid \mathrm{pa}(y)).$$

Thus $\mathcal{M}(\mathcal{G})$ is precisely distributions such that:

$$V_i \perp\!\!\!\perp \boldsymbol{V}_{\mathrm{pre}_<(i) \setminus \mathrm{pa}(i)} \mid \boldsymbol{V}_{\mathrm{pa}(i)}, \qquad\qquad i = 1, \ldots, |\boldsymbol{V}|.$$

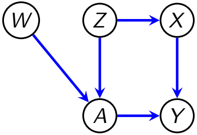
Ordered Markov Property

We say that p obeys the **ordered local Markov property** with respect to $\mathcal{G}$ and a topological ordering $<$ if:

$$V_i \perp\!\!\!\perp \boldsymbol{V}_{\mathrm{pre}_<(i)\setminus\mathrm{pa}(i)} \mid \boldsymbol{V}_{\mathrm{pa}(i)}, \qquad\qquad i = 1, \ldots, |\boldsymbol{V}|.$$

In our example, with the order W, Z, X, A, Y this means

$$Z \perp\!\!\!\perp W \qquad\qquad X \perp\!\!\!\perp W \mid Z$$
$$A \perp\!\!\!\perp X \mid W, Z \qquad Y \perp\!\!\!\perp W, Z \mid A, X.$$



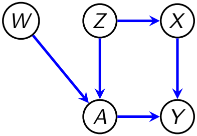
If we switch A and X, we get

$$Z \perp\!\!\!\perp W \qquad\qquad X \perp\!\!\!\perp A, W \mid Z \qquad\qquad Y \perp\!\!\!\perp W, Z \mid A, X,$$

which is equivalent (this may not be obvious, but you can check with semi-graphoids!)

d-Separation

Note that we can also obtain other independences using the graphoid axioms:



$$
\begin{aligned}
& X \perp\!\!\!\perp W, A \mid Z \quad \text{and} \quad Y \perp\!\!\!\perp W, Z \mid X, A \\
\implies \; & X \perp\!\!\!\perp W \mid Z, A \quad \text{and} \quad Y \perp\!\!\!\perp W \mid Z, X, A \\
\implies \; & X, Y \perp\!\!\!\perp W \mid Z, A \qquad \implies \qquad Y \perp\!\!\!\perp W \mid Z, A.
\end{aligned}
$$

Is there a way to deduce these directly?

Yes! We can use **d-separation**.

A **path** π is a sequence of adjacent edges, without repeating any vertex.

Examples:

$Z \to X \to Y$; (this is a **directed** path)

$W \to A \to Y \leftarrow X$.

d-Separation

For any path, the internal vertices are either:

- **colliders:** i.e. $\rightarrow V \leftarrow$; or
- **non-colliders:** i.e. $\rightarrow V \rightarrow$ or $\leftarrow V \rightarrow$ or $\leftarrow V \leftarrow$.

We say a path π from A to B is **open** given a set $\boldsymbol{C}$ if and only if

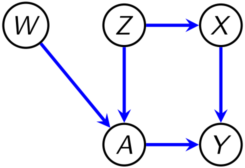
- no non-colliders on π are in $\boldsymbol{C}$; and
- every collider on π is an ancestor of something in $\boldsymbol{C}$.

Otherwise π is **blocked** (or **closed**).

Definition

We say that sets of vertices $\boldsymbol{A}$ and $\boldsymbol{B}$ are **d-separated** given $\boldsymbol{C}$ if **every** path from any $A \in \boldsymbol{A}$ to any $B \in \boldsymbol{B}$ is blocked by $\boldsymbol{C}$.

d-Separation Example



Is $\{W, A\}$ d-separated from $\{X\}$ by $\{Z\}$?

Is $\{W, A\}$ d-separated from $\{X\}$ by $\{Z, Y\}$?

Global Markov Property

Definition

A distribution P is said to obey the **global Markov property** with respect to a DAG $\mathcal{G}$ if whenever

$$\boldsymbol{A} \perp_d \boldsymbol{B} \mid \boldsymbol{C} \quad \text{in } \mathcal{G},$$

we have

$$V_{\boldsymbol{A}} \perp\!\!\!\perp V_{\boldsymbol{B}} \mid V_{\boldsymbol{C}} \quad \text{in } P.$$

In other words, d-separation implies conditional independence.

In addition to being ‘sound’ d-separation is **complete**: that is, any triple not d-separated is generally **not** independent.

That is: d-separation gives **all** independences implied by the model!

Markov Properties

There are three main **Markov properties** (models) which we can associate with DAGs.

Factorization. That is (if P has a density p) we have

$$p(\boldsymbol{v}) = \prod_{v \in \boldsymbol{v}} p(v \mid \mathrm{pa}_{\mathcal{G}}(v)).$$

(Ordered) Local Markov Property. For any topological ordering $\prec$, we have

$$V_i \perp\!\!\!\perp \boldsymbol{V}_{\mathrm{pre}(i;\prec) \setminus \mathrm{pa}(i)} \mid \boldsymbol{V}_{\mathrm{pa}(i)} \; [P].$$

Global Markov Property. Whenever $\boldsymbol{A}$ and $\boldsymbol{B}$ are d-separated by $\boldsymbol{C}$ in $\mathcal{G}$ then

$$V_{\boldsymbol{A}} \perp\!\!\!\perp V_{\boldsymbol{B}} \mid V_{\boldsymbol{C}} \; [P].$$

Structural Equations

An alternative model considers each variable to be generated from a **structural equation**:

$$X_v = f_v(X_{\mathrm{pa}(v)}, E_v)$$

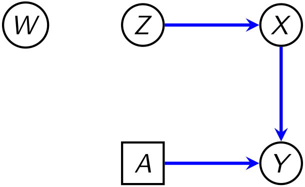
for a measurable function f_v and a noise term E_v.
The noise terms are assumed independent for DAGs.

Implications of this definition are completely equivalent to the others!

In a causal setting they are sometimes called **structural causal models** (Pearl, 2009; Peters et al., 2017), though we prefer the term **structural equation models**.

Causal Models

A DAG can also encode causal information:



If we intervene to experiment (*do*) on A, delete incoming edges.

In distribution, delete factor corresponding to A:

$$p(w, z, x, a, y) = p(w) \cdot p(z) \cdot p(x \mid z) \cdot p(a \mid w, z) \cdot p(y \mid x, a).$$
$$p(w, z, x, y \mid do(a)) = p(w) \cdot p(z) \cdot p(x \mid z) \qquad \times \qquad p(y \mid x, a).$$

All other factors are preserved (if causal DAG correctly specified).

Causal Effects

The function $p(\cdot \mid do(a))$ is just like any ordinary probability distribution, and obeys the same rules of conditioning and marginalization.

In particular, we can define expectations in the usual way:

$$\mathbb{E}[Y \mid do(A = a)] := \sum_{y} y \cdot p(y \mid do(a)).$$

Equipped with this distribution, we can now define the **average treatment effect** of A on Y:

$$\text{ATE} := \mathbb{E}[Y \mid do(A = 1)] - \mathbb{E}[Y \mid do(A = 0)].$$

This is sometimes called the **average causal effect** (ACE) or the **total effect**.

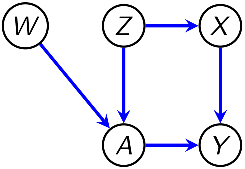
Confounding and Adjustment

1. Introduction

- Basic Causal Concepts
- Motivation
- Conditional Independence
- Directed Acyclic Graphs
- Confounding and Adjustment
- Selection Bias
- Potential Outcomes
- Other Causal Effects

Confounding

In this case there are other variables that causally affect both propensity to take the intervened A and our outcome Y.



For example, suppose older mothers ($Z = 1$) are more likely to take vitamin A (A), and their infants generally have worse health outcomes (X) which reduces their overall mental health level (Y).

A naïve estimate $\mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0]$ includes correlation due to this **confounding**.

This is **not** a causal quantity, since if we actually intervene to set A (e.g. by randomization), the contrast will (generally) be different.

Adjustment Using Parents

Note that we have

$$p(w, z, x, y \mid do(a)) = \frac{p(w, z, x, a, y)}{p(a \mid w, z)}.$$

Hence, to obtain (e.g.) $p(y \mid do(a))$ we just marginalize:

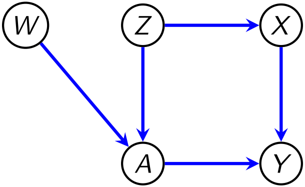
$$\begin{aligned}\sum_{w,z,x} p(y, w, z, x \mid do(a)) &= \sum_{w,z,x} \frac{p(y, w, z, x, a)}{p(a \mid w, z)} \\ &= \sum_{w,z,x} p(w, z) \cdot p(x, y \mid w, z, a) \\ &= \sum_{w,z} p(w, z) \cdot p(y \mid w, z, a).\end{aligned}$$

In this case we call $\{W, Z\}$ an **adjustment set** for the effect of A on Y.

The set of parents of a variable is **always** a valid adjustment set.

Adjustment sets are much more general that this, however.

Back-Door Paths



A **back-door** path from A to Y starts with an arrowhead at A.

Example. $A \leftarrow Z \rightarrow X \rightarrow Y$.

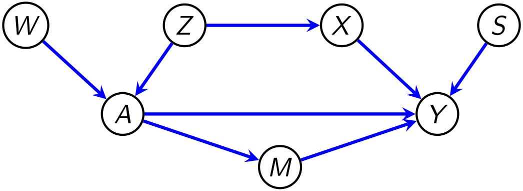
To identify $p(y \mid do(a))$ we must block all back-door paths **without** blocking any causal ones, nor inducing any selection bias.

Back-Door Criterion

Definition

A **back-door adjustment set** for the pair (A, Y) is one which:

- blocks all back-door paths from A to Y;
- does not contain any descendants of A.



Examples:

$\{Z\}$,	$\{X\}$,	$\{Z, X\}$
$\{W, Z\}$,	$\{W, X\}$,	$\{W, Z, X\}$
$\{S, Z\}$,	$\{S, X\}$,	$\{S, Z, X\}$
$\{S, W, Z\}$,	$\{S, W, X\}$,	$\{S, W, Z, X\}$.

The **optimal adjustment set** is just $\{X, S\}$.

Back-Door Adjustment

Theorem (Pearl, 1993)

Suppose that the pair p is causally Markov w.r.t. $\mathcal{G}$, and that we are interested in the causal effect of A on Y. Then this can be identified by

$$p(y \mid do(a)) = \sum_{x_C} p(x_C) \cdot p(y \mid a, x_C),$$

provided that X_C represents a back-door adjustment set for (A, Y).

Proof.

Since the back-door adjustment set contains no descendants of A, we have $A \perp\!\!\!\perp X_C \mid X_{\mathrm{pa}(a)}$.

It is also easy to see that $Y \perp\!\!\!\perp X_{\mathrm{pa}(a)} \mid A, X_C$ if X_C blocks all back-door paths and A all causal paths.

Back-Door Adjustment (ctd.)

Then:

$$\begin{aligned}
p(y \mid do(a)) &= \sum_{x_{\mathrm{pa}(v)}} p(x_{\mathrm{pa}(v)}) \cdot p(y \mid a, x_{\mathrm{pa}(v)}) \\
&= \sum_{x_{\mathrm{pa}(v)}} p(x_{\mathrm{pa}(v)}) \sum_{x_C} p(y \mid x_C, a, x_{\mathrm{pa}(v)}) \cdot p(x_C \mid a, x_{\mathrm{pa}(v)}) \\
&= \sum_{x_{\mathrm{pa}(v)}} p(x_{\mathrm{pa}(v)}) \sum_{x_C} p(y \mid x_C, a) \cdot p(x_C \mid x_{\mathrm{pa}(v)}) \\
&= \sum_{x_C} p(y \mid x_C, a) \sum_{x_{\mathrm{pa}(v)}} p(x_{\mathrm{pa}(v)}) \cdot p(x_C \mid x_{\mathrm{pa}(v)}) \\
&= \sum_{x_C} p(x_C) \cdot p(y \mid a, x_C).
\end{aligned}$$

□

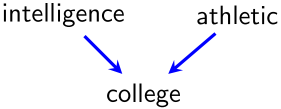
Selection Bias

1. Introduction

- Basic Causal Concepts
- Motivation
- Conditional Independence
- Directed Acyclic Graphs
- Confounding and Adjustment
- Selection Bias
- Potential Outcomes
- Other Causal Effects

Selection Bias

Bias can come from various sources; the most common is confounding, but **selection bias** is also a big concern.



intelligence

athletic

college

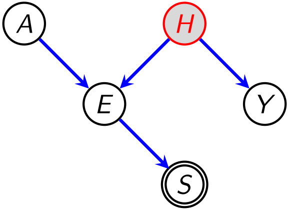
If we only observe people on a University Campus, we may incorrectly believe that intelligence and athletic ability are **negatively** correlated.

This is also referred to as **collider bias** or **Berkson's paradox**.

Side Effects

Suppose that patients may differentially drop out of a study due to side-effects.

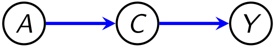
- H — general health;
- E — side effects;
- S — drop out (only observe $S = 1$).



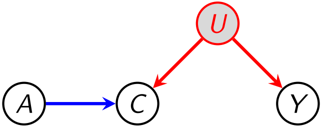
In this case we may erroneously think that the treatment is helpful, when really there is no effect.

Post-treatment Variables

It is generally a mistake to control for **post-treatment** variables, since it may block the causal effect:

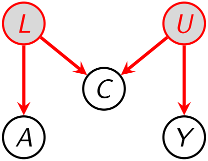


Indeed, the reverse problem can also occur!



M-bias

Another concern is so-called **M-bias**, which can arise if we try to condition on pre-treatment covariates but actually **open** a non-causal path by doing so.



Suppose that treatment A is smoking behaviour, C is childhood asthma and Y is adult asthma; L is parental smoking, U is underlying atopy.

Note that the back-door path is marginally blocked, but conditioning upon (only) C opens it!

The length of the path (four edges) means it is unlikely to be a strong bias in practice, however.

Potential Outcomes

1. Introduction
   - Basic Causal Concepts
   - Motivation
   - Conditional Independence
   - Directed Acyclic Graphs
   - Confounding and Adjustment
   - Selection Bias
   - Potential Outcomes
   - Other Causal Effects

Potential Outcomes

Consider a situation relevant to our running example:

- $A \in \{0, 1\}$, indicator of taking the vitamin A supplement;
- $Y \in \{0, 1\}$, indicator of no post-natal depression.

A woman takes the supplement ($A = 1$), and does not have post-natal depression ($Y = 1$).

It this **because** she took the treatment?

The question is begged: **what would have happened** if she had not taken it ($A = 0$)?

Potential Outcomes

Let us imagine two outcomes: $Y(a)$ for $a \in \{0,1\}$.

We observe $Y(A)$, but $Y(1-A)$ is always unseen.

$Y(0)$	$Y(1)$	type
0	0	never recover
1	0	hurt
0	1	helped
1	1	always recover

These pairs are known as **potential outcomes** (sometimes called counterfactuals). The **individual causal effect** (ICE) for me is

$$\mathrm{ICE} := Y(1) - Y(0).$$

$Y(1) = 1$, so ICE is either 1 or 0; but **cannot** deduce full 'type'.

This is the **fundamental problem of causal inference** (Holland, 1986).

Potential outcomes for causal inference is known as the Neyman-Rubin framework after Splawa-Neyman (1923/1990) and Rubin (1974).

Assumptions for Potential Outcomes

We generally make two important assumptions about potential outcomes.

- **No interference.** That is, the value of my outcome does not depend upon what anyone else decides to do.

  Counter-example: causal effect of a vaccine.

- **Single version of treatment.** For each treatment a and each individual there is a unique value of $Y(a)$.

  Counter-example: there are two different training courses run by different providers, but we consider them both to be $A = 1$.

Together these make up the **stable unit treatment value assumption** (or SUTVA for short), and allows us to assert that

$$A = a \qquad \Longrightarrow \qquad Y = Y(a).$$

This is known as the **consistency** property.

We can also weaken this to the **stable unit treatment distribution assumption** (SUTDA); only requires the *distribution* to be the same.

Average Causal Effects

The **average treatment effect** is now

$$\mathrm{ATE} := \mathbb{E}Y(1) - \mathbb{E}Y(0).$$

What is the connection to the naïve estimate?

$$\mathbb{E}[Y \mid A = x] = \mathbb{E}[Y(a) \mid A = a] \overset{!}{=} \mathbb{E}Y(a),$$

if the treatment is independent of the potential outcome: $Y(a) \perp\!\!\!\perp A$.

So, for example, *if treatment is assigned at random* then

$$\mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0] = \mathbb{E}Y(1) - \mathbb{E}Y(0)$$

and this difference is guaranteed to be causal.

Is this a good way of modelling causality? There has been contention (e.g. Dawid, 2000), but it is now generally accepted as a reasonable approach.

Conditonal Exchangeability

An alternative is to assume that there is a set of covariates that is sufficient to control for the confounding (suppose that we have X).

In other words we assume that $Y(a) \perp\!\!\!\perp A \mid X$.

Then we can use the g-formula:

$$\begin{aligned}\mathbb{E}_X\mathbb{E}[Y \mid A = a, X] &= \mathbb{E}_X\mathbb{E}[Y(a) \mid A = a, X] \\ &\overset{!}{=} \mathbb{E}_X\mathbb{E}[Y(a) \mid X] \\ &= \mathbb{E}Y(a),\end{aligned}$$

if the treatment is **conditionally** independent of each potential outcome.

So **all** we need to do is estimate $\mathbb{E}[Y \mid A = a, X]$ and average over levels of X. Unfortunately, this is hard to do well if X is multi-dimensional (let alone high-dimensional!)

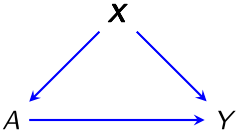
Also note that this **conditional exchangeability** is not a **testable assumption**.

Connection to *do*-Calculus

Note that

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{X}}\mathbb{E}[Y \mid A = a, \boldsymbol{X}] &= \int \left(\int y \cdot p(y \mid a, \boldsymbol{x}) \, dy \right) p(\boldsymbol{x}) \, d\boldsymbol{x} \\
&= \iint y \cdot p(y \mid \boldsymbol{x}, a) \cdot p(\boldsymbol{x}) \, dy \, d\boldsymbol{x} \\
&= \iint y \cdot p(y, \boldsymbol{x} \mid \, do(a)) \, dy \, d\boldsymbol{x} \\
&= \mathbb{E}[Y \mid \, do(A = a)]
\end{aligned}
$$

under the assumption that the causal graph is:



References

Dawid, A. P. Causal inference without counterfactuals. *Journal of the American statistical Association*, 95(450), 407-424, 2000.

Dawid, A. P. Decision-theoretic foundations for statistical causality. *Journal of Causal Inference*, 9(1), 39-77, 2021.

Holland, P. W. Statistics and causal inference. *Journal of the American statistical Association*, 81(396), 945-960, 1986.

Splawa-Neyman, J. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. (1923), translated by Dabrowska and Speed in *Statistical Science*, 1990.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5), 688, 1974.

Other Causal Effects

1. Introduction
   - Basic Causal Concepts
   - Motivation
   - Conditional Independence
   - Directed Acyclic Graphs
   - Confounding and Adjustment
   - Selection Bias
   - Potential Outcomes
   - Other Causal Effects

‘The’ Causal Effect

There is no such thing as *the* causal effect.

We have seen the **average causal effect** (ACE, or ATE):

$$\text{ACE} = \mathbb{E}Y(1) - \mathbb{E}Y(0) = \mathbb{E}[Y \mid do(A=1)] - \mathbb{E}[Y \mid do(A=0)],$$

but we may also be interested in a **conditional** ATE (or CATE) given some $X = x$:

$$\begin{aligned}\text{CATE}(x) &= \mathbb{E}[Y(1) \mid X = x] - \mathbb{E}[Y(0) \mid X = x] \\ &= \mathbb{E}[Y \mid X = x, do(A=1)] - \mathbb{E}[Y \mid X = x, do(A=0)],\end{aligned}$$

which considers subgroups defined by particular (pretreatment!) covariates.

Causal effects for treatment status

May also be interested in the **effect of treatment on the treated** (ETT or ATT), which is

$$\begin{aligned} \mathrm{ETT} &= \mathbb{E}[Y(1) \mid A = 1] - \mathbb{E}[Y(0) \mid A = 1] \\ &= \mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y(0) \mid A = 1]. \end{aligned}$$

This is easier to identify, since only need that everyone *could* have not been treated:

$$P(A = 0 \mid x) > 0 \qquad\qquad \text{for almost all } x.$$

Can analogously define the **effect of treatment on the controlled** (ETC or ATC).

Other Causal Effects

Some interventions are based on different contrasts (e.g. by a change in variance, or by adding noise).

Others may be **adaptive** to (e.g.) a patient's history.

The **individual causal effect** (ICE or ITE) requires that potential outcomes are well-defined:

$$\mathrm{ICE} = Y(1) - Y(0).$$

Similarly the **principal stratum effect** requires treatment 'types' to be well-defined:

$$\mathrm{PSE}_{\mathrm{CO}} = \mathbb{E}[Y(1) - Y(0) \mid \mathrm{Type} = \mathrm{CO}].$$

The study of **mediation** leads to various definitions of **direct** and **indirect** effects (controlled, pure, natural...)

Identification

In order to identify the ATE, we need:

- **SUTVA** (or SUDVA);
- **positivity**:

$$P(A = a \mid x) > 0 \qquad a \in \{0, 1\} \text{ and almost all } x.$$

- **conditional exchangeability**:

$$Y(a) \perp\!\!\!\perp A \mid X.$$

[This is not testable from a single dataset.]

For the ETT we only need

$$P(A = 0 \mid x) > 0 \qquad \text{for almost all } x.$$

Identification

Using these assumptions, we have:

$$
\begin{aligned}
P(Y(a)) &= \sum_x P(Y(a) \mid x) \cdot p(x) && \text{probability calculus} \\
&= \sum_x P(Y(a) \mid a, x) \cdot p(x) && \text{conditional exchangeability} \\
&= \sum_x P(Y \mid a, x) \cdot p(x) && \text{consistency+positivity.}
\end{aligned}
$$

This is an example of the **g-formula** (Robins, 1986).

Reference

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7 (9-12), 1393–1512.

Outline

1. Introduction

2. Single-World Intervention Graphs

3. Estimation Methods

4. Machine Learning Methods

NPSEM-IEs

2. Single-World Intervention Graphs

- NPSEM-IEs
- d-separation in SWIGs
- Adjustment for Confounding

Structural causal models

In machine learning it is common to use **structural causal models** (SCMs) to represent causal models.

These originate with the work of Sewall Wright in the 1920s, and he referred to them as **structural equation models** (SEMs).

Each variable (say X_v) is written as a function of its parents and a noise term:

$$X_v \leftarrow f_v(X_{\mathrm{pa}(v)}, \varepsilon_v).$$

Often the noise terms are assumed to be **independent**.

Note that we can also write this using potential outcome notation:

$$X_v = f_v(X_{\mathrm{pa}(v)}, \varepsilon_v) = Y(X_{\mathrm{pa}(v)}).$$

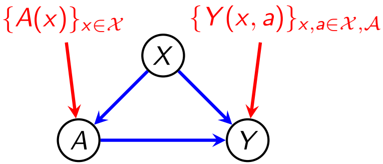
NPSEM-IEs

If the errors are independent, the model is referred to as a **non-parameteric SEM with independent errors** (NPSEM-IE) by Richardson and Robins (2013).

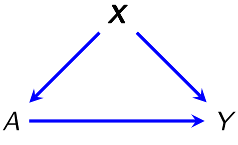
They note that it implicitly makes **cross-world** assumptions. For example, in the graph below, we would have (e.g.)

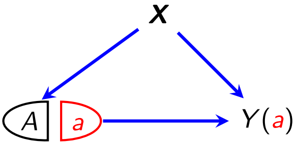
$$A(x') \perp\!\!\!\perp Y(x, a), \qquad \forall x, x'.$$

This is completely untestable **using any randomized trial**.


$\{A(x)\}_{x \in \mathcal{X}}$
$\{Y(x,a)\}_{x,a \in \mathcal{X},\mathcal{A}}$
X
A
Y

Single-World Intervention Graphs



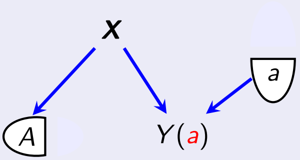


Single-World Intervention Graphs (SWIGs) combine graphs and potential outcomes so as to allow one to read off important conditions (see Richardson and Robins, 2013).

Note we can see by d-separation that the 'no unobserved confounding' assumption holds under this SWIG:

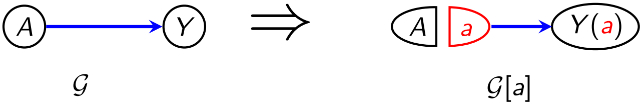
$$Y(a) \perp\!\!\!\perp A \mid \boldsymbol{X}.$$

Once nodes are split we can rearrange them:



Representing an intervention

$$P(A = a, Y = y) = P(A = a)P(Y = y \mid A = a)$$



The graph says that $Y(a) \perp_d A$, and hence:

$$P(Y(a)) = P(Y(a) \mid A = a) =^c P(Y \mid A = a), \qquad \forall a.$$

Notice that, for two distinct values a, a' of A, we **never** observe $Y(a)$ and $Y(a')$ on the same graph.

In particular, SWIGs will never say that $A \perp\!\!\!\perp \{Y(a), Y(a')\}$ if $a \neq a'$.

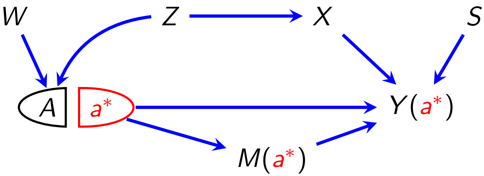
This is what is meant by **single-world** in the name of the class of graphs.

This has important consequences for the identification of direct effects.

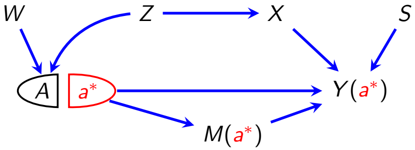
Node-splitting

What happens when we intervene in a SWIG?

1. Split the node(s) $\boldsymbol{V}_A$ being intervened on into $\boldsymbol{V}_A$ and $\boldsymbol{v}_A^*$.
2. Replace all descendants of $\boldsymbol{v}_A^*$ by $V(\boldsymbol{v}_A^*)$.
3. In the factorization, replace every instance of $\boldsymbol{v}_A$ with $\boldsymbol{v}_A^*$, and all descendants of $\boldsymbol{V}_A$ with $V(\boldsymbol{v}_A^*)$.



Node-splitting



Intervene to set $A = a^*$:

1. Add potential outcome to all descendants of A;
2. Remove any conditioning on $A = a$.

$$P(W, Z, X, S, A, M, Y) = P(W) \cdot P(Z) \cdot P(X \mid Z) \cdot P(S) \cdot \\ \times P(A \mid W, Z) \cdot P(M \mid A) \cdot P(Y \mid S, A, M)$$

$$P(W, Z, X, S, A, M(a^*), Y(a^*)) = P(W) \cdot P(Z) \cdot P(X \mid Z) \cdot P(S) \\ \times P(A \mid W, Z) \cdot P(M(a^*)) \cdot P(Y(a^*) \mid S, M(a^*)).$$

Note that we can replace **all** variables V with $V(a^*)$, but only affects the descendants of A.

Intuition behind node splitting

Question

How could we identify whether someone would choose to take treatment, i.e. have $A = 1$, and at the same time find out what happens to such a person if they don't take treatment $Y(a = 0)$?

Answer

Whenever a patient is observed to swallow the drug, we instantly intervene by administering a safe 'emetic' that causes the pill to be regurgitated before any drug can enter the bloodstream.

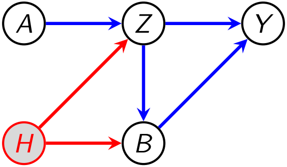
Since we assume the emetic has no side effects, the patient's recorded outcome is then $Y(a = 0)$.

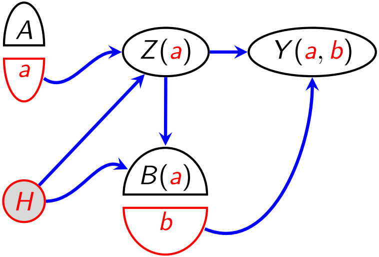
Hence the SWIG represents quantities that (at least in principle) are causally identifiable by an experiment; e.g.

$$\text{ETT} := \mathbb{E}[Y(1) \mid A = 1] - \mathbb{E}[Y(0) \mid A = 1].$$

(Robins et al. 2007)

Harder inferential problem





Query

Does this causal graph imply:

$$Y(a, b) \perp\!\!\!\perp B(a) \mid Z(a), A \qquad ?$$

Answer

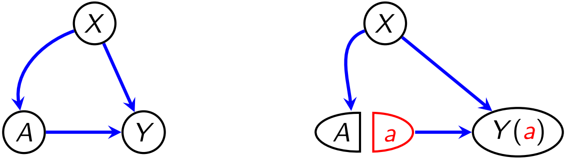
Yes! Applying d-separation to the SWIG on the right we see that there is no d-connecting path from $Y(a, b)$ given $Z(a)$.

Summary Adding Counterfactual Distributions to DAGs

Factorization of counterfactual variables: $P(\boldsymbol{V}(\boldsymbol{a}))$ factorizes with respect to the SWIG $\mathcal{G}[\boldsymbol{a}]$ (ignoring fixed nodes):

$$P\left(\boldsymbol{V}(\boldsymbol{a})\right) = \prod_{Y(\boldsymbol{a}) \in \boldsymbol{V}(\boldsymbol{a})} P\left(Y(\boldsymbol{a}) \middle| \operatorname{pa}_{\mathcal{G}[\boldsymbol{a}]}(Y(\boldsymbol{a})) \setminus \boldsymbol{a}\right).$$

Example



Suppose we want to identify the distribution of $Y(a)$ using these two SWIGs, but that we only observe $P(X, A, Y)$.

The previous slide tells us that

$$\begin{aligned} P(X, A, Y(a)) &= P(X) \cdot P(A \,|\, X) \cdot P(Y(a) \,|\, X) \\ &= P(X) \cdot P(A \,|\, X) \cdot P(Y(a) \,|\, X, A = a) \\ &= P(X) \cdot P(A \,|\, X) \cdot P(Y \,|\, X, A = a), \end{aligned}$$

and therefore

$$\begin{aligned} P(Y(a)) &= \sum_X P(X) \left\{ \sum_A P(A \,|\, X) \right\} P(Y \,|\, X, A = a) \\ &= \sum_X P(X) \cdot P(Y \,|\, X, A = a). \end{aligned}$$

d-separation in SWIGs

2. Single-World Intervention Graphs
   - NPSEM-IEs
   - d-separation in SWIGs
   - Adjustment for Confounding

Applying d-separation to the graph $\mathcal{G}[\boldsymbol{a}]$

We extend the definition of d-separation to SWIGs as follows:

- A **fixed** node is always **blocked** if it occurs as a non-endpoint on a path;
- A path on which one endpoint is a fixed node can d-connect that node to a random node if it satisfies the usual conditions on colliders and non-colliders;

In $\mathcal{G}[\tilde{a}]$ if subsets $\boldsymbol{B}(\tilde{a})$ and $\boldsymbol{C}(\tilde{a})$ of random nodes are d-separated by $\boldsymbol{D}(\tilde{a})$, then $\boldsymbol{B}(\tilde{a})$ and $\boldsymbol{C}(\tilde{a})$ are conditionally independent given $\boldsymbol{D}(\tilde{a})$ in the associated distribution $P(\boldsymbol{V}(\tilde{a}))$.

$$\begin{aligned} &\boldsymbol{B}(\tilde{a}) \text{ is d-separated from } \boldsymbol{C}(\tilde{a}) \text{ given } \boldsymbol{D}(\tilde{a}) \text{ in } \mathcal{G}[\tilde{a}] \\ &\Rightarrow \quad \boldsymbol{B}(\tilde{a}) \perp\!\!\!\perp \boldsymbol{C}(\tilde{a}) \mid \boldsymbol{D}(\tilde{a}) \quad [P(\boldsymbol{V}(\tilde{a}))]. \end{aligned} \qquad (*)$$

Applying d-separation to the graph $\mathcal{G}[\boldsymbol{a}]$

We extend the definition of d-connection to SWIGs as follows:

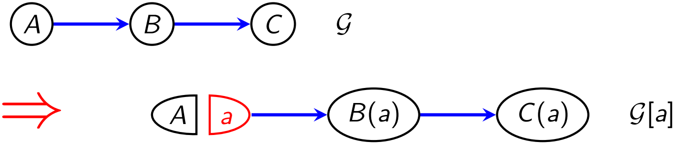
- A fixed node is always blocked if it occurs as a non-endpoint on a path;
- A path on which one endpoint is a fixed node **can d-connect** that node to a random node if it satisfies the usual conditions on colliders and non-colliders.

In $\mathcal{G}[\boldsymbol{a}, d]$, if fixed node d is d-separated from $\boldsymbol{B}(\boldsymbol{a}, d)$ given $\boldsymbol{C}(\boldsymbol{a}, d)$ then

$$P(\boldsymbol{B}(\boldsymbol{a}, d) \mid \boldsymbol{C}(\boldsymbol{a}, d)) = P(\boldsymbol{B}(\boldsymbol{a}, d') \mid \boldsymbol{C}(\boldsymbol{a}, d')).$$

In other words, the conditional distribution of $\boldsymbol{B}$ given $\boldsymbol{C}$ after intervening on $\boldsymbol{A}$ and $\boldsymbol{D}$ does not depend on the value assigned to $\boldsymbol{D}$.

Example of d-separation from fixed nodes



The fixed node a is d-separated from $C(a)$ given $B(a)$. Consequently it follows that

$$P(C(\tilde{a}) \mid B(\tilde{a})) = P(C(a^*) \mid B(a^*))$$

for any values $\tilde{a}$, a^*. This may alternatively be derived:

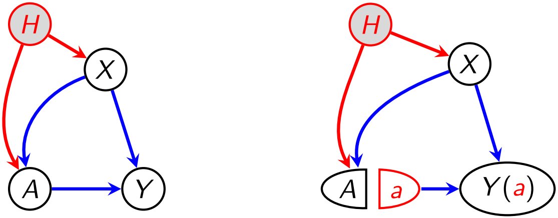
$$\begin{aligned} P(C(\tilde{a}) \mid B(\tilde{a})) &=^{d,\mathcal{G}[a]} P(C(\tilde{a}) \mid B(\tilde{a}), A = \tilde{a}) \\ &=^{c} P(C \mid B, A = \tilde{a}) \;=^{d,\mathcal{G}}\; P(C \mid B, A = a^*) \\ &=^{c} P(C(a^*) \mid B(a^*), A = a^*) \;=^{d,\mathcal{G}[a]}\; P(C(a^*) \mid B(a^*)) \end{aligned}$$

via consistency and d-separation in $\mathcal{G}[a]$ and $\mathcal{G}$.

Adjustment for Confounding

2. Single-World Intervention Graphs
   - NPSEM-IEs
   - d-separation in SWIGs
   - Adjustment for Confounding

Another Example



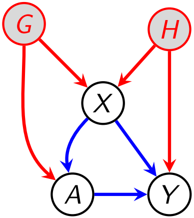
Here again we can read directly from the graph that

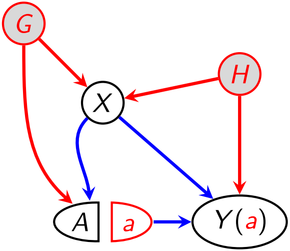
$$A \perp\!\!\!\perp Y(a) \mid X.$$

Hence

$$P(Y(a)) = \sum_x P(X = x) \cdot P(Y \mid A = a, X = x).$$

Exercise





Is it still the case that $A \perp\!\!\!\perp Y(a) \mid X$?

Summary

- SWIGs provide a simple way to unify graphs and counterfactuals via node-splitting
- The approach works via linking the factorizations associated with the SWIG to the distribution in the original DAG.
- The new graph represents a counterfactual distribution that is *identified* from the original joint distribution.
- (Not covered) Can combine information on the absence of individual and population level direct effects.
- (Not covered) Permits formulation of models where interventions on only some variables are well-defined.

References

Pearl, J. Causal diagrams for empirical research, *Biometrika* 82, 4, 669–709, 1995.

Richardson, TS, Robins, JM. Single World Intervention Graphs. *CSSS Tech. Report No. 128*
`http://www.csss.washington.edu/Papers/wp128.pdf`, 2013.

Richardson, TS, Robins, JM. SWIGs: A Primer. *UAI-13*, 2013.

Robins, JM A new approach to causal inference in mortality studies with sustained exposure periods – applications to control of the healthy worker survivor effect. *Mathematical Modeling* 7, 1393–1512, 1986.

Robins, JM, VanderWeele, TJ, Richardson TS. Discussion of "Causal effects in the presence of non compliance a latent variable interpretation" by Forcina, A. *Metron* LXIV (3), 288–298, 2007.

Outline

1. Introduction
2. Single-World Intervention Graphs
3. Estimation Methods
4. Machine Learning Methods

Causal estimation

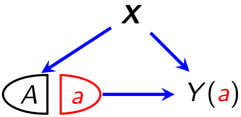
Given a particular causal effect there are myriad ways to estimate it!

Selecting the best one requires:

- experience and judgement of which causal methods work best in particular situations (e.g. data types, number of samples, number of variables);
- expert knowledge about the system.

We can classify some of the standard approaches into three groups:

- outcome regression / standardization / adjustment;
- propensity scores / inverse weighting / stratification / matching;
- hybrid approaches / doubly robust methods.



Assumptions

Recall the three basic assumptions needed for estimation in this context:

1. **SUTVA**: no interference and consistency;
2. **positivity**: i.e. that $P(A = a \mid \boldsymbol{X} = \boldsymbol{x}) > 0$ for suitable $a, \boldsymbol{x}$;
3. **conditional exchangeability**: i.e. that $Y(a) \perp\!\!\!\perp A \mid \boldsymbol{X}$.

Of these, 1 and 3 are not testable without additional information. Expert perspectives on the subject matter are crucial.

For 2, we can test this statistically by looking at estimates of $p(a \mid \boldsymbol{x})$.

There are also methods to characterize which ‘areas’ of $\boldsymbol{X}$ have good overlap.

Assumptions

Positivity can be assessed by estimating the **propensity score**. That is $\pi(\boldsymbol{x}) := P(A = 1 \mid \boldsymbol{X} = \boldsymbol{x})$.

Propensity Scores

3. Estimation Methods

- Propensity Scores
- Horvitz-Thompson Estimators
- Simulations
- Doubly-Robust Approaches

Outcome Regression

The simplest approach to using a propensity score is simply to add it to your regression model:

$$\mathbb{E}[Y \mid A, \pi(\boldsymbol{X})] = \beta A + \gamma \pi(\boldsymbol{X}).$$

Then the least squares estimator $\hat{\beta}$ will be consistent for the average causal effect.

Note that this relies on the form of $\mathbb{E}[Y \mid A, \pi(\boldsymbol{X})]$ being linear in both A and $\pi(\boldsymbol{X})$, which may not be the case.

It also relies on you specifying the propensity score model correctly, because what you usually fit is:

$$\arg\min_{\beta,\gamma} \sum_i \left(Y_i - \beta a_i - \gamma \hat{\pi}(x_i)\right)^2,$$

where $\hat{\pi}$ is an **estimate** of π.

Inverse Probability of Treatment Weighting (IPTW)

An alternative is to **reweight** observations by the reciprocal of the propensity of the treatment actually received.

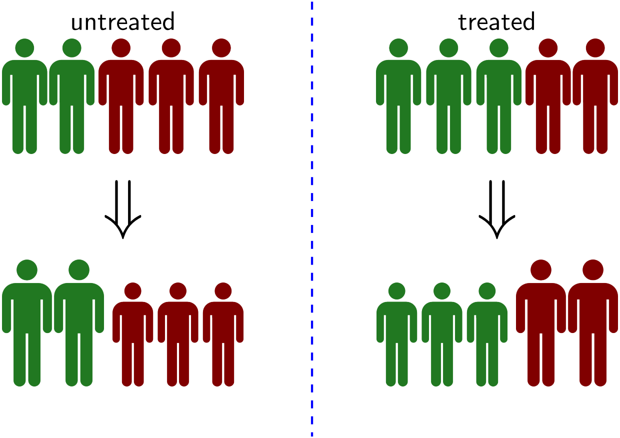
That is, by $\pi(x)$ if they were treated, and $1 - \pi(x)$ if they were not.

This creates a **pseudo-population** in which individuals are assigned treatment independently of any confounding variables.

We can choose to **stabilize** the weights by also multiplying by some arbitrary marginal distribution $p^*(a)$ (e.g. Bernoulli$(1/2)$). This is particularly useful for a continuous treatment.

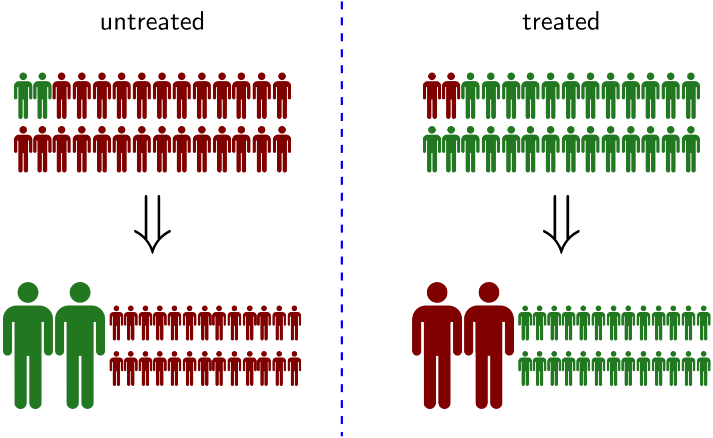
Pseudo-Population

Here is an illustration.



Pseudo-Population

But **beware of extreme weights**!



untreated

treated

Pseudo-Population

Take a simple example, with X, A and Y all being binary.

Suppose that $A \mid X \sim \text{Bernoulli}(0.4 + 0.3X)$.

Then:

X	A	Y	$p(A \mid X)^{-1}$
0	0	1	$5/3$
1	1	1	$10/7$
1	0	0	$10/3$
1	0	1	$10/3$
0	1	0	$5/2$
	$\vdots$		

Note that:

- rarer combinations (e.g. $X = 1$ and $A = 0$) are upweighted more;
- Y has no effect on the weight.

IPTW

Note that, in the pseudo-population, the distribution is given by

$$\begin{aligned} p^*(z,a,y) &= p(z,a,y) \cdot \frac{p^*(a)}{p(a \mid z)} \\ &= p(z) \cdot p(a \mid z) \cdot p(y \mid z,a) \cdot \frac{p^*(a)}{p(a \mid z)} \\ &= p(z) \cdot p^*(a) \cdot p(y \mid z,a), \end{aligned}$$

so now $X \perp\!\!\!\perp A$ marginally.

Exercise

Check that the marginal distribution for A under p^* is indeed $p^*(a)$.

$$p^*(y \mid a) = \sum_z p(z) \cdot p(y \mid z,a);$$

so p^* is the distribution of the model after we intervene to change the distribution of A given X to be $p^*(a)$.

Assumptions for IPTW

To actually perform the reweighting, we need an additional assumption.

- **Positivity.** We need $0 < \pi(x) < 1$ for every x.

If this doesn't hold, then reweighting is hopeless.

Positivity violations may happen for statistical or structural reasons:

statistical (e.g.) there are too many categories among your covariates;

structural (e.g.) amputees can't have a surgical procedure.

Horvitz-Thompson Estimators

3. Estimation Methods

- Propensity Scores
- Horvitz-Thompson Estimators
- Simulations
- Doubly-Robust Approaches

Horvitz-Thompson Estimators

Suppose we consider the following estimator for $\mathbb{E}Y(1)$:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{A_i Y_i}{\pi(\boldsymbol{X}_i)}.$$

Note that (if π is correctly specified) then this has mean

$$\begin{aligned}
\mathbb{E}\left[\frac{AY}{\pi(\boldsymbol{X})}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{AY(1)}{\pi(\boldsymbol{X})}\middle| Y(1), \boldsymbol{X}\right]\right] \\
&= \mathbb{E}\left[\frac{Y(1)}{\pi(\boldsymbol{X})}\mathbb{E}[A \mid Y(1), \boldsymbol{X}]\right] \\
&= \mathbb{E}\left[\frac{Y(1)}{\pi(\boldsymbol{X})}\mathbb{E}[A \mid \boldsymbol{X}]\right] \\
&= \mathbb{E}\left[\frac{Y(1)}{\pi(\boldsymbol{X})}\pi(\boldsymbol{X})\right] = \mathbb{E}Y(1).
\end{aligned}$$

Note, however, that this estimator may be outside valid range for Y!

Horvitz-Thompson Estimators

If we can estimate $\pi(x)$ (and it is bounded away from 0 and 1) then previous slide suggests that the Horvitz-Thompson estimator is a sensible way to estimate $\mathbb{E}Y(1)$ (and similarly for $\mathbb{E}Y(0)$).

Theorem

Given the correct family of distributions for π, we will have $\sqrt{n}(\hat{\pi}(\boldsymbol{X}_i) - \pi(\boldsymbol{X}_i)) = O_p(1)$, so then

$$\sqrt{n}\left|\frac{1}{n}\sum_i \frac{A_i Y_i}{\hat{\pi}(\boldsymbol{X}_i)} - \mathbb{E}Y(1)\right| = O_p(1).$$

Both the limiting distributions are Gaussian.

Regression IPTW

Suppose that we believe a regression formulation for the potential outcomes: e.g.

$$\mathbb{E}[Y(a) \mid \boldsymbol{W}] = \beta a + \gamma^T \boldsymbol{W}.$$

Inverse Probability Weighting

Then we can solve a **weighted** least squares formulation:

$$\arg\min_{\beta,\gamma} \sum_{i=1}^{n} \frac{1}{w_i} \left(Y_i - \beta a_i - \gamma \boldsymbol{W}_i \right)^2 ,$$

where $w_i = p(a_i \,|\, \boldsymbol{X}_i)$.

This can be achieved using the `weights` argument in `R`'s `lm`/`glm` functions.

Warning: standard errors are computed naïvely!

Simulations

3. Estimation Methods
    - Propensity Scores
    - Horvitz-Thompson Estimators
    - Simulations
    - Doubly-Robust Approaches

Demonstrations

The `R` package `causl`* allows one to simulate data from a parametrically specified causal model.

Suppose we want to have:

$$\begin{aligned} Z &\sim \text{Exponential}(\lambda) \\ A \mid Z = z &\sim \text{Bernoulli}\left(\text{logit}(\alpha_0 + \alpha_1 z)\right) \\ Y \mid do(A = a) &\sim N(\beta a,\ \sigma^2) \end{aligned}$$

with $\lambda = 2$, $\alpha_0 = 0$, $\alpha_1 = 1$ and $\beta = 1/2$.

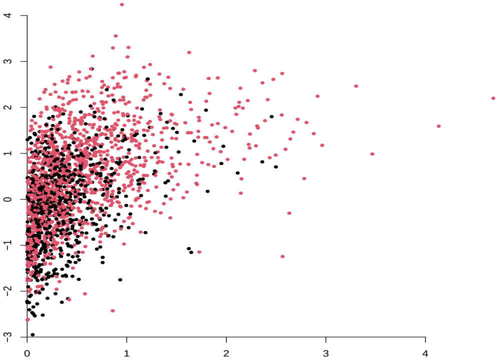
```
library(causl)
forms <- list(Z ~ 1,
              A ~ Z,
              Y ~ A,
              ~ 1)  ## for the copula
```

```
pars <- list(Z = list(beta = -log(2), phi=1),  ## we use log-link
             A = list(beta = c(0,1)),
             Y = list(beta = c(0,0.5), phi = 1),
             cop = list(beta = 1))
fam <- list(3,5,1,1)  # distributions: 1=normal, 3=Gamma, 5=binomial
```

Demonstrations

We can then use the rfrugalParam function to simulate our data:

```
set.seed(123)
dat <- rfrugalParam(1e4, formulas=forms, pars=pars, family=fam)
```



The plot shows the first 2000 data points.

Propensity Scores

Fit a binomial GLM to estimate the parameters in $\pi(Z)$.

```
## fit GLM to test conditional distribution of X
modX <- glm(A ~ Z, family=binomial, data=dat)
summary(modX)$coef[,1:2]


            Estimate Std. Error
(Intercept)   0.0258     0.0307
Z             1.0340     0.0540
```

We can obtain the estimated propensity scores using these fitted values.

```
ps <- fitted(modX)
head(ps)


    1     2     3     4     5     6
0.530 0.711 0.703 0.784 0.618 0.803


dat <- dplyr::mutate(dat, ps = ps)  # add est. propensity score
```

Outcome Regression

Given the manner in which we specified our model, it is hard to write down the 'correct' form for $\mathbb{E}[Y \mid A, Z]$, but we can try adding the propensity score instead:

```
## naive model as a baseline comparison
summary(lm(Y ~ A, data=dat))$coef[,1:2]
```

```
            Estimate Std. Error
(Intercept)   -0.115     0.0162
A              0.670     0.0205
```

```
## now add in propensity score
summary(lm(Y ~ A + ps, data=dat))$coef[,1:2]
```

```
            Estimate Std. Error
(Intercept)   -2.780     0.0575
A              0.487     0.0189
ps             4.473     0.0933
```

Recall that the true value was 0.5, so this works quite well.

Horvitz-Thompson

We can similarly obtain the Horvitz-Thompson estimator.

```
dat <- dplyr::mutate(dat, wts=A/ps + (1-A)/(1-ps))  # add weights

EY1 <- with(dat, mean(Y*A*wts))
EY1


[1] 0.485


EY0 <- with(dat, mean(Y*(1-A)*wts))
EY0


[1] -0.00513


EY1 - EY0


[1] 0.491
```

IPW

We can also implement IPW with regression to estimate the causal effect.

```
library(survey)  ## package that gives correct standard errors with weights

## fit weighted model
mod_w <- svyglm(Y ~ A, design=svydesign(~ 1, weights = ~ wts, data = dat))
summary(mod_w)$coef[,1:2]


            Estimate Std. Error
(Intercept) -0.00515     0.0176
A            0.49050     0.0218
```

Recall the naïve model for comparison.

```
## can compare to naive model
summary(glm(Y ~ A, data=dat))$coef[,1:2]


            Estimate Std. Error
(Intercept)   -0.115     0.0162
A              0.670     0.0205
```

Doubly-Robust Approaches

3. Estimation Methods
   - Propensity Scores
   - Horvitz-Thompson Estimators
   - Simulations
   - Doubly-Robust Approaches

When is the naïve estimate correct?

We know that, if $\boldsymbol{X}$ are sufficient to control for confounding, then

$$P(Y(a)) = \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \cdot P(Y \mid \boldsymbol{x}, A = a).$$

When does

$$P(Y(a)) = P(Y \mid A = a) \qquad ?$$

Suppose that $A \perp\!\!\!\perp \boldsymbol{X}$. Then:

$$\begin{aligned} P(Y \mid A = a) &= \sum_{\boldsymbol{x}} P(\boldsymbol{X}, Y \mid A = a) \\ &= \sum_{\boldsymbol{x}} P(\boldsymbol{X} \mid A = a) \cdot P(Y \mid \boldsymbol{X}, A = a) \\ &= \sum_{\boldsymbol{x}} P(\boldsymbol{X}) \cdot P(Y \mid \boldsymbol{X}, A = a) \\ &= P(Y(a)). \end{aligned}$$

When is the naïve estimate correct?

We know that, if $\boldsymbol{X}$ are sufficient to control for confounding, then

$$P(Y(a)) = \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \cdot P(Y \mid \boldsymbol{x}, A = a).$$

When does

$$P(Y(a)) = P(Y \mid A = a) \qquad ?$$

Or suppose that $Y \perp\!\!\!\perp \boldsymbol{X} \mid A$. Then

$$\begin{aligned} P(Y \mid A = a) &= \sum_{\boldsymbol{x}} P(\boldsymbol{X} = \boldsymbol{x}) \cdot P(Y \mid A = a) \\ &= \sum_{\boldsymbol{x}} P(\boldsymbol{X} = \boldsymbol{x}) \cdot P(Y \mid \boldsymbol{X} = \boldsymbol{x}, A = a) \\ &= P(Y(a)). \end{aligned}$$

When is the naïve estimate correct?

We know that, if $\boldsymbol{X}$ are sufficient to control for confounding, then

$$P(Y(a)) = \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \cdot P(Y \mid \boldsymbol{x}, A = a).$$

When does

$$P(Y(a)) = P(Y \mid A = a) \qquad ?$$

In summary, if **either** $A \perp\!\!\!\perp \boldsymbol{X}$ **or** $Y \perp\!\!\!\perp \boldsymbol{X} \mid A$ then $Y(a)$ and $Y \mid A = a$ have the same distributions.

This is perhaps unsurprising, given that in either of those cases, $\boldsymbol{X}$ is not really a confounder at all!

Doubly Robust Approaches

Note we've seen that if we specify

- the **outcome model** (i.e. $Y \mid A, \boldsymbol{X}$) correctly, we can obtain a consistent estimate of the ACE by averaging over the empirical $\boldsymbol{X}$ values;
- the **propensity score model** (i.e. $A \mid \boldsymbol{X}$) correctly, we can use the Horvitz-Thompson estimator which is also consistent.

Is there an estimator that uses both of these models, but only requires one of them to be correct?

Yes!

We can use the following approach: suppose we believe that

$$\mathbb{E}[Y \mid a, \boldsymbol{x}] = Q_a(\boldsymbol{x}; \beta, \gamma) \qquad \text{and} \qquad \pi(\boldsymbol{x}) = \pi(\boldsymbol{x}; \eta)$$

for **parametric** models Q_0, Q_1, and π.

These are sometimes called **working models**.

Doubly Robust Methods

Notice that the following function has expectation $Y(1)$ if **either** Q_1 or π is specified correctly:

$$\begin{aligned}\mu_1^{dr}(O) &= Q_1(\boldsymbol{X}) + \frac{A}{\pi(\boldsymbol{X})}\{Y - Q_1(\boldsymbol{X})\} \\ &= \frac{AY}{\pi(\boldsymbol{X})} + \left\{1 - \frac{A}{\pi(\boldsymbol{X})}\right\} Q_1(\boldsymbol{X}).\end{aligned}$$

So fit 'nuisance' models Q and π to the data (e.g. by maximum likelihood). This gives parameter estimates $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\eta}$.

Then consider the following estimator of $\mathbb{E}Y(1)$:

$$\hat{\mu}_1^{dr} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{A_i\{Y_i - Q_{A_i}(\boldsymbol{X}_i; \hat{\beta}, \hat{\gamma})\}}{\pi(\boldsymbol{X}_i; \hat{\eta})} + Q_1(\boldsymbol{X}_i; \hat{\beta}, \hat{\gamma})\right\}.$$

If **either** model is correctly specified, then by the above we can see that the estimate will be consistent.

This property is called **double robustness**.

Doubly Robust Methods

We can do something similar for $\hat{\mu}_0^{dr}$, and then

$$\hat{\beta}^{dr} := \hat{\mu}_1^{dr} - \hat{\mu}_0^{dr}. \qquad (\dagger)$$

We call this the **augmented** inverse probability weighted estimator (AIPW).

In addition, each $\hat{\mu}_a^{dr}$ is **semi-parametric efficient** if both parametric models are correct, so it achieves the same rate (asymptotically) as maximum likelihood estimation.

If Q_a is wrong then MLEs will be difficult to interpret.

In practice, even under moderate misspecifications of both models, the doubly robust estimator mostly performs well in practice.

Doubly Robust Methods

Let us suppose that $\mathbb{E}Y$ is linear in A and $\boldsymbol{X}$ separately, so

$$\mathbb{E}[Y \mid A = a, \boldsymbol{X} = \boldsymbol{x}] = \beta_A a + \beta_{\boldsymbol{X}} \boldsymbol{x}.$$

```
# get propensity score
ps <- fitted(glm(A ~ Z, data=dat, family="binomial"))
dat <- dplyr::mutate(dat, ps = ps)  # add est. propensity score

# outcome model
modY <- lm(Y ~ A + Z, data=dat)

dat0 <- dat1 <- dat    ## set 0 and 1 in mock datasets
dat0$A <- 0; dat1$A <- 1

## compute mu_x for x = {0,1}
mu1 <- mean(dat$A*(dat$Y - predict(modY))/dat$ps
            + predict(modY, dat1))

mu0 <- mean((1-dat$A)*(dat$Y - predict(modY))/(1-dat$ps)
            + predict(modY, dat0))

mu1 - mu0


[1] 0.481
```

Outline

1. Introduction
2. Single-World Intervention Graphs
3. Estimation Methods
4. Machine Learning Methods

References

Pearl, J. *Causality: Models, Reasoning, and Inference.* 3rd Ed. Cambridge, 2009.

Spirtes, P, Glymour, C, Scheines R. *Causation, Prediction, and Search.* Lecture Notes in Statistics 81, Springer-Verlag, 2000.

Wright, S. The theory of path coefficients. *Genetics*, 8: 239–255, 1923.

Wright, S. The method of path coefficients. *Annals of Mathematical Statistics*, 5(3): 161–215, 1934.