Causal Thinking

MATH-336

Mats J. Stensrud
Chair of Biostatistics
mats.stensrud@epfl.ch

Autumn 2022

Section 1

Structure of the course

Structure

- Lectures **Mondays 10h15**.
  I will use the iPad and the blackboard.
- I encourage you to ask questions along the way!
- Moodle is our main platform.
  - Announcements.
  - Problem sheets (and solutions).
  - Links to relevant literature.
  - Link to Ed Discussions.
  - All questions about the course should be asked on Ed Discussions.
- Slides and problem sheets will be uploaded every week.
  - Problem sheets will be made available on Mondays evenings.
  - Office hours with the teaching assistants: Monday 08h15-10h00.

Exam

- One graded homework. 20 % of the grade.
- Written exam. 80 % of the grade.

Goal

After the course, you should:

- understand the meaning and utility of causal models,
- understand and critically evaluate causal assumptions,
- recognize whether a research question concerns causal effect,
- design a study to answer a causal question,
- be able to translate a research question to a formal causal estimand,
- critically evaluate how causal inference is drawn in practice from data,
- suggest and implement suitable causal methods in practice.

Outline of the course

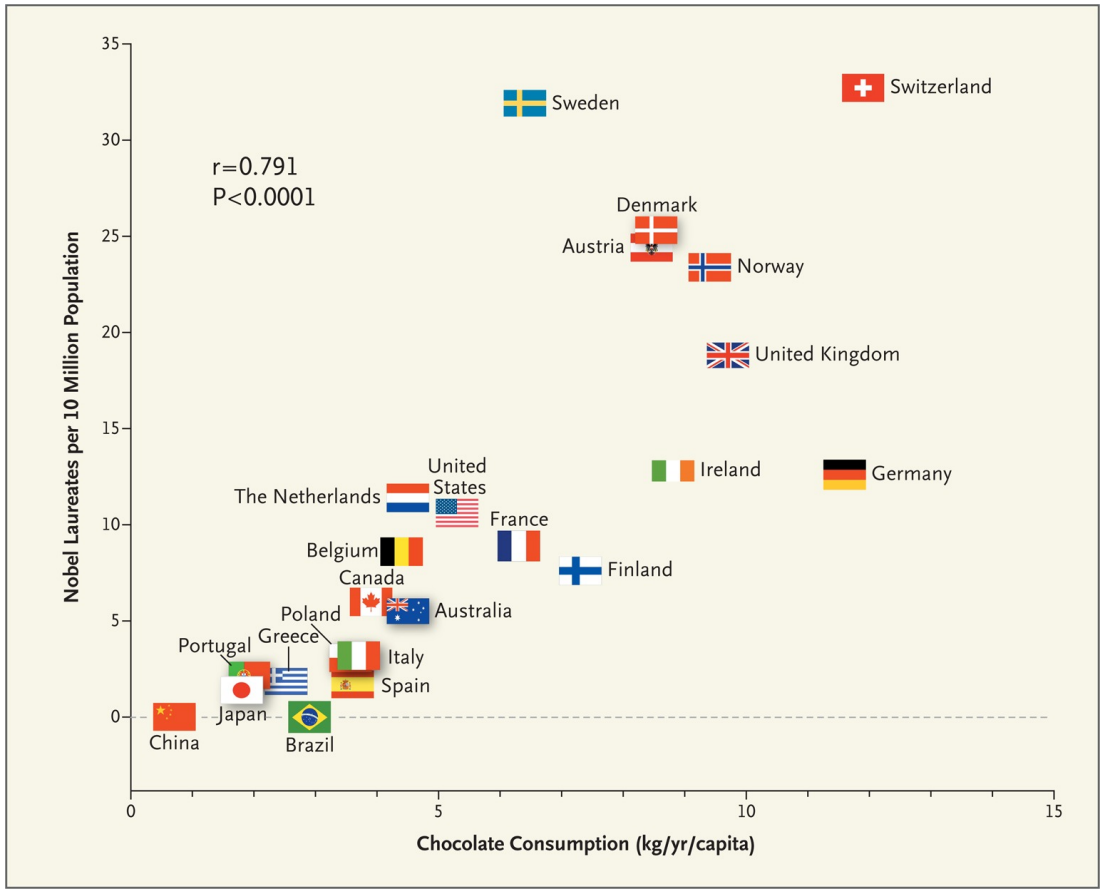
- Study the **theory for causal inference** using counterfactuals,
- See how this theory can be applied in practice,
- and study its close link to the **design of experiments**.
- Derive results for identification of causal parameters in different study designs,both experiments and observational studies.
- Causal graphs will play a key role here...
- Translate practical questions to counterfactual parameters.
- Look at examples

Why is this useful?

- Help you to gain scientific literacy
- Understand bias in data. This is very important for data science
- Pose good (causal) questions
- Think about validation

Section 2

Motivation


r=0.791
P<0.0001
Nobel Laureates per 10 Million Population
Chocolate Consumption (kg/yr/capita)
0
5
10
15
20
25
30
35
Sweden
Switzerland
Denmark
Austria
Norway
United Kingdom
Ireland
Germany
United States
The Netherlands
France
Belgium
Finland
Canada
Australia
Poland
Greece
Portugal
Italy
Spain
Japan
China
Brazil

Warm-up example: Race and death penalty

Consider a famous data set that records the race of the defendants (D) in murder cases in Florida between 1976 and 1987.[1] The outcome is death penalty (Y).

Defendant	White	Black
Yes	53	15
No	430	176

$P(Y = 1 \mid D = w) = \frac{53}{53+430} = 0.11 > P(Y = 1 \mid D = b) = \frac{15}{15+176} = 0.08.$

Now, consider death penalty conditional on the race of the victim (V):

Victim	White	
Defendant	White	Black
Yes	53	11
No	414	37

Victim	Black	
Defendant	White	Black
Yes	0	4
No	16	139

$P(Y = 1 \mid D = w, V = w) = \frac{1}{8} < P(Y = 1 \mid D = b, V = w) = \frac{1}{5}.$
$P(Y = 1 \mid D = w, V = b) = 0 < P(Y = 1 \mid D = b, V = b) = 0.03.$

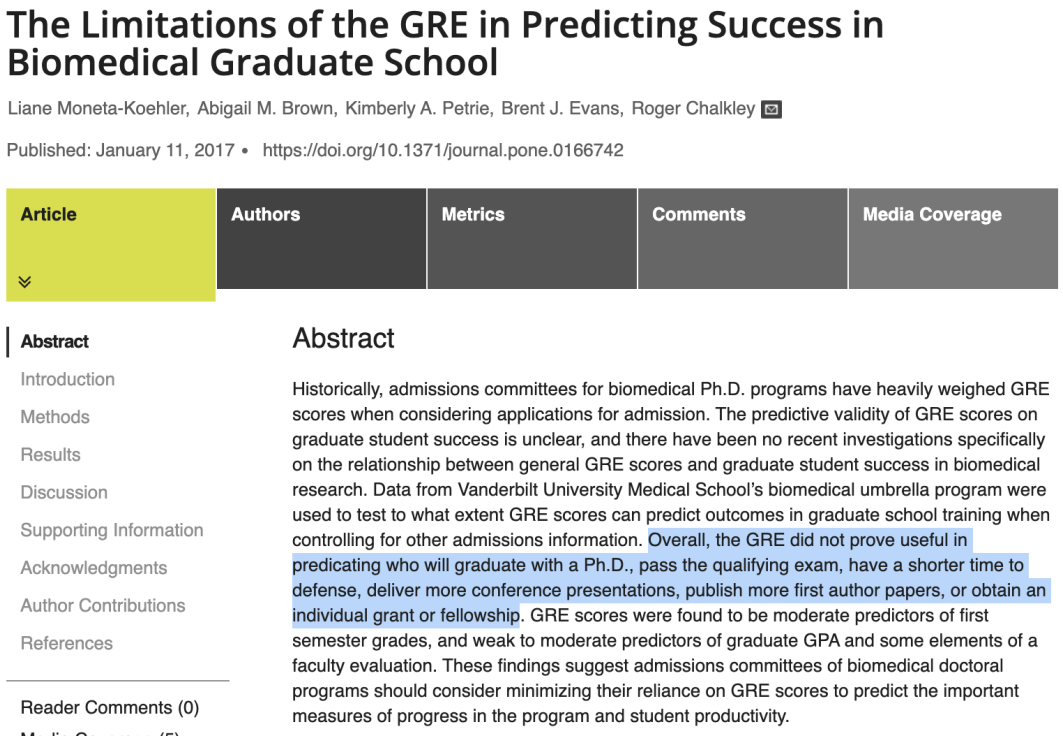
[1] From *Robin Evans*, Oxford, see also Agresti, 2002

Lessons learned from the warm-up example

- Example of Simpson's paradox (that you may be familiar with).
  By the way, paradoxes don't really exist...
- Be careful about interpreting *marginal* and *conditional* (in)dependencies.
  We will carefully (and formally) study conditional (in)dependencies in much detail in this course.
- The reason why we believe that the conditional estimates are more useful was due to **a causal story**.
  There is no statistical method that can determine the causal story from the data alone.
- How would design a study to answer the causal question "Are black defendants more likely to get death penalty just because they are black"?

Admission to universities in the USA

GRE is a test that is required to get into many PhD-programs in the USA.



The Limitations of the GRE in Predicting Success in Biomedical Graduate School

Liane Moneta-Koehler, Abigail M. Brown, Kimberly A. Petrie, Brent J. Evans, Roger Chalkley

Published: January 11, 2017 • https://doi.org/10.1371/journal.pone.0166742

Article | Authors | Metrics | Comments | Media Coverage

Abstract
Introduction
Methods
Results
Discussion
Supporting Information
Acknowledgments
Author Contributions
References

Reader Comments (0)

Abstract

Historically, admissions committees for biomedical Ph.D. programs have heavily weighed GRE scores when considering applications for admission. The predictive validity of GRE scores on graduate student success is unclear, and there have been no recent investigations specifically on the relationship between general GRE scores and graduate student success in biomedical research. Data from Vanderbilt University Medical School's biomedical umbrella program were used to test to what extent GRE scores can predict outcomes in graduate school training when controlling for other admissions information. Overall, the GRE did not prove useful in predicating who will graduate with a Ph.D., pass the qualifying exam, have a shorter time to defense, deliver more conference presentations, publish more first author papers, or obtain an individual grant or fellowship. GRE scores were found to be moderate predictors of first semester grades, and weak to moderate predictors of graduate GPA and some elements of a faculty evaluation. These findings suggest admissions committees of biomedical doctoral programs should consider minimizing their reliance on GRE scores to predict the important measures of progress in the program and student productivity.

Unlike the death penalty example, now **conditioning** (on admission) leads to an inappropriate comparison.

In this course, we will formalize how to *design studies* and *analyse data* to answer causal questions.

Suppose now that the units are individuals.

Table 1: Data from a study of A (an *exposure*) and Y (an *outcome*).

	$Y = 1$	$Y = 0$
$A = 0$	10	90
$A = 1$	5	95

$A = 1$ is getting surgery, $Y = 1$ indicates survival after 1 year.

- What does the table tell us now?
- Could we infer that surgery reduces the risk of death?

Suppose that we say this was a randomized controlled trial, where A was randomized?

Let's be explicit

Unfortunately, the scientific literature is plagued by studies in which the causal question is not explicitly stated and the investigators' unverifiable assumptions are not declared. This casual attitude towards causal inference has led to a great deal of confusion.

Questions

- Descriptive / predictive:
  - "Is this patient at high risk of developing complications during surgery?"

  Causal:
  - "Which type of anaesthetic should this patient receive to reduce the risk of complications during surgery?"
  - "How does the amount of anaesthetic affect the risk of complications during surgery?"
  - "What can be done to reduce the risk of complications during surgery for an average / a particular type of patient?"

Questions

- Descriptive / predictive:
  - "Which type of client will buy which kind of product?"

  Causal:
  - "Should advert be at the top or bottom of website to increase the probability of viewing product?"
  - "How does the size of advert affect the probability of viewing product?"
  - "How can I get a client to buy my product?"

Questions

- Descriptive / predictive:
  - "Who is most likely to become long-term unemployed?"

  Causal:
  - "Will a minimum wage legislation increase the unemployment rate of a country?"
  - "What can be done to prevent someone from becoming unemployed?"

What's the question (Hernan et al, Chance, 2019)

How can women aged 60–80 years with stroke history be partitioned in classes defined by their characteristics?
`Hernan et al, Chance (2019)`

This question is just about description.

What's the question

What is the probability of having a stroke next year for women with certain characteristics?

This question is just about prediction.

What's the question

Will starting a statin reduce, on average, the risk of stroke in women with certain characteristics?

This question is about **causal effects**, *i.e. counterfactual prediction.*

3 tasks of data scientists

- Description
- Prediction
- Counterfactual prediction (*What would happen if...*)

Section 3

Prediction vs. causal inference

Prediction and causal inference are different exercises

- Prediction: Learn about Y after observing $X = x$.
- Causal inference: Learn about Y after ~~observing~~ setting $X = x$.

Pearl

Figure 1: Judea Pearl

"All the impressive achievements of deep learning amount to just curve fitting"

Motivation

Albert Einstein (1953):

> *"Development of Western science is based on two great achievements: the invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and the discovery of the possibility to find out causal relationships by systematic experiment (during the Renaissance)".*

Experiments

- Experiment (biology).
- The randomized controlled trial (medicine).
- A/B testing (tech industry).

Scurvy - the first randomized trial?

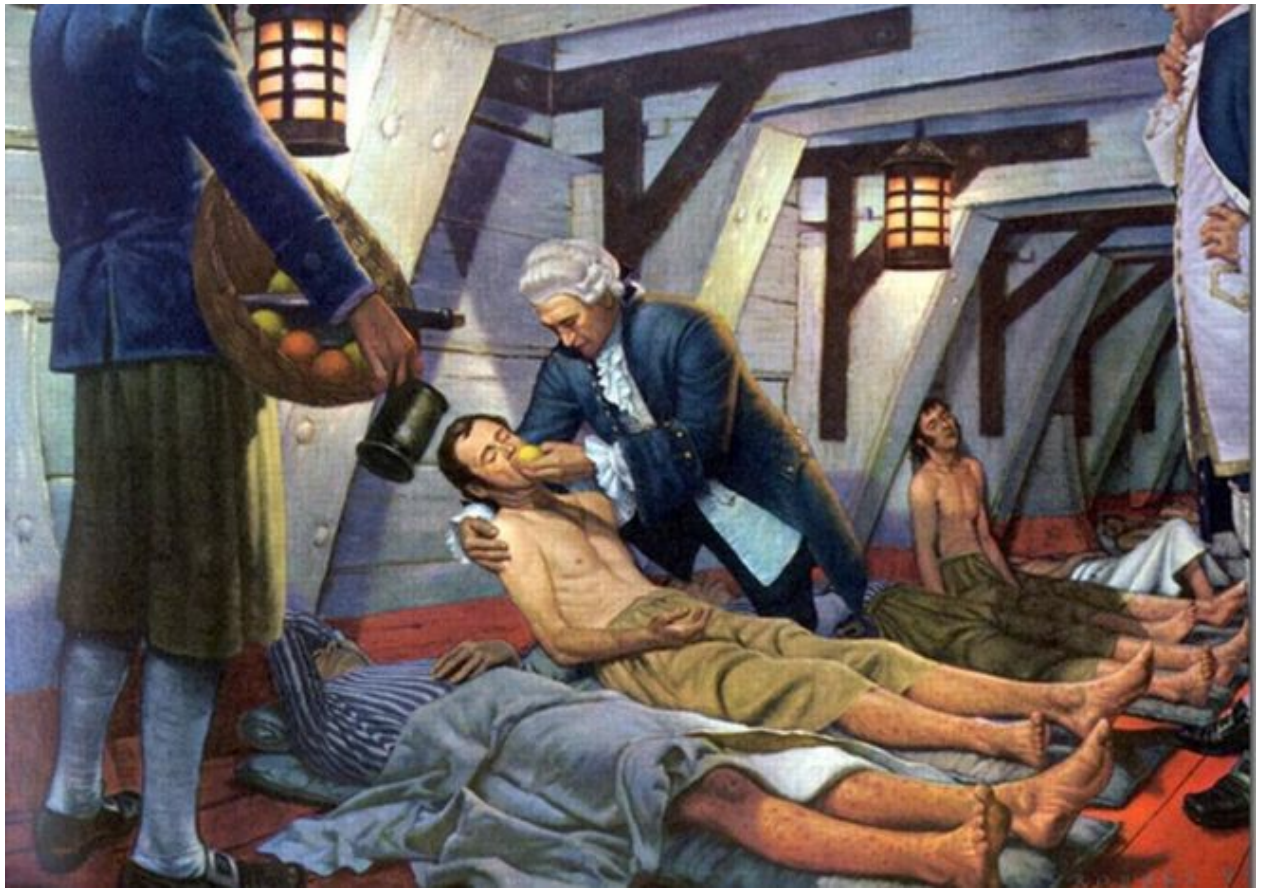


Figure 2: James Lind, the surgeon, 1753.

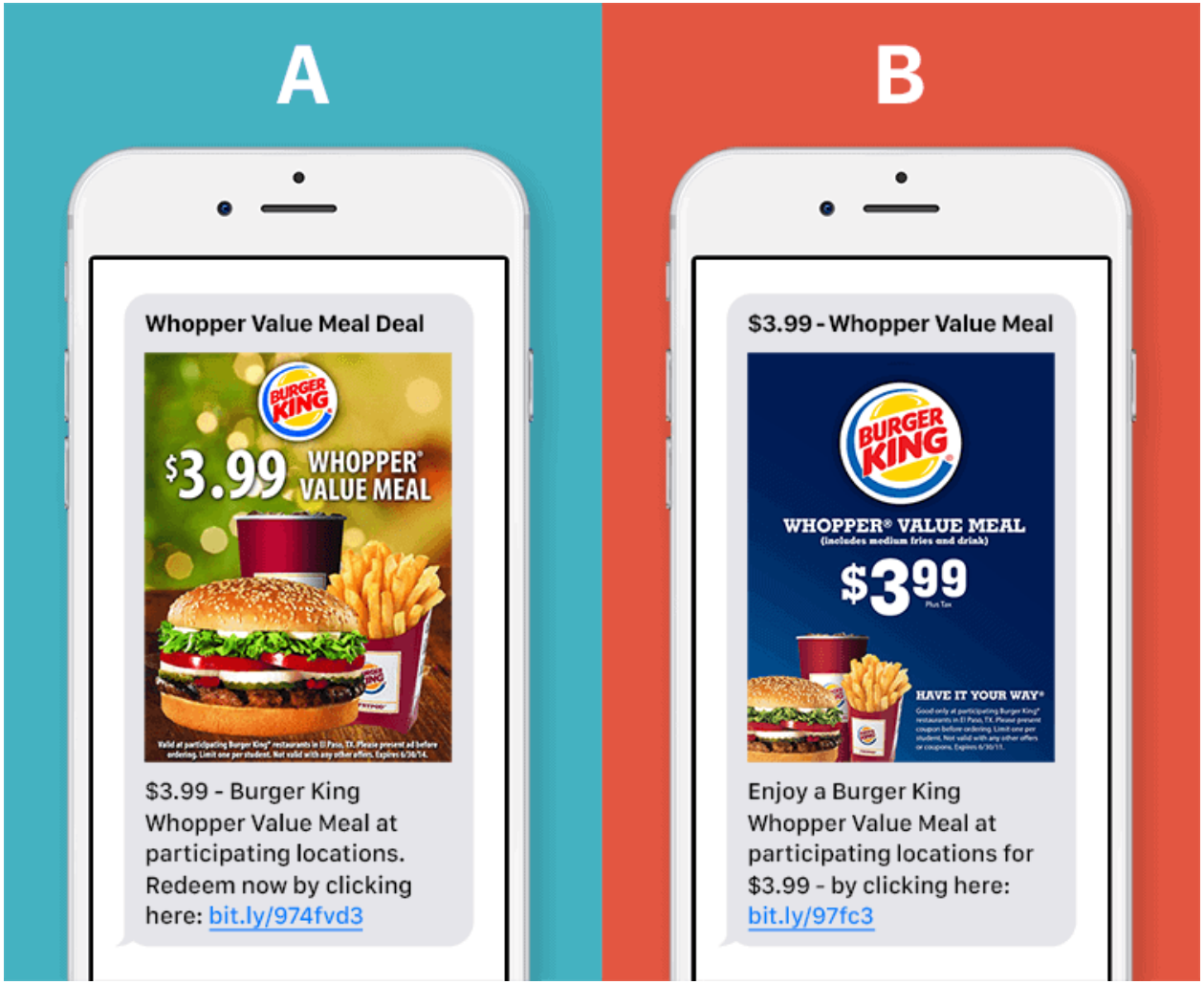
https://www.bbc.com/news/uk-england-37320399

Lind's experimental set-up (simplified)

- Recruit a bunch of sailors suffering from scurvy
- Flip a coin for each sailor to determine course of action A
  - Heads: $A = 1$ (a lemon a day)
  - Tails: $A = 0$ (elixir of vitriol a.k.a sulphuric acid
- For each sailor note down the outcomes denoted by Y : let's say $Y = 1$ is healthy and $Y = 0$ is sick with scurvy
- This is an example of a simple randomized controlled trial (RCT), also called A/B test.

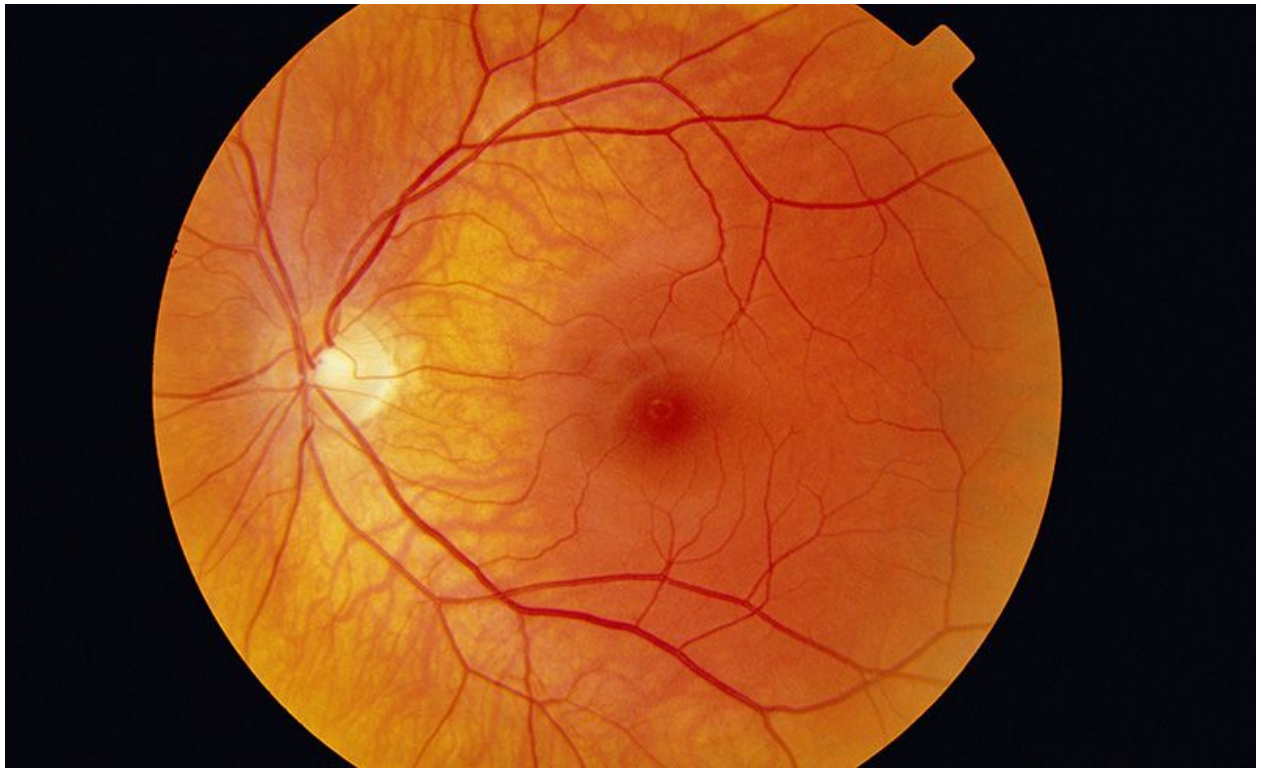
An aerial view of Rothamsted's Broadbalk field, site of the Broadbalk Wheat Experiment since 1843.

In Silicon Valley they call it AB testing.



Why bother?

Why bother?



"Can (...) predict whether someone is at risk of an impending heart attack",
Nature Biomedical Engineering, 2018

Decisions have to be made...

What if...

- Would starting treatment A prevent a heart attack?
- Is Drug A better than Drug B?
- Would the ad get more clicks if it were green instead of red?
- Would the election campain increase the number of votes?
- Would university education increase my future earnings?

What if questions can be assessed in experiments

... but experiments are often not available because they are

- impractical,
- expensive,
- time consuming,
- unethical,

... and experiments may not be perfectly executed.

So, what do we do?

Emulate the experiment of interest from available *observational* data.

some nuances?

https://becominghuman.ai/summary-of-the-alphago-paper-b55ce24d8a7c

Section 4

Counterfactuals

What would happen if...

Counterfactuals 1973:

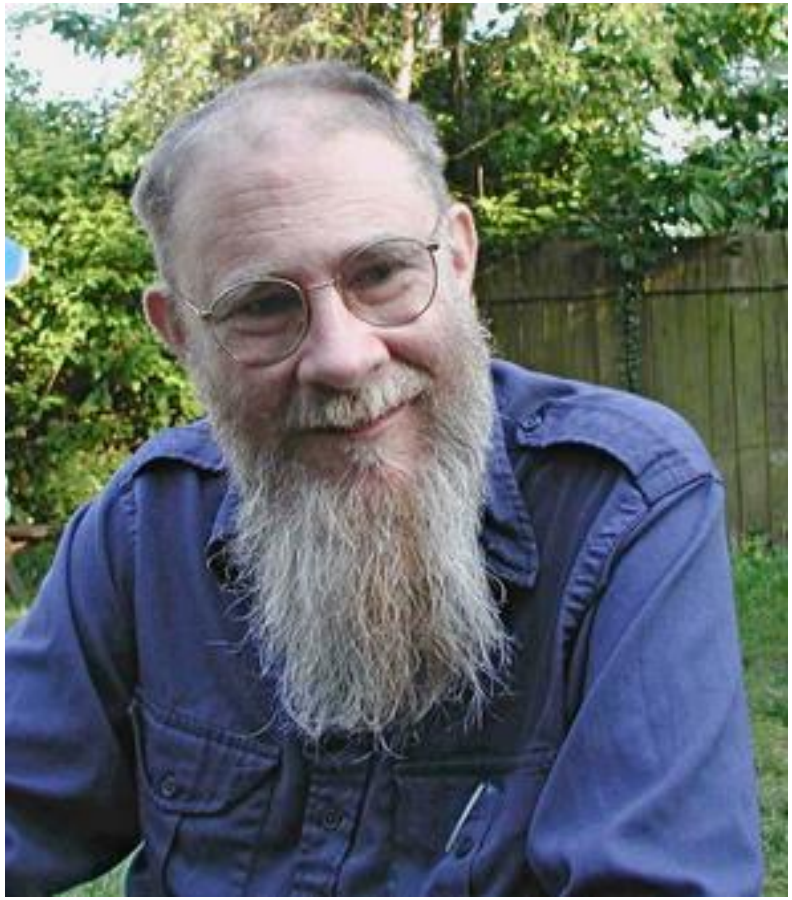


Figure 3: David Lewis

https://en.wikipedia.org/w/index.php?curid=58724625

Prediction and causal inference are different exercises

- Prediction: Learn about Y after observing $A = a$.
- Previously you have studied random variables *conditional* on parameters,

$$Y \sim \text{Distribution}\{g(a)\}.$$

Such (conditional) associations are not necessarily easy to interpret (see the next example).

In this course, we will study causal problems, which are different exercises.

Two fundamental questions of causality (Pearl, 2009)

1. What empirical evidence is required for legitimate inference of cause–effect relationships?
2. Given that we are willing to accept causal information about a phenomenon, what inferences can we draw from such information, and how?

In this course, we will consider **mathematical tools** for **casting causal questions** or **deriving causal answers**.

Section 5

Prediction vs. causal inference

Prediction and causal inference are different exercises

- Prediction: Learn about Y after observing $A = a$.
  That is, infer properties of the law P that generated the observations Y.
- Causal inference: Learn about Y after ~~observing~~ fixing $A = a$.
  That is, infer properties of a *counterfactual* law, say, P^a,that would generate data when a is fixed.

What is a causal effect (in a simple setting)

Consider the following observed random variables:

- A binary treatment $A \in \{0, 1\}$.
- An outcome $Y \in \mathcal{Y}$.
- A vector of baseline covariates $L \in \mathcal{L}$.

Define the *counterfactual* or *potential outcome* variables

- $Y^a \in \mathcal{Y}$.
  The outcome variable that would have been observed under the treatment value a (the superscript denotes the counterfactual).
- Often we will specifically instantiate a, i.e. set a to a value:
  $Y^{a=0} \in \mathcal{Y}$.
  The outcome variable that would have been observed under the treatment value $a = 0$.
  $Y^{a=1} \in \mathcal{Y}$.
  The outcome variable that would have been observed under the treatment value $a = 1$.

Individual level causal effect

Definition (Individual level causal effect)

A causal effect for individual (unit) i is $Y_i^{a=0}$ vs $Y_i^{a=1}$.

Remarks on $Y^{a=0}$ and $Y^{a=1}$

The fundamental problem of causal inference:

- Suppose $A = 1$. Then $Y = Y^{a=1}$ is observed, but $Y^{a=0}$ is unobserved...
- Suppose $A = 0$. Then $Y = Y^{a=0}$ is observed, but $Y^{a=1}$ is unobserved...

The consequence is that individual level effect cannot be identified.[2]

[2] We will consider a possible exception later.

Intervening is not the same as conditioning

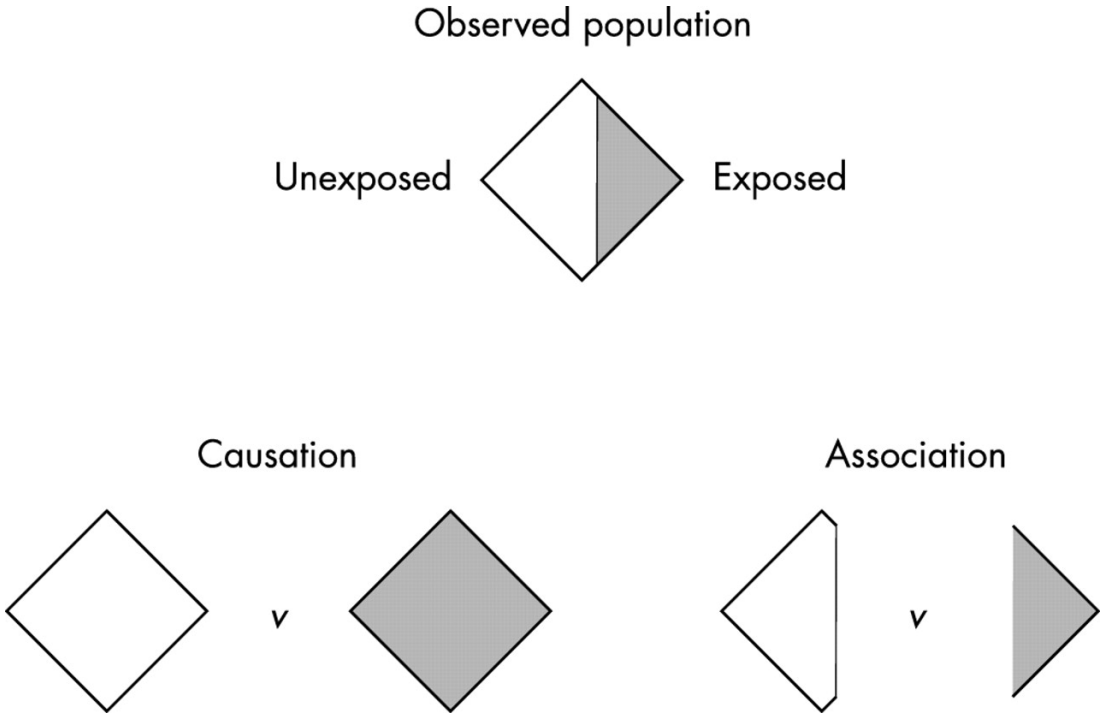


Figure taken from Hernan, 2014, BMJ.

Section 6

Lecture 2

We make decisions based on "what if" questions...

Some repetition from the first lecture.

- Would starting treatment A prevent a heart attack?
- Is Drug A better than Drug B?
- Would the election campaign increase the number of votes?
- Would university education increase my future earnings?
- What would happen if I went to UNIGE instead of EPFL?

Section 7

Defining a causal effect

Counterfactuals a.k.a. potential outcomes

- We will posit unobserved fixed potential or counterfactual outcomes[3] for each unit[4] under different treatments[5]

  Hint: It is helpful to think about a counterfactual random variable as a variable that **does exist** in this world, even before interventions take place, but we are not able to observe it.

- We will use superscripts to indicate that a random variable is counterfactual. For example consider a random variable Y. A counterfactual version Y^g is the value Y would have had under an intervention g (also called treatment regime or treatment strategy).
- To get started, in the first lectures, we will consider some simple interventions g which only fixes a binary treatment A to a value $a \in \{0, 1\}$.

[3] I will use the terms "counterfactuals" and "potential outcomes". interchangeably.
[4] I will use the terms "unit", "subject" and "individual" interchangeably.
[5] I will use the terms "treatment" and "exposure" interchangeably.

Illustrative experiment (trial) on heart transplant.

Assess the effect of $A \in \{0, 1\}$ (1 if heart transplant, 0 otherwise) on $Y \in \{0, 1\}$ (1 if dead, 0 otherwise)[6].

	A	Y	Y^0	Y^1
Rheia	0	0	0	?
Kronos	0	1	1	?
Demeter	0	0	0	?
Hades	0	0	0	?
Hestia	1	0	?	0
Poseidon	1	0	?	0
Hera	1	0	?	0
Zeus	1	1	?	1
Artemis	0	1	1	?
Apollo	0	1	1	?
Leto	0	0	0	?
Ares	1	1	?	1
Athena	1	1	?	1
Hephaestus	1	1	?	1
Aphrodite	1	1	?	1
Cyclope	1	1	?	1
Persephone	1	1	?	1
Hermes	1	0	?	0
Hebe	1	0	?	0
Dionysus	1	0	?	0

[6] Miguel A Hernan and James M Robins. *Causal inference: What if?* CRC Boca

Remark on counterfactuals

The definition of counterfactuals presupposes:

- $Y^a = Y$ for every unit with $A = a$. In other words, $Y^A = Y$.
  "Consistency". That is, $Y = I(A = 0)Y^{a=0} + I(A = 1)Y^{a=1}$.
  This "consistency" assumption requires that
  - The intervention on A is well-defined.
    No matter how unit i received treatment a, the outcome Y^a is the same.
  - The counterfactual outcome of unit i does not depend on the treatment values of other units j, that is, "no interference".
    Otherwise Y_i^a is not well-defined.[7]

[7] This use of consistency is different from the use in estimation.

More on consistency

An example of an ill-defined intervention:
Imagine A is a person's body mass index (BMI). Setting the BMI to a counterfactually different level can happen in many different ways - losing weight by running, loss of appetite due to chain smoking, liposuction etc. Depending on what way the intervention is implemented each time, we will have very different health outcomes, i.e., re-running the experiment will give inconsistent results.
Another example is infectious diseases

The task of identification

Definition (Identification)

A parameter is said to be **identified** under a particular collection of assumptions if it can be expressed uniquely as a function of the distribution (law) of the observed variables.

That is, a parameter (*estimand*) is identified under a particular collection of assumptions if these assumptions imply that the distribution of the observed data is compatible with a single value of the parameter.

Average causal effect

Definition (Average causal effect)

$\mathbb{E}(Y^{a=0})$ vs $\mathbb{E}(Y^{a=1})$.

- Average causal effects can sometimes be identified from data (we will study this extensively).
- In most of this course, average causal effects will be our parameters of interest, i.e. our target estimands.

A more formal argument why randomisation is the gold standard

- In a randomised experiment, the treatment is assigned independently of all other factors (e.g. by a coin flip or a random number generator).
- In a randomised experiment one of the counterfactual outcomes $Y^{a=0}$ or $Y^{a=1}$ is unobserved.
- However, randomisation ensures that it is *random* whether $Y^{a=0}$ or $Y^{a=1}$ is unobserved, that is,

$$P(Y^a = y \mid A = 1) = P(Y^a = y \mid A = 0), \forall a \in \{0, 1\}, \forall y \in \mathcal{Y}.$$

because the treatment assignment is independent of all other factors, including the counterfactual outcomes (Y^a). This conditional independence is called **exchangeability**.

Independence notation

Definition (Conditional independence)

$X \perp\!\!\!\perp Y \mid Z \iff F_{X,Y|Z=z}(x,y) = F_{X|Z=z}(x) \cdot F_{Y|Z=z}(y) \ \forall\ x, y, z,$
where $F_{X,Y|Z=z}(x,y) = P(X \leq x, Y \leq y \mid Z = z)$.

We say that X and Y are conditionally independent given Z.
In other words, when $Z = z$ is known, X provides no additional information that allows us to *predict* Y.

Exchangeability (re-visited)

In particular, we can re-write the condition from Slide 55,

$$P(Y^a = y \mid A = 1) = P(Y^a = y \mid A = 0), \forall a \in \{0, 1\}, \forall y \in \mathcal{Y},$$

as

$$Y^a \perp\!\!\!\perp A, \forall a \in \{0, 1\}.$$

Effect contrasts

- Additive effect: $\mathbb{E}(Y^{a=1}) - \mathbb{E}(Y^{a=0}) = \mathbb{E}(Y^{a=1} - Y^{a=0})$.
  The additive effect is an average over individual level causal effects. These are marginal quantities.
- Relative effect: $\frac{\mathbb{E}(Y^{a=1})}{\mathbb{E}(Y^{a=0})} \neq \mathbb{E}\left(\frac{Y^{a=1}}{Y^{a=0}}\right)$.
  The relative effect is not an average over individual level causal effects.

Causal effects in the population

More generally, we can consider population causal effects[8]:

Definition (Population causal effect)

A population causal effect can be defined as a contrast of any functional of the marginal distributions of counterfactual outcomes under different interventions.

- For example $\text{VAR}(Y^{a=1}) - \text{VAR}(Y^{a=0})$.
  Remember that we cannot identify $\text{VAR}(Y^{a=1} - Y^{a=0})$.
- From here on, I will often say *causal effect* when I talk about *average causal effect*.

[8] Hernan and Robins, *Causal inference: What if?*

Section 8

Randomisation

Example conditions that ensure identification of causal effects

Suppose that the following 3 conditions hold:

1. $Y^a \perp\!\!\!\perp A, \forall a \in \{0,1\}$ (exchangeability[9]).
2. $P(A = a) > 0 \ \forall a \in \{0,1\}$ (positivity[10]).
3. $Y^a = Y$ for every unit with $A = a$ (consistency[11]).
   that is, $Y = I(A = 0)Y^{a=0} + I(A = 1)Y^{a=1}$.

From (1)-(3), $\mathbb{E}(Y^a) = \mathbb{E}(Y \mid A = a)$.
That is, we have *identified* $\mathbb{E}(Y^a)$ as a functional of observed data.
Assumptions (1)-(3) are external to the data, but – importantly – they hold by design in a perfectly executed experiment.
Just to be clear: The counterfactual independence $Y^a \perp\!\!\!\perp A, \forall a \in \{0,1\}$ does NOT imply the factual independence $Y \perp\!\!\!\perp A$.

[9] Also called ignorability.
[10] Also called overlap. Note that this is a feature of the distribution, not the sample.
[11] Similar to the condition SUTVA: Stable Unit Treatment Value Assumption.

Side note: relation to previous statistics courses

- So far you have considered random variables, say, Y.
- Y has a law – i.e. distribution – and you have *inferred*, *i.e. estimated* features of this law: deterministic features of this random variable.
- In the regression courses, you went further and looked at random variables *conditional* on parameters. For example, linear regression is the best (min squared error) linear approximation of Y (or of $\mathbb{E}[Y \mid A]$). where x is a parameter.
- We consider the problem of **identifying** functionals $f(Y^a)$.
  If a functional is identified, then we can use what you have learned so far (and more) **to estimate** these functionals.

Terminology

Remember the difference between the following terms:

- **Estimand** (a parameter of interest, often a causal effect).
- **Estimator** (an algorithm / function that can be applied to data).
- **Estimate** (an output from applying the estimator to data).

We talk about bias of an estimator with respect to an estimand.
That is, the term *bias* (biased / unbiased) is defined with respect to an estimand.

Terminology

estimand

Ingredients

- 150g unsalted butter, plus extra for greasing
- 150g plain chocolate, broken into pieces
- 150g plain flour
- ½ tsp baking powder
- ½ tsp bicarbonate of soda
- 200g light muscovado sugar
- 2 large eggs

Method

1. Heat the oven to 160C/140C fan/gas 3. Grease and base line a 1 litre heatproof glass pudding basin and a 450g loaf tin with baking parchment.

2. Put the butter and chocolate into a saucepan and melt over a low heat, stirring. When the chocolate has all melted remove from the heat.

estimator

estimate

A simple example of estimation of causal effects

Because $\mathbb{E}(Y^a) = \mathbb{E}(Y \mid A = a)$, the simple difference-in-means estimator,

$$\hat{\delta} = \frac{1}{n_1} \sum_{A_i=1} Y_i - \frac{1}{n_0} \sum_{A_i=0} Y_i, \;\; n_a = \sum_{i=1}^{n} I(A = a),$$

is an unbiased estimator of the average (additive) causal effect of A in a randomised experiment.

We will discuss estimation in more detail later in this course.

Conditional randomisation

- Let $L \in \{0, 1\}$
  In the heart transplant example, let $L = 1$ if the individual is critically ill, 0 otherwise.
- Suppose A is *conditionally* randomised as a function of L such that $P(A = 1 \mid L = 0) = p_0$ and $P(A = 1 \mid L = 1) = p_1$,
  where $p_0 \neq p_1$ and $p_0, p_1 \in (0, 1)$.

How do we identify $\mathbb{E}(Y^a)$?

Illustrative *conditional* experiment (trial) on heart transplant

	L	A	Y
Rheia	0	0	0
Kronos	0	0	1
Demeter	0	0	0
Hades	0	0	0
Hestia	0	1	0
Poseidon	0	1	0
Hera	0	1	0
Zeus	0	1	1
Artemis	1	0	1
Apollo	1	0	1
Leto	1	0	0
Ares	1	1	1
Athena	1	1	1
Hephaestus	1	1	1
Aphrodite	1	1	1
Cyclope	1	1	1
Persephone	1	1	1
Hermes	1	1	0
Hebe	1	1	0
Dionysus	1	1	0

In this conditional randomised trial $p_0 = 0.5, p_1 = 0.75$
Compute an estimator based on the numbers above, and you will find that $\hat{\mathbb{E}}(Y^{a=1}) - \hat{\mathbb{E}}(Y^{a=0}) = 0$.

Identification in a conditional randomised experiment

A is *conditionally* randomised such that $P(A = 1 \mid L = 0) = p_0$ and $P(A = 1 \mid L = 1) = p_1$, where $p_0 \neq p_1$ and $p_0, p_1 \in (0, 1)$.

$Y^a \not\perp\!\!\!\perp A, \forall a \in \{0, 1\}$ (Exchangeability from Slide 62 may fail), but

1. $Y^a \perp\!\!\!\perp A \mid L, \forall a \in \{0, 1\}$ (Exchangeability).
2. $P(A = a \mid L = l) > 0 \ \forall a \in \{0, 1\}, \forall l$ s.t. $P(L = l) > 0$. (positivity).
3. $Y^a = Y$ for every unit with $A = a$ (consistency).

When 1-3 hold, then

$$\mathbb{E}(Y^a) = \sum_l \mathbb{E}(Y \mid L = l, A = a) P(L = l).$$

These conditions hold by design in a conditional randomised experiment.

Identification in a conditional randomised experiment

Proof.

$$\begin{aligned}
\mathbb{E}(Y^a) &= \sum_l \mathbb{E}(Y^a \mid L = l) P(L = l) \\
&= \sum_l \mathbb{E}(Y^a \mid L = l, A = a) P(L = l) \quad \text{(exchangeability)} \\
&= \sum_l \mathbb{E}(Y \mid L = l, A = a) P(L = l). \quad \text{(positivity and consistency)}
\end{aligned}$$

□

We say that the 3rd line is an identification formula for $\mathbb{E}(Y^a)$.
This is a special case of a so-called G-formula (or truncation formula)[12].

[12] James M Robins. "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". In: *Mathematical modelling* 7.9-12 (1986), pp. 1393–1512.

Alternative weighted identification formula

$$\mathbb{E}(Y^a) = \sum_l \mathbb{E}(Y \mid L = l, A = a) \Pr(L = l)$$
$$= \mathbb{E}\left[\frac{I(A = a)}{\pi(A \mid L)} Y\right].$$

where $\pi(a \mid l) = P(A = a \mid L = l)$.
Why bother with equivalent expressions?
Because they motivate different *estimators*.

Proof of IPW

Proof.

$$
\begin{aligned}
&\mathbb{E}\left[\frac{I(A=a)}{\pi(A \mid L)} Y\right] \\
=&\mathbb{E}\left[\frac{I(A=a)}{P(A=a \mid L)} Y^a\right] \quad \text{(consistency and positivity)} \\
=&\mathbb{E}\left[\mathbb{E}\left\{\frac{I(A=a)}{P(A=a \mid L)} Y^a \mid L\right\}\right] \\
=&\mathbb{E}\left\{\mathbb{E}\left[\frac{I(A=a)}{P(A=a \mid L)} \mid L\right] \mathbb{E}\left[Y^a \mid L\right]\right\} \quad \text{(exchangeability)} \\
=&\mathbb{E}\left\{\mathbb{E}\left[Y^a \mid L\right]\right\} = \mathbb{E}\left[Y^a\right].
\end{aligned}
$$

Section 9

Causal inference from observational data

Observational data

Definition (Observational data)

A sample from a population where the treatment (exposure) is not under the control of the researcher.

That is, the treatment (exposure) of interest is not randomly assigned.

Following Robins[14], let's be slightly more abstract

- A dataset is a string of numbers.
- These data represent empirical measurements (for example, for each study subject, a series of treatments and outcomes).
- In an analysis, calculations are performed on these numbers.
- Based on the calculations, causal inference is drawn.
- "*Since the numerical strings and the computer algorithm applied to them are well-defined mathematical objects, it would be important to provide formal mathematical definitions for the English sentences expressing the investigator's causal inferences that agree well with our informal intuitive understanding*"[13].

[13] James M Robins. "Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect"". In: *Computers & Mathematics with Applications* 14.9-12 (1987), pp. 923–945.

[14] Robins, "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect".

Observational studies

- In an observational study, treatment is not assigned according to randomisation, but according to someone's choice, for example the patient, the costumer or the medical doctor.
- People who choose to take treatment may be different from those who choose not to take treatment, in the sense that they have different risk of the outcome even before the decision is made. $Y^a \not\perp\!\!\!\perp A, \forall a \in \{0, 1\}$.
- The question is, can we find the characteristics L, which are associated with treatment and the outcome such that $Y^a \perp\!\!\!\perp A \mid L, \forall a \in \{0, 1\}$?
  In other words, exchangeability does no longer hold by design, but can we *assume* that it holds? What do we need to include in L for this to hold?
- Yet, humans have learned a lot from *observations*, and many scientific studies are not experiments. We have learned about *effects of* smoking, global warming, evolution, astrophysics etc.

Same data, different story

Suppose the data (identical numbers to the slide 69) were from an *observational* study (now A is not randomly assigned), where the doctors tended to provide transplants ($A = 1$) to those with most severe disease ($L = 1$)

	L	A	Y
Rheia	0	0	0
Kronos	0	0	1
Demeter	0	0	0
Hades	0	0	0
Hestia	0	1	0
Poseidon	0	1	0
Hera	0	1	0
Zeus	0	1	1
Artemis	1	0	1
Apollo	1	0	1
Leto	1	0	0
Ares	1	1	1
Athena	1	1	1
Hephaestus	1	1	1
Aphrodite	1	1	1
Cyclope	1	1	1
Persephone	1	1	1
Hermes	1	1	0
Hebe	1	1	0
Dionysus	1	1	0

Example continues

- Suppose first that L is the *only* outcome predictor unequally distributed between those with $A = 1$ and $A = 0$. Then $Y^a \perp\!\!\!\perp A \mid L, \forall a \in \{0, 1\}$.
- Now, suppose that the doctors not only used L to make treatment decisions, but also used smoking status, $S \in \{0, 1\}$, where smoking status is an outcome predictor. Then, $Y^a \not\perp\!\!\!\perp A \mid L, \forall a \in \{0, 1\}$.
- Thus, $Y^a \perp\!\!\!\perp A \mid L, \forall a \in \{0, 1\}$ may not hold in observational studies.
- Suppose the investigators did not measure S. Can they use the observed data to evaluate whether $Y^a \not\perp\!\!\!\perp A \mid L, \forall a \in \{0, 1\}$ holds? The answer is *no*.

More on consistency

- Consistency requires well-defined interventions.
- How do we reason about exchangeability for a treatment A that is ill-defined?
- Suppose now that our exposure (treatment) is obesity A.
- How can we identify common causes of obesity L and the outcome mortality Y? Difficult when we don't even have a sufficiently specified A
- And does positivity hold? There can be some Ls (say, related to exercise) for which nobody is obese.
- The target trial where obesity is the exposure seems to involve unreasonable interventions. How can we instantly make people non-obese? By forcing them to exercise? By doing surgery? By diet? All of these interventions may have different effects.

Section 10

Effect modification and conditional effects

THE PRECISION MEDICINE INITIATIVE

PRECISION MEDICINE | INITIATIVE | PRINCIPLES | STORIES

GO TO TOP

"Doctors have always recognized that every patient is unique, and doctors have always tried to tailor their treatments as best they can to individuals. You can match a blood transfusion to a blood type — that was an important discovery. What if matching a cancer cure to our genetic code was just as easy, just as standard? What if figuring out the right dose of medicine was as simple as taking our temperature?"

- President Obama, January 30, 2015

Effect modification

Definition (Effect modification)

We say that V is a modifier of the effect of A on Y when the average causal effect of A on Y varies across levels of V.

Since the average causal effect can be be defined on different scales, effect modification depends on the scale.

Definition (Qualitative effect modification)

We say there is qualitative effect modification if the average causal effects if there exist v, v' such that the effect given $V = v$ are in the opposite direction of effects given $V = v'$.

Note that:

- V may or may not be equal to L.
- *"Effect heterogeneity across strata of V"* is often used interchangeably with *"effect modification by V"*.

Why bother with effect modification?

- So far we have focused on average causal effects.
- However, effects will often be different in different subpopulations of individuals (between men and women, Greek and Romans etc.).
- It is often of practical interest to target future intervention to subsets of the full population (If the treatment has a positive effect in men and negative effect in women, we would like to give men and women different treatments).
- Some individuals will have different benefit of treatment than others (towards precision medicine and personalised medicine...).
- Later in the course, we will also see that this is important when we are going to *generalize* (or *transport*) effects from a study to other populations (for example, we have done an experiment in a selected population, and now we want to make decisions in another population. Therefore our question is how the intervention will work in this other population).

Illustrative experiment (trial) on heart transplant.

We may be interested in effects conditional on a baseline variable V.

	V	Y^0	Y^1
Rheia	1	0	1
Demeter	1	0	0
Hestia	1	0	0
Hera	1	0	0
Artemis	1	1	1
Leto	1	0	1
Athena	1	1	1
Aphrodite	1	0	1
Persephone	1	1	1
Hebe	1	1	0
Kronos	0	1	0
Hades	0	0	0
Poseidon	0	1	0
Zeus	0	0	1
Apollo	0	1	0
Ares	0	1	1
Hephaestus	0	0	1
Cyclope	0	0	1
Hermes	0	1	0
Dionysus	0	1	0

Here, $V = 1$ if woman, $V = 0$ if man.

Concrete example

Suppose that:

- $\mathbb{E}(Y^{a=1} \mid V = 1) = 0.6 > \mathbb{E}(Y^{a=0} \mid V = 1) = 0.4.$
- $\mathbb{E}(Y^{a=1} \mid V = 0) = 0.4 < \mathbb{E}(Y^{a=0} \mid V = 0) = 0.6.$

We conclude that there is qualitative effect modification by gender.
Treatment $A = 1$

- increases mortality in women, but
- reduces mortality in men.

Let $P(V = 0) = 0.5$. Then, the average causal effect
$\mathbb{E}(Y^{a=1}) - \mathbb{E}(Y^{a=0}) = 0.$

Identification of effects modified by V.

For simplicity suppose that V and L are disjoint (different random variables).

1. $Y^a \perp\!\!\!\perp A \mid L, V, \forall a \in \{0, 1\}$ (Exchangeability).
2. $P(A = a \mid L = l, V = v) > 0 \; \forall a \in \{0, 1\}, \forall l \in \mathcal{L}, \forall v \in \mathcal{V}$ (Positivity).
3. $Y^a = Y$ for every unit with $A = a$ (Consistency).

How to identify effect modification

- Strategy for identification:
  1. Stratify by V.
  2. Identify the effect within each level $V = v$.
- For example, in a conditional randomised trial, an identification formula for the average causal effect of $A = a$ in the stratum defined by $V = v$ is

$$\mathbb{E}(Y^a \mid V = v) = \sum_l \mathbb{E}(Y \mid L = l, V = v, A = a) P(L = l \mid V = v).$$

Romans vs Greeks.

Consider a conditional randomised study on Heart transplant, and let V indicate whether the individual is Roman ($V = 0$) or Greek ($V = 1$)[15]

Stratum $V = 0$

	L	A	Y
Cybele	0	0	0
Saturn	0	0	1
Ceres	0	0	0
Pluto	0	0	0
Vesta	0	1	0
Neptune	0	1	0
Juno	0	1	1
Jupiter	0	1	1
Diana	1	0	0
Phoebus	1	0	1
Latona	1	0	0
Mars	1	1	1
Minerva	1	1	1
Vulcan	1	1	1
Venus	1	1	1
Seneca	1	1	1
Proserpina	1	1	1
Mercury	1	1	0
Juventas	1	1	0
Bacchus	1	1	0

[15] Hernan and Robins, *Causal inference: What if?*

Concrete example from Slide 86

Suppose that:

- $\mathbb{E}(Y^{a=1}) = 0.55$ and $\mathbb{E}(Y^{a=0}) = 0.40$.
- $\mathbb{E}(Y^{a=1} \mid V = 1) = 0.5 = \mathbb{E}(Y^{a=0} \mid V = 1) = 0.5$ (in Greeks).
- $\mathbb{E}(Y^{a=1} \mid V = 0) = 0.6 > \mathbb{E}(Y^{a=0} \mid V = 0) = 0.3$. (in Romans)

We conclude that there is effect modification by nationality.

Section 11

Interaction is different from effect modification

Interaction requires multiple interventions

- Consider two binary treatments $A \in \{0, 1\}$ and $E \in \{0, 1\}$.
  For example, chemotherapy and surgery.
- For each individual we can imagine 4 potential outcomes, that is, $Y^{a=0,e=0}$, $Y^{a=1,e=0}$, $Y^{a=0,e=1}$ and $Y^{a=1,e=1}$.

Definition (Additive interaction)

There is additive interaction if

$$\mathbb{E}(Y^{a=0,e=0}) - \mathbb{E}(Y^{a=1,e=0}) \neq \mathbb{E}(Y^{a=0,e=1}) - \mathbb{E}(Y^{a=1,e=1}).$$

Additive interaction is symmetric wrt. A and E,

$$\begin{aligned} & \mathbb{E}(Y^{a=0,e=0}) - \mathbb{E}(Y^{a=1,e=0}) \neq \mathbb{E}(Y^{a=0,e=1}) - \mathbb{E}(Y^{a=1,e=1}) \\ \implies & \mathbb{E}(Y^{a=0,e=0}) - \mathbb{E}(Y^{a=0,e=1}) \neq \mathbb{E}(Y^{a=1,e=0}) - \mathbb{E}(Y^{a=1,e=1}). \end{aligned}$$

`Remember that, unlike interactions, effect heterogeneity did only involve interventions on` A, `not the modifier` V.

Multiplicative interaction

Definition (Multiplicative interaction)

There is multiplicative interaction if

$$\frac{\mathbb{E}(Y^{a=0,e=0})}{\mathbb{E}(Y^{a=1,e=0})} \neq \frac{\mathbb{E}(Y^{a=0,e=1})}{\mathbb{E}(Y^{a=1,e=1})}.$$

Example: Interaction

- A chemotherapy, E radiation therapy, Y being cured of cancer.
- Interaction question: Is there interaction between the effect of receiving both A chemotherapy and E radiation therapy?

	E = 0	E = 1
A = 0	0.02	0.05
A = 1	0.04	0.10

Table 2: Experiment where A and E are randomised[16]

[16] Tyler J VanderWeele and Mirjam J Knol. "A tutorial on interaction". In: *Epidemiologic Methods* 3.1 (2014), pp. 33–72.

Conceptual example

- Let Y indicate being cured. There is additive interaction because

$$\mathbb{E}(Y^{a=0,e=0}) - \mathbb{E}(Y^{a=1,e=0}) \neq \mathbb{E}(Y^{a=0,e=1}) - \mathbb{E}(Y^{a=1,e=1})$$
$$0.02 - 0.04 \neq 0.05 - 0.10,$$

  but no multiplicative interaction because $\frac{0.02}{0.04} = \frac{0.5}{0.10}$.
- Suppose we had 100 versions of drug E after A was randomly assigned. Then, we would expect to cure 3 additional persons if we used all of the drug supply among those with $A = 0$. However, we would expect to cure 6 additional people if we used all the supply among those with $A = 1$.

Interaction and its relation to factorial experiments[17]

- How would you conduct an experiment to evaluate interactions between varialbes?
- We need a *factorial* design.
  - Each treatment (A and E in our example) has different levels ($A, E \in \{0, 1\}$ in our example). A factorial design consists of an equal number of replicates of all possible combinations of the levels of the factors.
  - In our Example from Slide 95, there are $2^2 = 4$ different combination of treatment levels.

[17] David Roxbee Cox and Nancy Reid. *The theory of the design of experiments*. CRC Press, 2000.

Interaction summary

- Just to say that there is an interaction on some scale is uninteresting; all it means is that both exposures have some effect on the outcome.
- Additive interaction is more relevant to public health.

Plan for lecture 4

- Target trials (briefly)
- Structural Equation Models
- Causal graphs
  - Bayesian networks
  - Link to structural equations
  - D-separation
  - Examples

Section 12

Target trial

The target trial

- We have argued that contrast between average counterfactual outcomes under different treatments are often of substantial interest.
- We have also clarified that conducting an experiment guarantees identification of a causal effect. However, conducting an experiment is not always feasible.
- For each causal effect of interest, we can conceptualize a (hypothetical) randomised experiment to quantify it. This hypothetical randomised experiment is called the **target experiment** or **target trial**.
- Being explicit about specifying the target trial forces us to be explicit about the causal question of interest. We ask the question: "What randomised experiment are you trying to emulate?"

Specification of the target trial

To make a causal question practically interesting and useful, it is important to clarify the following, which is part of the specification of the target trial:

- Target population (eligibility criteria).
- Interventions (the treatment strategies).
- Outcome (what is the outcome and when will the outcome be measured)
- Statistical analysis (application of estimators and their statistical properties).

Also clarifies how the claims made can be falsified in the future (in principle), by conducting the target trial. This fits with a positivist (Popperian) view of science.

Motivation

- You have seen that conditional dependencies are hard to interpret.
  - Death penalty example
  - GRE example (you will see this again today)
- We have also seen that (average) causal effects are identified by design in experiments, but also can be identified under assumptions (exchangeability, consistency and positivity) in an observational study. However, reasoning about *counterfactual* (in)dependencies is at least as hard as observed (in)dependencies.
- We will now introduce graphs to clarify when:
  1. Observed independencies can be interpreted causally.
  2. Counterfactual independencies are plausible, which will allow identification of causal effects.
- Importantly, this allows us to study much more complex and realistic settings than those we have considered so far.

Section 13

Structural equations

Structural equation model

Definition

A structural equation model (SEM) is a model that describes how values are assigned to each variable in a system

Think about nature (God) assigning values to each variable in the system. This describes a generative story of how the data came to be as follows. Or think about each equation above represents a physical mechanism that determines the value of the variable on the left (output) from values of the variable on the right (inputs)

We motivate structural equation models (SEMs) with an example

Consider

$$
\begin{aligned}
L &= f_L(U_L) \\
A &= f_A(L, U_A) \\
Y &= f_Y(A, L, U_Y) = Y^{a=A, l=L} = \sum_{a,l} I(A = a, L = l) Y^{a,l}
\end{aligned} \tag{1}
$$

Here U_L, U_A, U_Y are external unmeasured factors that are mutually independent. Here, the generative story is as follows:

- The value of L is determined as a function of the value of U_L as given by the function f_L.
- The value of A is determined as a function of the value of L, U_A as given by the function f_A.
- The value of Y is determined as a function of the value of L, A, U_Y as given by the function f_Y.

We will accompany the structural equations with a picture

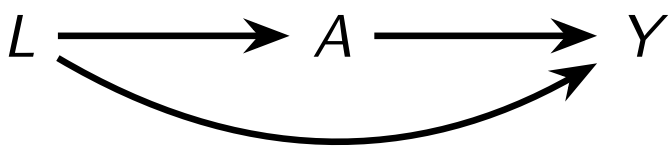
Structural equation models are typically accompanied with a corresponding picture known as a path diagram (as above): that is, a graph which makes explicit the directionality of the underlying process.
For a more compact representation, unmeasured factors that do not determine two or more variables in the system can be left out of the graph (I will repeat this point in later slides, and make the notion more formal).

SEM example (continued)

Consider the SEM $\mathcal{M}$

$$\begin{aligned} L &= f_L(U_L) \\ A &= f_A(L, U_A) \\ Y &= f_Y(A, L, U_Y) \end{aligned} \tag{2}$$

and the graph $\mathcal{G}$,



How does $\mathcal{M}$ induce an observed data distribution over $P(L = l, A = a, Y = y)$ and can this distribution be fully described in some way by simply looking at the graph $\mathcal{G}$?
And how about the distributions under interventions on A, that is, $P(L = l, A = a, Y^a = y)$?

Section 14

Graphs

What is a graph?

Definition (Graph)

A graph $\mathcal{G}$ is a collection of

- Nodes (vertices), $V = \{V_1, V_2, \cdots, V_m\}$.
- Edges $(V_i V_j)$ connecting some of the vertices.

We write $(V_i V_j)$ to denote an edge that connects V_i and V_j.
A **path** is a sequence of edges of the form
$\langle (V_1, V_2), (V_2, V_3), \cdots, (V_{k-1}, V_k) \rangle$,

What is a directed graph?

Definition (Directed Graph)

A directed graph is a graph with a set of nodes and *arrows* connecting some of the nodes. A graph $\mathcal{G}$ is a collection of

- Nodes (vertices) $V = \{V_1, V_2, \cdots, V_k\}$.
- **Directed** edges connecting some of the nodes.

We write $(V_i V_j)_{\rightarrow}$ to denote a directed edge from V_i to V_j.
It is directed, because the graphs A **directed path** is a sequence of edges of the form
$\langle (V_1, V_2)_{\rightarrow}, (V_2, V_3)_{\rightarrow}, \cdots, (V_{k-1}, V_k)_{\rightarrow} \rangle$,
A directed graph has a **cycle** if there exists a path
$\langle (V_1, V_2)_{\rightarrow}, (V_2, V_3)_{\rightarrow}, \cdots, (V_{k-1}, V_k)_{\rightarrow}, (V_k, V_1)_{\rightarrow} \rangle$.
A Directed Acyclic Graph is a directed graph with no cycles.
PS: Now the subscript does not longer indicate an individual. V_1 is now a random variable. From now on, I will use $V_1(\omega)$ when I talk about the value for a particular individual.

Basic definitions

In a DAG $\mathcal{G}$ we define the following sets (parents, children, ancestors and descendants):

- $\mathbf{pa}_G(V_i) \equiv \{V_t : V_t \to V_i \text{ exists in } \mathcal{G}\}$.
- $\mathbf{ch}_G(V_i) \equiv \{V_t : V_i \to V_t \text{ exists in } \mathcal{G}\}$.
- $\mathbf{an}_G(V_i) \equiv \{V_t : V_t \to V_a \to \cdots \to V_j \to V_i \text{ exists in } \mathcal{G}\} \cup V_i$.
- $\mathbf{de}_G(V_i) \equiv \{V_t : V_i \to V_a \to \cdots \to V_j \to V_t \text{ exists in } \mathcal{G}\}$.

Further terminology:

- A path where $V_a \to V_i \leftarrow V_b$ is called a collider path, and here V_i is a collider.
- A path where $V_a \leftarrow V_i \to V_b$ is called a fork.
- A path is *blocked* if it contains a collider. Otherwise it is *open*.
- A DAG is complete if there is an arrow between every pair of nodes.

Topological order with respect to a graph

Definition (Topological order of a DAG)

The nodes $V_1, V_2, \ldots$ follow a topological order relative to a DAG $\mathcal{G}$, if V_j is not ancestor of V_i whenever $j > i$.

Note that topological orders are not necessarily unique, but in the DAG in Figure 126 the only possible topological order is $\langle L, A, Y \rangle$.

Some preliminaries

- Consider a study population Ω.
- Let ω be an element (i.e. unit or individual) in Ω.
- Note that we used subscript i to denote an individual in the first lecture, but now the subscript just indicates a particular random variable, and we write $V_i(\omega)$ when we consider the value for individual ω.
- Consider a discrete random variable V_j.
- Let $V_j(\omega)$ be the value of V_j in ω.
- Let $\mathcal{G}$ be a DAG with nodes $V = \{V_1, V_2, \cdots, V_m\}$.
- We use overlines to denote histories of variables, e.g. $\overline{v}_j \equiv (v_1, v_2, \ldots, v_j) \in \mathcal{V}_1 \times \mathcal{V}_2 \times \cdots \times \mathcal{V}_m$.
- Let $PA_k = \{V_j : V_j \in \mathbf{pa}_G(V_k)\}$. A random variable
- Let $pa_k = \{v_j : V_j \in \mathbf{pa}_G(V_k)\}$ for a $\overline{v} \equiv (v_1, v_2, \ldots, v_m) \in \mathcal{V}_1 \times \mathcal{V}_2 \times \cdots \times \mathcal{V}_m$ A realisation of PA_k.
- From now on I will use $p(v_i \mid v_j)$ to denote conditional densities $P(V_i = v_i \mid V_j = v_j)$.

Non-parametric structural equation model (NSPEM) with respect to a graph

There exist unknown functions $f_1, \ldots, f_m$ such that the observed ("factual") variables $V_1, \ldots, V_m$ satisfy

$$
\begin{aligned}
V_1 &= f_1(U_1) \\
V_2 &= f_2(PA_2, U_2) \\
V_3 &= f_3(PA_3, U_3) \\
&\vdots \\
V_m &= f_m(PA_m, U_m)
\end{aligned} \tag{3}
$$

where:

- $f_0, f_1, \ldots$ are unknown deterministic functions.
- PA_i is the set of random variables that are parents of V_i.
- $U_0, U_1, \ldots$ are random variables ("*disturbances*" *or*" *errorterms*") (not drawn in the graph). Sometimes called exogenous variables.

NPSEM continues

For any treatment regime $g = (g_{j_1}, \ldots, g_{j_t})$, the counterfactual variables under g are generated by replacing the functions $(f_{j_1}, \ldots, f_{j_t})$ with the functions $(g_{j_1}, \ldots, g_{j_t})$, where $t \leq m$. Below is an illustration. This is called performing recursive substitution.

$$
\begin{aligned}
V_1^g &= f_1(U_1) \\
V_2^g &= f_2(PA_2^g, U_2) \\
&\vdots \\
V_{j_1}^{g+} &= g_{j_1}(PA_{j_1}^g, U_{j_1}) \\
&\vdots \\
V_m^g &= f_m(PA_m^g, U_m) \qquad (4)
\end{aligned}
$$

The superscript "g" indicates that V_i^g is a counterfactual variable (in other words, potential outcome variable). The superscript "$g+$" is given to the variables on which we intervene. A NPSEM requires (3) and (4) to hold.

Some remarks

- Structural: f_k not only generates observed (factual variables), but also variables in other counterfactual worlds where we have done interventions.
- Counterfactual: The variable $V_j^g, j \in \{0, \ldots, m\}$ are called counterfactual variables under treatment regime g.
- A cause: A variable A is a cause of a variable Y if a change in A can lead to a change in Y.

Example: Point intervention

Let the regime g be defined by the intervention that sets V_2 to a.

$$\begin{aligned} V_1^a &= f_1(U_1) \\ V_1^{a+} &= a \\ V_3^a &= f_3(PA_3^a, U_3) \\ &\vdots \\ V_m^a &= f_m(PA_m^a, U_m) \end{aligned} \tag{5}$$

The superscript "a" indicates that V_i^a is a counterfactual variable (or potential outcome variable)
where we have intervened to set a variable, here V_1 (now with a superscript $a+$) to a.

Let's interpret this model, specifically

- Only the arguments to the structural equation determine the value of a node.
  That is, the value of $V_j(\omega)$ does not depend on any other unit ω' in the population.
  (**No interference**)
- Suppose that a unit ω has $PA_k(\omega) = pa_k$. Then, under any intervention g that fixes $PA_k^g = pa_k$ we have that $V_k(\omega) = V_k^g(\omega)$.
  (**Consistency**)
- When PA_k is known, the value of other variables $\overline{V} \setminus PA_k$ do not determine V_k. (**Exclusion restriction**).

The causal inference part is an assumption about the errors!

We must say something about the dependencies between the U's to encode causal relations.

Definition (Independent error model)

A NPSEM wrt. a DAG $\mathcal{G}$ such that $U_0, \ldots, U_M$ are mutually independent.

`This is Pearl's NPSEM-IE`[18]`.`
`"IE" stands for independent errors.`
`NB: The independent error assumption is not really needed, and can be relaxed in the more general FFRCISTG model`[19]
The U_ks represent all other variables that are used by nature, the decision maker or anyone else to determine the value of V_k.

[18] Judea Pearl. *Causality: Models, Reasoning and Inference 2nd Edition*. Cambridge University Press, 2000.

[19] Robins, "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect".

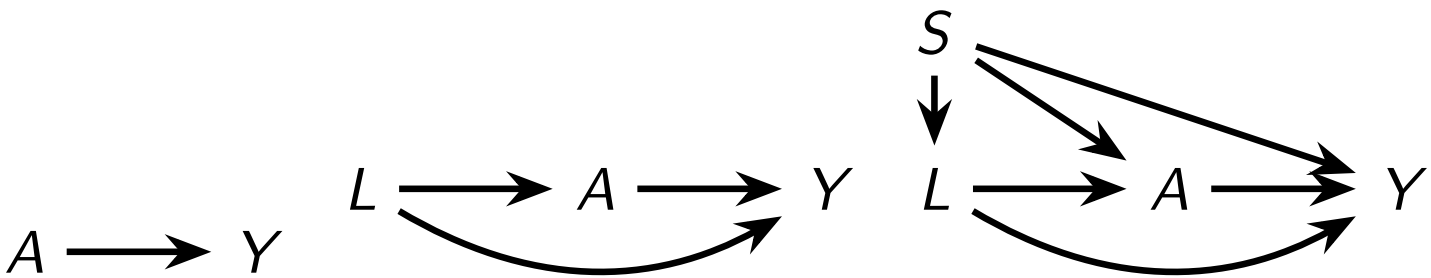
Section 15

Causal graphs

Let us start with some intuition

Suppose I were to explain what is going on in the experiment on heart transplant for my friend who studied literature. I will draw intuitive diagrams that can be formalised as causal graphs. We have previously discussed:

- Completely randomised experiment.
- Conditional randomised experiment.
- Observational study with smoking.



This way of building causal stories using diagrams can be formalised by graphs.

Next step

- In the previous slide, we just made these diagrams to encode qualitative subject matter knowledge.
- However, we shall see that the diagram can be formalised as a causal directed acyclic graph, DAG, which encodes information about causal and non-causal associations in a causal network: it allows us to represent both association and causation in the same graph.

Graphs

Some things you need to know about graphs

- Graphs encode conditional independendcies
- Graphs allow us to *represent* and *organize* assumptions and prior knowledge.
- Graphs make the assumptions transparent and explicit.

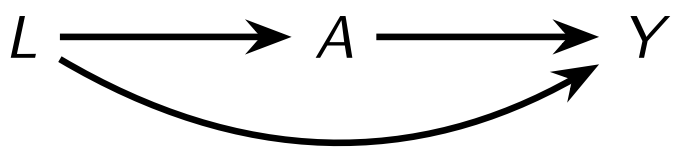
What is the role of causal graphs?

- Graphs help us to reason about independencies; that is, they help us reason about whether certain exchangeability assumptions (conditional independencies) hold.
- This agrees with the mantra: "draw your assumptions before your conclusions".[20]
- Graphs help us to conceptualize problems and have intuitive appeal, also for researchers who are illiterate in math.
- However, the intuitive graphical representations have a **mathematical justification**. Therefore you can translate the intuitive subject-matter expertise (from doctors, economists, social scientists) to precise mathematical statements.
- Graphs allow us to encode *causation* and *association*.

[20] Hernan and Robins, *Causal inference: What if?*

Example

We can now define the graph below as a causal DAG that describes the conditional randomised trial on heart transplants,



where $V_1 = L, V_2 = A, V_3 = Y$.
Here $\mathbf{pa}_G(Y) = (L, A)$.
The graph is complete because there is an arrow between every pair of nodes.

What is a model

Definition (Statistical model)

A statistical model $\mathcal{P}$ is a collection of laws, $\mathcal{P} = \{P_\eta : \eta \in \Gamma\}$.

Here Γ could be an infinite dimensional space. We will typically only restrict ourselves to the space of models with finite variance.

Bayesian network

Definition (Bayesian network)

A Bayesian Network with respect to a DAG $\mathcal{G}$ with nodes $V = (V_1, \ldots, V_m)$ is a statistical model for the random vector V specifying that V belongs to the collection of laws $\mathcal{B}$ satisfying the Markovian factorisation

$$p(v) = \prod_{j=1}^{m} p(v_j \mid pa_j)$$

Here, $p(x \mid y) \equiv P(X = x \mid Y = y)$.
We say that the DAG $\mathcal{G}$ represents the Bayesian Network $\mathcal{B}$.
For any law p in $\mathcal{B}$, we say that p factors according to $\mathcal{G}$, or that p is represented by $\mathcal{B}$.

Causal DAG

Definition (Robins EPI 207)

A causal model associated with a DAG satisfies:

1. The lack of an arrow from node V_i to V_j can be interpreted as the absence of a direct causal effect of V_i on V_j (relative to the other variables on the graph).
2. Any variable is a cause of all its descendants. Equivalently, any variable is caused by all its ancestors.
3. All common causes, even if unmeasured, of any pair of variables on the graph, are themselves on the graph.
4. The Causal Markov Assumption (CMA): The causal DAG is a statistical DAG, i.e., the distribution of V factors.
5. Because of the causal meaning of parents and descendants on a causal DAG, the Causal Markov Assumption is equivalent to the statement:
    - Conditional on its direct causes (i.e., parents), a variable V_i is independent of any variable it does not cause (i.e., any nondescendant).

Absence of common causes in the DAG (point 3)

The arguments here are analogous to the motivating example for the simple graph with A, L, Y and smoking S.

- Remember that U_k represents all other variables that determines (causes) V_k except the parents PA_k.
- Suppose that there exists a variable C that is a direct determinant of V_k relative to the DAG (i.e. it does not only determine V_k through variables in the DAG).
- This means that $U_k = m_k(C, U_k^*)$ for some function m_k.
- Suppose that C is also a direct determinant of a node j (but C is still not in the DAG).
- Thus, $U_j = m_j(C, U_j^*)$ for some function m_j.
- Thus, $U_k \not\perp\!\!\!\perp U_j$.

Factorisation of the NPSEM-IE (point 4)

Argument for Markov factorisation of causal model wrt. a DAG

$$p(v) = \prod_{j=1}^{m} p(v_j \mid pa_j).$$

Proof.

Consider $p(v_j \mid \overline{v}_{j-1})$ for any $j \in \{0, \ldots, m\}$. Here pa_j are the parents of v_j.

$$\begin{aligned}
&p(v_j \mid \overline{v}_{j-1}) \\
&= p(f_{v_j}(PA_j, U_{v_j}) = v_j \mid \overline{V}_{j-1} = \overline{v}_{j-1}) \\
&= p(f_{v_j}(pa_j, U_{v_j}) = v_j \mid \overline{V}_{j-1} = \overline{v}_{j-1}) \\
&= p(f_{v_j}(pa_j, U_{v_j}) = v_j \mid f_{v_{j-1}}(pa_{j-1}, U_{v_{j-1}}) = v_{j-1}, \ldots, f_{v_1}(pa_1, U_{v_1}) = v_1) \\
&= p(f_{v_j}(PA_j, U_{v_j}) = v_j \mid PA_j = pa_j).
\end{aligned}$$

□

No restrictions on $p(v)$ imposed by the NPSEM-IE

The only restriction imposed on the *observed* law is the factorisation

$$p(v) = \prod_{j=1}^{m} p(v_j \mid pa_j).$$

Proof.

Any further restriction must be a restriction on the form of $p(v_j \mid pa_j)$ for any $j \in \{0, \ldots, m\}$. But

$$P(V_j = v_j \mid PA_j = pa_j) = P(f_{v_j}(pa_j, U_{v_j}) = v_j),$$

and we have not put any restrictions on the marginal density of U_{v_j}. □

Factorisation of the nodes V

Lemma

If V follows a NPSEM-IE, then for any $p(\overline{v}_{j-1})$ with $p(\overline{v}_{j-1}) > 0$ we have that $p(v_j \mid \overline{v}_{j-1}) = p(v_j \mid pa_j)$ and therefore the joint density factorizes as

$$p(v) = \prod_{j=1}^{m} p(v_j \mid pa_j).$$

This factorisation is the only restriction that the causal model implies on the law of the observed data.

Thus, in our example from slide 137, the observed law factorizes as

$$p(v) = p(l, a', y) = p(l)p(a' \mid l)p(y \mid a', l),$$

which means that here we put absolutely no restrictions on the law $p(v) \equiv P(V = v)$. You do not have to prove this.

So we have an algorithm for creating causal graphs

We can create a causal DAG by:

1. Draw nodes for the exposure A and the outcome Y of interest.
   - Draw an arrow from A to Y.
2. If there exists a common cause C of A and Y, write C in the graph.
   - Draw arrows from C to A and from C to Y.
     These common causes must be drawn, even if they are unmeasured.
3. If there exists a common cause C' of any pair $W, W' \in (C, A, Y)$, write C' in the graph.
   - Draw arrows from C' to W and from C' to W'.
4. Continue in this way until there are no common causes...

Markov equivalence classes

Definition (Markov equivalence class)

A Markov equivalence class is a set of DAGs that encode the same set of conditional independencies.

Example of markov equivalent DAGs:

Implication: We cannot use data alone to distinguish between causal graphs.

Linear structural equation example

We have not imposed any parametric assumptions so far. However, just for the illustration, suppose we have a (partially) linear structural equation model with two variables satisfying

$$\begin{aligned} A &= f(U_A) \\ Y &= \alpha + \beta A + U_Y \end{aligned} \tag{6}$$

This structural equation model implies that the individual level causal effects is $Y^{a=1} - Y^{a=0} = \beta$!
We conclude that the linear equation model relies on extremely strong assumptions that usually will be implausible. In this course, we will not rely on such assumptions.

Modified non-parametric example

A different SEM $\mathcal{M}$

$$\begin{aligned} L &= f_L(U_L) \\ A &= f_A(L, U_A) \\ Y &= f_Y(A, U_Y) \end{aligned} \tag{7}$$

and the graph $\mathcal{G}$,

- Encodes that, changes in L leaves Y unchanged, provided that U_Y and A remain constant.
- Does this graph encode any restrictions on the distribution of (L, A, Y)?
  `We will formally study what kind of restrictions the structural models involve`

Section 16

Lecture 5

Plan for today

- D separation
- Examples
- The backdoor criterion

Properties of conditional independence

Theorem (Graphoid axioms)

Let X, Y, Z, W be random variables on a Cartesian product space. Conditional independence satsifies

1. $X \perp\!\!\!\perp Y \mid Z \implies Y \perp\!\!\!\perp X \mid Z$ *(Symmetry)*
2. $X \perp\!\!\!\perp Y, W \mid Z \implies X \perp\!\!\!\perp Y \mid Z$ *(Decomposition)*
3. $X \perp\!\!\!\perp Y, W \mid Z \implies X \perp\!\!\!\perp W \mid Y, Z$ *(Weak union)*
4. $X \perp\!\!\!\perp W \mid Y, Z$ *and* $X \perp\!\!\!\perp Y \mid Z \implies X \perp\!\!\!\perp Y, W \mid Z$ *(Contraction)*
5. *If* $p(x, y, z, w) > 0$, *then* $X \perp\!\!\!\perp W \mid Y, Z$ *and* $X \perp\!\!\!\perp Y \mid W, Z \implies X \perp\!\!\!\perp Y, W \mid Z$ *(Intersection)*

Proof of Graphoid axioms

I will not prove all of them here. I just state a brief proof of the first one here.

Proof.

1. Symmetry follows simply because

$$X \perp\!\!\!\perp Y \mid Z \leftrightarrow p(x \mid z)p(y \mid z) = p(x, y \mid z)$$
$$= p(y \mid z)p(x \mid z) \leftrightarrow Y \perp\!\!\!\perp X \mid Z \ .$$

D separation of a path

Now we will study a beautiful graphical condition on $\mathcal{G}$ that immediately tells if $X \perp\!\!\!\perp Y \mid Z$, where X, Y, Z are disjoint sets of nodes in V, is implied by the Markov factorisation.

Definition (d-separation of a path)

A path r is d-separated by a set of nodes Z iff

1. r contains a chain $V_i \to V_j \to V_k$ or a fork $V_i \leftarrow V_j \to V_k$ such that V_j is in Z, or
2. r contains a collider $V_i \to V_j \leftarrow V_k$ such that V_j *is not* in Z and such that no descendant of V_j is in Z.

Otherwise the path is d-connected.

D separation of two nodes

Definition (d-separation of two nodes)

Nodes V_i and V_k are d-separated by a set of nodes Z if all trails between V_i and V_k are d-separated by Z. We write d-separation as

$$(V_i \perp\!\!\!\perp V_k \mid Z)_G \,.$$

If V_i and V_k are not d-separated, they are d-connected and we write

$$(V_i \not\perp\!\!\!\perp V_k \mid Z)_G \,.$$

Theorem (Soundness of d-separation)

$(V_i \perp\!\!\!\perp V_k \mid Z)_G$ *implies the statistical independence*

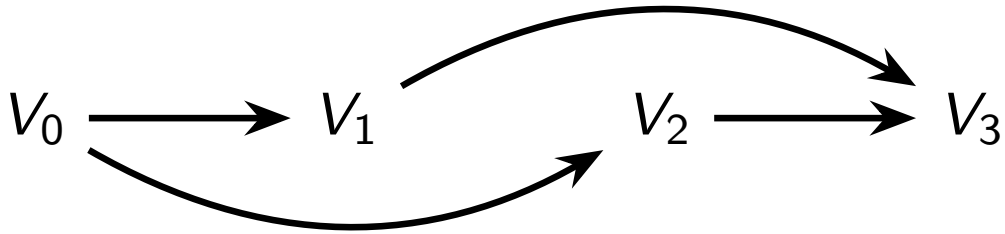
$$V_i \perp\!\!\!\perp V_k \mid Z.$$

A consequence of soundness is that d-separation in $\mathcal{G}$ implies conditional independence for any distribution that factorizes according to $\mathcal{G}$.

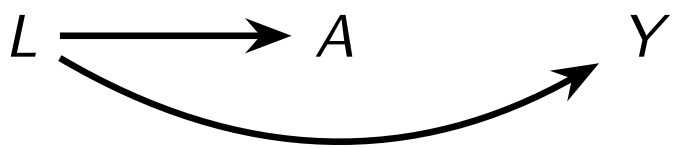
D-separation details and intuition

- D-separation can be shown solely using the Graphoid axioms (but the proof is tedious).
- d-separation allows us to determine independencies of a distribution from the structure of a statistical DAG that represents it.
- Heuristically, two variables are d-separated (independent) if there is no open path between them.

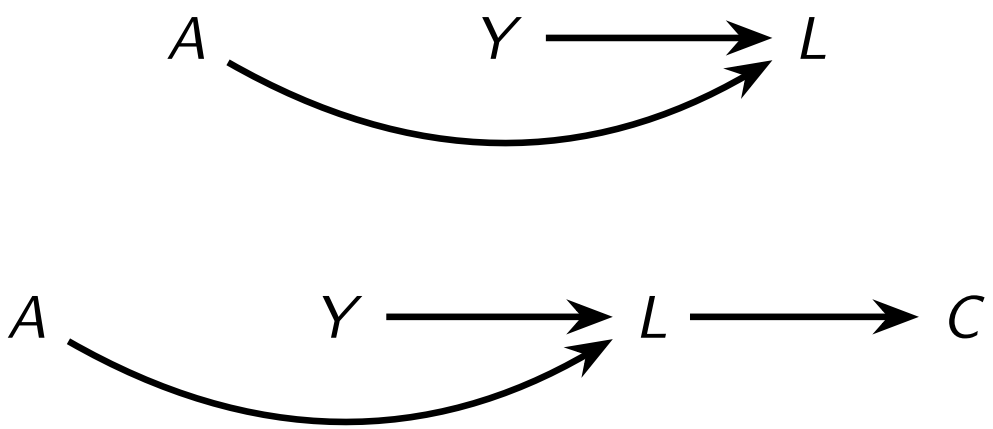
D-separation and some questions in class



Carrying a lighter A and the risk of lung cancer Y



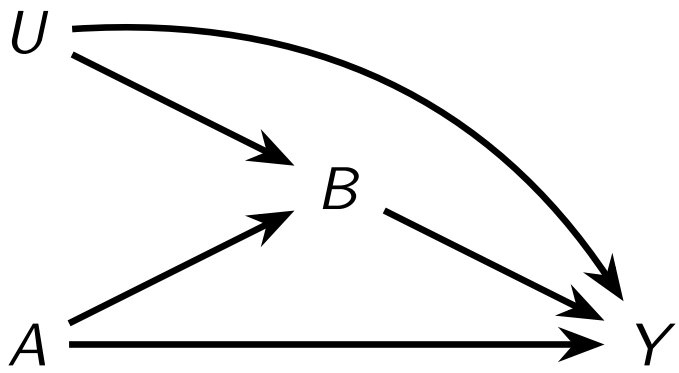
A gene A that causes heart disease L but not smoking Y, where C is taking aspirin (A cardiovascular drug)



Example: Birth weight paradox

- Birth weight predicts infant mortality.
- Investigators often stratify on birth weight when evaluating the effect of maternal smoking on infant mortality.
- Among infants with low birth weight, the mortality rate ratio for smoke exposed infants versus non-exposed infants is 0.79 (95% CI: 0.76, 0.82).
- This birth weight paradox has been a controversy for decades.
- One suggestion is that the effect of maternal smoking is modified by birth weight in such a way that smoking is beneficial for LBW babies.
- Is this indeed the likely explanation?

Example: Birth weight paradox



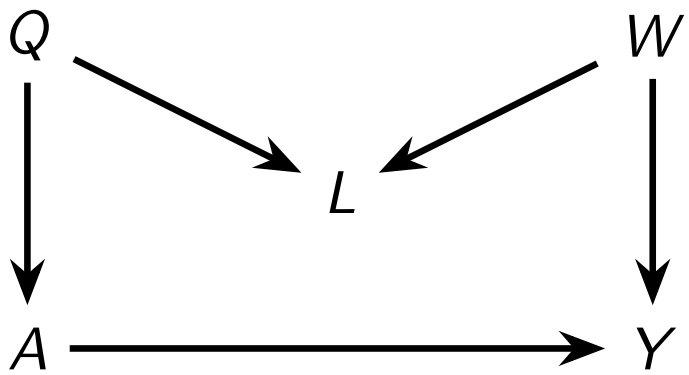
- A Smoking status of the mother
- B Birth weight
- U Unknown factor (e.g. genetic) causing low birth weight
- Y Infant mortality

PS: for this graph to be more plausible, we should also add common causes of A and Y.

A clinical story

- Suppose the graph on Slide 158 represents a study of HIV-positive individuals to estimate the effect of an antiretroviral treatment A on 3-year risk of death Y.
- The unmeasured variable $U \in \{0, 1\}$ indicates high level of immunosuppression. Those with $U = 1$ have a greater risk of death.
- Individuals who drop out from the study or are otherwise lost to follow-up are censored ($C = 1$).
- Individuals with $U = 1$ are more likely to be censored because the severity of their disease prevents them from participating in the study.
- The effect of U on censoring C is mediated by the presence of symptoms (fever, weight loss, diarrhea, and so on), CD4 count, and viral load in plasma, all included in L, which we suppose are measured.
- Individuals receiving treatment are at a greater risk of experiencing side effects, which could lead them to dropout, as represented by the arrow from A to C. We have to restrict the analysis to individuals who remained uncensored ($C = 0$) because those are the only ones in which Y can be assessed.

Berkson bias



- A Drink a glass of red wine a day.
- Y Nausea
- L Aspirin
- Q Family history of cardiovascular disease
- W Frequency of headache

Q: We measure Aspirin. Should we adjust for Aspirin in the analysis?

Faithfulness and completeness of d-separation

Definition

A law $\mathbb{P}$ is faithful to a DAG $\mathcal{G}$ if for any disjoint set of nodes A, B, C we have that $A \perp\!\!\!\perp C \mid B$ under $\mathbb{P}$ implies $(A \perp\!\!\!\perp C \mid B)_{\mathcal{G}}$.

Theorem (Completeness of d-separation)

In a Bayesian Network with respect to a direct acyclic graph $\mathcal{G}$ there exists a faithful law $\mathbb{P}$.

We will not prove this important result[21].
The completeness of faithfulness d-separation allows us to use d-separation to represent the conditional independence structure of a multivariate distribution.
You can look at the graph, and read off all independencies that hold in the entire class of distributions factorizing according to the DAG.

[21] Ann Becker, Dan Geiger, and Christopher Meek. "Perfect tree-like markovian distributions". In: *arXiv preprint arXiv:1301.3834* (2013); Pearl, *Causality: Models, Reasoning and Inference 2nd Edition*.

The causal Markov assumption and faithfulness (intuition and interpretation)

- d-separation implies statistical independence, but does not allow one to deduce that d-connection implies statistical dependence.
- However, d-connected variables will be independent only if there is an exact balancing of positive and negative causal effects.
- Because such precise balancing of effects is highly unlikely to occur, we shall henceforth generally assume that d-connected variables are dependent.

Backdoor adjustment

Definition (Backdoor path)

In a DAG $\mathcal{G}$ a backdoor path between two nodes V_i and V_j is a trail that starts in V_i and ends in V_j; and with initial edge being an arrow pointing into V_i

Example backdoor path between V_i and V_j is: $V_i \leftarrow V_k \rightarrow V_j$.

Backdoor theorem

Theorem (Backdoor theorem wrt. to a DAG)

In a causal model represented by a DAG $\mathcal{G}$ let X, Y and Z be three sets of nodes of $\mathcal{G}$, each comprised of one or more nodes. Suppose that X contains no descendants of Z and it blocks all back-door paths between any node in Z and any node in Y: Suppose that $g = (g_1, \ldots, g_t)$ is a regime for $Z = (Z_1, \ldots, Z_t)$ (for some $t \geq 1$) such that treatment assignments depend at most on X: Then, for any x in the support of X such that $p(Z = g(x) \mid x) \equiv Pr(Z = g(x) \mid X = x) > 0$; it holds that

$$\begin{aligned} P(Y^g = y, Z^{g+} = z, X = x) =& p^g(y, x, z) \\ =& p(y \mid x, z) I(g(x) = z) p(x) \\ =& P(Y = y \mid Z = z, X = x) I(g(x) = z) P(X = x), \end{aligned}$$

and in particular,

$$P(Y^g = y) = \sum_x \sum_z p(y \mid x, z) I(g(x) = z) p(x).$$

The backdoor theorem continues

See Pearl[22] for proof (not required for the exam etc). This theorem is very useful, because it allows us to identify causal effects even if certain nodes in the graph are unmeasured. The last part of the theorem, after "in particular", will be useful in the exercises of Lecture 5.

[22] Judea Pearl. "Causal diagrams for empirical research". In: *Biometrika* 82.4 (1995), pp. 669–688.

Implication from the Backdoor theorem

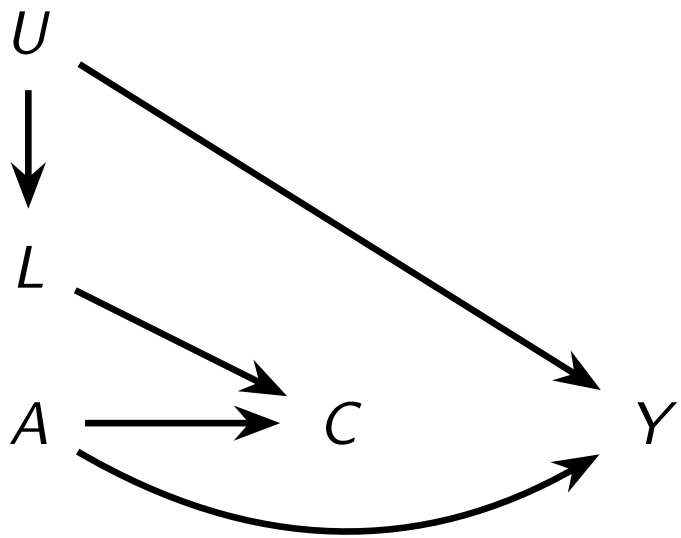
It follows from the backdoor theorem that if $Y^a \perp\!\!\!\perp A \mid L$ then

$$P(Y^a = y) = \sum_l P(Y = y \mid L = l, A = a)P(L = l).$$

However, we can also use it to identify causal effects in much more complicated settings, which also involve unmeasured variables.

Loss to follow-up example 1

A graph corresponding to the story from Slide 150



Factorisation according to the DAG with ordering $\langle A, U, L, C, Y \rangle$:

$$p(y, c, l, u, a) = p(y \mid u, a)p(c \mid l, a)p(l \mid u)p(u)p(a)$$

But how do we use this factorization to identify causal effects?

Backdoor theorem in the example on loss to follow-up

Consider the example from Slide 158.

- Note that
  - L blocks all backdoor paths between (A, C) and Y.
  - Thus,

$$\mathbb{E}(Y^{a,c=0}) = \sum_l \mathbb{E}(Y \mid A = a, C = 0, L = l) P(L = l),$$

which can be estimated simply by standardisation:

- Estimate $\mathbb{E}(Y \mid A = a, C = 0, L = l)$ by $\hat{\mathbb{E}}(Y \mid A = a, C = 0, L = l)$,
- Estimate $P(L = l)$ empirically.
- Standardise

PS: Many causal questions are more difficult

Realistic questions are often more difficult. Consider for example:

- when should we start a treatment?
- How long should we continue treatment?
- When to switch to different treatment?
- What event should guide us to switch treatment?

PS: Many causal questions are more difficult

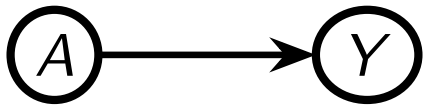
Realistic questions are often more difficult. Consider for example:

- when should we start a treatment?
- How long should we continue treatment?
- When to switch to different treatment?
- What event should guide us to switch treatment?

We will discuss such questions later in the course

Elephant in the room...

In a randomised study, the following graph is a causal DAG:



And we know that $Y^a \perp\!\!\!\perp A$ for $a \in \{0, 1\}$.
But the counterfactual independence cannot be read off from the graph!
This raises some questions:

- Can we construct graphs to read off such counterfactual independencies?
- Can we read off factorisations of *counterfactual* laws from graphs?

D-separation allows us to read off whether an association is causal

- We can graphically check – using d-separation – whether an observed association between two variables A and B conditional on C is (solely) due to a causal effect (i.e. that the association is unconfounded).
- However, we also want to use graph to evaluate if we can identify functionals of *counterfactual* variables, for example $\mathbb{E}(Y^a)$. Now, the elephant in the room is that there are no counterfactual variables on the DAG! And we did want to reason about counterfactual independencies. Thus, whereas we can evaluate independencies between *factual* variables in a DAG, we cannot study *counterfactual* independencies.
- Here we will study a recent and elegant[23] transformation of the DAG – the so-called Single World Intervention Graph (SWIG) – that does allow us to read off independencies between factual and counterfactual variables.

[23] Thomas S Richardson and James M Robins. "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality". In: *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper* 128.30 (2013).

Lecture 6

- Static SWIGs
- Dynamic SWIGs

Section 17

Single World Intervention Graphs (SWIGs)

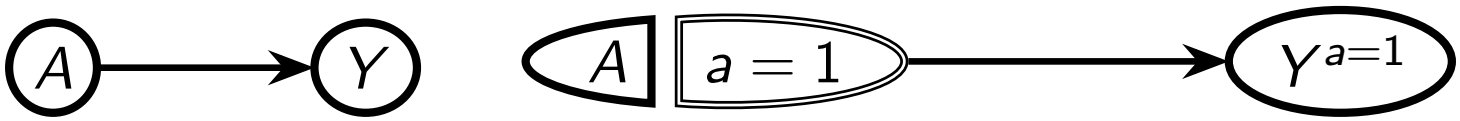
Creation of SWIGs

The SWIG $\mathcal{G}(a)$ is created as follows (it can be conceived as a function that transforms the original DAG into a new graph, which is still a DAG):

1. Draw the DAG that represents the causal model.
2. Split treatment variables into two nodes (indicated by semi-circles), left and right.
   - The left node encodes the random variable treatment that would have been observed in the absence of an intervention. This is called the *natural value of treatment* node. Natural value of treatment nodes should be treated as nodes of an ordinary DAG, i.e., ordinary random variables.
   - The right node encodes the value of treatment under the intervention. These nodes should be treated as constants, i.e. fixed nodes.
3. Re-label every non-manipulated descendant of an intervention node with superscript: the superscripts indicate the counterfactual.
   - Use consistency to obtain graphs with minimal labelling, i.e. the minimal set of counterfactuals in the superscript.

Example: SWIG in a simple randomised trial

SWIG under treatment $a = 1$:



We can read the independence $Y^{a=1} \perp\!\!\!\perp A$.
We also associate the new **factorisation**:

$$P(A = a', Y^{a=1} = y) = P(A = a')P(Y^{a=1} = y),$$

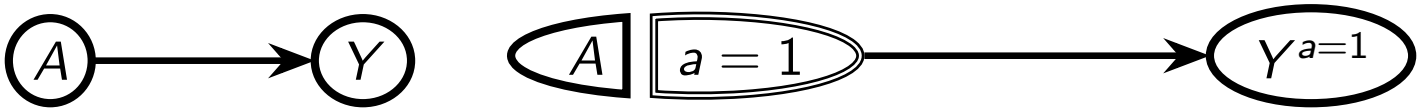
where we omit the fixed nodes from the conditioning set. Furthermore, we make a **modularity** assumption (which would be implied by the independent error assumption)

$$P(Y^{a=1} = y) = P(Y = y \mid A = 0),$$

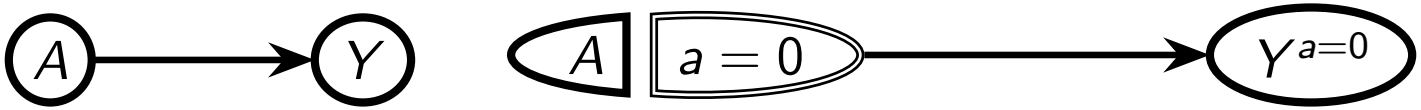
which links the original factorisation to the original DAG factorisation.

Single world

We can read the independence $Y^{a=1} \perp\!\!\!\perp A$ from the SWIG for treatment $a = 1$:



We can read the independence $Y^{a=0} \perp\!\!\!\perp A$ from the SWIG for treatment $a = 0$:



Why do we need both graphs? These are two different graphs that represent the factorisation of different margins: $P(A = a', Y^{a=1} = y)$ and $P(A = a', Y^{a=0} = y)$. None of the SWIGs encodes assumptions between counterfactuals from different worlds $Y^{a=0}$ and $Y^{a=1}$. This is a feature, not a bug.
It has to do with identification. Node splitting preserves identification. If I observe every node that I included in the original DAG, then the counterfactual laws defined by the node splittings are also going to be identified: If $P(A = a', Y = y)$ is identified, then $P(A = a', Y^{a=1} = y)$ is identified and so is $P(A = a', Y^{a=0} = y)$, but not $P(A = a', Y^{a=1} = y', Y^{a=0} = y)$.

Factorisation

Definition (SWIG factorisation)

The factorisation associated with a SWIG is

$$P(V^{\overline{a}} = v) = \prod_{V_i \in V} P(V_i^{\overline{a}_i} = v_i \mid (PA_{\mathcal{G}(\overline{a}),i} \setminus \overline{a}) = q)$$

where $q \subseteq pa_i \subset v$ and $\overline{a}_i \subseteq \overline{a}$ ($\overline{a}_i$ are the elements of $\overline{a}$ that are ancestors of V_i).

Modularity

Definition (Modularity)

The DAG pair $(\mathcal{G}, p(v))$ and the SWIG pair $(\mathcal{G}(\overline{a}), p^{\overline{a}}(v))$ under an intervention that sets $\overline{A} = (A_0, \ldots, A_k)$ to $\overline{a} = (a_0, \ldots, a_k)$ satisfy modularity for every $V_i \in V$ if

$$
\begin{aligned}
&P(V_i^{\overline{a}_i} = v_i \mid (PA_{\mathcal{G}(\overline{a}),i} \setminus \overline{a}) = q) \\
=&P(V_i = v_i \mid (PA_{\mathcal{G},i} \setminus \overline{A}) = q, (PA_{\mathcal{G},i} \cap \overline{A}) = \overline{a}_{PA_{\mathcal{G},i} \cap \overline{A}})
\end{aligned}
$$

This definition looks like a mouthful, but it is conceptually quite easy to understand. It bridges counterfactual densities to observable densities.
It is implied by the independent error assumption of the NPSEM-IE, and it holds under a weaker causal model, the FFRCISTG[24] (I have not shown this).

[24] Richardson and Robins, "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality".

Causal models, factorisation and modularity

Theorem

A NPSEM-IE model (and the FFRCISTG model that includes the NPSEM-IE model as a strict submodel) obeys factorisation and modularity.

We will not prove this result, but we will use it extensively.
In our saturated graph when we intervene to set $a = 1$, it implies that $P(Y^{a=1} = y) = P(Y = y \mid A = 1)$.

D separation of a path (minimal modification in SWIGs)

A slight twist of D-separation for SWIGs

Definition (d-separation of a path)

A path r is d-separated by a set of nodes Z iff

1. r contains a chain $V_i \to V_j \to V_k$ or a fork $V_i \leftarrow V_j \to V_k$ such that V_j is in Z, or
2. r contains a collider $V_i \to V_j \leftarrow V_k$ such that V_j *is not* in Z and such that no descendant of V_j is in Z.

If a path is not d-separated by Z and there is no fixed node on the path, then the path is d-connected given Z.

SWIT in a simple randomised trial

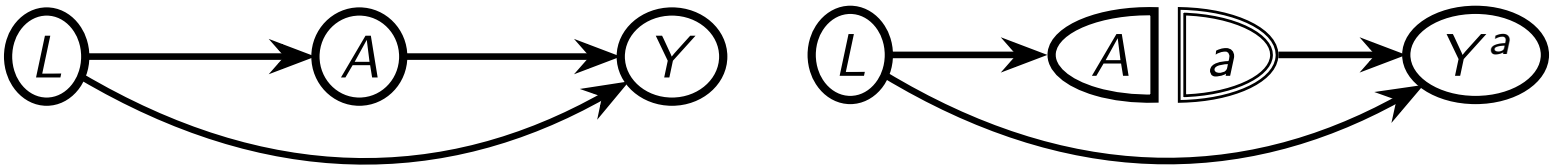
A SWIT is a SWIG template[25], i.e. a graph valued function:

- It takes a specific value a as input.
- Returns a SWIG $G(a)$.
- SWIG $G(0)$ represents $p(A = a', Y^{a=0} = y)$.
- SWIG $G(1)$ represents $p(A = a', Y^{a=1} = y)$.

The SWIT represents both the SWIGs from the previous slide. Hereafter we will use SWITs for simplicity, most of the time.

[25] Note that I am sometimes sloppy and use the word SWIG when I formally talk about a SWIT.

SWIG in a conditional randomised experiment

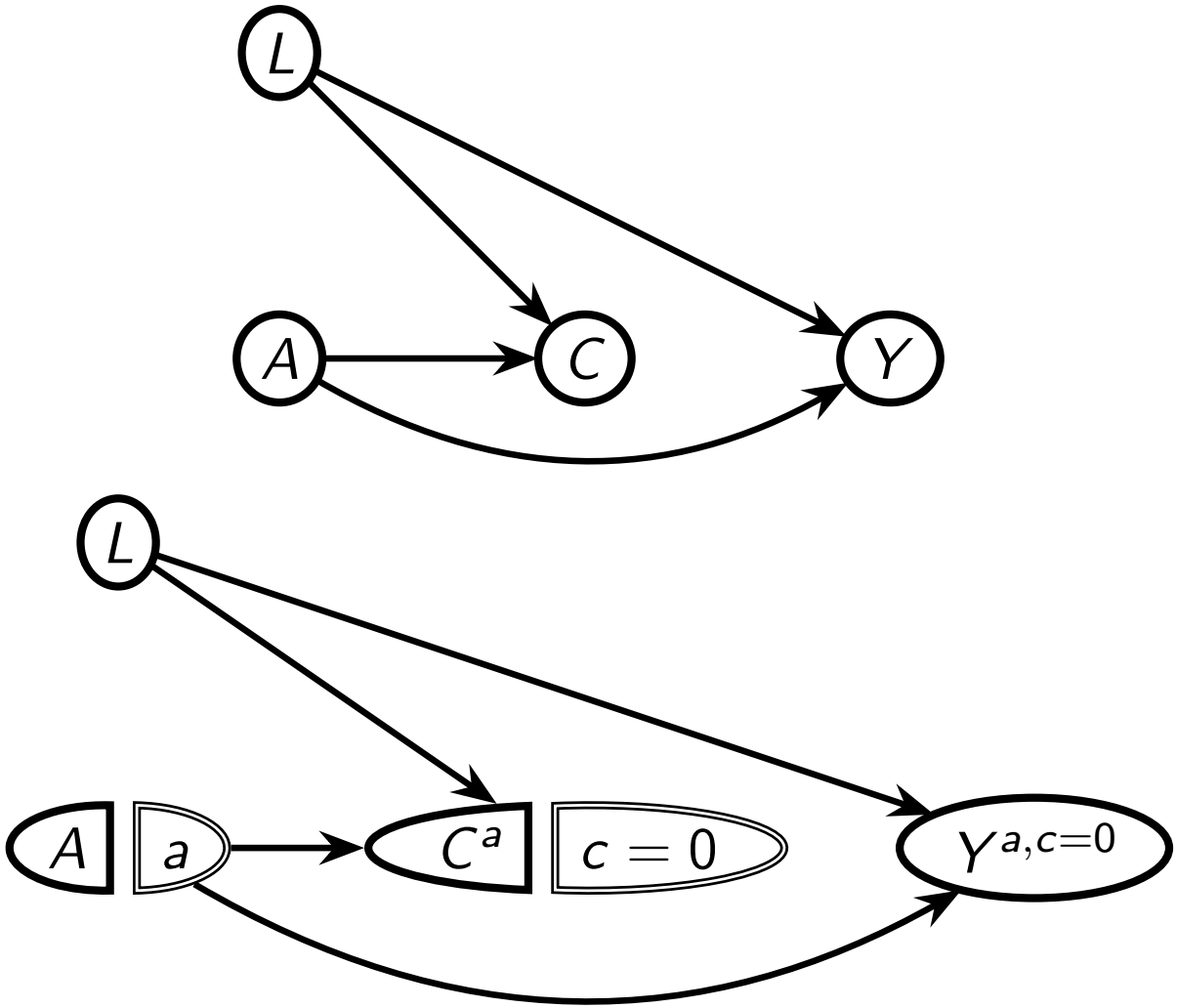


$$P(Y^a = y) = \sum_l P(Y^a = y \mid L = l)P(L = l) \text{ factorization}$$

$$= \sum_l P(Y = y \mid A = a, L = l)P(L = l). \text{ modularity}$$

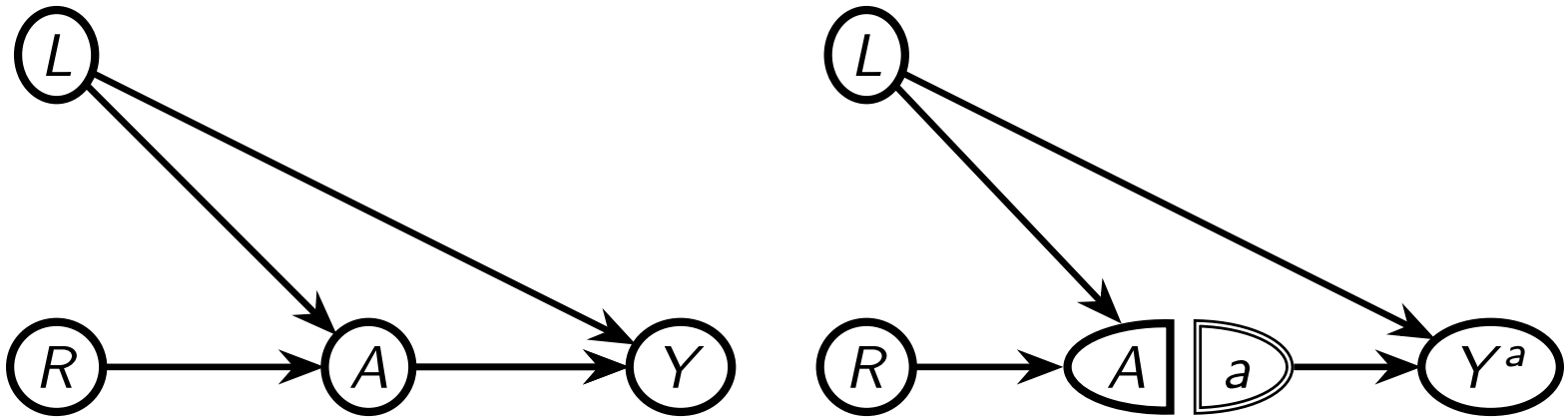
SWIG in an experiment with loss to follow-up (C)

A is treatment, C is censoring. The counterfactual outcome $Y^{a,c=0}$ is the outcome if we kept every individual uncensored ($c = 0$) under treatment a.



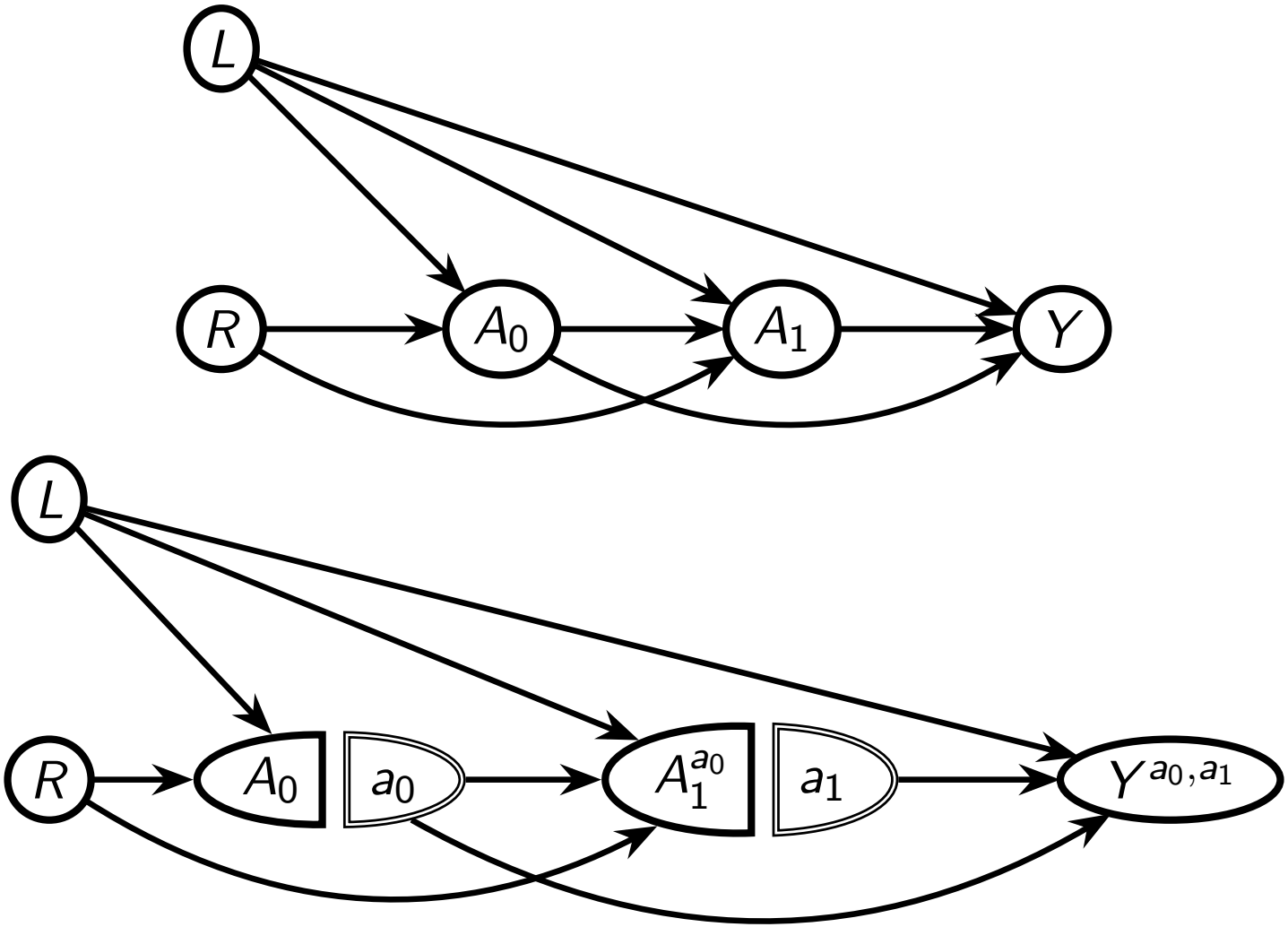
SWIG in an experiment with imperfect adherence

R is the strategy that was assigned, and A denotes taking treatment. Here, the counterfactual in the SWIG is the outcome had the patient taken treatment a. The lack of an arrow from R to Y^a encodes the assumption that randomisation only causes the outcome through the treatment A.



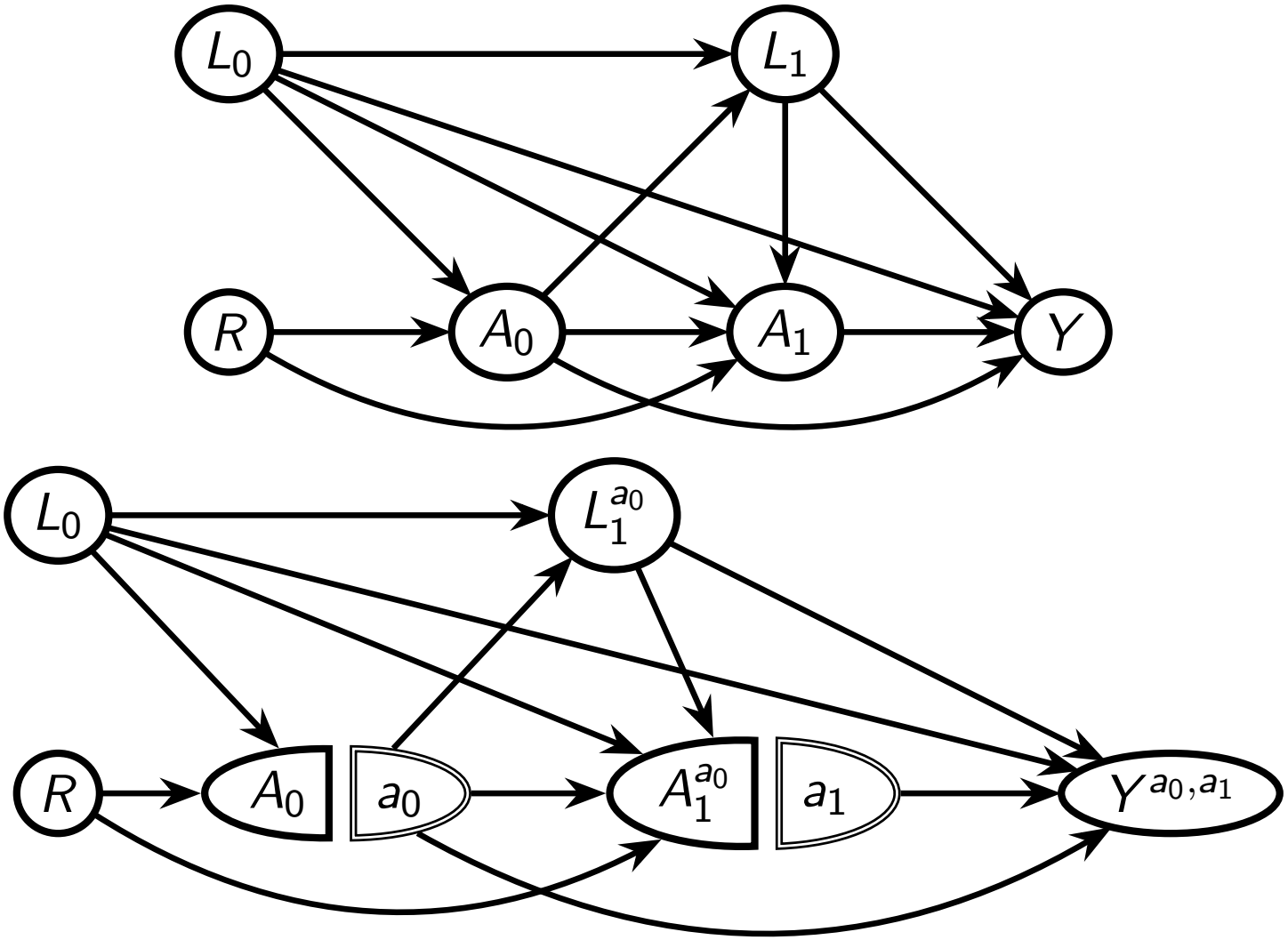
SWIG in an experiment with imperfect adherence

R is the strategy that was assigned, and A_k denotes taking treatment at time $k \in \{0, 1\}$.



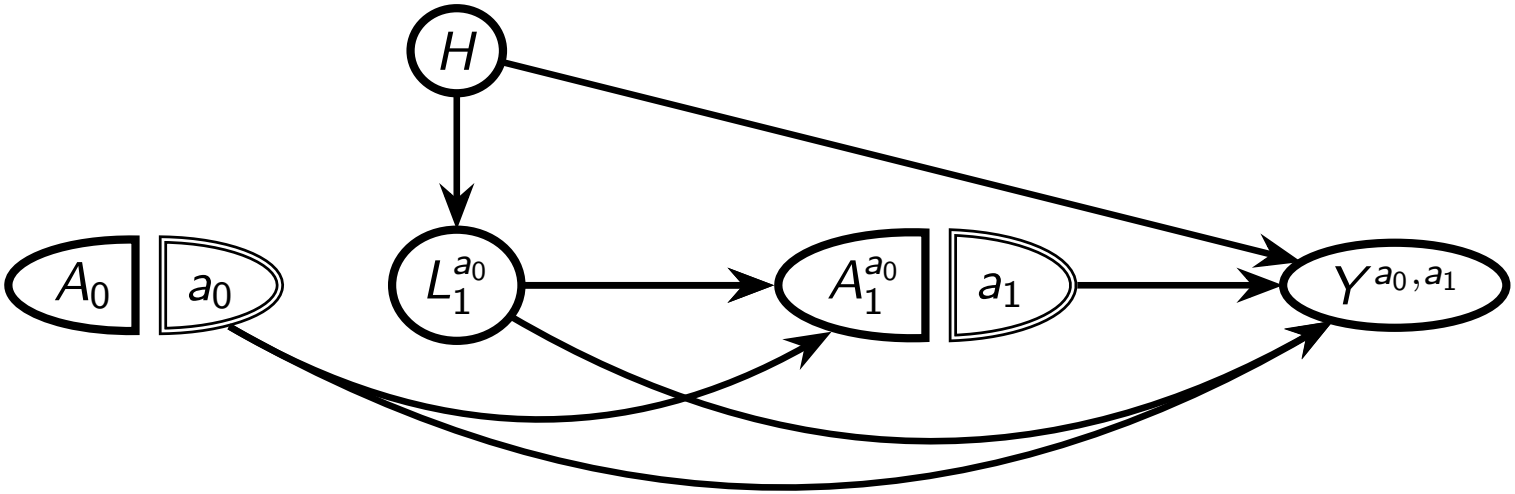
SWIG in an experiment with imperfect adherence

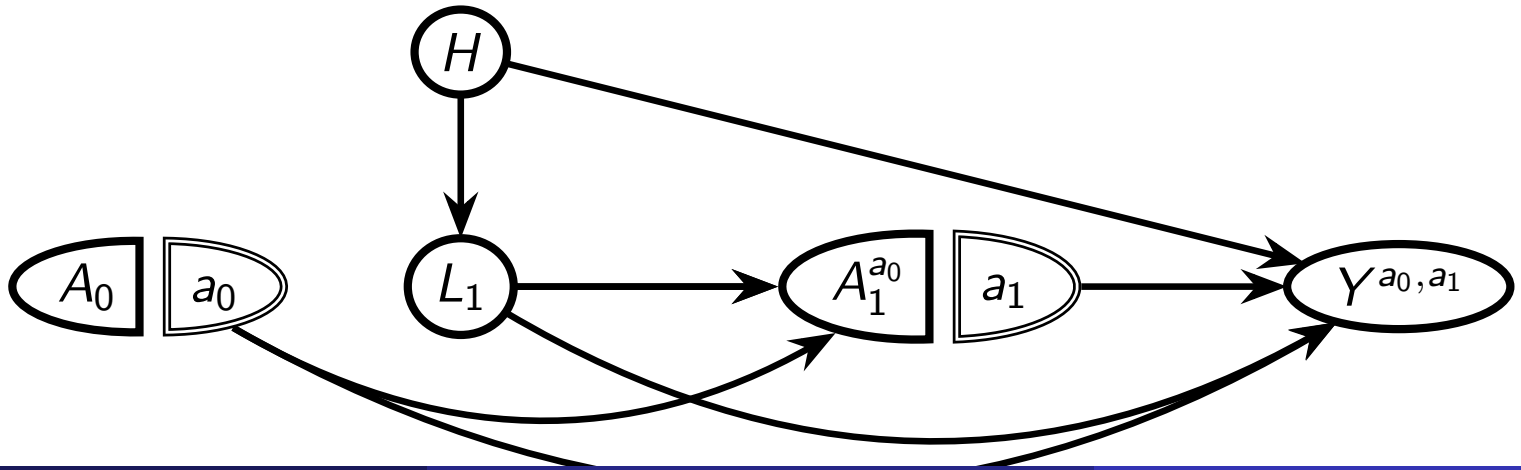
R is the strategy that was assigned, and A_k denotes taking treatment at time $k \in \{0, 1\}$.



SWIG and independencies

These graphs illustrate minimal labelling ($L_1^{a_0} = L_1$). The first graph is not minimally labelled, but encodes the same information as the second graph which is minimally labelled.





SWIG criterion for identification of effects

Consider the observed random variables $\overline{A}_K, \overline{L}_K, Y$.

Definition (g-formula)

The g-formula for the *marginal* of $Y \equiv Y_K$ under treatment assignment $\bar{a} = \bar{a}_K = (a_0, \ldots, a_K)$ is defined as

$$b_{\bar{a}}(y) = \sum_{\bar{l}_K} p(y \mid \bar{l}_K, \bar{a}_K) \prod_{j=0}^{K} p(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1}),$$

where $\bar{l}_k = (l_0, \ldots, l_k)$, $k \leq K$, are instantiations of **observed** variables $\overline{L}_k = (L_0, \ldots, L_k)$, $k \leq K$.

We define variables indexed by subscript "-1", e.g. L_{-1}, to be empty.[26]

[26] Robins, "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect"; Richardson and Robins, "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality".

Sufficient condition for identification

Theorem (Identification of static regimes)

Consider an intervention that sets $\bar{a} = \bar{a}_K = (a_0, \ldots, a_K)$. *Under positivity and consistency,*

$$P(Y^{\bar{a}} = y) = b_{\bar{a}}(y)$$

if and only if for $k \in \{0, \ldots, K\}$

$$Y^{\bar{a}} \perp\!\!\!\perp I(A_k = a_k) \mid L_0, \ldots, L_k, A_0 = a_0, \ldots, A_{k-1} = a_{k-1}.$$

This theorem follows from Robins[27] and Richardson and Robins[28], and is closely related to the backdoor theorem of Pearl[29].
The theorem establishes when we can use the g-formula to identify causal effects.

[27] Robins, "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect".

[28] Richardson and Robins, "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality".

[29] Pearl, "Causal diagrams for empirical research".

Proof of the "if" part in a simple case

Consider the case with two treatments (A_0, A_1) and a binary outcome $Y \in \{0, 1\}$. Suppose that $Y^{a_0,a_1} \perp\!\!\!\perp A_0$ and $Y^{a_0,a_1} \perp\!\!\!\perp A_1 \mid L_1, A_0 = a_0$

Proof.

$$
\begin{aligned}
\mathbb{E}(Y^{a_0,a_1}) &= \mathbb{E}(Y^{a_0,a_1} \mid A_0 = a_0) \text{ exchangeability} \\
&= \sum_{l_1} \mathbb{E}(Y^{a_0,a_1} \mid L_1 = l_1, A_0 = a_0) p(l_1 \mid a_0) \\
&= \sum_{l_1} \mathbb{E}(Y^{a_0,a_1} \mid A_1 = a_1, L_1 = l_1, A_0 = a_0) p(l_1 \mid a_0) \text{ exchangeability} \\
&= \sum_{l_1} \mathbb{E}(Y \mid A_1 = a_1, L_1 = l_1, A_0 = a_0) p(l_1 \mid a_0) \text{ consistency, positivity}
\end{aligned}
$$

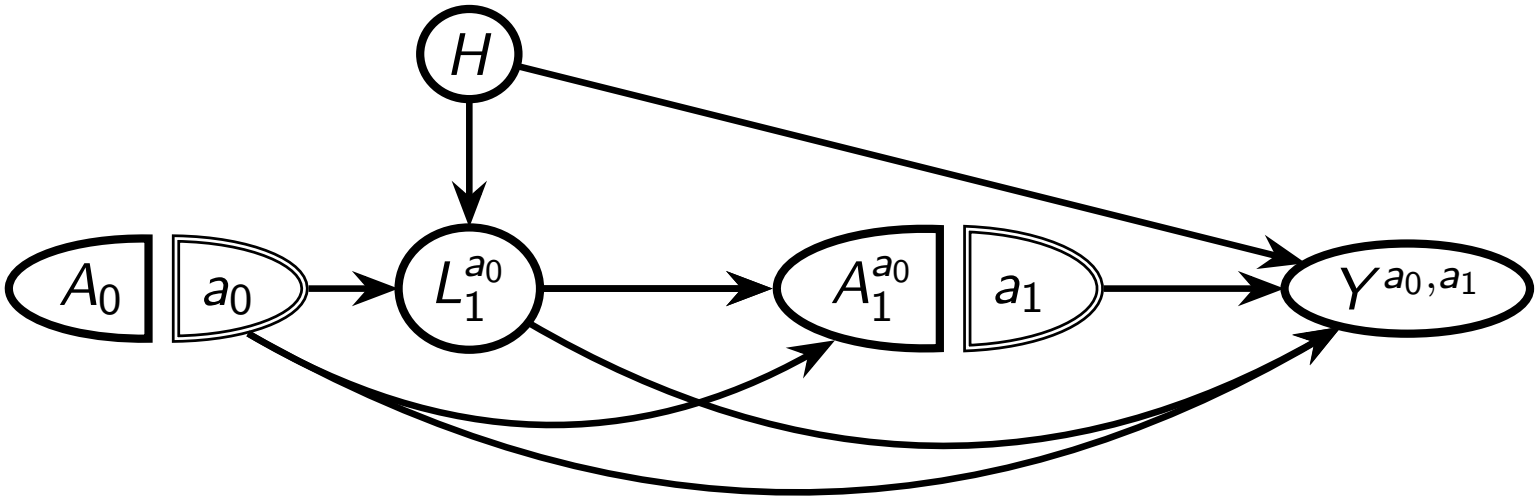
□

Comments to the g-formula

- The independence condition in the identification theorem cannot be read directly off of a SWIG. However, on the next slide we see how the identification condition is implied by an independence in the SWIG.
- Importantly, the g-formula allows identification in the presence of unmeasured variables.

Reading off independencies in SWIGs

Let H be a hidden (unmeasured) variable



We can read off $Y^{a_0,a_1} \perp\!\!\!\perp A_1^{a_0} \mid L_1^{a_0}, A_0$.
However, what we needed for using the g-formula is the independence $Y^{a_0,a_1} \perp\!\!\!\perp A_1 \mid L_1, A_0 = a_0$.
Use consistency: $A_1^{a_0} \mid L_1^{a_0}, A_0 = a_0$ is equal to $A_1 \mid L_1, A_0 = a_0$, i.e.,
$Y^{a_0,a_1} \perp\!\!\!\perp A_1^{a_0} \mid L_1^{a_0}, A_0 \implies Y^{a_0,a_1} \perp\!\!\!\perp A_1 \mid L_1, A_0 = a_0$.

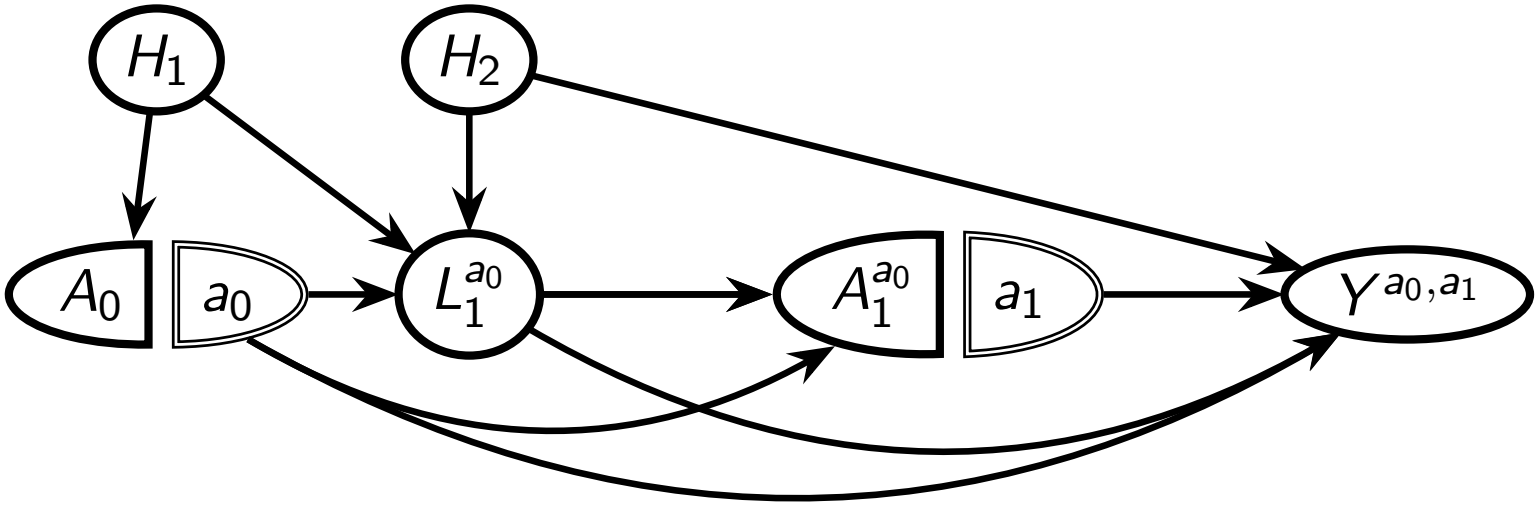
Using the identification theorem

Thus, we can identify the expected counterfactual outcome under the intervention that sets $A_0 = a_0$ and $A_1 = a_1$ in the graph in Slide 184 as

$$\mathbb{E}(Y^{a_0,a_1}) = \sum_{l_1} \mathbb{E}(Y \mid A_1 = a_1, L_1 = l_1, A_0 = a_0) P(L_1 = l_1 \mid A_0 = a_0).$$

Note that we have identified the counterfactual as a function of only the observed variables in the graph, even if there is a hidden variable H in the graph.

Additional SWIG

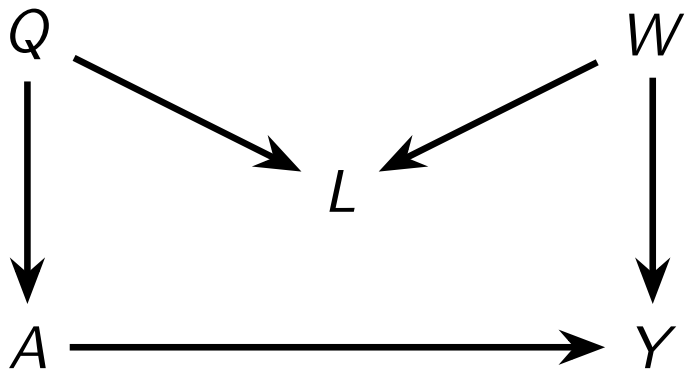


What is the g-formula? Compare to Figure 184. Indeed, the g-formula is just a function of observed data distributions, but here it does not identify the causal estimand because the identification conditions are violated.

Some insights

- We have studied identification from an "all or nothing" perspective.
  - We will later look at sensitivity analyses and bounds.
- The identification assumptions we have studied are non-parametric (PS: I consider this to be a feature, not a bug). We have not considered other assumptions that also can be used to justify identification, for example
  - monotone effects.
  - no effect modification.
- We have **not learned** the graphical structure. On the other hand, we have learned what we can infer from a given graphical structure; heuristically, we encode what we know and believe in the graph, and then we deduce what we can learn from this knowledge and assupmtions.
  - Learning the graphical structure itself from data is a very ambitious task.
  - In principle, the causal structure could be learned by doing a large amount of experiments (I am not discussing this in more detail here).

Berkson bias



- A Drink a glass of red wine a day.
- Y Nausea
- L Aspirin
- Q Family history of cardiovascular disease
- W Frequency of headache

Q: We measure Aspirin. Should we adjust for Aspirin in the analysis?
Draw the SWIG...

Section 18

Dynamic regimes

Dynamic regimes

Definition (Dynamic regime)

A dynamic regime $g = (g_0, \ldots, g_k)$, where $g_k : (\overline{A}_{k-1}, \overline{L}_k) \mapsto A_k$, is a policy that assigns treatment (possibly at multiple time points) based on the measured history $(\overline{A}_{k-1}, \overline{L}_k)$.

We will restrict ourselves to settings where

$$g_k : (\overline{L}_k) \mapsto A_k$$

.

Dynamic regime SWIGs

Definition (d-SWIG from Robins and Richardson)

Given a template $\mathcal{G}(a)$ and a dynamic regime g for $\bar{a}$, the d-SWIG $\mathcal{G}(g)$ is defined by applying the following transformation:

- Replace each fixed node a_j with a random node A_j^{g+} that inherits children from a_j. Include dashed directed edges from every variable that is an input to the function g_i that determines the variable A_i^{g+}.
- Each random node V_i that is a descendant of at least one variable A_i^{g+} is relabeled as V_i^g.

Time-varying exposures (treatments) are frequent

Examples:

- Smoking status, which depends on other events in life.
- A therapeutic drug, for which the dose is adjusted according to the response over time (patients take the drug every day, every week etc)
- Cancer screening, which e.g. depends on previous diagnostic tests.
- Surgical interventions (e.g. transplants) are given at a certain time after the diagnosis.
- Expression of genes.

Running example: HIV

Consider a 5-year follow-up study of individuals infected with the human immunodeficiency virus (HIV)[30].

- A_k takes value 1 if the individual receives antiretroviral therapy in month k, and 0 otherwise. Define $A_{-1} = 0$.
- Suppose Y measures health status at 5 years of follow-up.
- So far we have considered *deterministic* treatment rules, for example "always treat", where the outcome of interest is $Y^{a=1}$ vs "never treat", where the outcome of interest is $Y^{a=0}$.
  When $\overline{A} \equiv \overline{A}_K$, we can define 2^K such static regimes...
- However, often we want to make *dynamic* treatment decisions.
- Let $L_k \in \{0, 1\}$ be an indicator of low CD4 cell count measured at month k.
- Depending on the value of L_k, we may argue that it is good or bad to start treatment at time k.

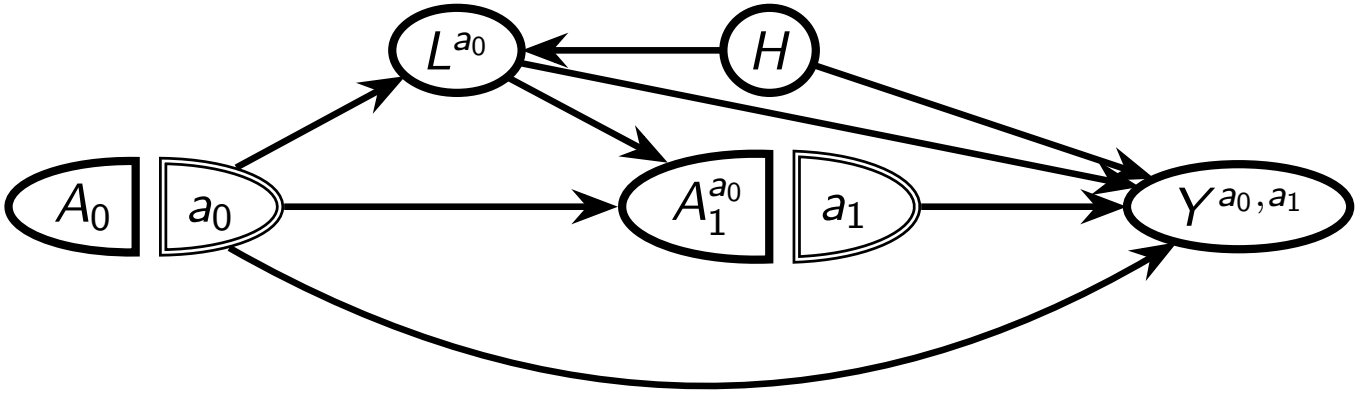
[30] Hernan and Robins, *Causal inference: What if?*

Example of Dynamic Regime

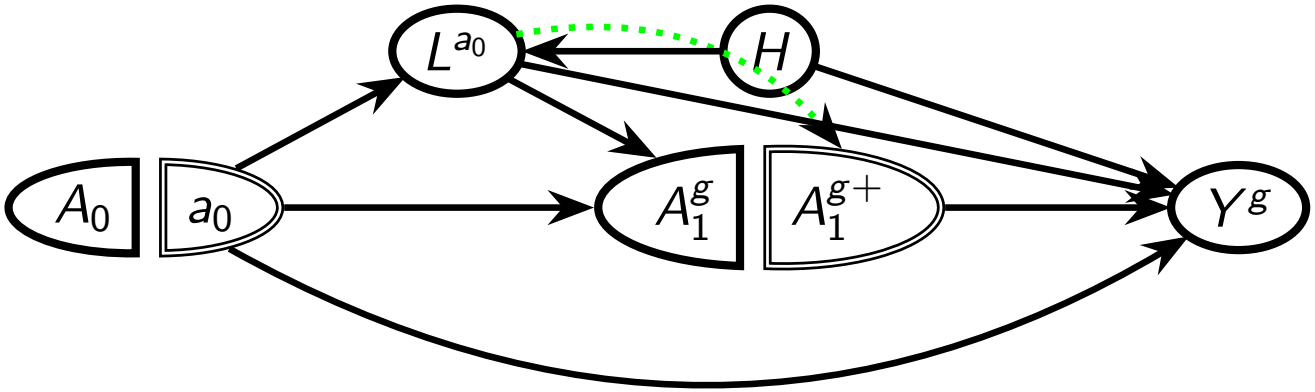
A simple example of a dynamic regime g for setting with two treatments is

- $A_0^{g+} = a_0$.
- $A_1^{g+} = L_1^{a_0}$
  In the HIV example this would mean that you are treated at time 1 if the CD4 cell count is low at that time.

Static vs dynamic



$Y^{a_0,a_1} \perp\!\!\!\perp A_0$ and $Y^{a_0,a_1} \perp\!\!\!\perp A_1^{a_0} \mid L_0^{a_0}, A_0$.



$Y^{g} \perp\!\!\!\perp A_0$ and $Y^{g} \perp\!\!\!\perp A_1^{a_0} \mid L_0^{a_0}, A_0$.
$Y^{g} \perp\!\!\!\perp A_0$ and, using the graph and consistency, $Y^{g} \perp\!\!\!\perp A_1 \mid L_0, A_0 = a_0$.

Sufficient condition for identification (Repetition of previous slide)

Theorem (Identification of static regimes)

Consider an intervention that sets $\overline{a} = \overline{a}_K = (a_0, \ldots, a_K)$. *Under positivity and consistency,*

$$P(Y^{\overline{a}} = y) = b_{\overline{a}}(y)$$

if and only if for $k \in \{0, \ldots, K\}$

$$Y^{\overline{a}} \perp\!\!\!\perp I(A_k = a_k) \mid L_0, \ldots, L_k, A_0 = a_0, \ldots, A_{k-1} = a_{k-1}.$$

This theorem follows from Robins[31] and Richardson and Robins[32], and is closely related to the backdoor theorem of Pearl[33]; Indeed, we can just call it "The SWIG backdoor criterion"

~~The theorem establishes when we can~~ use the g-formula to identify causal effects.

[31] Robins, "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect".

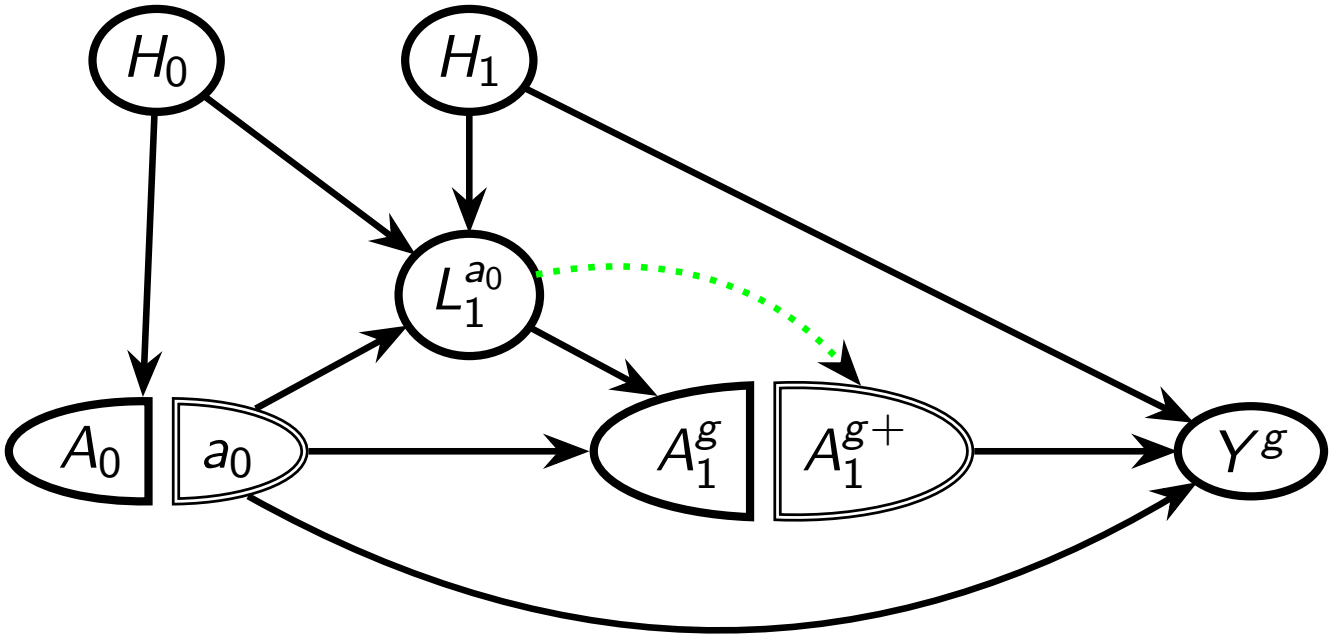
[32] Richardson and Robins, "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality".

[33] Pearl, "Causal diagrams for empirical research".

Identification results for dynamic regimes

- We can use the same identification conditions (independencies in Slide 181) as for static regimes, only if A_k^{g+} **is not** a function of A_j^{g+} for $j < k$. However, we need to use the extended g-formula as the identification formula (as defined in Slide 201).
- if A_k^{g+} **is** a function of A_j^{g+} for any $j < k$, we need slightly stronger conditions (we are not presenting them now). This is e.g. the case in the graph in Slide 199 (due to the red arrow).

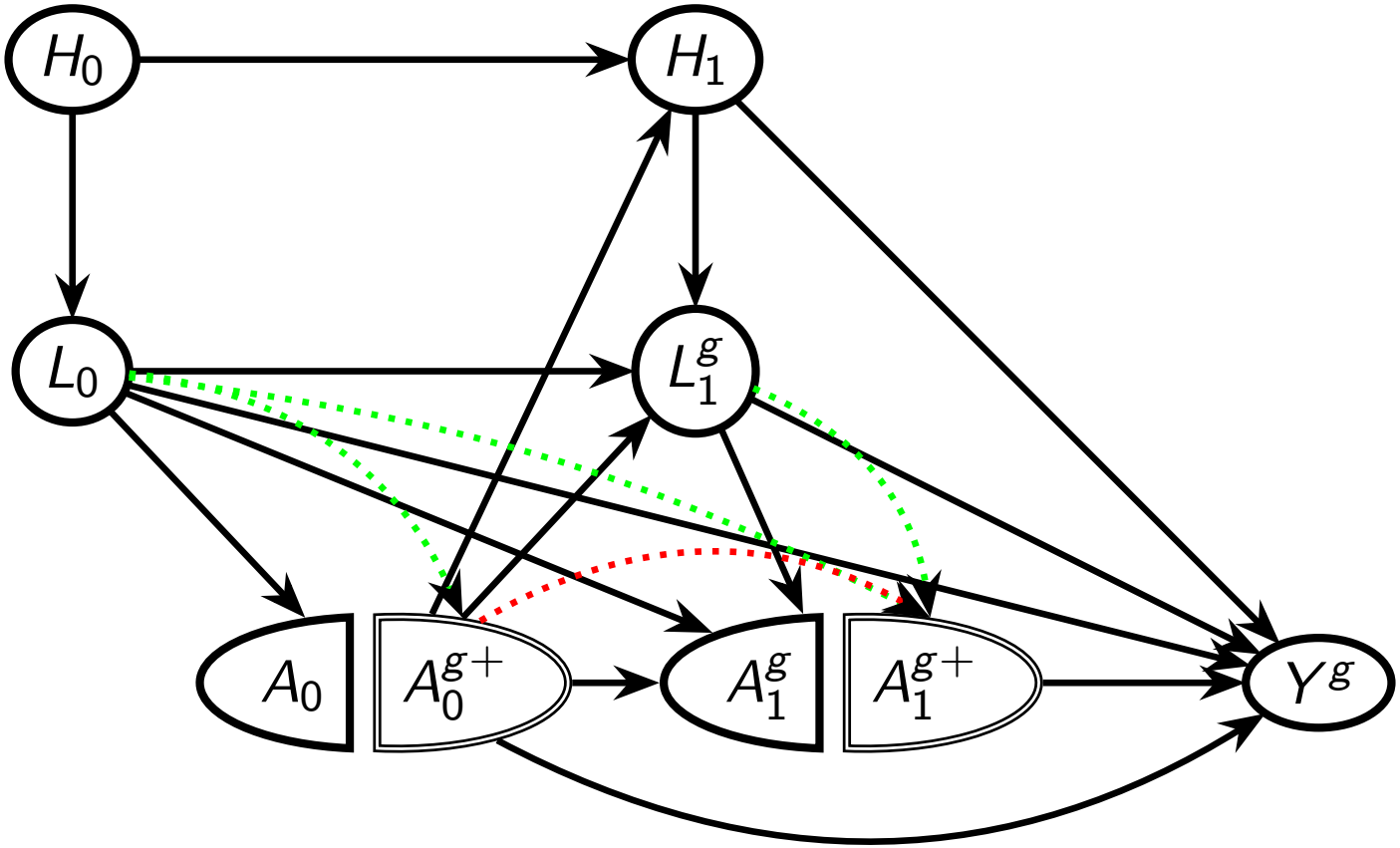
Does the identification conditions hold in the following Dynamic SWIG?



$Y^g \not\perp\!\!\!\perp A_0$ because $A_0 \leftarrow H_0 \rightarrow L_1^{a_0} \rightarrow A_1^{g+} \rightarrow Y^g$ is open. However, we would have identification in a static SWIG where $A_1^{g+} \equiv a_1$. So, in that sense, dynamic regimes require stronger conditions for identification, even though the independencies are stated in the same way.

HIV SWIG

A (busy) graph illustrating a conditional RCT, where H_0 and H_1 are hidden variables (e.g. the actual immune status of the patient).



Consistency for dynamic regimes

Now we generalize the consistency conditions such that it is valid for time-varying dynamic regimes. Indeed, it can simply be expressed as

$$\text{if } \overline{A}_K = \overline{A}_K^{g+}, \text{ then } Y^g = Y.$$

A special case for static regimes is if $\overline{A}_K = \overline{a}_K$, then $Y^{\overline{a}_K} = Y$.

Marginal extended g-formula under interventions that depend on $\overline{L}_k$

Suppose that g_k is only a function of $\overline{L}_k$. Then, the marginal extended g-formula is defined as the following function of observed random variables $\overline{A}_K, \overline{L}_k, Y$.

Definition (Marginal extended g-formula)

$$b_g(y) = \sum_{\bar{a}_K} \sum_{\bar{l}_K} p(y \mid \bar{l}_K, \bar{a}_K) \prod_{j=0}^{K} p(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1}) p^g(a_j \mid \bar{l}_j),$$

where $\bar{l}_k = (l_0, \ldots, l_k)$, $k \leq K$, are instantiations of **observed** variables and $p^g(a_j \mid \bar{l}_j)$ is the density of A_k^{g+} given $\overline{L}_k^g$, which is determined by g_k.

We let variables indexed by subscript -1, e.g. L_{-1} be empty.
Note that $p^g(a_k \mid \bar{l}_k)$ is a known function. It is determined by the investigator (even if it has a superscript g). If g_k is a deterministic function of $\bar{l}_k$, then

$$p^g(a'_k \mid \bar{l}_k) = \begin{cases} 1 \text{ if } a'_k = g_k(\bar{l}_k), \\ 0 \text{ if } a'_k \neq g_k(\bar{l}_k), \ k \in \{0, \ldots, K\}. \end{cases}$$

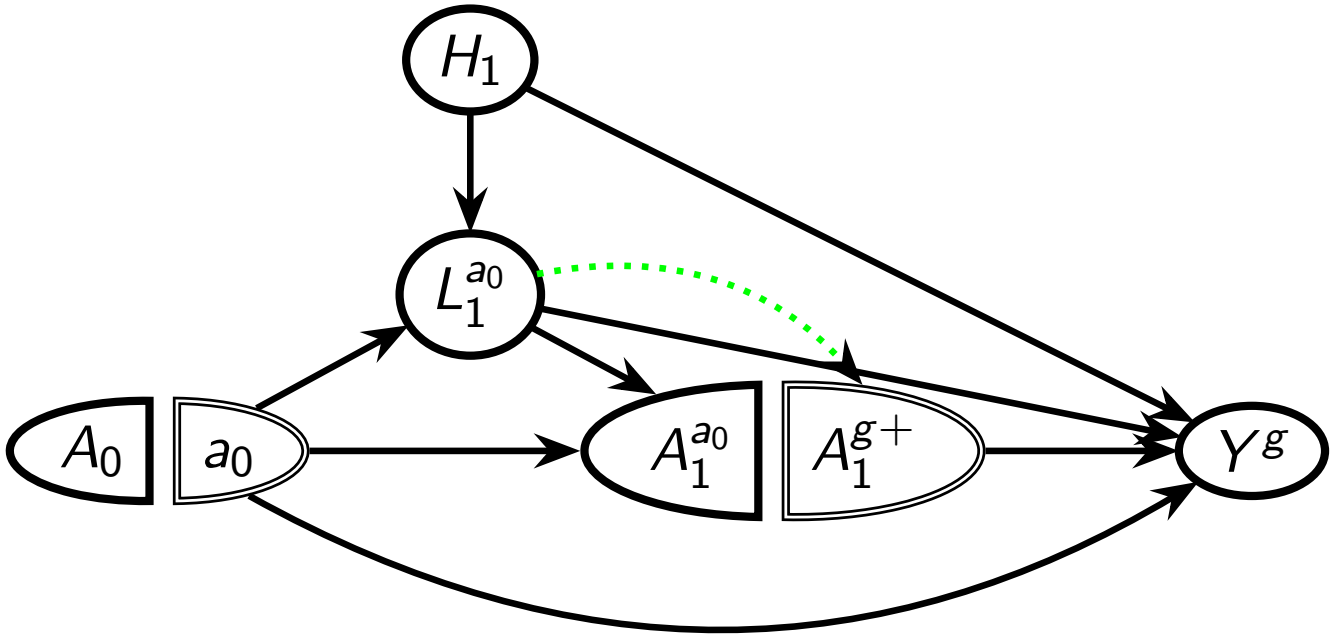
Relation to the g-formula for static regimes

The dynamic extended g-formula density generalizes the marginal g-formula from slide 181, because for a static intervention that sets $\bar{a} = (a_0, \ldots, a_K)$ we have that for $k \in \{0, \ldots, K\}$,

$$p^g(a'_k \mid \bar{l}_k) = \begin{cases} 1 \text{ if } a'_k = a_k, \\ 0 \text{ if } a'_k \neq a_k. \end{cases}$$

HIV example

Consider the example in Slide 193, and suppose the following SWIG:



let the dynamic regime g be

- $A_0^{g+} = a_0$.
- $A_1^{g+} = L_1^{a_0}$

That is, a patient is treated at time 1 if the CD4 cell count is low at that time.

HIV Example cont.

Then the g-formula reduces to

$$
\begin{aligned}
&b_g(y) \\
&= \sum_{\bar{a}_1', l_1} p(y \mid A_1 = a_1', L_1 = l_1, A_0 = a_0') I(a_1' = l_1) p(l_1 \mid A_0 = a_0) I(a_0' = a_0), \\
&= \sum_{l_1} p(y \mid A_1 = l_1, L_1 = l_1, A_0 = a_0) p(l_1 \mid A_0 = a_0).
\end{aligned}
$$

because

$$
p^g(a_1' \mid \bar{l}_1) = \begin{cases} 1 \text{ if } a_1' = l_1, \\ 0 \text{ if } a_1' \neq l_1. \end{cases}
$$

SWIG criterion to identify effects of dynamic regimes (you do not need to understand the extended g-formula density)

Definition (extended g-formula density)

The *dynamic extended g-formula density* for $Y \equiv Y_K$ under treatment regime g given by the functions $g_0, \ldots, g_K$ that determine $\overline{A}_K = (A_0, \ldots, A_K)$ is

$$f^g(y, \bar{l}_K, \bar{a}_K, \bar{a}_K^+) = p(y \mid \bar{l}_K, \bar{a}_K^+) \prod_{j=0}^{K} p(l_j, a_j \mid \bar{l}_{j-1}, \bar{a}_{j-1}^+) \prod_{t=0}^{K} p^g(a_t^+ \mid pa_{A_t^{g+}}),$$

where $\bar{l}_k = (l_0, \ldots, l_k)$, $k \leq K$, are **observed** variables, $p^g(a_t^+ \mid pa_{A_t^{g+}})$ is the density of A_t^{g+} given $PA_{A_t^{g+}}$ is the input to g_t, for $t \in \{0, K\}$.

James M Robins. "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". In: *Mathematical modelling* 7.9-12 (1986), pp. 1393–1512; Thomas S Richardson and James M Robins. "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality". In: *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper* 128.30 (2013).

Section 19

A brief note on estimation (learning)

Lecture 8: Plan for estimation (learning)

- Review foundations of estimation theory that are relevant to causal inference.
  - Statistical models (Parametric and non-parametric).
  - Correctly specified models.
- Motivate why we need to study certain estimation problems.
  - Convergence of conditional means.
- Introduce some commonly used estimators: Regression estimators and inverse probability weighted estimators.
  - Brief summary of linear models.
  - Logistic regression models.
  - M-estimators.
  - Link this back to counterfactuals.

My take on data science

1. Start with the question.
   (Design your target trial)
2. Formalize the question in mathematical language.
   (Define your estimand)
3. Display the assumptions that are needed to identify your estimand.
   (Present your identifiability conditions)
4. Compute estimates of your estimands from your data.
   (Do your estimation)

$\Longrightarrow$ we **never** start the process by considering a regression model (linear, logistic, Cox model, neural net, random forest, ..., whatever).

Finite sample inference: Where does randomness come from?

- We will mostly consider superpopulation inference, where the randomness comes from the fact that we have a random draw from the superpopulation.
- However, in a randomised trial, we do not necessarily need to consider a superpopulation at all.
- In these (simple) settings, we can do *finite* sample inference.
- Yet, we shall see that to generalize the results outside of the study – which is really what researcher would like to do in most settings – it is necessary to consider large sample extensions (which fundamentally ends up being superpopulations).

Superpopulation inference and finite sample inference

- We will most often suppose that our study population is sampled at random from an (essentially) infinite superpopulation, sometimes referred to as the target population.
- Broadly speaking, we aimed to generalize our results to this superpopulation.
- It is possible to take a different point of view in randomised trials, often called "design-based inference", which we will study now. This does not require the consideration of a superpopulation at all.[34]

Definition (Design-based inference)

Inference is drawn from a *finite* population, where the potential outcomes of the experimental units are fixed and the randomness comes solely from the treatment assignment.

[34] However, to generalize results from finite samples to settings outside of the experiment – even if we start in the design based setting – it is necessary to rely on superpopulation inference. Thus, if we are interested in using the results from the trials for decisions (or rigorous reasoning more broadly) outside of the experiment, it seems that we need to rely on superpopulation inference anyway.

Superpopulation inference and finite sample inference

- We will most often suppose that our study population is sampled at random from an (essentially) infinite superpopulation, sometimes referred to as the target population.
- Broadly speaking, we aimed to generalize our results to this superpopulation.
- It is possible to take a different point of view in randomised trials, often called "design-based inference", which we will study now. This does not require the consideration of a superpopulation at all.[35]

Definition (Design-based inference)

Inference is drawn from a *finite* population, where the potential outcomes of the experimental units are fixed and the randomness comes solely from the treatment assignment.

[35] However, to generalize results from finite samples to settings outside of the experiment – even if we start in the design based setting – it is necessary to rely on superpopulation inference. Thus, if we are interested in using the results from the trials for decisions (or rigorous reasoning more broadly) outside of the experiment, it seems that we need to rely on superpopulation inference anyway.

Fisher Randomisation inference

- Key idea: do inference based solely on the assignment mechanism.
- The counterfactuals $Y_i^{a=1}$, $Y_i^{a=0}$ are considered to be *fixed* variables.
- All the *randomness* comes from the random assignment of A.
- Fisher's aim was to test the sharp null hypothesis, using Fisher exact test.
- The idea is basically a stochastic proof by contradiction...
- Fisher's null hypothesis is $H_0 : Y_i^{a=1} \equiv Y_i^{a=0}$ for all $i \in \{1, 2, \ldots, n\}$. In words, the treatment has no effect of the outcomes in no individual. Under the null hypothesis (but of course not under the alternative) $Y_i^{a=1} = Y_i^{a=0} = Y_i$.
- This null hypothesis is called a **sharp** null hypothesis because it is specified such that it allows the researcher to fill in a hypothetical value for each unit's missing counterfactual outcome

Fisher's exact test: A test of individual effects

- Define the *sharp null hypothesis* $H_0 : Y_i^{a=1} = Y_i^{a=0}$ for all $i \in \{1, 2, \ldots, n\}$.
- Define a test statistic[36], e.g. $S^{diff} = \frac{1}{n_1} \sum_{i:A_i=1} Y_i - \frac{1}{n_0} \sum_{i:A_i=0} Y_i$.
- Let s^* be an observed test statistic. Then $P(S \geq s^*)$ is a p-value, where the probability is under the law that describes the null hypothesis.
- Fisher suggested an exact test.
  - The idea is to ask the following question: How unusual or extreme is the observed statistic (say, absolute difference), assuming that the null hypothesis is true?
- Intuitively, we want to have *power* against alternative hypotheses, but this is somehow complicated because there are many alternative hypotheses. It seems reasonable to have good power against alternative hypotheses that are substantively most interesting.

[36] A statistic is a known, real-valued function of the data (here, $Y_1, A_1, L_1, \ldots, Y_n, A_n, L_n$)

*Examples of statistics

- Averages (like above)
- Trimmed means
- Quantiles (medians)
- T-statistics
- Rank statistics (perhaps good when heavy-tailed distributions)

One example is the Kolmogorov-Smirnov Statistic. Define, the empirical distributions

$$\hat{F}_{a=1}(y) = \frac{1}{n_1} \sum_{i:A_i=1} I(Y_i \leq y) \quad \hat{F}_{a=0}(y) = \frac{1}{n_0} \sum_{i:A_i=1} I(Y_i \leq y).$$

The Kolmogorov-Smirnov Statistic is

$$S^{ks} = \sup_y |\hat{F}_{a=1}(y) - \hat{F}_{a=0}(y)| = \max_i |\hat{F}_{a=1}(Y_i) - \hat{F}_{a=0}(Y_i)|.$$

*We can combine statistics

- Fisher's exact p-value inference is valid when there is one test statistic and one null hypothesis.
- However, we can combine test statistics.
  - Consider two statistics S^1 and S^2.
  - The combine $S^{comb} = g(S^1, S^2)$. (e.g. $S^{comb} = \max(S^1, S^2)$)
  - Then we can calculate a p-value

$$P(S^{comb} \leq s^{\star,comb})$$

Illustration of Fisher's exact test

Under the sharp H_0, we can impute missing values of the counterfactuals

i	$Y_i^{a=1}$	$Y_i^{a=0}$	A_i	Y_i
1	-5	-5	1	-5
2	6	6	0	6
3	8	8	1	8
4	0	0	0	0

Table 3: Fisher's idea

The idea is resampling without replacement

Consider the estimator $\frac{1}{n_1}\sum_{i:A_i=1} Y_i - \frac{1}{n_0}\sum_{i:A_i=0} Y_i$. Because we have a completely randomised experiment, the following $\binom{4}{2} = 6$ scenarios are equally possible under H_0,

$$\boldsymbol{A} = (1,1,0,0),\ \hat{\tau} = \frac{-5+6-8-0}{2} = -3.5$$

$$\boldsymbol{A} = (1,0,1,0),\ \hat{\tau} = \frac{-5-6+8-0}{2} = -1.5$$

$$\boldsymbol{A} = (1,0,0,1),\ \hat{\tau} = \frac{-5-6-8+0}{2} = -9.5$$

$$\boldsymbol{A} = (0,1,1,0),\ \hat{\tau} = \frac{5+6+8-0}{2} = 9.5$$

$$\boldsymbol{A} = (0,1,0,1),\ \hat{\tau} = \frac{5+6-8+0}{2} = 1.5$$

$$\boldsymbol{A} = (0,0,1,1),\ \hat{\tau} = \frac{5-6+8+0}{2} = 3.5$$

One way of explaining Fisher's exact test

1. Do the randomization.
2. Calculate a statistic S, a function of the observed data.
3. Under the assumption of H_0, i.e. no individual level causal effect, fill in missing potential outcomes.
4. Under the assumption of H_0, generate many hypothetical replications of the randomization, and in each of which calculate a statistic S_{rep}.
5. Compare S with the values S_{rep}

This is an example of a permutation test.

More formally

- Define $H_0 : Y_i^{a=1} = Y_i^{a=0}$.
- Now, consider the randomisation distribution of two statistics S
- Define $\mathcal{F} = (\boldsymbol{Y^0}, \boldsymbol{Y^1})$. In this case, the randomization distributions of $S = S(\boldsymbol{A}, \boldsymbol{Y}, \boldsymbol{L})$ is

$$F(s) = P(S \leq s \mid \mathcal{F})$$

- Then the one-sided p-value of observing the same value or more extreme of the observed statistics S is $F(S)$.
- In our example, the one-sided p-value is $1 - F(-1.5) = 1 - 0.5$.

Fisher's randomization test formally

Theorem (Nominal coverage of the exact test)

Under consistency and H_0, $P(F(S) \leq \alpha \mid \mathcal{F}) \leq \alpha$ for all $\alpha \in (0, 1)$.

Proof.

This follows from some basic properties of the distribution function: indeed, $F^{-1}(\alpha) = \sup\{s : F(s) \leq \alpha\}$. Also $F(s)$ is non-decreasing and right-continuous and therefore

$$P(F(S) \leq \alpha) = P(S < F^{-1}(\alpha)) = \lim_{s \to F^{-1}(\alpha)} P(S \leq s) \leq \alpha.$$

□

PS: you may have seen the probability integral transform before, i.e. if X is continuous, then $Z = F(X) \sim U(0, 1)$

$$P(F(X) \leq \alpha) = P(X \leq F^{-1}(\alpha)) = F(F^{-1}(\alpha)) = \alpha.$$

Conservative or good?

Conservative does not necessarily mean appropriate. Consider a confidence interval formed by stating that a random 95% of the time, the interval is any positive or negative number, and that 5% of the time, the interval is the number 0. Such an interval would cover the true value of any quantity of interest at least 95% of the time, and thus would also be a "conservative" interval. It would not, however, be of any use....
Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015

Checking for no causal effect (hypothesis testing)

- Suppose we want to check if there is no causal effect.
- A classical frequentist approach goes as follows
  - Assume no effect (the null hypothesis).
  - Calculate a statistic,[37] and see how surprising the statistics is, under the assumption of no effect.
  - If it is very surprising, we reject.
- This is contrapositive logic, applied to probabilities.

[37] A statistic is a known, real-valued function of the data

We should be careful with this (Example from Shpitser)

Suppose we do cancer screening.

- Consider a rare cancer, our outcome Y, such that $P(Y = 1) = 0.00001$
- Consider also a test T. And suppose
  - Test false positive $P(T = 1 \mid Y = 0) = 0.01$.
  - Test false negative $P(T = 0 \mid Y = 1) = 0.001$.
- Suppose we had a positive test, $Y = 1$. Should we worry?
- Just use Bayes theorem,

$$P(Y = 1 \mid T = 1) = \frac{P(T = 1 \mid Y = 1)P(Y = 1)}{P(T = 1)} \approx 0.001.$$

- What would the Frequentist do? Assume $Y = 0$, and check how surprised we would be, that is, calculate $P(T = 1 \mid Y = 0) = 0.01$, which is surprising....
- Lesson learned, if hypothesis probabilities are uneven, hypothesis testing is not ideal..

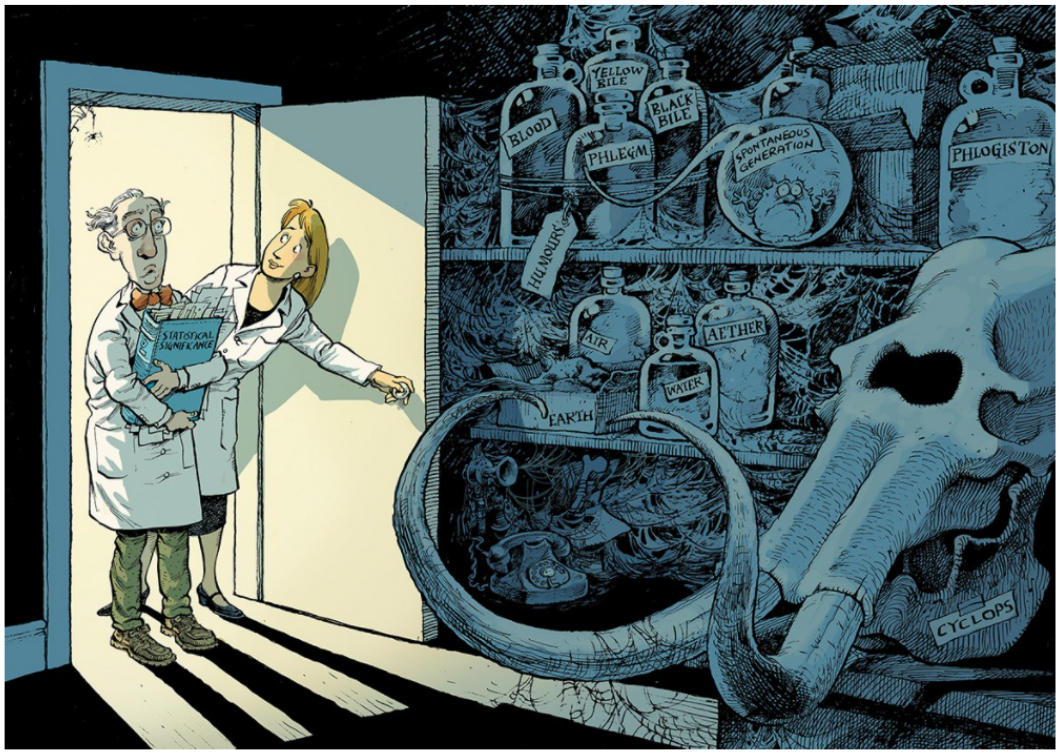
Abandon Statistical Significance?

COMMENT · 20 MARCH 2019

Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

Valentin Amrhein, Sander Greenland & Blake McShane



Reasons (may seem obvious, but worth a reminder)

- There is nothing wrong with the p-value itself, as a mathematical construct.
- However, it is often misused.
- $p < 0.05$ is an arbitrary threshold.
- P-hacking is frequently done in practice.

Blakeley B McShane et al. "Abandon statistical significance". In: *The American Statistician* 73.sup1 (2019), pp. 235–245

Estimation (learning) in causal inference settings (informal motivation)

- An identification formula motivates estimators.
- Estimation in causal inference settings is, in principle, identical to the inverse problem you have studied in previous machine learning or statistics classes.
- However, the functionals we are estimating are sometimes unusual, and therefore we sometimes need new estimators. Indeed, a lot of identification results in causal inference have motivated new estimation theory.
- Broadly speaking, causal inference researchers are concerned about bias.
  - After doing the hard work of deriving an identification formula, we do not want to induce bias in the estimation step.
- I remind you about how we divide the causal inference into different tasks: (i) Define your question of interest (estimand), (ii) Evaluate whether the estimand is identified, (iii) if it is identified, we proceed with estimation.

Estimation vs. identification

- We have considered identification assumptions that are necessary even if we had an infinite amount of data.
- The statistical modeling assumption we consider now are invoked because we do not have infinite amount of data.

PS: In this course we will mainly consider frequentist inference: probability is defined as a limiting frequency. An alternative is Bayesian inference,[38] which defines probability as a degree of belief.

[38] Again, this is not the same as a Bayesian network

Where does randomness come from?

In causal inference

- Sampling variability
  Like classical statistics
  - Sample from superpopulation (classical inference)
  - Sample of counterfactuals (e.g. Fisher Randomization test)
- Non-deterministic counterfactuals
  But we have assumed that the counterfactuals are deterministic. And, in practice, that doesn't change anything when we do superpopulation inference (we will get to it).

Where do we do inference

Suppose we estimate the proportion of treated individuals who develop the outcome (say, death) as

$$\hat{p} = \hat{P}(Y = 1 \mid A = 1) = 7/13,$$

and I get 95% confidence intervals in the usual way (called Wald intervals) as

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

When is this confidence interval valid and what does it mean?

`Example from Hernan & Robins, Chapter 10.3`

There are two options

1. Individuals are sampled at random from an essentially infinite super-population, sometimes referred to as the source or target population. Then, if we repeatedly draw random samples of size 13 from the treated individuals in the super-population, the number of individuals who develop the outcome among the 13 is a binomial random variable with success probability equal to the true $P(Y = 1 \mid A = 1)$.
   This is the model we will consider most of the time.
2. We are not considering a super-population; we are doing inference in the sample we have. We assume that every individual i has a non-deterministic probability $p_i^{a=1}$ of experiencing $Y = Y_i^{a=1} = 1$ (because we consider those with $A = 1$). However, for the confidence interval to be correct, we must assume that $p_i^{a=1}$ is constant in i, say, $p_i^{a=1} = p$. Think about the idea that $p_i^{a=1}$ is constant in i. This seems very contrived, as we would believe that individuals have different risk of the outcome, due to genetics, life style factors etc.

Motivation for regression modelling and the curse of dimensionality

Definition (Statistical model)

A statistical model $\mathcal{P}$ is a collection of laws, $\mathcal{P} = \{P_\eta : \eta \in \Gamma\}$.

PS: Statistical models are sometimes called probabilistic hypothesis classes in the machine learning literature.

Definition (Parametric statistical model)

A statistical model $\mathcal{P}$ is parametric $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^k$ for a positive integer k.

So far we have been non-parametric: we have not restricted ourselves to parametric models. This is arguably desirable, because then we *do not* impose parametric restrictions on the data generating mechanism.

Consistency of an estimator

Here is an informal definition. Consistency of an estimator with respect to a parameter (the estimand) means that, when the sample size increases, the estimates get arbitrarily close to the parameter.
PS: This definition is with respect to an estimator. We have previously discussed consistency as an identification conditions, concerning interventions, which is a different thing.

More formal definition of consistent estimator (not strictly needed, but for your information)

Let $\{P_\eta : \eta \in \Gamma\}$ is a family of distributions (laws), and $X_\eta = \{X_1, X_2, \ldots : X_i \sim P_\eta\}$ is an infinitely large sample from the law P_η. Let $\{\hat{\mu}_n(\eta)\}$ be a sequence of estimators for $\mu(\eta)$, where e.g. $\hat{\mu}_n$ is an estimator based on the first n observations of a sample. Then the sequence $\{\hat{\mu}_n(\eta)\}$ is said to be (weakly) consistent if

$$\operatorname*{plim}_{n\to\infty} \hat{\mu}_n(\eta) = \mu(\eta), \quad \text{for all } \eta \in \Gamma.$$

where plim denotes convergence in probability, that is,

$$P_\eta(\mid \hat{\mu}_n(\eta) - \mu(\eta) \mid > \epsilon) \to 0 \text{ as } n \to \infty \text{ for all } \epsilon > 0, \eta \in \Gamma.$$

Motivation: Simple mean estimation

- Suppose we are interested in estimating a parameter, say, $h(L, A, Y)$ from an observed sample of n observations, (L_i, A_i, Y_i), $i = 1, \ldots, n$.
- Suppose we would like to ignore the assumptions encoded in our model $\mathcal{P}$ when we study $h(L, A, Y)$; more precisely, we will only use the fact that we have draws from i.i.d. individuals where $\mathbb{E}(Y) = \mu$ and that Y is continuous with finite variance $\sigma^2 < \infty$.
- Our statistical model is non-parametric; $\mathcal{P} = \{P(Y = y) : \int y^2 f(y) dy < \infty\}$. For $h(L, A, Y) \equiv \mathbb{E}(Y)$, we would simply do the empirical mean (sample mean) $\mathbb{E}_n(Y) = \frac{1}{n} \sum_{i=1}^{n} Y_i$. By the weak law of large numbers (WLLN),

$$\lim_{n \to \infty} P(|\mathbb{E}_n(Y) - \mu| > \epsilon) = 0.$$

  So the estimator is consistent. Indeed, the estimator is $\sqrt{n}$-consistent, and by the CLT $\sqrt{n}(\mathbb{E}_n(Y) - \mu) \sim \mathcal{N}(0, \sigma^2)$.
- Because $\mathbb{E}_n(Y)$ has variance σ^2/n, which is $O_P(1/n)$, then $\sqrt{n}(\mathbb{E}_n(Y) - \mu)$ has variance σ^2 which is $O_P(1)$, i.e. "bounded in probability" or "uniformly tight": A sequence $\{Q_n\}$ is uniformly tight if for all $\epsilon > 0$ there exists an M s.t. $\sup_n P(|Q_n| > M) < \epsilon$.

Motivation continues

- Now, suppose L is continuous and our parameter of interest is the conditional mean $h(L, A, Y) \equiv \mathbb{E}(Y \mid L)$.
- In particular, to estimate $\mathbb{E}(Y \mid L = l)$ there exists at most one individual l with $L_i = l$ and $\mathbb{E}_n(Y \mid L = l) = Y_i$, regardless of n, and clearly we do not have $\sqrt{n}$-consistency.
- Thus, we have to do something else...

Parameteric modelling

- Can we really say that the distribution that generated the data belongs to a parametric model?
- The answer is **no** in most settings. Therefore many argue that non-parametric methods are more desirable. And this is why machine learning methods are blooming.
- However, it is often argued that studying parametric models is useful (i) because they can be good approximations, (ii) sometimes we have knowledge about the data generating mechanism and (iii) they provide the background for understanding non-parametric methods.

PS: a saturated model, because it does not impose restrictions on the data; we just call it a model because it looks like a model, but the model does not put any restrictions on the data generating mechanism.

What is bias

- Systematic bias: We say there is *systematic bias* if the causal estimand of interest is not identified.
  Informally, any structural association between the treatment and the outcome that does not arise from the causal effect of treatment on the outcome.
- Bias due to model misspecification: When we use a statistical model that is misspecified (I give a formal definition of model mis-specification in a later slide).

Section 20

Lecture 9

Reminder: Maximum Likelihood Estimation (MLE)

Consider a vector $\theta = [\theta_1, \theta_2, \ldots, \theta_k]^\mathsf{T}$ of parameters that indexes the distribution $\{f(\cdot\,;\theta) \mid \theta \in \Theta\}$, where Θ is a parameter space.
We evaluate the observed data sample $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$, which gives us the likelihood,

$$L_n(\theta) = L_n(\theta; \boldsymbol{Y}) = f_n(\boldsymbol{Y}; \theta),$$

where $f_n(\boldsymbol{Y}; \theta)$ is a product of n density functions evaluated at $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$.
MLE maximises the likelihood, i.e.

$$\theta = \arg\max_{\theta \in \Theta} L_n(\theta\,;\boldsymbol{Y}).$$

The logarithm is a monotone function, and thus it is more convenient to maximise the log-likelihood: $\ell(\theta\,;\boldsymbol{Y}) = \log L_n(\theta\,;\boldsymbol{Y})$. If $\ell(\theta\,;\boldsymbol{Y})$ is differentiable in θ, we solve $M(\boldsymbol{Y};\theta) = \frac{\delta \ell(\theta\,;\boldsymbol{Y})}{\delta\theta}$, , i.e. the score equations (also called likelihood equations)

$$p_1 \equiv \frac{\partial \ell}{\partial \theta_1} = 0, \quad \frac{\partial \ell}{\partial \theta_2} = 0, \quad \ldots, \quad \frac{\partial \ell}{\partial \theta_k} = 0.$$

.

*Reminder: Likelihood continues

We need local concavity. Thus, the Hessian matrix

$$\mathbf{H}\left(\widehat{\theta}\right) = \begin{bmatrix} \left.\frac{\partial^2 \ell}{\partial \theta_1^2}\right|_{\theta=\widehat{\theta}} & \left.\frac{\partial^2 \ell}{\partial \theta_1 \, \partial \theta_2}\right|_{\theta=\widehat{\theta}} & \cdots & \left.\frac{\partial^2 \ell}{\partial \theta_1 \, \partial \theta_k}\right|_{\theta=\widehat{\theta}} \\ \left.\frac{\partial^2 \ell}{\partial \theta_2 \, \partial \theta_1}\right|_{\theta=\widehat{\theta}} & \left.\frac{\partial^2 \ell}{\partial \theta_2^2}\right|_{\theta=\widehat{\theta}} & \cdots & \left.\frac{\partial^2 \ell}{\partial \theta_2 \, \partial \theta_k}\right|_{\theta=\widehat{\theta}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial^2 \ell}{\partial \theta_k \, \partial \theta_1}\right|_{\theta=\widehat{\theta}} & \left.\frac{\partial^2 \ell}{\partial \theta_k \, \partial \theta_2}\right|_{\theta=\widehat{\theta}} & \cdots & \left.\frac{\partial^2 \ell}{\partial \theta_k^2}\right|_{\theta=\widehat{\theta}} \end{bmatrix},$$

is negative semi-definite at $\widehat{\theta}$. The Fisher information matrix is defined as $\mathcal{I}(\theta) = \mathbb{E}\left[\mathbf{H}\left(\widehat{\theta}\right)\right]$.

Logistic regression

Suppose $Y \in \{0, 1\}$. Define $\beta = [\beta_1, \beta_2, \ldots, \beta_k]^\mathsf{T}$ as a vector of k parameter and consider a k dimensional covariate $\mathbf{X}$. Then the logistic model is defined as

$$\text{logit}(\mathbb{E}[Y_i \mid \mathbf{X}_i]) = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{\beta}^T \mathbf{X}_i,$$

or, equivalently, we can write that that Y follows a Bernoulli distribution,

$$P(Y_i = y \mid \mathbf{X}_i) = {p_i}^y (1-p_i)^{1-y} = \left(\frac{e^{\boldsymbol{\beta}^T \mathbf{X}_i}}{1+e^{\boldsymbol{\beta}^T \mathbf{X}_i}}\right)^y \left(1 - \frac{e^{\boldsymbol{\beta}^T \mathbf{X}_i}}{1+e^{\boldsymbol{\beta}^T \mathbf{X}_i}}\right)^{1-y}$$

$$= \frac{e^{\boldsymbol{\beta}^T \mathbf{X}_i \cdot y}}{1+e^{\boldsymbol{\beta}^T \mathbf{X}_i}}.$$

Thus the likelihood is $\mathcal{L}(\beta) = \prod_{i=1}^n {p_i}^{Y_i}(1-p_i)^{1-Y_i}$, which can be solved numerically, e.g. solving the score equations (you can derive this from the log-likelihood, take derivatives wrt. β).

$$\sum_{i=1}^n \begin{pmatrix} 1 \\ X_i \end{pmatrix} \left(Y_i - \frac{\exp(\beta^T X_i)}{1+\exp(\beta^T X_i)}\right) = 0.$$

M-estimation, preliminaries

You only need to know the basics of M-estimation. Some of the slides on M-estimation, labelled with *, are additional readings that you do not need to study in detail.

Consider a generic *statistical* model, and suppose we have i.i.d. random vectors $Z_1, \ldots, Z_n$ where $Z \sim \mathbb{P}_Z(z)$ from this model. Let θ be a k dimensional parameter. If θ fully characterizes $\mathbb{P}_Z(z)$, then we write $\mathbb{P}_Z(z;\theta)$. Let θ_0 denote the true value of θ. It follows that if θ fully characterizes $\mathbb{P}_Z(z)$, then the true density is $\mathbb{P}_Z(z;\theta_0)$.
We are considering the (classical) statistical problem of deriving an estimator for θ.

Definition of an M-estimator

Definition (M-estimator)

An M-estimator for θ is the solution $\hat{\theta}$ (assuming that it exists and is well defined) to the $(k \times 1)$ system of estimating equations

$$\sum_{i=1}^{n} M(Z_i; \hat{\theta}) = 0,$$

We say that $M(z; \theta) = \{M_1(z; \theta), \ldots, M_k(z; \theta)\}^T$ is an *unbiased estimating function for* $\mathbb{E}_\theta(M(Z_i; \theta)) = 0$. The expectation is taken wrt. to the distribution of Z at θ. From now on, we will suppress the subscript when we evaluate the expectation in the true value θ_0, i.e. $\mathbb{E}(M(Z_i; \theta)) \equiv \mathbb{E}_{\theta_0}(M(Z_i; \theta))$.

MLE is an M-estimator

Consider a fully parametric model $\mathbb{P}_Z(z;\theta)$. Define,

$$M(z;\theta) = \frac{\delta \log(\mathbb{P}_Z(z;\theta))}{\delta\theta},$$

where the right hand side is a k dimensional vector of derivatives. Solving an estimating equation with this $M(z;\theta)$ yields a maximum likelihood estimator (MLE), and thus the MLE is an M-estimator.

Methods of moment estimators are M-estimators

Consider a fully parametric model $\mathbb{P}_Z(z;\theta)$. Define,

$$M_m(Z_i;\theta) = Z_i^m - \mathbb{E}_\theta(Z_i^m),$$

where $m = 1, \ldots, k$, i.e. k is the dimension of θ.

Overview of properties of M-estimators

`This is for your information, not something we will go through in detail`

Theorem (M-estimator)

Under suitable regularity conditions, $\hat{\theta}$ is a consistent and asymptotically normal estimator,

$$\hat{\theta} \xrightarrow{P} \theta_0$$

and

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} \mathcal{N}(0, \Sigma),$$

where Σ is a covariance matrix.

*Sufficient regularity conditions for M-estimators

`This is for your information, not something we will go through in detail`

Suppose that the following regularity conditions hold.

1. $\sup_{\theta \in \Theta} |M_n(\theta) - M_0(\theta)| \xrightarrow{P} 0$.
2. For all $\epsilon > 0$, $\inf\{|M_0(\theta)| : d(\theta, \theta_0) \geq \epsilon\} > 0 = |M_0(\theta_0)|$.
   For this condition it is sufficient that there exists a unique solution, Θ is compact and M is continuous.
3. $M_n(\hat{\theta}_n) = o_P(1)$.

where $M_n(\theta) = \mathbb{E}_n(M(Z;\theta))$ is the expectation over the empirical distribution and $M_0(\theta) = \mathbb{E}(M(Z;\theta))$ over the true data generating law.

*Proof that the conditions above are sufficient for the consistency of M-estimators

Proof.

From the 2nd condition, for all $\epsilon > 0$ there is a $\delta > 0$ such that

$$
\begin{aligned}
& P(d(\hat{\theta}_n, \theta_0) \geq \epsilon) \\
\leq & P(|M_0(\hat{\theta}_n)| - |M_0(\theta_0)| \geq \delta) \\
= & P(|M_0(\hat{\theta}_n)| - |M_n(\hat{\theta}_n)| + |M_n(\hat{\theta}_n)| - |M_n(\theta_0)| + |M_n(\theta_0)| - |M_0(\theta_0)| \geq \delta) \\
\leq & P(|M_0(\hat{\theta}_n)| - |M_n(\hat{\theta}_n)| \geq \frac{\delta}{3}) + P(|M_n(\hat{\theta}_n)| - |M_n(\theta_0)| \geq \frac{\delta}{3}) + \\
& P(|M_n(\theta_0)| - |M_0(\theta_0)| \geq \frac{\delta}{3}).
\end{aligned}
$$

Condition 1 implies that the first and third probabilities go to zero.
Condition 3 implies that the second goes to zero. □

Plan

- Estimation with a point treatment.
  - Standardisation
  - Propensity methods
- Uncertainty quantification
  - Bootstrap
- Estimation with time-varying treatments

Example: Smoking Cessation A on weight gain Y.

1566 cigarette smokers aged 25-74 years. The outcome weight gain measured after 10 years.

Mean baseline characteristics	A = 1	A = 0
Age, years	46.2	42.8
Men, %	54.6	46.6
White, %	91.1	85.4
University, %	15.4	9.9
Weight, kg	72.4	70.3
Cigarettes/day	18.6	21.2
Years smoking	26.0	24.1
Little exercise, %	40.7	37.9
Inactive life, %	11.2	8.9

Miguel A Hernan and James M Robins. *Causal inference: What if?* CRC Boca Raton, FL:, 2018.

On estimation of causal effects

From slide 72, remember that from an experiment where A is randomised conditional on L, or more generally when consistency, positivity and exchangeability ($Y^a \perp\!\!\!\perp A \mid L$) hold, we have that

$$\begin{aligned}\mathbb{E}(Y^a) &= \sum_l \mathbb{E}(Y \mid L = l, A = a) \Pr(L = l) \\ &= \mathbb{E}\left[\frac{I(A = a)}{\pi(A \mid L)} Y\right].\end{aligned}$$

where $\pi(a \mid l) = P(A = a \mid L = l)$.
This equality motivates different estimators.

Regression estimator

We can also write

$$\begin{aligned}\mathbb{E}(Y^a) &= \sum_l \mathbb{E}(Y \mid L = l, A = a) \Pr(L = l) \\ &= \mathbb{E}(\mathbb{E}(Y \mid L, A = a)),\end{aligned}$$

where you should note that the outer expectation in the second line is with respect to the marginal of L. Denote

$$\mathbb{E}(Y \mid L = l, A = a) = Q(l, a).$$

$Q(l, a)$ is usually unknown, even in an experiment.

Regression estimator

Consider a parametric regression model $Q(I, a; \beta)$ of $Q(I, a)$; that is a linear or nonlinear function of (I, a) and the finite-dimensional parameter β.
We estimate β from the observed data. For example, we could in our conditional randomised trial pose a simple linear model

$$Q(I, a; \beta) = \beta_1 + \beta_2 a + \beta_3^T I,$$

which can be fitted with least squares methods.
If the outcome is binary ($Y \in \{0, 1\}$), we could fit a logistic regression model such as

$$\text{logit}\{Q(I, a; \beta)\} = \beta_1 + \beta_2 a + \beta_3^T I.$$

We can fit the logistic regression models with maximum likelihood estimators.

Definition (Correctly specified model)

A model is correctly specified if there exists a value β_0 such that $Q(I, a; \beta)$ evaluated at β_0 yields the true function $Q(I, a)$.

PS: As in any regression setting, the models we have posited may or may not be correctly specified.

Example continues

- We can estimate the conditional sample mean $\hat{\mathbb{E}}(Y \mid A = 1) = 4.5$ in quitters and $\hat{\mathbb{E}}(Y \mid A = 0) = 2.0$ in non-quitters. More specifically, the difference is

$$\hat{\mathbb{E}}(Y \mid A = 1) - \hat{\mathbb{E}}(Y \mid A = 0) = 2.5 \text{ (95\% CI : } 1.7, 3.4),$$

  but we will not assign a causal interpretation to the estimates.
- Let L include the baseline variables sex (0: male, 1: female), age (in years), race (0: white, 1: other), education (5 categories), intensity and duration of smoking (number of cigarettes per day and years of smoking), physical activity in daily life (3 categories), recreational exercise (3 categories), and weight (in kg).
- Suppose $A \perp\!\!\!\perp Y^a \mid L$.

Standardization: A natural way of estimating counterfactual outcomes

If we knew $Q(l, a)$, a natural way of estimating $\mathbb{E}(Y^a)$ is by the empirical average

$$\frac{1}{n}\sum_{i=1}^{n} Q(L_i, a),$$

motivated by the identification formula expression $\mathbb{E}(\mathbb{E}(Y \mid L, A = a))$. When we do not know $Q(l, a)$, but we assume that our model $Q(L_i, a; \beta)$ is correctly specified, we can use the outcome regression estimator to get the estimator

$$\hat{\mu}_{REG}(a) = \frac{1}{n}\sum_{i=1}^{n} Q(L_i, a; \hat{\beta}).$$

For example, using the linear estimator from the previous slide, we can estimate $\mathbb{E}(Y^{a=1})$ - $\mathbb{E}(Y^{a=0})$ by

$$\frac{1}{n}\sum_{i=1}^{n} Q(L_i, 1; \hat{\beta}) - \frac{1}{n}\sum_{i=1}^{n} Q(L_i, 0; \hat{\beta}) = \hat{\beta}_2,$$

that is, the regression parameter is the causal effect.

More broadly, our causal effects are not equal to regression coefficients

- Whereas the causal effect turned out to be equal to a regression coefficient in the previous slide, regression coefficients are not necessarily equal to our causal effect of interest.
- For example, the coefficients in the logistic regression model

$$\text{logit}\{Q(l, a; \beta)\} = \beta_1 + \beta_2 a + \beta_3^T l.$$

do not necessarily translate to a causal effect of interest.

Standardization (G-computation)

We say that standardization is a plug-in g-formula estimator because it simply replaces the conditional mean outcome in the g-formula by its estimates.

Section 21

Propensity score methods

Matching on the propensity score (intuitive motivation)

- In a homework you will see that, for all a,

$$Y^a \perp\!\!\!\perp A \mid L \implies Y^a \perp\!\!\!\perp A \mid \pi(a \mid L).$$

.

- We could, for each treated individual (i.e. individual with $A = 1$), match this individual with an untreated individual with *similar* propensity score.
- Then crudely compare the mean in the two groups.
- This crude comparison should be fine, but...
- Potential problems
  - What does similar propensity score mean? A conservative approach means that we "waste" data, but a loose approach mean that we compare people with different propensity scores...
  - How many matches should we choose?
  - Do we really get the average treatment effect?

Motivation for inverse probability weighting (IPW)

- We would like to adjust for confounding: imbalance between L's among those who are treated and untreated.
- Suppose that we find a treated subject i, who due to her confounders was *unlikely* to be treated. That is, $\pi(1 \mid L_i)$ is small.
- We *upweight* her, so that she represents herself but also the others like herself (in terms of L) who were unexposed.
- Similarly, we upweight untreated individuals with a small value of $\pi(0 \mid L_i)$.
- Heuristically, we can think about the weighted sample as a pseudopopulation where we observe each individual for each exposure level. In particular, $\pi^*(0 \mid L_i) = \pi^*(1 \mid L_i)$ for all i in the weighted population (which we indicate by the $*$).
- In this pseudopopulation, confounders are balanced between treatment groups, and a crude comparison estimates a causal effect (Intuitively, we get a new DAG for this pseudopopulation, where the arrow from L to A is omitted).

Motivating example

Suppose the counterfactual data are:

Group:	A			B			C		
Response Y^1:	1	1	1	2	2	2	3	3	3
Response Y^0:	0	0	0	1	1	1	2	2	2

and the average treatment effect $\mathbb{E}(Y^{a=1}) - \mathbb{E}(Y^{a=0}) = 1$.
but we observe:

Group:	A			B			C		
Response Y^1:	1	1	?	?	2	?	3	?	?
Response Y^0:	?	?	0	1	?	1	?	2	2

The naive contrast $\mathbb{E}(Y \mid A = 1) - \mathbb{E}(Y \mid A = 0) = \frac{7}{4} - \frac{6}{5} = 0.55$.

Example from Oliver Dukes.

Example continues

- However, from the table we see that,

$$\hat{\pi}(1, \text{group A}) = \frac{2}{3}$$
$$\hat{\pi}(1, \text{group B}) = \frac{1}{3}$$
$$\hat{\pi}(1, \text{group C}) = \frac{1}{3}$$

- Let us estimate $\mathbb{E}(Y^{a=1})$ by a weighted average, where each observation is weighted by $\frac{1}{\hat{\pi}(1,\text{group X})}$,Group X $\in \{\text{Group A}, \text{Group B}, \text{Group C}\}$,

$$\frac{(1+1)\frac{3}{2} + 2\frac{3}{1} + 3\frac{3}{1}}{\frac{3}{2} + \frac{3}{2} + \frac{3}{1} + \frac{3}{1}} = 2$$

and estimate $\mathbb{E}(Y^{a=0})$ by weighting each observation by $\frac{1}{\hat{\pi}(0,\text{Group X})}$, Group X $\in \{\text{Group A}, \text{Group B}, \text{Group C}\}$,

$$\frac{0\frac{3}{1} + (1+1)\frac{3}{2} + (2+2)\frac{3}{2}}{\frac{3}{1} + \frac{3}{2} + \frac{3}{2} + \frac{3}{2} + \frac{3}{2}} = 1.$$

Estimation when the propensity score is known

When $\pi(a \mid l)$ is a known function, the estimator of $\mathbb{E}(Y^a)$ is

$$\hat{\mu}_{IPW}(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_i = a) Y_i}{\pi(A_i \mid L_i)}.$$

The propensity score $\pi(a \mid l)$, unlike the function $Q(l, a)$, is known in randomised experiments (it is determined by the investigator). However, in most observational data settings, it is unknown.
PS: This estimator has been known for a long time and is often called the Horvitz Thompson estimator in survey sampling[39].

[39] Daniel G Horvitz and Donovan J Thompson. "A generalization of sampling without replacement from a finite universe". In: *Journal of the American statistical Association* 47.260 (1952), pp. 663–685.

Estimation when the propensity score is unknown

More generally, we can propose a regression model $\pi(A \mid L; \gamma)$ for $\pi(A \mid L)$, and we can consider the estimator

$$\hat{\mu}_{IPW}(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_i = a) Y_i}{\pi(A_i \mid L_i; \gamma)}.$$

For example, suppose that we fit a logistic regression model and find the MLE $\hat{\gamma}$ of γ, which is the solution to the estimating equation (See slide 242)

$$\sum_{i=1}^{n} \begin{pmatrix} 1 \\ L_i \end{pmatrix} \left(A_i - \frac{\exp(\gamma_1 + \gamma_2^T L_i)}{1 + \exp(\gamma_1 + \gamma_2^T L_i)} \right) = 0.$$

Section 22

Lecture 10

Marginal Structural Models

- We have learned that a *statistical* models puts restrictions on laws, that is, it puts restriction on conditional distributions (densities).
- Thus, the statistical model puts restrictions on the observed data distributions.
- A causal model puts restrictions on *counterfactual* densities, e.g. based on independence ($\perp\!\!\!\perp$) restrictions.
- We can make a causal (structural) model parametric by imposing parametric models for counterfactuals. Examples of such models are marginal structural models. Note that these models cannot be fitted directly to the data, because we don't directly observe the counterfactuals (see next slide)

Marginal structural models

An alternative way of weighting by the propensity scores is to define a so-called marginal structural model, which is a *statistical* model that parameterizes a functional of a *marginal* counterfactual Y^a (not the *joint* counterfactual $(Y^{a=1}, Y^{a=0})$).

- An example of a marginal structural model is

$$\mathbb{E}(Y^a) = \eta_0 + \eta_1 a.$$

- This model is saturated[40] for a binary A and implies that

$$\begin{aligned} \mathbb{E}(Y^0) &= \eta_0 \\ \mathbb{E}(Y^1) &= \eta_0 + \eta_1 \\ \mathbb{E}(Y^1) - \mathbb{E}(Y^0) &= \eta_1 \end{aligned}$$

- You can think about this as a regression model that is fitted to a (pseudo)population where A is randomly assigned.

[40] it does not impose restrictions on the data.

Estimator in marginal structural model

The estimator in a marginal structural model will look like

$$\hat{\mu}_{MSM}(a) = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i=a)Y_i}{\pi(A_i|L_i;\gamma)}}{\frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i=a)}{\pi(A_i|L_i;\gamma)}}.$$

I have omitted a proof.

PS: you can also try to show that, under our identifiability assumptions, $\hat{\mu}_{MSM}(a)$ is a consistent estimator of $\mathbb{E}(Y^a)$ by using results for weighted least square regressions. Both $\hat{\mu}_{IPW}(a)$ and $\hat{\mu}_{MSM}(a)$ are consistent. If Y is binary, only $\hat{\mu}_{MSM}(a)$ ensures that the estimate of $\mathbb{E}(Y^a)$ is in $[0, 1]$.

Further intuition on inverse probability weighting

- We can think of IPTW as creating an imaginary pseudopopulation in which there is no confounding: informally, we have a population where each individual i is represented by themselves and $w_i - 1$ other individuals, where w_i is the weight of individual i.
    - More formally, we consider a new law defined by a likelihood ratio (see next slide)
- Indeed, this is the way many applied researchers (including applied statisticians) think about this way of modelling. Formally, we do not need the concept of a pseudopopulation, but it is sometimes a useful motivation for the math and gives us some direction to come up with solutions.
- To be explicit, let us use the subscript "ps" to denote probability and expectation in the pseudopopulation (P_{ps} and $\mathbb{E}_{ps}$), while P and $\mathbb{E}$ without subscripts refer to the actual population. Consider the observed data $(Y\overline{A}, \overline{L})$.

Estimation when the propensity is unknown

Define $\theta = (\mu, \gamma^T)^T$, and solve the stacked estimating equations

$$\sum_{i=1}^{n} \left(\frac{I(A_i = a) Y_i}{\pi(A_i \mid L_i; \gamma)} - \mu \right) = 0$$

$$\sum_{i=1}^{n} \begin{pmatrix} 1 \\ L_i \end{pmatrix} \left(A_i - \frac{\exp(\gamma_1 + \gamma_2^T L_i)}{1 + \exp(\gamma_1 + \gamma_2^T L_i)} \right) = 0,$$

The solution $\hat{\mu}_{IPW}$ to this system is an M-estimator, and therefore it is consistent (under our regularity conditions). We can use M-estimator theory to argue that the estimator is asymptotically normal.
In the next slide, we will study an interesting special case.

Outcome prediction for predictive purposes

Outcome regression is often used for purely *predictive* purposes.

- Online stores would like to predict which customers are more likely to purchase their products. The goal is not to determine whether your age, sex, income, geographic origin, and previous purchases have a causal effect on your current purchase. Rather, the goal is to identify those customers who are more likely to make a purchase so that specific marketing programs can be targeted to them. It is all about association, not causation. Similarly, doctors use algorithms based on outcome regression to identify patients at high risk of developing a serious disease or dying.
- A study found that Facebook Likes predict sexual orientation, political views, and personality traits (Kosinski et al, 2013). Low intelligence was predicted by, among other things, a "Harley Davidson" Like. This is purely predictive, not necessarily causal.

From Hernan and Robins, *Causal inference: What if?*

Prediction and procedures for model selection

- Model selection is a different endeavour when the aim is prediction.
- Investigators who seek to do pure predictions may want to include any variables that, when used as covariates in the model, improve its predictive ability.
- This motivates the use of selection procedures, such as forward selection, backward elimination, stepwise selection and new developments in machine learning.
- However, using these procedures for causal inference tasks can be unnecessary and harmful. Both bias and inflated variance may be the result.
- For example, we do not fit a propensity score model to predict the treatment A as good as possible: we just fit the model to guarantee exchangeability. Indeed, covariates that strongly associated with treatment, but are not necessary to guarantee exchangeability, do not reduce bias. Adjustment for these variables can lead to larger variance...

Example: The IPW estimator and variance

Suppose that we are in the randomised experiment, such that γ is known: let $P(A = 1 \mid L) = 0.5$, so $A \perp\!\!\!\perp L$. Suppose also that we adapt the correctly specified model $\pi(1 \mid I; \gamma) = \gamma$. In particular, the truth is $\gamma_0 = 0.5$.

Statistician 1 suggests using the true value $\gamma_0 = 0.5$ because it is known.

Statistician 2 suggests using the MLE $\pi(1 \mid I; \hat{\gamma}) = \hat{\gamma} = \frac{1}{n} \sum_{i=1}^{n} A_i$.

Who selected the most efficient estimator?

*Statistician 1

The estimator for $\mu_1 = \mathbb{E}(Y^{a=1}) = \mathbb{E}(Y \mid A = 1)$ is

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{A_i Y_i}{\gamma_0} = \sum_{i=1}^{n}\frac{A_i Y_i}{n/2}$$

$\hat{\mu}_1$ is consistent because $\mathbb{E}(YA) = \mathbb{E}(A\mathbb{E}(Y \mid A)) = \frac{\mu_1}{2}$ and thus $n^{-1}\sum_{i=1}^{n} A_i Y_i \xrightarrow{P} \frac{\mu_1}{2}$. After some algebra,

$$\sqrt{n}(\hat{\mu}_1 - \mu_i) = 2n^{-1/2}\sum_{i=1}^{n}(A_i Y_i - \mu_1/2).$$

Define $\sigma_1^2 = var(Y \mid A = 1)$,

$$var(AY) = \mathbb{E}(var(AY \mid A)) + var(\mathbb{E}(AY \mid A)) \tag{8}$$

$$= \mathbb{E}(A\sigma_1^2) + var(A\mu_1) = \frac{\sigma_1^2}{2} + \frac{\mu_1^2}{4}. \tag{9}$$

CLT: $\sqrt{n}(\hat{\mu}_1 - \mu_i) \xrightarrow{D} \mathcal{N}(0, 2\sigma_1^2 + \mu_1^2)$.

*Statistician 2

The estimator for $\mu_1 = \mathbb{E}(Y^{a=1}) = \mathbb{E}(Y \mid A = 1)$ is

$$\hat{\mu}_1^* = \frac{1}{n} \sum_{i=1}^{n} \frac{A_i Y_i}{\hat{\gamma}} = \frac{\sum_{i=1}^{n} A_i Y_i}{\sum_{i=1}^{n} A_i}.$$

Indeed, $\hat{\mu}_1^*$ is consistent, $\mathbb{E}(YA) = \mathbb{E}(A\mathbb{E}(Y \mid A)) = \frac{\mu_1}{2}$, so that $n^{-1} \sum_{i=1}^{n} A_i Y_i \xrightarrow{P} \frac{\mu_1}{2}$ and $n^{-1} \sum_{i=1}^{n} A_i \xrightarrow{P} \frac{1}{2}$
After some algebra,

$$\sqrt{n}(\hat{\mu}_1^* - \mu_i) = \frac{n^{-1/2} \sum_{i=1}^{n} A_i (Y_i - \mu_1)}{n^{-1} \sum_{i=1}^{n} A_i}$$

$$var(A(Y - \mu_1)) = \mathbb{E}(Avar(Y - \mu_1) \mid A) + var(A\mathbb{E}(Y - \mu_1) \mid A) = \frac{\sigma_1^2}{2} + 0$$

CLT and Slutsky's theorem : $\sqrt{n}(\hat{\mu}_1^* - \mu_i) \xrightarrow{D} \mathcal{N}(0, 2\sigma_1^2)$.

Interesting insight from Statistician 1 vs Statistician 2

- It is more efficient to estimate the propensity score, even if the true propensity is known. (This is a more general result; not just a special case we have considered here.)
- Does this contradict what we know from MLE theory, where including more known information, leads to lower variance? **No**, this is not a contradiction because the IPW estimator is not an MLE for μ_1.

Standard error and variance for IPW estimators

- We can sometimes obtain variance estimators from M-estimator theory.
- However, I do suggest using the bootstrap for the settings we consider here (see next slide for a brief introduction to bootstrap).
  - Computer intensive but convenient.
  - Simple in practice, but rigorous theory behind

*On the variance of M-estimators

Under regularity conditions, the asymptotic properties of an M-estimator $\hat{\theta}$ can be derived from Taylor series approximations, the law of large numbers, and the central limit theorem. Here is a brief outline.

- Let θ_0 and $\dot{M}(Z_i, \theta) = \partial M(Z_i, \theta)/\partial \theta^\intercal$ (This is a $k \times k$ matrix).
- $C(\theta_0) = E[-\dot{M}(Z_i, \theta_0)]$, and
- $B(\theta_0) = E[M(Z_i, \theta_0)M(Z_i, \theta_0)^\intercal]$. Then under suitable regularity assumptions, $\hat{\theta}$ is consistent and asymptotically Normal, i.e.,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Sigma(\theta_0)) \text{ as } n \to \infty,$$

  where $\Sigma(\theta_0) = C(\theta_0)^{-1} B(\theta_0) \{C(\theta_0)^{-1}\}^\intercal$.
- This can be seen by a first-order Taylor series expansion of each row of the estimating equation $\sum_{i=1}^n M(Z_i; \hat{\theta}) = 0$ in $\hat{\theta}$ about θ_0,

$$0 = \sum_{i=1}^n M(Z_i; \theta_0) + \sum_{i=1}^n \left[\dot{M}(Z_i, \theta^*)\right](\hat{\theta} - \theta_0),$$

  where θ^* is a value between $\hat{\theta}$ and θ_0.

*Variance continues

- The sandwich form of $\Sigma(\theta_0)$ suggests several possible large sample variance estimators.
- For some problems, the analytic form of $\Sigma(\theta_0)$ can be derived and estimators of θ_0 and other unknowns simply plugged into $\Sigma(\theta_0)$.
- Alternatively, $\Sigma(\theta_0)$ can be consistently estimated by the empirical sandwich variance estimator, where the expectations in $C(\theta)$ and $B(\theta)$ are replaced with their empirical counterparts.
- Let $C_i = -\dot{M}(Z_i, \theta)|_{\theta=\hat{\theta}}$, $C_n = n^{-1}\sum_{i=1}^n C_i$, $B_i = M(Z_i, \hat{\theta})M(Z_i, \hat{\theta})^\intercal$, and $B_n = n^{-1}\sum_{i=1}^n B_i$. The empirical sandwich estimator of the variance of $\hat{\theta}$ is:

$$\hat{\Sigma} = C_n^{-1} B_n \{C_n^{-1}\}^\intercal / n.$$

Bootstrap

Bootstrap is a method for estimating the variance of a parameter.
Let $U_n = g(X_1, \ldots, X_n)$ be a statistic, i.e. a function of data. For example, $\hat{\mu}_{IPW}(a) = \frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i=a)Y_i}{\pi(A_i|L_i;\hat{\gamma})}$, where in this case $X_i = (L_i, A_i, Y_i)$.
We want to estimate $VAR(U_n)$, and the bootstrap is motivated by two steps

1. Estimate $VAR(U_n)$ by $VAR_{\hat{\mathbb{P}}_n}(U_n)$, where $\hat{\mathbb{P}}_n$ is the empirical distribution.
2. Approximate $VAR_{\hat{\mathbb{P}}_n}(U_n)$ using simulations.

Step 2 is very useful when it is hard to express the closed form solution to the variance of U_n. Bootstrap variance estimation is done as follows:

1. Draw $X_1^*, \ldots, X_n^* \sim \hat{\mathbb{P}}_n$. (Sample with replacement from $(X_1, \ldots, X_n)$)
2. Compute $U_n^* = g(X_1^*, \ldots, X_n^*)$.
3. Repeat step 1 and 2 K times to get $U_{n,1}^*, U_{n,2}^*, \ldots, U_{n,K}^*$.[41]
4. $v_{\text{boot}} = \frac{1}{K}\sum_{k=1}^{K}\left(U_{n,k}^* - \frac{1}{K}\sum_{l=1}^{K} U_{n,l}^*\right)^2$

[41] Usually ≥ 1000 times.

Bootstrap

- Bootstrap is based on two approximations

$$VAR(U_n) \approx VAR_{\hat{\mathbb{P}}_n}(U_n) \approx v_{\text{boot}}.$$

- Bootstrap is very useful in practice and simple to implement: You just draw $X_1^*, \ldots, X_n^*$ with replacement from $(X_1, \ldots, X_n)$.

Bootstrap confidence intervals

Bootstrap confidence intervals can be created in several ways.

1. The normal intervals: $U_n \pm \eta_{\alpha/2}\hat{\text{se}}_{boot}$, $\sqrt{v_{\text{boot}}} = \hat{\text{se}}_{boot}$, where $\eta_{\alpha/2}$ is the $\alpha/2$ quantile of a standard normal variable. this requires U_n to be close to normal.
2. Percentile intervals: Define the interval $C_n = (U^*_{\eta/2}, U^*_{1-\eta/2})$, where U^*_ρ is the ρ sample quantile of $(U^*_{n,1}, U^*_{n,2}, \ldots, U^*_{n,K})$.
3. Studentised pivot intervals: Often perform better. A pivot is a random variable whose distribution does not depend on unknowns.

There are also many other ways of obtaining bootstrap confidence intervals. One high-level disclaimer: The bootstrap can, under certain data generating mechanisms, fail. If we have i.i.d. data an we study functionals that are reasonably smooth, which we study in the course the bootstrap will usually work. We will not consider violations in depth here.
For a detailed theory on the bootstrap, see Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. 1. Cambridge university press, 1997

Section 23

Doubly robust estimators

Precision and IPW

- IPW estimators are often considered to be inefficient, that is, to have low precision.
- In principle, we can give two reasons why:
  - They give a more appropriate ("honest") reflection of the uncertainty, because they do not rely on implausible model assumptions.
  - They are truly inefficient, and we could impose the same model assumptions, and obtain a more efficient estimator.
- Asymptotic results from *semi-parametric efficiency theory* suggest that both these explanations can be true. We will not go into the details of semiparametric estimation theory, but we will show properties in some interesting examples.

Doubly robustness

- Natural way is to combine both regression and inverse probability weighting.
- Give a full factorization and see which terms are estimated in IPW and regression modelling.

Definition (Doubly robust estimator)

An estimator $\hat{\mu}$ of a parameter μ is doubly robust if it is a consistent estimator for μ if either of two models are correctly specified (e.g., the propensity model or the outcome regression model is correctly specified), but not necessarily both models are correctly specified.

Doubly robust estimator

Theorem (Doubly robust estimator of $\mathbb{E}(Y \mid L, A = a)$)

If either the propensity model $\pi(a \mid l; \gamma)$ or the outcome regression model $Q(l, a; \beta)$ is correctly specified, then

$$\mathbb{E}\left[\frac{I(A=a)Y}{\pi(a \mid L;\gamma)} + \left(1 - \frac{I(A=a)}{\pi(a \mid L;\gamma)}\right) Q(L, a; \beta)\right] = \mathbb{E}[\mathbb{E}(Y \mid L, A = a)].$$

Intuitively, the doubly robust estimator – unlike the simple inverse probability weighted estimator – exploits information from both treated and untreated. PS: note that we can re-write the expression in the theorem,

$$\begin{aligned}
&\mathbb{E}\left[\frac{I(A=a)Y}{\pi(a \mid L;\gamma)} + \left(1 - \frac{I(A=a)}{\pi(a \mid L;\gamma)}\right) Q(L, a; \beta)\right] \\
=&\mathbb{E}\left[Q(L, a; \beta) + \frac{I(A=a)}{\pi(a \mid L;\gamma)}\{Y - Q(L, a; \beta)\}\right]
\end{aligned}$$

Proof

Proof.

Suppose first that $\pi(a \mid l; \gamma)$ is correctly specified, but the outcome model $Q(l, a; \beta)$ is misspecified. Use iterative expectation,

$$
\begin{aligned}
\mathbb{E}\left\{\frac{I(A=a)Y}{\pi(a \mid L;\gamma)}\right\} =& \mathbb{E}\left\{\frac{I(A=a)}{\pi(a \mid L;\gamma)}E(Y \mid L, A)\right\} \\
=& \mathbb{E}\left\{\frac{I(A=a)}{\pi(a \mid L;\gamma)}E(Y \mid L, A=a)\right\} \\
=& \mathbb{E}\left\{\frac{\mathbb{E}(I(A=a) \mid L)}{\pi(a \mid L;\gamma)}E(Y \mid L, A=a)\right\} \\
=& \mathbb{E}\left\{\frac{(\pi(a \mid L)}{\pi(a \mid L;\gamma)}E(Y \mid L, A=a)\right\} \\
=& \mathbb{E}\left\{\mathbb{E}(Y \mid L, A=a)\right\}.
\end{aligned}
$$

Proof continues

Proof.

Next, consider the second term

$$\begin{aligned}
\mathbb{E}\left\{\left(1-\frac{I(A=a)}{\pi(a\mid L;\gamma)}\right)Q(L,a;\beta)\right\} =& \mathbb{E}\left\{\mathbb{E}\left[\left(1-\frac{I(A=a)}{\pi(a\mid L;\gamma)}\right)Q(L,a;\beta)\mid L\right]\right\} \\
=& \mathbb{E}\left\{\mathbb{E}\left(1-\frac{\mathbb{E}(I(A=a)\mid L)}{\pi(a\mid L;\gamma)}\right)Q(L,a;\beta)\right\} \\
=& \mathbb{E}\left\{(1-1)Q(L,a;\beta)\right\}=0.
\end{aligned}$$

Proof continues (note: no reference to counterfactuals)

Proof.

Suppose now that $\pi(a \mid I; \gamma)$ is mis-specified, but the outcome model $Q(I, a; \beta)$ is correctly specified. After some algebra,

$$\begin{aligned}
&\mathbb{E}\left[\frac{I(A=a)Y}{\pi(a \mid L;\gamma)} + \left(1 - \frac{I(A=a)}{\pi(a \mid L;\gamma)}\right) Q(L, a; \beta)\right] \\
=&\mathbb{E}\left[Q(L, a; \beta) + \frac{I(A=a)}{\pi(a \mid L;\gamma)}\{Y - Q(L, a; \beta)\}\right]
\end{aligned}$$

Due to the correct specification, we know that the first term $\mathbb{E}[Q(L, a; \beta)] = \mathbb{E}[\mathbb{E}(Y \mid L, A = a)]$. Furthermore, using iterative expectation on the second term (conditional on L, similar to part 1 of the proof)

$$\begin{aligned}
&\mathbb{E}\left[\frac{I(A=a)}{\pi(a \mid L;\gamma)}\{Y - Q(L, a; \beta)\}\right] \\
=&\mathbb{E}\left[\frac{E(I(A=a) \mid L)}{\pi(a \mid L;\gamma)}\{E(Y \mid L, A=a) - Q(L, a; \beta)\}\right] = 0.
\end{aligned}$$

Some practical thoughts on estimation

- If we cannot guarantee that our model is correctly specified, we should in principle try to use different estimators (In practice it can be difficult).
- If all estimators give similar results, then there is some evidence (but not a guarantee!!) that we have modelled the problem correctly.
- If the estimators do not give the same results, try to understand why...
- In practice some degree of misspecification is inescapable in all models, and model misspecification will introduce some bias. But the misspecification of the treatment model (IP weighting) and the outcome model (standardization) will not generally result in the same magnitude and direction of bias in the effect estimate. Therefore the IP weighted estimate will generally differ from the standardised estimate because unavoidable model misspecification will affect the point estimates differently.
- The main advantage of doubly robust estimators is that they can have small bias, even when $Q(l, a)$ and $\pi(a \mid l)$ are estimated with machine learning methods. This has to do with the fact that the bias of the doubly robust estimator is a product of the errors in estimating $Q(l, a)$ and $\frac{1}{\pi(a|l)}$.

Section 24

Lecture 11

Section 25

Time-varying treatments

Extension: time-varying treatments

- When the identification conditions hold, we target the g-formula, here for a static regime,

$$\mathbb{E}(Y^{\bar{a}}) = \sum_{y} y b_{\bar{a}}(y) = \sum_{\bar{l}_K} \mathbb{E}(Y \mid \bar{l}_K, \bar{a}_K) \prod_{j=0}^{K} p(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1}),$$

  see slide 180.
- We have considered the case with $K = 1$, and we either modelled
  - the outcome mean (parametric g-formula, also called standardization)
  - IPW
- This can be generalized to any K

G-formula for time-varying treatment

- Suppose we target:

$$\mathbb{E}(Y^{\bar{a}}) = \sum_{y} y b_{\bar{a}}(y) = \sum_{\bar{l}_K} \mathbb{E}(Y \mid \bar{l}_K, \bar{a}_K) \prod_{j=0}^{K} p(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1}),$$

- The generalization of standardisation, which is often called the parametric g-formula or g-computation (*but not* g-estimation) is to model $p(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1})$ for all $j \leq K$ and $p(y \mid \bar{l}_K, \bar{a}_K)$.
- The practical problem is that $p(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1})$ migh be densities, and density estimation is much harder than mean estimation.
- Indeed, the sums, or in the continuous case integrals, are intractable for large k

G-formula algorithm

Given n individuals with observed variables $\overline{A}_K, \overline{L}_K, Y$.

1. Assume statistical models for
   - $\mathbb{E}(Y \mid \overline{l}_K, \overline{a}_K; \beta)$, and
   - $p(l_j \mid \overline{l}_{j-1}, \overline{a}_{j-1}; \alpha_j)$.
2. Fit each models by MLE, which would give us $\hat{\beta}$ and $\hat{\alpha}_j$ for all $j \leq K$.
3. Obtain an estimate of $\sum_y y b_{\overline{a}}(y)$ by
   - for each individual i and time j, sequentially sample r draws, where a draw m_i for individual i is
   - Sample $L_{j,m_i} \sim p(L_j \mid L_{j-1,m_i}, L_{j-2,m_i}, \ldots, L_{0,i}, \overline{a}_{j-1}; \hat{\alpha}_j)$
   - Compute $Y_{m_i} \equiv \mathbb{E}(Y \mid L_{K,m_i}, L_{K-1,m_i}, \ldots, L_{0,i}, \overline{a}_K; \hat{\beta})$
   - return
$$\frac{1}{nr} \sum_{i=1}^{n} \sum_{m_i=1}^{r} Y_{m_i}$$
4. Give bootstrap confidence intervals.
   Note that this means sampling twice: one time to evaluate the evaluate the big sum in the "return" statement, second time to get confidence intervals.

Pros/Cons of g-computationo (from Shpitser)

Positives:

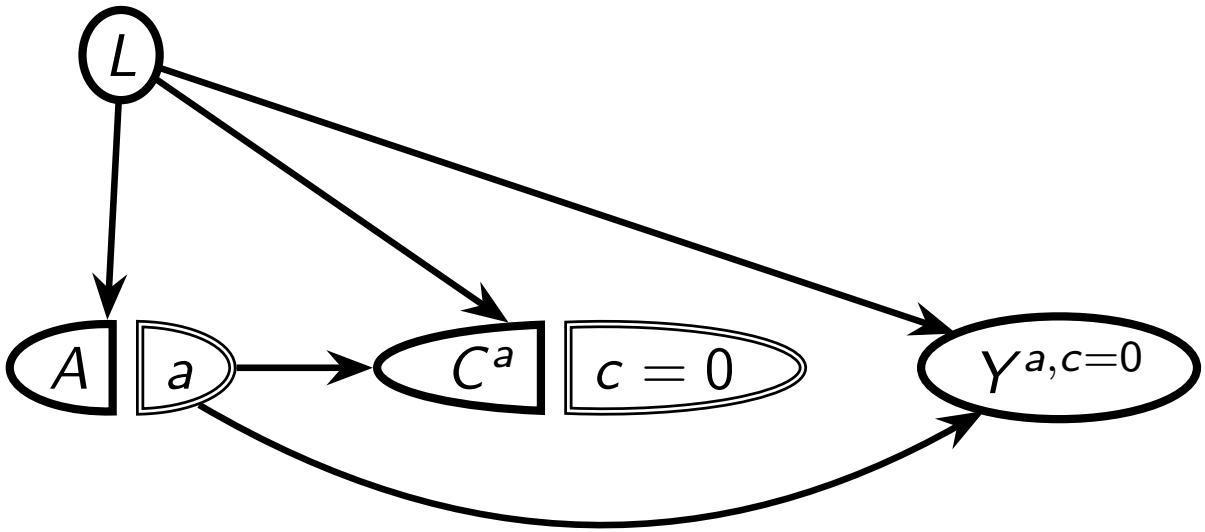
- Efficient if the models are correctly specified.
- In practice, people have reported that the approach is fairly robust to misspecification in practice.
- Conceptually, this is simple.

Negatives:

- Have to do a lot of (parametric) modeling, which means a risk of model misspecification.
- In general settings, this could be intractable and very slow.
- Sampling is computationally intensive
- Sampling trajectories can be unstable.

Study on weight gain continues

Slightly extended graph



Example: Smoking Cessation A on weight gain Y.

1566 cigarette smokers aged 25-74 years. The outcome weight gain measured after 10 years.

Mean baseline characteristics	A = 1	A = 0
Age, years	46.2	42.8
Men, %	54.6	46.6
White, %	91.1	85.4
University, %	15.4	9.9
Weight, kg	72.4	70.3
Cigarettes/day	18.6	21.2
Years smoking	26.0	24.1
Little exercise, %	40.7	37.9
Inactive life, %	11.2	8.9

Hernan and Robins, *Causal inference: What if?*

Example; Censoring: weight gain study continues

- Suppose that there were 63 additional individuals who met our eligibility criteria but were excluded from the analysis because their weight in 1982 was not known. That is, their outcome was *censored*.
- Excluding the censored individuals will lead to selection bias due to conditioning on a collider.
- Then, the naive estimate can be correctly described as

$$\hat{\mathbb{E}}(Y \mid A = 1, C = 0) - \hat{\mathbb{E}}(Y \mid A = 0, C = 0) = 2.5 \text{ (95\% CI : } 1.7, 3.4),$$

- On the other hand, the causal effect of interest is

$$\hat{\mathbb{E}}(Y^{a=1,c=0}) - \hat{\mathbb{E}}(Y^{a=0,c=0})$$

- We derived an identification formula $E[Y^{a,c=0}] = \sum_l E[Y \mid A = a, C = 0, L = l]P(L = l)$, that motivates a -formula estimator, see the next slide.

Estimation using the g-formula in the smoking example

We can estimate $\hat{\mathbb{E}}(Y^{a,c=0})$ by a plug-in g-formula estimator or a parametric g-formula estimator,

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbb{E}}(Y \mid A = a, C = 0, L_i)$$

where $\hat{\mathbb{E}}(Y \mid A = a, C = 0, L_i)$ is a regression model, like $Q(l, a; \beta)$ which is fitted to those who are uncensored $(C = 0)$.

- Suppose that included a product term between smoking cessation A and intensity of smoking, but otherwise only main terms. This implies that our model imposes the restriction that each covariate's contribution to the mean is independent of that of the other covariates, except that the contribution of smoking cessation varies linearly.
- If we were interested in the average causal effect in a particular subset of the population, say characterised by V, we could have restricted our calculations to that subset.

Section 26

More on IPW

Censoring: weight gain study continues with IPW

- We can consider an IPW estimator in the presence of censoring
- We multiply the original IPW weight with an inverse probability of censoring weight,

$$\pi_C(c, a, l) \equiv P(C = c \mid A = a, L = l).$$

  The proof that this work is essentially identical to the proof that IPW weighting works. Just replace $\pi(a, l)$ in the original proof with the product $\pi(a, l)\pi_C(0, a, l) = P(A = a, C = 0 \mid L = l)$.
- Explicitly,

$$\hat{\mu}_{IPW}(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_i = a, C_i = 0) Y_i}{\pi(A_i \mid L_i; \gamma_1)\pi_C(0, a, L_i; \gamma_2)}.$$

- How would you obtain an estimate of $\pi_C(0, a, l)$?

General positivity definition

Here is a more general definition of positivity that I include for your reference. The function $g_{j_l}(\cdot)$ is the function that gives a value to a_{j_l} under the counterfactual regime g of interest.

Definition (Positivity)

for each $k \in \{0, \ldots, K\}$, suppose

$$p(v_{j_k} \mid \overline{v}_{j_k-1}) > 0 \; \forall \; \overline{v}_{j_k} \text{ s.t.}$$
$$p(\overline{v}_{j_k-1}) > 0 \text{ and } \overline{v}_{j_l} = g_{j_l}(\overline{v}_{j_l-1}), l = 1, \ldots, k.$$

The intuition is that covariates that will have positive probability in the counterfactual world must also have positive probability in the observed world. Otherwise, we cannot identify outcomes in the counterfactual world from the observed data distributions.

IPW more explicitly

- We define the law $P_{ps}(Y = y, \overline{A}_K = \overline{a}_K, \overline{L}_K = \overline{l}_K)$ by the likelihood ratio

$$\frac{p_{ps}(Y, \overline{A}_K, \overline{L}_K)}{p(Y, \overline{A}_K, \overline{L}_K)} = \frac{g(\overline{A}_K)}{\prod_{k=0}^{K} p(A_k \mid \overline{L}_k, \overline{A}_{k-1})},$$

  where we sometimes use the short hand notation $\overline{A}_K = \overline{A}$ and $\overline{L}_K = \overline{L}$. Thus
  - $g(\overline{A}) = \prod_{k=0}^{K} p_{ps}(A_k \mid \overline{L}_k, \overline{A}_{k-1})$,
  - $p(Y \mid \overline{L}_K, \overline{A}_K) = p_{ps}(Y \mid \overline{L}_K, \overline{A}_K)$
  - $\prod_{j=0}^{K} p(L_j \mid \overline{L}_{j-1}, \overline{A}_{j-1}) = \prod_{j=0}^{K} p_{ps}(L_j \mid \overline{L}_{j-1}, \overline{A}_{j-1})$

  That is, most of the conditional densities are identical in the pseudopopulation and the observed population, and, importantly, $g(\overline{A})$ is not a function of L
- Intuitively, We can think of IPTW as a procedure to cut the arrows (in a DAG) from the covariate history ($\overline{L}_k$) into treatment (A_k). Indeed, *many* applied researchers like this heuristic way of thinking about the problems.

IPW continues: 2 features

Now we state two features of IPW.

Feature 1:

- When using unstabilised weights,
  $P_{ps}(A_k = a_k \mid \overline{A}_{k-1} = \overline{a}_{k-1}, \overline{L}_k = \overline{l}_k) = 0.5$.
  In the pseudopopulation, we have that

$$(A_k \perp\!\!\!\perp \overline{A}_{k-1}, \overline{L}_k)_{ps}$$

- When using stabilised weights,
  $P_{ps}(A_k = a_k \mid \overline{A}_{k-1} = \overline{a}_{k-1}, \overline{L}_k = \overline{l}_k) = P(A_k = a_k)$.
  In the pseudopopulation, we have that

$$(A_k \perp\!\!\!\perp \overline{L}_k \mid \overline{A}_{k-1})_{ps}.$$

`PS: A pseudopopulation is defined differently than a counterfactual population, but the results in the next slide shows how they are related.`

IPW continues: 2 important features

Feature 2:
Suppose that exchangeability, positivity and consistency hold. Then, IPW creates a pseudopopulation characterised by the following:

- Regardless of whether we use unstabilised or stabilised weights or not,

$$\mathbb{E}(Y^{\bar{a}}) = \mathbb{E}_{ps}(Y^{\bar{a}}) = \mathbb{E}_{ps}(Y \mid \overline{A} = \bar{a}).$$

- Thus, the average causal effect is equal to association in the pseudopopulation, and we have that

$$\mathbb{E}(Y^{\bar{a}}) - \mathbb{E}(Y^{\bar{a}'}) = \mathbb{E}_{ps}(Y \mid \overline{A} = \bar{a}) - \mathbb{E}_{ps}(Y \mid \overline{A} = \bar{a}').$$

IPW theorem

We will give a theorem that shows feature 2 [42]:
Remember that the g-formula for the *marginal* of $Y \equiv Y_K$ under treatment assignment $\bar{a} \equiv \bar{a}_K = (a_0, \ldots, a_K)$ is defined as

$$b_{\bar{a}}(y) = \sum_{\bar{l}_K} p(y \mid \bar{l}_K, \bar{a}_K) \prod_{j=0}^{K} p(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1}).$$

Theorem (IPW theorem)

Under positivity,

$$\int y b_{\bar{a}}(y) dy = \mathbb{E}_{ps}(Y \mid \overline{A} = \bar{a}).$$

You will see that the theorem is very similar to other IPW results we have already shown.

[42] Feature 1 follows from some of the steps in the proof of feature 2, but I haven't written out all the details here

*Lemma of IPW theorem

Lemma

If the weights take the form

$$\frac{g(\overline{A})}{\prod_{k=0}^{K} p(A_k \mid \overline{L}_k, \overline{A}_{k-1})},$$

then

$$b_{\overline{a}}(y) = \frac{1}{g(\overline{a})} \mathbb{E}\left\{ \frac{g(\overline{A}) I(\overline{A} = \overline{a})}{\prod_{k=0}^{K} p(A_k \mid \overline{L}_k, \overline{A}_{k-1})} p(y \mid \overline{L}_K, \overline{A}_K) \right\}.$$

Proof.

$$
\begin{aligned}
& b_{\bar{a}}(y) \\
&= \sum_{\bar{l}_K} p(y \mid \bar{l}_K, \bar{a}_K) \prod_{j=0}^{K} p(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1}) \\
&= \sum_{\bar{l}_K} p(y \mid \bar{l}_K, \bar{a}_K) \frac{\prod_{k=0}^{K} p(a_k \mid \bar{l}_k, \bar{a}_{k-1})}{\prod_{k=0}^{K} p(a_k \mid \bar{l}_k, \bar{a}_{k-1})} \prod_{j=0}^{K} p(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1}) \\
&= \sum_{\bar{l}_K} \frac{1}{\prod_{k=0}^{K} p(a_k \mid \bar{l}_k, \bar{a}_{k-1})} p(y \mid \bar{l}_K, \bar{a}_K) \prod_{k=0}^{K} p(a_k \mid \bar{l}_k, \bar{a}_{k-1}) \prod_{j=0}^{K} p(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1}) \\
&= \sum_{\bar{l}_K} \frac{1}{\prod_{k=0}^{K} p(a_k \mid \bar{l}_k, \bar{a}_{k-1})} p(y, \bar{l}_K, \bar{a}_K).
\end{aligned}
$$

Proof.

$$= \sum_{\bar{l}_K} \frac{1}{\prod_{k=0}^{K} p(a_k \mid \bar{l}_k, \bar{a}_{k-1})} p(y, \bar{l}_K, \bar{a}_K)$$

$$= \sum_{\bar{l}_K} \sum_{\bar{a}^*} \frac{I(\bar{a}^* = \bar{a})}{\prod_{k=0}^{K} p(a_k^* \mid \bar{l}_k, \bar{a}_{k-1}^*)} p(y, \bar{l}_K, \bar{a}_K^*)$$

$$= \frac{1}{g(\bar{a})} \sum_{\bar{l}_K} \sum_{\bar{a}^*} \frac{g(\bar{a}^*) I(\bar{a}^* = \bar{a})}{\prod_{k=0}^{K} p(a_k^* \mid \bar{l}_k, \bar{a}_{k-1}^*)} p(y \mid \bar{l}_K, \bar{a}_K^*) p(\bar{l}_K, \bar{a}_K^*)$$

$$= \frac{1}{g(\bar{a})} \mathbb{E}\left\{ \frac{g(\overline{A}) I(\overline{A} = \bar{a})}{\prod_{k=0}^{K} p(A_k \mid \overline{L}_k, \overline{A}_{k-1})} p(y \mid \overline{L}_K, \overline{A}_K) \right\}.$$

where the expectation is taken over $\overline{A}_K, \overline{L}_K$ under the distribution that generated the observed data, and positivity is used in the last line. □

So the lemma from Slide 309 is shown.

*A corollary

Proof.

$$\int y b_{\bar{a}}(y) dy$$

$$= \int y \frac{1}{g(\bar{a})} \mathbb{E}\left\{ \frac{g(\bar{A}) I(\bar{A} = \bar{a})}{\prod_{k=0}^{K} p(A_k \mid \bar{L}_k, \bar{A}_{k-1})} p(y \mid \bar{L}_K, \bar{A}_K) \right\} dy$$

$$= \frac{1}{g(\bar{a})} \int \mathbb{E}\left\{ \frac{g(\bar{A}) I(\bar{A} = \bar{a})}{\prod_{k=0}^{K} p(A_k \mid \bar{L}_k, \bar{A}_{k-1})} y p(y \mid \bar{L}_K, \bar{A}_K) \right\} dy$$

$$= \frac{1}{g(\bar{a})} \mathbb{E}\left\{ \frac{g(\bar{A}) I(\bar{A} = \bar{a})}{\prod_{k=0}^{K} p(A_k \mid \bar{L}_k, \bar{A}_{k-1})} Y \right\} \text{ (by def of expectation)}$$

*Another (simple) lemma

Lemma (individuals with $\overline{A} = \bar{a}$ in the psedopopulation)

$$\mathbb{E}\left\{ \frac{g(\overline{A})I(\overline{A} = \bar{a})}{\prod_{k=0}^{K} p(A_k \mid \overline{L}_k, \overline{A}_{k-1})} \right\} = g(\bar{a}).$$

Proof.

We use that the g-formula is a density, i.e. that $\int b_{\bar{a}}(y)dy = 1$,

$$1 = \int b_{\bar{a}}(y)dy = \int \frac{1}{g(\bar{a})} \mathbb{E}\left\{ \frac{g(\overline{A})I(\overline{A} = \bar{a})}{\prod_{k=0}^{K} p(A_k \mid \overline{L}_k, \overline{A}_{k-1})} p(y \mid \overline{L}_K, \overline{A}_K) \right\} dy$$

$$g(\bar{a}) = \mathbb{E}\left\{ \frac{g(\overline{A})I(\overline{A} = \bar{a})}{\prod_{k=0}^{K} p(A_k \mid \overline{L}_k, \overline{A}_{k-1})} \right\},$$

where we used that integrals of sums are sums of integrals. □

*PS: Pseudopopulation vs observed population

Just a PS: the lemma allows us to characterize the number of treated in the pseudopopulation vs the original population. Recall that $\mathbb{E}(I(\overline{A} = \overline{a}))$ is the fraction of individuals with $\overline{A} = \overline{a}$ in the observed population. Let n be the total size of the observed population. Then

$$n \times \mathbb{E}(I(\overline{A} = \overline{a}))$$

is the expected number of individuals with $\overline{A} = \overline{a}$ in the observed population and

$$n \times \mathbb{E}\left\{ \frac{g(\overline{A})I(\overline{A} = \overline{a})}{\prod_{k=0}^{K} p(A_k \mid \overline{L}_k, \overline{A}_{k-1})} \right\} = n \times g(\overline{a})$$

is the expected number of individuals with $\overline{A} = \overline{a}$ in the pseudopopulation.

*Finally: A poof of the Theorem

Proof.

plugging in for $g(\bar{a})$ in the Expression from the Corollary on slide 312,

$$= \frac{\mathbb{E}\left\{ \frac{g(\overline{A})I(\overline{A}=\bar{a})}{\prod_{k=0}^{K} p(A_k|\overline{L}_k,\overline{A}_{k-1})} Y \right\}}{\mathbb{E}\left\{ \frac{g(\overline{A})I(\overline{A}=\bar{a})}{\prod_{k=0}^{K} p(A_k|\overline{L}_k,\overline{A}_{k-1})} \right\}} \quad \text{(i.e. an IPW formula)}$$

$$= \frac{\mathbb{E}_{ps}(I(\overline{A}=\bar{a})Y)}{P_{ps}(\overline{A}=\bar{a})}$$

$$= \mathbb{E}_{ps}(Y \mid \overline{A}=\bar{a}).$$

□

This allows us to say "association is causation" in the pseudopopulation.

We can encode various assumptions in MSMs

- Suppose we hypothesize that the causal effect of treatment history $\bar{a}$ on the mean of Y is a linear function of the cumulative exposures, i.e.

$$\text{cum}(\bar{a}) = \sum_{k=0}^{K} a_k.$$

- This hypothesis is included in the MSM

$$\mathbb{E}(Y^{\bar{a}}) = \mathbb{E}_{ps}(Y \mid \overline{A} = \bar{a}) = \eta_0 + \eta_1 \text{cum}(\bar{a}).$$

  That is, we model the marginal mean of the counterfactuals $Y^{\bar{a}}$. Whereas there are 2^K treatment combinations (unknowns on the left-hand side of the equation), we have now reduced the model such that there are only two unknowns on the right-hand side of the equation.

- Obviously, like a statistical model, this model could also be misspecified, e.g. if the counterfactual outcome depends on some other function of the regime or if the outcome depends nonlinearly on the cumulative exposure.

*Motivating the weighted regressions

Lemma (Result for weighted least squares)

Suppose excheangeability, consistency and positivity hold. Then $\mathbb{E}_{ps}(Y \mid \overline{A} = \overline{a}) = \int yb(\overline{a})dy = \mathbb{E}(Y^{\overline{a}})$. *Then,*

$$\mathbb{E}\left\{ \frac{g(\overline{A})}{\prod_{k=0}^{K} p(A_k \mid \overline{L}_k, \overline{A}_{k-1})} [Y - \mathbb{E}(Y^{\overline{A}})] \right\}$$

$$\mathbb{E}_{ps}\left\{ [Y - \mathbb{E}(Y^{\overline{A}})] \right\}$$

$$= \mathbb{E}_{ps}\left\{ \mathbb{E}_{ps}\left\{ [Y - \mathbb{E}(Y^{\overline{A}})] \mid \overline{A} \right\} \right\}$$

$$= 0, \text{ because the inner expectation above is 0.}$$

Consider now the estimating equations

We use the results from the previous slide and the parameterisation

$$\mathbb{E}(Y^{\bar{a}}) = \eta_0 + \eta_1 \text{cum}(\bar{a}).$$

Now, consider the (two-dimensional) estimating equation

$$\sum_{i=1}^{n} M(\bar{L}_{k,i}, \bar{A}_i; \eta_0, \eta_1) = 0,$$

where

$$M(\bar{L}_k, \bar{A}; \eta_0, \eta_1) = \frac{g(\bar{A})}{\prod_{k=0}^{K} p(A_k|\bar{L}_k, \bar{A}_{k-1}; \gamma)} \begin{pmatrix} 1 \\ \text{cum}(\bar{A}) \end{pmatrix} [Y - \eta_0 - \eta_1 \text{cum}(\bar{A})].$$

This is an estimating equation for the weighted least squares estimator, where we simultaneously also solve the estimating equations for the propensities. Together, we denote the estimating equations for the counterfactual model and the propensity scores a "stacked estimating equation".

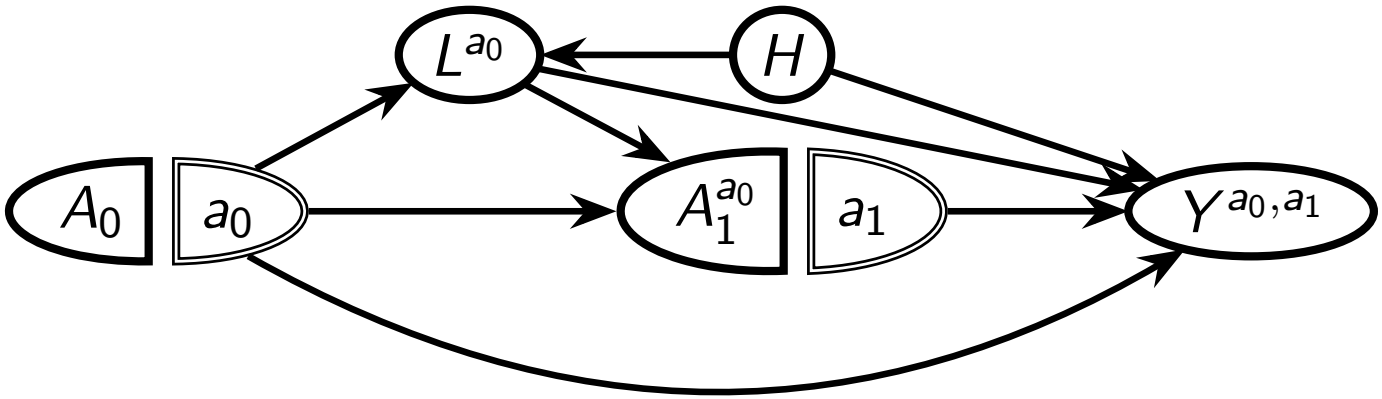
Null hypotheses in MSMs

Note that under the null hypothesis of no effect of any a_k, the MSM is correctly specified with

$$\mathbb{E}(Y^{\bar{a}}) = \eta_0.$$

However, the standardisation estimator (parametric g-formula estimator) suffers from the so-called "g-null-paradox". That is, it is possible to show that it will always reject the null hypothesis – even if the null hypothesis is true – when the sample size grows.

Treatment-confounder feedback



- we cannot adjust for L using traditional methods, like stratification, outcome regression, and matching.
- But we read off that $Y^{a_0,a_1} \perp\!\!\!\perp A_0$ and $Y^{a_0,a_1} \perp\!\!\!\perp A_1^{a_0} \mid L_0^{a_0}, A_0 = a_0$, and we can fit MSMs, like the one on Slide 318.

MSMs and effect modification

Suppose that an investigator believes that for a particular component V of the vector of baseline covariates L_0, there might exist qualitative effect modification with respect to V. For example, suppose that $A = 1$ is harmful to subjects with $V = 0$ and beneficial to those with $V = 1$.
To examine this hypothesis, we would elaborate the MSM,

$$\mathbb{E}(Y^{\bar{a}} \mid V) = \eta_0 + \eta_1 \text{cum}(\bar{a}) + \eta_2 V + \eta_3 \text{cum}(\bar{a}) V.$$

Then we have qualitative effect modification if $\text{sign}(\eta_1) \neq \text{sign}(\eta_1 + \eta_3)$.
We can e.g. use the weights,

$$\frac{\prod_{k=0}^{K} p(A_k \mid V, \overline{A}_{k-1})}{\prod_{k=0}^{K} p(A_k \mid \overline{L}_k, \overline{A}_{k-1})}$$

in a weighted least squares regression model.
One thing to notice: Here, IPW is used to adjust for confounding and regression modelling is used to study effect modification.

MSMs and direct effects

To illustrate a point, consider the saturated MSM for two binary treatments A_0, A_1,

$$\mathbb{E}(Y^{\bar{a}}) = \mathbb{E}(Y^{a_0,a_1}) = \eta_0 + \eta_1 a_0 + \eta_2 a_1 + \eta_3 a_0 a_1.$$

Now, the direct effect of A_0 when A_1 is set to 1 is $\mathbb{E}(Y^{1,1}) - \mathbb{E}(Y^{0,1})$.
How do we articulate the hypothesis that $\mathbb{E}(Y^{1,1}) = \mathbb{E}(Y^{0,1})$?

$$\begin{aligned} \mathbb{E}(Y^{1,1}) &= \mathbb{E}(Y^{0,1}) \\ \eta_0 + \eta_1 + \eta_2 + \eta_3 &= \eta_0 + \eta_2 \\ 0 &= \eta_1 + \eta_3 \end{aligned}$$

Optimal regimes and dynamic MSMs

Suppose that we aim to find the optimal treatment regime g^* in a given class of regimes $\{g = x : x \in \mathcal{X}\}$, where $|\mathcal{X}| = m$. Suppose that $x \in \{0, 1, \ldots, 999\}$. Let $n = 2000$ individuals.

- Suppose I come up with the following strategy: Run an experiment and randomly assign the regime g (In the experiment, we know association is causation)
- Maximize $\hat{\mathbb{E}}(Y \mid X = x)$
- Problem: We have m regimes, but only 2000 people so $\hat{\mathbb{E}}(Y \mid X = x)$ will be too variable...we will expect to have two people receiving the regime.
- Running example: Once we have started treatment (say, antiretroviral therapy in patients with HIV), then we never stop treatment. The question is: what is the best X to start treatment?

Dynamic MSMs

- Constructing an MSM allows us to impose assumptions, and then borrow strength across the regimes g, for example by assuming that $\mathbb{E}(Y \mid X = x) = \mathbb{E}(Y^x)$ is smooth in x.
- Note that we have to do this even if the data are from an experiment.
- Idea: for example, suppose we fit the model

$$\mathbb{E}(Y^x) = \eta_0 + \eta_1 x + \eta_2 x^2 + \eta_3 x^3.$$

- Then, we find the optimal regime g^* by maximising $\eta_1 x + \eta_2 x^2 + \eta_3 x^3$ over x.
- However, because there may be qualitative effect modification, we can expand the model to

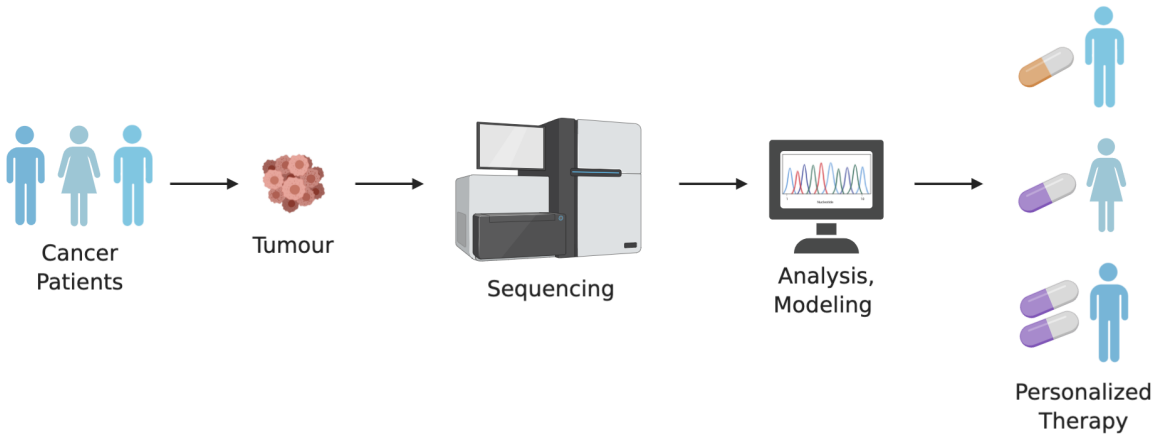
$$\mathbb{E}(Y^x \mid V) = \eta_0 + \eta_1 x + \eta_2 x^2 + \eta_3 x^3 + \eta_4 x V,$$

and for each value of V maximize $\eta_1 x + \eta_2 x^2 + \eta_3 x^3 + \eta_4 x V$ over x,
$g(v) = \underset{x \in \mathcal{X}}{\arg\max}\ \eta_1 x + \eta_2 x^2 + \eta_3 x^3 + \eta_4 x v$

Advantages of MSMs

- Easy to understand
- Can be fitted with simple (weighted models) in standard statistical software

Precision medicine is a buzz word



My claim:

Modelling the disease process is of secondary importance in precision medicine, except when it helps support the identification (and estimation) of optimal regimes.

Precision medicine is a buzz word, and the idea is simple

- The idea is to tailor treatment decisions to patient characteristics.
- The premise: *individual heterogeneity* can be leveraged to *individualize therapy*.
- Work on causal inference gives us theory for *optimizing* individual decisions.
  - What if patient i receives treatment A vs. treatment B?
    That is, what is the causal effect of taking A vs. B...

Algorithmic vs. human decisions

- Decision rules might be algorithmically individualized.[43]
- Yet these rules will be implemented under supervision of humans (e.g., doctors).[44]
- Are optimal algorithmic regimes better than human-decision rules?
  - Care providers may have information that is not recorded in the observed data.
    $\Longrightarrow$ unmeasured confounding in the data.
  - So, when should we let humans override algorithmic treatment recommendations?

[43] **topol2019high**.
[44] **matheny2019artificial**.

...but causal inference requires strong assumptions, no?

- We need to take the causal question seriously.
  Scientists who choose not to give up causal inference must understand that, without selecting a definition of a causal effect, it is impossible to evaluate whether we have reasonably estimated one.
- Can we deal with unmeasured confounding?
  - Sometimes we can point identify effects in the presence of unmeasured confounding.
    Instrumental variables, front-door variables, negative controls (proximal inference) ...
  - Other times we can bound the causal effects.

  Transparency about study goals and the assumptions we make to justify an analysis are required to discuss bias, refine our questions and improve our answers.

Section 27

Unmeasured confounding and instrumental variables

We have derived results under identification assumptions, but what do we do when these assumptions are violated?

- We have relied on the key assumptions that we have:
  - measured a sufficient set of variables to adjust for confounding, and
  - we have avoided selection bias.
- If these assumptions are incorrect, our estimation strategies will yield bias.
- Now we will discuss *alternative* strategies that can validly estimate causal effects under an alternative set of assumptions that do not require that our conventional exchangeability conditions hold.
- Our first example is instrumental variable (IV) methods.
  - Instrumental variables are *very* popular in economics and the social sciences. Angrist, Imbens and Card were awarded the 2021 Nobel medal in Economics for their work on instrumental variables.

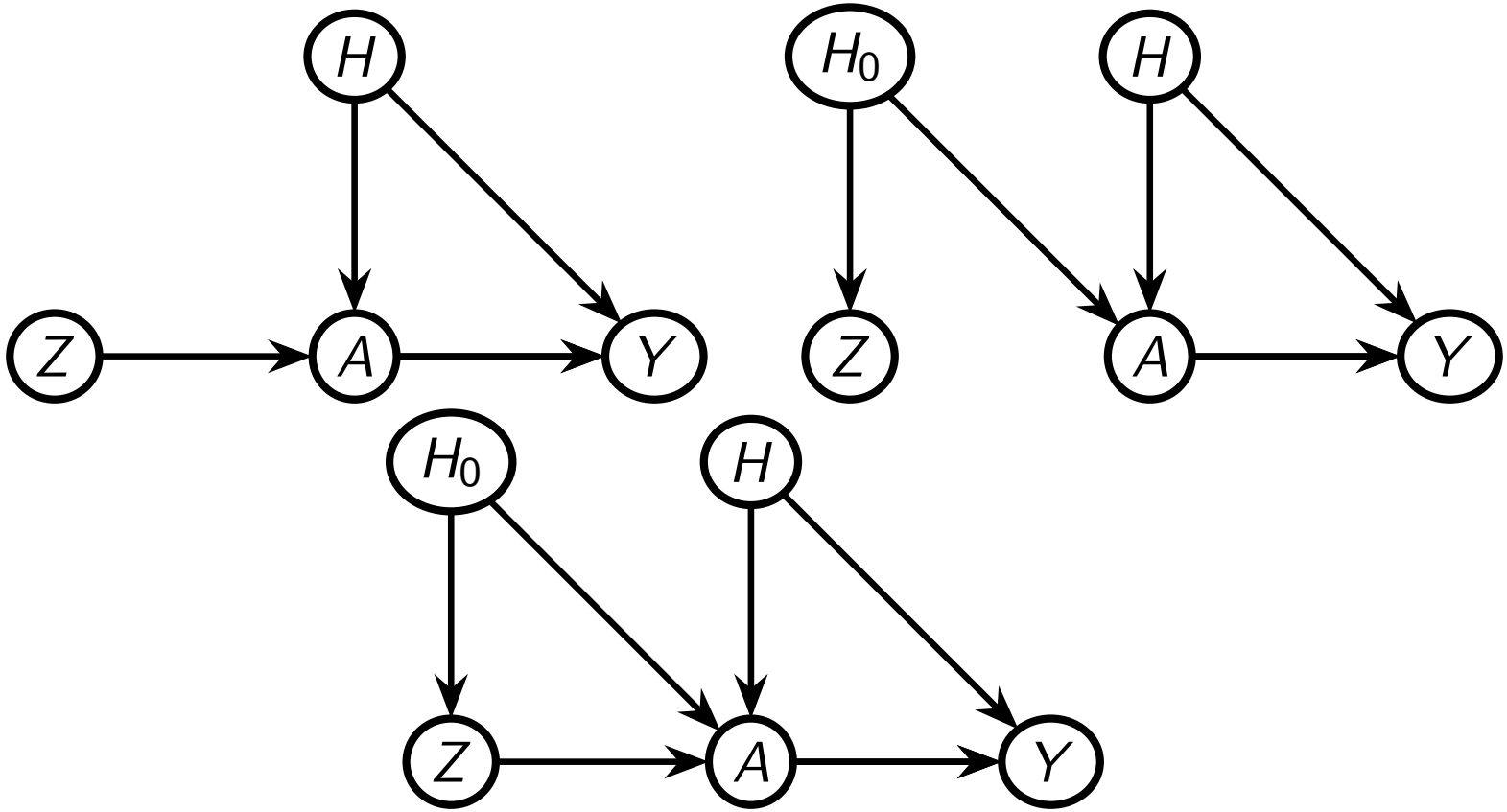
Definition of IVs

Definition (Main IV assumptions)

1. $\text{cor}(Z, A) \neq 0$ (instrument strength)
2. $Y^{z,a} = Y^a$ for all a, z (exclusion restriction)
3. $Z \perp\!\!\!\perp Y^a$ for all a (unconfoundedness of Z)

- The main IV assumptions are not themselves sufficient to identify the average effect of A on Y; thus, we need additional assumptions, and people have suggested several different ones; these different conditions are often called *homogeneity conditions*.
- If the unconfoundedness assumption holds, then there are no common causes of Z and Y in the DAG.

3 graphs that satisfy the main IV assumptions



In the first graph, we have a *causal instrument*, in the second graph we have a *proxy instrument* and the third graph is a combination.

Example of IV studies

- Several studies in economics.
  - A seminal example is on the effect of education on future earnings by Joshua D Angrist and Alan B Krueger. "Estimating the payoff to schooling using the Vietnam-era draft lottery". In: *NBER working paper* w4067 (1992);
    It seems very difficult to adjust for common causes of education and future earnings, but the authors used the result of a lottery that determined priority for the US military during the Vietnam war.
- Randomised controlled trials when treatment is blinded.
- Some non-blinded studies. For example, American economist gave families vouchers Z to reduce the costs from moving from a neighbourhood with high poverty to a neighbourhood with low poverty. A denotes moving. Y is psychological stress.
- Mendelian Randomization
  - A genetic variant Z that is associated with treatment A and is not associated with the outcome Y. outside of A.
- Applied researchers use them in a range of other settings too.

More on examples

- In our smoking cessation study: The price of cigarettes in the population could be an instrument if:
  - Cigarette price affects the decision to quit smoking,
  - cigarette price affects weight change only through its effect on smoking cessation, and
  - no common causes of cigarette price and weight change exist.

Hernan and Robins, *Causal inference: What if?*

Additional IV assumptions: linear structural equation model

Suppose that the structural equation for Y is linear,

Definition (Linear SEM)

$$Y = f_y(A, H, \epsilon_Y) = \beta A + g(H, \epsilon_Y).$$

Clearly, this linear structure is stronger than what we have previously imposed when we have done identification (think about all the things we did on non-parametric structural equations, DAGs and SWIGs).
However, some disciplines have almost only considered linear models. And there is a lot of disagreement about whether these assumptions are justified.
We will leave these issues aside for a moment (but we will get back to them), and notice that if follows from the linear SEM that

$$Y^{a'} - Y^a = \beta(a' - a).$$

IV theorem

Theorem (IV theorem 1)

If $cor(Z, A) \neq 0$, $\mathbb{E}(Y^{a=0} \mid Z) = \mathbb{E}(Y^{a=0})$, *and the linear SEM hold, then*

$$\psi \equiv \frac{cov(Y, Z)}{cov(A, Z)} = \beta$$

The first two assumptions are implied by the main IV assumptions
We will call ψ the IV functional.

Proof of IV theorem 1

Proof.

Under the linear SEM,

$$Y - Y^{a=0} = \beta A + g(H, \epsilon_Y) - g(H, \epsilon_Y) = \beta A, \ \textit{i.e.} \ Y - \beta A = Y^{a=0}$$

Thus, $\mathbb{E}(Y - \beta A \mid Z) = \mathbb{E}(Y^{a=0} \mid Z) = \mathbb{E}(Y^{a=0})$, that is, $\mathbb{E}(Y - \beta A \mid Z)$ is independent of Z. Thus,

$$\begin{aligned}
&\mathbb{E}[\{Z - \mathbb{E}(Z)\}(Y - \beta A)] \\
=&\mathbb{E}[\{Z - \mathbb{E}(Z)\}\mathbb{E}\{(Y - \beta A) \mid Z\}] \text{ iterative expectation} \\
=&\mathbb{E}[\{Z - \mathbb{E}(Z)\}\mathbb{E}\{Y - \beta A\}] \text{ the independence above} \\
=&\mathbb{E}\{Z - \mathbb{E}(Z)\}\mathbb{E}\{Y - \beta A\}, \text{ the independence above} \\
=&0,
\end{aligned}$$

and therefore $\text{cov}(Y - \beta A, Z) = 0$ (you'll see this by using the definition of covariance), and $0 = \text{cov}(Y - \beta A, Z) = \text{cov}(Y, Z) - \beta \text{cov}(A, Z)$. □

An intuitive interpretation

In the "causal instrument" graph, the IV functional has an intuitive interpretation:

- Consider the coefficient of the population least squares of a dependent variable W on $(1, Z)$. This coefficient say β_{WZ}, is indeed $\beta_{WZ} = \frac{\text{cov}(W,Z)}{\text{var}(Z)}$.
  That is, we can think about β_{WZ} as the limit in probability of the least squares coefficient in the regression model $\mathbb{E}(W \mid Z; \alpha, \beta) = \alpha + \beta Z$.
- By dividing the definition of ψ by $\text{var}(Z)$ in the numerator and denominator it follows that

$$\psi = \frac{\beta_{YZ}}{\beta_{AZ}}.$$

- If the instrument $Z \in \{0, 1\}$ then

$$\psi = \frac{\mathbb{E}(Y \mid Z = 1) - \mathbb{E}(Y \mid Z = 0)}{\mathbb{E}(A \mid Z = 1) - \mathbb{E}(A \mid Z = 0)},$$

  in words, the average additive effect of Z on Y divided by average effect of Z on A in our IV graph.

The problem with the linear SEM

- The linear SEM is a very strong restriction. Essentially, we are saying that all individuals have the same effect of the treatment, which is very unlikely.
  - In fact, the key idea of "personalised medicine" is that different people respond different to treatments.
  - In the smoking cessation example, this assumption would only hold if smoking cessation made every individual in the population gain (or lose) the same amount of body weight!
- In the homework you will show that ψ has a causal interpretation even under a relaxation of the linear SEM assumption, where

$$Y = f_y(A, H, \epsilon_Y) = h(\epsilon_Y)A + g(H, \epsilon_Y),$$

where h and g are unspecified functions. However, now we've made the strong assumption that H does not modify the causal effect of A (on the additive scale) because h does not have H as argument.

A further relaxation (we will not study this one in detail)

Definition (Robins's IV assumptions)

1. $\text{cor}(Z, A) \neq 0$ (instrument strength) (as before)
2. $Y^{z,a} = Y^a$ for all a, z (exclusion restriction) (as before)
3. $Z \perp\!\!\!\perp Y^{a=0}$ (unconfoundedness of Z) (Slightly weaker)
4. There exists a β such that

$$\mathbb{E}(Y \mid Z, A) - \mathbb{E}(Y^{a=0} \mid Z, A) = \beta A$$

On Robins's IV assumptions

Theorem

Under Robins's IV assumptions, $\beta = \psi$

Proof.

Using assumption 4,

$$\mathbb{E}(Y - \beta A \mid Z, A) = \mathbb{E}(Y^{a=0} \mid Z, A),$$

and thus we integrate out A,

$$\mathbb{E}(Y - \beta A \mid Z) = \mathbb{E}(Y^{a=0} \mid Z),$$

and follow the steps in the proof of the first theorem on IVs. □

*Robins's IV assumptions (continuation)

- The last assumption is an example of a so-called structural nested model,

$$\mathbb{E}(Y \mid Z, A) - \mathbb{E}(Y^{a=0} \mid Z, A) = h(A, Z; \beta)$$

  satisfying $h(0, Z; \beta) = 0$ for all β.

- The model from the previous slide does not assume effect homogeneity, but it does (only) assume no effect modification by Z on the additive scale.

- However, β does not (without extra assumptions) have the interpretation as the average (additive) treatment effect, but when A is binary it quantifies an average treatment effect of the treated (use consistency to prove this),

$$\mathbb{E}(Y^{a=1} \mid Z, A = 1) - \mathbb{E}(Y^{a=0} \mid Z, A = 1) = \mathbb{E}(Y^{a=1} - Y^{a=0} \mid A = 1) = \beta.$$

- Does this model generalize the linear structural equation model? Yes.

But is it a plausible assumption that we can reason about?

- How can a scientist (or other expert) argue in support of a constant average causal effect within levels of the proposed instrument Z and the treatment A in any particular study? Hernan and Robins, *Causal inference: What if?*
- Yet another possibility is to assume that, for any level of the unmeasured variable H, the effect of A on Y is the same, i.e.

$$\mathbb{E}(Y^{a=1} \mid H) - \mathbb{E}(Y^{a=0} \mid H) = \mathbb{E}(Y^{a=1}) - \mathbb{E}(Y^{a=0}),$$

  but this assumption is not plausible either because the unmeasured variables can often be effect modifiers.

  - For example, weight gain after smoking cessation can vary with prior intensity of smoking, which may itself be an unmeasured confounder for the effect of smoking cessation on weight gain.[45]

[45] Hernan and Robins, *Causal inference: What if?*

Another common alternative

The final criterion we will study is not a criterion about homogeneity.

Definition (Imbens and Angrist's IV assumptions)

1. $\text{cor}(Z, A) \neq 0$ (valid instrument)
2. $Y^{z,a} = Y^a$ for all a, z (exclusion restriction)
3. $Z \perp\!\!\!\perp Y^a$ and $Z \perp\!\!\!\perp A^z$ (strong unconfoundedness of Z)
4. $A^{z=1} \geq A^{z=0}$ (Monotonicity)

- These assumptions are often used in practice
- Note that the 3rd assumption is violated in two of our example graphs
- The 4th assumption is strong but sometimes plausible. I will give some intuition why.

Intuition (Robins)

We can only can estimate the effect of treatment on those whose behavior was actually affected by the instrument, so Compliers and Defiers are the only relevant sets. If we have both, then we get mixed up. If there are only Compliers, things work out OK

More on monotonicity

Suppose $Z, A \in \{0, 1\}$. Then we can divide the population into 4 mutually exclusive groups

- $(A^{z=1} = 0, A^{z=0} = 0)$, the never-takers
- $(A^{z=1} = 1, A^{z=0} = 1)$, the always-takers
- $(A^{z=1} = 0, A^{z=0} = 1)$, the defiers
- $(A^{z=1} = 1, A^{z=0} = 0)$, the compliers

Monotonicity assumes no defiers in the entire population. That is, nobody does exactly the opposite of what they are told to do.

Definition (Local Average Treatment Effect)

The local average treatment effect in stratum $A^{z=1} = a, A^{z=0} = a'$

$$\text{LATE} = \mathbb{E}(Y^{a=1} - Y^{a=0} \mid A^{z=1} = a, A^{z=0} = a')$$

In particular, the complier average treatment effect is

$$\text{CACE} = \mathbb{E}(Y^{a=1} - Y^{a=0} \mid A^{z=1} = 1, A^{z=0} = 0)$$

Monotonicity example

As a simplified example, consider a physician who generally prefers Treatment A, but prescribes Treatment B for more physically active patients (e.g., because Treatment A is associated with risk of motor-skill impairment), and another physician who generally prefers Treatment B, but makes exceptions for patients with a family history of diabetes (e.g., because a new study suggests such patients might respond better to Treatment A). Any physically active patient with a family history of diabetes who could potentially have seen either of these providers would "defy" both preferences and thus violate the monotonicity assumption Sonja A Swanson et al. "Definition and evaluation of the monotonicity condition for preference-based instruments". In: *Epidemiology (Cambridge, Mass.)* 26.3 (2015), p. 414

Returning to our smoking example

- the compliers are those who would quit smoking when the high cigarette price is high and who would not quit smoking when the cigarette price is low. Conversely, the defiers are those who would *not* quit smoking when the high cigarette price is high and who would quit smoking when the cigarette price is low.[46]

[46] Hernan and Robins, *Causal inference: What if?*

Result on Imbens and Angrist's IV

Theorem

Under conditions 1-4 of Angrist and Imbens,

$$CACE = \psi$$

*Proof of Angrist and Imbens

Proof.

$$
\begin{aligned}
& Y^{z=1} - Y^{z=0} \\
&= Y^{z=1, A^{z=1}} - Y^{z=0, A^{z=0}} \text{ consistency} \\
&= Y^{A^{z=1}} - Y^{A^{z=0}} \text{ exclusion restriction} \\
&= (Y^{a=1} - Y^{a=0}) A^{z=1} + Y^{a=0} - \{(Y^{a=1} - Y^{a=0}) A^{z=0} + Y^{a=0}\} \text{ consistency} \\
&= (Y^{a=1} - Y^{a=0})(A^{z=1} - A^{z=0})
\end{aligned}
$$

Note that because $A^{z=1} \geq A^{z=0}$, $(A^{z=1} - A^{z=0}) \in \{0, 1\}$. Thus,

$$
\begin{aligned}
\mathbb{E}(Y^{z=1} - Y^{z=0}) &= \mathbb{E}[(Y^{a=1} - Y^{a=0})(A^{z=1} - A^{z=0})] \\
&= \mathbb{E}[(Y^{a=1} - Y^{a=0}) \mid A^{z=1} - A^{z=0} = 1] P(A^{z=1} - A^{z=0} = 1) \\
&= \mathbb{E}[(Y^{a=1} - Y^{a=0}) \mid A^{z=1} > A^{z=0}] P(A^{z=1} > A^{z=0}).
\end{aligned}
$$

Thus, $\mathbb{E}[(Y^{a=1} - Y^{a=0}) \mid A^{z=1} > A^{z=0}] = \frac{\mathbb{E}(Y^{z=1} - Y^{z=0})}{P(A^{z=1} > A^{z=0})}$. □

*Proof of Angrist and Imbens

Proof.

Furthermore, by assumption 3,

$$\mathbb{E}(Y^{z=1} - Y^{z=0}) = \mathbb{E}(Y \mid Z = 1) - \mathbb{E}(Y \mid Z = 0)$$

and

$$\begin{aligned}
P(A^{z=1} > A^{z=0}) &= P(A^{z=1} = 1, A^{z=0} = 0) \\
&= P(A^{z=1} = 1) - P(A^{z=1} = 1, A^{z=0} = 1) \text{ law of total probability} \\
&= P(A^{z=1} = 1) - P(A^{z=0} = 1) \text{ monotonicity} \\
&= P(A = 1 \mid Z = 1) - P(A = 1 \mid Z = 0) \text{ assumption 3 and consist.}
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}[(Y^{a=1} - Y^{a=0}) \mid A^{z=1} > A^{z=0}] &= \frac{\mathbb{E}(Y^{z=1} = Y^{z=0})}{P(A^{z=1} > A^{z=0})} \\
&= \frac{\mathbb{E}(Y \mid Z = 1) - \mathbb{E}(Y \mid Z = 0)}{P(A = 1 \mid Z = 1) - P(A = 1 \mid Z = 0)}
\end{aligned}$$

IV estimation

"IV estimation requires modeling assumptions (such as monotonicity) even if infinite data were available. This is not the case for previous methods like IP weighting or standardization: If we had treatment, outcome, and confounder data from all individuals in the superpopulation"[47].

[47] Hernan and Robins, *Causal inference: What if?*

The monotonicity assumption was considered to be a salvation, but...

- Hard to use the complier average treatment effect by decision makers, because it only tells us something about a subset of the population. Suppose, for example, 10% of the population are compliers. Then,
  - can we justify to make recommendations based on the CATE to everyone in the population?
  - Unfortunately, we cannot observe the compliers, so we cannot target the intervention to the compliers.
  - What is the right thing to do if the treatment is not beneficial in always-takers and never-takers?
  - I agree with Hernan & Robins that it is often better to be more honest and accept that "interest in this estimand is not the result of its practical relevance, but rather of the (often erroneous) perception that it is easy to identify..."

Angus Deaton

"This goes beyond the old story of looking for an object where the light is strong enough to see; rather, we have control over the light, but choose to let it fall where it may and then proclaim that whatever it illuminates is what we were looking for all along."
Angus Deaton

In practice

- "Second, relatively minor violations of conditions (i)-(iv) for IV estimation may result in large biases of unpredictable or counterintuitive direction. The foundation of IV estimation is that the denominator blows up the numerator. Therefore, when the conditions do not hold perfectly or the instrument is weak, there is potential for explosive bias in either direction."
- "As a result, an IV estimate may often be more biased than an unadjusted estimate. In contrast, previous methods tend to result in slightly biased estimates when their identifiability conditions are only slightly violated, and adjustment is less likely to introduce a large bias. The exquisite sensitivity of IV estimates to departures from its identifiability conditions makes the method especially dangerous"

Hernan and Robins, *Causal inference: What if?*

More positively

- "IV estimation is better reserved for settings with lots of unmeasured confounding, a truly dichotomous and time-fixed treatment A, a strong and causal proposed instrument Z, and in which either effect homogeneity is expected to hold, or one is genuinely interested in the effect in the compliers and monotonicity is expected to hold."[48]
- Causal inference relies on transparency of assumptions and results from analyses that rely on different assumptions. In that sense, IV is an attractive approach because it depends on a different set of assumptions than other methods.

[48] Hernan and Robins, *Causal inference: What if?*

Plan for the last lecture

- Some more on IVs
- Bounds
- Sensitivity analysis

Section 28

IV inequalities

Motivation: IV inequalities

Theorem (IV inequalities)

Suppose $Z \perp\!\!\!\perp Y^a$, positivity and consistency hold. Then,

$$
\begin{aligned}
P[Y=0, A=0 \mid Z=0] + P[Y=1, A=0 \mid Z=1] \leq 1; \\
P[Y=0, A=1 \mid Z=0] + P[Y=1, A=1 \mid Z=1] \leq 1; \\
P[Y=1, A=0 \mid Z=0] + P[Y=0, A=0 \mid Z=1] \leq 1; \\
P[Y=1, A=1 \mid Z=0] + P[Y=0, A=1 \mid Z=1] \leq 1.
\end{aligned}
$$

The idea is that the instrumental variable assumptions put constraints on the joint law $p(y, a, z)$. This is interesting, because , in principle, we can use these logical bounds to use evaluate the IV assumptions: we can derive a test of whether the IV assumption $Z \perp\!\!\!\perp Y^a$ holds. If any of the above inequalities fail, then the core conditions must be violated; however, it is possible that the core IV conditions are violated without failing the inequalities.

Cont

Proof.

For $i, j, k \in \{0, 1\}$,

$$
\begin{aligned}
& P[Y^{a=i} = j] \\
& = P[Y^{a=i} = j \mid Z = k] \quad \text{bc. } (Z \perp\!\!\!\perp Y^a) \\
& = P[Y^{a=i} = j, A = i \mid Z = k] + P[Y^{a=i} = j, A = 1 - i \mid Z = k] \quad \text{laws of prob.} \\
& = P[Y = j, A = i \mid Z = k] + P[Y^{a=i} = j, A = 1 - i \mid Z = k] \quad \text{const.} \\
& \leq P[Y = j, A = i \mid Z = k] + P[A = 1 - i \mid Z = k] \\
& = 1 - P[Y = 1 - j, A = i \mid Z = k];
\end{aligned}
$$

Thus

$$
\begin{aligned}
& \max_k P[Y = 1, A = i \mid Z = k] \leq P[Y^{a=i} = 1] \\
& \leq \min_{k^*} 1 - P[Y = 0, X = i \mid Z = k^*],
\end{aligned}
$$

where the lower bounds follows by taking $j = 0$ in the exp. for $P[Y^{a=i} = j]$ □

According to Pearl

"The instrumental inequality can be used in the detection of undesirable side- effects. Violations of this inequality can be attributed to one of two possibilities: either there is a direct causal effect of the assignment (Z) on the response (Y), unmediated by the treatment (A), or there is a common causal factor influencing both variables. If the assignment is carefully randomized, then the latter possibility is ruled out and any violation of the instrumental inequality (even un- der conditions of imperfect compliance) can safely be attributed to some direct influence of the assignment process on subjects' response (e.g., psychological aversion to being treated). Alternatively, if one can rule out any direct effects of Z on Y , say through effective use of a placebo, then any observed violation of the instrumental inequality can safely be attributed to spurious dependence between Z and Y, namely, to selection bias.

Section 29

Motivation for bounds

Bounds

- Motivation: Can we derive *partial* identification results (i.e. bounds) under weaker assumptions (than those imposed so far)?
- Anyway are bounds useful? I think the answer is yes.
  The following text is from Robins and Greenland:
  - "Some argue against reporting bounds for nonidentifiable parameters, because bounds are often so wide as to be useless for making public health decisions.
  - But we view the latter problem as a reason for reporting bounds in conjunction with other analyses: Wide bounds make clear that the degree to which public health decisions are dependent on merging the data with strong prior beliefs.
  - Even when the ITT[49] null hypothesis of equality of treatment arm-specific means is rejected, the bounds may appropriately include zero. If treatment benefits some subjects and harms others, the ATE parameter may be zero even though both the sharp and ITT null hypotheses are false

[49] say, the effect of Z in our considerations

According to Pearl

When conditions for identification are not met, the best one can do is derive bounds for the quantities of interest—namely, a range of possible values that represents our ignorance about the data-generating process and that cannot be improved with increasing sample size.

Bounds on the ATE

$\mathbb{E}[Y^1 - Y^0] = \mathbb{E}[Y^1] - \mathbb{E}[Y^0]$ can be decomposed as

$$\sum_{a=0}^{1} E\left[Y^1 \mid A = a\right] P[A = a] - \sum_{a=0}^{1} E\left[Y^0 \mid A = a\right] P[A = a]. \tag{10}$$

- $\mathbb{E}[Y^a \mid A = a] = \mathbb{E}[Y \mid A = a]$ by consistency.
- $\mathbb{E}[Y^a \mid A = a]$ and $P[A = a]$ are identifiable and can be consistently estimated by their empirical counterparts.
- the observed data provide no information about $\mathbb{E}[Y^a \mid A = 1 - a]$, such that (10) is only partially identifiable without additional assumptions (such as exchangeability).

Bounds on the ATE

- $\mathbb{E}[Y^1 - Y^0]$ is bounded by smallest and largest possible values for $\mathbb{E}[Y^a \mid A = 1 - a]$.
- If Y^1 and Y^0 are not bounded then bounds on $\mathbb{E}[Y^1 - Y^0]$ will be ranging from $-\infty$ to ∞.
- Informative bounds are only possible if Y^0 and Y^1 are bounded.
- Because any bounded variable can be rescaled to take values in the unit interval, without loss of generality assume $Y^a \in [0, 1]$ for $a = 0, 1$. Then $0 \leq \mathbb{E}[Y^a \mid A = 1 - a] \leq 1$ and from (10) it follows that $\mathbb{E}[Y^1 - Y^0]$ is bounded below by setting $\mathbb{E}[Y^1 \mid A = 0] = 0$ and $\mathbb{E}[Y^0 \mid A = 1] = 1$, which yields the lower bound

$$E\big[Y^1 \mid A = 1\big]P[A = 1] - E\big[Y^0 \mid A = 0\big]P[A = 0] - P[A = 1].$$

Similarly, $\mathbb{E}[Y^1 - Y^0]$ is bounded above by setting $\mathbb{E}[Y^1 \mid A = 0] = 1$ and $\mathbb{E}[Y^0 \mid A = 1] = 0$, which yields the upper bound

$$E\big[Y^1 \mid A = 1\big]P[A = 1] - E\big[Y^0 \mid A = 0\big]P[A = 0] + P[A = 0].$$

Width of bounds

Determining treatment effect bounds can be viewed as a constrained optimization problem. The assumptions we make, for example exchangeabilities, determine the constraints.

- The bounds from the previous slide have width 1 and are contained in $[-1, 1]$, and are called the Manski-Robins bounds.

Motivating example 2: bounds

- We will consider a setting where Z, A, Y are all binary. This could for example be plausible in a randomized controlled trial, where
  - Z is treatment assignment
  - A is the treatment taken
  - Y is the outcome
- In our motivation, we will assume no defiers (suppose the treatment is only available among those with $Z = 1$). However, importantly, we will relax this assumption; let's think about an RCT with one-sided compliance.
- What do we know about the average treatment effect?
  - We will explore this (and build some intuition) in the next slides.

Motivating example 2 (cont.): always-takers

- Suppose monotonicitiy (no defiers).
  - Then we can simply identify always-takers by $A^{z=0} = 1$.
  - The fraction of always-takers is $P(A = 1 \mid Z = 0)$
  - $\mathbb{E}(Y^{a=1} \mid A = 1, Z = 0) = \mathbb{E}(Y \mid A = 1, Z = 0)$
    $= \mathbb{E}(Y \mid A^{z=0} = 1, A^{z=1} = 1)$.
  - $\mathbb{E}(Y^{a=1} - Y^{a=0} \mid A = 1, Z = 0) \leq \mathbb{E}(Y \mid A = 1, Z = 0)$
    with equality when all always-takers have $Y^{a=0} = 0$.

Motivating example 2 (cont.): never-takers

- Suppose monotonicitiy (no defiers).
  - Then we can simply identify never-takers by $A^{z=1} = 0$.
  - The fraction of never-takers is $P(A = 0 \mid Z = 1)$
  - $\mathbb{E}(Y^{a=0} \mid A = 0, Z = 1) = \mathbb{E}(Y \mid A = 0, Z = 1) = \mathbb{E}(Y \mid A^{z=0} = 0, A^{z=1} = 0)$.
  - $\mathbb{E}(Y^{a=1} - Y^{a=0} \mid A = 0, Z = 1) \leq 1 - \mathbb{E}(Y^{a=0} \mid A = 0, Z = 1)$ with equality when all never-takers have $Y^{a=1} = 1$.

Suppose no effect in compliers

Combine the simple results from the two previous slides to gain some insight:

- Suppose monotonicitiy (no defiers).
  - Suppose no effect in compliers $\implies \mathbb{E}(Y^{z=1} = Y^{z=0}) = 0$, in other words no intention to treat effect (ITT). Think about it, if it isn't clear!
  - Then the maximum possilbe effect of actually taking treatment is

$$\begin{aligned}
&\mathbb{E}(Y^{a=1} - Y^{a=0}) \\
\leq &\mathbb{E}(Y^{a=1} \mid A = 1, Z = 0) P(A = 1 \mid Z = 0) \\
&+ [1 - \mathbb{E}(Y^{a=0} \mid A = 0, Z = 1)] P(A = 0 \mid Z = 1),
\end{aligned}$$

  even if the intention to treat (ITT) effect $\mathbb{E}(Y^{z=1} = Y^{z=0}) = 0$.
  - Thus, even if the ITT effect is zero, there could a be considerable causal effects of taking treatment. In other words, even if the ITT is null, the ATE can be nonzero, which seriously complicate the interpretation of hypothesis tests of the ITT in settings with (a substantial amount of) noncompliance.

Section 30

Bounds and decision making

Bounds on conditional average treatment effects and optimal decisions

When $\mathbb{E}(Y^a \mid L = l)$ is point identified, we simply identify the optimal rule

$$g_{\mathbf{opt}}(l) \equiv \underset{a \in \{0,1\}}{\arg\max}\, \mathbb{E}(Y^a \mid L = l).$$

When $\mathbb{E}(Y^a \mid L = l)$ partially identified, then

$$\begin{aligned} \mathcal{L}^a(l) \leq & \mathbb{E}(Y^a \mid L = l) \leq \mathcal{U}^a(l), \;\; a = 0, 1, \\ \mathcal{L}(l) \leq & \mathbb{E}(Y^1 - Y^0 \mid L = l) \leq \mathcal{U}(l), \end{aligned}$$

where $\mathcal{L}(l) = \mathcal{L}^1(l) - \mathcal{U}^0(l)$ and $\mathcal{U}(l) = \mathcal{U}^1(l) - \mathcal{L}^0(l)$.

Are bounds uninformative?

- Bounds are often considered "uninformative" when

$$\mathcal{L}(l) \leq 0 \leq \mathcal{U}(l),$$

  that is, when the sign of $\mathbb{E}(Y^1 - Y^0 \mid L = l)$ is unidentified.
- But there exist formal decision theory results...

*Consider a generalized lower bound on the $\mathbb{E}(Y^a \mid L = l)$

Let $g : \mathcal{L} \to \{0, 1\}$ be a treatment rule (dynamic wrt. to the covariate $l \in \mathcal{L}$), that is a function which assigns treatment.

- Define the "bounds optimal rule":

$$g_{\mathbf{bopt}}(l) = \arg\max_{a \in \{0,1\}} \left[\{1 - w(l)\}\{\mathcal{L}(l)a + \mathcal{L}^0(l)\} + w(l)\{-\mathcal{U}(l)(1-a) + \mathcal{L}^1(l)\}\right],$$

where $0 \leq w(l) \leq 1$ for all l. This is maximisation of a lower bound in the following sense:

Lemma

For a decision rule $g(l)$,

$$h^g(l) := \left[\{1 - w(l)\}\{\mathcal{L}(l)I(g(l) = 1) + \mathcal{L}^0(l)\} + w(l)\{-\mathcal{U}(l)I(g(l) = 0) + \mathcal{L}^1(l)\}\right] \leq \mathbb{E}(Y^g \mid L = l).$$

*Proof of $g_{\mathbf{bopt}}$ being a lower bound (previous slides)

Proof.

Let $g : \mathcal{L} \to \{0, 1\}$ be a decision rule.

$$
\begin{aligned}
&\mathbb{E}(Y^g \mid L = l) \\
&= \mathbb{E}(Y^{a=1} \mid L = l) I(g(l) = 1) + \mathbb{E}(Y^{a=0} \mid L = l) I(g(l) = 0) \\
&= \mathbb{E}(Y^{a=1} - Y^{a=0} \mid L = l) I(g(l) = 1) + \mathbb{E}(Y^{a=0} \mid L = l) \\
&= \mathbb{E}(Y^{a=0} - Y^{a=1} \mid L = l) I(g(l) = 0) + \mathbb{E}(Y^{a=1} \mid L = l)
\end{aligned}
$$

Because $\mathcal{L}^a(l) \leq \mathbb{E}(Y^a \mid L = l) \leq \mathcal{U}^a(l),\ a = 0, 1$, and we can use these bounds to find that

$$
\begin{aligned}
&\{1 - w(l)\} \{\mathcal{L}(l) I(g(l) = 1) + \mathcal{L}^0(l)\} \\
&\quad + w(l) \{-\mathcal{U}(l) I(g(l) = 0) + \mathcal{L}^1(l)\} \\
&\leq \mathcal{L}^1(l) I(g(l) = 1) + \mathcal{L}^0(l) I(g(l) = 0) \leq \mathbb{E}(Y^g \mid L = l),
\end{aligned}
$$

where $0 \leq w(l) \leq 1$ for all $l \in \mathcal{L}$. □

Classical decision criteria

Classical criteria are special cases of $g_{\mathbf{bopt}}$, for example

- Minimax regret (Opportunist) $\min_g \max[\mathbb{E}(Y^{g_{\mathbf{opt}}}) - \mathbb{E}(Y^g)]$. $(w(l) = 0.5 \ \forall l)$
- Healthcare decision-making. $\max_g \mathbb{E}\{\mathbb{E}(Y^0 \mid L) + \mathcal{L}(L)g(L)\}$ $(w(l) = 0 \ \forall l)$
- Maximax utility (Optimist) $\max_g \max \mathbb{E}(Y^g)$.
- Maximin utility (Pessimist) $\max_g \min \mathbb{E}(Y^g)$.

*Features of the bounds

- If $\mathbb{E}(Y^a \mid L = l)$, $a = 0, 1$ is point identified, then $g_{\mathbf{bopt}} = g_{\mathbf{opt}}$.
- If there is no uncertainty about the optimal decision, that is $0 \notin (\mathcal{L}(l), \mathcal{U}(l))$, then $g_{\mathbf{bopt}} = g_{\mathbf{opt}}$ regardless of the choice of $w(l)$.

Section 31

Sensitivity analysis

Sensitivity analysis

"Over recent decades recognition has grown that the conventional statistical models used to analyze epidemiological data cannot be reasonably claimed to be correct in the way most textbooks treat them to be. In particular, conventional models for epidemiological data-generating processes cannot be credibly taken to represent targets of primary scientific interest." This is a quote form Sander Greenland.

- Also called bias analysis.
- Different from bounds: bounds are derived under *minimal* assumptions, whereas sensitivity analysis rely on assumptions that the investigator find plausible (but the notion of plausibility is subjective...).
- In some sense, bounds are therefore more desirable as they rely on less (subjective) assumptions, but they are often wide.

Motivation

- Intuition for sensitivity analysis: We want to evaluate how strong the unmeasured confounder would have to be associated with the treatment and outcome for the treatment-outcome association not to be causal.
- The data themselves, however, do not give an indication whether there is no unmeasured confounding.
- Thus, we study how robust the estimated associations are to potential unmeasured or uncontrolled confounding.
- Thus, a sensitivity analysis usually suggests the existence of an unmeasured confounder H and introduces a model where either the H-A association or the H-A association or both.
- Is this a science or an art?

Sensitivity analysis: example of a strategy

Motivation

- Suppose that we study a binary treatment A. Then, using law of total expectation and consistency (no exchangeability here),

$$\begin{aligned}
&\mathbb{E}(Y^a) \\
&= \mathbb{E}[\mathbb{E}(Y^a \mid L)] \\
&= \mathbb{E}[\mathbb{E}(Y \mid A = a, L)\pi(a, L)] + \mathbb{E}[\mathbb{E}(Y^a \mid A = 1 - a, L)\pi(1 - a, L)],
\end{aligned}$$

  which is the same argument as in Slide 366.

- The only counterfactual is the right hand term. If $Y^a \perp\!\!\!\perp A \mid L$, we have that

$$\mathbb{E}(Y^a \mid A = 1 - a, L = l) = \mathbb{E}(Y^a \mid A = a, L = l) = \mathbb{E}(Y \mid A = a, L = l),$$

  for all l.

Example continues

- However, if $Y^a \not\perp\!\!\!\perp A \mid L$, we cannot use the argument above, but we could specify a sensitivity parameter $\delta_a(l)$ as

$$\delta_a(l) = \mathbb{E}(Y^a \mid A = 1, L = l) - \mathbb{E}(Y^a \mid A = 0, L = l),$$

  which clearly is 0 if $Y^a \perp\!\!\!\perp A \mid L$.
  A simple example of a sensitivity function is

$$\delta_a(l) = \gamma_a a.$$

  If γ_a is positive, this function would say that individuals who received a would have higher risk of the outcome Y, even when adjusting for L.
  For example there could be an unmeasured (hidden) variable $A \leftarrow H \rightarrow Y$.

- In practice, we could do a sensitivity analysis by choosing a large number of values for γ_a.

- Then, when $\delta_a(l)$ is specified, we can identify $\mathbb{E}(Y^a)$ (next slide exercise)

Exercises to the sensitivity analysis

- We can use $\delta_a(l)$ for identification. To see this, suppose $a = 1$, use consistency,

$$\delta_1(l) = \mathbb{E}(Y \mid A = 1, L = l) - \mathbb{E}(Y^1 \mid A = 0, L = l),$$

and using the form of $\delta_1(l)$ we have

$$\begin{aligned}\mathbb{E}(Y^1) = &\mathbb{E}[\mathbb{E}(Y \mid A = 1, L)\pi(1, L)] \\ &+ \mathbb{E}[\{\mathbb{E}(Y \mid A = 1, L) - \delta_1(L)\}\pi(0, L)]\end{aligned}$$

- Indeed, it can be shown that $\delta_a(l)$ puts no restrictions on the observed data law $p(y, a, l)$.

Different approaches to sensitivity analysis

- In the spirit of Cornfield (1959), specify how a collection of unmeasured variables H affects the outcome of interest Y and how H affects A.
- In the suggested approach above we only specified a single relation, that is, a mean counterfactual outcome conditional on L:
  - We used few (in our simple example one) sensitivity parameter
  - We were agnostic about the structure of the hidden confounders H (i.e. whether they are binary, continuous, etc etc).
- On the other hand, Conrfield-like approaches can be useful when
  - H is a known confounder (say, smoking) that was not measured in the study
  - We somehow have reasons to know the association between H and the outcome and the treatment.