EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Causal Machine Learning

Causal Inference basis

Michael Knaus

March 13, 2023

I ask you for one more week of patience

Without causal inference, the outputs of causal ML methods are just numbers

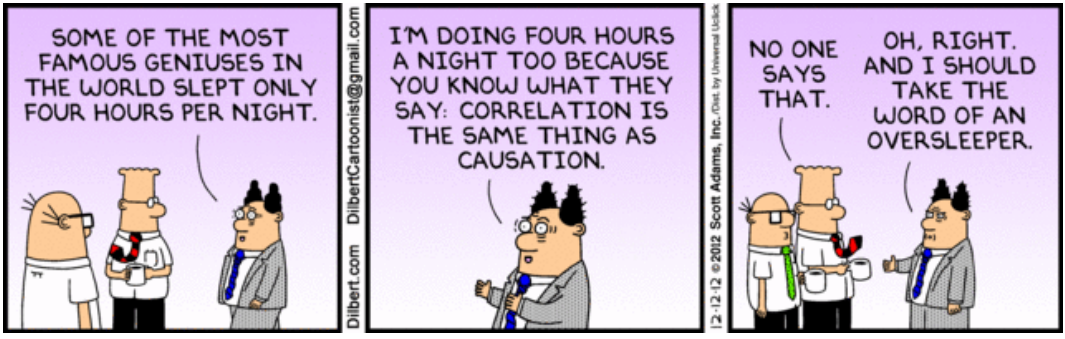
Plan for today

Crash course/recap in causal inference

1. Correlation vs. causation
2. We need a framework … or two
3. Potential outcomes
4. Experiments
5. Structural causal models
6. Single world intervention graphs
7. Outlook

Correlation vs. causation

We are the guys on the left



What's the deal?

Intuition by pictures

Source

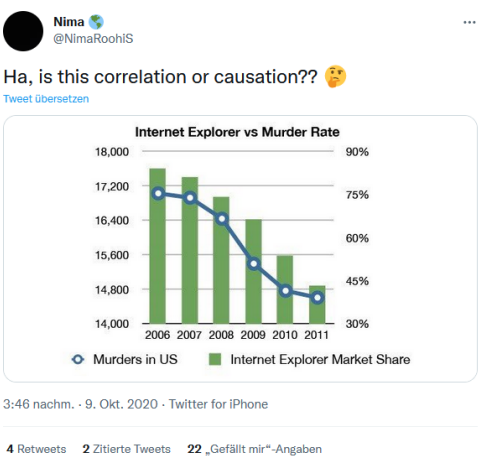
Intuition by pictures

Source

Intuition by pictures

Source

A bit more serious looking



Source

An admittedly very naive policy conclusion from this data analysis would be to abolish Internet Explorer to further decrease murder rates

More for potential entertainment

My favourite clip

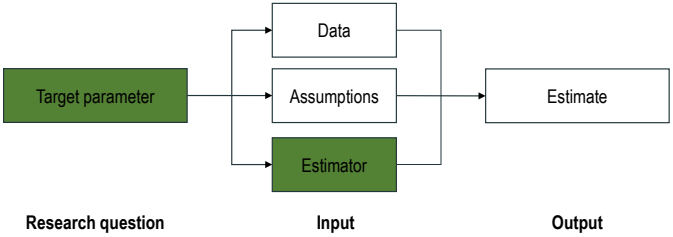
Spurious correlations

More birds

Plane

Recall from first week

I told you we will focus on extending the menu of parameters and estimators



BUT, we still need to understand the assumptions that are required for credible parameter estimates such that we can critically assess them in analyses

Only pushing buttons to produce numbers and stories around these numbers can be dangerous

Bad or no news for me?

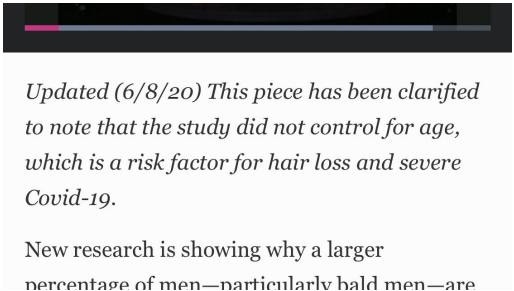
CORONAVIRUS | 130,196 views | Jun 6, 2020, 11:26am EDT

Bald Men At Higher Risk Of Severe Coronavirus Symptoms

Marla Milling Contributor ⓘ

Healthcare

I am a Forbes.com Contributor specializing in geriatric health and women's health articles.



Updated (6/8/20) This piece has been clarified to note that the study did not control for age, which is a risk factor for hair loss and severe Covid-19.

New research is showing why a larger percentage of men—particularly bald men—are

Source

Fortunately, serious data analysts know that we should not build decisions on such analyses

And so should you

We need a framework … or two

Frameworks

There are two prominent frameworks for causal inference and a bridge between them:

- Potential Outcomes (POs): Prevalent in economics and statistics
- Structual Causal Models (SCM) often represented as Directed Acyclic Graphs (DAGs): Prevalent in computer science and industry
- Single World Intervention Graphs (SWIGs): A (for me) very useful bridge between the two

In this course we focus on the (for economists currently) more familiar POs, but also introduce basic ideas of SCMs and SWIGs along the way

Causal inference pipeline on a high level

Regardless of the framework, two steps lead to an estimate of a causal effect:

1. **Identification:** Impose assumptions to express target parameter in terms of observable distributions
2. **Estimation:** Select and apply a suitable estimator to estimate target parameter in a sample of the observable distribution

For more nuanced pipelines, see e.g. Kate Hoffman or Peter Hull

But the general mantra is that **identification always comes before estimation**

This is holds in general, but is important to emphasize in this course to avoid tempting traps like ”I use a causal forest, so I am estimating a causal effect”

Be careful! We only estimate a causal effect if we identified it in the first step

Potential outcomes

Potential outcomes toy example (1/2)

Potential outcomes start with a thought experiment that there are different states of the world depending on the decisions taken

For the introduction, we focus on the case of

- two individuals $i = 1, 2$
- having access to two pills of aspirin $W \in \{no, P1, P2\}$
- where we are interested in the effect of aspirin on headache (Y)

Without additional assumptions the potential outcome of individual i depends on the treatment status of both individuals and the three possible treatment states

$$Y_i \left(\begin{array}{c} W_1 \\ W_2 \end{array} \right)$$

Potential outcomes toy example (2/2)

This results in seven potential outcomes per individual

$$Y_i\begin{pmatrix} no \\ no \end{pmatrix}; Y_i\begin{pmatrix} no \\ P1 \end{pmatrix}; Y_i\begin{pmatrix} no \\ P2 \end{pmatrix}; Y_i\begin{pmatrix} P1 \\ no \end{pmatrix}; Y_i\begin{pmatrix} P1 \\ P2 \end{pmatrix}; Y_i\begin{pmatrix} P2 \\ no \end{pmatrix}; Y_i\begin{pmatrix} P2 \\ P1 \end{pmatrix}$$

In total we conceive already 14 potential outcomes in this very simple example

The combinatoric explodes if we add more individuals and pills

$\Rightarrow$ We need more structure to make generalizable statements

Stable Unit Treatment Value Assumption

The Stable Unit Treatment Value Assumption (SUTVA) states that the potential outcomes do not depend

1. on the treatment status of other individuals (no interference)
2. on the particular treatment received (homogeneous treatments)

Then,

- The treatment can be collapsed into a binary indicator: $W \in \{0, 1\}$
- The potential outcome only depends on the treamtent status of the individual: $Y(W)$, i.e. in our example $Y_1(W_1)$ and $Y_2(W_2)$

SUTVA is standard

We need SUTVA to be able to define the standard target parameters

$\Rightarrow$ It is fundamental for what we do

The literature most of the time operates under the assumption that it holds

$\Rightarrow$ Most of the resources you'll find take it as given

We will do so as well from now on

However, be aware that it might be violated, e.g. if LinkedIn runs experiments to boost private messages

Potential outcomes (dream) world under SUTVA

- Binary treatment: $W \in \{0, 1\}$
- Potential outcome under treatment w: $Y(w)$
- Individual Treatment Effect (ITE): $\Delta = Y(1) - Y(0)$

i	$Y_i(1)$	$Y_i(0)$	Δ_i
1	0	1	-1
2	3	2	1
3	1	1	0
4	2	1	1
⋮	⋮	⋮	⋮

$\Rightarrow$ Under SUTVA, each individual has as many POs as there are treatment states (two in our case) $\Rightarrow$ it buys us a relatively simple dream world

Target parameters

Imposing SUTVA enables us to define parameter classics as (un-)conditional expectations of (differences) of POs

- Average potential outcome (APO): $\gamma_w \equiv \mathbb{E}[Y(w)]$
  - What is the expected outcome if everybody receives treatment w?
- Average Treatment Effect (ATE): $\tau_{ATE} \equiv \mathbb{E}[Y(1) - Y(0)] = \gamma_1 - \gamma_0$
  - What is the expected treatment effect in the population?
- Average Treatment Effect on the Treated (ATT): $\tau_{ATT} \equiv \mathbb{E}[Y(1) - Y(0) \mid W = 1]$
  - What is the expected treatment effect in the subpopulation actually receiving the treatment?
- Conditional Average Treatment Effect (CATE): $\tau(x) \equiv \mathbb{E}[Y(1) - Y(0) \mid X = x]$
  - What is the expected treatment effect for somebody with characteristics $X = x$?

Reality check

The parameters are defined with respect to hypothetical potential outcome distributions that exist in our imagination (or do they really exist? 🤷‍♂️)

In reality, we only observe $Y \Rightarrow$ no (w) $\Rightarrow$ no potential outcome (yet)

So, how can the observed outcomes help us to learn something about the unobservable causal effects? 🤔

SUTVA serves as link between potential outcomes and observable world

Under SUTVA, we can write

$$Y = Y(W) \text{ or } Y = (1 - W)Y(0) + WY(1) \tag{1}$$

$\Rightarrow$ We observe at least one potential outcome 🥳

$\Rightarrow$ Offers a glimpse into the dream world

(Hard) Reality

Only one potential outcome is observable:

	Partly observed		Unobserved	Observed	
i	$Y_i(1)$	$Y_i(0)$	Δ_i	W_i	Y_i
1	0	1	-1	0	1
2	3	2	1	1	3
3	1	1	0	0	1
4	2	1	1	1	2
⋮	⋮	⋮	⋮	⋮	⋮

$\Rightarrow$ The counterfactual potential outcome is missing $\Rightarrow$ ITE is never observed $\Rightarrow$ "fundamental problem of causal inference" (Holland, 1986)

Debunk (Naive) Group Comparisons

POs help us, e.g., to unpack why comparing means of two groups might not provide a causal effect (correlation vs. causation reloaded/formal):

First see how SUTVA allows us to link observed and hypothetical distributions

$$
\begin{aligned}
&\mathbb{E}[Y \mid W = 1] - \mathbb{E}[Y \mid W = 0] \\
&\overset{(1)}{=} \mathbb{E}[\overbrace{(1 - W)}^{=0} Y(0) + \overbrace{W}^{=1} Y(1) \mid W = 1] - \mathbb{E}[\overbrace{(1 - W)}^{=1} Y(0) + \overbrace{W}^{=0} Y(1) \mid W = 0] \\
&= \mathbb{E}[Y(1) \mid W = 1] - \mathbb{E}[Y(0) \mid W = 0]
\end{aligned}
$$

Debunk (Naive) Group Comparisons

Without further assumptions, we can decompose the mean comparison as

$$
\begin{aligned}
&\mathbb{E}[Y \mid W = 1] - \mathbb{E}[Y \mid W = 0] \\
&\quad = \mathbb{E}[Y(1) \mid W = 1] - \mathbb{E}[Y(0) \mid W = 0] \\
&\quad = \underbrace{\mathbb{E}[Y(1) - Y(0)]}_{\text{ATE}} \\
&\qquad + \underbrace{\mathbb{E}[Y(0) \mid W = 1] - \mathbb{E}[Y(0) \mid W = 0]}_{\text{confounding bias}} \\
&\qquad + \underbrace{(1 - \mathbb{E}[W])(\mathbb{E}[Y(1) - Y(0) \mid W = 1] - \mathbb{E}[Y(1) - Y(0) \mid W = 0])}_{\text{heterogeneous effect bias}}
\end{aligned}
$$

where $\tau_{ATU} \equiv \mathbb{E}[Y(1) - Y(0) \mid W = 0]$ is the ATE ON THE UNTREATED (ATU)

For compactness, we switch below to $\mu_{ww} = \mathbb{E}[Y(w) \mid W = w]$, e.g. $\mu_{10} = \mathbb{E}[Y(1) \mid W = 0]$

Note that by the law of iterated expectations

$$ATE = \mathbb{E}[Y(1)-Y(0)] = p\,\mathbb{E}[Y(1)-Y(0) \mid W = 1]+(1-p)\,\mathbb{E}[Y(1)-Y(0) \mid W = 0] = p\mu_{11}-p\mu_{01}+(1-p)\mu_{10}-(1-p)\mu_{00} \tag{2}$$

Then,

$$\begin{aligned}
\mathbb{E}[Y \mid W = 1] - \mathbb{E}[Y \mid W = 0] &\overset{(\dagger)}{=} \mathbb{E}[\underbrace{(1-W)}_{=0} Y(0) + \underbrace{W}_{=1} Y(1) \mid W = 1] - \mathbb{E}[\underbrace{(1-W)}_{=1} Y(0) + \underbrace{W}_{=0} Y(1) \mid W = 0] \\
&= \mathbb{E}[Y(1) \mid W = 1] - \mathbb{E}[Y(0) \mid W = 0] + ATE - ATE \\
&\overset{(\dagger)}{=} ATE + \mu_{11} - \mu_{00} - p\mu_{11} + \underbrace{p\mu_{01} + (1-p)\mu_{01}}_{=\mu_{01}} - (1-p)\mu_{10} + (1-p)\mu_{00} - (1-p)\mu_{01} \\
&= ATE + \mu_{01} - \mu_{00} + \underbrace{\mu_{11} - p\mu_{11}}_{(1-p)\mu_{11}} - (1-p)\mu_{10} + (1-p)\mu_{00} - (1-p)\mu_{01} \\
&= ATE + \mu_{01} - \mu_{00} + (1-p)[\underbrace{\mu_{11} - \mu_{01}}_{=ATT} \underbrace{-\mu_{10} + \mu_{00}}_{-ATU}] \quad \text{🤓}
\end{aligned}$$

Example: Covid vaccinations

Beginning of Covid vaccination campaigns where old and vulnerable are treated 💉 $= 1$, while young and healthy remained untreated 💉 $= 0$

$$\mathbb{E}[🏥 \mid 💉 = 1] - \mathbb{E}[🏥 \mid 💉 = 0]$$

$$= \underbrace{\mathbb{E}[🏥(1) - 🏥(0)]}_{<0}$$

$$+ \underbrace{\mathbb{E}[🏥(0) \mid 💉 = 1] - \mathbb{E}[🏥(0) \mid 💉 = 0]}_{>0}$$

$$+ \underbrace{(1 - \mathbb{E}[💉])(\mathbb{E}[🏥(1) - 🏥(0) \mid 💉 = 1] - \mathbb{E}[🏥(1) - 🏥(0) \mid 💉 = 0])}_{>0}$$

> 0 or at least $> \mathbb{E}[🏥(1) - 🏥(0)]$

Experiments

Experiments

Whether you call it randomized controlled trial (RCT) or A/B testing, randomizing the treatment is arguably the cleanest way to obtain causal effect estimates

Why does randomization work so well?

Randomness does not care about anything $\Rightarrow$ W can not be predicted

$\Rightarrow \mathbb{E}[W \mid X] = \mathbb{E}[W]$, i.e. any variables X that are not affected by the treatment are uninformative about the treatment assignment

We say W is independent of other variables

Most important for us, randomization of W implies that it is independent of potential outcomes

$$Y(w) \perp\!\!\!\perp W \text{ for all } W \in \{0, 1\} \tag{3}$$

Then,

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

In words: whether we condition on the factual, the counterfactual treatment or on nothing does not change the expected potential outcome

Identification of average effect in experiments

Under randomization of W, the mean comparison of slide 22 identifies the ATE:

$$\begin{aligned}\mathbb{E}[Y \mid W=1] - \mathbb{E}[Y \mid W=0] &\overset{(1)}{=} \mathbb{E}[Y(1) \mid W=1] - \mathbb{E}[Y(0) \mid W=0] \\ &\overset{(3)}{=} \underbrace{\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]}_{=\tau_{ATE}}\end{aligned}$$

We have expressed the inherently unobservable ATE in terms of observable quantities $\Rightarrow$ identification of target parameter

We need to make assumptions to remove confounding and heterogeneity bias

Remark: In an experiment $\tau_{ATE} = \tau_{ATT} = \tau_{ATU}$ because (3) implies that $\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1) - Y(0) \mid W=1] = \mathbb{E}[Y(1) - Y(0) \mid W=0]$

Estimation of average effect

The next step after identification where we need to argue that randomization was successful and that SUTVA holds is to estimate the effect

The easiest way to estimate the parameter is a simple mean comparison:

$$\hat{\tau}_{ATE} = \frac{1}{\sum_i W_i} \sum_i W_i Y_i - \frac{1}{\sum_i (1 - W_i)} \sum_i (1 - W_i) Y_i$$

or equivalently an OLS regression of the form

$$Y = \alpha + \tau W + U_{Y \sim W}$$

Identification of heterogeneous effects in experiments

Also conditional independence holds in a proper randomized experiment

$$Y(w) \perp\!\!\!\perp W \mid X \text{ for all } W \in \{0, 1\} \tag{4}$$

such that the Conditional Average Treatment Effect (CATE) is also identified

$$\tau(x) \equiv \mathbb{E}[Y(1) - Y(0) \mid X = x] = \mathbb{E}[Y \mid W = 1, X = x] - \mathbb{E}[Y \mid W = 0, X = x]$$

Note that we condition on variables that are not affected by the treatment

This is implied by writing X and not $X(w)$

We will learn estimators that target CATEs later in this course

Structural causal models

Start with a model

As the name suggest, we start with a model that is expressed as a system of equations

In the case of an RCT, the model reads like this:

$$W = f_W(U_W)$$
$$Y = f_Y(W, U_Y)$$

W is a function of random noise that is, e.g., generated by a 🪙 flip

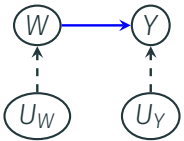
Y is a function of the treatment and noise that is independent of the coin flip U_Y

Note that SUTVA is already encoded in this structure as W is a scalar (homogeneous treatment) and Y is not a function of others (no interference)

Graphical representation

The Directed Acyclic Graph (DAG) of an RCT looks like this

where the independent noise components are usually surpressed, but for completeness you could draw them also explicitly



See Hünermund & Barenboim (2019) for an introduction targeting economists

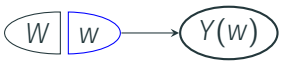
Single world intervention graphs

A synthesis

Richardson & Robins (2013) provide a tool to find the potential outcomes that are implied by a structural causal model (see also the primer for an introduction)

A SWIG splits the treatment node and sets the treatment to a particular value

All outgoing arrows become now potential outcomes



Most importantly we can plug in the machinery of d-separation developed for DAGs to find the implied (conditional) independences

See identification notebook "DAG and SWIG for RCTs" for an illustration

I like SWIGs

A thorough introduction of SWIGs is beyond the scope of this course

However, I encourage you to have a look if you (like me) think in POs and are confused by claims that DAGs encode the (conditional) independences we are familiar with in the PO framework

I never saw them because DAGs do not contain POs

SWIGs make POs living in DAGs visible and allow to plug-in their powerful machinery to find (conditional) independences with respect to POs also in settings that are not as easy as an RCT

Outlook

Outlook

We will focus on two more research designs in the following:

- Conditional independence / selection-on-observables / unconfoundedness / exogeneity /...
- Instrumental variables

For them a bunch of causal ML are available

We will not look into difference-in-differences and regression discontinuity designs as causal ML methods for them are still in their infancies

However, they most likely build on the same principles that we will establish in the following

If causal inference is new for you (and even if not), I highly recommend to read Chapters 3 & 4 of Causal Inference - The Mixtape by Scott Cunningham

Ceterum censeo a fancy method alone is not a credible identification strategy

$\Rightarrow$ separate identification and estimation