Causal inference, week 2
Potential outcomes

Daniel Stegmueller

Duke University

Causal relations are....

- Asymmetrically structured
- Counterfactual definition
  - D is a cause of Y if Y would not have occurred, had D not been present
- Comparative definition:
  - The causal effect of D on Y is the change in Y due to D
- Defined in terms of (possibly hypothetical) manipulations

Effects of causes

- Effects of causes (“forward causal inference”), not causes of effects
- Note the (possibly unfamiliar) perspective:
  - You don’t study *determinants* of Y, but the *effect* of X on Y
  - “Never ask Why? Only ask What if” (Rubin)
  - One important implication is that such designs usually focus on one (few) quantities of causal interest

One framework, two formalisms

- Potential outcomes
- Graphical models
- They are formally equivalent (Pearl 2000 ch.7)

Potential outcomes

Potential outcomes framework

- Assume (for now) two different, precisely defined causal states
- A *potential* outcome is the true value of the outcome of interest that *would* result from exposure to the alternative causal state
- Denote the potential outcome states of a unit i as y_{i0} and y_{i1} for "treatment" and "control", respectively
- Both hypothetical/potential outcomes exist at the same time for the same unit
- Conceptually, we can thus define unit-level treatment effects by contrasting the treatment states
- Usually this is $y_{i1} - y_{i0}$

Discussion points:

- What are the causal states of "bureaucracy" or "economic status"?
- Ceteris paribus causal states & structural invariance

Potential outcomes framework

Language / notation:

- Units (i)
- Treatment (D)
- Outcomes (Y)
- Outcomes in treatment states, (Y_0, Y_1)
- Covariates/Confounders (X)
- Treatment effect (Δ)

An aside on language: what is an *experiment*?

- We will use language borrowed from the experimental tradition (e.g., control group)
- But a large fraction of the models/methods discussed later will be using observational data
- What distinguish experiments from observational data is the element of *control*
    - *An experiment*: system under study is under the control of the investigator. This includes the materials/subject studies, the assignment of treatments/manipulations, the measurement procedures
    - *An observational study*: some features (esp., but not limited to, treatment assigned) are not under investigator's control

[This follows the definition in Cox and Reid, 2000. The Theory of the Design of Experiments. Chapman & Hall.]

Counterfactual Theory: Potential outcomes

Group	Y_1	Y_0
$D=1$	Observed Y	counterfactual
$D=0$	counterfactual	Observed Y

$$Y = DY_1 + (1-D)Y_0$$

or, for any given unit i:

$$y_i = d_i y_{i1} + (1-d_i) y_{i0}$$

The fundamental problem of causal inference

- Each unit only observed in one treatment state

$$d_i = 0 \text{ or } d_i = 1$$

- Each unit has one observed outcome

$$y_{i0} \text{ or } y_{i1}$$

- We only ever observe

$$y_i = d_i y_{i1} + (1 - d_i) y_{i0}$$

- But treatment effect is difference for same unit

$$\delta_i \equiv y_{i0} - y_{i1}$$

This is the fundamental problem of causal inference aptly discussed in Holland, 1986. Statistics and Causal Inference. JASA 81, p. 945–970.

Common aggregate quantities of interest

- Focus is usually shifted to aggregate causal effects
- Most well-known & widely used: average treatment effect in the population of interest

$$E[\delta] = E[Y_1 - Y_0] = E[Y_1] - E[Y_0] \qquad \textbf{(ATE)}$$

- Note that δ is a random variable (and not necessarily constant in the population)
- Two conditional treatment effects are often sought:

$$E[\delta|D=1] = E[Y_1 - Y_0|D=1] \qquad \textbf{(ATT)}$$

$$E[\delta|D=0] = E[Y_1 - Y_0|D=0] \qquad \textbf{(ATC)}$$

Other quantities can be defined (and may even be of greater importance), e.g., proportion of individual-treatment effect distribution that is less than zero

SUTVA

- PO approach gains its simplicity/elegance by maintaining the fundamental assumption of "stable unit treatment value"

  See: Rubin, 1986. "Which Ifs Have Causal Answers (Comment on Holland)" JASA 81, p.961–962.

- Assumes that the *potential outcomes* of individual A are not affected by *potential changes* in the treatment exposure of individual B

- Paraphrasing Rubin

  > the value of Y for a unit when exposed to a treatment will be the same no matter (i) what mechanism is used to assign the treatment to the unit, (ii) what treatments the other units receive

SUTVA

- Example: simple experiment, assigning three individuals to treatment and control under two different assignment rules:
  - 1 treated, 2 control
  - 2 treated, 1 contol

	Patterns			y_d	
	1	2	3	T	C
1 treated					
Individual 1	T	C	C	1	0
Individual 2	C	T	C	1	0
Individual 3	C	C	T	1	0
2 treated					
Individual 1	T	C	T	2	0
Individual 2	T	T	C	2	0
Individual 3	C	T	T	2	0

SUTVA

What to do??

- Be explicit about the assumption and likely limitations in any given application
- Think about impact of "no-macro-effects" assumption. Limited intervention? Limited effect sizes? etc.
- Alternative approaches model causal effect explicitly as function of treatment assignment patterns [HARD]

Naive estimation of treatment effects in an observational study

- Denote by π the proportion of individuals taking/selecting the treatment (and $1-\pi$ control)
- π is fixed in the population (by the sum of individual choices) and unknown to the investigator
- Take a random sample of size N of this population
- Denote by $E_N[x]$ the sample mean of quantity x (i.e., $N^{-1}\sum_{i=1}^{N} x$)
- A naive estimate of the average causal effect is the difference in sample means of the treated and the control individuals:

$$E_N[y_i|d_i=1]-E_N[y_i|d_i=0]$$

- Your intuition tells you that this is not consistent. Why?

Naive estimation of treatment effects in an observational study

- The naive estimator converges to the difference

$$E[Y_1|D=1]-E[Y_0|D=0]$$

which is usually not equal to the causal average effect we seek

- Remember our definition of the ATE, $E[\delta]=E[Y_1]-E[Y_0]$
- Rewrite this as the decomposition

$$\begin{aligned}E[\delta]=&\left(\pi E[Y_1|D=1]+(1-\pi)E[Y_1|D=0]\right)-\\&\left(\pi E[Y_0|D=1]+(1-\pi)E[Y_0|D=0]\right)\end{aligned}$$

- We have 5 unknowns
  - proportion who selects/takes treatment, π
  - conditional expectations of potential outcomes

Naive estimation of treatment effects in an observational study

What can estimate three ...

- The proportion π because (for large samples)

$$E_N[d_i] \xrightarrow{p} \pi$$

- The average outcome under treatment for those in the treatment group

$$E_N[y_i|d_i = 1] \xrightarrow{p} E[Y_1|D = 1]$$

- The average outcome under control for those in the control group

$$E_N[y_i|d_i = 0] \xrightarrow{p} E[Y_0|D = 0]$$

... and miss two

- The *counterfactual* conditional expectations

$$E[Y_1|D = 0] \qquad \text{and} \qquad E[Y_0|D = 1]$$

Dangers of naive estimation

- Say you are interested in estimating the ATE. What bias can you expect?
- Take our decomposition from above and rewrite

$$\begin{aligned} E[Y_1|D=1]-E[Y_0|D=0] =& E[\delta]+ \\ & (E[Y_0|D=1]-E[Y_0|D=0])+ \\ & (1-\pi)(E[\delta|D=1]-E[\delta|D=0]) \end{aligned}$$

- Let's consider the two sources of expected bias
  - Baseline differences: the difference in outcomes at baseline (absent the treatment) for those in the treatment group and the control group
  - Effect differences: differential treatment effect for those in the treatment versus the control group. It is scaled by the proportion of untreated individuals.

Dangers of naive estimation

- Consider the following simple example of college attendance and earnings (again...)

Group	$E[Y_1]$	$E[Y_0]$
College (D=1)	**\$1,000**	\$600
Not college (D=0)	\$800	**\$500**

Fraction of indiv. with college degrees: 0.25

- We have the following causal quantities

$$ATT = \$400, ATC = \$300, ATT = 0.25 * 400 - 0.75 * 300 = \$325$$

- The naive observational estimator yields an estimated effect of \$500
- It is thus biased for all three causal quantities
  - ATE: \$500 vs \$325
  - ATT: \$500 vs \$400
  - ATC: \$500 vs \$300

Dangers of naive estimation

- The expected bias for the average causal effect is \$175
- In terms of our decomposition equation above:
  - $E[\delta] = \$325$
  - $(E[Y_0|D=1] - E[Y_0|D=0]) = \$600 - \$500 = \$100$
  - $(1-\pi)(E[\delta|D=1] - E[\delta|D=0]) = 0.75 * (\$400 - \$300) = \$75$

Group	$E[Y_1]$	$E[Y_0]$
College (D=1)	**\$1,000**	\$600
Not college (D=0)	\$800	**\$500**

Fraction of indiv. with college degrees: 0.25

Remark: Some identified causal quantities

- ► Unbiased estimates of ATE can be produced under assumptions

  A1 $E[Y_1|D=1]=E[Y_1|D=0]$

  A2 $E[Y_0|D=1]=E[Y_0|D=0]$

- ► *Very* unlikely to be met in practice

- ► In some situations, one of the two assumptions might be defensible. Then some quantities are identified:

  - A1 true, A2 false: naive estimator biased for ATE, but unbiased for ATC

    (because estimator converges to $E[Y_1|D=0]-E[Y_0|D=0]$; insert equality into decomposition above to see)

  - A1 false, A2 true: naive estimator biased for ATE, but unbiased for ATT

    (because estimator converges to $E[Y_1|D=1]-E[Y_0|D=1]$)

Potential outcomes are missing data

Potential outcomes are missing data

D	Y^1	Y^0	Y	δ
1	10	?	10	?
1	10	?	10	?
1	5	?	5	?
1	5	?	5	?
0	?	7	7	?
0	?	7	7	?
0	?	4	4	?
0	?	4	4	?

- Missingness mechanisms
  - Missing completely at random (MCAR)
  - Missing at random (MAR)
  - Not missing at random (NMAR)

MCAR mechanism

- Nothing (except for the observed Y_0s or Y_1s) that tells me what a missing Y_0 or Y_1 would have been
- nothing (except the observed Y_0s or Y_1s) that tells me how to impute them
- use just those observed values for inference
- ==> average unobserved outcomes can be inferred from observed outcomes

  (average: there might be effect size variation, but we can’t know what they are)

MCAR mechanism

- Example: causal effect heterogeneity but no baseline differences

D	Y^1	Y^0	Y	δ
1	10	7	10	3
1	10	7	10	3
1	5	4	5	1
1	5	4	5	1
0	10	7	7	3
0	10	7	7	3
0	5	4	4	1
0	5	4	4	1

- Naive estimate= 7.5-5.5, ATE=ATT=ATC=2

MCAR mechanism

- Observed outcomes inform us about unobserved outcomes (on average)

D	Y^1	Y^0	Y	δ
1	10		10	
1	10	$E[Y^0]=5.5$	10	$E[\delta \mid D=1]=2$
1	5	(from $Y^0 \mid D=0$)	5	(the ATT)
1	5		5	
0		7	7	
0	$E[Y^1]=7.5$	7	7	$E[\delta \mid D=0]=2$
0	(from $Y^1 \mid D=1$)	4	4	(the ATC)
0		4	4	

MAR mechanism

- Something predicts what those missing values would have been
- Missing at Random
  - Covariate C contains information about missing Y_0 or Y_1
  - Imputation: “fill in” missing Y_0 or Y_1 using C
  - Causal inference: (sometimes) “condition on C”. Analyze within values of C and average results according to $P(C)$
  - Question is: when to condition? Two examples...

MAR mechanism

True causal effects as before, but now with C. Condition on it?

D	C	Y^1	Y^0	Y	δ
1	1	10	$E[Y^0 \mid C=1]=7$	10	$E[\delta \mid C=1]=3$
1	1	10	(from $Y^0 \mid C=1, D=0$)	10	
1	0	5	$E[Y^0 \mid C=0]=4$	5	$E[\delta \mid C=0]=1$
1	0	5	(from $Y^0 \mid C=0, D=0$)	5	
0	1	$E[Y^1 \mid C=1]=10$	7	7	$E[\delta \mid C=1]=3$
0	1	(from $Y^1 \mid C=1, D=1$)	7	7	
0	0	$E[Y^1 \mid C=0]=5$	4	4	$E[\delta \mid C=0]=1$
0	0	(from $Y^1 \mid C=0, D=1$)	4	4	

MAR mechanism

- We could but we don't need to
- Condition anyway...

$$\begin{aligned} ATE &= E(\delta|C=1)\times P(C=1)+E(\delta|C=0)\times P(C=0) \\ &= 3\times 0.5+1\times 0.5=2 \end{aligned}$$

- Why?
- Because C is distributed the same way over D ('treatment' and 'control' groups)
- Other treatment effects generated by averaging over $P(C|D=1)$ or $P(C|D=0)$

MAR mechanism [should condition]

A different situation:

D	C	Y^1	Y^0	Y	δ
1	1	10	7	10	3
1	1	10	7	10	3
1	0	5	4	5	1
0	1	10	7	7	3
0	0	5	4	4	1
0	0	5	4	4	1

naive estimate, ATE, ATT, ATC are all different

MAR mechanism [should condition]

D	C	Y^1	Y^0	Y	δ
1	1	10	7	10	3
1	1	10	7	10	3
1	0	5	4	5	1
0	1	10	7	7	3
0	0	5	4	4	1
0	0	5	4	4	1

$\text{ATE} = 0.5 \times 3 + 0.5 \times 1 = 2$ $\qquad P(C=1)$ still 0.5 overall

$\text{ATT} = 0.66 \times 3 + 0.33 \times 1 = 2.33$ $\qquad P(C=1|D=1) = 2/3$

$\text{ATC} = 0.33 \times 3 + 0.66 \times 1 = 1.66$ $\qquad P(C=1|D=0) = 1/3$

Naive = 8.33-5 = 3.33

We *must* condition on C to identify the causal effects

NMAR mechanism

- Something informs us about missing Y_0 or Y_1
- We don't know what it is (or it is D itself).
- Thus, can't impute missing Y_0, Y_1
- What to do?
    - Find some C to make problem MAR
    - Build seperate model for D
    - Run an experiment, find an "instrument"
    - Present naive estimate and admit defeat....

In a nutshell...

- Treatment assignment ideally independent of (functions of) potential outcomes

$$(Y_1, Y_0) \perp D$$

  - Equivalent to MCAR missing data
  - Most straightforwardly achieved by having D independent of everything (i.e., randomized assignment)

- The next best thing is conditional independence given C (i.e., randomized within some values of C)

$$(Y_1, Y_0) \perp D|C$$

  - Equivalent to MAR missing data

All in on controls??

- It seems as if it never hurts (and often helps) to condition on any available C.
- **This is incorrect.**
- Argument is easy to see in other formalism for thinking about inference ..... next week.