ETH zürich

Counterfactuals, potential outcomes and estimation

Causality

Christina Heinze-Deml

Spring 2019

Announcements

- Series 4 will be uploaded later today
- Next week:
  - Normal lecture from 10-11
  - In-class exercise from 11-12
    - Jupyter notebook and R
    - Can also ask questions about the course material and the series
- Semester feedback

Last week

- Instrumental variables
- Transportability

- Course outline
  - Background and framework
  - Using the known causal graph structure to identify and estimate causal effects
  - Causal structure learning

Today

- Counterfactuals
- Potential outcomes
- Estimation

- Course outline
  - Background and framework
  - Using the known causal graph structure to identify and estimate causal effects
  - Causal structure learning

Counterfactuals

- Not all causal questions can be expressed with $p(y|do(x))$
  - E.g., what fraction of the healthy untreated population would have gotten the disease had they been treated?
  - Retrospective thinking

- Consider SEM

$$Z \leftarrow f_Z(N_Z)$$
$$X \leftarrow f_X(Z, N_X)$$
$$Y \leftarrow f_Y(X, N_Y)$$

- Unit-level counterfactual reasoning analyzes relations such as "Y would be y had X been x in situation $N = n$"

Examples

Counterfactuals

- Noise as “unobserved uncertainty-producing variables” or “background variables”
- Unit-level counterfactuals may earn predictive value
  - When noise remains constant; or
  - When noise can be observed sometime in the future
    - See eye-doctor example

- Often this is not the case and many unit-level counterfactual statements cannot be falsified
  - Cannot observe $Y|do(X = 1)$ and $Y|do(X = 0)$ for the same individual at the same time

Potential outcomes

- Also known as “Rubin causal model”
- Often used in applied data analysis
  - Biostatistics
  - Econometrics

- Often used without graphs
- Frameworks can be combined with so-called Single-World Intervention Graphs (SWIGs) (Richardson and Robins (2013))

- Now: look at basic notation and concepts

Potential outcomes with binary treatment

- For binary treatment X and response Y, two potential outcome variables:
  - $Y_i(x = 0)$ (shorthand: $Y_i(x_0)$):
    value of Y that would be observed for a given unit i if assigned $X = 0$
  - $Y_i(x = 1)$ (shorthand: $Y_i(x_1)$):
    value of Y that would be observed for a given unit i if assigned $X = 1$
  - Unit-level reasoning

- Example:
  - X: flu vaccine
  - Y_i: time until individual gets the flu
  - $Y_i(x = 0)$: time until individual would get the flu if they did not receive the flu vaccine
  - $Y_i(x = 1)$: time until individual would get the flu if they received the flu vaccine

Potential outcomes with binary treatment

- For binary treatment X and response Y, two potential outcome variables:
  - $Y_i(x = 0)$ (shorthand: $Y_i(x_0)$):
    value of Y that would be observed for a given unit i if assigned $X = 0$
  - $Y_i(x = 1)$ (shorthand: $Y_i(x_1)$):
    value of Y that would be observed for a given unit i if assigned $X = 1$

- $P(Y(x) = y) \neq P(y \mid x)$ in general – causation is not association

- Counterfactual outcomes are the ones that would have been observed, had the treatment been different
  - If my treatment was $x = 1$, my counterfactual outcome is $Y_i(x = 0)$

Example

- X: flu vaccine
- Y_i: time until individual gets the flu

- I got the vaccine and did not get sick.
  - My actual treatment was $x = 1$.
  - My observed outcome was $Y_i = Y_i(x = 1)$.

- Had I not gotten the vaccine, would I have gotten sick?
  - My counterfactual treatment is $x = 0$.
  - My counterfactual outcome is $Y_i(x = 0)$.

Potential outcomes vs counterfactual outcomes

- Before the treatment decision is made, any outcome is a potential outcome
- After the study, there is an observed outcome and a counterfactual outcome

- Sometimes the terms potential outcome and counterfactual outcome are used interchangeably

Potential outcomes with binary treatment

- Unit-level causal effect: $Y_i(x_1) - Y_i(x_0)$
- Average of unit-level causal effects:

$$\frac{1}{n}\sum_i Y_i(x_1) - Y_i(x_0)$$

  - Cannot be computed directly

- Fundamental problem of causal inference:
  Can only observe one potential outcome for each unit.
- With certain assumptions, can estimate population level causal effects

Potential outcomes with binary treatment and binary response

- Binary treatment X and binary response Y, $Y = 1$ indicates recovery
- Have four response "types":

$Y_i(x_0)$	$Y_i(x_1)$	**Name**
0	0	Never recover
0	1	Helped
1	0	Hurt
1	1	Always recover

Potential outcomes with binary treatment and binary response

- Consider 5 units (individuals)
- Assignment to treatments

Unit	Potential outcomes		Observed	
	$Y_i(x_0)$	$Y_i(x_1)$	X	Y_i
1	0	1	1	
2	0	1	0	
3	0	0	1	
4	1	1	1	
5	1	0	0	

Potential outcomes with binary treatment and binary response

- Observed outcomes

Unit	Potential outcomes		Observed	
	$Y_i(x_0)$	$Y_i(x_1)$	X	Y_i
1	0	1	1	1
2	0	1	0	0
3	0	0	1	0
4	1	1	1	1
5	1	0	0	1

Potential outcomes with binary treatment and binary response

- Can only observe one potential outcome for each unit
- Causal inference as missing data problem

Unit	Potential outcomes		Observed	
	$Y_i(x_0)$	$Y_i(x_1)$	X	Y_i
1	?	1	1	1
2	0	?	0	0
3	?	0	1	0
4	?	1	1	1
5	1	?	0	1

Potential outcomes with binary treatment and binary response

- Average causal effect

$$ACE = E\big(Y(x_1) - Y(x_0)\big) = p(\text{Helped}) - p(\text{Hurt})$$

  - Difference in % recovery if everyone treated ($X = 1$) vs. if no one treated ($X = 0$)

- If treatment X assigned randomly, then $X \perp\!\!\!\perp Y(x_0)$ and $X \perp\!\!\!\perp Y(x_1)$. Hence

$$\begin{aligned} E\big(Y(x_1) - Y(x_0)\big) &= E(Y(x_1)|X = 1) - E(Y(x_0)|X = 0) \\ &= E(Y|X = 1) - E(Y|X = 0) \end{aligned}$$

Assumptions

- SUTVA: Stable Unit Treatment Value Assumption
- Consistency
- Ignorability: “no unmeasured confounders”
- Positivity

Assumptions

- SUTVA: Stable Unit Treatment Value Assumption
  - No interference between units, no contagion
  - Allows to write potential outcome for the i-th person in terms of only that person's treatment
- Consistency
  - $Y = Y(x)$ if $X = x \ \forall x$
- Ignorability: "no unmeasured confounders"
  - Given covariates Z, treatment assignment is independent from the potential outcomes
    - $X \perp\!\!\!\perp Y(x_0) \mid Z$ and $X \perp\!\!\!\perp Y(x_1) \mid Z$
- Positivity
  - For every z, treatment assignment is not deterministic
    - $P(X = x|Z = z) > 0 \ \forall x, z$

Observed data and potential outcomes

- $E(Y(x)|\ Z = z) = E(Y(x)|X = x,\ Z = z)$ by ignorability
- $E(Y(x)|X = x,\ Z = z) = E(Y|X = x,\ Z = z)$ by consistency

- So $E(Y(x)|\ Z = z) = E(Y|X = x,\ Z = z)$
  - $E(Y|X = x,\ Z = z)$ involves only observed data

Estimation methods – Setting

- X is binary: treatment vs control
- Interested in average causal effect
  - In *do*-notation:

$$ACE = E(Y|do(X = 1)) - E(Y|do(X = 0))$$

  - In potential outcomes notation:

$$ACE = E\big(Y(x_1) - Y(x_0)\big)$$

- Assume observed control variables Z form a valid adjustment set
  - The following methods do not address the question of identification
  - They are about estimation provided that observed covariates Z form a valid adjustment set

Matching

- Idea:
  - Match individuals in the treatment group ($X = 1$) to individuals in the control group ($X = 0$) on the covariates Z
  - Create a dataset with these matched pairs and perform outcome analysis
- Example:
  - Say older people are more likely to receive the treatment $X = 1$
    - At younger ages, there are more people with $X = 0$
    - At older ages, there are more people with $X = 1$
  - In a randomized trial, for any age, there should be about the same number of treated and untreated people
  - By matching treated people to control people of the same age, there will be about the same number of treated and controls at any age

Matching

* Suppose Z = hypertension (high blood pressure)

* Suppose hypertensive people more likely to be treated than people wo hypertension

After matching:

$P(X=1 \mid \text{"red"}) = 0.5$

$P(X=1 \mid \text{"blue"}) = 0.5$

Pre-matching

Treated	Control
$\frac{2}{3}$ red	$\frac{2}{7} \approx 29\%$ red

⇒ imbalance

Post-matching

Treated	Control
$\frac{2}{3}$ red	$\frac{2}{3}$ red

⇒ balance

Matching

- If exact matching not possible, need metric of closeness to find matches
  - E.g., use Mahalanobis distance

$$D(Z_i, Z_j) = \sqrt{(Z_i - Z_j)^T S^{-1} (Z_i - Z_j)}$$

where S is the sample covariance matrix of Z and Z_i are the covariates for subject i

  - Accounts for different scales of the covariates

Matching

- Greedy matching
  1. Randomly order list of treated subjects and control subjects
  2. Start with the first treated subject and find the control subject with the smallest distance
  3. Remove the matched pair from the lists of available subjects
  4. Move to the next treated subject and find the control subject with the smallest distance
  5. Repeat steps 3 + 4 until all treated subjects are matched

- Optimal matching
  - Greedy matching does not lead to the smallest total distance
  - Optimal matching computationally demanding
    - Still feasible for ~1 millon treatment-control pairings (e.g., 1000 treated subjects, 1000 controls)

Matching

- After matching, need to assess "covariate balance"
  - "Table 1" with mean and standard deviations and standardized mean differences
  - Standardized mean difference

$$\text{smd} = \frac{\bar{Z}_{\text{treatment}} - \bar{Z}_{\text{control}}}{\sqrt{\frac{s^2_{\text{treatment}} + s^2_{\text{control}}}{2}}}$$

    - Rules of thumb:
      - Values < 0.1 indicate "adequate balance"
      - Values $0.1 - 0.2$ are "not too alarming"
      - Values > 0.2 indicate "serious imbalance"

Matching

- Advantages:
  - Matching will reveal lack of overlap in covariate distribution (violations of positivity)
    - Positivity requires $P(X = x|Z = z) > 0 \; \forall x, z$
  - Once matched, can be treated as if from a randomized trial

- Disadvantages:
  - Won't be able to match exactly on the full set of covariates, especially when these are high-dimensional
  - Discards data

- Many extensions exist

Propensity scores

- Assume
  - Z is a valid adjustment set
  - Treatment X is binary

- Probability of receiving treatment given Z: $\pi = P(X = 1|Z = z)$
  - π is called the propensity score

- Can show:
  - $X \perp\!\!\!\perp Z|\pi$
  - If Z satisfies the adjustment criterion, then π also satisfies the adjustment criterion

Rosenbaum and Rubin (1983)

Propensity scores

- Reduces an arbitrarily large set of control variables Z to a single number
- If π has much lower dimension than Z, might be better to use

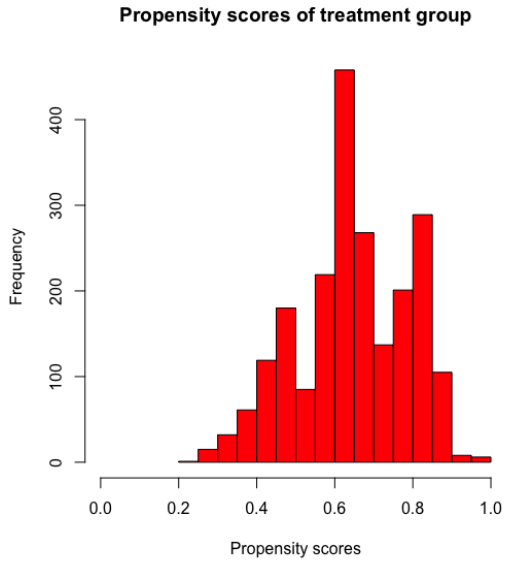
- $\pi = P(X = 1|Z = z)$ needs to be modeled and estimated
  - Most commonly: logistic regression
  - Modelling $P(X = 1|Z = z)$ can be a high-dimensional regression problem itself if Z is high-dimensional

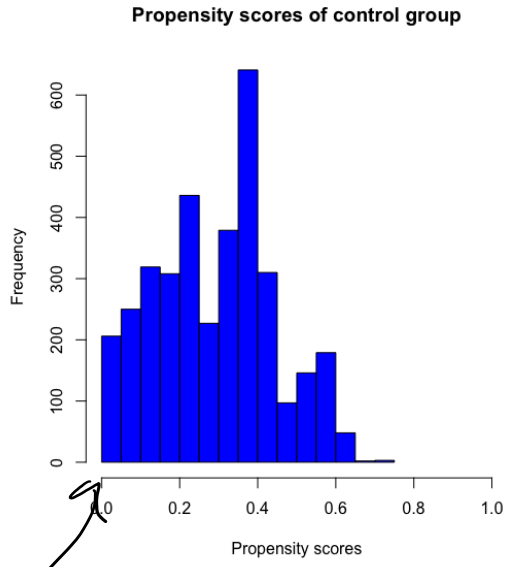
Propensity score matching

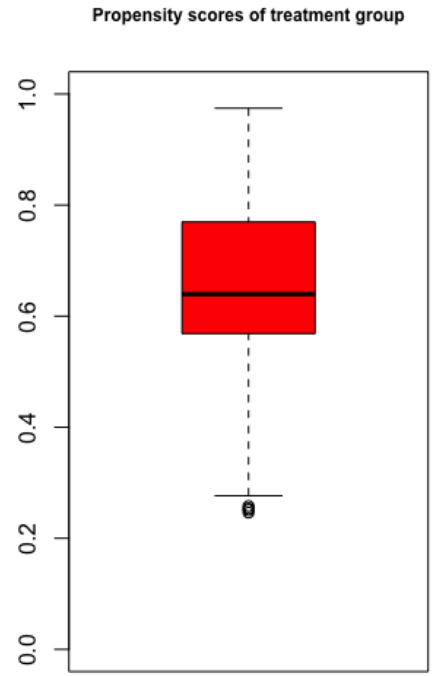
- Disadvantage of matching on Z
  - If Z is large, many values of Z will have few or no individuals at all → no exact matches
- Match on propensity scores π
  - Compares each treated individual with one who was just as likely to have received the treatment but did not
  - On average, the differences between such matched individuals must be due to the treatment
- Typically easier to find matches on π than on Z
  - Could have same value of π but different values of Z

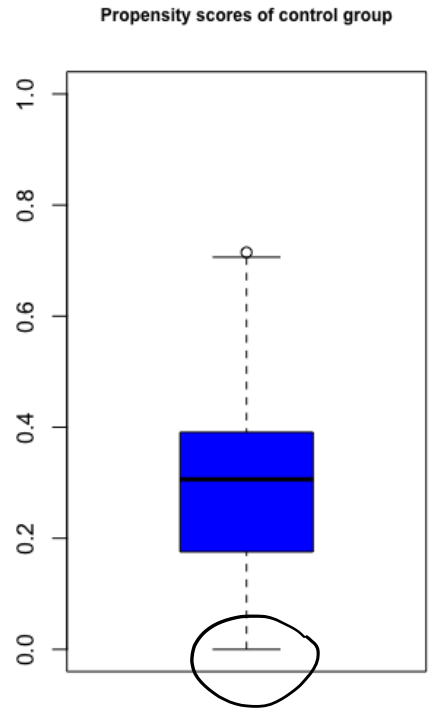
Clicker question – Propensity scores

- Positivity: For every z, treatment assignment is not deterministic
  - $P(X = x|Z = z) > 0 \; \forall x, z$









$P(X=1|Z=z)=0$ for some z

Inverse probability weighting

- Example
  - Single binary confounder Z
  - Suppose propensity score $P(X = 1|Z = 1) = 0.1$
    - Among people with $Z = 1$, only 10% receive the treatment
  - Suppose propensity score $P(X = 1|Z = 0) = 0.8$
    - Among people with $Z = 0$, 80% receive the treatment

Inverse probability weighting

- Idea: rather than match, use all data but downweigh/upweigh observations
  - Weighting by the inverse of the probability of treatment received
  - For treated: weigh by the inverse of the propensity score $\pi = P(X = 1|Z)$
  - For control: weigh by the inverse of $1 - \pi = P(X = 0|Z)$

- Known as inverse probability of treatment weighting (IPTW)

Example

Inverse probability weighting

- Estimator

$$\hat{E}(Y|do(X=1)) = \frac{1}{n}\sum_i Y_i \, 1\{X_i = 1\} w_i$$

  where $w_i = \frac{1}{\hat{\pi}_i} = \frac{1}{\hat{P}(X=1|Z_i)}$

- Equivalently for $\hat{E}(Y|do(X=0))$

- If propensity score $\pi_i = P(X=1|Z_i)$ is very small, weight will be very large
- Small estimation errors in $\hat{\pi}_i$ can lead to large estimation errors in $\hat{E}(Y|do(X=x))$

Inverse probability weighting

- More generally, consider
  - Observational distribution $p(x_{\boldsymbol{V}})$
  - Interventional distribution $p(x_{\boldsymbol{V}}|do(X_k \leftarrow \tilde{N}_k))$ with imperfect intervention
    - Shorthand: $p(x_{\boldsymbol{V}}|do(X_k \leftarrow \tilde{N}_k)) = \tilde{p}(x_{\boldsymbol{V}})$

- Factorizations agree except for the term of the intervened variable
  - Assume strictly positive densities

Inverse probability weighting

- Observational distribution $p(x_{\boldsymbol{V}})$
- Interventional distribution $p\left(x_{\boldsymbol{V}} \middle| do\left(X_k \leftarrow \tilde{N}_k\right)\right) = \tilde{p}(x_{\boldsymbol{V}})$
- Interested in certain aspect $l(x_{\boldsymbol{V}})$, then:

$$E\left(l(x_{\boldsymbol{V}}) | do\left(X_k \leftarrow \tilde{N}_k\right)\right)$$

$$= \int l(x_{\boldsymbol{V}}) \tilde{p}(x_{\boldsymbol{V}}) dx_{\boldsymbol{V}}$$

$$= \int l(x_{\boldsymbol{V}}) \frac{\tilde{p}(x_{\boldsymbol{V}})}{p(x_{\boldsymbol{V}})} p(x_{\boldsymbol{V}}) dx_{\boldsymbol{V}}$$

$$= \int l(x_{\boldsymbol{V}}) \frac{\tilde{p}\left(x_k | x_{\mathrm{pa}(k)}\right)}{p\left(x_k | x_{\mathrm{pa}(k)}\right)} p(x_{\boldsymbol{V}}) dx_{\boldsymbol{V}}$$

Inverse probability weighting

- Given observations $x_{\boldsymbol{V}}^1, \ldots, x_{\boldsymbol{V}}^n$ drawn from observational distribution $p(x_{\boldsymbol{V}})$, can construct estimator for expectation under interventional distribution:

$$\hat{E}\left(l(x_{\boldsymbol{V}})|do\left(X_k \leftarrow \tilde{N}_k\right)\right) = \frac{1}{n}\sum_i l\left(x_{\boldsymbol{V}}^i\right) w_i$$

where $w_i = \frac{\tilde{p}\left(x_k^i | x_{\mathrm{pa}(k)}^i\right)}{p\left(x_k^i | x_{\mathrm{pa}(k)}^i\right)}$

  - [See Series 4.]
  - Related to survey sampling, importance sampling, reinforcement learning
    - See Elements of Causal Inference, Chapter 8.2.

Recap

- Concepts to know:
  - Counterfactuals
  - Potential outcomes
  - Matching
  - Propensity score matching
  - Inverse probability weighting

References and acknowledgments

- Counterfactuals
  - Peters, Janzing and Schölkopf (2017). Elements of Causal Inference. Chapters 3.3., 6.4
- Potential outcomes
  - Peters, Janzing and Schölkopf (2017). Elements of Causal Inference. Chapter 6.9
- Estimation
  - Shalizi (2019). Advanced Data Analysis from an Elementary Point of View. Chapter 23.1.3-23.1.5
  - Peters, Janzing and Schölkopf (2017). Elements of Causal Inference. Chapter 8.2.1

- Optional reading:
  - Richardson and Robins (2013). Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality.
  - Rosenbaum and Rubin (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects.