Machine learning and causal inference (observational studies)

Robert Castelo

robert.castelo@upf.edu

@robertclab

Dept. of Medicine and Life Sciences
Universitat Pompeu Fabra

upf.

Barcelona

Barcelona School of Economics
Data Science Methodology Program
Winter Term 2024-25

BSE
Barcelona School of Economics

Causal inference in observational studies

- In observational studies we have no control on the treatment/exposure of the units.
- Treatment assignment mechanism is not random, but some sort of selection made by the unit itself, the environment or any mechanism other than an experiment: no manipulation of causes.
- Approach: attempt to approximate a randomized experiment within the observational study.
- Note that treatment is not going to be independent from potential outcomes anymore. We need to work without strong ignorability $X \perp\!\!\!\perp Y(x)$ for $x = \{0, 1\}$.

Barcelona School of Economics

Causal inference in observational studies

- Rewrite $\text{ATE} := \text{E}[Y(1)] - \text{E}[Y(0)]$ using the law of total probability:

$$\begin{aligned} \text{ATE} &:= \{\text{E}[Y(1)|X=0]\Pr(X=0) + \text{E}[Y(1)|X=1]\Pr(X=1)\} \\ &- \{\text{E}[Y(0)|X=0]\Pr(X=0) + \text{E}[Y(0)|X=1]\Pr(X=1)\}, \end{aligned}$$

but notice that $\text{E}[Y(1)|X=0]$ and $\text{E}[Y(0)|X=1]$ are not identifiable.

= counterfactual

The hypothetical average labor market score if the treatment group hadn't received treatment (counterfactual

- It can be shown (Morgan and Winship, 2015, pg. 59) that the $\widehat{\text{ATE}}$ estimator converges in probability to

$$\text{E}[Y(1)|X=1] - \text{E}[Y(0)|X=0] = \text{ATE} + \text{bias},$$

where

This measures how individuals in the treatment group (X=1) and control group (X=0) would differ in the absence of treatment effect/obersvation

ATT=E[Y(1)Y(0)X=1]AT

$$\text{bias} := \{\text{E}[Y(0)|X=1] - \text{E}[Y(0)|X=0]\} + \Pr(X=0)\{ATT - ATC\},$$

and the first term is the *baseline bias* and the second is the *differential treatment effect bias*.

This bias accounts for differences in treatment effects (ATTATC) and how often

Causal inference in observational studies

- Example adapted from Morgan and Winship (2015, pg. 59-60): *the effect of obtaining a master's degree on labor market outcome*. We observe that individuals who have obtained a master's degree score higher than those who didn't.

- **ATE:** *obtaining a master's degree makes individuals more successful in the labor market, reflected in some kind of numerical score, the higher the better.*

- **Baseline bias** ($\mathrm{E}[Y(0)|X=1] - \mathrm{E}[Y(0)|X=0]$): *Individuals who obtain master's degrees would have been done better in the labor market than those who didn't obtain them, in the counterfactual state in which they did not in fact obtain master's degrees.*

- **Differential treatment effect bias** ($\Pr(\mathrm{X}{=}0)\{\mathrm{ATT\text{-}ATC}\}$): *Those who didn't obtain master's degrees would not have done as well as those who did obtain them in the counterfactual state in which they did in fact obtain master's degrees.*

Causal inference in observational studies

- Suppose 30% of the people obtains a master's degree ($P(X=1)=0.3$) and we observe the labor market outcome scores specified below.

Group	$\mathrm{E}[Y(1)\|X]$	$\mathrm{E}[Y(0)\|X]$
Treatment ($X=1$)	10	6
Control ($X=0$)	8	5

- **ATE:=** $\{\mathrm{E}[Y(1)|X=0]\Pr(X=0)+\mathrm{E}[Y(1)|X=1]\Pr(X=1)\}-\{\mathrm{E}[Y(0)|X=0]\Pr(X=0)+\mathrm{E}[Y(0)|X=1]\Pr(X=1)\}=8\cdot 0.7+10\cdot 0.3-5\cdot 0.7-6\cdot 0.3=3.3$

5 > 3.3

- **Baseline bias:=** $\mathrm{E}[Y(0)|X=1]-\mathrm{E}[Y(0)|X=0]=6-5=1$
- **DTE bias:=** $\Pr(X=0)(ATT-ATC)=0.7\cdot(10-6-8+5)=0.7$
- Note that $\widehat{\mathrm{ATE}}:=\mathrm{E}[Y(1)|X=1]-\mathrm{E}[Y(0)|X=0]=10-5=5$ is upwardaly biased from the actual ATE, 3.3, which follows also from substracting baseline and DTE bias from $\widehat{\mathrm{ATE}}$ (5-1-0.7=3.3).

Causal inference in observational studies: bounds

- Assuming SUTVA (consistency, no-interference) and observed data being a random sample, we can derive **bounds** on ATE for a bounded outcome Y. For instance, consider a binary outcome $Y = \{0, 1\}$, i.e., $-1 \leq \text{ATE} \leq 1$ is a risk difference (Robins[1], 1989; Manski[2], 1990).
- Recall ATE:

$$\begin{aligned} \text{ATE} &:= \{\text{E}[Y(1)|X=0]\Pr(X=0) + \text{E}[Y(1)|X=1]\Pr(X=1)\} \\ &- \{\text{E}[Y(0)|X=0]\Pr(X=0) + \text{E}[Y(0)|X=1]\Pr(X=1)\}, \end{aligned}$$

- ~~Upper bound by making ATE as large as possible:~~

Since E[Y(1)X=0]E[Y(1)X=0] and E[Y(0)X=1]E[Y(0)X=1] are counterfactuals and cannot be observed, we d

$$\Pr(X=0) + \text{E}[Y(1)|X=1]\Pr(X=1) - \text{E}[Y(0)|X=0]\Pr(X=0).$$

- Lower bound by making ATE as small as possible:

$$\text{E}[Y(1)|X=1]\Pr(X=1) - \text{E}[Y(0)|X=0]\Pr(X=0) - \Pr(X=1).$$

[1] Robins, J.M. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. *Health Ser. Res. Met.*, 1989.

[2] Manski, C.F. Nonparametric bounds on treatment effects. *Am. Econ. Rev.*, 1990. https://www.jstor.org/stable/2006592.

Causal inference in observational studies: bounds

- The resulting interval derived from those bounds is contained in $[-1, 1]$.

- Upper bound on ATE:

$$\Pr(X = 0) + \mathrm{E}[Y(1)|X = 1]\Pr(X = 1) - \mathrm{E}[Y(0)|X = 0]\Pr(X = 0)\,.$$

- Lower bound on ATE:

$$-\Pr(X = 1) + \mathrm{E}[Y(1)|X = 1]\Pr(X = 1) - \mathrm{E}[Y(0)|X = 0]\Pr(X = 0)\,.$$

- The interval has width $\Pr(X = 0) + \Pr(X = 1) = 1$ and, consequently, (1) it cannot exclude $\mathrm{ATE} = 0$ and (2) the sign of the ATE cannot be determined from the observed data.

- These bounds are sharp, in the sense that narrower bounds are not possible without additional assumptions.

Causal inference in observational studies: bounds

- Consider the previous example on the effect of obtaining a master's degree and labor market, but now the outcome is binary, where $Y = 1$ means the student got a job, while $Y = 0$ means the student went unemployed.

- We have a random sample of 40 students, among which 25 obtained a master's degree, thus 15 did not. Among the 40 students, 20 found a job and 20 went unemployed. Among those 20 who found a job, 15 had obtained a master's degree and 5 didn't.

- By causal consistency we can write the previous upper bound as:

$$\underset{\frac{15}{40}}{\Pr(X=0)} + \underset{\frac{15}{25}\cdot\frac{25}{40}}{\mathrm{E}[Y|X=1]\Pr(X=1)} - \underset{\frac{5}{15}\cdot\frac{15}{40}}{\mathrm{E}[Y|X=0]\Pr(X=0)} = 0.625$$

- Lower bound is one unit lower, thus bounds are [-0.375, 0.625].

Causal inference in observational studies: bounds

- Upper bound estimate on ATE derived by guessing numbers to best case:

	Group		Whole
	$X=0$	$X=1$	sample
$Y(1)=1$	15	15	30
$Y(1)=0$	0	10	10
Total	15	25	40
$Y(0)=1$	5	0	5
$Y(0)=0$	10	25	35
Total	15	25	40

- Upper bound estimate on ATE:

$$\text{ATE}^{\uparrow} := \text{E}[Y(1)] - \text{E}[Y(0)] = \frac{30}{40} - \frac{5}{40} = 0.625\,.$$

- Lower bound estimate on ATE. Guess numbers to worst case and then:

$$\text{ATE}^{\downarrow} := \text{E}[Y(1)] - \text{E}[Y(0)] = \frac{15}{40} - \frac{30}{40} = -0.375\,.$$

Causal inference in observational studies: bounds

- We have attempted to work without assuming ignorability, i.e., $X \perp\!\!\!\perp Y(0)$ and $X \perp\!\!\!\perp Y(1)$, because the lack of control on the treatment/exposure of the units may lead to a dependence with the potential outcome.

  ignorability: Whether a unit receives treatment (X=1X=1) or not (X=0X=0) does not depend on the potential outcomes (Y(1),Y(0)Y(1),Y(0)).

- How may actually arise this dependence? For instance, in the example about labor market outcome and obtaining a master's degree.

  -->It introduces bias when estimating the causal effect (ATEATE).

- When treatment/exposure and potential outcomes share a common cause, a phenomenon known as **confounding**.

- For instance, a competitive admission process for a master's degree may be selecting students who would be successful in the job market anyway.

Barcelona School of Economics

Propensity scores

- Approach: if we knew the factors (covariates) Z that drive the confounding phenomenon, we could assume ignorability **conditional on baseline covariates** Z:

$$X \perp\!\!\!\perp \{Y(0), Y(1)\} | Z \,,$$

also known in the literature as the *"conditional exchangeability"* or *"no unmeasured confounders"* assumption.

- Under this assumption, $\Pr(X = 1 | Y(0), Y(1), Z) = \Pr(X = 1 | Z)$. The term $\Pr(X = 1 | Z)$ is the probability of treatment given baseline covariates and it is also known as the **propensity score** (Rosenbaum and Rubin, 1983)[3].

[3] Rosenbaum, P.R. and Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 1983. https://doi.org/10.1093/biomet/70.1.41.

Propensity scores

- Let $e(Z) := \Pr(X = 1|Z)$ denote the propensity score on covariates Z. If we assume that there are no unmeasured confounders other than Z, i.e., $\Pr(X = 1|\, Y(0), Y(1), Z) = \Pr(X = 1|Z)$, then

$$\Pr(X = 1|\, Y(0), Y(1), e(Z)) = \Pr(X = 1|e(Z))\,,$$

  i.e., it is sufficient to adjust/stratify/control for the scalar $e(Z)$ rather than for the possibly multi-dimensional Z.

- Given a sample of n units, we may have K different propensity scores with $K \leq n$. Let $\{\pi_1, \ldots, \pi_K\}$ be the set of those different propensity scores.

- If we stratify units according to their propensity score, the size n_k of stratum $k \in \{1 \ldots K\}$ corresponds to:

$$n_k := \sum_{i=1}^{n} \mathrm{I}[e(Z_i) = \pi_k]\,.$$

The summation i=1nI[e(Zi)=k]i=1nI[e(Zi)=k] counts individuals whose propensity scores fall within the range of the stratum.

Propensity scores

- Within each stratum, units have the same propensity score and we can assume ignorability like in a randomized experiment.

- A consistent estimator of ATE is then

$$\widehat{\mathrm{ATE}}_{str} := \sum_k \left(\frac{n_k}{n}\right) \widehat{\mathrm{ATE}}_k \,,$$

where

$$\widehat{\mathrm{ATE}}_k := \frac{\sum_i^n Y_i \mathrm{I}[X_i = 1, e(Z_i) = \pi_k]}{\sum_i^n \mathrm{I}[X_i = 1, e(Z_i) = \pi_k]} - \frac{\sum_i^n Y_i \mathrm{I}[X_i = 0, e(Z_i) = \pi_k]}{\sum_i^n \mathrm{I}[X_i = 0, e(Z_i) = \pi_k]} \,.$$

Propensity scores: inverse probability weighting

- Consider the number m of individuals in stratum k that are treated:

$$m_k := \sum_{i=1}^{n} \mathrm{I}[X_i = 1, e(Z_i) = \pi_k] \,.$$

- Recall the ATE estimator for stratum k:

$$\widehat{\mathrm{ATE}}_k := \frac{\sum_i^n Y_i \mathrm{I}[X_i = 1, e(Z_i) = \pi_k]}{\sum_i^n \mathrm{I}[X_i = 1, e(Z_i) = \pi_k]} - \frac{\sum_i^n Y_i \mathrm{I}[X_i = 0, e(Z_i) = \pi_k]}{\sum_i^n \mathrm{I}[X_i = 0, e(Z_i) = \pi_k]} \,.$$

- Rewrite it as follows:

$$\widehat{\mathrm{ATE}}_k := \frac{\sum_i^n Y_i \mathrm{I}[X_i = 1, e(Z_i) = \pi_k]}{m_k} - \frac{\sum_i^n Y_i \mathrm{I}[X_i = 0, e(Z_i) = \pi_k]}{n_k - m_k} \,.$$

Propensity scores: inverse probability weighting

- Recall the overall ATE estimator $\widehat{\text{ATE}}_{str}$:

$$\widehat{\text{ATE}}_{str} := \sum_k \left(\frac{n_k}{n}\right) \widehat{\text{ATE}}_k \,,$$

- We can obtain the following equivalent form:

$$\widehat{\text{ATE}}_{str} := \frac{1}{n} \sum_k \left\{ \frac{\sum_i^n X_i \, Y_i \text{I}[X_i = 1, e(Z_i) = \pi_k]}{m_k / n_k} - \frac{\sum_i^n (1 - X_i) \, Y_i \text{I}[X_i = 0, e(Z_i) = \pi_k]}{(n_k - m_k)/n_k} \right.$$

- The previous expression is approximately equal to:

$$\widehat{\text{ATE}}_{ipw} := \frac{1}{n} \sum_i \left\{ \frac{X_i \, Y_i}{e(Z_i)} - \frac{(1 - X_i) \, Y_i}{1 - e(Z_i)} \right\} .$$

- We weight individuals by the inverse of the probability of being assigned the treatment actually received.

Propensity scores: regression

- Another approach is to use a regression model:

$$\mathrm{E}[\,Y|X = x, e(Z) = e(z)] = \alpha_0 + \alpha_1 x + \alpha_2 e(z)\,.$$

- By consistency and ignorability within each stratum,

$$\mathrm{E}[\,Y(x)|e(Z) = e(z)] = \alpha_0 + \alpha_1 x + \alpha_2 e(z)\,,$$

- Consequently, for $\mathrm{ATE} := \mathrm{E}[\,Y(2)] - \mathrm{E}[\,Y(0)]$, $\alpha_1 = \mathrm{ATE}$ and thus we can define $\widehat{\mathrm{ATE}}_{reg} = \hat{\alpha}_1$.

Propensity scores: estimation

- Propensity scores are unknown in observational studies. They need to be estimated using logistic regression or some supervised machine learning method (Lee et al., 2010)[4].

- In the case of logistic regression, we consider the binary treatment variable X as response and the covariates $Z_1, \ldots, Z_k$ as explanatory variables:

$$\log \frac{e(Z)}{1-e(Z)} = \log \frac{\Pr(X=1|Z)}{1-\Pr(X=1|Z)} = \beta_0 + \beta_1 Z_1 + \cdots + \beta_k Z_k \,.$$

- Once the $\hat{\beta}_i$ cofficients have been estimated, we can obtain the propensity scores by using the model formula:

$$\hat{e}(Z) = \frac{e^{\hat{\beta}_0 + \cdots + \hat{\beta}_k}}{1 + e^{\hat{\beta}_0 + \cdots + \hat{\beta}_k}} \,.$$

[4] Lee et al. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29:337-346. https://doi.org/10.1002/sim.3782

Causal inference in observational studies

- Elephant in the room: **how do we choose Z among observed variables?**
- Second elephant in the room: **are there unmeasured confounders?**.
- Sensitivity analysis methods to assess robustness of causal inference to the violation of no unmeasured confounders.
- Techniques that allow for unmeasured confounders, e.g., instrumental variables.

Graphical models !!!!!

Barcelona School of Economics

Concluding remarks

- In randomized studies association is causation.
- In observational studies we can work with bounds or ignorability given covariates (which covariates? do we measure all potential confounders?).
- Causal inference requires working with precise terminology, explicit assumptions and a strong subject-matter knowledge. Remember Rubin's quote: *"assumptions are the strands that link statistics to science"*.
- Causal inference is a vast field, many things we haven't seen: randomization-based inference using Fisher's exact test and permutation tests, causal inference with non-compliance and/or interference, Bayesian methods, etc.

Barcelona School of Economics