

Causal Estimation

Karla Diaz-Ordaz

UCL

APTS Week 4,
September 2026

Previously...

- ▶ We introduced several causal languages (e.g. do-calculus, potential outcomes, POs)
- ▶ we will now use them to define causal effects
- ▶ then discuss the assumptions needed for identification
- ▶ We have also previously used **causal diagrams** to:
 - (a) encapsulate our causal knowledge and
 - (b) determine whether or not conditional exchangeability holds (given the assumptions embedded in the diagram).

Causal effects

Average causal effects can be defined for **different populations**:

- (1) the whole population
- (2) the treated/exposed
- (3) the *compliers*
- (·) ...

Each can be defined on **different scales**, *i.e.* using different contrasts.

We will only consider the first one, for simplicity

Also, we focus on

- ▶ A = binary (point-)treatment
- ▶ Y = some (numeric) outcome –not survival / time-to-event as this needs special considerations
- ▶ X = sufficient to adjust for confounding: ‘valid’ adjustment set.

Estimands

Contrasts for the whole population (called **marginal**) that we could consider are:

- ▶ **Average Treatment Effect** (ATE):

$$E(Y(1)) - E(Y(0))$$

- ▶ For binary Y :

- ▶ **Risk ratio**:

$$E(Y(1))/E(Y(0))$$

- ▶ **Odds ratio**:

$$\frac{E(Y(1))}{(1 - E(Y(1)))} / \frac{E(Y(0))}{(1 - E(Y(0)))}$$

Identification

- ▶ Consider $ATE = E(Y(1)) - E(Y(0))$.
- ▶ It can be identified from observational data under certain assumptions.

Those most often invoked are:

1. **No interference**
2. **Consistency**
3. **Conditional (mean) exchangeability**

see Hernan & Robins, part I for a discussion:

www.hsph.harvard.edu/faculty/miguel-hernan/causal-inference-book/

Identifying assumption

1- No interference :

the exposure of one individual does not affect the potential outcome of another. Examples where it may not hold:

- ▶ vaccination status of one individual may affect disease status of another
- ▶ extra tuition received by one individual may affect the exam performance of another
- ▶ In both examples the exposure is somehow shared.
- ▶ Departures make the definitions of exposure and POs more complex.

For the rest of the lecture will assume this assumption is met.

Identifying assumptions

2 - Consistency: *the potential outcome $Y(a)$ of those observed to have exposure value $A = a$ is equal to the observed outcome:*

$$Y(a) = Y \text{ for those with } A = a$$

- ▶ In other words, if treatment A could have been assigned in many ways, all would have resulted in the same observed outcome
- ▶ Example where it may not hold: A = training, Y = employment: being assigned may not have the same effect as choosing A .

- Cole and Frangakis, Epidemiology. 2009; 20(1): 3
- VanderWeele, Epidemiology. 2009; 20(6): 880
- Pearl, Epidemiology. 2010; 21(6): 872

Identifying assumptions

3 - Exchangeability

- ▶ **(Mean) exchangeability** (in randomized experiments): *the mean potential outcome is the same in exposed and unexposed (and this holds for both POs):*

$$E(Y(a) | A = 1) = E(Y(a) | A = 0), \text{ for } a = 0, 1.$$

Formally: $Y(a) \perp\!\!\!\perp A$, for $a = 0, 1$

- ▶ **Conditional (mean) exchangeability:**

Within strata of $\mathbf{X}$ the mean potential outcome is the same in exposed and unexposed (and this holds for both POs):

$$E(Y(a) | A = 1, \mathbf{X} = \mathbf{x}) = E(Y(a) | A = 0, \mathbf{X} = \mathbf{x}), \text{ for } a = 0, 1, \forall \mathbf{x}.$$

Formally: $Y(a) \perp\!\!\!\perp A | \mathbf{X} = \mathbf{x}$, for $a = 0, 1, \forall \mathbf{x}$.

Basic Identification: 'ideal' randomized experiments

(double-bind, no losses, full adherence)

We aim to identify $ATE = E(Y(1)) - E(Y(0))$ from observed (or observable) data

- (1) By **exchangeability**,
 $E(Y(a))$ can be replaced by $E(Y(a) | A = a)$, for $a = 0, 1$.
- (2) By **consistency**,
 $E(Y(a) | A = a)$ can be replaced by $E(Y | A = a)$.
 - ▶ Hence, $E(Y(1)) = E(Y | A = 1)$
 - ▶ and $E(Y(0)) = E(Y | A = 0)$ hence

$$ATE = E(Y | A = 1) - E(Y | A = 0).$$

Motivation: observational studies

- ▶ Consistency and exchangeability are not guaranteed in observational studies.
- ▶ If individuals are conditionally exchangeable within strata of X , (i.e. X is a sufficient adjustment set)
we can focus on each stratum and repeat the previous steps:
 - ▶ **conditional exchangeability**:
 $E(Y(a)|X = x)$ can be replaced by $E(Y(a)|A = a, X = x)$,
for $a = 0, 1$.
 - ▶ **by consistency**
 $E(Y(a)|A = a, X = x)$ can be replaced by
 $E(Y|A = a, X = x), \forall a, x$.
- ▶ Hence
$$E(Y(a)|X = x) = E(Y|A = a, X = x)$$
- ▶ Interpretation: within values of X , whether $A = a$ obtained by intervention or observation makes no difference wrt. distribution of Y .

Identification revisited: average treatment effect (ATE)

- ▶ Assume no interference, consistency and conditional exchangeability holds given valid adjustment set X .
- ▶ Let us identify $Y(1)$. By the law of total probability:

$$E\{Y(1)\} = E_X[E\{Y(1)|X\}]$$

- ▶ By conditional exchangeability (A indep. of $Y(1)$ given X)

$$= E_X[E\{Y(1)|A=1, X\}]$$

- ▶ By consistency,

$$= E_X[E(Y|A=1, X)]$$

- ▶ In general, $E[Y(a)] = E_X[E[Y|A=a, X]]$, hence

$$ATE = E_X[E[Y|A=1, X] - E[Y|A=0, X]]$$

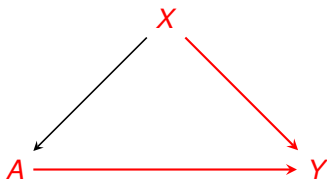
Estimation

- ▶ We have reduced the problem of estimating a causal quantity to estimating cond. expectations of observed data
- ▶ We now need to come up with estimation strategies for those conditional expectations
- ▶ We come back to the causal assumptions afterwards to discuss their plausibility in the context of our applied work, when interpreting the results
- ▶ but for now, we have a pure statistical problem

Methods based on Outcome Modelling

Basic idea (1)

The identification steps suggest we model the cond. expectations involved, i.e. $E[Y|X, A]$:



Estimation 1: G-computation

- ▶ Assume for now, X is discrete, so $ATE =$

$$\sum_x \left\{ E(Y|A=1, X=x) - E(Y|A=0, X=x) \right\} Pr(X=x)$$

- ▶ fit flexible parametric regression model for $E(Y|A, X)$ to data
- ▶ average over empirical X -distribution
- ▶ this is known also as *standardization*, or “outcome modelling”. It is a special case of **G-computation**

Estimation 1: G-computation

1. **Model** Y given X **separately** by treatment level a , e.g. in the treated and then the untreated

$$E(Y | A = a, X = x)$$

- ▶ Example: in the continuous outcome case

$$E(Y | a, X) = \alpha_a + \gamma_a^T X \quad (1)$$

2. Use these fitted models to **predict** the POs $Y(a)$ of each individual on the basis of their X .
3. Average these POs over the observed distribution of X .
4. Take the difference of these mean POs to estimate ATE.
 - ▶ assumes the regression model eq (1) is **correctly specified**
 - ▶ this is an example of a **plug-in** estimator: we obtain $\hat{\theta}$ by plugging fitted models directly into the formula defining the estimand θ

A simple example with continuous Y

What is the causal effect of maternal smoking on birth weight? Data used by Cattaneo (Journal of Econometrics, 2010) on singleton babies born in Pennsylvania between 1989 and 1991.

Y : birth weight (g) `bweight` ;

A : maternal smoking (No/Yes) `smoker` ;

X : baby was 1st born (No/Yes) `fbaby`, marital status, maternal alcohol intake, paternal education, maternal age

- ▶ 4,642 records in total

- ▶ Fit a model for Y on A controlling for X

```
m0<-lm(bweight ~ fbaby + mmarried + alcohol +  
fedu + mage, data=data[mbsmoke==0,])  
Y0<-predict(m0, newdata = data)  
m1<-lm(bweight ~ fbaby + mmarried + alcohol +  
fedu + mage, data=data[mbsmoke==1,])  
Y1<-predict(m1, newdata = data)  
ATE<-mean(Y1)-mean(Y0)
```

ATE= -231.14

Assumptions

- ▶ Parametric regression models need to be correctly specified
- ▶ this includes assumptions about effect modification
- ▶ assuming the model is correct introduces extrapolation
- ▶ an alternative (and necessary) assumption for non-parametric estimation is to assume *positivity*

Positivity

- ▶ Assume that we only have 2 categorical X , `fbaby`, `mmarried`
- ▶ Fit a saturated model

```
lm(formula = bweight ~ mbsmoke + mmarrried + fbaby +  
mbsmoke:mmarrried + mbsmoke:fbaby +mmarrried:fbaby +  
mbsmoke:mmarrried:fbaby, data = data)
```

- ▶ To estimate each β_x , none of the 8 categories can be empty.

mbsmoke	mmarrried	fbaby=0	fbaby=1
0	0	409	530
0	1	1657	1182
1	0	265	190
1	1	278	131

- ▶ **positivity**: for each possible level $X = x$,

$$0 < Pr(A = 1|X = x) < 1$$

- ▶ In a sufficiently large sample, $\forall$ observed x , positivity guarantees there is exposed and unexposed individuals

Pitfalls of outcome modelling estimators

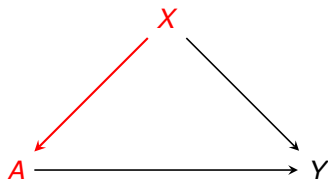
- ▶ We can only check how well the model fits the **observed** data, but regression-based estimators rely on the **extrapolation** of these fitted relationships to regions where there is little data to support the model choice.
- ▶ The parameters of regression models are typically estimated by maximum likelihood.
- ▶ For some models (logistic and Cox regression), maximum likelihood estimators are biased in small samples.
- ▶ In particular, bias increases as the number of events per parameter decreases.
- ▶ The higher dimensionality of X (the valid adjustment set), the larger this finite-sample bias.

Peduzzi *et al* (J Clin Epidemiol, 1995 & 1996) gave a rule of thumb of “10 or more events per variable”.

Modelling the treatment

Alternative strategy: Methods based on Treatment Modelling

Instead of modelling $E[Y | A, X]$ as before, we now model A given $\mathbf{X}$:
 $Pr[A = a | X]$, called the propensity score



Key reference Rosenbaum and Rubin (Biometrika, 1983) “The central role of the propensity score in observational studies for causal effects”.

- ▶ There are many methods based on PS:
 - ▶ PS stratification,
 - ▶ PS matching
 - ▶ PS inverse weights

Inverse probability weighting: Alternative Identification

Formally, we motivate this estimation strategy by showing (under our identification assumptions):

$$E[Y(a)] = E \left[\frac{I(A=a)}{Pr[A=a|X]} Y \right]$$

Proof

- ▶ $E \left[\frac{I(A=a)}{Pr[A=a|X]} Y \right]$ is equal to $E \left[\frac{I(A=a)}{Pr[A=a|X]} Y(a) \right]$ by consistency.
- ▶ by positivity, $Pr[A=a|X] \neq 0$, by iterated expectations we have

$$E \left[\frac{I(A=a)}{Pr[A=a|X]} Y(a) \right] = E \left\{ E \left[\frac{I(A=a)}{Pr[A=a|X]} Y(a) \middle| X \right] \right\}$$

- ▶ by conditional exchangeability

$$= E \left\{ E \left[\frac{I(A=a)}{Pr[A=a|X]} \middle| X \right] E[Y(a) | X] \right\}$$

- ▶ and since $E \left[\frac{I(A=a)}{Pr[A=a|X]} \middle| X \right] = 1$ we have $= E \{ E[Y(a) | X] \} = E[Y(a)]$

Estimation 2: Horvitz-Thompson estimator

- ▶ $E[Y(a)] = E\left[\frac{I(A=a)}{Pr[A=a|X]} Y\right]$ suggest an estimation strategy
- ▶ Specify a PS model: $\pi(X) = Pr(A = 1 | X)$
- ▶ to estimate $E[Y(1)]$:
 1. predict $\hat{\pi}(X)$
 2. average outcomes in the exposed, weighting by $\hat{\pi}(X)^{-1}$

$$\frac{1}{n} \sum_{i=1}^n \frac{A_i Y_i}{\hat{\pi}(X_i)}.$$

3. For the ATE, do the corresponding steps for $E[Y(0)]$ and take the difference
- ▶ One disadvantage is that it is not guaranteed to respect the range of the outcome. Use “Hajek’s” (normalize by the sum of weights)

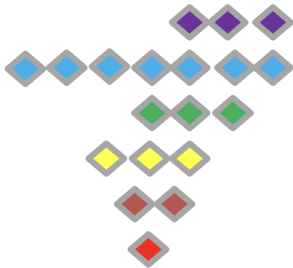
$$\tilde{w}_i = \frac{w_i}{\sum_{j:A_j=a} w_j} \quad \Rightarrow \quad \hat{\psi}_{Hajek} = \sum_{i:A_i=a} \tilde{w}_i Y_i$$

IPW Intuition— pseudo population for the ATE

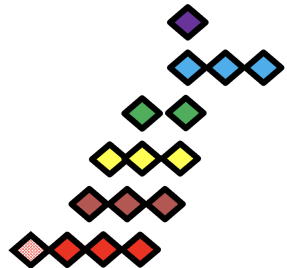


- ▶ We have individuals prescribed statins (or not) according to X , a score (represented here by colour)
- ▶ We want to know what the outcome would have been if all individuals in each colour remained untreated vs all treated

Not prescribed statins

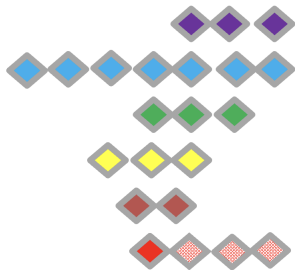


Prescribed statins

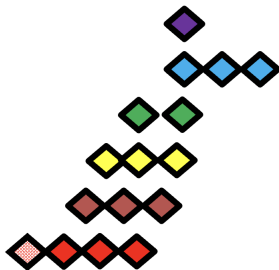


- ▶ Let's focus on the red row, starting with the **treated**: we need an extra red person to get statins
- ▶ we have 3 (out of 4) treated, so each of them needs to represent themselves and a little bit more : weight each by $1/0.75 = 1.333333$
- ▶ $0.33 \text{ extra each} \times 3 = 1$

Not prescribed
statins

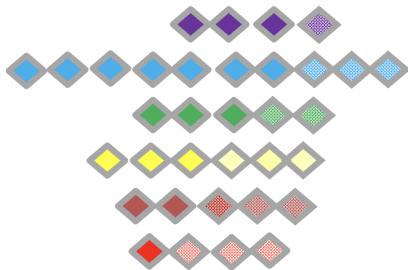


Prescribed
statins

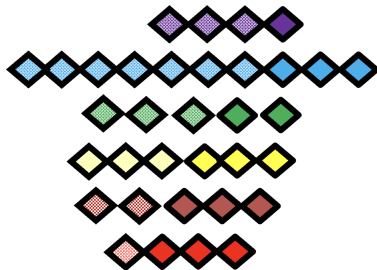


- ▶ The one untreated red (grey border) needs to represent 3 others. This means multiplying it by $1/0.25$ this is the inverse of the prob of getting its own treatment conditional on its level of X
- ▶ To remove the association between A and X , we want to *balance* the numbers of each colour getting (or not) statins (black or grey borders)

Not prescribed statins



Prescribed statins



- ▶ We created a pseudo-population where A is no longer associated with X (represented here by colour)
- ▶ To obtain ATE, we can just average the outcomes in the pseudo-populations and contrast

Assumptions for IPW

- ▶ relies on the PS model being **correctly specified**
- ▶ assumes **positivity** $0 < Pr(A_i = 1 | X_i = x) < 1$, for all x .
- ▶ If this doesn't hold, we have infinite weights.
- ▶ Positivity violations may happen for statistical or structural reasons:

statistical (e.g.) there are too many categories among your covariates;

structural (e.g.) contra-indications

- ▶ Estimated propensity score close to 1 or 0 are problematic, since they imply that a few individuals will receive a very large weight, leading to imprecise and unstable estimates.

Stabilized weights

- ▶ The goal of IPW is to create a pseudo-population where there is imbalance in the covariates by treatment A
- ▶ we could have used a pseudo-population in which all individuals have a probability of receiving either treatment equal to 0.5 $Pr(A = a) = 0.5$, so the weights become

$$w = \frac{0.5}{Pr(A|X)}$$

- ▶ This is useful to **stabilize** the weights: multiplying by some arbitrary marginal distribution $Pr(a)$
- ▶ A common choice

$$w_1 = \frac{Pr[A = 1]}{Pr(A|X)}$$

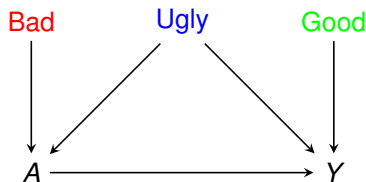
- ▶ Reduces variance, less dominated by a few extreme weight observations

Practical considerations

- ▶ A parametric propensity score model is specified, usually using logistic regression.
- ▶ The weights are calculated from the predictions from this model.
- ▶ **Robust** estimates of SE are needed to account for the lack of independence after re-weighting.
- ▶ Furthermore, it is possible to reflect in sandwich estimators of SE, that the weights were estimated from the data. Or use bootstrap.
- ▶ Can use R packages, such as `ipw`

Which variables should be included?

- ▶ We want include a sufficient adjustment set, as determined by our analysis of the DAG
- ▶ but what about adding other *extra* variables?



- ▶ We want to adjust for the 'ugly': (a sufficient adjustment set)
- ▶ including predictors of the outcome (the good) can improve precision
- ▶ BUT adjusting for the 'bad' variables (only associated with treatment)
 - ▶ will always inflate variance
 - ▶ can amplify any little bias that may have been present

How to know if my PS is good?

- ▶ Propensity score modelling is a **very unusual** modelling exercise!
- ▶ Our main thought should **not** be “what predicts A well?”
- ▶ but what leads to better balance
 - ▶ Check positivity / overlap
 - ▶ Check balancing property (various diagnostics).

What do we do if there is poor overlap / very large weights?

- ▶ Poor overlap shows up as some observations having much more influence than others (very large weights) — happens when units are treated contrary to expectation: indication of unmeasured confounding??
- ▶ Investigate why: who are they? are they some combination of variables that makes them “rare”? do they have extreme values for some confounders?
- ▶ **Trim** the tails of the PS (e.g. 0.01 or 0.02, Lee 2011): removes subjects with extreme PS values, making positivity more plausible but the resulting causal effects apply to a subset of the population
- ▶ or **truncate** the weights introduces some bias, as you're not weighting properly, but reduces the variance
- ▶ better still: try to diagnose the source of poor overlap and recode/group variables to avoid sparse values

Example

Revisiting the Cattaneo dataset

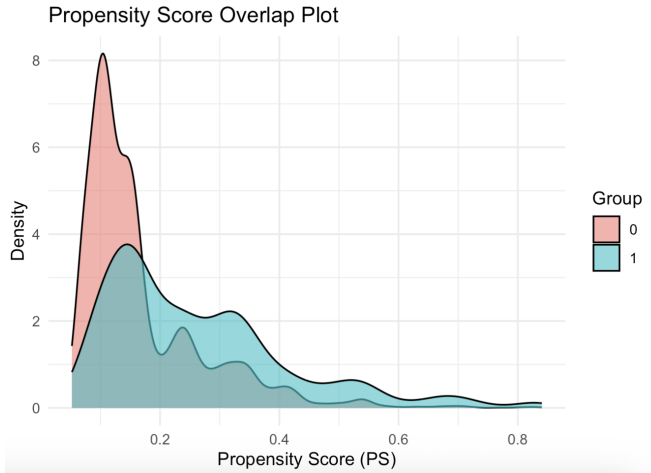
- ▶ ATE of maternal smoking on baby birth weight
- ▶ X: baby was 1st born, marital status, maternal alcohol intake, paternal education, maternal age

```
##PS
ps.mod<-glm(mbsmoke ~ fbaby + mmarried + alcohol +
            fedu + mage, data=data, family=binomial())
ps = predict(ps.mod,type="response")
w = mbsmoke/ps + (1-mbsmoke)/(1-ps)
mod = lm(bweight ~ mbsmoke,data=data,weights=w)
```

- ▶ Estimated ATE is -230.49

```
> diag(sandwich::vcovHC(mod, type = "HC1"))^0.5
(Intercept)      mbsmoke
  9.712371      24.568119
```

Looking at the overlap



Checking balance

- ▶ recall that the goal was to achieve balance in the variable distribution for the 2 groups
- ▶ after weighting, check if we have good balance
- ▶ Calculate the standardised mean difference

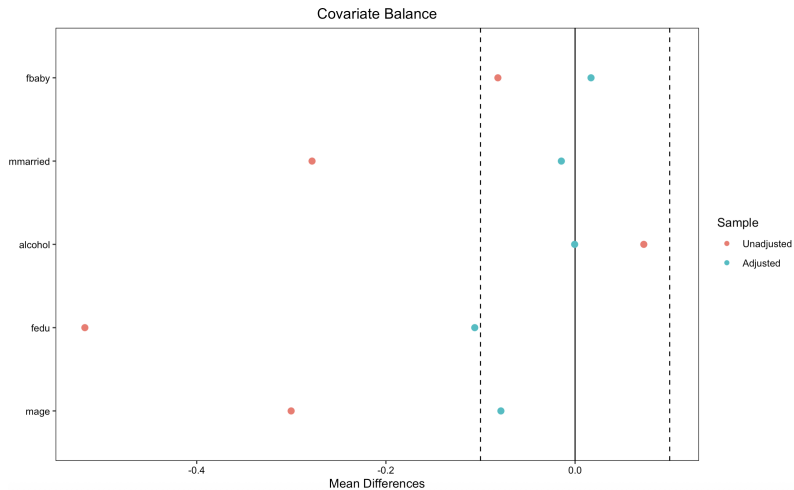
$$d = \frac{\bar{X}_{\text{treatment}} - \bar{X}_{\text{control}}}{\sqrt{\frac{s_{\text{treatment}}^2 + s_{\text{control}}^2}{2}}}$$

- ▶ report the absolute standardised differences (either a “Table 1” or a plot)
- ▶ rule-of-thumb, differences should be < 0.1
- ▶ in R, can use the package `cobalt`

Imbalance?

- ▶ refine the PS model, this is part of the design stage, and thus **it is not “cheating”**, as we have not yet looked at the outcome, only the treatment model
 - ▶ include interactions? other non-linear terms?
 - ▶ re-assess balance (back and forth)
- ▶ the final PS model is the one that achieves good balance

Example: the Cattaneo dataset



IPTW vs Outcome regression

- ▶ IPTW-estimators tend to have inflated imprecision relative to traditional regression adjustment.
- ▶ This is partly because - in the presence of model building - they give a more honest reflection of the uncertainty about the causal effect.
- ▶ This is also partly because of IPTW being inefficient
- ▶ We know that adjustment for variables associated with the outcome improves efficiency...so this gives us an idea

Notation

- ▶ let $O = (X, A, Y) \sim P$ denote a single observation (covariates, treatment, outcome), with $O_1, \dots, O_n \stackrel{iid}{\sim} P$ the observed data
- ▶ to denote an expectation for a new X , we use

$$P\{f(X)\} = \int f(X) dP(X)$$

(following the convention: a distribution treated as a linear operator on functions)

- ▶ the sample average is

$$P_n\{f(X)\} = \frac{1}{n} \sum_{i=1}^n f(X_i)$$

- ▶ we will use this compact notation for the doubly robust estimators below

Doubly Robustness

So, if we specify correctly

- ▶ the **outcome model** (i.e. $E(Y | A, X)$), the g-computation is consistent
- ▶ the **propensity score model** (i.e. $E(A | X)$), the Horvitz-Thompson estimator which is also consistent

Is there an estimator that uses both of these models, but only requires one of them to be correct?

Yes! There are several, in fact, and give rise to the term “doubly robust”

Doubly Robust Methods

- ▶ We posit **parametric** *working models* for

$$E[Y \mid A, X] = Q_a(X; \beta, \gamma) \quad \text{and} \quad E[A \mid X] = \pi(X; \eta)$$

- ▶ As we will see, the following function has expectation $E[Y(1)]$ if **either** Q_1 or π is specified correctly:

$$\begin{aligned} \mu_1^{AIPW}(O) &= Q_1(X) + \frac{A}{\pi(X)} \{Y - Q_1(X)\} \\ &= \frac{AY}{\pi(X)} + \left\{1 - \frac{A}{\pi(X)}\right\} Q_1(X). \end{aligned}$$

- ▶ So fit models Q_a and π to the data (e.g. by maximum likelihood), to obtain parameter estimates $\hat{Q}$ and $\hat{\pi}$
- ▶ consistent if **either** model is correctly specified: **double robustness**
- ▶ This correction term $\mu_1^{AIPW}(O) - Q_1(X)$ is a first example of what is called an **efficient influence function** — we will see next lecture why it takes exactly this form

proof sketch

- ▶ To estimate $E\{Y(1)\}$ consistently we have

$$\hat{\mu}_1^{AIPW} = P_n \left\{ \frac{A_i Y_i}{\hat{\pi}(X_i)} + \left(1 - \frac{A_i}{\hat{\pi}(X_i)} \right) \hat{Q}_1(X_i) \right\}$$

- ▶ If $\hat{\pi}(X_i)$ is correctly specified, (re-arranging terms)

$$\hat{\mu}_1^{AIPW} = P_n \left\{ \frac{A_i Y_i}{\hat{\pi}(X_i)} - \left(\frac{A_i - \hat{\pi}(X_i)}{\hat{\pi}(X_i)} \right) \hat{Q}_1(X_i) \right\}$$

- ▶ If $\hat{Q}_1(X_i)$ is correctly specified, (re-arranging terms)

$$\hat{\mu}_1^{AIPW} = P_n \left\{ \frac{A_i (Y_i - \hat{Q}_1(X_i))}{\hat{\pi}(X_i)} + \hat{Q}_1(X_i) \right\}$$

AIPW for ATE

We do something similar for $\hat{\mu}_0^{AIPW}$, and then

$$\hat{\tau}^{AIPW} := \hat{\mu}_1^{AIPW} - \hat{\mu}_0^{AIPW}.$$

- ▶ known as **augmented** inverse probability weighted estimator (AIPW)
- ▶ In addition, each $\hat{\mu}_a^{AIPW}$ is **semi-parametric efficient** if both parametric models are correct, so it achieves the same rate (asymptotically) as maximum likelihood estimation.
- ▶ If Q_a is wrong then MLEs will be difficult to interpret.
- ▶ In practice, even under moderate misspecifications of both models, the doubly robust estimator mostly performs well

Recap: parametric estimation and model misspecification

- ▶ The methods covered today (g-computation and Horvitz-Thompson/Hajek's IPW) rely on the outcome or exposure model being correctly specified
- ▶ AIPW is more robust, as it uses both models, but we still need at least one of them to be correct
- ▶ This misspecification results in biases (we no longer estimate the correct thing)

Data-adaptive estimators

- ▶ In routine practice, we are advised to check model fit (or covariate balance)
- ▶ and add non-linear terms and interactions to our parametric model as needed
- ▶ Why not just estimate these models non-parametrically ?
- ▶ Example: use (machine learning) ML to estimate $Q_1(X) \equiv E(Y|A = 1, X)$ as $\hat{Q}_{1n}(X)$, subindex n emphasises it was learnt on sample size n (usually dropped)
- ▶ giving rise:

$$\bar{Q}_1 = \frac{1}{n} \sum_{i=1}^n \hat{Q}_{1n}(X_i)$$

- ▶ But they **are** problematic

Challenge 1: Bias

- ▶ Data-adaptive methods are optimally tuned to **minimise prediction error**
- ▶ this can introduce **bias** due to **oversmoothing** over the observed data range, which is useful for prediction over that range, but may induce severe bias in causal effect estimates;
- ▶ In addition, if using an algorithm that does variable selection, it is possible to mistakenly throw out important confounders.
- ▶ **Undersmoothing** can reduce some of these biases.
- ▶ However, it is not readily clear how this should be done...

Challenge 2

no valid uncertainty!

plug-in estimators based on ML predictions are 'easily' obtained

- ▶ They may have complex distributions and variability is difficult to quantify.
- ▶ It is not clear how the uncertainty (in the ML estimation of the nuisance models) propagates into the SEs of the final estimator
- ▶ In general, bootstrap is not valid for these estimators

Summary

- ▶ Outcome modelling (g-comp) is valid to infer causal effects (provided we have conditional exchangeability given a sufficient adjustment set):
 - ▶ it is vulnerable to extrapolation
 - ▶ does not make extrapolation visible or explicit
- ▶ Alternatively, IPW methods are not vulnerable to extrapolation **but** result in larger standard errors
- ▶ naive use of machine learning estimation leads to bias and invalid uncertainty quantification
- ▶ **Looking ahead:** Double-robust procedures are the basis for using data-adaptive methods appropriately