Causal Estimation

Robin Evans, Vanessa Didelez and Karla Diaz-Ordaz

University of Oxford, BIPS, University of Bremen, and UCL

APTS Week 4, Southampton
September 2025

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 previously used causal diagrams to 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):

$$E\left(Y(1)\right) - E\left(Y(0)\right)$$

- For binary Y:
  - Risk ratio:

$$E\left(Y(1)\right) / E\left(Y(0)\right)$$

  - Odds ratio:

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

Identification

- Consider ATE $= E\left(Y(1)\right) - E\left(Y(0)\right).$
- 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):*

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

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

- **Conditional (mean) exchangeability**:

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

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

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

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

$$\text{ATE} = E(Y \mid A = 1) - E(Y \mid A = 0).$$

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:
    $E\left(Y(a)|X=x\right)$ can be replaced by $E\left(Y(a)\,|A=a,X=x\right)$, for $a=0,1$.
  - by consistency
    $E\left(Y(a)\,|A=a,X=x\right)$ can be replaced by $E\left(Y\,|A=a,X=x\right),\forall a,x$.
- Hence
$$E\left(Y(a)|X=x\right)=E\left(Y\mid A=a,X=x\right)$$
- 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 $Y(1)$)

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

- By consistency,

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

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

$$\text{ATE} = E_X[E[Y|A=1,X] - E[Y|A=0,X]]$$

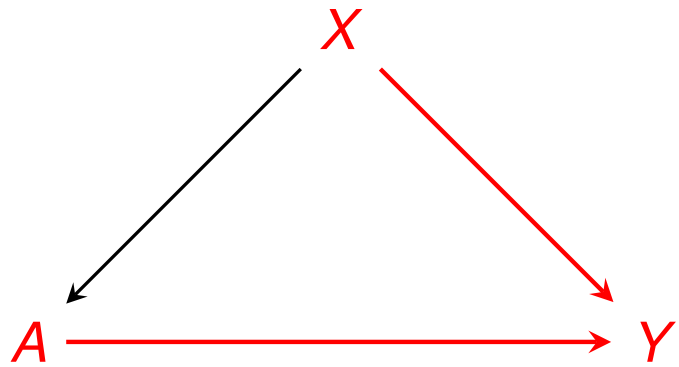
Estimation

- 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 $=$

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

- 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

plug-in estimation general X

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

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

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 is correctly specified

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`
- we can fit a saturated model

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

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

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

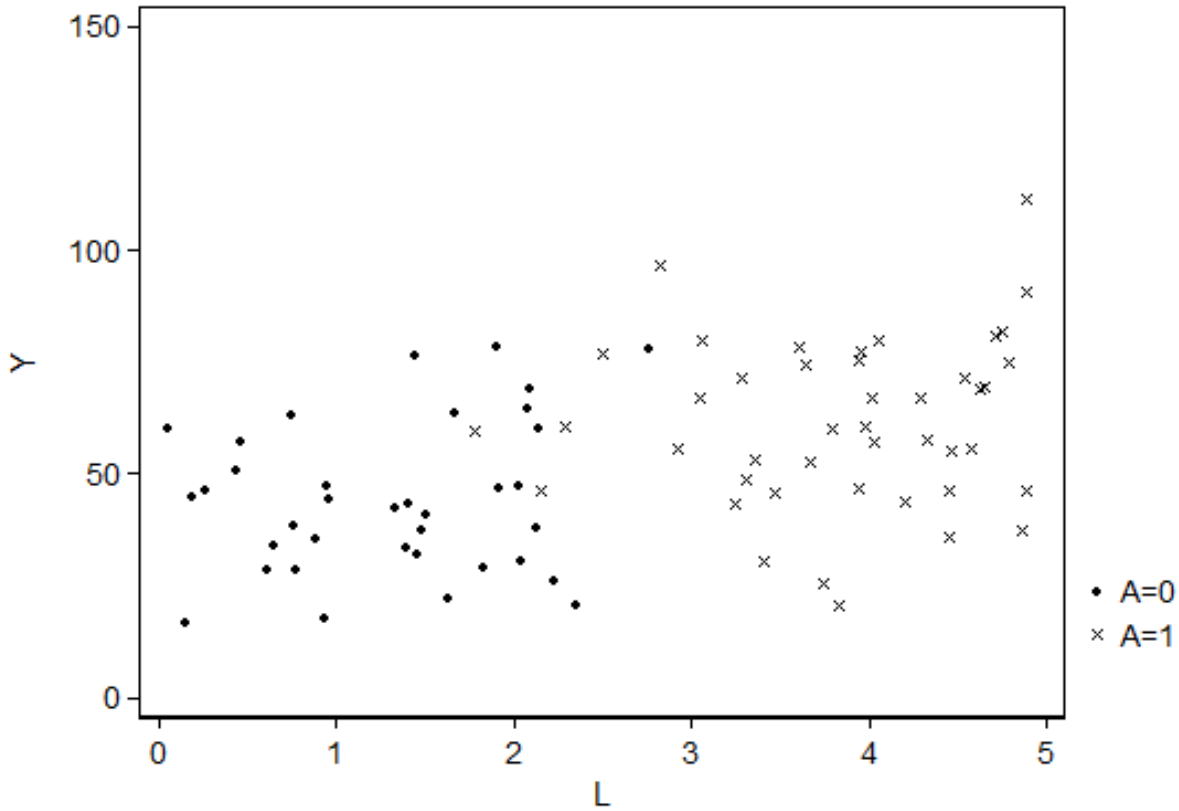
- This is related to positivity, which states that, for each possible level $X = x$,

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

- In a sufficiently large sample, for every observed value of X, positivity guarantees that there will be both exposed and unexposed individuals

Pitfalls of regression methods 1: Extrapolation

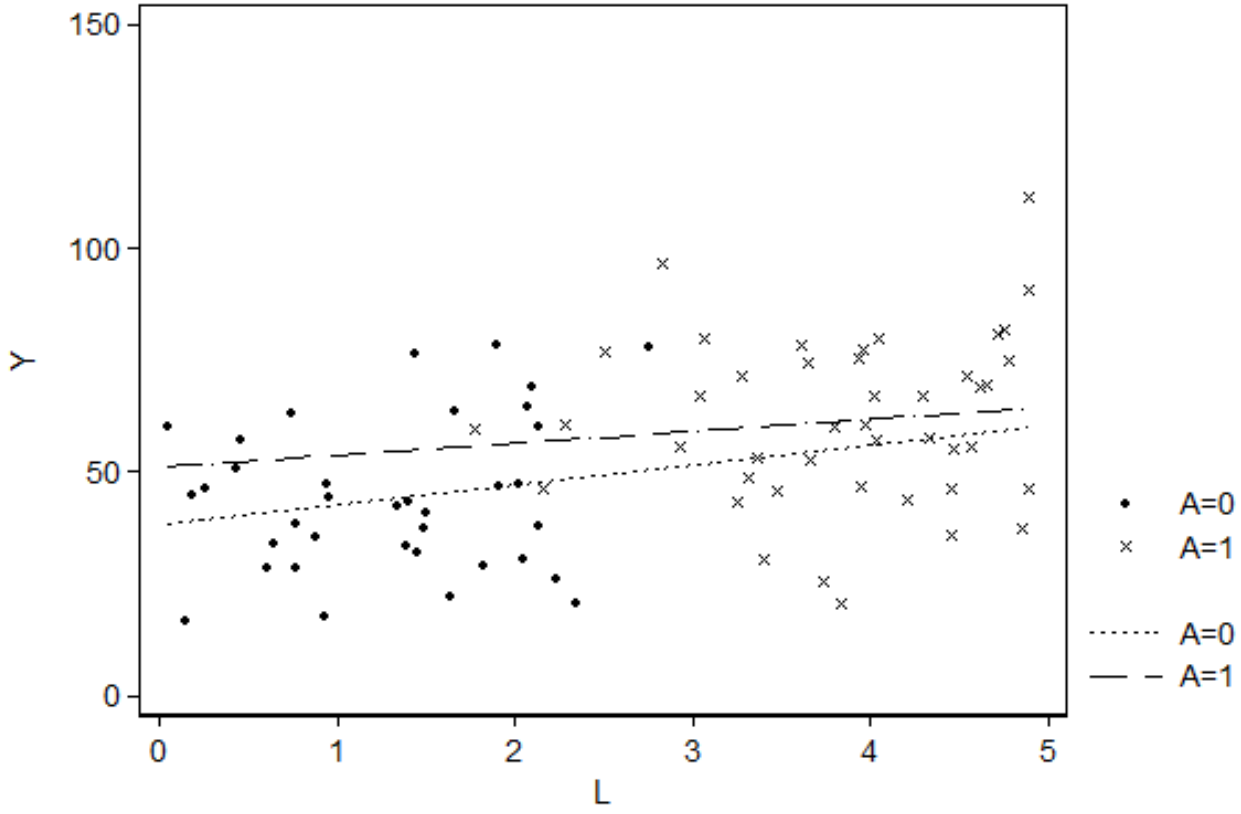
lack of positivity



Consider this simulated example, with one continuous variable L in the sufficient set. There is little overlap between the L-values of the $A = 0$ and $A = 1$ groups. This means that positivity is violated. For high values of L, everyone is exposed; for low values of L, everyone is unexposed.

Positivity violation

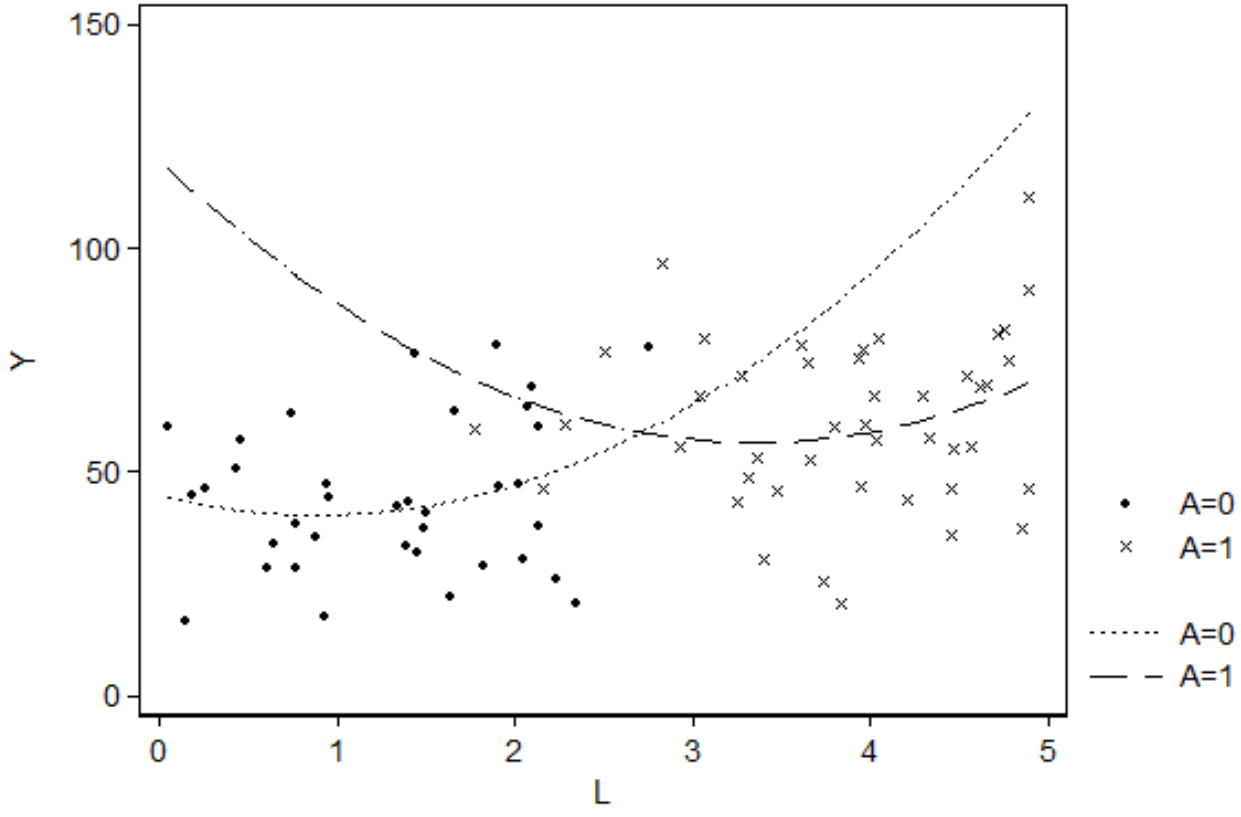
A simulated example (2)



There is little information in the data to choose between this linear model (with effect modification)...

Positivity violation

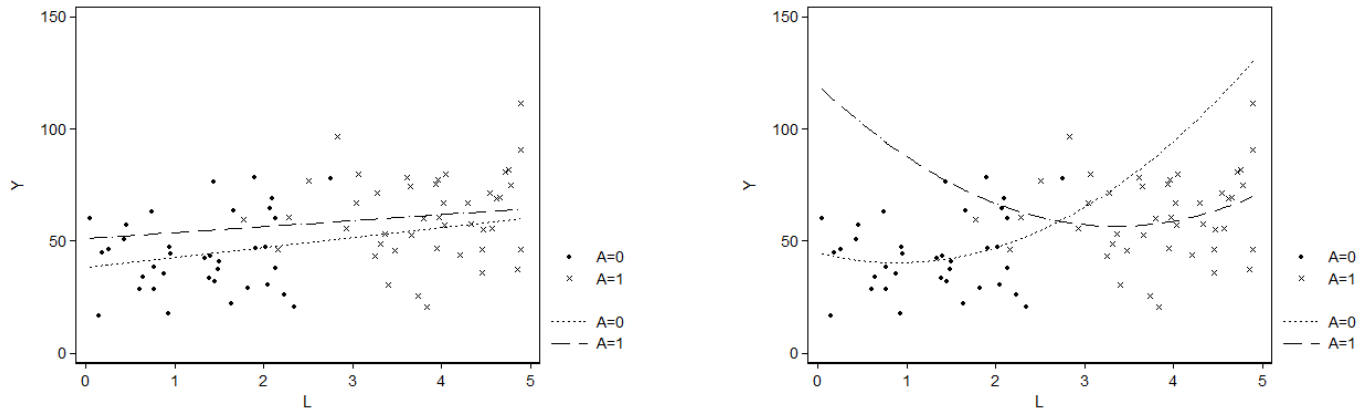
A simulated example (3)



. . . and this quadratic model, even though the estimated ATEs are v different.

Positivity violation

A simulated example (4)



- Both models fit the data almost equally well, but lead to different estimates of the causal effect.
- We can only check how well the model fits the observed data, but, whenever there is lack of positivity, the regression-based estimator of the causal effect relies on the extrapolation of these fitted relationships to regions where there is little data to support the model choice.
- Regression methods flag this only very mildly (via slightly increased SEs) and proceed to give 'precise' estimates based on extrapolations.

Pitfall 2: finite-sample bias

Logistic and Cox regression

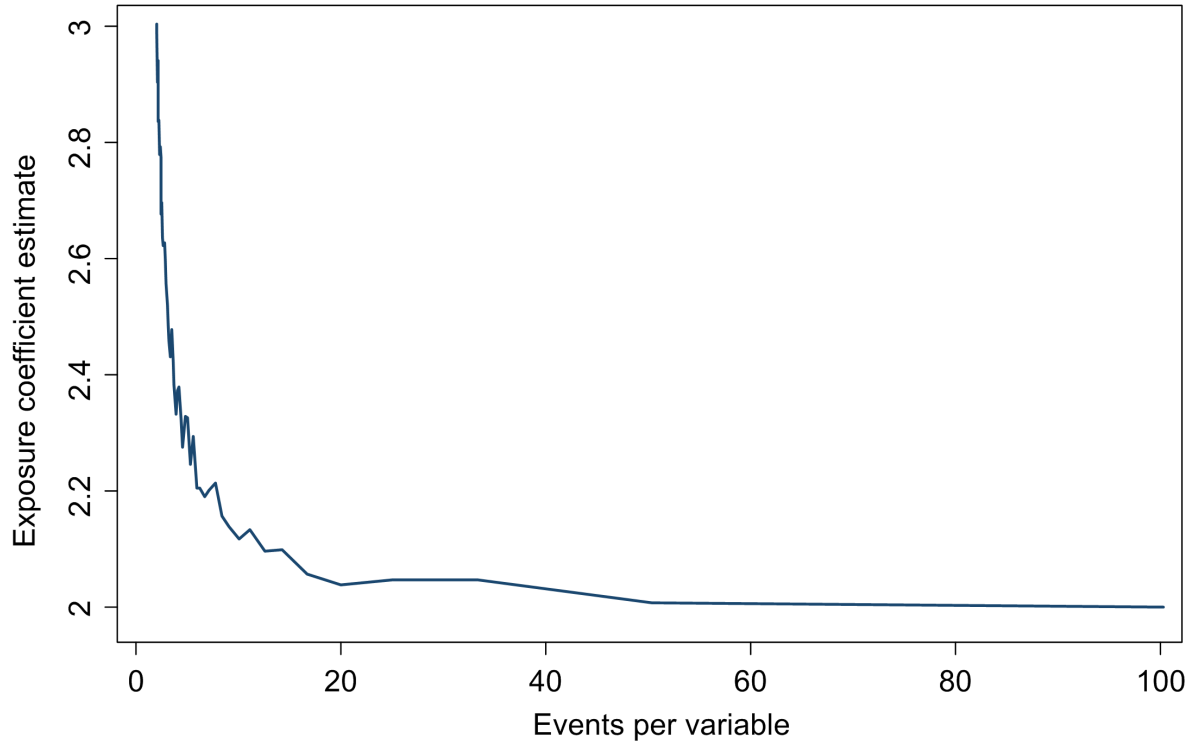
- The parameters of regression models are typically estimated by maximum likelihood.
- ML estimators are, in general, only asymptotically unbiased.
- For logistic and Cox regression, ML estimators can be noticeably biased in small samples.
- In particular, bias increases as the number of events per parameter decreases.
- The higher dimensionality of X (the valid adjustment set), the larger this finite-sample bias.

Reference

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

Finite sample bias in logistic regression

A simulated example

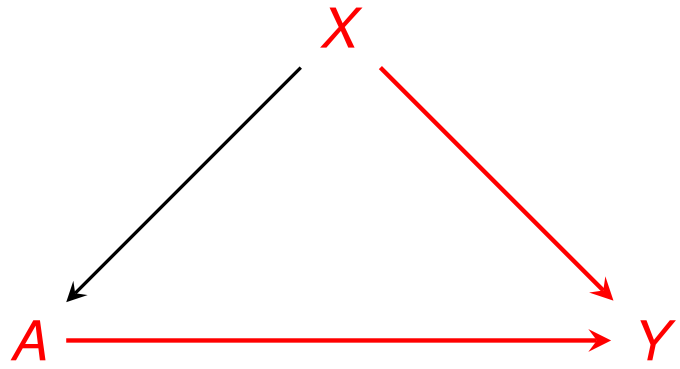


Here are the standard logistic regression estimates of an exposure log odds ratio (true value = 2) when adjusting for different numbers of (independent) covariates.

Alternative estimation streategy

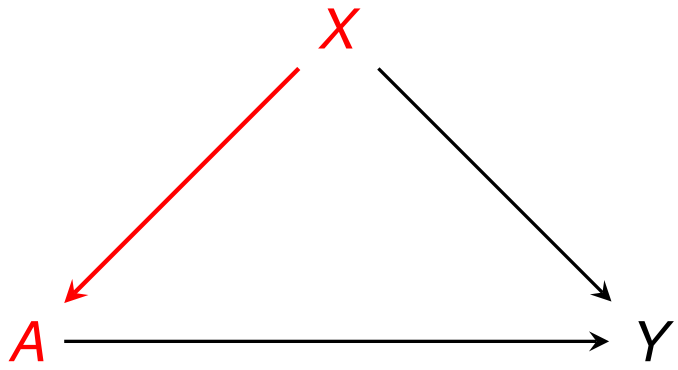
Basic idea (1)

Instead of modelling this:



Methods based on Treatment Modelling

We model A given $\mathbf{X}$



Key reference

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

Alternative Identification:

Inverse probability weighting

- we can show (under our identification assumptions):

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

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, $Pr[A=a \mid X] \neq 0$, by iterated expectations we have

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

- by conditional exchangeability

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

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

Estimation 2: IPW

Horvitz-Thompson estimator

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

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

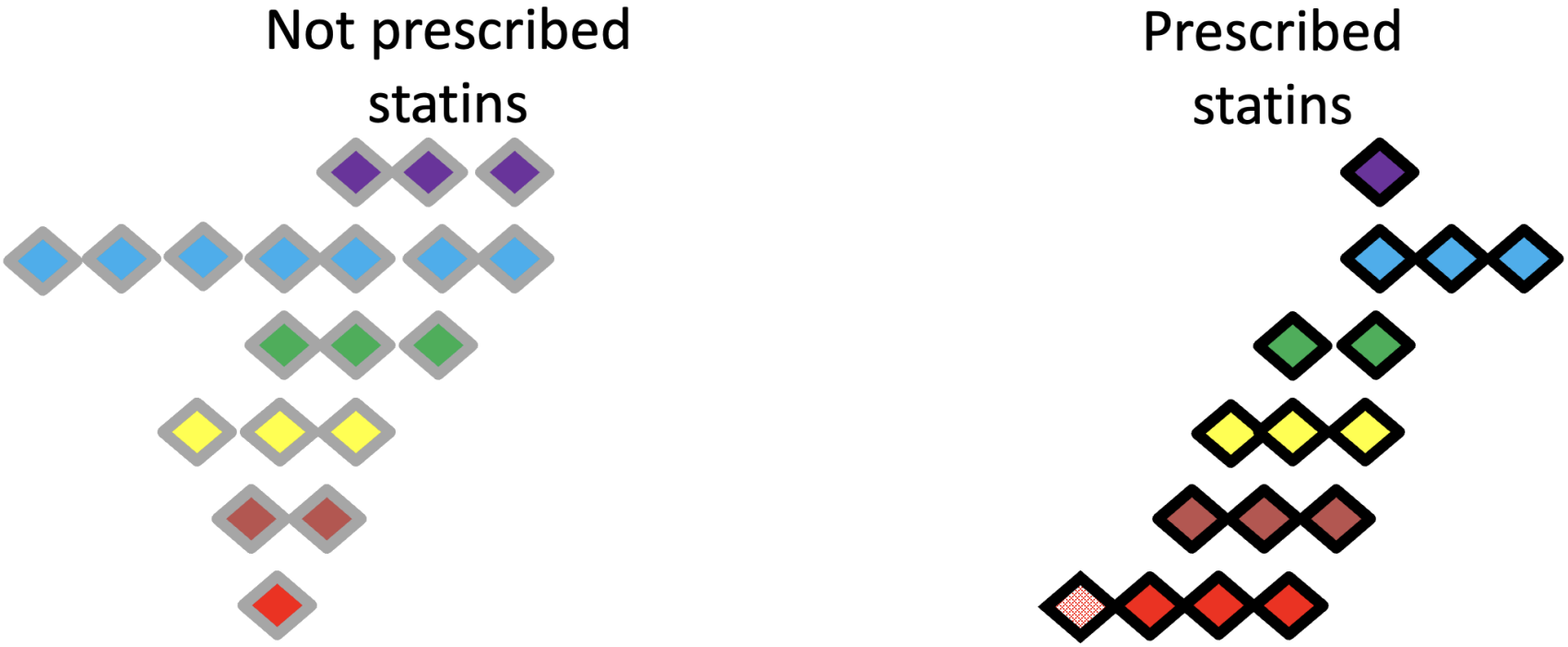
  3. For the ATE, do the corresponding steps for $E[Y(0)]$ and take the difference
- One disadvantage is that it is not guaranteed to respect the range of the outcome.

Intuition– pseudo-population

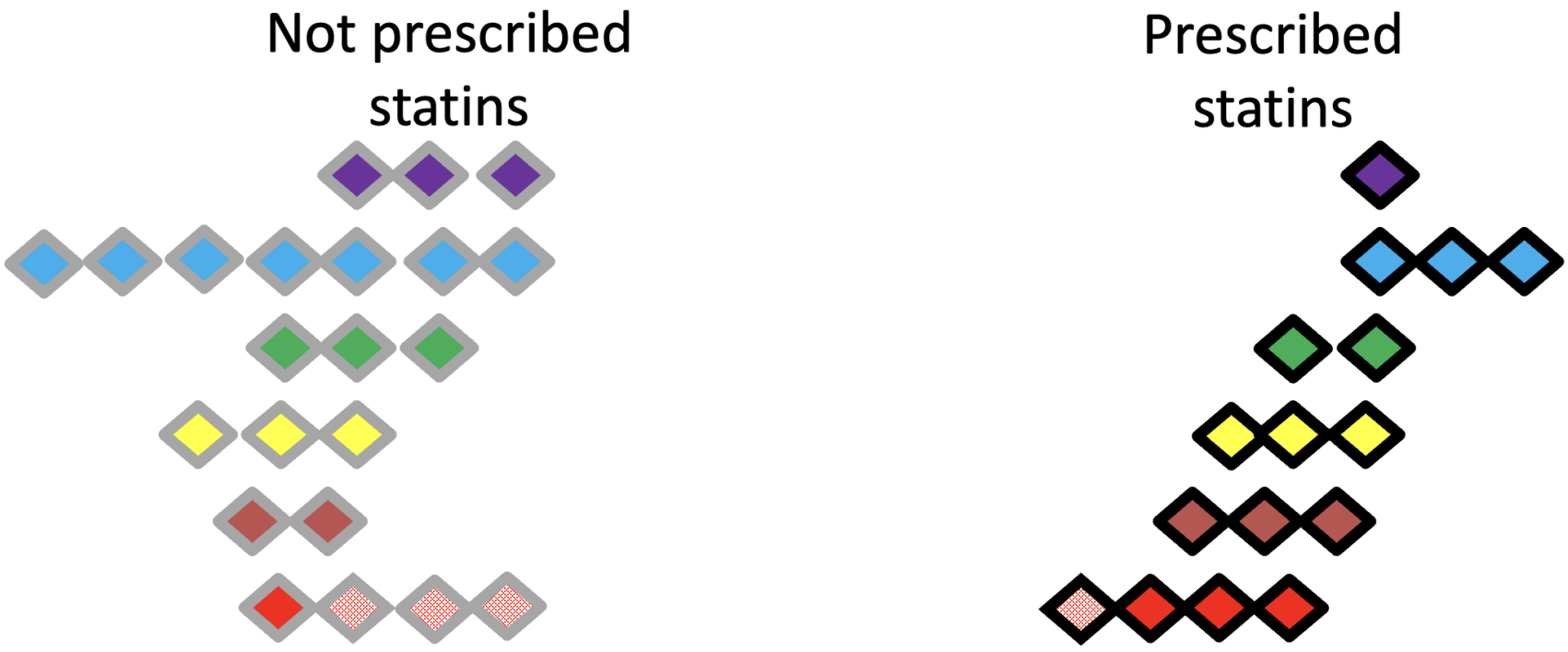
- So, we **reweight** observations by the reciprocal of the propensity of the treatment actually received.
- this creates a **pseudo-population** in which individuals are assigned treatment independently of any confounding variables.
- let's see an illustration

Basic idea

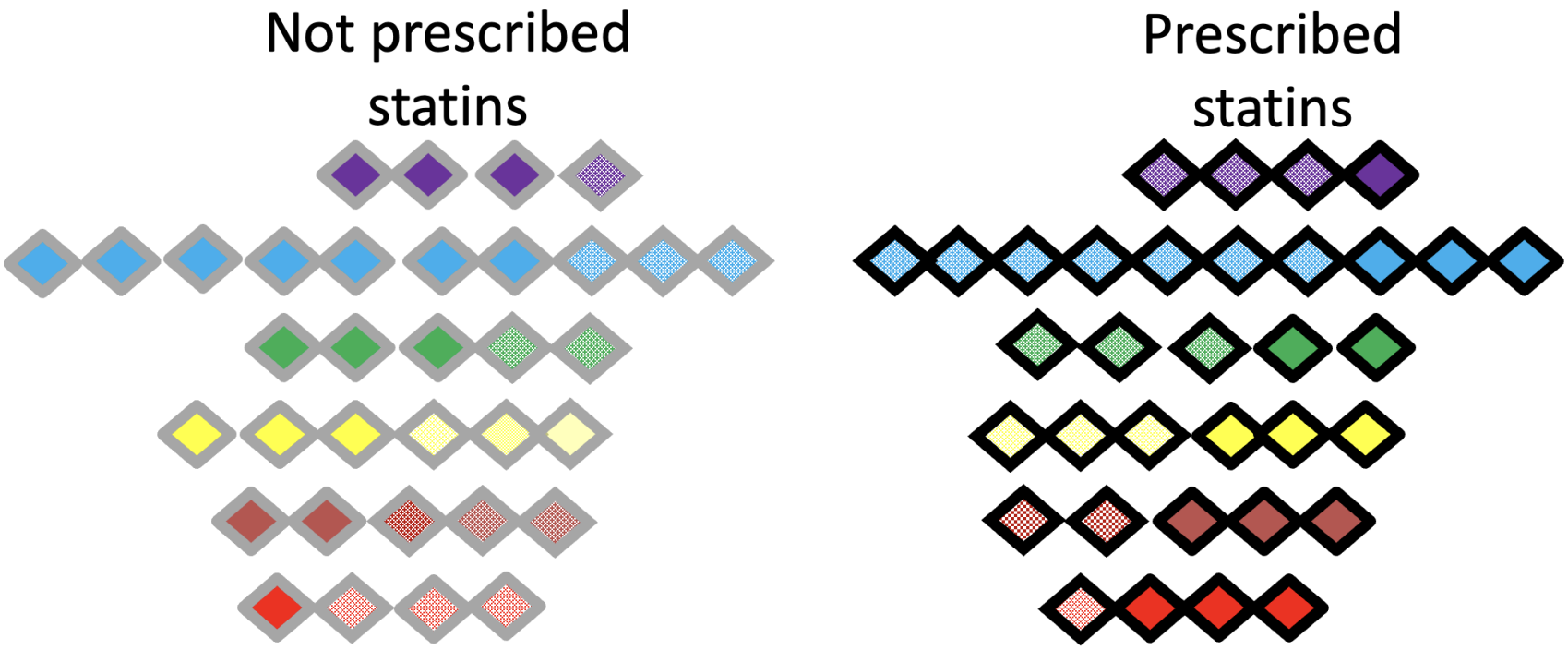
- We have individuals prescribed statins (or not) according to X, a score (represented here by colour)
- We want to know what the outcome would have been if all individuals in each colour remained untreated vs all treated



- 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 extra each $\times 3 = 1$

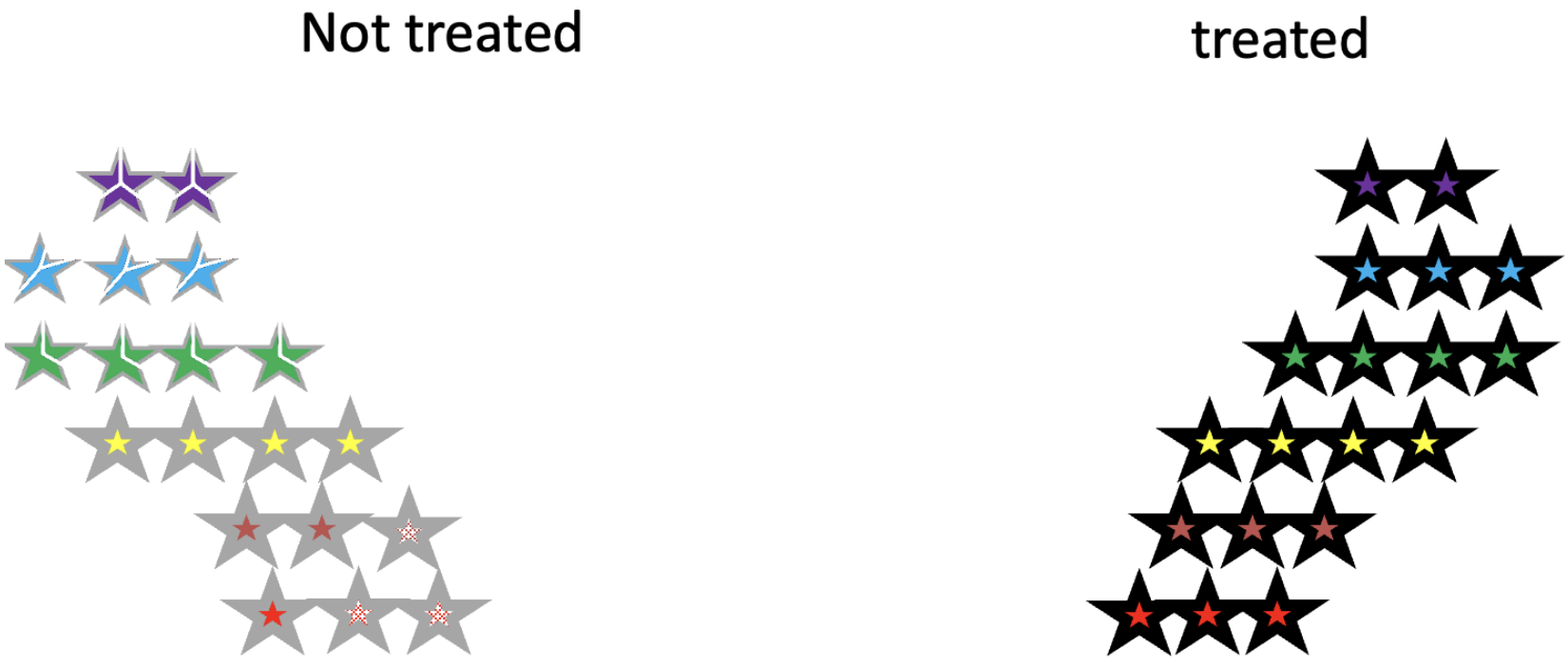


- The one untreated red (grey border) needs to represent 3 others. This means multiplying it by 1/0.25 this is the inverse of the prob of getting its own treatment conditional on its level of *X*
- To remove the association between *A* and *X*, we want to *balance* the numbers of each colour getting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



- 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

Pseudo-population for the ATT



- Alternatively, we **reweight** untreated observations by the reciprocal of the propensity of the treatment actually received, $w = (1 - \hat{\pi}(X))^{-1}$
- but the treated retain a weight =1
- and take the difference of the resulting averages, we obtain ATT

Stabilized weights

- The goal of IP weighting is to create a pseudo-population in which there is no association between the covariates X and treatment A
- there are other ways to create such a pseudo-population:
- i.e. a pseudo-population in which all individuals have a probability of receiving $A = 1$ equal to 0.5 and also equal $Pr(A = 0 = 0.5$, regardless of their values of X
  $w = 0.5/Pr(A|X)$
- This is useful to **stabilize** the weights, by multiplying by some arbitrary marginal distribution $Pr(a)$
- A common choice is to assign to the treated the probability of receiving treatment $Pr[A = 1]$ in the original population, and to the untreated the probability of not receiving treatment $Pr[A = 0]$

Assumptions for IPW

- assumes positivity $0 < P(A_i = 1|X_i = x) < 1$, for all x.
- If this doesn't hold, then reweighting is hopeless.
- 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

- relies on the PS model being correctly specified

IPW continued

- 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`

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)
```

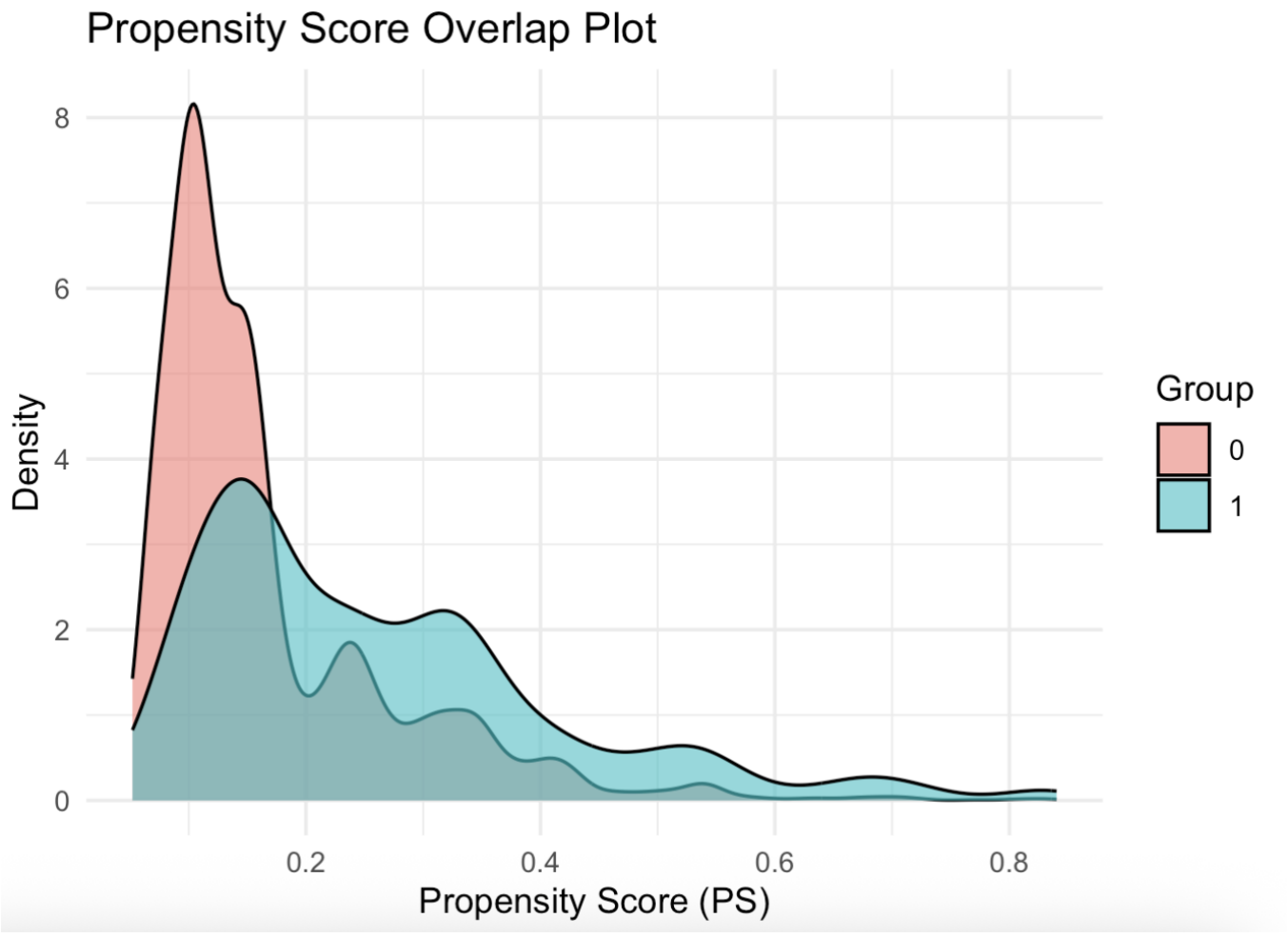
- Estimated ATE is -230.49

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

If there are some very large weights:

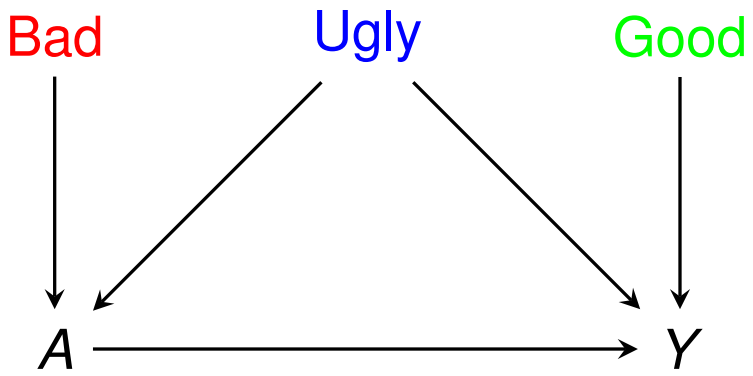
- The positivity assumption is required, otherwise we have infinite weights.
- Estimated propensity score close to 1 or 0 are problematic, since they imply that a few individuals will receive a very large weight, leading to imprecise and unstable estimates.
- Some observations have much more influence in the analysis than others
- Happens when units are treated contrary to expectation: Indication of unmeasured confounding??
- How to handle large weights:
  - Investigate why: who are they? are they some combination of variables that makes them "rare"? do they have extreme values for some confounders?
  - You can trim (0.01 or 0.02) (e.g. Lee, 2011) : it changes the population to whom inference applies
  - or truncate the weights introduces some bias, as you're not weighting properly, but reduces the variance

Looking at the overlap



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?

- trimming the tails of the PS
- this removes subjects who have extreme PS values
- makes the Positivity assumption more plausible
- but the resulting causal effects apply to a subset of the population
- truncating is also done
- better to try to diagnose the source of poor overlap
- recode /group variables to avoid sparse values

Checking balance

- We have calculated a PS, and we have chosen our estimator (e.g. Matching or IPW)
- recall that the goal was to achieve balance in the distribution of the confounders for the 2 groups that we use to calculate the causal effects
- after matching or weighting we need to check that we are happy with the achieved balance
- this can be done by reporting a Table 1 or a plot of the characteristics of the confounders stratified by the exposure in the match sample or the re-weighted sample
- you should report the balance (table or plot, or table of standardised mean differences)

Standardised differences

- A standardised difference is the difference in means between groups, divided by the (pooled) standard deviation.

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

- typically, report the absolute standardised differences
- rule-of-thumb, differences should be < 0.1

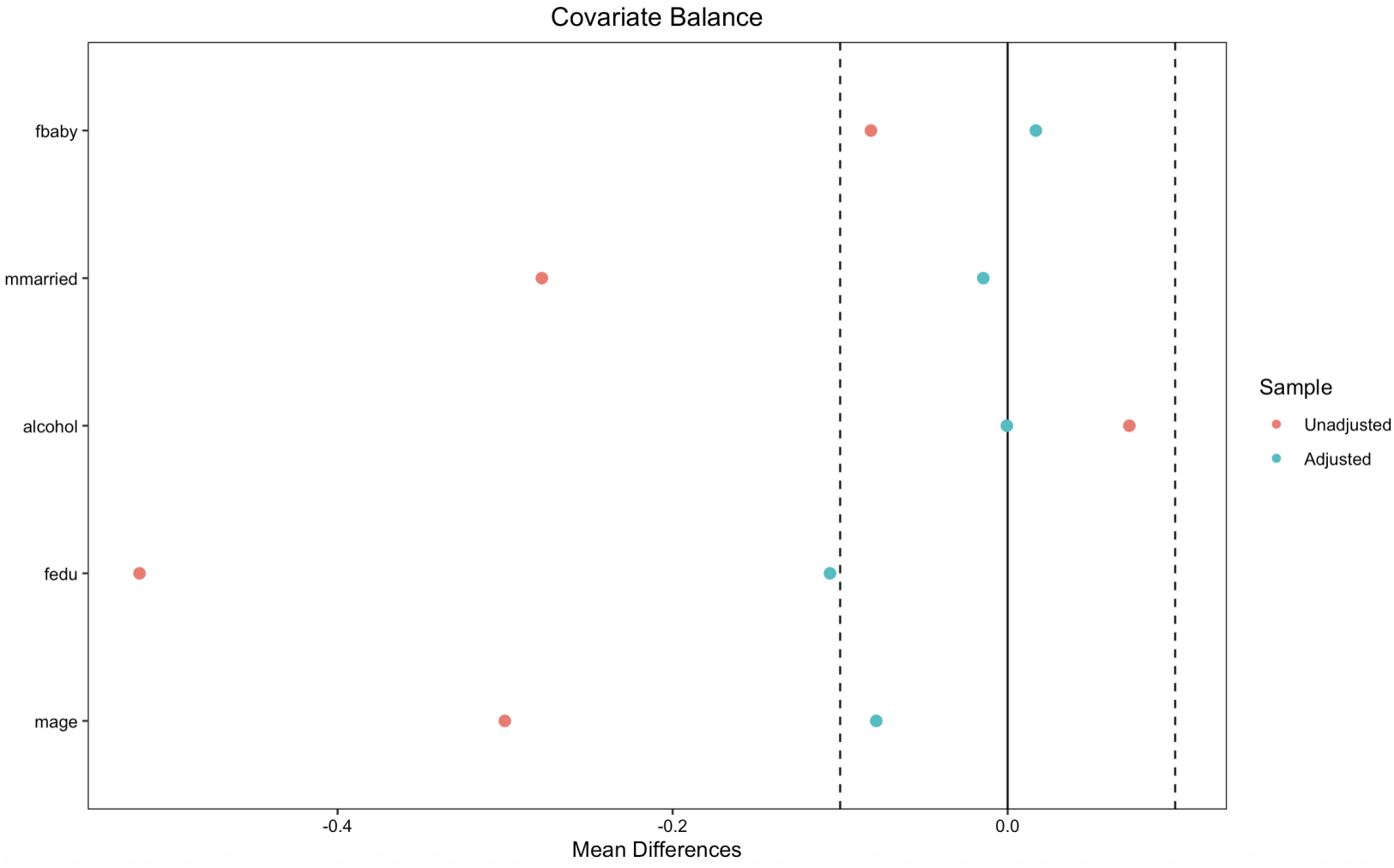
Weighted Standardised differences

- Calculate the std differences on weighted means and weighted variances.
- Stratify on treatment group, find weighted mean and weighted variance for each group
- This can be done by hand or with software used for surveys (e.g., svydesign in R)
- Take difference in weighted means and divide by an estimate of the pooled (weighted) standard deviation

Imbalance?

- Can we 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
- in R, can use the package `cobalt`

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

Doubly Robust Approaches

Note we've seen that if we specify

- the **outcome model** (i.e. $E(Y \mid A, X)$) correctly, we can obtain a consistent estimate of the ATE by averaging over the empirical X values;
- the **propensity score model** (i.e. $E(A \mid X)$) correctly, we can use 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!

We posit **parametric** *working models* (also called nuisance models)

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

Doubly Robust Methods

- 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 'nuisance' models Q_a and π to the data (e.g. by maximum likelihood), to obtain parameter estimates $\hat{Q}$ and $\hat{\pi}$
- The estimate is consistent if **either** model is correctly specified: this property is called **double robustness**

DR: consistency if either model is correctly specified–sketch

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

$$\mu_1^{AIPW} = \sum_i \left\{ \frac{A_i Y_i}{\pi(X_i)} + \left(1 - \frac{A_i}{\pi(X_i)}\right) E(Y_i | A_i = 1, X_i) \right\}$$

- If $\pi(X_i)$ is correctly specified, (re-arranging terms)

$$\mu_1^{AIPW} = \sum_i \left\{ \frac{A_i Y_i}{\pi(X_i)} - \left(\frac{A_i - \pi(X_i)}{\pi(X_i)}\right) E(Y_i | A_i = 1, X_i) \right\}$$

- If $Q_1(X)$ is correctly specified, (re-arranging terms)

$$\mu_1^{AIPW} = \sum_i \left\{ \frac{A_i \left(Y_i - Q_1(X)\right)}{\pi(X_i)} + Q_1(X) \right\}$$

AIPW for ATE

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

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

- We call this the **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 in practice.

Summary

In this session we have:

- Reviewed the assumptions needed to identify causal effects.
- Noted how, under these assumptions, the conditional mean POs can be replaced by conditional expectations of observed outcomes.
- Shown how, under appropriate assumptions, regression models deliver estimates of expectations of observed outcomes and these can be used to derive estimates of conditional causal effects.
- Found that using standardisation of conditional effects (and of conditional mean POs) we can derive marginal causal effects.
- These marginal effects are marginal with respect to the distribution of X: hence they refer to populations which are characterised by that distribution.

Summary (continued)

- 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
- Double-robust procedures can increase precision again, and are less vulnerable to extrapolation

Summary 3

- Augmented IPW estimators, which are doubly robust, are therefore particularly appealing since they allow us to get ‘the best of both worlds’, improving efficiency and gaining greater robustness to model misspecification.
- IPW and DR estimators generalise to more complex situations e.g. to estimate the effects of time-varying exposures in longitudinal studies.
- Double-robust procedures are the basis for using data-adaptive methods appropriately
- this helps alleviate model misspecification concerns