Causal Learning for Data Science

Vanessa Didelez
with much help from Ronja Foraita and Christine W Bang

Leibniz Institute for Prevention Research & Epidemiology – BIPS

December 2021

Data Train — UBRA Bremen

Part 2: Estimating a Causal Effect

Basic Setting

$X =$ binary (point-)treatment

$Y =$ some (numeric) outcome

(not survival / duration — that's special)

$C =$ sufficient adjustment set of pre-treatment covariates

Keeping it simple to focus on principles!

REFERENCE: Goetghebeur, E, le Cessie, S, De Stavola, B, Moodie, EE, Waernbaum, I. Formulating causal questions and principled statistical answers. Statistics in Medicine. 2020; 39: 4922– 4948.

Defining '(Point-)Treatment'

Oscar-winners live longer

$X =$ binary (point-)treatment

Well-defined?

Beware of **immortal-time bias**:

$X =$ 'did patient ever receive drug ABC? (yes/no)'

$\longrightarrow$ not a point treatment!

Target trial:
define unique time of eligibility and treatment assignment!

(Total) Causal Effects

In words

Total marginal (or population) effect:

what is the overall effect of intervening in X on Y?

Target trial: randomise X, regression of Y on X

Contrast setting do$(X = 1)$ versus setting do$(X = 0)$ by some well-defined (but possibly hypothetical) intervention.

(Total) Causal Effects

Can consider subgroups

Total conditional (or subgroup) effect:

what is the overall effect of intervening in X on Y within a *subgroup*, e.g. women aged 50-60?

Target trial: restrict to subgroup, randomise X; regress Y on X.

Note: subgroups relevant if we expect effect heterogeneity
$\Rightarrow$ nothing to do with confounding!
Finding such subgroups: active research

(Total) Causal Effect

Will focus on:

Formally: average causal effect

$$ACE = E(Y|\text{do}(X = 1)) - E(Y|\text{do}(X = 0))$$

or, with potential responses

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

aka: average treatment (ATE) or total causal effect (TCE), etc.

Key Assumptions

Consistency

Consistency Assumption:

If we observe $X^i = x$ then $Y^i = Y^i(x)$ (for individual i)

i.e. the outcome we observe under the observed treatment is the potential response had the treatment been *set* to what it was observed to be.

Violated if manipulation of X not well defined or so 'invasive' that observational setting not informative.

Example: X is 'BMI' — how to manipulate BMI itself?

Often: if violated, need more elaborate model and suitably detailed data.

Consistency

Under consistency and binary X:

$$Y^i = Y(1)^i X^i + Y(0)^i (1 - X^i)$$

Note:

consistency implicit in graphical / do$(\cdot)$ approaches

$\rightarrow$ invariance

No-Interference

Common assumption: **no-interference**:

Vector $\mathbf{x} =$ treatment values for **all** n units, then $Y^i(\mathbf{x}) = Y^i(x^i)$,

i.e. PR does not depend on treatment other units received.

Violation: e.g. vaccines, social networks.

Stable unit–treatment value (SUTVA):
consistency + no-interference.

No-Unmeasured-Confounding

Assumption of **no unmeasured confounding:**
(aka: random treatment assignment, or cond. exchangeability, ignorability, or ...)

Set C of observed (measured) pre–treatment covariates exists such that

$$Y(x) \perp\!\!\!\perp X \mid C$$

for all x to be considered as treatment values

Interpretation: within values of C, can consider X like randomised wrt Y

Denote: C is *sufficient* to adjust (control) for confounding;
or 'valid adjustment set'

Pre-Treatment Covariates?

What makes C **pre–treatment covariates**?

$\Rightarrow$ must be known not to be affected by intervention in treatment X!

Sufficient: C prior in time to X — but not necessary.

Often: C and X contemp. & share themselves common causes through past history, e.g. patient's medical history.

Graphically: C non-descendants of X.

(Overview: methods for causal covariate selection see Witte & Didelez, 2018)

Positivity Assumption (Overlap)

All methods for effect estimation essentially require

Assumption of positivity:

$$p(x \mid c) > 0 \text{ for all } x, c \quad (p(c) > 0)$$

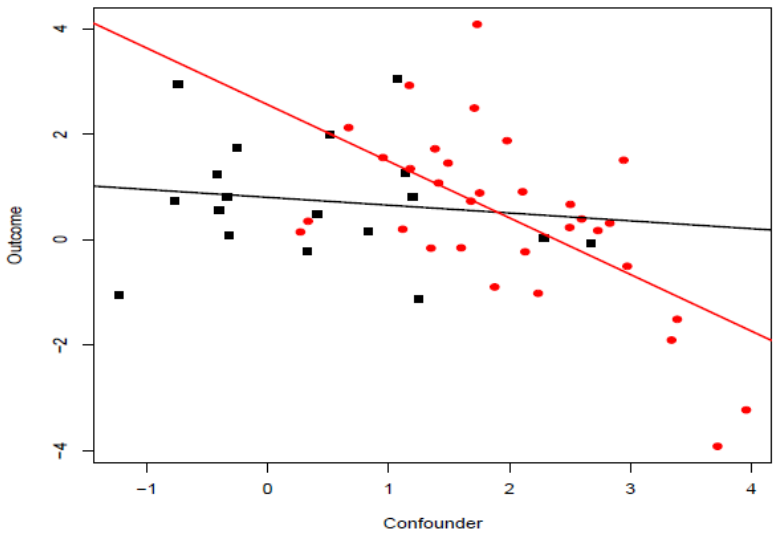
Interpretation: for all possible confounder values, it must be possible that a subject receives any value of treatment.

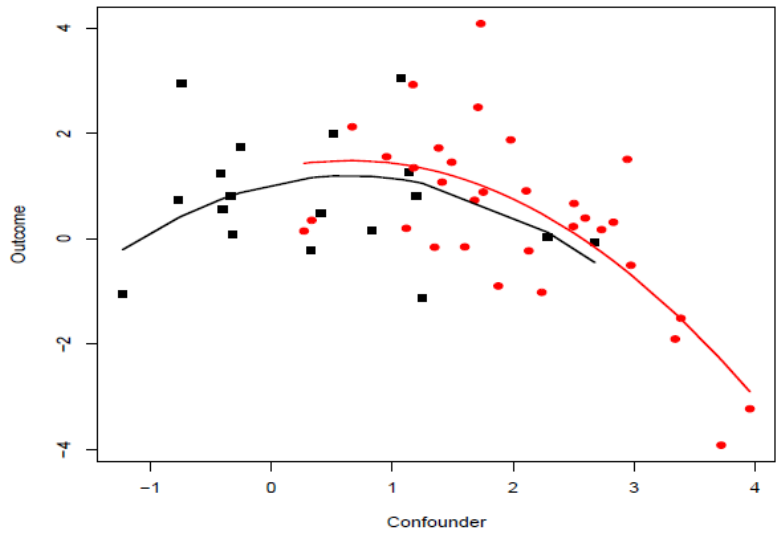
Positivity Assumption — Notes

- Positivity can be violated either by coincidence (small sample size), or **structurally**: certain combinations of C and X may not make sense!
- Target trial: can / should define eligibility criteria so that positivity is satisfied — requires domain knowledge
- (Lack of) positivity can be evaluated empirically (look at $p(x|c)$ or $p(c|x)$) — high-dim C becomes challenging Methods exist to characterise 'area of overlap' (Oberst et al, 2020)
- Do not include superfluous variables in C, especially: strong predictors of X that do not affect Y — can lead to apparent lack of positivity despite not being a problem

Positivity — Extrapolation

Regression-based approaches (based on fitting $p(y|x, c)$) may **mask** lack of positivity as regession models **extrapolate** $\Rightarrow$ can lead to vastly different causal effect estimates





Checking Assumptions?

- Consistency / no-interference:
  domain knowledge, study design
- No-unmeasured-confounding: compare analysis of observational data with actual randomised trial —
  Example: HRT-controversy
  Also: negative controls and similar designs
- Positivity (overlap) check:
  - basic: boxplot of each variable in C by treatment group;
  - advanced: consider **propensity score**, i.e. assess
    $P(X = 1|C = c)$
    obtain fitted values $\hat{p}^i = \hat{P}(X = 1|C = c^i)$ for each unit i,
    check $\hat{p}^i$ near zero in treated / controls, respectively.

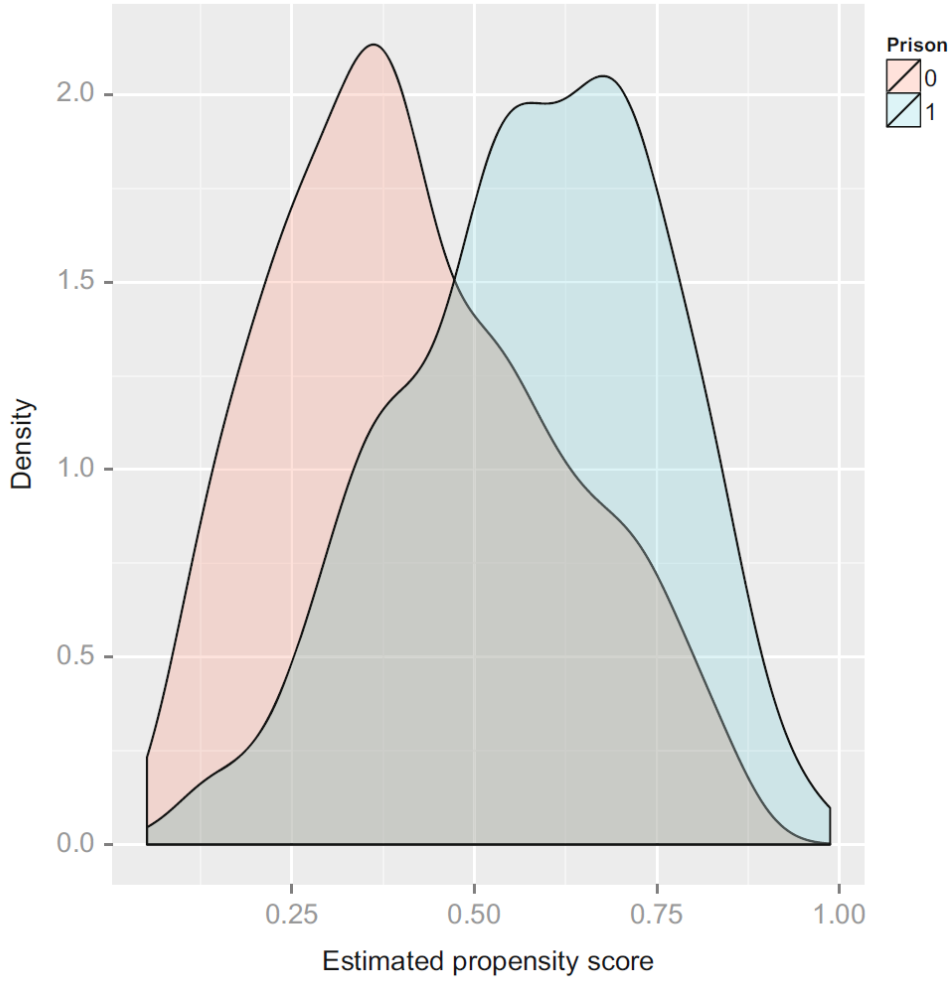
Propensity Score: Checking Positivity

Example:

$n = 1022$ offenders sentenced to either probation $X = 0$ or prison $X = 1$;
$C = 17$ covariates;
$Y =$ recidivism (yes/no);

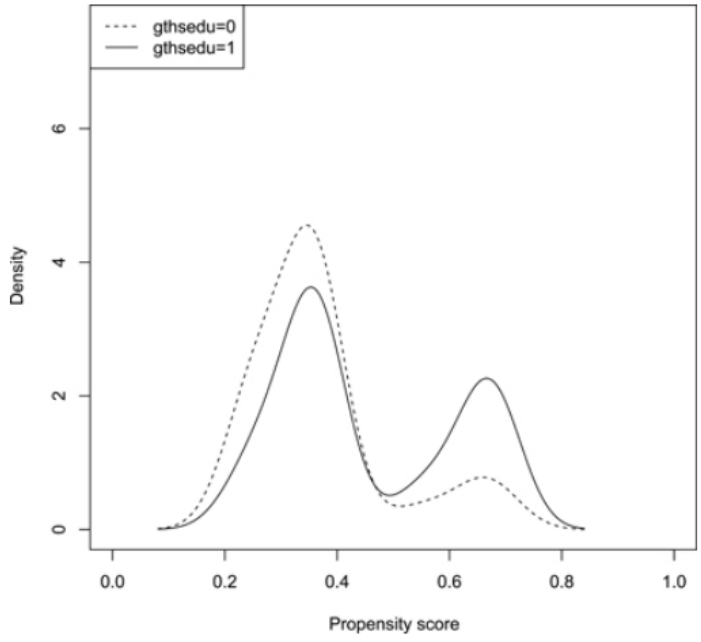
$\Rightarrow$ reasonable overlap.

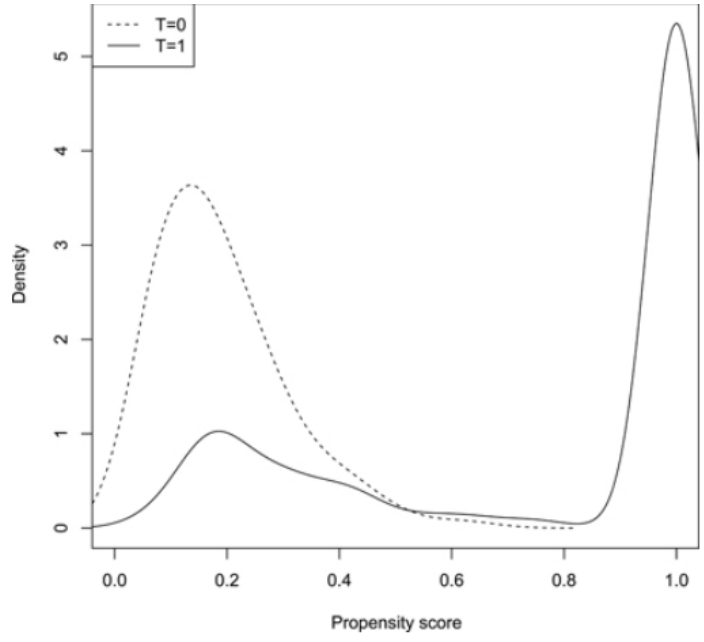
(Example taken from Guo et al., 2016)



Propensity Score: Checking Positivity

Which is good / bad overlap?

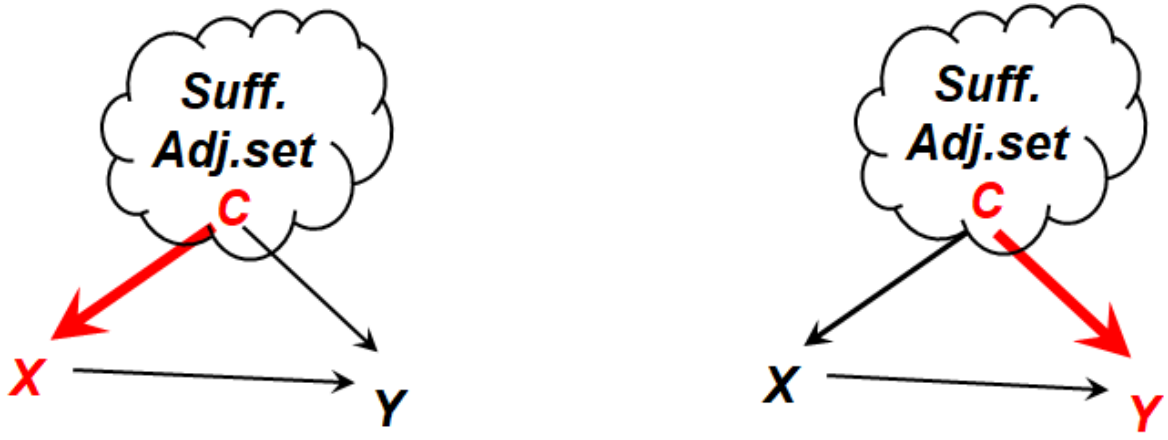




From: Brumback (2021, book)

Methods of Estimation

Based on treatment model $p(x|c)$ (propensity score)
or outcome model $p(y|x,c)$ or both



Principles:
– regression + standardisation,
– inverse-probability weighting (IPTW),
– stratification / matching,
– hybrid: doubly-robust estimation (double-ML)

Regression + Standardisation

aka: G-Formula *(Robins, 1986)*

Reminder: if C is sufficient set of covariates

$$E(Y \,|\, \text{do}(X = x)) = \sum_c E(Y|x, c)p(c)$$

An obvious way to use this is:

– fit flexible regression model for $E(Y|x, c; \beta)$ to data

– average over empirical C-distribution: $\sum_i E(Y|x, c^i; \hat{\beta})/n$

– R package `stdReg` (Sjolander and Dahlqwist, 2021)

Standardisation — Example

$Y =$ 'low birth weight' (binary); $X =$ 'mother smokes' (binary),
$C = \{$'age', 'race'$\}$ (Sjolander, 2016)

```
> fit2 <- glm(formula=lbw~(smoker+race+age)^2,
    family="binomial", data=clslowbwt)
```

flexible outcome model

Standardisation — Example

$Y =$ 'low birth weight' (binary); $X =$ 'mother smokes' (binary),
$C = \{$'age', 'race'$\}$ (Sjolander, 2016)

```
> fit2 <- glm(formula=lbw~(smoker+race+age)^2,
   family="binomial", data=clslowbwt)
> fit.std <- stdGlm(fit=fit2, data=clslowbwt, X="smoker",
   clusters="id")
> summary(fit.std)

  Estimate Std. Error lower 95 upper 95
0    0.279     0.0406    0.199    0.358
1    0.407     0.0555    0.298    0.516
```

flexible outcome model

standardised means
control / treatment groups

Standardisation — Example

$Y =$ 'low birth weight' (binary); $X =$ 'mother smokes' (binary),
$C = \{$'age', 'race'$\}$ (Sjolander, 2016)

```
> fit2 <- glm(formula=lbw~(smoker+race+age)^2,
   family="binomial", data=clslowbwt)
> fit.std <- stdGlm(fit=fit2, data=clslowbwt, X="smoker",
   clusters="id")
> summary(fit.std)

  Estimate Std. Error lower 95 upper 95
0    0.279     0.0406    0.199    0.358
1    0.407     0.0555    0.298    0.516

> summary(fit.std, contrast="difference", reference=0)

  Estimate Std. Error lower 95 upper 95
0    0.000     0.0000  0.00000    0.000
1    0.128     0.0681 -0.00544    0.262
```

flexible outcome model

standardised means
control / treatment groups

difference, i.e.
estimated ACE

Regression, but Why Standardisation?

Why not just look at coefficient ('effect') of X in a regression model for $E(Y|X, C)$?

- Marginal effect sensible summary also with arbitrary interactions / complex models
- Contrast of marginal $E(Y \,|\, \text{do}(X = x))$ corresponds to randomised trial where covariates C can be / are ignored
- Further issue: **non-collapsibility**! logistic regression / odds ratios not collapsible.
    - If set of sufficient covariates C not unique, cond. effects may depend on choice of C, but not marginal ones.

Regression + Standardisation

- Consistency (asy. unbiasedness) of estimation relies on correctly specified model for $p(y|x, c)$.
- Danger of extrapolation: it can happen that the regression relation $p(y|x, c)$ is determined primarily by treated subjects in one region of C and control subjects in another...
- ... should not happen under positivity — must be checked!

Regression + Standardisation

- The method is special case of **G–formula** for sequential treatments (Robins, 1986).
- Population effect $E(Y \mid \text{do}(X = x))$ depends on *distribution* $p(c)$ of covariates target population

  $\Rightarrow$ not necessarily the same in different populations (e.g. age distribution). If $p(y|x, c)$ regarded as 'stable' across populations, then can just replace $\hat{p}(c)$ in the above by different covariate distribution for different populations

  (e.g. UK versus USA covariate distribution).

  $\Rightarrow$ 'Transportability'

Partial Dependency Plots

(Zhao & Hastie, 2021)

- Close relation between Friedman's **partial dependence plot (PDP)** for visualising black-box prediction methods and back-door adjustment / standardisation
- Under the causal assumptions, can interpret PDP like ACE even for *continuous* treatment $\approx$ standardising over adjustment set C
- But: positivity hard to justify for entire range of a continuous treatment...
- PDP as basis for estimation of total causal effect very unstable and erratic asymptotic behaviour
  $\Rightarrow$ double-machine-learning! (later)

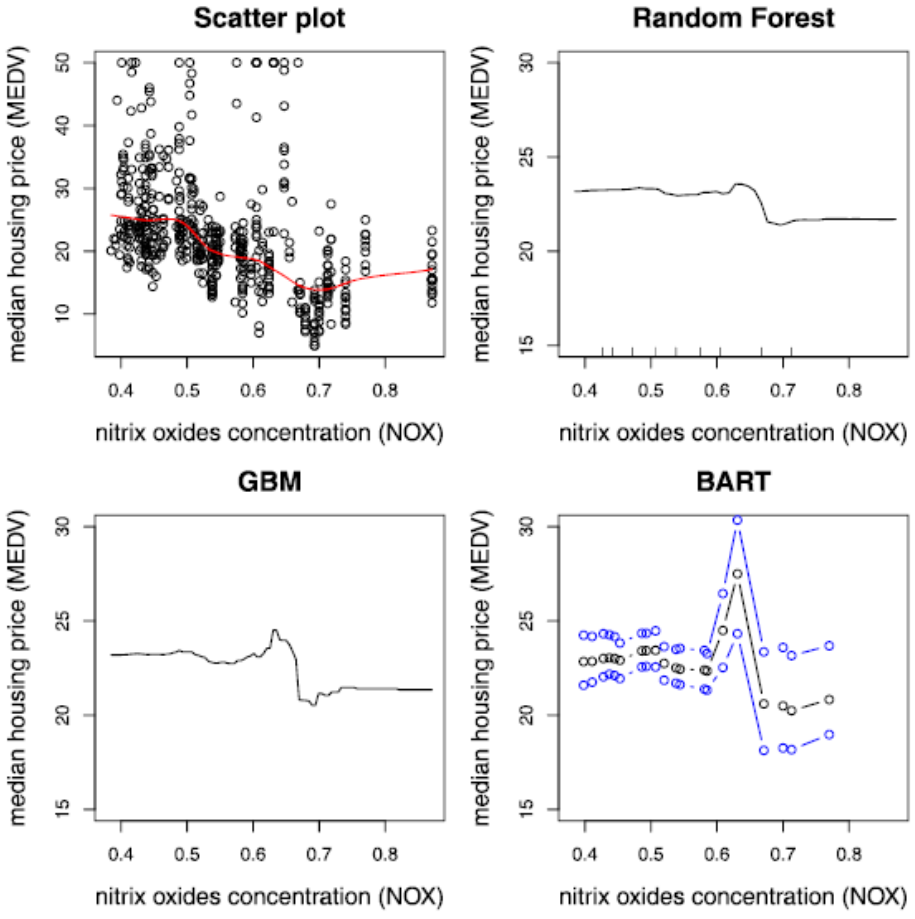
Partial Dependency Plots

(Zhao & Hastie, 2021)

Effect of nitrix oxides concentration on median value of owner-occupied homes

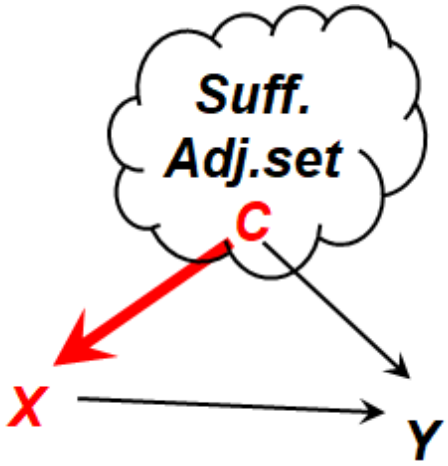
‘adjusted’ for: crime rate,
prop. residential/industrial zones,
av. # of rooms per dwelling,
age of the houses,
distance to city / highways,
pupil-teacher ratio,
% of blacks and
% of lower class

— valid adjustment set?



Treatment Modelling Approaches

The following methods are all based on models for X given C instead of modelling Y given C.



Marginal Structural Models (MSMs)

The adjustment formula

$$\sum_c p(y|x,c)p(c) \quad \text{or} \quad \sum_c E(Y|x,c)p(c)$$

might be 'awkward', $E(Y|x,c)$ non–linear with interactions, or C high dimensional and/or partly continuous.

$\Rightarrow$ Parameterise $E(Y \,|\, \text{do}(X = x))$ itself?!

$\Rightarrow$ **Marginal structural models (MSM)**

$\Rightarrow$ fitted by inverse probability of treatment weighting (IPTW)

Marginal Structural Models

(Hernán et al, 2001)

MSM: *semiparametric* model for

$p(y \mid \text{do}(X = x))$ or more typically $E(Y \mid \text{do}(X = x))$

e.g. linear, logistic, CoxPH, loglinear, probit etc.

Marginal: refers to time-varying covariates $\rightarrow$ not covered

Structural: model under intervention in X (not observational)

Note: term 'structural' is used in many different ways — here it always refers to modelling the underlying causal relationships.

Inverse Probability Weighting

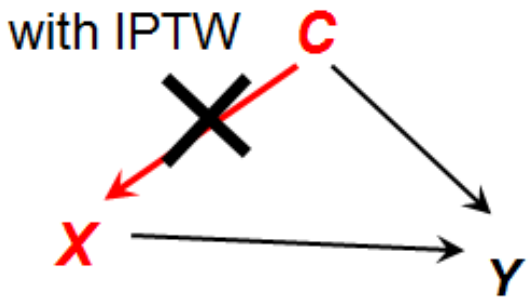
(IPTW)

Idea is based on:

$$p(y \,|\, \text{do}(X = x))$$

$$= \sum_c p(y|x, c)p(c)$$

$$= \sum_c \frac{p(y,x,c)}{p(x|c)}$$



In words: the population is re-weighted so that X becomes independent of C.

Covariate balance check:
success of re-weighting can be assessed empirically in data!

Inverse Probability Weighting

(IPTW)

Idea is based on:

$$p(y \,|\, \text{do}(X = x)) = \sum_c p(y|x, c)p(c) = \sum_c \frac{p(y, x, c)}{p(x|c)}$$

$\Rightarrow$ fit MSM with individuals' weights $w^i = p(x^i|c^i)^{-1}$

$\Rightarrow$ creates 'pseudo sample' in which C is not confounding

$\Rightarrow$ unbiased estimating equations for parameters of MSM $E(Y \,|\, \text{do}(X = x); \beta)$

Here, $w^i = p(x^i|c^i)^{-1}$ is the inverse of the probability that individual i receives 'treatment' x^i given they have covariates c^i.

IPTW Estimator

Define $\pi(c) = P(X = 1|C = c)$ — propensity score.

Can show (under our assumptions):

$$E\left(\frac{X}{\pi(C)}Y\right) = E(Y \mid \text{do}(X = 1))$$

and similarly

$$E\left(\frac{1-X}{1-\pi(C)}Y\right) = E(Y \mid \text{do}(X = 0))$$

Proof: iterated conditional expectation.

IPTW Estimator

With model $\pi(C;\alpha) \Rightarrow$ plug-in $\pi(C;\hat{\alpha})$

IPTW yields consistent estimator for ACE

- if $\pi(C;\alpha)$ correctly specified
- can obtain sandwich standard errors or bootstrap, or theoretical asymptotical standard errors
- IPTW often large variance, wide CIs — reflects lack of information in areas with 'extreme weights'
- ... extreme weights indicate possible near violation of positivitiy
- Solution: restriction of relevant population and / or truncation of weights (e.g. at 99%-percentile).

MSM / IPTW — Implementation

Easy to implement with standard software for regression models by specifying weights

```
p.i <- glm(x.trt~c1*c2*c3,family="binomial")$fitted
w.i <- 1/(x.trt*p.i+(1-x.trt)*(1-p.i))
msm.out <- glm(y.out~x.trt,family="binomial",weights=w.i)
Coefficients:
            Estimate Std. Error  z value Pr(>|z|)
(Intercept) -2.02055    0.01270 -159.118   <2e-16
x.trt        0.13202    0.01753    7.532    5e-14
cov.out <- sandwich(msm.out)
sqrt(cov.out[2,2])  [1] 0.03218826
```

Note: default standard errors ignore variability in (estimated!) weights

$\Rightarrow$ sandwich st.error

Notes on IPTW

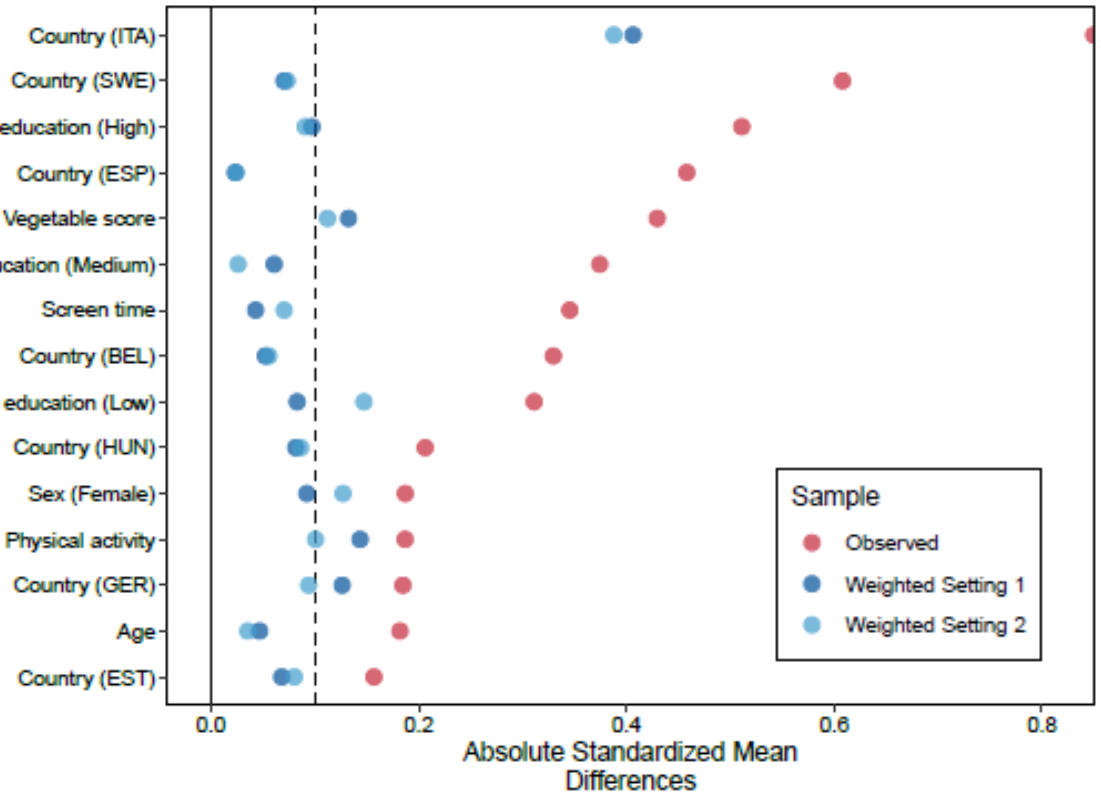
- Consistent when both, models for $E(Y \mid \text{do}(X = x))$ and $\pi(c)$ correctly specified...
- ... avoids modelling of Y-C relation
- Extension: 'overlap weights' — **new estimand** with most weights on subpopulations with higher overlap
- IPTW especially useful when study design (or other) supplies background knowledge to model weights $p(x|c)$.
- Problem: estimation of weights $p(x|c)^{-1}$ not obvious when X continuous.

Notes on IPTW

- MSM mimics a model for Y given X in the situation of a trial where X was randomised
- Even if X *actually* randomised, there may be changes to treatment status over time (non-adherence) — this then becomes a problem of **time-dependent treatment**
- MSMs with IPTW mostly used in longitudinal situations / time-dependent / sequential treatments with time-varying confounding
  $\Rightarrow$ 'marginal' over time-dep. confounders / covariates.

Checking Assumptions: Balance

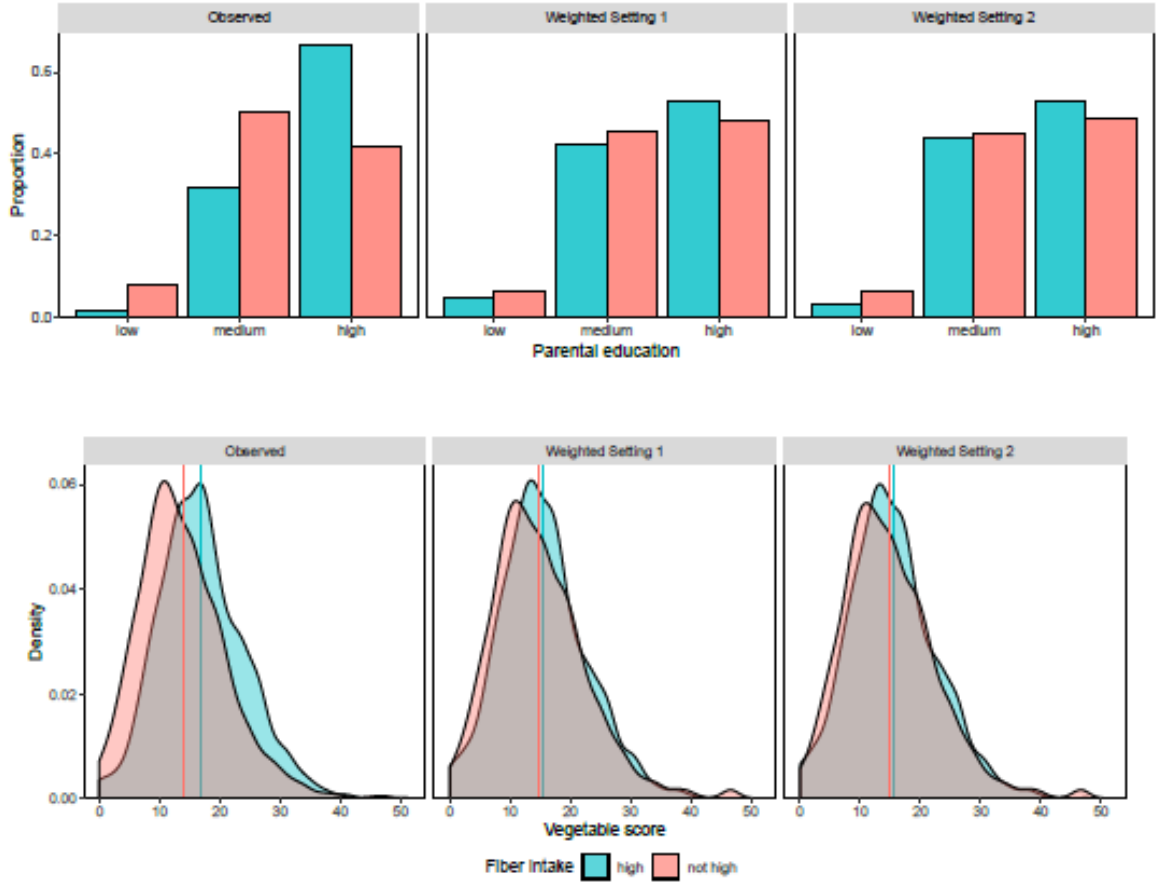
Example: causal effect of ‘fibre intake’ on children’s BMI; large adjustment set (country, parental edu, vege-score, etc.)
Plot: std.mean-diff b/w re-weighted treated / untreated sample



(Nöhren, 2021, MA thesis)

Checking Assumptions: Balance

Example(ctd.), checking balance of whole distribution of covariates



Double Robustness (DR)

(Robins & Rotnitzky, 2001)

- Regression-standardisation relies on correct outcome model
- MSM / IPTW relies on correct treatment model
- Danger (especially with high-dim C): models will be 'misspecified'

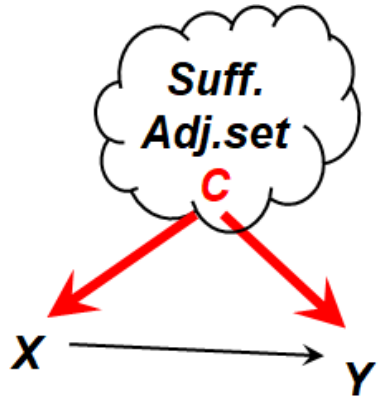
  $\Rightarrow$ want to fit them data-adaptively

  $\Rightarrow$ known to yield unstable (irregular) effect estimators!

  $\Rightarrow$ Better: **double-machine learning** of causal effects based on doubly-robust estimation.

Double Robustness (DR)

(Robins & Rotnitzky, 2001)



Can find class of augmented IPTW (AIP(T)W) estimators:

consistent for ACE if

- either model $\pi(C;\alpha) = p(X = 1 \mid C;\alpha)$ correctly specified

  or

- model for $p(y \mid x, c;\beta)$ correctly specified;
- but one of them can be wrong.

AIPTW Estimator

Basic Idea

Still binary exposure X
For *arbitrary* functions $m(C)$ and $\pi(C)$, define

$$\hat{\mu}_1 = m(C) + \frac{X}{\pi(C)}(Y - m(C)) = \frac{X}{\pi(C)}Y + \left[1 - \frac{X}{\pi(C)}\right]m(C)$$

Property (under our assumptions): if

either $m(C) = E(Y \mid X = 1, C)$ or $\pi(c) = P(X = 1|C = c)$, then

$$E(\hat{\mu}_1) = E(Y \mid \text{do}(X = 1))$$

Analogously for $\hat{\mu}_0$ and $E(Y \mid \text{do}(X = 0))$.

AIPTW Comments

- Kang and Shafer (2007) find: AIPTW with parametric models can be quite bad if *both models slightly misspecified* $\rightarrow$ much research on improvements
- **NEW:** Statistical properties of doubly-robust estimators allow the use of machine learning for treatment and outcome models
  - $\Rightarrow$ DR minimises slow convergence rates / overfitting typical for machine learning with sample-splitting or cross-fitting
- `AIPW` R package (Zhong et al., 2021) or `npcausal` R package (Kennedy, 2021)

'Demystifying statistical learning based on efficient influence functions', Hines et al (2021; arXiv:2107.00681)

Super Learner??

- To fit $m(C)$ and $\pi(c)$ can use data-adaptive methods developed for prediction!
  - `AIPW` package uses the Super Learner
  - ... an ensemble method allowing combination of several prediction algorithms into one
  - ... uses k-fold cross-validation to build the optimal weighted combination of predictions from a library of candidate algorithms
    – choice of library quite important (active research)
- double machine learning methods avoid strong modeling assumptions

  ... and can still achieve optimal $\sqrt{n}$ rate of convergence for causal effect estimation under *some* conditions

Propensity Score (PS) — Other Usage

(Rosenbaum & Rubin, 1983)

Have used $\pi(c) = P(X = 1|C = c)$

$\Rightarrow$ also known as **propensity score**.

Note: $\pi := \pi(C)$ is random variable.

MSM: used $\pi(C)$ for weighting.

But: can also use $\pi(C)$ for adjustment-type approaches, due to it being a balancing score...

Propensity / Balancing Score

(Still assuming: C sufficient set of covariates; X binary.)

Use of propensity scores (vs. IPTW) is based on

$$X \perp\!\!\!\perp C \,|\, \pi \qquad \text{i.e. } \pi \text{ balances } C$$

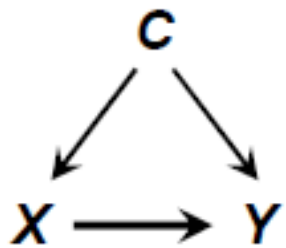
Hence (with properties of C):

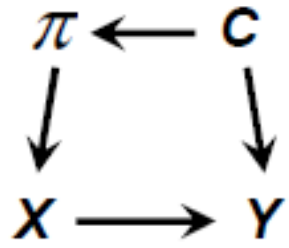
$$Y(x) \perp\!\!\!\perp X \,|\, \pi$$

making π a minimal sufficient reduction of C (univariate $\in [0, 1]$).

Propensity Score — Graphically

Propensity score $\pi := \pi(C) = P(X = 1|C)$ satisfies these conditional independencies:





Left: assumption of C being sufficient set of covariates.

Right: π is deterministic function of C and $X \perp\!\!\!\perp C \,|\, \pi$.

Propensity Score in Practice

- Estimate propensity score $\hat{\pi}$ with model for $\pi(C; \alpha)$.
- Required: correctly specified model $\pi(C; \alpha)$.
  Non-parametric approaches: random forests etc.
- Note: predictive quality of $\pi(C; \alpha)$ for X not important because need C to be (X, Y)-confounders, not nec. strong predictors of X.

  In fact: strong X-predictors $\Rightarrow$ bias amplification

  (Pearl, 2011).
- Check balancing property (see IPTW).
- Check positivity / overlap $\Rightarrow$ if necessary: restrict / prune!

Propensity Score in Practice

Methods for using PS (other than weighting):

- PS **stratification**: divide into strata (often quintiles) and fit $p(y|x)$ to each stratum separately.

  (Strata specific effects can reveal effect modification.)

  Then weighted average to obtain overall population effect.

- Alternative: **matching** on propensity score, i.e. match each treated with k untreated with similar propensity score — this estimates effect of treatment on the treated (ETT)!

- Sometimes: **PS adjustment** — specify model for $p(y|x, \pi)$ and fit with $\hat{\pi}$ plugged in.

- Extrapolation is automatically avoided.

Survival of Cancer Patients

Example

US National Cancer Institute's SEER data base; observational study.

Covariates: year of diagnosis, tumor size, geogr. registry, race, marital status

Propensity score	Treatment	No.	5-Year-Surv.	Difference
1st quintile	A	56	85.6%	
	B	1008	86.7%	-1.1%
2nd quintile	A	106	82.8%	
	B	964	83.4%	-0.6%
3rd quintile	A	193	85.2%	
	B	866	88.8%	-3.6%
4th quintile	A	289	88.7%	
	B	978	87.3%	1.4%
top quintile	A	462	89.0%	
	B	604	88.5%	0.5%

Overall estimated (weighted average) $ACE = -0.68$.

From strata specific results: slight suggestion that treatment B is better for those who are more likely to receive it.

Notes on Propensity Score

- Best with binary treatment / exposure.
- PS stratification consistent if $\pi(C;\alpha)$ correctly specified, but can be markedly biased due to residual confounding within strata possible.
- Consistency can be achieved by increasing number of strata when sample size is 'large' or by additional modelling of $E(Y|C=c;\mathrm{do}(X=x))$ within strata.
- PS popular especially for matching: π is 'one-dimensional reduction' of covariates — but at cost of first modelling / estimating $\pi = p(x|c;\alpha)$.
- PS matching / stratification not really suitable for sequential treatments.

Notes on Propensity Score

- Danger: modelling $\pi(C;\alpha)$ may focus on strong predictors of $X \Rightarrow$ can amplify bias! $\Rightarrow$ selection of C as adjustment set should be separate process from fitting $\pi(C;\alpha)$.
- Interpretation of PS-analyses sometimes regarded as difficult compared to actual covariate values.
- Simulations suggest that IPTW with π superior to stratification. (Lunceford & Davidian, 2004)
- Critique of PS-matching: King & Nielsen (working paper)

Estimating Causal Effects

Summary (no unobs. conf.)

Given suff. adjustment set C (& other structural asspts):

- Traditional: regression adjustment to be supplemented by...
- ... standardisation to obtain population effect (g-formula in time-varying context) – underused in practice
- or MSMs fitted by IPTW – easy to use, also with time-varying data – but can be inefficient
- propensity score methods (stratification / matching) – overused?
- Combination leads to *doubly robust* estimation procedures $\Rightarrow$ promising new methods use double-machine learning
- Always check positivity/overlap & balance with **all** methods!

Validation?

- Causal (counterfactual) conclusions from observational data cannot be validated on that same data!
  - Can check balance on *observed* confounders, but more important for *unobserved* factors
  - Need: experimental validation on different / new (ideally experimental) data $\longrightarrow$ some example tomorrow
- Recent studies: compare randomised trials with real-world (observational) studies

$\Rightarrow$ Often: evidence that much bias is due to inappropriate analysese, more than to lack of randomisation

Validation?

- Helpful: compare very different methods for estimating the same estimand

  $\Rightarrow$ if not in agreement some assumptions are violated
- Sometimes: different study designs can be used to check if same conclusion is obtained
  - Natural experiments, pragmatic trials ...
- **Negative controls:** similar exposure or similar outcome with same source of confounding but known zero-effect $\Rightarrow$ assess unobserved confounding
- Instrumental variables — topic of its own...

Quantitative Bias Analysis

- Can investigate:
  "How much would our conclusions change if there was an unobserved confounder with certain properties"
  $\Rightarrow$ Sensitivity / bias analysis (Lash et al, book)
- Formal approaches: based on Bayesian models (Greenland, Handbook Epidemiology chapter!)
- Ad-hoc method: **E-value** (Ding & Vanderweele)
  "How strongly must an unobserved confounder be associated with X and Y to explain away the causal effect (in the worst case)?"

Extensions / Outlook

- Multiple treatments (X has more than two levels), continuous treatments (positivity?)
- Different interventions: nudging / shift-interventions
- Different estimands: effect of treatment on the treated, (in)direct effects, 'principal-stratum' effect etc.
- Different outcomes: survival / time-to-event (censoring), multivariate outcomes
- Sequential / time-dependent treatments (dealing with 'switching', 'when-to-start?')
  $\longrightarrow$ time-dependent confounding!
- Effect heterogeneity, individualised / adaptive / optimal treatments $\longrightarrow$ (optimal) dynamic treatments