Causal Inference

Vanessa Didelez and Robin Evans
BIPS, University of Bremen (Germany), and University of Oxford (UK)

September 2024

APTS — Oxford

Part 3: Estimating a Causal Effect (basics)

Basic Setting

$A =$ binary **(point-)treatment**

$Y =$ some (numeric) outcome
(not survival / duration — that's special)

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

Keeping it simple to focus on essentials!

General 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.

(Total) Causal Effect

Will focus on: Average causal effect

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

or, with potential outcomes

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

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

Key Assumptions

Consistency

Consistency Assumption:

If we observe $A^i = a$ then $Y^i = Y^i(a)$ (for individual i)

i.e. the outcome we observe under the actual treatment is the potential outcome had the treatment been *set* to what it actually was.

Violated, e.g., if manipulation of A not well defined or so 'invasive' that observational setting not informative.

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

Often: if violated, need more elaborate model; e.g. intervene in physical activity to change BMI.

Consistency

Under consistency, and e.g. for binary A, can write

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

Note:

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

$\rightarrow$ invariance (at distributional level)

No-Interference

Common assumption: **no-interference**:

Vector $\mathbf{a} =$ treatment assignments for n units, then $Y^i(\mathbf{a}) = Y^i(a^i)$,

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

Violation: e.g. vaccines, social networks.

Stable unit–treatment value (SUTVA):

consistency + no-interference.

Conditional Exchangeability

Assumption of **cond. exchangeability** or **no unmeasured confounding** (or **random treatment assignment**, **ignorability** or ...):
Let X be (subset of) measured pre–treatment covariates, then

$$Y(a) \perp\!\!\!\perp A \mid X$$

Interpretation: within values of X, can consider A like randomised wrt outcome.

Denote: X is *sufficient* to adjust for confounding;

or 'valid adjustment set'.

Conditional Exchangeability

with do$(\cdot)$

Assumption of **cond. exchangeability** or **no unmeasured confounding** & **'consistency'** with do–notation:

$$p(y \mid x; \mathsf{do}(A = a)) = p(y \mid x, a)$$

Interpretation: within values of X, whether $A = a$ obtained by intervention or observation makes no difference wrt. distribution of Y.

Useful software for querying DAGs: `DAGitty` (Textor et al, 2016)

Pre-Treatment Covariates?

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

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

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

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

Graphically: X non-descendants of A.

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

Positivity Assumption

checking overlap

Often, methods for effect estimation require

Assumption of positivity:

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

Interpretation: for all (suff.) covariate values, it must be possible that a subject receives any value of treatment.

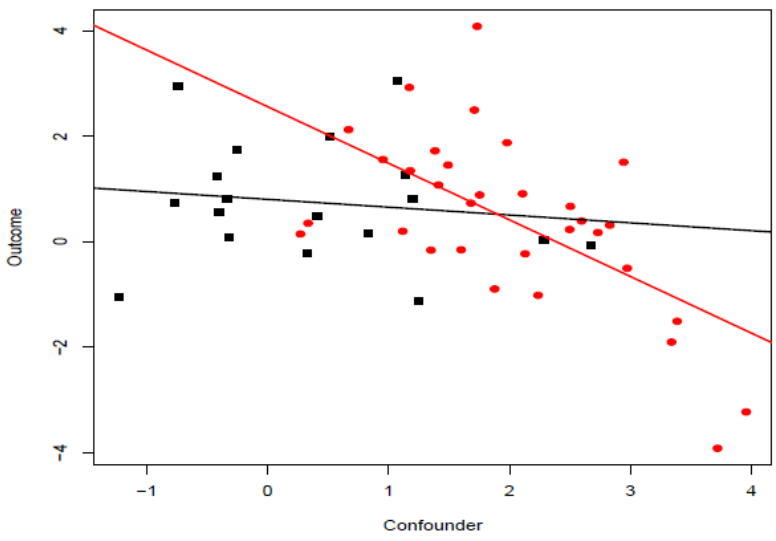
In practice: often empirically (nearly) violated — modifications and adaptations of methods exist sometimes.

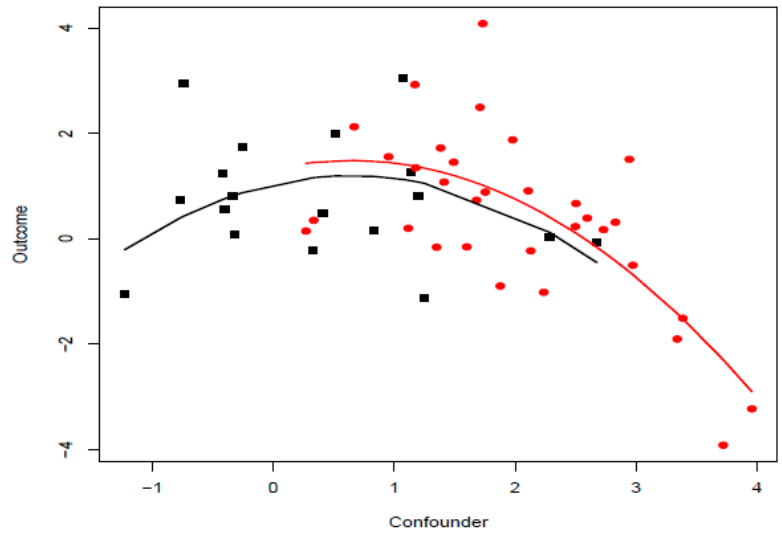
Positivity Assumption — Notes

- Target trial: can / should define eligibility criteria so that positivity is satisfied — requires domain knowledge
- (Lack of) positivity can be evaluated empirically (look at $p(a|x)$ or $p(x|a)$) — high-dim X becomes challenging Methods exist to characterise 'area of overlap'
- Do not include superfluous variables in X, especially: strong predictors of A that do not affect Y — can lead to apparent violations of positivity
- Regression-based approaches may mask lack of positivity as fitted models allow **extrapolation**.

Positivity Assumption

Extrapolation





Identifiability Revisited

We consider $p(y \,|\, \text{do}(A = a))$ or equivalently $p(Y(a))$.

With the above assumptions:

$$p(Y(a)) \overset{(i)}{=} \sum_x p(Y(a)|x)p(x) \overset{(ii)}{=} \sum_x p(Y(a)|a,x)p(x)$$

$$\ldots \overset{(iii)}{=} \sum_x p(y|a,x)p(x)$$

(i) probability calculus
(ii) valid adjustment set
(iii) causal consistency & positivity
(iv) no-interference was needed for well-definedness of causal effect

Checking Assumptions?

- Consistency / no-interference:
  domain knowledge, study design
- No-unmeasured-confounding: compare analysis of observational data with actual randomised trial — often not possible; triangulation, e.g. negative controls etc.
- Positivity
  - basic: boxplot of each variable in X by treatment group;
  - advanced: fit model for $\pi(x) = P(A = 1|X = x)$ — the **propensity score**
    obtain fitted values $\hat{\pi}^i = \hat{P}(A = 1|X = x^i)$ for each unit i and check for near zero/one in treated and control group, respectively.

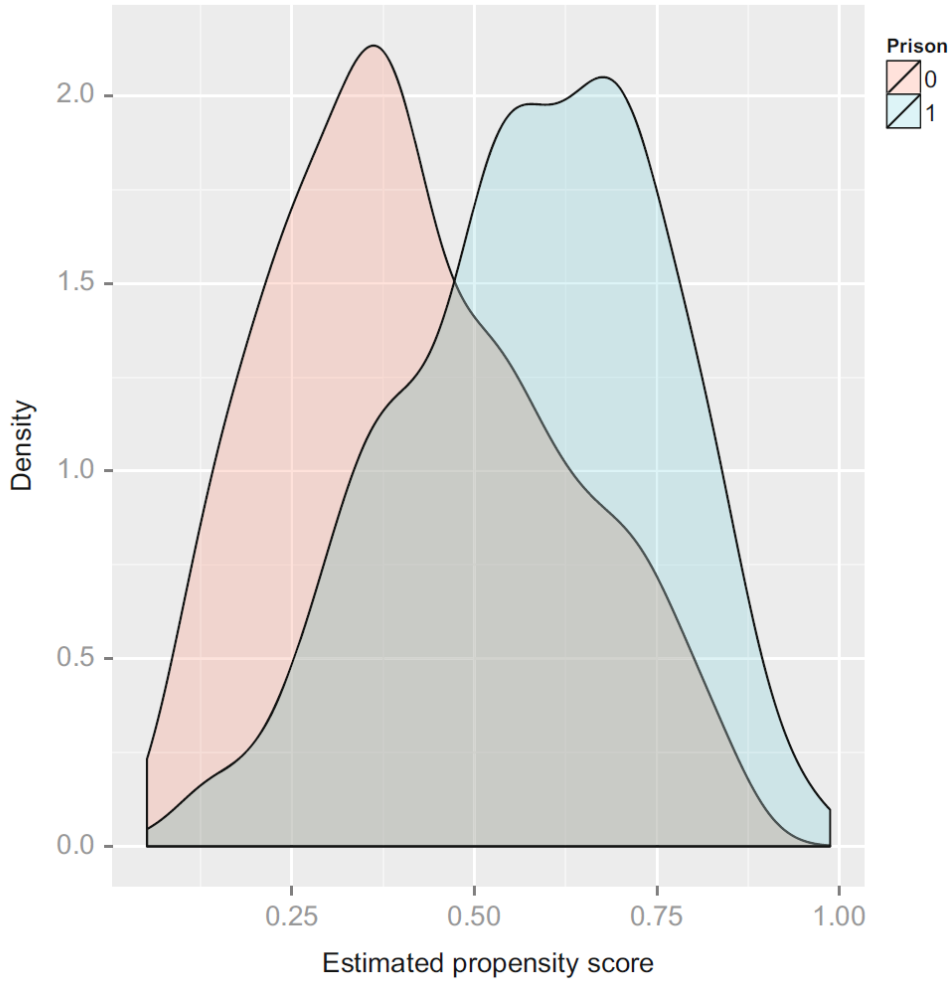
Propensity Score: Checking Positivity

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

$\Rightarrow$ reasonable overlap.

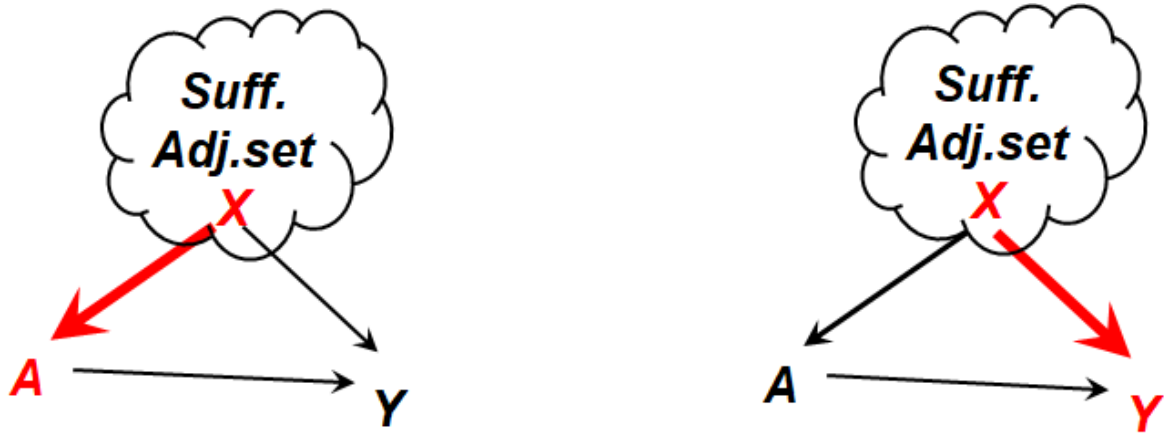
Example taken from Guo et al. (2016)

Covariate balance:
`R` package `cobalt`



Methods

Methods use either treatment model $p(a|x)$ (propensity score)
or outcome model $p(y|a,x)$ or both



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

Standardisation

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

Reminder: if X is sufficient set of covariates

$$E(Y \,|\, \text{do}(A = a)) = \sum_x E(Y|a, x)p(x)$$

An obvious way to use this is:

– fit flexible regression model for $E(Y|a, x; \beta)$ to data
to avoid 'g-null paradox'! (Evans & Didelez, 2024:JRSSB)

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

– e.g. with `R` package `stdReg` (Sjølander and Dahlqwist, 2017)

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

Using Standardisation

Why not just look at regression model $E(Y|A, X)$?

- Consider: marginal versus conditional causal effect and collapsible versus non-collapsible parameters;
- Logistic regression / odds ratios not collapsible.
- If set of sufficient covariates X not unique, cond. effects may depend on choice of X (e.g. COR), but not marginal ones.
- Marginal $E(Y \,|\, \text{do}(A = a))$ corresponds to randomised trial where covariates X can be /are ignored.

Note: Consistency of effect estimation relies on correctly specified model for $p(y|a, x)$.

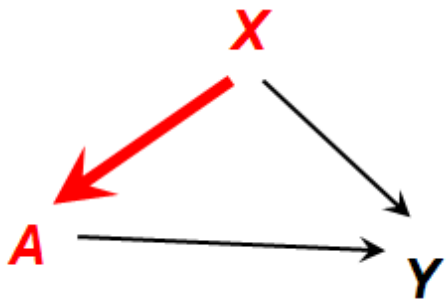
Using Standardisation

- R package `stdReg` (Sjølander and Dahlqwist, 2017)
- The method is special case of **G–formula** for sequential treatments (Robins, 1986).
- Population effect $E(Y \mid \text{do}(A = a))$ depends on *distribution* $p(x)$ of covariates in target population

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

Treatment Modelling Approaches

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



From Standardisation to MSM

The functionals

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

might be 'awkward', especially if A continuous,
$E(Y|a,x)$ non–linear with interactions,
or X high dimensional and/or partly continuous.

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

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

Marginal Structural Models

(Hernán et al, 2001)

MSM: *semiparametric* model for

$p(y \,|\, \text{do}(A = a))$ or more typically $E(Y \,|\, \text{do}(A = a))$

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

Marginal: refers to time-varying covariates $\rightarrow$ later

Structural: model under intervention in A (not observational)

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

Inverse Probability Weighting

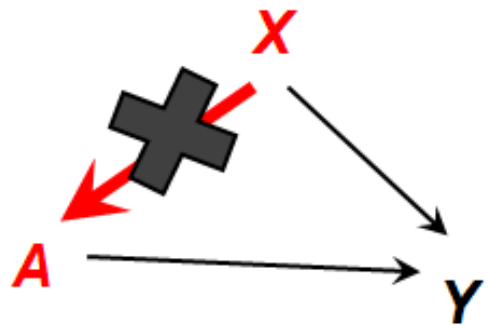
(IPTW)

Idea is based on:

$$p(y \mid \mathsf{do}(A = a))$$

$$= \sum_x p(y|a, x)p(x)$$

$$= \sum_x \frac{p(y,a,x)}{p(a|x)}$$



Inverse Probability Weighting

(IPTW)

Idea is based on:

$$p(y \,|\, \text{do}(A = a)) = \sum_x p(y|a, x)p(x) = \sum_x \frac{p(y, a, x)}{p(a|x)}$$

$\Rightarrow$ fit model for $E(Y \,|\, \text{do}(A = a))$ with weights $w^i = p(a^i|x^i)^{-1}$

$\Rightarrow$ creates 'pseudo sample' where X is balanced

$\Rightarrow$ unbiased estimating equations for parameters of $E(Y \,|\, \text{do}(A = a))$.

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

IPTW Estimator

Basic Idea

Simple situation: binary exposure A;

define $\pi(x) = P(A = 1|X = x)$.

Can show (under our assumptions):

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

and similarly

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

Proof: iterated conditional expectation (exercise!)

$\longrightarrow$ see 'Horvitz–Thompson' principle

IPTW Estimator

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

IPTW yields consistent estimator for ACE:

– if $\pi(X;\alpha)$ correctly specified;

– can obtain sandwich standard errors or bootstrap, or theoretical asymptotical standard errors.

IPTW — Implementation

Easy to implement with standard software for regression models by specifying weights:
first, obtain weights, then fit chosen model.

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

Note: default standard errors ignore variability in (estimated!) weights $\Rightarrow$ sandwich st.error (or: bootstrap)

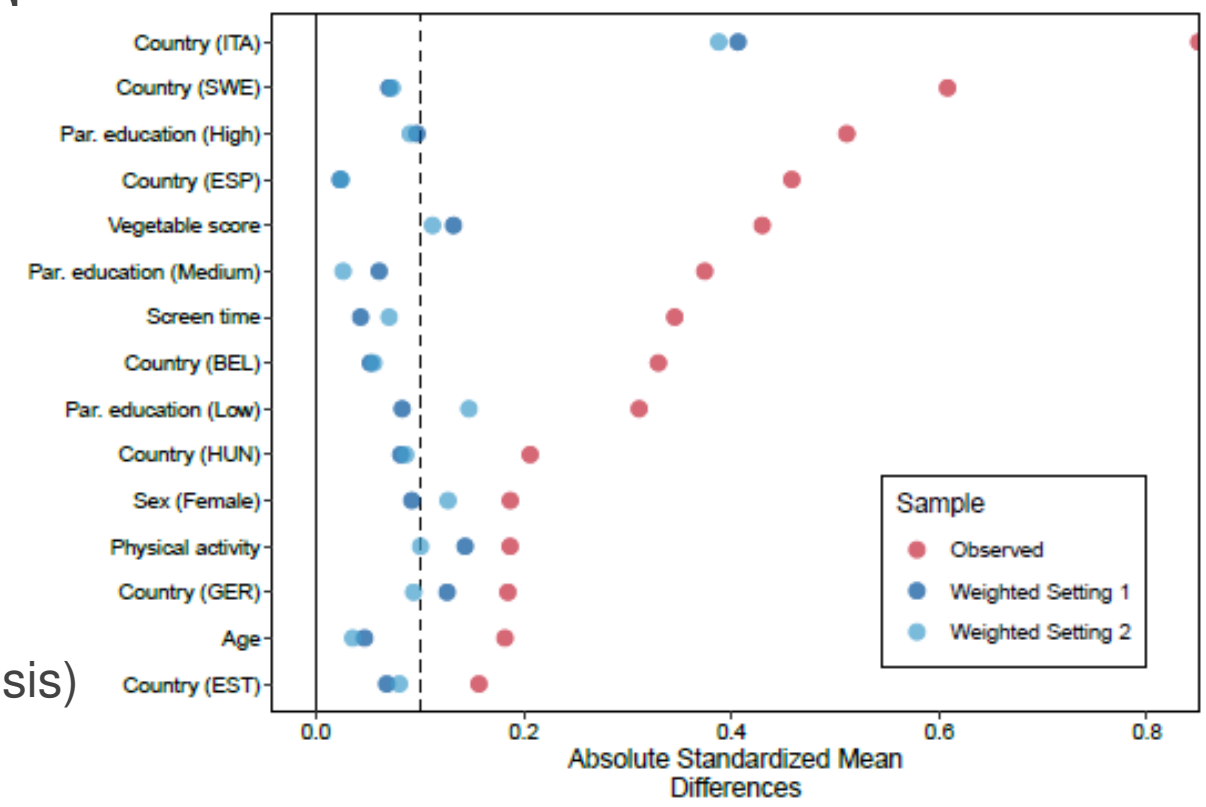
Notes on IPTW

- Consistent when both, models for $E(Y \mid \text{do}(A = a))$ and $\pi(x)$ correctly specified.
- If $p(a|x) \approx 0 \Rightarrow$ large weights $\Rightarrow$ use 'stabilised' weights, e.g. $w = \tilde{p}(a)/p(a|x)$, where $\tilde{p}(a)$ some distribution for A $\Rightarrow$ more efficient estimators.
- Check assumptions: in weighted population, observed covariates must be 'balanced' — e.g. package `cobalt` for balance plots.

Checking Assumptions: Balance

Example: causal effect of ‘fibre intake’ on children’s BMI; large adjustment set (country, parental edu, vege-score, etc.)

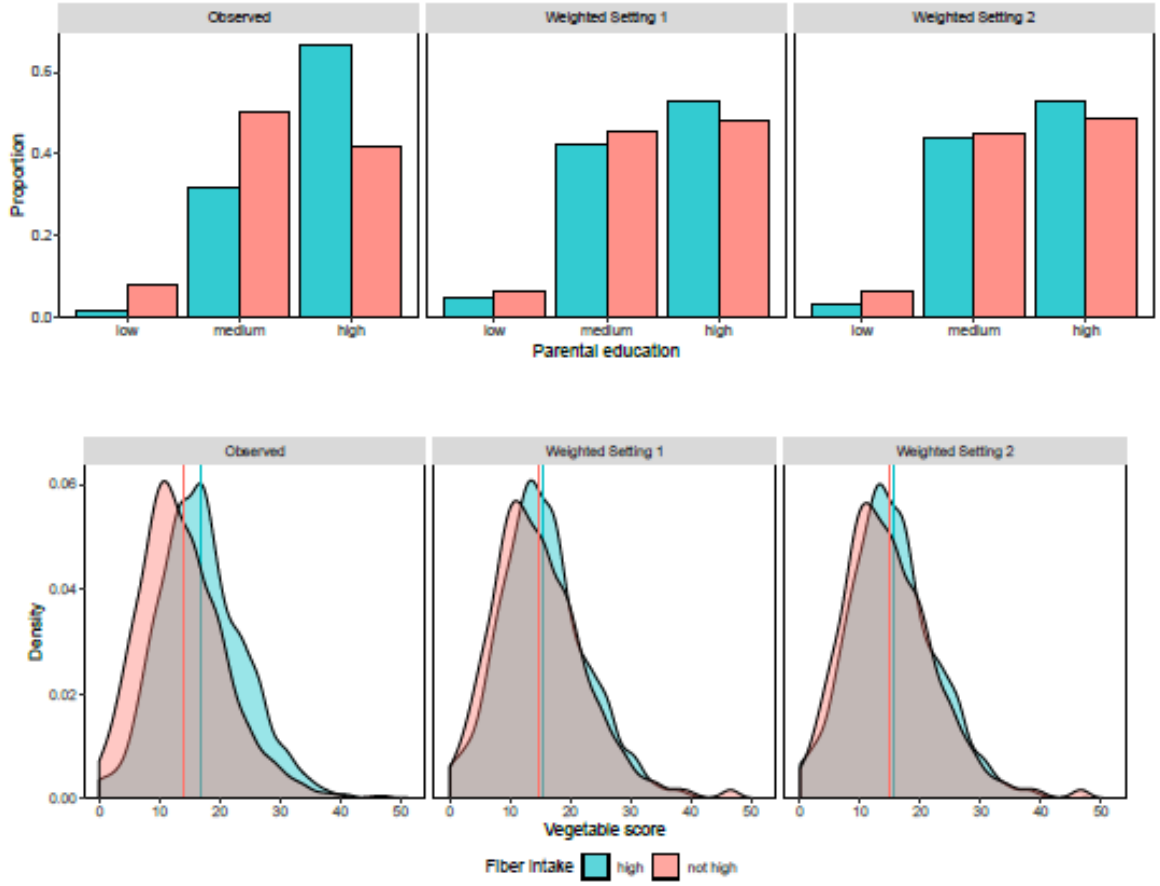
Plot: std.mean-diff (abs)
treated / untreated
red: unweighted
blue: weighted



(Nöhren, 2021:MAthesis)

Checking Assumptions: Balance

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



Notes on IPTW

- MSMs with IPTW mostly used in longitudinal situations / sequential treatments with time-varying confounding $\Rightarrow$ 'marginal' over time-dependent confounders / covariates.
- IPTW especially useful when study design (or other) supplies background knowledge to model weights $p(a|x)$.
- Software: `R` package `ipw` for longitudinal data and correct standard errors
- Problem: estimation of weights $p(a|x)^{-1}$ not obvious, but possible, when A continuous.

Continuous Treatments?

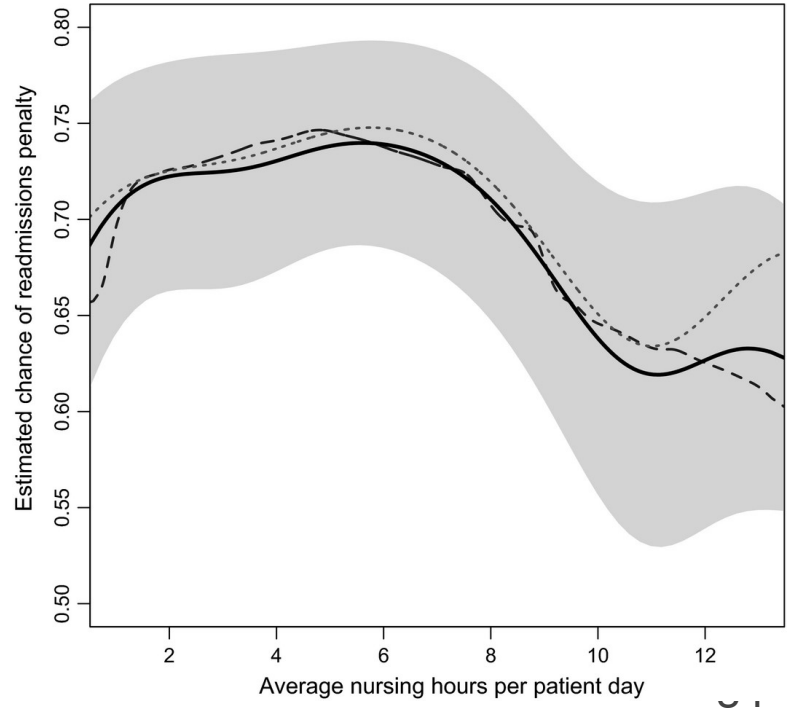
Wanted: $E(Y(a))$ or $E(Y \mid \text{do}(a))$ as a function of a
– assume a (semi-)parametric model;
– but non-parametric methods exist

(e.g. Kennedy et al, 2017:JRSSB)

Issues:
– positivity?
– meaningful interventions?

Example: in hospitals,
$A =$ nurse staffing hours per day
$Y =$ excess re-admission

dashed = regr.std
dotted = IPTW
full = “double-robust”



Propensity Score (PS)

(Rosenbaum & Rubin, 1983)

Have used $\pi(x) = P(A = 1|X = x)$

$\Rightarrow$ **propensity score**.

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

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

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

Propensity / Balancing Score

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

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

$$A \perp\!\!\!\perp X \mid \pi \qquad \text{i.e. } \pi \text{ balances } X$$

Hence (with properties of X):

$$Y(a) \perp\!\!\!\perp A \mid \pi$$

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

Propensity Score — Graphically

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

Left: assumption of X being sufficient set of covariates.

Right: π is deterministic function of X and $A \perp\!\!\!\perp X \mid \pi$.

Propensity Score in Practice

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

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

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

Propensity Score in Practice

Methods for using PS (other than weighting):

- PS **stratification**: divide into strata (often quintiles) and fit $p(y|a)$ 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 ETT!

- Sometimes: **PS adjustment** — specify model for $p(y|a, \pi)$ 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 $\pi(X;\alpha)$ 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 $E(Y|X=x;\mathrm{do}(A=a))$ within strata.
- PS popular especially for matching: π is 'one-dimensional reduction' of covariates — but at cost of first modelling / estimating $\pi = p(a|x;\alpha)$.
- PS matching / stratification not really suitable for sequential treatments.

Notes on Propensity Score

- Danger: modelling $\pi(X;\alpha)$ may focus on strong predictors of $A \Rightarrow$ can amplify bias! $\Rightarrow$ selection of X as adjustment set should be separate process from fitting $\pi(X;\alpha)$.
- Interpretation of PS sometimes regarded as difficult compared to actual covariate values.
- Simulations suggest that IPTW use of π superior to stratification. (Lunceford & Davidian, 2004)
- Critique of PS matching: King & Nielsen (2019)

Estimating Causal Effects

Summary (no unobs. conf.)

Given sufficient obs. confounders X (& positivity):

- Traditional: regression adjustment or...
- ... standardise 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 of above approaches leads to *doubly robust* estimation procedures
- Should always check positivity / overlap!

Thank You!

www.leibniz-bips.de/en

Contact
Vanessa Didelez
Leibniz Institute for Prevention Research
and Epidemiology – BIPS
Achterstraße 30
D-28359 Bremen
didelez@leibniz-bips.de