Inverse Probability of Treatment Weighting for Control of Counfounding

RYAN COOK

UCLA FIELDING SCHOOL OF PUBLIC HEALTH, DEPARTMENT OF EPIDEMIOLOGY

4/30/19

Overview

- What is IPTW?
- Why do I want to use it?
- How does it work?
- How do I do it?

Notation

Keeping it as simple as possible:

X = treatment/exposure

Y = outcome

C = set of confounder variables C1, C2, C3…Cn

- At least 1, but possibly many more

Definition of confounder

C is a confounder if…

- It is causally related to X and (at least) correlated with Y

OR

- It is causally related to Y and correlated with X

AND

- It is not part of the causal pathway between X and Y (it is not a mediator)

What is IPTW?

Weighting-based method of confounder control: Transforms study sample into one in which confounders are balanced across treatment/exposure groups

Two-stage analysis

- Association between confounders and treatment/exposure
- Association between treatment and outcome

“Marginal” outcome model

- E(Y|X) instead of E(Y|X,C)

What IPTW is NOT

IPTW is NOT more valid than multiple regression

IPTW is NOT a solution to missing confounder data

IPTW is NOT more efficient than multiple regression (usually)

Where is IPTW useful?

High number of covariates AND

- Low number of events/stratum
- Bad distribution of Y

If you are more comfortable estimating the relationship between C and X than C and Y, consider IPTW

If you want your paper to be more likely to be accepted…consider IPTW

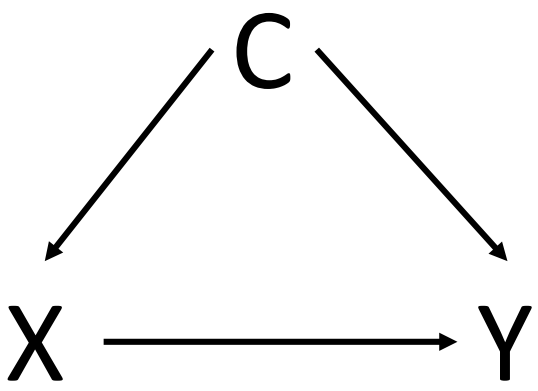
How does IPTW work?

In an IPTW population, treatment/exposure groups are balanced on covariates used to create the weights

- Importantly, they are NOT balanced on variables NOT used in the IPTW estimation (this is not an RCT)

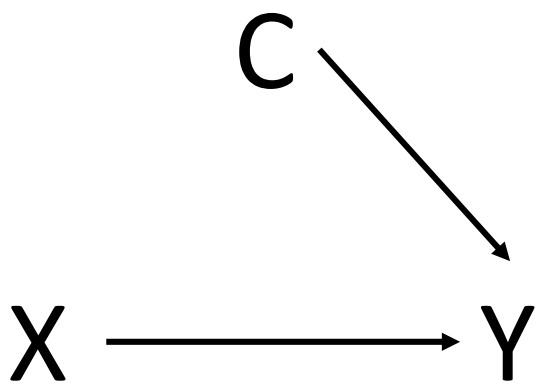
If these variables are balanced, they are independent from treatment, and cannot confound the X-Y relationship

Observed data

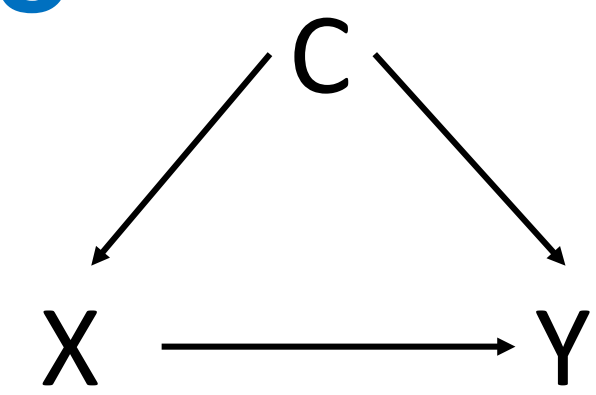


IPTW

What we want



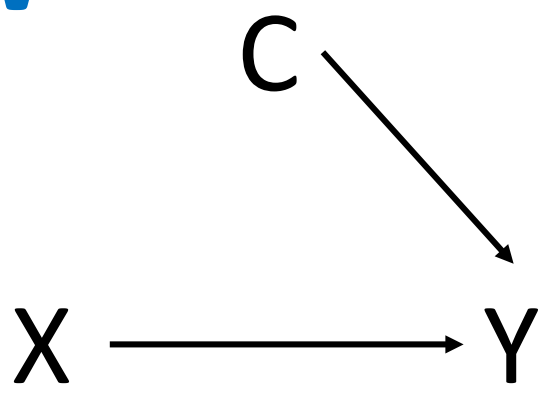
O



O = f(y,x,c)

= f(y|x,c)f(x|c)f(c)

W



W = f(y,x,c)

= f(y|x,c)f(x)f(c)

X and C are statistically independent if
f(x|c) = f(x)

$$W = O * \frac{W}{O}$$

$$W = f(y|x,c)f(x|c)f(c) * \frac{f(y|x,c)f(x)f(c)}{f(y|x,c)f(x|c)f(c)}$$

$$W = \underbrace{f(y|x,c)f(x|c)f(c)}_{\text{Observation}} * \underbrace{\frac{f(x)}{f(x|c)}}_{\text{Weight}}$$

Definition of IPTW

$$IPTW = \frac{P(X=x)}{P(X=x|C)}$$

With a binary treatment or exposure,

$$\frac{P(X=1)}{P(X=1|C)} \text{ if treated} \qquad \frac{P(X=0)}{P(X=0|C)} \text{ if control}$$

How to obtain $P(X = 1)$

Frequency table(x)

Treated	Not treated
N, %	N, %

*Will be constant across observations

How to obtain $P(X = 1 \mid C)$

1: Logistic Regression(x = c)

2: Output predicted values on probability scale, or output on log odds scale and transform with expit function $\frac{exp(\log \text{odds})}{(exp(\log \text{odds})+1)}$

*Will be different for observations with different values of C

*This is a propensity score

How to obtain numerator and denominator

IPTW = conditional variable:

- If treated:
  - Numerator = P(X = 1)
  - Denominator = P(X = 1|C)

- If not treated:
  - Numerator = 1 - P(X = 1)
  - Denominator = 1 – P(X = 1|C)

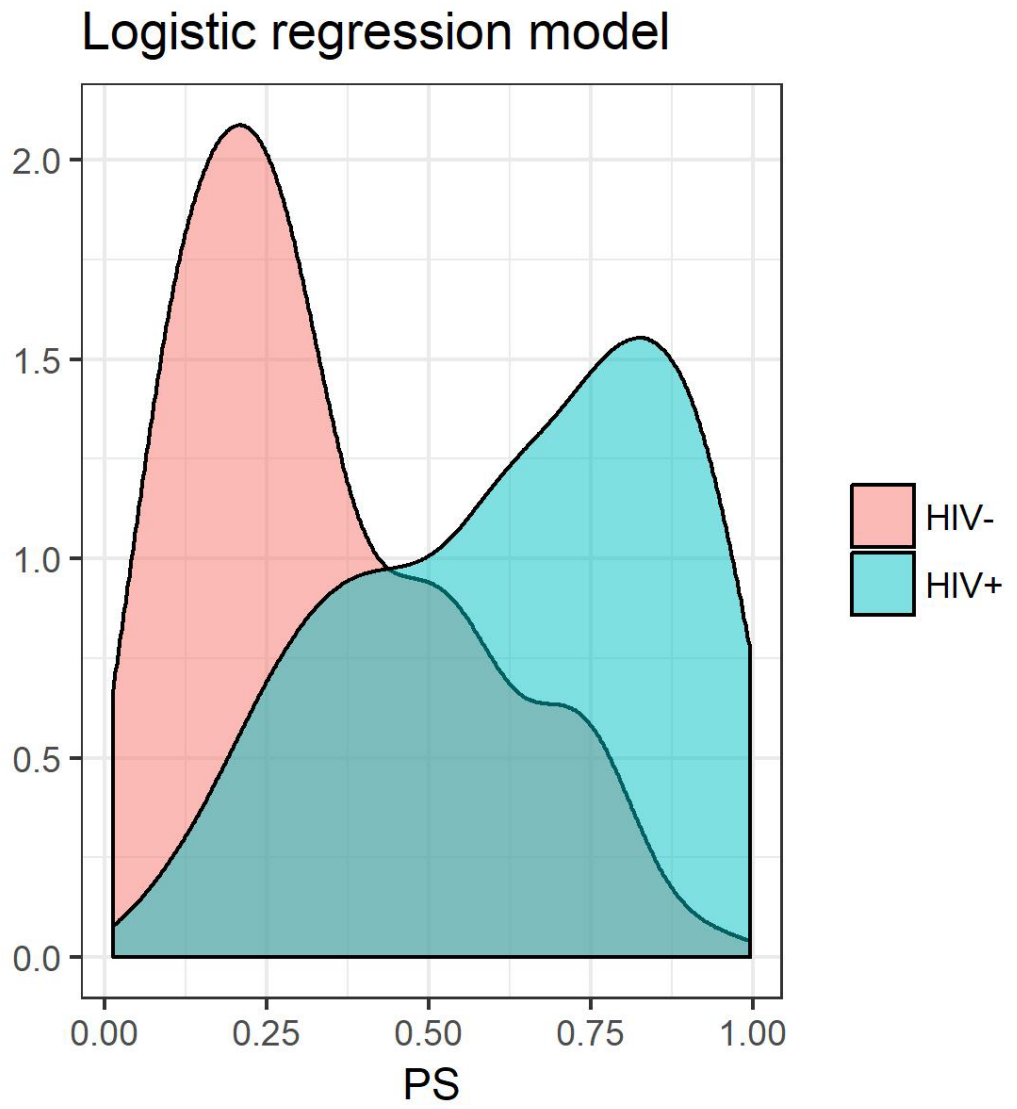
How to evaluate quality of weights

1. ALWAYS run a summary of the weight variable
   - The mean of the weights should be close to 1
   - Consider trimming observations with very large or very small weights

Evaluating weights

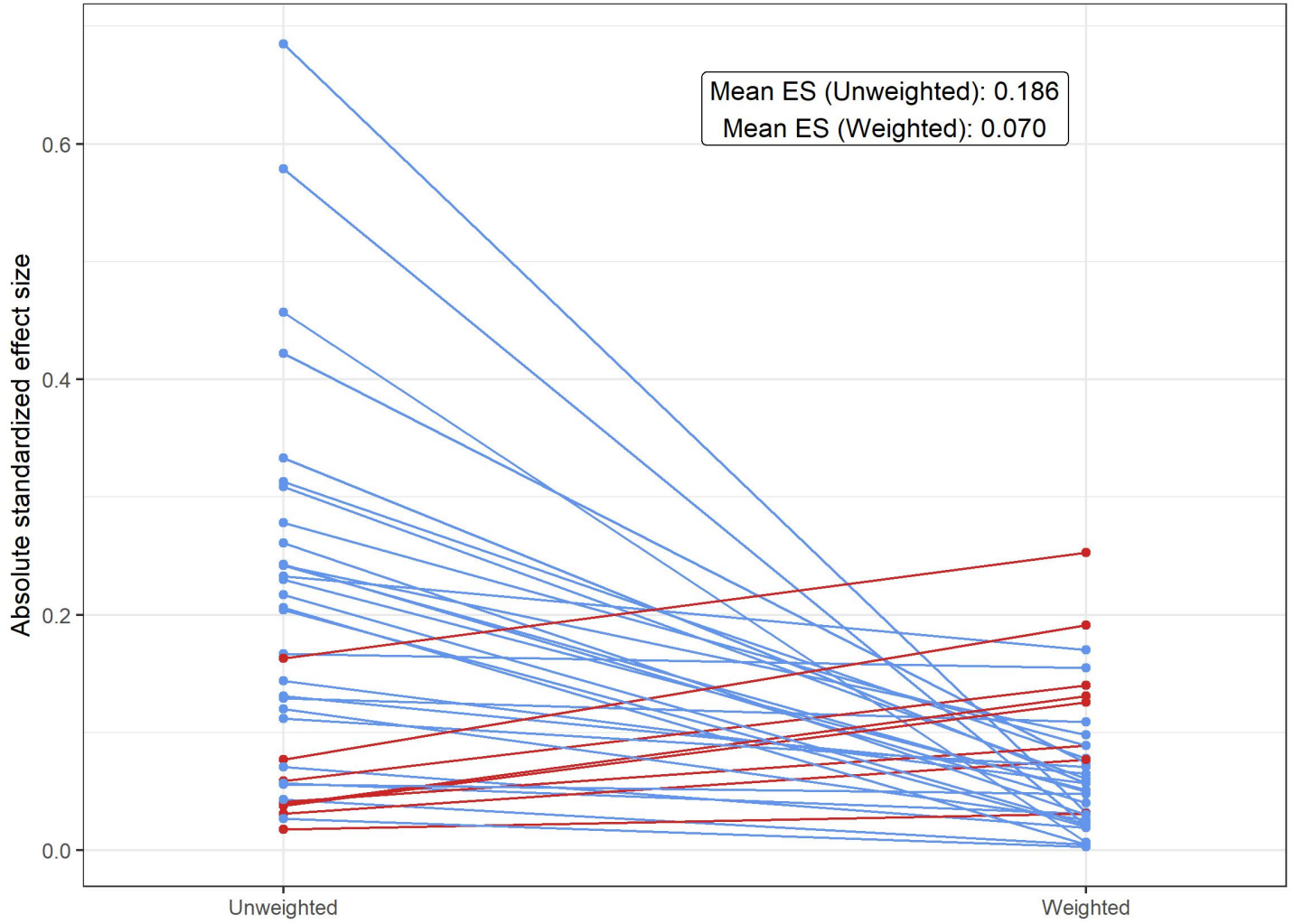
Look at a histogram of P(X = 1|C)*, split by treatment group:

*fitted values from the logistic regression



Evaluating weights

Compute a measure of balance (e.g. the standardized mean difference) in each confounder variable before and after weighting:

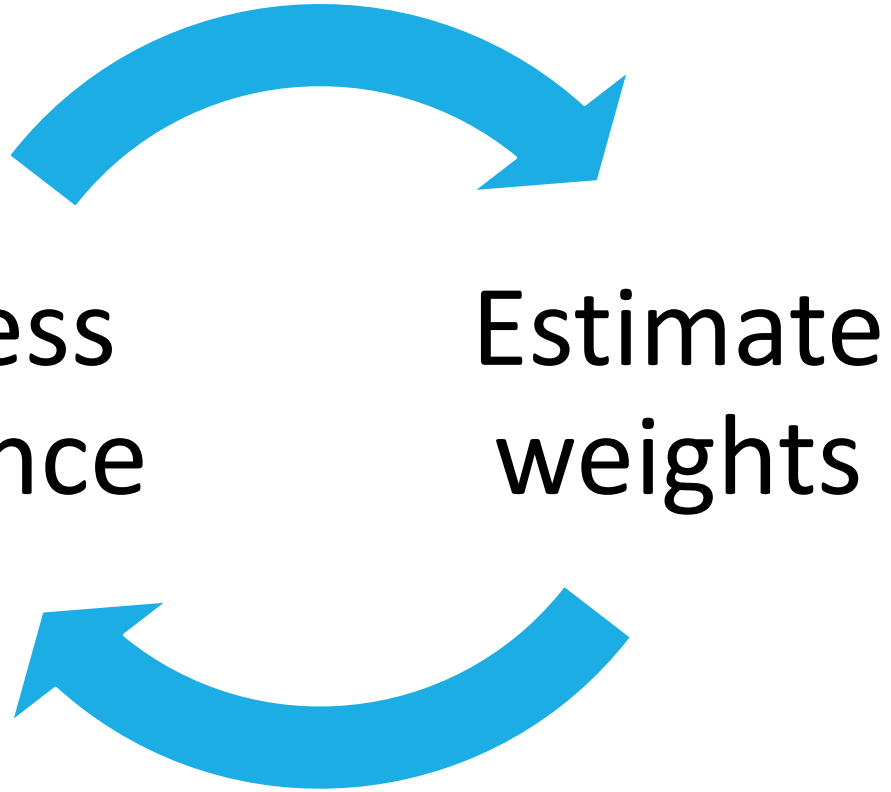


Evaluating weights

Compare “Table 1” *p* values before and after weighting

*Looking at effect sizes is better, but this can be informative

IPTW estimation and evaluation is an iterative process



Fitting outcome model

Regression(y = x, weights = IPTW)

BUT

You must correct the standard error!

- Use a “survey” analysis procedure
- Use a tool for robust SE estimation (e.g. the sandwich estimator)
- Bootstrap: best option because it accounts for the fact that the IPTW themselves are not known and add additional uncertainty to the analysis

Example – smoking and lung cancer

smk	lc	c1	c2	c3
0	0	0	0	0
0	0	0	0	0
0	0	0	0	1
1	1	0	0	0
0	0	0	0	0
1	1	0	0	0

True $OR_{smk} = 2.7$

Biased $OR_{smk} = 3.6$

Step 1: Obtain P(X = 1)

```
prop.table(table(df$smk))

     0      1
0.4249 0.5751
```

smk	lc	c1	c2	c3	px
0	0	0	0	0	0.5751
0	0	0	0	0	0.5751
0	0	0	0	1	0.5751
1	1	0	0	0	0.5751
0	0	0	0	0	0.5751
1	1	0	0	0	0.5751

Step 2: Obtain P(X = 1|C)

```
pxc <- fitted(glm(smk ~ c1 + c2 + c3, family = "binomial"))
```

smk	lc	c1	c2	c3	px	pxc
0	0	0	0	0	0.5751	0.533
0	0	0	0	0	0.5751	0.533
0	0	0	0	1	0.5751	**0.138**
1	1	0	0	0	0.5751	0.533
0	0	0	0	0	0.5751	0.533
1	1	1	0	0	0.5751	**0.741**

Step 3: Create IPTW

```
df$iptw <- ifelse(df$smk == 1, df$px/df$pxc, (1-df$px)/(1-df$pxc))
```

smk	lc	c1	c2	c3	px	pxc	iptw
0	0	0	0	0	0.5751	0.533	0.909
0	0	0	0	0	0.5751	0.533	0.909
0	0	0	0	1	0.5751	0.138	0.493
1	1	0	0	0	0.5751	0.533	1.08
0	0	0	0	0	0.5751	0.533	0.909
1	1	1	0	0	0.5751	0.741	0.776

Step 4: Diagnostics

```
summary(df$iptw)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.4929  0.7761  0.9099  0.9991  1.0790  4.1674
```

Step 5: Outcome model

```
svydes <- svydesign(id ~ 1, weights = ~iptw, data = df)
mod <- svyglm(lc ~ smk, design = svydes, family = "binomial")
summary(mod)
```

```
Coefficients:
             Estimate   Std. Error t value Pr(>|t|)
(Intercept)  0.24452    0.03289    7.434   1.14e-13 ***
smk          0.99069    0.04942   20.047   < 2e-16 ***
```

$OR_{smk} = 2.7$

IPTW vs. Multiple regression

IPTW

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.24452	0.03289	7.434	1.14e-13	***
smk	**0.99069**	**0.04942**	**20.047**	**< 2e-16**	*******

Multiple Regression

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.03924	0.03665	1.071	0.284	
smk	**1.03826**	**0.04753**	**21.846**	**<2e-16**	*******
c1	0.48533	0.05358	9.059	<2e-16	***
c2	0.93259	0.06647	14.030	<2e-16	***
c3	-0.93734	0.07320	-12.805	<2e-16	***

R code

1. px <- prop.table(table(df$x))[2]
2. pxc <- fitted(glm(x ~ c, data = df, family = “binomial”))
3. df$IPTW <- ifelse(df$x == 1, px/pxc, (1-px)/(1-pxc))
4. summary(df$IPTW)
5. svydes <- svydesign(id ~ 1, weights = ~IPTW, data = df)
6. mod <- svyglm(y ~ x, design = svydes, family = “gaussian”)

SAS “code”

1. Proc freq; table x;
   - Output the percent treated to a new dataset, remember it, or save it as a macro variable.
2. In data step, create variable px = percent treated (same for everybody)
3. Proc logistic; model x = c; output pred = pxc;
4. Merge pxc into dataset
5. In data step, create variable
   - if x = 1 then IPTW = px/pxc
   - else if x = 0 then IPTW = (1-px)/(1-pxc)
6. Proc summary on IPTW
7. Proc genmod; class x; model y = x; weight = IPTW;

Details and extensions

The numerator, P(X = x) is actually not necessary for bias reduction

IPTW may be estimated as $\frac{1}{P(X=x|C)}$, known as the "unstabilized" version

These are less efficient and I see no reason to exclude the numerator

Extensions

Formula presented estimates the ATE, the “average treatment effect”

- What reduction in MI rates would we expect to see if we gave everyone in our population statins as opposed to giving everyone in our population placebo

ATT = average treatment effect among the treated

- Among those receiving statins, what reduction in MI rates are actually attributable to statins?
- If treated, IPTW = 1
- In untreated, $\text{IPTW} = \frac{P(X=1\,|C)}{P(X=0\,|C)} * \frac{P(X=0)}{P(X=1)}$

ATU = average treatment effect among the untreated

Extensions

Treatment with >2 categories

- $IPTW = \frac{P(X=x)}{P(X=x|C)}$ for x ϵ {0,1, 2, …. n}

- Use multinomial logistic regression to estimate P(X=x|C)

Extensions

Multiple treatments X and Z (interaction analysis)

$$IPTW = \frac{P(X=x)}{P(X=x|C_1)} * \frac{P(Z=z)}{P(Z = z|C_2)} = \frac{P(X=x)P(Z=z)}{P(X = x|C_1)P(Z=z|C_2)}$$

C1 and C2 can be the same set or different

More

Epi 204, 211, 212 – Onyebuchi Arah