

SISCER Module 13

Lecture 4: g-computation, IPW and AIPW

Ting Ye & Qingyuan Zhao

University of Washington & University of Cambridge

July 2026

Plan

- ▶ g-computation
- ▶ Inverse probability weighting
- ▶ AIPW
- ▶ Propensity score

Motivation: smoking and lung cancer

Motivation: smoking and lung cancer

- ▶ In Lecture 2, we discussed studying the causal effect of smoking on lung cancer using observational data.

Motivation: smoking and lung cancer

- ▶ In Lecture 2, we discussed studying the causal effect of smoking on lung cancer using observational data.
- ▶ The observational data contain rich covariate information: age, race, nativity, rural versus urban residence, occupational exposures to dust and fumes, religion, education, marital status, ...

Motivation: smoking and lung cancer

- ▶ In Lecture 2, we discussed studying the causal effect of smoking on lung cancer using observational data.
- ▶ The observational data contain rich covariate information: age, race, nativity, rural versus urban residence, occupational exposures to dust and fumes, religion, education, marital status, ...
- ▶ These variables are called **confounders** if they affect both treatment assignment and the potential outcomes.

Motivation: smoking and lung cancer

- ▶ In Lecture 2, we discussed studying the causal effect of smoking on lung cancer using observational data.
- ▶ The observational data contain rich covariate information: age, race, nativity, rural versus urban residence, occupational exposures to dust and fumes, religion, education, marital status, ...
- ▶ These variables are called **confounders** if they affect both treatment assignment and the potential outcomes.
- ▶ Again, recall that we are interested in inferring

$$\text{ATE} := E[Y(1) - Y(0)].$$

No unmeasured confounding

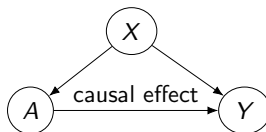
No unmeasured confounding

Suppose all confounders are measured, which we denote as X . In other words, *people who look comparable are comparable*.

No unmeasured confounding

Suppose all confounders are measured, which we denote as X . In other words, *people who look comparable are comparable*.

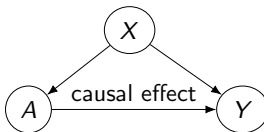
► This is called **NUCA** (no unmeasured confounding assumption) (or ‘conditional exchangeability’).



No unmeasured confounding

Suppose all confounders are measured, which we denote as X . In other words, *people who look comparable are comparable*.

► This is called **NUCA** (no unmeasured confounding assumption) (or ‘conditional exchangeability’).



► **NUCA** requires the treatment is as good as randomly assigned in each stratum of X :

$$A \perp\!\!\!\perp Y(0), Y(1) \mid X.$$

👉 “natural experiment”

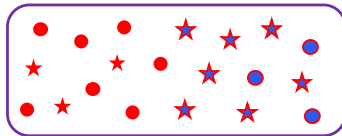
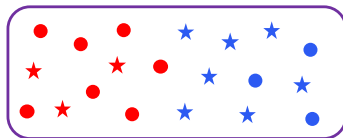
👉 This assumption cannot be falsified with data.

Positivity

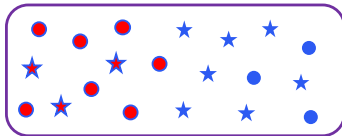
Recall that

$$\text{ATE} = \underbrace{\mathbb{E} Y(1)}_{\text{everyone receives treatment}} - \underbrace{\mathbb{E} Y(0)}_{\text{everyone receives control}}.$$

OVERALL POPULATION



TREAT ALL IN OVERALL
POPULATION?



TREAT NONE IN OVERALL
POPULATION?

Positivity

Recall that

$$\text{ATE} = \underbrace{\mathbb{E} Y(1)}_{\text{everyone receives treatment}} - \underbrace{\mathbb{E} Y(0)}_{\text{everyone receives control}}.$$

👉 Imagine there is a stratum of $X = x$ where everyone only receives treatment $P(A = 1 \mid X = x) = 1$, then we cannot know their outcomes under control; similarly, if $P(A = 0 \mid X = x) = 1$, then we cannot know their outcomes under treatment.

Positivity

Recall that

$$\text{ATE} = \underbrace{\mathbb{E} Y(1)}_{\text{everyone receives treatment}} - \underbrace{\mathbb{E} Y(0)}_{\text{everyone receives control}}.$$

👉 Imagine there is a stratum of $X = x$ where everyone only receives treatment $P(A = 1 \mid X = x) = 1$, then we cannot know their outcomes under control; similarly, if $P(A = 0 \mid X = x) = 1$, then we cannot know their outcomes under treatment.

► Positivity

$$\boxed{\forall x : P(X = x) > 0 \implies 0 < P(A = 1 \mid X) < 1.}$$

Identification

Identification

- ▶ ATE can be identified from observational data under both **NUCA** and **positivity**.

Identification

- ▶ ATE can be identified from observational data under both **NUCA** and **positivity**.
- ▶ We focus on three most important identification formulae
 1. g-computation / standardization
 2. inverse probability weighting (IPW)
 3. augmented inverse probability weighting (AIPW)

Identification

- ▶ ATE can be identified from observational data under both **NUCA** and **positivity**.
- ▶ We focus on three most important identification formulae
 1. g-computation / standardization
 2. inverse probability weighting (IPW)
 3. augmented inverse probability weighting (AIPW)
- ▶ Now that we are in observational studies, in general, we need **correct models** (unlike in RCTs).

g-computation

Method 1: g-computation

👉 Also known as: g-formula, standardization, outcome regression.

Method 1: g-computation

👉 Also known as: g-formula, standardization, outcome regression.

1. Within stratum x , treatment is randomly assigned by **NUCA**, so the average treatment effect within stratum x is

$$\text{ATE}(x) := \mathbb{E}[Y(1) - Y(0) | X = x] = \mathbb{E}[Y(1) | X = x] - \mathbb{E}[Y(0) | X = x]$$

Method 1: g-computation

👉 Also known as: g-formula, standardization, outcome regression.

1. Within stratum x , treatment is randomly assigned by **NUCA**, so the average treatment effect within stratum x is

$$\begin{aligned} \text{ATE}(x) &:= \mathbb{E}[Y(1) - Y(0) | X = x] = \mathbb{E}[Y(1) | X = x] - \mathbb{E}[Y(0) | X = x] \\ &= \underbrace{\mathbb{E}[Y | A = 1, X = x]}_{\mu_1(x)} - \underbrace{\mathbb{E}[Y | A = 0, X = x]}_{\mu_0(x)}. \end{aligned}$$

association = causation

Method 1: g-computation

👉 Also known as: g-formula, standardization, outcome regression.

1. Within stratum x , treatment is randomly assigned by **NUCA**, so the average treatment effect within stratum x is

$$\begin{aligned} \text{ATE}(x) &:= \mathbb{E}[Y(1) - Y(0) | X = x] = \mathbb{E}[Y(1) | X = x] - \mathbb{E}[Y(0) | X = x] \\ &= \underbrace{\mathbb{E}[Y | A = 1, X = x]}_{\mu_1(x)} - \underbrace{\mathbb{E}[Y | A = 0, X = x]}_{\mu_0(x)}. \end{aligned}$$

association = causation

2. Then we average over strata

$$\text{ATE} = \sum_x \text{ATE}(x)P(X = x) = \sum_x (\mu_1(x) - \mu_0(x))P(X = x).$$

Method 1: g-computation

Formal derivation:

$$\begin{aligned}\mathbb{E}[Y(a)] &= \sum_x \mathbb{E}[Y(a) \mid X = x]P(X = x) && \text{(law of total expectation)} \\ &= \sum_x \mathbb{E}[Y(a) \mid X = x, A = a]P(X = x) && \text{(no unmeasured confounding)} \\ &= \sum_x \mathbb{E}[Y \mid X = x, A = a]P(X = x) && \text{(consistency)}\end{aligned}$$

Method 1: g-computation

$$ATE = \sum_x ATE(x)P(X = x) = \sum_x (\mu_1(x) - \mu_0(x))P(X = x).$$

Method 1: g-computation

$$\text{ATE} = \sum_x \text{ATE}(x)P(X = x) = \sum_x (\mu_1(x) - \mu_0(x))P(X = x).$$

👉 When X is continuous, $\text{ATE} = \int (\mu_1(x) - \mu_0(x))p(x) dx$.

Method 1: g-computation

$$\text{ATE} = \sum_x \text{ATE}(x)P(X = x) = \sum_x (\mu_1(x) - \mu_0(x))P(X = x).$$

- When X is continuous, $\text{ATE} = \int (\mu_1(x) - \mu_0(x))p(x) dx$.
- Regardless of the space of X , can also be written as

$$\text{ATE} = \mathbb{E}[\mu_1(X) - \mu_0(X)] = \mathbb{E}[\mathbb{E}[Y \mid A = 1, X]] - \mathbb{E}[\mathbb{E}[Y \mid A = 0, X]].$$

Method 1: g-computation

$$\text{ATE} = \sum_x \text{ATE}(x)P(X = x) = \sum_x (\mu_1(x) - \mu_0(x))P(X = x).$$

👉 When X is continuous, $\text{ATE} = \int (\mu_1(x) - \mu_0(x))p(x) dx$.

👉 Regardless of the space of X , can also be written as

$$\text{ATE} = \mathbb{E}[\mu_1(X) - \mu_0(X)] = \mathbb{E}[\mathbb{E}[Y | A = 1, X]] - \mathbb{E}[\mathbb{E}[Y | A = 0, X]].$$

👉 Compare this with the naive contrast

$$\mathbb{E}[Y | A = 1] - \mathbb{E}[Y | A = 0] = \mathbb{E}[\mathbb{E}[Y | A = 1, X] | A = 1] - \mathbb{E}[\mathbb{E}[Y | A = 0, X] | A = 0].$$

Method 1: g-computation

The idea can be extended to other (related) causal estimands:

- ▶ Average treatment effect among the treated (ATT):

$$ATT := \mathbb{E}[Y(1) - Y(0) \mid A = 1] = \sum_x \{\mu_1(x) - \mu_0(x)\} P(X = x \mid A = 1)$$

- ▶ Average treatment effect among the control:

$$\mathbb{E}[Y(1) - Y(0) \mid A = 0] = \sum_x \{\mu_1(x) - \mu_0(x)\} P(X = x \mid A = 0)$$

- ▶ Average treated effect in a target population (transported causal effect):

$$\sum_x \{\mu_1(x) - \mu_0(x)\} \tilde{P}(X = x),$$

where $\tilde{P}$ is the distribution of X in the target population.

A toy observational study: all binary variables

- ▶ Y = death (1: yes; 0: no)
- ▶ A = surgery (1: yes; 0: no)
- ▶ X = injury (1: severe; 0: non-severe)
- ▶ Naive contrast

$$\mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0] = 0.4 - 0.6 = -0.2.$$
- ▶ Note: prob. of receiving surgery are
 - 31% for non-severe injury
 - 86% for severe injury
- ▶ We assume there is no confounder beyond X .
 What is the causal effect of surgery on mortality?

$X = 0$		Outcome Y		Total
		0	1	
Treatment A	0	4	5	9
	1	3	1	4
Total		7	6	13

$X = 1$		Outcome Y		Total
		0	1	
Treatment A	0	0	1	1
	1	3	3	6
Total		3	4	7

Quiz 1: g-computation in our toy example

$$\begin{aligned}
 \text{ATE} &= \text{ATE}(0)P(X=0) + \text{ATE}(1)P(X=1) \\
 &= (E[Y(1)|X=0] - E[Y(0)|X=0])P(X=0) \\
 &\quad + (E[Y(1)|X=1] - E[Y(0)|X=1])P(X=1)
 \end{aligned}$$

$X = 0$		Outcome Y		Total
		0	1	
Treatment A	0	4	5	9
	1	3	1	4
Total		7	6	13

$X = 1$		Outcome Y		Total
		0	1	
Treatment A	0	0	1	1
	1	3	3	6
Total		3	4	7

Quiz 1: g-computation in our toy example

► $\mu_1(x=0) = 1/4, \mu_0(x=0) = 5/9$

► $\mu_1(x=1) = 3/6, \mu_0(x=1) = 1$

$$\begin{aligned} \text{ATE} &= \text{ATE}(0)P(X=0) + \text{ATE}(1)P(X=1) \\ &= (E[Y(1)|X=0] - E[Y(0)|X=0])P(X=0) \\ &\quad + (E[Y(1)|X=1] - E[Y(0)|X=1])P(X=1) \\ &= \left(\frac{1}{4} - \frac{5}{9}\right)\frac{13}{20} + \left(\frac{3}{6} - \frac{1}{1}\right)\frac{7}{20} \\ &= (-0.31) \times 0.65 + (-0.5) \times 0.35 \\ &= -0.37 \end{aligned}$$

$X = 0$		Outcome Y		Total
		0	1	
Treatment A	0	4	5	9
	1	3	1	4
Total		7	6	13

$X = 1$		Outcome Y		Total
		0	1	
Treatment A	0	0	1	1
	1	3	3	6
Total		3	4	7

Method 2: Inverse probability weighting (IPW)

👉 IPW assigns every unit a weight to create a pseudo-population in which A no longer depends on X .

Method 2: Inverse probability weighting (IPW)

👉 IPW assigns every unit a weight to create a pseudo-population in which A no longer depends on X .

$$\text{ATE} = \mathbb{E} \left[\frac{AY}{\pi(X)} - \frac{(1-A)Y}{1-\pi(X)} \right]$$

where $\pi(x) = P(A = 1|X = x)$ is the **propensity score**.

$X = x_1$



$X = x_2$



Method 2: Inverse probability weighting (IPW)

👉 IPW assigns every unit a weight to create a pseudo-population in which A no longer depends on X .

$$\text{ATE} = \mathbb{E} \left[\frac{AY}{\pi(X)} - \frac{(1-A)Y}{1-\pi(X)} \right]$$

where $\pi(x) = P(A = 1|X = x)$ is the **propensity score**.

- ▶ This is simply a weighted average of the outcome, reweighed according to their propensity of receiving their treatments.

 $X = x_1$  $X = x_2$ 

Method 2: Inverse probability weighting (IPW)

👉 IPW assigns every unit a weight to create a pseudo-population in which A no longer depends on X .

$$\text{ATE} = \mathbb{E} \left[\frac{AY}{\pi(X)} - \frac{(1-A)Y}{1-\pi(X)} \right]$$

where $\pi(x) = P(A = 1 | X = x)$ is the **propensity score**.

- ▶ This is simply a weighted average of the outcome, reweighed according to their propensity of receiving their treatments.
- ▶ Suppose 5 patients have $X = x$, among those 1 received the treatment, 4 received the control, i.e., $P(A = 1 | X = x) = 0.2$ and $P(A = 0 | X = x) = 0.8$. For $\mathbb{E} Y(1)$, the treated patient must stand in for the other 4 and has weight $1/0.2 = 5$; for $\mathbb{E} Y(0)$, the 4 control patients must stand in for the other 1 and each has weight $1/0.8 = 1.25$.

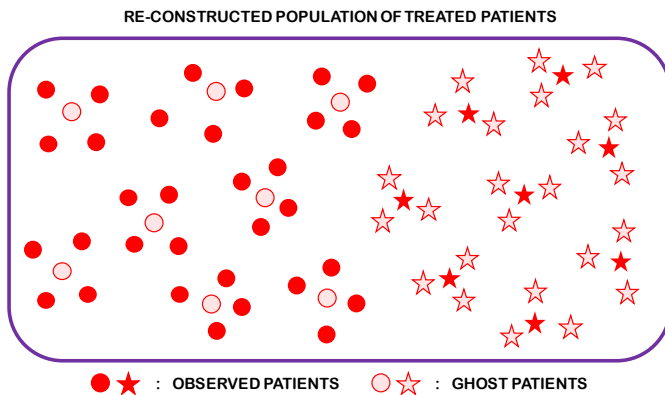
$X = x_1$



$X = x_2$



Method 2: Inverse probability weighting (IPW)



$$P(A = 1 \mid W = \star) = 0.25 \quad P(A = 1 \mid W = \bullet) = 0.80$$

Method 2: Inverse probability weighting (IPW)

- The IPW formula can be derived by repeated use of the law of total expectation:

$$\begin{aligned} & \mathbb{E} \left[\frac{AY}{P(A=1 | X)} \right] \\ &= \mathbb{E} \left[\frac{AY(1)}{P(A=1 | X)} \right] \quad (\text{consistency}) \\ &= \mathbb{E} \left[\frac{\mathbb{E}\{AY(1) | X\}}{P(A=1 | X)} \right] \\ &= \mathbb{E} \left[\frac{\mathbb{E}\{A | X\} \mathbb{E}\{Y(1) | X\}}{P(A=1 | X)} \right] \quad (\text{no unmeasured confounding}) \\ &= \mathbb{E} [\mathbb{E}\{Y(1) | X\}] \\ &= \mathbb{E} Y(1). \end{aligned}$$

Positivity: $P(X=x) > 0 \implies P(A=1 | X=x) > 0$.

Two seemingly different formula

$$\mathbb{E} Y(1) = \underbrace{\mathbb{E} \left[\frac{AY}{P(A=1 | X)} \right]}_{\text{IPW}} = \underbrace{\mathbb{E}[\mathbb{E}[Y | A=1, X]]}_{\text{g-computation}}.$$

Two seemingly different formula

$$\mathbb{E} Y(1) = \underbrace{\mathbb{E} \left[\frac{AY}{P(A=1 | X)} \right]}_{\text{IPW}} = \underbrace{\mathbb{E}[\mathbb{E}[Y | A=1, X]]}_{\text{g-computation}}.$$

👉 But this identity (in population) is no coincidence:

$$\begin{aligned} \mathbb{E} \left[\frac{AY}{P(A=1 | X)} \right] &= \mathbb{E} \left[\mathbb{E} \left\{ \frac{AY}{P(A=1 | X)} \mid X \right\} \right] = \mathbb{E} \left[\frac{\mathbb{E}[AY | X]}{P(A=1 | X)} \right] \\ &= \mathbb{E} \left[\frac{0 + P(A=1 | X) \mathbb{E}[Y | A=1, X]}{P(A=1 | X)} \right] \\ &= \mathbb{E}[\mathbb{E}[Y | A=1, X]]. \end{aligned}$$

Two seemingly different formula

$$\mathbb{E} Y(1) = \underbrace{\mathbb{E} \left[\frac{AY}{P(A=1 | X)} \right]}_{\text{IPW}} = \underbrace{\mathbb{E}[\mathbb{E}[Y | A=1, X]]}_{\text{g-computation}}.$$

👉 But this identity (in population) is no coincidence:

$$\begin{aligned} \mathbb{E} \left[\frac{AY}{P(A=1 | X)} \right] &= \mathbb{E} \left[\mathbb{E} \left\{ \frac{AY}{P(A=1 | X)} \mid X \right\} \right] = \mathbb{E} \left[\frac{\mathbb{E}[AY | X]}{P(A=1 | X)} \right] \\ &= \mathbb{E} \left[\frac{0 + P(A=1 | X) \mathbb{E}[Y | A=1, X]}{P(A=1 | X)} \right] \\ &= \mathbb{E}[\mathbb{E}[Y | A=1, X]]. \end{aligned}$$

👉 But it can make a difference when $P(A=1 | X)$ and $\mathbb{E}[Y | A=1, X]$ are replaced by estimates.

Quiz 2: IPW in our toy example

► $\pi(0) = P(A = 1 \mid X = 0)$

► $\pi(1) = P(A = 1 \mid X = 1)$

$$\begin{aligned} \text{ATE} &= \frac{1}{n} \sum_{i: X_i=0} \frac{A_i Y_i}{\pi(0)} - \frac{1}{n} \sum_{i: X_i=0} \frac{(1 - A_i) Y_i}{1 - \pi(0)} \\ &+ \frac{1}{n} \sum_{i: X_i=1} \frac{A_i Y_i}{\pi(1)} - \frac{1}{n} \sum_{i: X_i=1} \frac{(1 - A_i) Y_i}{1 - \pi(1)} \end{aligned}$$

$X = 0$		Outcome Y		Total
		0	1	
Treatment A	0	4	5	9
	1	3	1	4
Total		7	6	13

$X = 1$		Outcome Y		Total
		0	1	
Treatment A	0	0	1	1
	1	3	3	6
Total		3	4	7

Quiz 2: IPW in our toy example

► $\pi(0) = P(A = 1 \mid X = 0) = 4/13$

► $\pi(1) = P(A = 1 \mid X = 1) = 6/7$

$$\begin{aligned} \text{ATE} &= \frac{1}{n} \sum_{i:X_i=0} \frac{A_i Y_i}{4/13} - \frac{1}{n} \sum_{i:X_i=0} \frac{(1 - A_i) Y_i}{9/13} \\ &+ \frac{1}{n} \sum_{i:X_i=1} \frac{A_i Y_i}{6/7} - \frac{1}{n} \sum_{i:X_i=1} \frac{(1 - A_i) Y_i}{1/7} \\ &= \frac{1}{20} \left(\frac{13}{4} \times 1 - \frac{13}{9} \times 5 + \frac{7}{6} \times 3 - 7 \times 1 \right) \\ &= -0.37 \end{aligned}$$

$X = 0$		Outcome Y		Total
		0	1	
Treatment A	0	4	5	9
	1	3	1	4
Total		7	6	13

$X = 1$		Outcome Y		Total
		0	1	
Treatment A	0	0	1	1
	1	3	3	6
Total		3	4	7

Estimation

causal inference $\approx$ causal identification + statistical inference.

G-computation and IPW give causal identification formulas. The remaining is statistics!

Estimation with g-computation

G-computation

$$\text{ATE} = \sum_x \text{ATE}(x)P(X = x) = \sum_x (\mu_1(x) - \mu_0(x))P(X = x).$$

Estimation with g-computation

G-computation

$$\text{ATE} = \sum_x \text{ATE}(x)P(X = x) = \sum_x (\mu_1(x) - \mu_0(x))P(X = x).$$

1. Fit an outcome model $\hat{\mu}_1(x)$ using data from the treated group
2. Fit another outcome model $\hat{\mu}_0(x)$ using data from the control group
3. The estimator is

$$\widehat{\text{ATE}}_g := \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(X_i) - \frac{1}{n} \sum_{i=1}^n \hat{\mu}_0(X_i).$$

Estimation with g-computation

G-computation

$$\text{ATE} = \sum_x \text{ATE}(x)P(X = x) = \sum_x (\mu_1(x) - \mu_0(x))P(X = x).$$

1. Fit an outcome model $\hat{\mu}_1(x)$ using data from the treated group
2. Fit another outcome model $\hat{\mu}_0(x)$ using data from the control group
3. The estimator is

$$\widehat{\text{ATE}}_g := \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(X_i) - \frac{1}{n} \sum_{i=1}^n \hat{\mu}_0(X_i).$$

- Advantages: more efficient; usually computationally stable
- Disadvantages: sensitive to model misspecification; potential danger of extrapolation

Estimation with IPW

IPW

$$\text{ATE} = \mathbb{E} \left[\frac{AY}{\pi(X)} - \frac{(1-A)Y}{1-\pi(X)} \right], \quad \pi(x) = P(A=1|X=x).$$

Estimation with IPW

IPW

$$\text{ATE} = \mathbb{E} \left[\frac{AY}{\pi(X)} - \frac{(1-A)Y}{1-\pi(X)} \right], \quad \pi(x) = P(A=1|X=x).$$

1. Fit a propensity score model $\hat{\pi}(x)$ (e.g., logistic regression)
2. The IPW estimator (Horvitz–Thompson type)

$$\widehat{\text{ATE}}_{\text{IPW}}^{\text{HT}} := \frac{1}{n} \sum_{i=1}^n \frac{A_i Y_i}{\hat{\pi}(X_i)} - \frac{1}{n} \sum_{i=1}^n \frac{(1-A_i) Y_i}{1-\hat{\pi}(X_i)}.$$

Estimation with IPW

IPW

$$\text{ATE} = \mathbb{E} \left[\frac{AY}{\pi(X)} - \frac{(1-A)Y}{1-\pi(X)} \right], \quad \pi(x) = P(A=1|X=x).$$

1. Fit a propensity score model $\hat{\pi}(x)$ (e.g., logistic regression)
2. The IPW estimator (Horvitz–Thompson type)

$$\widehat{\text{ATE}}_{\text{IPW}}^{\text{HT}} := \frac{1}{n} \sum_{i=1}^n \frac{A_i Y_i}{\hat{\pi}(X_i)} - \frac{1}{n} \sum_{i=1}^n \frac{(1-A_i) Y_i}{1-\hat{\pi}(X_i)}.$$

- Advantages: sometimes the propensity score may be easier to model
- Disadvantages: sensitive to model misspecification; less efficient; can be unstable if some $\hat{\pi}(X_i)$ are close to zero

Estimation with IPW

Practical recommendations

- ▶ A usually more stable option is the stabilized, Hájek-type IPW estimator

$$\widehat{\text{ATE}}_{\text{IPW}}^{\text{Hájek}} = \frac{1}{n} \sum_{i=1}^n \frac{A_i Y_i}{\hat{\pi}(X_i)} / \left\{ \frac{1}{n} \sum_{i=1}^n \frac{A_i}{\hat{\pi}(X_i)} \right\} - \frac{1}{n} \sum_{i=1}^n \frac{(1 - A_i) Y_i}{1 - \hat{\pi}(X_i)} / \left\{ \frac{1}{n} \sum_{i=1}^n \frac{(1 - A_i)}{1 - \hat{\pi}(X_i)} \right\}$$

- ▶ It is helpful to examine $\hat{\pi}(X_i)$ separately for the treated and control groups. If $\hat{\pi}(X_i)$ can get very close to 0 or 1, we can either truncate $\hat{\pi}(X_i)$ at 0.1 and 0.9, or drop units with $\hat{\pi}(X_i)$ outside $[0.1, 0.9]$.
 - ▶ Caution: population change

Example: School meal program and body mass index

Chan et al. (2016) used a subsample of National Health and Nutrition Examination Survey (NHANES) 2007-2008 to study whether participation in school meal programs led to an increase in BMI for school children.

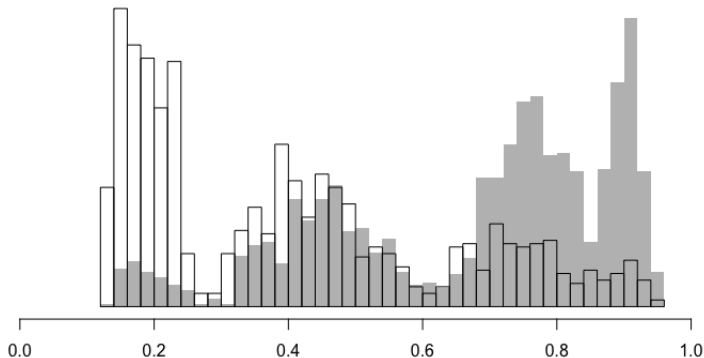
The dataset has the following key covariates: age, sex, race, ethnicity, family above 200% federal poverty level, participation in the special supplemental nutrition program, participation in food stamp program, childhood food security, any insurance, sex of the adult respondent, age of the adult respondent.

Given the nature of observational studies, covariates are not balanced between the treated and control groups. For example, children who receive treatment tend to come from families with incomes below 200% of the federal poverty level and are more likely to participate in other food programs.

A naive comparison (simple outcome mean difference) says that the treated children on average have 0.53 higher BMI compared to the control children.

Example: School meal program and body mass index

Estimated propensity scores for the treated (gray) and control (white)



Example: School meal program and body mass index

We apply the IPW and stabilized IPW, with the propensity scores truncated at $(0, 1)$, $(0.01, 0.99)$, $(0.05, 0.95)$, and $(0.1, 0.9)$, with bootstrap SE:

\$trunc0			\$trunc.01			\$trunc.05		
est	se		est	se		est	se	
IPW	-1.516	0.494	IPW	-1.516	0.468	IPW	-1.499	0.468
SIPW	-0.156	0.254	SIPW	-0.156	0.250	SIPW	-0.152	0.250

- ▶ The IPW: large SE, and is sensitive to the truncation values.
This is an example showing the instability of the IPW estimator
- ▶ Stabilized IPW: smaller SE and more stable

AIPW

Method 3: Augmented inverse probability weighting (AIPW)

👉 Why not use both the outcome model and the propensity score model?

$$\text{ATE} = \mathbb{E} \left[\underbrace{\mu_1(X) - \mu_0(X)}_{\text{g-computation}} \right].$$

Method 3: Augmented inverse probability weighting (AIPW)

👉 Why not use both the outcome model and the propensity score model?

$$\text{ATE} = \mathbb{E} \left[\underbrace{\mu_1(X) - \mu_0(X)}_{\text{g-computation}} + \underbrace{\frac{A\{Y - \mu_1(X)\}}{\pi(X)} - \frac{(1 - A)\{Y - \mu_0(X)\}}{1 - \pi(X)}}_{\text{augmentation, with mean zero}} \right].$$

Method 3: Augmented inverse probability weighting (AIPW)

👉 Why not use both the outcome model and the propensity score model?

$$\text{ATE} = \mathbb{E} \left[\underbrace{\mu_1(X) - \mu_0(X)}_{\text{g-computation}} + \underbrace{\frac{A\{Y - \mu_1(X)\}}{\pi(X)} - \frac{(1 - A)\{Y - \mu_0(X)\}}{1 - \pi(X)}}_{\text{augmentation, with mean zero}} \right].$$

- ▶ Augmentation seeks to rectify any incorrect estimation of $\mu_1(X), \mu_0(X)$ in g-computation.
- ▶ Suppose $\hat{\mu}_1(x)$ overestimates $\mu_1(x) = E(Y \mid A = a, X = x)$ throughout, then the g-computation estimator overshoots the target but the augmentation term is negative and brings it back down on target.

Method 3: Augmented inverse probability weighting (AIPW)

AIPW can also be rewritten as

$$\text{ATE} = \mathbb{E} \left[\underbrace{\frac{AY}{\pi(X)} - \frac{(1-A)Y}{1-\pi(X)}}_{\text{IPW}} + \underbrace{\left(1 - \frac{A}{\pi(X)}\right) \mu_1(X) - \left(1 - \frac{1-A}{1-\pi(X)}\right) \mu_0(X)}_{\text{augmentation, with mean zero}} \right].$$

- ▶ Augmentation seeks to rectify any incorrect estimation of $\pi(X)$ in IPW.
- ▶ Suppose Y is non-negative and $\hat{\pi}(x)$ underestimates $\pi(x) = P(A = 1 \mid X = x)$ throughout, then the IPW estimator overshoots the target but the augmentation term is negative and brings it back down on target.

Method 3: Double robustness of AIPW

Method 3: Double robustness of AIPW

Suppose $\hat{\mu}_a(x) \xrightarrow{P} \mu_a^*(x)$ for $a = 0, 1$, and $\hat{\pi}(x) \xrightarrow{P} \pi^*(x)$, where $\mu_0^*(x), \mu_1^*(x), \pi^*(x)$ are **not necessarily** equal to the true values $\mu_0(x), \mu_1(x), \pi(x)$.

Method 3: Double robustness of AIPW

Suppose $\hat{\mu}_a(x) \xrightarrow{P} \mu_a^*(x)$ for $a = 0, 1$, and $\hat{\pi}(x) \xrightarrow{P} \pi^*(x)$, where $\mu_0^*(x), \mu_1^*(x), \pi^*(x)$ are **not necessarily** equal to the true values $\mu_0(x), \mu_1(x), \pi(x)$.

Then,

$$\begin{aligned} & \mathbb{E} \left[\mu_1^*(X) - \mu_0^*(X) + \frac{A\{Y - \mu_1^*(X)\}}{\pi^*(X)} - \frac{(1-A)\{Y - \mu_0^*(X)\}}{1 - \pi^*(X)} \right] \\ &= \mathbb{E} \left[\mu_1^*(X) - \mu_0^*(X) + \frac{\pi(X)\{\mu_1(X) - \mu_1^*(X)\}}{\pi^*(X)} - \frac{(1 - \pi(X))\{\mu_0(X) - \mu_0^*(X)\}}{1 - \pi^*(X)} \right] \quad (\text{why?}) \end{aligned}$$

Method 3: Double robustness of AIPW

Suppose $\hat{\mu}_a(x) \xrightarrow{P} \mu_a^*(x)$ for $a = 0, 1$, and $\hat{\pi}(x) \xrightarrow{P} \pi^*(x)$, where $\mu_0^*(x), \mu_1^*(x), \pi^*(x)$ are **not necessarily** equal to the true values $\mu_0(x), \mu_1(x), \pi(x)$.

Then,

$$\begin{aligned} & \mathbb{E} \left[\mu_1^*(X) - \mu_0^*(X) + \frac{A\{Y - \mu_1^*(X)\}}{\pi^*(X)} - \frac{(1-A)\{Y - \mu_0^*(X)\}}{1 - \pi^*(X)} \right] \\ &= \mathbb{E} \left[\mu_1^*(X) - \mu_0^*(X) + \frac{\pi(X)\{\mu_1(X) - \mu_1^*(X)\}}{\pi^*(X)} - \frac{(1 - \pi(X))\{\mu_0(X) - \mu_0^*(X)\}}{1 - \pi^*(X)} \right] \quad (\text{why?}) \end{aligned}$$

Double robustness: the above equation equals ATE if either

- (i) $\mu_1^*(x) = \mu_1(x)$ and $\mu_0^*(x) = \mu_0(x)$, or
- (ii) $\pi^*(x) = \pi(x)$.

👉 **Two chances to get it right!**

Estimation after causal identification

AIPW

1. Fit an outcome model $\hat{\mu}_1(x)$ using data from the treated group
2. Fit another outcome model $\hat{\mu}_0(x)$ using data from the control group
3. Fit a propensity score model $\hat{\pi}(x)$
4. The estimator is

$$\widehat{\text{ATE}}_{\text{AIPW}} := \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(X_i) - \frac{1}{n} \sum_{i=1}^n \hat{\mu}_0(X_i) + \frac{1}{n} \sum_{i=1}^n \frac{A_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} - \frac{1}{n} \sum_{i=1}^n \frac{(1 - A_i)(Y_i - \hat{\mu}_0(X_i))}{1 - \hat{\pi}(X_i)}$$

Estimation after causal identification

AIPW

1. Fit an outcome model $\hat{\mu}_1(x)$ using data from the treated group
2. Fit another outcome model $\hat{\mu}_0(x)$ using data from the control group
3. Fit a propensity score model $\hat{\pi}(x)$
4. The estimator is

$$\widehat{\text{ATE}}_{\text{AIPW}} := \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(X_i) - \frac{1}{n} \sum_{i=1}^n \hat{\mu}_0(X_i) + \frac{1}{n} \sum_{i=1}^n \frac{A_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} - \frac{1}{n} \sum_{i=1}^n \frac{(1 - A_i)(Y_i - \hat{\mu}_0(X_i))}{1 - \hat{\pi}(X_i)}$$

- Advantages: doubly robust; more efficient and stable compared to IPW; can be used in combination with machine learning algorithms
- Disadvantages: still biased if all models are wrong; can be unstable if some $\hat{\pi}(X_i)$ are close to zero

Robust covariate adjustment for RCT

Recall that in an RCT, we can use **baseline covariates** to improve efficiency.

$$\begin{aligned}\widehat{\text{ATE}}_{\text{AIPW}} &= \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(X_i) - \frac{1}{n} \sum_{i=1}^n \hat{\mu}_0(X_i) + \frac{1}{n} \sum_{i=1}^n \frac{A_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} - \frac{1}{n} \sum_{i=1}^n \frac{(1 - A_i)(Y_i - \hat{\mu}_0(X_i))}{1 - \hat{\pi}(X_i)} \\ &= \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(X_i) - \frac{1}{n} \sum_{i=1}^n \hat{\mu}_0(X_i) + \frac{1}{n} \sum_{i:A_i=1} \frac{Y_i - \hat{\mu}_1(X_i)}{\hat{\pi}(X_i)} - \frac{1}{n} \sum_{i:A_i=0} \frac{(Y_i - \hat{\mu}_0(X_i))}{1 - \hat{\pi}(X_i)}.\end{aligned}$$

Robust covariate adjustment for RCT

Recall that in an RCT, we can use **baseline covariates** to improve efficiency.

$$\begin{aligned}\widehat{\text{ATE}}_{\text{AIPW}} &= \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(X_i) - \frac{1}{n} \sum_{i=1}^n \hat{\mu}_0(X_i) + \frac{1}{n} \sum_{i=1}^n \frac{A_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} - \frac{1}{n} \sum_{i=1}^n \frac{(1 - A_i)(Y_i - \hat{\mu}_0(X_i))}{1 - \hat{\pi}(X_i)} \\ &= \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(X_i) - \frac{1}{n} \sum_{i=1}^n \hat{\mu}_0(X_i) + \frac{1}{n} \sum_{i:A_i=1} \frac{Y_i - \hat{\mu}_1(X_i)}{\hat{\pi}(X_i)} - \frac{1}{n} \sum_{i:A_i=0} \frac{(Y_i - \hat{\mu}_0(X_i))}{1 - \hat{\pi}(X_i)}.\end{aligned}$$

👉 Replacing $\hat{\pi}(X_i)$ by n_1/n , we get

► This ensures consistency (why?)

$$\begin{aligned}\widehat{\text{ATE}}_{\text{AIPW}} &= \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(X_i) - \frac{1}{n} \sum_{i=1}^n \hat{\mu}_0(X_i) + \frac{1}{n_1} \sum_{i:A_i=1} (Y_i - \hat{\mu}_1(X_i)) - \frac{1}{n_0} \sum_{i:A_i=0} (Y_i - \hat{\mu}_0(X_i)) \\ &= \bar{Y}_1 - \bar{Y}_0 - \underbrace{\left\{ \frac{1}{n_1} \sum_{i:A_i=1} \hat{\mu}_1(X_i) - \frac{1}{n_0} \sum_{i:A_i=0} \hat{\mu}_0(X_i) \right\}}_{=0 \text{ for logistic, Poisson, etc.}} + \underbrace{\left\{ \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(X_i) - \frac{1}{n} \sum_{i=1}^n \hat{\mu}_0(X_i) \right\}}_{\text{g-computation}}.\end{aligned}$$

AIPW in our toy example¹

- ▶ $\mu_1(x=0) = 1/4, \mu_0(x=0) = 5/9$
- ▶ $\mu_1(x=1) = 3/6, \mu_0(x=1) = 1$
- ▶ $P(A=1|X=0) = 4/13$
- ▶ $P(A=1|X=1) = 6/7$

$$\begin{aligned} \text{ATE} &= \frac{1}{n} \sum_{i:X_i=0} \frac{A_i(Y_i - 1/4)}{A_i} - \frac{1}{n} \sum_{i:X_i=0} \frac{(1 - A_i)(Y_i - 5/9)}{1 - A_i} \\ &+ \frac{1}{n} \sum_{i:X_i=1} \frac{A_i(Y_i - 3/6)}{A_i} - \frac{1}{n} \sum_{i:X_i=1} \frac{(1 - A_i)(Y_i - 1)}{1 - A_i} \\ &+ \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(X_i) - \frac{1}{n} \sum_{i=1}^n \hat{\mu}_0(X_i) \end{aligned}$$

$X = 0$		Outcome Y		Total
		0	1	
Treatment A	0	4	5	9
	1	3	1	4
Total		7	6	13

$X = 1$		Outcome Y		Total
		0	1	
Treatment A	0	0	1	1
	1	3	3	6
Total		3	4	7

¹When X is discrete and $\hat{\pi}(x) \in (0, 1)$ for all x , all three estimators are equal.

AIPW in our toy example¹

- ▶ $\mu_1(x=0) = 1/4, \mu_0(x=0) = 5/9$
- ▶ $\mu_1(x=1) = 3/6, \mu_0(x=1) = 1$
- ▶ $P(A=1|X=0) = 4/13$
- ▶ $P(A=1|X=1) = 6/7$

$$\begin{aligned}
 \text{ATE} &= \frac{1}{n} \sum_{i:X_i=0} \frac{A_i(Y_i - 1/4)}{4/13} - \frac{1}{n} \sum_{i:X_i=0} \frac{(1 - A_i)(Y_i - 5/9)}{9/13} \\
 &+ \frac{1}{n} \sum_{i:X_i=1} \frac{A_i(Y_i - 3/6)}{6/7} - \frac{1}{n} \sum_{i:X_i=1} \frac{(1 - A_i)(Y_i - 1)}{1/7} \\
 &+ \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(X_i) - \frac{1}{n} \sum_{i=1}^n \hat{\mu}_0(X_i) \\
 &= 0 + \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(X_i) - \frac{1}{n} \sum_{i=1}^n \hat{\mu}_0(X_i) = -0.37
 \end{aligned}$$

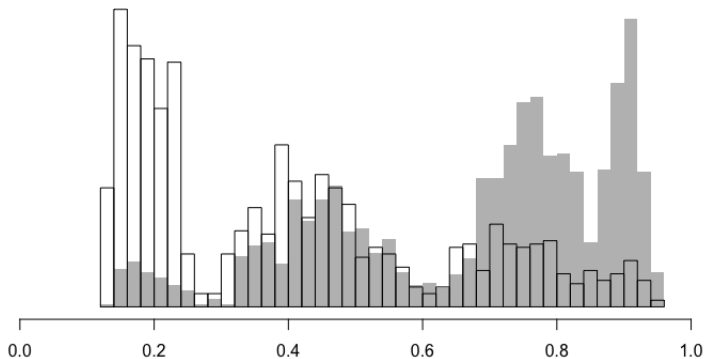
$X = 0$		Outcome Y		Total
		0	1	
Treatment A	0	4	5	9
	1	3	1	4
Total		7	6	13

$X = 1$		Outcome Y		Total
		0	1	
Treatment A	0	0	1	1
	1	3	3	6
Total		3	4	7

¹When X is discrete and $\hat{\pi}(x) \in (0, 1)$ for all x , all three estimators are equal.

Example: School meal program and body mass index

Estimated propensity scores for the treated (gray) and control (white)



Example: School meal program and body mass index

Comparing all the estimators (bootstrap SE, no propensity score truncation):

	reg	IPW	SIPW	DR
est	-0.017	-1.516	-0.156	-0.019
se	0.231	0.513	0.256	0.235

Comparing all the estimators (bootstrap SE, propensity score truncated at [0.1, 0.9]):

	reg	IPW	SIPW	DR
est	-0.017	-0.713	-0.054	-0.043
se	0.226	0.418	0.239	0.235

Statistical inference

1. When $\hat{\mu}_0(x), \hat{\mu}_1(x), \hat{\pi}(x)$ are estimated using **parametric models**:

Statistical inference

1. When $\hat{\mu}_0(x)$, $\hat{\mu}_1(x)$, $\hat{\pi}(x)$ are estimated using **parametric models**:
 - ▶ Bootstrap for estimating standard errors, computing confidence intervals, and hypothesis testing
 - The lazy statistician's method
 - Sample with replacement to create a new sample of the same size as the study sample, estimate the effect estimate in that sample, repeat many (e.g., 1000) times, find 2.5 and 97.5 percentiles of the 1000 estimates as the 95% confidence interval
 - ▶ Sandwich variance estimator (implemented in the CausalGAM package in R)

Statistical inference

1. When $\hat{\mu}_0(x)$, $\hat{\mu}_1(x)$, $\hat{\pi}(x)$ are estimated using **parametric models**:
 - ▶ Bootstrap for estimating standard errors, computing confidence intervals, and hypothesis testing
 - The lazy statistician's method
 - Sample with replacement to create a new sample of the same size as the study sample, estimate the effect estimate in that sample, repeat many (e.g., 1000) times, find 2.5 and 97.5 percentiles of the 1000 estimates as the 95% confidence interval
 - ▶ Sandwich variance estimator (implemented in the CausalGAM package in R)
2. When $\hat{\mu}_0(x)$, $\hat{\mu}_1(x)$, $\hat{\pi}(x)$ are estimated using **machine learning algorithms** (like random forest):
 - ▶ Use AIPW + cross fitting (implemented in the AIPW package in R)
 - ▶ Influence-function based variance estimator

Statistical inference

1. When $\hat{\mu}_0(x)$, $\hat{\mu}_1(x)$, $\hat{\pi}(x)$ are estimated using **parametric models**:
 - ▶ Bootstrap for estimating standard errors, computing confidence intervals, and hypothesis testing
 - The lazy statistician's method
 - Sample with replacement to create a new sample of the same size as the study sample, estimate the effect estimate in that sample, repeat many (e.g., 1000) times, find 2.5 and 97.5 percentiles of the 1000 estimates as the 95% confidence interval
 - ▶ Sandwich variance estimator (implemented in the CausalGAM package in R)
2. When $\hat{\mu}_0(x)$, $\hat{\mu}_1(x)$, $\hat{\pi}(x)$ are estimated using **machine learning algorithms** (like random forest):
 - ▶ Use AIPW + cross fitting (implemented in the AIPW package in R)
 - ▶ Influence-function based variance estimator

👉 The primary consideration is to choose an approach such that $\hat{\mu}_1(x)$, $\hat{\mu}_0(x)$, $\hat{\pi}(x)$ are close to the truth.

Propensity score

The central role of the propensity score

- ▶ Under the **NUCA** $A \perp\!\!\!\perp Y(0), Y(1) \mid X$, we have introduced the propensity score $\pi(X) = P(A = 1 \mid X)$, which is the probability of receiving treatment given covariate value X .
- ▶ To remove confounding, we need to adjust for covariates X (e.g., age, race, ...)
- ▶ **Key observation:** $\pi(X)$ is a scalar and coarsest summary of the observed covariates X that can make the treated and control groups comparable (Rosenbaum and Rubin, 1983)

$$A \perp\!\!\!\perp Y(0), Y(1) \mid \pi(X).$$

- ▶ In practice, we estimate $\pi(X)$, commonly by fitting a logistic regression of A_i on X_i . Denote the estimated propensity score as $\hat{\pi}(X)$.

The central role of the propensity score

Other approaches motivated by $A \perp\!\!\!\perp Y(0), Y(1) \mid \pi(X)$.

- ▶ **Subclassification (propensity score stratification)**: We stratify by the estimated propensity score $\hat{\pi}(X)$, e.g. with five subclasses. Within each subclass, the true propensity score $\pi(X)$ is approximately constant. We can estimate the outcome mean difference within each stratum and combine them by a weighted average.
- ▶ **g-computation (outcome regression) with the propensity score as a covariate**: the above formula shows that we can just use $\pi(X)$ as a “derived covariate”.
- ▶ **Propensity score matching**

- Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(3):673–700.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.