INTRODUCTION TO CAUSAL INFERENCE

Michael Kühhirt

November 27, 2017

v.2017-11-27-1:24pm

Estimation of total effects II

LAST WEEK: COVARIATE ADJUSTMENT USING REGRESSION

Q: What does covariate adjustment do?

LAST WEEK: COVARIATE ADJUSTMENT USING REGRESSION

Q: What does covariate adjustment do?

A: It equalizes (or balances) the distribution of measured covariates Z between different values of the treatment.

LAST WEEK: COVARIATE ADJUSTMENT USING REGRESSION

Q: What does covariate adjustment do?

A: It equalizes (or balances) the distribution of measured covariates Z between different values of the treatment.

Q: What for?

LAST WEEK: COVARIATE ADJUSTMENT USING REGRESSION

Q: What does covariate adjustment do?

A: It equalizes (or balances) the distribution of measured covariates Z between different values of the treatment.

Q: What for?

A: In order to get rid of the portion of the association between X and Y that's based on differences in Z (and thereby, hopefully, to identify the ATE of X on Y).

LAST WEEK: COVARIATE ADJUSTMENT USING REGRESSION

Q: What does covariate adjustment do?

A: It equalizes (or balances) the distribution of measured covariates Z between different values of the treatment.

Q: What for?

A: In order to get rid of the portion of the association between X and Y that's based on differences in Z (and thereby, hopefully, to identify the ATE of X on Y).

Q: So, how do standardization and regression do that (in priciple)?

LAST WEEK: COVARIATE ADJUSTMENT USING REGRESSION

Q: What does covariate adjustment do?

A: It equalizes (or balances) the distribution of measured covariates Z between different values of the treatment.

Q: What for?

A: In order to get rid of the portion of the association between X and Y that's based on differences in Z (and thereby, hopefully, to identify the ATE of X on Y).

Q: So, how do standardization and regression do that (in priciple)?

A: By first calculating the x-specific means of Y *within* values of Z (and thus "holding Z constant") and then "pretending" that Z has the same distribution in all x by multiplying the respective mean by the marginal distribution of Z, $P(Z = z)$.

LAST WEEK: COVARIATE ADJUSTMENT USING REGRESSION

Q: What does covariate adjustment do?

A: It equalizes (or balances) the distribution of measured covariates Z between different values of the treatment.

Q: What for?

A: In order to get rid of the portion of the association between X and Y that's based on differences in Z (and thereby, hopefully, to identify the ATE of X on Y).

Q: So, how do standardization and regression do that (in priciple)?

A: By first calculating the x-specific means of Y *within* values of Z (and thus "holding Z constant") and then "pretending" that Z has the same distribution in all x by multiplying the respective mean by the marginal distribution of Z, $P(Z = z)$.

Q: Say that again!?!?

LAST WEEK: COVARIATE ADJUSTMENT USING REGRESSION

Q: What does covariate adjustment do?

A: It equalizes (or balances) the distribution of measured covariates Z between different values of the treatment.

Q: What for?

A: In order to get rid of the portion of the association between X and Y that's based on differences in Z (and thereby, hopefully, to identify the ATE of X on Y).

Q: So, how do standardization and regression do that (in priciple)?

A: By first calculating the x-specific means of Y *within* values of Z (and thus "holding Z constant") and then "pretending" that Z has the same distribution in all x by multiplying the respective mean by the marginal distribution of Z, $P(Z = z)$.

Q: Say that again!?!?

A: It simply calculates the association between X and Y for a scenario in which Z has the same distribution in all x. In other words, it *simulates* the association between X and Y if X was independent from Z.

LAST WEEK:
COVARIATE ADJUSTMENT USING REGRESSION

Q: What does covariate adjustment do?

A: It equalizes (or balances) the distribution of measured covariates Z between different values of the treatment.

Q: What for?

A: In order to get rid of the portion of the association between X and Y that's based on differences in Z (and thereby, hopefully, to identify the ATE of X on Y).

Q: So, how do standardization and regression do that (in priciple)?

A: By first calculating the x-specific means of Y *within* values of Z (and thus "holding Z constant") and then "pretending" that Z has the same distribution in all x by multiplying the respective mean by the marginal distribution of Z, $P(Z = z)$.

Q: Say that again!?!?

A: It simply calculates the association between X and Y for a scenario in which Z has the same distribution in all x. In other words, it *simulates* the association between X and Y if X was independent from Z.

Q: Hmmm, that's exactly what we want if we want to learn about the ATE, right?

LAST WEEK: COVARIATE ADJUSTMENT USING REGRESSION

Q: What does covariate adjustment do?

A: It equalizes (or balances) the distribution of measured covariates Z between different values of the treatment.

Q: What for?

A: In order to get rid of the portion of the association between X and Y that's based on differences in Z (and thereby, hopefully, to identify the ATE of X on Y).

Q: So, how do standardization and regression do that (in priciple)?

A: By first calculating the x-specific means of Y *within* values of Z (and thus "holding Z constant") and then "pretending" that Z has the same distribution in all x by multiplying the respective mean by the marginal distribution of Z, $P(Z = z)$.

Q: Say that again!?!?

A: It simply calculates the association between X and Y for a scenario in which Z has the same distribution in all x. In other words, it *simulates* the association between X and Y if X was independent from Z.

Q: Hmmm, that's exactly what we want if we want to learn about the ATE, right?

A: Yup! Important caveat though: adjustment can only equalize the distribution of measured covariates (unlike randomization, remember?).

TODAY: COVARIATE ADJUSTMENT USING IPT WEIGHTING

1. Nonparametric IPT weighting
2. Parametric estimation of IPT weights
3. Comparing the results of different estimators
4. Checks for model misspecification and positivity violations

NONPARAMETRIC IPT WEIGHTING

WHAT IS INVERSE PROBABILITY OF TREATMENT (IPT) WEIGHTING?

IPT weighting is an *alternative* technique for covariate adjustment and thus provides an additional estimator for ATEs.

WHAT IS INVERSE PROBABILITY OF TREATMENT (IPT) WEIGHTING?

IPT weighting is an *alternative* technique for covariate adjustment and thus provides an additional estimator for ATEs.

So it has the same goal as the techniques from last week but it takes a different route to achieve it.

WHAT IS INVERSE PROBABILITY OF TREATMENT (IPT) WEIGHTING?

IPT weighting is an *alternative* technique for covariate adjustment and thus provides an additional estimator for ATEs.

So it has the same goal as the techniques from last week but it takes a different route to achieve it.

Unlike standardization and regression it doesn't calculate/predict means for Y within values of Z.

WHAT IS INVERSE PROBABILITY OF TREATMENT (IPT) WEIGHTING?

IPT weighting is an *alternative* technique for covariate adjustment and thus provides an additional estimator for ATEs.

So it has the same goal as the techniques from last week but it takes a different route to achieve it.

Unlike standardization and regression it doesn't calculate/predict means for Y within values of Z.

Instead, it weights each unit of observation with the inverse probability of the value of X (the "treatment") that is actually observed for that unit.

WHAT IS INVERSE PROBABILITY OF TREATMENT (IPT) WEIGHTING?

IPT weighting is an *alternative* technique for covariate adjustment and thus provides an additional estimator for ATEs.

So it has the same goal as the techniques from last week but it takes a different route to achieve it.

Unlike standardization and regression it doesn't calculate/predict means for Y within values of Z.

Instead, it weights each unit of observation with the inverse probability of the value of X (the "treatment") that is actually observed for that unit.

This *also* balances the distribution of (measured!) covariates Z, but *without* looking at the association between X and Y within the values of Z (i.e., *without conditioning* on Z).

WHAT IS INVERSE PROBABILITY OF TREATMENT (IPT) WEIGHTING?

IPT weighting is an *alternative* technique for covariate adjustment and thus provides an additional estimator for ATEs.

So it has the same goal as the techniques from last week but it takes a different route to achieve it.

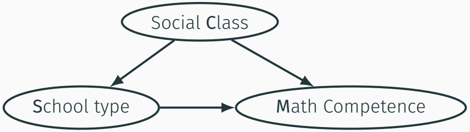
Unlike standardization and regression it doesn't calculate/predict means for Y within values of Z.

Instead, it weights each unit of observation with the inverse probability of the value of X (the "treatment") that is actually observed for that unit.

This *also* balances the distribution of (measured!) covariates Z, but *without* looking at the association between X and Y within the values of Z (i.e., *without conditioning* on Z).

This doesn't make much of a difference for estimating total effects, but it can be important for estimating direct effects and cumulative effects later on.

TOY DATA FOR DEMONSTRATION OF NONPARAMETRIC ADJUSTMENT



Data generating process for `07estte1a.dta` ($N = 100,000$):

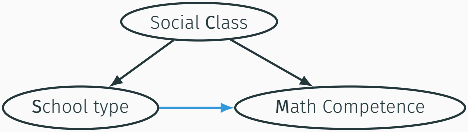
$$C = \varepsilon_C$$

$$S = 0.2C + \varepsilon_S$$

$$M = 20C - 5S + 5SC + \varepsilon_M$$

C is binary (1=higher class; 0=lower class); S is binary (1=private; 0=public); M is continuous

TOY DATA FOR DEMONSTRATION OF NONPARAMETRIC ADJUSTMENT



Data generating process for `07estte1a.dta` ($N = 100,000$):

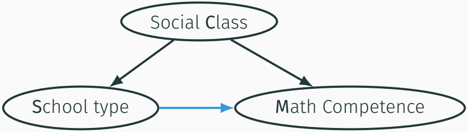
$$C = \varepsilon_C$$
$$S = 0.2C + \varepsilon_S$$
$$M = 20C - 5S + 5SC + \varepsilon_M$$

C is binary (1=higher class; 0=lower class); S is binary (1=private; 0=public); M is continuous

Causal parameter of interest (unobservable):

$$\text{ATE}_{S \to M} = E(M^{S=\text{private}}) - E(M^{S=\text{public}}) = 45 - 49 = -4$$

TOY DATA FOR DEMONSTRATION OF NONPARAMETRIC ADJUSTMENT



Data generating process for `07estte1a.dta` ($N = 100,000$):

$$C = \varepsilon_C$$
$$S = 0.2C + \varepsilon_S$$
$$M = 20C - 5S + 5SC + \varepsilon_M$$

C is binary (1=higher class; 0=lower class); S is binary (1=private; 0=public); M is continuous

Causal parameter of interest (unobservable):

$$\text{ATE}_{S\rightarrow M} = E(M^{S=\text{private}}) - E(M^{S=\text{public}}) = 45 - 49 = -4$$

Statistical parameter, δ, that identifies $\text{ATE}_{S\rightarrow M}$:

$$\delta = \sum_c E(M|S = \text{private}, C = c)P(C = c) - \sum_c E(M|S = \text{public}, C = c)P(C = c)$$
$$= E\left(M \frac{U_{S=\text{private}}}{P(S = \text{private}|C = c)}\right) - E\left(M \frac{U_{S=\text{public}}}{P(S = \text{public}|C = c)}\right)$$

(Hernán and Robins, 2018, Part I, p. 24)

NONPARAMETRIC COVARIATE ADJUSTMENT: IPT WEIGHTING ESTIMATOR

$$\begin{aligned}\hat{\delta} &= \hat{E}\left(M\frac{U_{S=\text{private}}}{\hat{P}(S=\text{private}|C=c)}\right) - \hat{E}\left(M\frac{U_{S=\text{public}}}{\hat{P}(S=\text{public}|C=c)}\right)\\ &= \hat{E}_w(M|S=\text{private}) - \hat{E}_w(M|S=\text{public})\end{aligned}$$

with $\hat{E}_W(.)$ being the average after weighting each unit u with $\widehat{IPTW} = \frac{1}{\hat{P}(S=s|C=c)}$

1. In each c, calculate the probability that S=s, $\hat{P}(S=s|C=c)$.

(Hernán and Robins, 2018, Ch. 2.4)

NONPARAMETRIC COVARIATE ADJUSTMENT: IPT WEIGHTING ESTIMATOR

$$\hat{\delta} = \hat{E}\left(M \frac{U_{S=\text{private}}}{\hat{P}(S = \text{private}|C = c)}\right) - \hat{E}\left(M \frac{U_{S=\text{public}}}{\hat{P}(S = \text{public}|C = c)}\right)$$
$$= \hat{E}_w(M|S = \text{private}) - \hat{E}_w(M|S = \text{public})$$

with $\hat{E}_W(.)$ being the average after weighting each unit u with $\widehat{IPTW} = \frac{1}{\hat{P}(S=s|C=c)}$

1. In each c, calculate the probability that S=s, $\hat{P}(S = s|C = c)$.
2. Calculate the inverse of that probability, $\frac{1}{\hat{P}(S=s|C=c)}$.

(Hernán and Robins, 2018, Ch. 2.4)

NONPARAMETRIC COVARIATE ADJUSTMENT: IPT WEIGHTING ESTIMATOR

$$\hat{\delta} = \hat{E}\left(M \frac{U_{S=\text{private}}}{\hat{P}(S = \text{private}|C = c)}\right) - \hat{E}\left(M \frac{U_{S=\text{public}}}{\hat{P}(S = \text{public}|C = c)}\right)$$
$$= \hat{E}_w(M|S = \text{private}) - \hat{E}_w(M|S = \text{public})$$

with $\hat{E}_W(.)$ being the average after weighting each unit u with $\widehat{IPTW} = \frac{1}{\hat{P}(S=s|C=c)}$

1. In each c, calculate the probability that S=s, $\hat{P}(S = s|C = c)$.
2. Calculate the inverse of that probability, $\frac{1}{\hat{P}(S=s|C=c)}$.
3. Weight (i.e., multiply) each unit of observation with the inverse probability of their observed S.

(Hernán and Robins, 2018, Ch. 2.4)

NONPARAMETRIC COVARIATE ADJUSTMENT: IPT WEIGHTING ESTIMATOR

$$\hat{\delta} = \hat{E}\left(M \frac{U_{S=\text{private}}}{\hat{P}(S = \text{private}|C = c)}\right) - \hat{E}\left(M \frac{U_{S=\text{public}}}{\hat{P}(S = \text{public}|C = c)}\right)$$
$$= \hat{E}_w(M|S = \text{private}) - \hat{E}_w(M|S = \text{public})$$

with $\hat{E}_W(.)$ being the average after weighting each unit u with $\widehat{IPTW} = \frac{1}{\hat{P}(S=s|C=c)}$

1. In each c, calculate the probability that S=s, $\hat{P}(S = s|C = c)$.
2. Calculate the inverse of that probability, $\frac{1}{\hat{P}(S=s|C=c)}$.
3. Weight (i.e., multiply) each unit of observation with the inverse probability of their observed S.
4. Calculate the mean difference in M by s in the weighted pseudo-population.

(Hernán and Robins, 2018, Ch. 2.4)

NONPARAMETRIC COVARIATE ADJUSTMENT: IPT WEIGHTING ESTIMATOR

$$\begin{aligned}\hat{\delta} &= \hat{E}\left(M\frac{U_{S=\text{private}}}{\hat{P}(S=\text{private}|C=c)}\right) - \hat{E}\left(M\frac{U_{S=\text{public}}}{\hat{P}(S=\text{public}|C=c)}\right)\\ &= \hat{E}_w(M|S=\text{private}) - \hat{E}_w(M|S=\text{public})\end{aligned}$$

with $\hat{E}_W(.)$ being the average after weighting each unit u with $\widehat{IPTW} = \frac{1}{\hat{P}(S=s|C=c)}$

1. In each c, calculate the probability that S=s, $\hat{P}(S=s|C=c)$.
2. Calculate the inverse of that probability, $\frac{1}{\hat{P}(S=s|C=c)}$.
3. Weight (i.e., multiply) each unit of observation with the inverse probability of their observed S.
4. Calculate the mean difference in M by s in the weighted pseudo-population.
5. Result: estimate of $\text{ATE}_{S\rightarrow M}$

(Hernán and Robins, 2018, Ch. 2.4)

CALCULATE $\hat{P}(S = s|C = c)$

```
. tab pschool hiclass, col

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

           |        hiclass
   pschool |         0          1 |     Total
-----------+----------------------+----------
         0 |    71,683     14,163 |    85,846 
           |     89.85      70.05 |     85.85 
-----------+----------------------+----------
         1 |     8,099      6,055 |    14,154 
           |     10.15      29.95 |     14.15 
-----------+----------------------+----------
     Total |    79,782     20,218 |   100,000 
           |    100.00     100.00 |    100.00 
```

Source: `08estte2.do#1A`

CALCULATE $\widehat{\text{IPTW}} = 1/\hat{P}(S = s|C = c)$

```
. gen iptw=.
(100,000 missing values generated)
. replace iptw=1/.1015  if hiclass==0 & pschool==1
(8,099 real changes made)
. replace iptw=1/.8985  if hiclass==0 & pschool==0
>
(71,683 real changes made)
. replace iptw=1/.2995  if hiclass==1 & pschool==1
(6,055 real changes made)
. replace iptw=1/.7005  if hiclass==1 & pschool==0
(14,163 real changes made)
```

Source: `08estte2.do#1B`

APPLY $\widehat{\text{IPTW}}$ AND CALCULATE $\hat{E}_w(M|S = \text{private}) - \hat{E}_w(M|S = \text{public})$

```
. tabstat mathcomp [aw=iptw], by(pschool) notot
Summary for variables: mathcomp
     by categories of: pschool
 pschool |      mean
---------+----------
       0 |  49.00899
       1 |  45.10707
---------+----------

. reg mathcomp pschool [pw=iptw], nohead nopvalues
(sum of wgt is   2.0001e+05)
-------------------------------------------------------------
             |               Robust
    mathcomp |      Coef.   Std. Err.     [95% Conf. Interval]
-------------+-----------------------------------------------
     pschool |      -3.90       0.12         -4.14       -3.67
       _cons |      49.01       0.04         48.92       49.10
-------------------------------------------------------------
```

Source: `08estte2.do#1C`

$$\hat{\delta} = -3.9$$

After weighting, conditioning on C is no longer necessary!

S AND C INDEPENDENT IN THE WEIGHTED PSEUDO-POPULATION

```
. tab hiclass pschool [aw=iptw], nofreq col
```

hiclass	pschool 0	1	Total
0	79.78	79.79	79.78
1	20.22	20.21	20.22
Total	100.00	100.00	100.00

Source: `08estte2.do#1D`

Weighting balances the distribution of covariates over values of the treatment by downweighting (upweighting) units with a large (small) probability to receive the treatment value with which they're observed.

S AND C DEPENDENT IN THE POPULATION

```
. tab hiclass pschool, nofreq col

           |        pschool
   hiclass |         0          1 |     Total
-----------+----------------------+----------
         0 |     83.50      57.22 |     79.78 
         1 |     16.50      42.78 |     20.22 
-----------+----------------------+----------
     Total |    100.00     100.00 |    100.00 
```

Source: `08estte2.do#1E`

So in these data, units from higher (lower) social class are upweighted (downweighted) in the group of public school students and units from lower (higher) social class are upweighted (downweighted) in the group of private school students.

INTERPRETATION OF ESTIMATES

statistical *On average, private school students have a 3.9 [95% CI:−4.14, −3.67] points lower math competence than public school students after adjusting for differences in social class (in the population from which the sample was drawn).*

counterfactual *Had every student attended private school instead of public school, average math competence would have been 3.9 [95% CI:−4.14, −3.67] points lower (in the population from which the sample was drawn).*
Or:
Attending private school instead of public school leads to an average decrease in math competence by 3.9 points (in the population from which the sample was drawn).

INTERPRETATION OF ESTIMATES

Assumptions for interpretation (after nonparametric adjustment):

statistical
- No random bias
- No measurement bias
  (see Hernán and Robins, 2018, Ch.9)

counterfactual the previous plus
- No positivity violation
  There are private school students and public school students in each social class.
- No confounding bias, no overcontrol bias, no endogenous selection bias
  We d-separated every noncausal path, we didn't d-separate any causal path, we didn't d-connect any noncausal path from school type to math competence.

IPT WEIGHTING OF MARGINAL STRUCTURAL MODELS

Recall that we can use a marginal structural (mean) model (MSM) to define counterfactual means of Y and their dependence on the values of X:

$$E(Y^{X=x}) = \theta_0 + \theta_1 x$$

(Hernán and Robins, 2018, Ch. 12.4)

IPT WEIGHTING OF MARGINAL STRUCTURAL MODELS

Recall that we can use a marginal structural (mean) model (MSM) to define counterfactual means of Y and their dependence on the values of X:

$$E(Y^{X=x}) = \theta_0 + \theta_1 x$$

For a binary X:

$$\begin{aligned} \text{ATE}_{X\to Y} &= E(Y^{X=1}) - E(Y^{X=0}) \\ &= (\theta_0 + \theta_1) - \theta_0 \\ &= \theta_1 \end{aligned}$$

(Hernán and Robins, 2018, Ch. 12.4)

IPT WEIGHTING OF MARGINAL STRUCTURAL MODELS

Recall that we can use a marginal structural (mean) model (MSM) to define counterfactual means of Y and their dependence on the values of X:

$$E(Y^{X=x}) = \theta_0 + \theta_1 x$$

For a binary X:

$$\begin{aligned} \text{ATE}_{X \to Y} &= E(Y^{X=1}) - E(Y^{X=0}) \\ &= (\theta_0 + \theta_1) - \theta_0 \\ &= \theta_1 \end{aligned}$$

Given the identification conditions for covariate adjustment, IPT weighting allows us to express the MSM in terms of an observable regression model in the pseudo-population, which can be estimated from (sample) data,

$$E(Y^{X=x}) = E_w(Y|X = x) = \beta_0 + \beta_1 x$$

with $\text{ATE}_{X \to Y} = \beta_1$.

(Hernán and Robins, 2018, Ch. 12.4)

THE CURSE OF DIMESIONALITY STRIKES AGAIN!

Like the number of means for standardization, the number of IPTWs to be estimated increases with the number of covariates and the number of their values, e.g.,

THE CURSE OF DIMESIONALITY STRIKES AGAIN!

Like the number of means for standardization, the number of IPTWs to be estimated increases with the number of covariates and the number of their values, e.g.,

- with binary covariate and binary treatment: $2 \times 2 = 4$ IPTWs

THE CURSE OF DIMESIONALITY STRIKES AGAIN!

Like the number of means for standardization, the number of IPTWs to be estimated increases with the number of covariates and the number of their values, e.g.,

- with binary covariate and binary treatment:
  $2 \times 2 = 4$ IPTWs
- with 1 covariate with 100 values and treatment with 100 values:
  $100 \times 100 = 10,000$ IPTWs

THE CURSE OF DIMESIONALITY STRIKES AGAIN!

Like the number of means for standardization, the number of IPTWs to be estimated increases with the number of covariates and the number of their values, e.g.,

- with binary covariate and binary treatment: $2 \times 2 = 4$ IPTWs
- with 1 covariate with 100 values and treatment with 100 values: $100 \times 100 = 10,000$ IPTWs
- with 20 binary covariates and binary treatment: $2^{20} \times 2 = 2,097,152$ IPTWs

THE CURSE OF DIMESIONALITY STRIKES AGAIN!

Like the number of means for standardization, the number of IPTWs to be estimated increases with the number of covariates and the number of their values, e.g.,

- with binary covariate and binary treatment:
  $2 \times 2 = 4$ IPTWs
- with 1 covariate with 100 values and treatment with 100 values:
  $100 \times 100 = 10,000$ IPTWs
- with 20 binary covariates and binary treatment: $2^{20} \times 2 = 2,097,152$ IPTWs

With high-dimensional and finite sample data, there'll be very few (or even no) observations for many possible combinations of x and z (i.e., violation of positivity).

THE CURSE OF DIMESIONALITY STRIKES AGAIN!

Like the number of means for standardization, the number of IPTWs to be estimated increases with the number of covariates and the number of their values, e.g.,

- with binary covariate and binary treatment: $2 \times 2 = 4$ IPTWs
- with 1 covariate with 100 values and treatment with 100 values: $100 \times 100 = 10,000$ IPTWs
- with 20 binary covariates and binary treatment: $2^{20} \times 2 = 2,097,152$ IPTWs

With high-dimensional and finite sample data, there'll be very few (or even no) observations for many possible combinations of x and z (i.e., violation of positivity).

Consequence:

THE CURSE OF DIMESIONALITY STRIKES AGAIN!

Like the number of means for standardization, the number of IPTWs to be estimated increases with the number of covariates and the number of their values, e.g.,

- with binary covariate and binary treatment: $2 \times 2 = 4$ IPTWs
- with 1 covariate with 100 values and treatment with 100 values: $100 \times 100 = 10,000$ IPTWs
- with 20 binary covariates and binary treatment: $2^{20} \times 2 = 2,097,152$ IPTWs

With high-dimensional and finite sample data, there'll be very few (or even no) observations for many possible combinations of x and z (i.e., violation of positivity).

Consequence:

- Small *n* in many combinations of x and z increases extent of random bias in IPTWs or their calculation becomes unfeasible altogether.

THE CURSE OF DIMESIONALITY STRIKES AGAIN!

Like the number of means for standardization, the number of IPTWs to be estimated increases with the number of covariates and the number of their values, e.g.,

- with binary covariate and binary treatment: $2 \times 2 = 4$ IPTWs
- with 1 covariate with 100 values and treatment with 100 values: $100 \times 100 = 10,000$ IPTWs
- with 20 binary covariates and binary treatment: $2^{20} \times 2 = 2,097,152$ IPTWs

With high-dimensional and finite sample data, there'll be very few (or even no) observations for many possible combinations of x and z (i.e., violation of positivity).

Consequence:

- Small n in many combinations of x and z increases extent of random bias in IPTWs or their calculation becomes unfeasible altogether.
- But without covariate adjustment $\rightarrow$ structural bias.

THE CURSE OF DIMESIONALITY STRIKES AGAIN!

Like the number of means for standardization, the number of IPTWs to be estimated increases with the number of covariates and the number of their values, e.g.,

- with binary covariate and binary treatment: $2 \times 2 = 4$ IPTWs
- with 1 covariate with 100 values and treatment with 100 values: $100 \times 100 = 10,000$ IPTWs
- with 20 binary covariates and binary treatment: $2^{20} \times 2 = 2,097,152$ IPTWs

With high-dimensional and finite sample data, there'll be very few (or even no) observations for many possible combinations of x and z (i.e., violation of positivity).

Consequence:

- Small n in many combinations of x and z increases extent of random bias in IPTWs or their calculation becomes unfeasible altogether.
- But without covariate adjustment $\rightarrow$ structural bias.

Meet the curse of dimensionality, again!

THE CURSE OF DIMESIONALITY STRIKES AGAIN!

Like the number of means for standardization, the number of IPTWs to be estimated increases with the number of covariates and the number of their values, e.g.,

- with binary covariate and binary treatment: $2 \times 2 = 4$ IPTWs
- with 1 covariate with 100 values and treatment with 100 values: $100 \times 100 = 10,000$ IPTWs
- with 20 binary covariates and binary treatment: $2^{20} \times 2 = 2,097,152$ IPTWs

With high-dimensional and finite sample data, there'll be very few (or even no) observations for many possible combinations of x and z (i.e., violation of positivity).

Consequence:

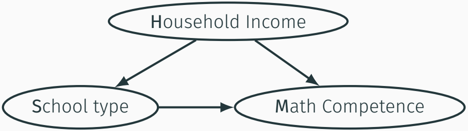
- Small *n* in many combinations of x and z increases extent of random bias in IPTWs or their calculation becomes unfeasible altogether.
- But without covariate adjustment $\rightarrow$ structural bias.

Meet the curse of dimensionality, again!

This time, however, we don't escape it by modelling E(Y|X,Z) (i.e., the outcome) but by modelling P(X=x|Z) (i.e., the treatment).

PARAMETRIC ESTIMATION OF IPT WEIGHTS

TOY DATA FOR DEMONSTRATION OF PARAMETRIC ADJUSTMENT



Data generating process for `07estte1b.dta` ($N = 100,000$):

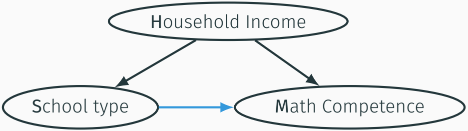
$$H = \varepsilon_H$$

$$S = 0.05H + \varepsilon_S$$

$$M = 7H - 0.3H^2 - 5S + .3SH + \varepsilon_M$$

H is continuous; S is binary (1=private; 0=public); M is continuous

TOY DATA FOR DEMONSTRATION OF PARAMETRIC ADJUSTMENT



Data generating process for `07estte1b.dta` ($N = 100,000$):

$$H = \varepsilon_H$$

$$S = 0.05H + \varepsilon_S$$

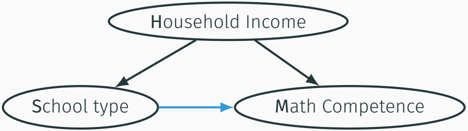
$$M = 7H - 0.3H^2 - 5S + .3SH + \varepsilon_M$$

H is continuous; S is binary (1=private; 0=public); M is continuous

Causal parameter of interest (unobservable):

$$\text{ATE}_{S \to M} = E(M^{S=\text{private}}) - E(M^{S=\text{public}}) = 44.61 - 48.61 = -4$$

TOY DATA FOR DEMONSTRATION OF PARAMETRIC ADJUSTMENT



Data generating process for `07estte1b.dta` $(N = 100,000)$:

$$H = \varepsilon_H$$

$$S = 0.05H + \varepsilon_S$$

$$M = 7H - 0.3H^2 - 5S + .3SH + \varepsilon_M$$

H is continuous; S is binary (1=private; 0=public); M is continuous

Causal parameter of interest (unobservable):

$$\text{ATE}_{S\rightarrow M} = E(M^{S=\text{private}}) - E(M^{S=\text{public}}) = 44.61 - 48.61 = -4$$

Statistical parameter, δ, that identifies $\text{ATE}_{S\rightarrow M}$:

$$\delta = \sum_h E(M|S = \text{private}, H = h)P(H = h) - \sum_h E(M|S = \text{public}, H = h)P(H = h)$$

$$= E\Big(M\frac{U_{S=\text{private}}}{P(S = \text{private}|H = h)}\Big) - E\Big(M\frac{U_{S=\text{public}}}{P(S = \text{public}|H = h)}\Big)$$

(Hernán and Robins, 2018, Part I, p. 24)

ESTIMATE $\hat{P}(S = \text{private}|H = h)$ USING LOGISTIC REGRESSION

```
. logit pschool hhinc, nohead nopval
Iteration 0:   log likelihood = -37035.778
Iteration 1:   log likelihood = -32045.528
Iteration 2:   log likelihood = -30931.093
Iteration 3:   log likelihood = -30923.622
Iteration 4:   log likelihood = -30923.618
Iteration 5:   log likelihood = -30923.618
```

pschool	Coef.	Std. Err.	[95% Conf.	Interval]
hhinc	0.60	0.01	0.59	0.61
_cons	-4.14	0.03	-4.19	-4.09

Source: `08estte2.do#2A`

These model parameters are used to estimate each unit's probability to attend private school, conditional on household income:

$$\begin{aligned}\widehat{P}(S = \text{private}|H = h) &= \frac{exp(\hat{\gamma}_0 + \hat{\gamma}_1 h)}{1 + exp(\hat{\gamma}_0 + \hat{\gamma}_1 h)} \\ &= \frac{exp(-4.14 + 0.6h)}{1 + exp(-4.14 + 0.6h)}\end{aligned}$$

(Hernán and Robins, 2018, Ch. 12.2)

CALCULATE $1/\hat{P}(S = \text{private}|H = h)$ AND $1/\hat{P}(S = \text{public}|H = h)$

```
. * predict P(X=1|Z) from model
. predict pden, pr
.
.
. * generate IPT weight
. gen iptw = .
(100,000 missing values generated)
. replace iptw = 1/pden     if pschool == 1  // for obs. with x=1
(12,173 real changes made)
. replace iptw = 1/(1-pden) if pschool == 0  // for obs. with x=0
(87,827 real changes made)
```

Source: `08estte2.do#2B`

$$\widehat{IPTW} = \frac{1}{\widehat{P}(S = s|H = h)}$$

with $\widehat{IPTW} = \frac{1}{\widehat{P}(S = \text{private}|H = h)}$ for private school students

and $\widehat{IPTW} = \frac{1}{\widehat{P}(S = \text{public}|H = h)} = \frac{1}{1 - \widehat{P}(S = \text{private}|H = h)}$ for public school students.

```
. reg mathcomp pschool [pw = iptw], nohead nopval
(sum of wgt is   1.9777e+05)
```

mathcomp	Coef.	Robust Std. Err.	[95% Conf.	Interval]
pschool	-3.78	0.15	-4.07	-3.48
_cons	47.89	0.05	47.80	47.99

Source: `08estte2.do#2C`

$$\widehat{E}_w(M|S=s) = \hat{\beta}_0 + \hat{\beta}_1 s$$

The estimates of the regression parameters can be used to estimate the ATE:

$$\begin{aligned}\hat{\delta} &= (\hat{\beta}_0 + \hat{\beta}_1) - \hat{\beta}_0 = \hat{\beta}_1 \\ &= (47.89 - 3.78) - 47.89 \\ &= -3.78\end{aligned}$$

INTERPRETATION OF ESTIMATES

statistical On average, private school students have a 3.78 [95% CI: -4.07,-3.48] points lower math competence than public school students after adjusting for differences in household income (in the population from which the sample was drawn).

counterfactual Had every student attended private school instead of public school, average math competence would have been 3.78 [95% CI: -4.07,-3.48] points lower (in the population from which the sample was drawn).
Or:
Attending private school instead of public school leads to an average decrease in math competence by 3.78 [95% CI: -4.07,-3.48] points (in the population from which the sample was drawn).

INTERPRETATION OF ESTIMATES

Assumptions for interpretation (after parametric adjustment):

statistical
- No random bias
- No measurement bias
  (see Hernán and Robins, 2018, Ch.9)
- No (treatment) model misspecification

counterfactual the previous plus
- No positivity violation
  There is (sufficient) overlap between private school students and public school students in the distribution of household income.
- No confounding bias, no overcontrol bias, no endogenous selection bias
  Adjusting for household income d-separates every noncausal path, doesn't d-separate any causal path, and doesn't d-connect any noncausal path from school type to math competence.

STABILIZED TREATMENT WEIGHTS (SIPTW)

$$\widehat{SIPTW} = \frac{\widehat{P}(X = x)}{\widehat{P}(X = x|Z)}$$

Weighting with the SIPTW as defined above also creates a pseudo-population in which X is independent of measured covariates Z.

(Hernán and Robins, 2018, Ch. 12.3)

STABILIZED TREATMENT WEIGHTS (SIPTW)

$$\widehat{SIPTW} = \frac{\widehat{P}(X = x)}{\widehat{P}(X = x|Z)}$$

Weighting with the SIPTW as defined above also creates a pseudo-population in which X is independent of measured covariates Z.

It also leads to the same estimate for the ATE as weighting with the nonstabilized IPTW from before.

(Hernán and Robins, 2018, Ch. 12.3)

STABILIZED TREATMENT WEIGHTS (SIPTW)

$$\widehat{SIPTW} = \frac{\widehat{P}(X = x)}{\widehat{P}(X = x|Z)}$$

Weighting with the SIPTW as defined above also creates a pseudo-population in which X is independent of measured covariates Z.

It also leads to the same estimate for the ATE as weighting with the nonstabilized IPTW from before.

But SIPTW have some desirable statistical properties compared to nonstabilized weights:

(Hernán and Robins, 2018, Ch. 12.3)

STABILIZED TREATMENT WEIGHTS (SIPTW)

$$\widehat{SIPTW} = \frac{\widehat{P}(X = x)}{\widehat{P}(X = x|Z)}$$

Weighting with the SIPTW as defined above also creates a pseudo-population in which X is independent of measured covariates Z.

It also leads to the same estimate for the ATE as weighting with the nonstabilized IPTW from before.

But SIPTW have some desirable statistical properties compared to nonstabilized weights:

- The pseudo-population has the same size as the original population.

(Hernán and Robins, 2018, Ch. 12.3)

STABILIZED TREATMENT WEIGHTS (SIPTW)

$$\widehat{SIPTW} = \frac{\widehat{P}(X = x)}{\widehat{P}(X = x|Z)}$$

Weighting with the SIPTW as defined above also creates a pseudo-population in which X is independent of measured covariates Z.

It also leads to the same estimate for the ATE as weighting with the nonstabilized IPTW from before.

But SIPTW have some desirable statistical properties compared to nonstabilized weights:

- The pseudo-population has the same size as the original population.
- The confidence intervals around estimates are typically narrower than for weighting with IPTW.

(Hernán and Robins, 2018, Ch. 12.3)

ESTIMATE $\hat{P}(S = \text{private})$ AND $\hat{P}(S = \text{public})$

```
. logit pschool, nohead nopvalues
Iteration 0:   log likelihood = -37035.778
Iteration 1:   log likelihood = -37035.778
------------------------------------------------------------------
     pschool |      Coef.   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
       _cons |      -1.98        0.01         -2.00        -1.96
------------------------------------------------------------------

.
.
. * predict P(X=1) from model
. predict pnum, pr
.
.
. * calculate stabilized IPTW, with P(X=0) = 1 - P(X=1)
. gen siptw = .
(100,000 missing values generated)
. replace siptw = pnum*iptw     if pschool == 1  // for obs. with x=1
(12,173 real changes made)
. replace siptw = (1-pnum)*iptw if pschool == 0  // for obs. with x=0
(87,827 real changes made)
```

Source: `08estte2.do#2D`

$$\widehat{SIPTW} = \widehat{P}(S = s) \times \widehat{IPTW}$$

USING $\widehat{SIPTW}$ LEADS TO THE SAME RESULT AS BEFORE

```
. reg mathcomp pschool [pw = siptw], robust nohead nopvalues
(sum of wgt is   9.9994e+04)
```

mathcomp	Coef.	Robust Std. Err.	[95% Conf.	Interval]
pschool	-3.78	0.15	-4.07	-3.48
_cons	47.89	0.05	47.80	47.99

Source: `08estte2.do#2E`

COMPARING THE RESULTS OF DIFFERENT ESTIMATORS

LET'S COMPARE (MOST OF) OUR ESTIMATES FROM `07estte1b.dta`

Table 1: Estimates for ATE of school type on math competence based on `07estte1b.dta`

	Unadj.	Simple RA	Flex. RA	SIPTW	SIPTW-RA
1.pschool	4.556	-4.123	-4.166	-3.778	-3.978
	(0.122)	(0.0963)	(0.108)	(0.149)	(0.117)
N	100000	100000	100000	100000	100000

Standard errors in parentheses

Source: `08estte2.do#3B`

With the exception of the estimate based on the unadjusted difference in mean math competence between private and public school students all estimates are relatively similar (around -4). Our identification analysis based on the graphical causal model told us to adjust for household income. So we prefer the estimates after adjustment and consider the unadjusted estimate structurally biased (here: confounding bias).

The estimates in the column 'SIPTW-RA' come from a regression model of M on S and H (i.e., the 'simple' model) that was additionally weighted with the SIPTW. This estimator is thus based on an outcome model *and* a treatment model and is considered 'doubly robust', because it's unbiased if either model is correctly specified (Hernán and Robins, 2018, Part II, p. 29).

WHAT SUCH A COMPARISON CAN (AND CAN'T) TELL US

Comparing the results of estimators for the ATE based on outcome models (like regression) and those based on treatment models (like IPT weighting) is valuable to detect (severe) model misspecification and should therefore be done whenever feasible.

(Hernán and Robins, 2018, Ch. 13.4, 13.5).

WHAT SUCH A COMPARISON CAN (AND CAN'T) TELL US

Comparing the results of estimators for the ATE based on outcome models (like regression) and those based on treatment models (like IPT weighting) is valuable to detect (severe) model misspecification and should therefore be done whenever feasible.

That's because both types of estimators rest on the assumption that we adjust for the *right* covariates (selected based on our causal model).

(Hernán and Robins, 2018, Ch. 13.4, 13.5).

WHAT SUCH A COMPARISON CAN (AND CAN'T) TELL US

Comparing the results of estimators for the ATE based on outcome models (like regression) and those based on treatment models (like IPT weighting) is valuable to detect (severe) model misspecification and should therefore be done whenever feasible.

That's because both types of estimators rest on the assumption that we adjust for the *right* covariates (selected based on our causal model).

Any difference in the resulting estimates must then come from misspecification of the outcome model, the treatment model, or both.

(Hernán and Robins, 2018, Ch. 13.4, 13.5).

WHAT SUCH A COMPARISON CAN (AND CAN'T) TELL US

Comparing the results of estimators for the ATE based on outcome models (like regression) and those based on treatment models (like IPT weighting) is valuable to detect (severe) model misspecification and should therefore be done whenever feasible.

That's because both types of estimators rest on the assumption that we adjust for the *right* covariates (selected based on our causal model).

Any difference in the resulting estimates must then come from misspecification of the outcome model, the treatment model, or both.

Unfortunately, in real-world applications we can never be certain which model is (in)correct (because the true data generating process is unobservable).

(Hernán and Robins, 2018, Ch. 13.4, 13.5).

WHAT SUCH A COMPARISON CAN (AND CAN'T) TELL US

Comparing the results of estimators for the ATE based on outcome models (like regression) and those based on treatment models (like IPT weighting) is valuable to detect (severe) model misspecification and should therefore be done whenever feasible.

That's because both types of estimators rest on the assumption that we adjust for the *right* covariates (selected based on our causal model).

Any difference in the resulting estimates must then come from misspecification of the outcome model, the treatment model, or both.

Unfortunately, in real-world applications we can never be certain which model is (in)correct (because the true data generating process is unobservable).

If results are similar, however, severe model misspecification in either model is unlikely and thus provides some confidence that the estimates aren't biased in that way.

(Hernán and Robins, 2018, Ch. 13.4, 13.5).

WHAT SUCH A COMPARISON CAN (AND CAN'T) TELL US

Comparing the results of estimators for the ATE based on outcome models (like regression) and those based on treatment models (like IPT weighting) is valuable to detect (severe) model misspecification and should therefore be done whenever feasible.

That's because both types of estimators rest on the assumption that we adjust for the *right* covariates (selected based on our causal model).

Any difference in the resulting estimates must then come from misspecification of the outcome model, the treatment model, or both.

Unfortunately, in real-world applications we can never be certain which model is (in)correct (because the true data generating process is unobservable).

If results are similar, however, severe model misspecification in either model is unlikely and thus provides some confidence that the estimates aren't biased in that way.

Importantly, similar results *don't* guarantee that we have adjusted for the right variables (something which can only be judged on the basis of a causal model). So similar results from different estimators don't somehow confirm that there is no structural bias.

(Hernán and Robins, 2018, Ch. 13.4, 13.5).

CHECKS FOR MODEL MISSPECIFICATION AND POSITIVITY VIOLATIONS

SOME CHECKS FOR MODEL SPECIFICATION AND POSITIVITY

There are a number of ways in which we can detect model misspecification and positivity violations (unlike for structural bias, which can only be excluded based on a theory/causal model).

SOME CHECKS FOR MODEL SPECIFICATION AND POSITIVITY

There are a number of ways in which we can detect model misspecification and positivity violations (unlike for structural bias, which can only be excluded based on a theory/causal model).

To detect model misspecification we can

SOME CHECKS FOR MODEL SPECIFICATION AND POSITIVITY

There are a number of ways in which we can detect model misspecification and positivity violations (unlike for structural bias, which can only be excluded based on a theory/causal model).

To detect model misspecification we can

- compare estimates from the same adjustment technique using simple and more complex models (this works for both outcome and treatment models),

SOME CHECKS FOR MODEL SPECIFICATION AND POSITIVITY

There are a number of ways in which we can detect model misspecification and positivity violations (unlike for structural bias, which can only be excluded based on a theory/causal model).

To detect model misspecification we can

- compare estimates from the same adjustment technique using simple and more complex models (this works for both outcome and treatment models),
- compare estimates across different adjustment techniques (in our case: regression and IPT weighting).

SOME CHECKS FOR MODEL SPECIFICATION AND POSITIVITY

There are a number of ways in which we can detect model misspecification and positivity violations (unlike for structural bias, which can only be excluded based on a theory/causal model).

To detect model misspecification we can

- compare estimates from the same adjustment technique using simple and more complex models (this works for both outcome and treatment models),
- compare estimates across different adjustment techniques (in our case: regression and IPT weighting).

To detect positivity violations we can check whether

SOME CHECKS FOR MODEL SPECIFICATION AND POSITIVITY

There are a number of ways in which we can detect model misspecification and positivity violations (unlike for structural bias, which can only be excluded based on a theory/causal model).

To detect model misspecification we can

- compare estimates from the same adjustment technique using simple and more complex models (this works for both outcome and treatment models),
- compare estimates across different adjustment techniques (in our case: regression and IPT weighting).

To detect positivity violations we can check whether

- the distribution of Z (largely) overlaps for different values of X,

SOME CHECKS FOR MODEL SPECIFICATION AND POSITIVITY

There are a number of ways in which we can detect model misspecification and positivity violations (unlike for structural bias, which can only be excluded based on a theory/causal model).

To detect model misspecification we can

- compare estimates from the same adjustment technique using simple and more complex models (this works for both outcome and treatment models),
- compare estimates across different adjustment techniques (in our case: regression and IPT weighting).

To detect positivity violations we can check whether

- the distribution of Z (largely) overlaps for different values of X,
- the distribution of $\hat{P}(X = 1|Z)$ (largely) overlaps for different values of X,

SOME CHECKS FOR MODEL SPECIFICATION AND POSITIVITY

There are a number of ways in which we can detect model misspecification and positivity violations (unlike for structural bias, which can only be excluded based on a theory/causal model).

To detect model misspecification we can

- compare estimates from the same adjustment technique using simple and more complex models (this works for both outcome and treatment models),
- compare estimates across different adjustment techniques (in our case: regression and IPT weighting).

To detect positivity violations we can check whether

- the distribution of Z (largely) overlaps for different values of X,
- the distribution of $\hat{P}(X = 1|Z)$ (largely) overlaps for different values of X,
- the mean of the $\widehat{SIPTW}$ is close to 1 and its standard deviation is (way) below 1 (deviations from this may also indicate severe misspecification of the treatment model).

CHECKING THE DISTRIBUTION OF THE $\widehat{SIPTW}$

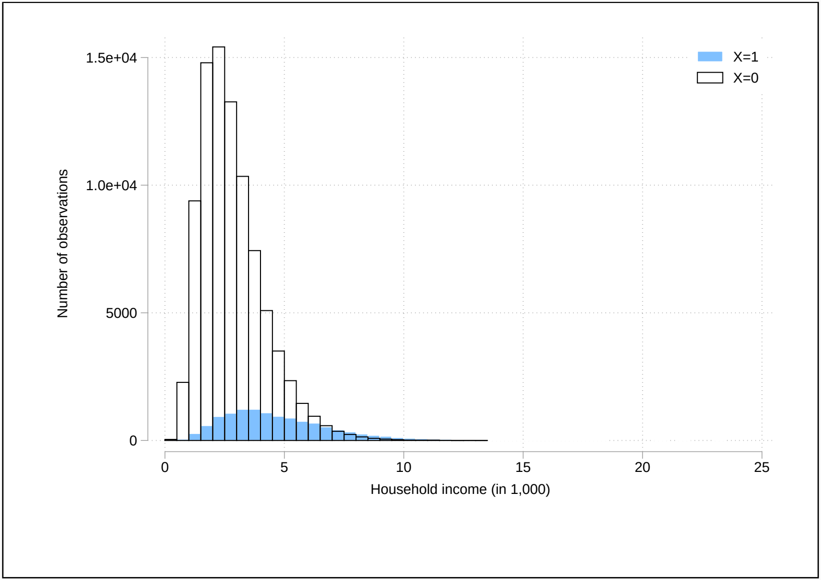
```
. sum siptw, det
                               siptw
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .1757277       .1217373
 5%     .5339589       .1217443
10%      .904268       .1217621       Obs             100,000
25%     .9173374       .1218755       Sum of Wgt.     100,000
50%     .9417001                      Mean           .9999398
                        Largest       Std. Dev.      .4615444
75%      .999531         20.916
90%     1.159085       30.12489       Variance       .2130233
95%     1.430252       34.18182       Skewness       18.96017
99%     2.676924       35.33284       Kurtosis        946.447
```

Source: `08estte2.do#4A`

Here, the mean is very close to 1 and the standard deviation is way below 1. Therefore, this test doesn't provide evidence for severe positivity violations (or misspecification of the treatment model).

If this wasn't the case, we'd need to limit the estimation to the covariate areas for which positivity holds and/or specify a more flexible treatment model (e.g., including nonlinearities and product terms between variables).

CHECKING OVERLAP OF COVARIATE DISTRIBUTIONS



Source: `08estte2.do#4B`

CHECKING OVERLAP OF COVARIATE DISTRIBUTIONS

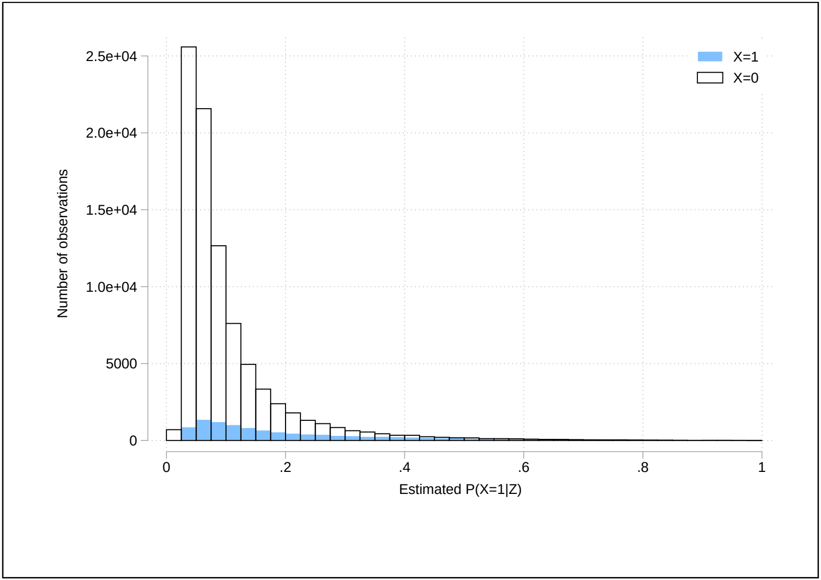
We can see that although the distribution of household income is shifted to the right for private school students, we find students from both school types across all values of income except the most extreme.

Generally, positivity is violated for areas of the covariate distribution for which there is no overlap.

Our estimates can only be generalized to areas of the covariate distribution for which there is overlap, because only these observations contribute to estimation.

In other words, for areas without overlap we can't compare the outcome for different values of X but similar values of Z.

CHECKING OVERLAP OF DISTRIBUTION OF $\hat{P}(X = 1|Z)$



Source: `08estte2.do#4C`

CHECKING OVERLAP OF DISTRIBUTION OF $\hat{P}(X = 1|Z)$

$\hat{P}(X = 1|Z)$ is a function of all covariates Z (here: only of houseold income) and can therefore be used as a summary measure of Z.

Overlap in the distribution of $\hat{P}(X = 1|Z)$ between values of X can thus also be used to assess positivity. It is particularly helpful if there are many covariates and it is tedious to check overlap for all of them separately.

Unsurprisingly, we again see that there is overlap across the distribution except for the extremes.

SUMMARY:
STRENGTHS & WEAKNESSES OF REGRESSION AND IPT WEIGHTING

	Regression	IPT weighting

SUMMARY:
STRENGTHS & WEAKNESSES OF REGRESSION AND IPT WEIGHTING

Regression	IPT weighting
Narrower CI (i.e., more efficient)	Wider CI (i.e., less efficient)

SUMMARY: STRENGTHS & WEAKNESSES OF REGRESSION AND IPT WEIGHTING

Regression	IPT weighting
Narrower CI (i.e., more efficient)	Wider CI (i.e., less efficient)
Rather robust to positivity violations	Quite sensitive to positivity violations

SUMMARY:
STRENGTHS & WEAKNESSES OF REGRESSION AND IPT WEIGHTING

Regression	IPT weighting
Narrower CI (i.e., more efficient)	Wider CI (i.e., less efficient)
Rather robust to positivity violations	Quite sensitive to positivity violations
Widely inaccurate extrapolation possible	Widely inaccurate extrapolation unlikely

SUMMARY: STRENGTHS & WEAKNESSES OF REGRESSION AND IPT WEIGHTING

Regression	IPT weighting
Narrower CI (i.e., more efficient)	Wider CI (i.e., less efficient)
Rather robust to positivity violations	Quite sensitive to positivity violations
Widely inaccurate extrapolation possible	Widely inaccurate extrapolation unlikely
Sensitive to misspecification of the outcome model	Sensitive to misspecification of the treatment model

SUMMARY: STRENGTHS & WEAKNESSES OF REGRESSION AND IPT WEIGHTING

Regression	IPT weighting
Narrower CI (i.e., more efficient)	Wider CI (i.e., less efficient)
Rather robust to positivity violations	Quite sensitive to positivity violations
Widely inaccurate extrapolation possible	Widely inaccurate extrapolation unlikely
Sensitive to misspecification of the outcome model	Sensitive to misspecification of the treatment model
No routine positivity check	Routine positivity check
Can handle continuous treatments	Can only handle categorical treatments well

SUMMARY:
STRENGTHS & WEAKNESSES OF REGRESSION AND IPT WEIGHTING

Regression	IPT weighting
Narrower CI (i.e., more efficient)	Wider CI (i.e., less efficient)
Rather robust to positivity violations	Quite sensitive to positivity violations
Widely inaccurate extrapolation possible	Widely inaccurate extrapolation unlikely
Sensitive to misspecification of the outcome model	Sensitive to misspecification of the treatment model
No routine positivity check	Routine positivity check
Can handle continuous treatments	Can only handle categorical treatments well
Can't handle more complex settings e.g., time-varying treatments	Can handle more complex settings e.g., time-varying treatments

SUMMARY:
"HOW SERIOUSLY DO WE TAKE OUR ESTIMATES?"

Effect estimates can only be taken as seriously as the underlying assumptions.

(Hernán and Robins, 2018, Ch. 13.5)

SUMMARY: “HOW SERIOUSLY DO WE TAKE OUR ESTIMATES?”

Effect estimates can only be taken as seriously as the underlying assumptions.

If any assumption is (severly) violated, our estimates may differ (largely) from the true effect.

(Hernán and Robins, 2018, Ch. 13.5)

SUMMARY: "HOW SERIOUSLY DO WE TAKE OUR ESTIMATES?"

Effect estimates can only be taken as seriously as the underlying assumptions.

If any assumption is (severly) violated, our estimates may differ (largely) from the true effect.

The task of the analyst is to ensure that all assumptions hold (at least approximately).

(Hernán and Robins, 2018, Ch. 13.5)

SUMMARY:
"HOW SERIOUSLY DO WE TAKE OUR ESTIMATES?"

Effect estimates can only be taken as seriously as the underlying assumptions.

If any assumption is (severly) violated, our estimates may differ (largely) from the true effect.

The task of the analyst is to ensure that all assumptions hold (at least approximately).

Scientific discussion should systematically assess the plausibility of each assumption.

(Hernán and Robins, 2018, Ch. 13.5)

SUMMARY: “HOW SERIOUSLY DO WE TAKE OUR ESTIMATES?”

Effect estimates can only be taken as seriously as the underlying assumptions.

If any assumption is (severly) violated, our estimates may differ (largely) from the true effect.

The task of the analyst is to ensure that all assumptions hold (at least approximately).

Scientific discussion should systematically assess the plausibility of each assumption.

On the other hand, criticism of estimates merely stating that assumptions *may* be violated is insufficient and uninformative.

(Hernán and Robins, 2018, Ch. 13.5)

SUMMARY: "HOW SERIOUSLY DO WE TAKE OUR ESTIMATES?"

Effect estimates can only be taken as seriously as the underlying assumptions.

If any assumption is (severly) violated, our estimates may differ (largely) from the true effect.

The task of the analyst is to ensure that all assumptions hold (at least approximately).

Scientific discussion should systematically assess the plausibility of each assumption.

On the other hand, criticism of estimates merely stating that assumptions *may* be violated is insufficient and uninformative.

Any criticism, therefore, is required to outline a concrete reason for a (severe) violation (e.g., mention of a specific potential confounder not adjusted for in the analysis).

(Hernán and Robins, 2018, Ch. 13.5)

SUMMARY: IDENTIFICATION VS. ESTIMATION

For identification, the number of observations is of no concern (assumption of infinite population data).

(Hernán and Robins, 2018, Ch. 10.1)

SUMMARY: IDENTIFICATION VS. ESTIMATION

For identification, the number of observations is of no concern (assumption of infinite population data).

Therefore, large samples (even “big data”) can’t save identification if there is *confounding bias*, *overcontrol bias*, or *endogenous selection bias*.

(Hernán and Robins, 2018, Ch. 10.1)

SUMMARY: IDENTIFICATION VS. ESTIMATION

For identification, the number of observations is of no concern (assumption of infinite population data).

Therefore, large samples (even “big data”) can’t save identification if there is *confounding bias*, *overcontrol bias*, or *endogenous selection bias*.

Estimation, in contrast, becomes easier with more observations, because large samples reduce

(Hernán and Robins, 2018, Ch. 10.1)

SUMMARY: IDENTIFICATION VS. ESTIMATION

For identification, the number of observations is of no concern (assumption of infinite population data).

Therefore, large samples (even “big data”) can’t save identification if there is *confounding bias*, *overcontrol bias*, or *endogenous selection bias*.

Estimation, in contrast, becomes easier with more observations, because large samples reduce

- the extent of *random bias*,

(Hernán and Robins, 2018, Ch. 10.1)

SUMMARY: IDENTIFICATION VS. ESTIMATION

For identification, the number of observations is of no concern (assumption of infinite population data).

Therefore, large samples (even "big data") can't save identification if there is *confounding bias*, *overcontrol bias*, or *endogenous selection bias*.

Estimation, in contrast, becomes easier with more observations, because large samples reduce

- the extent of *random bias*,
- the risk of *misspecification* of functional form of associations between variables (when using parametric models).

(Hernán and Robins, 2018, Ch. 10.1)

SUMMARY:
IDENTIFICATION VS. ESTIMATION

For identification, the number of observations is of no concern (assumption of infinite population data).

Therefore, large samples (even "big data") can't save identification if there is *confounding bias*, *overcontrol bias*, or *endogenous selection bias*.

Estimation, in contrast, becomes easier with more observations, because large samples reduce

- the extent of *random bias*,
- the risk of *misspecification* of functional form of associations between variables (when using parametric models).

We use statistical inference (mainly confidence intervals) to quantify uncertainty due to random bias.

(Hernán and Robins, 2018, Ch. 10.1)

SUMMARY: IDENTIFICATION VS. ESTIMATION

For identification, the number of observations is of no concern (assumption of infinite population data).

Therefore, large samples (even "big data") can't save identification if there is *confounding bias*, *overcontrol bias*, or *endogenous selection bias*.

Estimation, in contrast, becomes easier with more observations, because large samples reduce

- the extent of *random bias*,
- the risk of *misspecification* of functional form of associations between variables (when using parametric models).

We use statistical inference (mainly confidence intervals) to quantify uncertainty due to random bias.

Flexible modelling and comparison of estimators can be helpful to assess misspecification of estimation models.

(Hernán and Robins, 2018, Ch. 10.1)

NEXT WEEK: DEFINITION AND IDENTIFICATION OF DIRECT & INDIRECT EFFECTS

1. Terminology for mediation analysis
2. Counterfactual definition of direct and indirect effects
3. Graphical identification of direct and indirect effects
4. Treatment-induced mediator-outcome confounding

THANK YOU FOR YOUR ATTENTION!

REFERENCES

Hernán, M. A. and Robins, J. M. (2018). *Causal Inference (v. 04-10-17)*. Chapman & Hall/CRC, Boca Raton, FL. URL `http://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/`.

Get these slides from

`https://osf.io/jgrsy/`

The slides were created using the METROPOLIS theme by Matthias Vogelgesang available at

`https://github.com/matze/mtheme`