INTRODUCTION TO CAUSAL INFERENCE

Michael Kühhirt

November 20, 2017

v.2017-11-23-11:44am

Estimation of total effects I

LAST WEEK: IDENTIFICATION OF TOTAL EFFECTS BY COVARIATE ADJUSTMENT

- model based selection of adjustment variables Z
- identification of average total effect by adjusted mean difference:

$$\begin{aligned}\mathrm{ATE}_{X \to Y} &= E(Y^x) - E(Y^{x'}) \\ &= \sum_z E(Y|X = x, Z = z)P(Z = z) - \sum_z E(Y|X = x', Z = z)P(Z = z)\end{aligned}$$

- estimation: calculation of numerical value for mean difference (with finite sample data)

IDENTIFICATION CONDITIONS FOR TOTAL EFFECTS

1. d-separation of noncausal paths and d-connection of causal paths between X and Y (based on plausible causal model)
    - no confounding bias
    - no endogenous selection bias
    - no overcontrol bias
2. positivity (can be checked empirically)

$$P(X = x|Z = z) > 0 \text{ for all } z \text{ with } P(Z = z) > 0$$

Violation of any of these conditions leads to failed identification.

TODAY: ESTIMATION OF TOTAL EFFECTS USING REGRESSION

1. Nonparametric covariate adjustment
2. The curse of dimensionality
3. Parametric regression for covariate adjustment
4. Statistical significance and confidence intervals

ROADMAP FOR CAUSAL INFERENCE

1. Specify the causal model
2. Define the causal parameter of interest
   (along with the target population)
3. Link the causal model to the available empirical data
4. Assess whether the causal parameter of interest can be identified with the available data and define the respective statistical parameter
5. Specify the statistical model used to estimate the statistical parameter
6. Estimate the statistical parameter
7. Interpret the results and discuss assumptions

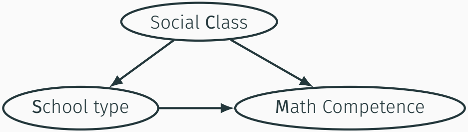
(Petersen and van der Laan, 2014)

ROADMAP FOR CAUSAL INFERENCE

1. Specify the causal model
2. Define the causal parameter of interest
   (along with the target population)
3. Link the causal model to the available empirical data
4. Assess whether the causal parameter of interest can be identified with the available data and define the respective statistical parameter
5. Specify the statistical model used to estimate the statistical parameter
6. Estimate the statistical parameter
7. Interpret the results and discuss assumptions

(Petersen and van der Laan, 2014)

NONPARAMETRIC COVARIATE ADJUSTMENT

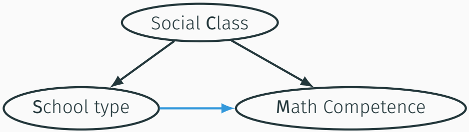


Data generating process for `07estte1a.dta` ($N = 100,000$):

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

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` ($N = 100,000$):

$$C = \varepsilon_C$$

$$S = 0.2C + \varepsilon_S$$

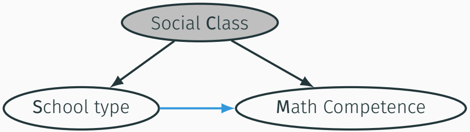
$$M = 20C - 5S + 5SC + \varepsilon_M$$

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):

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

Statistical parameter that identifies $\text{ATE}_{S \to M}$:

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

ESTIMATION WITH SAMPLE DATA: TERMINOLOGY

estimand parameter of interest, δ, in the target population

Here: *difference in mean math competence between private school students and public school students, adjusted for social class*

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

(Hernán and Robins, 2018, Ch. 11.1)

ESTIMATION WITH SAMPLE DATA: TERMINOLOGY

estimand parameter of interest, δ, in the target population

Here: *difference in mean math competence between private school students and public school students, adjusted for social class*

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

estimator mathematical rule that produces (estimates!) numerical value for estimand, $\hat{\delta}$, from sample data of size N

Here: *difference in sample averages of math competence between private school students and public school students, adjusted for social class*

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

$$= \sum_c \left(\frac{1}{n_{\text{private};c}} \sum_u^{U_{\text{private};c}} m_u \right) \frac{n_c}{N} - \sum_c \left(\frac{1}{n_{\text{public};c}} \sum_u^{U_{\text{public};c}} m_u \right) \frac{n_c}{N}$$

estimate numerical value, $\hat{\delta}$, produced by estimator with specific sample

- $\hat{\delta}$ unlikely to exactly equal estimand because of random bias
  random bias quantifiable through confidence intervals
- in case of no structural bias: $E(\hat{\delta}) = \delta$

(Hernán and Robins, 2018, Ch. 11.1)

ESTIMATION WITH SAMPLE DATA: TERMINOLOGY

estimand parameter of interest, δ, in the target population
Here: *difference in mean math competence between private school students and public school students, adjusted for social class*

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

estimator mathematical rule that produces (estimates!) numerical value for estimand, $\hat{\delta}$, from sample data of size N
Here: *difference in sample averages of math competence between private school students and public school students, adjusted for social class*

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

$$= \sum_c \left(\frac{1}{n_{\text{private};c}} \sum_u^{U_{\text{private};c}} m_u \right) \frac{n_c}{N} - \sum_c \left(\frac{1}{n_{\text{public};c}} \sum_u^{U_{\text{public};c}} m_u \right) \frac{n_c}{N}$$

estimate numerical value, $\hat{\delta}$, produced by estimator with specific sample

- $\hat{\delta}$ unlikely to exactly equal estimand because of random bias
- random bias quantifiable through confidence intervals
- in case of no structural bias: $E(\hat{\delta}) = \delta$

(Hernán and Robins, 2018, Ch. 11.1)

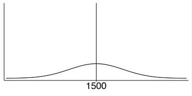
Target population
$N = 80,000,000$
Estimand: $\delta = 1,500$

Sample A
$n = 1,000$
Estimate A:
$\hat{\delta}_{n=1,000} = 1,150$

Sample B
$n = 10,000$
Estimate B:
$\hat{\delta}_{n=10,000} = 1,650$

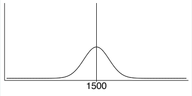
Sample C
$n = 1,000,000$
Estimate C:
$\hat{\delta}_{n=1,000,000} = 1,505$

Distribution of estimates after drawing ∞ samples with $n = 1,000$:



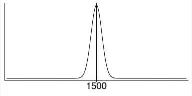
$E[\hat{\delta}_n] = \delta$

Distribution of estimates after drawing ∞ samples with $n = 10,000$:



$E[\hat{\delta}_n] = \delta$

Distribution of estimates after drawing ∞ samples with $n = 1,000,000$:



$E[\hat{\delta}_n] = \delta$

UNADJUSTED MEAN DIFFERENCE DOESN'T IDENTIFY ATE

$$E(M^{\text{private}}) - E(M^{\text{public}}) \neq E(M|S = \text{private}) - E(M|S = \text{public})$$

```
. tabstat mathcomp, by(pschool) notot
Summary for variables: mathcomp
     by categories of: pschool
 pschool |       mean
---------+-----------
       0 |   48.26859
       1 |   50.69592
---------+-----------

. reg mathcomp pschool, nohead nopvalues
------------------------------------------------------------------
    mathcomp |      Coef.   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
     pschool |       2.43        0.11          2.20         2.65
       _cons |      48.27        0.04         48.18        48.35
------------------------------------------------------------------
```

Source: `07estte1.do#2`

Average math competence is 2.43 points higher for private school students than for public school students.

NONPARAMETRIC COVARIATE ADJUSTMENT: STANDARDIZATION ESTIMATOR

$$\hat{\delta} = \sum_{c} \left(\frac{1}{n_{\text{private};c}} \sum_{u}^{U_{\text{private};c}} m_u \right) \frac{n_c}{N} - \sum_{c} \left(\frac{1}{n_{\text{public};c}} \sum_{u}^{U_{\text{public};c}} m_u \right) \frac{n_c}{N}$$

1. Divide observations into two groups: S=private and S=public.
2. Further divide both groups into two groups: C=lower class and C=higher class.
3. Within groups S×C, calculate average M.
4. Multiply (i.e., weight) each average with sample proportion of C.
5. Sum up weighted averages to adjusted mean within S=private and S=public.
6. Calculate difference between both adjusted means.
7. Result: estimate of $ATE_{S \rightarrow M}$.

(Hernán and Robins, 2018, Ch. 2.3)

NONPARAMETRIC COVARIATE ADJUSTMENT: STANDARDIZATION ESTIMATOR

$$\hat{\delta} = \sum_{c} \left(\frac{1}{n_{\text{private};c}} \sum_{u}^{U_{\text{private};c}} m_u \right) \frac{n_c}{N} - \sum_{c} \left(\frac{1}{n_{\text{public};c}} \sum_{u}^{U_{\text{public};c}} m_u \right) \frac{n_c}{N}$$

1. Divide observations into two groups: S=private and S=public.
2. Further divide both groups into two groups: C=lower class and C=higher class.
3. Within groups S×C, calculate average M.
4. Multiply (i.e., weight) each average with sample proportion of C.
5. Sum up weighted averages to adjusted mean within S=private and S=public.
6. Calculate difference between both adjusted means.
7. Result: estimate of $ATE_{S \to M}$.

(Hernán and Robins, 2018, Ch. 2.3)

NONPARAMETRIC COVARIATE ADJUSTMENT: STANDARDIZATION ESTIMATOR

$$\hat{\delta} = \sum_{c} \Big(\frac{1}{n_{\text{private};c}} \sum_{u}^{U_{\text{private};c}} m_u \Big) \frac{n_c}{N} - \sum_{c} \Big(\frac{1}{n_{\text{public};c}} \sum_{u}^{U_{\text{public};c}} m_u \Big) \frac{n_c}{N}$$

1. Divide observations into two groups: S=private and S=public.
2. Further divide both groups into two groups: C=lower class and C=higher class.
3. Within groups s×c, calculate average M.
4. Multiply (i.e., weight) each average with sample proportion of c.
5. Sum up weighted averages to adjusted mean within S=private and S=public.
6. Calculate difference between both adjusted means.
7. Result: estimate of ATE

(Hernán and Robins, 2018, Ch. 2.3)

NONPARAMETRIC COVARIATE ADJUSTMENT: STANDARDIZATION ESTIMATOR

$$\hat{\delta} = \sum_{c} \Big(\frac{1}{n_{\text{private};c}} \sum_{u}^{U_{\text{private};c}} m_u \Big) \frac{n_c}{N} - \sum_{c} \Big(\frac{1}{n_{\text{public};c}} \sum_{u}^{U_{\text{public};c}} m_u \Big) \frac{n_c}{N}$$

1. Divide observations into two groups: S=private and S=public.
2. Further divide both groups into two groups: C=lower class and C=higher class.
3. Within groups s×c, calculate average M.
4. Multiply (i.e., weight) each average with sample proportion of c.
5. Sum up weighted averages to adjusted mean within S=private and S=public.
6. Calculate difference between both adjusted means.
7. Result: estimate of $ATE_{S \rightarrow M}$

(Hernán and Robins, 2018, Ch. 2.3)

NONPARAMETRIC COVARIATE ADJUSTMENT: STANDARDIZATION ESTIMATOR

$$\hat{\delta} = \sum_{c} \left(\frac{1}{n_{\text{private};c}} \sum_{u}^{U_{\text{private};c}} m_u \right) \frac{n_c}{N} - \sum_{c} \left(\frac{1}{n_{\text{public};c}} \sum_{u}^{U_{\text{public};c}} m_u \right) \frac{n_c}{N}$$

1. Divide observations into two groups: S=private and S=public.
2. Further divide both groups into two groups: C=lower class and C=higher class.
3. Within groups s×c, calculate average M.
4. Multiply (i.e., weight) each average with sample proportion of c.
5. Sum up weighted averages to adjusted mean within S=private and S=public.
6. Calculate difference between both adjusted means.
7. Result: estimate of $\text{ATE}_{S,M}$.

(Hernán and Robins, 2018, Ch. 2.3)

NONPARAMETRIC COVARIATE ADJUSTMENT: STANDARDIZATION ESTIMATOR

$$\hat{\delta} = \sum_{c} \left(\frac{1}{n_{\text{private};c}} \sum_{u}^{U_{\text{private};c}} m_u \right) \frac{n_c}{N} - \sum_{c} \left(\frac{1}{n_{\text{public};c}} \sum_{u}^{U_{\text{public};c}} m_u \right) \frac{n_c}{N}$$

1. Divide observations into two groups: S=private and S=public.
2. Further divide both groups into two groups: C=lower class and C=higher class.
3. Within groups s×c, calculate average M.
4. Multiply (i.e., weight) each average with sample proportion of c.
5. Sum up weighted averages to adjusted mean within S=private and S=public.
6. Calculate difference between both adjusted means.
7. Result: estimate of $ATE_{S \to M}$.

(Hernán and Robins, 2018, Ch. 2.3)

NONPARAMETRIC COVARIATE ADJUSTMENT: STANDARDIZATION ESTIMATOR

$$\hat{\delta} = \sum_{c} \Big(\frac{1}{n_{\text{private};c}} \sum_{u}^{U_{\text{private};c}} m_u \Big) \frac{n_c}{N} - \sum_{c} \Big(\frac{1}{n_{\text{public};c}} \sum_{u}^{U_{\text{public};c}} m_u \Big) \frac{n_c}{N}$$

1. Divide observations into two groups: S=private and S=public.
2. Further divide both groups into two groups: C=lower class and C=higher class.
3. Within groups s×c, calculate average M.
4. Multiply (i.e., weight) each average with sample proportion of c.
5. Sum up weighted averages to adjusted mean within S=private and S=public.
6. Calculate difference between both adjusted means.
7. Result: estimate of $\text{ATE}_{S \to M}$.

(Hernán and Robins, 2018, Ch. 2.3)

$\hat{P}(C = c)$ AND $\sum_c \hat{E}(M|S = \text{private}, C = c)\hat{P}(C = c)$

```
. tab hiclass

   hiclass |      Freq.     Percent        Cum.
-----------+-----------------------------------
         0 |     79,782       79.78       79.78
         1 |     20,218       20.22      100.00
-----------+-----------------------------------
     Total |    100,000      100.00

. tabstat mathcomp if pschool==1, by(hiclass) notot

Summary for variables: mathcomp
     by categories of: hiclass

 hiclass |      mean
---------+----------
       0 |  40.10014
       1 |  64.86853
--------------------

. disp 64.86853 * .20218 + 40.10014 * .79782
45.107813
```


Source: 07estte1.do#3A

$\sum_c \hat{E}(M|S = \text{public}, C = c)\hat{P}(C = c)$ AND $\hat{\delta}$

```
. tabstat mathcomp if pschool==0, by(hiclass) notot

Summary for variables: mathcomp
     by categories of: hiclass

 hiclass |      mean
---------+----------
       0 |   44.9853
       1 |  64.88624
--------------------

. disp 64.88624 * .20218 + 44.9853 * .79782
49.008872

. disp 45.107813 - 49.008872
-3.901059
```

Source: `07estte1.do#3B`

ALTERNATIVE NONPARAMETRIC COVARIATE ADJUSTMENT: SATURATED REGRESSION ESTIMATOR

1. Estimate regression of M on C (dummy) in two groups by S:

$$\hat{E}(M|S = \text{private}, C = c) = \hat{\phi}_0 + \hat{\phi}_1 c$$
$$\hat{E}(M|S = \text{public}, C = c) = \hat{\gamma}_0 + \hat{\gamma}_1 c$$

2. Insert sample average for C in both models.
3. Calculate difference between predicted averages:

$$\hat{\delta} = [\hat{\phi}_0 + \hat{\phi}_1 \hat{E}(C)] - [\hat{\gamma}_0 + \hat{\gamma}_1 \hat{E}(C)]$$

4. Result: estimate of $\text{ATE}_{S \rightarrow M}$

(Hernán and Robins, 2018, Ch. 11.3)

ALTERNATIVE NONPARAMETRIC COVARIATE ADJUSTMENT: SATURATED REGRESSION ESTIMATOR

1. Estimate regression of M on C (dummy) in two groups by S:

$$\hat{E}(M|S = \text{private}, C = c) = \hat{\phi}_0 + \hat{\phi}_1 c$$
$$\hat{E}(M|S = \text{public}, C = c) = \hat{\gamma}_0 + \hat{\gamma}_1 c$$

2. Insert sample average for C in both models.
3. Calculate difference between predicted averages:

$$\hat{\delta} = [\hat{\phi}_0 + \hat{\phi}_1 \hat{E}(C)] - [\hat{\gamma}_0 + \hat{\gamma}_1 \hat{E}(C)]$$

4. Result: estimate of $ATE_{S \to M}$

(Hernán and Robins, 2018, Ch. 11.3)

ALTERNATIVE NONPARAMETRIC COVARIATE ADJUSTMENT: SATURATED REGRESSION ESTIMATOR

1. Estimate regression of M on C (dummy) in two groups by S:

$$\hat{E}(M|S = \text{private}, C = c) = \hat{\phi}_0 + \hat{\phi}_1 c$$
$$\hat{E}(M|S = \text{public}, C = c) = \hat{\gamma}_0 + \hat{\gamma}_1 c$$

2. Insert sample average for C in both models.
3. Calculate difference between predicted averages:

$$\hat{\delta} = \left[\hat{\phi}_0 + \hat{\phi}_1 \hat{E}(C)\right] - \left[\hat{\gamma}_0 + \hat{\gamma}_1 \hat{E}(C)\right]$$

4. Result: estimate of $ATE_{S \to M}$

(Hernán and Robins, 2018, Ch. 11.3)

ALTERNATIVE NONPARAMETRIC COVARIATE ADJUSTMENT: SATURATED REGRESSION ESTIMATOR

1. Estimate regression of M on C (dummy) in two groups by S:

$$\hat{E}(M|S = \text{private}, C = c) = \hat{\phi}_0 + \hat{\phi}_1 c$$
$$\hat{E}(M|S = \text{public}, C = c) = \hat{\gamma}_0 + \hat{\gamma}_1 c$$

2. Insert sample average for C in both models.
3. Calculate difference between predicted averages:

$$\hat{\delta} = \left[\hat{\phi}_0 + \hat{\phi}_1 \hat{E}(C)\right] - \left[\hat{\gamma}_0 + \hat{\gamma}_1 \hat{E}(C)\right]$$

4. Result: estimate of $\text{ATE}_{S \to M}$

(Hernán and Robins, 2018, Ch. 11.3)

$\sum_c \hat{E}(M|S = \textbf{private}, C = c)\hat{P}(C = c)$

```
. reg mathcomp hiclass if pschool==1, nohead nopval
------------------------------------------------------------
    mathcomp |     Coef.   Std. Err.      [95% Conf. Interval]
-------------+----------------------------------------------
     hiclass |     24.77       0.16         24.45        25.09
       _cons |     40.10       0.11         39.89        40.31
------------------------------------------------------------

. lincom _b[_cons] + _b[hiclass] * .20218, nopval
 ( 1)  .20218*hiclass + _cons = 0
------------------------------------------------------------
    mathcomp |     Coef.   Std. Err.      [95% Conf. Interval]
-------------+----------------------------------------------
         (1) |     45.11       0.09         44.94        45.28
------------------------------------------------------------
```

Source: `07estte1.do#4A`

$$\hat{\delta} = 45.11 - 49.01 = -3.9$$

$\sum_c \hat{E}(M|S = \textbf{public}, C = c)\hat{P}(C = c)$

```
. reg mathcomp hiclass if pschool==0, nohead nopval
------------------------------------------------------------------------
    mathcomp |      Coef.   Std. Err.      [95% Conf. Interval]
-------------+----------------------------------------------------------
     hiclass |      19.90       0.09           19.73       20.07
       _cons |      44.99       0.04           44.92       45.05
------------------------------------------------------------------------

. lincom _b[_cons] + _b[hiclass] * .20218, nopval
 ( 1)  .20218*hiclass + _cons = 0
------------------------------------------------------------------------
    mathcomp |      Coef.   Std. Err.      [95% Conf. Interval]
-------------+----------------------------------------------------------
         (1) |      49.01       0.03           48.95       49.07
------------------------------------------------------------------------
```

Source: `07estte1.do#4B`

$$\hat{\delta} = 45.11 - 49.01 = -3.9$$

INTERPRETATION OF ESTIMATES (OF AVERAGE TOTAL EFFECTS)

Respective interpretations for our estimates:

statistical *On average, private school students have a 3.9 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 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 (OF AVERAGE TOTAL EFFECTS)

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.

THE CURSE OF DIMENSIONALITY

The number of means to be estimated increases with the number of covariates and the number of their values, e.g.,

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

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 adjusted means or their calculation becomes unfeasible altogether
- But without covariate adjustment → structural bias

Meet the curse of dimensionality!

(Hernán and Robins, 2018, Ch. 10.5)

COVARIATE ADJUSTMENT WITH HIGH-DIMENSIONAL DATA

The number of means to be estimated increases with the number of covariates and the number of their values, e.g.,

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

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 adjusted means or their calculation becomes impossible altogether
- But without covariate adjustment → structural bias

Meet the curse of dimensionality!

(Hernán and Robins, 2018, Ch. 10.5)

The number of means to be estimated increases with the number of covariates and the number of their values, e.g.,

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

(Hernán and Robins, 2018, Ch. 10.5)

COVARIATE ADJUSTMENT WITH HIGH-DIMENSIONAL DATA

The number of means to be estimated increases with the number of covariates and the number of their values, e.g.,

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

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 adjusted means or their calculation becomes unfeasible altogether
- But without covariate adjustment $\rightarrow$ structural bias

Meet the curse of dimensionality!

(Hernán and Robins, 2018, Ch. 10.5)

COVARIATE ADJUSTMENT WITH HIGH-DIMENSIONAL DATA

The number of means to be estimated increases with the number of covariates and the number of their values, e.g.,

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

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 adjusted means or their calculation becomes unfeasible altogether
- But without covariate adjustment $\rightarrow$ structural bias

Meet the curse of dimensionality!

(Hernán and Robins, 2018, Ch. 10.5)

The number of means to be estimated increases with the number of covariates and the number of their values, e.g.,

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

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 adjusted means or their calculation becomes unfeasible altogether
- But without covariate adjustment → structural bias

Meet the curse of dimensionality!

(Hernán and Robins, 2018, Ch. 10.5)

COVARIATE ADJUSTMENT WITH HIGH-DIMENSIONAL DATA

The number of means to be estimated increases with the number of covariates and the number of their values, e.g.,

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

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 adjusted means or their calculation becomes unfeasible altogether.
- But without covariate adjustment → structural bias.

Meet the curse of dimensionality!

(Hernán and Robins, 2018, Ch. 10.5)

COVARIATE ADJUSTMENT WITH HIGH-DIMENSIONAL DATA

The number of means to be estimated increases with the number of covariates and the number of their values, e.g.,

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

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 adjusted means or their calculation becomes unfeasible altogether.
- But without covariate adjustment → structural bias.

Meet the curse of dimensionality!

(Hernán and Robins, 2018, Ch. 10.5)

COVARIATE ADJUSTMENT WITH HIGH-DIMENSIONAL DATA

The number of means to be estimated increases with the number of covariates and the number of their values, e.g.,

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

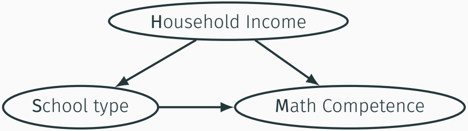
With high-dimensional and finite sample data, there'll be very few (or even no) observations for many possible combinations of x and z (i.e., violation of positivity).

Consequence:

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

Meet the curse of dimensionality!

(Hernán and Robins, 2018, Ch. 10.5)



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

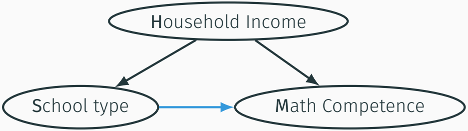
$$H = \varepsilon_H$$

$$S = 0.05H + \varepsilon_S$$

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

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` ($N = 100,000$):

$$H = \varepsilon_H$$

$$S = 0.05H + \varepsilon_S$$

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

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

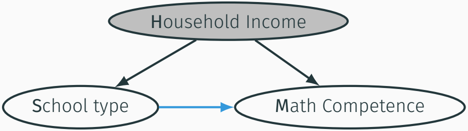
Causal parameter of interest (unobservable):

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

Statistical parameter that identifies $\text{ATE}_{S\to M}$

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

TOY DATA FOR DEMONSTRATION OF PARAMETRIC ADJUSTMENT



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

$$H = \varepsilon_H$$
$$S = 0.05H + \varepsilon_S$$
$$M = 7 * H - 0.3H^2 - 5S + .3SH + \varepsilon_M$$

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

Causal parameter of interest (unobservable):

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

Statistical parameter that identifies $\text{ATE}_{S\to M}$:

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

SAMPLE DISTRIBUTION OF HOUSEHOLD INCOME

```
 . sum hhinc, det
                                 hhinc
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .8485717       .3078159
 5%     1.183861       .3112431
10%     1.418475       .3463157       Obs             100,000
25%     1.924917       .3726511       Sum of Wgt.     100,000
50%     2.709498                      Mean           3.064972
                        Largest       Std. Dev.      1.632335
75%     3.786846       18.11737
90%     5.148115       20.63799       Variance       2.664517
95%     6.161319       21.98266       Skewness        1.70071
99%     8.685948       23.11648       Kurtosis       8.277186
. capture noisily tab hhinc
too many values
```

Source: `07estte1.do#6A`

Too many values to produce frequency table.

VALUES OF SCHOOL TYPE AT LOWER END OF INCOME DISTRIBUTION

```
. tab pschool hhinc if hhinc<.373, col

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

           |                     hhinc
   pschool |  .3078159   .3112431   .3463157   .3726511 |     Total
-----------+--------------------------------------------+----------
         0 |         1          1          1          1 |         4
           |    100.00     100.00     100.00     100.00 |    100.00
-----------+--------------------------------------------+----------
     Total |         1          1          1          1 |         4
           |    100.00     100.00     100.00     100.00 |    100.00
```

Source: `07estte1.do#6B`

Positivity violation: $\hat{P}(S = \text{private}|H = h) = 0$, for some h

VALUES OF SCHOOL TYPE AT HIGHER END OF INCOME DISTRIBUTION

```
. tab pschool hhinc if hhinc>18, col

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

           |                      hhinc
   pschool |  18.11737   20.63799   21.98266   23.11648 |     Total
-----------+--------------------------------------------+----------
         1 |         1          1          1          1 |         4 
           |    100.00     100.00     100.00     100.00 |    100.00 
-----------+--------------------------------------------+----------
     Total |         1          1          1          1 |         4 
           |    100.00     100.00     100.00     100.00 |    100.00 
```

Source: `07estte1.do#6B`

Positivity violation: $\hat{P}(S = \text{public}|H = h) = 0$, for some h

CAN WE ESCAPE THE CURSE OF DIMENSIONALITY?

Unfortunately, there is only an incomplete, technical escape by specifying low-dimensional statistical models to *predict* adjusted means.

These models allow the calculation of all adjusted means based on much fewer model parameters.

This comes at the cost of potential bias from misspecified statistical models.

In case of model misspecification, estimates can be systematically biased even if we adjust for all relevant covariates.

There are different types of statistical models to estimate causal effects:

1. outcome models (e.g., regression adjustment),
2. treatment models (e.g., inverse probability of treatment weighting),
3. doubly robust models (e.g., IPT-weighted regression adjustment)

(Hernán and Robins, 2019, Ch. 10.5)

Unfortunately, there is only an incomplete, technical escape by specifying low-dimensional statistical models to *predict* adjusted means.

These models allow the calculation of all adjusted means based on much fewer model parameters.

This comes at the cost of potential bias from misspecified statistical models.

In case of model misspecification, estimates can be systematically biased even if we adjust for all relevant covariates.

There are different types of statistical models to estimate causal effects:

1. outcome models (e.g., regression adjustment),
2. treatment models (e.g., inverse probability of treatment weighting),
3. doubly robust models (e.g., IPT-weighted regression adjustment)

(Hernán and Robins, 2018, Ch. 10.5)

CAN WE ESCAPE THE CURSE OF DIMENSIONALITY?

Unfortunately, there is only an incomplete, technical escape by specifying low-dimensional statistical models to *predict* adjusted means.

These models allow the calculation of all adjusted means based on much fewer model parameters.

This comes at the cost of potential bias from misspecified statistical models.

In case of model misspecification, estimates can be systematically biased even if we adjust for all relevant covariates.

There are different types of statistical models to estimate causal effects:

1. outcome models (e.g., regression adjustment),
2. treatment models (e.g., inverse probability of treatment weighting),
3. doubly robust models (e.g., IPT-weighted regression adjustment).

(Hernán and Robins, 2018, Ch. 10.9)

CAN WE ESCAPE THE CURSE OF DIMENSIONALITY?

Unfortunately, there is only an incomplete, technical escape by specifying low-dimensional statistical models to *predict* adjusted means.

These models allow the calculation of all adjusted means based on much fewer model parameters.

This comes at the cost of potential bias from misspecified statistical models.

In case of model misspecification, estimates can be systematically biased even if we adjust for all relevant covariates.

There are different types of statistical models to estimate causal effects:

1. outcome models (e.g., regression adjustment)
2. treatment models (e.g., inverse probability of treatment weighting)
3. doubly robust models (e.g., IPT-weighted regression adjustment)

(Hernán and Robins, 2018, Ch. 10.5)

CAN WE ESCAPE THE CURSE OF DIMENSIONALITY?

Unfortunately, there is only an incomplete, technical escape by specifying low-dimensional statistical models to *predict* adjusted means.

These models allow the calculation of all adjusted means based on much fewer model parameters.

This comes at the cost of potential bias from misspecified statistical models.

In case of model misspecification, estimates can be systematically biased even if we adjust for all relevant covariates.

There are different types of statistical models to estimate causal effects:

1. outcome models (e.g., regression adjustment),
2. treatment models (e.g., inverse probability of treatment weighting),
3. doubly robust models (e.g., IPT-weighted regression adjustment).

(Hernán and Robins, 2018, Ch. 10.5)

CAN WE ESCAPE THE CURSE OF DIMENSIONALITY?

Unfortunately, there is only an incomplete, technical escape by specifying low-dimensional statistical models to *predict* adjusted means.

These models allow the calculation of all adjusted means based on much fewer model parameters.

This comes at the cost of potential bias from misspecified statistical models.

In case of model misspecification, estimates can be systematically biased even if we adjust for all relevant covariates.

There are different types of statistical models to estimate causal effects:

1. outcome models (e.g., regression adjustment),
2. treatment models (e.g., inverse probability of treatment weighting),
3. doubly robust models (e.g., IPT-weighted regression adjustment).

(Hernán and Robins, 2018, Ch. 10.5)

CAN WE ESCAPE THE CURSE OF DIMENSIONALITY?

Unfortunately, there is only an incomplete, technical escape by specifying low-dimensional statistical models to *predict* adjusted means.

These models allow the calculation of all adjusted means based on much fewer model parameters.

This comes at the cost of potential bias from misspecified statistical models.

In case of model misspecification, estimates can be systematically biased even if we adjust for all relevant covariates.

There are different types of statistical models to estimate causal effects:

1. outcome models (e.g., regression adjustment),
2. treatment models (e.g., inverse probability of treatment weighting),
3. doubly robust models (e.g., IPT-weighted regression adjustment).

(Hernán and Robins, 2018, Ch. 10.5)

CAN WE ESCAPE THE CURSE OF DIMENSIONALITY?

Unfortunately, there is only an incomplete, technical escape by specifying low-dimensional statistical models to *predict* adjusted means.

These models allow the calculation of all adjusted means based on much fewer model parameters.

This comes at the cost of potential bias from misspecified statistical models.

In case of model misspecification, estimates can be systematically biased even if we adjust for all relevant covariates.

There are different types of statistical models to estimate causal effects:

1. outcome models (e.g., regression adjustment),
2. treatment models (e.g., inverse probability of treatment weighting),
3. doubly robust models (e.g., IPT-weighted regression adjustment).

(Hernán and Robins, 2018, Ch. 10.5)

PARAMETRIC REGRESSION FOR COVARIATE ADJUSTMENT

PARAMETRIC COVARIATE ADJUSTMENT: REGRESSION ESTIMATOR

1. Estimate regression of M on H in two groups of S:

$$\hat{E}(M|S = \text{private}, H = h) = \hat{\phi}_0 + \hat{\phi}_1 h$$
$$\hat{E}(M|S = \text{public}, H = h) = \hat{\gamma}_0 + \hat{\gamma}_1 h$$

2. Insert sample average of H for h in both models.
3. Calculate difference between predicted averages

$$\hat{\delta} = [\hat{\phi}_0 + \hat{\phi}_1 \hat{E}(H)] - [\hat{\gamma}_0 + \hat{\gamma}_1 \hat{E}(H)]$$

4. Result: estimate of $\text{ATE}_{S \to M}$

(Hernán and Robins, 2018, Ch. 15.1)

Here: estimate is biased, because the true association between H and M is nonlinear (model misspecification).

PARAMETRIC COVARIATE ADJUSTMENT: REGRESSION ESTIMATOR

1. Estimate regression of M on H in two groups of S:

$$\hat{E}(M|S = \text{private}, H = h) = \hat{\phi}_0 + \hat{\phi}_1 h$$
$$\hat{E}(M|S = \text{public}, H = h) = \hat{\gamma}_0 + \hat{\gamma}_1 h$$

2. Insert sample average of H for h in both models.
3. Calculate difference between predicted averages:

$$\hat{\delta} = [\hat{\phi}_0 + \hat{\phi}_1 \hat{E}(H)] - [\hat{\gamma}_0 + \hat{\gamma}_1 \hat{E}(H)]$$

4. Result: estimate of $ATE_{S \rightarrow M}$

(Hernán and Robins, 2018, Ch. 15.1)

Here: estimate is biased, because the true association between H and M is nonlinear (model misspecification).

PARAMETRIC COVARIATE ADJUSTMENT: REGRESSION ESTIMATOR

1. Estimate regression of M on H in two groups of S:

$$\hat{E}(M|S = \text{private}, H = h) = \hat{\phi}_0 + \hat{\phi}_1 h$$
$$\hat{E}(M|S = \text{public}, H = h) = \hat{\gamma}_0 + \hat{\gamma}_1 h$$

2. Insert sample average of H for h in both models.
3. Calculate difference between predicted averages:

$$\hat{\delta} = \left[\hat{\phi}_0 + \hat{\phi}_1 \hat{E}(H)\right] - \left[\hat{\gamma}_0 + \hat{\gamma}_1 \hat{E}(H)\right]$$

4. Result: estimate of ATE

(Hernán and Robins, 2018, Ch. 15.1)

Here: estimate is biased, because the true association between H and M is nonlinear (model misspecification).

PARAMETRIC COVARIATE ADJUSTMENT: REGRESSION ESTIMATOR

1. Estimate regression of M on H in two groups of S:

$$\hat{E}(M|S = \text{private}, H = h) = \hat{\phi}_0 + \hat{\phi}_1 h$$
$$\hat{E}(M|S = \text{public}, H = h) = \hat{\gamma}_0 + \hat{\gamma}_1 h$$

2. Insert sample average of H for h in both models.
3. Calculate difference between predicted averages:

$$\hat{\delta} = \left[\hat{\phi}_0 + \hat{\phi}_1 \hat{E}(H)\right] - \left[\hat{\gamma}_0 + \hat{\gamma}_1 \hat{E}(H)\right]$$

4. Result: estimate of $\text{ATE}_{S \rightarrow M}$

(Hernán and Robins, 2018, Ch. 15.1)

Here: estimate is biased, because the true association between H and M is nonlinear (model misspecification).

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

```
. reg mathcomp hhinc if pschool==1, nohead nopval
─────────────┬──────────────────────────────────────────────────
    mathcomp │       Coef.   Std. Err.      [95% Conf. Interval]
─────────────┼──────────────────────────────────────────────────
       hhinc │        3.71        0.04           3.64       3.78
       _cons │       33.73        0.19          33.35      34.10
─────────────┴──────────────────────────────────────────────────

. lincom _b[_cons]+_b[hhinc]*3.065, nopval
 ( 1)  3.065*hhinc + _cons = 0
─────────────┬──────────────────────────────────────────────────
    mathcomp │       Coef.   Std. Err.      [95% Conf. Interval]
─────────────┼──────────────────────────────────────────────────
         (1) │       45.10        0.11          44.89      45.30
─────────────┴──────────────────────────────────────────────────
```

Source: `07estte1.do#7A`

$\sum_h \hat{E}(M|S = \textbf{public}, H = h)\hat{P}(H = h)$

```
. reg mathcomp hhinc if pschool==0, nohead nopval
------------------------------------------------------------
    mathcomp |      Coef.   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
       hhinc |       4.81       0.02          4.77        4.86
       _cons |      33.23       0.07         33.09       33.37
------------------------------------------------------------

. lincom _b[_cons]+_b[hhinc]*3.065, nopval
 ( 1)  3.065*hhinc + _cons = 0
------------------------------------------------------------
    mathcomp |      Coef.   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
         (1) |      47.98       0.03         47.92       48.04
------------------------------------------------------------
```

Source: `07estte1.do#7B`

$$\hat{\delta} = 45.1 - 47.98 = -2.88$$

$\sum_h \hat{E}(M|S = \textbf{private}, H = h)\hat{P}(H = h)$

With correct model specification, including quadratic term for H:

```
. reg mathcomp c.hhinc##c.hhinc if pschool==1, nohead nopval
------------------------------------------------------------------
       mathcomp |      Coef.   Std. Err.     [95% Conf. Interval]
----------------+-------------------------------------------------
          hhinc |       7.46       0.11          7.24        7.68
                |
c.hhinc#c.hhinc |      -0.31       0.01         -0.32       -0.29
                |
          _cons |      24.49       0.32         23.87       25.11
------------------------------------------------------------------

. lincom _b[_cons] + _b[hhinc] * 3.065 + _b[c.hhinc#c.hhinc] * 3.065 * 3.065, nopv
> al
 ( 1)  3.065*hhinc + 9.394225*c.hhinc#c.hhinc + _cons = 0
---------------------------------------------------------------
    mathcomp |      Coef.   Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------
         (1) |      44.48       0.10         44.29       44.68
---------------------------------------------------------------
```

Source: `07estte1.do#8A`

$\sum_h \hat{E}(M|S = \textbf{public}, H = h)\hat{P}(H = h)$

With correct model specification, including quadratic term for H:

```
. reg mathcomp c.hhinc##c.hhinc if pschool==0, nohead nopval
------------------------------------------------------------------
        mathcomp |      Coef.   Std. Err.     [95% Conf. Interval]
-----------------+------------------------------------------------
           hhinc |       7.10        0.08          6.95       7.25
                 |
 c.hhinc#c.hhinc |      -0.31        0.01         -0.33      -0.29
                 |
           _cons |      29.84        0.13         29.59      30.09
------------------------------------------------------------------

. lincom _b[_cons] + _b[hhinc] * 3.065 + _b[c.hhinc#c.hhinc] * 3.065 * 3.065, nopv
> al
 ( 1)  3.065*hhinc + 9.394225*c.hhinc#c.hhinc + _cons = 0
---------------------------------------------------------------
    mathcomp |      Coef.   Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------
         (1) |      48.65        0.04         48.58      48.72
---------------------------------------------------------------
```

Source: `07estte1.do#8B`

$$\hat{\delta} = 44.48 - 48.65 = -4.17$$

To achieve the same result with a single model, we need to include product terms for X (here: S) and all covariates Z (here: H).

1. Estimate (correctly specified) regression of M on S and H:

$$\hat{E}(M|S = s, H = h) = \hat{\beta}_0 + \hat{\beta}_1 s + \hat{\beta}_2 sh + \hat{\beta}_3 sh^2 + \hat{\beta}_4 h + \hat{\beta}_5 h^2$$

(Hernán and Robins, 2018, Ch. 15.1)

SINGLE REGRESSION MODEL

To achieve the same result with a single model, we need to include product terms for X (here: S) and all covariates Z (here: H).

1. Estimate (correctly specified) regression of M on S and H:

$$\hat{E}(M|S=s, H=h) = \hat{\beta}_0 + \hat{\beta}_1 s + \hat{\beta}_2 sh + \hat{\beta}_3 sh^2 + \hat{\beta}_4 h + \hat{\beta}_5 h^2$$

2. Insert sample average of H for h for calculation of averages for M:

$$\begin{aligned}\hat{\delta} &= \left[\hat{\beta}_0 + \hat{\beta}_1 + \beta_2\hat{E}(H) + \hat{\beta}_3\hat{E}(H)^2 + \hat{\beta}_4\hat{E}(H) + \hat{\beta}_5\hat{E}(H)^2\right] \\ &\quad - \left[\hat{\beta}_0 + \hat{\beta}_4\hat{E}(H) + \hat{\beta}_5\hat{E}(H)^2\right] \\ &= \hat{\beta}_1 + \beta_2\hat{E}(H) + \beta_3\hat{E}(H)^2\end{aligned}$$

3. Result: estimate of $ATE_{S \to M}$

(Hernán and Robins, 2018, Ch. 15.1)

SINGLE REGRESSION MODEL

To achieve the same result with a single model, we need to include product terms for X (here: S) and all covariates Z (here: H).

1. Estimate (correctly specified) regression of M on S and H:

$$\hat{E}(M|S = s, H = h) = \hat{\beta}_0 + \hat{\beta}_1 s + \hat{\beta}_2 sh + \hat{\beta}_3 sh^2 + \hat{\beta}_4 h + \hat{\beta}_5 h^2$$

2. Insert sample average of H for h for calculation of averages for M:

$$\begin{aligned}\hat{\delta} &= \left[\hat{\beta}_0 + \hat{\beta}_1 + \beta_2 \hat{E}(H) + \hat{\beta}_3 \hat{E}(H)^2 + \hat{\beta}_4 \hat{E}(H) + \hat{\beta}_5 \hat{E}(H)^2\right] \\ &- \left[\hat{\beta}_0 + \hat{\beta}_4 \hat{E}(H) + \hat{\beta}_5 \hat{E}(H)^2\right] \\ &= \hat{\beta}_1 + \beta_2 \hat{E}(H) + \beta_3 \hat{E}(H)^2\end{aligned}$$

3. Result: estimate of $\text{ATE}_{S \to M}$

(Hernán and Robins, 2018, Ch. 15.1)

$\hat{E}(M|S = s, H = h) = \hat{\beta}_0 + \hat{\beta}_1 s + \hat{\beta}_2 sh + \hat{\beta}_3 sh^2 + \hat{\beta}_4 h + \hat{\beta}_5 h^2$

```
. reg mathcomp i.pschool##c.hhinc##c.hhinc, nohead nopval
```

mathcomp	Coef.	Std. Err.	[95% Conf.	Interval]
pschool				
0	0.00	(base)		
1	-5.35	0.34	-6.03	-4.68
hhinc	7.10	0.08	6.95	7.25
pschool#c.hhinc				
1	0.36	0.14	0.10	0.63
c.hhinc#c.hhinc	-0.31	0.01	-0.33	-0.29
pschool#c.hhinc#c.hhinc				
1	0.01	0.01	-0.02	0.03
_cons	29.84	0.13	29.59	30.09

Source: `07estte1.do#9A`

$\hat{\delta} = \hat{\beta}_1 + \beta_2 \hat{E}(H) + \beta_3 \hat{E}(H)^2$

```
. lincom _b[1.pschool] + _b[1.pschool#c.hhinc] * 3.065    ///
>      + _b[1.pschool#c.hhinc#c.hhinc] * 3.065 * 3.065, nopval
 ( 1)  1.pschool + 3.065*1.pschool#c.hhinc + 9.394225*1.pschool#c.hhinc#c.hhinc =
       0
------------------------------------------------------------------
    mathcomp |       Coef.   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
         (1) |       -4.17        0.11          -4.38       -3.95
------------------------------------------------------------------
```

Source: `07estte1.do#9A`

INTERPRETATION OF ESTIMATES

statistical On average, private school students have a 4.17 [95% CI: -4.38,-3.95] 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 4.17 [95% CI: -4.38,-3.95] 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 4.17 [95% CI: -4.38,-3.95] 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 model misspecification (new!)

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
  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.

Model without nonlinearities and product terms:

1. Estimate (misspecified) regression of M on S and H:

$$\hat{E}(M|S = s, H = h) = \hat{\beta}_0 + \hat{\beta}_1 s + \hat{\beta}_2 h$$

2. Insert sample average of H for h for calculation of averages for M:

$$\begin{aligned}\hat{\delta} &= [\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 \hat{E}(H)] - [\hat{\beta}_0 + \hat{\beta}_2 \hat{E}(H)] \\ &= \hat{\beta}_1\end{aligned}$$

3. Result: estimate of $ATE_{S \to M}$

Here: only small bias due to model misspecification.

SIMPLEST MODEL

Model without nonlinearities and product terms:

1. Estimate (misspecified) regression of M on S and H:

$$\hat{E}(M|S = s, H = h) = \hat{\beta}_0 + \hat{\beta}_1 s + \hat{\beta}_2 h$$

2. Insert sample average of H for h for calculation of averages for M:

$$\begin{aligned}\hat{\delta} &= [\hat{\beta}_0 + \hat{\beta}_1 + \beta_2 \hat{E}(H)] - [\hat{\beta}_0 + \hat{\beta}_2 \hat{E}(H)] \\ &= \hat{\beta}_1\end{aligned}$$

3. Result: estimate of $ATE_{S \to M}$

Here: only small bias due to model misspecification.

SIMPLEST MODEL

Model without nonlinearities and product terms:

1. Estimate (misspecified) regression of M on S and H:

$$\hat{E}(M|S=s,H=h)=\hat{\beta}_0+\hat{\beta}_1 s+\hat{\beta}_2 h$$

2. Insert sample average of H for h for calculation of averages for M:

$$\begin{aligned}\hat{\delta}&=\left[\hat{\beta}_0+\hat{\beta}_1+\beta_2\hat{E}(H)\right]-\left[\hat{\beta}_0+\hat{\beta}_2\hat{E}(H)\right]\\&=\hat{\beta}_1\end{aligned}$$

3. Result: estimate of $\text{ATE}_{S\rightarrow M}$

Here: only small bias due to model misspecification.

$\hat{E}(M|S = S, H = H) = \hat{\beta}_0 + \hat{\beta}_1 S + \hat{\beta}_2 H$

```
. reg mathcomp pschool hhinc, nohead nopval
```

mathcomp	Coef.	Std. Err.	[95% Conf.	Interval]
pschool	-4.12	0.10	-4.31	-3.93
hhinc	4.48	0.02	4.44	4.52
_cons	34.17	0.06	34.04	34.29

Source: `07estte1.do#9B`

$$\hat{\delta} = \hat{\beta}_1 = -4.12$$

MISSPECIFIED SIMPLE MODEL CAN BE SUBSTANTIALLY BIASED

Simple model with the first toy data, `07estte1a.dta`, in which social class moderated the effect of school type:

```
. reg mathcomp pschool hiclass, nohead nopval
─────────────────────────────────────────────────────────────
    mathcomp │      Coef.   Std. Err.     [95% Conf. Interval]
─────────────┼───────────────────────────────────────────────
     pschool │      -3.09        0.09         -3.27      -2.92
     hiclass │      21.00        0.08         20.85      21.15
       _cons │      44.80        0.03         44.73      44.87
─────────────────────────────────────────────────────────────
```

Source: `07estte1.do#10A`

True ATE was -4.

CORRECTLY SPECIFIED MODEL FOR FIRST TOY DATA

This model includes a product term for school type and social class:

```
. reg mathcomp i.pschool##i.hiclass, nohead nopval
------------------------------------------------------------------------
        mathcomp |      Coef.   Std. Err.      [95% Conf. Interval]
-----------------+------------------------------------------------------
         pschool |
               0 |       0.00  (base)
               1 |      -4.89        0.11          -5.10        -4.67
                 |
         hiclass |
               0 |       0.00  (base)
               1 |      19.90        0.09          19.73        20.07
                 |
 pschool#hiclass |
             1 1 |       4.87        0.18           4.51         5.23
                 |
           _cons |      44.99        0.04          44.92        45.05
------------------------------------------------------------------------
```

Source: `07estte1.do#10B`

CORRECTLY SPECIFIED MODEL FOR FIRST TOY DATA

This model includes a product term for school type and social class:

```
. lincom _b[1.pschool] + _b[1.pschool#1.hiclass] * .20218, nopval
( 1)  1.pschool + .20218*1.pschool#1.hiclass = 0
```

mathcomp	Coef.	Std. Err.	[95% Conf.	Interval]
(1)	-3.90	0.09	-4.08	-3.72

Source: `07estte1.do#10B`

$$\hat{\delta} = \hat{\beta}_1 + \hat{\beta}_2 \hat{E}(C) = -3.9$$

With finite samples, parametric modelling is generally unavoidable (curse of dimensionality).

Downside: misspecified parametric models lead to biased estimates (even if these models adjust for all relevant covariates).

To minimize misspecification bias, it's better to use flexible models (with more parameters, e.g., polynomials, splines, product terms) that can adapt to the data.

Trade-off: additional model parameters usually lead to wider confidence intervals (i.e., more uncertainty due to random bias).

Pragmatic solutions:

- compare estimates from simple and more flexible models,
- compare estimates from different estimators (next week),
- test for model misspecification and positivity violations (next week).

(Hernán and Robins, 2018, Ch. 11.5.)

MODEL SPECIFICATION AND THE BIAS-VARIANCE-TRADEOFF

With finite samples, parametric modelling is generally unavoidable (curse of dimensionality).

Downside: misspecified parametric models lead to biased estimates (even if these models adjust for all relevant covariates).

To minimize misspecification bias, it's better to use flexible models (with more parameters, e.g. polynomials, splines, product terms) that can adapt to the data.

Trade-off: additional model parameters usually lead to wider confidence intervals (i.e., more uncertainty due to random bias).

Pragmatic solutions:

- compare estimates from simple and more flexible models,
- compare estimates from different estimators (next week),
- test for model misspecification and positivity violations (next week).

(Hernán and Robins, 2018, Ch. 11.5.)

MODEL SPECIFICATION AND THE BIAS-VARIANCE-TRADEOFF

With finite samples, parametric modelling is generally unavoidable (curse of dimensionality).

Downside: misspecified parametric models lead to biased estimates (even if these models adjust for all relevant covariates).

To minimize misspecification bias, it's better to use flexible models (with more parameters, e.g, polynomials, splines, product terms) that can adapt to the data.

Trade-off: additional model parameters usually lead to wider confidence intervals (i.e., more uncertainty due to random bias).

Pragmatic solutions:

- compare estimates from simple and more flexible models,
- compare estimates from different estimators (next week),
- test for model misspecification and positivity violations (next week).

(Hernán and Robins, 2018, Ch. 11.5.)

MODEL SPECIFICATION AND THE BIAS-VARIANCE-TRADEOFF

With finite samples, parametric modelling is generally unavoidable (curse of dimensionality).

Downside: misspecified parametric models lead to biased estimates (even if these models adjust for all relevant covariates).

To minimize misspecification bias, it's better to use flexible models (with more parameters, e.g, polynomials, splines, product terms) that can adapt to the data.

Trade-off: additional model parameters usually lead to wider confidence intervals (i.e., more uncertainty due to random bias).

Pragmatic solutions:

- compare estimates from simple and more flexible models;
- compare estimates from different estimators (next week);
- test for model misspecification and positivity violations (next week).

(Hernán and Robins, 2018, Ch. 11.5.)

MODEL SPECIFICATION AND THE BIAS-VARIANCE-TRADEOFF

With finite samples, parametric modelling is generally unavoidable (curse of dimensionality).

Downside: misspecified parametric models lead to biased estimates (even if these models adjust for all relevant covariates).

To minimize misspecification bias, it's better to use flexible models (with more parameters, e.g, polynomials, splines, product terms) that can adapt to the data.

Trade-off: additional model parameters usually lead to wider confidence intervals (i.e., more uncertainty due to random bias).

Pragmatic solutions:

- compare estimates from simple and more flexible models;
- compare estimates from different estimators (next week);
- test for model misspecification and positivity violations (next week).

(Hernán and Robins, 2018, Ch. 11.5.)

MODEL SPECIFICATION AND THE BIAS-VARIANCE-TRADEOFF

With finite samples, parametric modelling is generally unavoidable (curse of dimensionality).

Downside: misspecified parametric models lead to biased estimates (even if these models adjust for all relevant covariates).

To minimize misspecification bias, it's better to use flexible models (with more parameters, e.g, polynomials, splines, product terms) that can adapt to the data.

Trade-off: additional model parameters usually lead to wider confidence intervals (i.e., more uncertainty due to random bias).

Pragmatic solutions:

- compare estimates from simple and more flexible models,
- compare estimates from different estimators (next week),
- test for model misspecification and positivity violations (next week).

(Hernán and Robins, 2018, Ch. 11.5.)

MODEL SPECIFICATION AND THE BIAS-VARIANCE-TRADEOFF

With finite samples, parametric modelling is generally unavoidable (curse of dimensionality).

Downside: misspecified parametric models lead to biased estimates (even if these models adjust for all relevant covariates).

To minimize misspecification bias, it's better to use flexible models (with more parameters, e.g, polynomials, splines, product terms) that can adapt to the data.

Trade-off: additional model parameters usually lead to wider confidence intervals (i.e., more uncertainty due to random bias).

Pragmatic solutions:

- compare estimates from simple and more flexible models,
- compare estimates from different estimators (next week),
- test for model misspecification and positivity violations (next week).

(Hernán and Robins, 2018, Ch. 11.5.)

MODEL SPECIFICATION AND THE BIAS-VARIANCE-TRADEOFF

With finite samples, parametric modelling is generally unavoidable (curse of dimensionality).

Downside: misspecified parametric models lead to biased estimates (even if these models adjust for all relevant covariates).

To minimize misspecification bias, it's better to use flexible models (with more parameters, e.g, polynomials, splines, product terms) that can adapt to the data.

Trade-off: additional model parameters usually lead to wider confidence intervals (i.e., more uncertainty due to random bias).

Pragmatic solutions:

- compare estimates from simple and more flexible models,
- compare estimates from different estimators (next week),
- test for model misspecification and positivity violations (next week).

(Hernán and Robins, 2018, Ch. 11.5.)

STATISTICAL SIGNIFICANCE AND CONFIDENCE INTERVALS

COMMON MISCONCEPTIONS ABOUT STATISTICAL SIGNIFICANCE

1. "Significant" regression coefficient as evidence for causal effect
2. "Nonsignificant" regression coefficient as evidence for absence of causal effect
3. Statistical significance (low p-value) interpreted as measure for substantive significance of association or effect

(Shalizi, 2016, Ch. 2.4)

The heuristic, $p < 0.05 \rightarrow$ "importance/truth", isn't very useful for accumulating scientific knowledge and informing real-world decisions.
(see Statement of the American Statistical Association on Statistical Significance and P-Values)

The standard p-value (stars in regression tables) provides the probability that a given sample estimate (e.g., a regression coefficient) is drawn, when the estimand is really 0 [p-value $= P(\hat{\delta}|\delta = 0)$].

If the estimand is different from zero, this standard p-value becomes completely uninformative, because it only pertains to a scenario in which $\delta = 0$.

Furthermore, for the same numerical estimate, the p-value decreases ("becomes more significant") as the sample size increases.

Therefore, even *substantively* significant (i.e., large) estimates can be statistically insignificant in small samples and *substantively* insignificant (i.e., tiny) estimates can be statistically significant in large samples.

But, of course, the importance/strength of some effect is independent from the number of observations we decided to sample.

For gauging the strength of an effect (once it is assumed to be identified), we need to focus on the effect size and use confidence intervals as a measure of random variability over samples (of size n).

THE MEANING OF P-VALUES

The standard p-value (stars in regression tables) provides the probability that a given sample estimate (e.g., a regression coefficient) is drawn, when the estimand is really 0 [p-value $= P(\hat{\delta}|\delta = 0)$].

If the estimand is different from zero, this standard p-value becomes completely uninformative, because it only pertains to a scenario in which $\delta = 0$.

Furthermore, for the same numerical estimate, the p-value decreases ("becomes more significant") as the sample size increases.

Therefore, even *substantively* significant (i.e., large) estimates can be statistically insignificant in small samples and *substantively* insignificant (i.e., tiny) estimates can be statistically significant in large samples.

But, of course, the importance/strength of some effect is independent from the number of observations we decided to sample.

For gauging the strength of an effect (once it is assumed to be identified), we need to focus on the effect size and use confidence intervals as a measure of random variability over samples (of size n).

THE MEANING OF P-VALUES

The standard p-value (stars in regression tables) provides the probability that a given sample estimate (e.g., a regression coefficient) is drawn, when the estimand is really 0 [p-value $= P(\hat{\delta}|\delta = 0)$].

If the estimand is different from zero, this standard p-value becomes completely uninformative, because it only pertains to a scenario in which $\delta = 0$.

Furthermore, for the same numerical estimate, the p-value decreases ("becomes more significant") as the sample size increases.

Therefore, even substantively significant (i.e., large) estimates can be statistically insignificant in small samples and substantively insignificant (e.g., tiny) estimates can be statistically significant in large samples.

But of course, the importance/strength of some effect is independent from the number of observations we decided to sample.

For gauging the strength of an effect (once it is assumed to be identified), we need to focus on the effect size and use confidence intervals as a measure of random variability over samples (of size n).

THE MEANING OF P-VALUES

The standard p-value (stars in regression tables) provides the probability that a given sample estimate (e.g., a regression coefficient) is drawn, when the estimand is really 0 [p-value $= P(\hat{\delta}|\delta = 0)$].

If the estimand is different from zero, this standard p-value becomes completely uninformative, because it only pertains to a scenario in which $\delta = 0$.

Furthermore, for the same numerical estimate, the p-value decreases (“becomes more significant”) as the sample size increases.

Therefore, even *substantively* significant (i.e., large) estimates can be statistically insignificant in small samples and *substantively* insignificant (i.e., tiny) estimates can be statistically significant in large samples.

But, of course, the importance/strength of some effect is independent from the number of observations we decided to sample.

For gauging the strength of an effect (once it is assumed to be identified), we need to focus on the effect size and use confidence intervals as a measure of random variability over samples (of size n).

THE MEANING OF P-VALUES

The standard p-value (stars in regression tables) provides the probability that a given sample estimate (e.g., a regression coefficient) is drawn, when the estimand is really 0 [p-value $= P(\hat{\delta}|\delta = 0)$].

If the estimand is different from zero, this standard p-value becomes completely uninformative, because it only pertains to a scenario in which $\delta = 0$.

Furthermore, for the same numerical estimate, the p-value decreases (“becomes more significant”) as the sample size increases.

Therefore, even *substantively* significant (i.e., large) estimates can be statistically insignificant in small samples and *substantively* insignificant (i.e., tiny) estimates can be statistically significant in large samples.

But, of course, the importance/strength of some effect is independent from the number of observations we decided to sample.

For gauging the strength of an effect (once it is assumed to be identified), we need to focus on the effect size and use confidence intervals as a measure of random variability over samples (of size n).

THE MEANING OF P-VALUES

The standard p-value (stars in regression tables) provides the probability that a given sample estimate (e.g., a regression coefficient) is drawn, when the estimand is really 0 [p-value $= P(\hat{\delta}|\delta = 0)$].

If the estimand is different from zero, this standard p-value becomes completely uninformative, because it only pertains to a scenario in which $\delta = 0$.

Furthermore, for the same numerical estimate, the p-value decreases ("becomes more significant") as the sample size increases.

Therefore, even *substantively* significant (i.e., large) estimates can be statistically insignificant in small samples and *substantively* insignificant (i.e., tiny) estimates can be statistically significant in large samples.

But, of course, the importance/strength of some effect is independent from the number of observations we decided to sample.

For gauging the strength of an effect (once it is assumed to be identified!), we need to focus on the effect size and use confidence intervals as a measure of random variability over samples (of size n).

CONFIDENCE INTERVALS

CIs quantify uncertainty of the estimation due to random differences of samples from the target population.

Wider CIs signal greater uncertainty due to random bias.

Interpretation: The 95% CI covers the estimand in 95 of 100 samples of size n.

Whether the CI in a given sample covers the estimand remains unknown (because we could have one of the 5% of samples in which the CI doesn't cover the estimand).

When estimating causal effects: In case of violation of the identification conditions and model misspecification the probability that the 95% CI covers the true effect may be less than 95%.

When bias is severe, this probability may approach 0.

CIs quantify uncertainty of the estimation due to random differences of samples from the target population.

Wider CIs signal greater uncertainty due to random bias.

CONFIDENCE INTERVALS

CIs quantify uncertainty of the estimation due to random differences of samples from the target population.

Wider CIs signal greater uncertainty due to random bias.

Interpretation: The 95% CI covers the estimand in 95 of 100 samples of size n.

Whether the CI in a given sample covers the estimand remains unknown (because we could have one of the 5% of samples in which the CI doesn't cover the estimand).

When estimating causal effects: In case of violation of the identification conditions and model misspecification, the probability that the 95% CI covers the true effect may be less than 95%.

When bias is severe, this probability may approach 0.

CONFIDENCE INTERVALS

CIs quantify uncertainty of the estimation due to random differences of samples from the target population.

Wider CIs signal greater uncertainty due to random bias.

Interpretation: The 95% CI covers the estimand in 95 of 100 samples of size n.

Whether the CI in a given sample covers the estimand remains unknown (because we could have one of the 5% of samples in which the CI doesn’t cover the estimand).

When estimating causal effects, in case of violation of the identification conditions and model misspecification, the probability that the 95% CI covers the true effect may be less than 95%.

When bias is severe, this probability may approach 0.

CONFIDENCE INTERVALS

CIs quantify uncertainty of the estimation due to random differences of samples from the target population.

Wider CIs signal greater uncertainty due to random bias.

Interpretation: The 95% CI covers the estimand in 95 of 100 samples of size n.

Whether the CI in a given sample covers the estimand remains unknown (because we could have one of the 5% of samples in which the CI doesn't cover the estimand).

When estimating causal effects: In case of violation of the identification conditions and model misspecification the probability that the 95% CI covers the true effect may be less than 95%.

When bias is severe, this probability may approach 0.

CONFIDENCE INTERVALS

CIs quantify uncertainty of the estimation due to random differences of samples from the target population.

Wider CIs signal greater uncertainty due to random bias.

Interpretation: The 95% CI covers the estimand in 95 of 100 samples of size n.

Whether the CI in a given sample covers the estimand remains unknown (because we could have one of the 5% of samples in which the CI doesn't cover the estimand).

When estimating causal effects: In case of violation of the identification conditions and model misspecification the probability that the 95% CI covers the true effect may be less than 95%.

When bias is severe, this probability may approach 0.

REMEMBER: UNADJUSTED MEAN DIFFERENCE IN FIRST TOY DATA

```
. tabstat mathcomp, by(pschool) notot
Summary for variables: mathcomp
     by categories of: pschool
 pschool |      mean
---------+----------
       0 |  48.26859
       1 |  50.69592
--------------------

. reg mathcomp pschool, nohead nopvalues
------------------------------------------------------------------
    mathcomp |      Coef.   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
     pschool |       2.43        0.11          2.20         2.65
       _cons |      48.27        0.04         48.18        48.35
------------------------------------------------------------------
```

Source: `07estte1.do#2`

The regression coefficient is statistically significant but far away from the true ATE ($= -4$). The 95%CI doesn't cover the true effect either.

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

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

If any one assumption is (strongly) violated, 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 (strong) violation (e.g., mention of a specific potential confounder not adjusted for in the analysis).

(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 one assumption is (strongly) violated, 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 (strong) violation (e.g., mention of a specific potential confounder not adjusted for in the analysis).

(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 one assumption is (strongly) violated, 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 (strong) violation (e.g., mention of a specific potential confounder not adjusted for in the analysis).

(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 one assumption is (strongly) violated, 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 (strong) violation (e.g., mention of a specific potential confounder not adjusted for in the analysis).

(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 one assumption is (strongly) violated, 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 (strong) violation (e.g., mention of a specific potential confounder not adjusted for in the analysis).

(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 one assumption is (strongly) violated, 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 (strong) violation (e.g., mention of a specific potential confounder not adjusted for in the analysis).

(Hernán and Robins, 2018, Ch. 13.5)

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 adjustment).

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)

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 adjustment).

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)

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 adjustment).

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)

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 adjustment).

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)

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 adjustment).

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)

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 adjustment).

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)

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 adjustment).

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: ESTIMATION OF TOTAL EFFECTS USING IPT WEIGHTING

1. Nonparametric inverse probability of treatment (IPT) weighting
2. Parametric estimation of treatment weights
3. Comparing estimators
4. Checks for model misspecification and positivity violations

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/`.

Petersen, M. L. and van der Laan, M. J. (2014). Causal models and learning from data: Integrating causal modeling and statistical estimation. *Epidemiology*, 25(3), 418–426.

Shalizi, C. R. (2016). *Advanced Data Analysis from an Elementary Point of View*. Cambridge University Press, New York. URL `http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/ADAfaEPoV.pdf`.

Get these slides from

https://osf.io/jgrsy/

The METROPOLIS theme was created by Matthias Vogelgesang and is available from

https://github.com/matze/mtheme