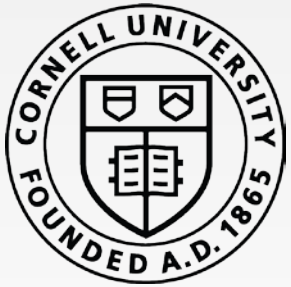
Identification and estimation of causal effects

Felix Thoemmes

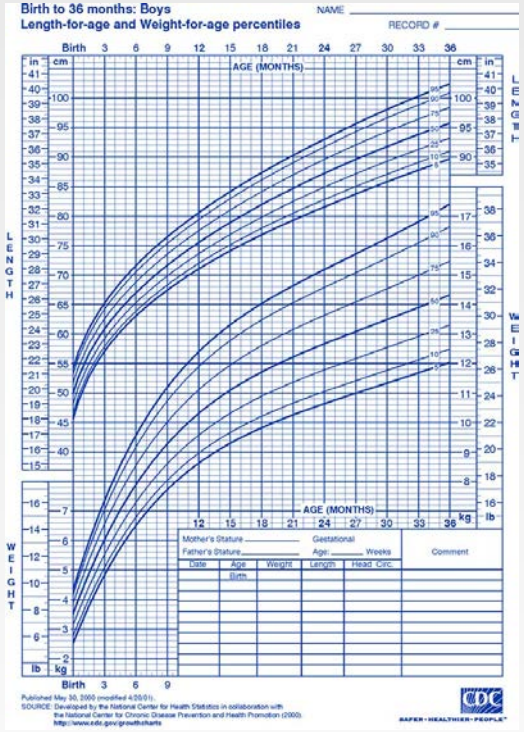


Github resources

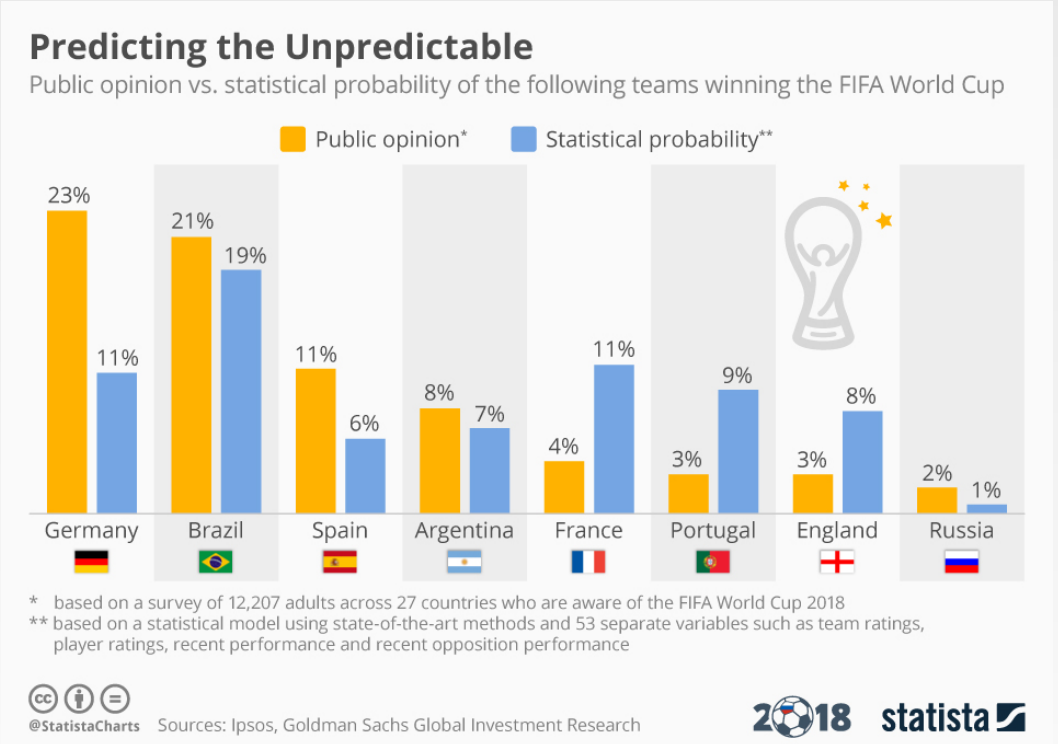
https://github.com/felixthoemmes/IPN_workshop

We do not always need causal inference

Description



Prediction



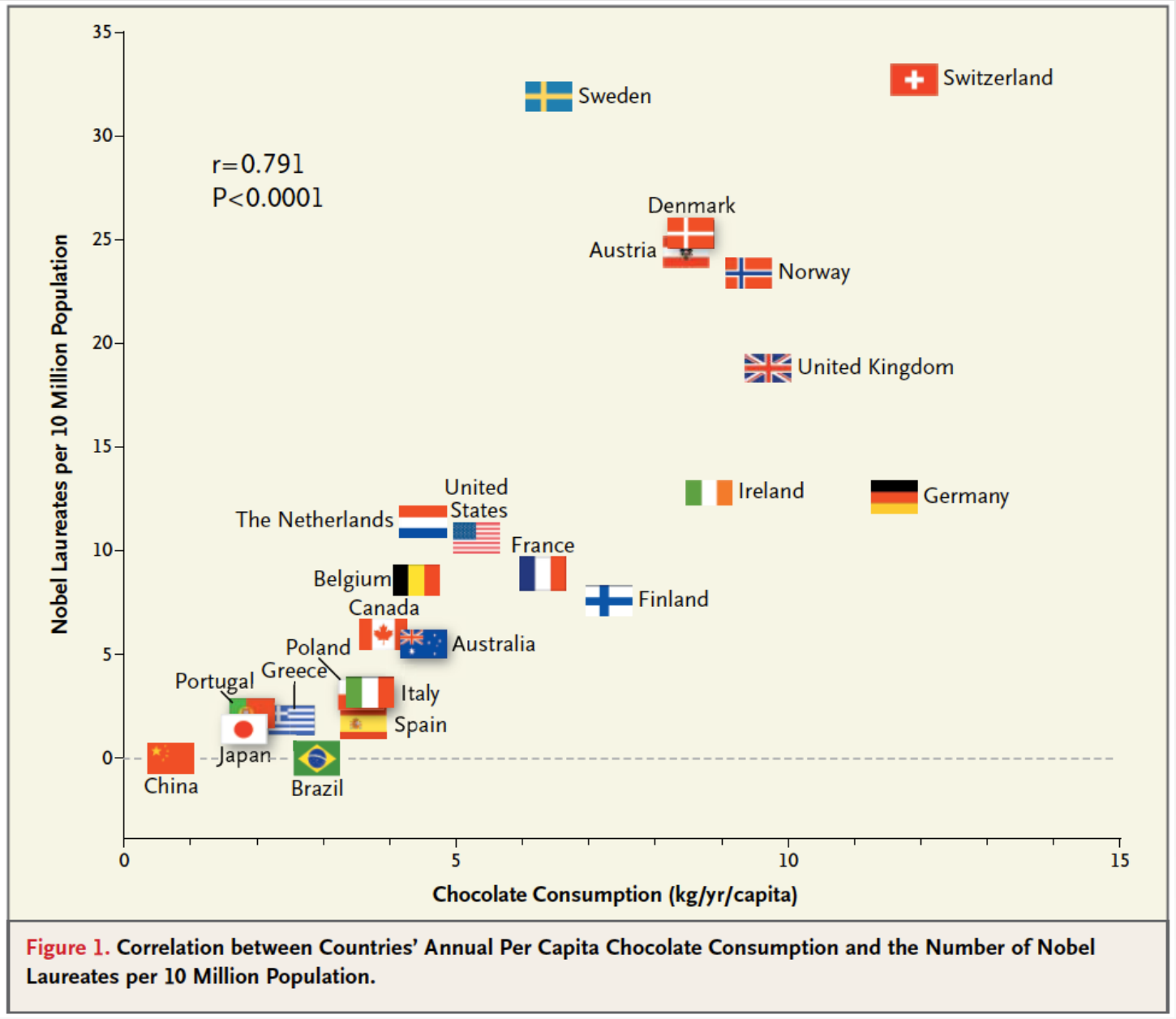
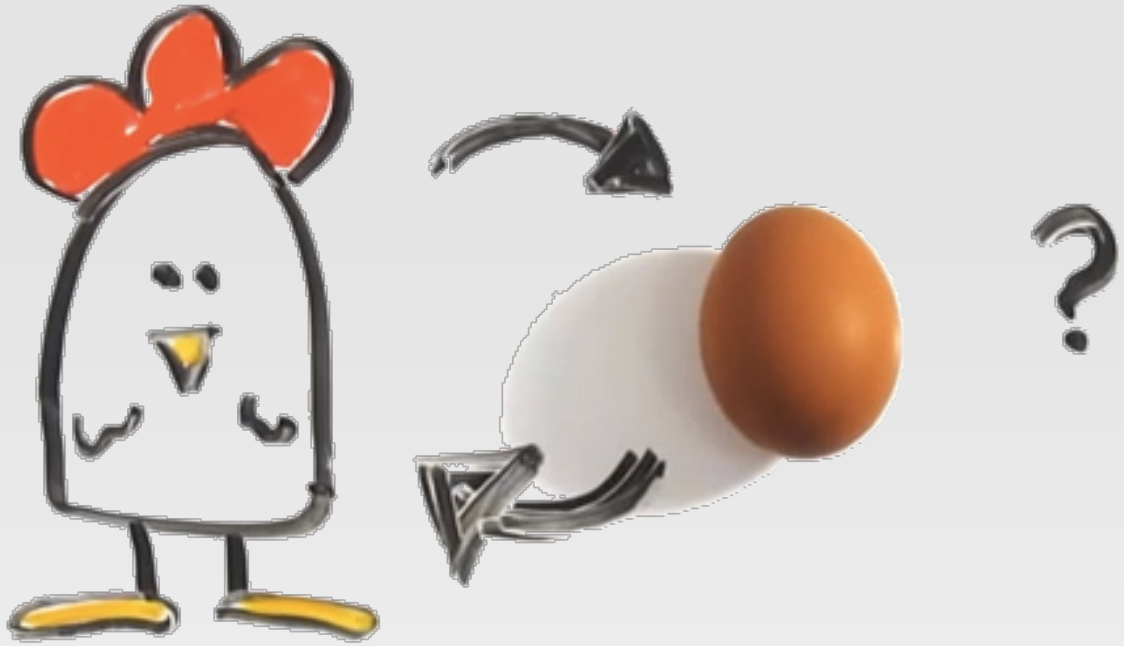


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**



When do we need causality?

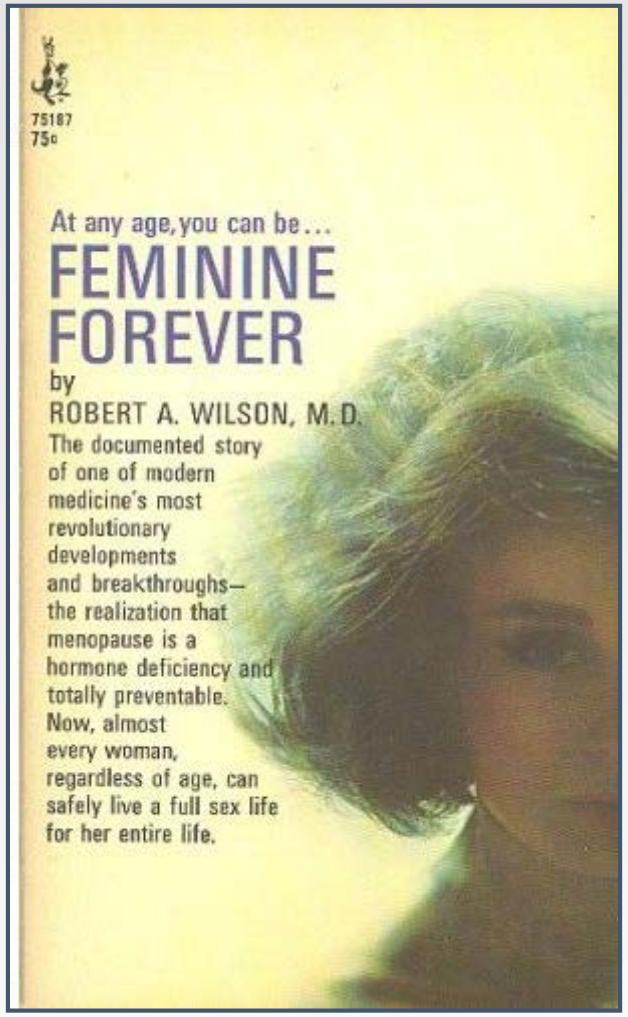
When do we need causality?

- To answer “causal” questions?
- Will changing one variable produce changes in another variable?
- Will implementing policy / treatment generate certain outcomes?
- To build theory, we need causality

Hormone Replacement therapy

- 1968 “Feminine Forever“

- 2002
  Women’s Health Initative
  trial on hormone replacement therapy



Example

- Two variables X and Y are correlated with each other – why?

1. X causes Y
2. Y causes X
3. Common cause C causes both X and Y
4. X and Y share a common effect that was conditioned on
5. Random chance

Example

- Based on *statistical evidence alone,* only the last alternative (chance) can be ruled out

1. X causes Y
2. Y causes X
3. Common cause C causes both X and Y
4. X and Y share a common effect that was conditioned on
5. Random chance

Example

- The remaining four alternative are *statistically indistinguishable*

1. X causes Y
2. Y causes X
3. Common cause C causes both X and Y
4. X and Y share a common effect that was conditioned on

Example

- Common cause C causes both X and Y
- Suppose we observe that stains on hand and lung cancer are correlated
- Tobacco use causes stains on hand and lung cancer

Example

- X and Y share a common effect that was conditioned on
- Suppose wealth and intelligence are unrelated to each other
- We restrict our observations to students admitted to Harvard – here wealth and intelligence will be negatively related to each other

Example

- Among the most intelligent, it does not matter whether you are rich or not. You will be admitted.
- Among the least intelligent, only the richest will be admitted (because their family donated a building to Harvard).
- Conditional on admittance to Harvard, wealth and intelligence appear to be negatively correlated

Causal effects

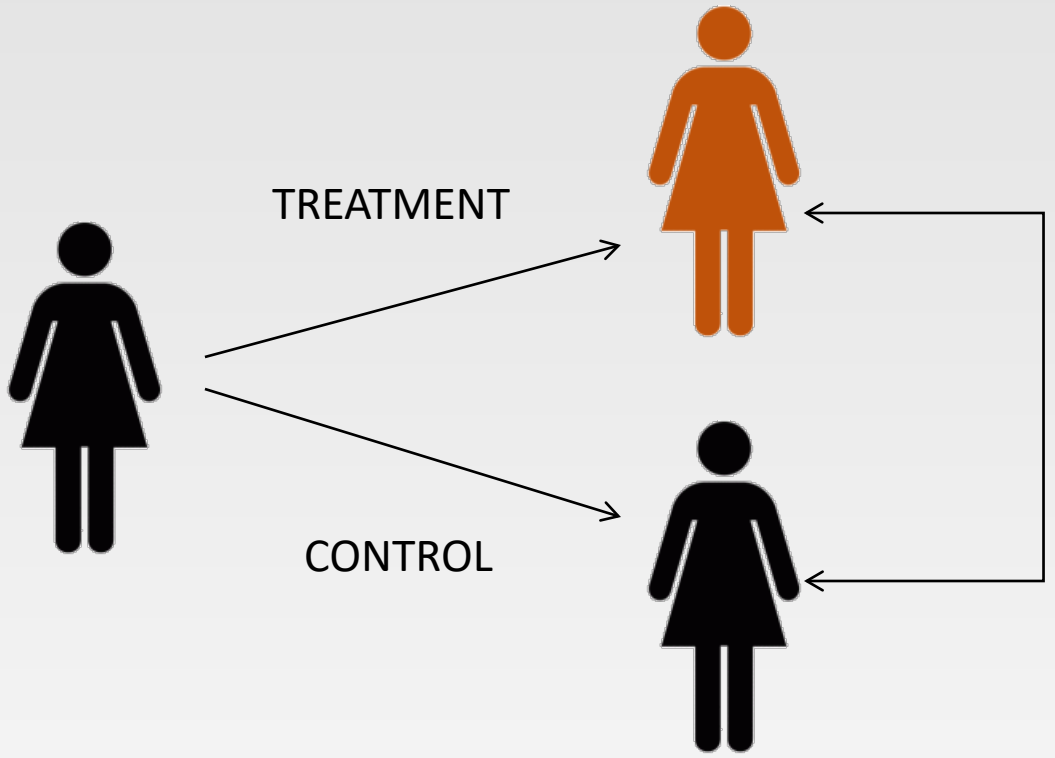
"No causes in, no causes out."
Nancy Cartwright, "Hunting Causes and Using Them"

No statistical model alone yields causal effects

It is always necessary to makes some *untestable* assumptions.

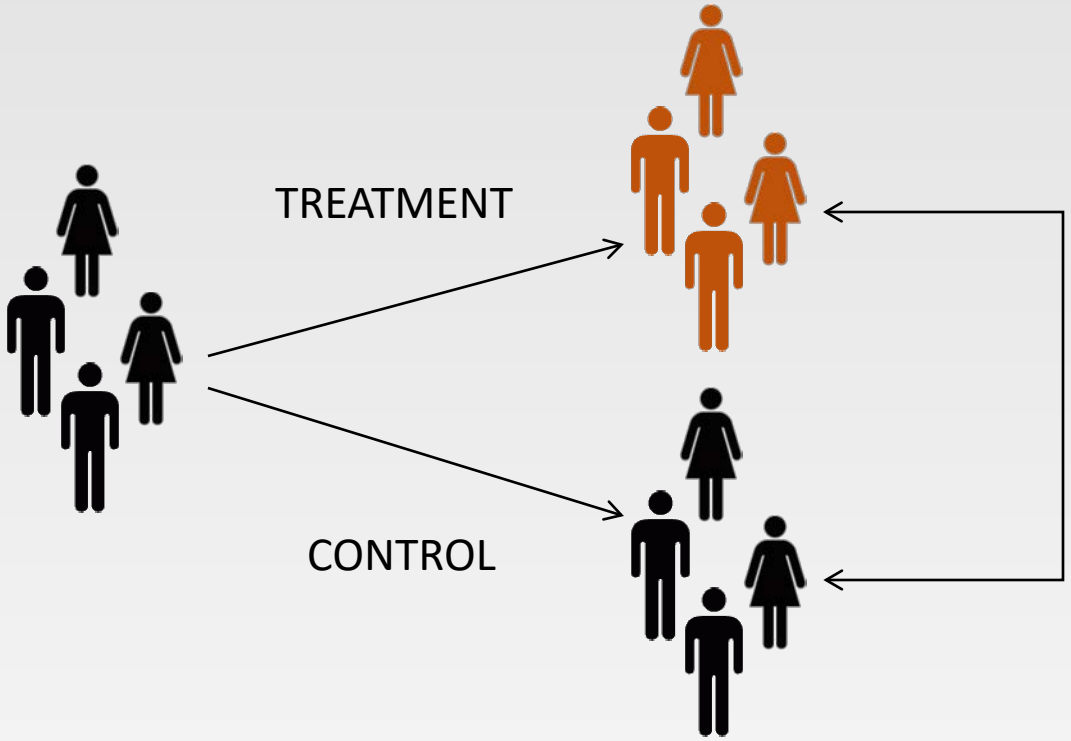
Brainstorm exercise

Some theory and definitions



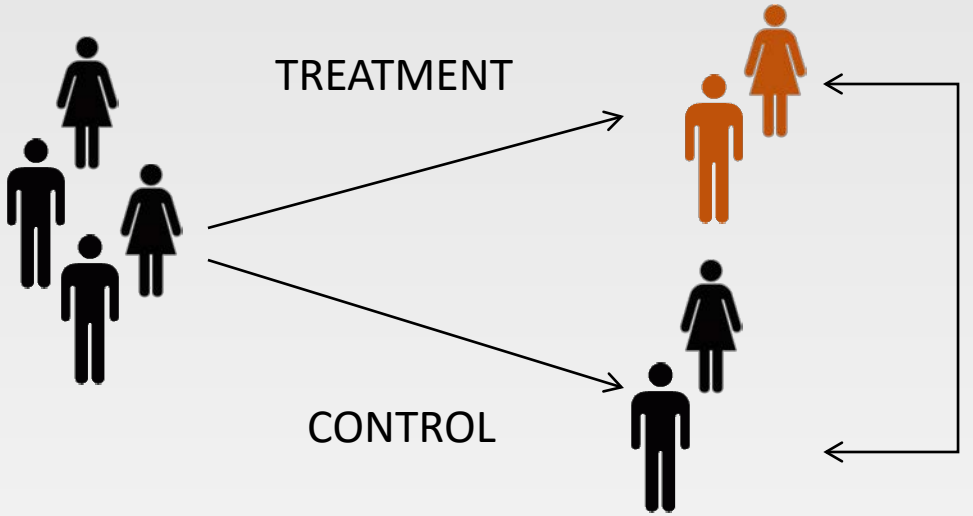
$$\tau_i = Y_{i1} - Y_{i0}$$

Unit –level Causal Effect



$$\tau = \frac{1}{n}\sum_{i=1}^{n} Y_{i1} - \frac{1}{n}\sum_{i=1}^{n} Y_{i0}$$

Average Causal Effect



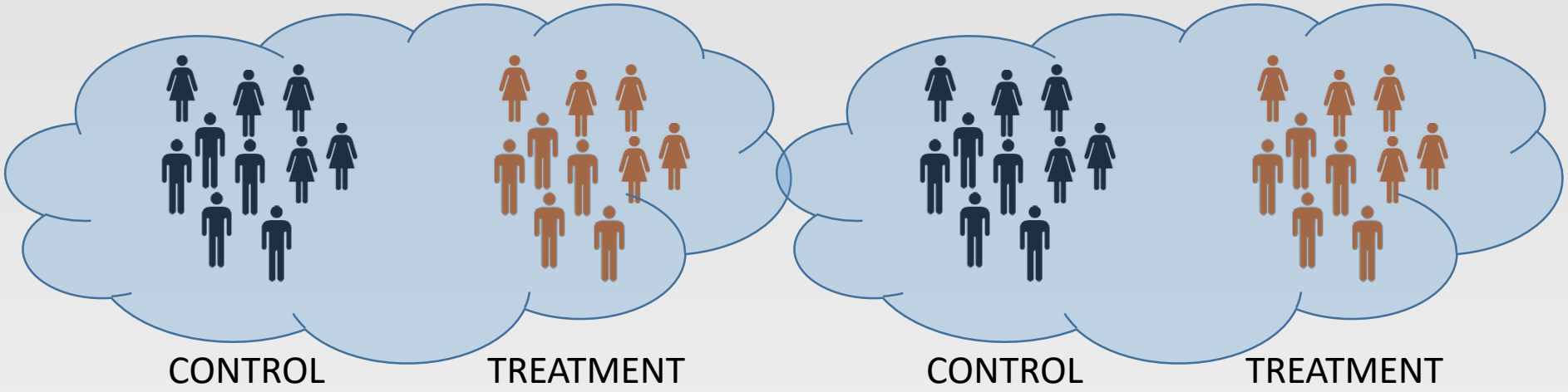
$$\tau^* = \frac{1}{n}\sum_{i=1}^{n}(Y_{i1} \mid Z_1) - \frac{1}{n}\sum_{i=1}^{n}(Y_{i0} \mid Z_0)$$

Estimate of the Average Causal Effect

$\tau = \tau^*$?

"Causation ≠ Association"

Causal identification



$E(Y_{i1}) = E(Y_{i1} \mid z_i = 1)$
$E(Y_{i0}) = E(Y_{i0} \mid z_i = 0)$

$E(Y_{i1}) \neq E(Y_{i1} \mid z_i = 1)$
$E(Y_{i0}) \neq E(Y_{i0} \mid z_i = 0)$

$E(Y_{i1}) = E(Y_{i1} \mid z_i = 1)$
$E(Y_{i0}) = E(Y_{i0} \mid z_i = 0)$

Randomized experiment

$E(Y_{i1}) \neq E(Y_{i1} \mid z_i = 1)$
$E(Y_{i0}) \neq E(Y_{i0} \mid z_i = 0)$

Non-randomized experiment

$E(Y_{i1}) = E_x\{E(Y_{i1} \mid z_i = 1, x)\}$ Non-randomized

$E(Y_{i0}) = E_x\{E(Y_{i0} \mid z_i = 0, x)\}$ experiment with unconfoundedness assumption

X contains all confounding covariates

Conditional ignorabilty (Rubin, 194)

U	Z	$P(U=u)$	$P(X=0 \mid U)$	$P(X=1 \mid U)$	$\tau_0 = E(Y \mid X=0, U)$	$\tau_1 = E(Y \mid X=1, U)$	$\tau_1 - \tau_0$
u_1	1	1/4	1/2	1/2	0	2	2
u_2	1	1/4	1/2	1/2	0	2	2
u_3	2	1/4	1/2	1/2	0	1/4	1/4
u_4	2	1/4	1/2	1/2	0	1/4	1/4

https://www.metheval.uni-jena.de/projekte/book_causal-effects/tools.php

U	Z	$P(U=u)$	$P(X=0 \mid U)$	$P(X=1 \mid U)$	$\tau_0 = E(Y \mid X=0, U)$	$\tau_1 = E(Y \mid X=1, U)$	$\tau_1 - \tau_0$
u_1	1	1/4	1/2	1/2	0	2	2
u_2	1	1/4	1/2	1/2	0	2	2
u_3	2	1/4	1/2	1/2	0	1/4	1/4
u_4	2	1/4	1/2	1/2	0	1/4	1/4

Z	T	Y0	Y1
1	0	0	●
1	1	●	2
1	0	0	●
1	1	●	2
2	0	0	●
2	1	●	1/4
2	0	0	●
2	1	●	1/4
		0	1.125

U	Z	$P(U=u)$	$P(X=0 \mid U)$	$P(X=1 \mid U)$	$\tau_0 = E(Y \mid X=0, U)$	$\tau_1 = E(Y \mid X=1, U)$	$\tau_1 - \tau_0$
u_1	1	1/4	1/4	3/4	0	2	2
u_2	1	1/4	1/4	3/4	0	2	2
u_3	2	1/4	3/4	1/4	0	1/4	1/4
u_4	2	1/4	3/4	1/4	0	1/4	1/4

U	Z	$P(U=u)$	$P(X=0 \mid U)$	$P(X=1 \mid U)$	$\tau_0 = E(Y \mid X=0, U)$	$\tau_1 = E(Y \mid X=1, U)$	$\tau_1 - \tau_0$
u_1	1	1/4	1/4	3/4	0	2	2
u_2	1	1/4	1/4	3/4	0	2	2
u_3	2	1/4	3/4	1/4	0	1/4	1/4
u_4	2	1/4	3/4	1/4	0	1/4	1/4

Z	T	Y0	Y1
1	1	●	2
1	1	●	2
1	1	●	2
1	0	0	●
2	0	0	●
2	0	0	●
2	0	0	●
2	1	●	1/4
		0	1.562

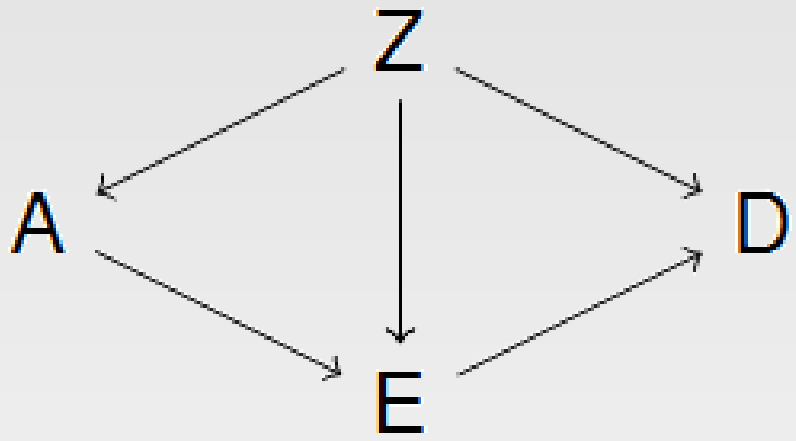
$\tau = \tau^*$?

"Causation ≠ Association"

Causal identification – ***but how?***

Brainstorm exercise

Graphical “language”

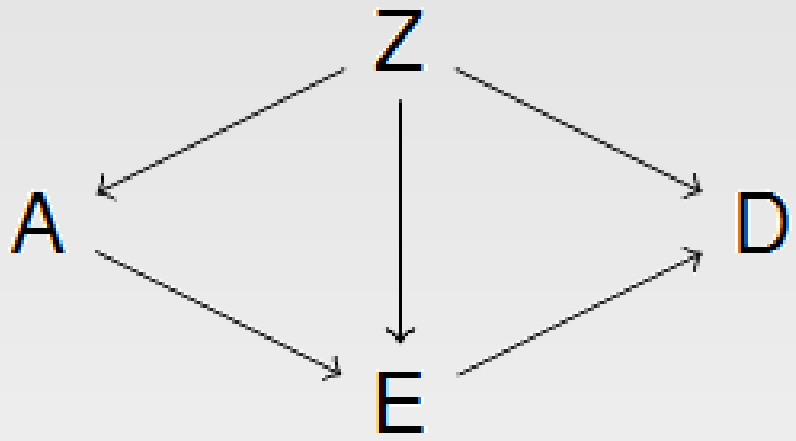


- Letters in graphs are **variables** and often referred to as nodes or vertices

- Connections between nodes are referred to as **arrows** or edges

Credit: J. Textor

Graphical “language”

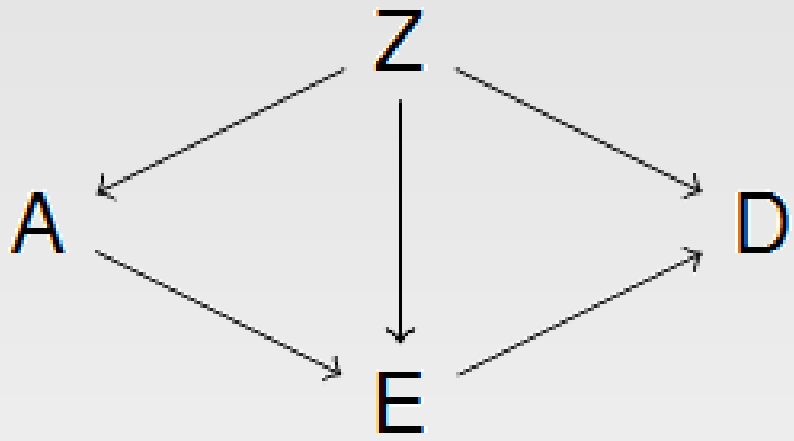


- Relationships between nodes

- Z is parent of A, D, E
- D is a child of Z and E
- D is a descendant of A
- A is an ancestor of D

Credit: J. Textor

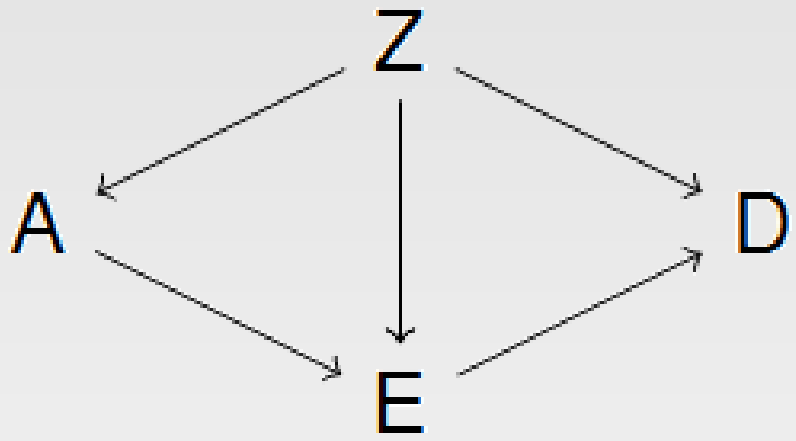
Graphical “language”



- Nodes with arrows (directly) emanating into other nodes are referred to as parents (or ancestors)

- Nodes with arrows pointing (directly) into them are children (or descendants)

Graphical “language”



- A connection via several arrows is referred to as a **path** (or trail)

- A path can have arrows going into different directions, e.g., A ← Z → E → D

Credit: J. Textor

Causal model

- A directed arrow between two nodes represents an assumed causal effect between two variables

- This effect may be linear, non-linear, deterministic – completely non-parametric

Causal model

- A bi-directed arrow between two nodes represents an unobserved cause

- We may either draw a latent variable with directed paths or use bi-directed arrows

Causal model

- The absence of an arrow between two variables denotes the absence of any direct effect or latent confounding

- It is the ***absence*** of arrows that is most critical and that must be argued for

Causal model

- There might be strong disagreement about how the DAG should look like

- However, once researchers agree on the structure of the DAG, there should also be agreement about which effects can be causally interpreted and which model should be used to estimate these effects

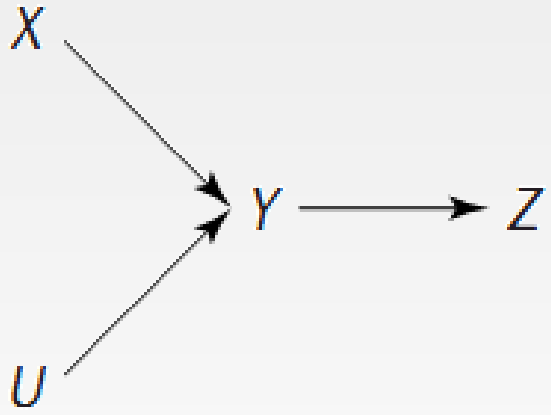
Causal model

- DAGs put our theoretical assumptions about the relationships between variables of interest in a graphical format

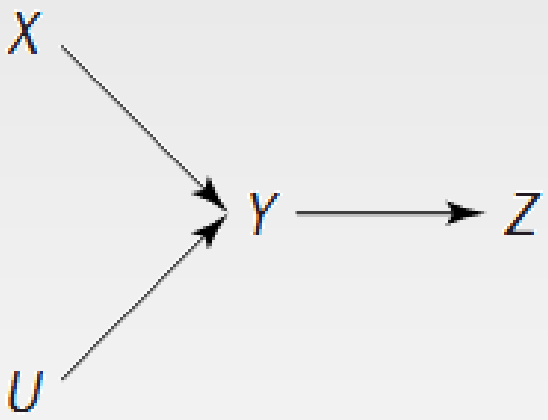
- DAGs as a manifestation of your literature review and best known theory

Causal model

- What kind of assumptions are advertised in the following DAG?



Credit: M. Glymour



- *X* and *U* are each direct causes of *Y* (direct with respect to other variables in the DAG).
- *Y* is a direct cause of *Z*.
- *X* is not a direct cause of *Z*, but *X* is an indirect cause of *Z* via *Y*.
- *X* is not a cause of *U* and *U* is not a cause of *X*.
- *U* is not a direct cause of *Z*, but *U* is an indirect cause of *Z* via *Y*.
- No two variables in the DAG (*X*, *U*, *Y*, or *Z*) share a prior cause not shown in the DAG, e.g., no variable causes both *X* and *Y*, or both *X* and *U*.

Credit: M. Glymour

Causal model

- The back-door criterion relies on a concept called d-separation and the idea of “blocking” paths that would otherwise induce bias

Type of paths

- We define a path that has an arrow going out of the treatment variable, as a front-door path
- We define a path that has an arrow going into the treatment variable, as a back-door path
- We define a path that is not blocked, as open
- We define a path that is blocked, as closed
- All four combinations of paths can exist

Type of paths

- Front-door + open → **causal path**
- Front-door + closed → neutral, but biasing (for total effect) if opened
- Back-door + open → **biasing path**
- Back-door + closed → neutral, but biasing if opened

Blocking a path

- Blocking refers to holding a variable constant in a graph
- In practice, this may mean regression adjustment, or other techniques
- Whether or not a variable blocks a particular path is dependent on other variables in the path, and the direction of the arrows

Blocking a path

- Common cause (fork)

- Mediation (chain)

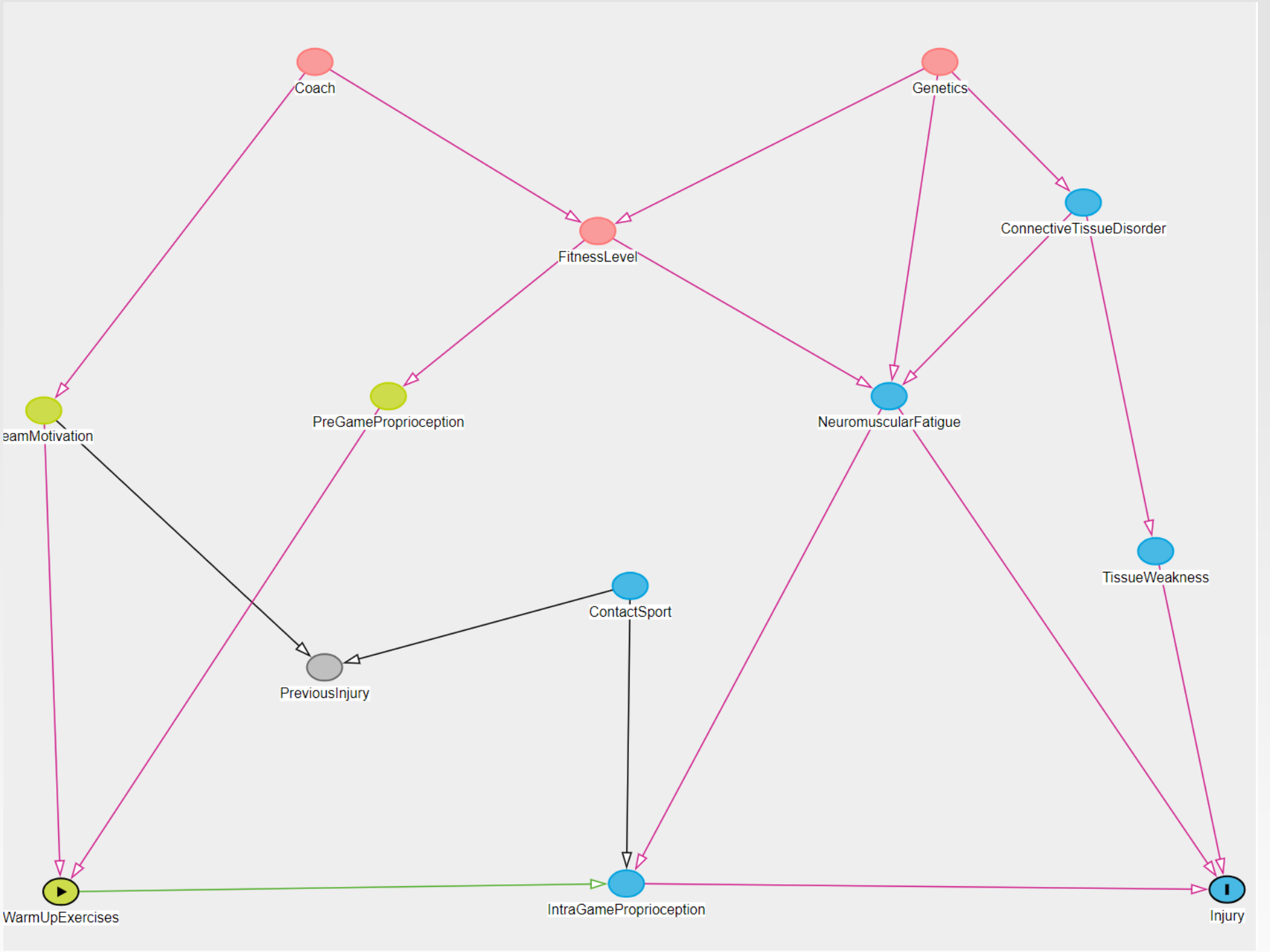
- Mediation (inverted chain)

- Collider (inverted fork)

Back-door

- Check all paths from your supposed cause X (treatment) to outcome
- Block all biasing paths, without blocking any causal paths, or opening any closed front-door paths (leave all front-door paths as they were)
- Note that in this process of blocking we might open previously closed back-door paths which could turn into biasing paths, which we will need to close as well

Brainstorm exercise


Coach
Genetics
FitnessLevel
ConnectiveTissueDisorder
eamMotivation
PreGameProprioception
NeuromuscularFatigue
TissueWeakness
ContactSport
PreviousInjury
WarmUpExercises
IntraGameProprioception
Injury

Exercise 1

Exercise 1

- Download the “ex1.pdf” file and follow the instructions given in the document

- If you have trouble copying and pasting from a PDF, you can also download the source file “ex1.Rmd”

Regression adjustment

U	Z	$P(U=u)$	$P(X=0\mid U)$	$P(X=1\mid U)$	$\tau_0 = E(Y \mid X=0, U)$	$\tau_1 = E(Y \mid X=1, U)$	$\tau_1 - \tau_0$
u_1	1	1/4	1/4	3/4	0	2	2
u_2	1	1/4	1/4	3/4	0	2	2
u_3	2	1/4	3/4	1/4	0	1/4	1/4
u_4	2	1/4	3/4	1/4	0	1/4	1/4

Z	T	Y0	Y1
1	1	●	2
1	1	●	2
1	1	●	2
1	0	0	●
2	0	0	●
2	0	0	●
2	0	0	●
2	1	●	1/4
		0	1.562

U	Z	$P(U=u)$	$P(X=0\mid U)$	$P(X=1\mid U)$	$\tau_0 = E(Y \mid X=0, U)$	$\tau_1 = E(Y \mid X=1, U)$	$\tau_1 - \tau_0$
u_1	1	1/4	1/4	3/4	0	2	2
u_2	1	1/4	1/4	3/4	0	2	2
u_3	2	1/4	3/4	1/4	0	1/4	1/4
u_4	2	1/4	3/4	1/4	0	1/4	1/4

Z	T	Y0	Y1
1	1	●	2
1	1	●	2
1	1	●	2
1	0	0	●

U	Z	$P(U=u)$	$P(X=0\mid U)$	$P(X=1\mid U)$	$\tau_0 = E(Y \mid X=0, U)$	$\tau_1 = E(Y \mid X=1, U)$	$\tau_1 - \tau_0$
u_1	1	1/4	1/4	3/4	0	2	2
u_2	1	1/4	1/4	3/4	0	2	2
u_3	2	1/4	3/4	1/4	0	1/4	1/4
u_4	2	1/4	3/4	1/4	0	1/4	1/4

Z	T	Y0	Y1
2	0	0	●
2	0	0	●
2	0	0	●
2	1	●	1/4

U	Z	$P(U=u)$	$P(X=0 \mid U)$	$P(X=1 \mid U)$	$\tau_0 = E(Y \mid X=0, U)$	$\tau_1 = E(Y \mid X=1, U)$	$\tau_1 - \tau_0$
u_1	1	1/4	1/4	3/4	0	2	2
u_2	1	1/4	1/4	3/4	0	2	2
u_3	2	1/4	3/4	1/4	0	1/4	1/4
u_4	2	1/4	3/4	1/4	0	1/4	1/4

Z	T	Y0	Y1
2	0	0	1/4
2	0	0	1/4
2	0	0	1/4
2	1	0	1/4
		0	.25

Z	T	Y0	Y1
1	1	0	2
1	1	0	2
1	1	0	2
1	0	0	2
		0	2

(.25 + 2) / 2 = 1.125

U	Z	$P(U=u)$	$P(X=0\mid U)$	$P(X=1\mid U)$	$\tau_0 = E(Y \mid X=0, U)$	$\tau_1 = E(Y \mid X=1, U)$	$\tau_1 - \tau_0$
u_1	1	1/4	1/4	3/4	0	2	2
u_2	1	1/4	1/4	3/4	0	2	2
u_3	2	1/4	3/4	1/4	0	1/4	1/4
u_4	2	1/4	3/4	1/4	0	1/4	1/4

Z	T	Y0	Y1
1	1	0	2
1	1	0	2
1	1	0	2
1	0	0	2
		0	2

U	Z	$P(U=u)$	$P(X=0 \mid U)$	$P(X=1 \mid U)$	$\tau_0 = E(Y \mid X=0, U)$	$\tau_1 = E(Y \mid X=1, U)$	$\tau_1 - \tau_0$
u_1	1	1/4	1/4	3/4	0	2	2
u_2	1	1/4	1/4	3/4	0	2	2
u_3	2	1/4	3/4	1/4	0	1/4	1/4
u_4	2	1/4	3/4	1/4	0	1/4	1/4

Z	T	Y0	Y1
2	0	0	1/4
2	0	0	1/4
2	0	0	1/4
2	1	0	1/4
		0	.25

U	Z	$P(U=u)$	$P(X=0 \mid U)$	$P(X=1 \mid U)$	$\tau_0 = E(Y \mid X=0, U)$	$\tau_1 = E(Y \mid X=1, U)$	$\tau_1 - \tau_0$
u_1	1	1/4	1/4	3/4	0	2	2
u_2	1	1/4	1/4	3/4	0	2	2
u_3	2	1/4	3/4	1/4	0	1/4	1/4
u_4	2	1/4	3/4	1/4	0	1/4	1/4

Z	T	Y0	Y1
2	0	0	1/4
2	0	0	1/4
2	0	0	1/4
2	1	0	1/4
		0	.25

Z	T	Y0	Y1
1	1	0	2
1	1	0	2
1	1	0	2
1	0	0	2
		0	2

(.25 + 2) / 2 = 1.125

Adjustment

- In the previous slide we saw the fundamental idea behind adjustment
- Using a model (here just modeling means within strata of a covariate) and estimating effects within strata of units that are identical
- We may also use this model to estimate potential outcomes
- Conditional ignorability must hold

Statistical aspects of adjustment

- Research manuscripts often mention that variables were adjusted on, or controlled for

- What exactly does it mean to control for something?

Statistical aspects of adjustment

- In many instances adjustment or controlling for is synonymous with adding a covariate (a potential confounder) to a linear regression model
- But other adjustment techniques also exist (matching, weighting)
- We will cover all three of them

Statistical aspects of adjustment

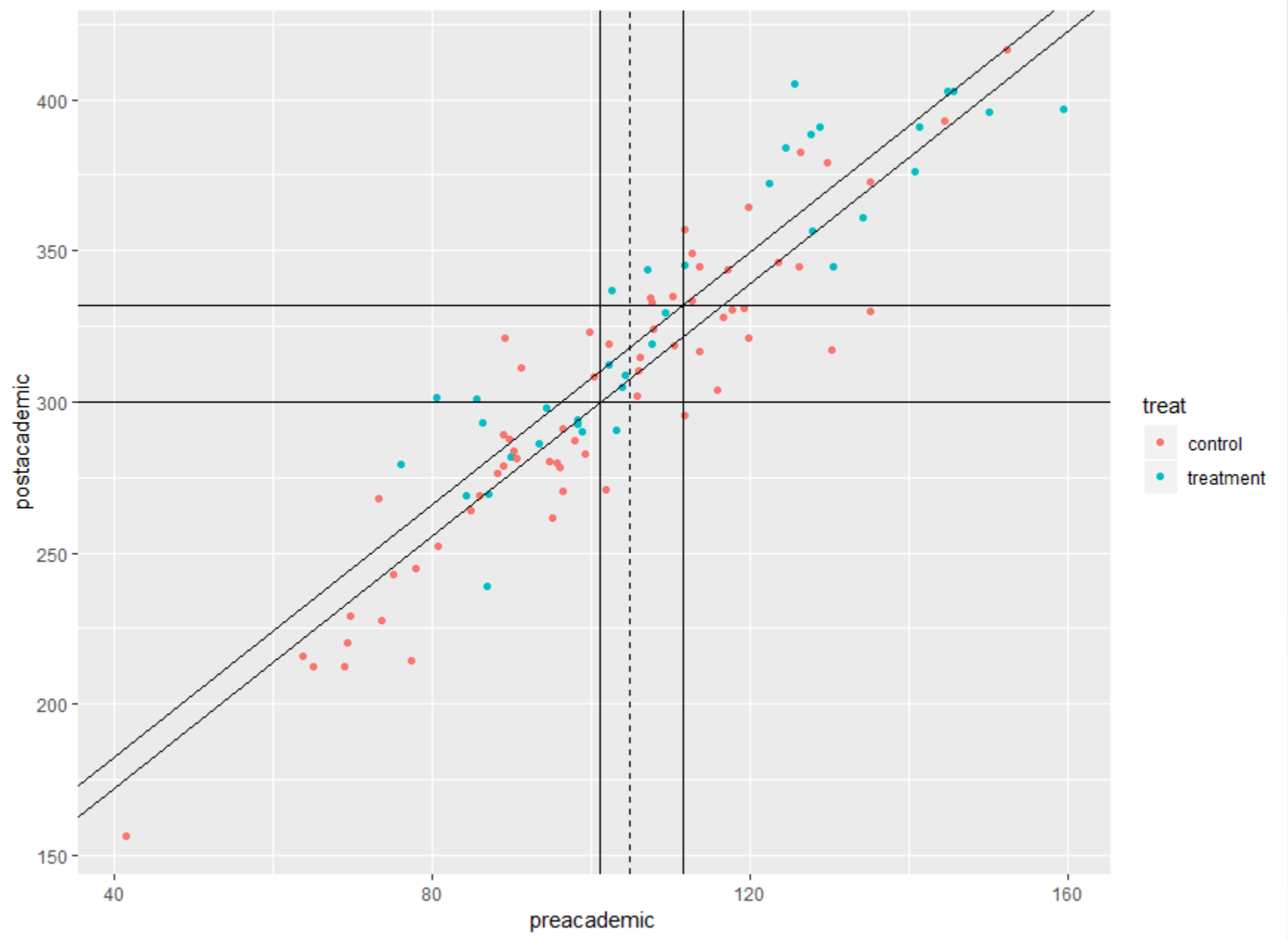
- Consider that we are interested in the causal effect of a non-randomized treatment

- We have covariates at our disposal, and through the use of theory (and maybe a DAG) we feel confident that we have a set of covariates that would fulfill ignorability

- How do we adjust on these variables?

Simple example

- A non-randomized treatment that hopes to increase academic achievement is offered to a group of students
- Students select into treatments and consequently end up differing on important pre-test covariates
- One of them is prior academic achievement


400
350
300
250
200
150
postacademic
40
80
120
160
preacademic
treat
control
treatment

Simple example

- Comparison of groups that differ in treatment status AND at the same time differ on their pre-treatment achievement

- How would the means of the two groups look like if the pre-treatment achievement scores were identical? (e.g., both were on the overall mean)

Simple example

- Using a model (e.g., a parametric linear regression model) we can predict what the post-test score of a unit would be if assigned to the treatment group, or the control group, given values on pre-academic achievement

- These predictions emerge from a model, hence this is referred to as *model-based adjustment*

Simple example

- We can use the regression equation to predict the post-test academic achievement for each person, under both control and treatment

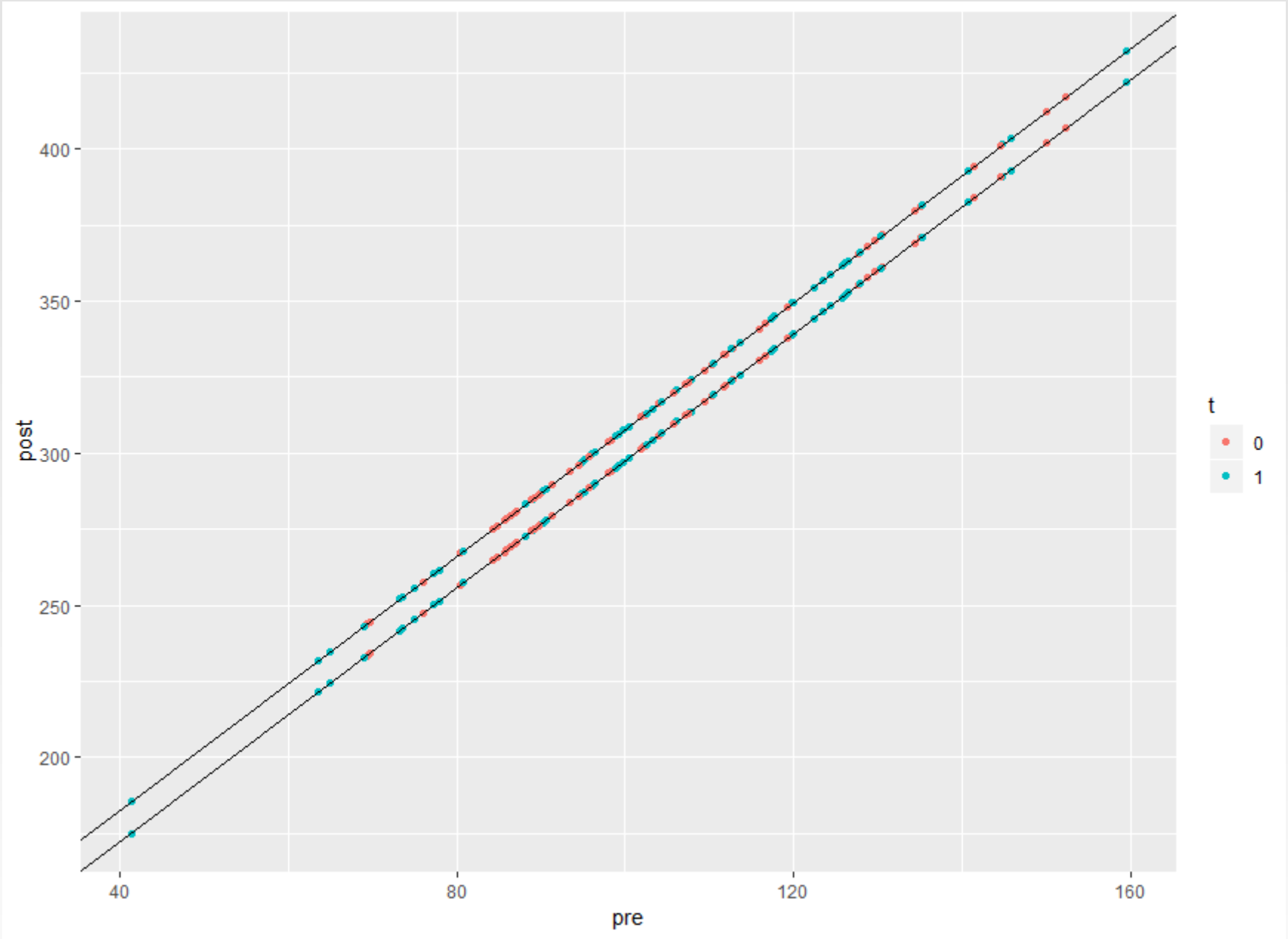
- From these individual values, we can aggregate means, which are referred to as “adjusted means”

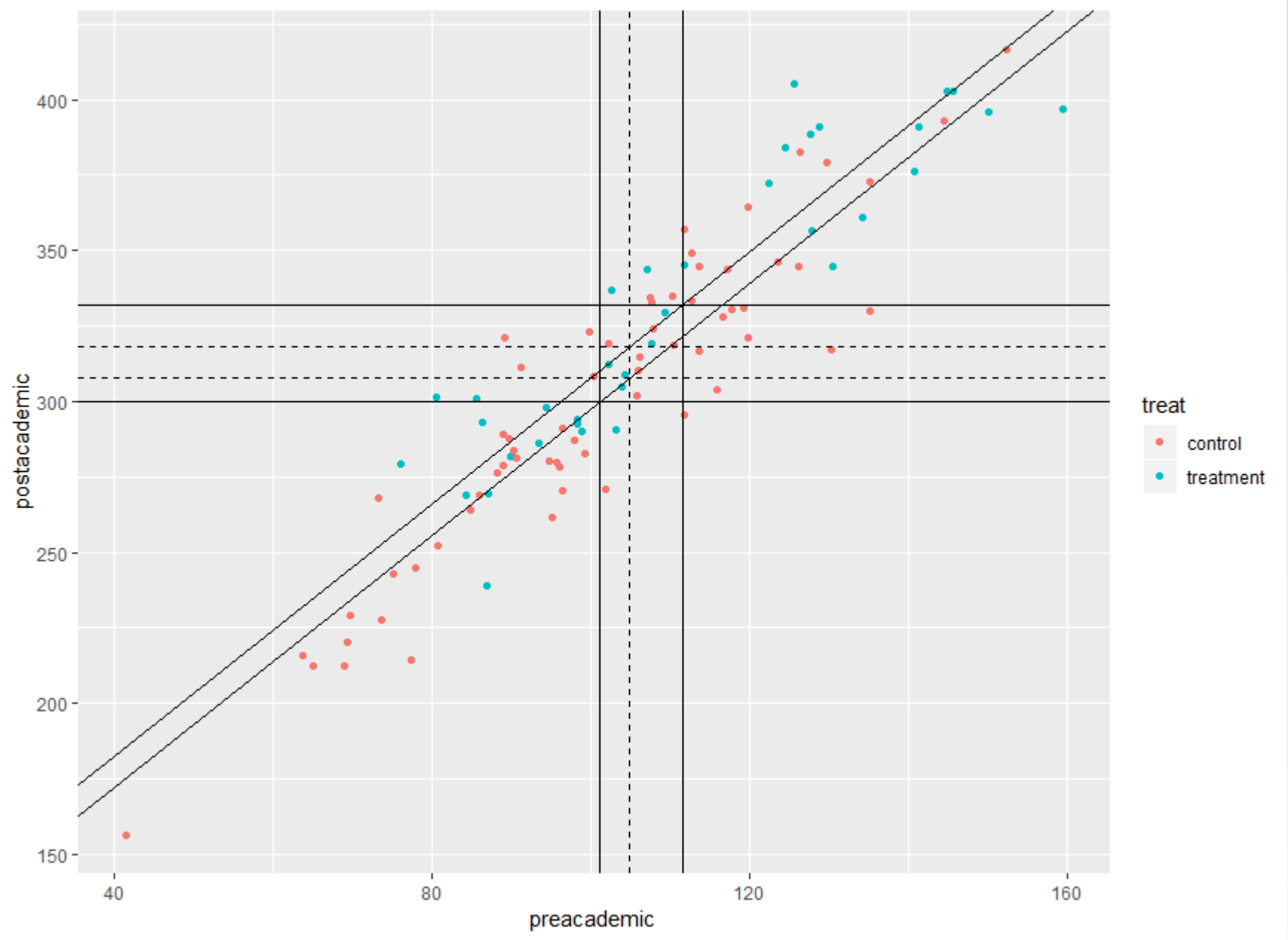
- This model is essentially identical to the ANCOVA model

Simple example

- Differences between these adjusted means are *adjusted treatment effects*

- They answer the question what the treatment effect would be, if we compared similar units *or* observed every potential outcome (which we do not)


post
400
350
300
250
200
40
80
120
160
pre
t
0
1


400
350
300
250
200
150
postacademic
40
80
120
160
preacademic
treat
control
treatment

Simple example

- It is possible to obtain these treatment effects *directly* from the (summary) output of a regression model
- This however only works in simple (linear, non-interactive) models
- More complicated models (with interactions, non-linear effects) require some type of post-processing of the regression estimates

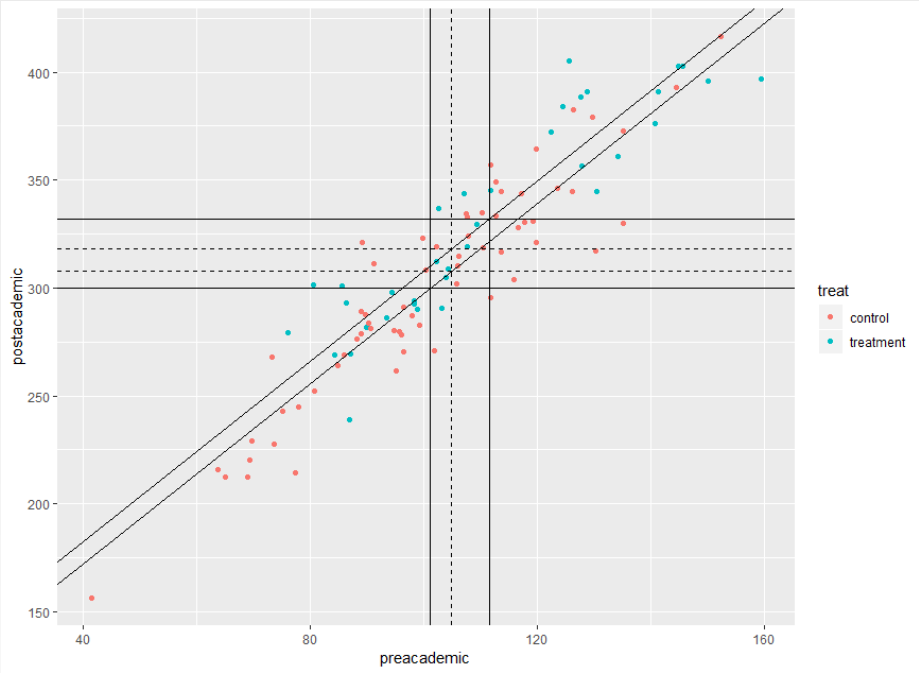
Brainstorm exercise

Simple example

- R has various packages that perform this type of analysis
  - *margins* – general package for average effects (based on Stata's margins command)
  - *EffectLiteR* – powerful package for causal effects using multi-group SEM
  - *tlme* – package for G-estimation, especially useful for doubly robust methods and flexible estimation
  - *emmeans* – general purpose package for post-processing regression results with categorical treatment variables

emmeans example

- Same example of post-academic achievement, and the effect of treatment, adjusted on pre-academic achievement



emmeans

```
> #unadjusted model
> lm.u ← lm(postacademic~treat)
> summary(emmeans(lm.u,"treat",contr="pairwise",weights="proportinal"),infer=TRUE)
$`emmeans`
 treat       emmean       SE df lower.CL upper.CL t.ratio p.value
 control   299.7158 6.120155 98 287.5706 311.8611  48.972  <.0001
 treatment 332.0407 8.160206 98 315.8470 348.2344  40.690  <.0001

Confidence level used: 0.95

$contrasts
 contrast             estimate       SE df  lower.CL upper.CL t.ratio p.value
 control - treatment -32.32488 10.20026 98 -52.56696 -12.0828  -3.169  0.0020

Confidence level used: 0.95

>
```

Regression and predicted values

```
> #linear adjustment on pre-test
> lm.a ← lm(postacademic~treat+preacademic)
> summary(lm.a)

Call:
lm(formula = postacademic ~ treat + preacademic)

Residuals:
    Min      1Q  Median      3Q     Max
-43.610 -11.458  -1.177  13.692  46.115

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    88.57682    9.34147   9.482 1.76e-15 ***
treattreatment 10.33836    4.08992   2.528   0.0131 *
preacademic     2.08961    0.08938  23.379  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.11 on 97 degrees of freedom
Multiple R-squared:  0.8633, Adjusted R-squared:  0.8605
F-statistic: 306.3 on 2 and 97 DF,  p-value: < 2.2e-16
```

Regression and predicted values

```
> #treatment effect by hand using predictions
> mean(predict(lm.a,newdata = data.frame(treat=factor(rep("control",100))))) - mean(predict(lm.a,newdata = data.frame(treat=factor(rep("treatment",100)))))
[1] -10.33836
>
```

emmeans

```
> #linear adjustment on pre-test
> lm.a ← lm(postacademic~treat+preacademic)
> summary(emmeans(lm.a,"treat",contr="pairwise",weights="proportinal"),infer=TRUE) #this will work
$`emmeans`
 treat       emmean       SE df lower.CL upper.CL t.ratio p.value
 control   307.6310 2.412077 97 302.8437 312.4183 127.538  <.0001
 treatment 317.9693 3.240649 97 311.5375 324.4011  98.119  <.0001

Confidence level used: 0.95

$contrasts
 contrast             estimate       SE df  lower.CL  upper.CL t.ratio p.value
 control - treatment -10.33836 4.089921 97 -18.45572 -2.220999  -2.528  0.0131

Confidence level used: 0.95
```

Exercise 2

Exercise 2

- Download the “ex2.pdf” file and follow the instructions given in the document

- If you have trouble copying and pasting from a PDF, you can also download the source file “ex2.Rmd”

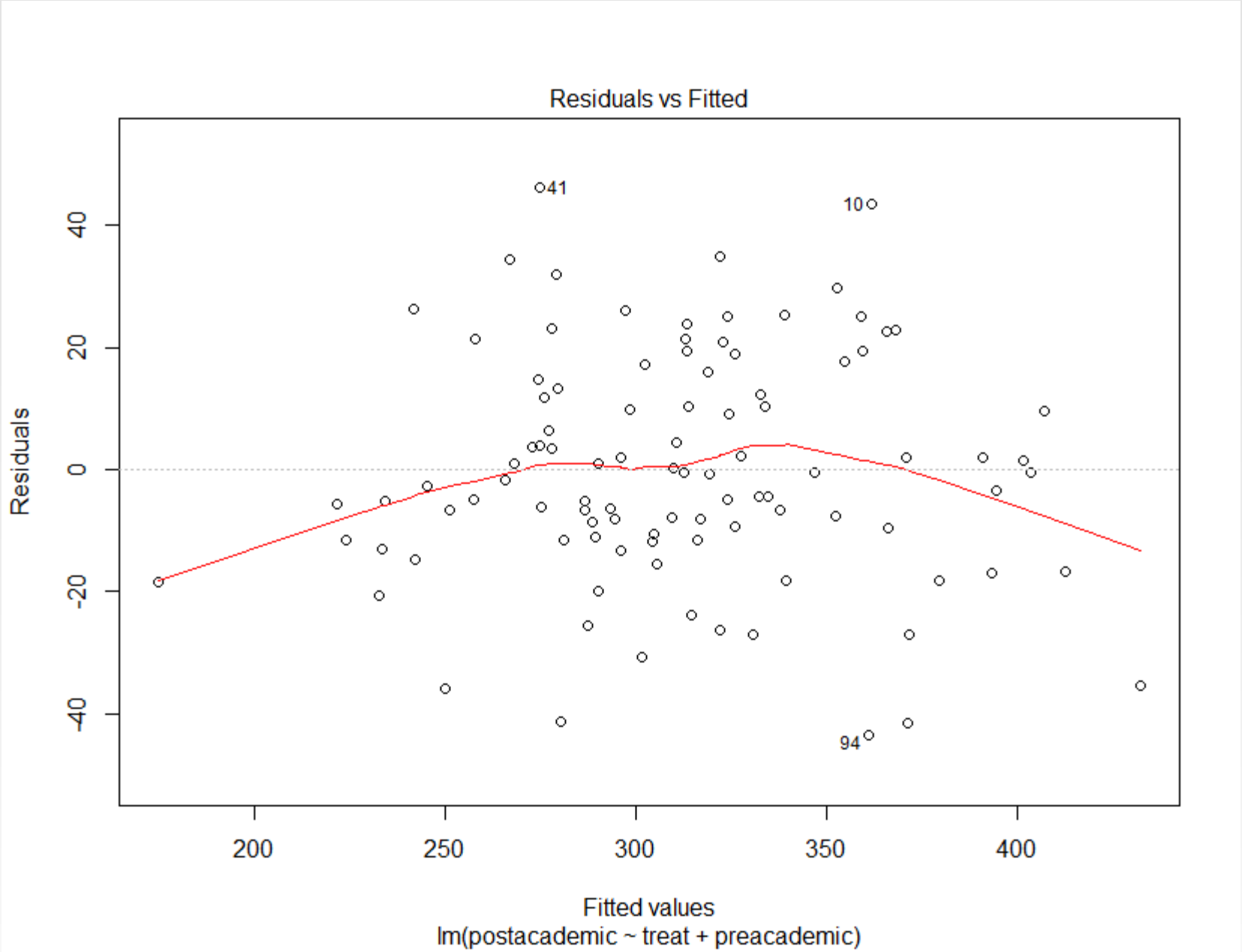
ANCOVA

- Linearity assumption needs to be confirmed
  - But can be relaxed

- No-interaction assumption needs to be confirmed
  - But can be relaxed

ANCOVA

- How do we know whether a non-linear effect or an interaction is needed?
- Regression diagnostics
- Machine learning to circumvent the whole process of selecting functional forms and checking them (*tmle* SuperLearner)


Residuals vs Fitted
41
10
94
Residuals
40
20
0
-20
-40
200
250
300
350
400
Fitted values
lm(postacademic ~ treat + preacademic)

Effect estimation

- In the presence of non-linear effects, and interactions, typical regression summaries do not show the treatment effect
- The treatment effect can still be derived by generating predicted values for each unit under treatment and under control (and then computing differences, and aggregates of differences)
- It is also possible to obtain the same estimates by evaluating group mean differences at the mean value of all covariates
- emmeans can do this for us

```
> lm.d ← lm(postacademic~treat+preacademic+preacademic2+treat:preacademic+treat:preacademic2)
> summary(lm.d)

Call:
lm(formula = postacademic ~ treat + preacademic + preacademic2 +
    treat:preacademic + treat:preacademic2)

Residuals:
    Min      1Q  Median      3Q     Max
-43.495 -11.148  -1.312  13.509  45.322

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                25.890942  37.083877   0.698    0.487
treattreatment             -7.084379  95.985875  -0.074    0.941
preacademic                 3.303751   0.749575   4.407 2.77e-05 ***
preacademic2               -0.005633   0.003723  -1.513    0.134
treattreatment:preacademic  0.375878   1.739344   0.216    0.829
treattreatment:preacademic2 -0.001890   0.007713  -0.245    0.807
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.88 on 94 degrees of freedom
Multiple R-squared:  0.8706, Adjusted R-squared:  0.8637
F-statistic: 126.5 on 5 and 94 DF,  p-value: < 2.2e-16

> summary(emmeans(lm.d,"treat",contr="pairwise",weights="proportinal"),infer=TRUE) #this will work
NOTE: Results may be misleading due to involvement in interactions
$`emmeans`
 treat       emmean       SE df lower.CL upper.CL t.ratio p.value
 control   307.6069 2.410992 94 302.8198 312.3940 127.585  <.0001
 treatment 318.2394 3.358040 94 311.5719 324.9068  94.769  <.0001

Confidence level used: 0.95

$contrasts
 contrast             estimate       SE df  lower.CL  upper.CL t.ratio p.value
 control - treatment -10.63247 4.133922 94 -18.84047 -2.424475  -2.572  0.0117

Confidence level used: 0.95
```

Categorical covariates

- The same principle applies if the covariate happens to be categorical and not continuous
- We can still predict outcomes under all combinations of covariate levels, and treatment assignments
- emmeans treats categorical predictors properly (for causal inference purposes) if we use the weights="proportional" option

ANCOVA

- Additional assumption that there is sufficient overlap (positivity)
- Adjusted means represent predictions based on observed regression slopes
- These predictions can be far outside the “region of common support”

ANCOVA

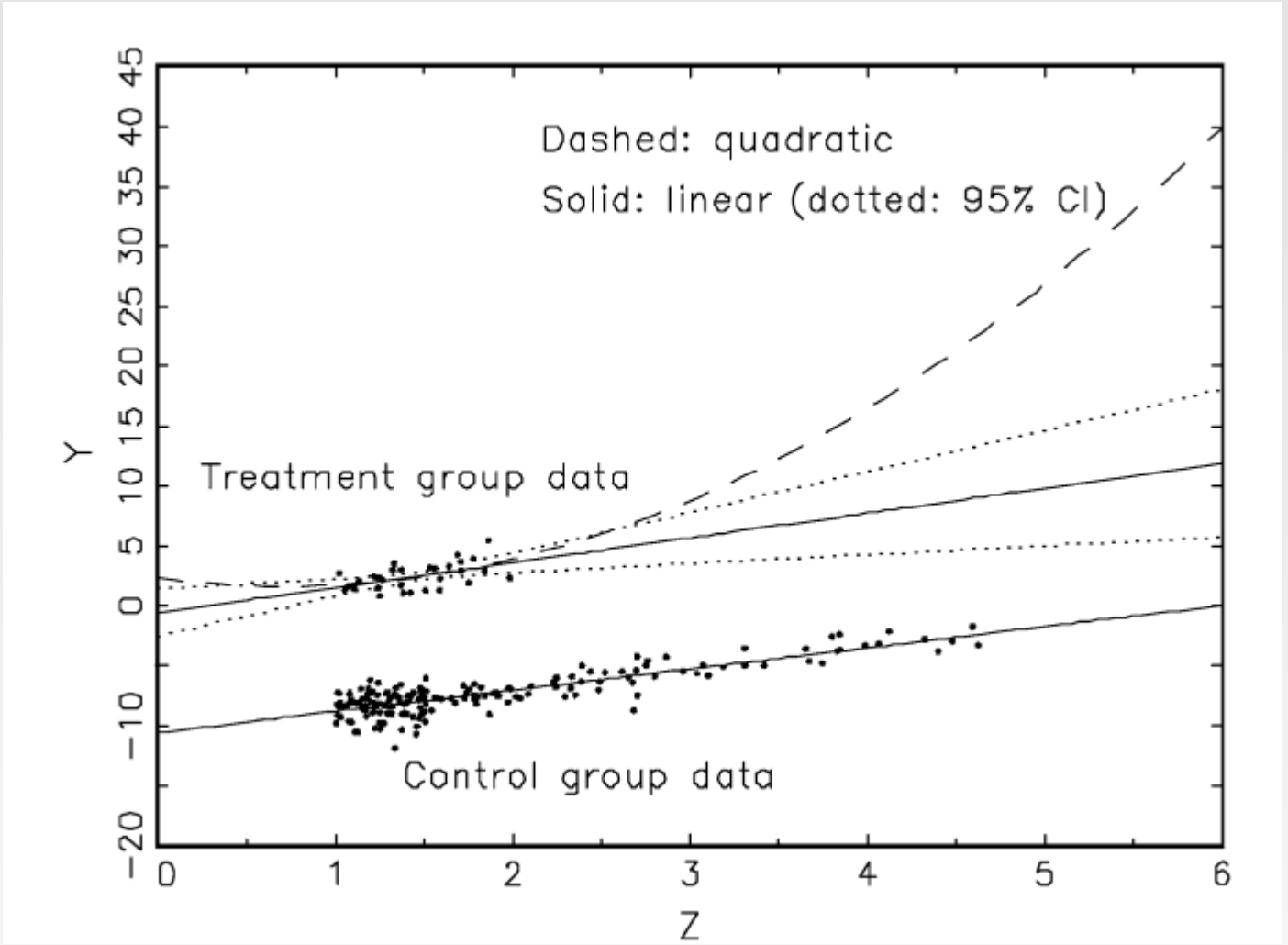
- Imagine adjusting for many variables
- Groups should have some overlap otherwise adjusted means are extrapolated into regions where potentially no data is observed
- One way to confront this is to restrict estimation of a causal effect for region in which there is overlap

Convex hull

- The convex hull is an area that is defined by the outer limits of a point cloud in multi-dimensional space
- We can restrict our analysis to those units that are assigned to one treatment, but still fall in the convex hull of the other units
- That way, we never extrapolate in regions that are completely sparse and only exclusively inhabited by one or the other group

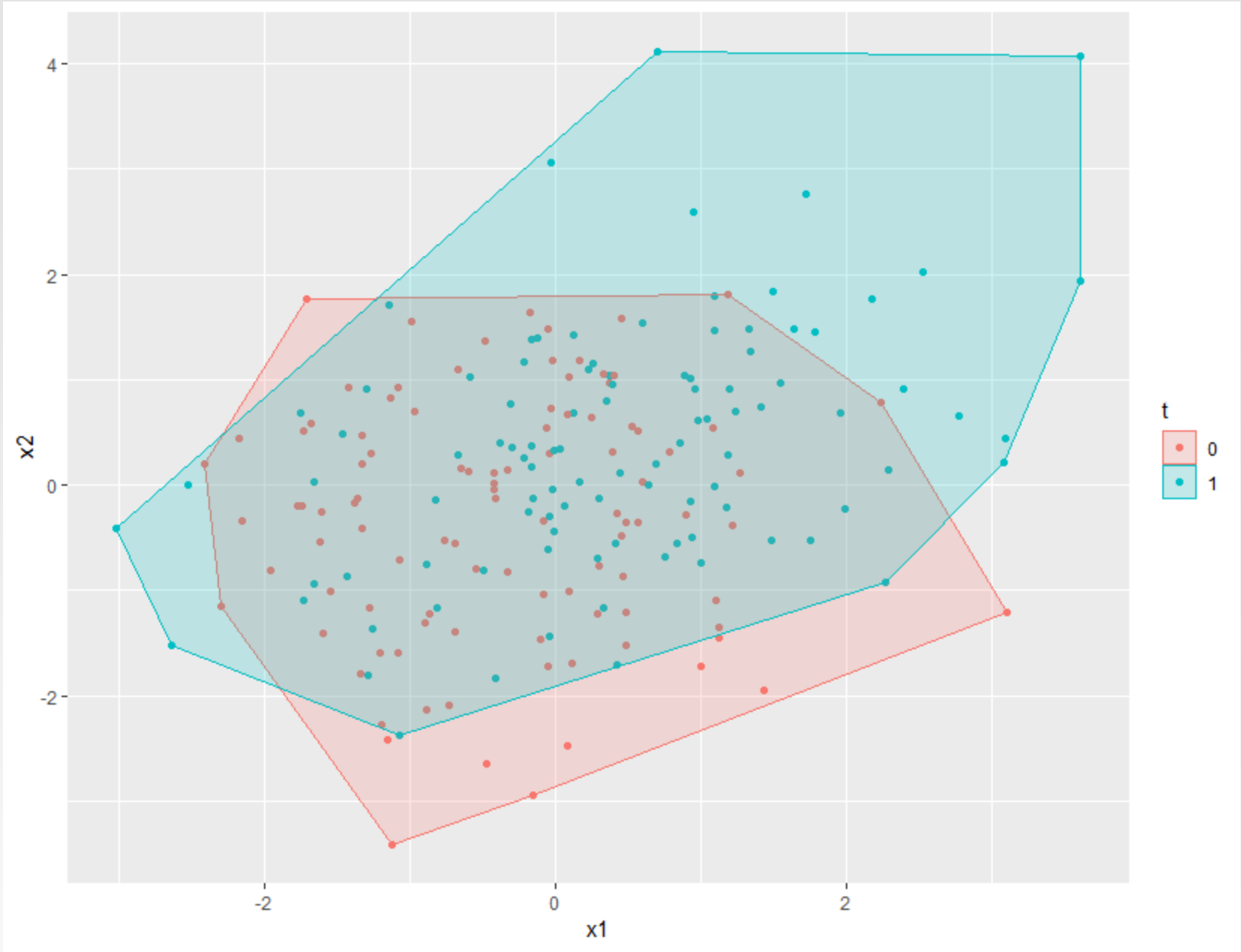
Convex hull

- Gary King gives an example on the causal effect of effects of democracies vs autocracies
- The counterfactual of autocratic Poland in 1990 lies within the range of other democracies
- The counterfactual of democratic Canada in 1995 lies far outside the range of other autocracies

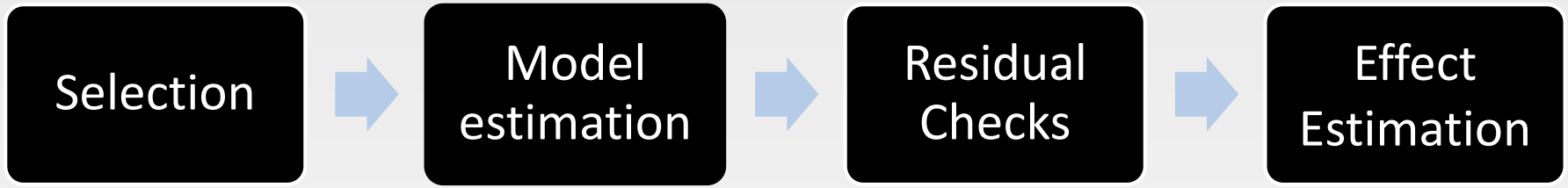

Dashed: quadratic
Solid: linear (dotted: 95% CI)
Treatment group data
Control group data
Y
Z
45
40
35
30
25
20
15
10
5
0
−10
−20
0
1
2
3
4
5
6

Convex hull

- Restricting the analysis to points that lie inside the region of common support reduces model dependency
- That means that slightly mis-specified models do not have very severe consequences (more robust)
- It also forces us to realize that there are points for which causal inference is unstable


4
2
0
-2
x2
-2
0
2
x1
t
0
1

Regression adjustment workflow



Exercise 3

Exercise 3

- Download the “ex3.pdf” file and follow the instructions given in the document

- If you have trouble copying and pasting from a PDF, you can also download the source file “ex3.Rmd”

Matching

Matching

- In the previous section on adjustment we used a parametric model to predict potential outcome

- Instead of using a model, we may choose to try to find pairs of variables that are identical (on covariates) but differ on treatment assignment, so that they can serve as the missing potential outcome for each other

U	Z	$P(U=u)$	$P(X=0\mid U)$	$P(X=1\mid U)$	$\tau_0 = E(Y \mid X=0, U)$	$\tau_1 = E(Y \mid X=1, U)$	$\tau_1 - \tau_0$
u_1	1	1/4	1/4	3/4	0	2	2
u_2	1	1/4	1/4	3/4	0	2	2
u_3	2	1/4	3/4	1/4	0	1/4	1/4
u_4	2	1/4	3/4	1/4	0	1/4	1/4

Z	T	Y0	Y1
1	1	●	2
1	1	●	2
1	1	●	2
1	0	0	●
2	0	0	●
2	0	0	●
2	0	0	●
2	1	●	1/4
		0	1.562

U	Z	$P(U=u)$	$P(X=0 \mid U)$	$P(X=1 \mid U)$	$\tau_0 = E(Y \mid X=0, U)$	$\tau_1 = E(Y \mid X=1, U)$	$\tau_1 - \tau_0$
u_1	1	1/4	1/4	3/4	0	2	2
u_2	1	1/4	1/4	3/4	0	2	2
u_3	2	1/4	3/4	1/4	0	1/4	1/4
u_4	2	1/4	3/4	1/4	0	1/4	1/4

Z	T	Y0	Y1
1	1	●	2
1	0	0	●
2	0	0	●
2	1	●	1/4
		0	1.125

U	Z	$P(U=u)$	$P(X=0\mid U)$	$P(X=1\mid U)$	$\tau_0 = E(Y \mid X=0, U)$	$\tau_1 = E(Y \mid X=1, U)$	$\tau_1 - \tau_0$
u_1	1	1/4	1/4	3/4	0	2	2
u_2	1	1/4	1/4	3/4	0	2	2
u_3	2	1/4	3/4	1/4	0	1/4	1/4
u_4	2	1/4	3/4	1/4	0	1/4	1/4

Z	T	Y0	Y1
1	1	0	2
1	0		
2	0	0	1/4
2	1		
		0	1.125

Matching

- An obstacle to matching is that if we have many covariates (and usually we want that, because otherwise ignorability does not hold), the region of common support gets very small

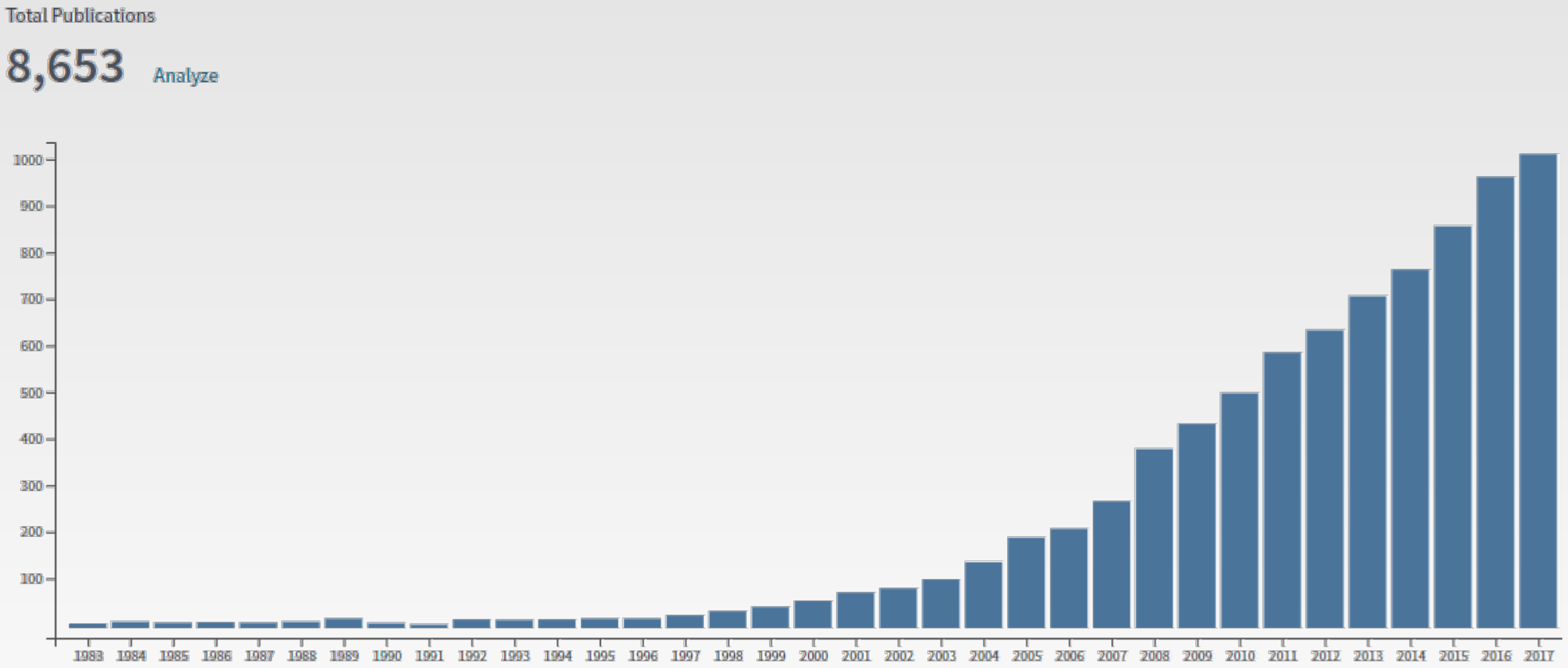
- Same issue with convex hull that restricted adjustment to region of common support

- Some consider this an advantage, because that means with a matching estimator you virtually never extrapolate

Matching

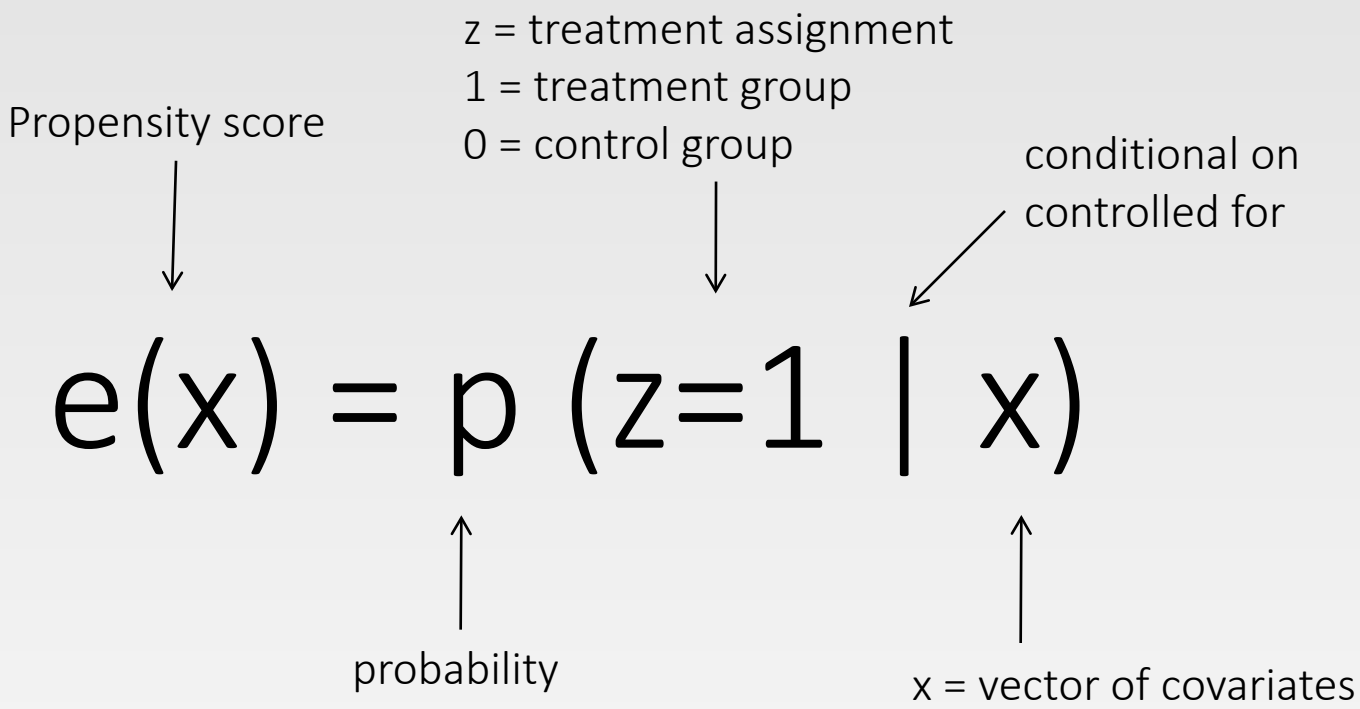
- Consider that we have 16 covariates
- 8 of them are binary which means 2^8 = 256 combinations
- The remaining 8 are continuous and in order to match them, we discretize them into 3 categories each (low, medium, high), so 6561 combinations
- Together we have well over a million different combinations of covariate levels – if at a minimum we want at least 1 treated and 1 untreated person in each combination we need a sample size of at least 2 – 3 millions

Increasing use of Propensity Scores



Source: Web of Science

Propensity scores



Propensity scores

$$e(x) = p\,(z=1 \mid x)$$

A single number summary based on all available covariates that expresses the probability that a given subject is assigned to the treatment condition, based on the values of the set of observed covariates

Propensity scores

Example of balance property

original sample

a	b	z	e(x)
0	0	0	.5
0	0	1	.5
1	0	0	.33
1	0	0	.33
1	0	1	.33
0	1	0	.66
0	1	1	.66
0	1	1	.66
1	1	1	1
1	1	1	1

$e(x) = p(z=1 \mid x=\{0\ 0\}) = .5$

$e(x) = p(z=1 \mid x=\{1\ 0\}) = .33$

$e(x) = p(z=1 \mid x=\{0\ 1\}) = .66$

$e(x) = p(z=1 \mid x=\{1\ 1\}) = 1$

$(a=1 \mid z=0) = .5$ $\quad (b=1 \mid z=0) = 1/4$

$(a=1 \mid z=1) = .5$ $\quad (b=1 \mid z=1) = .66$

Example of balance property

matched sample

a	b	z	e*(x)
0	0	0	.5
0	0	1	.5
1	0	0	.5
1	0	1	.5
0	1	0	.5
0	1	1	.5

(a=1 | z=0) = .5 (b=1 | z=0) = .5

(a=1 | z=1) = .5 (b=1 | z=1) = .5

p(z, x|e(x)) = p(z |e(x)) p(x |e(x))

Examples for z=1 and x = {0 1}

p(z=1, x={0 1}|e(x)) = 1/6

p(z=1 |e(x)) = .5

p(x={0 1} |e(x)) = .33

p(z |e(x)) p(x |e(x)) = (.5)(.33) = 1/6

Propensity scores

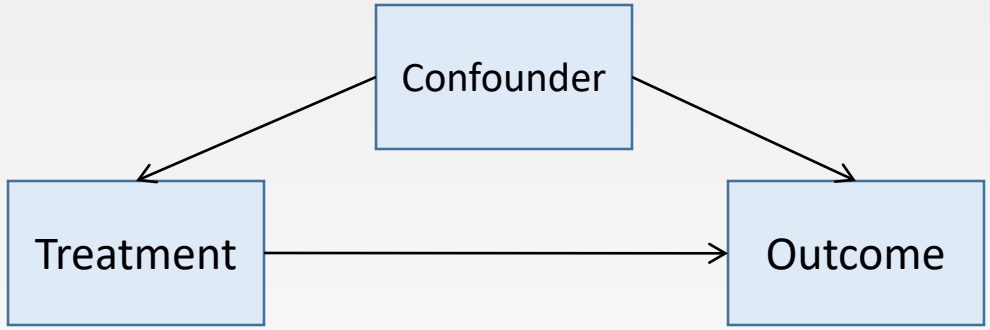
- Balance on the propensity score implies on average balance on all **observed** covariates

- Importantly, PS matching generates balance in the distributions of covariates, and not necessarily for each single pair of unit

Propensity score

- Propensity score models influence of confounders on treatment assignment

- In comparisons, ANCOVA models influence of confounders on outcome



Propensity scores	Regression adjustment
Tool to strengthen causal conclusions	Tool to strengthen causal conclusions
Models relationship between confounders and treatment	Models relationship between confounders and outcome
No assumption about functional form of propensity score	Classic ANCOVA assumes lineartiy and absence of interaction, but can be extended
Outcome variable unknown during propensity score analysis	Outcome variable always part of the adjustment
Sample size can be diminished, loss of power	Sample size stays constant, power can increase due to covariates

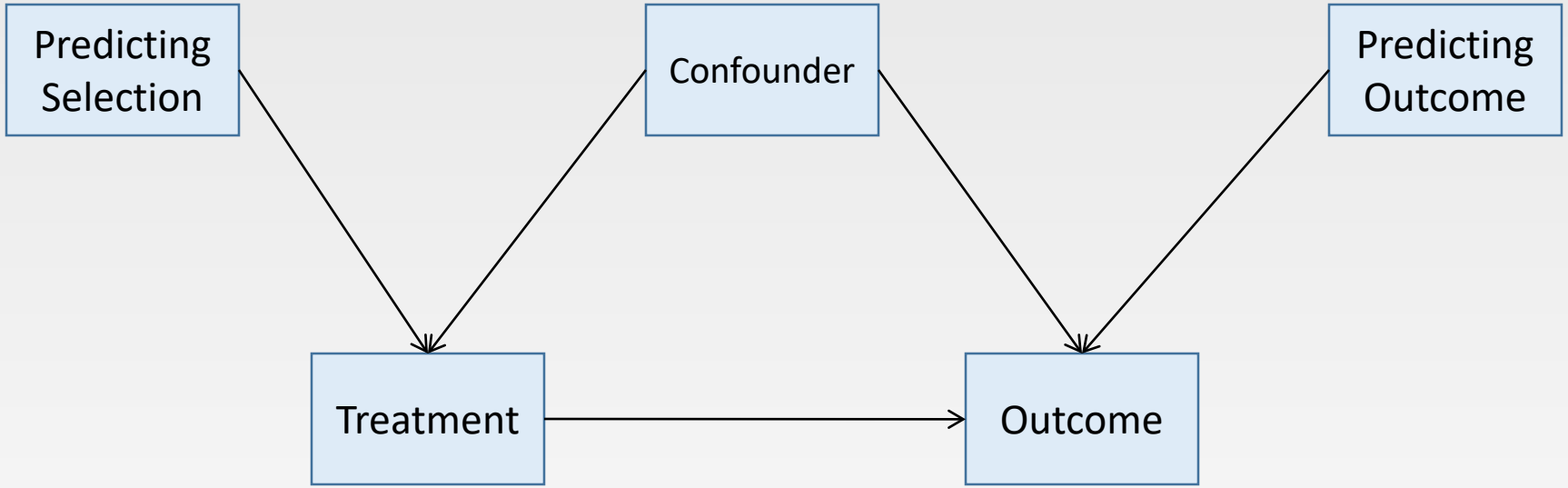
Brainstorm exercise

Matching

- Just like there are modeling choices in the estimation of the outcome model in ANCOVA, so are there choices in the estimation of the propensity score, and the type of matching that is performed

- The propensity score is typically estimated using a logistic regression (predicting treatment assignment from covariates), but one could also use machine learning algorithms

Propensity Score



Select true confounders and covariates predictive of outcome (back-door criterion, ignorability)

Propensity Score

- Estimation of propensity scores can be achieved in numerous ways
  - Logistic regression
  - Discriminant analysis
  - (Boosted) regression trees

Propensity Score

- Logistic regression model
  - Outcome is treatment assignment
  - Predictors are covariates
  - can be overfitted to the sample, e.g. include interactions, higher order terms
  - only interest is prediction and covariate balance

$$Log\left(\frac{e(x)}{1-e(x)}\right) = \beta_0 + \underline{\beta_i}\boldsymbol{X}$$

Propensity Score

- Various matching algorithm (full matching, optimal matching, etc.)
- Too many to discuss, but here are some of the classic ones
- In our exercise, we will mostly use classic matching methods

NN - matching

- Nearest-neighbor matching (NN)
- Participants are ordered and then one after the other is matched to the unit with the closest propensity score
- With or without replacement (sample size bias tradeoff)
- 1:1 or 1:k matching
- With our without caliper (avoids bad matches)

Kernel - matching

- Match every unit in one condition to a single unit in the other condition, but weight the matched units by their distance

- Weighting is defined through the kernel function (essentially bandwidth parameter)

Trade-offs

	Bias	Variance
1:1	↓	↑
1:k	↑	↓
Caliper	↓	↑
No caliper	↑	↓
Replacement	↓	↑
No Replacement	↑	↓
NN	↓	↑
Kernel	↑	↓
Small bandwidth	↓	↑
Large bandwidth	↑	↓

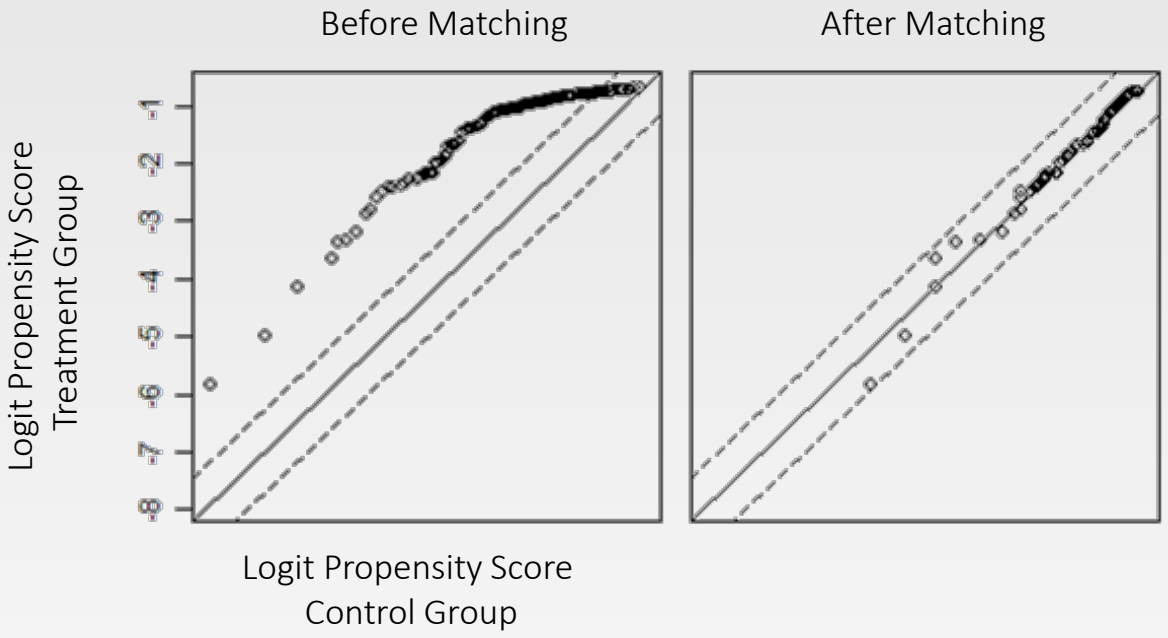
Source: Caliendo and Kopeinig, 2005

Propensity Score

- Check of covariate balance
  - standardized difference
  - graphical assessment (e.g. Q-Q plot)

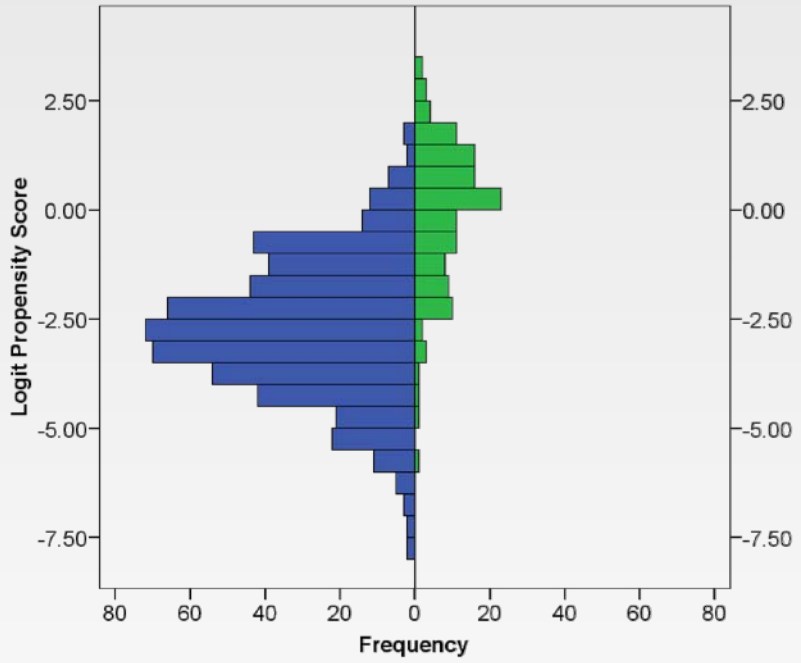
- Region of common support (distributional overlap)
  - graphical assessment (e.g. histograms)

Propensity Score

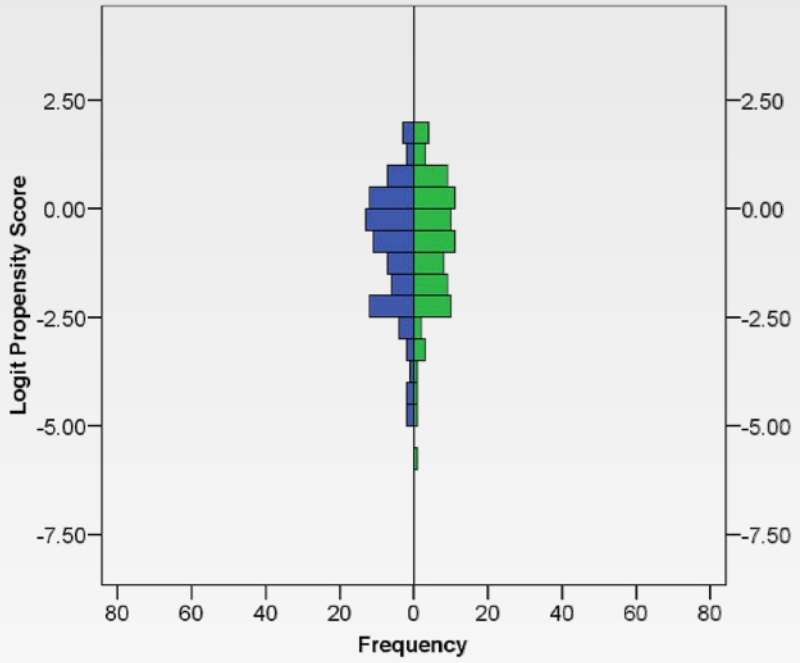


Quantiles of both distributions are plotted against each other

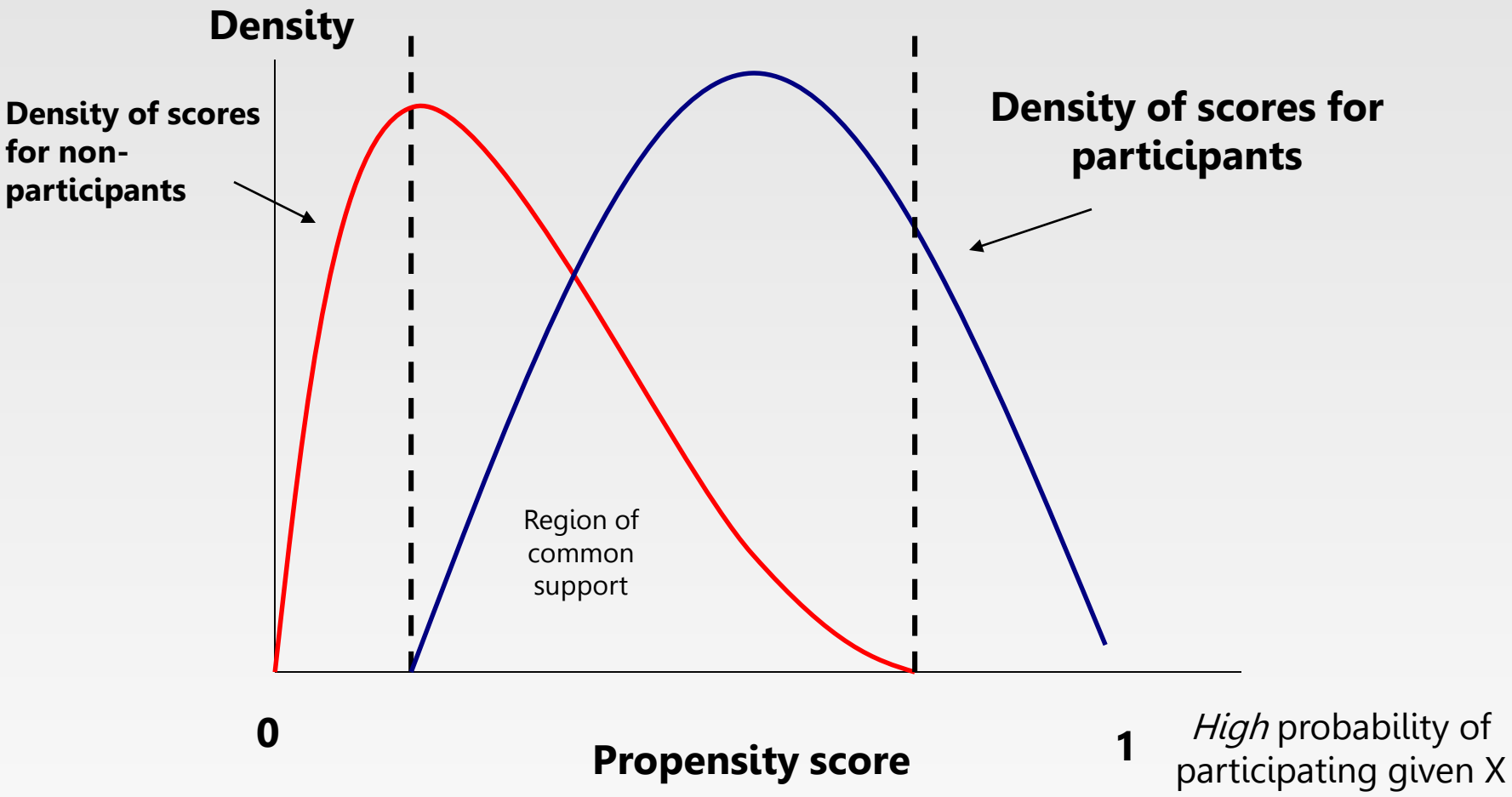
Propensity Score



Before Matching



After Matching


Density
Density of scores for non-participants
Density of scores for participants
Region of common support
0
Propensity score
1
High probability of participating given X

Propensity Score

- Estimate of treatment effect
  - Mean difference
  - Standard error dependent on conditioning scheme (dependent sample standard error or bootstrapping)

Packages

- R has various packages that perform this type of analysis
  - *MatchIt* – general and flexible package for matching estimators
  - *EffectLiteR* – allows inclusion of propensity score in estimation
  - *tlme* – allows inclusion of propensity score in estimation
  - *PSAgraphics* – package for visualization

Propensity Score Workflow

Exercise 4

Weighting

Weighting

- Another idea to adjust for confounding influences is to weight the observed population in a way that creates a “pseudo-population” in which the covariate and the treatment are independent of each other

- In such a pseudo-population, there is no more confounding

- This could mean that one person gets counted 4 times in a sample, while another person only gets counted ¼ of a time

Propensity scores	Regression adjustment	IPTW
Tool to strengthen causal conclusions	Tool to strengthen causal conclusions	Tool to strengthen causal conclusions
Models relationship between confounders and treatment	Models relationship between confounders and outcome	Construct weights based on confounder and treatment
No assumption about functional form of propensity score	Classic ANCOVA assumes lineartiy and absence of interaction, but can be extended	No assumption about functional form of weight equation
Outcome variable unknown during propensity score analysis	Outcome variable always part of the adjustment	Outcome variable unknown during weight construction
Sample size can be diminished, loss of power	Sample size stays constant, power can increase due to covariates	Sample size stays constant, but weights induce uncertainty in estimate

U	Z	$P(U=u)$	$P(X=0\mid U)$	$P(X=1\mid U)$	$\tau_0=E(Y\mid X=0,U)$	$\tau_1=E(Y\mid X=1,U)$	$\tau_1-\tau_0$
u_1	1	1/4	1/4	3/4	0	2	2
u_2	1	1/4	1/4	3/4	0	2	2
u_3	2	1/4	3/4	1/4	0	1/4	1/4
u_4	2	1/4	3/4	1/4	0	1/4	1/4

Z	T	Y0	Y1
1	1	●	2
1	1	●	2
1	1	●	2
1	0	0	●
2	0	0	●
2	0	0	●
2	0	0	●
2	1	●	1/4
		0	1.562

U	Z	$P(U=u)$	$P(X=0 \mid U)$	$P(X=1 \mid U)$	$\tau_0 = E(Y \mid X=0, U)$	$\tau_1 = E(Y \mid X=1, U)$	$\tau_1 - \tau_0$
u_1	1	1/4	1/4	3/4	0	2	2
u_2	1	1/4	1/4	3/4	0	2	2
u_3	2	1/4	3/4	1/4	0	1/4	1/4
u_4	2	1/4	3/4	1/4	0	1/4	1/4

Z	w	T	Y0	Y1
1	1.33	1	●	2
1	1.33	1	●	2
1	1.33	1	●	2
1	4.00	0	0	●
2	1.33	0	0	●
2	1.33	0	0	●
2	1.33	0	0	●
2	4.00	1	●	1/4
			0	1.125

$((2 \times 1.33) \times 3 + (.25 \times 4)) / 8 = 1.125$

$(0 * 1.33) \times 3 + (0 \times 4) / 8 = 0$

Weighting

- Each unit is weighted by the inverse of the probability of being assigned to its treatment, given the covariate values

$$w_i = \frac{1}{P(A_i = a_i | \boldsymbol{C}_i = \boldsymbol{c}_i)}.$$

- The denominator of this formula is the propensity score

Weighting

- A unit that has covariate values that make it very likely to be in the actual assigned group ends up getting a small weight
- A unit that has covariate values that make it very unlikely to be in the actual assigned group ends up getting a large weight
- Rare units in the control are up-weighted, and common units down-weighted, and likewise in the treatment
- This creates pseudo-populations in which the covariates are equally distributed

Statistical aspects of weighting

- Using weights can have some undesirable consequences in small samples

- Very rare units can get tremendous weight, and thus making the estimate highly dependent on that particular unit

Statistical aspects of weighting

- The use of so-called stabilized weights, and weight truncation is encouraged

$$sw_i = \frac{P(A_i = a_i)}{P(A_i = a_i | \boldsymbol{C}_i = \boldsymbol{c}_i)}.$$

U	Z	$P(U=u)$	$P(X=0 \mid U)$	$P(X=1 \mid U)$	$\tau_0 = E(Y \mid X=0, U)$	$\tau_1 = E(Y \mid X=1, U)$	$\tau_1 - \tau_0$
u_1	1	1/4	1/4	3/4	0	2	2
u_2	1	1/4	1/4	3/4	0	2	2
u_3	2	1/4	3/4	1/4	0	1/4	1/4
u_4	2	1/4	3/4	1/4	0	1/4	1/4

Z	w	T	Y0	Y1
1	.666	1	●	2
1	.666	1	●	2
1	.666	1	●	2
1	2.00	0	0	●
2	.666	0	0	●
2	.666	0	0	●
2	.666	0	0	●
2	2.00	1	●	1/4
			0	1.125

$((2 \times .666) \times 3 + (.25 \times 2)) / 4 = 1.125$

$(0 * .666) \times 3 + (0 \times 2) / 4 = 0$

Note that stabilized weights preserve the sample size.

Brainstorm exercise

Statistical aspects of weighting

- Truncation of the weights means that very large weights are scaled down to some pre-specified percentile
- This essentially protects against outliers
- Common used percentiles are the 1% and 99% percentile

Statistical aspects of weighting

- The outcome analysis is then performed by using a parametric model with the weights
- E.g., weighted least squares
- Because the estimation of the weights is itself subject to random variability, it is recommended to use robust standard errors for frequentist inference

Statistical aspects of weighting

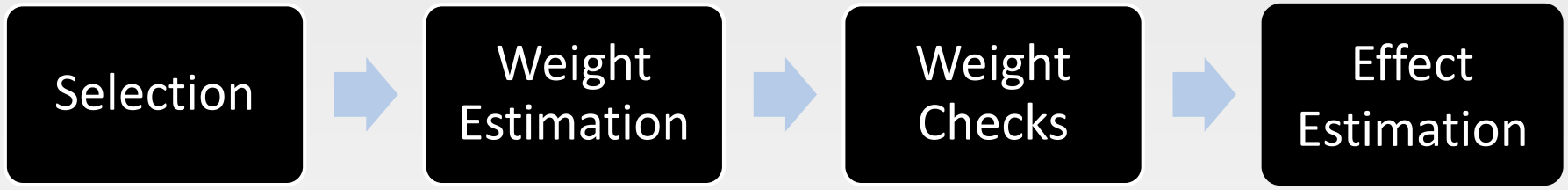
- What makes inverse-probability weighting such a powerful technique is that it can also be used for time-varying treatments

- The use of these weights in longitudinal data is in the literature often referred to as a marginal structural model

Packages

- R has one package that perform this type of analysis
    - *ipw* – general and flexible package for weighting estimators

Weighting workflow



Exercise 5

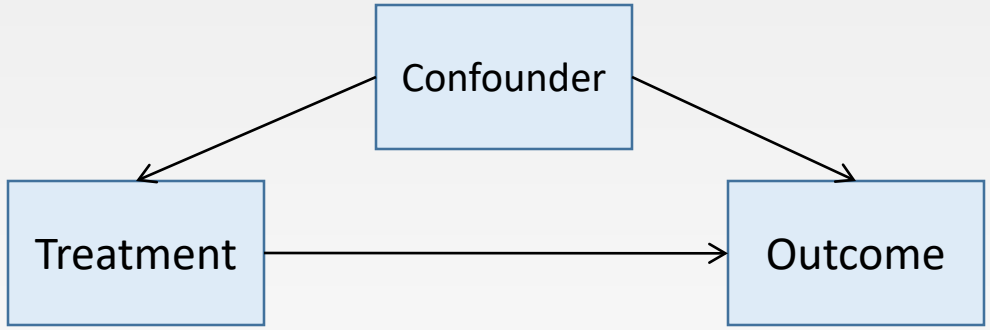
Combining methods

Combining methods

- The discussed methods can be used in conjunction with each other
- It is possible to first match participants, and then use regression adjustment to model the outcome
- Likewise it is also possible to construct weights, and then use regression adjustment to model the outcome

Combining methods

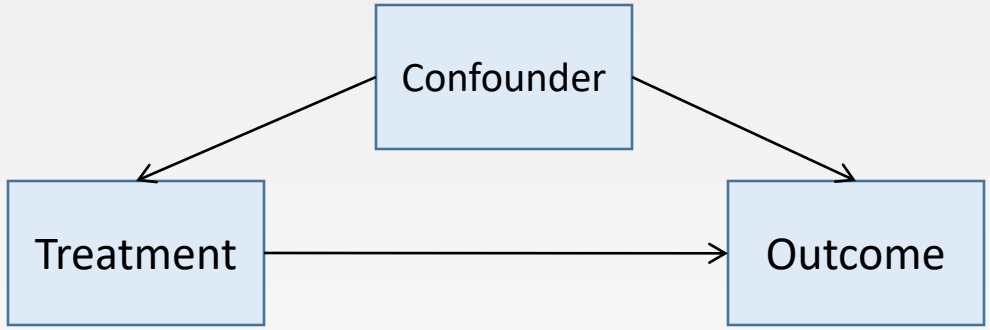
- An interesting property of this approach is that BOTH the relationship between confounders and treatment assignment AND the relationship between confounders and the outcome is modeled



Combining methods

- It turns out that if only one of the two models is correct, we still get unbiased estimates

- This property is known as “doubly-robust”



Exercise 6

Other identification strategies

Selection on observables

- So far all of our models were based on the assumption that we can achieve ignorability (back-door criterion)

- But what if that assumption is not really credible?

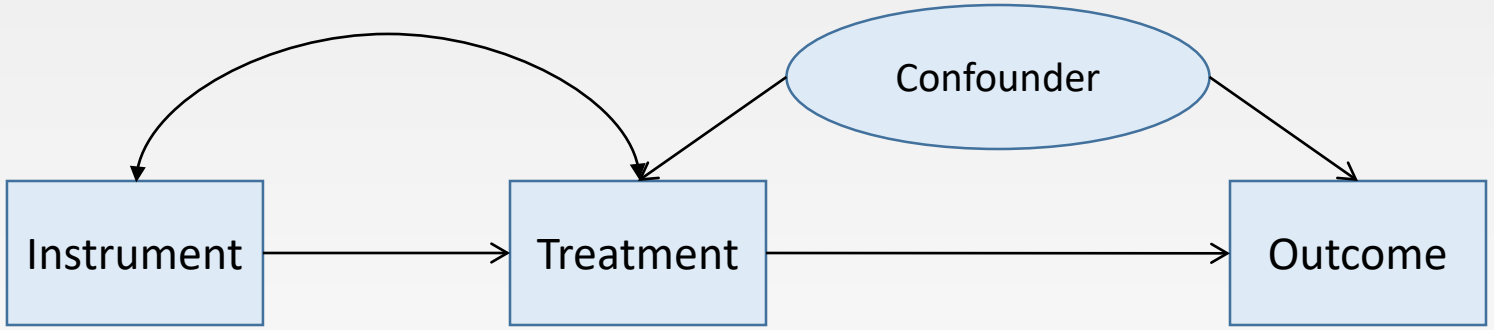
Other assumptions

- We can try to make other causal assumptions that do not rely on having observed all confounders

- Whether these assumptions are more or less plausible is always a matter of theory and debate

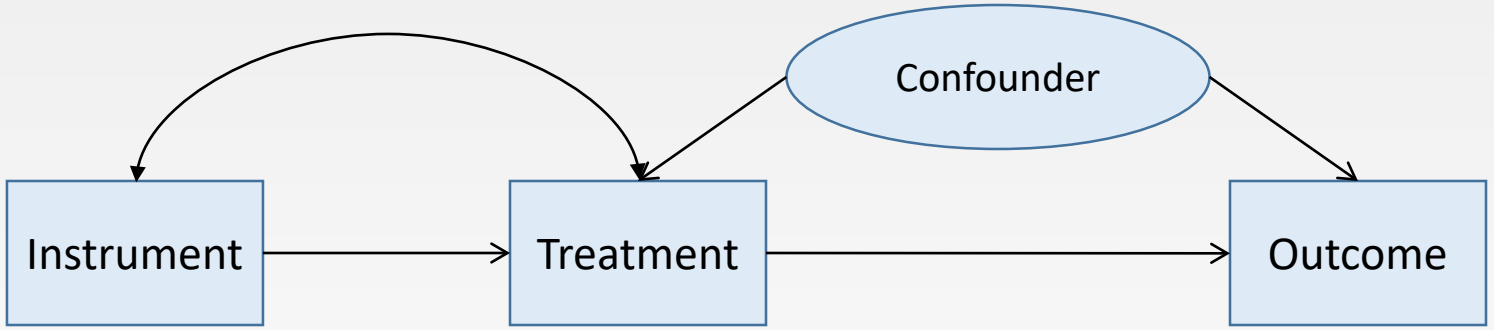
Instrumental variables

- One such causal assumption is that we are able to identify one (or more) variables that are so-called *instruments*



Instrumental variables

- An instrument has an effect on the treatment (may or may not be causal)
- An instrument is unrelated to the confounders
- An instrument does not have a direct effect on the outcome (exclusion restriction)



Instrumental variables

- Examples:
  - Randomized treatment assignment in which not everybody responds to treatment. Smoking cessation program -> Smoking frequency -> birth weight
  - Distance from hospital as instrument for causal relationship between home birth and birth complications

Instrumental variables

- IV estimates identify a causal effect for individuals “who can be induced to change [treatment] status by a change in the instrument”

- Local treatment effect is instrument-dependent, meaning that one and same effect with different instruments will give different results

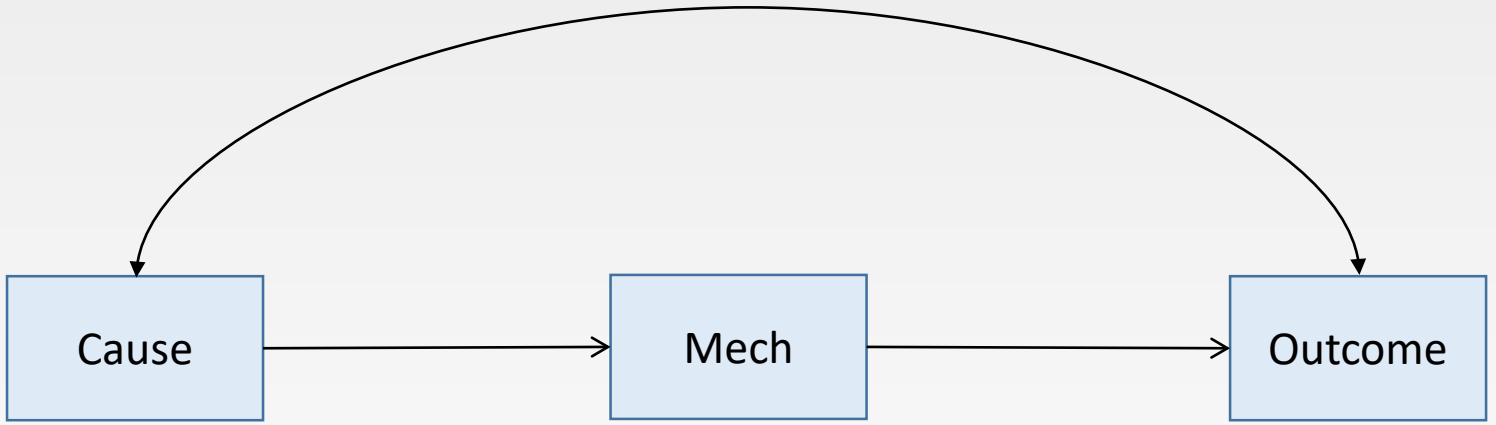
Instrumental variables

- IV can be estimated through 2-stage least-squares regression or using structural equation models
- It is customary to statistically compare estimates from an unadjusted model with those from the IV model (Hausman test)
- Also report the significance of the relationship between instrument and putative cause
- If more than one IV is available, then one can test whether the IV is correlated with the error of the outcome (Sargan test)

Brainstorm exercise

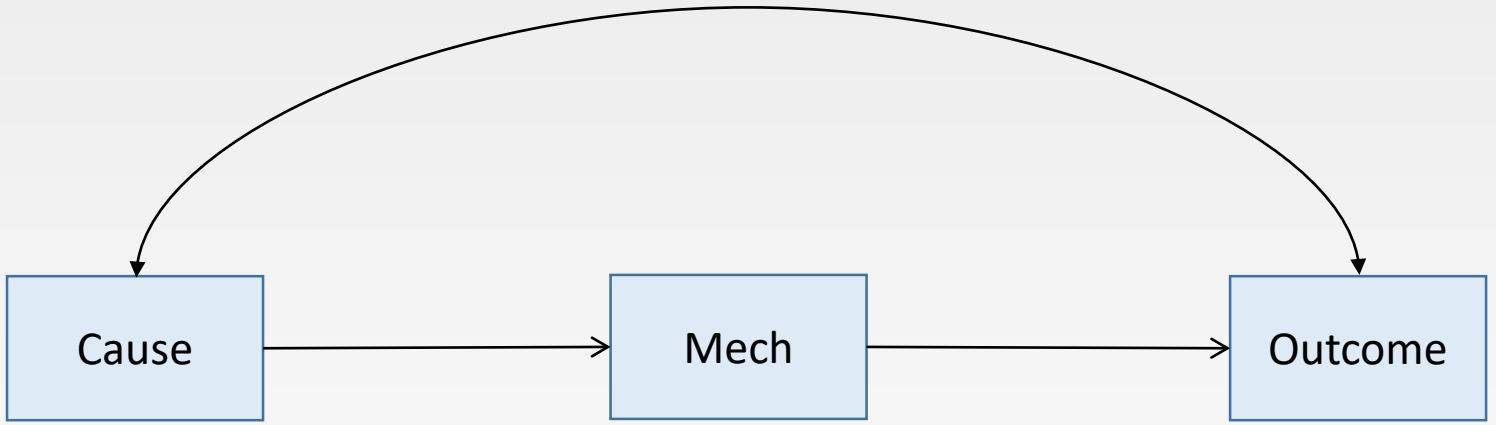
Mechanisms

- Similar in spirit to instrumental variables, one may find a unique mechanism (psychologists would call it full mediation)



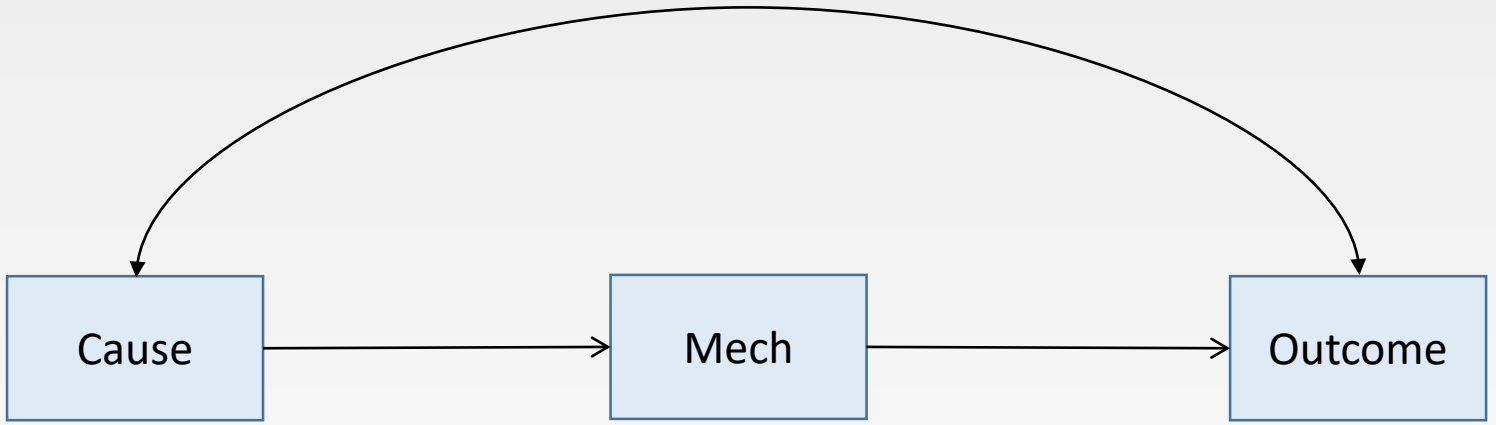
Mechanisms

- Despite unobserved confounding between the cause and the outcome, the presence of a mechanism allows us to compute the total causal effect



Mechanisms

- The mechanism must not be related to the unobserved confounders
- The cause must not directly cause the outcome, only via the mechanism



Brainstorm exercise

Statistical aspects of mechanisms

- One way to estimate the causal effect via a mechanism is to simply use a structural equation model (path model)
- We define a full mediation model, but allow the error terms of treatment and outcome to be correlated
- This yields a just-identified model from which we can derive the total effect as the product of the two paths (assuming linear models)

Exercise 7


Here be dragons
Calandradua

Complications

- Latent variables
- Missing data
- Machine learning
- Longitudinal data
- Indirect effects

Latent variables

- All of our examples so far used manifest variables
- What if some of our constructs are latents and we want to model them this way
- Use of full structural equation model (with measurement model) necessary, but many questions remain
- *EffectLiteR* currently only package that tackles these issues

Missing data

- All of our examples assumed complete data
- Missing data can induce unique biases even in perfectly unconfounded effects
- Identification of missingness mechanism (m-graphs) and techniques to recover effects (imputation) rely on additional assumptions

Machine learning

- All of our exercises were fitted “manually”, entering covariates and doing some model checks
- In complex models this seems hopeless
- Machine learning algorithms could potentially help here
- Some success stories with Bayesian Regression Trees, and SuperLearner (in *tmle* package)

Longitudinal data

- All of our exercises considered a treatment being administered at one point in time
- Longitudinal studies with time-varying treatments, and time-varying confounders pose additional problems
- Regression adjustment and matching often infeasible, weighting and marginal structural models possible (in *ipw* package)

Indirect effects

- All of our exercises considered the total effect of a treatment
- Sometimes we are interested in mechanisms of a treatment (opening the black box)
- This translates into indirect (mediated) effects
- Flurry of literature on this topic, incuding necessary assumptions, estimation strategies, and designs (in *Mediation* package)

Summary

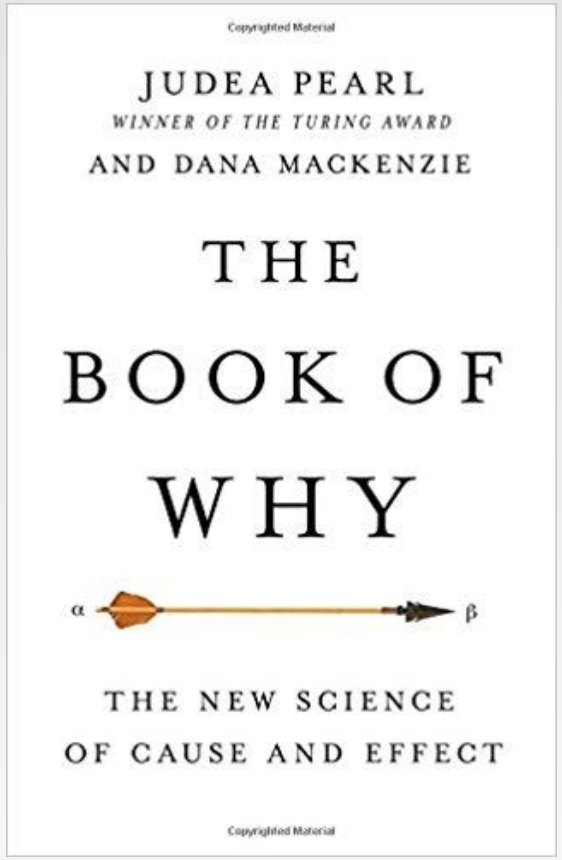
Respect assumptions

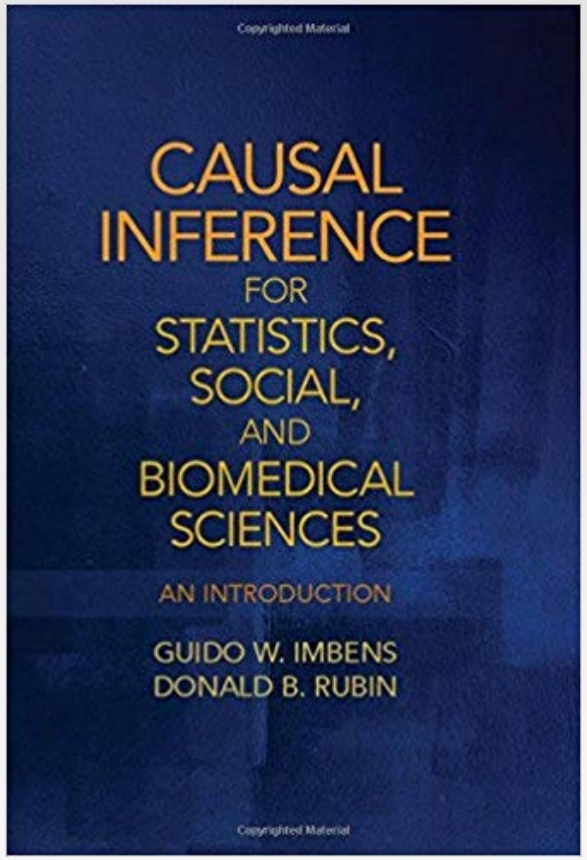
Causal conclusions are hard to get

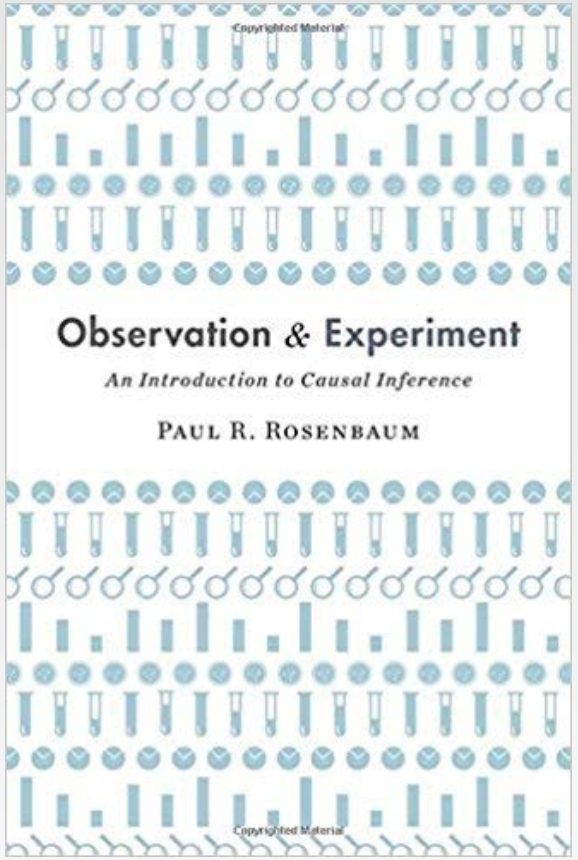
Many roads lead to Rome

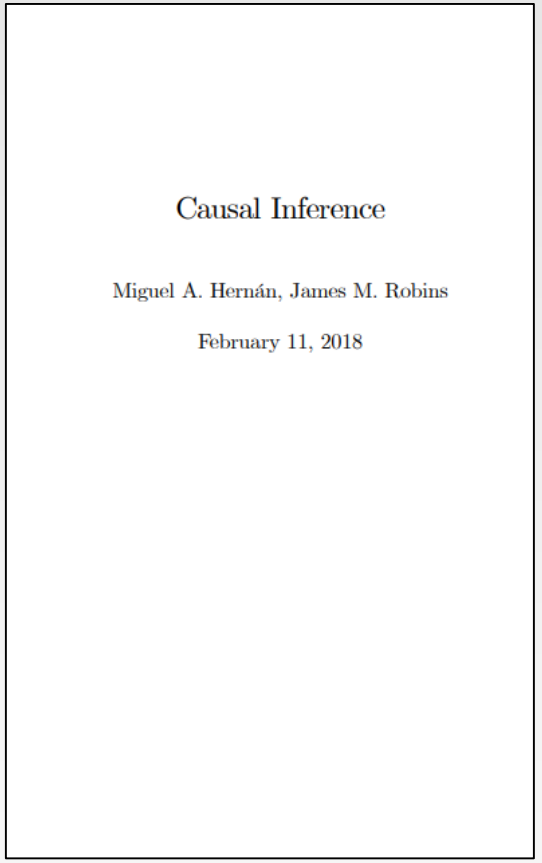
Brainstorm exercise

References


Copyrighted Material
JUDEA PEARL
WINNER OF THE TURING AWARD
AND DANA MACKENZIE
THE
BOOK OF
WHY
α
β
THE NEW SCIENCE
OF CAUSE AND EFFECT
Copyrighted Material


Copyrighted Material
CAUSAL
INFERENCE
FOR
STATISTICS,
SOCIAL,
AND
BIOMEDICAL
SCIENCES
AN INTRODUCTION
GUIDO W. IMBENS
DONALD B. RUBIN
Copyrighted Material


Copyrighted Material
Observation & Experiment
An Introduction to Causal Inference
PAUL R. ROSENBAUM
Copyrighted Material


Causal Inference
Miguel A. Hernán, James M. Robins
February 11, 2018

felix.thoemmes@cornell.edu

felixthoemmes.com

 @felixthoemmes