PSYCH 252: Statistical Methods for Behavioral and Social Sciences

Linear mixed effects models 1

02/21/2024

Logistics

Cannot access GitHub classroom #48

Howard Chiu
3 days ago in **Final Project**

PIN STAR WATCH 51 VIEWS

Apologies if I missed the announcement, but I just tried the link for the GitHub classroom and it's still not working for me.

Anyone else having the same issue?

Thanks!

Comment Edit Delete Endorse •••

2 Answers

Tobi Gerstenberg STAFF
3 days ago

Hi Howard,

I'm sorry to hear that it's not working. Can you post a screenshot of what error message you're getting? Here is the signup link: https://classroom.github.com/a/twW5Gcpf

Cheers,

Tobi

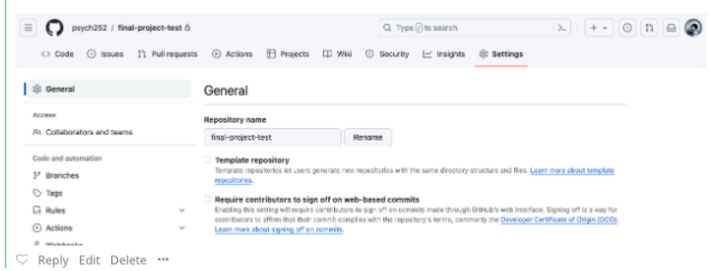
Comment Edit Delete Endorse •••

Add comment

Howard Chiu 1h
Hi Tobi,

I was wondering if you'd like us to just clone the repository locally, or that we'd eventually have to commit our files to the GitHub Classroom?

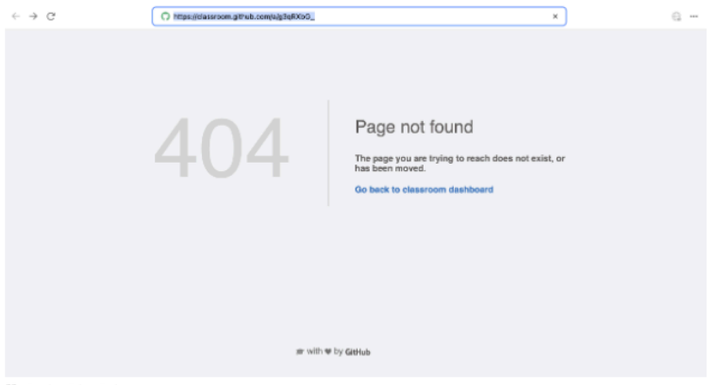
I was trying to make a fork but noticed that the settings to do so were not enabled.



Reply Edit Delete •••

Howard Chiu 3d
Thanks, the direct link worked for me!

I was otherwise getting a page not found from the Canvas assignment link.



Reply Edit Delete •••

Tobi Gerstenberg STAFF 3d
Great! And I realized that I hadn't updated the link on the assignment on canvas (I've done that now). Thanks a lot for posting here!

Reply Edit Delete •••

Homework 5

available after class today

due Thursday, February 19th, 8pm

Part 1: Causal graphs (2 points)

For each graph, determine whether different variables are independent of each other. In addition to writing *Yes* or *No*, please also write the resulting graph of performing d-separation by listing the vertices and edges in alphabetical order, ignoring redundancies. For instance, the graph below in *Figure 1* would be described as

Vertices: A, B, C, D

Edges: A-B, A-C, A-D, B-D

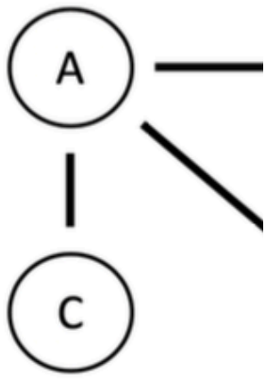


Figure 1: Figu

Part 2: Correlation constrains Causation (2.5 points)

It's intuitive to believe that additional years of compulsory education would increase yearly earnings, but a causal relationship is difficult to establish due to both practical and ethical concerns of randomly assigning years of required education. In this section, we will explore the dataset of a seminal work in economics by Angrist and Krueger that established a causal link between education and income through some very clever use of an instrumental variable. Here's the first paragraph from the paper.

> Every developed country in the world has a compulsory schooling requirement, yet little is known about the effect these laws have on educational attainment and earnings. This paper exploits an unusual natural experiment to estimate the impact of compulsory schooling laws in the United States. The experiment stems from the fact that children born in different months of the year start school at different ages, while compulsory schooling laws generally require students to remain in school until their sixteenth or seventeenth birthday. In effect, the interaction of school-entry requirements and compulsory schooling laws compel students born in certain months to attend school longer than students born in other months. Because one's birthday is unlikely to be correlated with pers
> exogenous variation i
> on education and ea

```
df.qob = read_tsv("dat
  head(50000) # We don
head(df.qob) %>%
  kable()
```

log_weekly_wage	educ
5.790019	
5.952494	
5.315949	
5.595926	
6.068915	
5.793871	

Part 3: Cross-Validation (11 points)

In this section, we will be using some US Census data to find relationships between demographic data and average income within cities.

Although there are many different models we could construct from all these variables, we will be focusing on just the following 4 models:

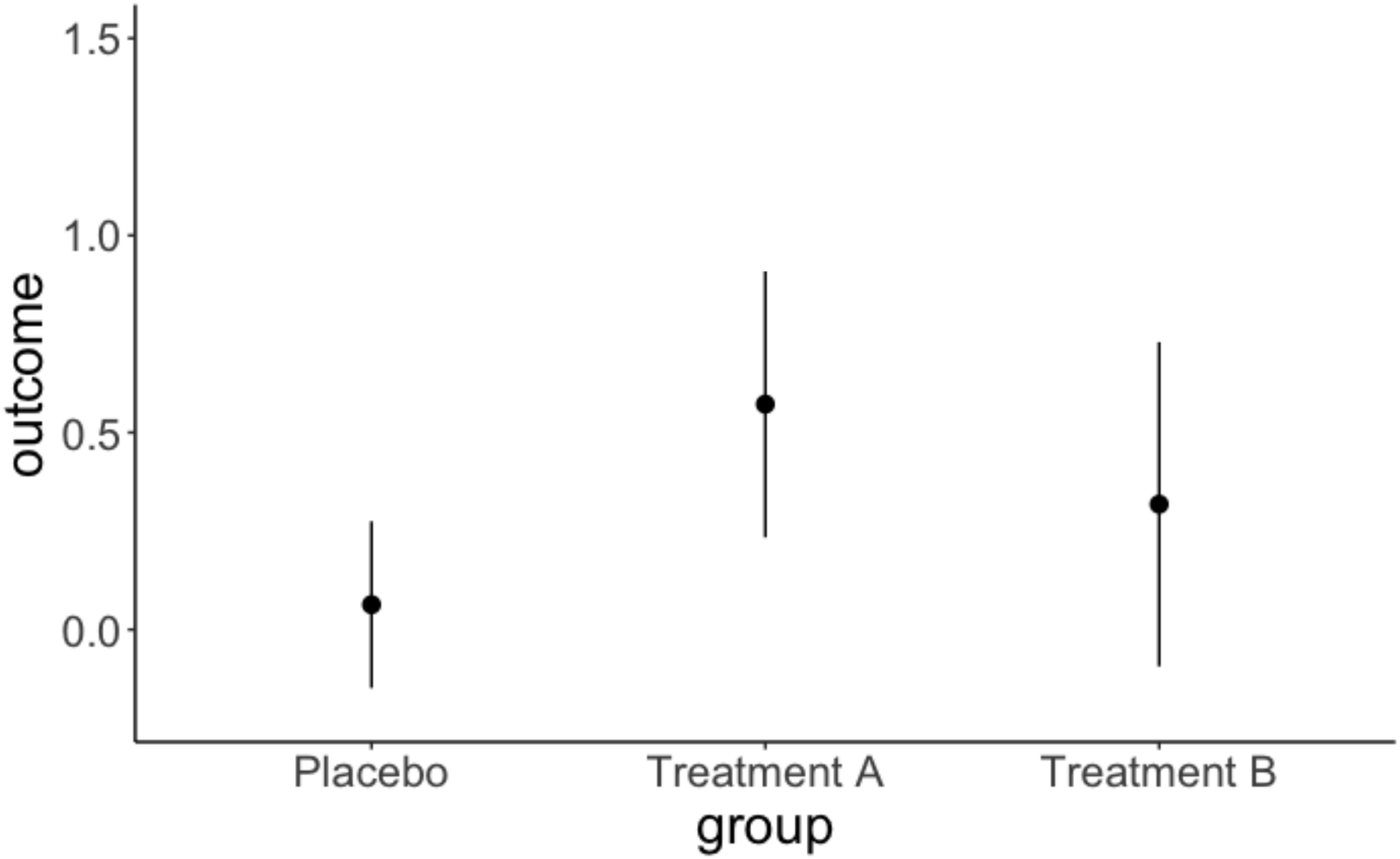
1. `model_med_age: mean_income ~ median_age`
2. `model_med_age10: mean_income ~ median_age + median_age^2 + median_age^3 + ... + median_age^10`
3. `model_edu: mean_income ~ less9thgrade + grade9to12 + highschool + somecollege + assoc + bachelors + grad`
4. `model_race: mean_income ~ percent_white + percent_black + percent_amindian_alaskan + percent_asian + percent_nativeandother + percent_other_nativeandother + percent_hisp + percent_race_other`

Things that came up

difference in significance vs. significant differences

significant differences

"Our tendency to look for a difference in significance should be replaced by a check for the significance of the difference."



“We compared treatments A and B with a placebo. Treatment A showed a significant benefit over placebo, while treatment B had no statistically significant benefit. Therefore, treatment A is better than treatment B.”

https://www.statisticsdonewrong.com/significant-differences.html

significant differences

Significant difference between Group 1 and 2?

what do the error bars mean?

standard deviation?

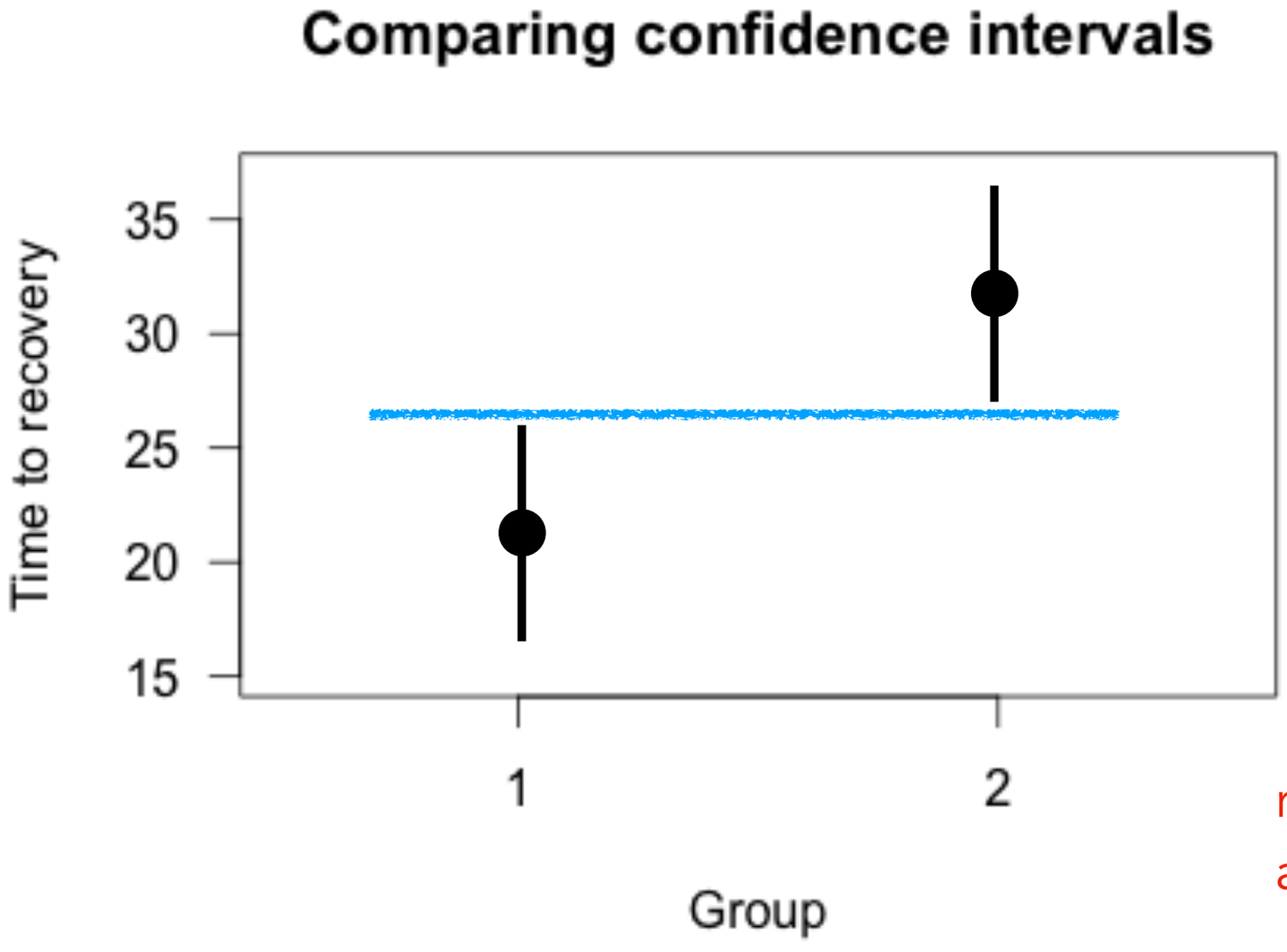
standard error of the mean?

confidence intervals?

What we would like to know, does the confidence interval of the difference between groups exclude 0?

significant differences

Significant difference between Group 1 and 2?



95% confidence intervals don't overlap

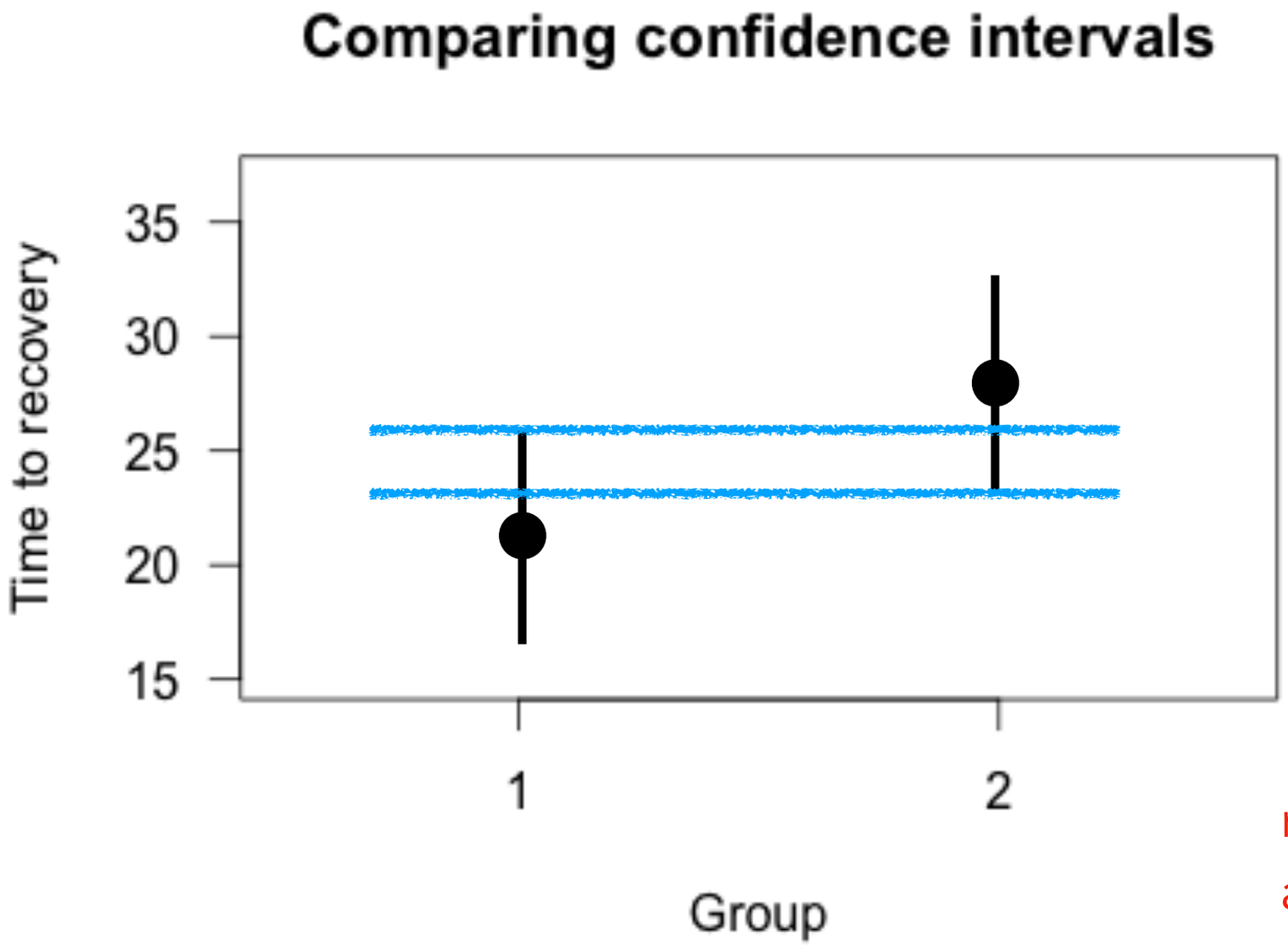
conservative test of significance

rejects H_0 **less often** than the appropriate statistical procedure

Schenker & Gentleman (2001) On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*

significant differences

Significant difference between Group 1 and 2?



95% confidence intervals don't overlap with mean

anti-conservative test of significance

rejects H_0 **more often** than the appropriate statistical procedure

Schenker & Gentleman (2001) On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*

Plan for today

- Quick recap
- Controlling for variables
- Mediation
- Moderation
- Linear mixed effects model
  - Modeling dependence in data

Quick recap

Quick recap: Model comparison

More complex models fit the data better.

We need to trade-off **model fit** and model **complexity**.

Quick recap: Model comparison

1. compare the proportional reduction in error using `anova()`
    - only works for nested models
2. cross-validation
    - works generally, but can take some time ...
    - different cross-validation procedures (LOO, k-fold, Monte Carlo)
3. AIC, BIC
    - works as long as we can compute the likelihood of the data
    - some discussion whether number of parameters is a good measure of model complexity
4. Bayesian model comparison

Quick recap: AIC and BIC

- AIC = Akaike Information Criterion
- BIC = Bayesian Information Criterion

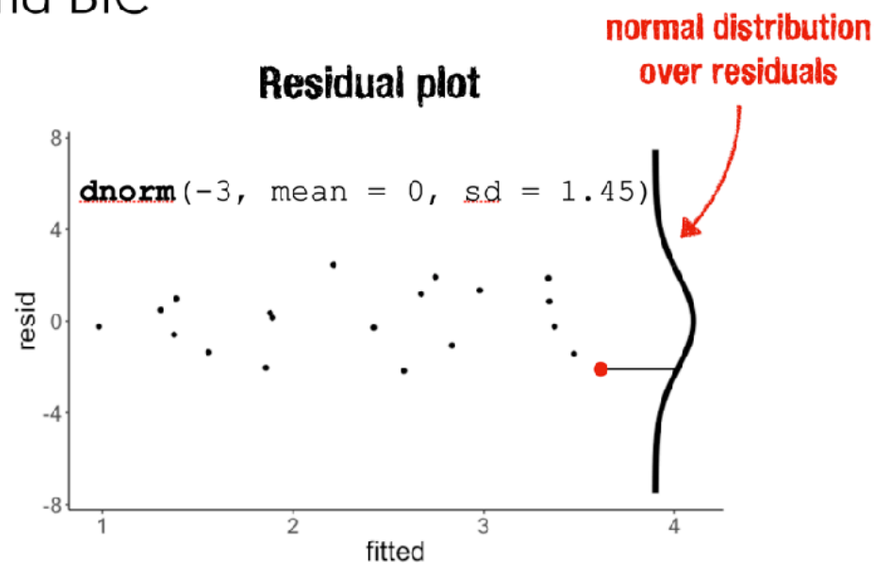
not that much Bayesian about it ...

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

$$\text{BIC} = \ln(n)k - 2\ln(\hat{L})$$

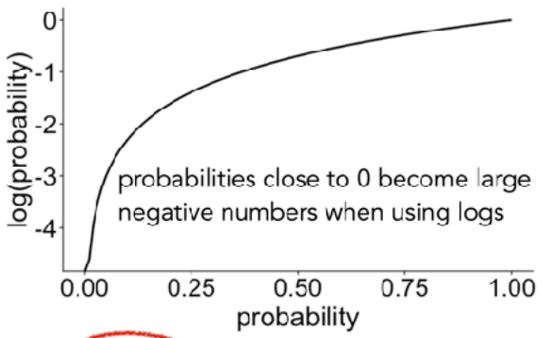
$\hat{L}$ = maximized value of the likelihood function of the model
k = number of parameters in the model
n = number of observations

AIC and BIC



since the data points are independent, we can calculate the overall likelihood by multiplying the likelihood of each observation

log() is your friend!



multiplying probabilities

$$0.01 \cdot 0.01 \cdot 0.01 \cdot 0.01 = 0.00000001$$

number becomes extremely small quickly

take log()

```
log(0.01) = -4.60517
```

number becomes large but that's ok

summing logs

$$(-4.60517) + (-4.60517) + (-4.60517) + (-4.60517) = -18.42068$$

transform back into probability

```
exp(-18.42068) = 0.00000001
```

often not necessary since we just use logLikelihood

AIC and BIC

```
dnorm(1.88, mean = 0, sd = 1.45) = 0.12
```

```
1  # generate some data
2  df.like = tibble(
3    x = runif(20, min = 0, max = 1),
4    y = 1 + 3 * x + rnorm(20, sd = 2)
5  )
6
7  # fit the model
8  fit = lm(formula = y ~ x,
9           data = df.like)
10
11 # model summary
12 fit %>%
13   glance()
```

x	y	fitted	resid	likelihood
0.90	5.22	3.34	1.88	0.12
0.27	0.20	1.56	-1.36	0.18
0.37	-0.18	1.86	-2.04	0.10
0.57	2.14	2.42	-0.28	0.27
0.91	3.13	3.37	-0.24	0.27
0.20	0.78	1.38	-0.59	0.25
0.90	4.20	3.34	0.86	0.23
0.94	2.05	3.47	-1.42	0.17
0.66	3.85	2.67	1.18	0.20
0.63	0.41	2.58	-2.17	0.09

inferred standard deviation of the error

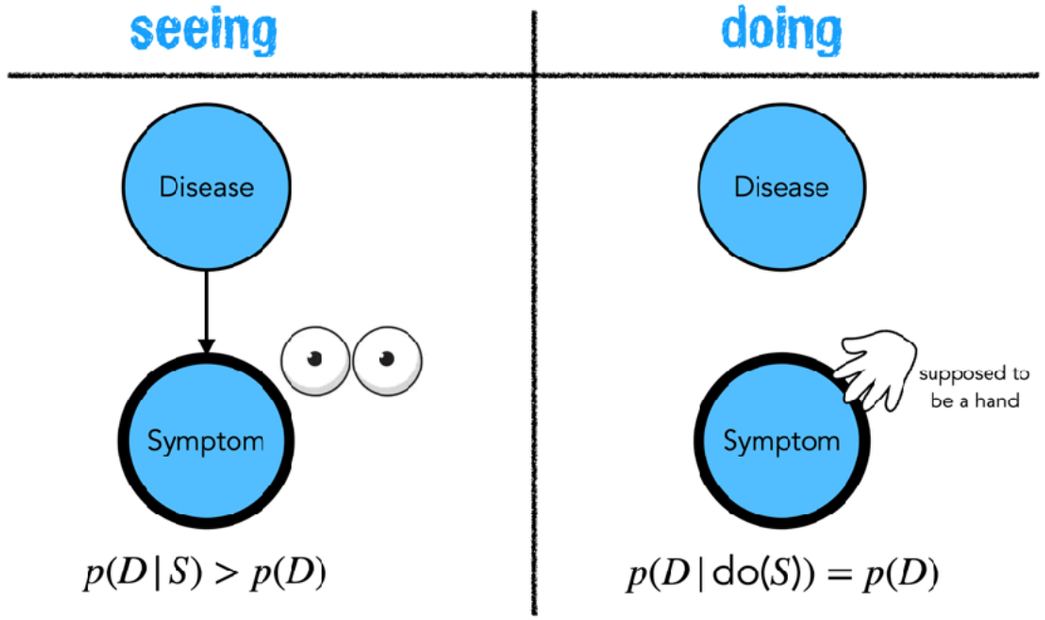
$$\sum_{i=1}^{n} \ln(\text{likelihood})$$

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual
0.25	0.21	1.45	6.16	0.02	2	-34.74	75.47	78.46	37.77	18

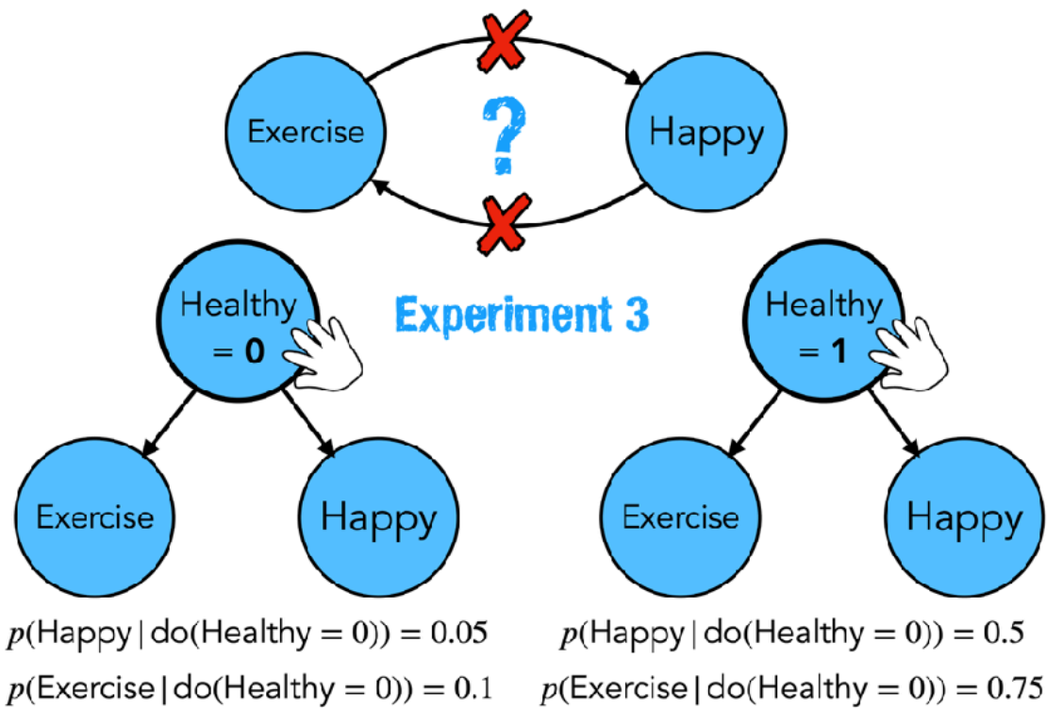
$e \sim \mathcal{N}(\text{mean} = 0, \text{sd} = 1.45)$

Quick recap: causation vs. correlation

Observation vs. Intervention

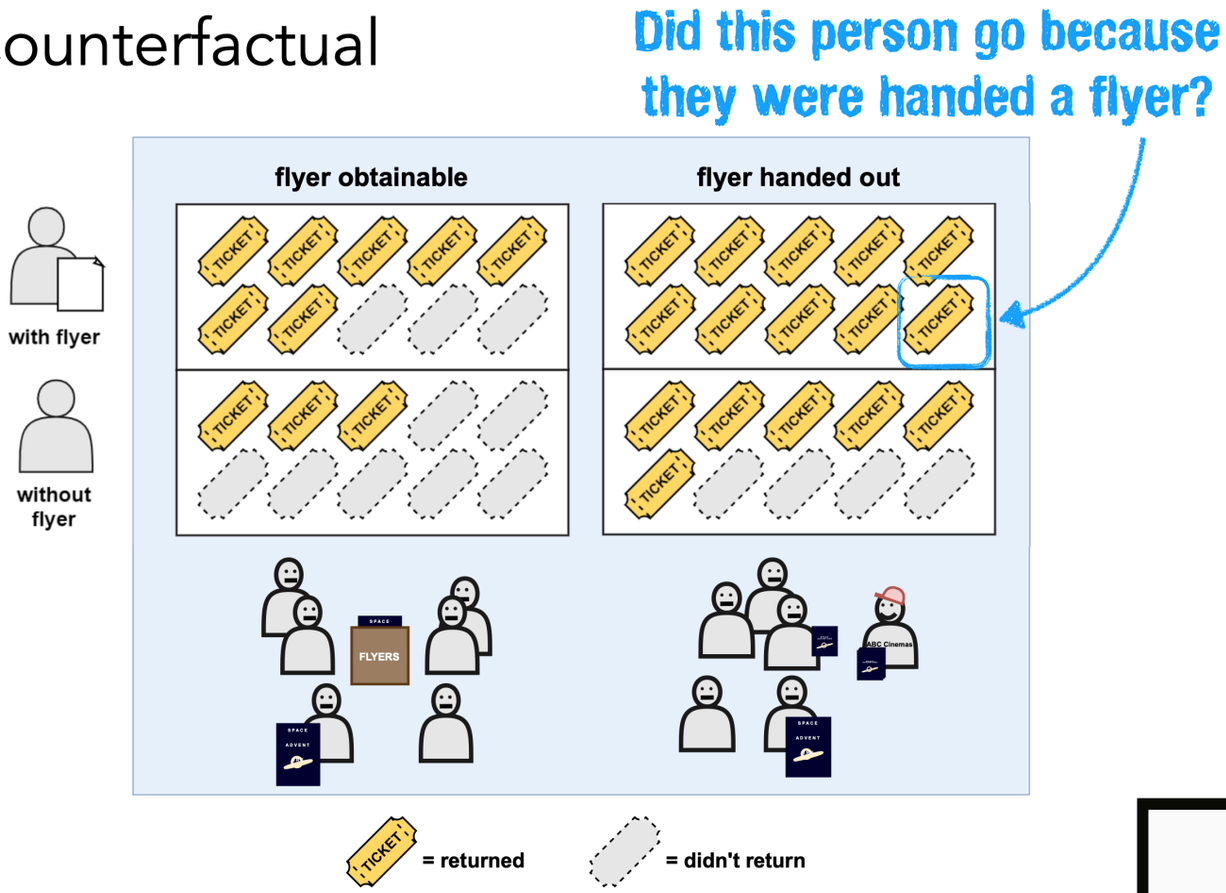


Inferring causal structure through intervention

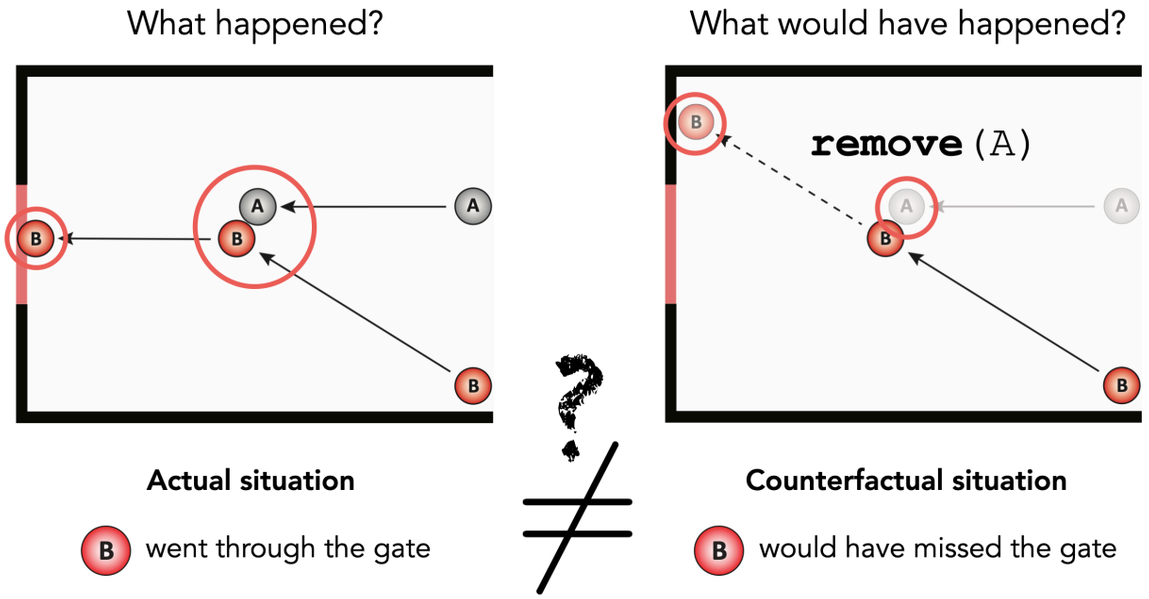


Quick recap: Observation, intervention, counterfactual

Counterfactual



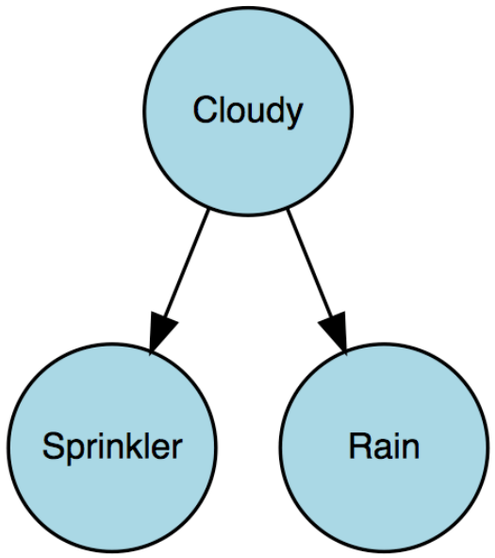
Counterfactual Simulation Model



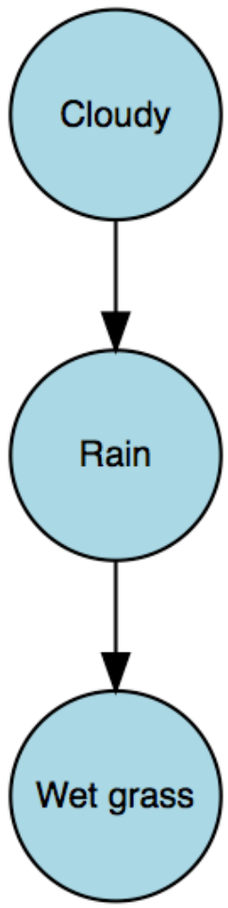
Gerstenberg, Goodman, Lagnado, & Tenenbaum (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*

Quick recap: Patterns of inference

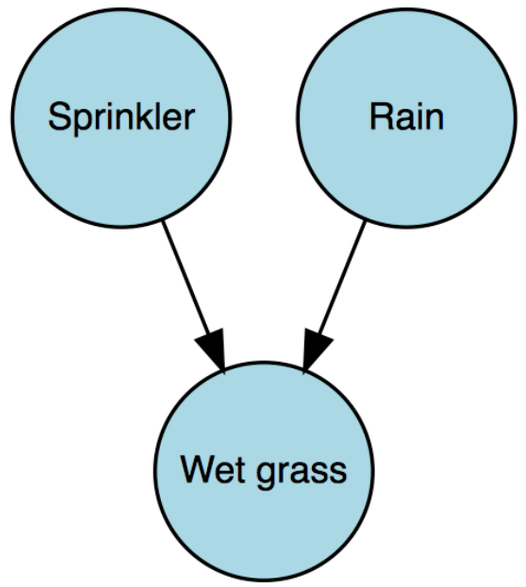
common cause



causal chain



common effect



Should I control?

When should I control for variables?

recent advances in graphical models have produced a way to help distinguish good from bad controls

***d*-separation**

directional

decide from a causal graph whether a set of variables `X` is independent of another set `Y`, given a third set `Z`

Goal: we want a precise (and unbiased) estimate of the predictive relationship between `X` and `Y`

we want to block all other paths from X to Y

http://bayes.cs.ucla.edu/BOOK-09/ch11-1-2-final.pdf

When should I control for variables?

How can I tell whether two variables are independent?

Recipe for independence

1. Draw the ancestral graph
2. "Moralize" the graph by "marrying" the parents
3. "Disorient" the graph by replacing arrows with edges
4. Delete the givens and their edges
5. Read the answer off the graph

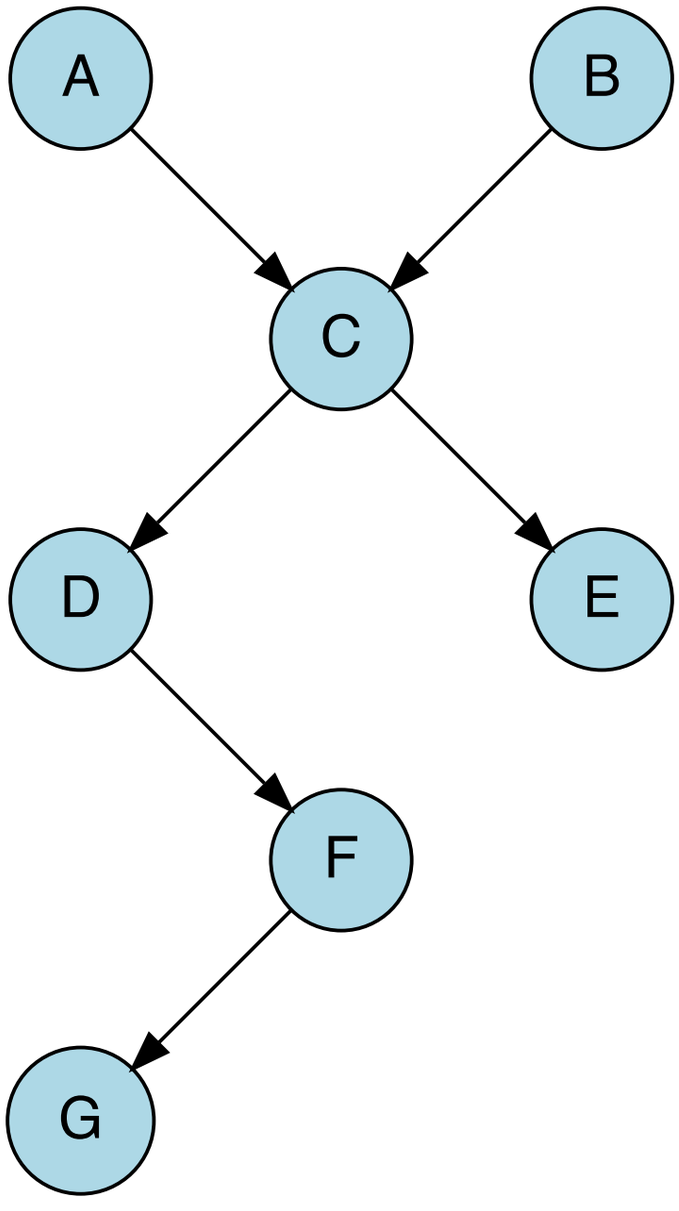
   - if variables are **disconnected** they are independent
   - if variables are connected (have a path between them) they are not guaranteed to be independent

http://web.mit.edu/jmn/www/6.034/d-separation.pdf

When should I control for variables?

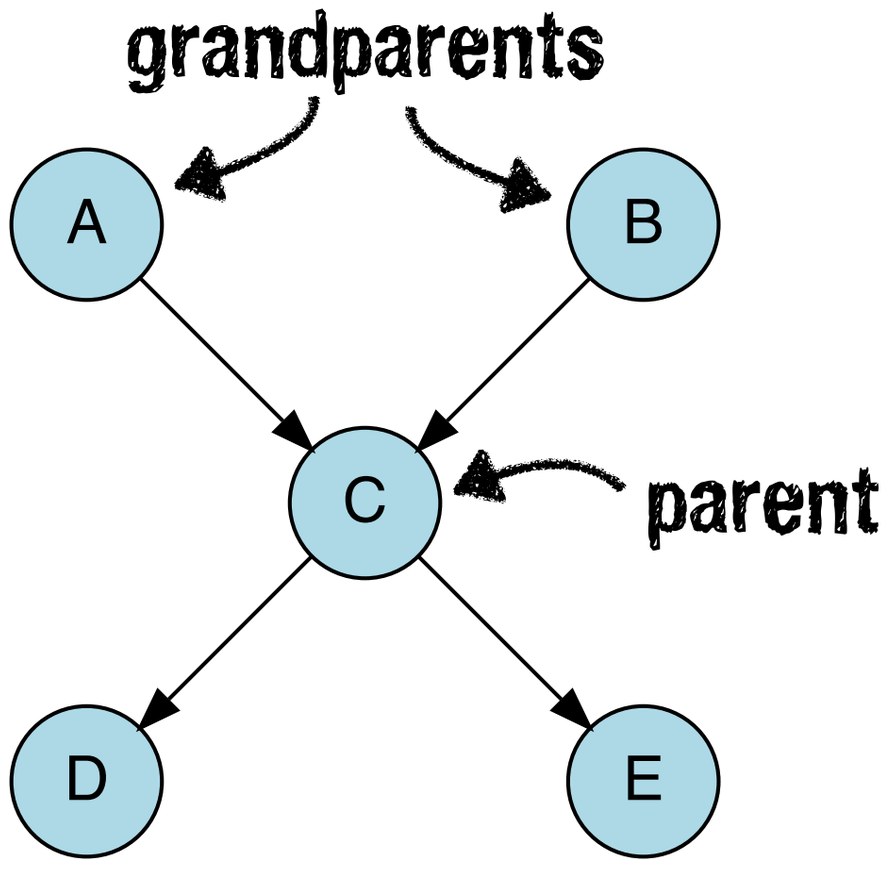
Are D and E independent?

$$p(D \mid E) = p(D) \ ?$$



1. Draw the ancestral graph

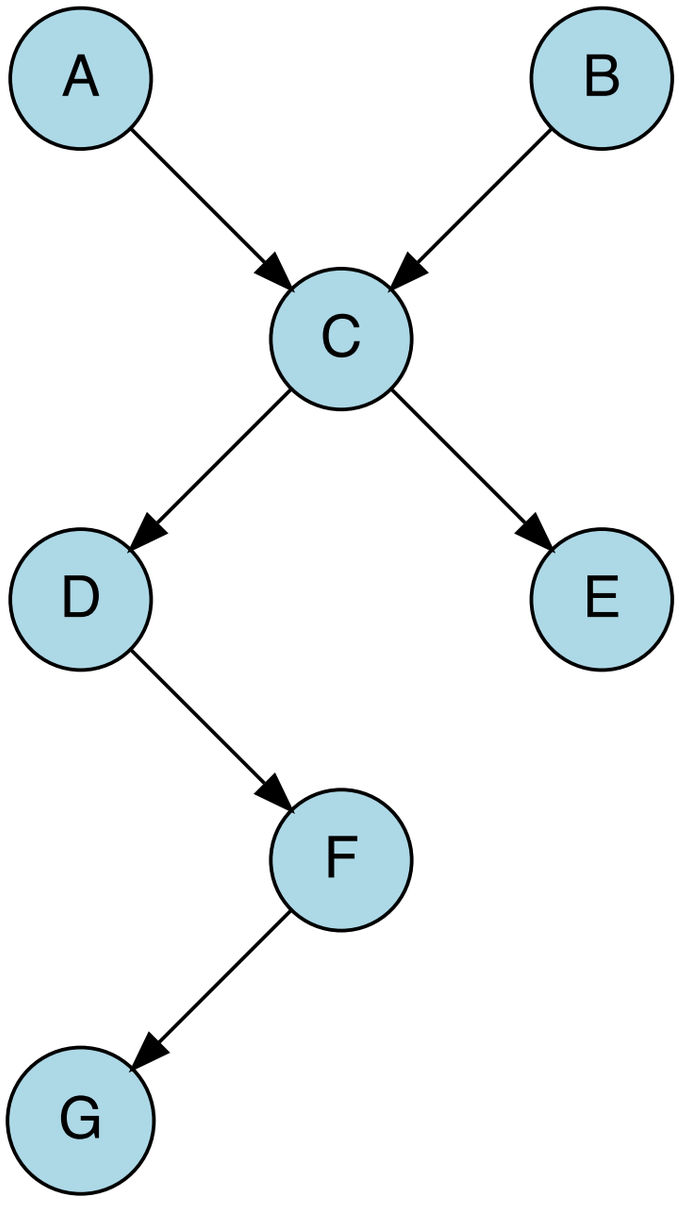
Construct the "ancestral graph" of all variables mentioned in the probability expression. This is a reduced version of the original net, consisting only of the variables mentioned and all of their ancestors (parents, parents' parents, etc.)



When should I control for variables?

Are D and E independent?

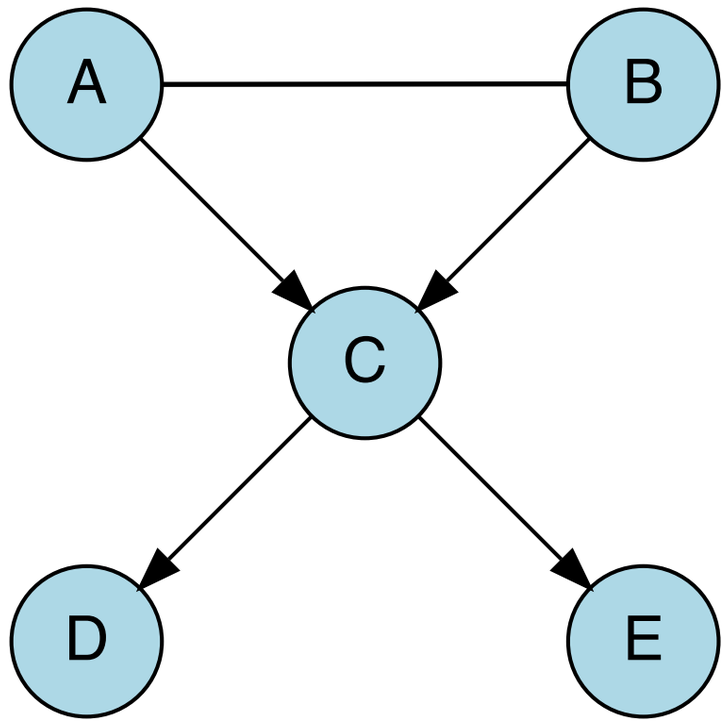
$$p(D \mid E) = p(D) \ ?$$



2. "Moralize" the graph

let's get married!

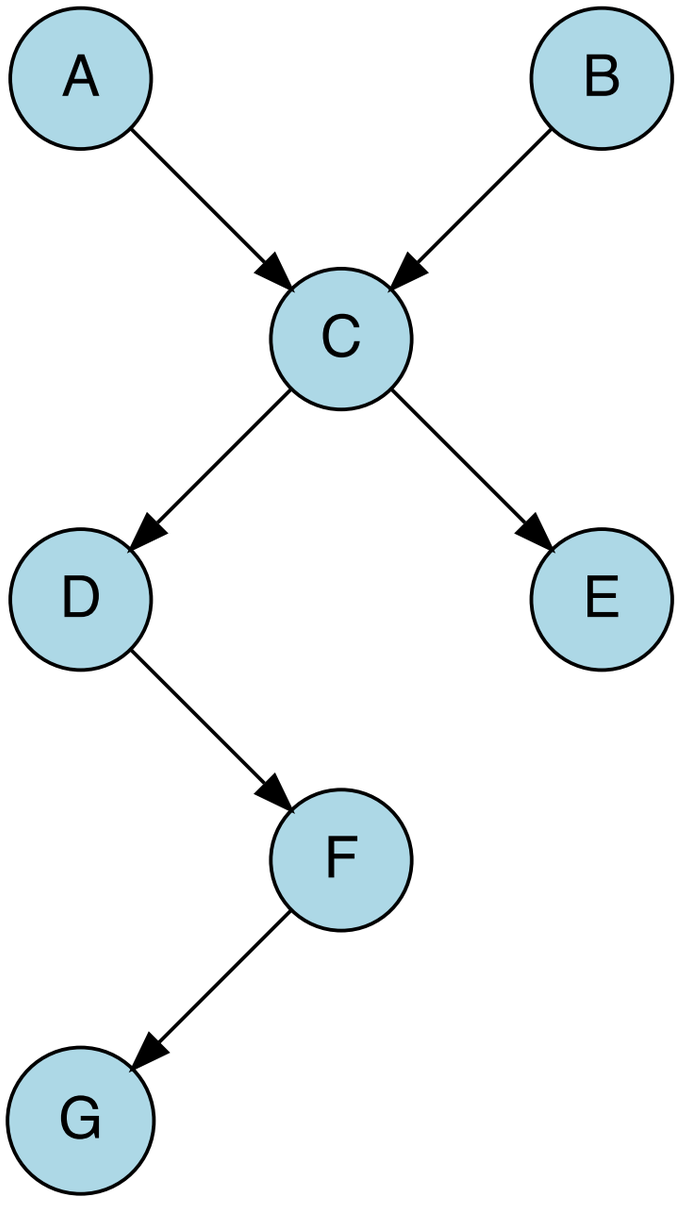
For each pair of variables with a common child, draw an undirected edge (line) between them. (If a variable has more than two parents, draw lines between every pair of parents.)



When should I control for variables?

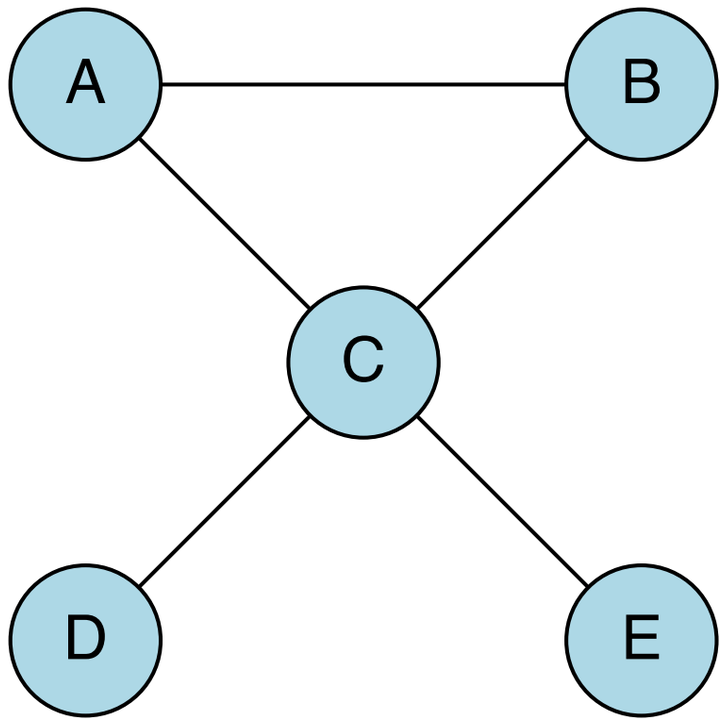
Are D and E independent?

$$p(D \mid E) = p(D) \ ?$$



3. "Disorient" the graph

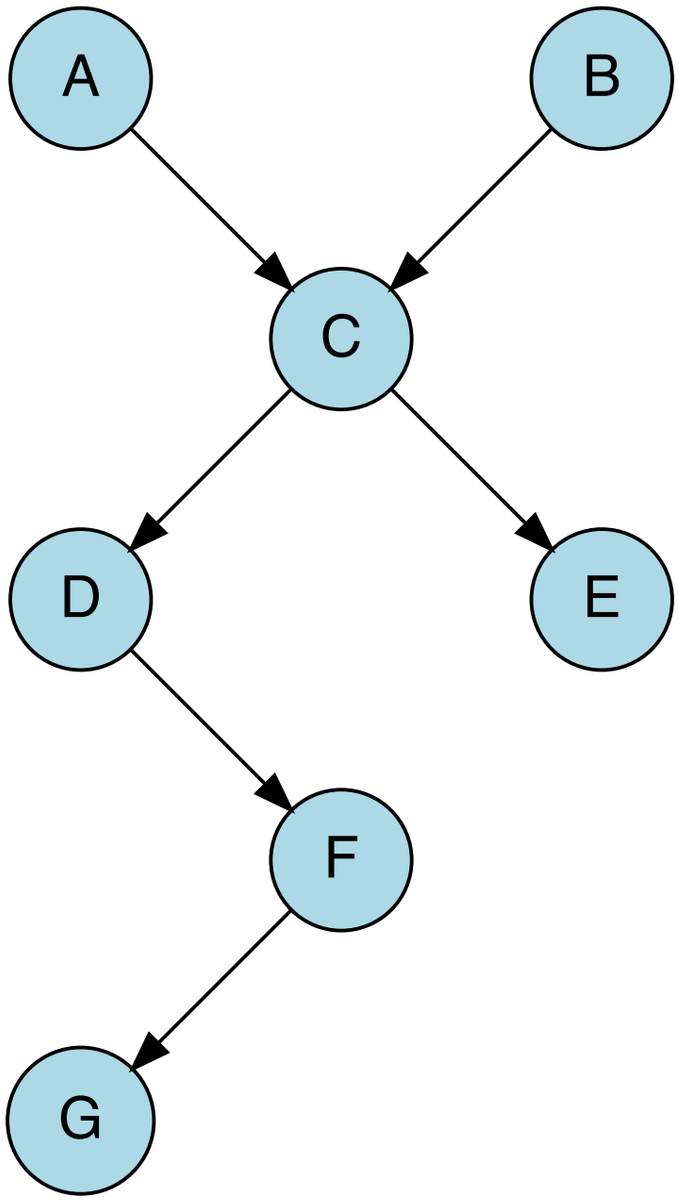
Replace arrows arrows with lines



When should I control for variables?

Are D and E independent?

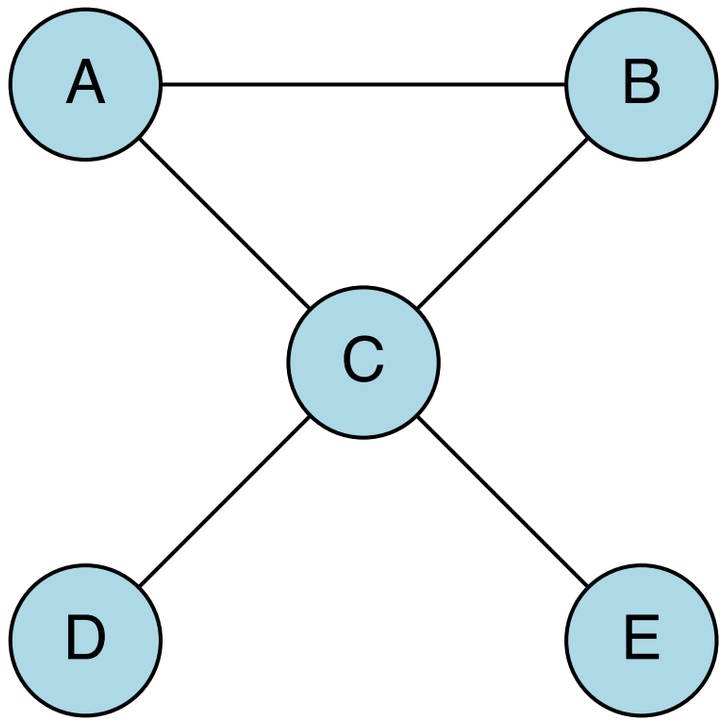
$$p(D \mid E) = p(D) \ ?$$



4. Delete the givens

Remove the variables that we condition on, as well as their edges

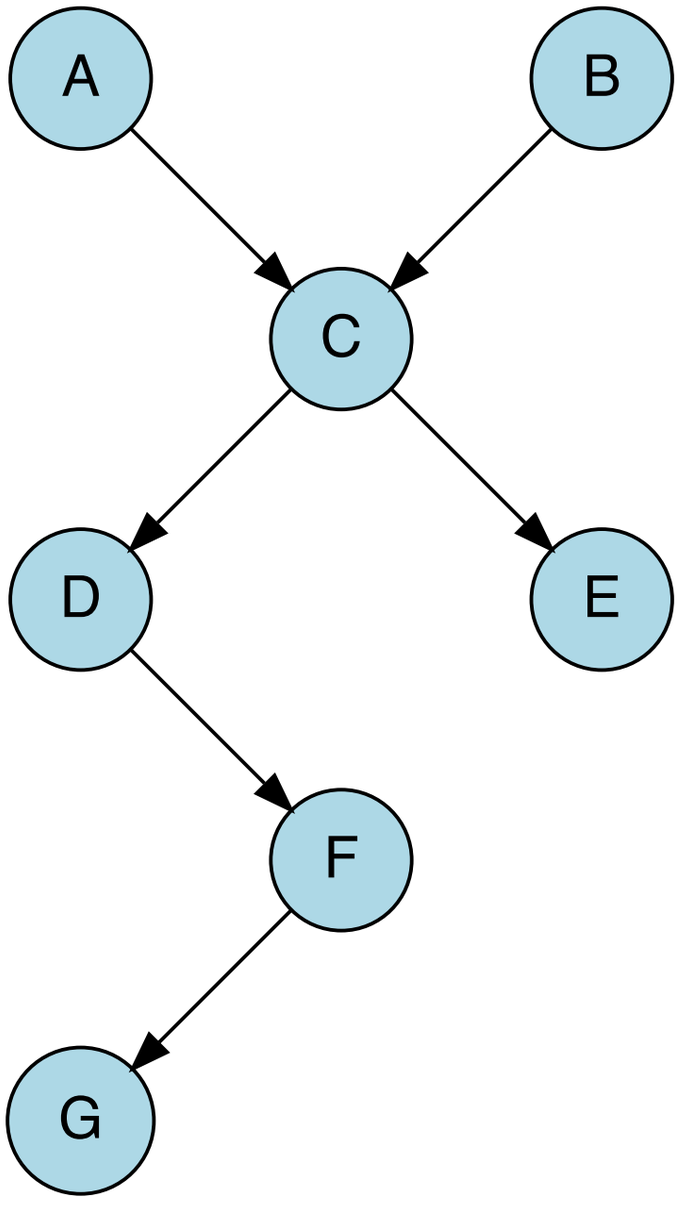
we didn't condition on anything, so there is nothing to delete



When should I control for variables?

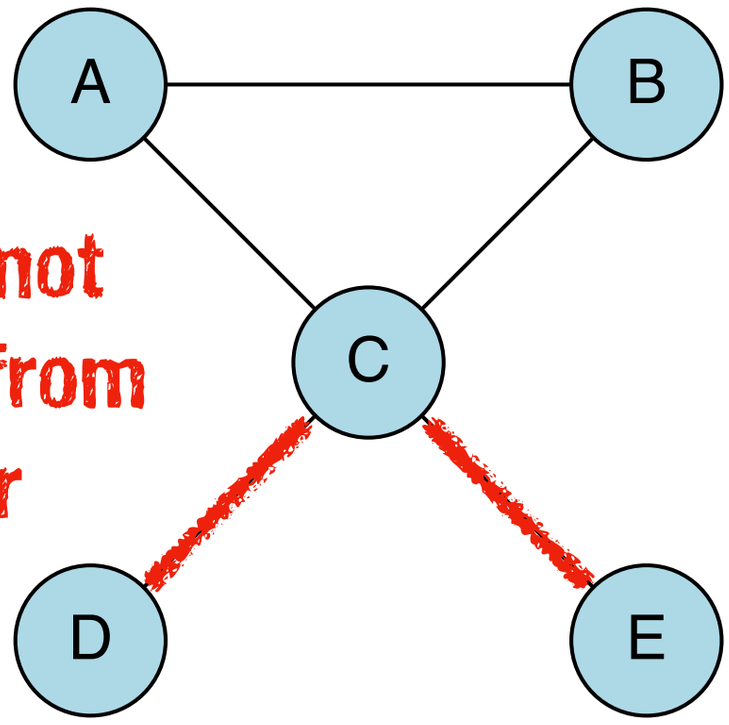
Are D and E independent?

$$p(D \mid E) = p(D) \ ?$$



5. Read answer off the graph

- if variables are **disconnected** they are independent
- if variables are connected (have a path between them) they are not guaranteed to be independent



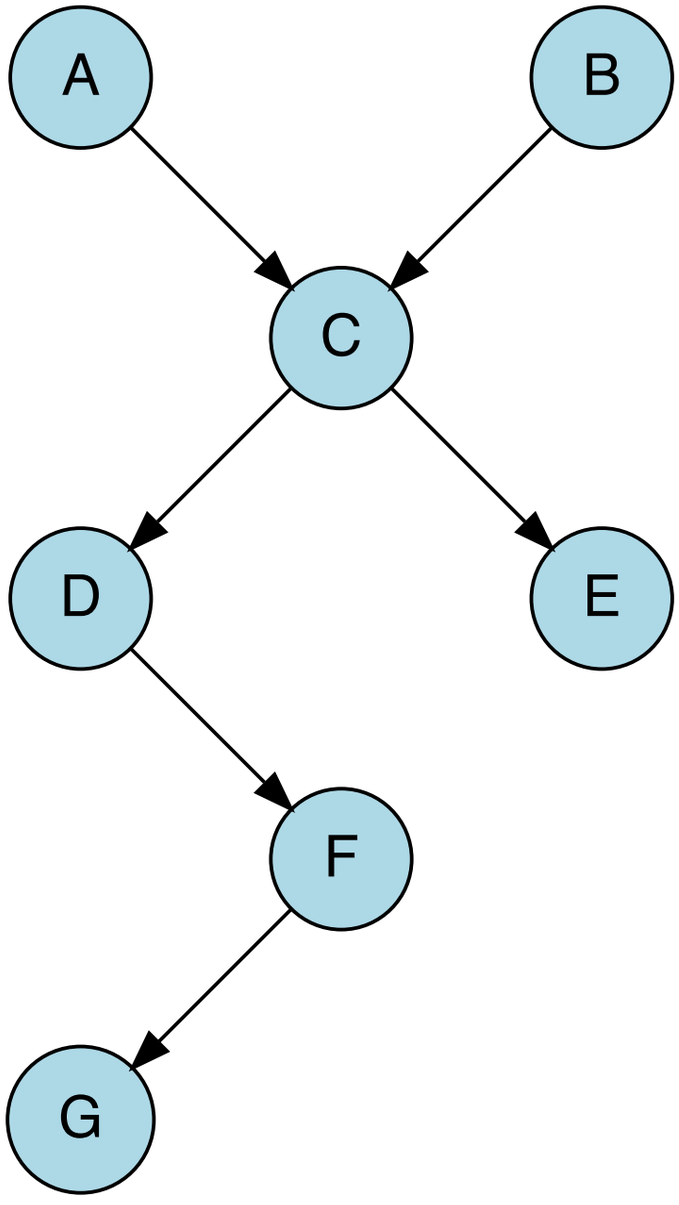
D and E are not independent from each other

they are connected via at least one path

When should I control for variables?

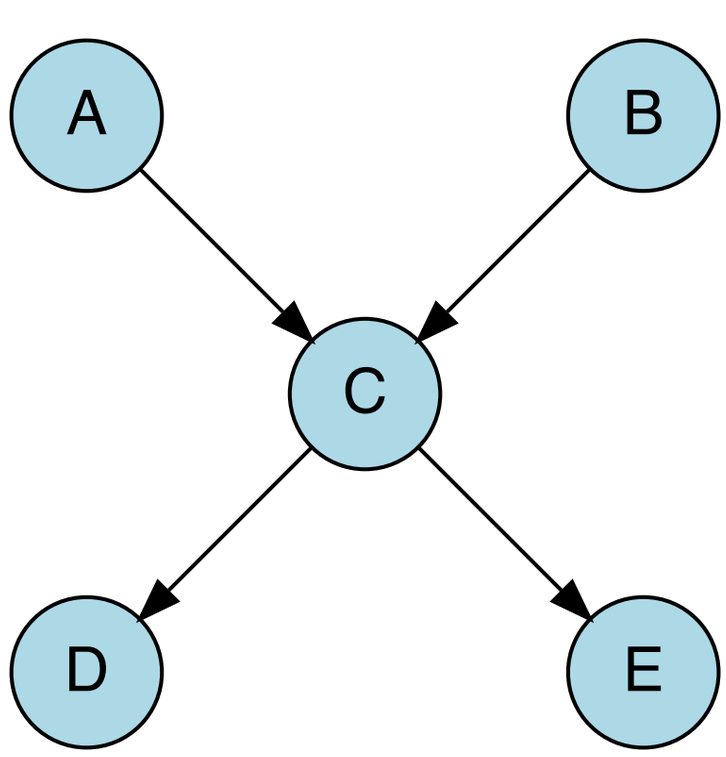
Are D and E independent, given C?

$$p(D \mid E, C) = p(D \mid C) \ ?$$



1. Draw the ancestral graph

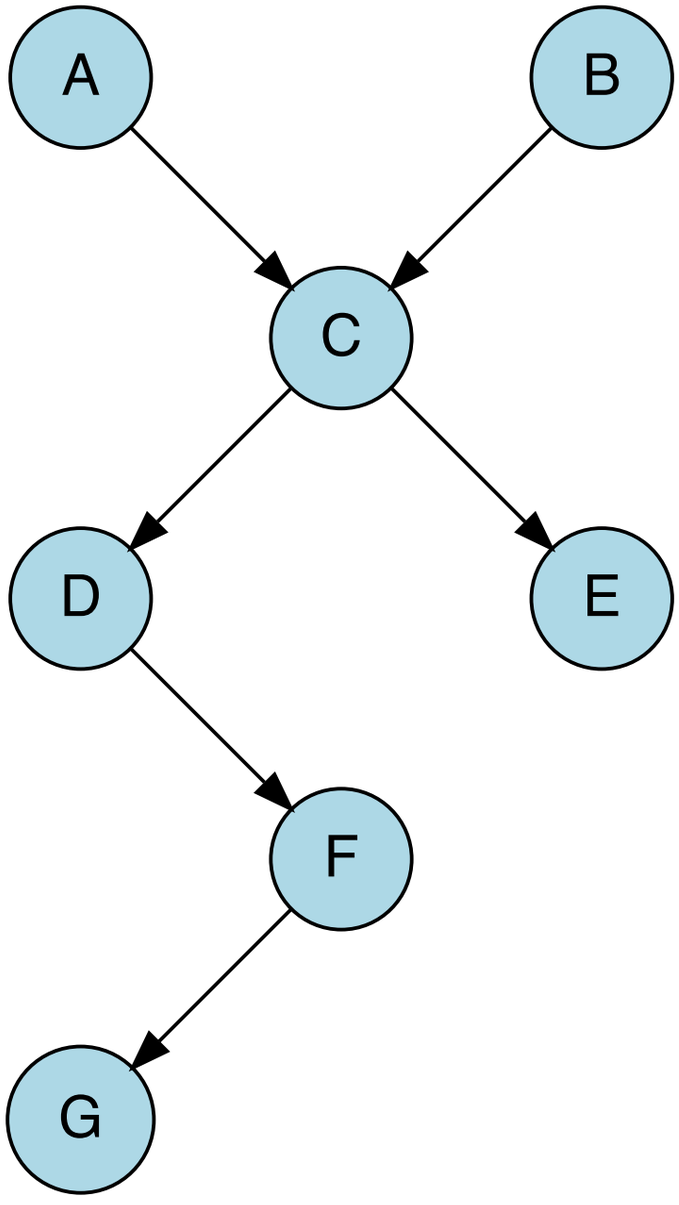
Construct the "ancestral graph" of all variables mentioned in the probability expression. This is a reduced version of the original net, consisting only of the variables mentioned and all of their ancestors (parents, parents' parents, etc.)



When should I control for variables?

Are D and E independent, given C?

$$p(D \mid E, C) = p(D \mid C) \ ?$$



2. "Moralize" the graph

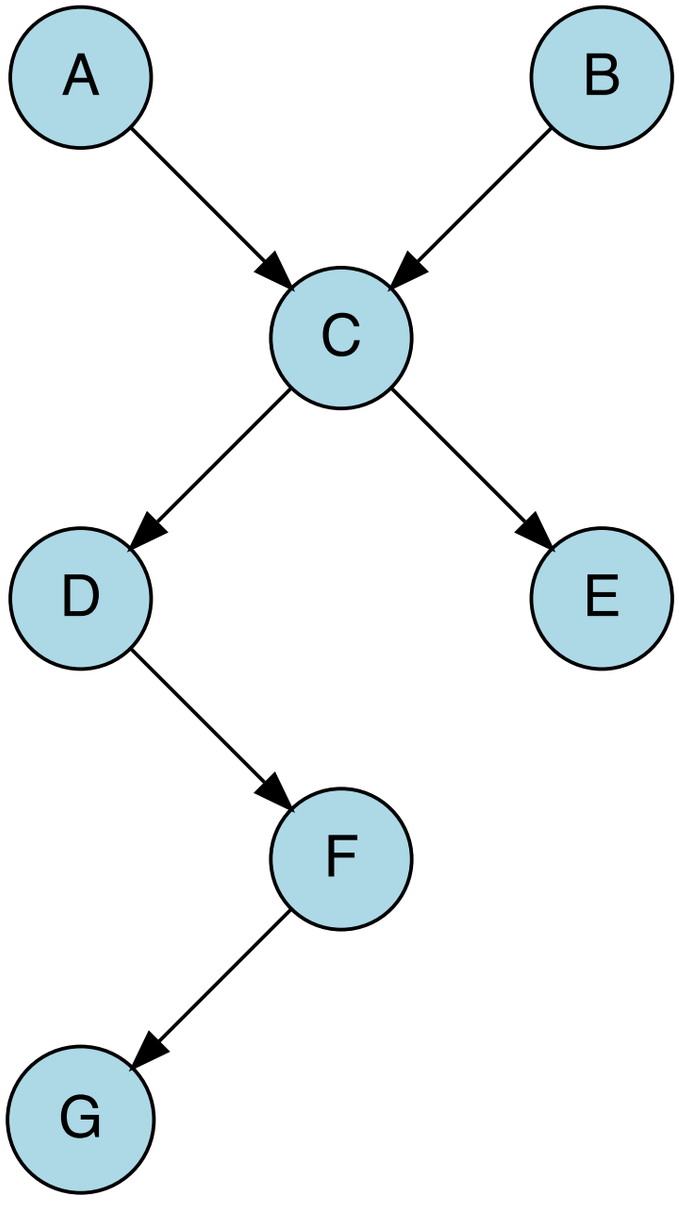
let's get married!

For each pair of variables with a common child, draw an undirected edge (line) between them. (If a variable has more than two parents, draw lines between every pair of parents.)

When should I control for variables?

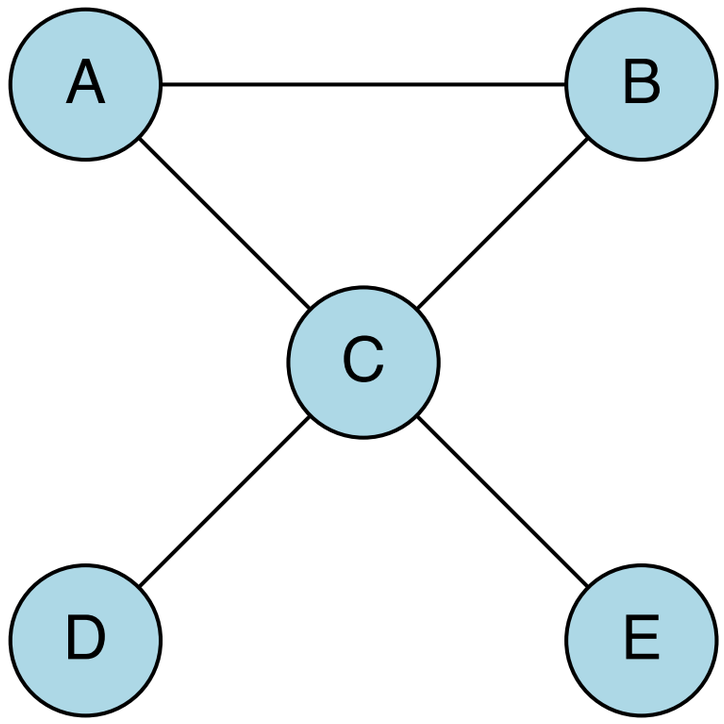
Are D and E independent, given C?

$$p(D \mid E, C) = p(D \mid C) \ ?$$



3. "Disorient" the graph

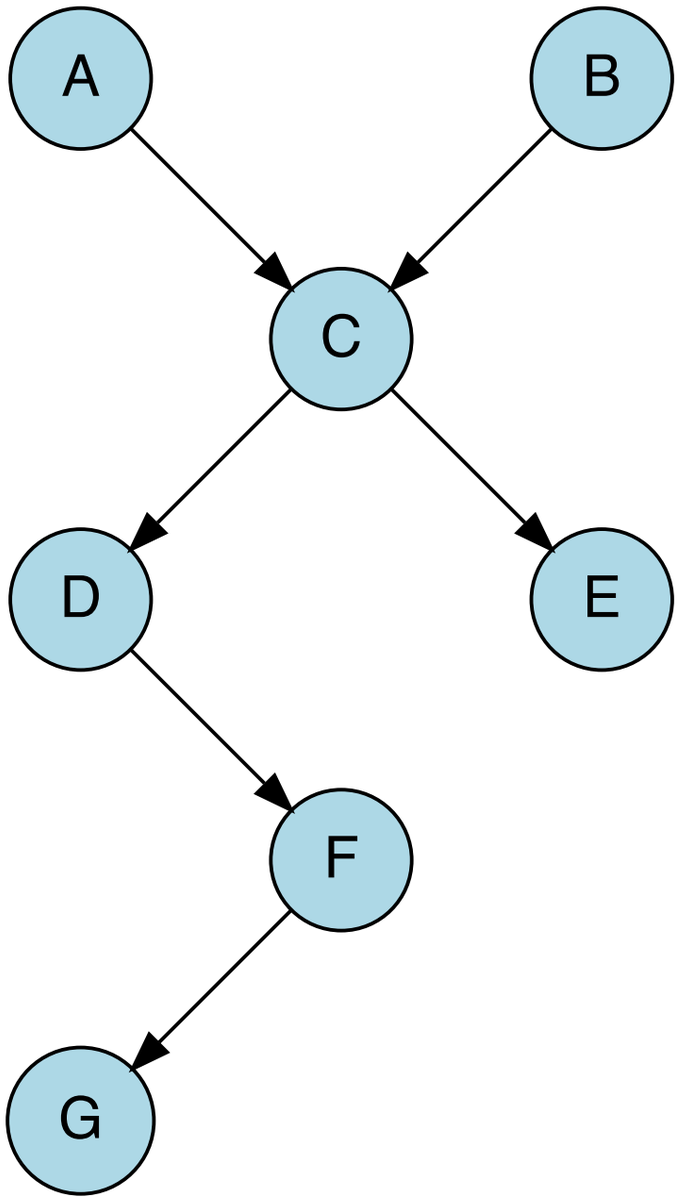
Replace arrows arrows with lines



When should I control for variables?

Are D and E independent, given C?

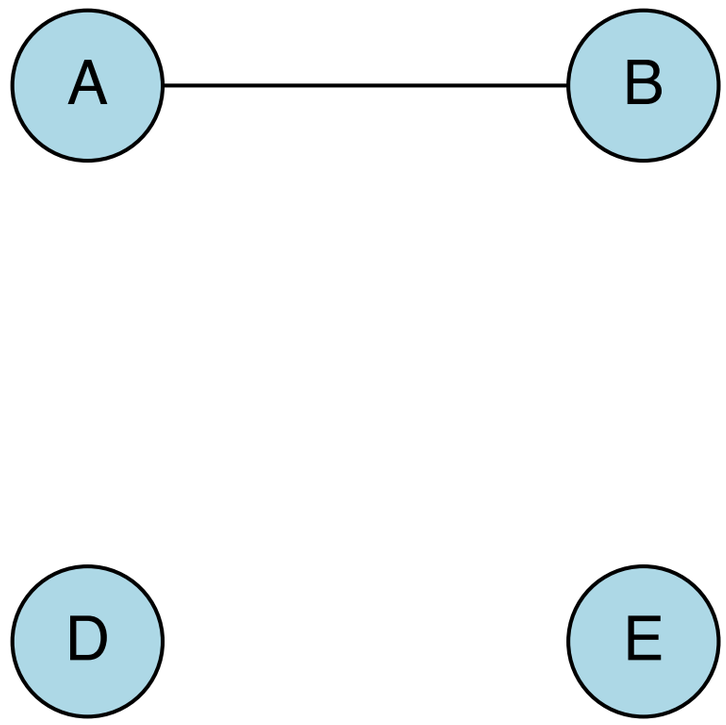
$$p(D \mid E, C) = p(D \mid C) \; ?$$



4. Delete the givens

Remove the variables that we condition on, as well as their edges

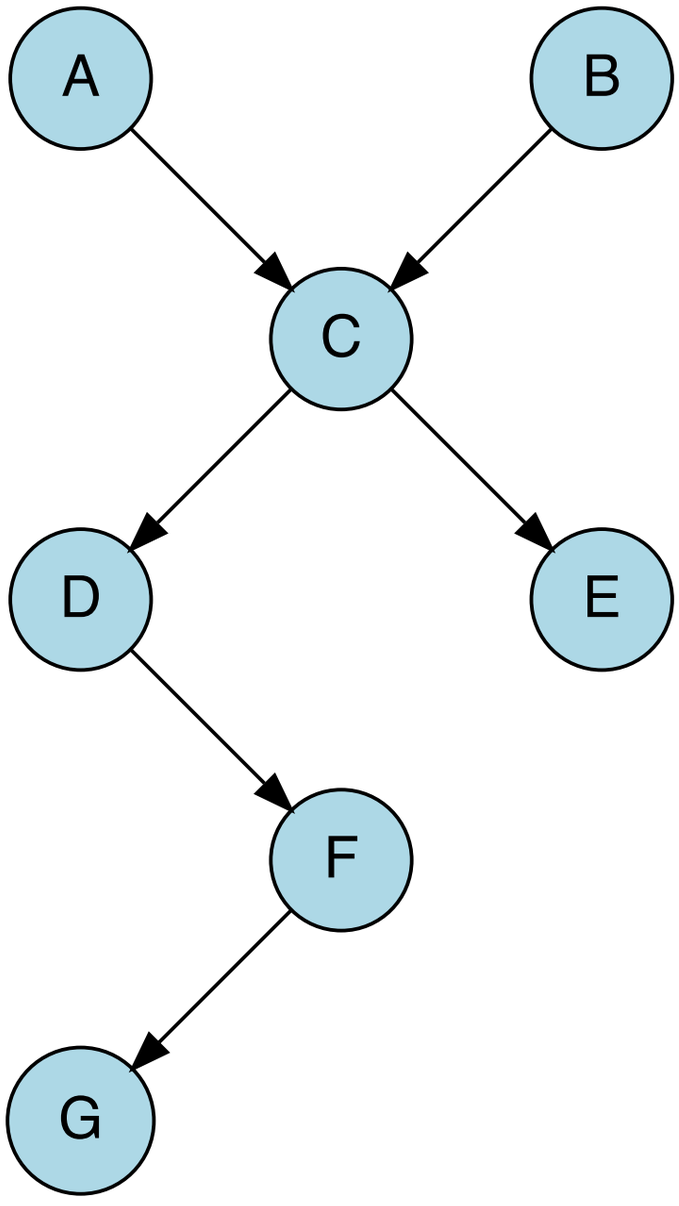
we conditioned on C!



When should I control for variables?

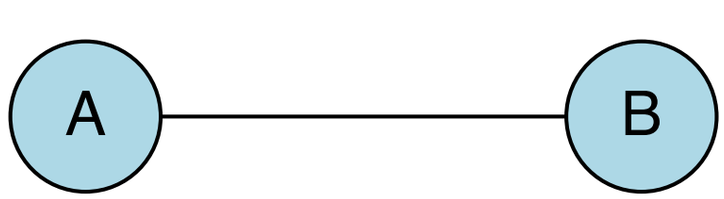
Are D and E independent, given C?

$$p(D \mid E, C) = p(D \mid C) \; ?$$



5. Read answer off the graph

- if variables are **disconnected** they are independent
- if variables are connected (have a path between them) they are not guaranteed to be independent



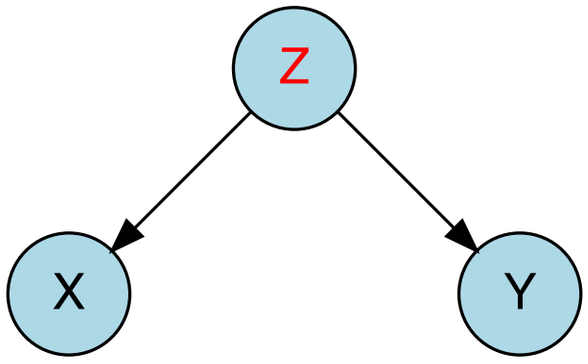
D and E are independent from each other conditioned on C

they aren't connected via a path

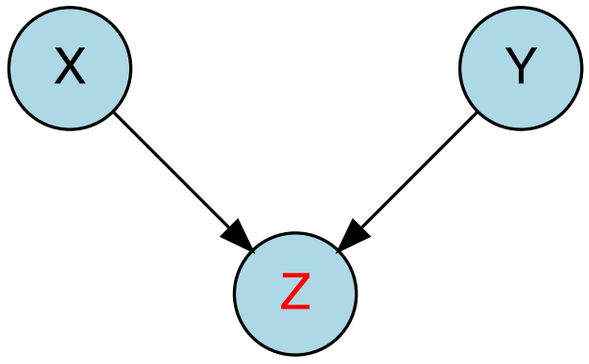
So what?

Patterns of inference

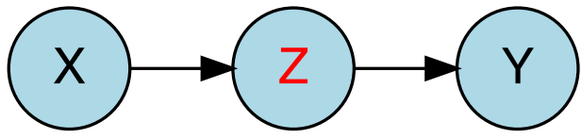
We want to estimate the (causal) relationship between `X` and `Y`



common cause



common effect



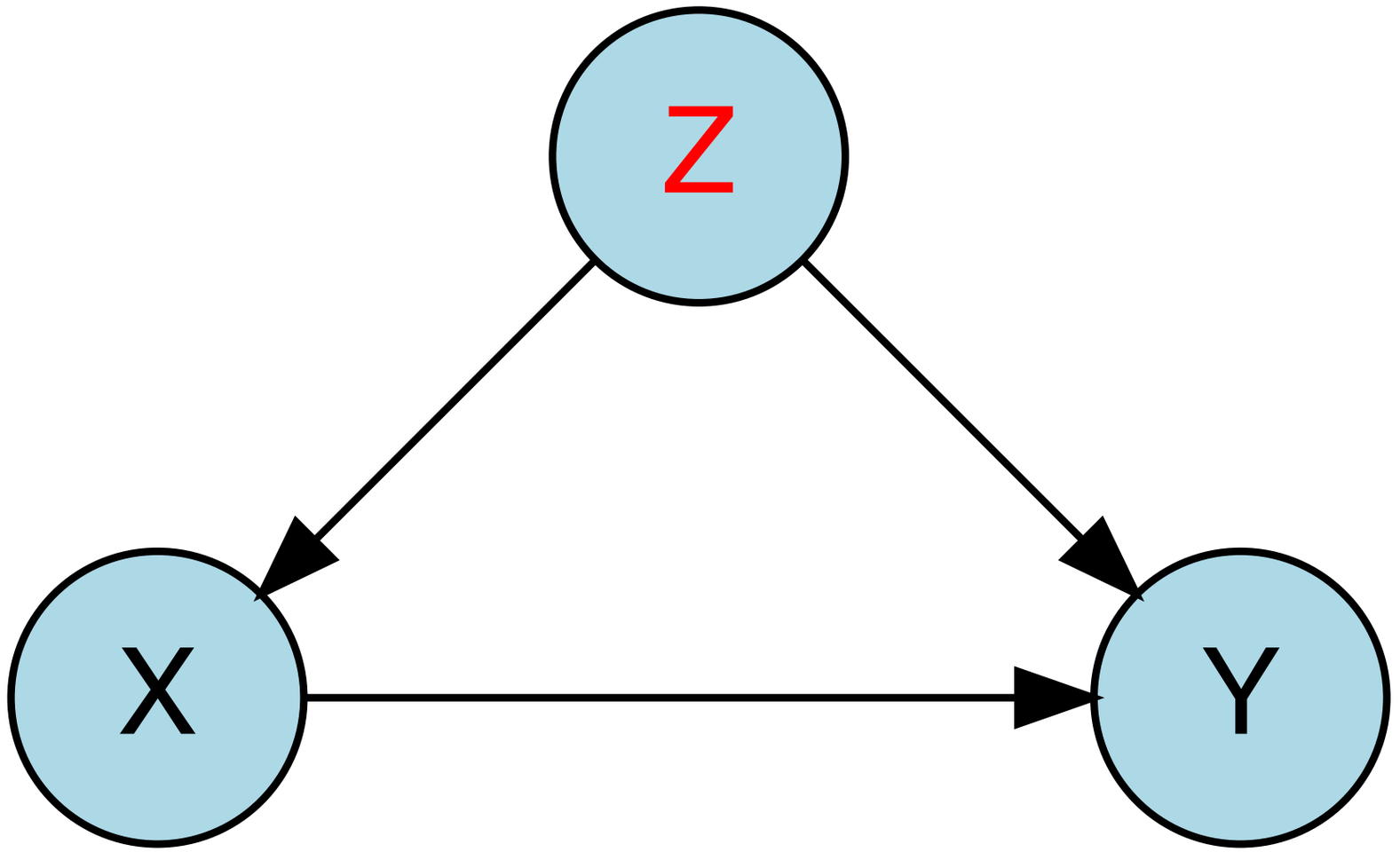
causal chain

by controlling for `Z` we hope to get a better estimate of the relationship between `X` and `Y`

d-separation helps us tell apart **good controls** from **bad controls**

When should I control for variables?

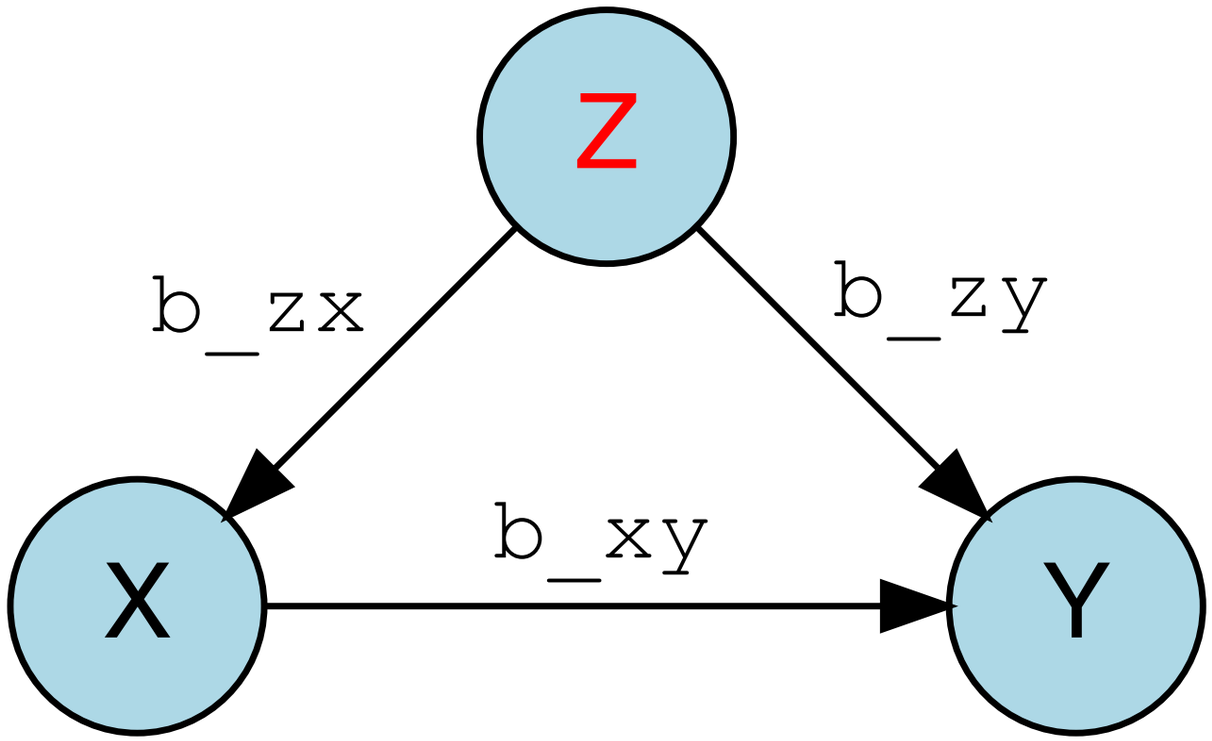
I want to estimate the effect that X has on Y



Is Z a **good** or a **bad** control here?

When should I control for variables?

```r
set.seed(1)

n = 1000
b_zx = 2
b_xy = 2
b_zy = 2
sd = 1

df = tibble(z = rnorm(n = n, sd = sd),
            x = b_zx * z + rnorm(n = n, sd = sd),
            y = b_zy * z + b_xy * x + rnorm(n = n, sd = sd))
```



$$Y = b_0 + b_1 \cdot X + e$$

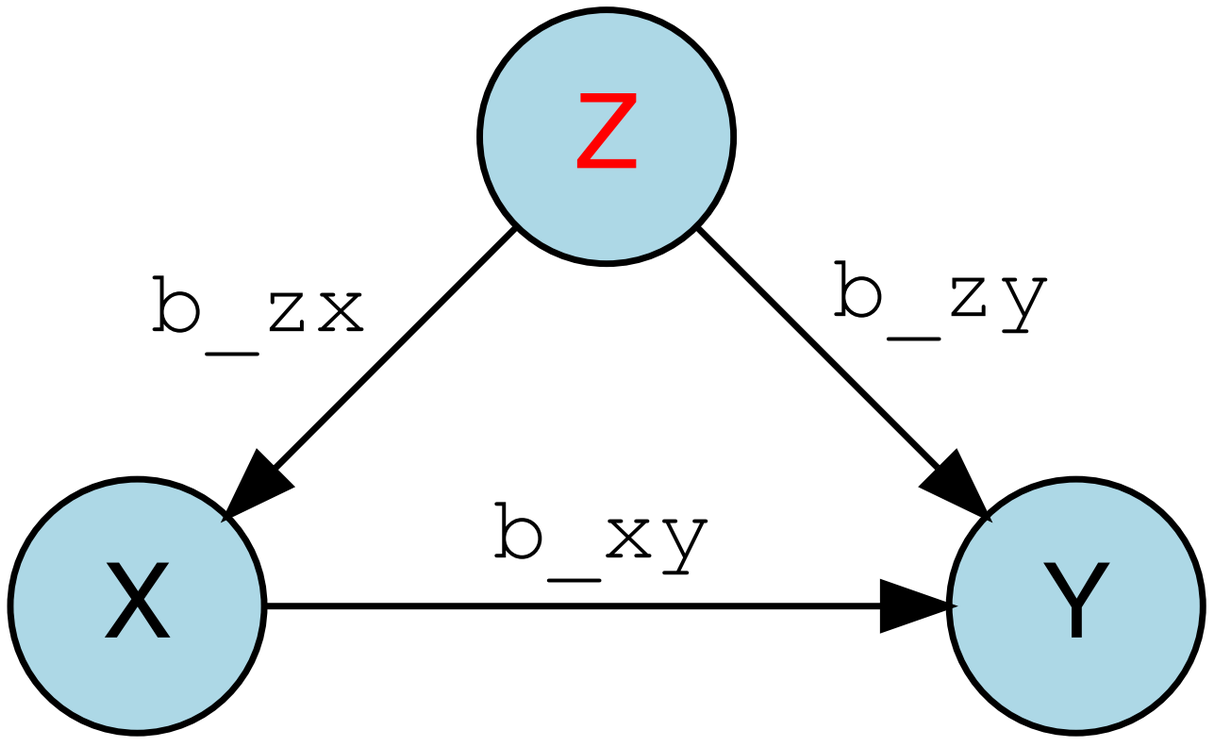
```r
# without control
lm(formula = y ~ x,
   data = df) %>%
  summary()
```

```
Call:
lm(formula = y ~ x, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6011 -0.9270 -0.0506  0.9711  4.0454

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02449    0.04389   0.558    0.577
x            2.82092    0.01890 149.225   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.388 on 998 degrees of freedom
Multiple R-squared:  0.9571,    Adjusted R-squared:  0.9571
F-statistic: 2.227e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

overestimating X's effect on Y

When should I control for variables?

```r
set.seed(1)

n = 1000
b_zx = 2
b_xy = 2
b_zy = 2
sd = 1

df = tibble(z = rnorm(n = n, sd = sd),
            x = b_zx * z + rnorm(n = n, sd = sd),
            y = b_zy * z + b_xy * x + rnorm(n = n, sd = sd))
```



$$Y = b_0 + b_1 \cdot X + b_2 \cdot Z + e$$

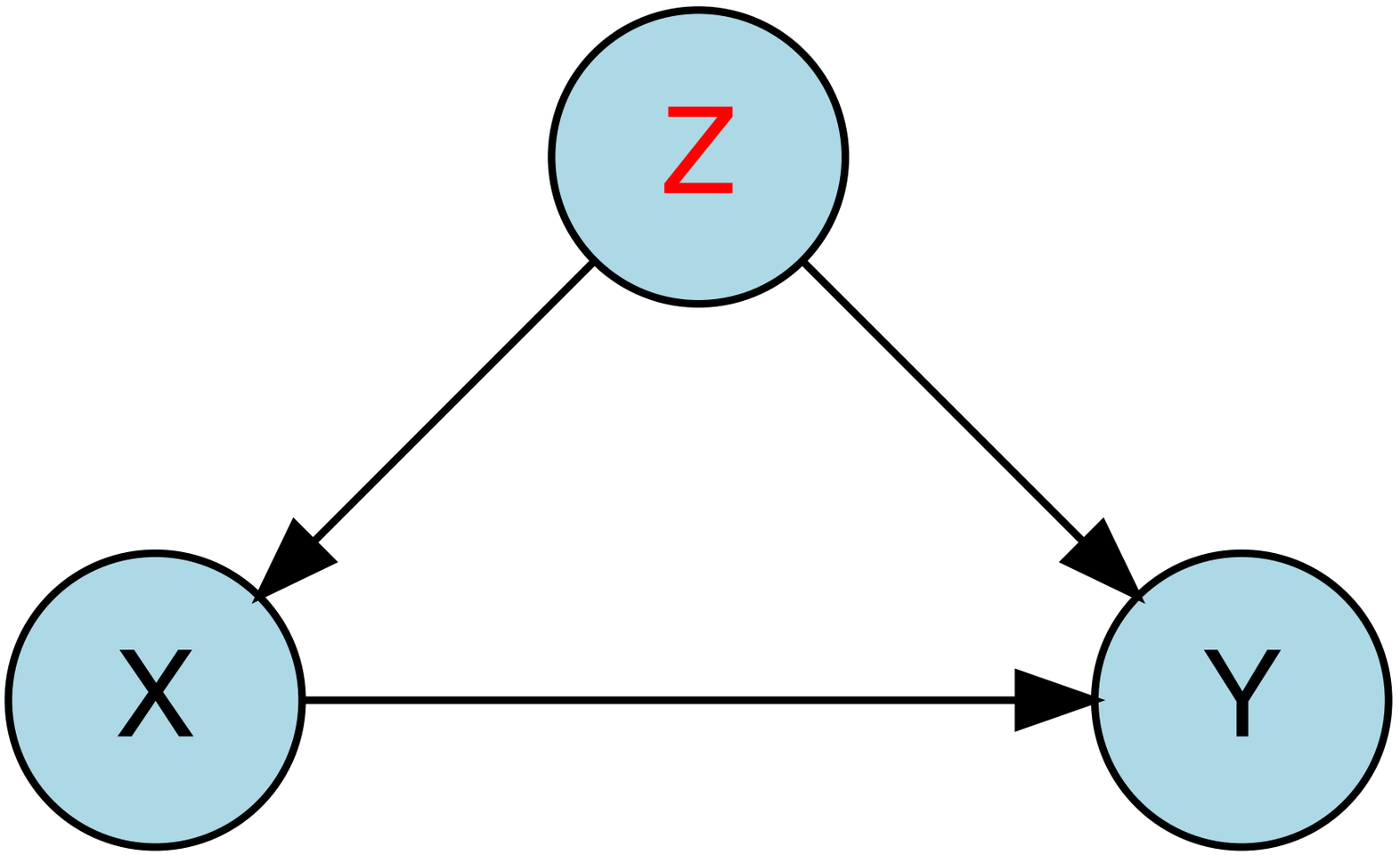
```r
# with control
lm(formula = y ~ x + z,
   data = df) %>%
  summary()
```

```
Call:
lm(formula = y ~ x + z, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6151 -0.6564 -0.0223  0.6815  2.8132

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01624    0.03260   0.498    0.618
x            2.02202    0.03135  64.489   <2e-16 ***
z            2.00501    0.07036  28.497   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.031 on 997 degrees of freedom
Multiple R-squared:  0.9764,    Adjusted R-squared:  0.9763
F-statistic: 2.059e+04 on 2 and 997 DF,  p-value: < 2.2e-16
```

accurate estimate of X's effect on Y

When should I control for variables?

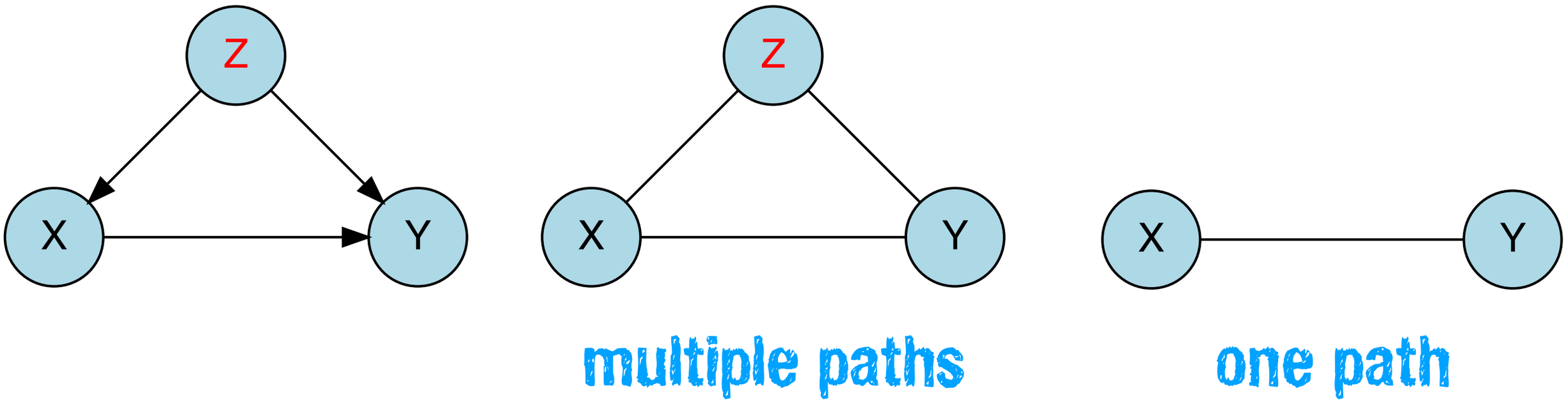
I want to estimate the effect that X has on Y



Z is a **good** control here!

When should I control for variables?

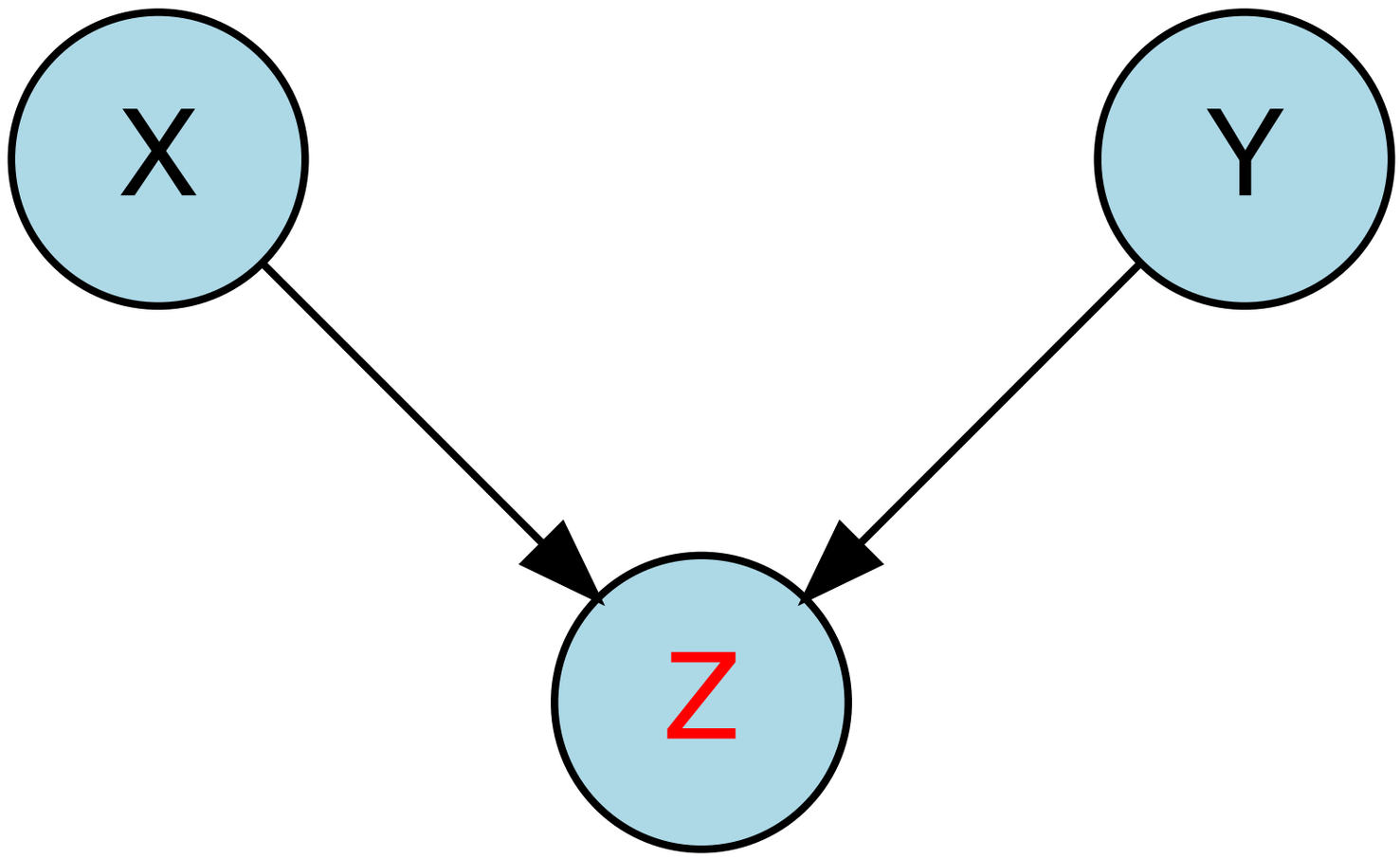
I want to estimate the effect that X has on Y



Z is a **good** control here!

When should I control for variables?

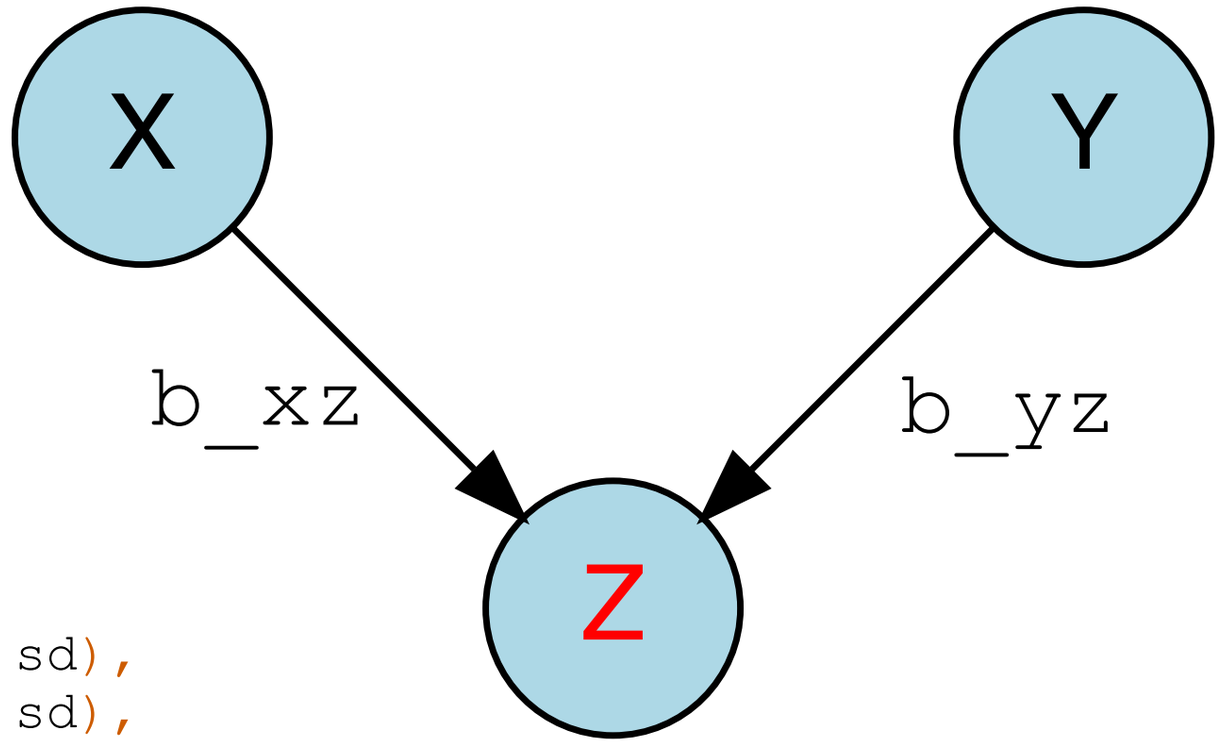
I want to estimate the effect that X has on Y



Is Z a **good** or a **bad** control here?

When should I control for variables?

```r
set.seed(1)

n = 1000
b_xz = 2
b_yz = 2
sd = 1

df = tibble(x = rnorm(n = n, sd = sd),
            y = rnorm(n = n, sd = sd),
            z = x * b_xz + y * b_yz + rnorm(n = n, sd = sd))
```



$$Y = b_0 + b_1 \cdot X + e$$

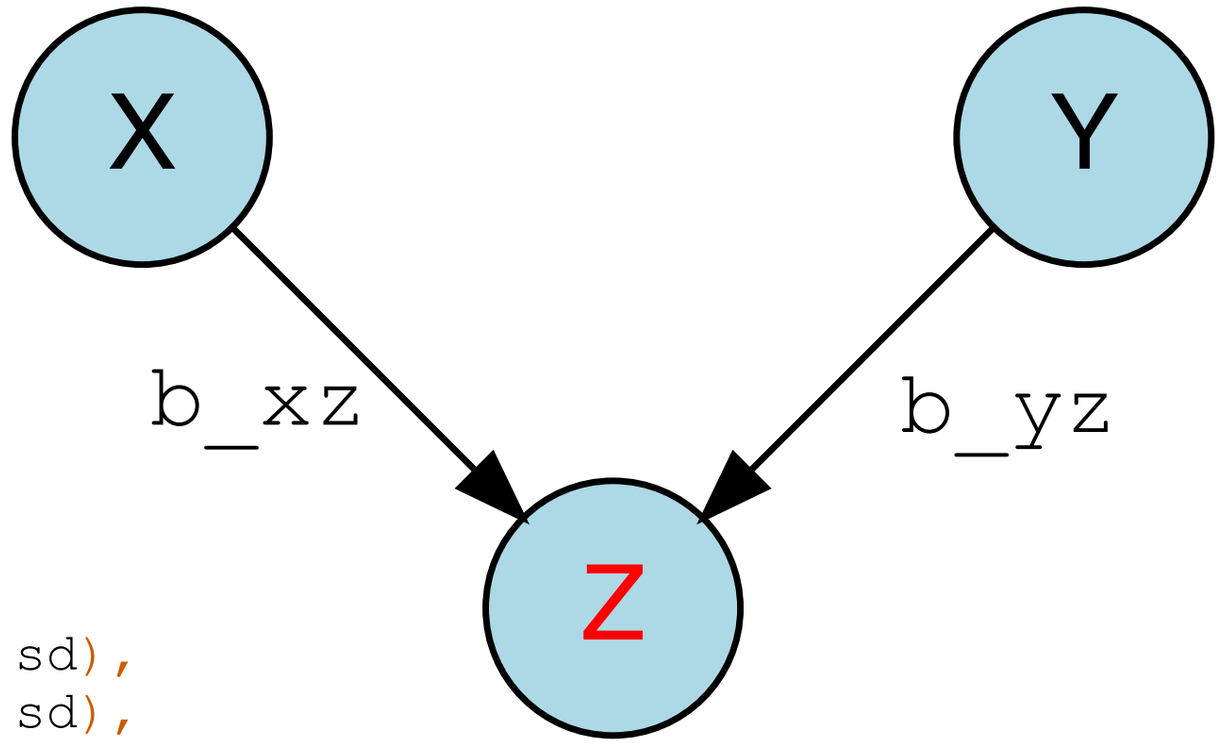
```r
# without control
lm(formula = y ~ x,
   data = df) %>%
  summary()
```

accurate estimate of X's effect on Y

```
Call:
lm(formula = y ~ x, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2484 -0.6720 -0.0138  0.7554  3.6443

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.016187   0.032905  -0.492    0.623
x            0.006433   0.031809   0.202    0.840

Residual standard error: 1.04 on 998 degrees of freedom
Multiple R-squared:  4.098e-05, Adjusted R-squared:  -0.000961
F-statistic: 0.0409 on 1 and 998 DF,  p-value: 0.8398
```

When should I control for variables?

```r
set.seed(1)

n = 1000
b_xz = 2
b_yz = 2
sd = 1

df = tibble(x = rnorm(n = n, sd = sd),
            y = rnorm(n = n, sd = sd),
            z = x * b_xz + y * b_yz + rnorm(n = n, sd = sd))
```



$$Y = b_0 + b_1 \cdot X + b_2 \cdot Z + e$$

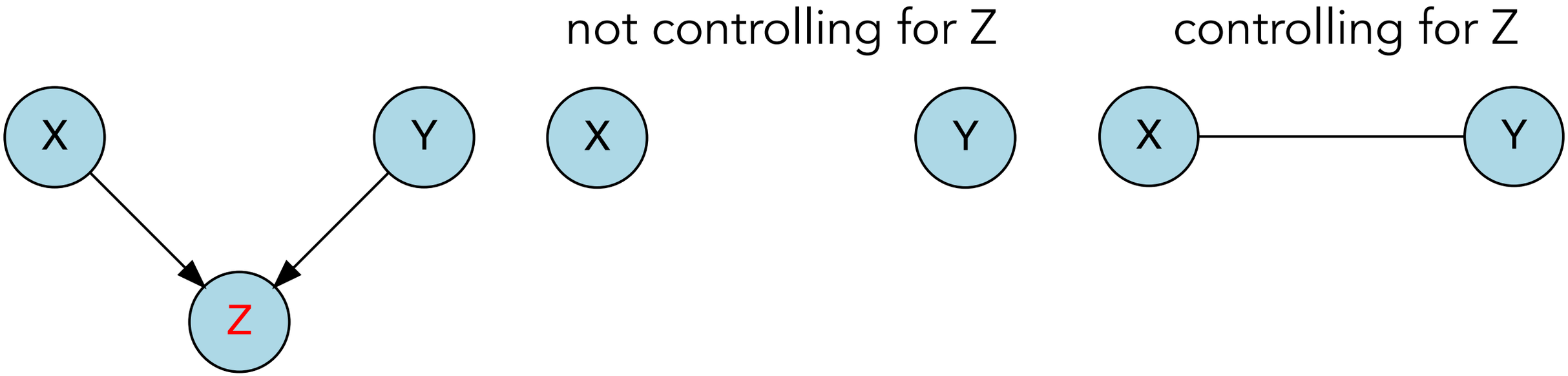
```r
# with control
lm(formula = y ~ x + z,
   data = df) %>%
  summary()
```

```
Call:
lm(formula = y ~ x + z, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.35547 -0.30016  0.00298  0.31119  1.73408

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.009608   0.014477  -0.664    0.507
x           -0.816164   0.018936 -43.102   <2e-16 ***
z            0.398921   0.006186  64.489   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4578 on 997 degrees of freedom
Multiple R-squared:  0.8066,    Adjusted R-squared:  0.8062
F-statistic:  2079 on 2 and 997 DF,  p-value: < 2.2e-16
```

inaccurate estimate of X's effect on Y

When should I control for variables?

I want to estimate the effect that X has on Y



no path **one path**

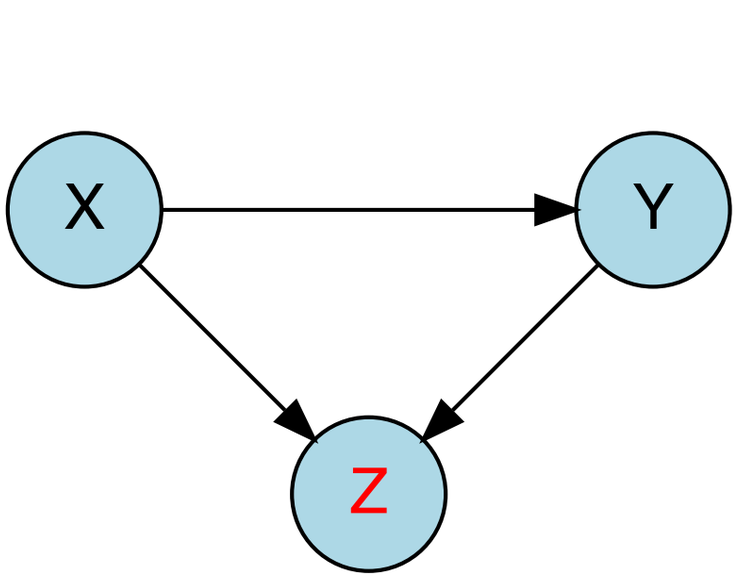
Z is a **bad** control here!

When should I control for variables?

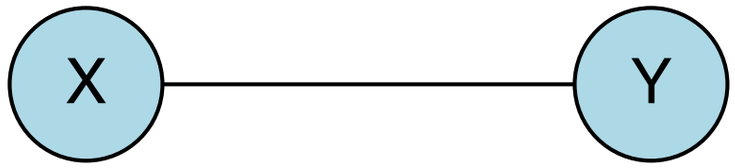
- checking for **d-separation** tells us whether or not variables are (conditionally) independent
- it also tells us whether paths of dependence "open up", or get "closed down"
- the graphical procedure doesn't necessarily reveal whether the dependence between variables changes: it reveals the **structure** of dependence but not the **strength**
- you can always double check via running simulations in R

When should I control for variables?

I want to estimate the effect that X has on Y

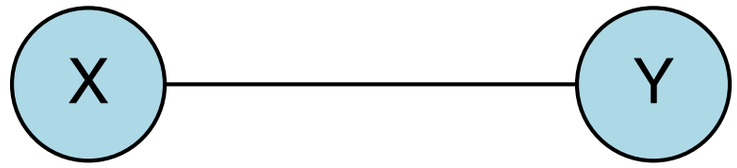


not controlling for Z



one path

controlling for Z



change in dependence

not revealed in the undirected graph

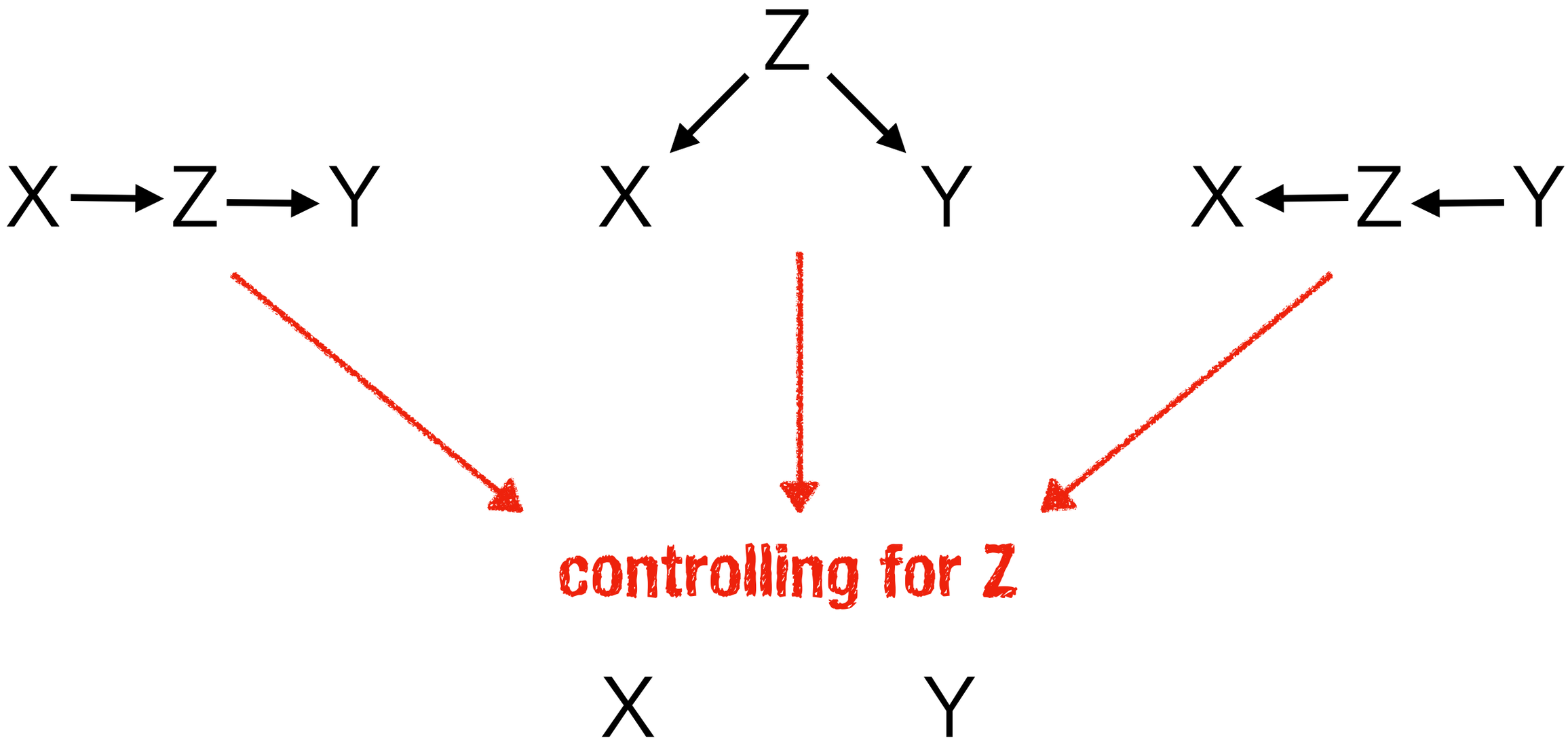
Z is a **bad** control here!

When should I control for variables?

- **good controls** reduce additional paths from `X` to `Y` apart from the direct path we are interested in estimating
- **bad controls** introduce additional paths (or change existing ones) that lead to a biased estimate of the direct path between `X` and `Y`

When should I control for variables?

Problem: We don't know the ground truth ...



we need to manipulate X experimentally to tell these apart*

* sort of (see next slide)

When should I control for variables?

- **causal discovery** is a very active field

what causal claims can we make from observational data?

Identifiability of Gaussian structural equation models with equal error variances

Jonas Peters*
Seminar for Statistics
ETH Zurich
Switzerland

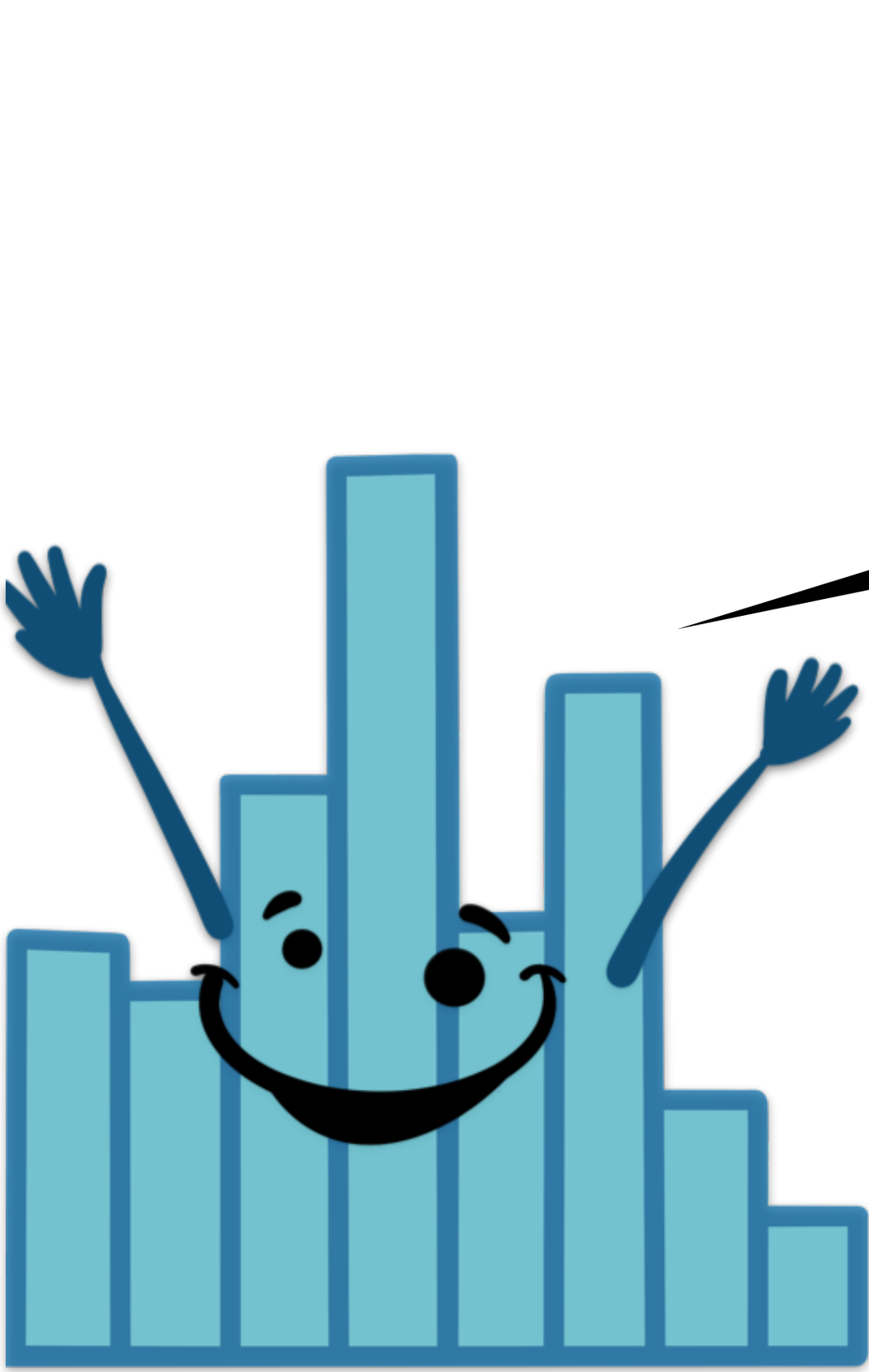
Peter Bühlmann*
Seminar for Statistics
ETH Zurich
Switzerland

October 29, 2018

causal model is fully identifiable if all noise variables have the same variances, and all variables are observed

beyond the scope of our class ...

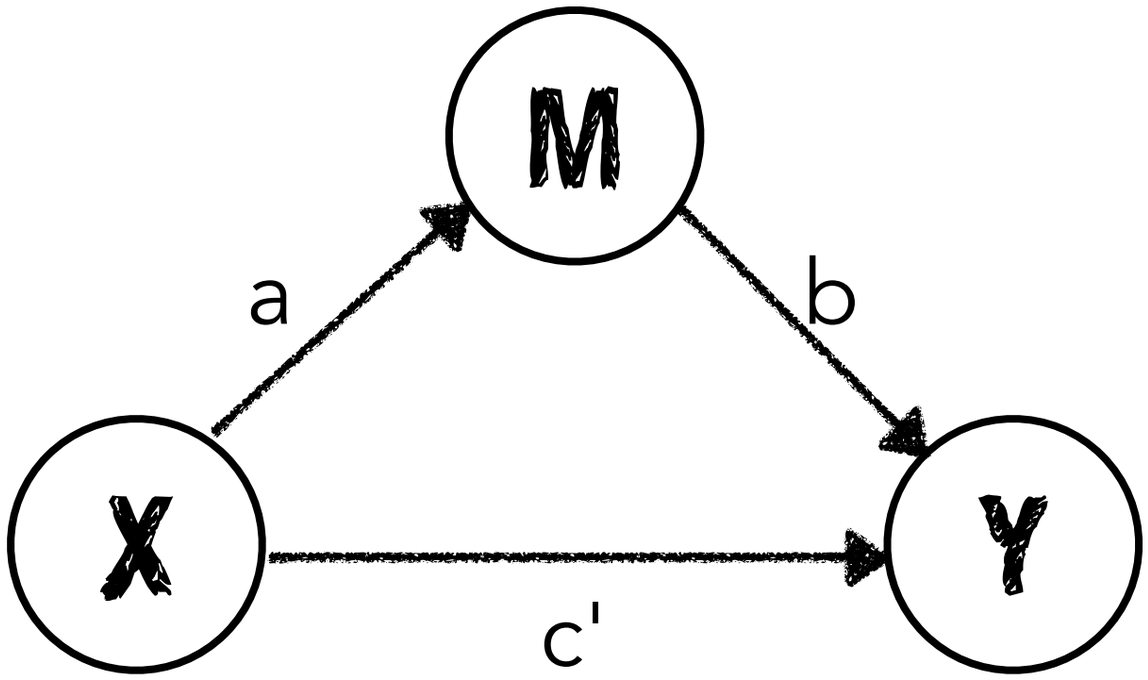
Peters & Buhlmann (2018) Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*


02:00
stretch break!


kino

Mediation

Definition

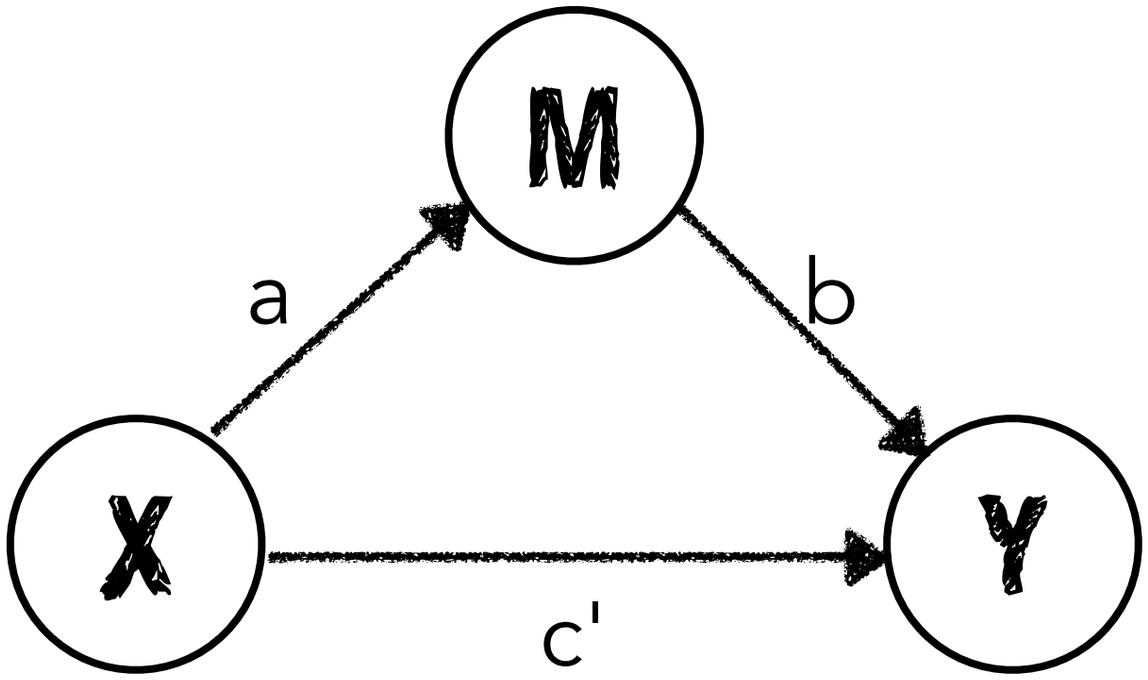


Rather than a direct causal relationship between **X** and **Y**, a mediation model proposes that **X** influences the mediator variable **M**, which in turn influences **Y**. Thus, the mediator variable serves to clarify the nature of the relationship between **X** and **Y**.

Adapted from Wikipedia

https://en.wikipedia.org/wiki/Mediation_(statistics)

Example



X = grades in Psych 252

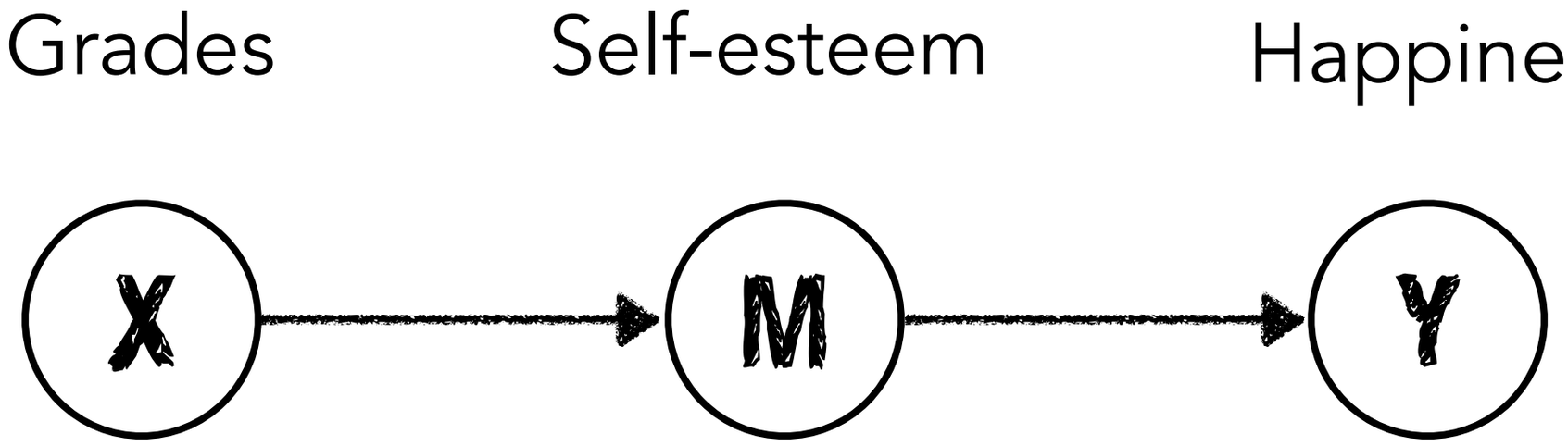
M = feelings of self-esteem

Y = happiness

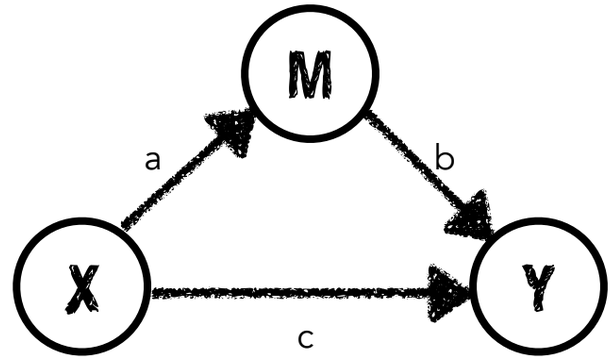
Is the relationship between grades in Psych 252 and happiness mediated by feelings of self-esteem?

Simulate a mediation analysis

```
# number of participants
n = 100

# generate data
df.mediation = tibble(
  x = rnorm(n, 75, 7),              # grades
  m = 0.7 * x + rnorm(n, 0, 5), # self-esteem
  y = 0.4 * m + rnorm(n, 0, 5)   # happiness
)
```



Bootstrapping



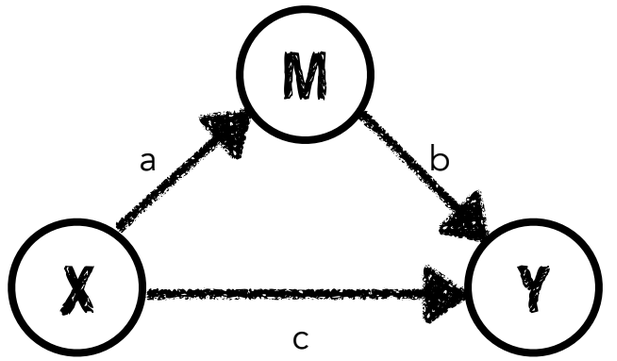

```
1  library("mediation")
2
3  # bootstrapped mediation
4  fit.mediation = mediate(model.m = fit.m_x,
5                          model.y = fit.y_mx,
6                          treat = "x",
7                          mediator = "m",
8                          boot = T)
9
10 # summarize results
11 fit.mediation %>% summary()
```

$$\hat{m} = b_0 + b_1 \cdot x$$

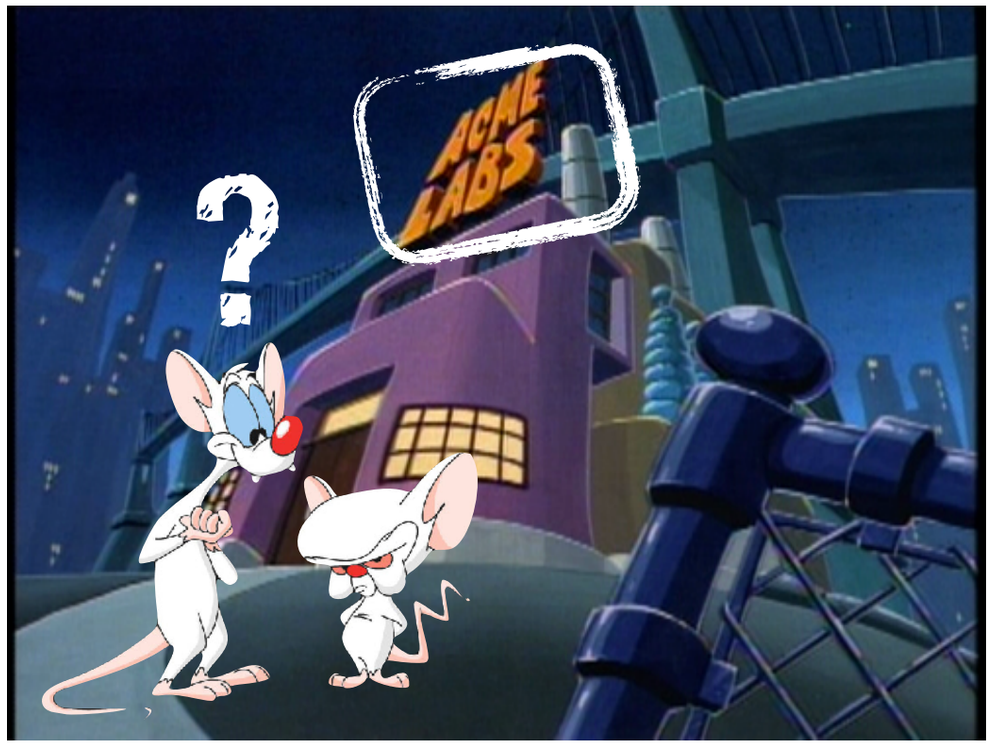
$$\hat{y} = b_0 + b_1 \cdot m + b_2 \cdot x$$

```
Causal Mediation Analysis

Nonparametric Bootstrap Confidence Intervals with the Percentile Method

               Estimate 95% CI Lower 95% CI Upper p-value
ACME            0.28078      0.14059         0.42  <2e-16 ***
ADE            -0.11179     -0.29276         0.10   0.272
Total Effect    0.16899     -0.00415         0.34   0.064 .
Prop. Mediated  1.66151     -3.22476        11.46   0.064 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Sample Size Used: 100


Simulations: 1000
```

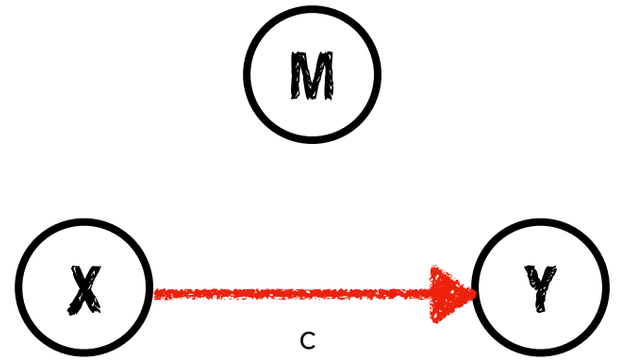
Bootstrapping




```
Causal Mediation Analysis

Nonparametric Bootstrap Confidence Intervals with the Percentile Method

                Estimate 95% CI Lower 95% CI Upper p-value
ACME             0.28078      0.14059         0.42  <2e-16 ***
ADE             -0.11179     -0.29276         0.10   0.272
Total Effect     0.16899     -0.00415         0.34   0.064 .
Prop. Mediated   1.66151     -3.22476        11.46   0.064 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Sample Size Used: 100


Simulations: 1000
```



Bootstrapping

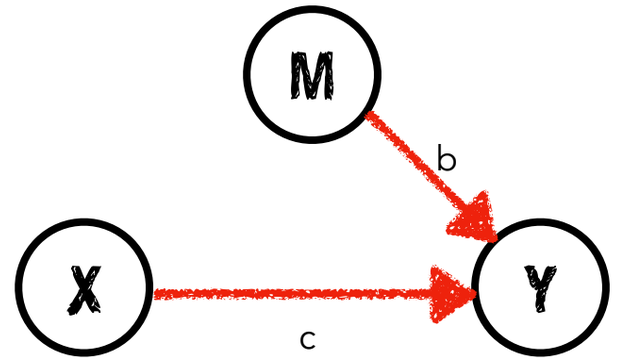



```
Causal Mediation Analysis

Nonparametric Bootstrap Confidence Intervals with the Percentile Method

               Estimate 95% CI Lower 95% CI Upper p-value
ACME            0.28078      0.14059          0.42  <2e-16 ***
ADE            -0.11179     -0.29276          0.10   0.272
Total Effect    0.16899     -0.00415          0.34   0.064 .
Prop. Mediated  1.66151     -3.22476         11.46   0.064 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Sample Size Used: 100


Simulations: 1000
```

$$\hat{y} = b_0 + b_1 \cdot x$$

```
Call:
lm(formula = y ~ 1 + x, data = df.mediation)

Residuals:
    Min      1Q  Median      3Q     Max
-10.917  -3.738  -0.259   2.910  12.540

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.78300    6.16002   1.426   0.1571
x            0.16899    0.08116   2.082   0.0399 *
```

Bootstrapping




```
Causal Mediation Analysis

Nonparametric Bootstrap Confidence Intervals with the Percentile Method

                Estimate 95% CI Lower 95% CI Upper p-value
ACME             0.28078      0.14059          0.42  <2e-16 ***
ADE             -0.11179     -0.29276          0.10   0.272
Total Effect     0.16899     -0.00415          0.34   0.064 .
Prop. Mediated   1.66151     -3.22476         11.46   0.064 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Sample Size Used: 100


Simulations: 1000
```

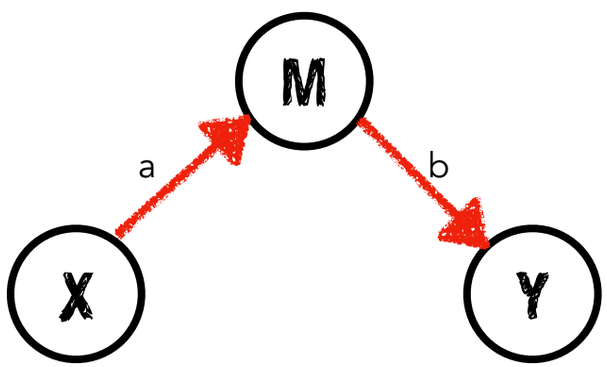
$$\hat{y} = b_0 + b_1 \cdot m + b_2 \cdot x$$ **ADE:** Average direct effect

```
Call:
lm(formula = y ~ 1 + m + x, data = df.mediation)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3651 -3.3037 -0.6222  3.1068 10.3991

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.80952    5.68297   1.374    0.173
m            0.42381    0.09899   4.281 4.37e-05 ***
x           -0.11179    0.09949  -1.124    0.264
```

Bootstrapping

```
Causal Mediation Analysis

Nonparametric Bootstrap Confidence Intervals with the Percentile Method

               Estimate 95% CI Lower 95% CI Upper p-value
ACME            0.28078      0.14059         0.42  <2e-16 ***
ADE            -0.11179     -0.29276         0.10   0.272
Total Effect    0.16899     -0.00415         0.34   0.064 .
Prop. Mediated  1.66151     -3.22476        11.46   0.064 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Sample Size Used: 100


Simulations: 1000
```

ACME



indirect effect: a * b

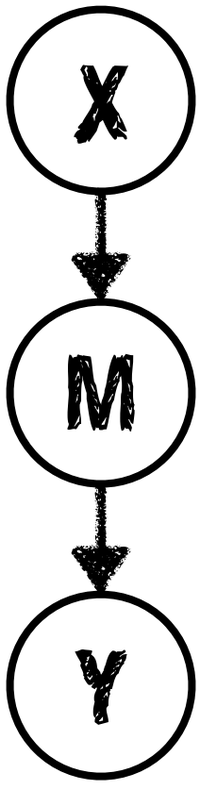
ADE: Average direct effect

ACME: Average causal mediation effect

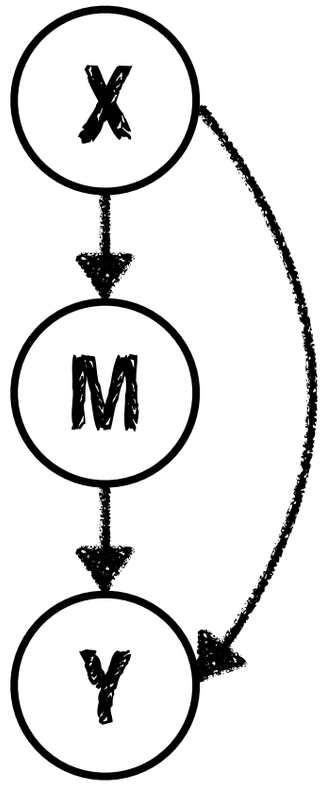
ACME = Total effect - ADE

Underlying causal model

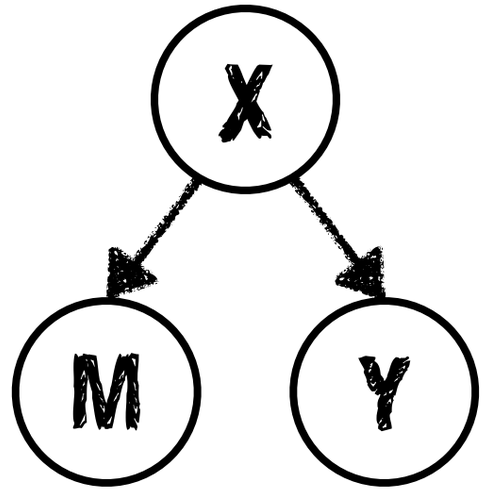
Full mediation



Partial mediation



No mediation

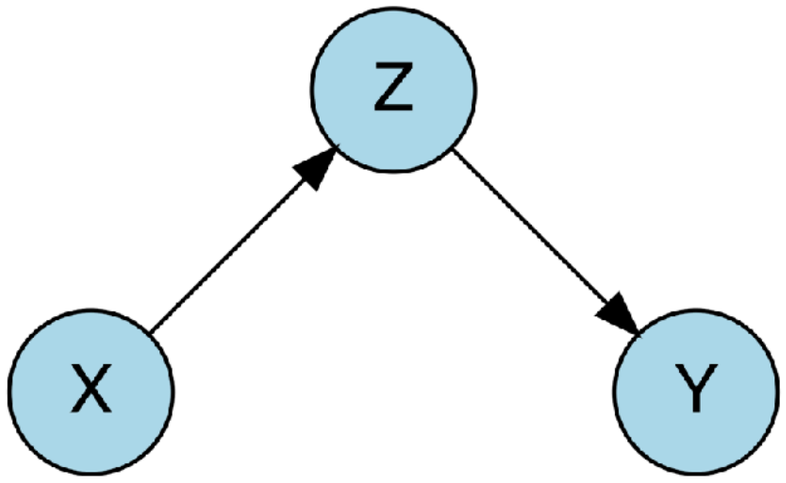


Full mediation: When the effect of **X** on **Y** completely disappears, **M** fully mediates between **X** and **Y**.

Partial mediation: When the effect of **X** on **Y** still exists, but in a smaller magnitude, **M** partially mediates between **X** and **Y**.

Beware of mediation analyses!!!

```r
set.seed(1)

n = 100 # number of observations

# causal chain
df.causal_chain = tibble(x = rnorm(n, 0, 1),
                         z = 2 * x + rnorm(n, 0, 1),
                         y = 2 * z + rnorm(n, 0, 1))
```

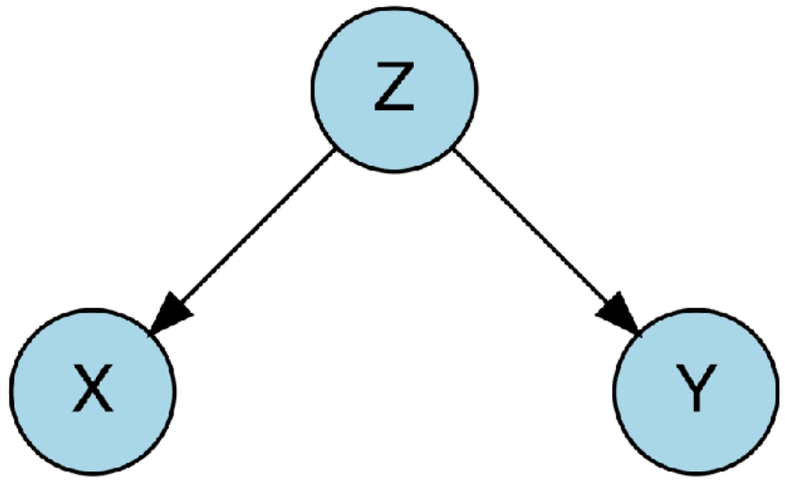



```
Causal Mediation Analysis

Nonparametric Bootstrap Confidence Intervals with the Percentile Method

               Estimate 95% CI Lower 95% CI Upper p-value
ACME             0.8287       0.6234         1.05  <2e-16 ***
ADE             -0.0535      -0.2548         0.15    0.55
Total Effect     0.7752       0.6391         0.90  <2e-16 ***
Prop. Mediated   1.0690       0.8131         1.35  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Sample Size Used: 100


Simulations: 1000
```

nice mediation result!

Beware of mediation analyses!!!

```r
set.seed(1)

n = 100 # number of observations

# common cause
df.common_cause = tibble(z = rnorm(n, 0, 1),
                         x = 2 * z + rnorm(n, 0, 1),
                         y = 2 * z + rnorm(n, 0, 1))
```

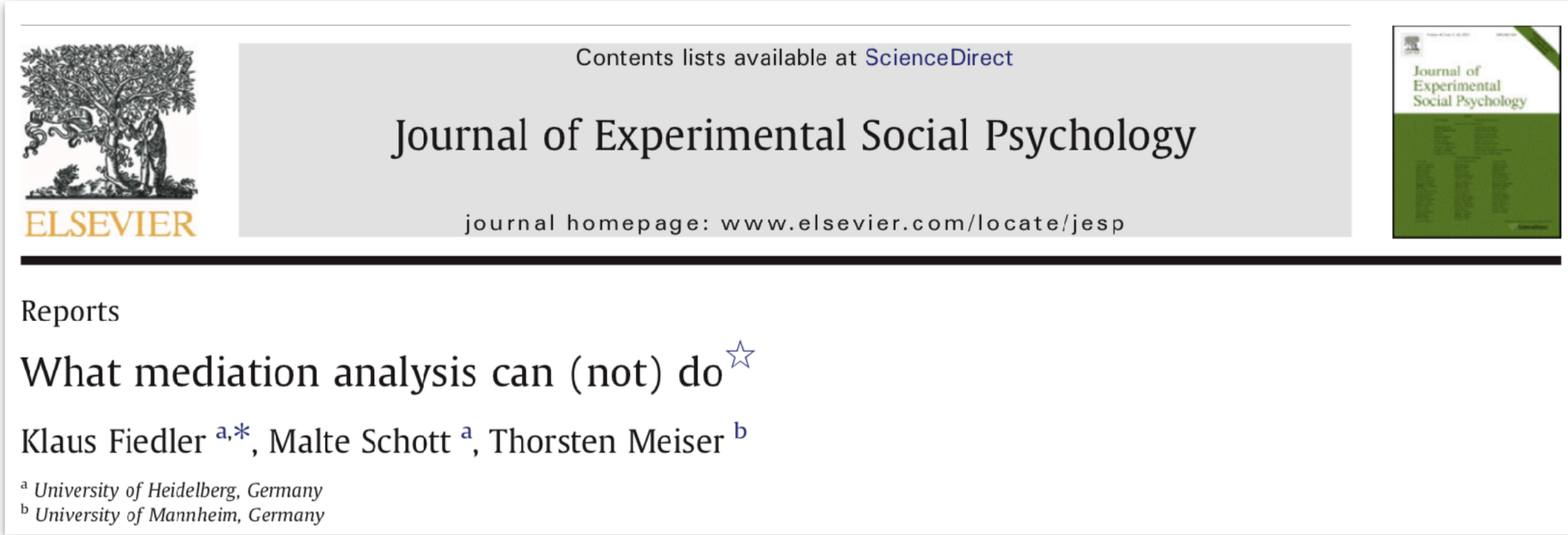



```
Causal Mediation Analysis

Nonparametric Bootstrap Confidence Intervals with the Percentile Method

               Estimate 95% CI Lower 95% CI Upper p-value
ACME             0.8287       0.6065         1.04  <2e-16 ***
ADE             -0.0535      -0.2675         0.16    0.56
Total Effect     0.7752       0.6353         0.90  <2e-16 ***
Prop. Mediated   1.0690       0.8134         1.37  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Sample Size Used: 100


Simulations: 1000
```

(not) nice mediation result!

Limitations

- correlational analysis
  - we need theories / experiments to tease apart causes and effects to properly map our variables onto the diagram



Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Reports

What mediation analysis can (not) do ☆

Klaus Fiedler [a,*], Malte Schott [a], Thorsten Meiser [b]

[a] University of Heidelberg, Germany
[b] University of Mannheim, Germany

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, 47(6), 1231-1236.

Limitations

many-to-one mapping

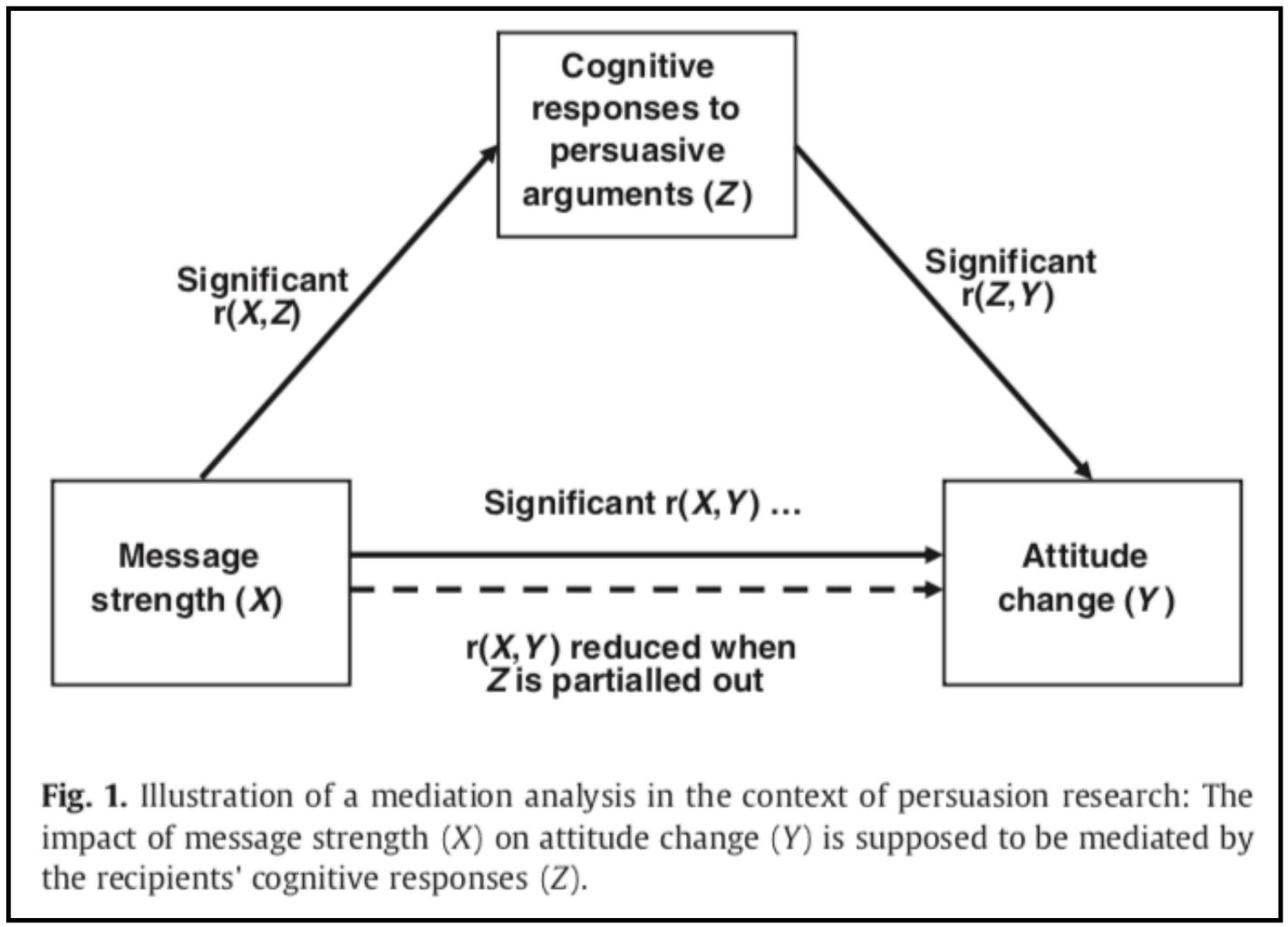


Fig. 1. Illustration of a mediation analysis in the context of persuasion research: The impact of message strength (*X*) on attitude change (*Y*) is supposed to be mediated by the recipients' cognitive responses (*Z*).

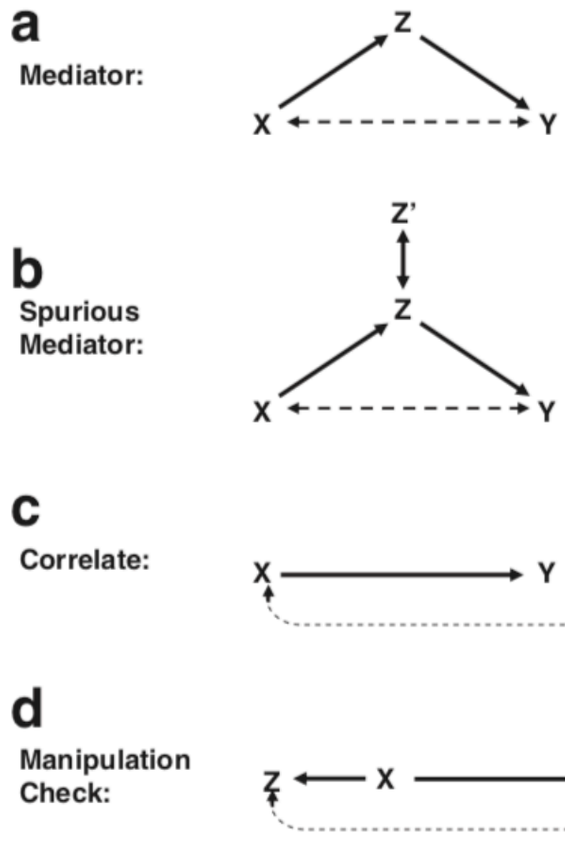


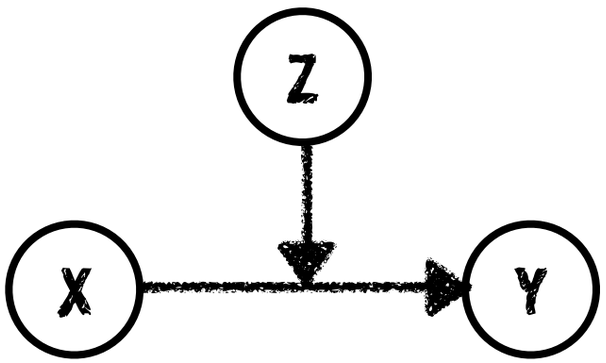
Fig. 2. Different causal models for the interpretation of the same intercorrelations between three variables, *X*, *Y*, and *Z*.

only experiments allow us to tell apart possible causal structures

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, 47(6), 1231-1236.

Moderation

Definition

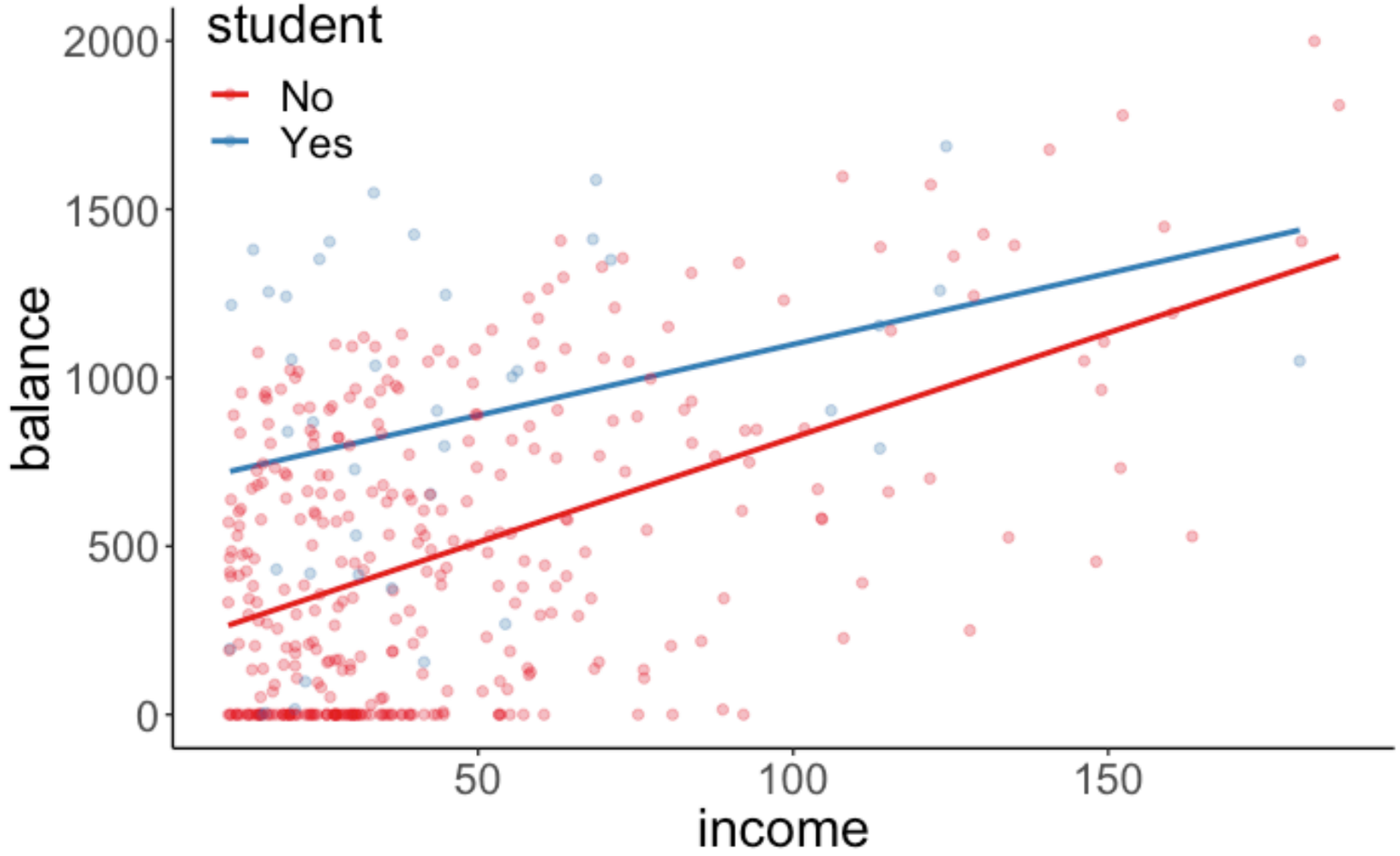


Moderation means that the effect of a predictor depends on the value of another.

Here, the nature of the relationship between **X** and **Y** depends on **Z**.

Have we come across moderation already?

Relationship between credit card balance, income, and whether the person is a student.



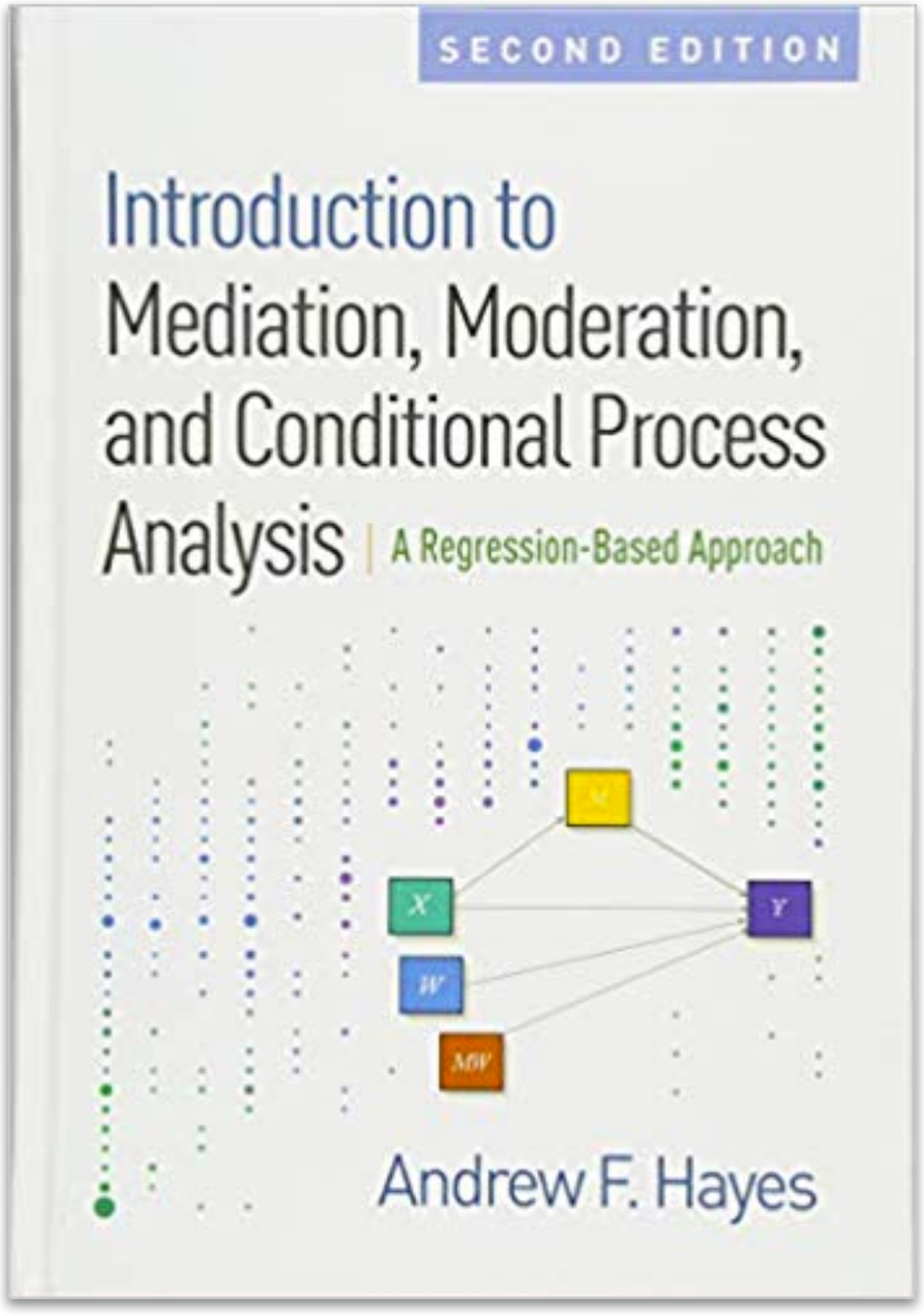
$$\widehat{\text{balance}}_i = 200.62 + 6.22 \cdot \text{income}_i + 476.68 \cdot \text{student}_i - 2.00 \cdot (\text{income}_i \times \text{student}_i)$$

if student = "No" $\widehat{\text{balance}}_i = 200.62 + 6.22 \cdot \text{income}_i$

if student = "Yes"

$$\begin{aligned}\widehat{\text{balance}}_i &= 200.62 + 6.22 \cdot \text{income}_i + 476.68 \cdot 1 - 2.00 \cdot (\text{income}_i \times 1) \\ &= 677.3 + 6.22 \cdot \text{income}_i - 2.00 \cdot \text{income}_i \\ &= 677.3 + 4.22 \cdot \text{income}_i\end{aligned}$$

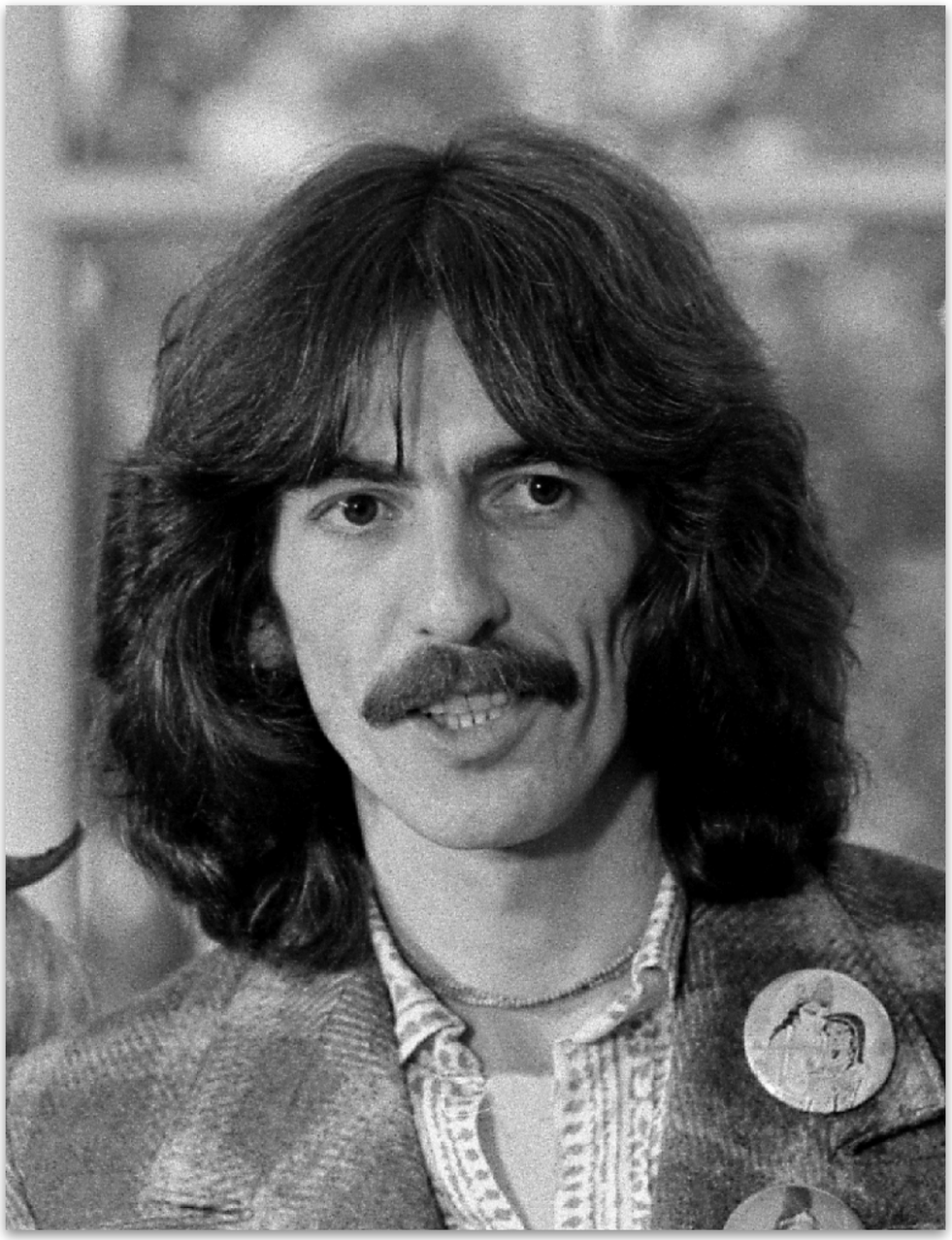
Learn more about mediation and moderation



Recoded with `brms` by Solomon Kurz here: https://bookdown.org/connect/#/apps/1523/access

Linear mixed effects models

Are faces of people born on February 25th more trustworthy than faces of people born on February 26th?



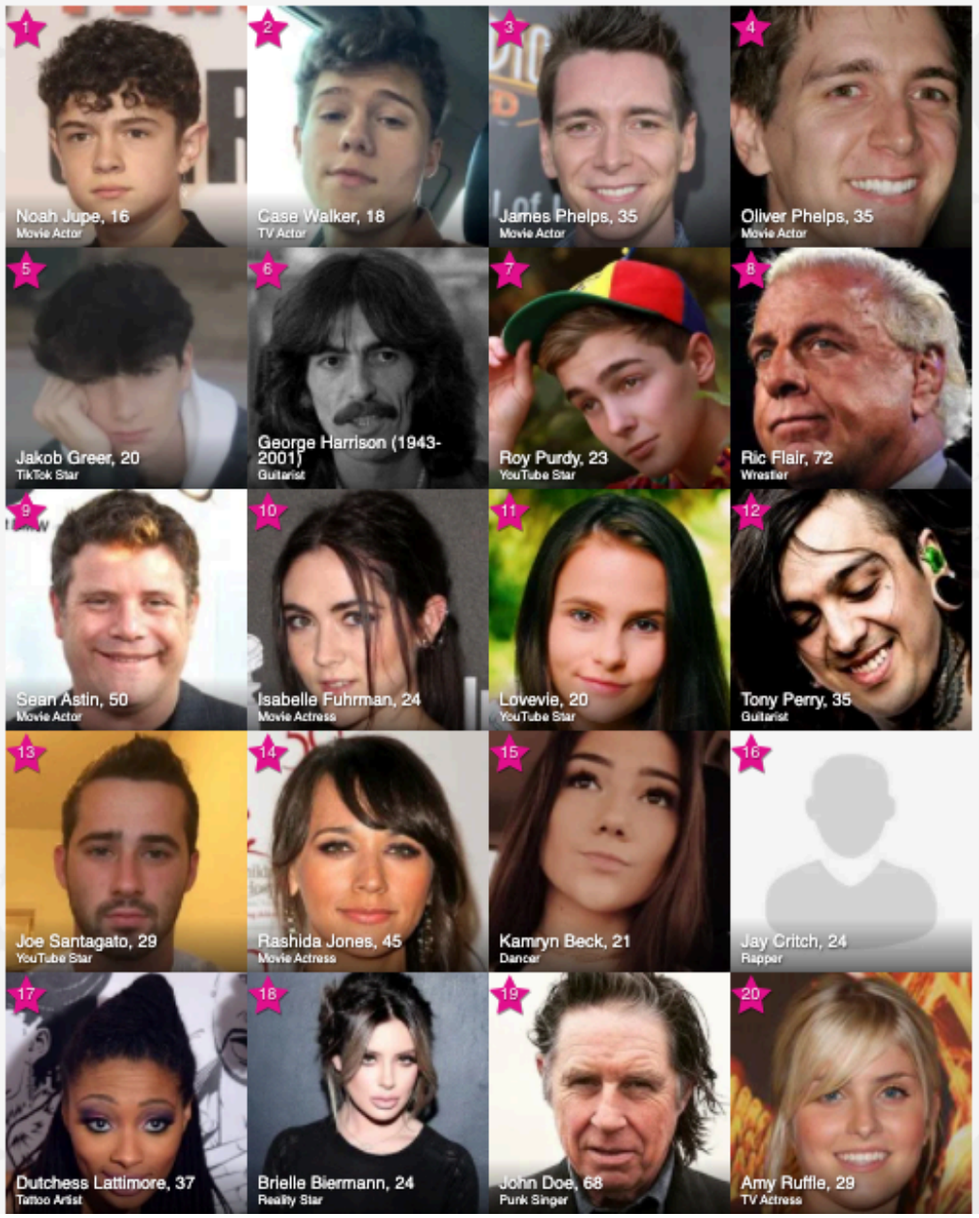
born on February 25th

born on February 26th

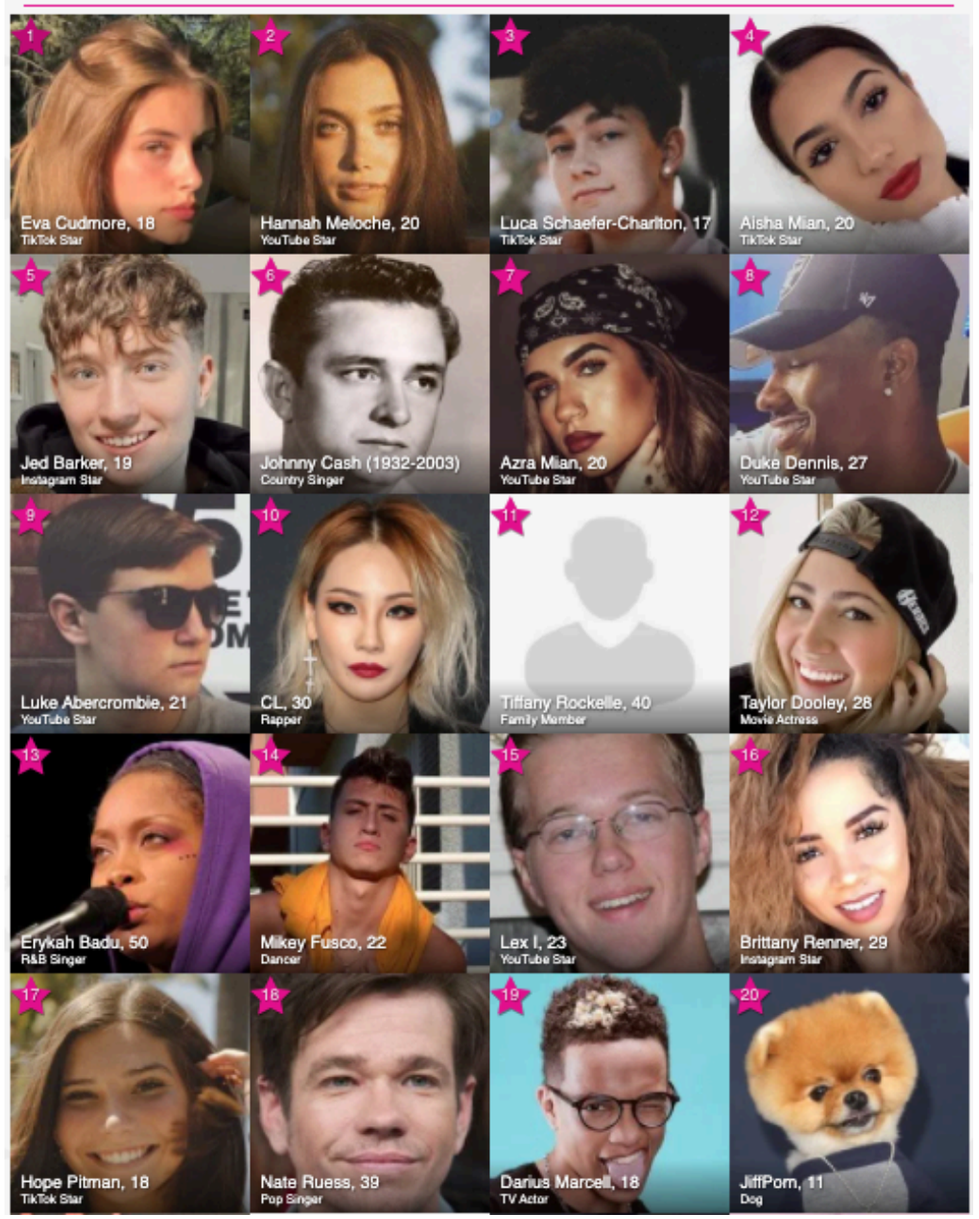
N = 100 participants

```
1 lm(formula = trustworthy ~ 1 + birthday,
2     data = df.birthday)
```

Are faces of people born on February 25th more trustworthy than faces of people born on February 26th?



born on February 25th

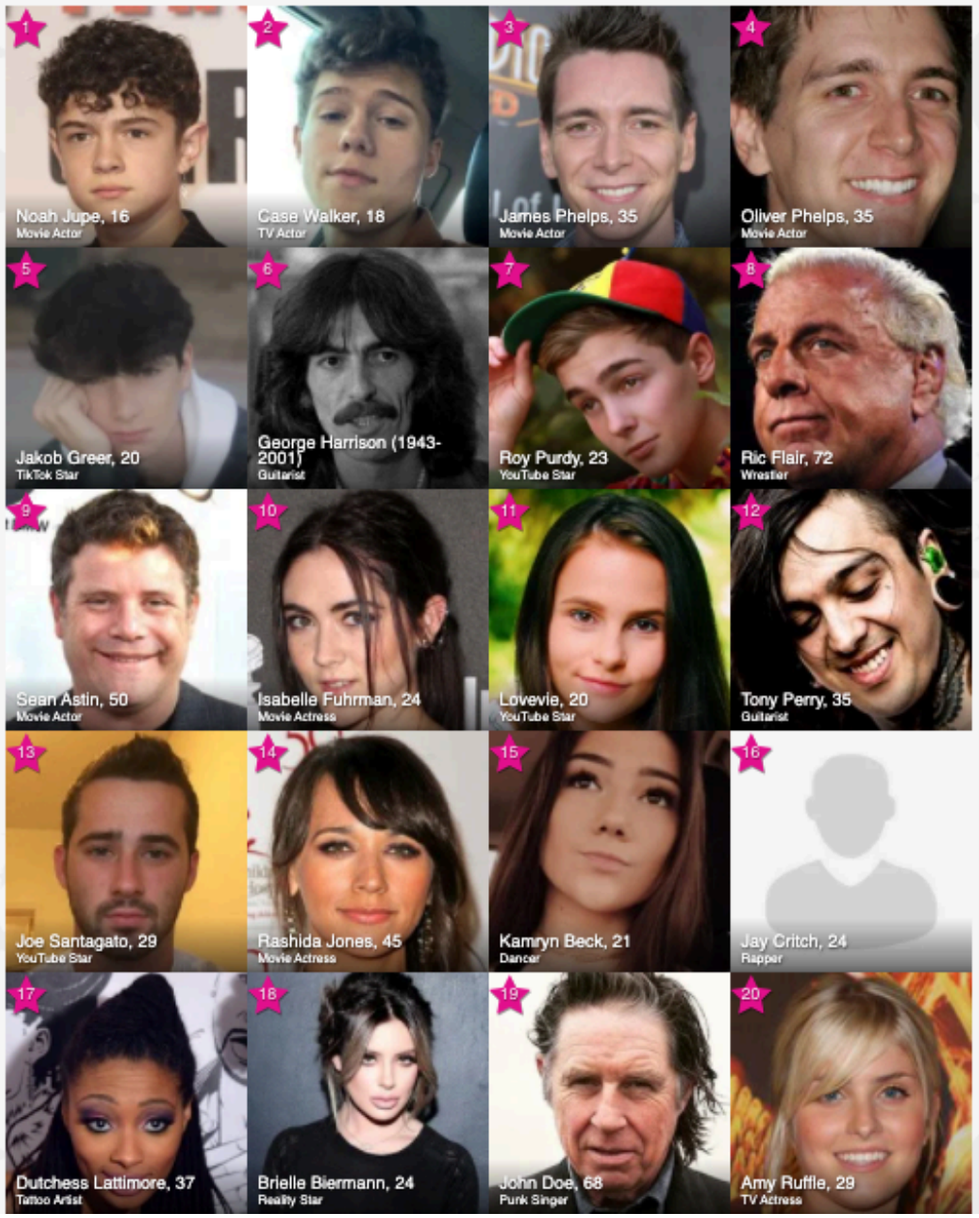


born on February 26th

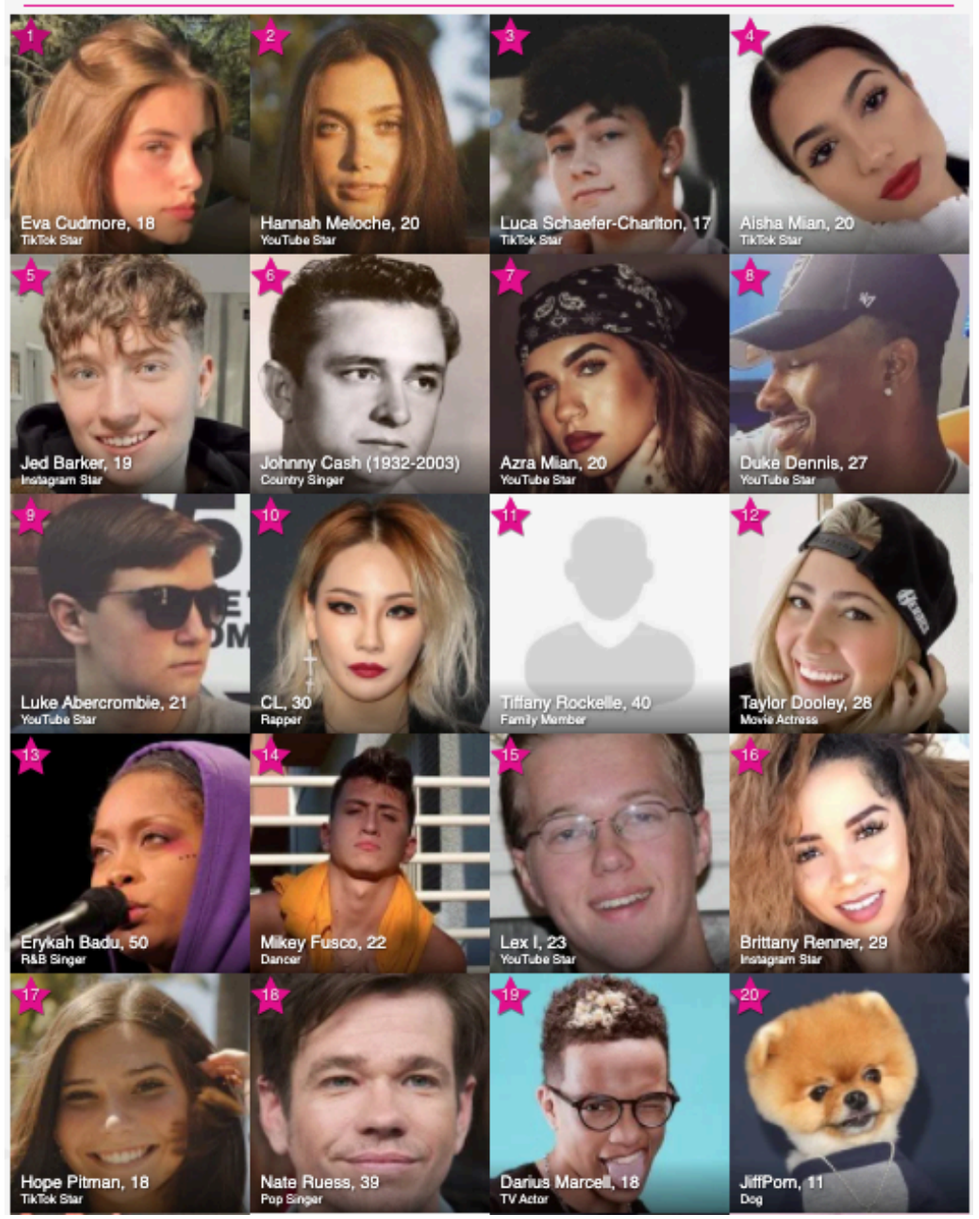
N = 2 participants

```
1 lm(formula = trustworthy ~ 1 + birthday,
2     data = df.birthday)
```

Are faces of people born on February 25th more trustworthy than faces of people born on February 26th?



born on February 25th



born on February 26th

N = 100 participants

```
1 lmer(formula = trustworthy ~ 1 + birthday +
                    (1 | item) + (1 | participant),
2       data = df.birthday)
```

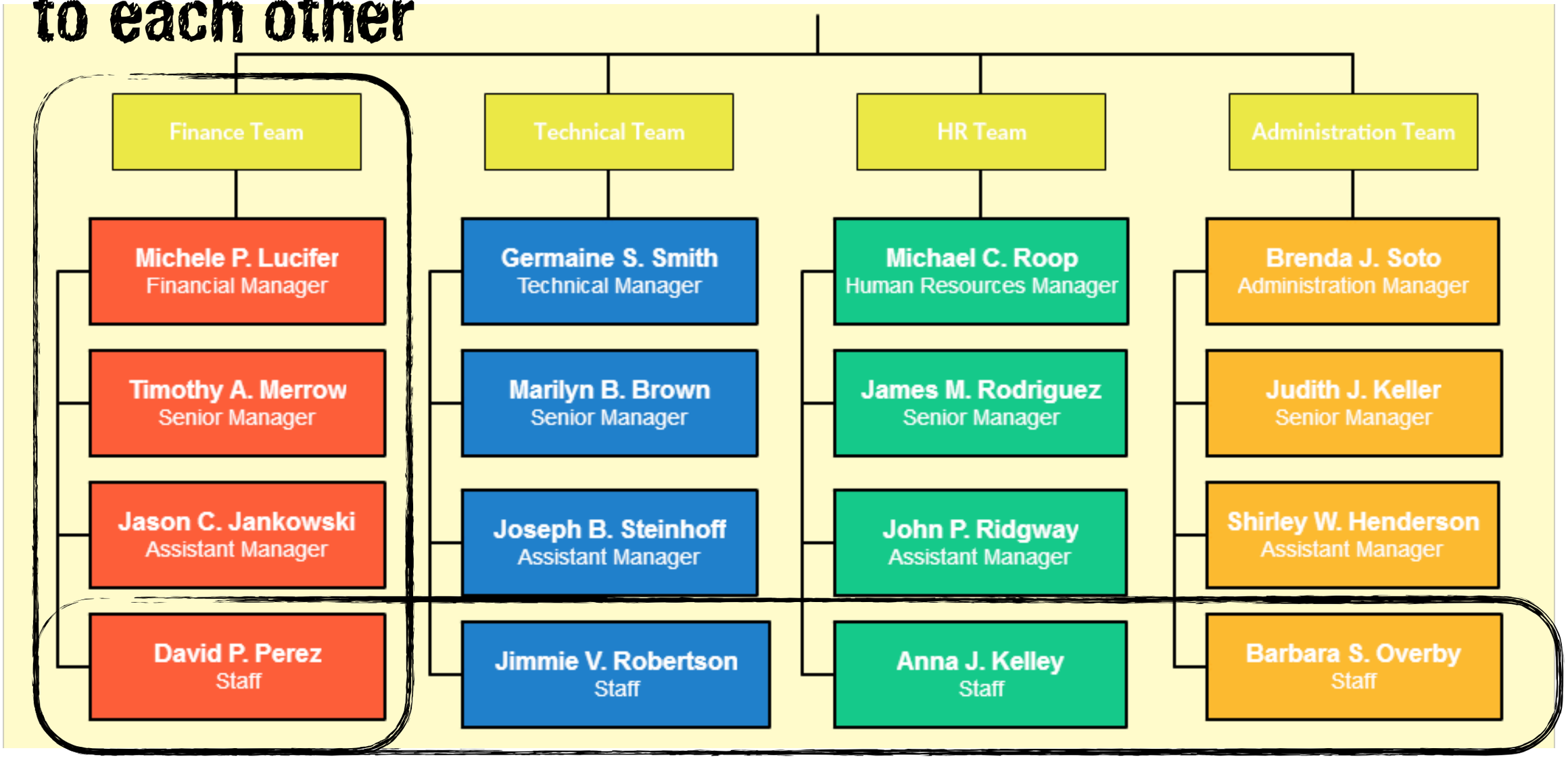
Dependence

- so far, all the models that we've discussed (linear model with different kinds of predictors and contrasts) make the assumption that the data are **iid** (independent, and identically distributed)

- often this assumption is violated
  - **psychology experiments**: many observations from the same participants
  - **survey data**: different populations between different states in the US
  - **time series**: distribution at $t + 1$ depends on t

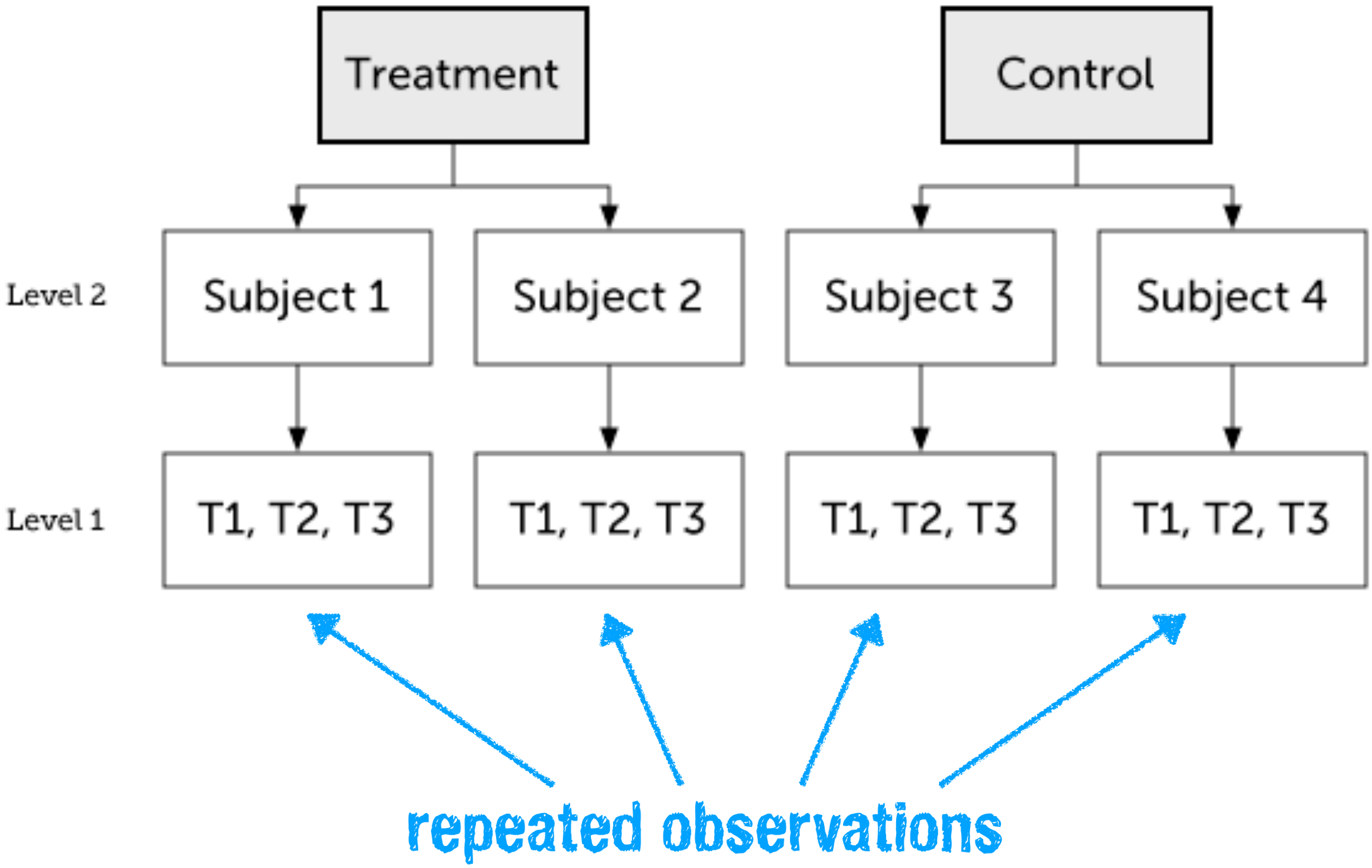
Main use cases: **Hierarchical models**

more similar to each other



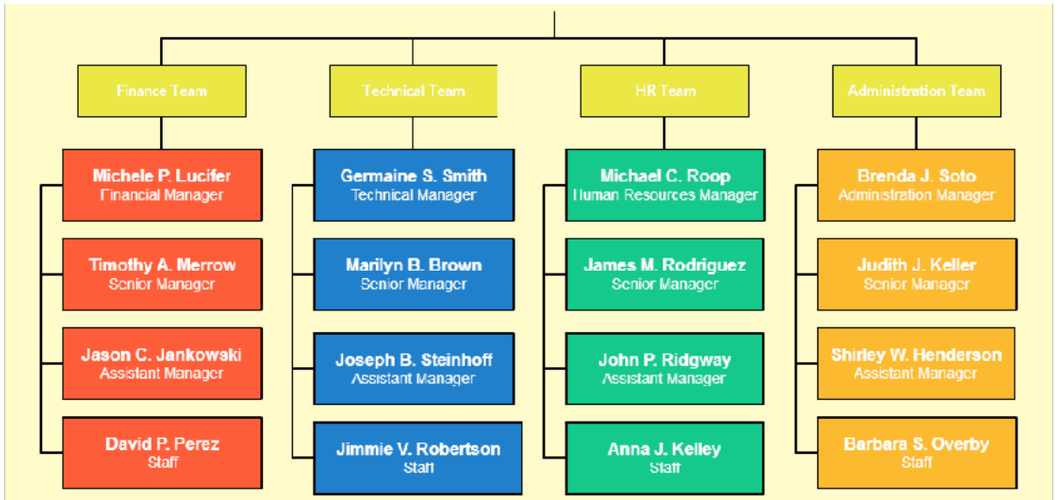
less similar to each other

Main use cases: **Longitudinal models**

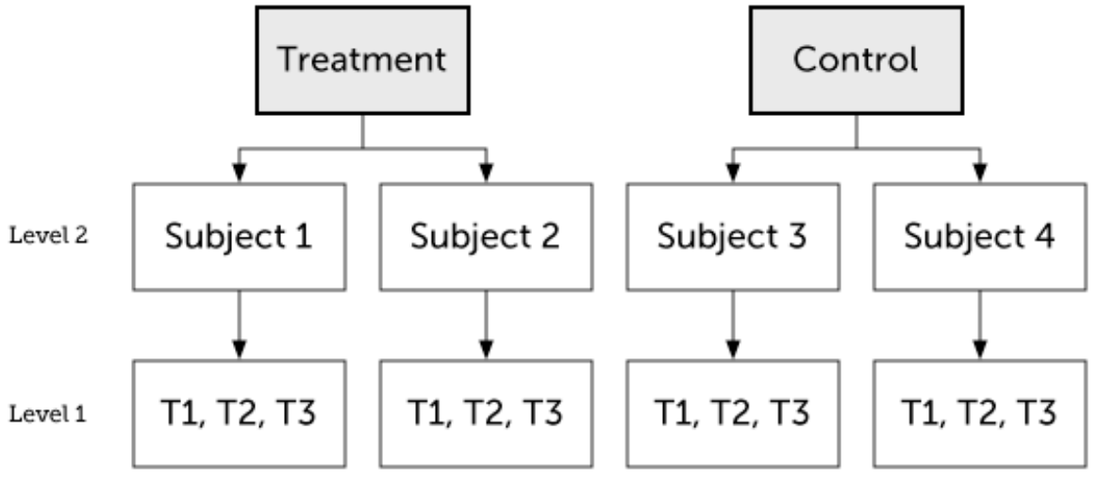


Linear mixed effects models

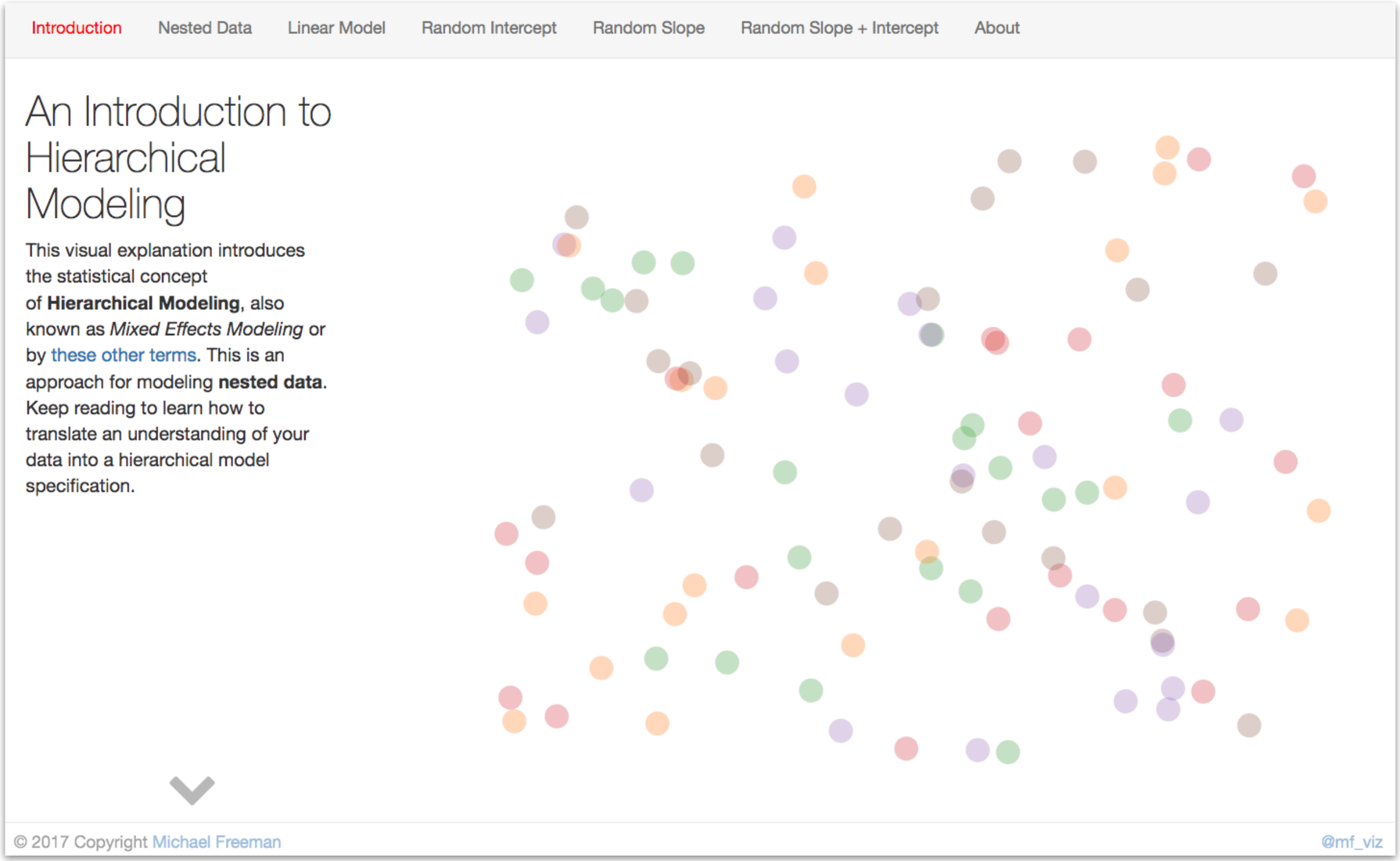
Hierarchical models



Longitudinal models



- allow us to account for dependencies in our data
- **hierarchical models**: schools > teachers > students
- **longitudinal models**: repeated observations from the same people



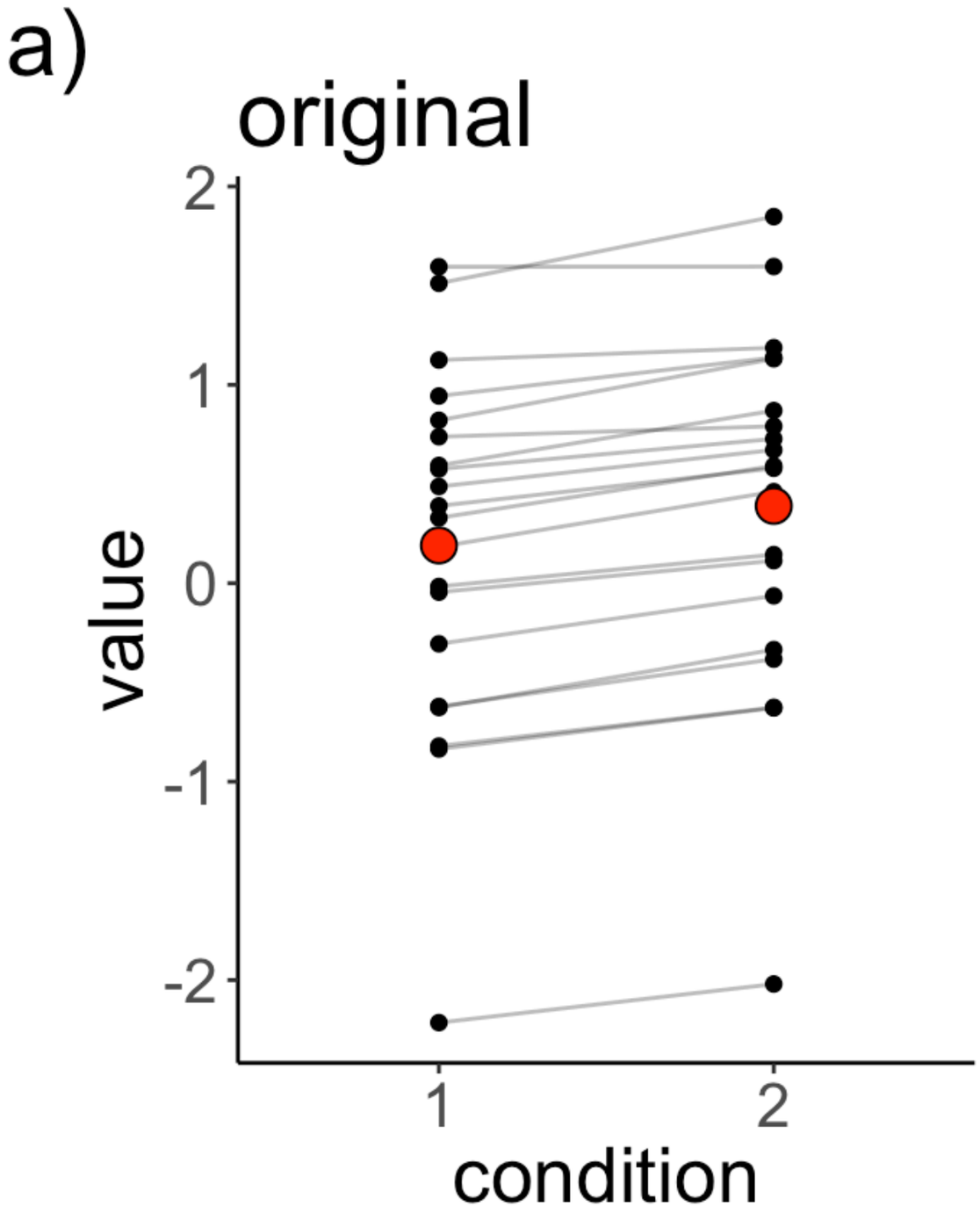
http://mfviz.com/hierarchical-models/

Modeling dependence in data

Dependence

Does it really matter?

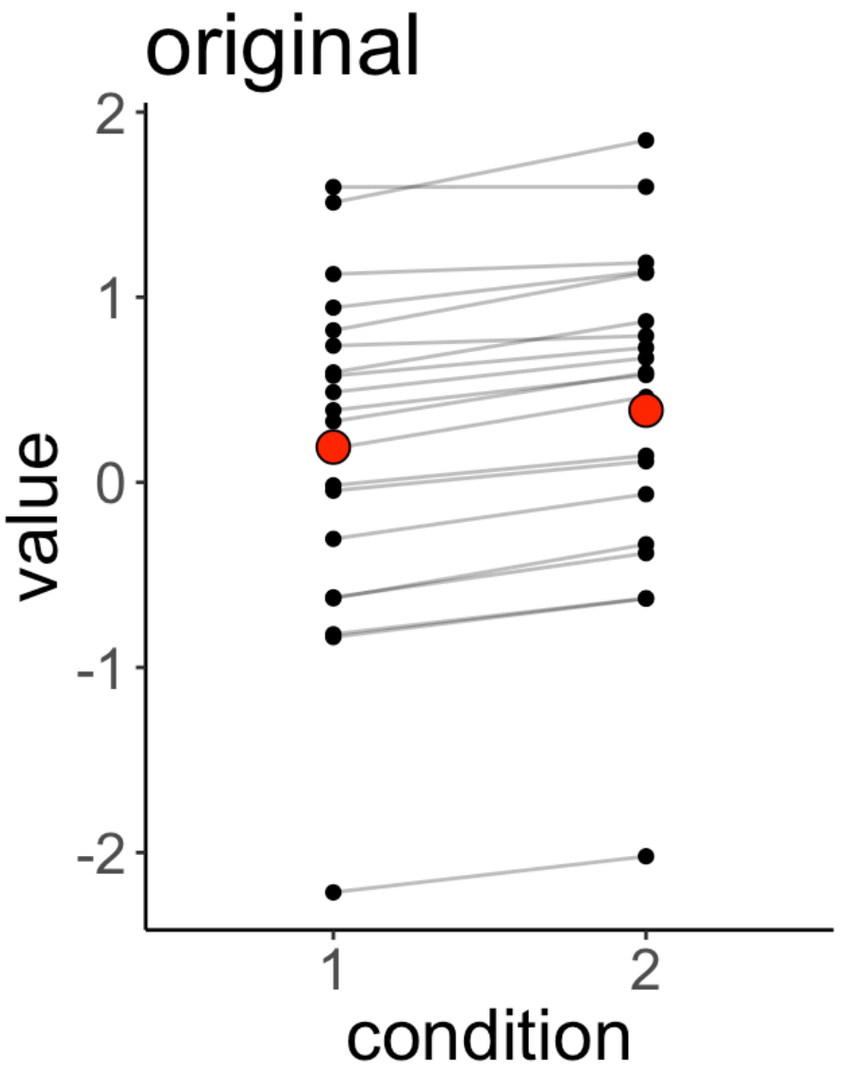
Is there a significant difference between conditions 1 and 2?



Dependence

assuming independence!

```
1 # linear model
2 lm(formula = value ~ condition,
3     data = df.original) %>%
4   summary()
```




```
Call:
lm(formula = value ~ condition, data = df.original)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4100 -0.5530  0.1945  0.5685  1.4578

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1905     0.2025   0.941    0.353
condition2    0.1994     0.2864   0.696    0.491

Residual standard error: 0.9058 on 38 degrees of freedom
Multiple R-squared:  0.01259,    Adjusted R-squared:  -0.0134
F-statistic: 0.4843 on 1 and 38 DF,  p-value: 0.4907
```

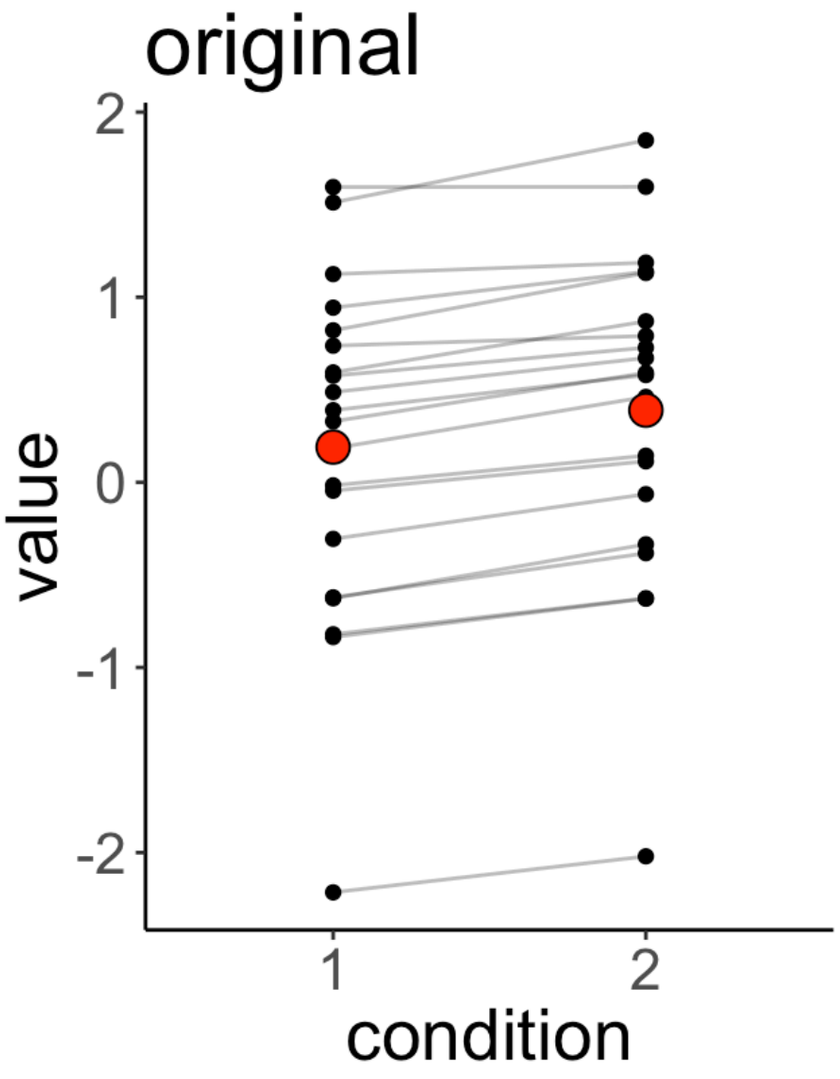
- we ignore the fact that we have repeated observations from the same participants
- in the data it looks like there is a small but consistent effect of condition

meet **`lmer()`**

Dependence

new syntax

```
1 # fit a linear mixed effects model
2 lmer(formula = value ~ condition + (1 | participant),
3                    data = df.original) %>%
4    summary()
```




```
Linear mixed model fit by REML ['lmerMod']
Formula: value ~ condition + (1 | participant)
   Data: df.original

REML criterion at convergence: 17.3

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.55996 -0.36399 -0.03341  0.34400  1.65823

Random effects:
 Groups       Name        Variance Std.Dev.
 participant (Intercept) 0.816722 0.90373
 Residual                 0.003796 0.06161
Number of obs: 40, groups:  participant, 20

Fixed effects:
            Estimate Std. Error t value
(Intercept)  0.19052    0.20255   0.941
condition2   0.19935    0.01948  10.231

Correlation of Fixed Effects:
            (Intr)
condition2 -0.048
```

no p-value!

NO P-VALUE

Dependence

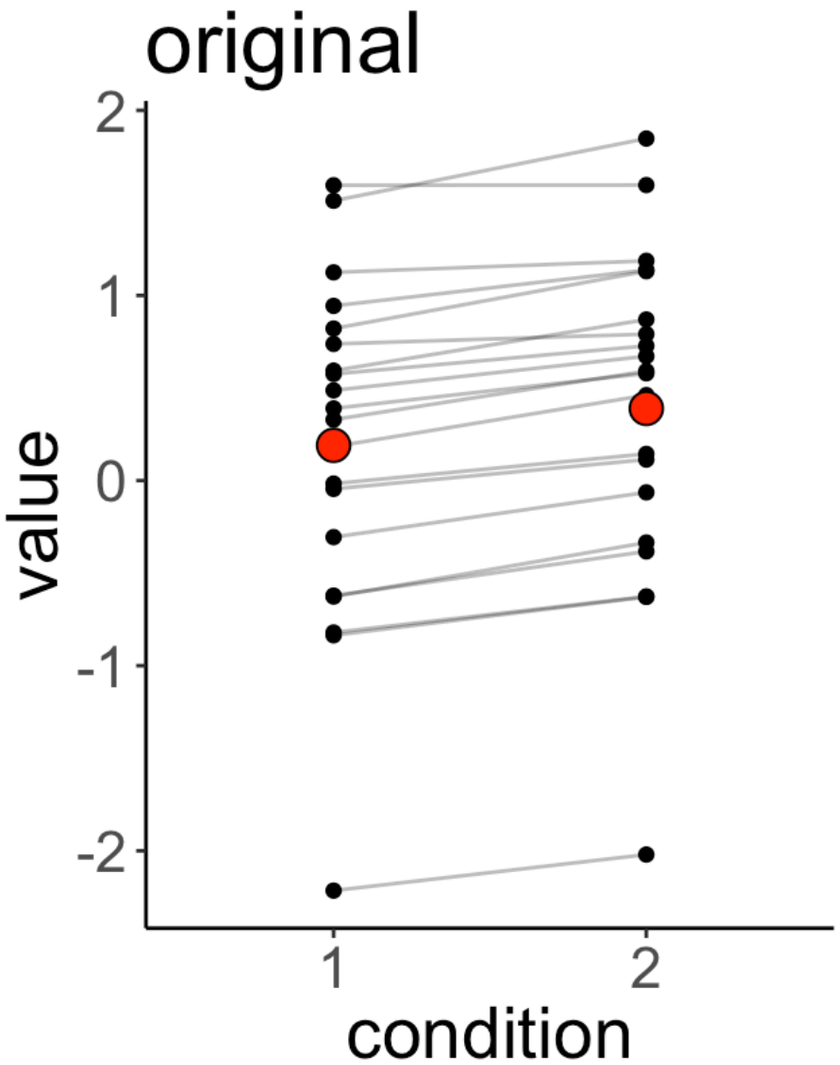
we can still do our good ol model comparison trick

```
1 # fit models
2 fit.compact = lmer(formula = value ~ 1 + (1 | participant),
3                    data = df.original)
4 fit.augmented = lmer(formula = value ~ 1 + condition + (1 | participant),
5                    data = df.original)
6
7 # compare via Chisq-test
8 anova(fit.compact, fit.augmented)
```

```
refitting model(s) with ML (instead of REML)
Data: df.original
Models:
fit.compact: value ~ 1 + (1 | participant)
fit.augmented: value ~ 1 + condition + (1 | participant)
                Df    AIC    BIC   logLik deviance  Chisq Chi Df Pr(>Chisq)
fit.compact      3 53.315 58.382 -23.6575   47.315
fit.augmented    4 17.849 24.605  -4.9247    9.849 37.466      1  9.304e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Dependence

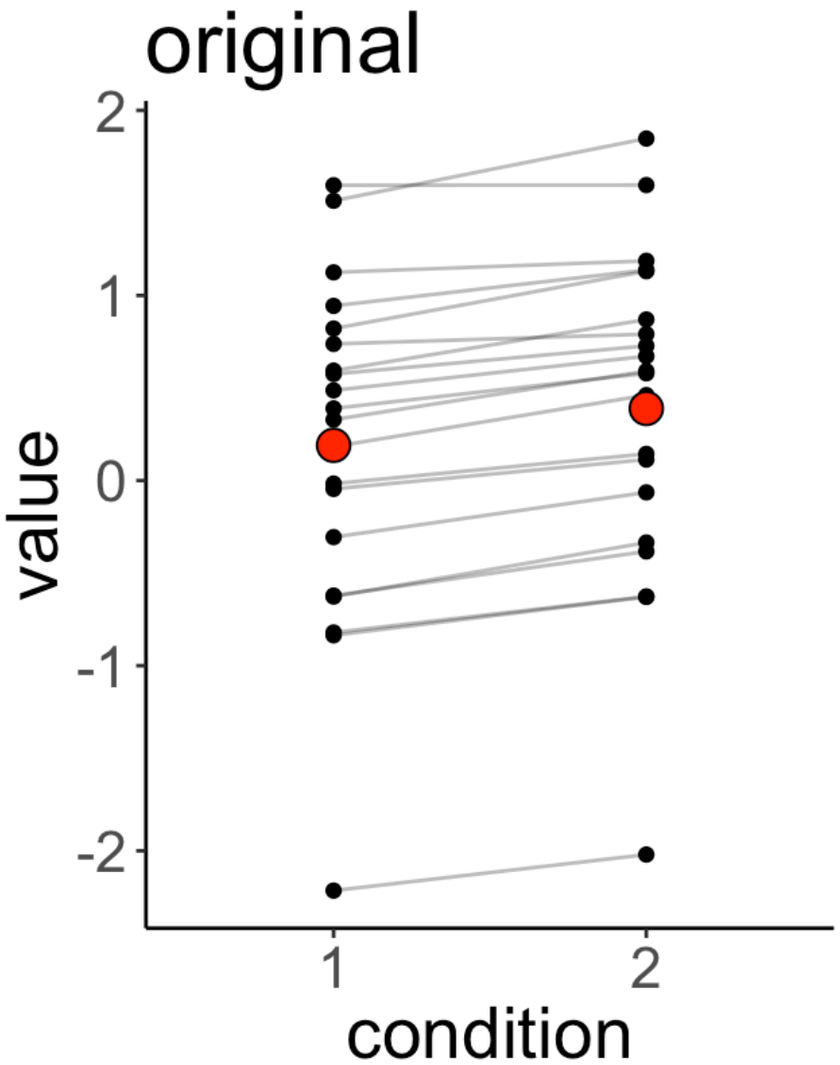
Why is the effect of condition significant when we account for the dependence in the data?



- there are large interindividual differences in the baseline
- the variance explained by the effect of condition is (much) smaller than the interindividual variance
- **but**: the effect of condition is highly consistent

Dependence

Why is the effect of condition significant when we account for the dependence in the data?



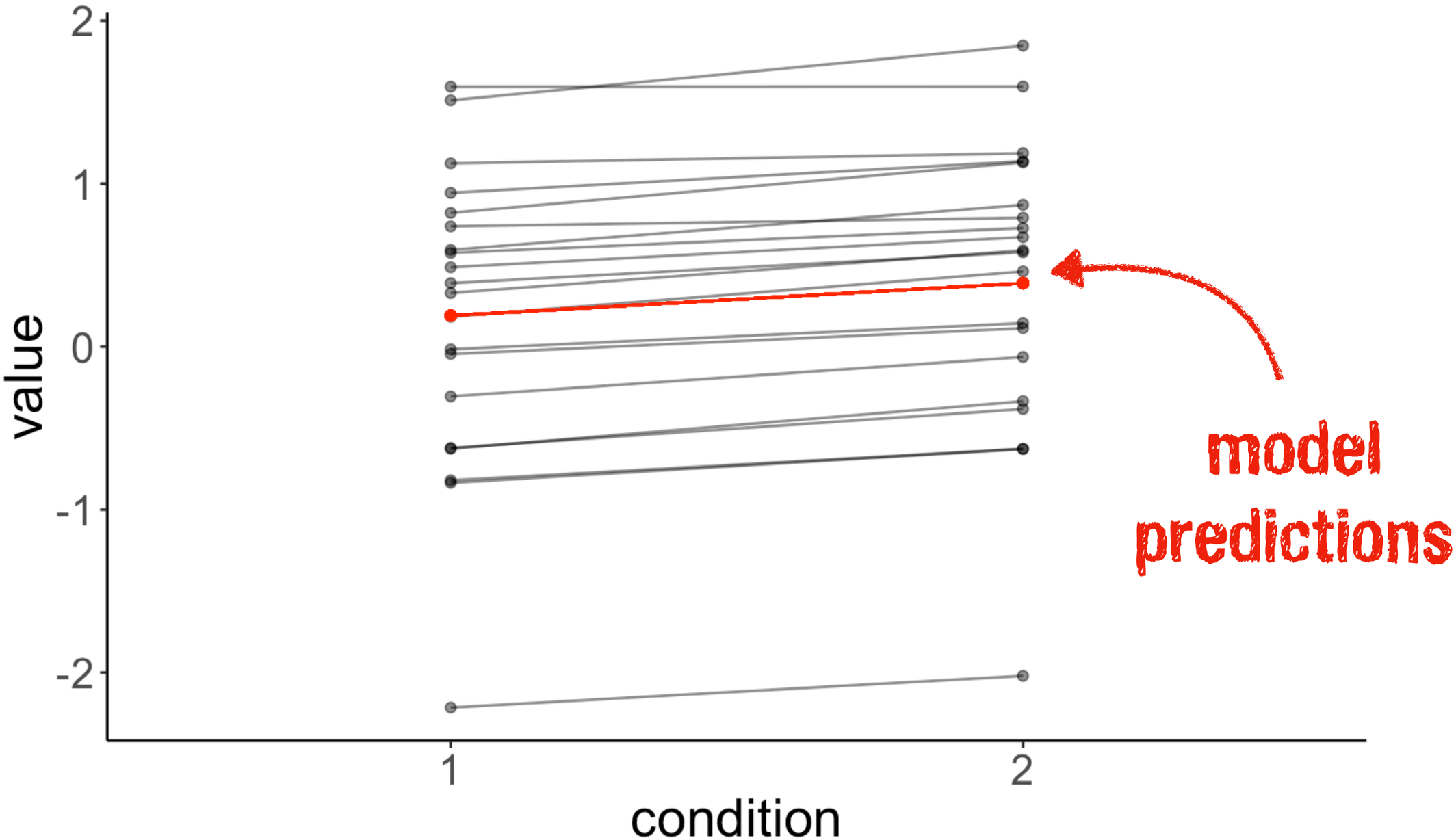
- by explicitly modeling the dependence in the data, we account for the interindividual differences

let's visualize the model predictions!

Linear model (assuming independence)

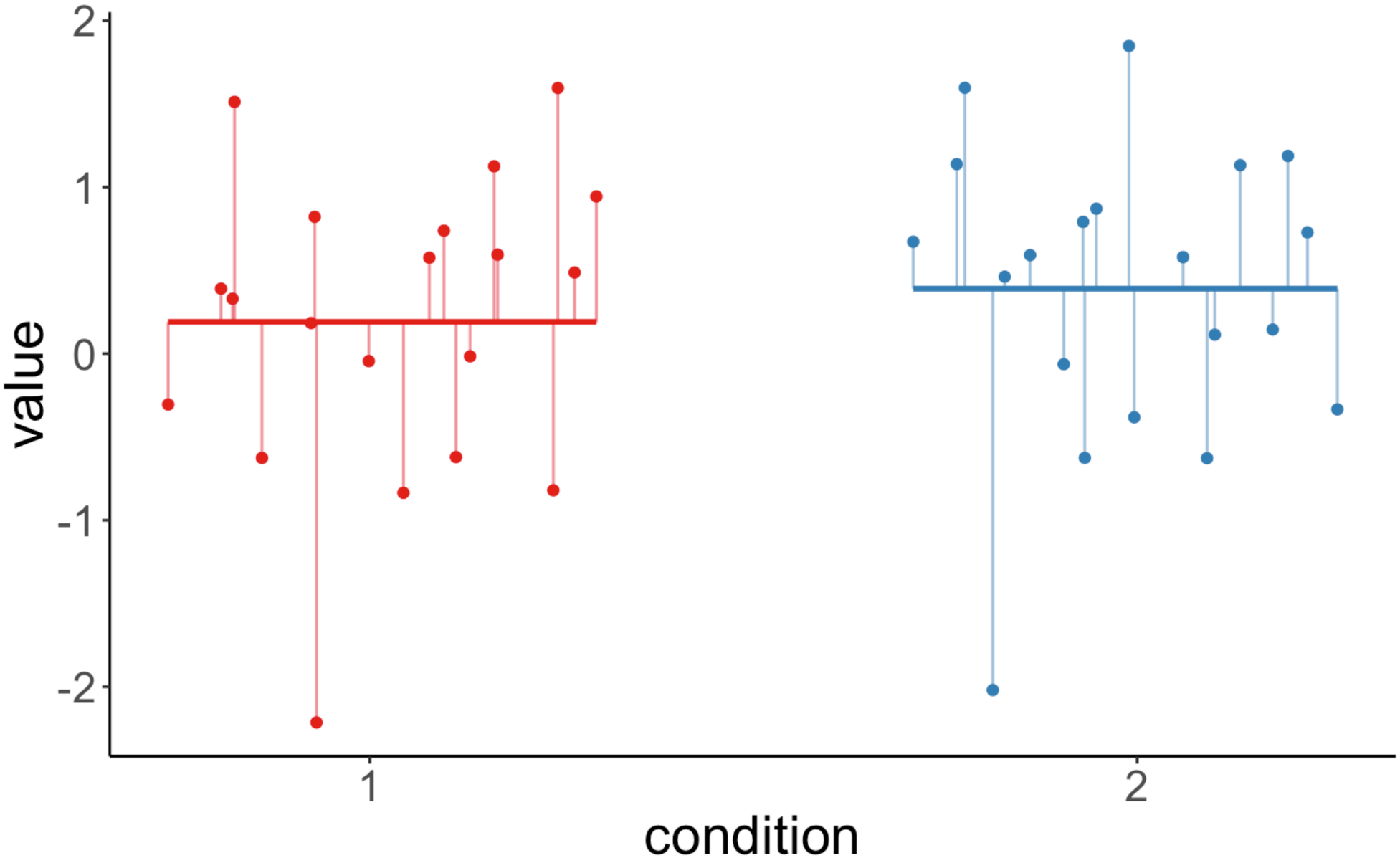
Predictions by the linear model which assumes independence

```
lm(formula = value ~ condition,
    data = df.original)
```



Linear model (assuming independence)

Residuals of the model

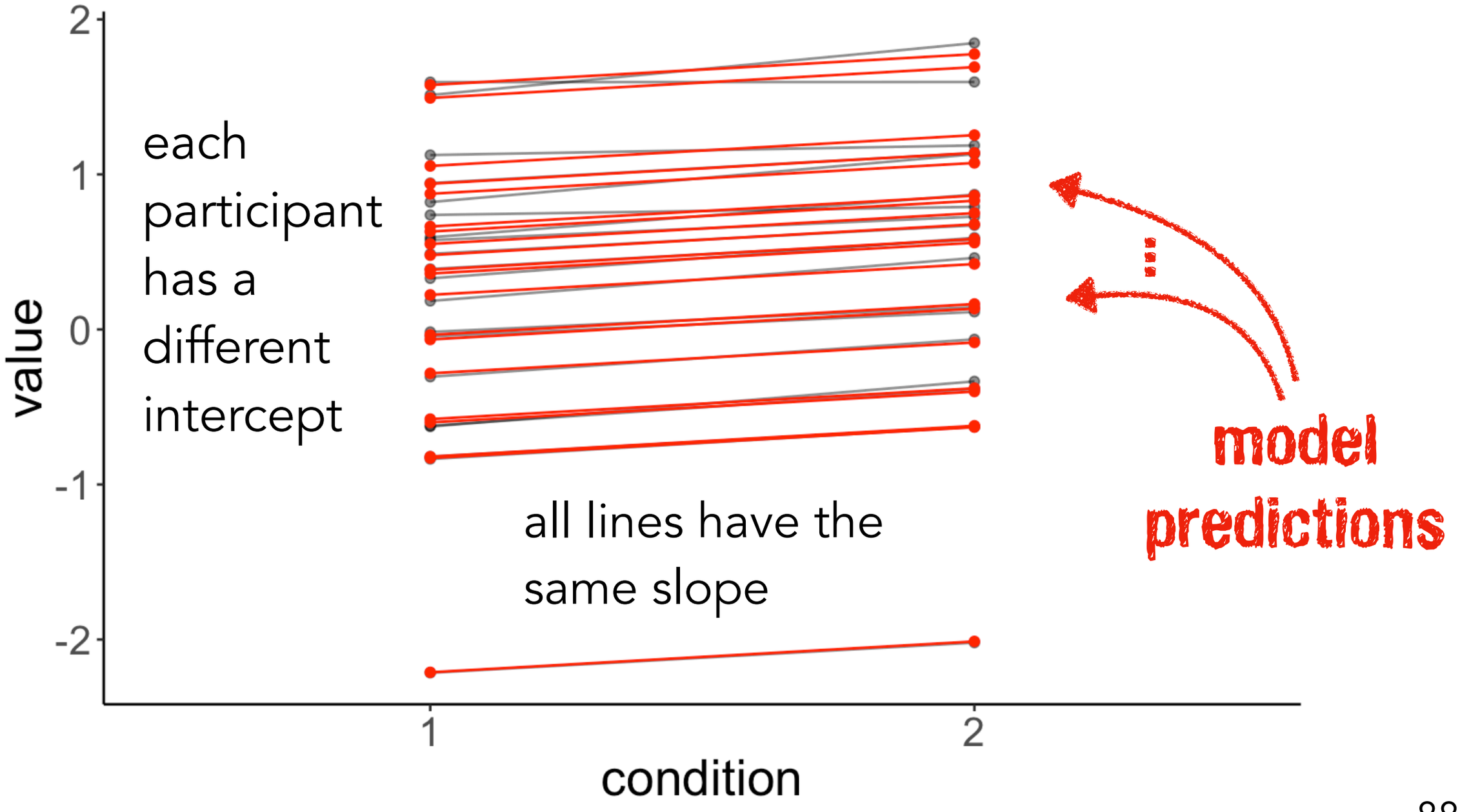


This is not much better than fitting a single line (point).

Linear mixed effects model (accounting for dependence)

Predictions by the linear mixed effects model

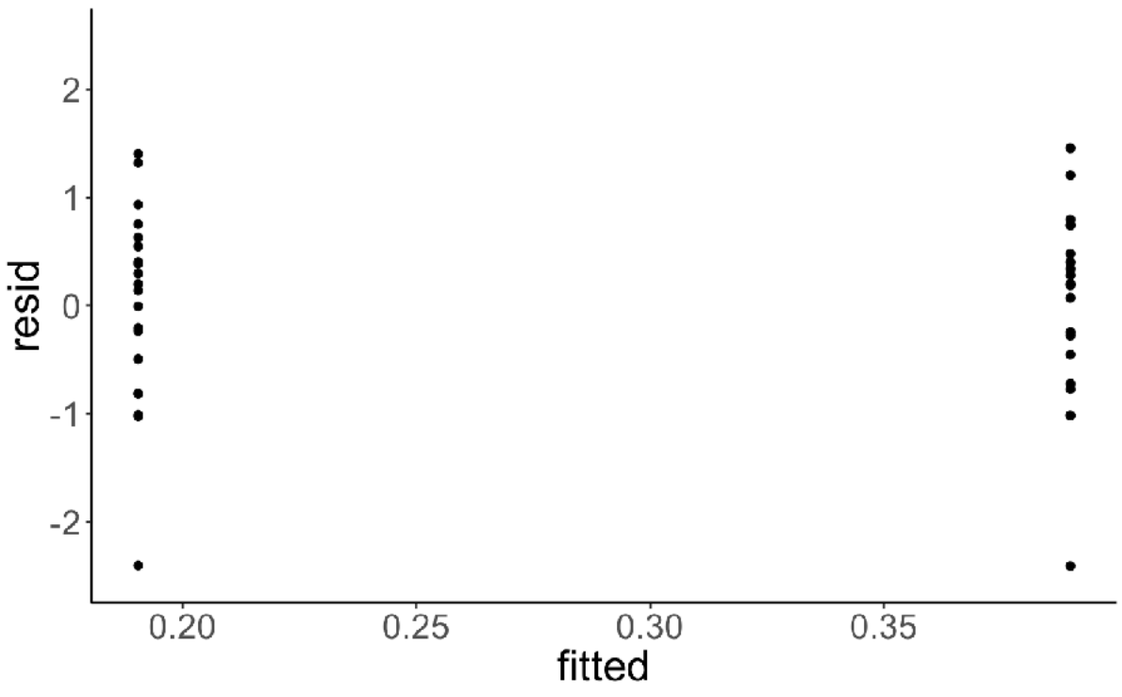
```
lmer(formula = value ~ condition + (1 | participant),
     data = df.original)
```



Model comparison

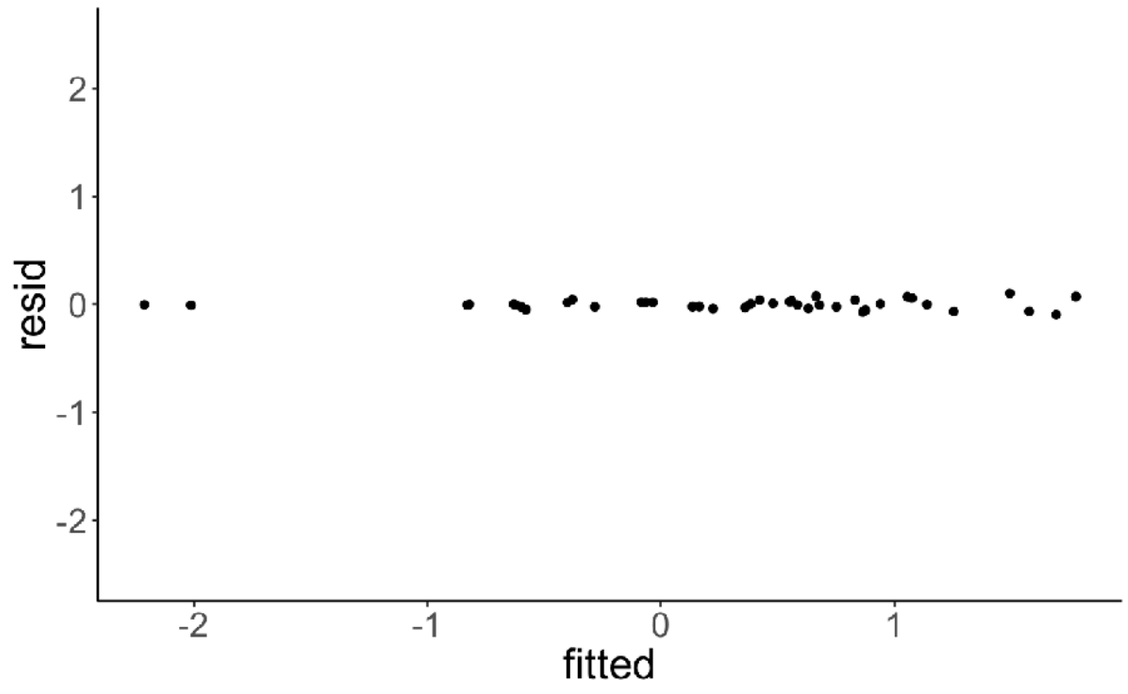
Residual plots

```
lm(formula = value ~ 1 + condition,
   data = df.original)
```



much variance left to be explained

```
lmer(formula = value ~ 1 + condition +
                (1 | participant),
     data = df.original)
```



almost all variance explained

Model comparison

Hypothesis test

Is taking into account individual differences worth it?

```
1  # fit models (without and with dependence)
2  fit.compact = lm(formula = value ~ 1 + condition,
3                   data = df.original)
4
5  fit.augmented = lmer(formula = value ~ 1 + condition + (1 | participant),
6                       data = df.original)
7
8  # compare models
9  # note: the lmer model has to be supplied first
10 anova(fit.augmented, fit.compact)
```

```
refitting model(s) with ML (instead of REML)
Data: df.original
Models:
fit.compact: value ~ 1 + condition
fit.augmented: value ~ 1 + condition + (1 | participant)
              Df     AIC     BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
fit.compact    3 109.551 114.617 -51.775  103.551
fit.augmented  4  17.849  24.605  -4.925    9.849 93.701      1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Linear model

```
lm(formula = value ~ 1 + condition,
   data = df.original)
```

$$\text{value}_i = b_0 + b_1 \cdot \text{condition}_i + e_i$$

i = observation

$$e_i \sim \mathcal{N}(\text{mean} = 0,\ \text{sd} = s_{\text{error}})$$

3 parameters: b_0, b_1, s_{error}

Linear mixed effects model

```
lmer(formula = value ~ 1 + condition +
                (1 | participant),
     data = df.original)
```

$$\text{value}_{i,j} = b_0 + b_1 \cdot \text{condition}_{i,j} + U_i + e_i$$

i = participant,
j = time point

$$e_i \sim \mathcal{N}(\text{mean} = 0,\ \text{sd} = s_{\text{error}})$$

$$U_i \sim \mathcal{N}(\text{mean} = 0,\ \text{sd} = s_{\text{U}})$$

b_0, b_1 = fixed effects
U_i = random effect

here: random intercept

4 parameters: b_0, b_1, s_{error}, s_{U}

Model coefficients

Linear model

```
fit = lm(formula = value ~ 1 + condition,
         data = df.original)
coef(fit)
```

```
(Intercept)  condition2
  0.1905239   0.1993528
```

- one intercept
- one slope for condition

Linear mixed effects model

```
fit = lmer(formula = value ~ 1 + condition +
                     (1 | participant),
           data = df.original)
coef(fit)
```

```
$participant
   (Intercept) condition2
1  -0.57839428  0.1993528
2   0.22299824  0.1993528
3  -0.82920677  0.1993528
4   1.49310938  0.1993528
5   0.36042775  0.1993528
6  -0.82060123  0.1993528
7   0.47929171  0.1993528
8   0.66401020  0.1993528
9   0.55135879  0.1993528
10 -0.28306703  0.1993528
11  1.57681676  0.1993528
12  0.38457642  0.1993528
13 -0.59969682  0.1993528
14 -2.21148391  0.1993528
15  1.05439374  0.1993528
16 -0.06476643  0.1993528
17 -0.03505690  0.1993528
18  0.93945348  0.1993528
19  0.87495531  0.1993528
20  0.63135911  0.1993528

attr(,"class")
[1] "coef.mer"
```

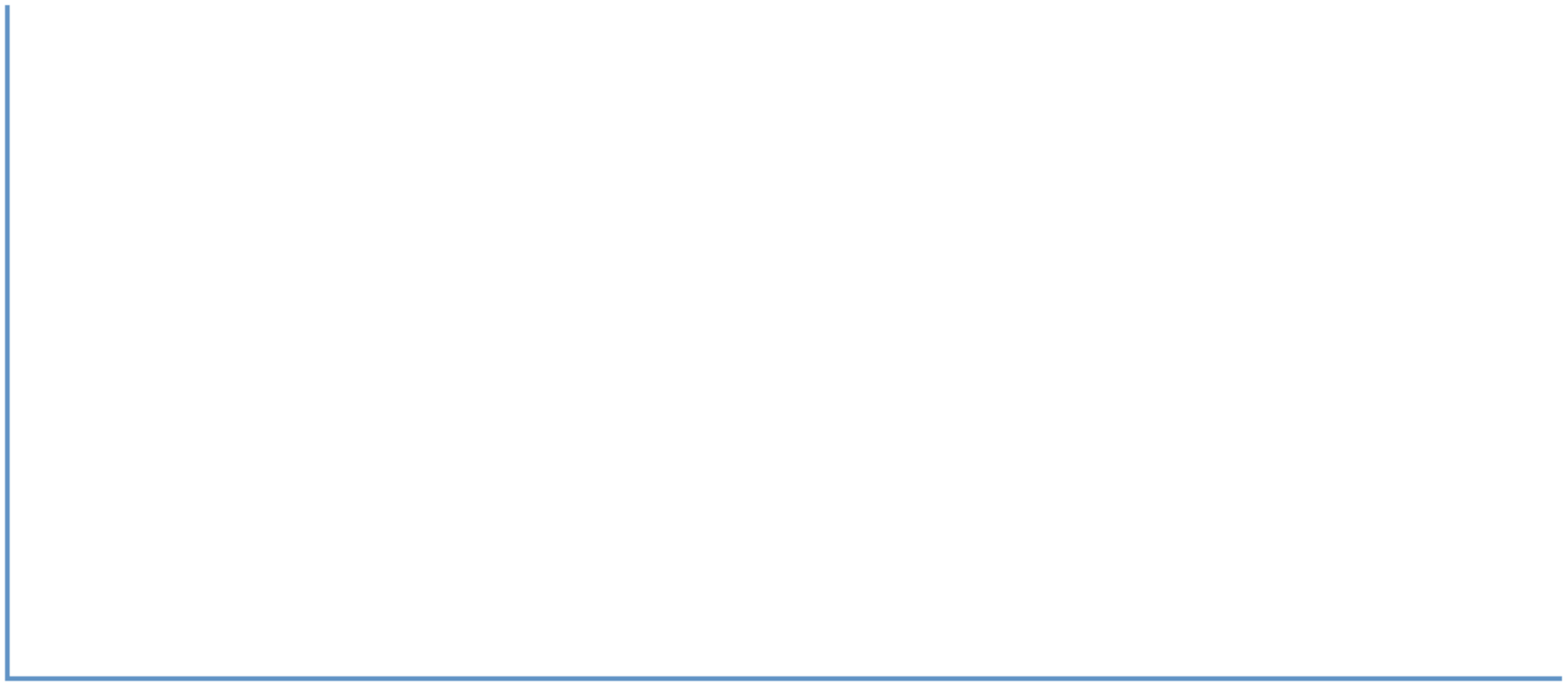
- different intercept for each participant
- one slope for condition

Plan for today

- Quick recap
- Controlling for variables
- Mediation
- Moderation
- Linear mixed effects model
  - Modeling dependence in data

Feedback

How was the pace of today's class?



much too slow

a little too slow

just right

a little too fast

much too fast

How happy were you with today's class overall?

What did you like about today's class? What could be improved next time?