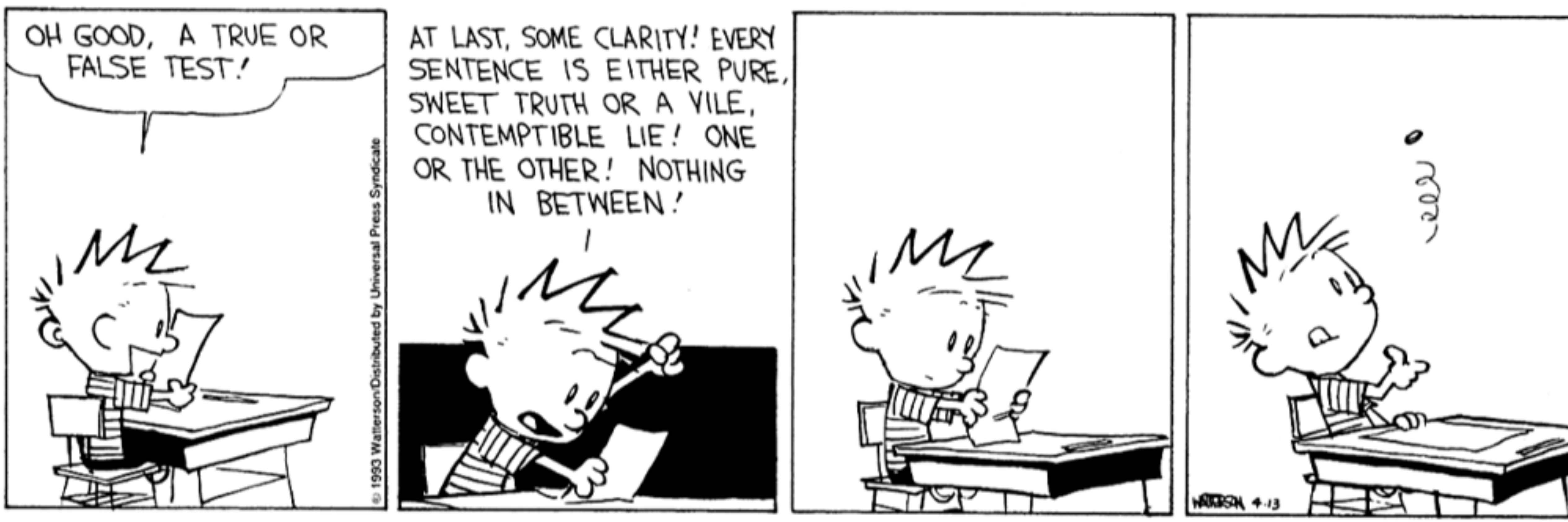
Model comparison



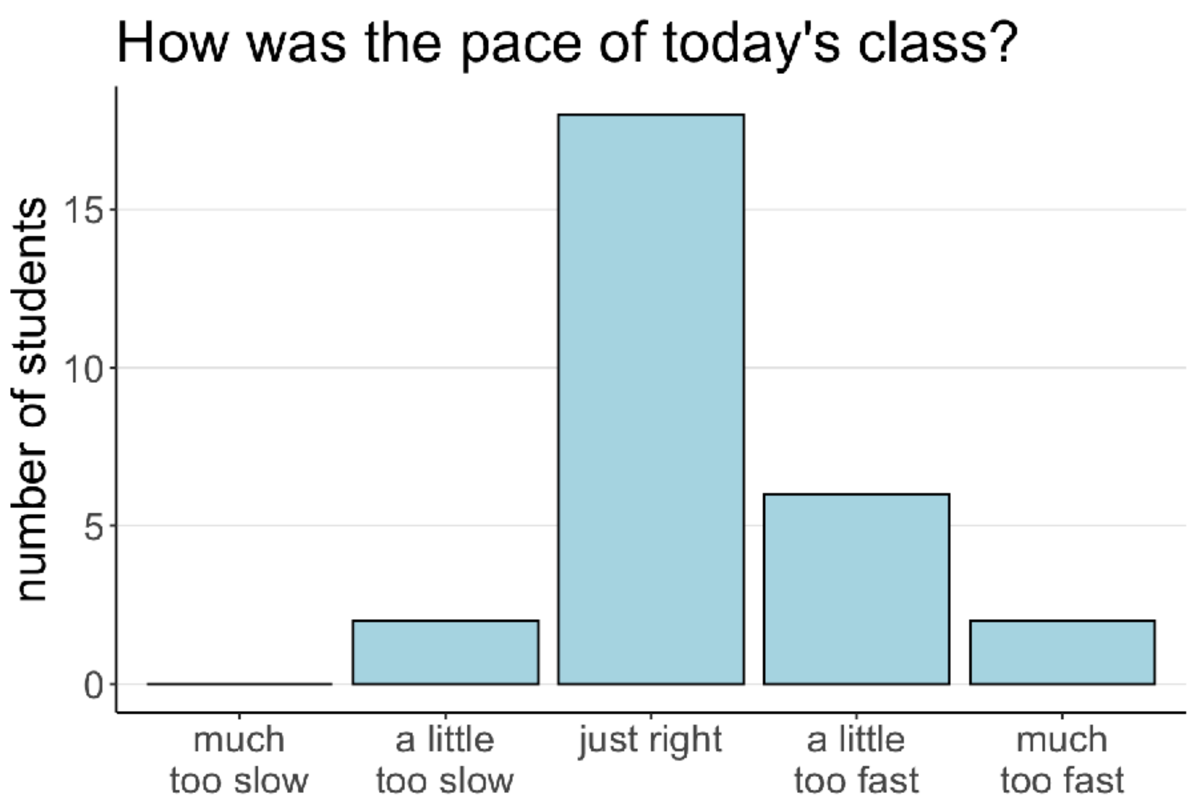
02/14/2020

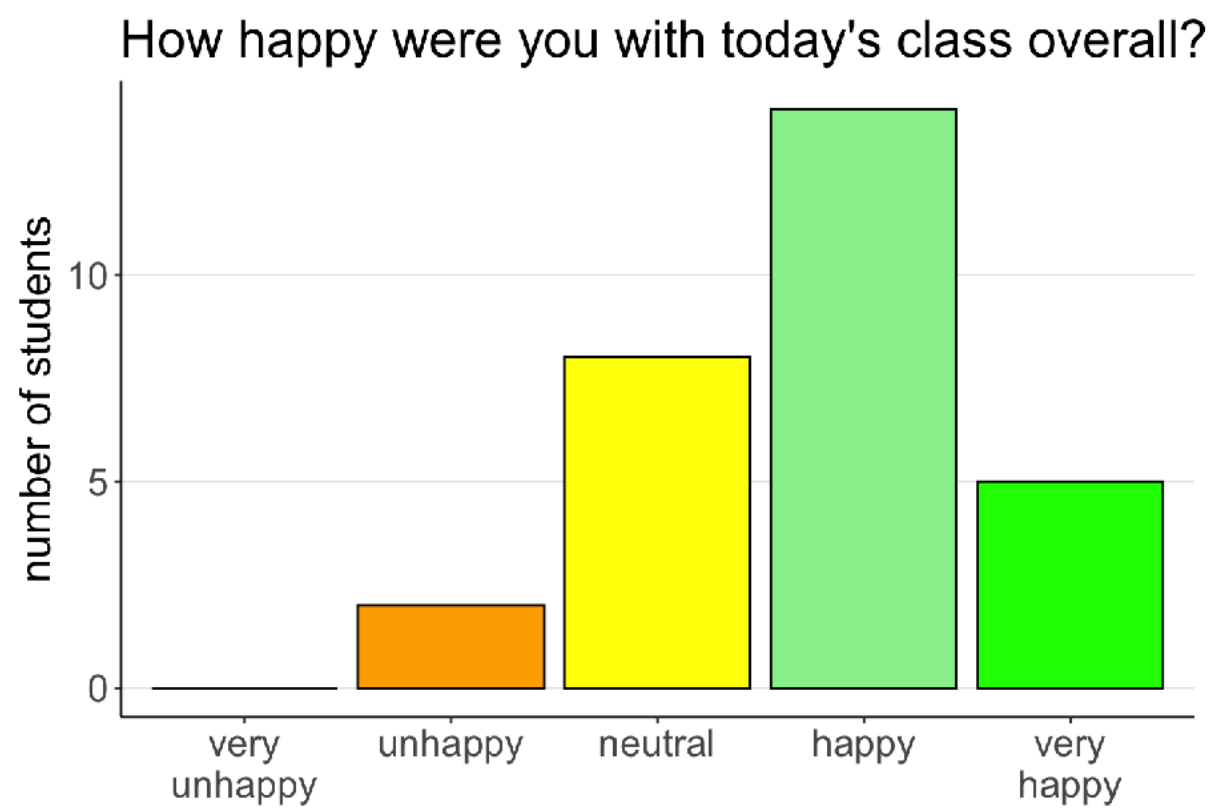
Feedback

Feedback

How many hours did you spend on homework 4?
number of students
0
2
4
6
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
or more

Feedback





class needs some work --
thanks for your feedback!

Feedback

Some signposts during the lecture would be appreciated (clearer indication that we are transitioning between topics and clearer articulation of the "why" behind concepts)

will try and do this better!

Feedback

The R is always confusing to learn in lecture style. I did like the visualization of the theoretical concepts. Still very very confused about "controlling" for variables (probably more confused then before the lecture, but I guess that's a good thing…)

will review controlling variables

Logistics

Final project

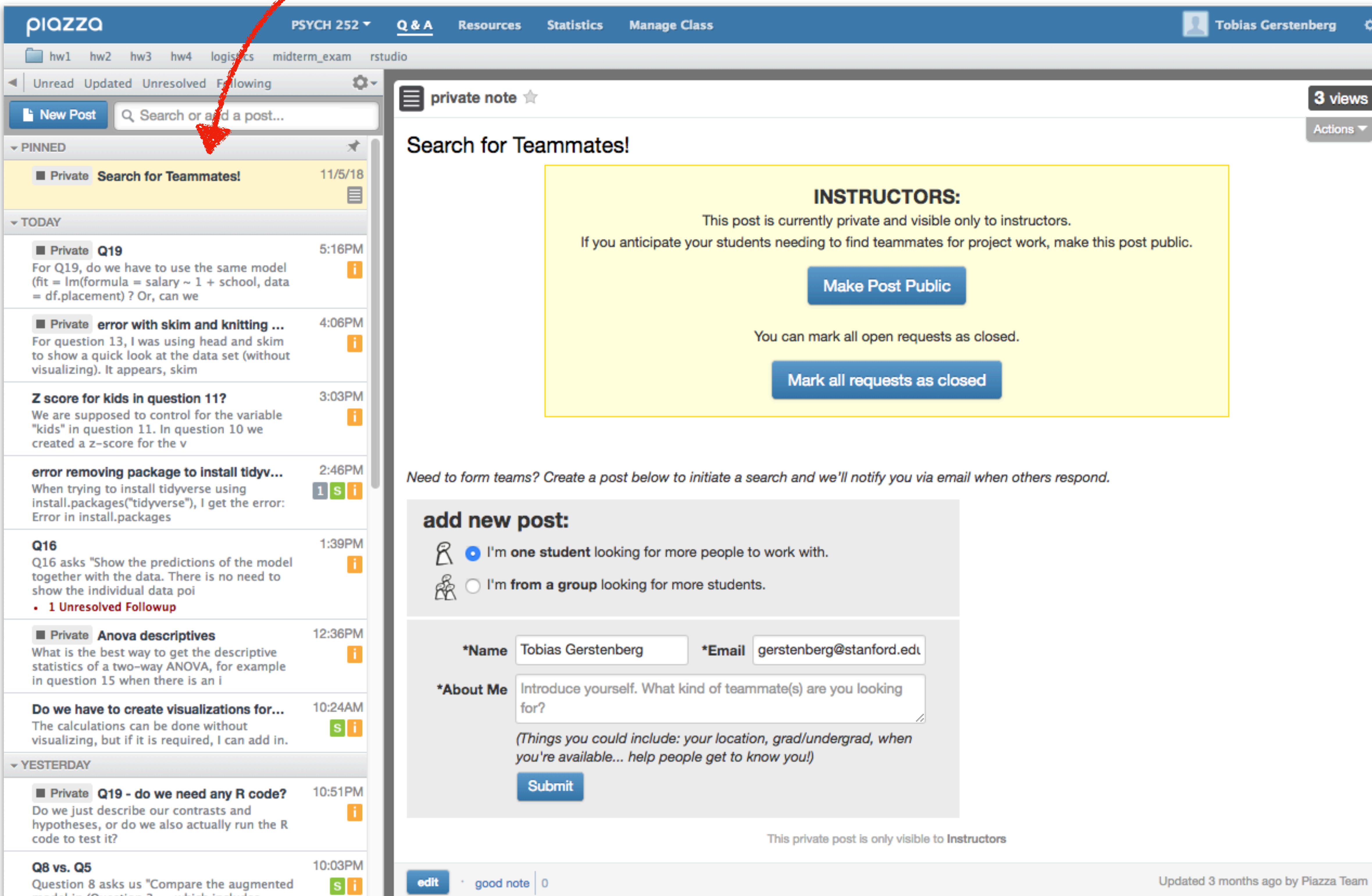
Project proposal

The project proposal is due on
Thursday, February 20th at 8pm

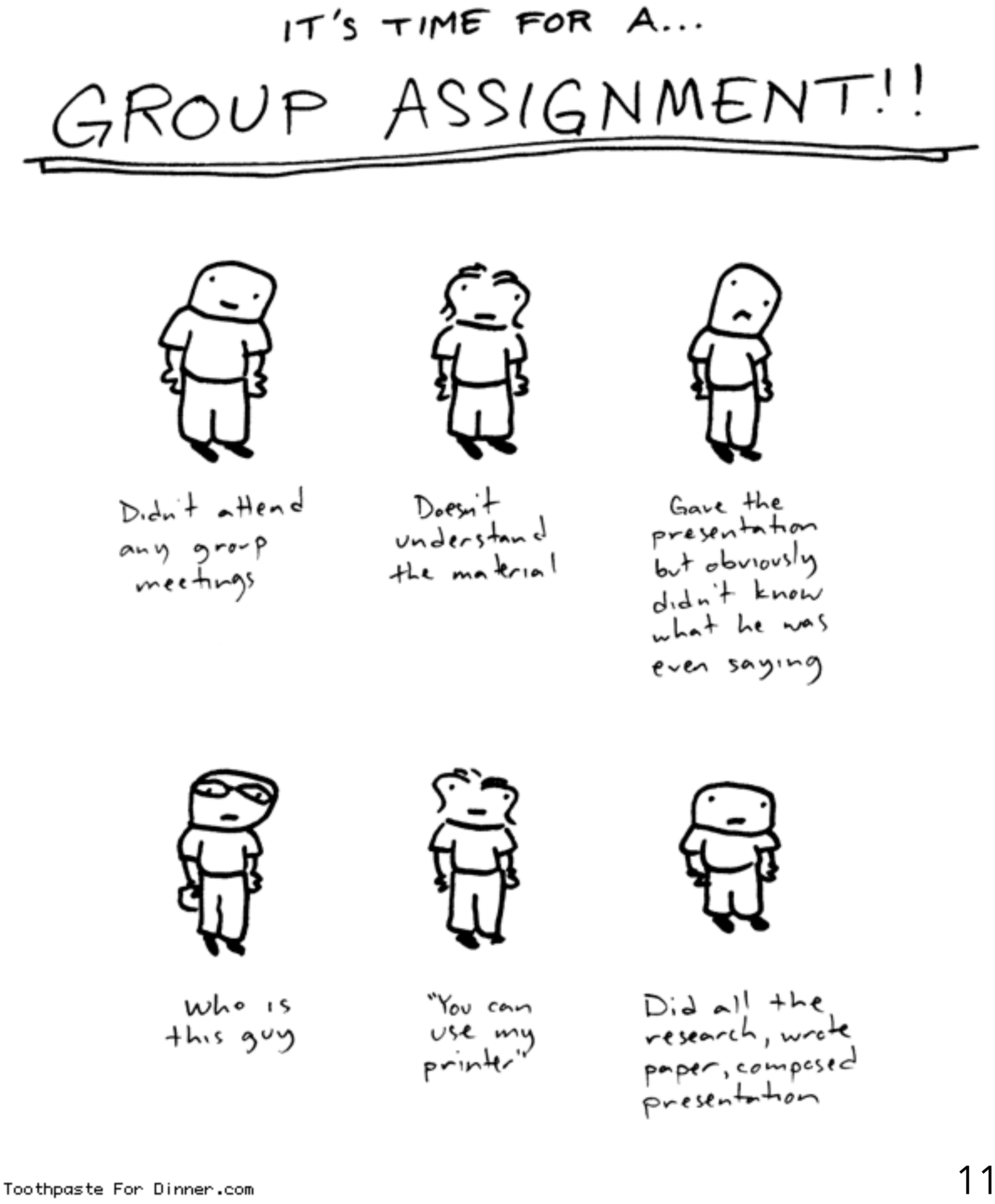
instructions of how to submit
will be released later today

Project proposal

find some teammates!



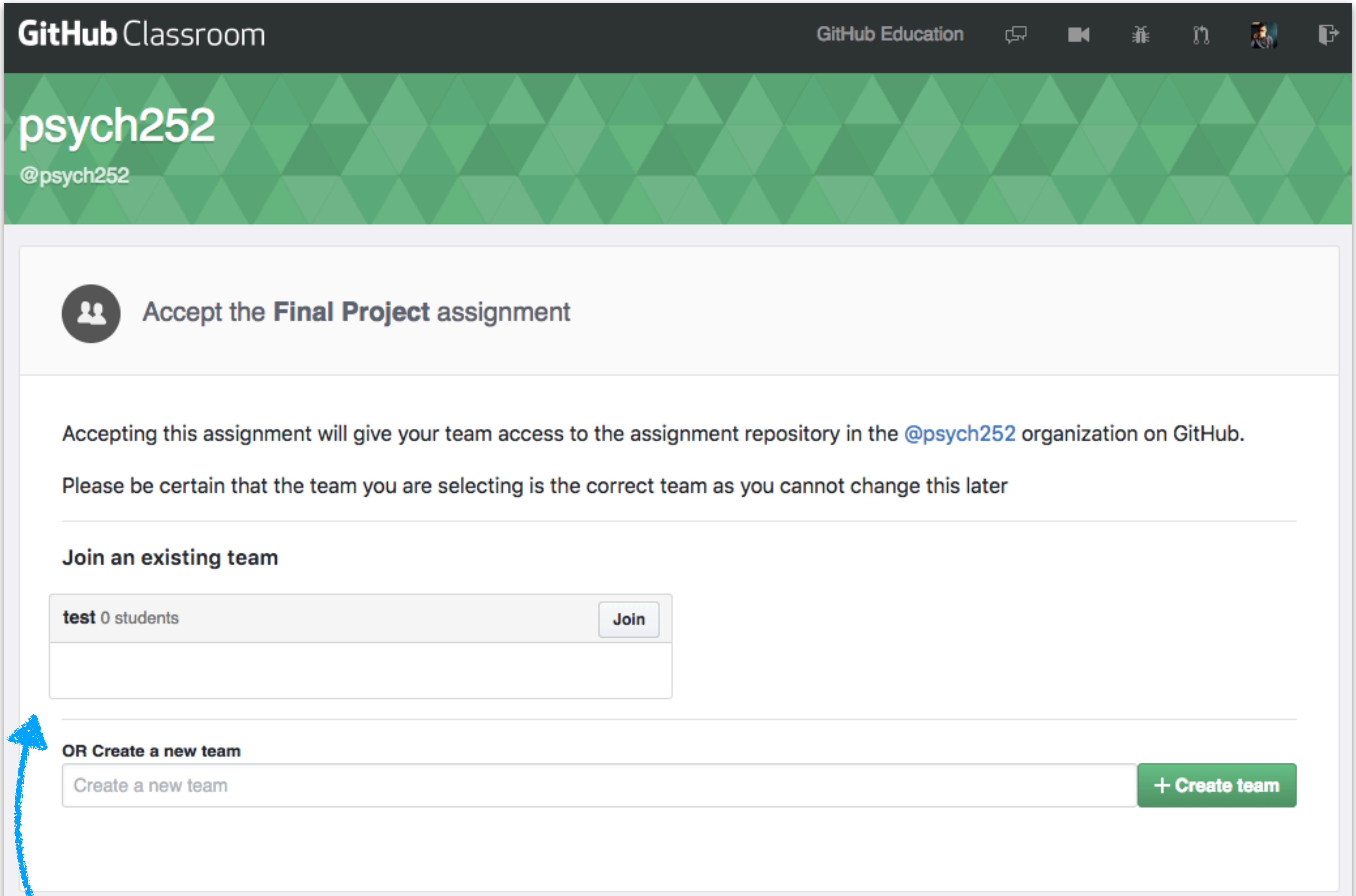
Project proposal



Members of teams will all the get same grade!

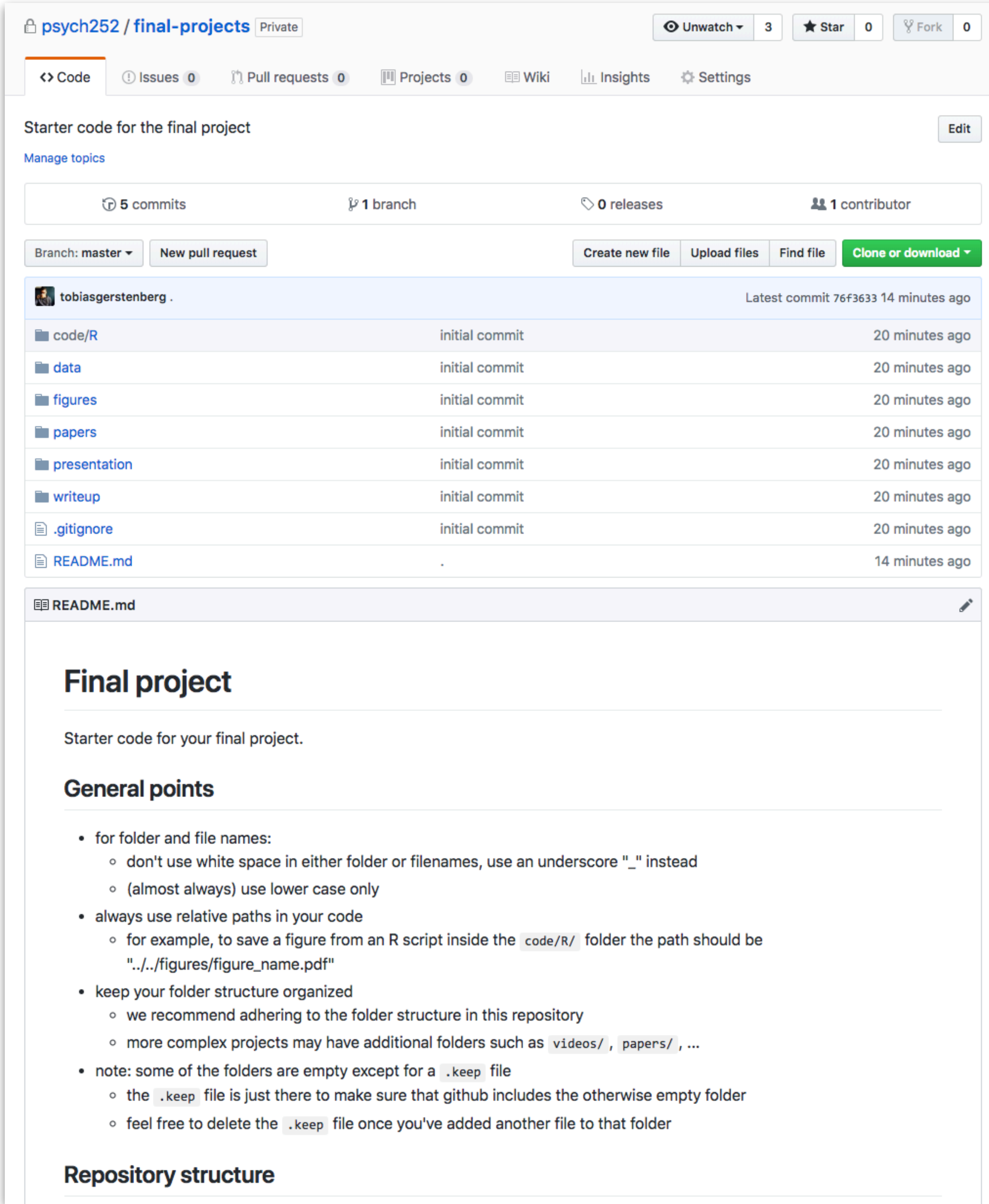
maximum 3 team members

Project proposal



join a team or make a new one

Project proposal



- each team will have their own private github repository
- all work on your final project should happen within this repository
- you can get **github** help in homework section
- post on Piazza in case you experience any problems getting set up

Project proposal

RMarkdown template

```
---
title: "A catchy project title goes here"
subtitle: "My team's name goes here"
author: "The team members' names go here"
date: "`r Sys.time()`"
urlcolor: blue # to show hyperlinks in blue when printed as pdf

# edit the output format below
# output: html_document # use this to render to html
output: pdf_document # use this to render to pdf

---

# Instructions

The project proposal is due on __Thursday, February 20th at 8pm__. It should not be longer
than __500 words__. It may contain code (code doesn't count toward the word limit).

## Research question

- What's your main research question?
- Which hypotheses are you trying to test?

## Methods

- Describe the data that you have. Are they based on an experiment, or will you work with
an existing data set?

## Analysis

- What analyses are you planning on using?

## Report

- What figures and/or tables are you planning on showing?
```

Upload the pdf to canvas

Midterm

How many hours did it take you to complete the midterm? (just write one number: e.g. 9 if it took you 9 hours)

Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**

"Late" late submission policy

we will subtract 2.5 points per hour

max points = 120

Plan for today

- Quick review: Controlling for variables
- Some more questions:
  - standardizing predictors
  - dummy coding vs. effect coding
- Mediation
- Moderation
- Model comparison
  - Cross-validation
  - AIC and BIC

Plan for today

- **Quick review: Controlling for variables**
- Some more questions:
  - standardizing predictors
  - dummy coding vs. effect coding
- Mediation
- Moderation
- Model comparison
  - Cross-validation
  - AIC and BIC

Quick review: controlling for variables

What does controlling for variables mean?

we are not actually "**controlling**" the variable

instead, we are taking the variable into consideration when making predictions

the hope is that we get a better estimate of the parameter that we are interested in by taking into account other factors

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

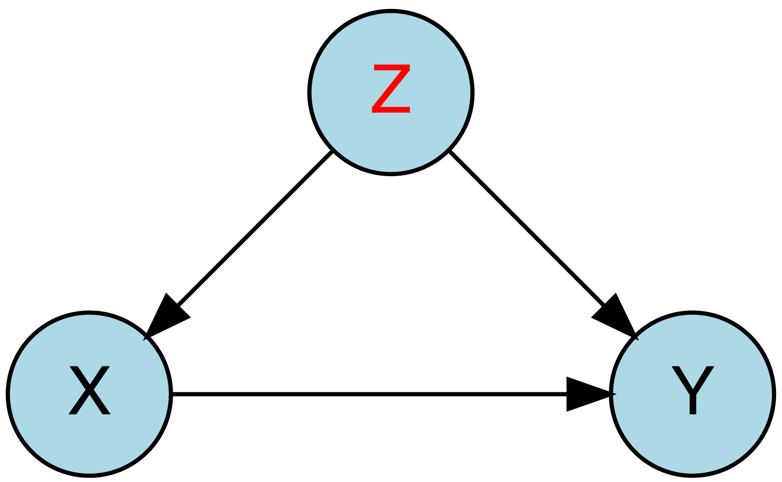
http://web.mit.edu/jmn/www/6.034/d-separation.pdf

When should I control for variables?

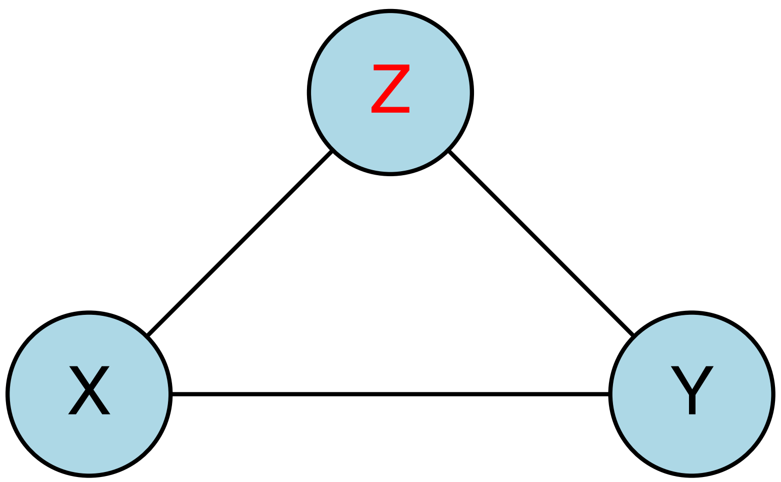
- checking for **d-separation** tells us whether or not variables are (conditionally) independent
- it also tells us whether paths of dependence "open up", or get "closed down"
- the graphical procedure doesn't necessarily reveal whether the dependence between variables changes: it reveals the **structure** 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

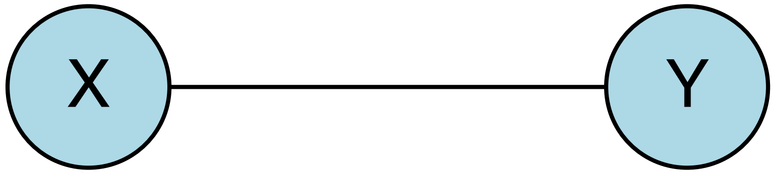


not controlling for Z



multiple paths

controlling for Z



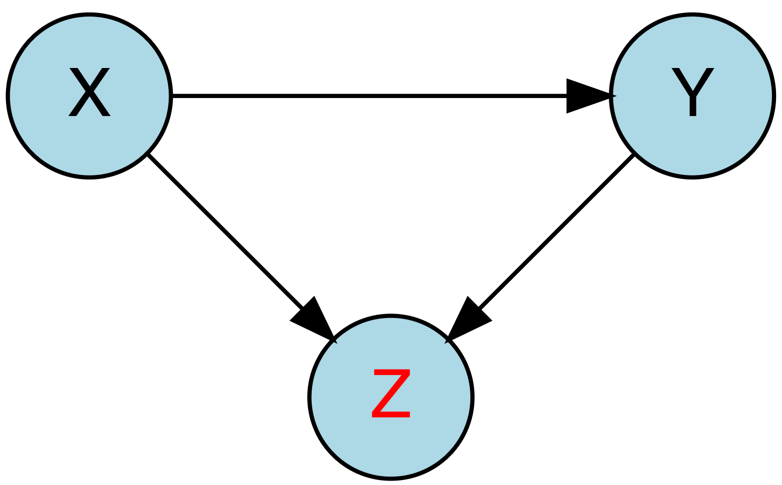
one path

additional paths from X to Y are removed

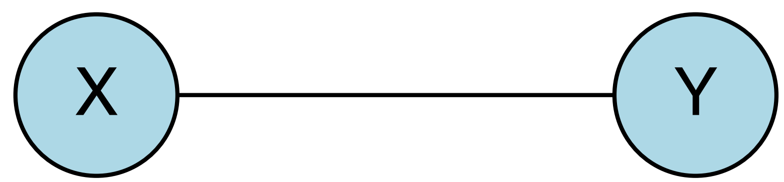
Z is a **good** control here!

When should I control for variables?

I want to estimate the effect that X has on Y

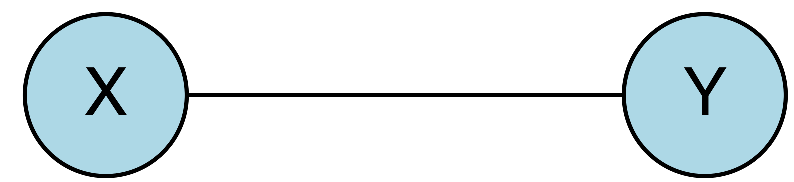


not controlling for Z



one path

controlling for Z



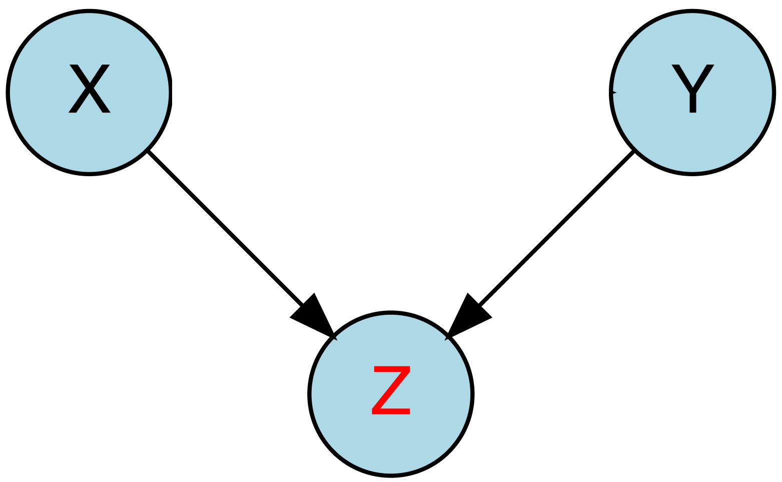
change in dependence

not revealed in the undirected graph

Z is a **bad** control here!

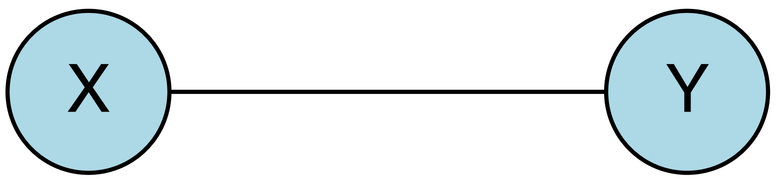
When should I control for variables?

I want to estimate the effect that X has on Y



not controlling for Z

controlling for Z



no path

one path

when controlling for Z, we might draw the wrong inference that X (directly) affects Y

Z is a **bad** control here!

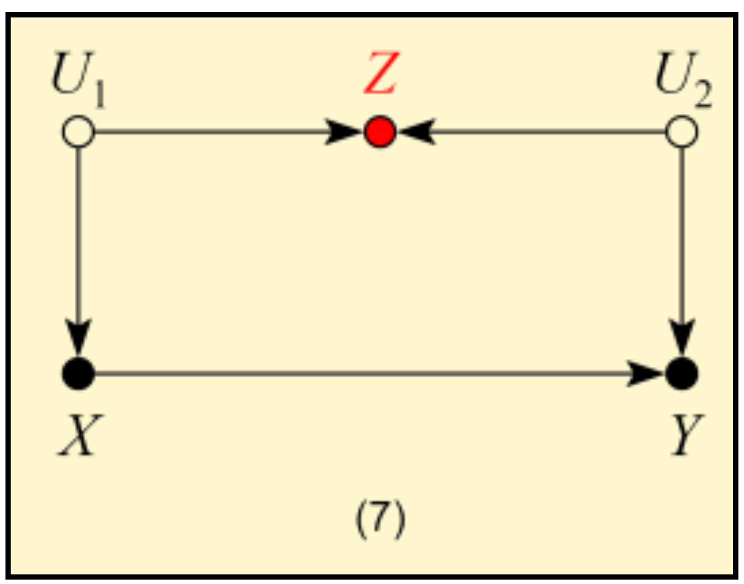
When should I control for variables?

- **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?

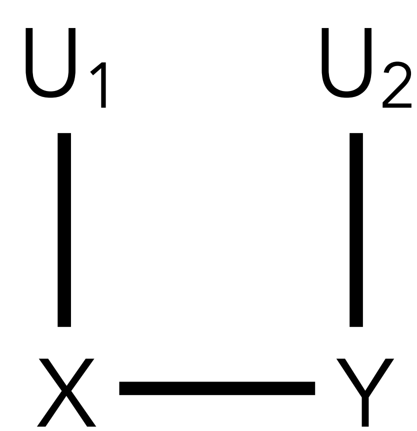
"Important topic. I feel like this procedure for determining good and bad controls is a bit convoluted. I feel like there's a simpler shorthand. Does another factor influence both x and y? Does another factor influence x? What's wrong with that simpler algorithm?"

not enough unfortunately

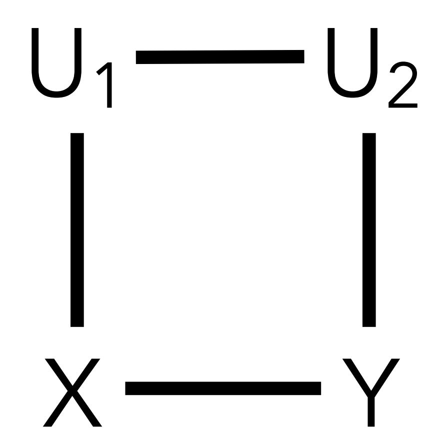


Z is a **bad** control here!

not controlling for Z



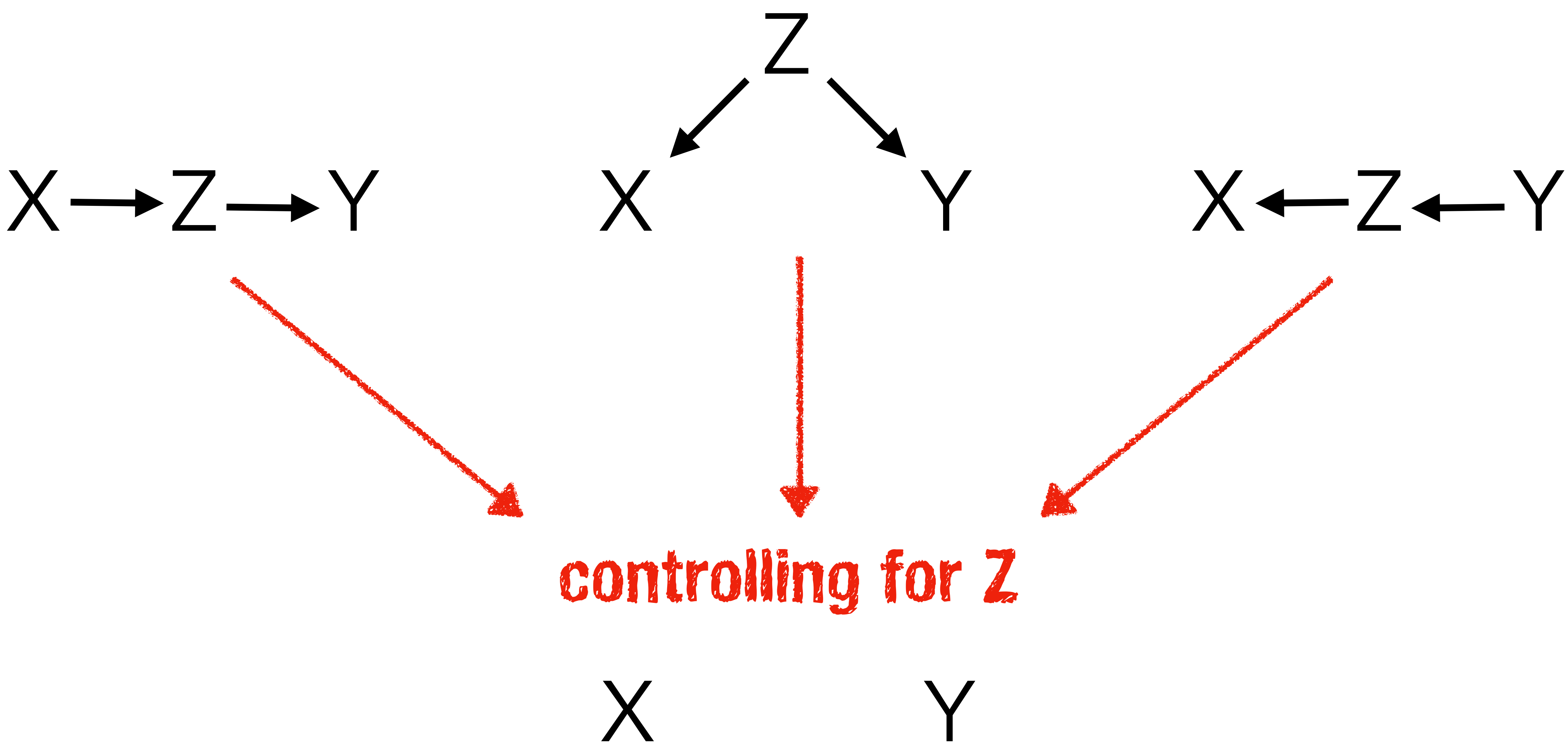
controlling for Z



additional path

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 ...

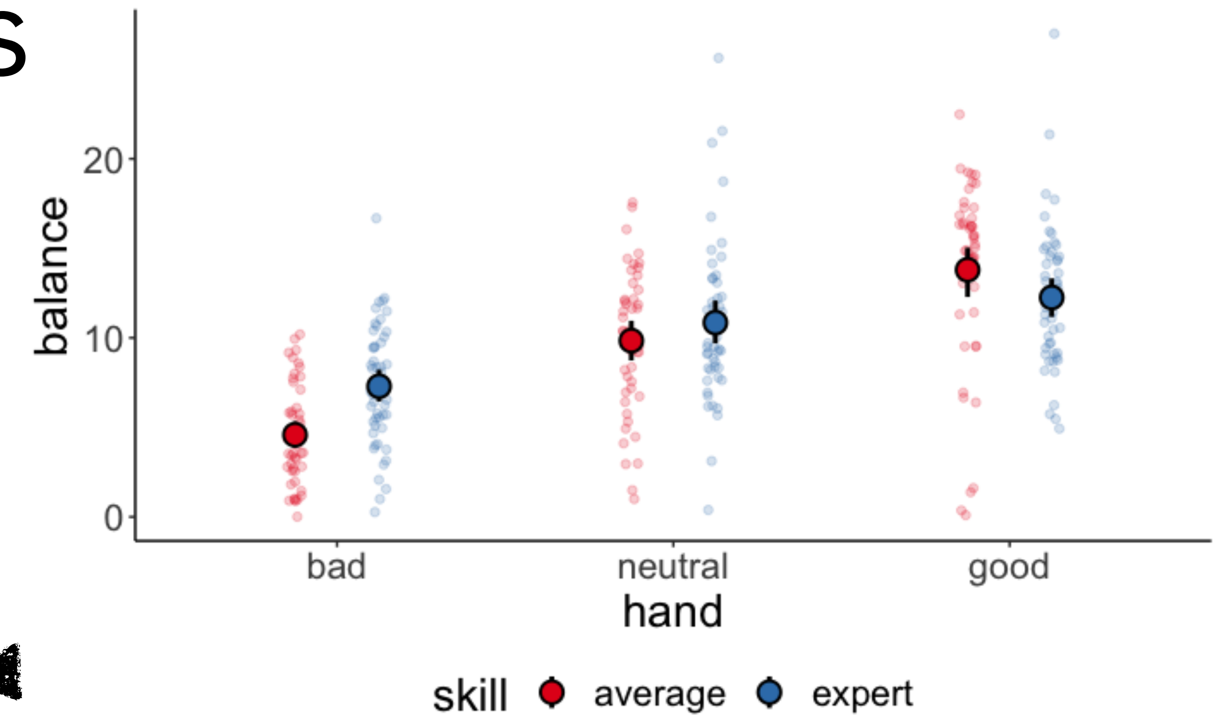
Peters & Buhlmann (2018) Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*

Plan for today

- Quick review: Controlling for variables
- **Some more questions:**
  - **standardizing predictors**
  - dummy coding vs. effect coding
- Mediation
- Moderation
- Model comparison
  - Cross-validation
  - AIC and BIC

When should I standardize my variables?

standardizing predictors



using continuous predictors

```
fit1 = lm(formula = balance ~ 1 + hand + skill,
    data = df.poker)
summary(fit1)
```

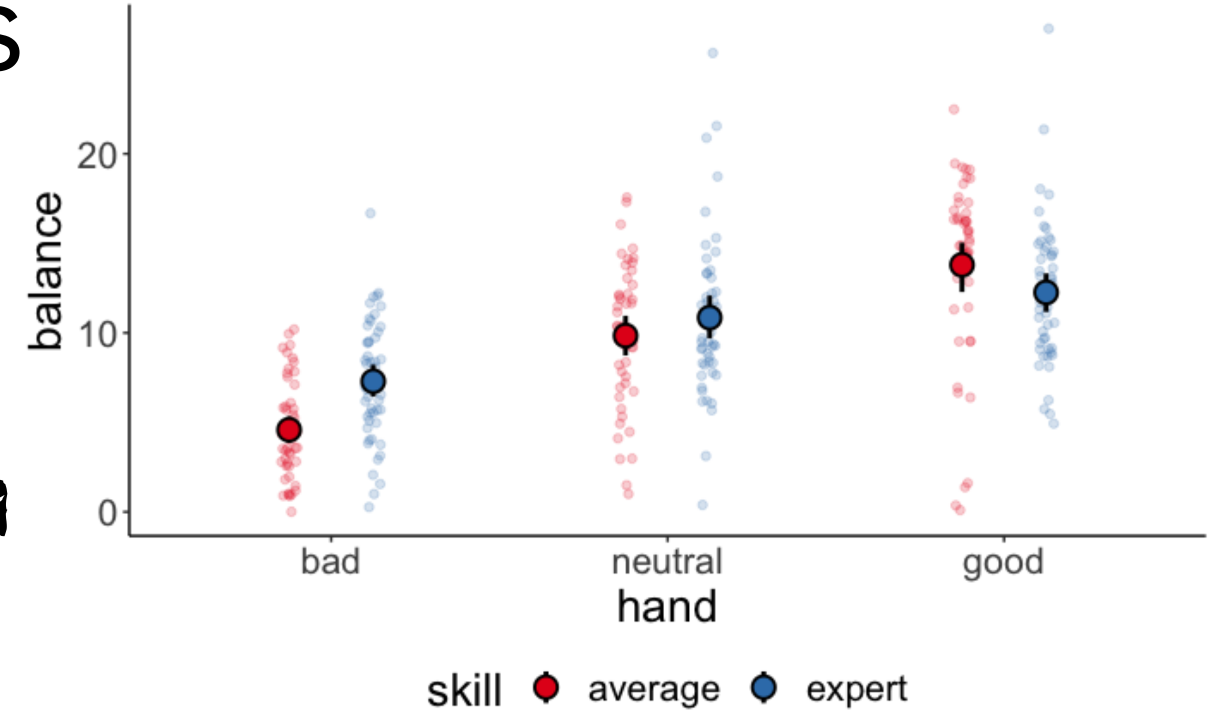
```
Call:
lm(formula = balance ~ 1 + hand + skill, data = df.poker)

Residuals:
     Min       1Q   Median       3Q      Max
-12.8518  -2.5874  -0.0441   2.7213  15.4963

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.7731     0.9505   3.970 9.03e-05 ***
hand          3.5424     0.2910  12.172  < 2e-16 ***
skill        -0.7243     0.4752  -1.524    0.129
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.116 on 297 degrees of freedom
Multiple R-squared:  0.3363,	Adjusted R-squared:  0.3318
F-statistic: 75.25 on 2 and 297 DF,  p-value: < 2.2e-16
```

hand	skill	hand_fct	skill_fct	balance
2	2	neutral	average	4.47
2	2	neutral	average	10.06
3	1	good	expert	9.15
1	2	bad	average	8.59
1	2	bad	average	2.81

⋮

hand is a significant predictor

standardizing predictors



with an interaction term

```r
fit2 = lm(formula = balance ~ 1 + hand * skill,
    data = df.poker)
summary(fit2)
```

```
Call:
lm(formula = balance ~ 1 + hand * skill, data = df.poker)

Residuals:
     Min       1Q   Median       3Q      Max
-13.9148  -2.3366   0.1438   2.4315  15.4963

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.1514     1.9462   5.216 3.44e-07 ***
hand          0.3533     0.9009   0.392 0.695221
skill        -4.9765     1.2309  -4.043 6.73e-05 ***
hand:skill    2.1261     0.5698   3.731 0.000228 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.029 on 296 degrees of freedom
Multiple R-squared:  0.3661,	Adjusted R-squared:  0.3597
F-statistic: 56.99 on 3 and 296 DF,  p-value: < 2.2e-16
```

hand	skill	hand_fct	skill_fct	balance
2	2	neutral	average	4.47
2	2	neutral	average	10.06
3	1	good	expert	9.15
1	2	bad	average	8.59
1	2	bad	average	2.81

⋮

hand is not a significant predictor anymore!

standardizing predictors

- interaction terms between continuous predictors introduce high correlations!

hand*skill

rowname	hand	skill	hand_x_skill	balance
hand				
skill	.00			
hand_x_skill	.75	.61		
balance	.58	-.07	.43	

nd	skill	hand_x_skill	balance
2	2	4	10.15
3	2	6	15.05
1	2	2	8.90
2	1	2	6.94

⋮

high correlations between interaction and each predictor

Title Text

hand	skill	hand_x_skill	balance
2	2	4	10.15
3	2	6	15.05
1	2	2	8.90
2	1	2	6.94

⋮

standardizing predictors

- centering (or standardizing) predictors removes the correlation that's otherwise introduced by the interaction

rowname	hand_c	skill_c	hand_x_skill_c	balance
hand_c				
skill_c	.00			
hand_x_skill_c	.00	.00		
balance	.58	-.07	.17	

interaction is not correlated with centered predictors

center (or standardize) the predictors when you fit a linear model with multiple continuous predictors including their interaction!

Plan for today

- Quick review: Controlling for variables
- **Some more questions:**
  - standardizing predictors
  - **dummy coding vs. effect coding**
- Mediation
- Moderation
- Model comparison
  - Cross-validation
  - AIC and BIC

dummy coding vs. effect coding

dummy coding vs. effect coding

```r
lm(formula = balance ~ hand, data = df.poker)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.9415     0.4111  14.451  < 2e-16 ***
handneutral   4.4051     0.5815   7.576 4.55e-13 ***
handgood      7.0849     0.5815  12.185  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

reference category

bad	neutral	good
5.94	10.35	13.03

```r
lm(formula = balance ~ hand, data = df.poker,
   contrasts = list(hand = "contr.sum"))
```

effect coding

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.7715     0.2374  41.165   <2e-16 ***
hand1        -3.8300     0.3357 -11.409   <2e-16 ***
hand2         0.5751     0.3357   1.713   0.0877 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

reference

grand mean = 9.77

what does hand1 and hand2 mean?

dummy coding vs. effect coding

dummy coding

	(Intercept)	handneutral	handgood
bad	1	0	0
neutral	1	1	0
good	1	0	1

- `intercept` is the mean for the reference category
- `handneutral` = difference between reference and neutral hand
- `handgood` = difference between reference and good hand

effect coding

	(Intercept)	hand1	hand2
bad	1	1	0
neutral	1	0	1
good	1	-1	-1

- `intercept` is the grand mean (because each linear contrast sums to 0)
- `hand1` = difference between bad hand and grand mean
- `hand2` = difference between neutral hand and grand mean

dummy coding vs. effect coding

dummy coding

	(Intercept)	handneutral	handgood
bad	1	0	0
neutral	1	1	0
good	1	0	1

hand	mean
bad	5.94
neutral	10.35
good	13.03

(Intercept)	handneutral	handgood
5.94	4.41	7.08

betas

$$\widehat{\text{balance}} \sim b_0 + b_1 \cdot \text{hand}_{\text{neutral}} + b_2 \cdot \text{hand}_{\text{good}}$$

hand = bad: $\widehat{\text{balance}} \sim b_0 + b_1 \cdot 0 + b_2 \cdot 0 = 5.94$

hand = neutral: $\widehat{\text{balance}} \sim b_0 + b_1 \cdot 1 + b_2 \cdot 0 = 5.94 + 4.41 = 10.35$

hand = good: $\widehat{\text{balance}} \sim b_0 + b_1 \cdot 0 + b_2 \cdot 1 = 5.94 + 7.08 = 13.03$

dummy coding vs. effect coding

effect coding

hand	mean
bad	5.94
neutral	10.35
good	13.03

	(Intercept)	hand1	hand2
bad	1	1	0
neutral	1	0	1
good	1	-1	-1

betas

(Intercept)	hand1	hand2
9.77	-3.83	0.58

$$\widehat{\text{balance}} \sim b_0 + b_1 \cdot \text{hand}_1 + b_2 \cdot \text{hand}_2$$

hand = bad: $\widehat{\text{balance}} \sim b_0 + b_1 \cdot 1 + b_2 \cdot 0 = 9.77 - 3.83 = 5.94$

hand = neutral: $\widehat{\text{balance}} \sim b_0 + b_1 \cdot 0 + b_2 \cdot 1 = 9.77 + 0.58 = 10.35$

hand = good: $\widehat{\text{balance}} \sim b_0 + b_1 \cdot (-1) + b_2 \cdot (-1) = 9.77 + 3.83 - 0.58 = 13.03$

dummy coding vs. effect coding

```r
lm(formula = balance ~ hand, data = df.poker)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.9415     0.4111  14.451  < 2e-16 ***
handneutral   4.4051     0.5815   7.576 4.55e-13 ***
handgood      7.0849     0.5815  12.185  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

is the difference between a neutral and bad hand significantly different from 0?

```r
lm(formula = balance ~ hand, data = df.poker,
   contrasts = list(hand = "contr.sum"))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   9.7715     0.2374  41.165   <2e-16 ***
hand1        -3.8300     0.3357 -11.409   <2e-16 ***
hand2         0.5751     0.3357   1.713   0.0877 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

is the difference between a bad hand and the grand mean significantly different from 0?

dummy coding vs. effect coding

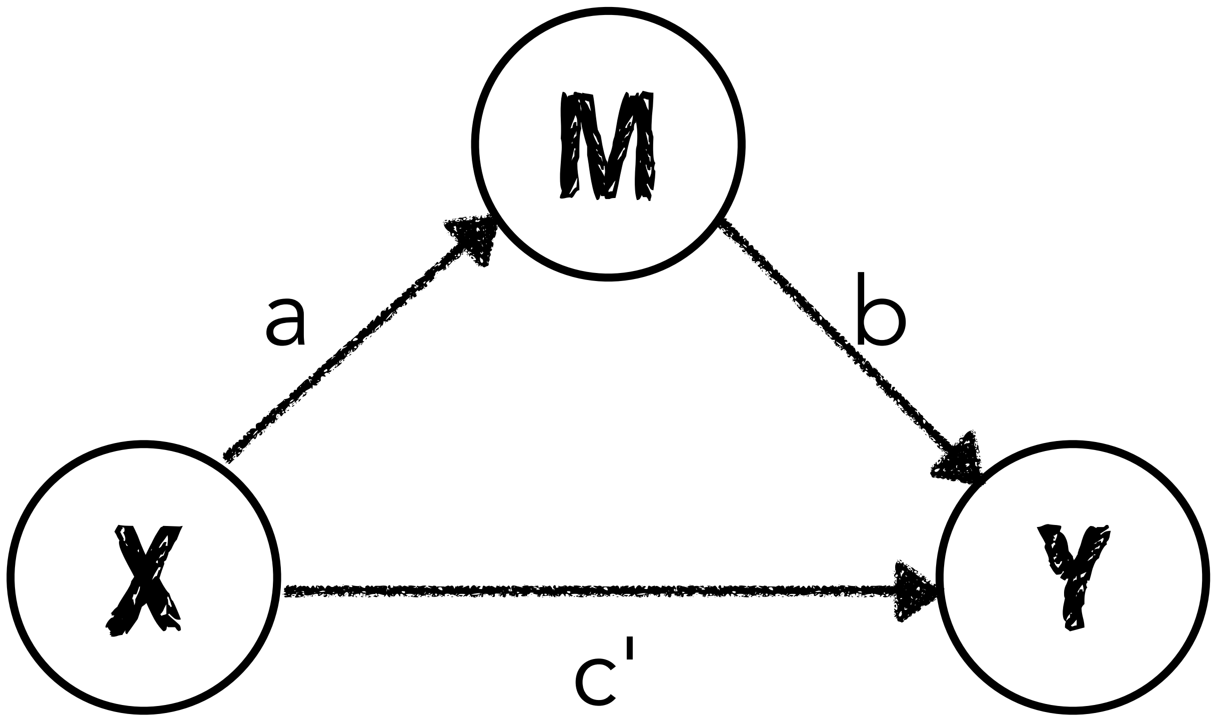
- **in practice**:
  - multiple categorical predictors (with interactions): **effect-coding**
  - one categorical and one (or several) continuous predictors: **dummy-coding**
  - several categorical and continuous predictors with interactions: **things get complicated ...**
  - parameters mean: how is the outcome variable predicted to change if
- **in practice in `R`:**
  - fit the linear model first in R using `lm()`
  - if interested in main effects, use the `anova()` function
  - if interested in specific comparisons, use the `emmeans` package and define the comparisons as linear contrasts

Plan for today

- Quick review: Controlling for variables
- Some more questions:
    - standardizing predictors
    - dummy coding vs. effect coding
- **Mediation**
- Moderation
- Model comparison
    - Cross-validation
    - AIC and BIC

Mediation

Definition

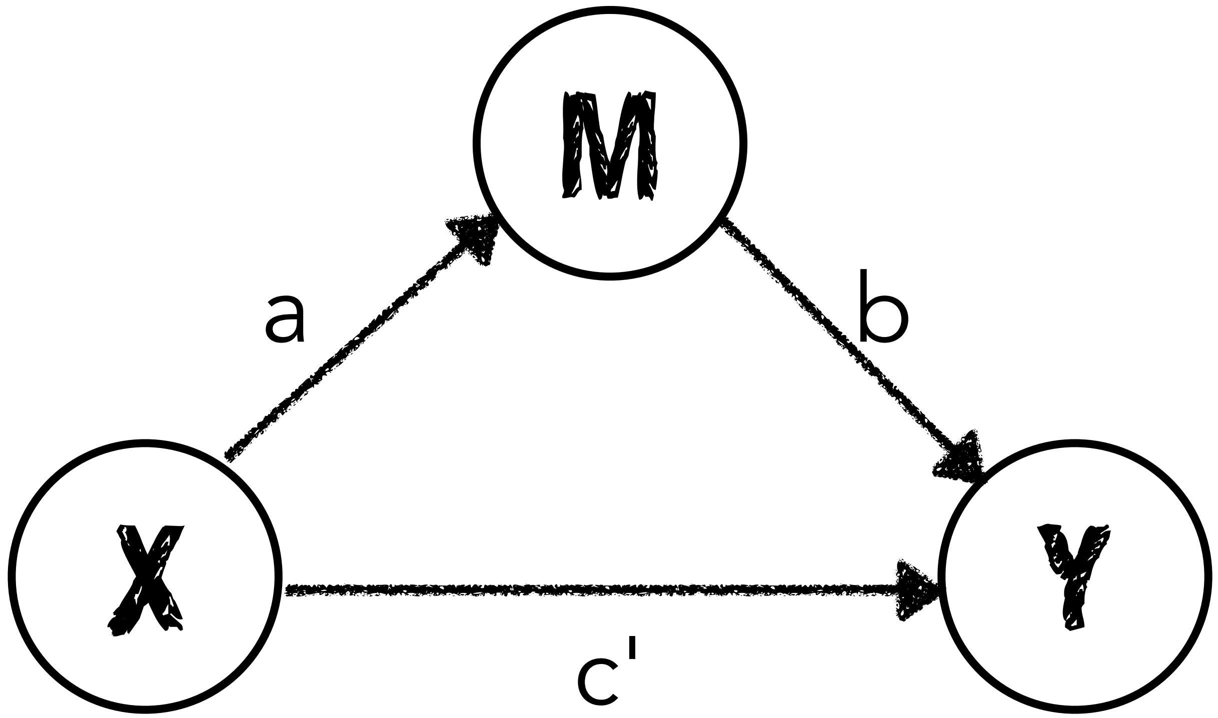


Rather than a direct causal relationship between **X** and **Y**, a mediation model proposes that **X** influences the mediator variable **M**, which in turn influences **Y**. Thus, the mediator variable serves to clarify the nature of the relationship between **X** and **Y**.

Adapted from Wikipedia

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

Example



X = grades in Psych 252

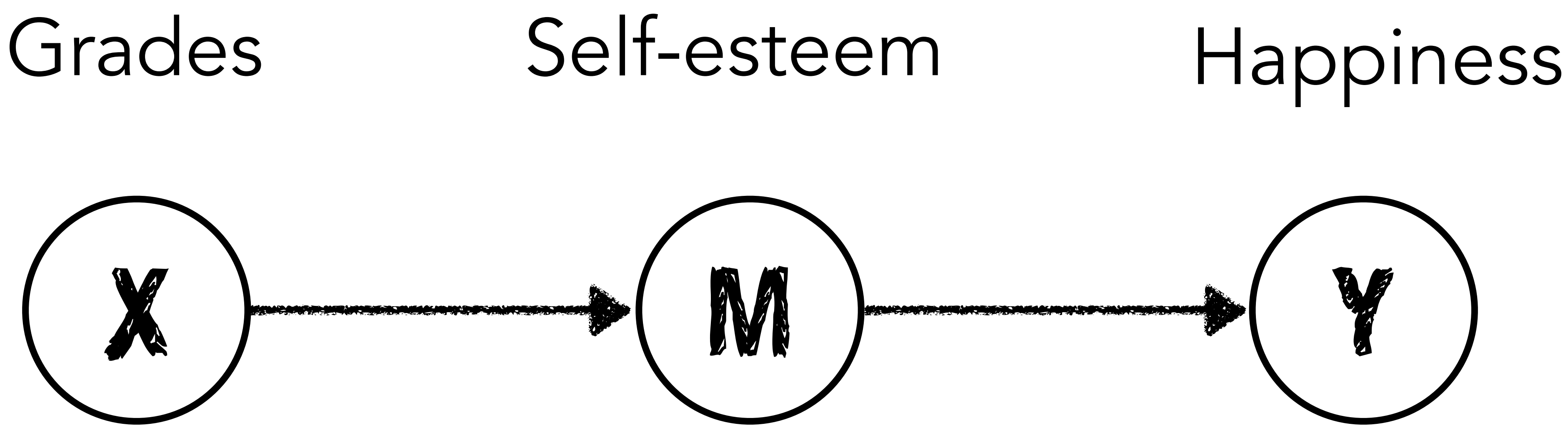
M = feelings of self-esteem

Y = happiness

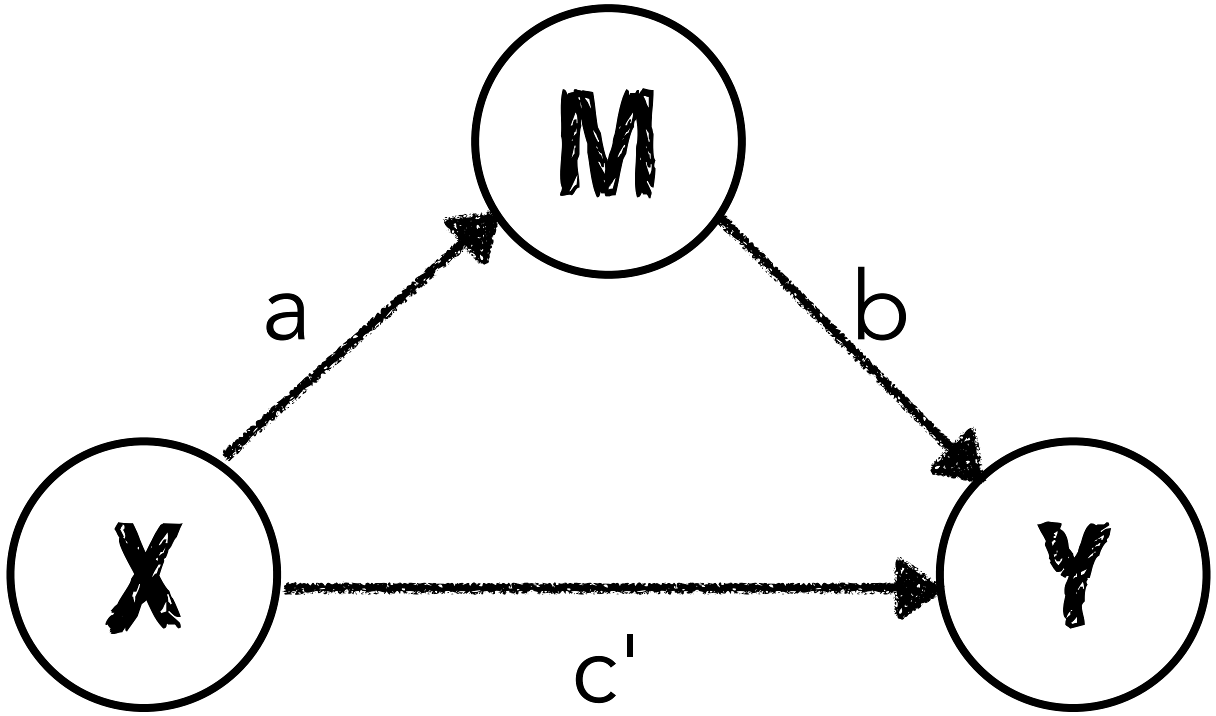
Is the relationship between grades in Psych 252 and happiness mediated by feelings of self-esteem?

Simulate a mediation analysis

```r
# 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
)
```



Baron and Kenny's (1986) steps for mediation



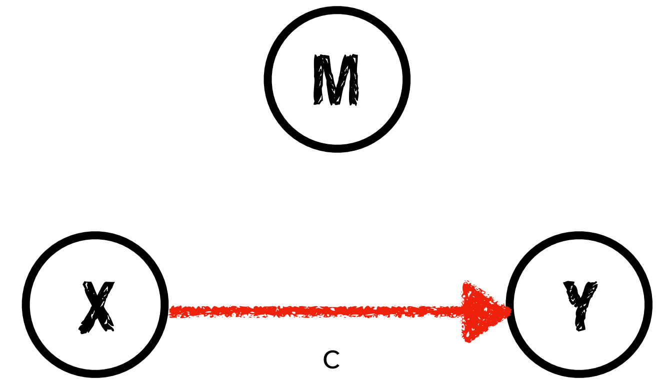
Sequence of regression models

1. Is there a relationship between **X** and **Y**?
2. Is there a relationship between **X** and **M**?
3. Does the relationship between **X** and **Y** change, once we control for **M**?

Baron, R. M. & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173-1182.

Is there a relationship between **X** and **Y**?

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



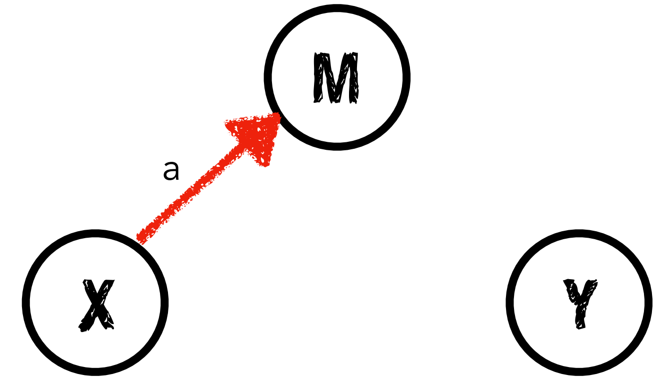

```r
# fit the model
fit.y_x = lm(formula = y ~ 1 + x,
             data = df.mediation)

# summarize the results
fit.y_x %>% summary()
```

```
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 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.16 on 98 degrees of freedom
Multiple R-squared:  0.04237,	Adjusted R-squared:  0.0326 
F-statistic: 4.336 on 1 and 98 DF,  p-value: 0.03993
```

significant relationship

Is there a relationship between **X** and **M**?

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



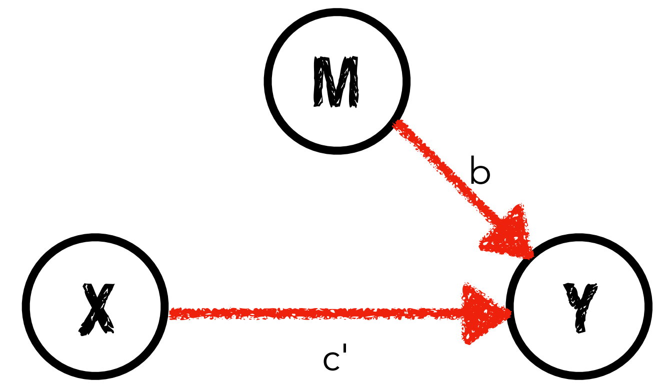

```r
# fit the model
fit.m_x = lm(formula = m ~ 1 + x,
             data = df.mediation)

# summarize the results
fit.m_x %>% summary()
```

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

Residuals:
    Min      1Q  Median      3Q     Max 
-9.5367 -3.4175 -0.4375  2.9032 16.4520 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.04494   13.41692   0.451    0.653    
x            0.66252    0.07634   8.678 8.87e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.854 on 98 degrees of freedom
Multiple R-squared:  0.4346,	Adjusted R-squared:  0.4288 
F-statistic: 75.31 on 1 and 98 DF,  p-value: 8.872e-14
```

significant relationship

Is there a relationship between **X** and **Y,** controlling for **M**?

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

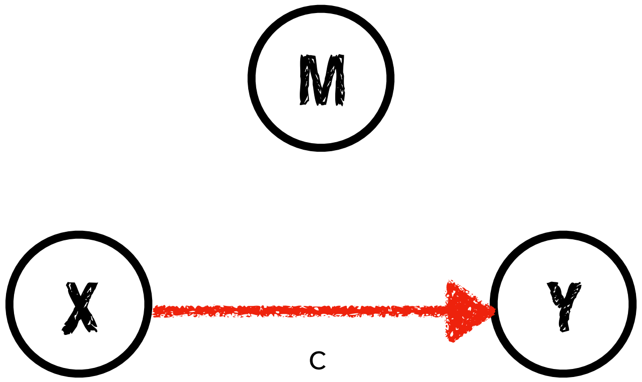



```r
# fit the model
fit.y_mx = lm(formula = m ~ 1 + m + x,
              data = df.mediation)

# summarize the results
fit.y_mx %>% summary()
```

```
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
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.756 on 97 degrees of freedom
Multiple R-squared:  0.1946,	Adjusted R-squared:  0.1779
F-statistic: 11.72 on 2 and 97 DF,  p-value: 2.771e-05
```

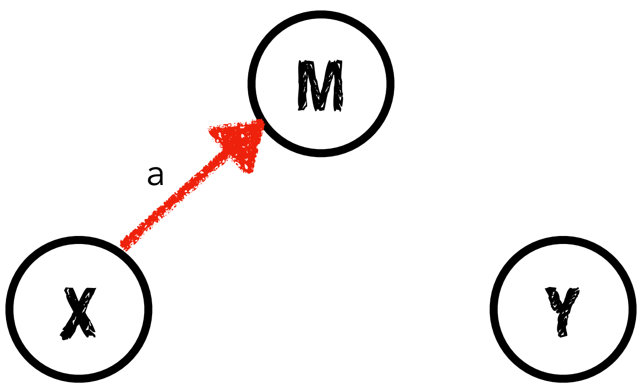
not significant

3 Step procedure



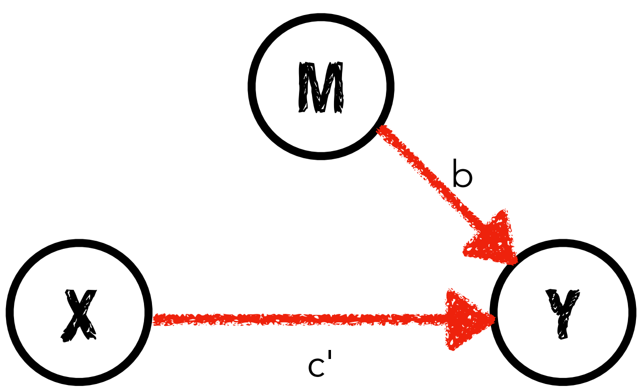
Relationship between **X** and **Y**?

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



Relationship between **X** and **M**?

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



Relationship between **X** and **Y**, controlling for **M**?

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

significant change?

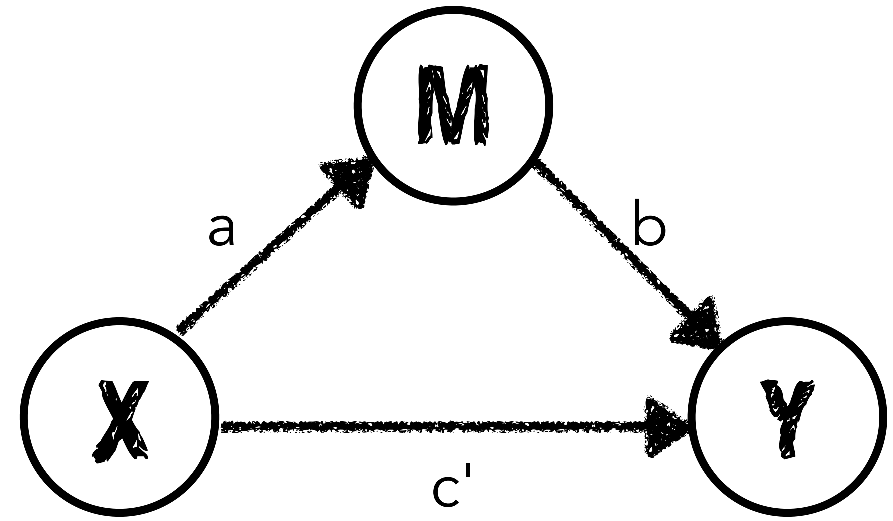
just because it changes from significant to not significant, does not mean the change was significant!

Is the mediation significant?

1. **Sobel test**
   - assumes normally distributed data
   - has low power
2. **Bootstrapping**
   - no assumption about how the data is distributed
   - has higher power
3. **Bayesian mediation analysis**

later in class!

1. Sobel test

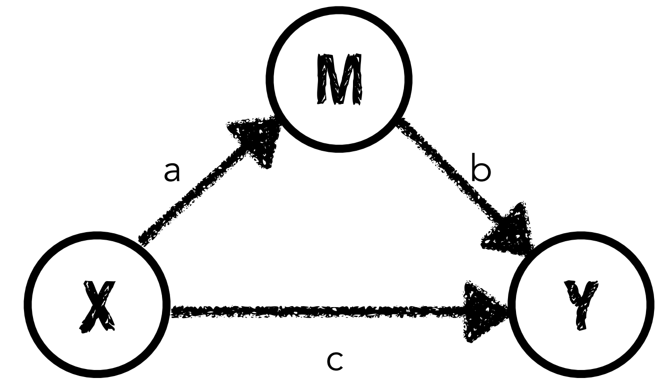



```r
library("multilevel")

# run the sobel test
fit.sobel = sobel(pred = df.mediation$x,
                  med = df.mediation$m,
                  out = df.mediation$y)

# calculate the p-value
(1 - pnorm(fit.sobel$z.value))*2
```

"It is becoming increasingly more difficult to publish tests of mediation based purely on the Baron and Kenny method or tests that make distributional assumptions such as the Sobel test."

A Wikipedia person

2. Bootstrapping



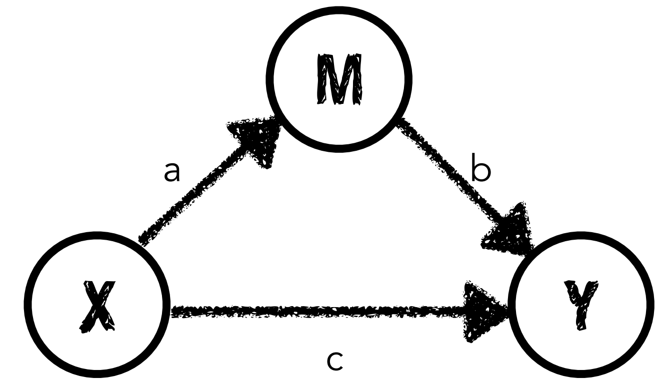

```r
library("mediation")

# bootstrapped mediation
fit.mediation = mediate(model.m = fit.m_x,
                        model.y = fit.y_mx,
                        treat = "x",
                        mediator = "m",
                        boot = T)

# summarize results
fit.mediation %>% summary()
```

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

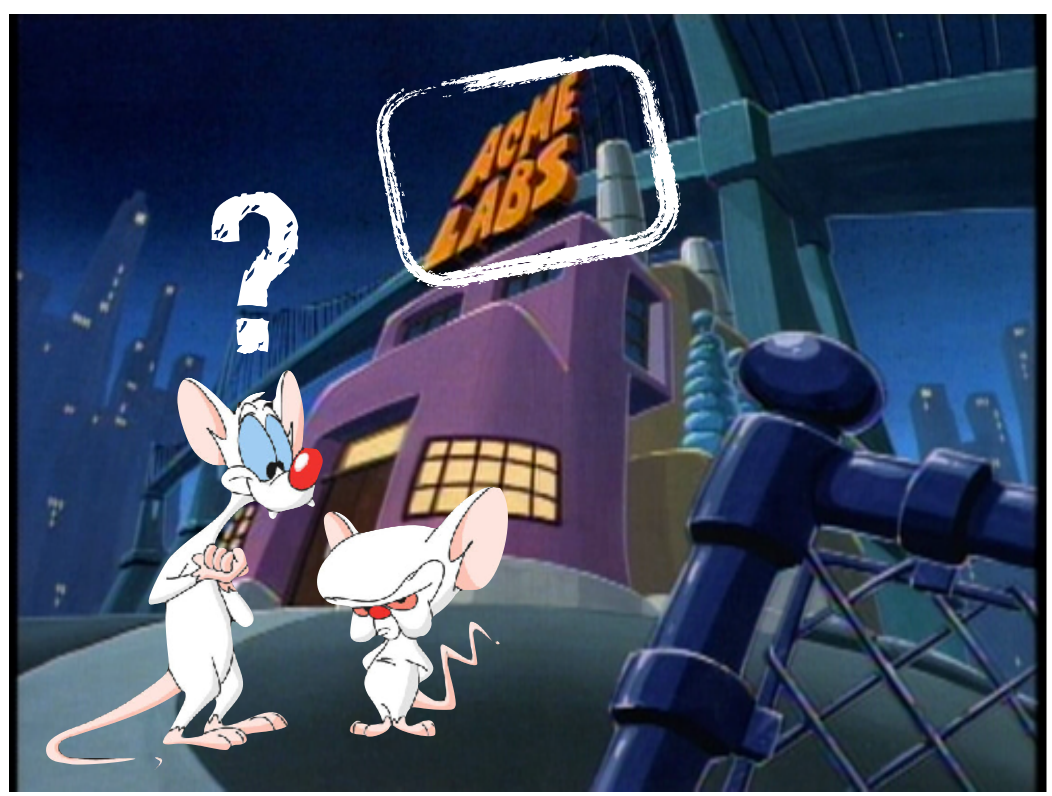
$\hat{y} = b_0 + b_1 \cdot m + b_2 \cdot x$

```
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
```

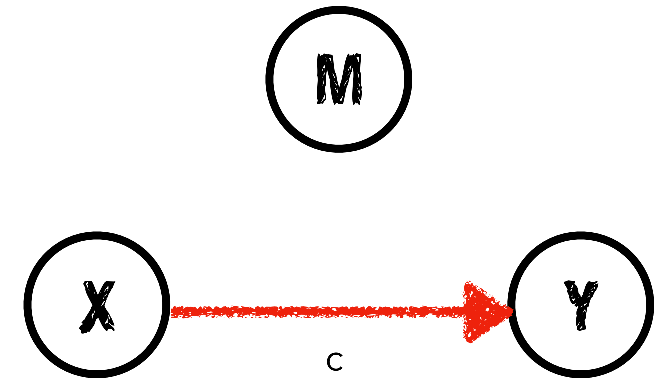
2. Bootstrapping




```
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
```



2. Bootstrapping

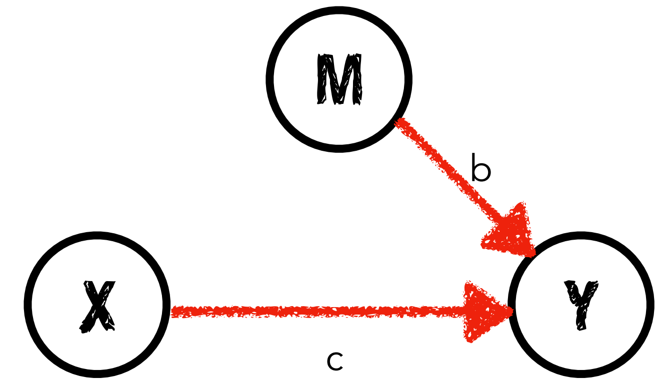



```
Causal Mediation Analysis

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 *
```

2. 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
```

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

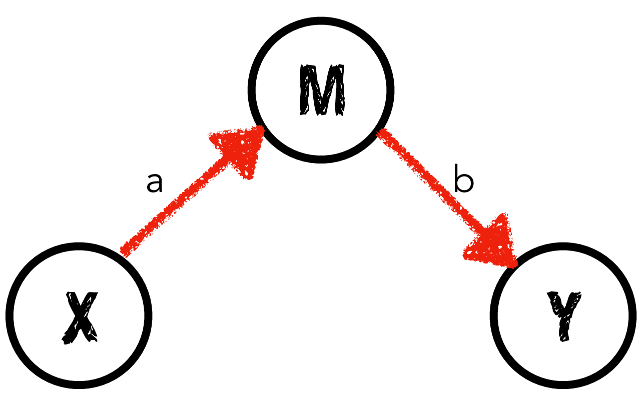
ADE: Average direct effect

```
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
```

2. 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



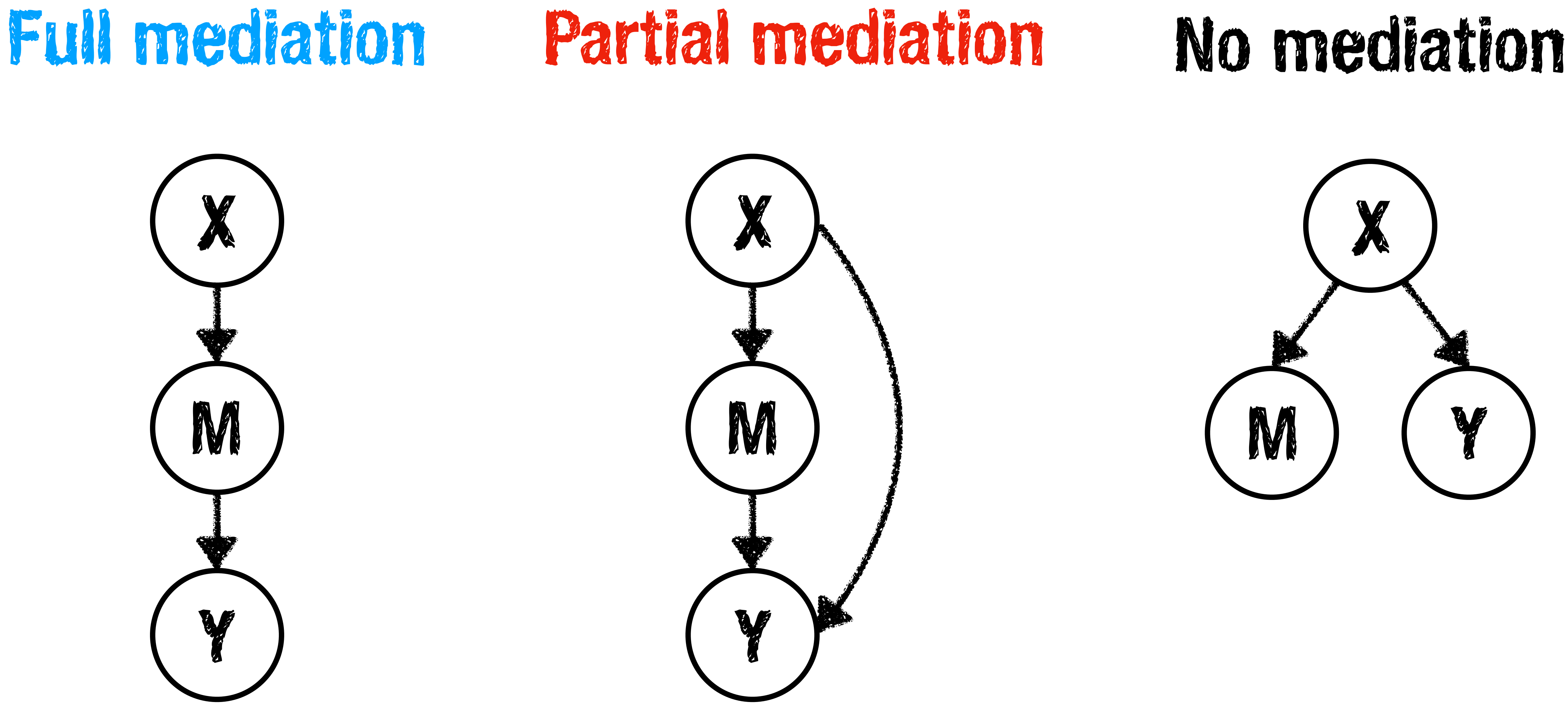
indirect effect: a * b

ADE: Average direct effect

ACME: Average causal mediation effect

ACME = Total effect - ADE

Underlying causal model



Full mediation: 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**.

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

ELSEVIER

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

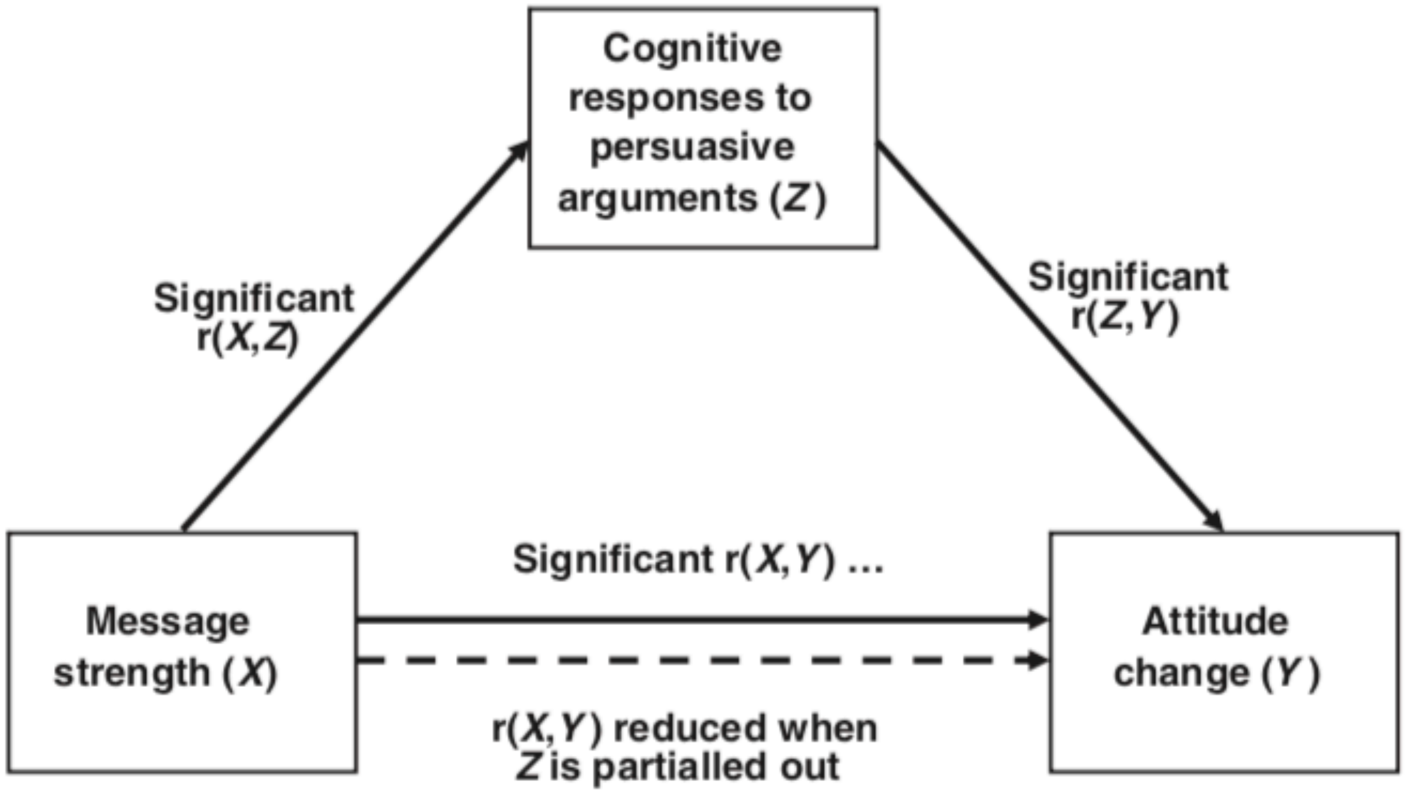


Fig. 1. Illustration of a mediation analysis in the context of persuasion research: The impact of message strength (X) on attitude change (Y) is supposed to be mediated by the recipients' cognitive responses (Z).

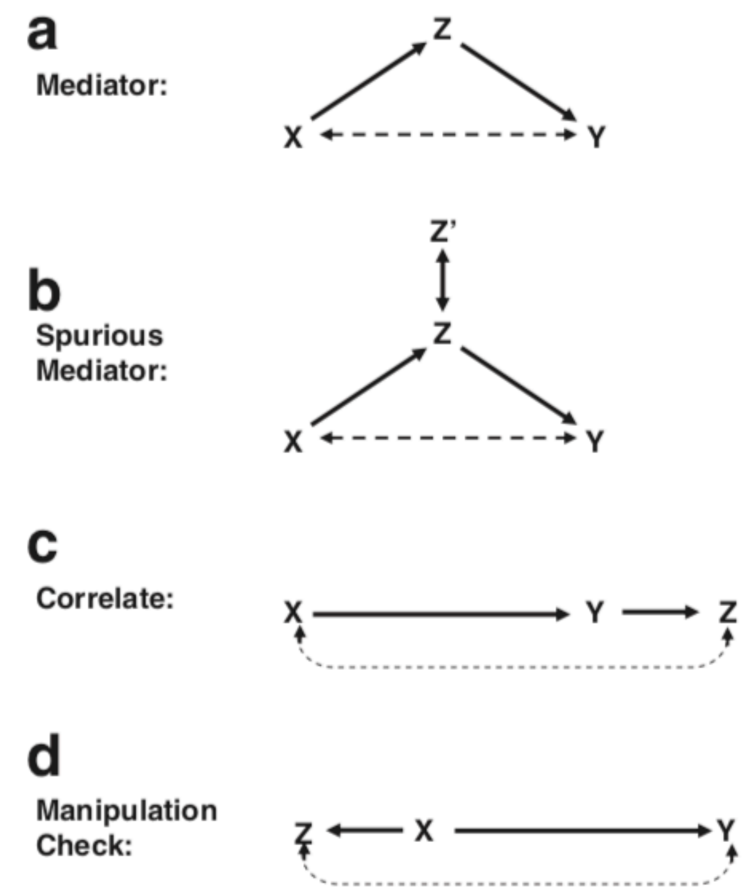


Fig. 2. Different causal models for the interpretation of the same intercorrelations between three variables, X, Y, and Z.

we need experiments to tell apart possible causal structures

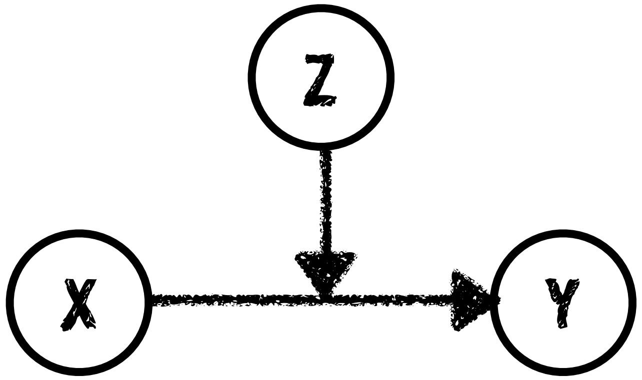
Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, 47(6), 1231-1236.

Plan for today

- Quick review: Controlling for variables
- Some more questions:
  - standardizing predictors
  - dummy coding vs. effect coding
- Mediation
- **Moderation**
- Model comparison
  - Cross-validation
  - AIC and BIC

Moderation

Definition

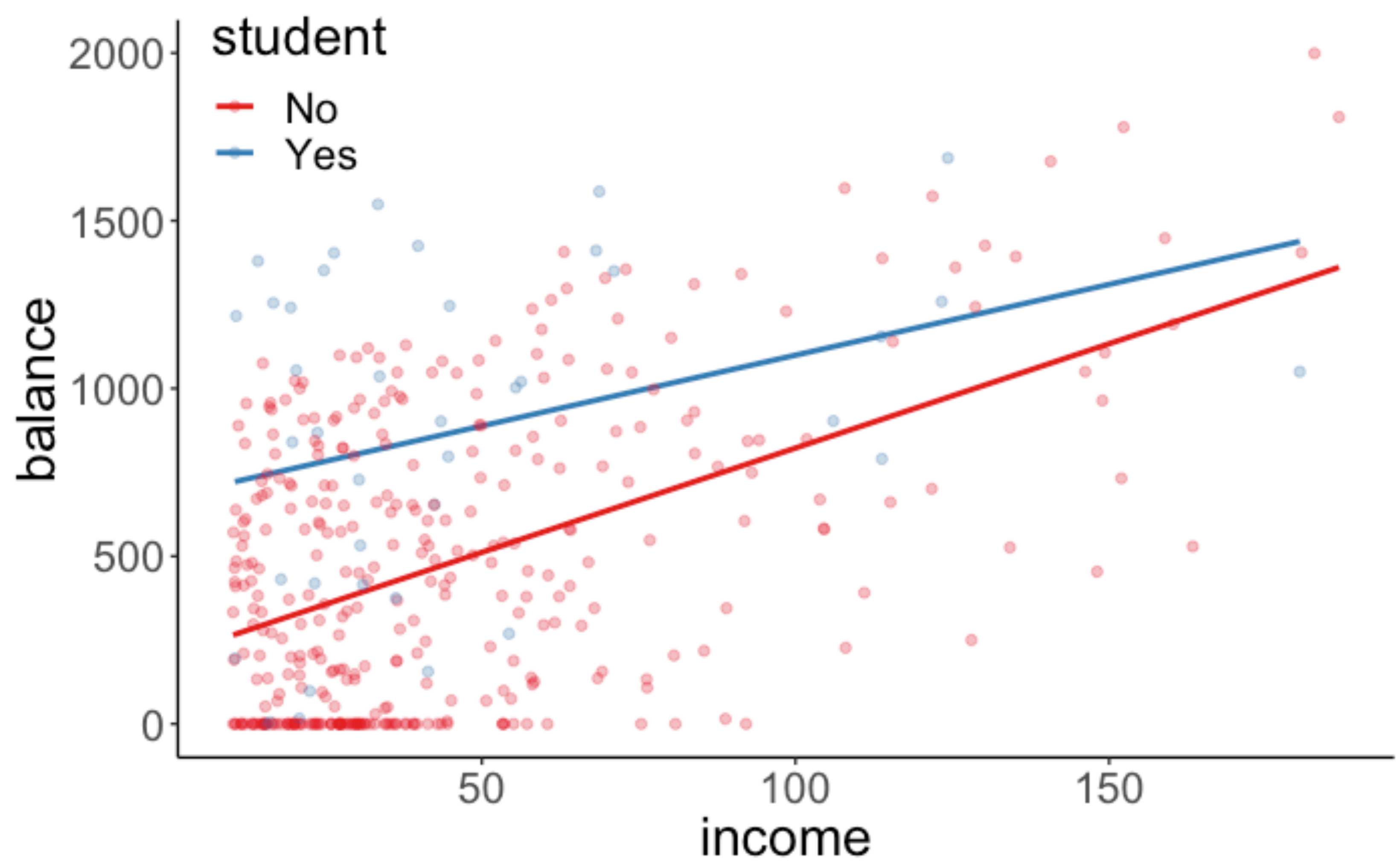


Moderation means that the effect of a predictor depends on the value of another.

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.



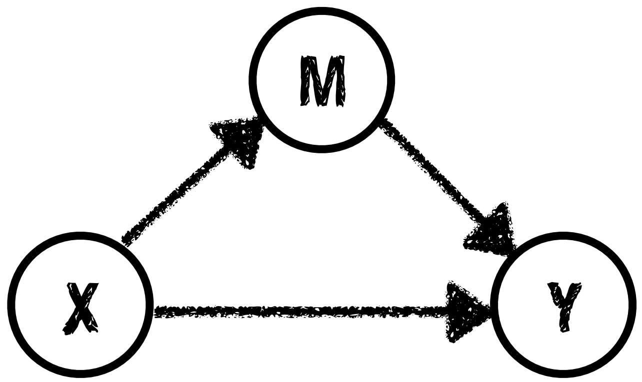
$$\widehat{\text{balance}}_i = 200.62 + 6.22 \cdot \text{income}_i + 476.68 \cdot \text{student}_i - 2.00 \cdot (\text{income}_i \times \text{student}_i)$$

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

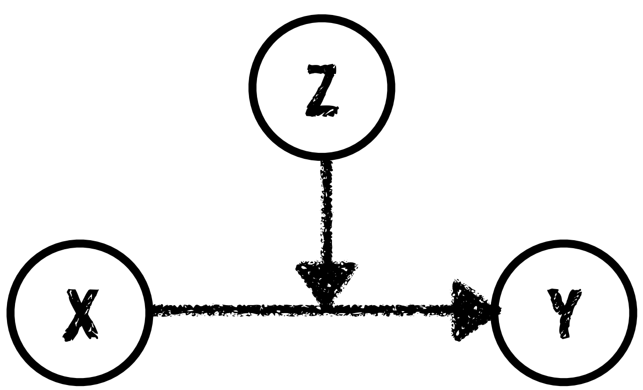
if student = "Yes"

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

Mediation and moderation

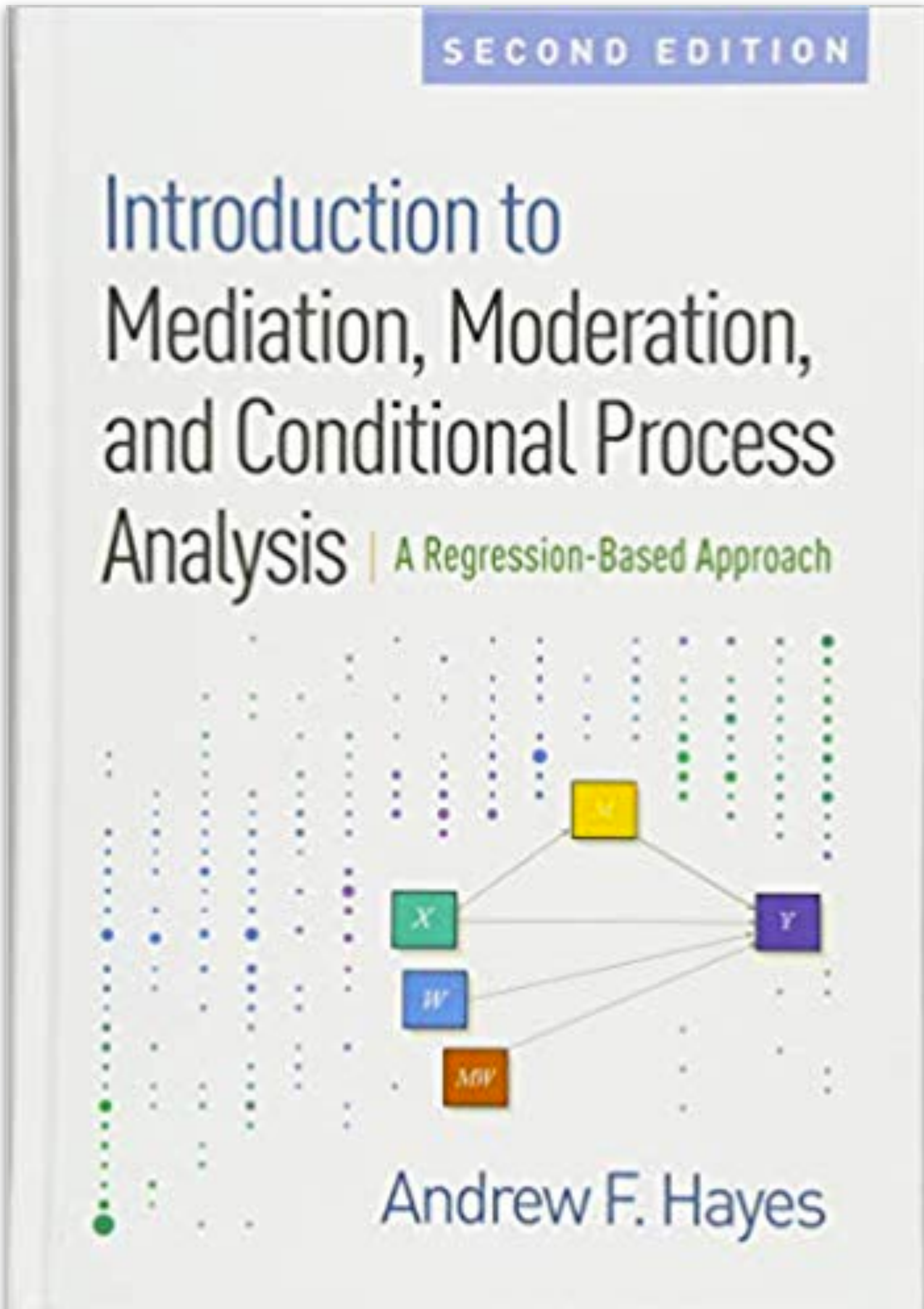


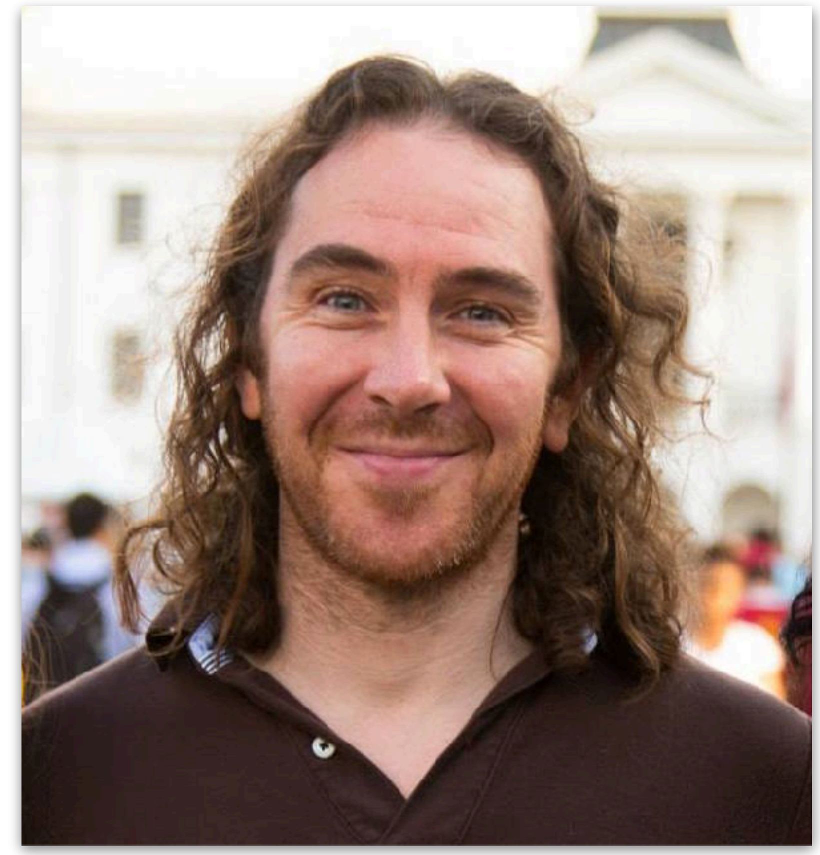
Mediation means that the effect of `X` on `Y` is (partially) indirect via another variable `M`.



Moderation means that the effect of a `X` on `Y` depends on the value of another variable `Z`.

Learn more about mediation and moderation





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

Plan for today

- Quick review: Controlling for variables
- Some more questions:
  - standardizing predictors
  - dummy coding vs. effect coding
- Mediation
- Moderation
- **Model comparison**
  - **Cross-validation**
  - AIC and BIC

Model comparison

The general procedure

1. Define H_0 as Model C (compact) and H_1 as Model A (augmented)
2. Fit model parameters to the data
3. Calculate the proportional reduction of error (PRE) in our sample
4. Decide whether the augmented model is **worth it** by comparing the observed PRE in our sample to the sampling distribution of PRE (assuming that H_0 is true)

Any problems with our approach?

sometimes it doesn't work ...

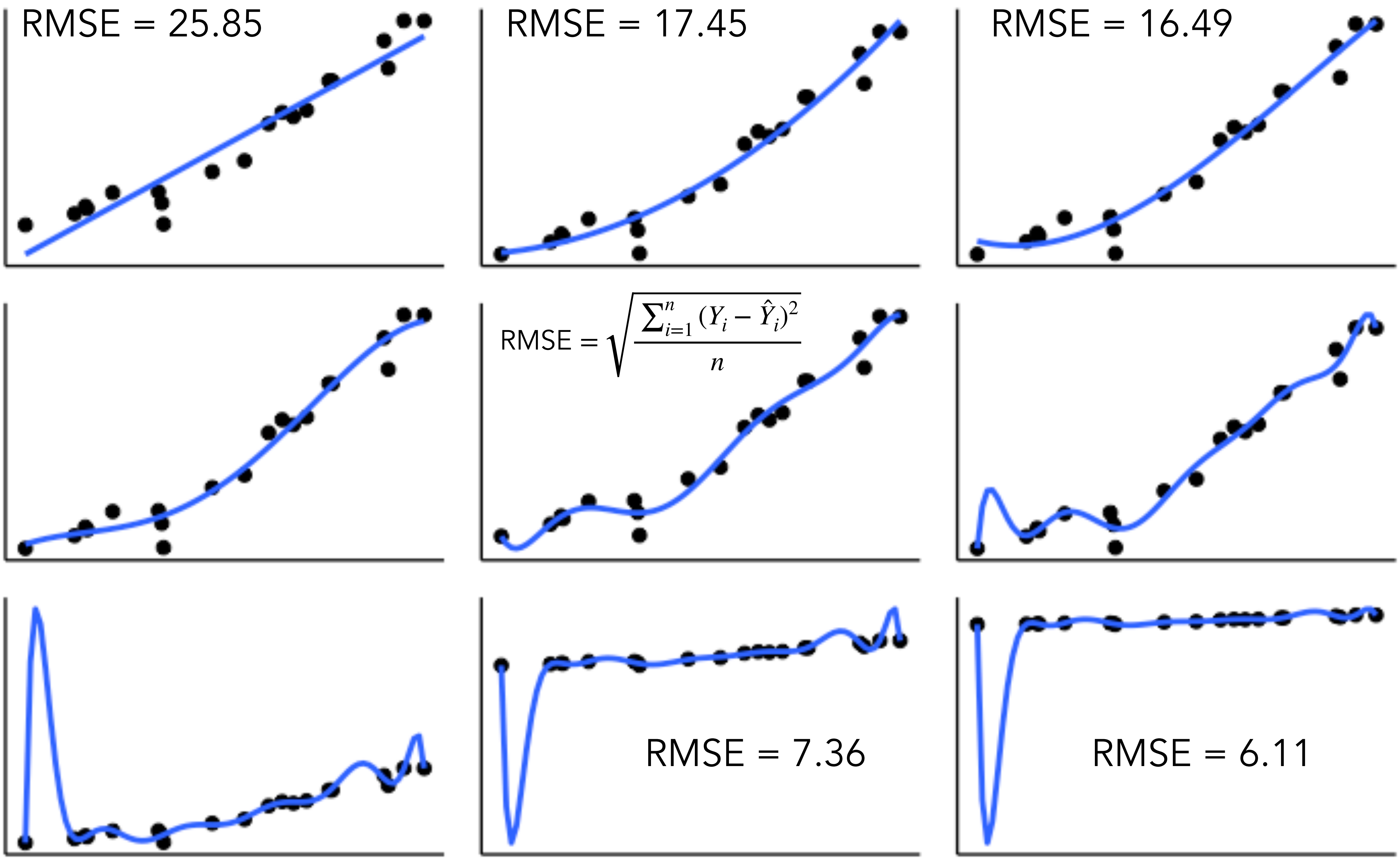
Model C	Model A
$\text{balance}_i = \beta_0 + \epsilon_i$	$\text{balance}_i = \beta_0 + \beta_1 \cdot \text{student}_i + \beta_2 \cdot \text{age}_i + \epsilon_i$ ✓
$\text{balance}_i = \beta_0 + \beta_1 \cdot \text{student}_i + \epsilon_i$	$\text{balance}_i = \beta_0 + \beta_1 \cdot \text{student}_i + \beta_2 \cdot \text{age}_i + \epsilon_i$ ✓
$\text{balance}_i = \beta_0 + \beta_1 \cdot \text{student}_i + \epsilon_i$	$\text{balance}_i = \beta_0 + \beta_1 \cdot \text{age}_i + \epsilon_i$ ✗
$\text{balance}_i = \beta_0 + \beta_1 \cdot \text{student}_i + \epsilon_i$	$\text{balance}_i = \beta_0 + \beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{degree}_i + \epsilon_i$ ✗

Tools for model comparison

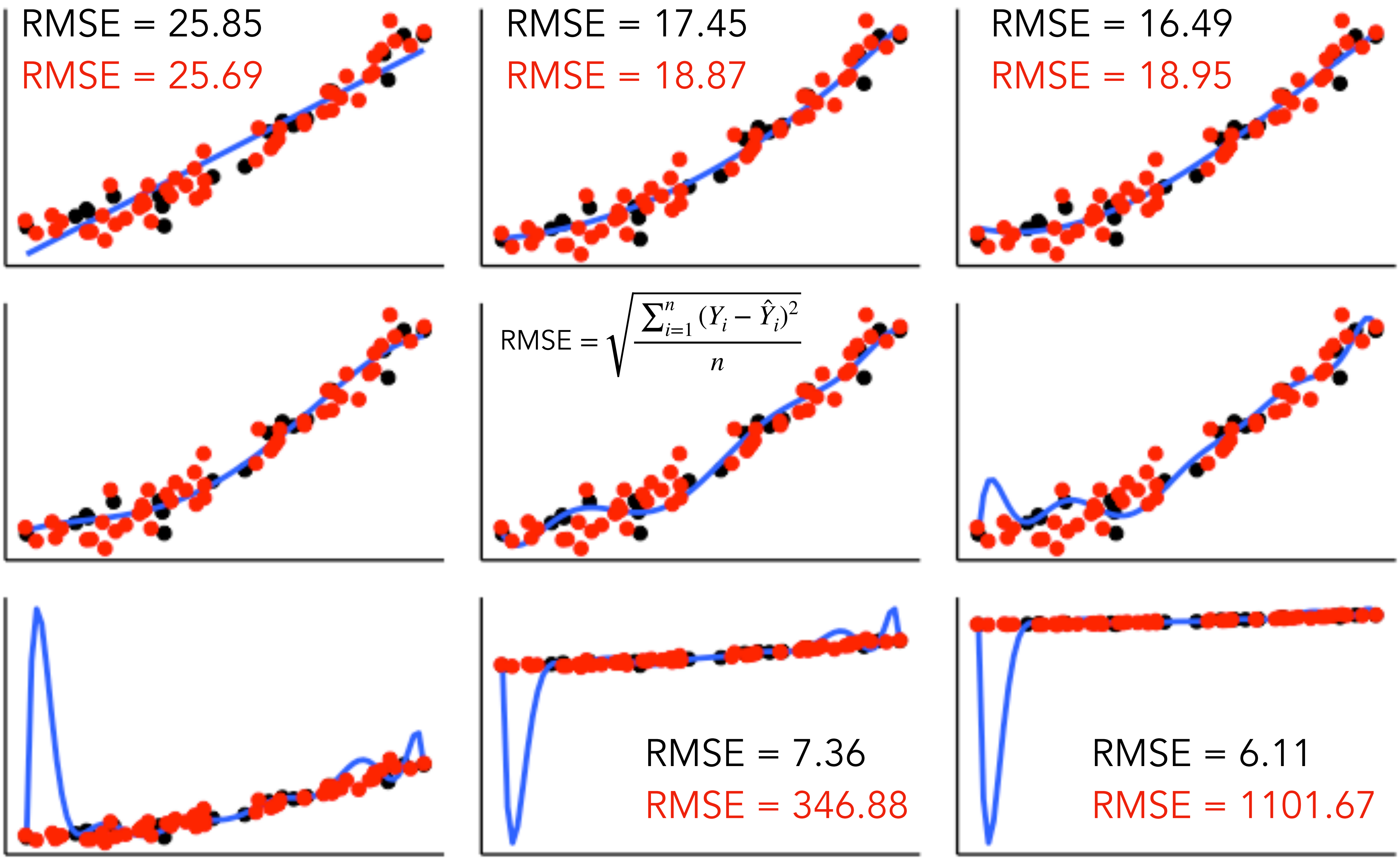
- **`anova()`:** compare a compact model with an augmented model via the F-test
  - problem: only works for nested models (where the augmented model contains all the predictors of the compact model and more)
- **What if we want to compare models that aren't nested?**
  - Cross-validation
  - AIC and BIC
  - Bayesian data analysis (we'll get there soon)

Cross-validation

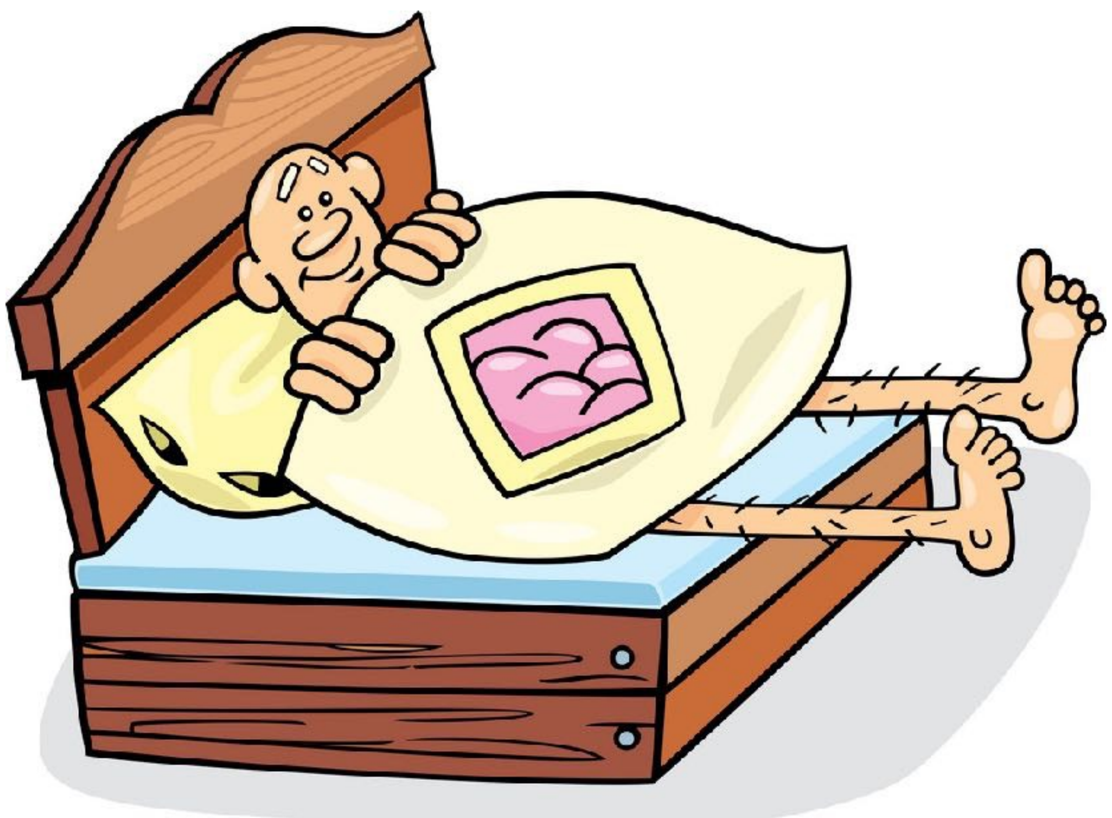
Which model describes the data best?



Which model describes the data best?



Underfitting vs. Overfitting



ONE WAY TO EXPLAIN UNDERFITTING

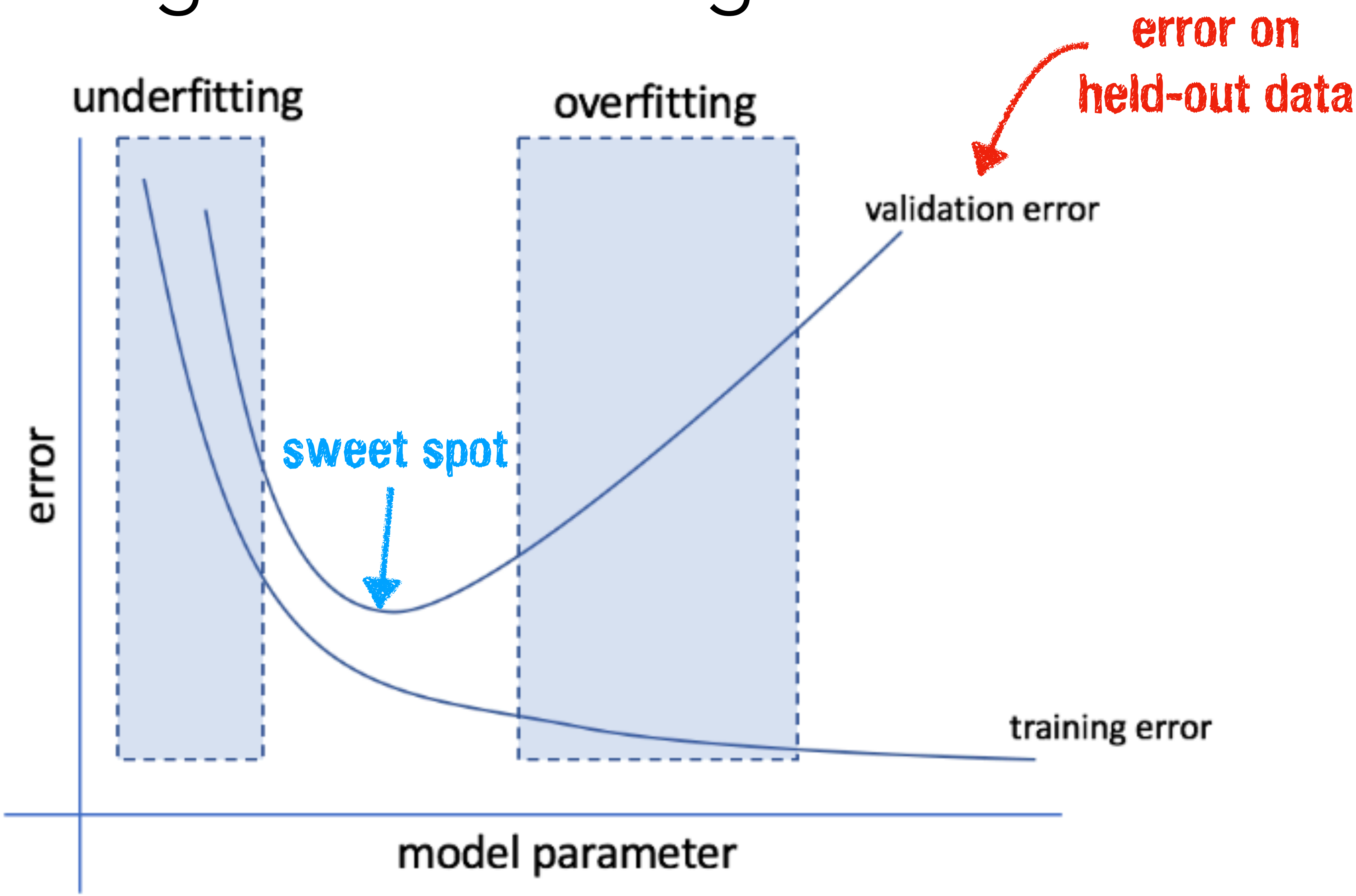
Underfitting vs. Overfitting

- **a good model should**:
  - explain the actual data well
  - predict future data well
- **bias-variance tradeoff**:
  - **bias** = error from erroneous assumptions in the model, high bias can cause a model to miss the relevant relations between predictors and outcome

    underfitting
  - **variance** = error from sensitivity to small fluctuations in the data, high variance can cause a model to fit the random noise in the data

    overfitting

Underfitting vs. Overfitting



in machine learning, the goal is often to find the sweet spot between underfitting and overfitting

Leave-one-out crossvalidation

non-inspirational quote

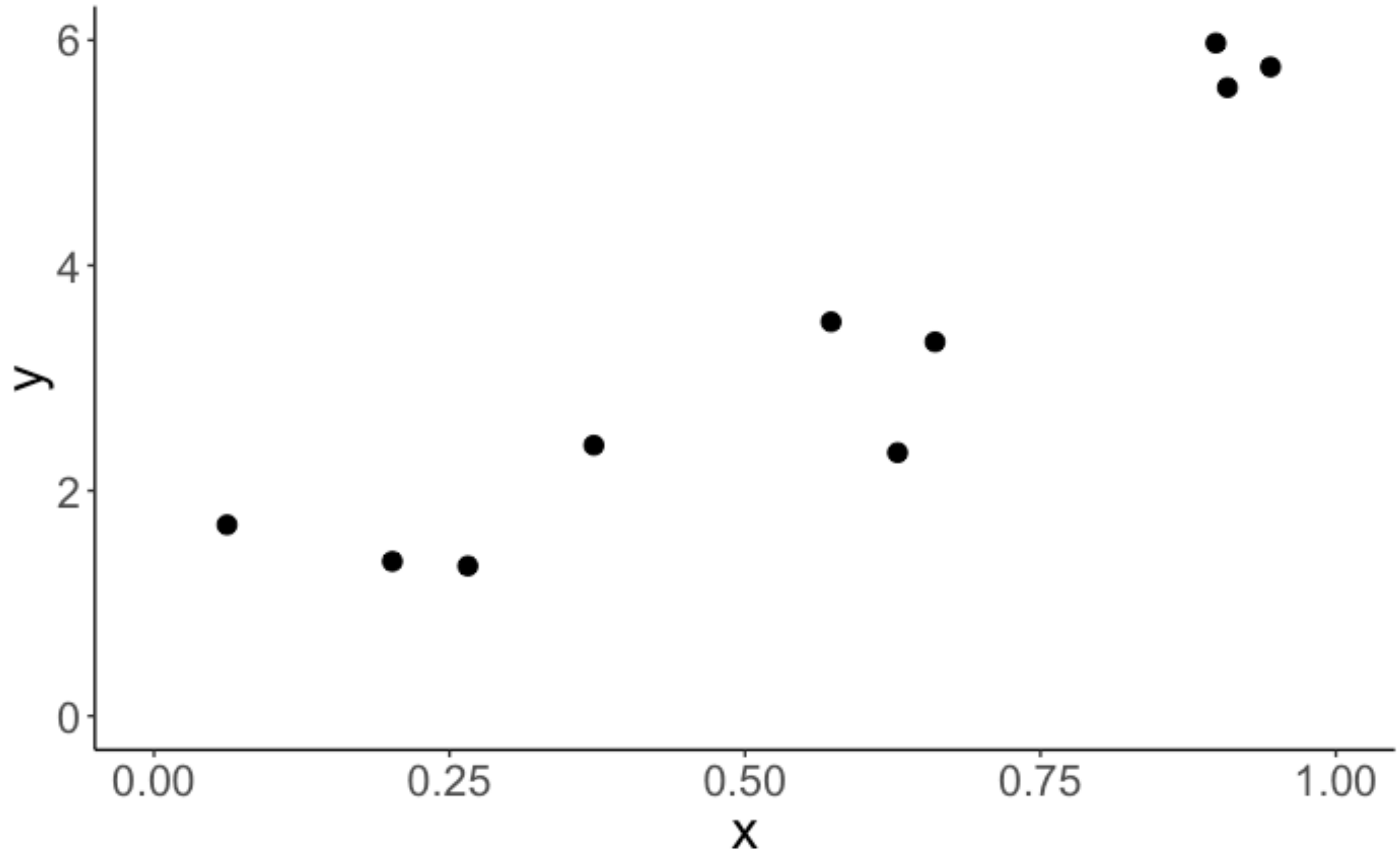
LOO

Leave one out cross-validation

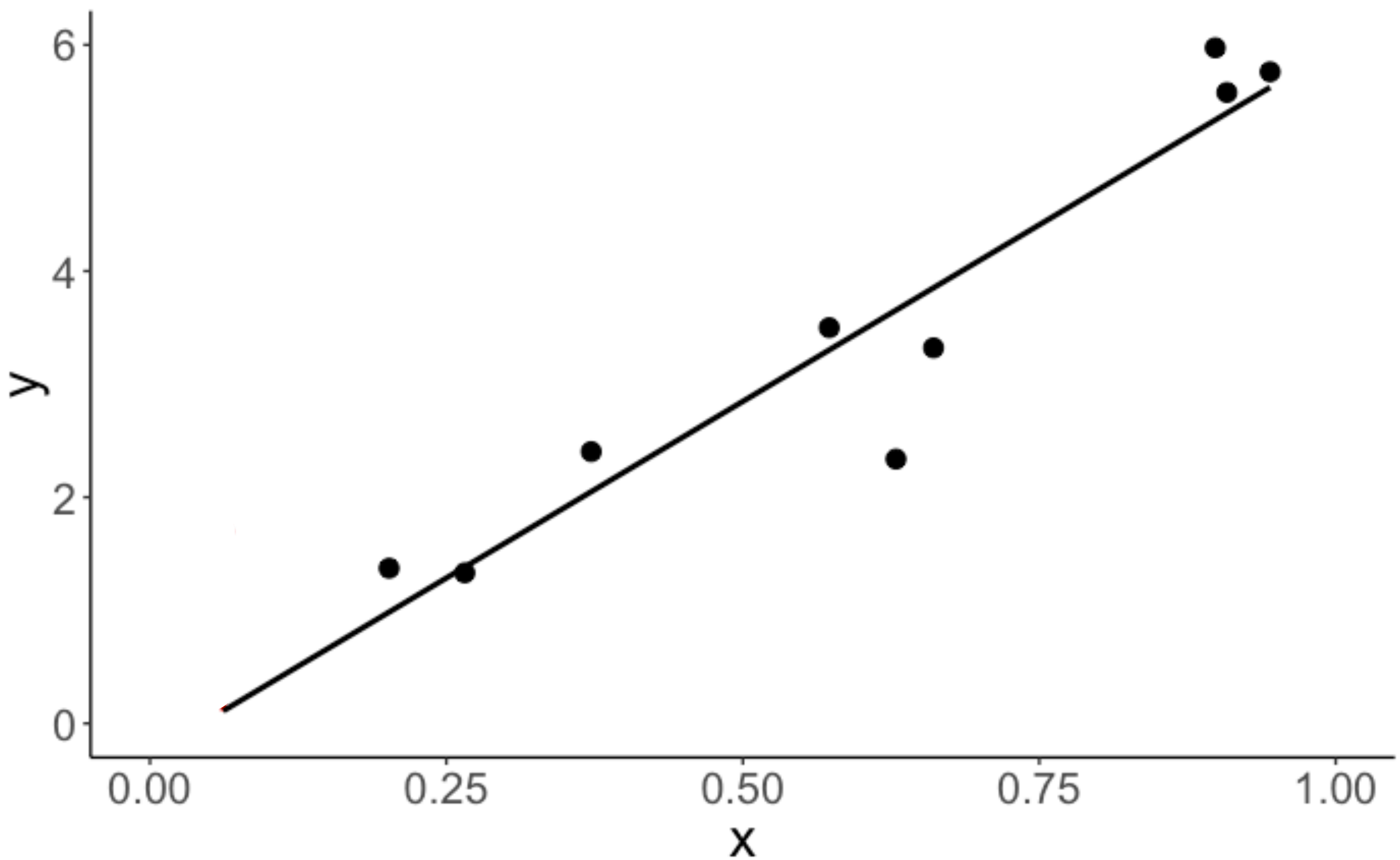
- train the model on all the data points except for one
- calculate the prediction error for the held-out data point

repeat for all data points

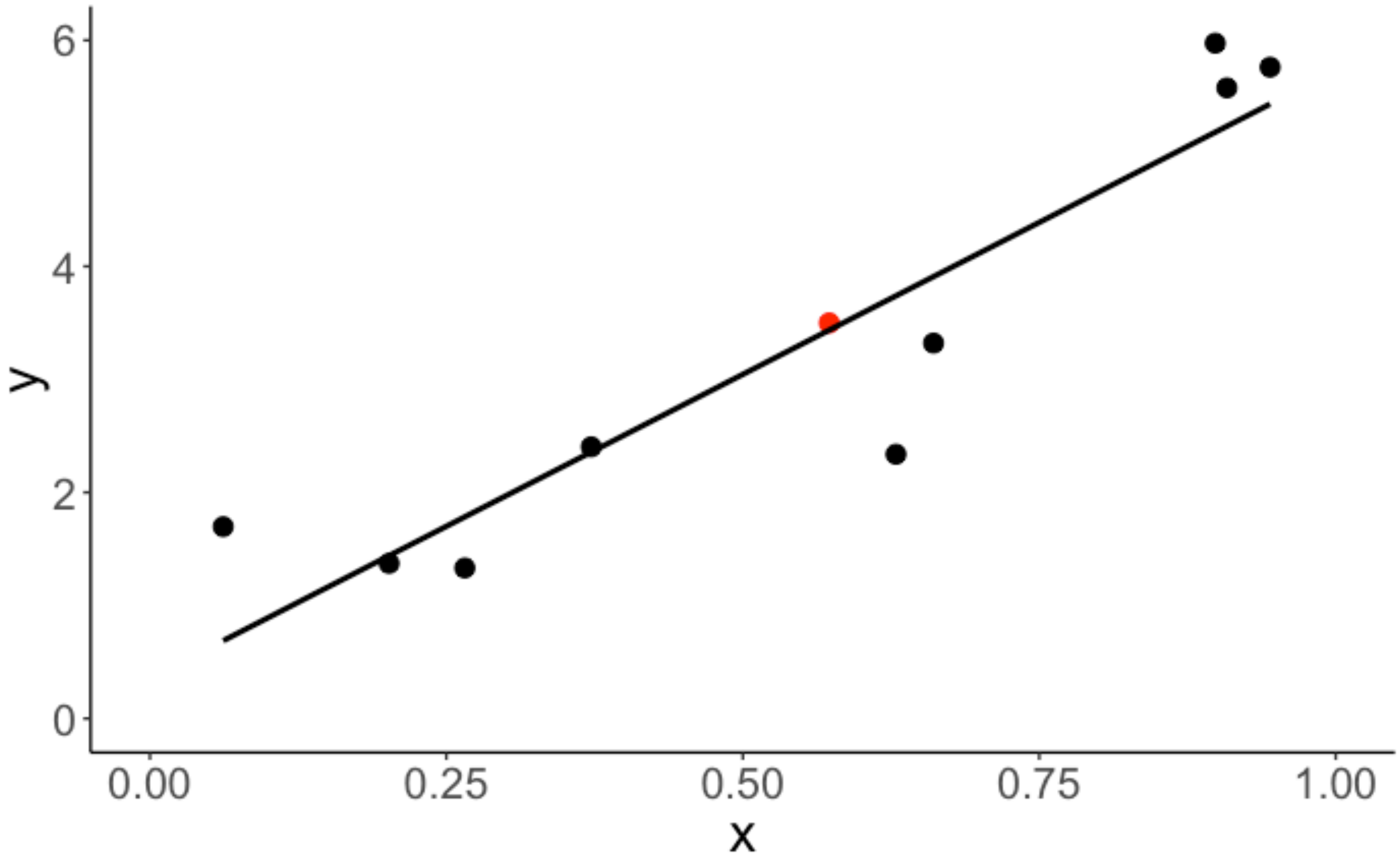
Leave one out cross-validation



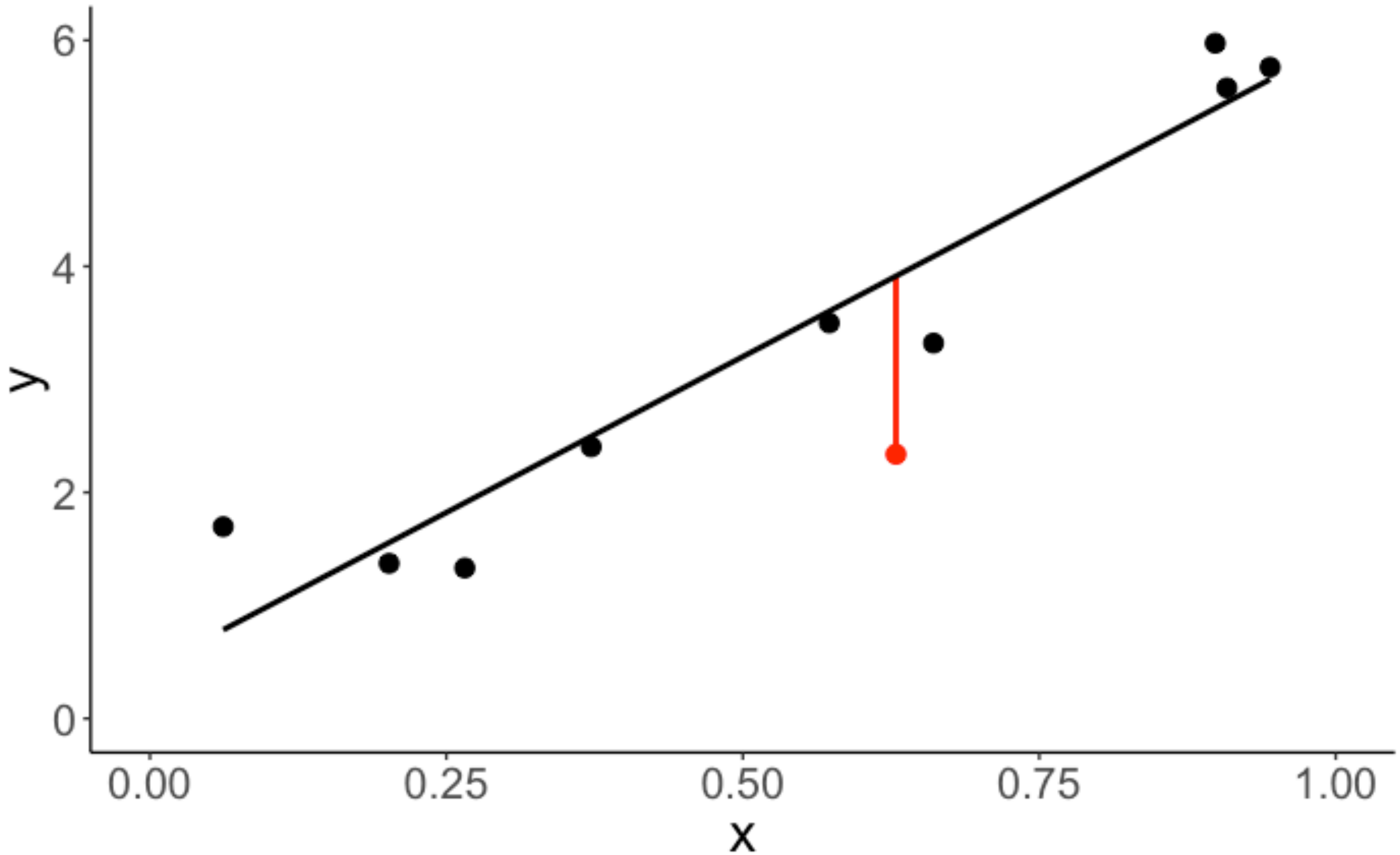
Leave one out cross-validation



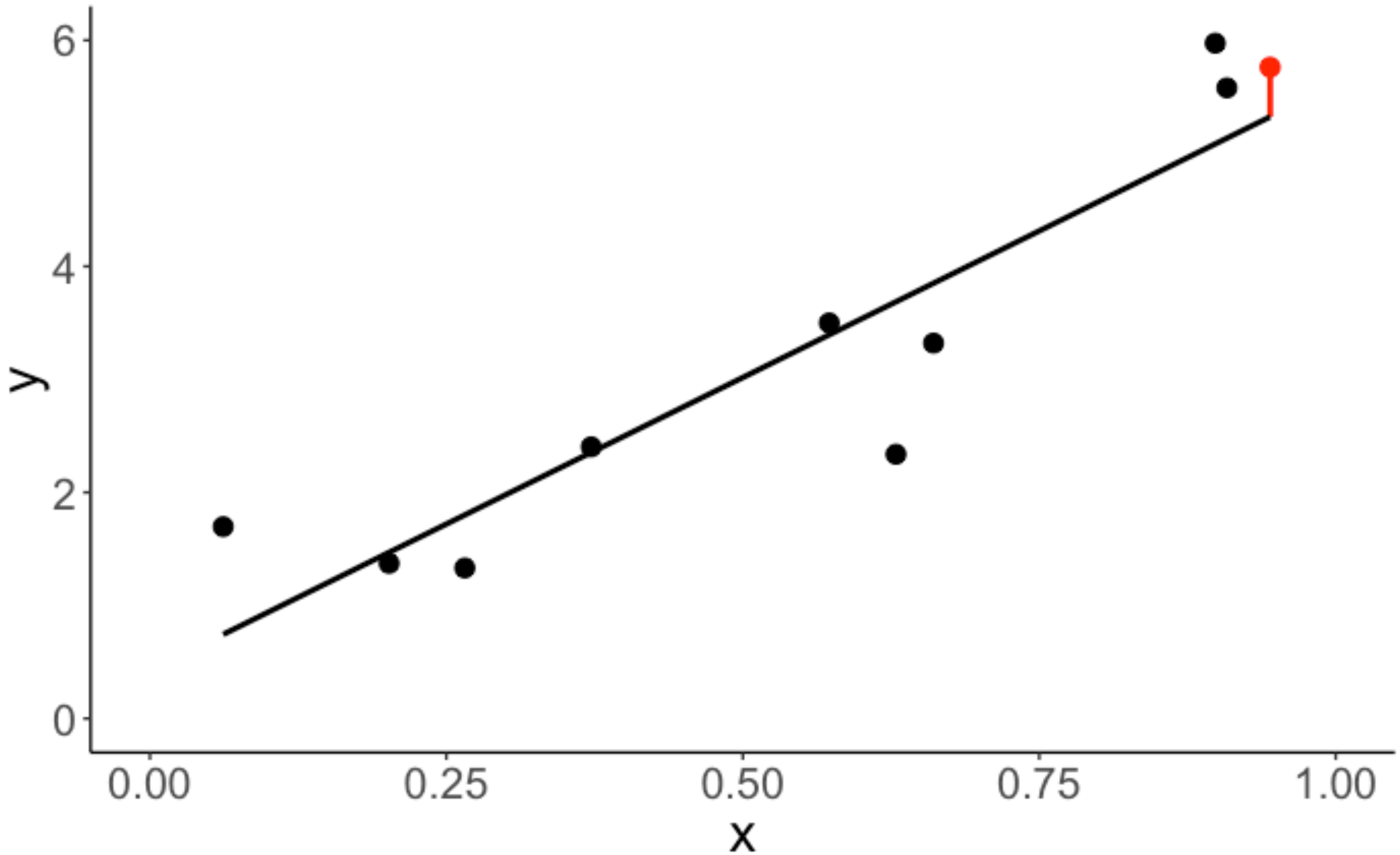
Leave one out cross-validation



Leave one out cross-validation

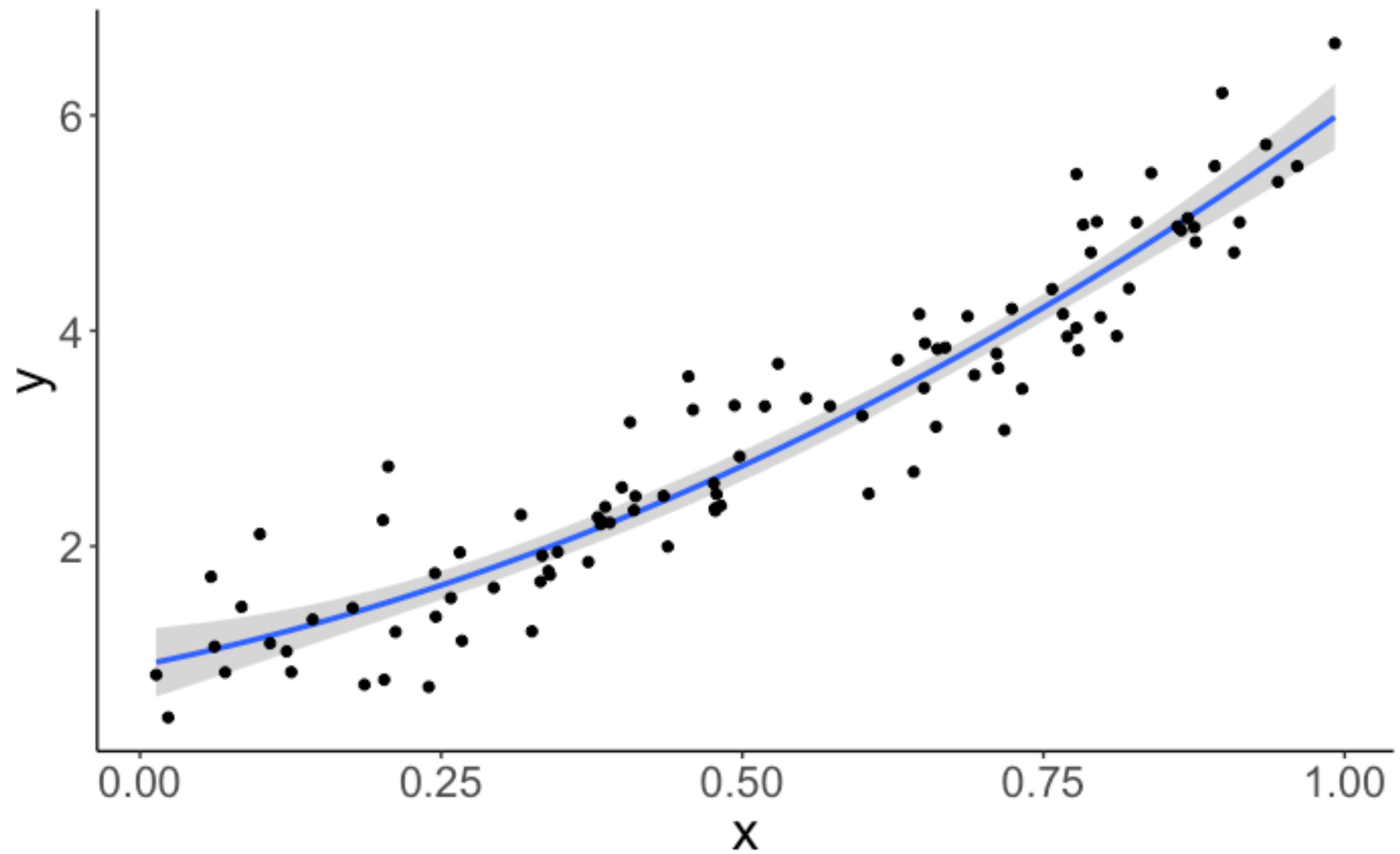


Leave one out cross-validation



Leave-one out crossvalidation

```r
# make example reproducible
set.seed(1)

# parameters
sample_size = 100
b0 = 1
b1 = 2
b2 = 3
sd = 0.5

# sample
df.data = tibble(
  participant = 1:sample_size,
  x = runif(sample_size, min = 0, max = 1),
  y = b0 + b1*x + b2*x^2 + rnorm(sample_size, sd = sd)
)
```

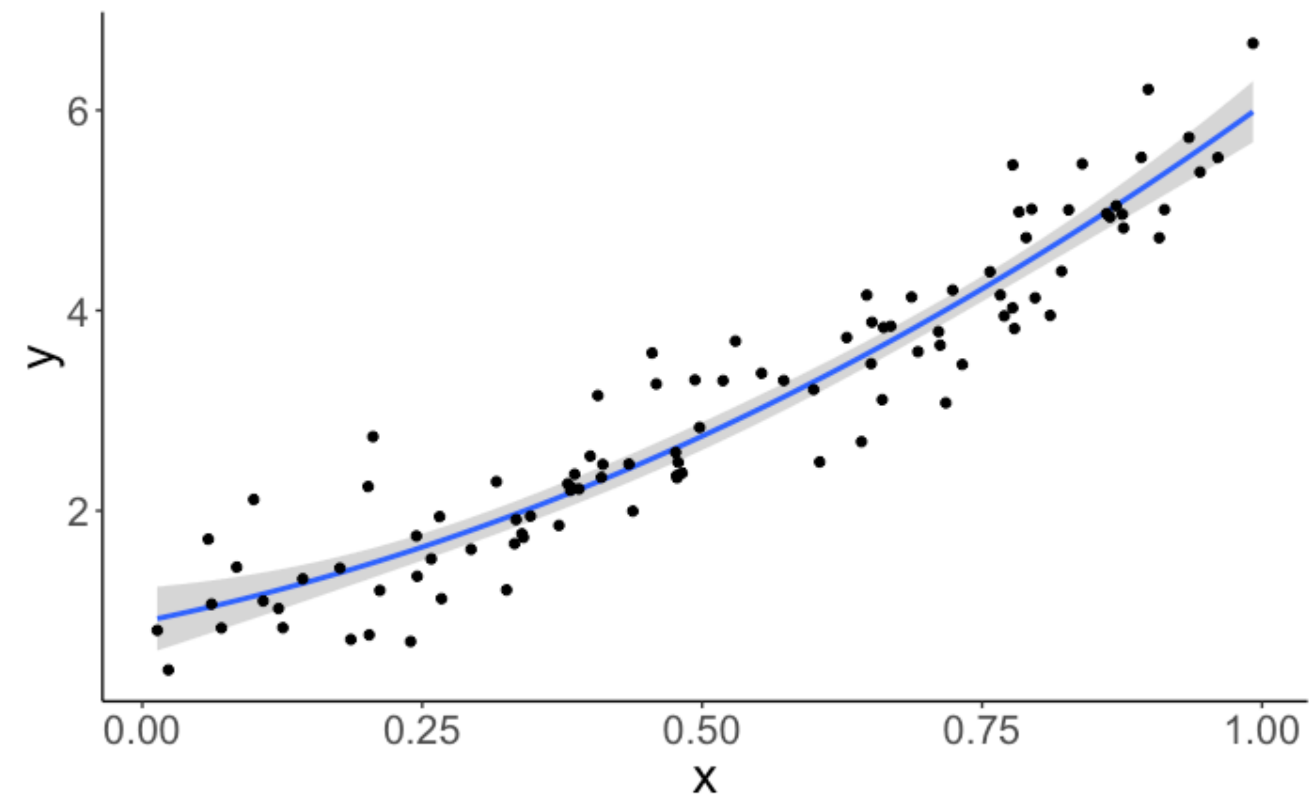


Leave-one out crossvalidation

ground truth

$$y_i = 1 + 2 \cdot x_i + 3 \cdot x_i^2 + e$$

$$e \sim \mathcal{N}(\text{mean} = 0,\ \text{sd} = 0.5)$$



candidate models

simple $\widehat{y}_i = b_0 + b_1 \cdot x_i$

correct $\widehat{y}_i = b_0 + b_1 \cdot x_i + b_2 \cdot x_i^2$

complex $\widehat{y}_i = b_0 + b_1 \cdot x_i + b_2 \cdot x_i^2 + b_3 \cdot x_i^3$

we could do an F-test here since the models are nested ...

Leave-one out crossvalidation

```r
library("modelr")

df.cross = df.data %>%
  crossv_loo() %>% # function which generates training and test data sets
  mutate(model_simple = map(train, ~ lm(y ~ 1 + x, data = .)),
         model_correct = map(train, ~ lm(y ~ 1 + x + I(x^2), data = .)),
         model_complex = map(train, ~ lm(y ~ 1 + x + I(x^2) + I(x^3), data = .))) %>%
  pivot_longer(cols = contains("model"),
               names_to = "model",
               values_to = "fit") %>%
  mutate(rmse = map2_dbl(.x = fit, .y = test, ~ rmse(.x, .y)))
```

nice package for basic cross-validation

model	mean_rmse
simple	0.65
correct	0.42
complex	0.41

complex model has the lowest error on the training data

Leave-one out crossvalidation

```r
library("modelr")

df.cross = df.data %>%
  crossv_loo() %>%
```

splits the data set into training and test

	train	test	.id
1	list(data = list(participant = 1:10, x = c(0.265508663...	list(data = list(participant = 1:10, x = c(0.265508663...	1
2	list(data = list(participant = 1:10, x = c(0.265508663...	list(data = list(participant = 1:10, x = c(0.265508663...	2
3	list(data = list(participant = 1:10, x = c(0.265508663...	list(data = list(participant = 1:10, x = c(0.265508663...	3
4	list(data = list(participant = 1:10, x = c(0.265508663...	list(data = list(participant = 1:10, x = c(0.265508663...	4
5	list(data = list(participant = 1:10, x = c(0.265508663...	list(data = list(participant = 1:10, x = c(0.265508663...	5
6	list(data = list(participant = 1:10, x = c(0.265508663...	list(data = list(participant = 1:10, x = c(0.265508663...	6
7	list(data = list(participant = 1:10, x = c(0.265508663...	list(data = list(participant = 1:10, x = c(0.265508663...	7
8	list(data = list(participant = 1:10, x = c(0.265508663...	list(data = list(participant = 1:10, x = c(0.265508663...	8
9	list(data = list(participant = 1:10, x = c(0.265508663...	list(data = list(participant = 1:10, x = c(0.265508663...	9
10	list(data = list(participant = 1:10, x = c(0.265508663...	list(data = list(participant = 1:10, x = c(0.265508663...	10

each entry is simply a pointer to the data set plus some indices `.idx` for the filtered data points

Leave-one out crossvalidation

```
library("modelr")

df.cross = df.data %>%
  crossv_loo() %>%
  mutate(model_simple = map(train, ~ lm(y ~ 1 + x, data = .)),
         model_correct = map(train, ~ lm(y ~ 1 + x + I(x^2), data = .)),
         model_complex = map(train, ~ lm(y ~ 1 + x + I(x^2) + I(x^3), data = .))) %>%
```

fit three different models to the training data set

Leave-one out crossvalidation

```
library("modelr")

df.cross = df.data %>%
  crossv_loo() %>%
  mutate(model_simple = map(train, ~ lm(y ~ 1 + x, data = .)),
         model_correct = map(train, ~ lm(y ~ 1 + x + I(x^2), data = .)),
         model_complex = map(train, ~ lm(y ~ 1 + x + I(x^2) + I(x^3), data = .))) %>%
  pivot_longer(cols = contains("model"),
               names_to = "model",
               values_to = "fit") %>%
  mutate(rmse = map2_dbl(.x = fit, .y = test, ~ rmse(.x, .y)))
```

calculate the root mean squared error for each model on the test data set

```
df.cross %>%
  group_by(model) %>%
  summarize(mean_rmse = mean(rmse))
```

model	mean_rmse
simple	0.65
correct	0.48
complex	0.70

the correct model has the lowest prediction error

Leave-one out crossvalidation

Any potential problems with LOO?

can be computationally expensive since it requires fitting the model `n` times (once for each data point) ...

k-fold cross validation

k-fold crossvalidation

Full data set

Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test

k-fold crossvalidation

k-fold crossvalidation

```r
df.cross = df.data %>%
  crossv_kfold(k = 10) %>%
  mutate(model_simple = map(train, ~ lm(y ~ 1 + x, data = .)),
         model_correct = map(train, ~ lm(y ~ 1 + x + I(x^2), data = .)),
         model_complex = map(train, ~ lm(y ~ 1 + x + I(x^2) + I(x^3), data = .))) %>%
  pivot_longer(cols = contains("model"),
               names_to = "model",
               values_to = "fit") %>%
  mutate(rsquare = map2_dbl(fit, test, rsquare))
```

using R^2 as a measure

why didn't we use R^2 for LOO?

this wouldn't work for LOO since we only have one data point in the test data...

model	median_rsquare
simple	0.839
correct	0.865
complex	0.860

the correct model accounts for the most variance in the test data

k-fold vs. leave-one-out crossvalidation

- **LOO**:
  - trained on **more** data
    - more variance
    - less bias
- **k-fold**:
  - trained on **less** data
    - less variance
    - more bias

Monte Carlo crossvalidation

Monte Carlo crossvalidation

random splits into training and test data

```
df.cross = df.data %>%
  crossv_mc(n = 50, test = 0.5) %>% # number of samples, and percentage of test
  mutate(model_simple = map(train, ~ lm(y ~ 1 + x, data = .x)),
         model_correct = map(train, ~ lm(y ~ 1 + x + I(x^2), data = .x)),
         model_complex = map(train, ~ lm(y ~ 1 + x + I(x^2) + I(x^3), data = .))) %>%
  gather("model", "fit", contains("model")) %>%
  mutate(rmse = map2_dbl(fit, test, rmse))
```

- splits can also be done in a **stratified** way
- for example, generate training data that has the same percentage of cases from each group
- fit data from some participants, test on data from other participants, ...

Plan for today

- Quick review: Controlling for variables
- Some more questions:
  - standardizing predictors
  - dummy coding vs. effect coding
- Mediation
- Moderation
- **Model comparison**
  - Cross-validation
  - **AIC and BIC**

AIC and BIC

AIC and BIC

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

not that much Bayesian about it ...

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

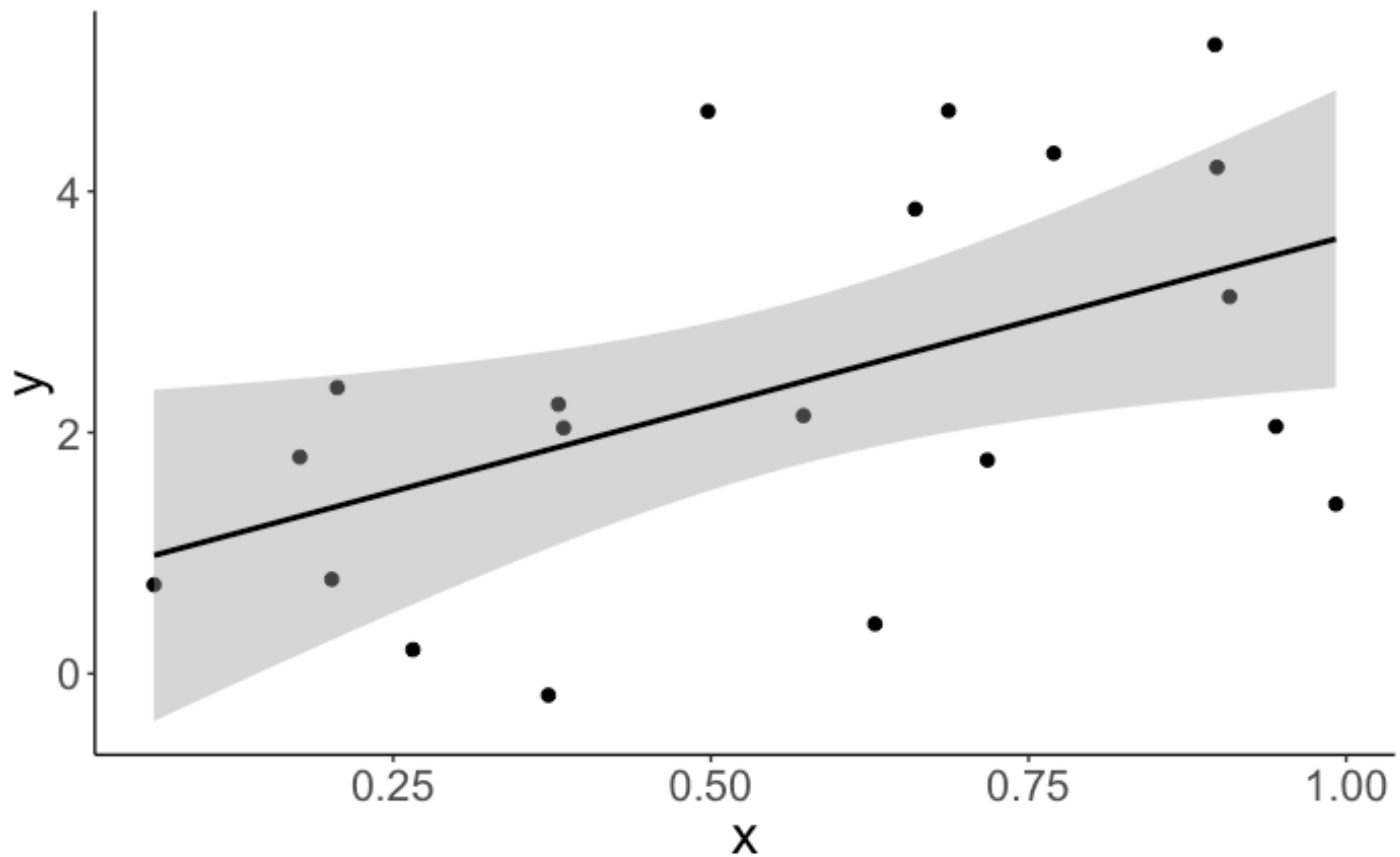
$$\text{BIC} = \ln(n)k - 2\ln(\hat{L})$$

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

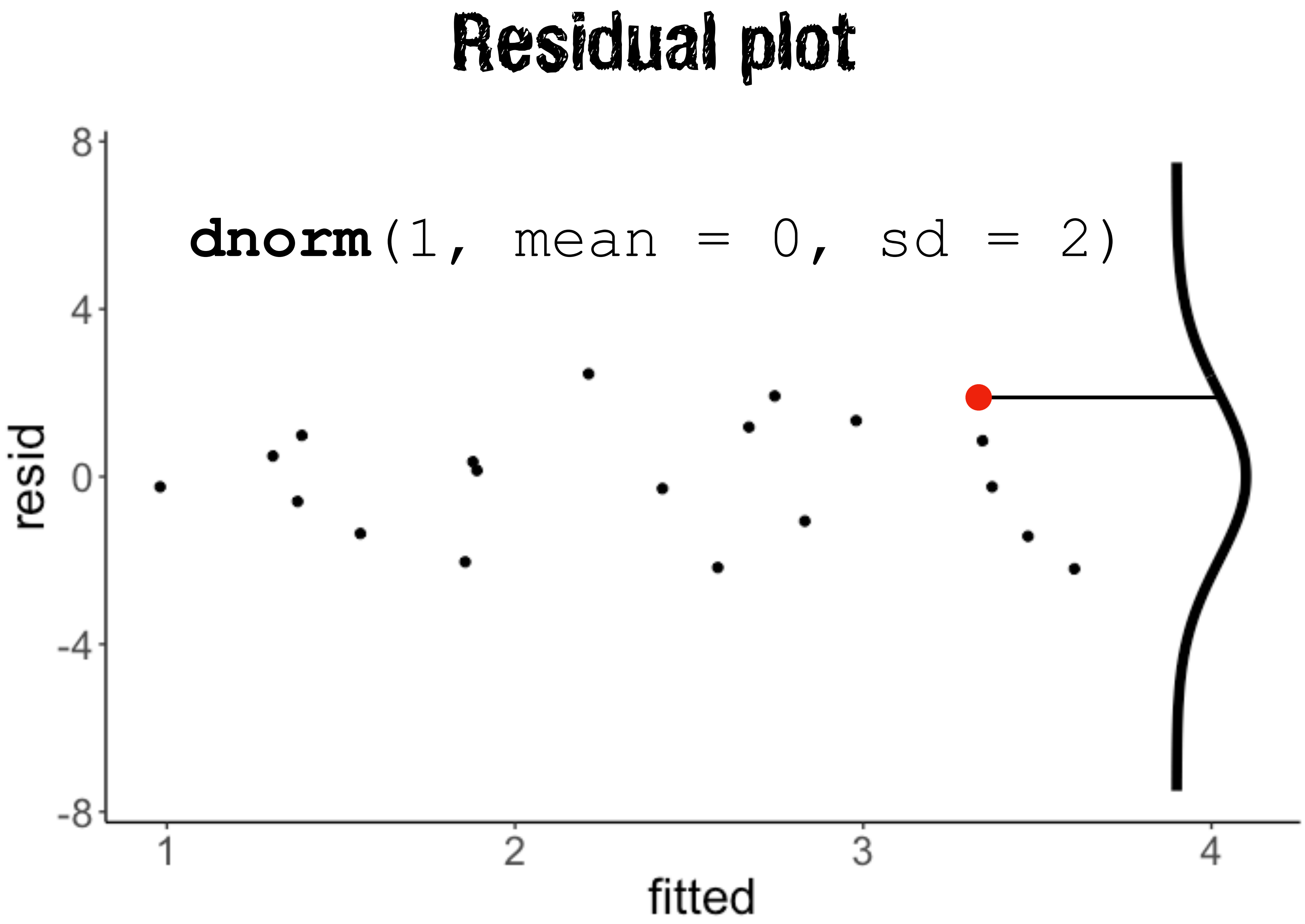
AIC and BIC

- How do we get the likelihood of our model?
  - in a linear regression, minimizing least squares is equivalent to maximizing the likelihood of the data given the model
- Assumption of the model:
  - residuals are normally distributed with:
    - `mean = 0` and `sd = sigma`
  - calculate overall likelihood by computing the likelihood of each residual, and then multiplying

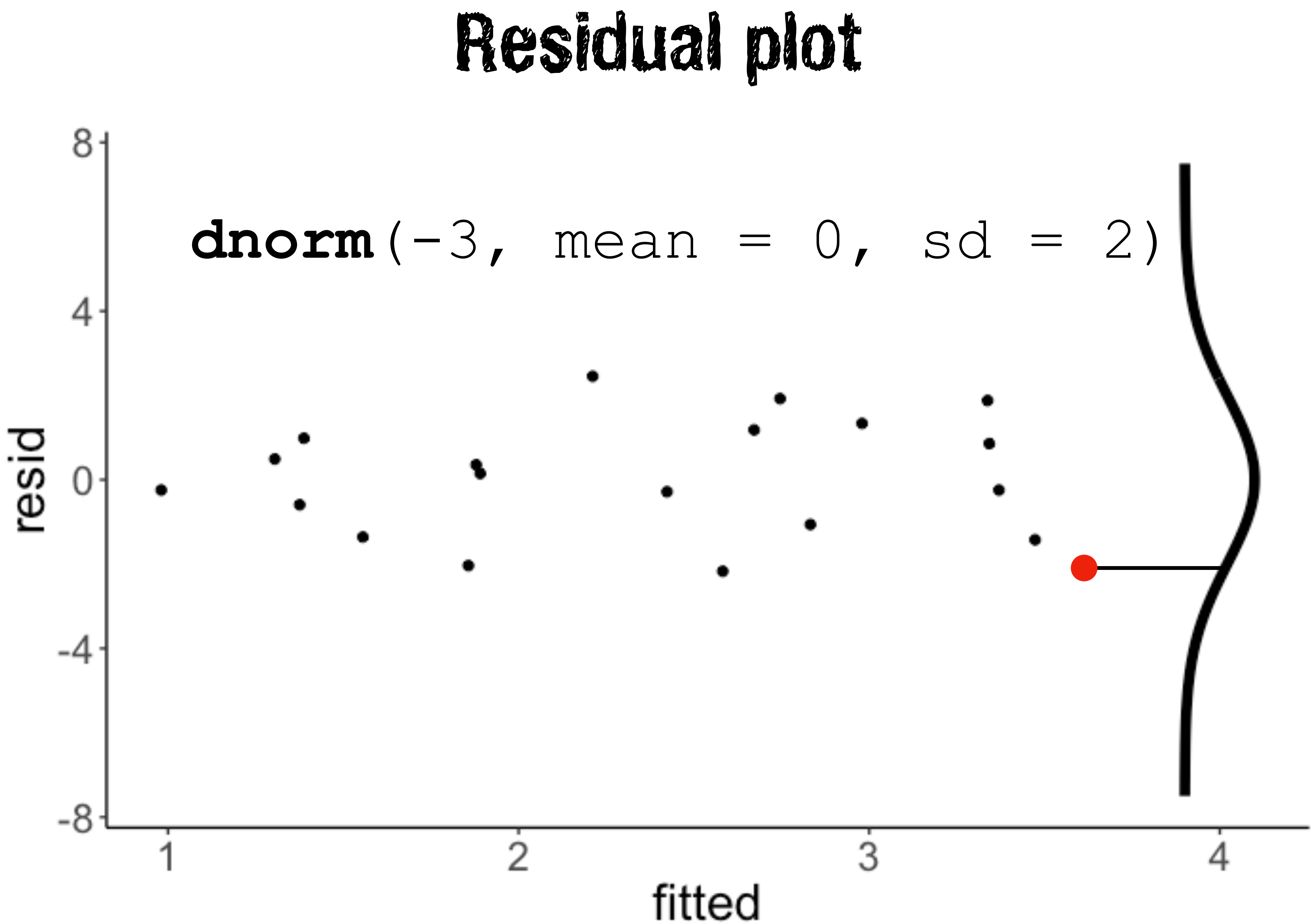
AIC and BIC



AIC and BIC



AIC and BIC



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

AIC and BIC

dnorm(4.20, mean = 0, sd = 2.17) = 0.02

```r
# generate some data
df.like = tibble(
  x = runif(20, min = 0, max = 1),
  y = 1 + 3 * x + rnorm(20, sd = 2)
)

# fit the model
fit = lm(formula = y ~ x,
         data = df.like)

# model summary
fit %>%
  glance()
```

x	y	fitted	resid	likelihood
0.41	5.74	1.53	4.20	0.02
0.91	4.86	4.80	0.06	0.18
0.29	0.98	0.80	0.18	0.18
0.46	0.71	1.87	-1.16	0.16
0.33	-0.34	1.05	-1.39	0.15
0.65	0.82	3.11	-2.29	0.11
0.26	-1.35	0.57	-1.92	0.12
0.48	4.75	2.00	2.75	0.08
0.77	4.96	3.86	1.11	0.16
0.08	0.80	-0.55	1.35	0.15

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
0.38	0.35	2.17	11.2	0	2	-42.78	91.56	94.55	84.43	18

$$e \sim \mathcal{N}(\text{mean} = 0,\ \text{sd} = 2.17)$$

AIC and BIC

ΔBIC	Evidence against higher BIC
0 to 2	Not worth more than a bare mention
2 to 6	Positive
6 to 10	Strong
>10	Very Strong

- for both AIC and BIC, *lower* is better!
- neither provide a test of a model in the sense of testing a null hypothesis
  - AIC or BIC tell us nothing about the absolute quality of a model, only the quality relative to other models
- The BIC generally penalizes free parameters more strongly than the Akaike information criterion, though it depends on the size of *n* and relative magnitude of *n* and *k*.

What shall I use when?

- Use it all!
- ideally, the different measures provide converging evidence

Table 2
Summary of the model results. Values for *r* and RMSE indicate means (with 5% and 95% quantiles in parentheses) based on 100 split-half cross-validation runs. BIC scores are based on running the models on the full data set.

Model	*r*	RMSE	BIC
Difference & pivotality	.86 (.66, .95)	10.56 (6.17, 17.21)	158.59
Difference	.70 (.30, .90)	26.92 (16.4, 40.6)	209.74
Pivotality	.63 (.41, .77)	14.23 (11.39, 17.54)	199.53
Optimality	.66 (.42, .84)	14.55 (10.54, 17.91)	199.47

Note: BIC = Bayesian Information Criterion (lower values indicate better model performance).

Summary

- Quick review: Controlling for variables
- Some more questions:
  - standardizing predictors
  - dummy coding vs. effect coding
- Mediation
- Moderation
- Model comparison
  - Cross-validation
  - AIC and BIC

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?

Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**

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

Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**

Thank you!