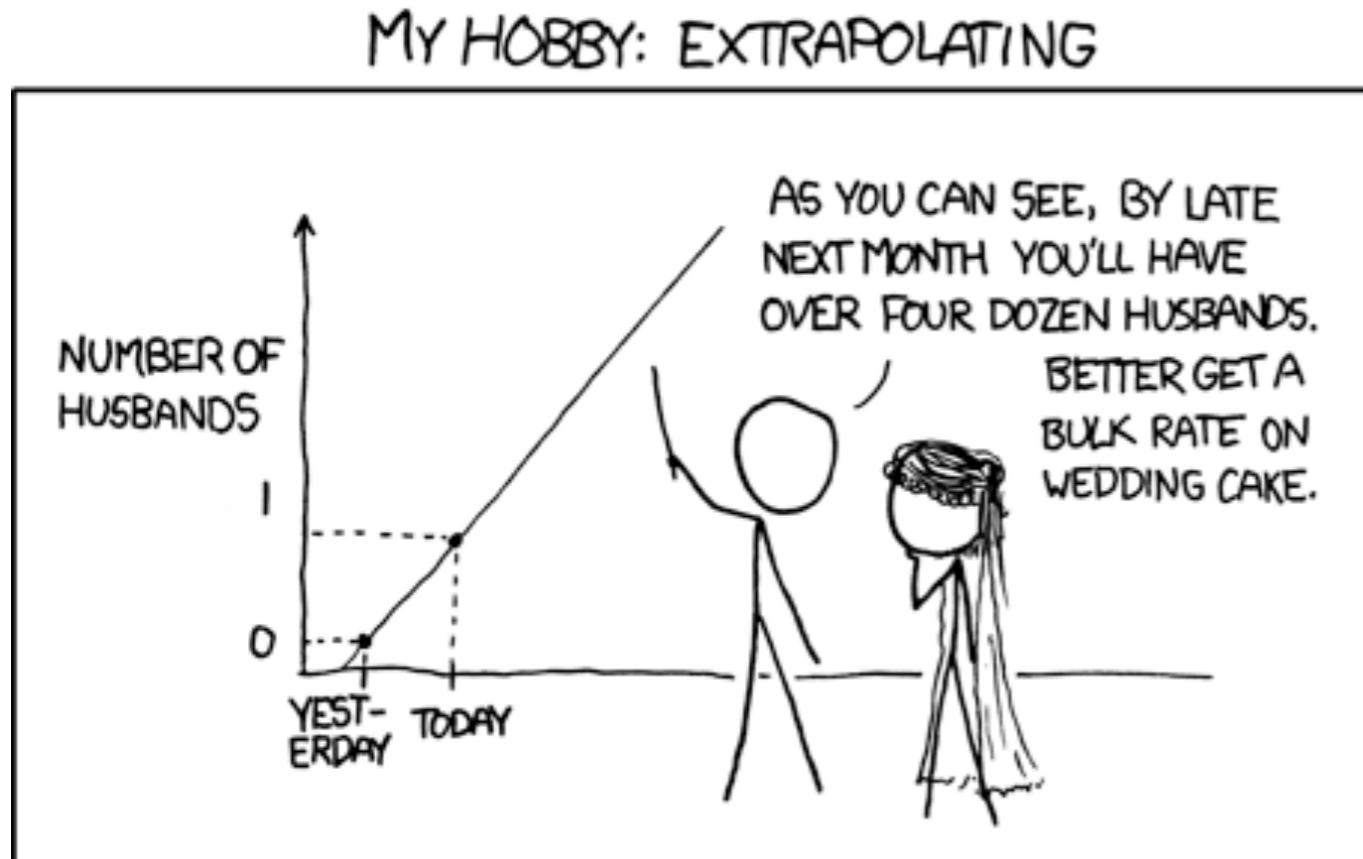


Linear model 1



Chat

Which side are you on: Frequentist vs. Bayesian?

To: Everyone ▾ More ▾

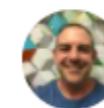
Type message here...



We're listening to
"Tamiditine" by
"Bombino" submitted
by Tobi

01/26/2022

Things that came up



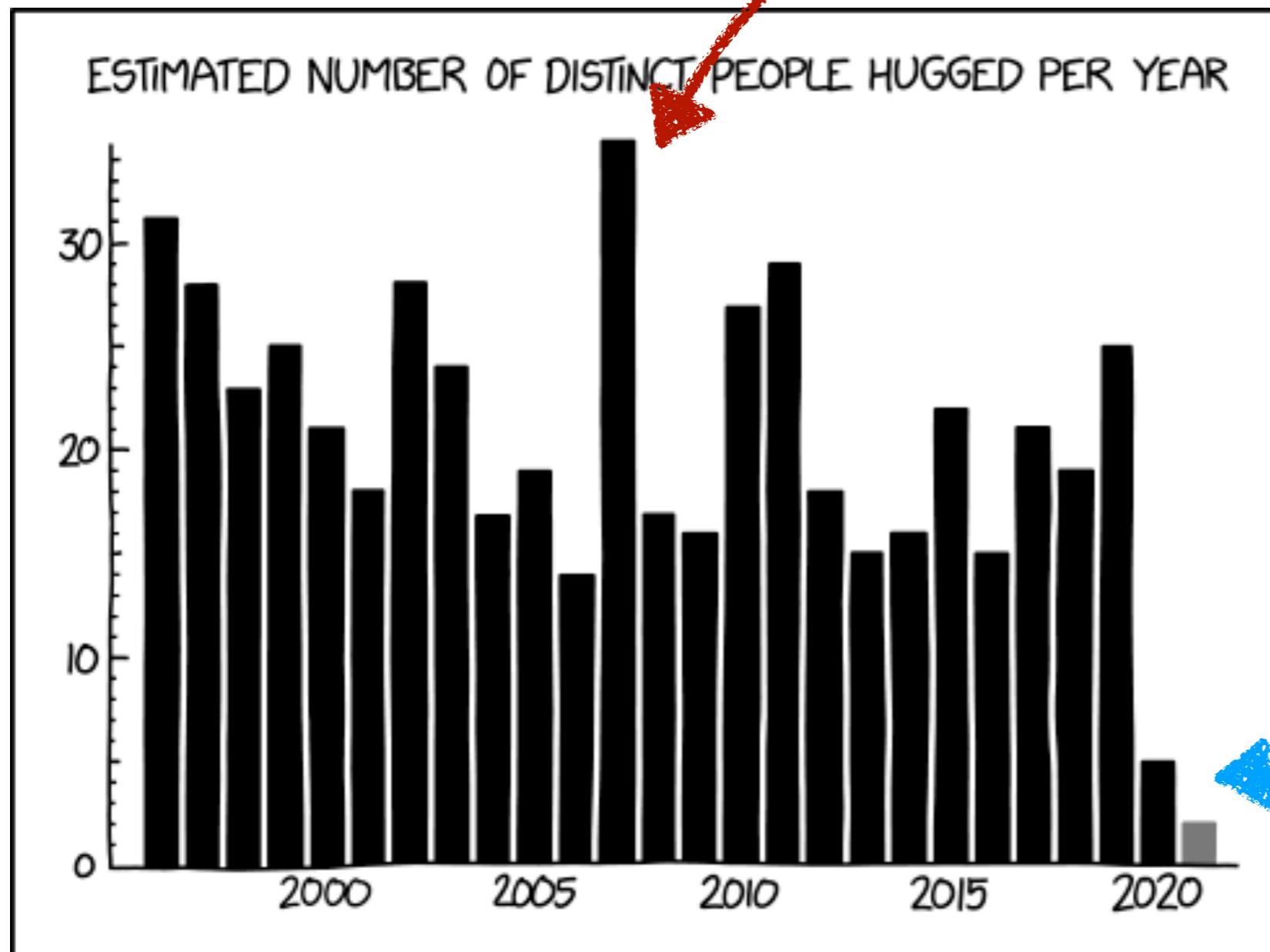
David Butler
@DavidKButlerUoA

...

Replying to @xkcdComic

Ok. So what happened in 2007?

3:05 PM · Feb 1, 2021 · Twitter Web App

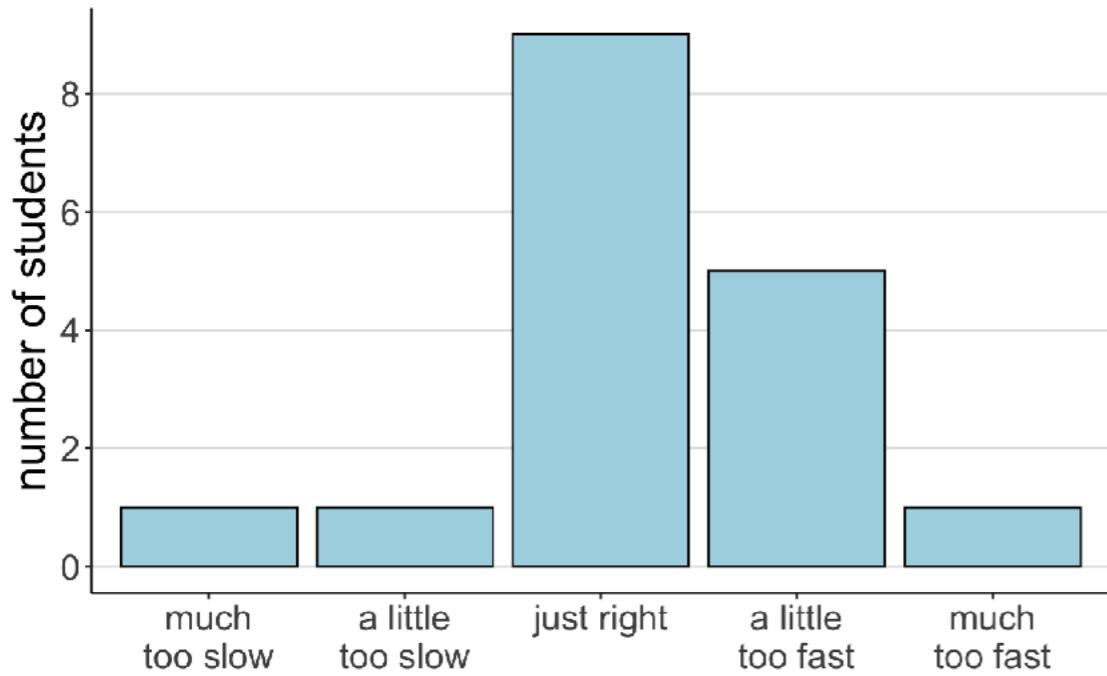


sad but probably true

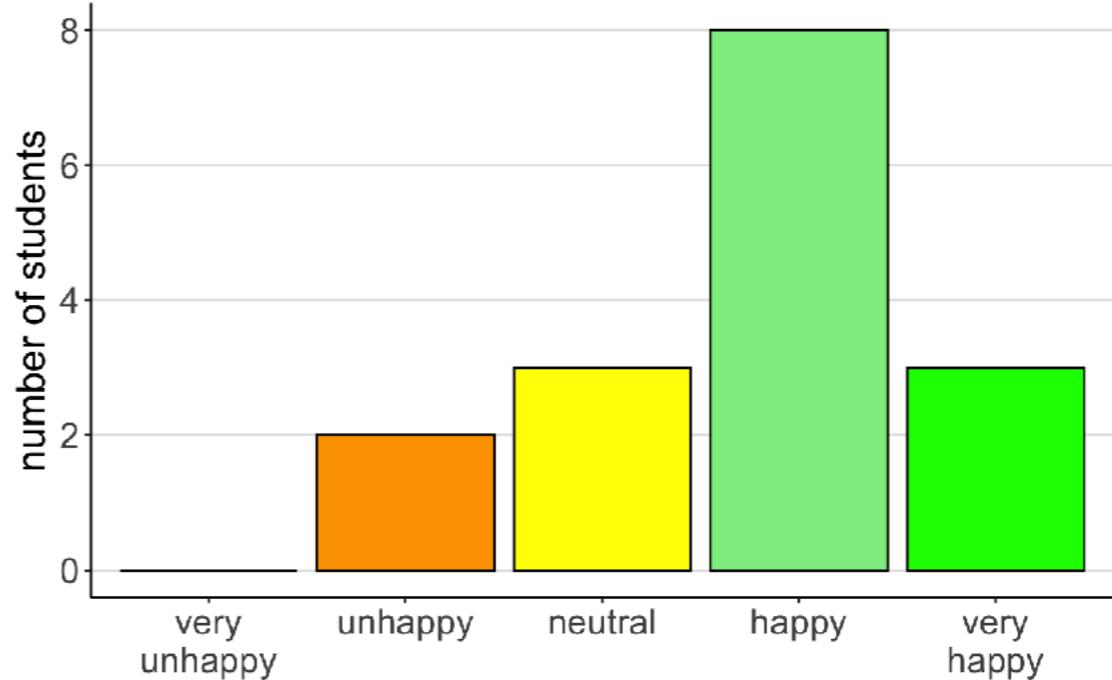
Your feedback

Your feedback

How was the pace of today's class?



How happy were you with today's class overall?



Could you review bootstrapping once more at the beginning of next class? Or could we have time in groups/breakout rooms to work through a quick example of bootstrapping?

I very much appreciate the review of prior materials at the beginning of the class. Could you please walk through code more slowly. I am not sure why and when we use "{}"

please release the homework before the classes that will review the material (right now it's released after the classes that contain the relevant material), this allows us to have a framework for applying the material. Thank you!

I'm excited we're finally getting to modeling!

Plan for today

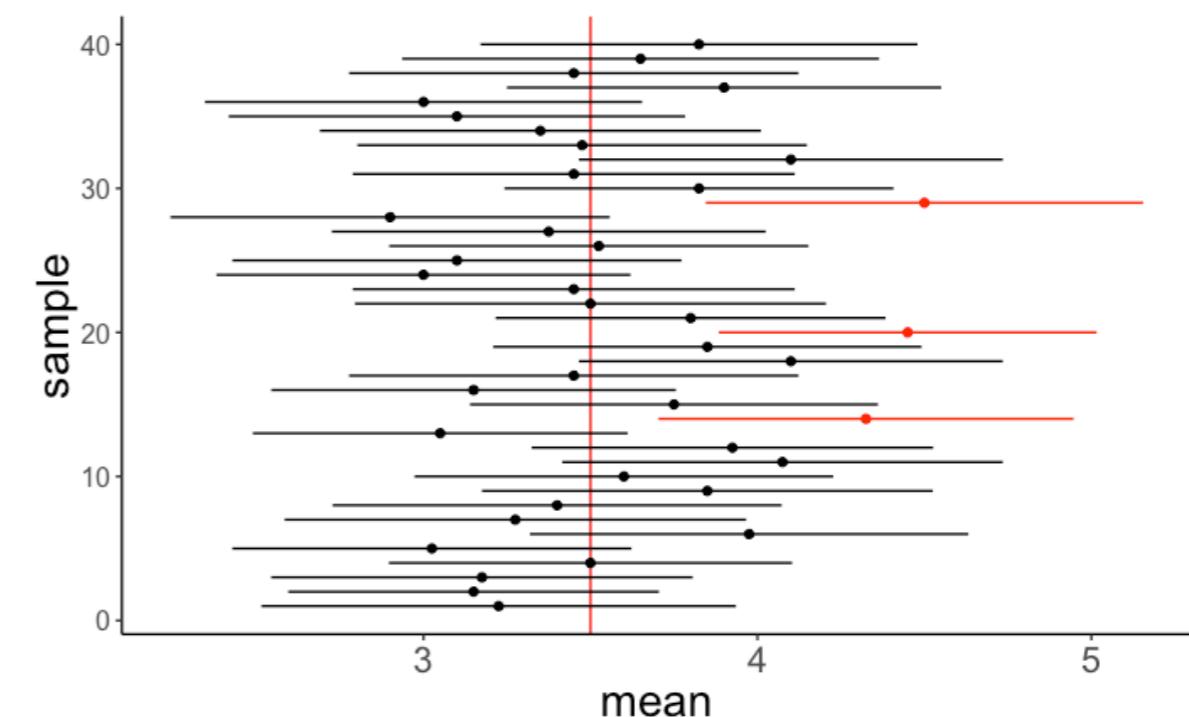
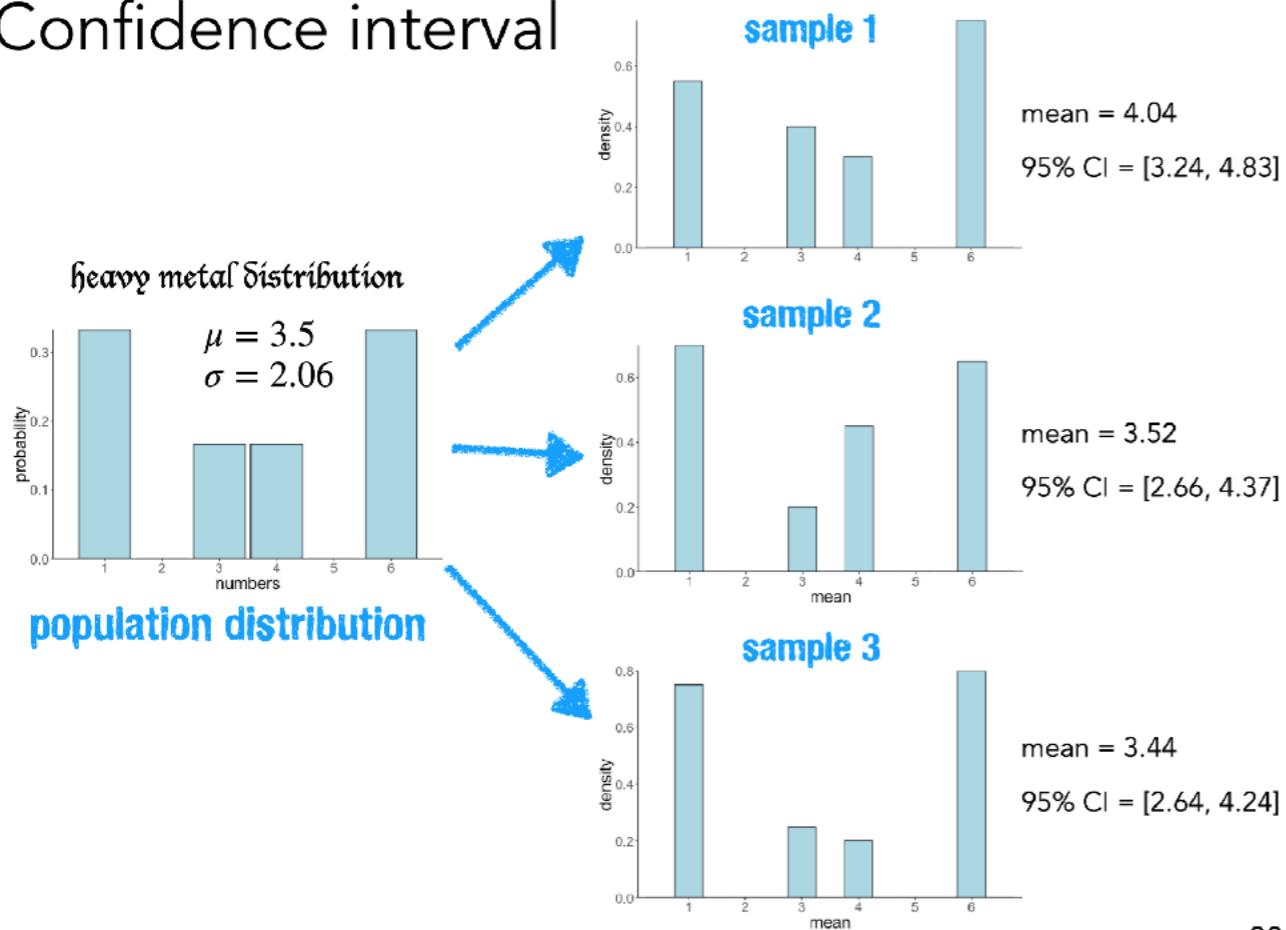
- Quick recap
- Modeling data
- Correlation
 - Pearson's moment correlation
 - Spearman's rank correlation
- Regression
 - The conceptual tour
 - The R route

Quick recap

Quick recap: Confidence intervals

"If we were to repeat the experiment over and over, then 95% of the time the confidence intervals contain the estimate of interest."

Confidence interval

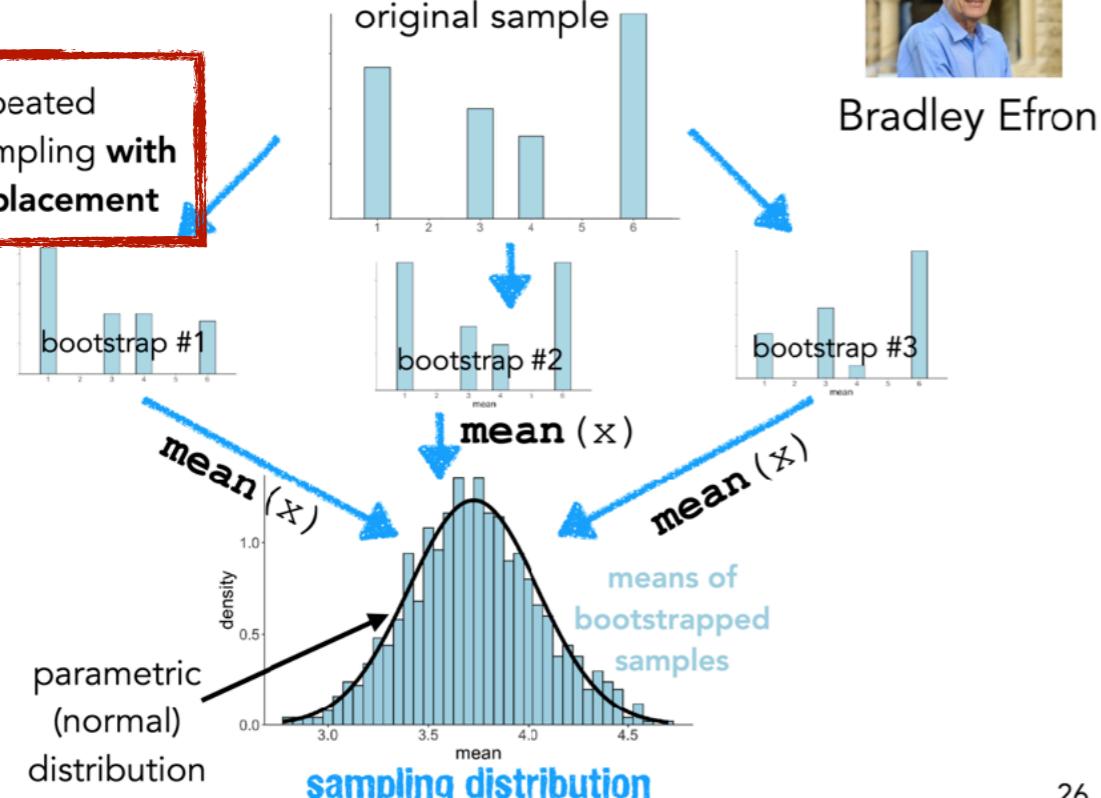


20

Quick recap: Bootstrapping

Bootstrap

repeated sampling **with replacement**

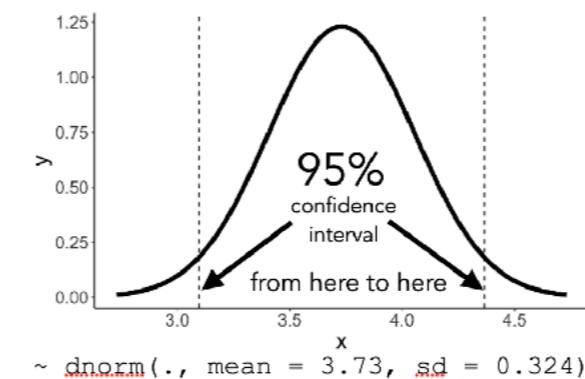


Bootstrap

How can I get the confidence interval of a statistical estimate (such as the mean)?

make assumptions

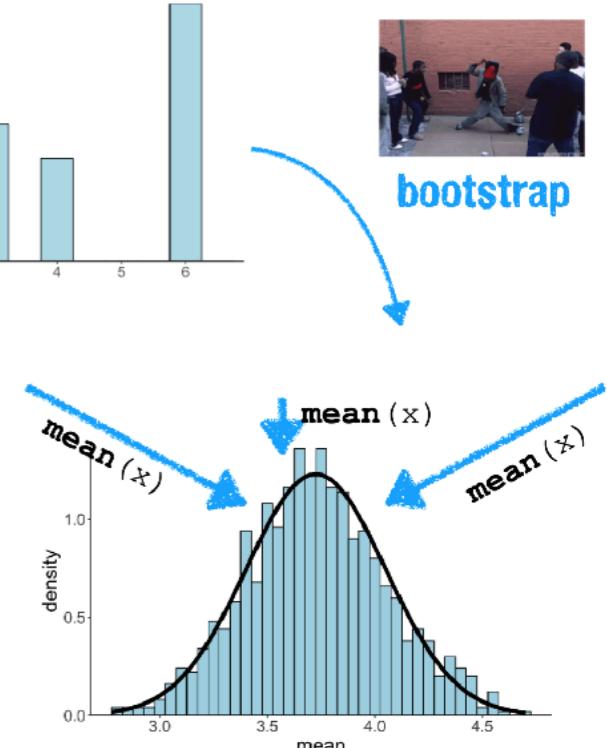
sampling distribution of the mean



26



Bradley Efron



27

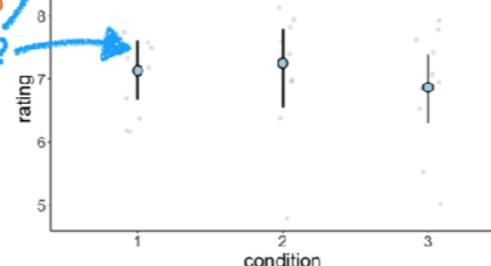
mean_cl_boot() explained

```

1 set.seed(1)
2
3 n = 10 # sample size per group
4 k = 3 # number of groups
5
6 df.data = tibble(participant = 1:(n*k),
7   condition = as.factor(rep(1:k, each = n)),
8   rating = rnorm(n*k, mean = 7, sd = 1))
9
10 ggplot(data = df.data,
11   mapping = aes(x = condition,
12     y = rating)) +
13   geom_point(alpha = 0.1,
14     position = position_jitter(width = 0.1, height = 0)) +
15   stat_summary(fun.data = "mean_cl_boot",
16     shape = 21,
17     size = 1,
18     fill = "lightblue")
  
```

what is this magic?

participant	condition	rating
1	1	6.37
2	1	7.18
3	1	6.16
4	1	6.50
5	1	7.33



29

9

Quick recap: Modeling data

$$\text{Data} = \text{Model} + \text{Error}$$

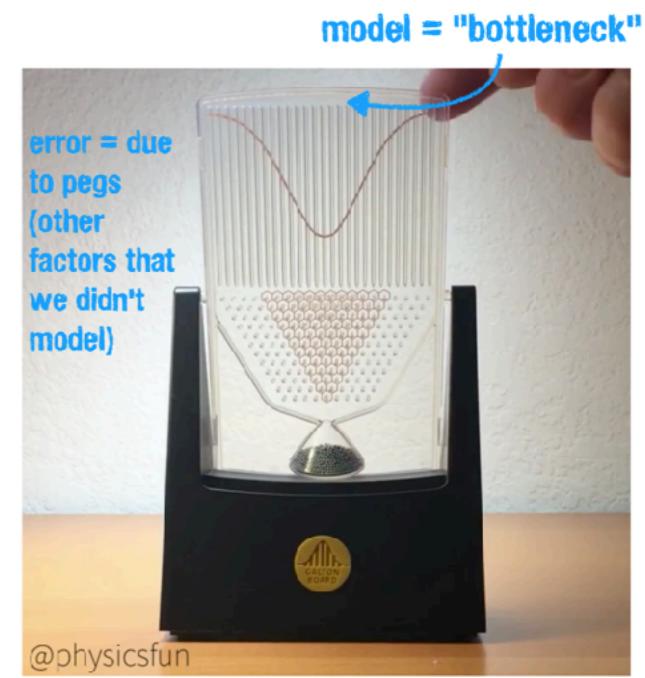
↑
**what makes for
a good model?**

- we build models with parameters, and fit those parameters to **minimize error**
- adding additional parameters to the model will always improve the model fit and reduce error
- fundamental trade-off between **simplicity** and **accuracy**

ERROR

1. We assume that the error between model and data is due to (a potentially large number of) factors that we didn't take into account.

2. We assume that each of these factors influences the data in **an additive way** (some pulling in one, others pulling in another direction).



Result: normal distribution

46

Assumption of normal distribution

$$\text{Error} = \text{Data} - \text{Model}$$

↑
**assumed to be
normally
distributed**

↑
**don't need to
be normally
distributed!!**

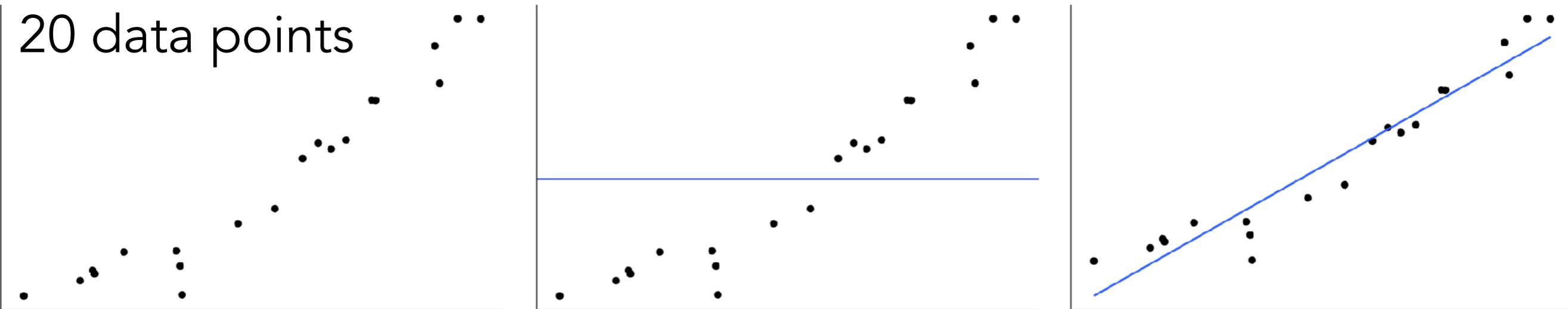
very common misconception!!!

**I've added some
resources on Ed
Discussion**

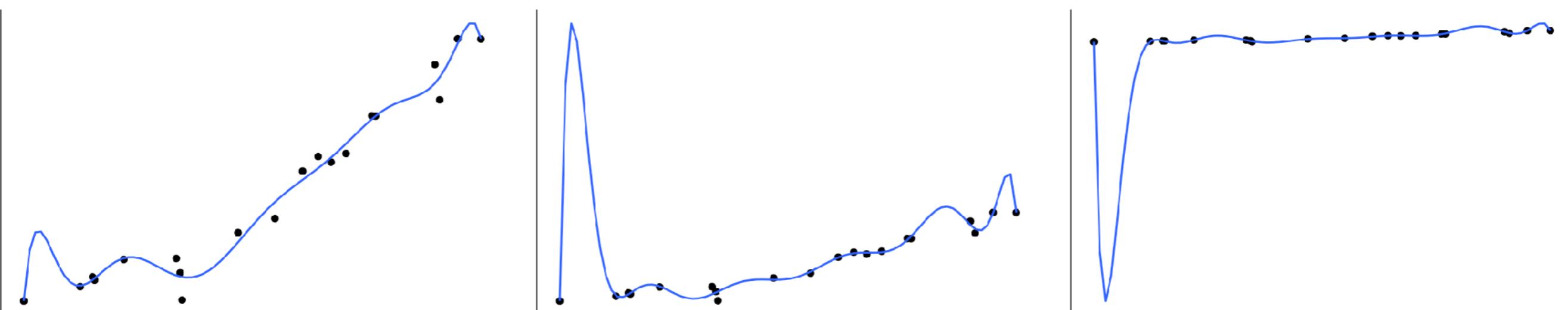
10

Modeling data

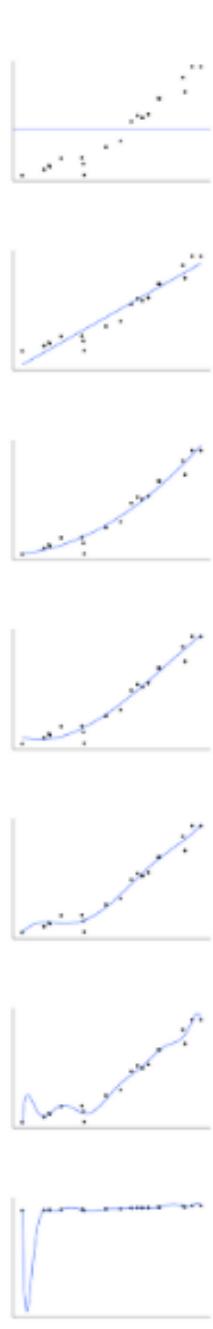
20 data points



Which model describes the data best?



Which model describes the data best

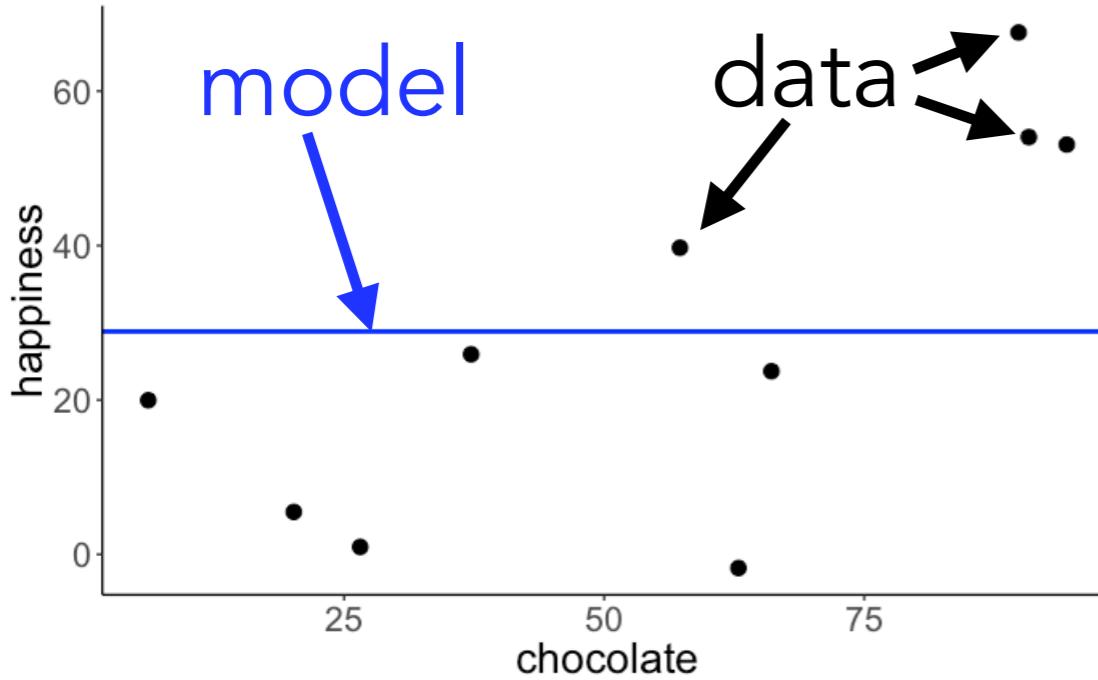




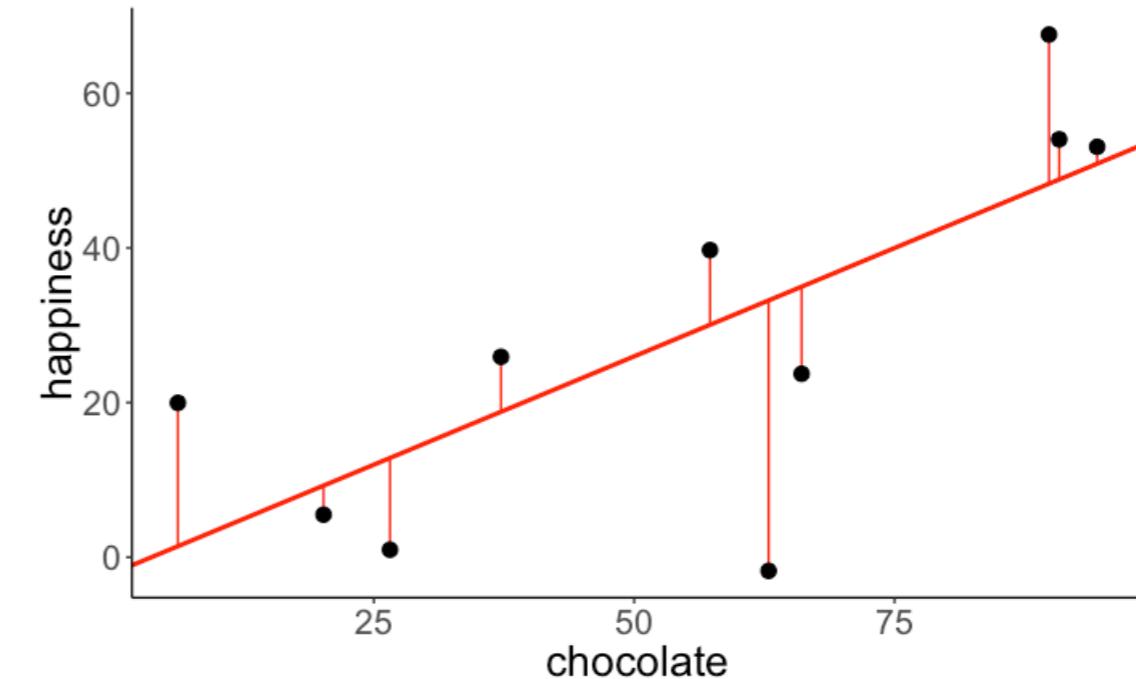
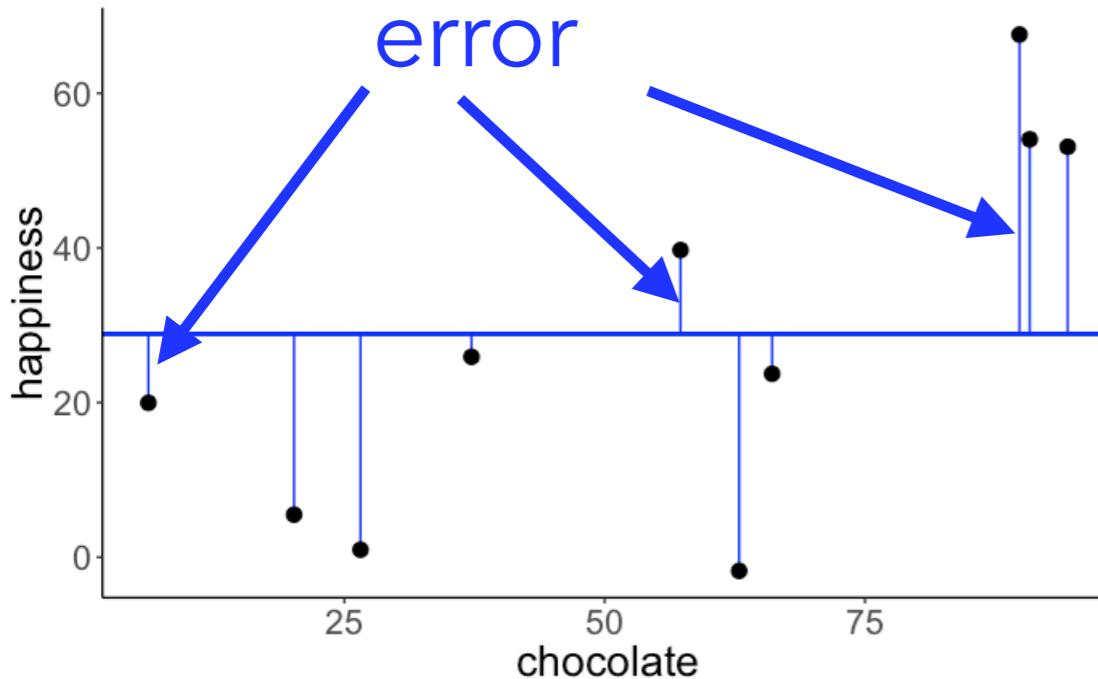
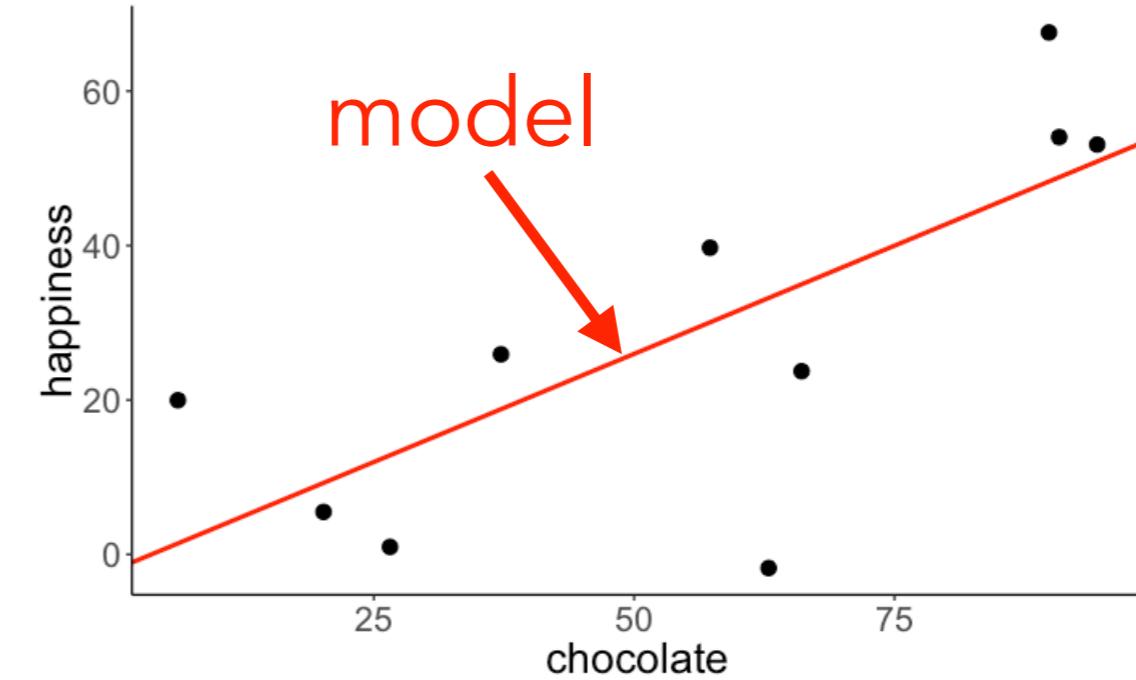


Data = Model + Error

H_0 : Chocolate consumption and happiness are unrelated.



H_1 : Chocolate consumption and happiness are related.



Example

Percentage of households that had internet access in the year 2013 by US state

i	internet	state	college	auto	density
1	79.0	AK	28.0	1.2	1.2
2	63.5	AL	23.5	1.3	94.4
3	60.9	AR	20.6	1.7	56.0
4	73.9	AZ	27.4	1.3	56.3
5	77.9	CA	31.0	0.8	239.1
6	79.4	CO	37.8	1.0	48.5
7	77.5	CT	37.2	1.0	738.1
8	74.5	DE	29.8	1.1	460.8
9	74.3	FL	27.2	1.2	350.6
10	72.2	GA	28.3	1.1	168.4

set before looking at the data

$$\text{model}_1: Y_i = 75 + \text{ERROR}$$

worth it?

fit to the data

$$\text{model}_2: Y_i = \beta_0 + \text{ERROR}$$

worth it?

additional predictor

$$\text{model}_3: Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$$

fit to the data

Compact model

model_C: $Y_i = \beta_0 + \text{ERROR}$

Augmented model

model_A: $Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$

$$\text{ERROR}(C) \geq \text{ERROR}(A)$$

Proportional reduction in error (PRE)

$$\text{PRE} = \frac{\text{ERROR}(C) - \text{ERROR}(A)}{\text{ERROR}(C)}$$

Compact model

$$\text{model}_C: Y_i = \beta_0 + \text{ERROR}$$

$$\text{ERROR}(C) = 50$$

Augmented model

$$\text{model}_A: Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$$

$$\text{ERROR}(A) = 30$$

Proportional reduction in error (PRE)

$$\begin{aligned} \text{PRE} &= 1 - \frac{\text{ERROR}(A)}{\text{ERROR}(C)} \\ &= 1 - \frac{30}{50} = .40 \end{aligned}$$

Increasing the complexity of the model reduced the error by 40%. **worth it?**

worth it?

Compact model

model_C: $Y_i = \beta_0 + \text{ERROR}$

Augmented model

model_A: $Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$

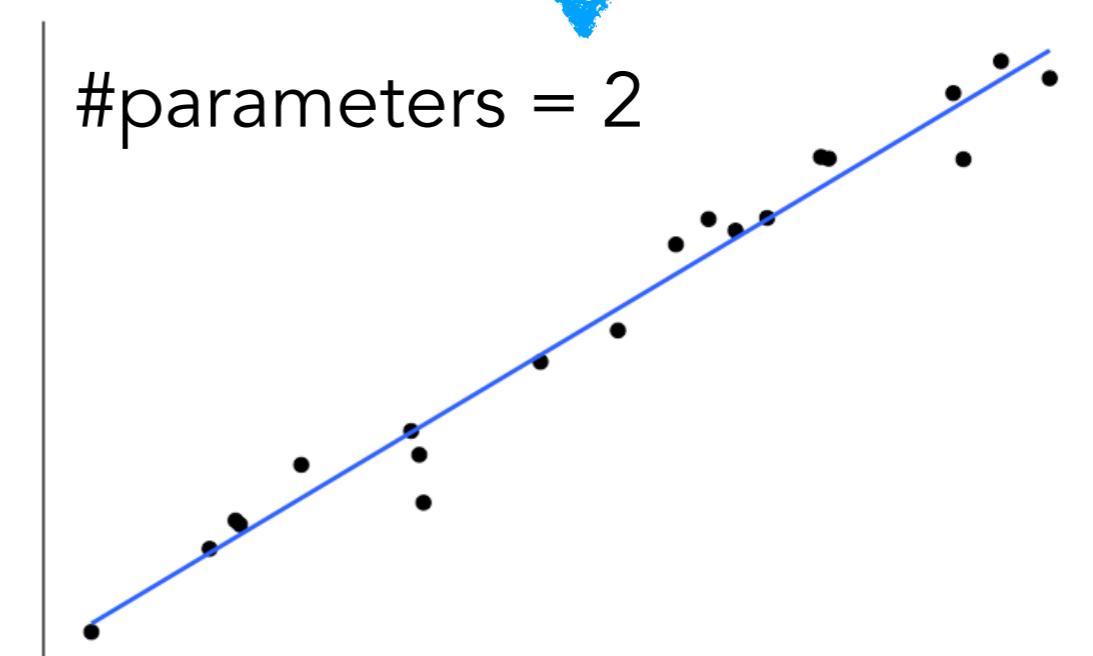
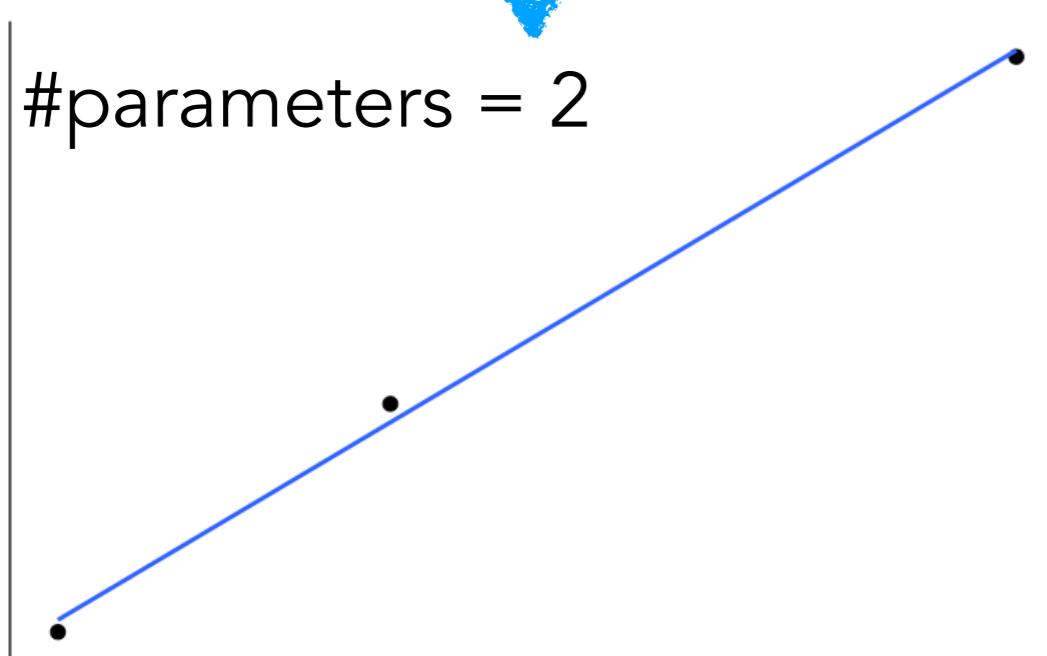
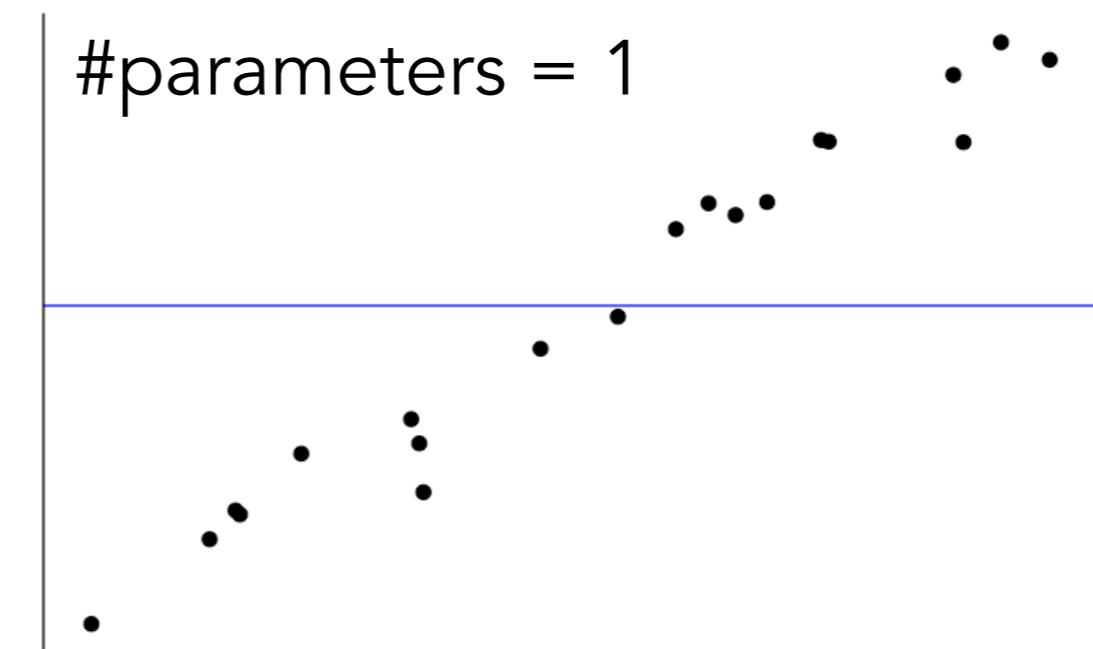
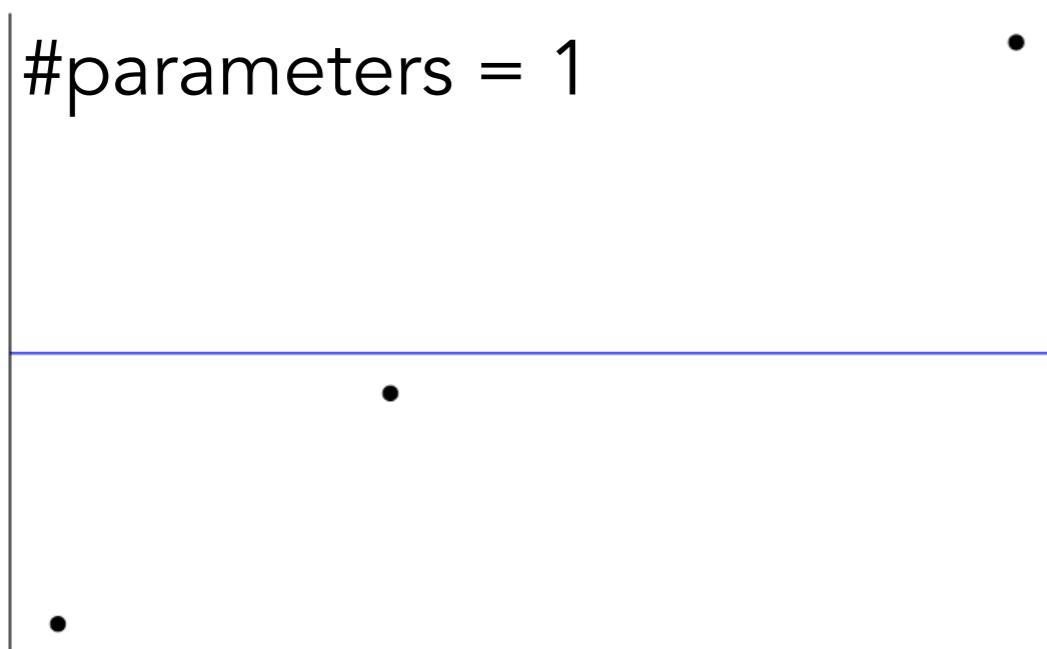
Proportional reduction in error (PRE)

$$\begin{aligned}\text{PRE} &= 1 - \frac{\text{ERROR}(A)}{\text{ERROR}(C)} \\ &= 1 - \frac{30}{50} = .40\end{aligned}$$

- to answer the **worth it?** question we need inferential statistics (define ERROR, sampling distributions, etc.)
- more likely to be **worth it** if:
 1. **PRE** is high
 2. the number of additional parameters in A compared to C is low (**PRE per additional parameter**)
 3. the number of parameters that could have been added to model_C to create model_A but were not is high

more impressed if the number of observations n is much greater than the number of parameters

PRE per parameter for different n



neato!

impressive!

General procedure

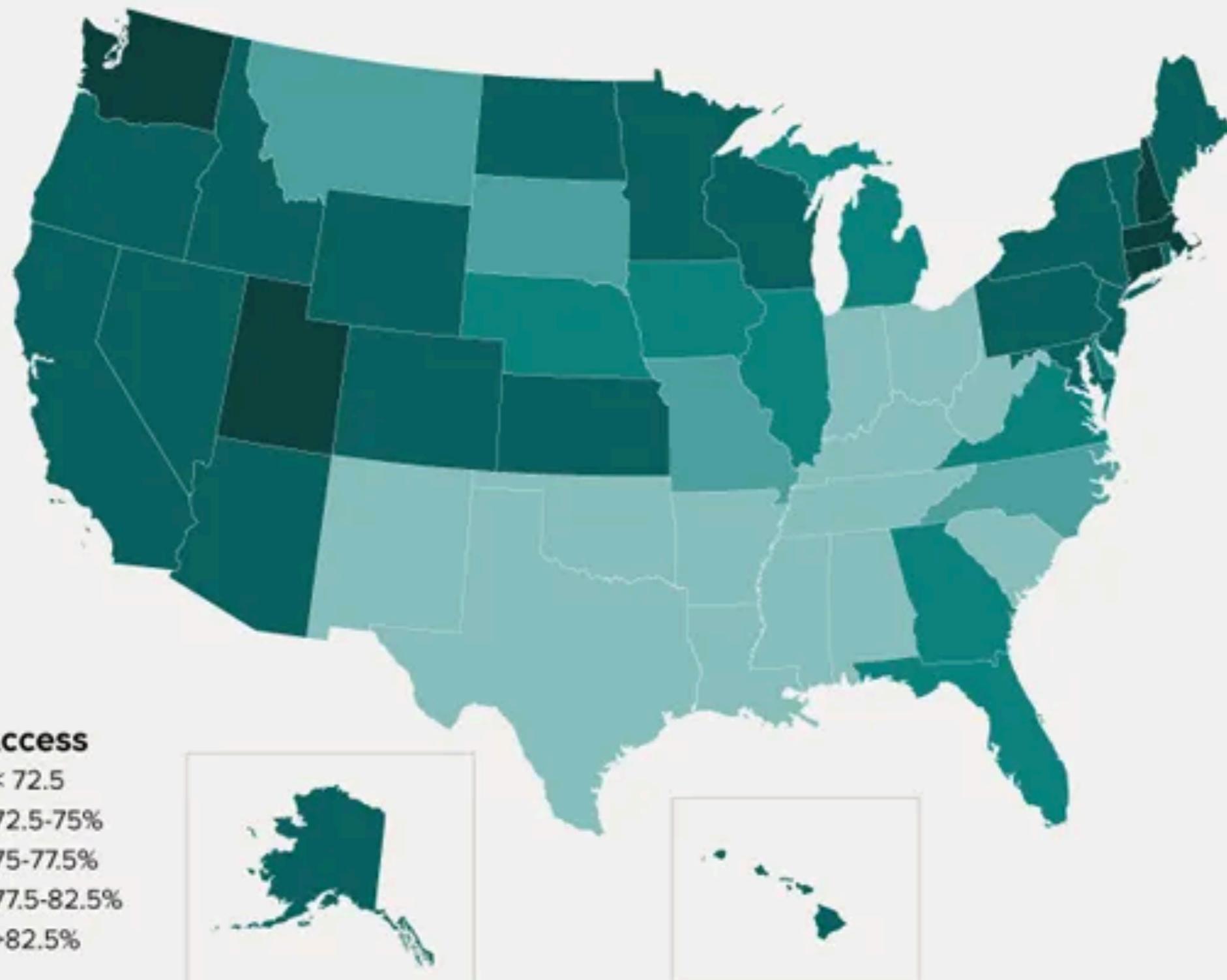
- for any question we want to ask about our DATA
 - we define model_C and model_A
 - compare the models using PRE
 - determine whether PRE is **worth it**
 - in standard frequentist lingo:
 - model_C = H_0 (null hypothesis) 
 - model_A = H_1 (alternative hypothesis) 
 - hypothesis test:
 - H_0 : **all** the parameters that are included in model_A but not in model_C are 0
 - H_1 : **not all** the parameters that are included in model_A but not in model_C are 0
- model comparison**

Statistical inferences about parameter values

Procedure

1. Start with research question
2. Formulate hypothesis as a comparison between compact and augmented model
3. Fit parameters in each model
4. Calculate the proportional reduction of error (PRE)
5. Decide whether PRE is **worth it**

Internet Access At Home



<http://thedataviz.com/2012/07/19/mapping-internet-access-in-u-s-homes/>

Notation

DATA = MODEL + ERROR

$$Y_i = \beta_0 + \epsilon_i \text{ simple model (true parameters)}$$

$$Y_i = b_0 + e_i \text{ simple model (estimated parameters)}$$

$$\hat{Y}_i = b_0$$

density

$$Y_i = b_0 + b_1 X_{i1} + e_i \text{ more complex model}$$

Percentage of households that had internet access in the year 2013 by US state

i	internet	state	college	auto	density
1	79.0	AK	28.0	1.2	1.2
2	63.5	AL	23.5	1.3	94.4
3	60.9	AR	20.6	1.7	56.0
4	73.9	AZ	27.4	1.3	56.3
5	77.9	CA	31.0	0.8	239.1
6	79.4	CO	37.8	1.0	48.5
7	77.5	CT	37.2	1.0	738.1
8	74.5	DE	29.8	1.1	460.8
9	74.3	FL	27.2	1.2	350.6
10	72.2	GA	28.3	1.1	168.4



Greek letters β or ϵ represent the true but unknowable parameters in the population.

Roman letters b or e represent estimates of these parameters using our DATA.

Research question and hypotheses

Is the average percentage of internet users per state significantly different from 75%?

Model_C: $Y_i = B_0 + \epsilon_i$

0 parameters

$$Y_i = 75 + e_i$$

Model_A: $Y_i = \beta_0 + \epsilon_i$

1 parameter

$$Y_i = b_0 + e_i$$

$$= \bar{Y} + e_i$$

i	internet	state	college	auto	density
1	79.0	AK	28.0	1.2	1.2
2	63.5	AL	23.5	1.3	94.4
3	60.9	AR	20.6	1.7	56.0
4	73.9	AZ	27.4	1.3	56.3
5	77.9	CA	31.0	0.8	239.1
6	79.4	CO	37.8	1.0	48.5
7	77.5	CT	37.2	1.0	738.1
8	74.5	DE	29.8	1.1	460.8
9	74.3	FL	27.2	1.2	350.6
10	72.2	GA	28.3	1.1	168.4

Fit parameters and calculate PRE

$$\mathbf{C: } Y_i = 75 + e_i \quad \mathbf{A: } Y_i = \bar{Y} + e_i$$

i	state	internet	compact_b	compact_se	augmented_b	augmented_se
1	AK	79.0	75	16.00	72.81	38.37
2	AL	63.5	75	132.25	72.81	86.60
3	AR	60.9	75	198.81	72.81	141.75
4	AZ	73.9	75	1.21	72.81	1.20
5	CA	77.9	75	8.41	72.81	25.95
6	CO	79.4	75	19.36	72.81	43.48
7	CT	77.5	75	6.25	72.81	22.03
8	DE	74.5	75	0.25	72.81	2.87
9	FL	74.3	75	0.49	72.81	2.23
10	GA	72.2	75	7.84	72.81	0.37

$$\begin{aligned}\text{PRE} &= 1 - \frac{\text{ERROR}(A)}{\text{ERROR}(C)} \\ &= 1 - \frac{1355}{1595} \approx .15\end{aligned}$$

Model A has
15% less error
than Model C.

$$\text{SSE(C)} = 1595 \quad \text{SSE(A)} = 1355$$

Decide whether it's **worth it**

- PRE is the estimate of an unknown true reduction of error η^2
- we need a sampling distribution of PRE
 - a distribution of what PRE would look like if Model C (our H_0) were true
 - we could just simulate such a sampling distribution ...
- PRE is closely related to the F statistic!

Decide whether it's **worth it**

- To compute the F statistic, we need:

- PRE
- number of parameters in Model C (PC) and Model A (PA)
- number of observations n

- more likely to be **worth it** if:
 1. PRE is high
 2. the number of additional parameters in A compared to C is low (**PRE per additional parameter**)
 3. the number of parameters that could have been added to model_C to create model_A but were not

**difference in parameters
between models A and C**

$$F = \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})}$$

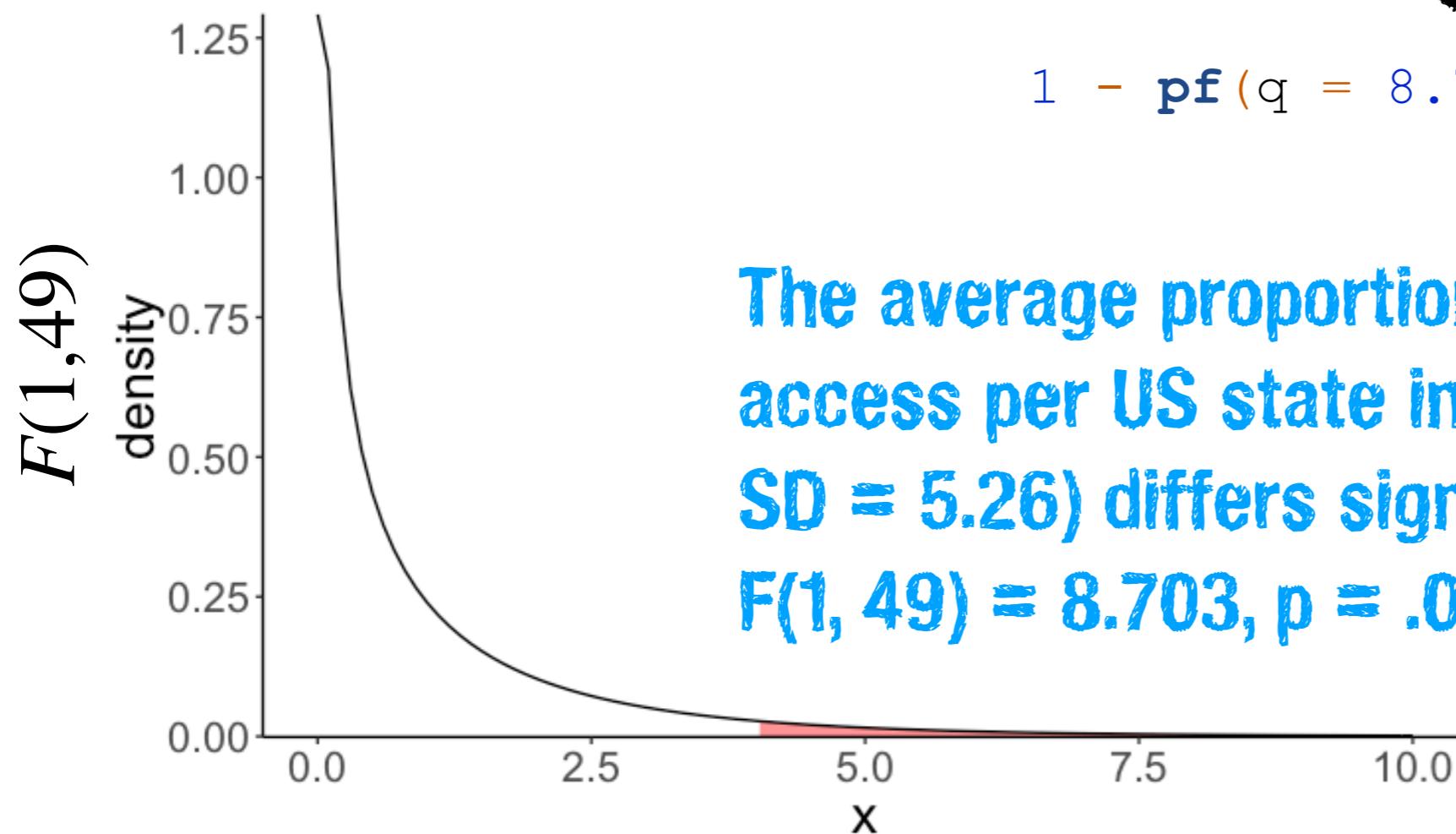


**number of observations
vs. parameters in Model A**

Decide whether it's **worth it**

$$F = \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})}$$

$$= \frac{.15/(1 - 0)}{(1 - .15)/(50 - 1)} = 8.703 \quad p = .00486$$



Note: I've used the approximated PRE here (PRE = 0.15), but I'm reporting the F value for the non-approximated PRE value.

we just performed a one sample t-test ...

```
t.test(df.internet$internet, mu = 75)
```

One Sample t-test

```
data: df.internet$internet  
t = -2.9502, df = 49, p-value = 0.00486  
alternative hypothesis: true mean is not equal to 75
```

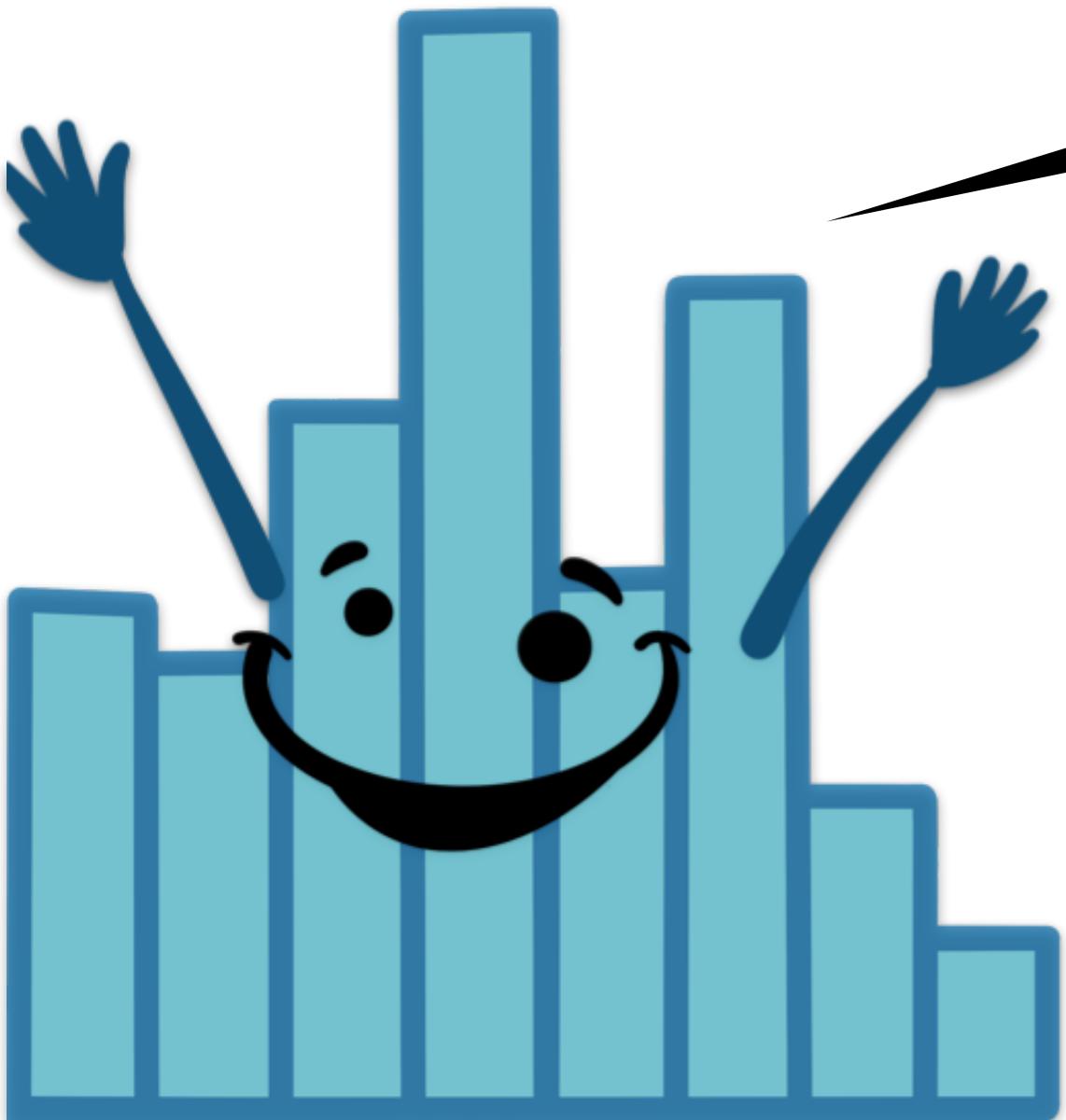
but ...

- we will apply the general procedure we went through to day to a large range of different situations
- thinking of hypothesis testing as model comparison is (hopefully more) intuitive
- this approach still works even if we can't find a recipe in our favorite stats cook book

We're listening to
"Everything's Good" by
"Phil Good" submitted
by Sarah Wu

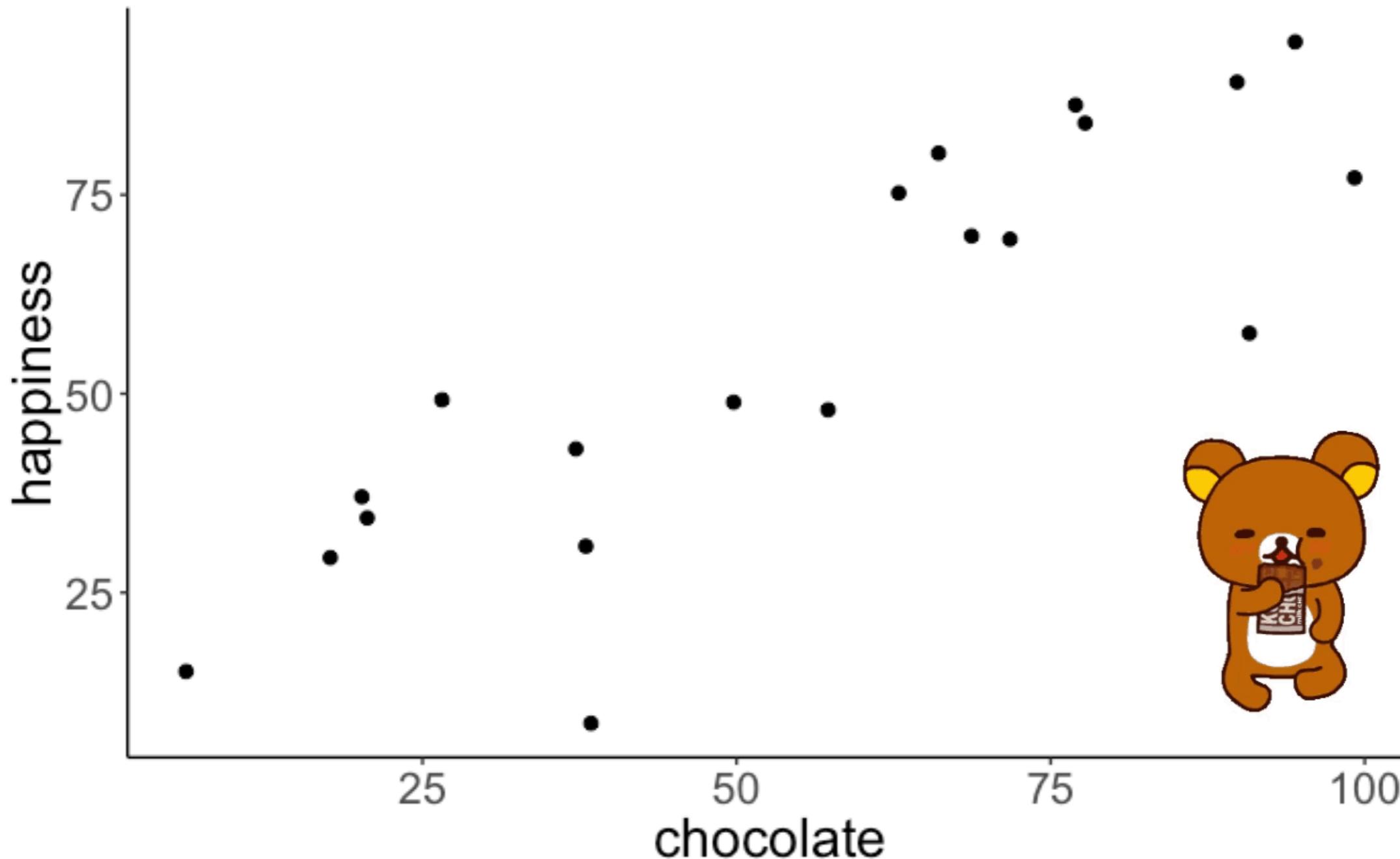
02:00

stretch break!



Correlation

How to best characterize the relationship between x and y by a single number?

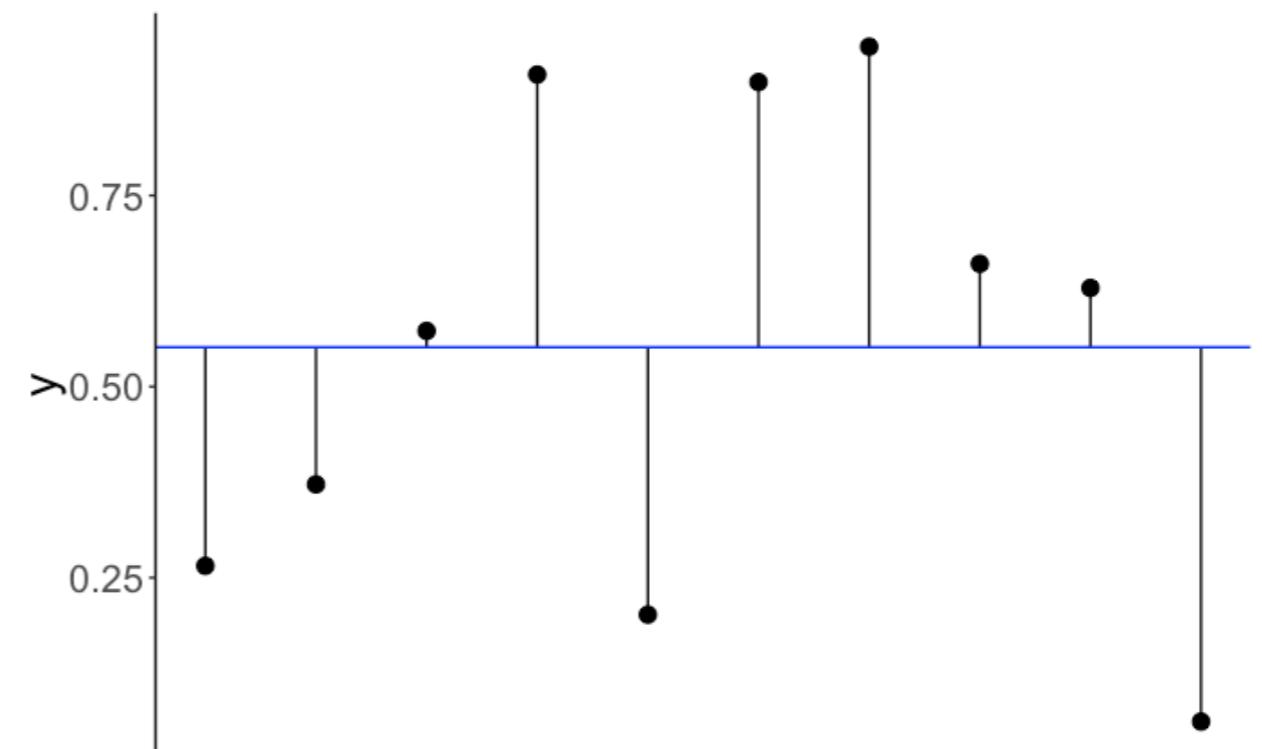
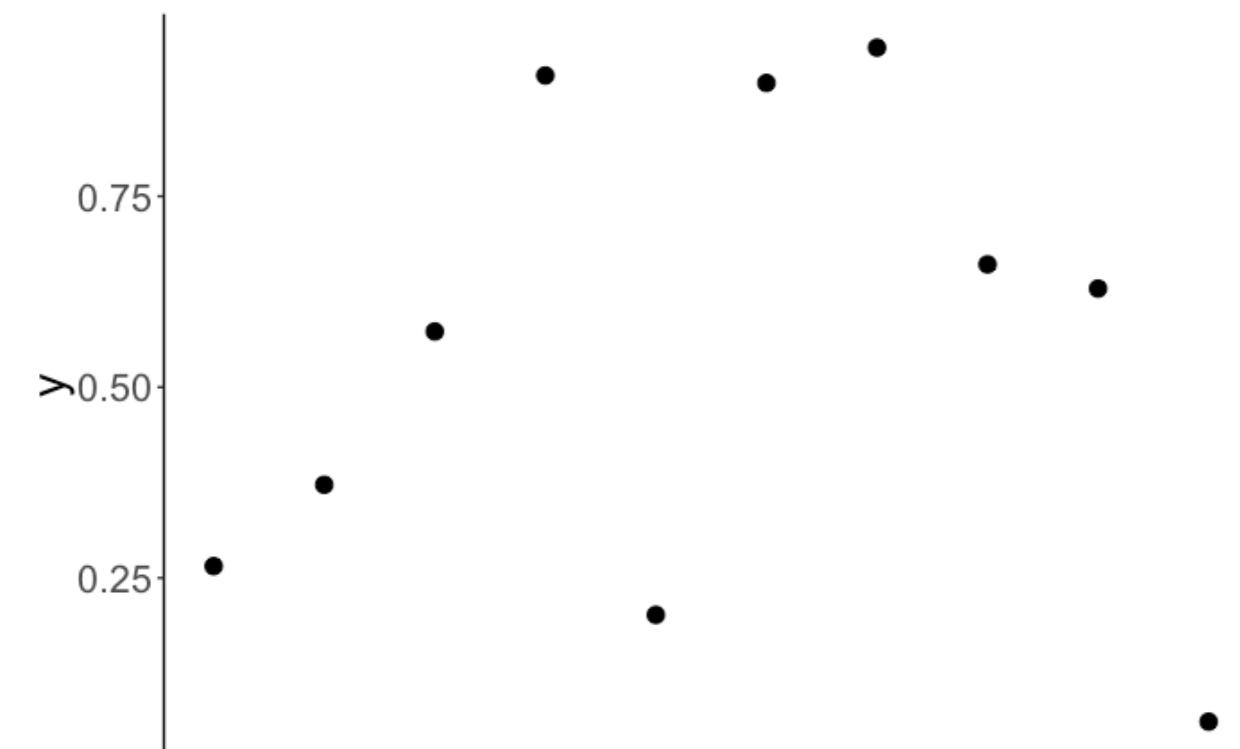


correlation = a measure of the relationship
between two variables

sample variance

$$Var(Y) = \sum_{i=1}^n \frac{(Y_i - \bar{Y})^2}{n - 1}$$

sum of squared errors



(I was too lazy to draw rectangles ...)

How well does the mean capture the data?

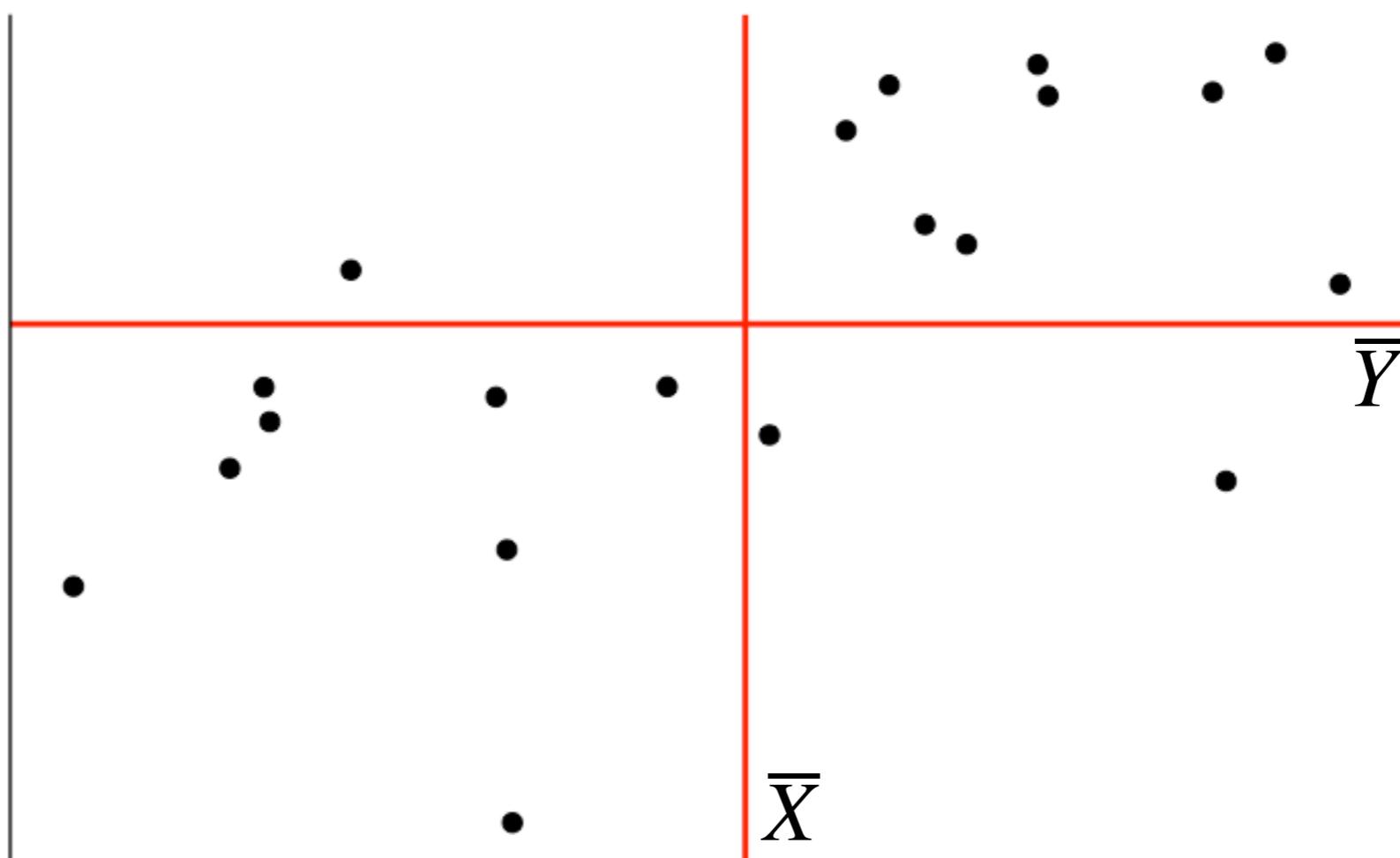
sample covariance

$$Cov(X, Y) = \sum_{i=1}^n \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$



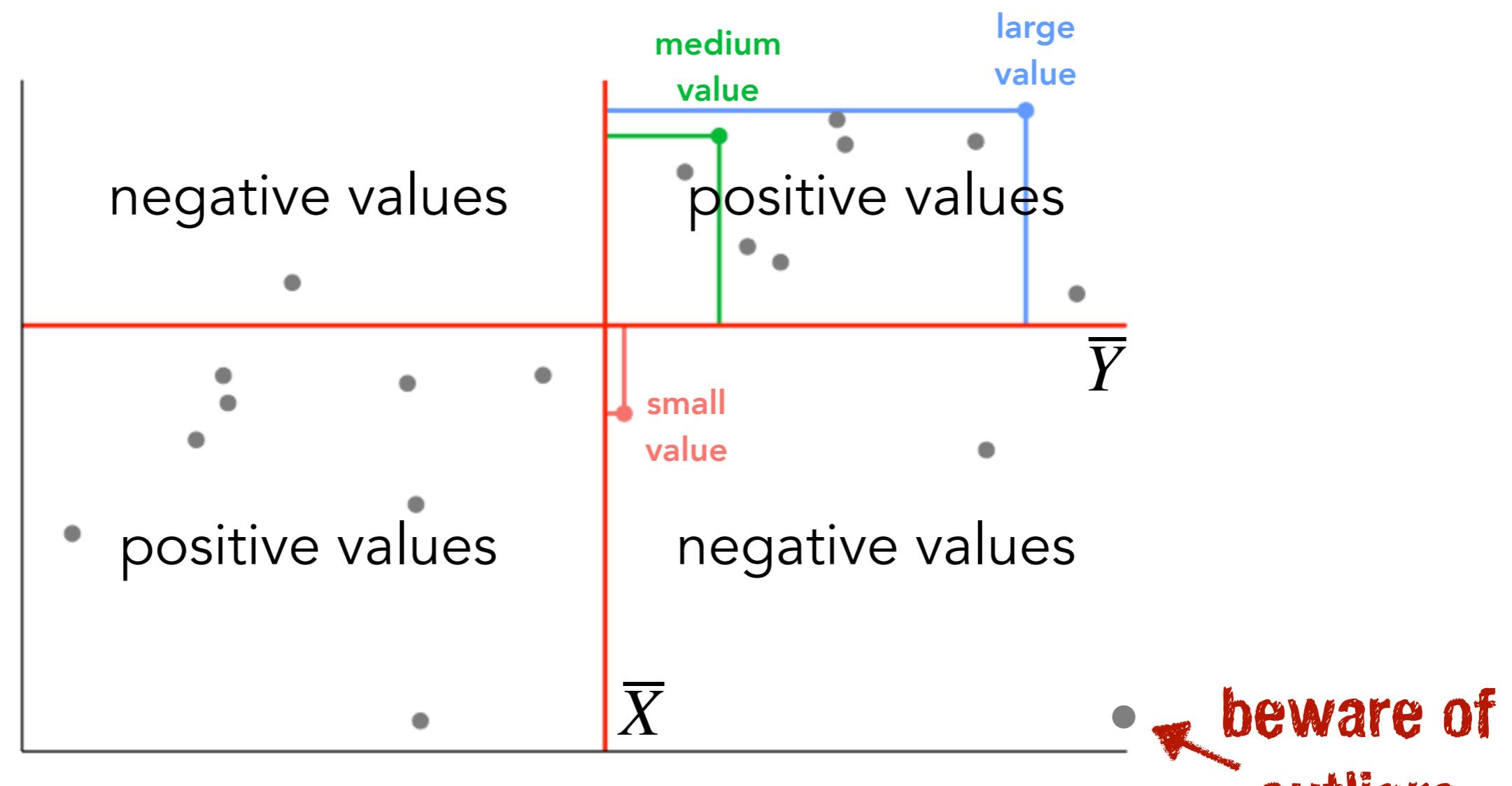
sample covariance

$$Cov(X, Y) = \sum_{i=1}^n \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$



sample covariance

$$Cov(X, Y) = \sum_{i=1}^n \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$



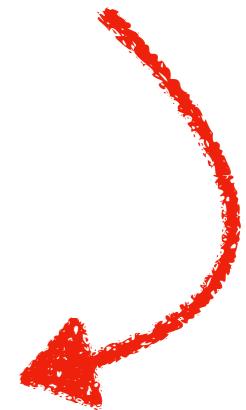
sample covariance

$$Cov(X, Y) = \sum_{i=1}^n \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

depends on the scale of the variables

the $n - 1$ s cancel out

sample correlation coefficient

$$r(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$


standardized covariation
(dividing by the standard deviations)

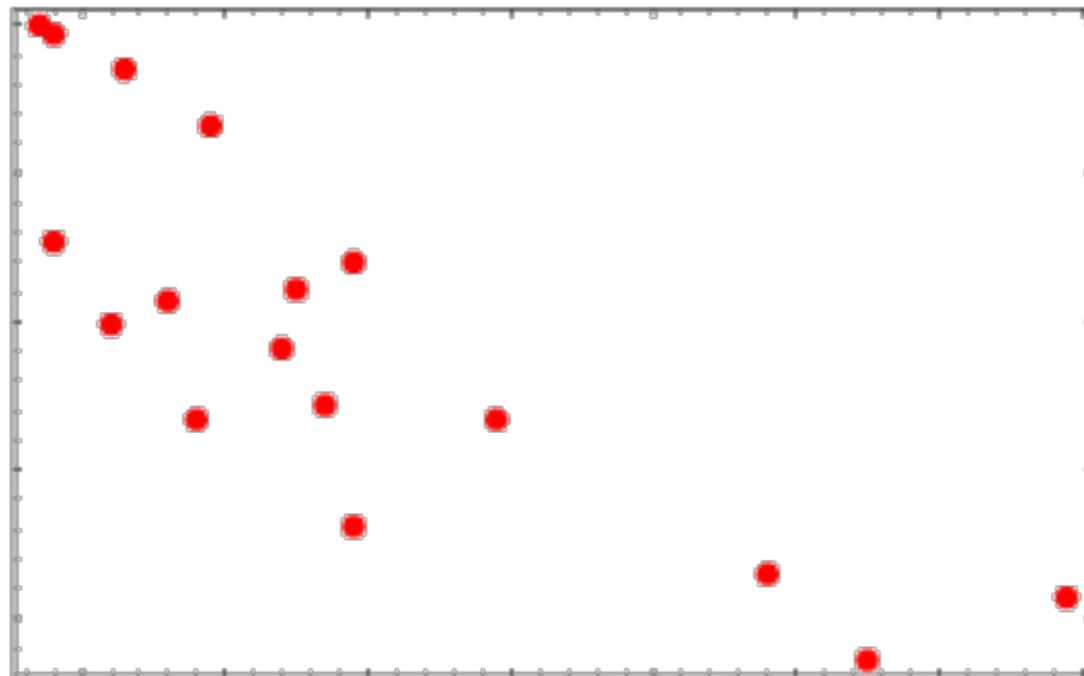
Properties of the Pearson correlation

- standardized: $-1 \leq r \leq 1$
- scale independent (for both X and Y)
- commutativity: $r(X, Y) = r(Y, X)$
- sign determines the direction of dependence
- captures **linear dependence** only

association not
causation



Who is the correlation champion?



Winner gets chocolate!

-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

-0.25 : 0

0 : 0.25

0.25 : 0.5

0.5 : 0.75

0.75 : 1

Who is the correlation champion?

Get ready to compete!

In what range is the correlation coefficient?

-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

-0.25 : 0

0 : 0.25

0.25 : 0.5

0.5 : 0.75

0.75 : 1

In what range is the correlation coefficient?

-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

-0.25 : 0

0 : 0.25

0.25 : 0.5

0.5 : 0.75

0.75 : 1

In what range is the correlation coefficient?

-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

-0.25 : 0

0 : 0.25

0.25 : 0.5

0.5 : 0.75

0.75 : 1

Leaderboard

In what range is the correlation coefficient?

-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

-0.25 : 0

0 : 0.25

0.25 : 0.5

0.5 : 0.75

0.75 : 1

In what range is the correlation coefficient?

-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

-0.25 : 0

0 : 0.25

0.25 : 0.5

0.5 : 0.75

0.75 : 1

In what range is the correlation coefficient?

-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

-0.25 : 0

0 : 0.25

0.25 : 0.5

0.5 : 0.75

0.75 : 1

Leaderboard

Widderen

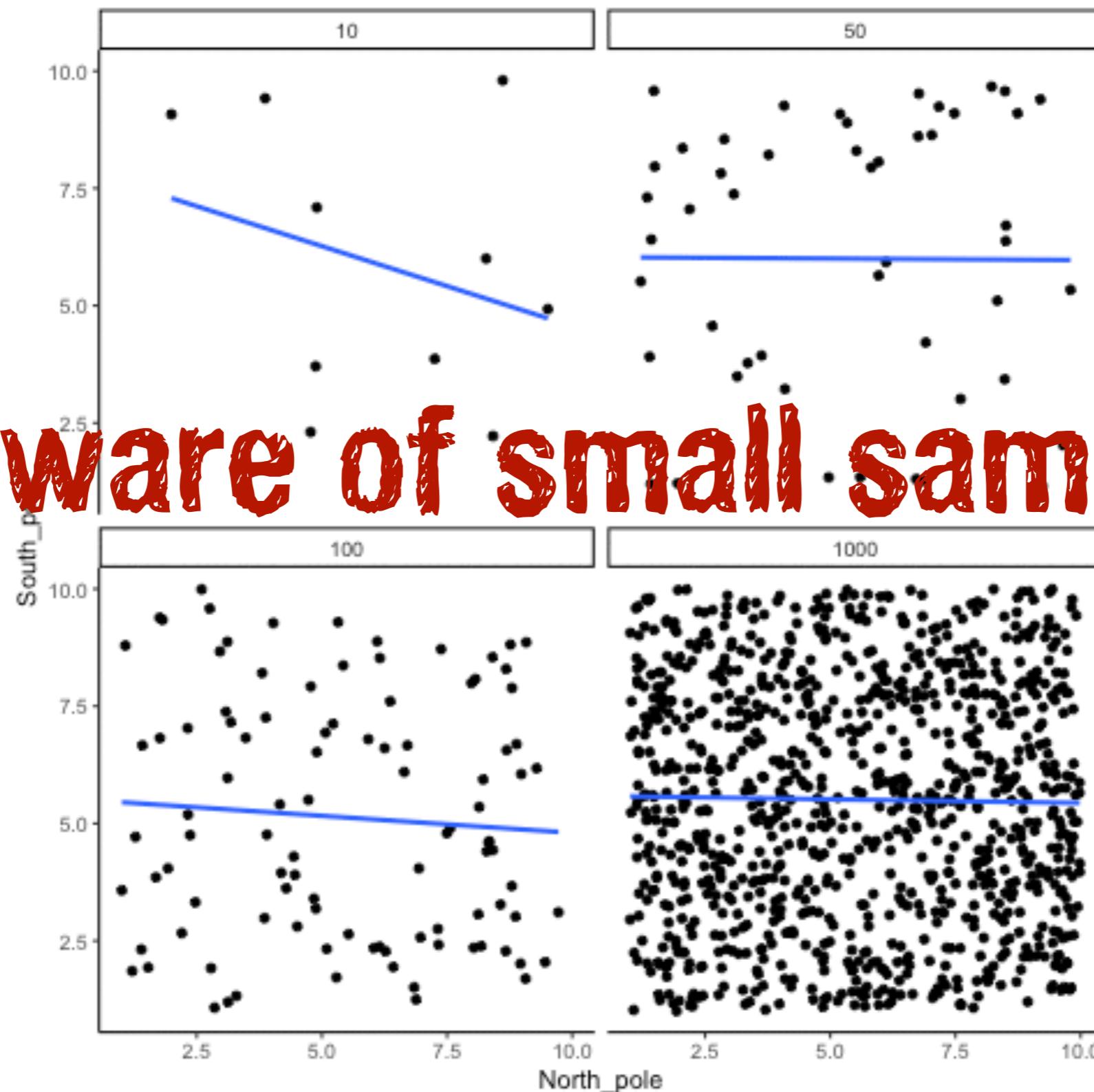
Be careful about interpreting correlations!



always visualize the data ...

$n = [10, 50, 100, 1000]$

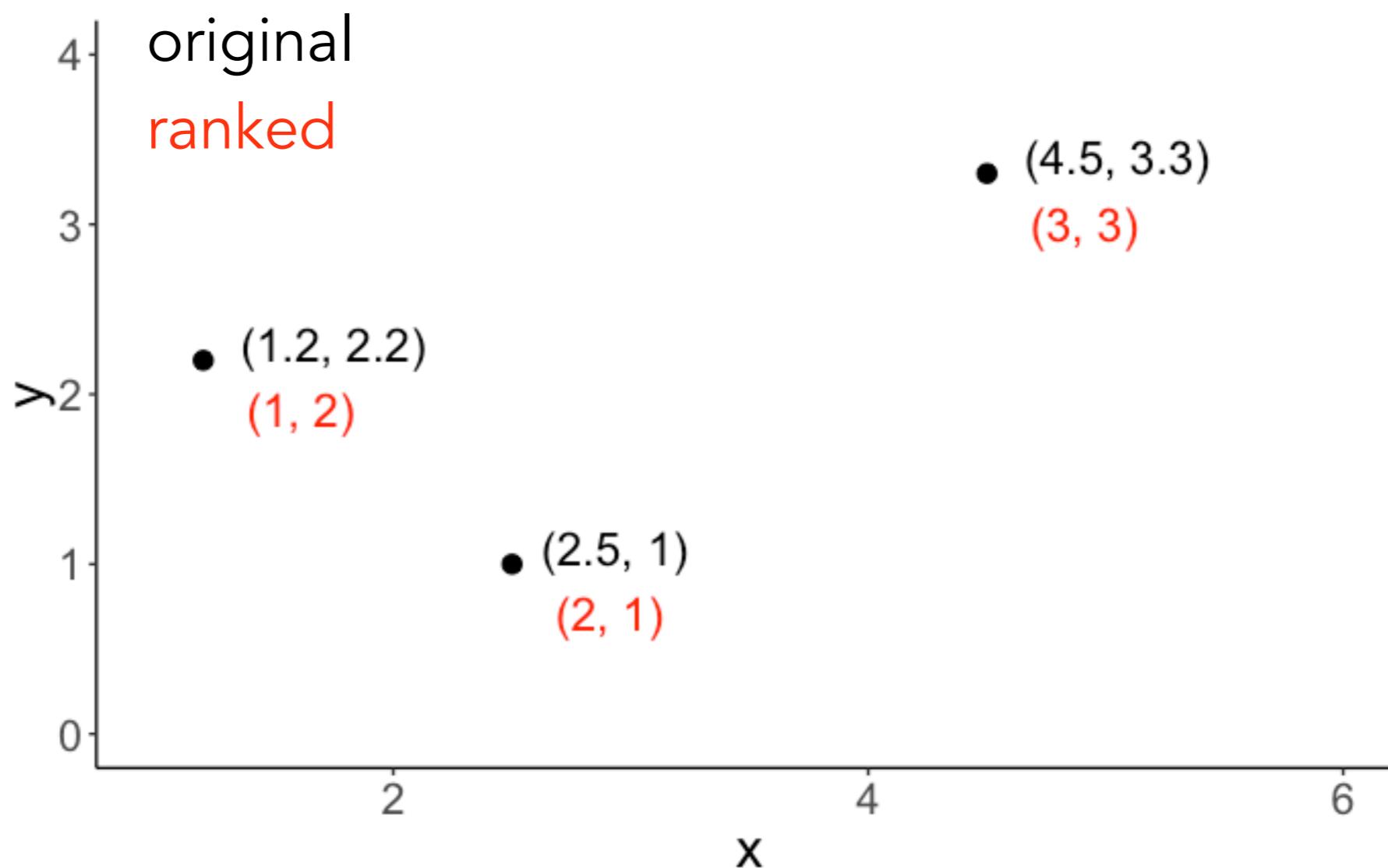
$X \sim \mathcal{U}(\min = 0, \max = 10)$
 $Y \sim \mathcal{U}(\min = 0, \max = 10)$



Beware of small samples!

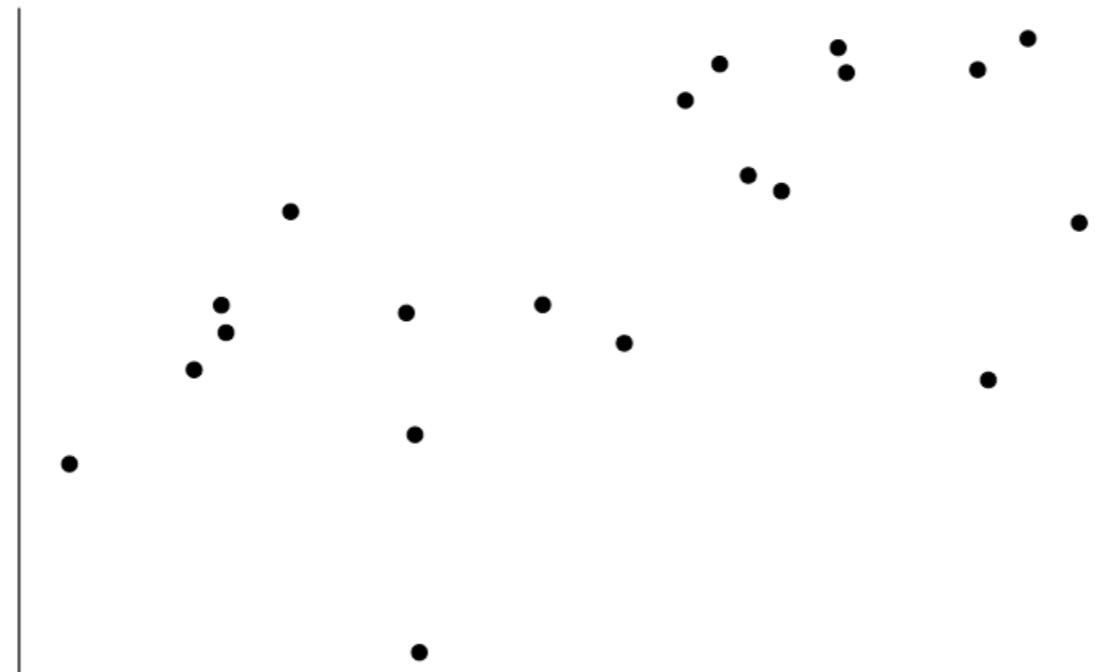
Spearman rank order correlation

- transform original data into ranks
- calculate correlation on the ranked data



Spearman rank order correlation

- transform original data into ranks
- calculate correlation on the ranked data



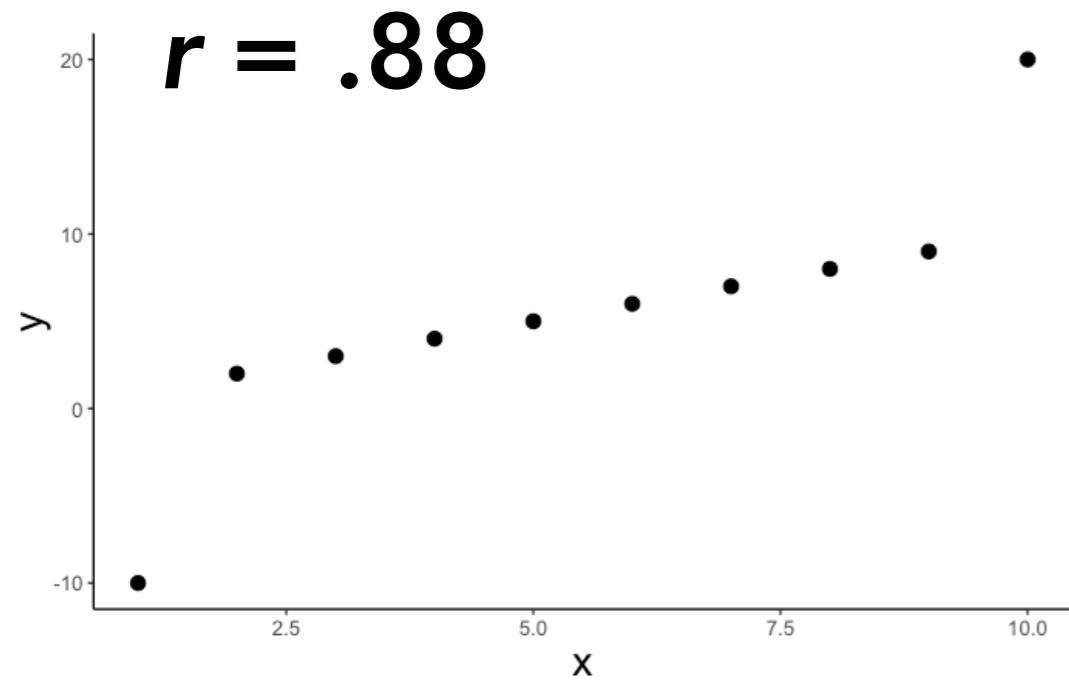
x	y	x_rank	y_rank
0.27	1.14	5	12
0.37	0.97	6	8
0.57	0.92	10	6
0.91	0.85	18	4
0.20	0.98	3	9
0.90	1.39	17	17
0.94	1.44	19	20
0.66	1.40	12	18
0.63	1.33	11	15
0.06	0.71	1	2

r	spearman	r_ranks
0.609	0.595	0.595

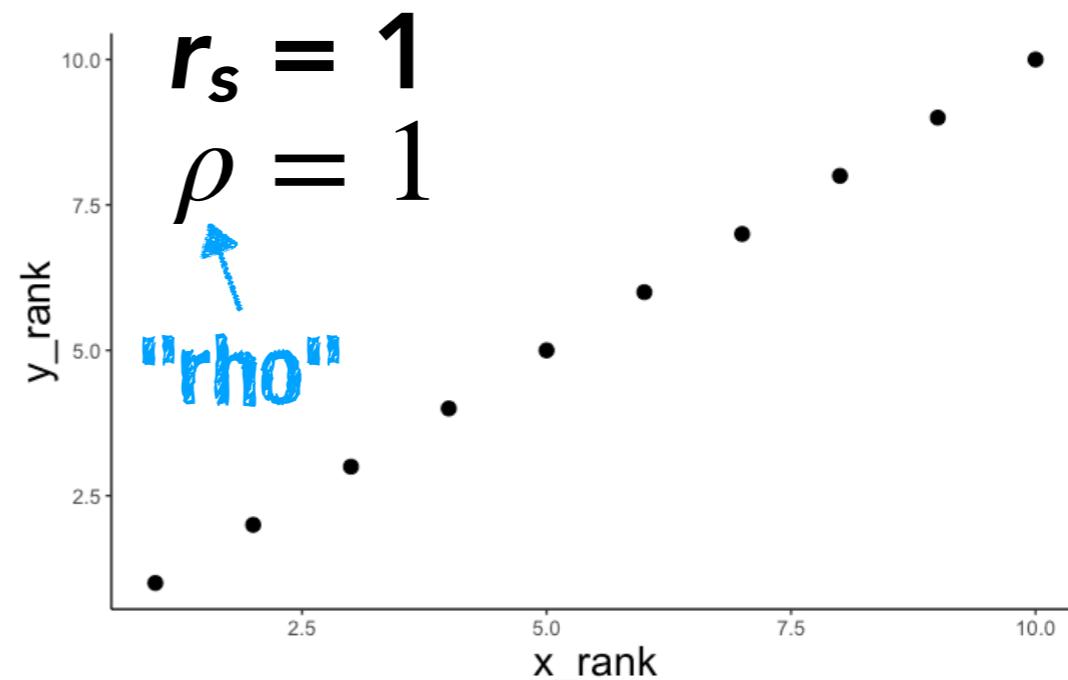
```
1 # correlation
2 df.spearman %>%
3   summarize(r = cor(x, y, method = "pearson"),
4             spearman = cor(x, y, method = "spearman"),
5             r_ranks = cor(x_rank, y_rank))
```

Spearman rank order correlation

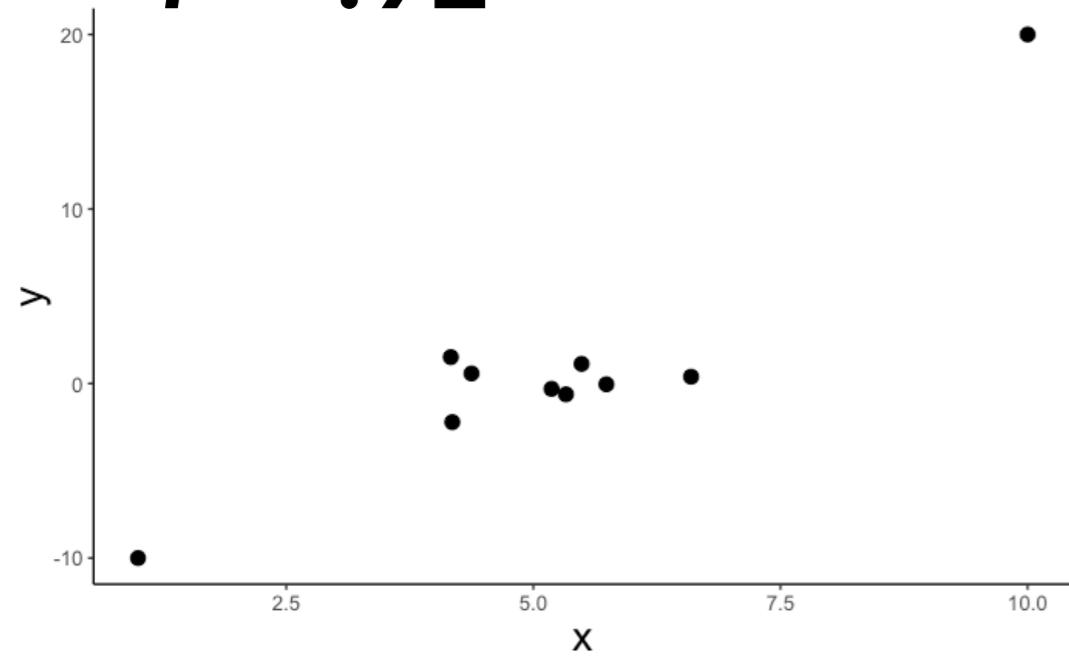
original



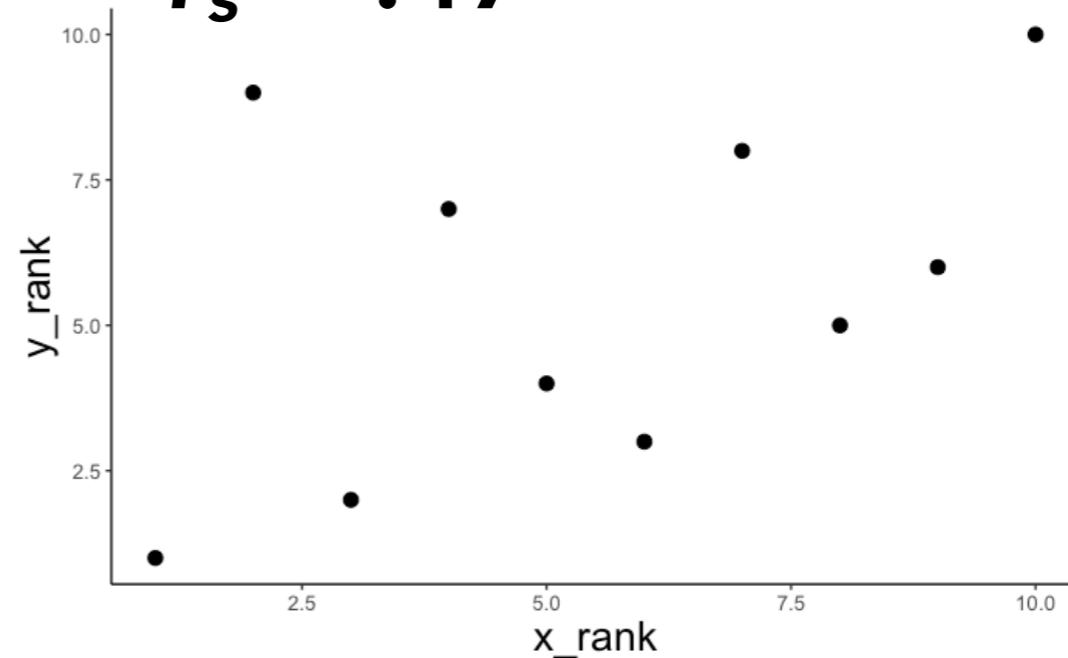
ranked



$r = .92$



$r_s = .47$



Pearson vs. Spearman

- Pearson's r captures the extent to which the relationship between two variable is **linear**
- Spearman's ρ captures the extent to which the relationship between two variables is **monotonic**
- What's better?
 - depends on the context
 - Spearman is robust to outliers, but it throws away (potentially useful) information

CORRELATION IS NOT CAUSATION



NYT Health
@NYTHealth

Want to live longer? Try going to the opera. Researchers in Britain have found that people who reported going to a museum or concert even once a year lived longer than those who didn't.



Another Benefit to Going to Museums? You May Live Longer

Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.

[nytimes.com](#)

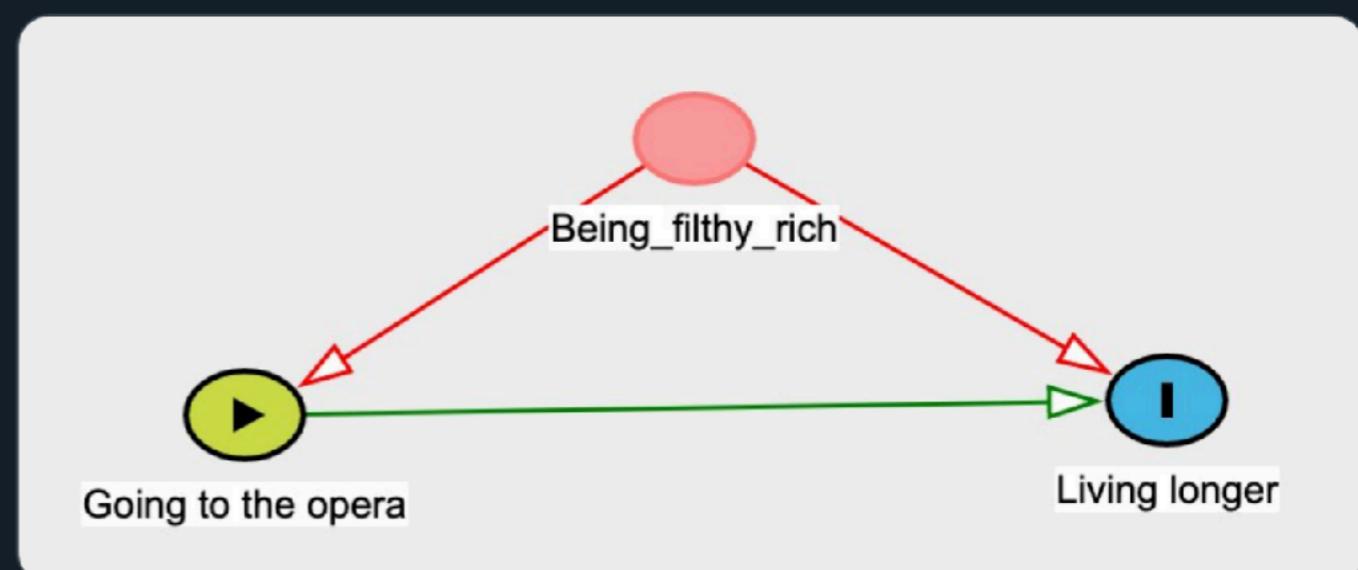
9:19 AM · Dec 22, 2019 · SocialFlow

336 Retweets 1.3K Likes



Andrew Heiss
@andrewheiss

ooh ooh i can draw the dag for this one!



NYT Health @NYTHealth · Dec 22, 2019

Want to live longer? Try going to the opera. Researchers in Britain have found that people who reported going to a museum or concert even once a year lived longer than those who didn't. [nyti.ms/2Q9AmZV](#)

2:47 PM · Dec 22, 2019 · Twitter Web App

[View Tweet activity](#)

837 Retweets 3.9K Likes

I USED TO THINK
CORRELATION IMPLIED
CAUSATION.



THEN I TOOK A
STATISTICS CLASS.
NOW I DON'T.



SOUNDS LIKE THE
CLASS HELPED.
WELL, MAYBE.



- correlations suggest that there is some causal relationship
- but this relationship need not be a direct causal relationship from A to B (or from B to A)

more about causation in a later class

Regression

The conceptual tour

Linear model: Simple regression

Data = Model + Error

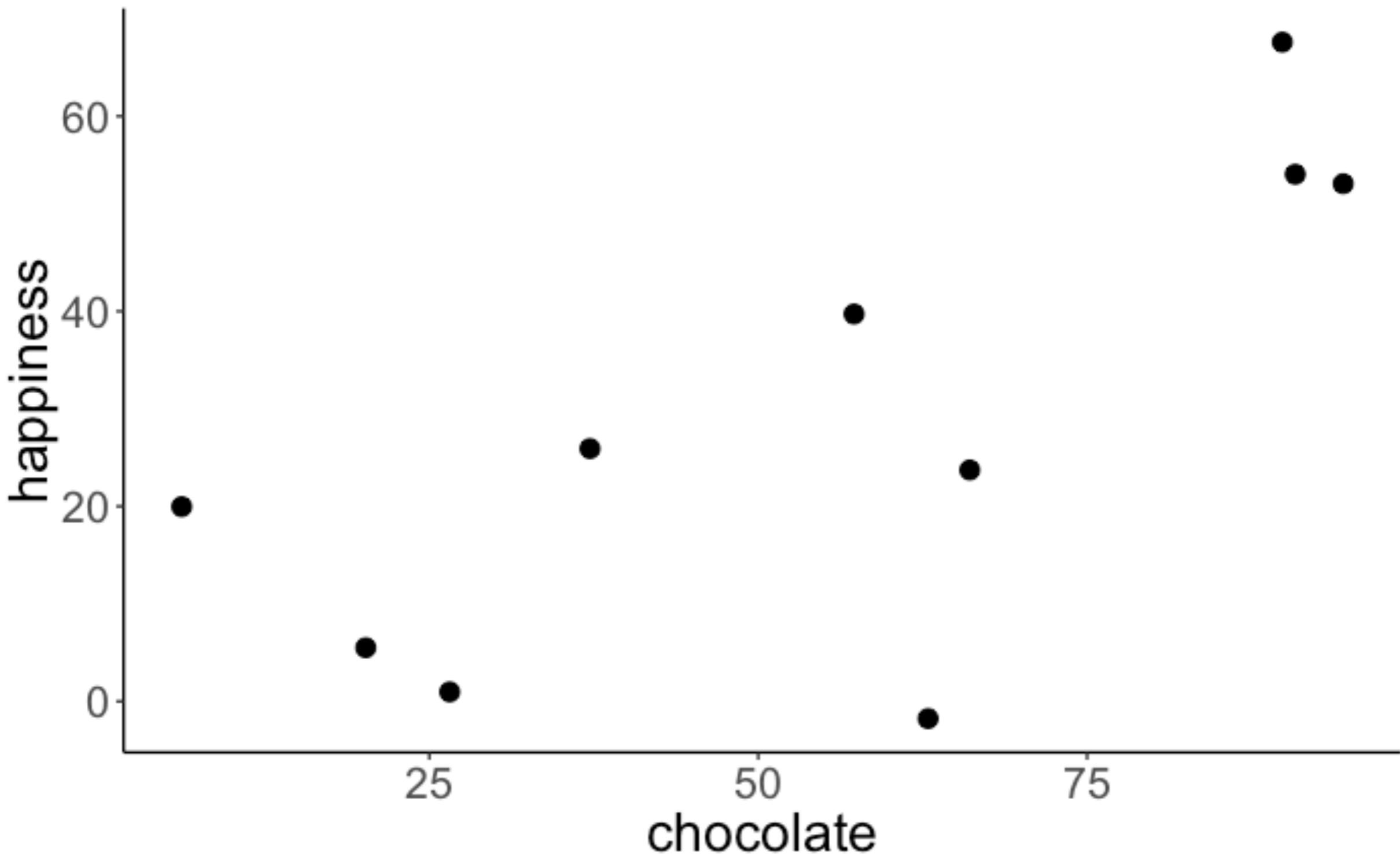
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

$$\hat{Y}_i = \beta_0 + \beta_1 X_i$$

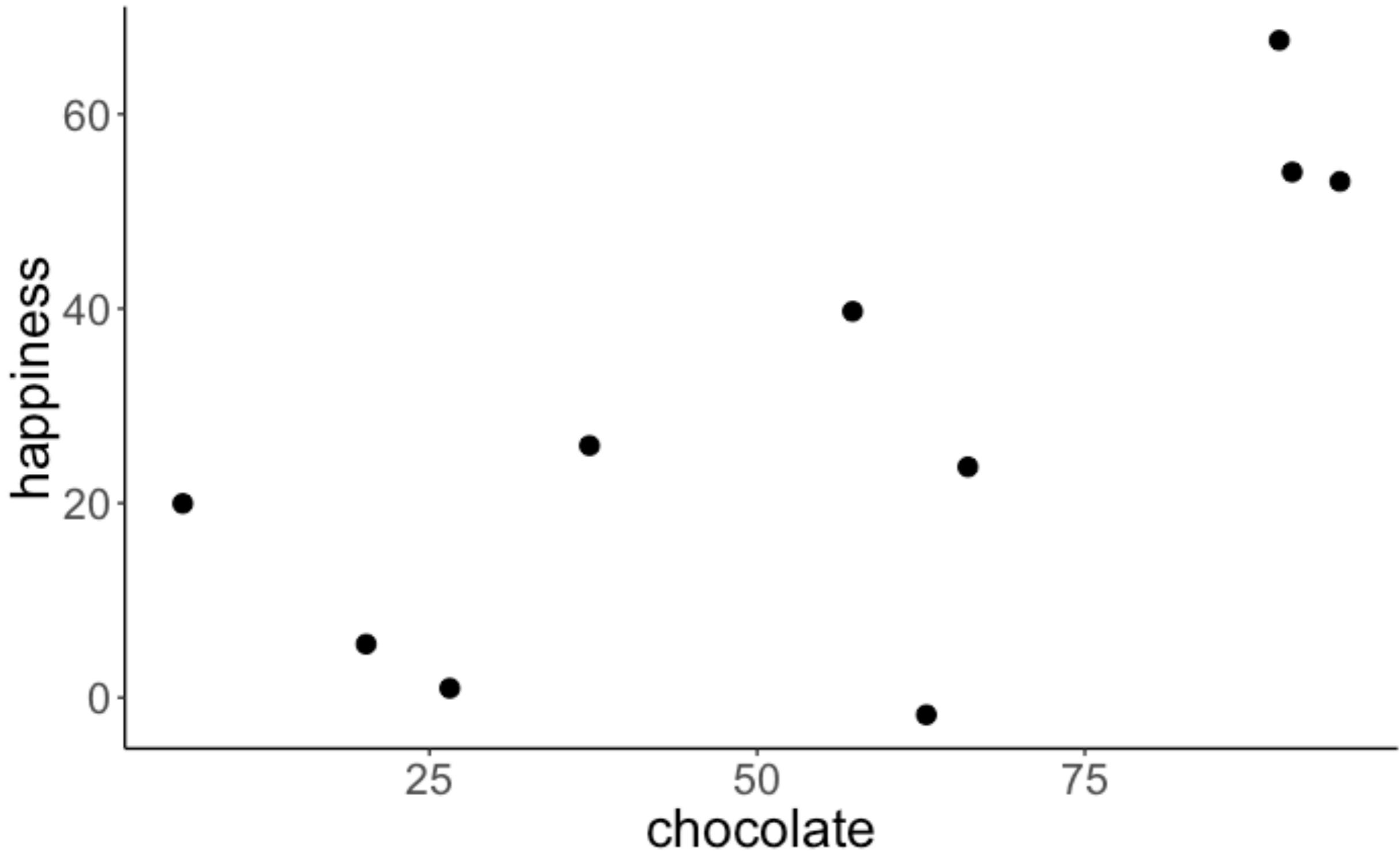


the model is a linear
combination of predictors

Does chocolate make us happy?



Is there a relationship between chocolate consumption and happiness?



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)

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)

H_0 : Chocolate consumption and happiness are unrelated.

Model C

$$Y_i = \beta_0 + \epsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

and

$$\beta_1 = 0$$

H_1 : Chocolate consumption and happiness are related.

Model A

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$



chocolate
consumption

The general procedure

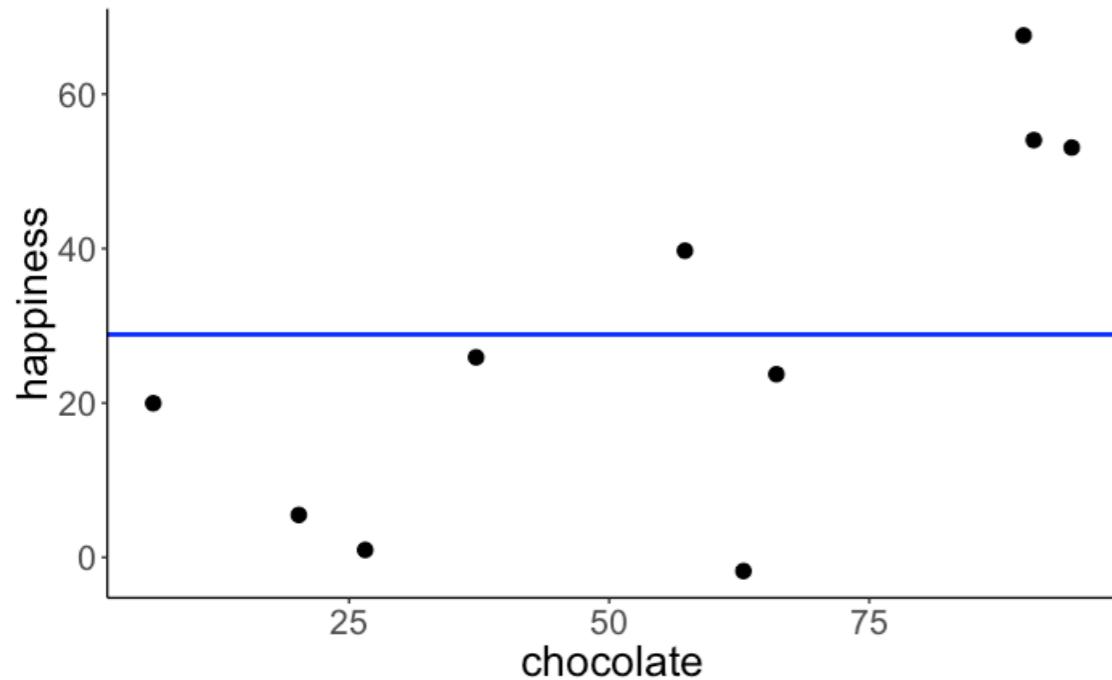
1. Define H_0 as Model C (compact) and H_1 as Model A (augmented)
- 2. Fit model parameters to the data**
3. Calculate the proportional reduction of error (PRE) in our sample
4. Decide whether the augmented model is **worth it** by comparing the observed PRE in our sample to the sampling distribution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

H_0 : Chocolate consumption and happiness are unrelated.

Model C

$$Y_i = \beta_0 + \epsilon_i$$

Model prediction



Fitted model

$$Y_i = 28.88 + e_i$$

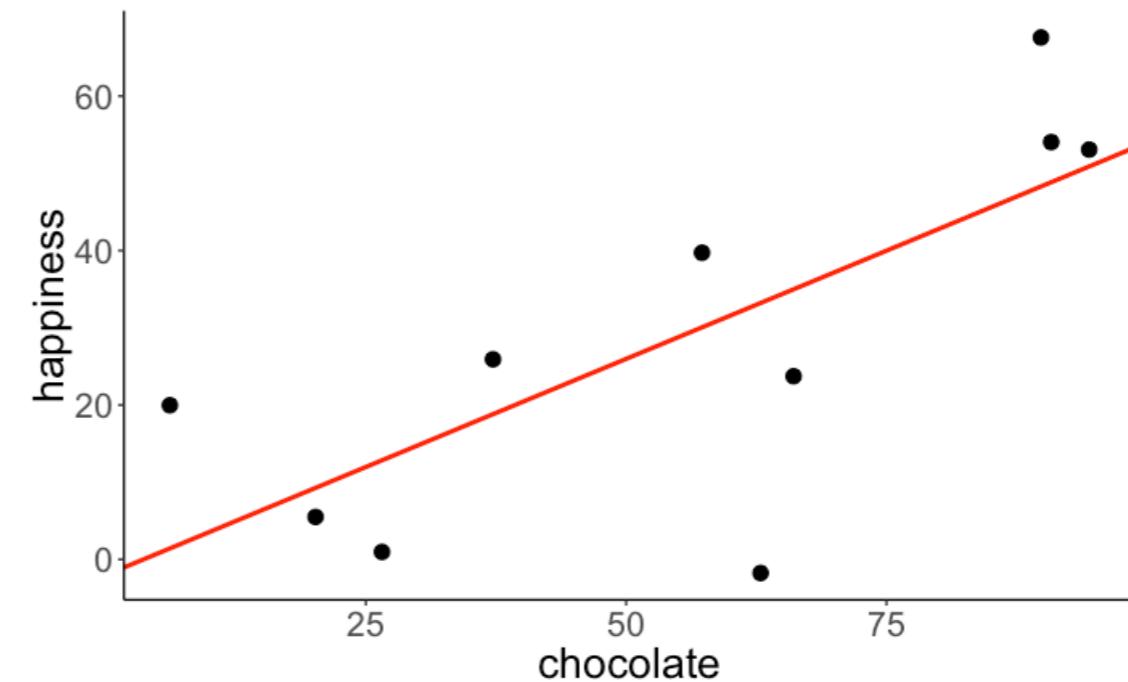
H_1 : Chocolate consumption and happiness are related.

Model A

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

chocolate consumption

Model prediction



Fitted model

$$Y_i = -2.04 + 0.56X_i + e_i$$

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)

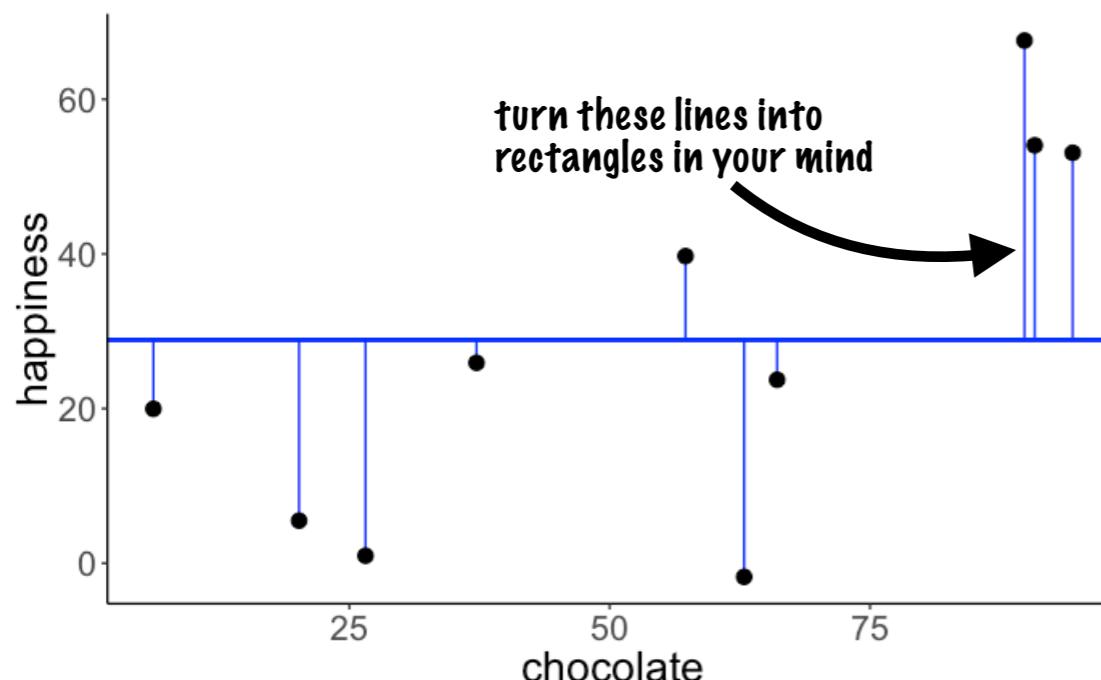
Calculate PRE

$$PRE = 1 - \frac{\text{ERROR}(A)}{\text{ERROR}(C)}$$

Both models were fit to minimize the sum of squared errors

OLS = Ordinary **least squares** regression

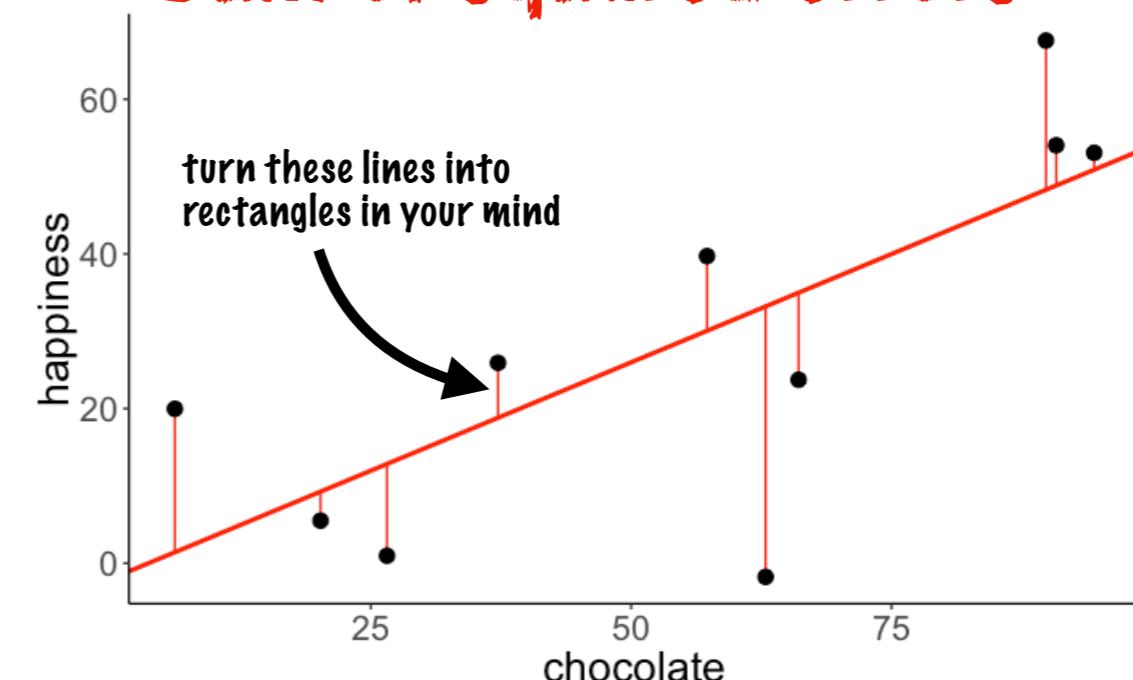
Sum of squared errors



$$\text{SSE}(C) = 5215.016$$

$$PRE = 1 - \frac{2396.946}{5215.016} \approx 0.54$$

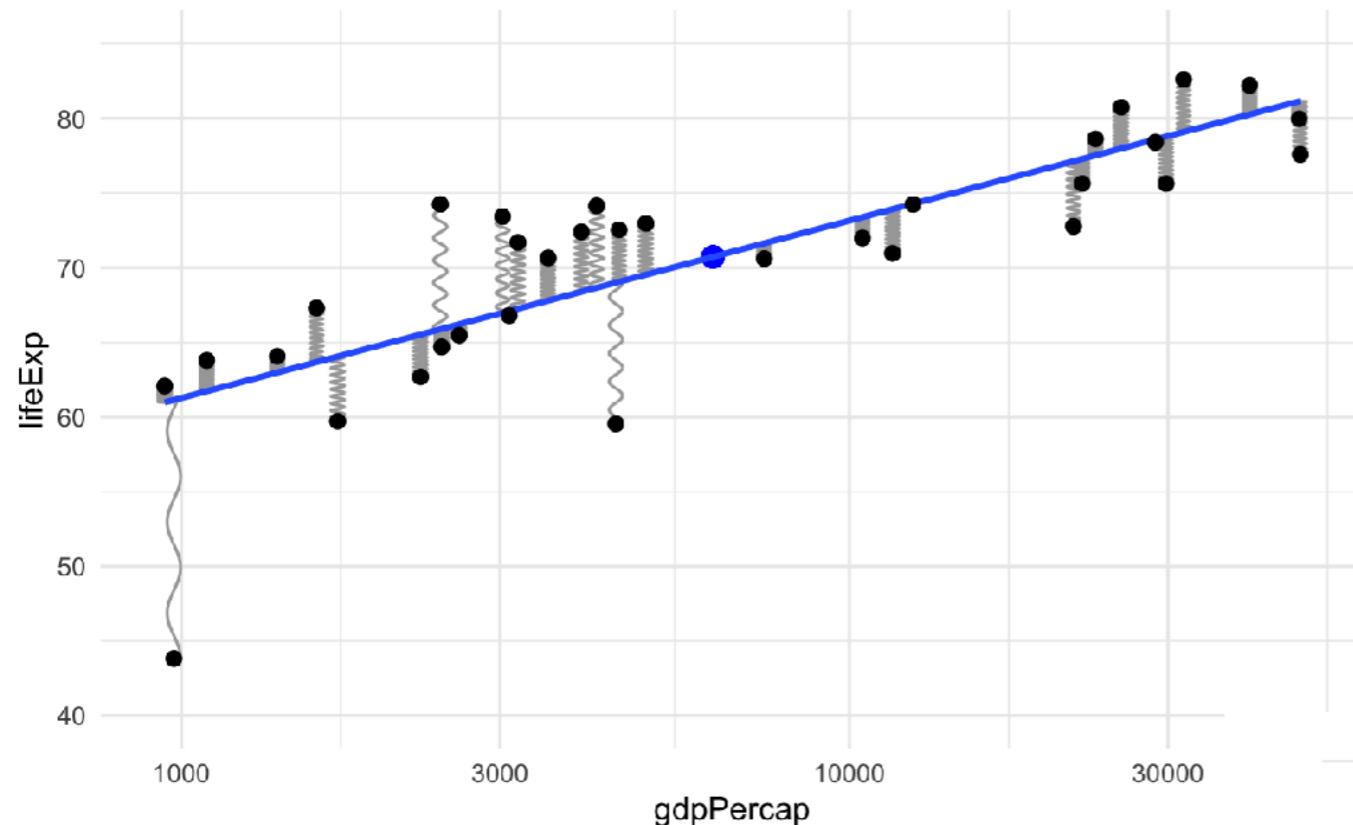
Sum of squared errors



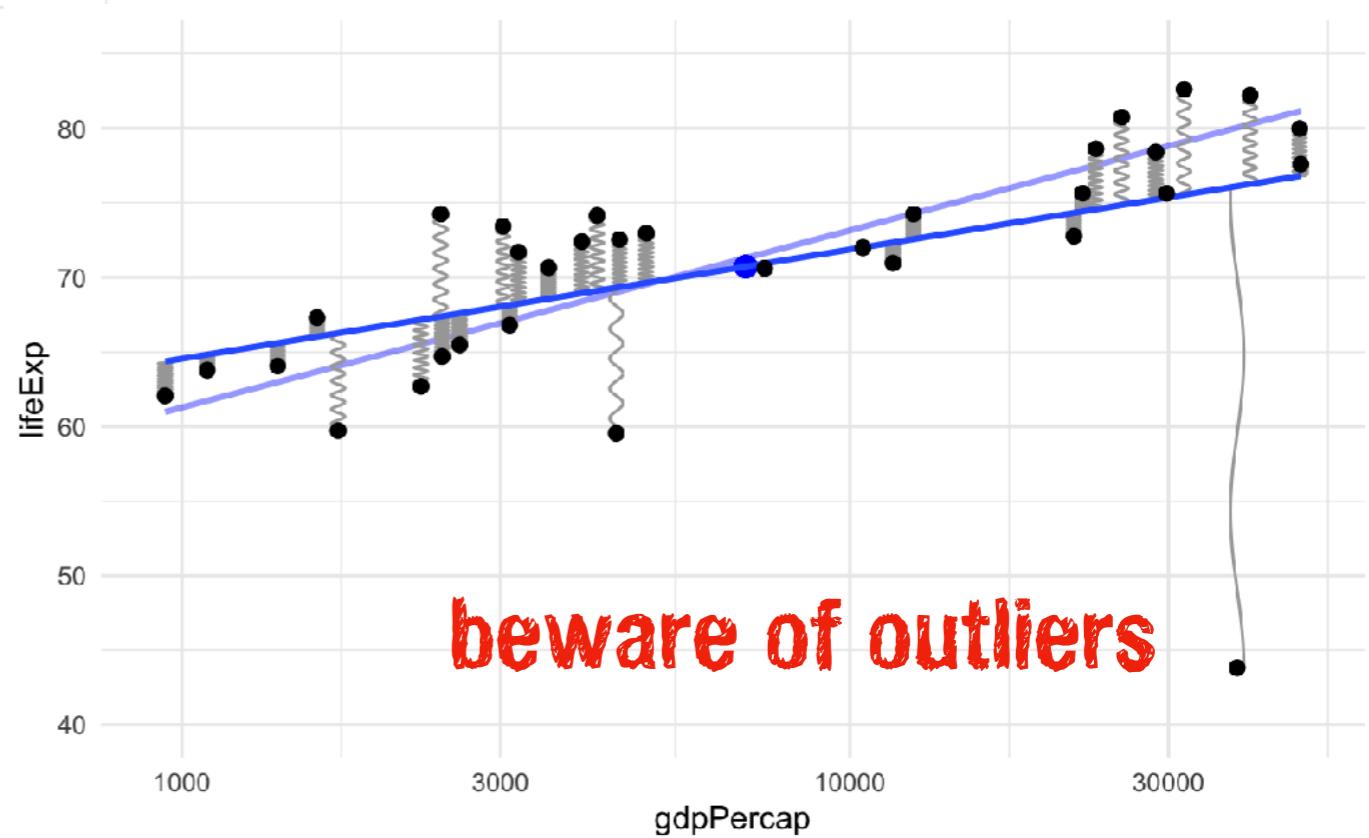
$$\text{SSE}(A) = 2396.946$$

The augmented model
reduces the error by 54%.

Least squares as springs



each point is
attached to the
line with an
identical spring



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)

Decide whether it's **worth it**

- To compute the F statistic, we need:
 - PRE
 - number of parameters in Model C (PC) and Model A (PA)
 - number of observations n

- more likely to be **worth it** if:
 1. PRE is high
 2. the number of additional parameters in A compared to C is low (**PRE per additional parameter**)
 3. the number of parameters that could have been added to model_C to create model_A but were not

**difference in parameters
between models A and C**

$$F = \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})}$$

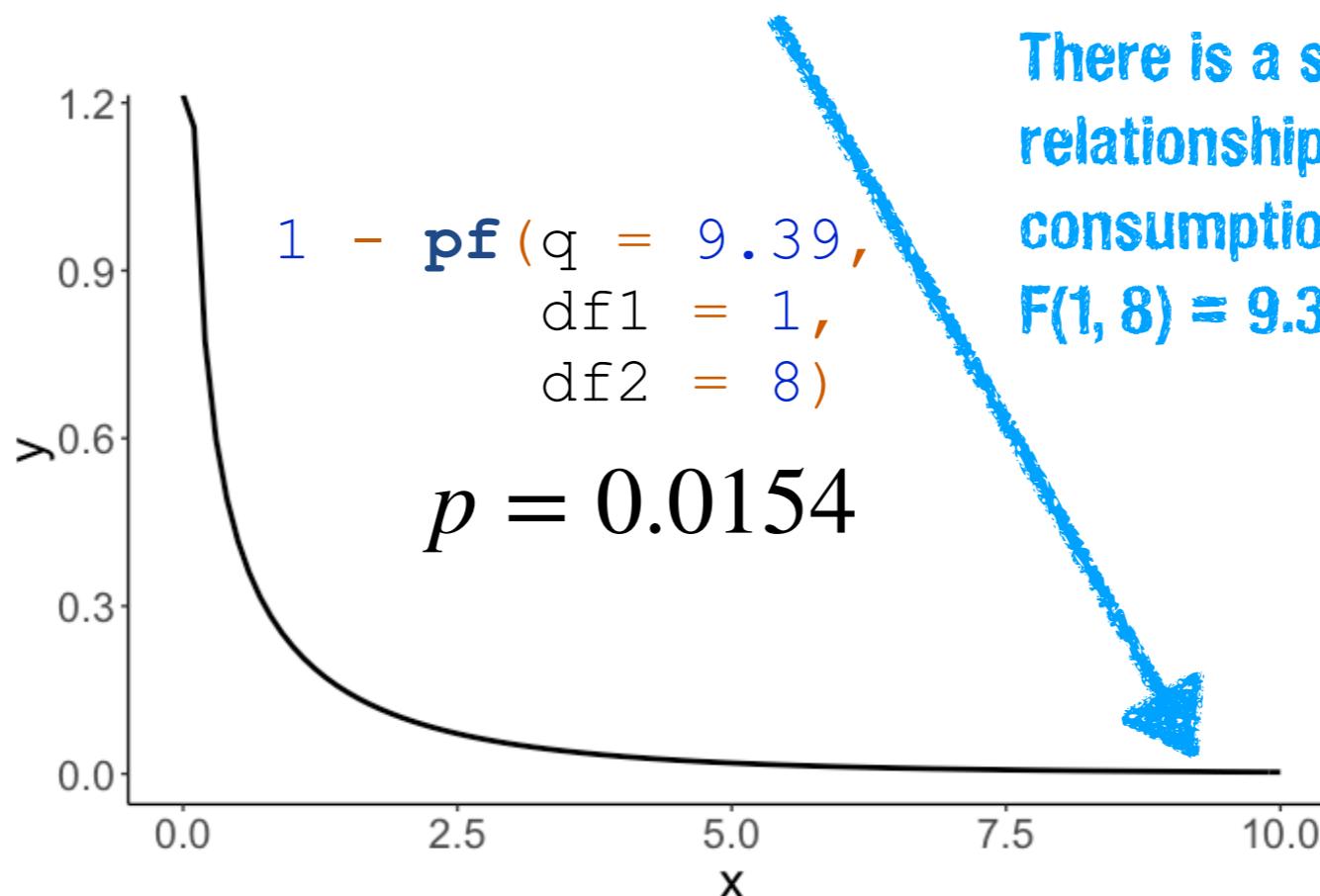
**number of observations
vs. parameters in Model A**

Decide whether it's **worth it**

- To compute the F statistic, we need:

- PRE = 0.54
- PC = 1
- PA = 2
- $n = 10$

$$\begin{aligned} F &= \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})} \\ &= \frac{0.54/(2 - 1)}{(1 - 0.54)/(10 - 2)} \\ &= 9.39 \end{aligned}$$



The R route

Credit card debt



Credit data set

df.credit

index	income	limit	rating	cards	age	education	gender	student	married	ethnicity	balance
1	14.89	3606	283	2	34	11	Male	No	Yes	Caucasian	333
2	106.03	6645	483	3	82	15	Female	Yes	Yes	Asian	903
3	104.59	7075	514	4	71	11	Male	No	No	Asian	580
4	148.92	9504	681	3	36	11	Female	No	No	Asian	964
5	55.88	4897	357	2	68	16	Male	No	Yes	Caucasian	331
6	80.18	8047	569	4	77	10	Male	No	No	Caucasian	1151
7	21.00	3388	259	2	37	12	Female	No	No	African American	203
8	71.41	7114	512	2	87	9	Male	No	No	Asian	872
9	15.12	3300	266	5	66	13	Female	No	No	Caucasian	279
10	71.06	6819	491	3	41	19	Female	Yes	Yes	African American	1350

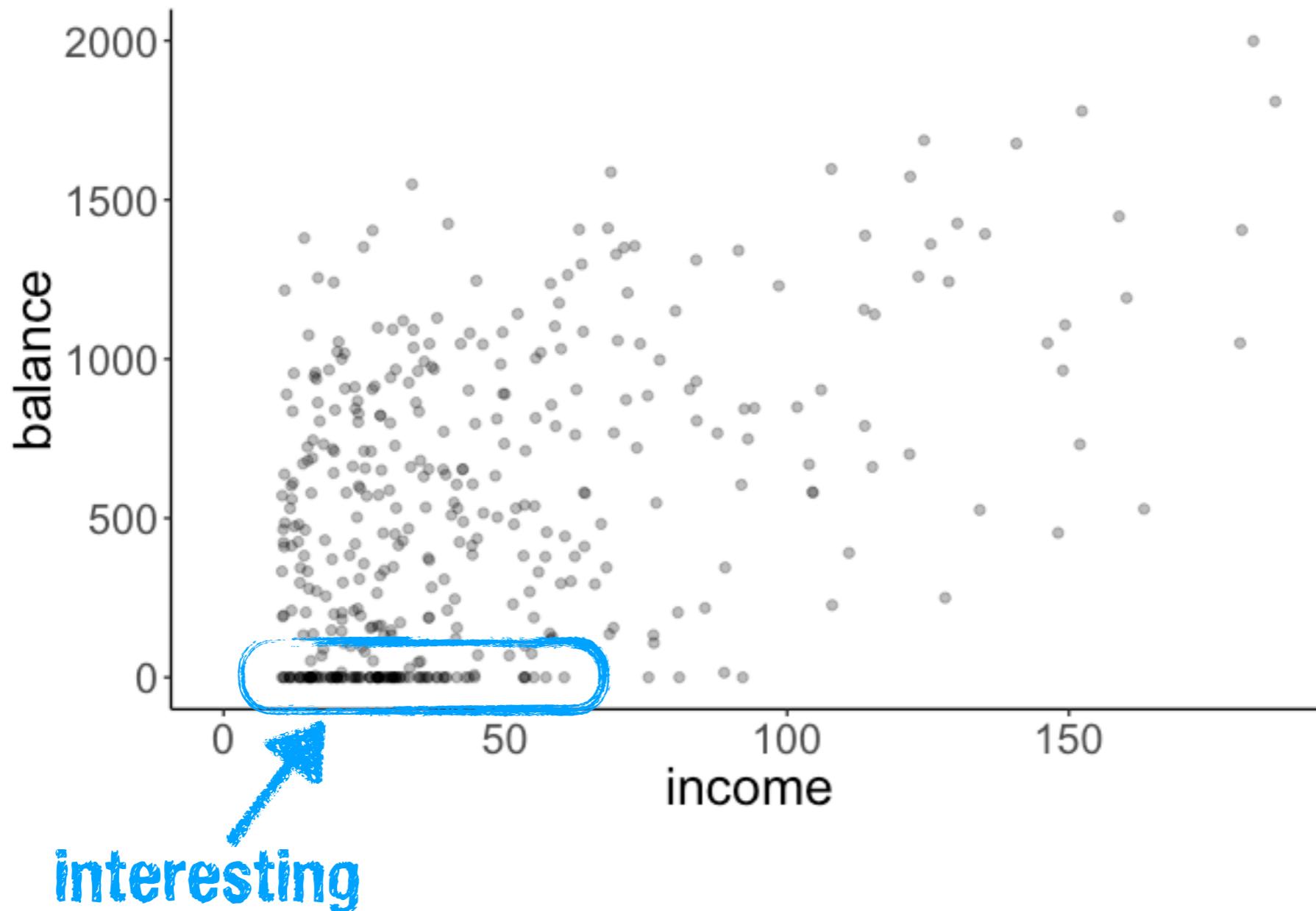
nrow(df.credit) = 400

**Is there a relationship between income
and the average credit card debt?**

variable	description
income	in thousand dollars
limit	credit limit
rating	credit rating
cards	number of credit cards
age	in years
education	years of education
gender	male or female
student	student or not
married	married or not
ethnicity	African American, Asian, Caucasian
balance	average credit card debt in dollars

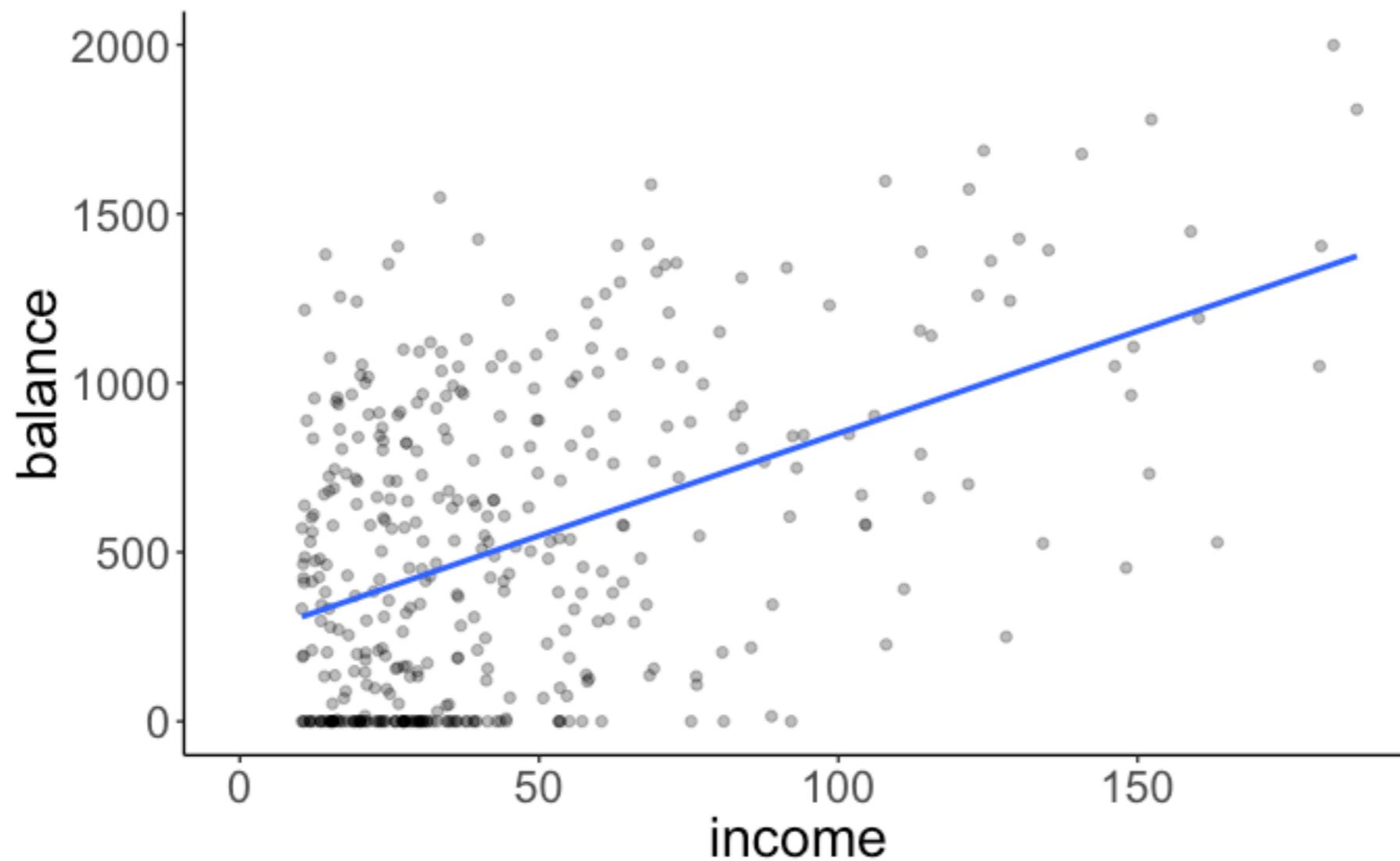
Always plot the data first ...

```
1 ggplot(data = df.credit,  
2         mapping = aes(x = income,  
3                             y = balance)) +  
4     geom_point(alpha = 0.3)
```



Always plot the data first ...

```
1 ggplot(data = df.credit,  
2         mapping = aes(x = income,  
3                             y = balance)) +  
4     geom_point(alpha = 0.3) +  
5     geom_smooth(method = "lm", se = F)
```

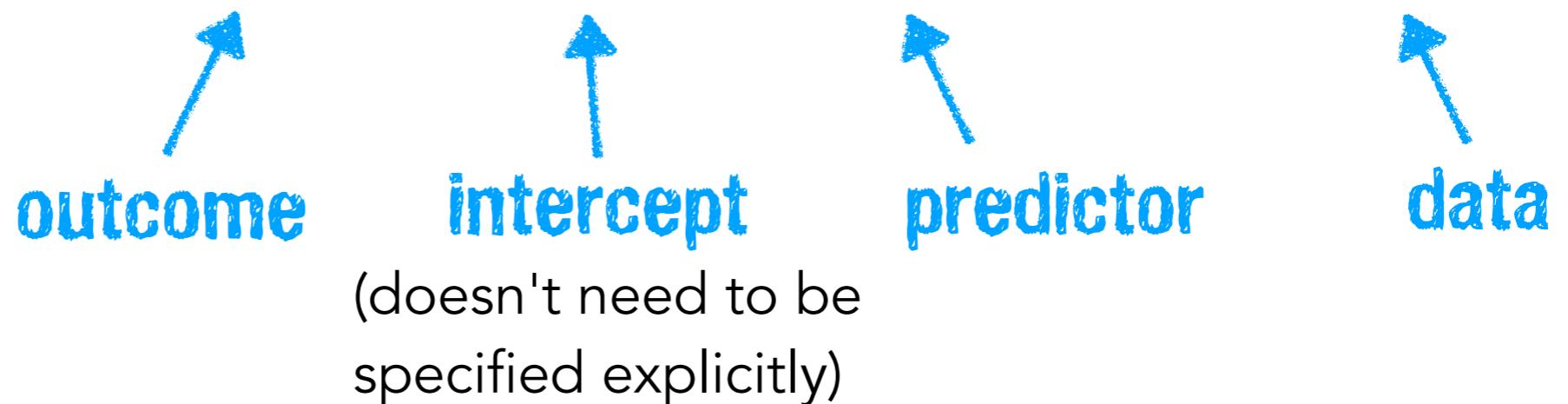


Linear model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

$$\text{balance}_i = b_0 + b_1 \cdot \text{income}_i + e_i$$

```
fit = lm(formula = balance ~ 1 + income, data = df.credit)
```



lm()

```
fit = lm(balance ~ 1 + income, data = df.credit)
```

```
print(fit)
```

Call:

```
lm(formula = balance ~ 1 + income, data = df.credit)
```

Coefficients:

(Intercept)	income
246.515	6.048

parameter estimates

which minimize the squared error between model and data

Interpreting regression parameters

Coefficients:

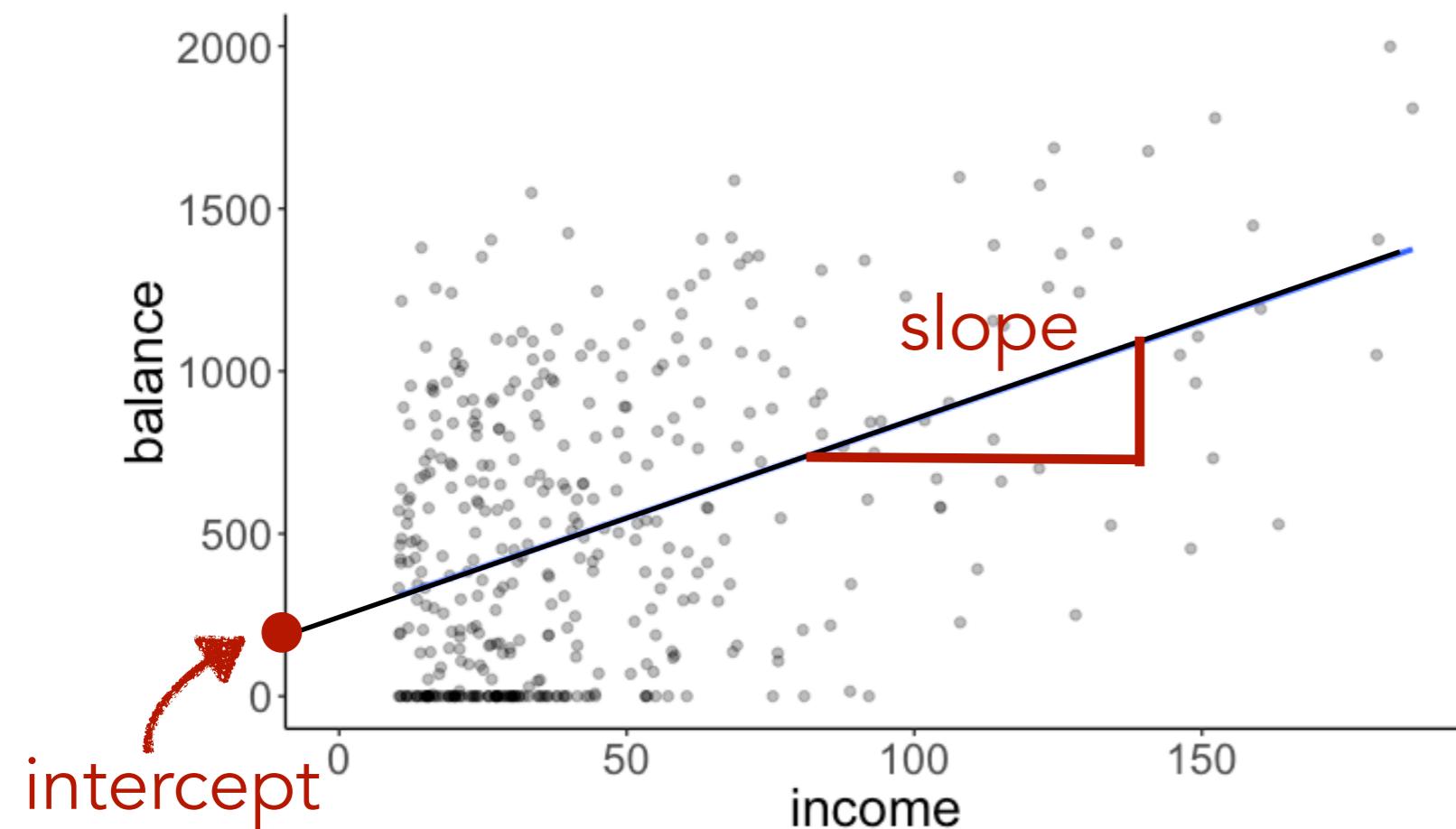
(Intercept) 246.515

income 6.048

variable	description
income	in thousand dollars
balance	average credit card debt in dollars

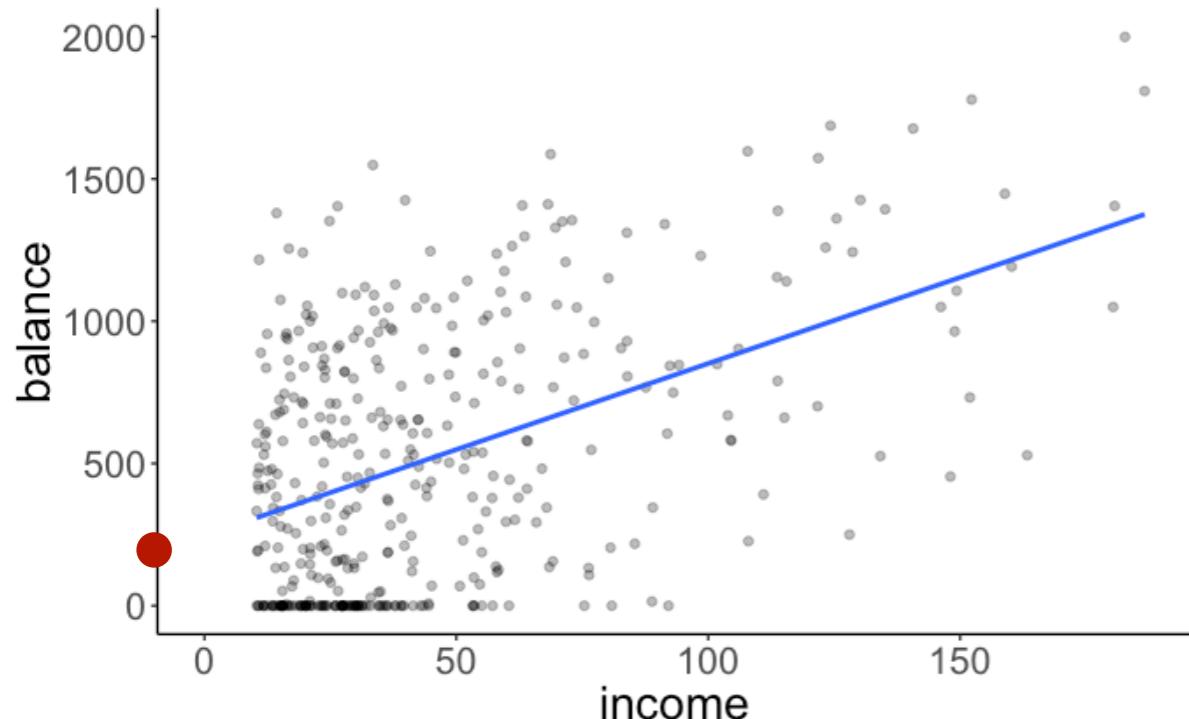
$$\text{balance}_i = b_0 + b_1 \cdot \text{income}_i + e_i$$

$$\text{balance}_i = 246.515 + 6.048 \cdot \text{income}_i + e_i$$



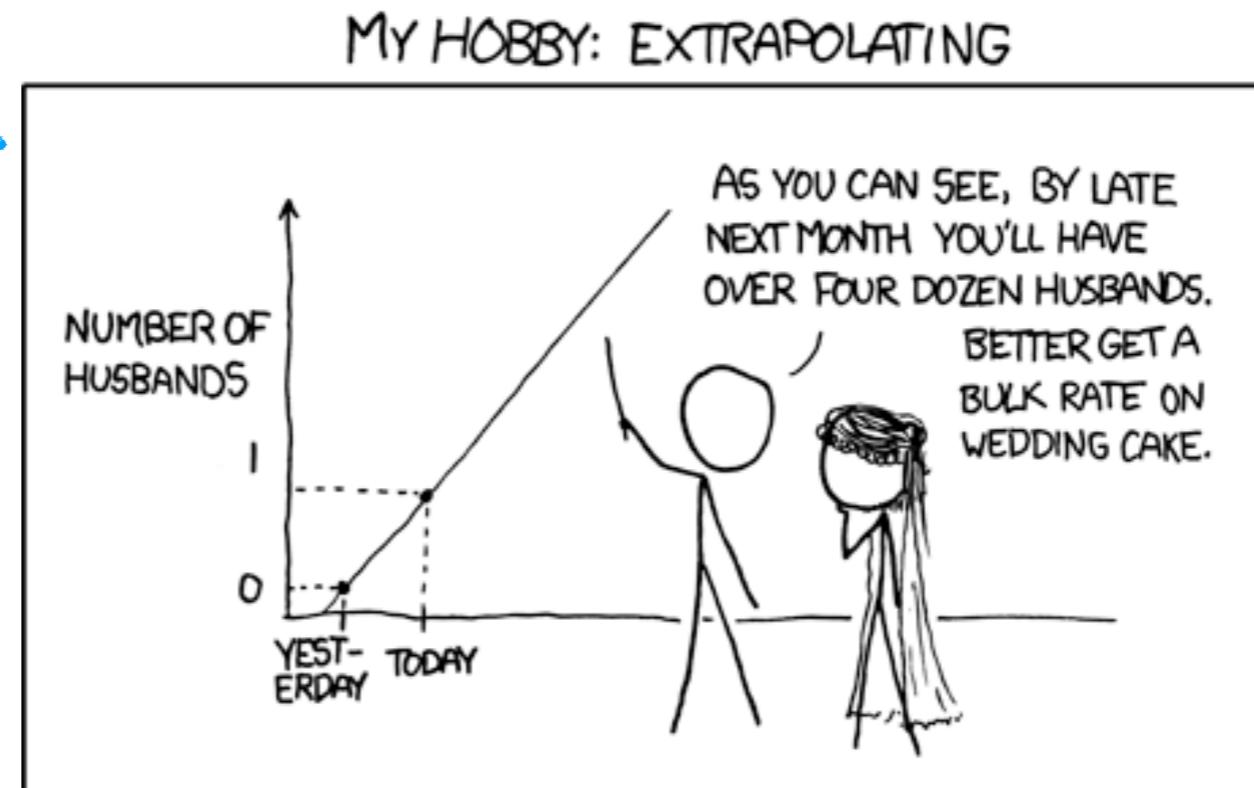
For each additional thousand dollars income, a person's average credit card is predicted to increase by \$6.05.

Be careful about extrapolating predictions



- intercept is often outside the range of predictor values
- sometimes doesn't make sense (e.g. age = 0, height = 0, ...)

comic from
slide 1



Centering the predictor

```
1 df.credit %>%
2   mutate(income_centered = income - mean(income)) %>%
3   select(balance, income, income_centered)
```

balance	income	income_centered
333	14.89	-30.33
903	106.03	60.81
580	104.59	59.37
964	148.92	103.71
331	55.88	10.66
1151	80.18	34.96
203	21.00	-24.22
872	71.41	26.19
279	15.12	-30.09
1350	71.06	25.84

```
library ("broom")
```



helps with tidying up
model objects in R

augment() adds columns to the original data such as predictions, residuals and cluster assignments

tidy() summarizes a model's statistical findings such as coefficients of a regression

glance() provides a one-row summary of model-level statistics

broom: turn messy model outputs
into **tidy** TIBBLES!



@allison_horst

summary()

Residuals:					
Min	1Q	Median	3Q	Max	
-803.64	-348.99	-54.42	331.75	1100.25	

fit %>%

augment()

balance	income	.fitted	.se.fit	.resid	.hat	.sigma	.cooksdi	.std.resid
333	14.89	336.58	26.92	-3.58	0.00	408.38	0.00	-0.01
903	106.03	887.79	40.71	15.21	0.01	408.38	0.00	0.04
580	104.59	879.13	39.99	-299.13	0.01	408.10	0.00	-0.74
964	148.92	1147.26	63.45	-183.26	0.02	408.27	0.00	-0.45
331	55.88	584.51	21.31	-253.51	0.00	408.18	0.00	-0.62
1151	80.18	731.47	28.74	419.53	0.00	407.83	0.00	1.03
203	21.00	373.51	24.76	-170.51	0.00	408.29	0.00	-0.42
872	71.41	678.42	25.42	193.58	0.00	408.26	0.00	0.48
279	15.12	338.00	26.83	-59.00	0.00	408.37	0.00	-0.14
1350	71.06	676.32	25.30	673.68	0.00	406.97	0.01	1.65

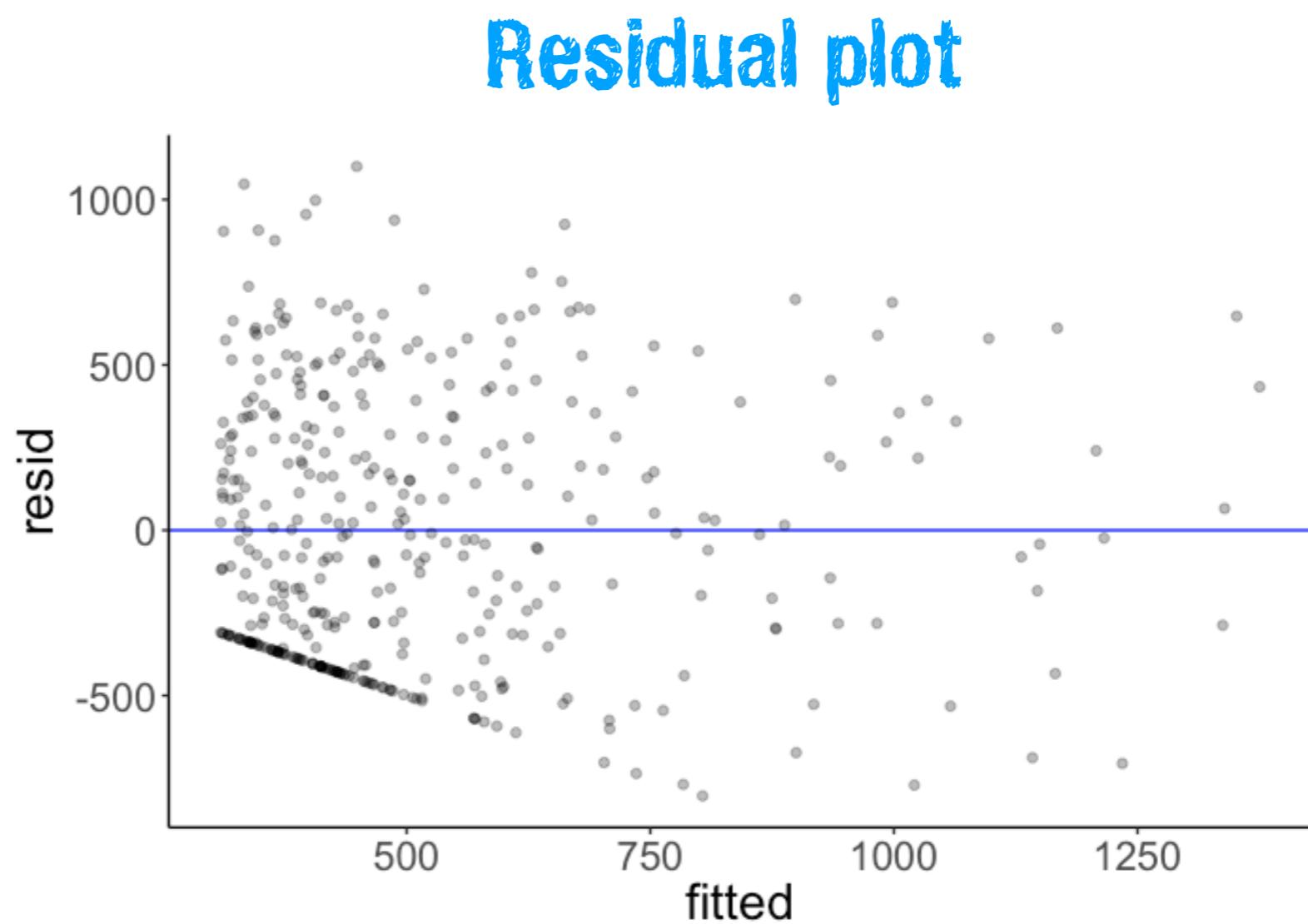
summary()

Residuals:					
Min	1Q	Median	3Q	Max	
-803.64	-348.99	-54.42	331.75	1100.25	

fit %>%

augment()

balance	income	.fitted	.resid
333	14.89	336.58	-3.58
903	106.03	887.79	15.21
580	104.59	879.13	-299.13
964	148.92	1147.26	-183.26
331	55.88	584.51	-253.51
1151	80.18	731.47	419.53
203	21.00	373.51	-170.51
872	71.41	678.42	193.58
279	15.12	338.00	-59.00
1350	71.06	676.32	673.68



summary()

```
1 lm(balance ~ 1 + income, data = df.credit) %>%
2   summary()
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	246.5148	33.1993	7.425	6.9e-13 ***
income	6.0484	0.5794	10.440	< 2e-16 ***

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

```
1 fit %>%
```

```
2   tidy(conf.int = TRUE)
```

a data frame, yay!

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	246.51	33.20	7.43	0	181.25	311.78
income	6.05	0.58	10.44	0	4.91	7.19

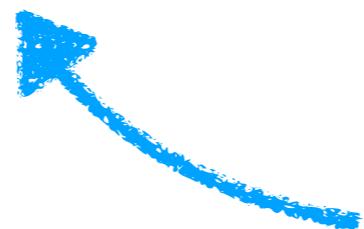
summary()

```
1 lm(balance ~ 1 + income, data = df.credit) %>%
2   summary()
```

```
Residual standard error: 407.9 on 398 degrees of freedom
Multiple R-squared:  0.215, Adjusted R-squared:  0.213
F-statistic: 109 on 1 and 398 DF,  p-value: < 2.2e-16
```

```
1 fit %>%
2   glance()
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual
0.21	0.21	407.86	108.99	0	2	-2970.95	5947.89	5959.87	66208745	398



useful model summary
(we will learn later what
the different values mean)

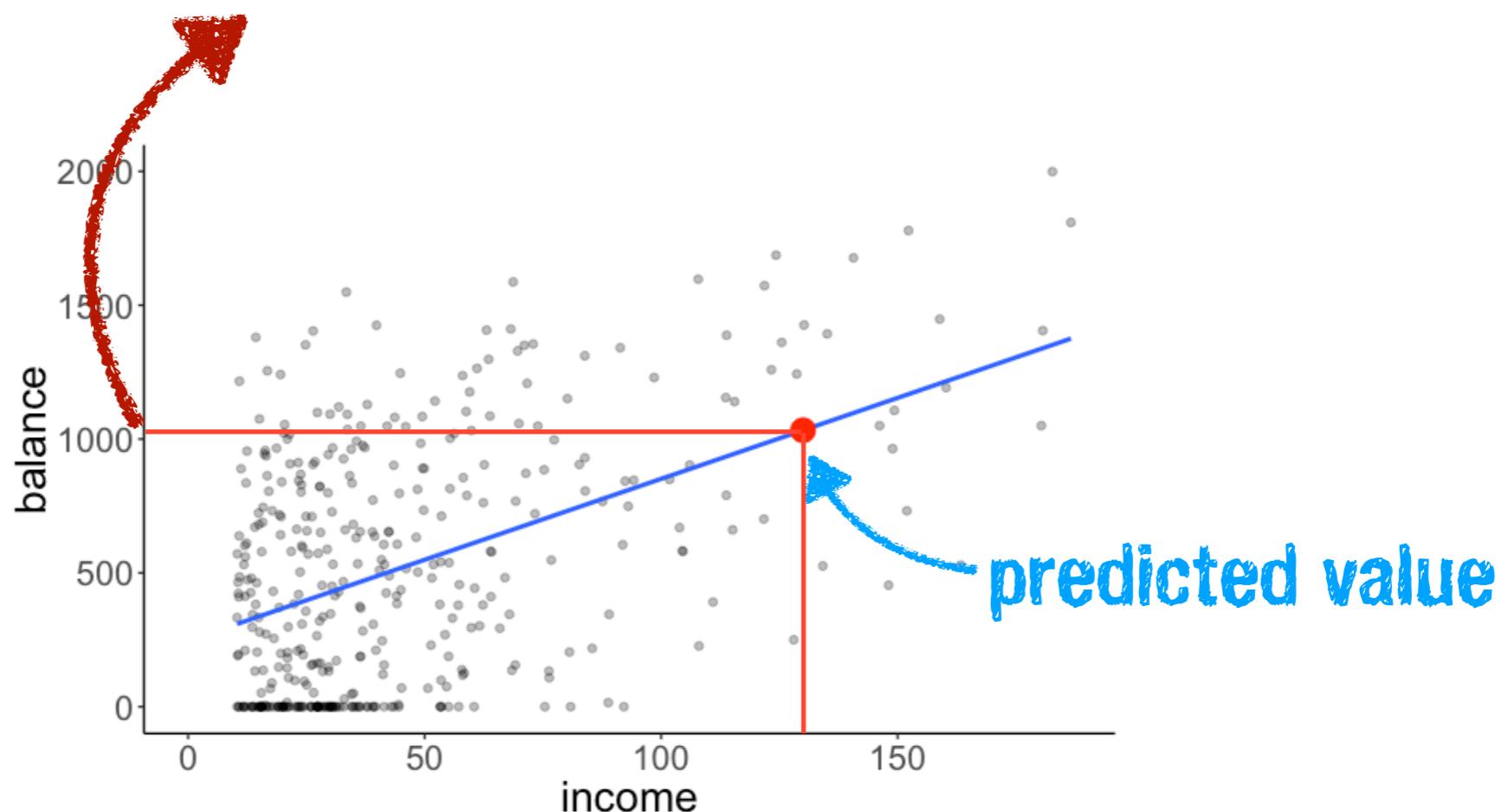
Making predictions

```
fit = lm(balance ~ 1 + income, data = df.credit)
```

$$\text{balance}_i = 246.515 + 6.048 \cdot \text{income}_i + e_i$$

```
augment(fit, newdata = tibble(income = 130))
```

$$\widehat{\text{balance}} = 246.515 + 6.048 \cdot 130$$



Hypothesis test

Compact Model

$$\text{balance}_i = \beta_0 + \epsilon_i$$

```
fit_c = lm(formula = balance ~ 1,  
           data = df.credit)
```

Augmented Model

$$\text{balance}_i = \beta_0 + \beta_1 \text{income}_i + \epsilon_i$$

```
fit_a = lm(formula = balance ~ 1 + income,  
           data = df.credit)
```

anova(fit_c, fit_a)

Analysis of Variance Table

Model 1: balance ~ 1

Model 2: balance ~ 1 + income

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	399	84339912			
2	398	66208745	1	18131167 108.99 < 2.2e-16 ***	

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

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)

```
fit_c = lm(formula = balance ~ 1,  
           data = df.credit)
```

```
fit_a = lm(formula = balance ~ 1 + income,  
           data = df.credit)
```

```
anova(fit_c, fit_a)
```

Hypothesis test

anova(fit_c, fit_a)

Analysis of Variance Table

Model 1: balance ~ 1

Model 2: balance ~ 1 + income

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	399	84339912			
2	398	66208745	1	18131167 108.99	< 2.2e-16 ***

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

$$\text{PRE} = 1 - \frac{66208745}{84339912} \approx 0.215$$

The augmented model reduces the error by 21.5%.

```
lm(balance ~ 1 + income, data = df.credit) %>%  
  summary()
```

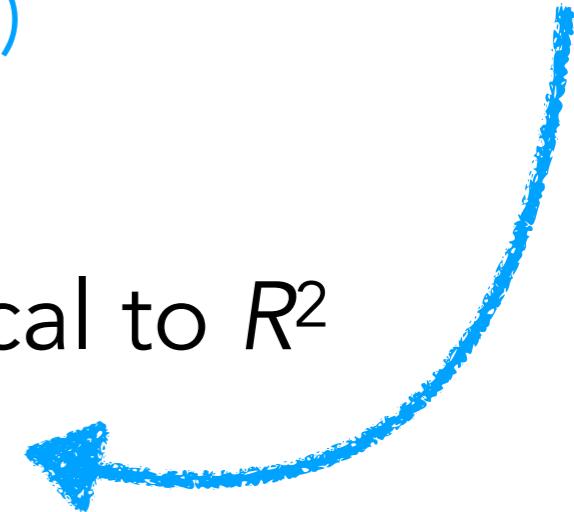
R^2

```
Residual standard error: 407.9 on 398 degrees of freedom  
Multiple R-squared: 0.215, Adjusted R-squared: 0.213  
F-statistic: 109 on 1 and 398 DF, p-value: < 2.2e-16
```

Hypothesis test

the **compact model** predicts the mean (which doesn't explain any of the variance)

- in the case of a simple regression PRE (proportion of reduced error) is identical to R^2 (variance explained)
- and R^2 is directly related to the correlation coefficient r



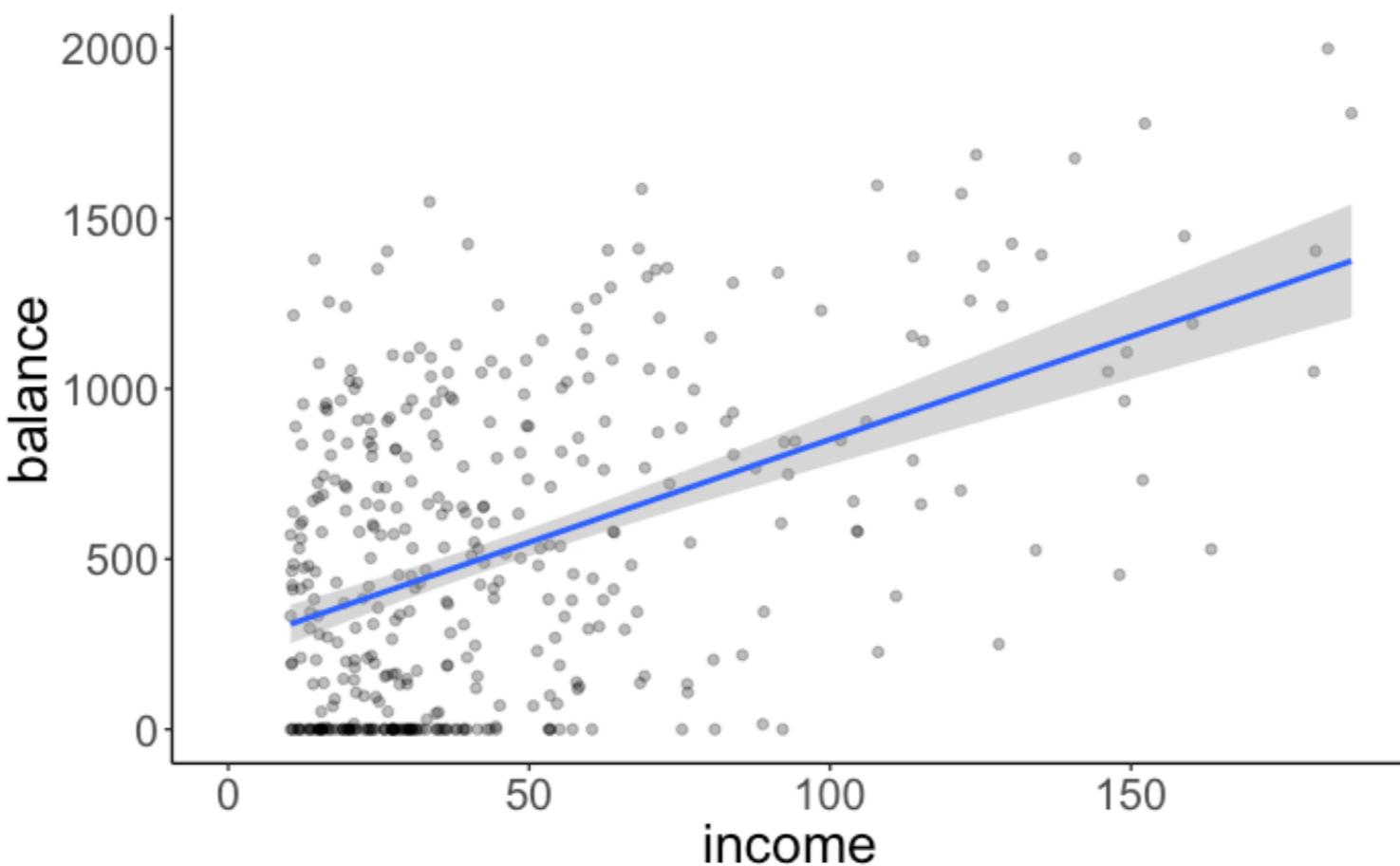
```
cor(df.credit$balance,  
df.credit$income)
```

$$R^2 = 0.215$$

$$r = .463$$

effect size measure

Reporting the results

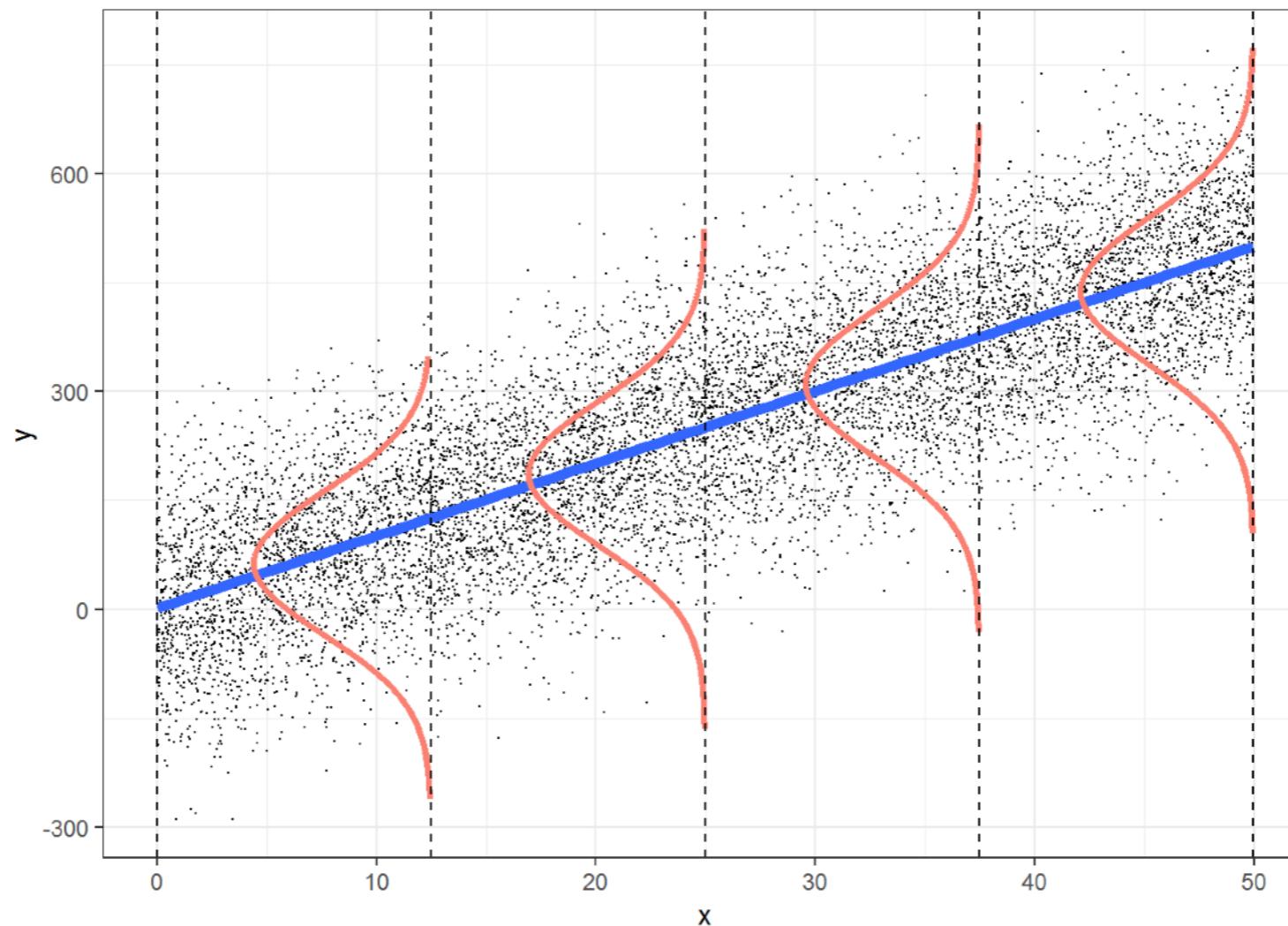


There is a significant relationship between a person's income and the average balance on their credit cards
 $F(1, 389) = 108.99, p < .001, r = .463$.

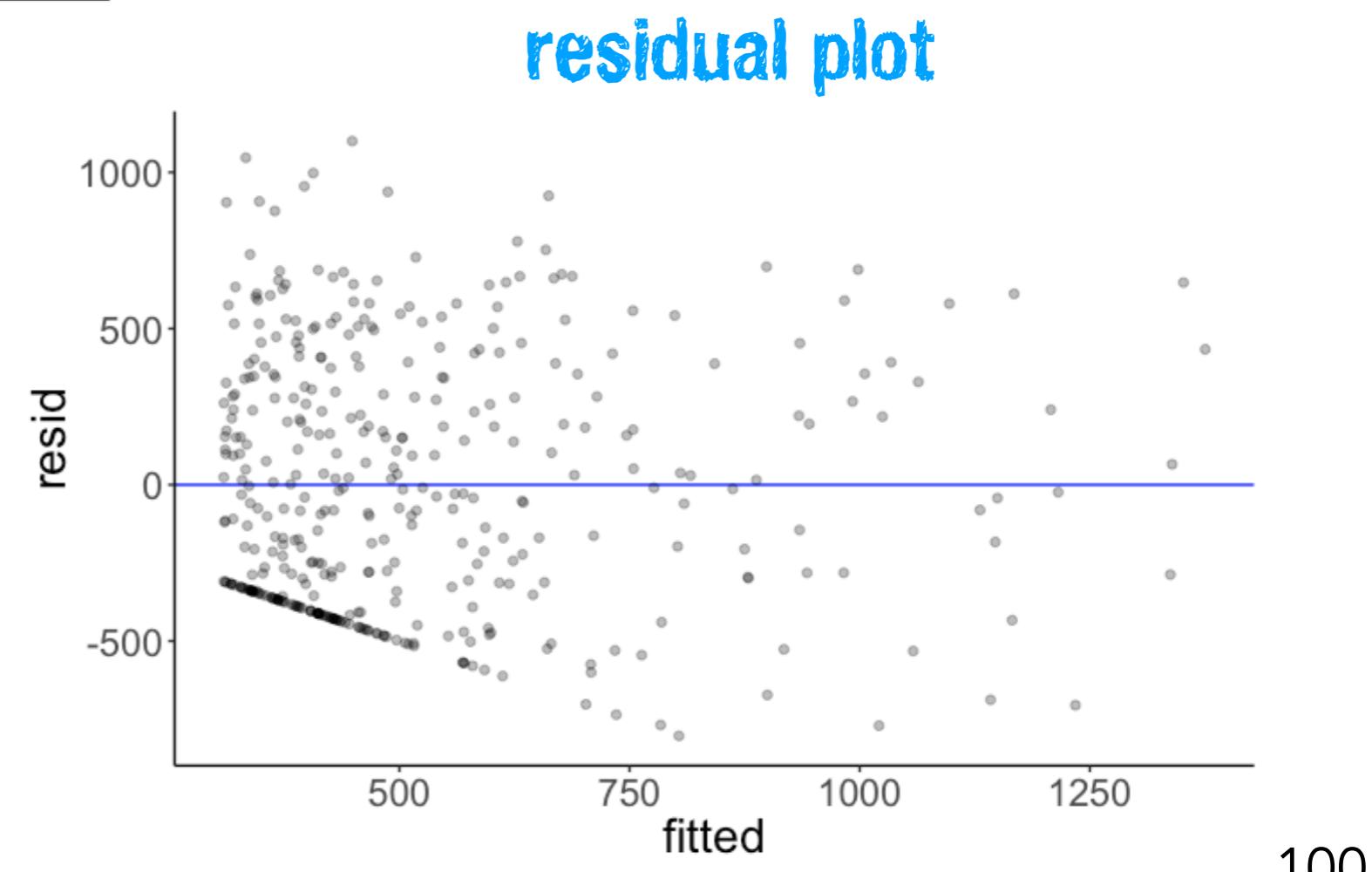
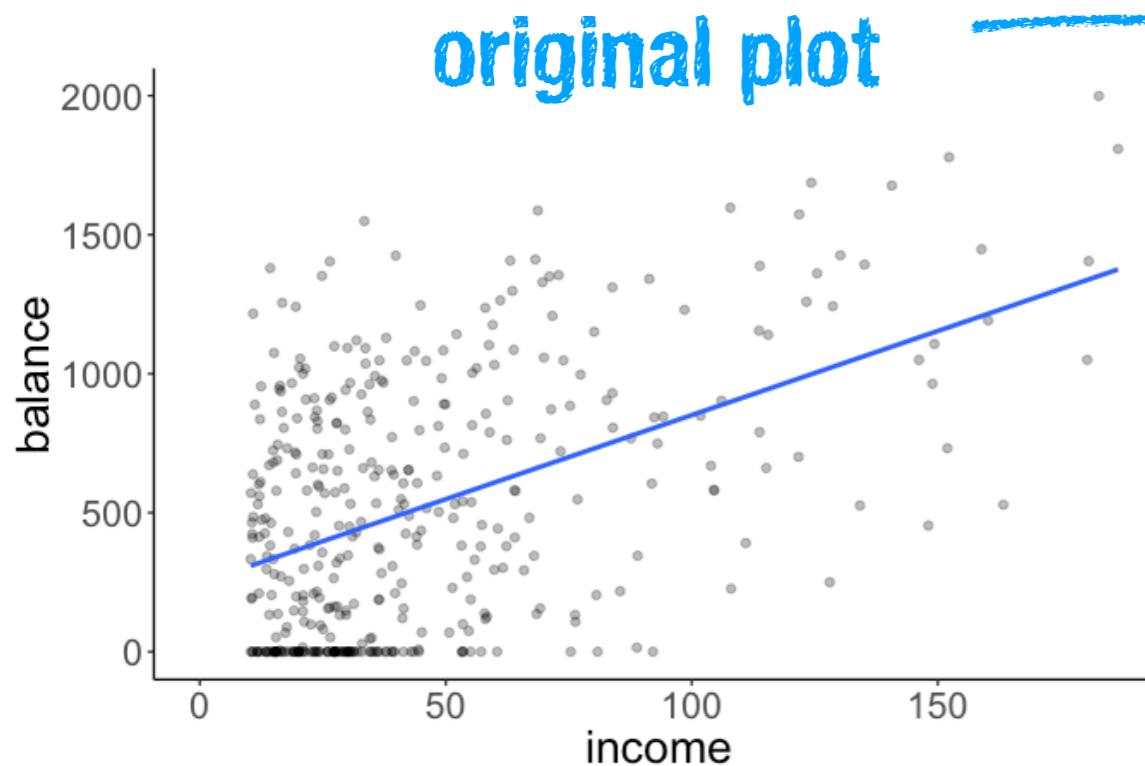
With each additional \$1000 of income, the average balance is predicted to increase by \$6.05 [4.91, 7.19] (95% CI).

Model assumptions

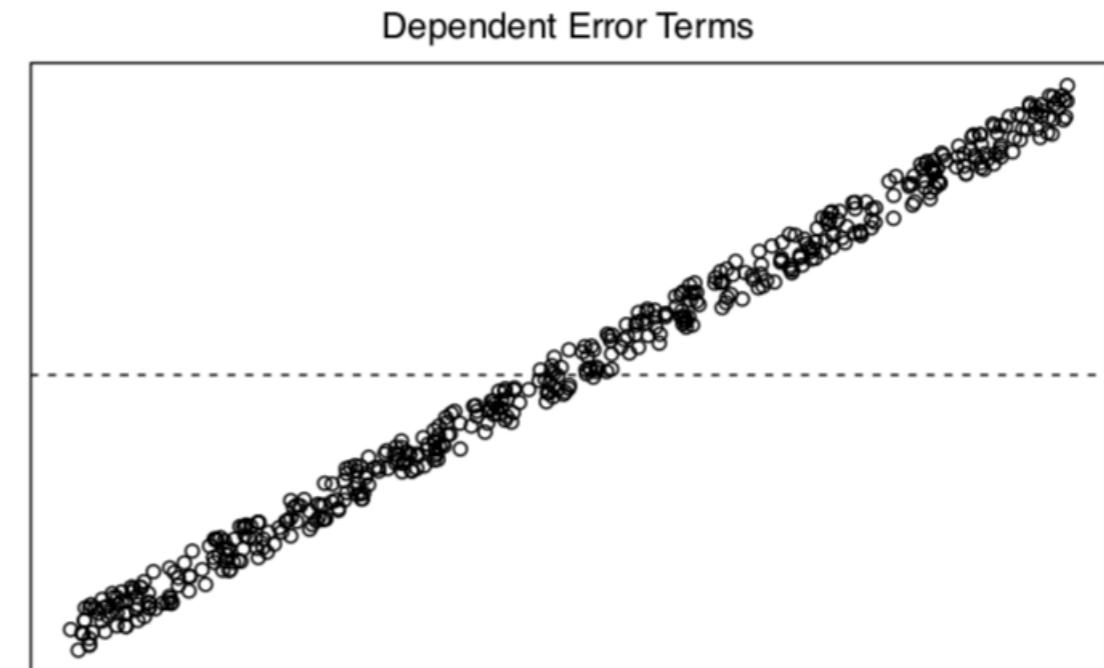
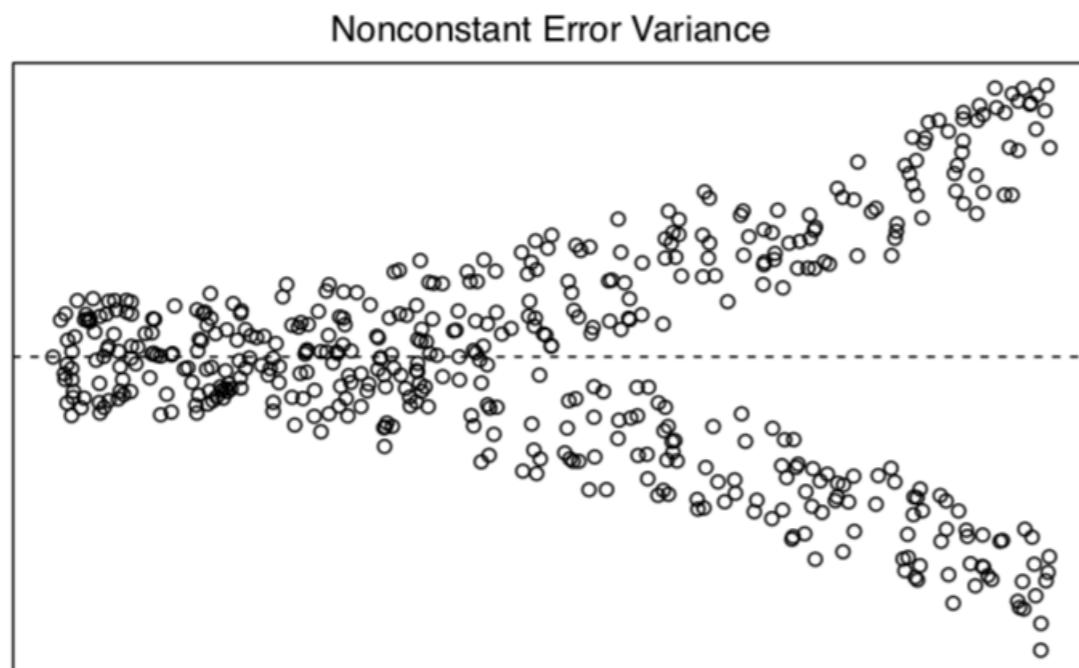
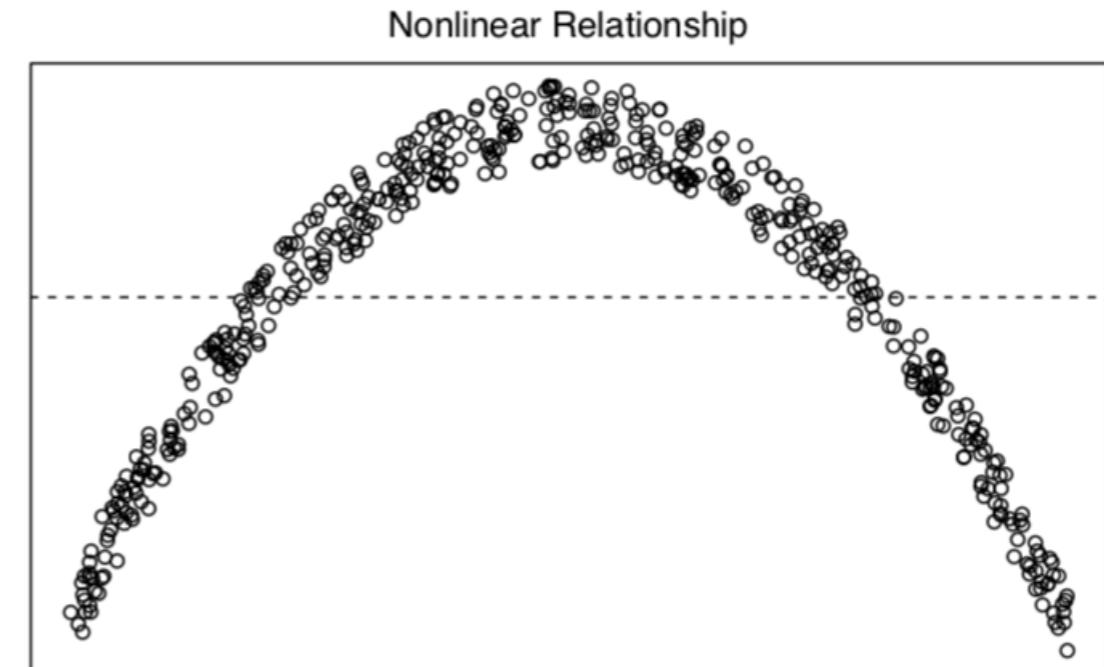
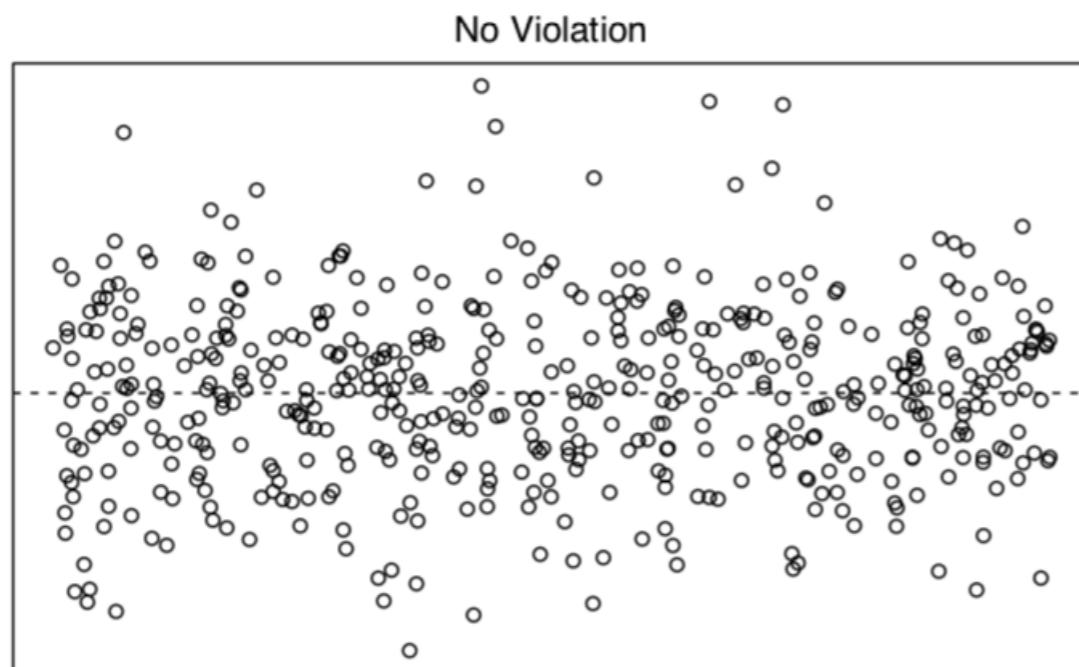
- independent observations
- Y is continuous
- errors are normally distributed
- errors have constant variance
- error terms are uncorrelated



Model assumptions



Model assumptions

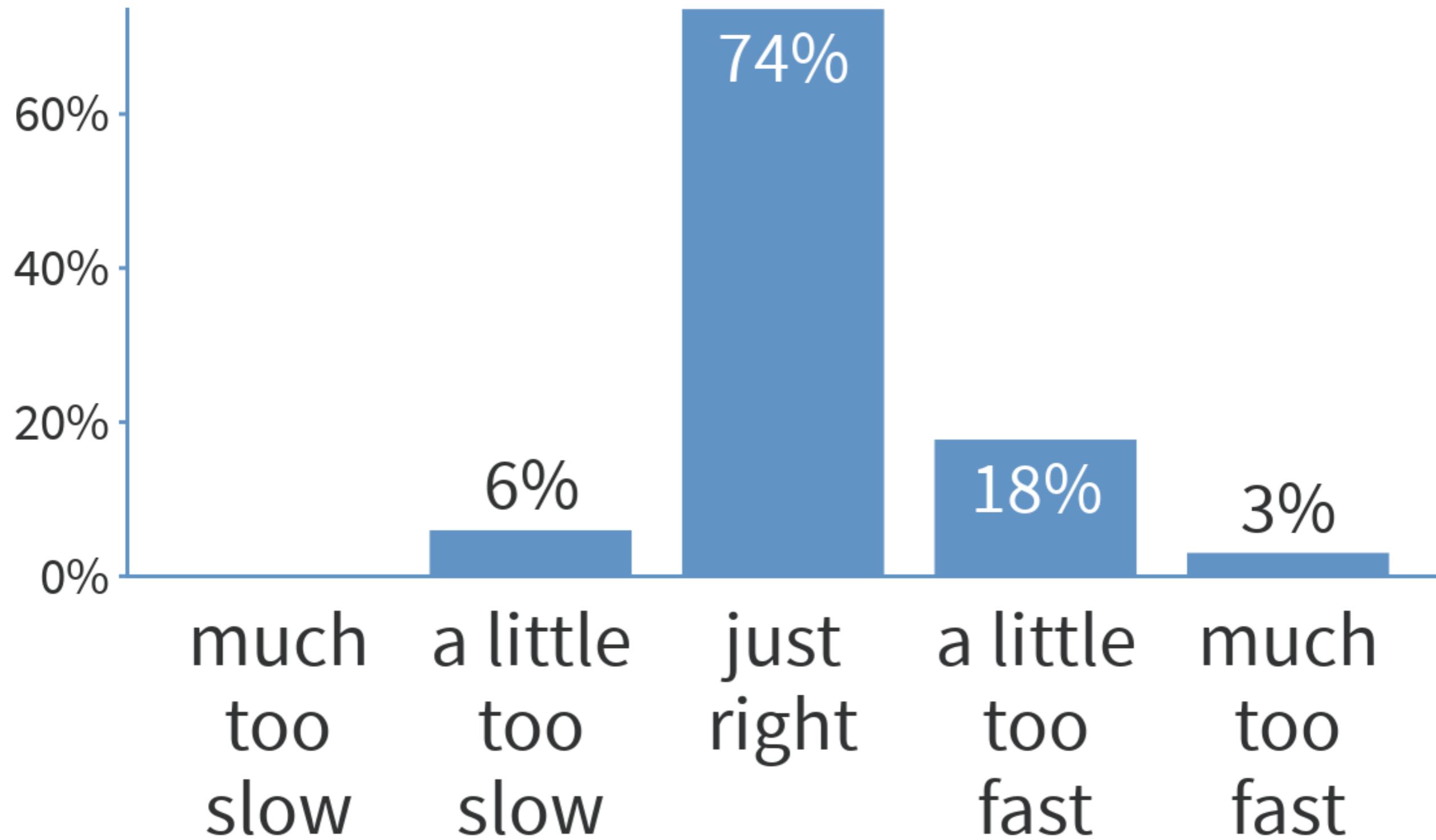


Summary

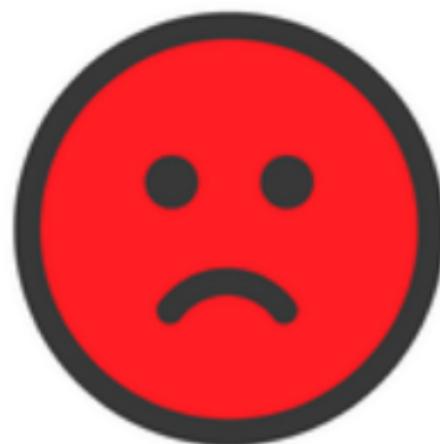
- Quick recap
- Modeling data
- Correlation
 - Pearson's moment correlation
 - Spearman's rank correlation
- Regression
 - The conceptual tour
 - The R route

Feedback

How was the pace of today's class?



How happy were you with today's class overall?



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

cover
process
foreground
important
limit
lead
junk
understand
explain
interaction
better
poll
questions
concepts
class
really
lecture
interaction
time
fully
us
fix
assume
low
clear
well
tobi
things

Thank you!