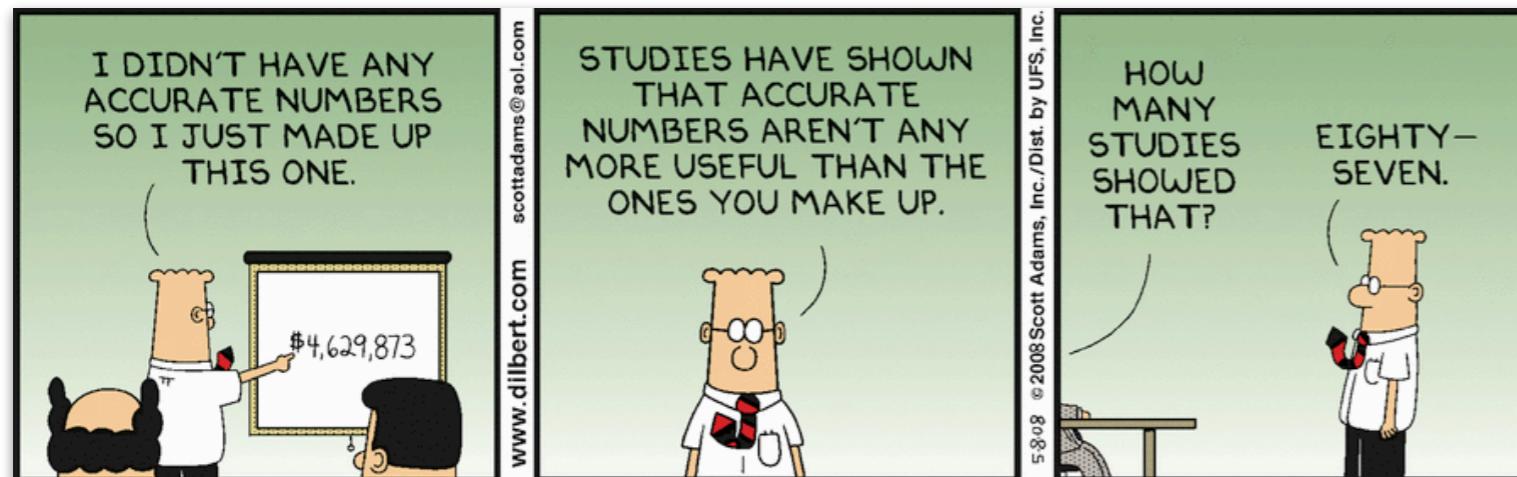


# Linear model 2



We're listening to "Pool Hall" by "The Wave Pictures" submitted by Tobi

Chat

What are your plans for the weekend?

To: Everyone ▾

Type message here...

More ▾

COLLABORATIVE PLAYLIST

**psych252**

<https://tinyurl.com/psych252spotify22>

PLAY

01/28/2022

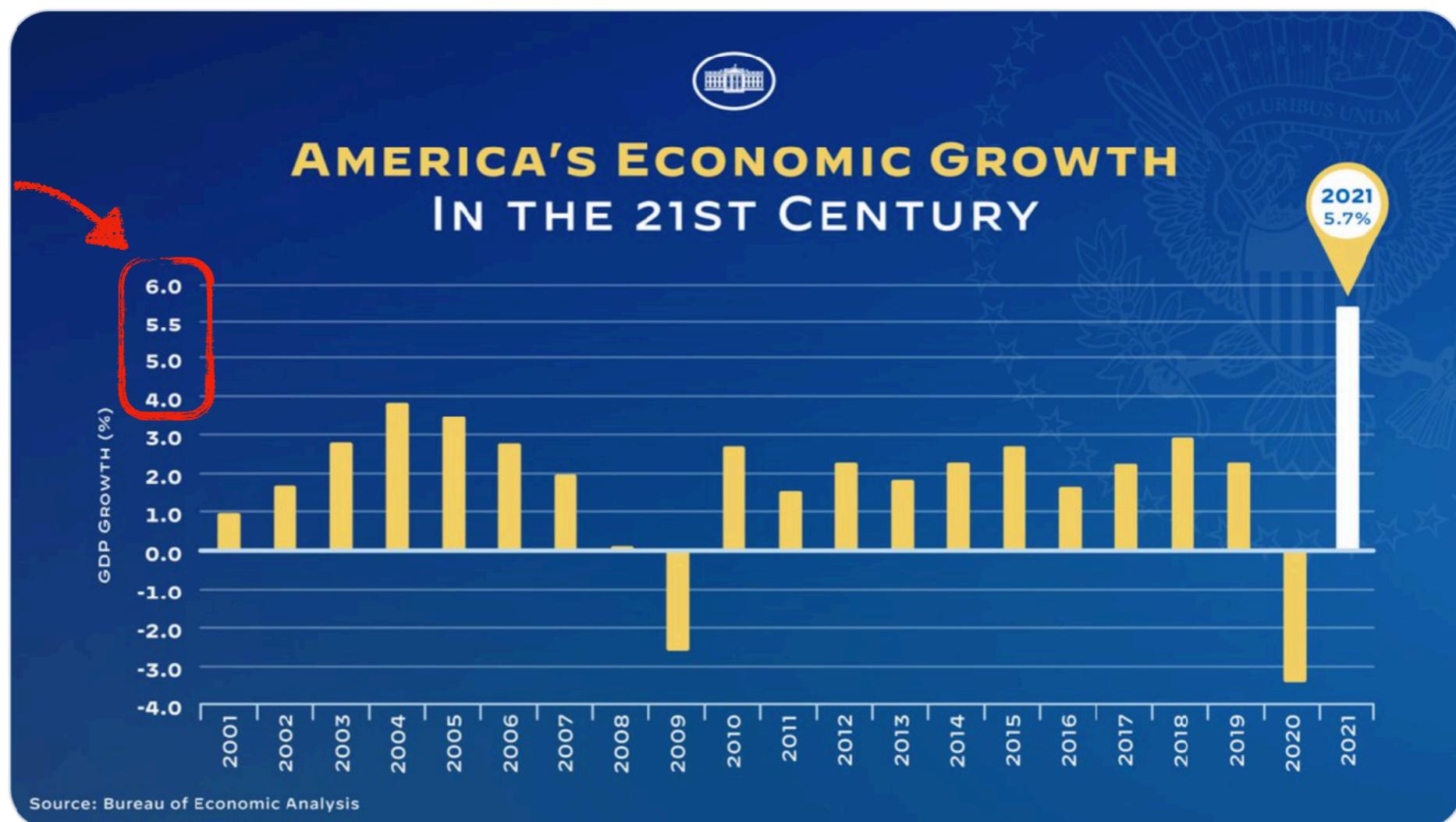
# **Things that came up**

# Graph crimes



The White House @WhiteHouse

We just learned that President Biden's first year in office was the strongest year for economic growth since 1984.



6:35 AM · Jan 27, 2022 · The White House

5,011 Retweets

1,779 Quote Tweets

17.4K Likes

# Graph crime drama continues



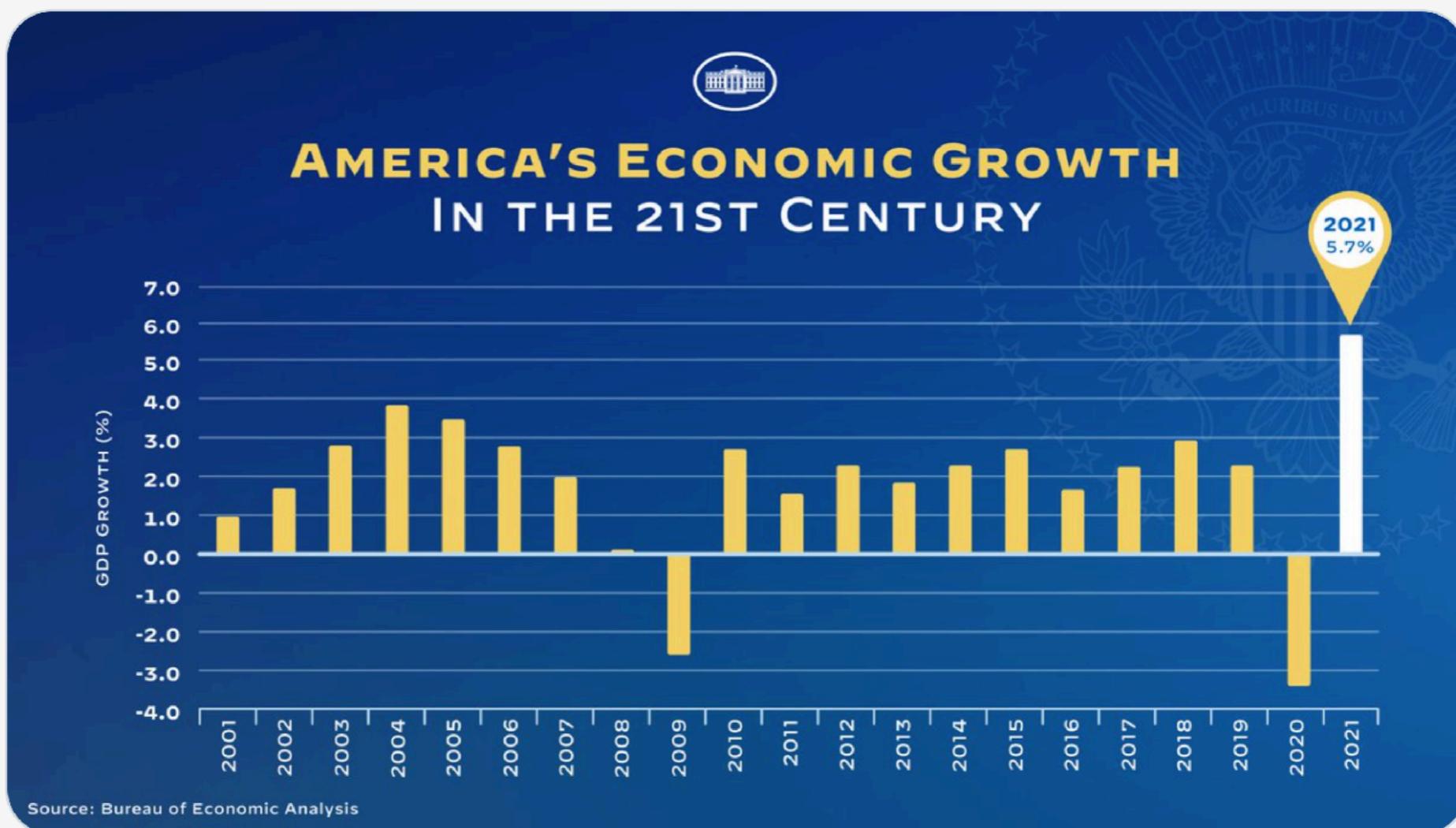
The White House ✅ @WhiteHouse · 6h

Replying to @WhiteHouse

This is y you proofread.

...

If you want a corrected chart to share the news of our record economic growth, use this one instead.



310



619



2.2K



# Graph crime drama continues

## ggplot2: change y axis interval but only for part of graph

Asked 1 day ago Active 1 day ago Viewed 340 times

Ok this is hypothetical. So hypothetically I have a data frame of years and GDP growth numbers. Hypothetically, let's say I want the y axis interval to be 1 everywhere, except at towards the top, where I want the y axis interval to be 0.5, **but still equally space so it seems like it's still 1 apart**. So I want to decrease the interval for 2 ticks (just to make numbers look better).  
Hypothetically.

```
gdp <- data.frame( # Hypothetical example, not real data, demonstration only
  year = c(2018, 2019, 2020, 2021)
  growth = c(4.5, 5.0, 5.4, 7.5)
)
ggplot(gdp, aes(x = year, y = growth)) +
  geom_line() +
  scale_y_continuous(limits = c(1, 6), breaks = 1:6)
# How to make breaks at 5, 5.5, 6, but the space between the ticks is the same as
# the others at 1, 2, 3, 4, 6
```

Please answer soon my boss is waiting

r ggplot2 graph scales

Share Edit Follow Close Flag

edited Jan 26 '22 at 3:34



elemnyf

24.5k 5 42 82

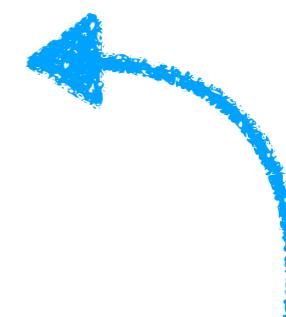
asked Jan 26 '22 at 12:36



NotWH

379 4 19

6 ↑ I don't even think this is possible in ggplot? Why would you want to do this? – pofebas Jan 26 '22 at  
16:42



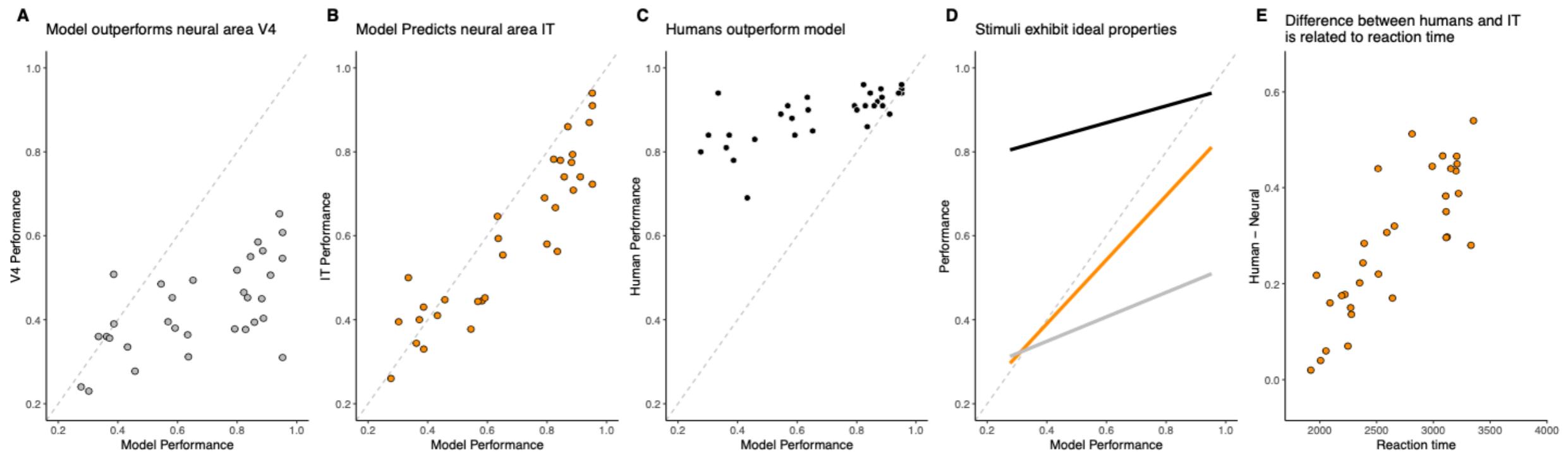
pretty sure this was a joke :)

# **Logistics**

# **Homework 2**

# Homework 2

Grades are posted, and solutions are on Canvas.



# **Homework 3**

# How many hours did it take you to complete Homework 3?



# **Homework 4**

# Homework 4

My name goes here

The names of the people I have worked with go here

2022-01-27 21:27:01

This homework is due by **Thursday, February 3rd, 8:00pm**. Upload a pdf file to Canvas called `4_modeling_data.pdf`

## I. Simple linear regression and prediction

In this section and the next, we'll be revisiting the credit dataset.

```
# Load data
df.credit = read_csv("data/credit.csv") %>%
  clean_names()
```

### 1. Explore and visualize the data

Use the `ggpairs()` function from the `GGally` package to make scatterplots of all pairwise combinations of the numeric variables.

Tip: use `progress = FALSE` within the function to suppress unwanted progress bars of each plot

```
### YOUR CODE HERE ###
```

```
#####
```

That's perhaps a bit too congested to be useful. Recreate the plot, focusing on just ~5 of the variables that seem interesting to you.

```
### YOUR CODE HERE ###
```

```
#####
```

Does your plot tell you anything interesting about the data? Briefly describe one observation.

YOUR ANSWER HERE

### 2. Simple linear regression

You decide you want to test the relationship between credit limit (`limit`) and average credit card debt (`balance`).

a) First, let's visualize this relationship. Create a scatterplot of `balance` (on the y-axis) as a function of `limit` (on the x-axis). Set the transparency of the points to 0.3.

```
### YOUR CODE HERE ###
```

## II. Multiple linear regression and controls

In psychological research, people often run linear regressions in which the goal is to assess the relationship between two variables while "controlling" for other variables. These control variables could, for example, be age and gender. But how should we decide whether and which variables to control for? In this exercise, we will see what potential effects controlling for variables can have in different situations.

3

Now you are interested in whether age is a significant predictor of credit limit.

### 4. Interpreting model parameters

a) Build a simple linear regression model to predict `limit` from `age`. Is age a significant predictor of credit limit?

```
# Simple regression without control variables
### YOUR CODE HERE ###
fit.lm1 =
#####
```

YOUR ANSWER HERE

Then you realize that age is actually related to income, which is a strong predictor of credit limit, so you are interested in seeing whether age is related to credit limit controlling for income.

b) Build a multiple regression model to predict `limit` using both `age` and `income` as predictors. Is age still a significant predictor? What could be an explanation for the change, if there was any?

```
# Multiple regression with control variables
### YOUR CODE HERE ###
fit.lm2 =
fit.lm3 =
#####
```

YOUR ANSWER HERE

## III. Interactions

We will be using the following dataset.

`families.csv`:

Data from a study of 68 companies, examining relationships between the quality of family-friendly programs at each company, the percentage of employees with families who use these programs, and employee satisfaction (all continuous variables).

# Maybe find a homework buddy to team up with!?



**Xi Jia** 7:33 PM

Gonna host Co-working virtual meetings every M/W after class 7:30-8:30;  
people can stay longer if they like. join if you'd like

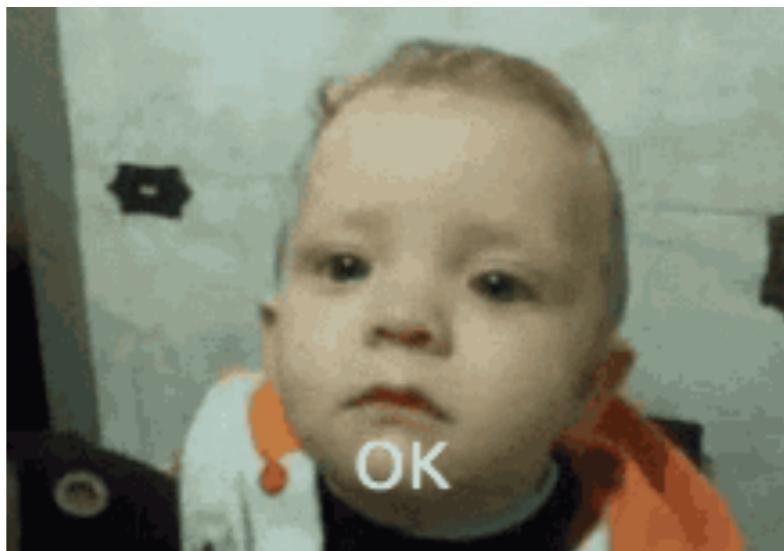
<https://zoom.us/j/2278618193?pwd=a2VtazYzbHIYOVc5MXZ5dENLL2FVZz09>

Link for today

# **Next week**

# Next week

- no class on Friday (I'm away ...)
- I will record the lecture and share the recording with you



# **Midterm**

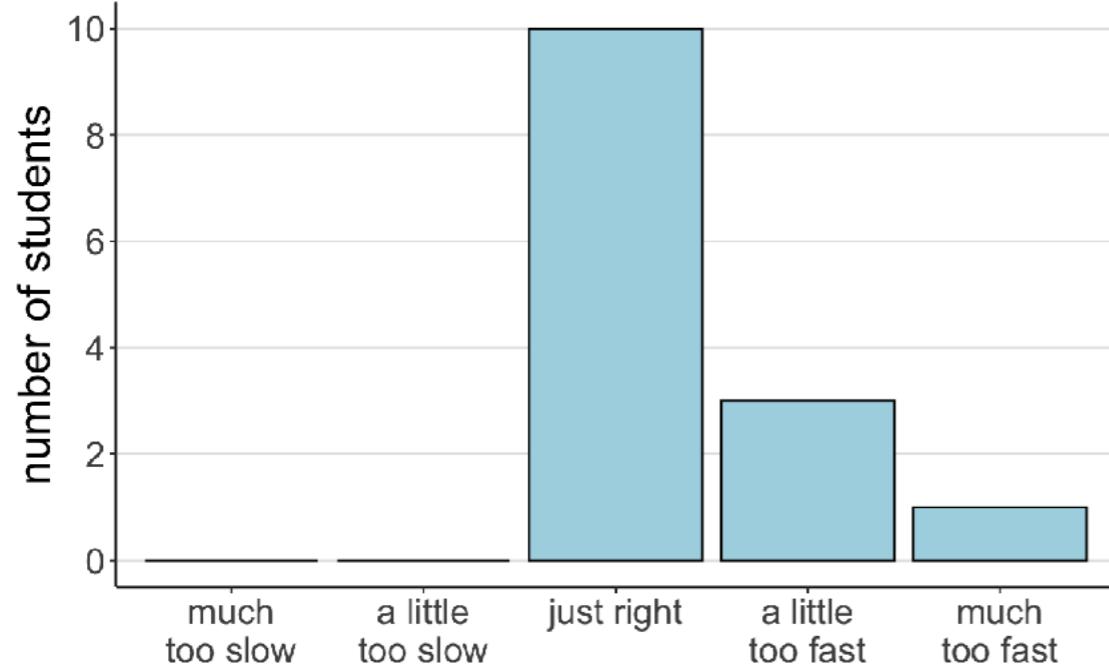
# Midterm

- will be released Friday next week
- midterm is take-home, open exam (you can use all the course notes, all of the internet, etc.)
- is due on **Thursday February 10th at 8pm** (stretch days can't be used for the midterm)
- midterm is like the other homeworks but:
  - it's longer (there won't be class on Wednesday February 8th)
  - you'll need to work on it on your own
- there won't be office hours or sections in the week that the midterm is due
- ask questions about the midterm on Ed Discussion (please ask private first -- we will then respond and, if appropriate, check with you whether we can share with the rest of the class)

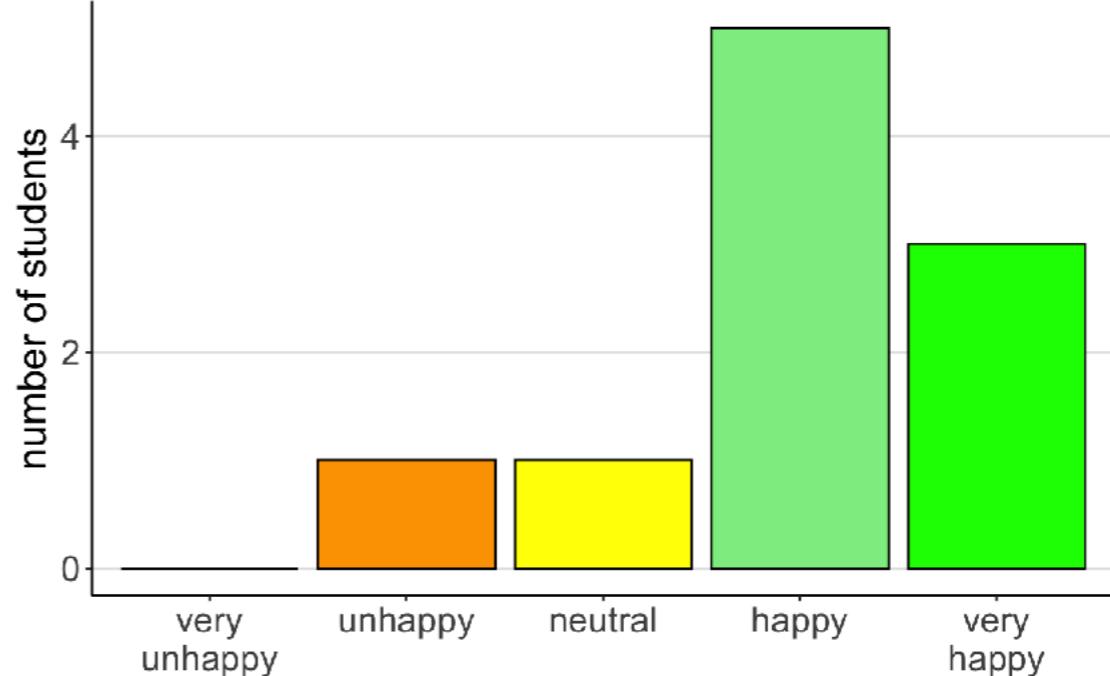
# **Feedback**

# Your feedback

How was the pace of today's class?



How happy were you with today's class overall?



Class today cleared up so many questions I had about pearson vs. spearman, what goes on under the hood of a t-test in R, etc. Plus the game was fun! And I liked the bear eating chocolate on one of your slides – really thankful this course makes learning stats/R feel accessible and fun.

is there space to use our own data set for assignments

The pace of the last 3-4 classes has been great, and the slides help to explain the statistics. However, a huge challenge is how this translates to the code used in R. When I am reviewing the rmd files it has been difficult for me to know what each of the commands and arguments are doing, and how they are linked to the statistics taught in the class. To help with this, could you please provide more notes in the rmd files that explains what the commands and arguments are doing.

Let us try/ see the process of how things are derived

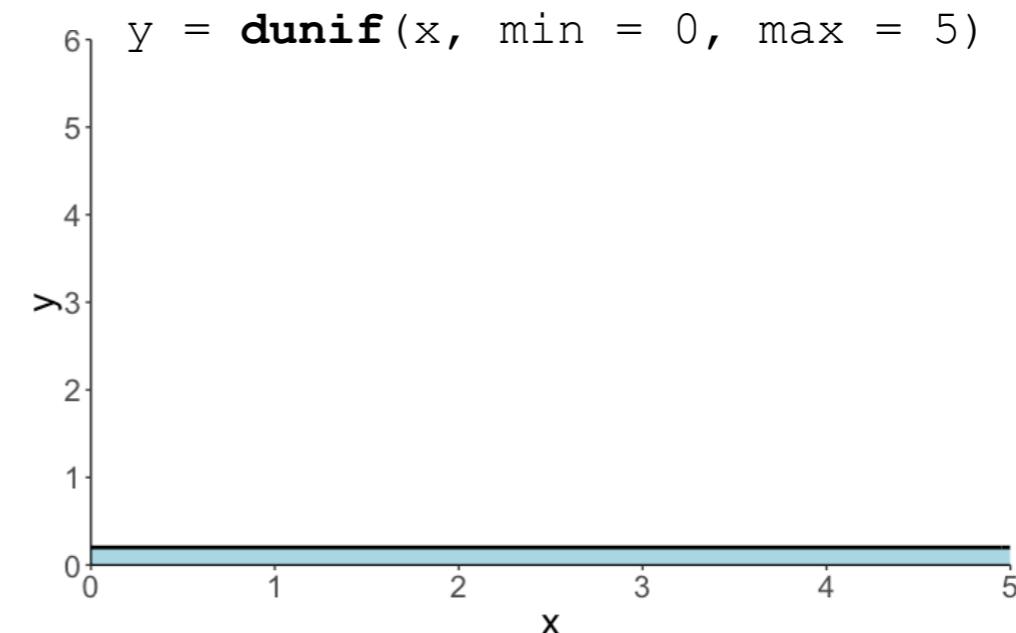
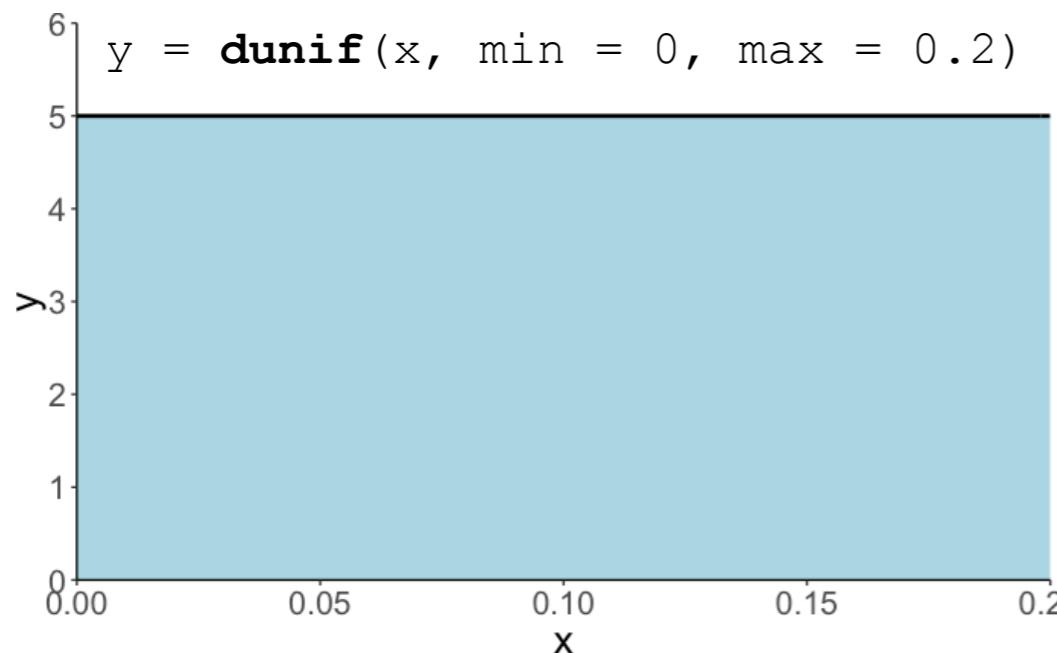
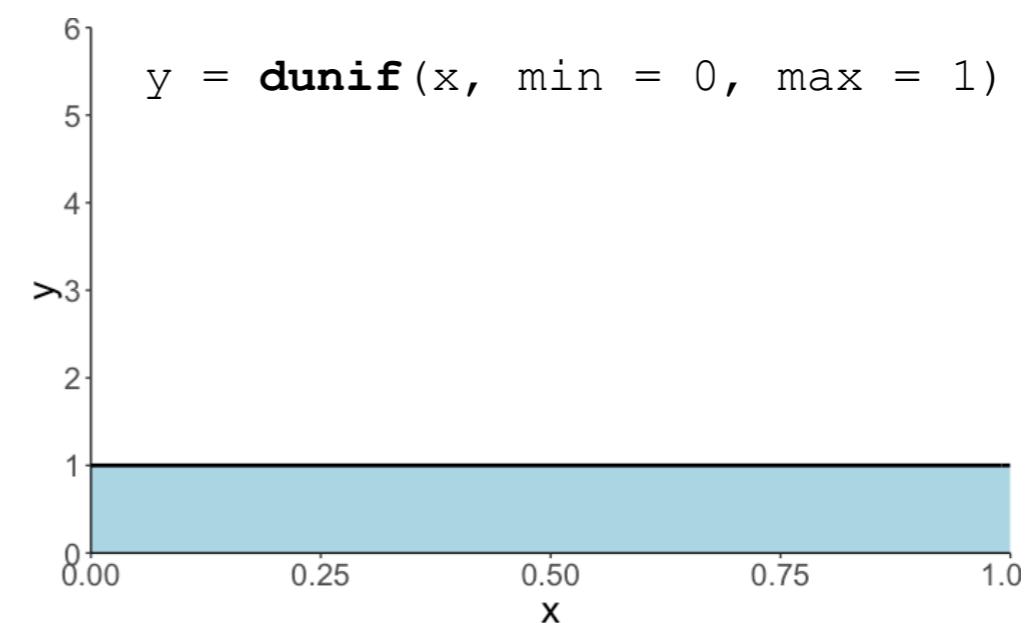
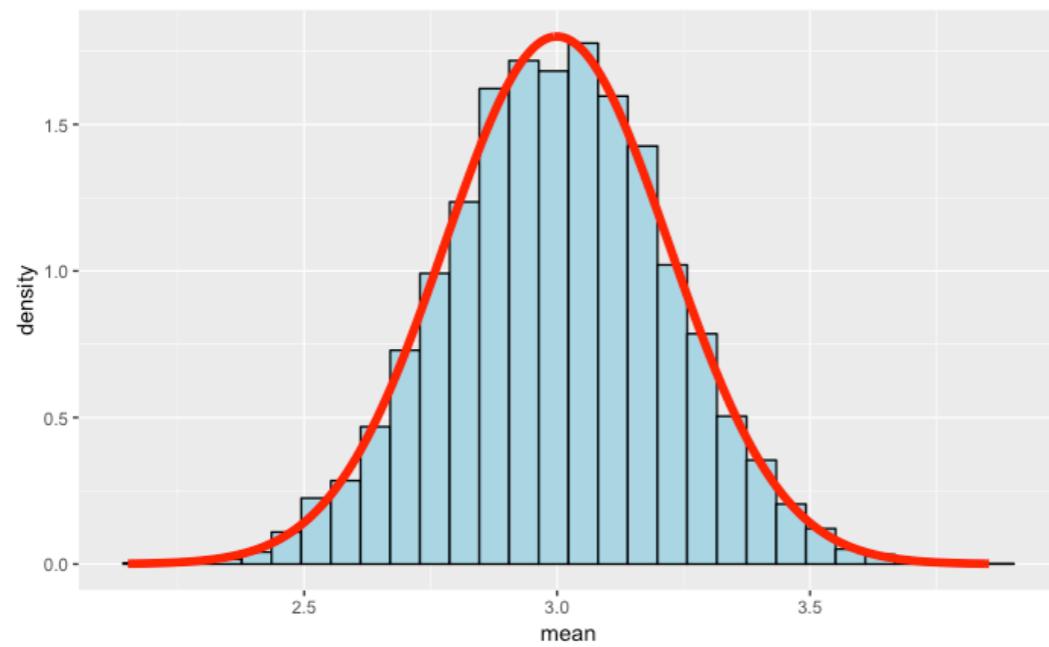
# Plan for today

- Quick recap
- Regression
  - The conceptual tour
  - The R route

# Quick recap

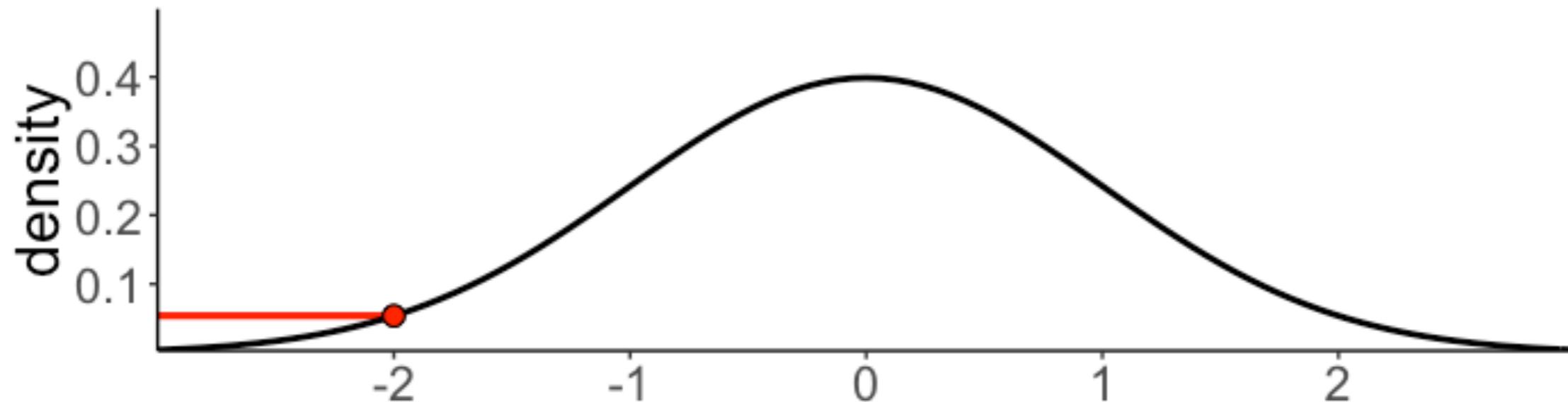
# Quick recap: Probabilities

My understanding is that the area under the density curve should add up to 1, or 100%. I'm very confused when I plot the density of the rating variable and see that the y-axis has labels that are higher than 1. How can density take on a value higher than 1? Thank you!

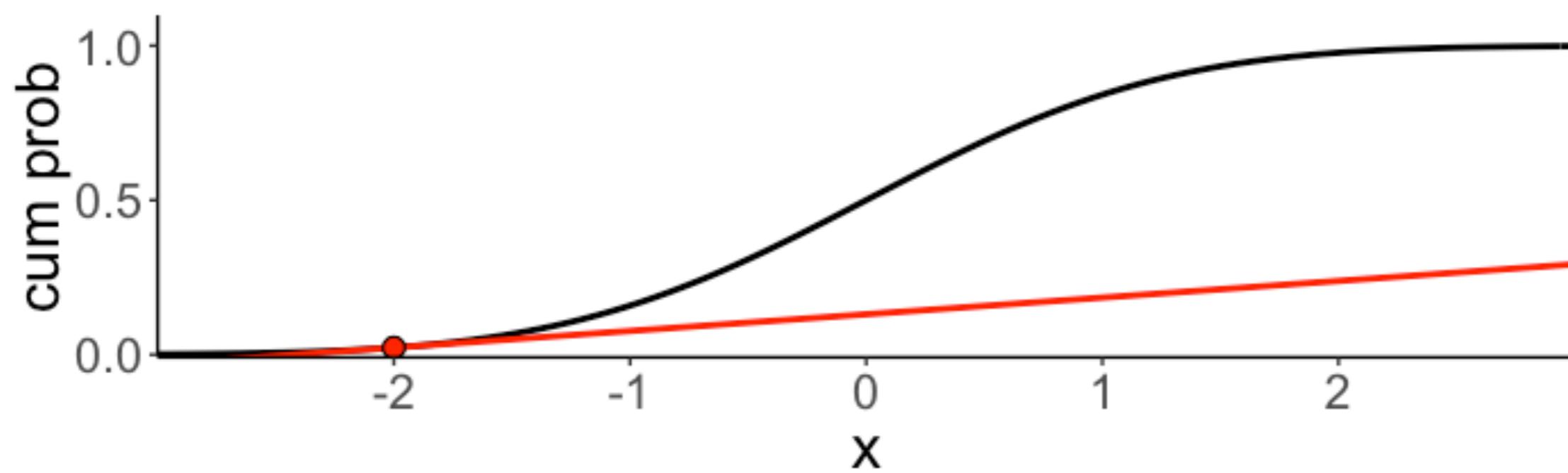


# Quick recap: Probabilities

**dnorm ( )**



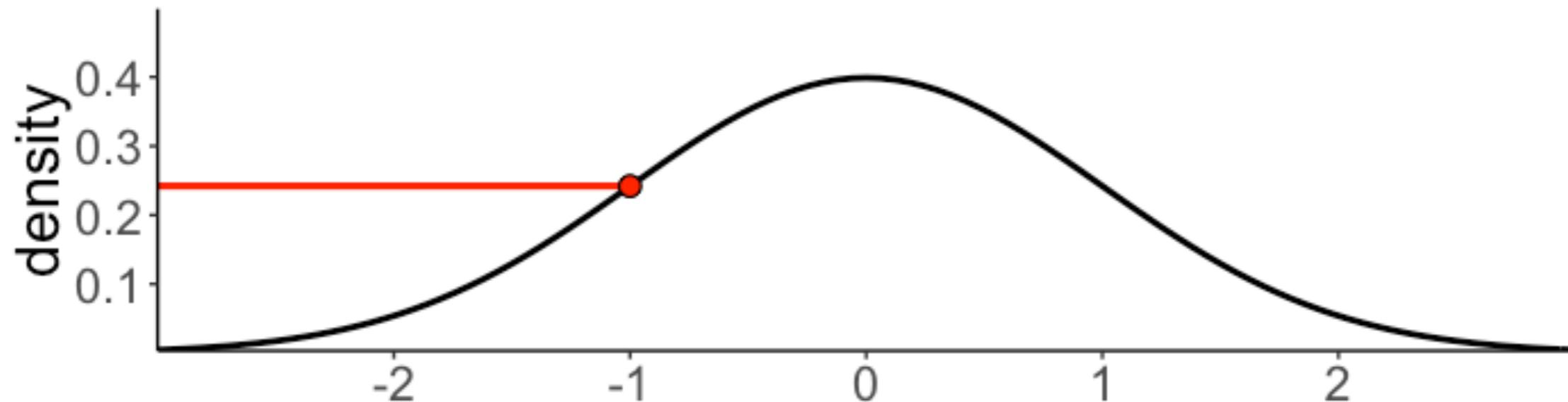
**pnorm ( )**



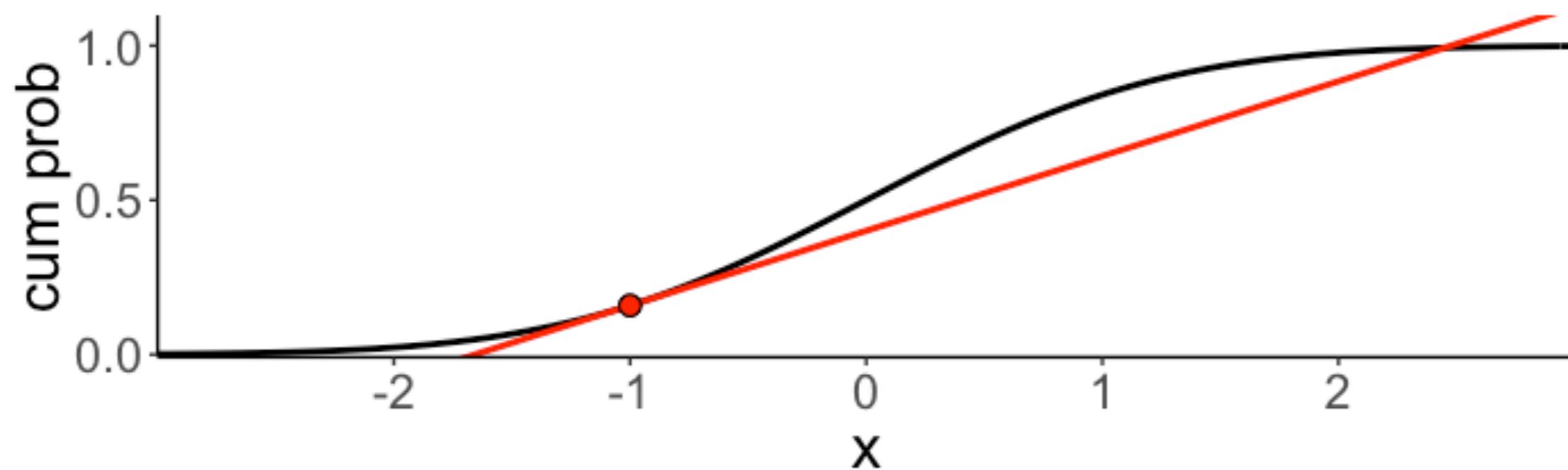
**dnorm ( )** is the first derivative of **pnorm ( )**

# Quick recap: Probabilities

**dnorm ( )**



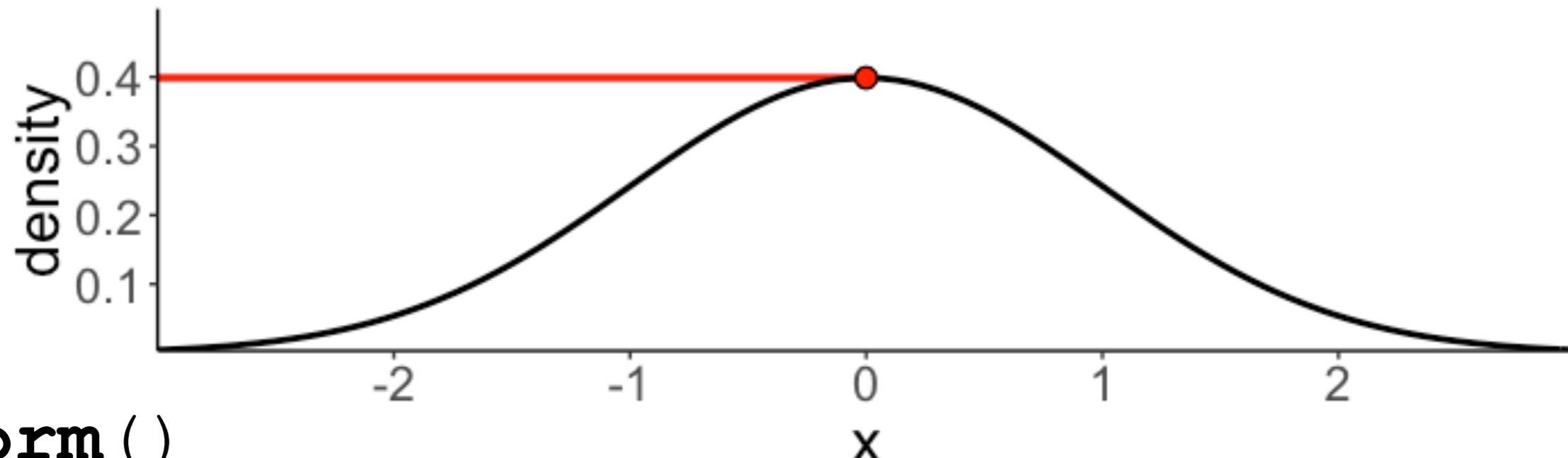
**pnorm ( )**



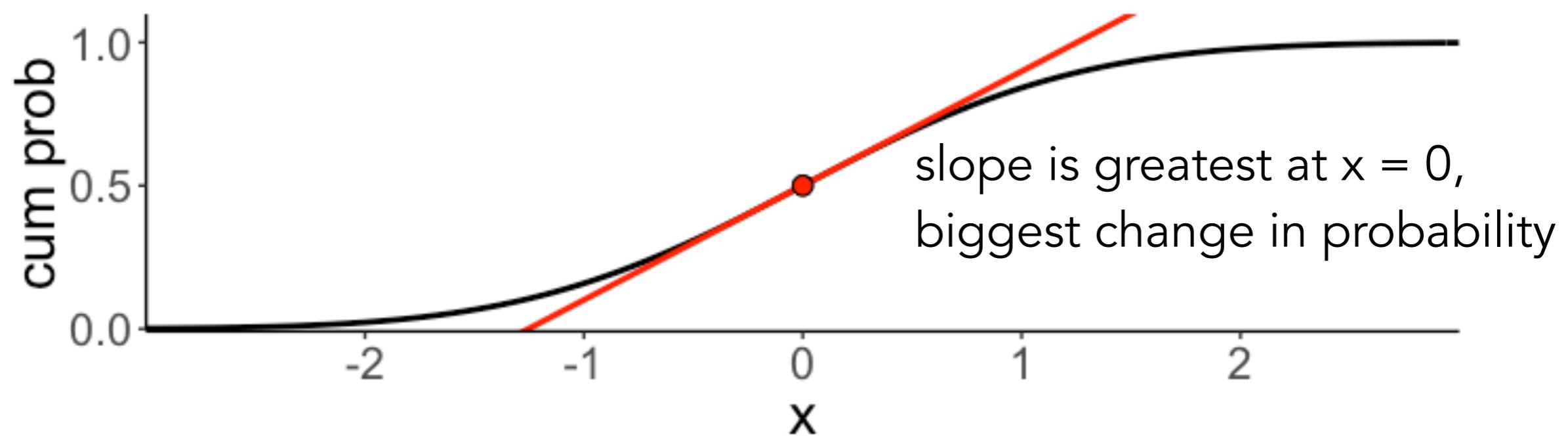
**dnorm ( )** is the first derivative of **pnorm ( )**

# Quick recap: Probabilities

**dnorm ( )**

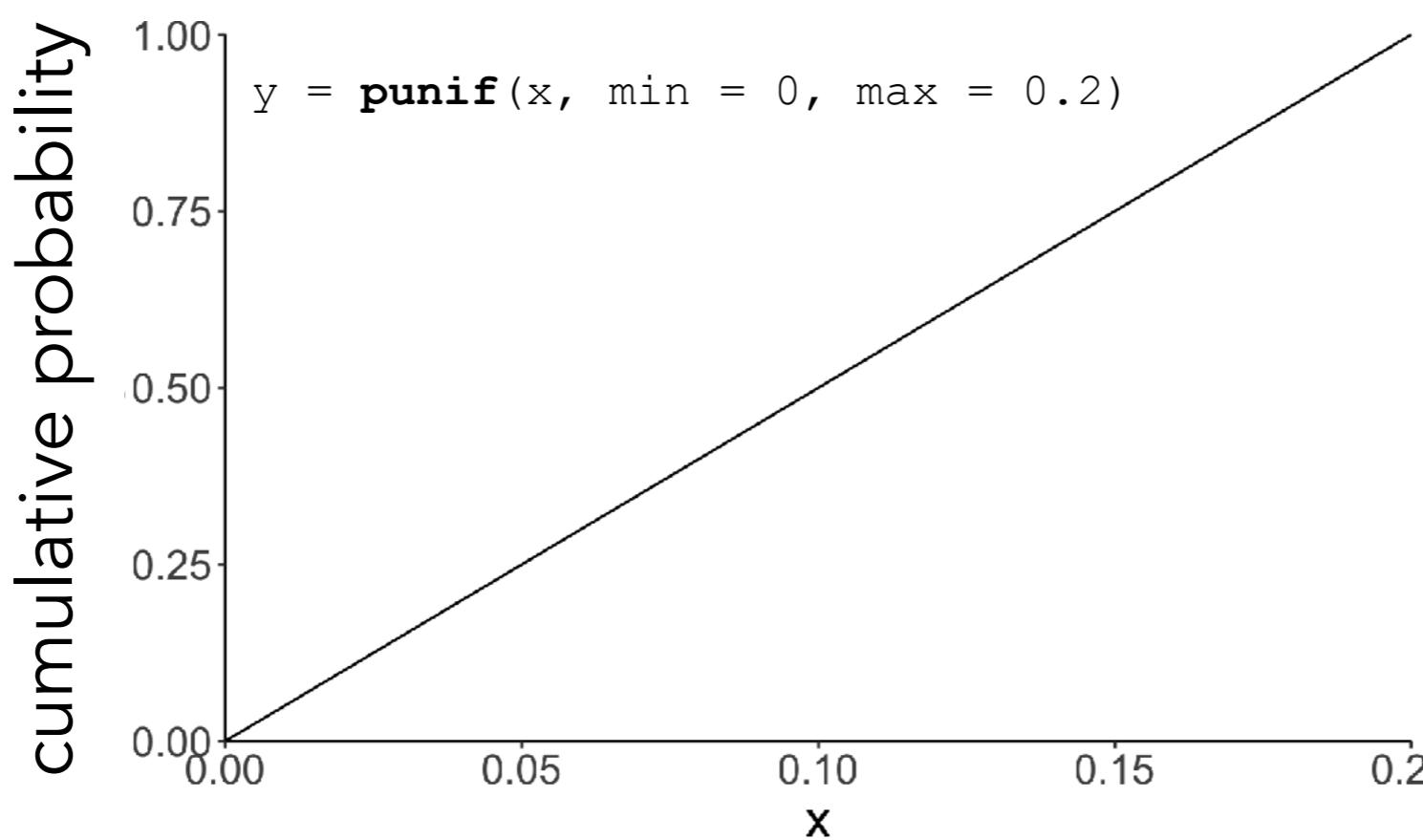
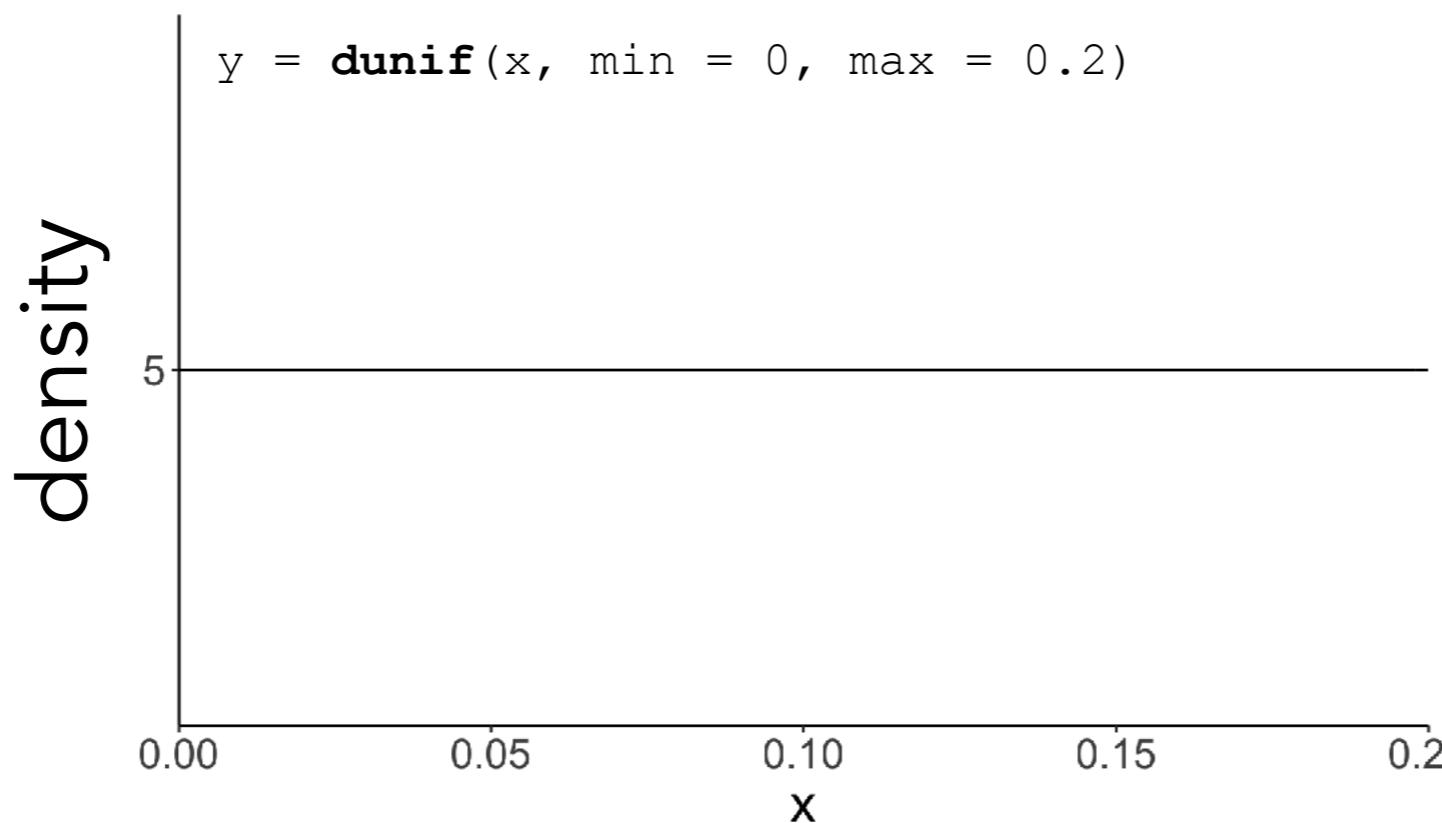


**pnorm ( )**

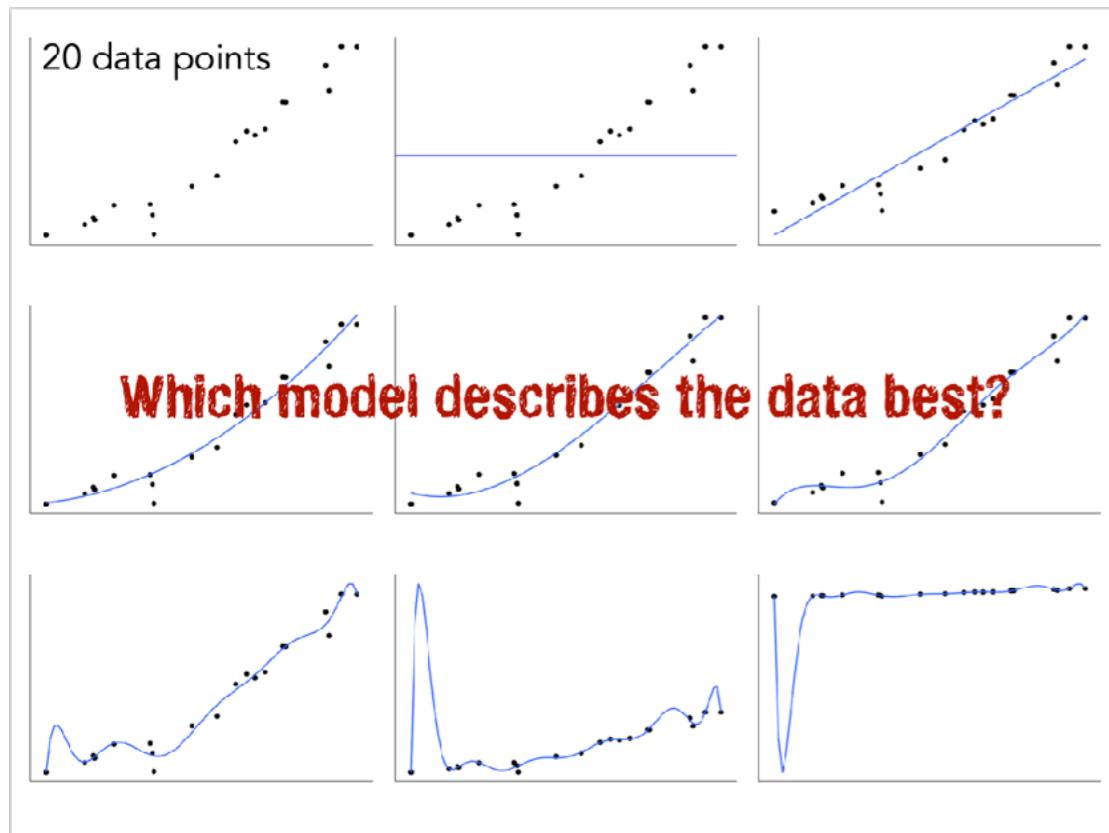


**dnorm ( )** is the first derivative of **pnorm ( )**

# Quick recap: Probabilities



# Quick recap: Modeling data



Compact model

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

Augmented model

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

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

Proportional reduction in error (PRE)

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

- 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  $\text{model}_C$  to create  $\text{model}_A$  but were not is high

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

# Quick recap: Modeling data

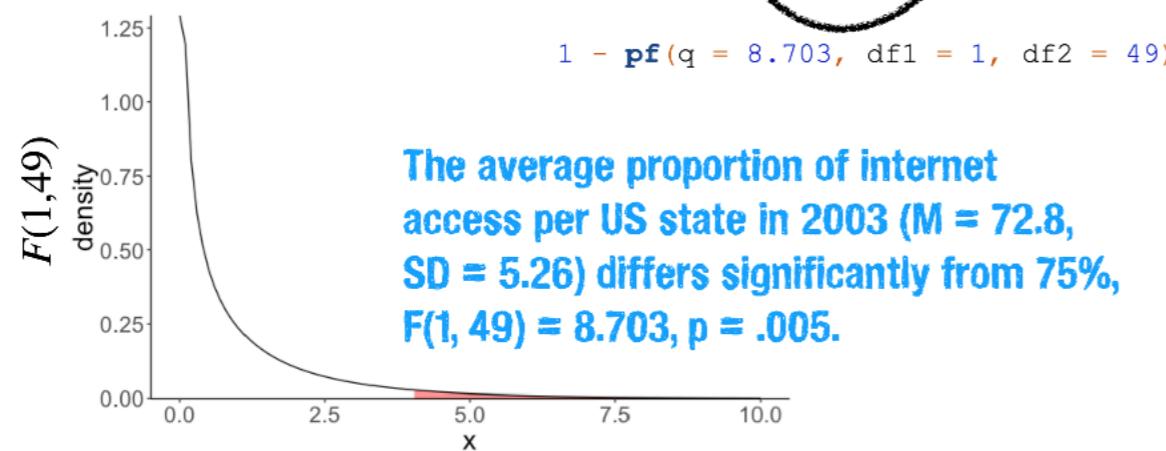
## Procedure

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**

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.



## Research question and hypotheses

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

Model<sub>C</sub>:  $Y_i = B_0 + \epsilon_i$   
**0 parameters**

$$Y_i = 75 + e_i$$

Model<sub>A</sub>:  $Y_i = \beta_0 + \epsilon_i$   
**1 parameter**

$$Y_i = b_0 + \epsilon_i \\ = \bar{Y} + \epsilon_i$$

`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

# **Generating a sampling distribution for PRE**

# 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**

- we have to construct a sampling distribution of PRE assuming that  $H_0$  is true
- and then compare the observed value of PRE to that distribution

## Population distribution

$$Y_i = 75 + \epsilon_i \quad \epsilon_i \sim \mathcal{N}(\mu = 0, \sigma = 5)$$

### Model C

$$Y_i = 75 + e_i$$

0 parameters

### Model A

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

1 parameter

# Sampling distribution of PRE

```
1 # simulation parameters
2 n_samples = 1000
3 sample_size = 50
4 mu = 75 # true mean of the distribution
5 sigma = 5 # true standard deviation of the errors
6
7 # function to draw samples from the population distribution
8 fun.draw_sample = function(sample_size, mu, sigma) {
9   sample = mu + rnorm(sample_size, mean = 0, sd = sigma)
10 }
11
12 # draw samples
13 samples = n_samples %>%
14   replicate(fun.draw_sample(sample_size, mu, sigma)) %>%
15   t() # transpose the resulting matrix (i.e. flip rows and columns)
```

sample	index	number
1	1	75.30
1	2	72.06
1	3	77.66
1	4	67.41
1	5	76.53
1	6	67.32
1	7	73.50
1	8	72.36
1	9	71.74
1	10	74.72

⋮

# Sampling distribution of PRE

```
1 # put samples in data frame and compute PRE
2 df.samples = samples %>%
3   as_tibble(.name_repair = ~ 1:ncol(samples)) %>%
4   mutate(sample = 1:n()) %>%
5   pivot_longer(cols = -sample,
6                 names_to = "index",
7                 values_to = "value") %>%
8   mutate(compact = 75) %>%
9   group_by(sample) %>%
10  mutate(augmented = mean(value))
```

sample	index	value	compact	augmented
1	1	73.43	75	74.75
	2	76.38	75	74.75
	3	79.92	75	74.75
	4	72.33	75	74.75
	5	77.75	75	74.75
2	1	79.84	75	73.92
	2	78.44	75	73.92
	3	79.49	75	73.92
	4	71.81	75	73.92
	5	79.57	75	73.92
3	1	78.99	75	74.93
	2	67.28	75	74.93
	3	77.74	75	74.93
	4	73.73	75	74.93
	5	73.49	75	74.93

# Sampling distribution of PRE

```
1 # put samples in data frame and compute PRE
2 df.samples = samples %>%
3   as_tibble(.name_repair = ~ 1:ncol(samples)) %>%
4   mutate(sample = 1:n()) %>%
5   pivot_longer(cols = -sample,
6                 names_to = "index",
7                 values_to = "value") %>%
8   mutate(compact = 75) %>%
9   group_by(sample) %>%
10  mutate(augmented = mean(value)) %>%
11  summarize(sse_compact = sum((value - compact)^2),
12             sse_augmented = sum((value - augmented)^2),
13             pre = 1 - sse_augmented/sse_compact)
```

calculate SSE  
for each model



calculate PRE

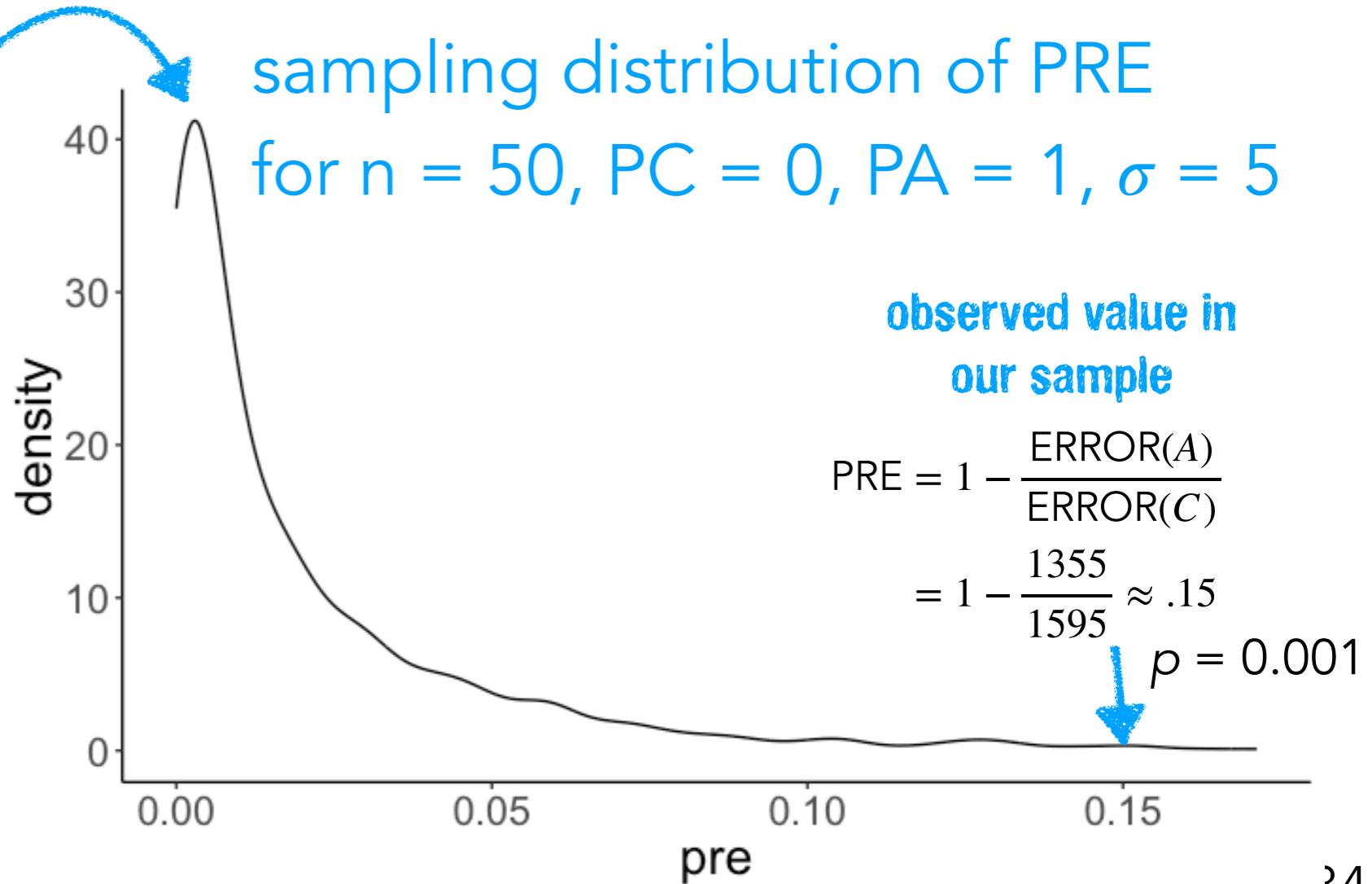
$$\text{PRE} = 1 - \frac{\text{SSE}_A}{\text{SSE}_C}$$

sample	sse_compact	sse_augmented	pre
1	1244.66	1241.52	0.00
2	953.25	894.95	0.06
3	1083.60	1083.32	0.00
4	888.06	798.19	0.10
5	1246.57	1233.25	0.01

# Sampling distribution of PRE

```
29 # sampling distribution for PRE  
30 ggplot(data = df.samples,  
31         mapping = aes(x = pre)) +  
32         stat_density(geom = "line")  
33  
34 # p-value for our sample  
35 df.samples %>%  
36 summarize(p_value = sum(pre >= df.summary$pre) / n())
```

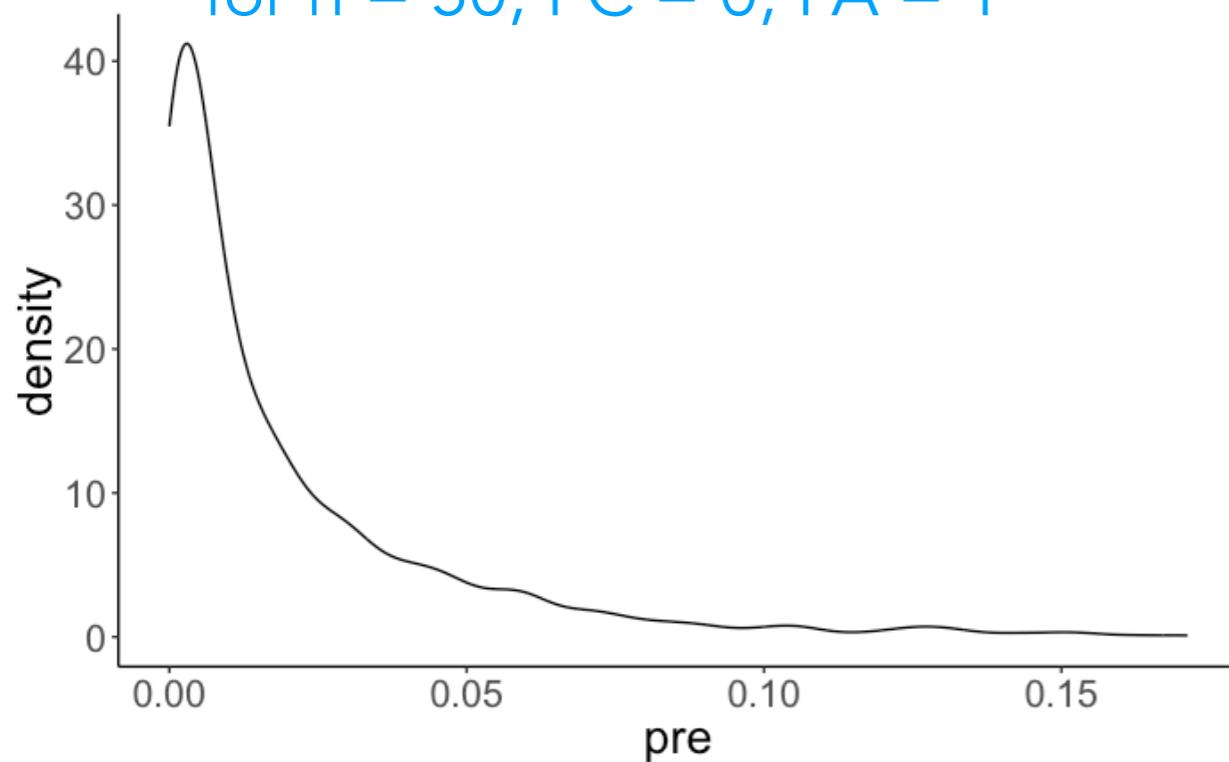
sample	sse_compact	sse_augmented	pre
1	1244.66	1241.52	0.00
2	953.25	894.95	0.06
3	1083.60	1083.32	0.00
4	888.06	798.19	0.10
5	1246.57	1233.25	0.01



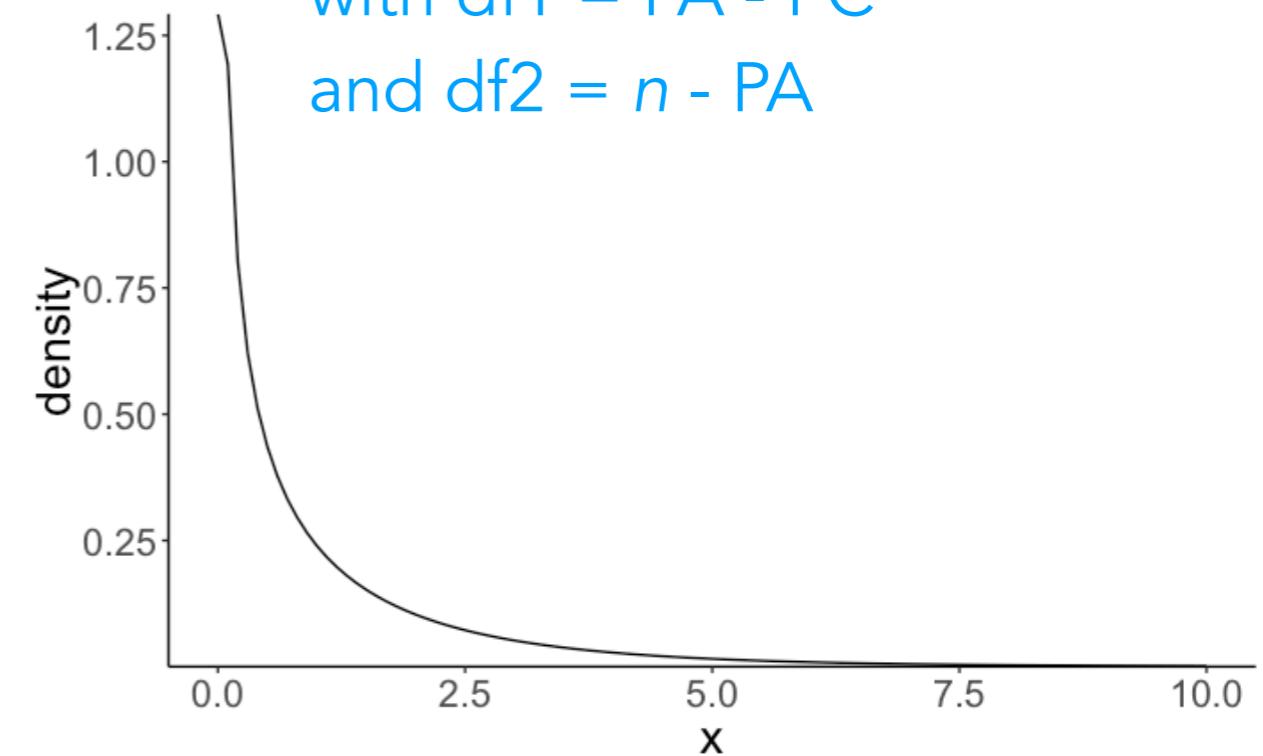
# Sampling distribution of PRE

deterministic mapping

sampling distribution of PRE  
for  $n = 50$ ,  $PC = 0$ ,  $PA = 1$



$F(df1, df2)$  distribution  
with  $df1 = PA - PC$   
and  $df2 = n - PA$

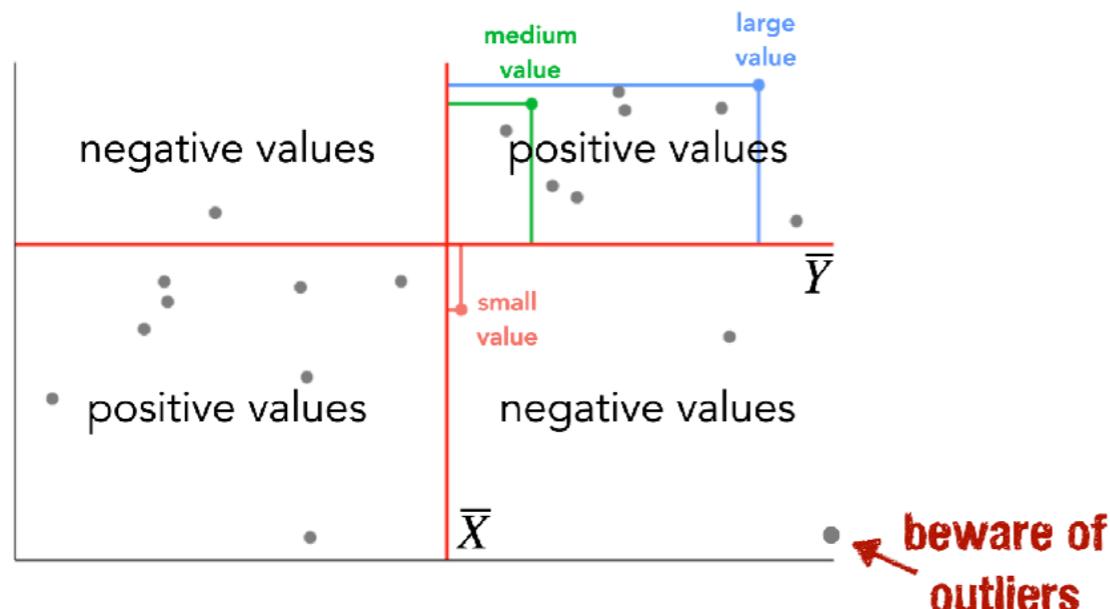


we use the F-distribution since it comes with R (and is the standard statistic to report)

# Quick recap: Correlation

## sample covariance

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



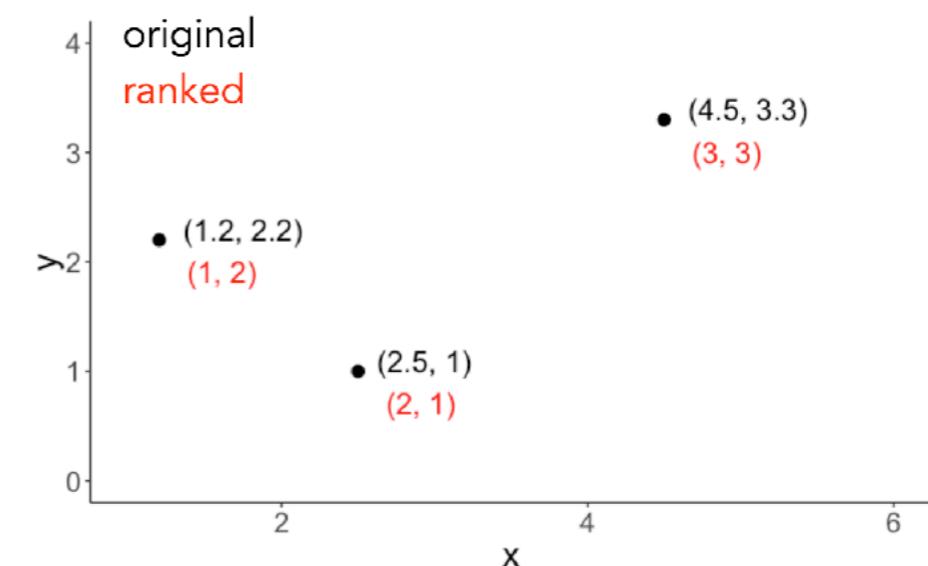
## sample correlation coefficient

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

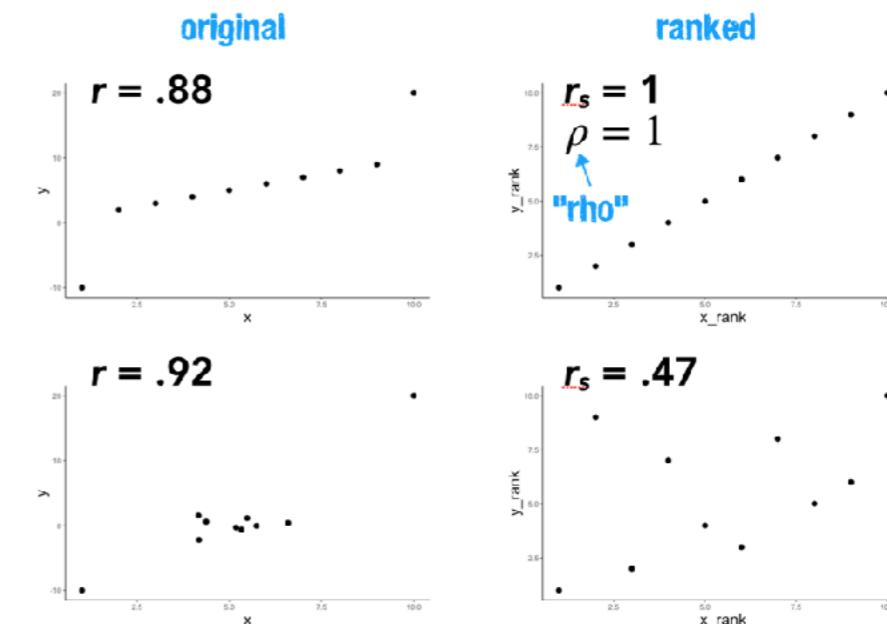
standardized covariation  
(dividing by the standard deviations)

## Spearman rank order correlation

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



## Spearman rank order correlation



# Regression

# The conceptual tour

# Linear model: Simple regression

Data = Model + Error

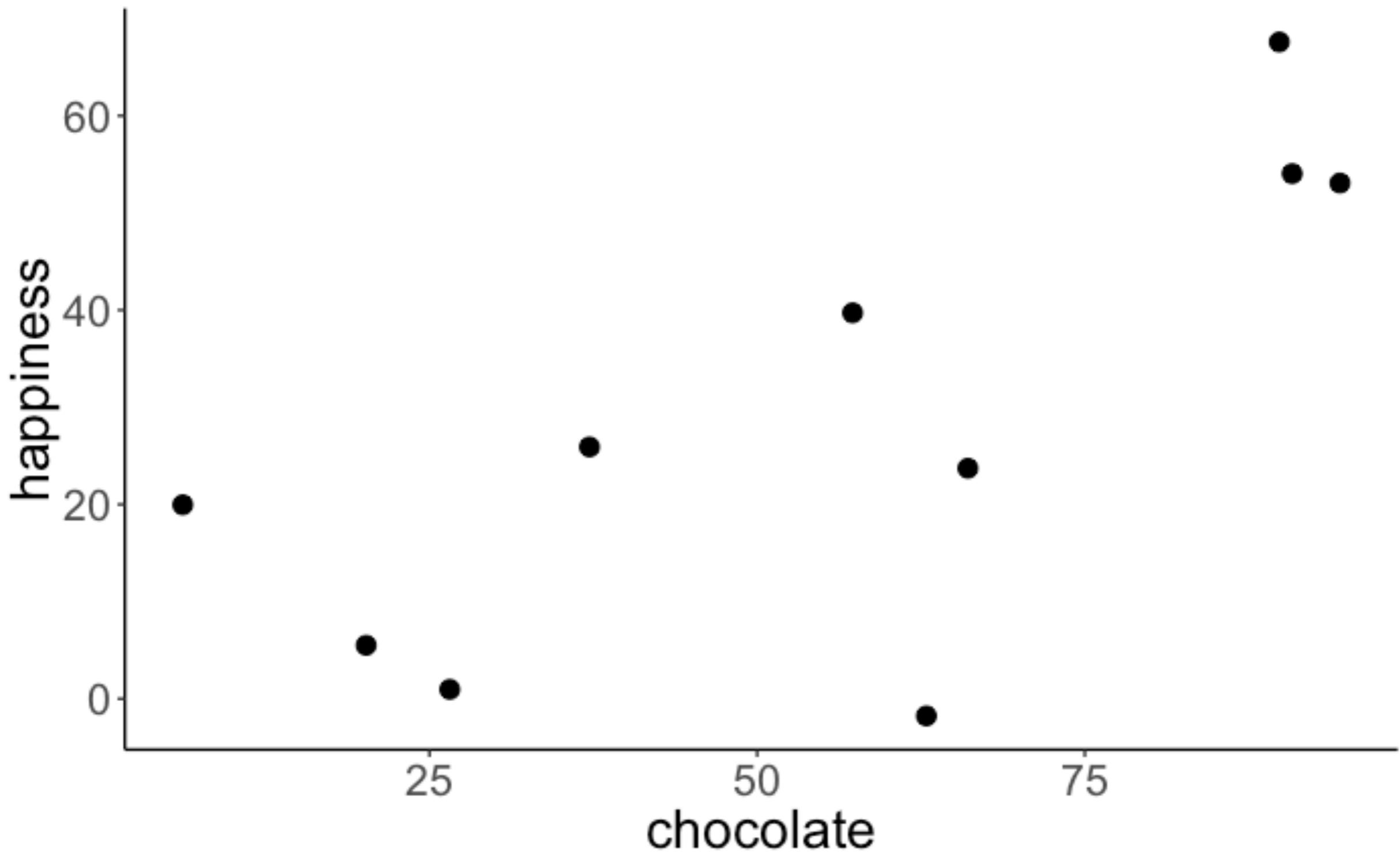
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

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



the model is a linear  
combination of predictors

# 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

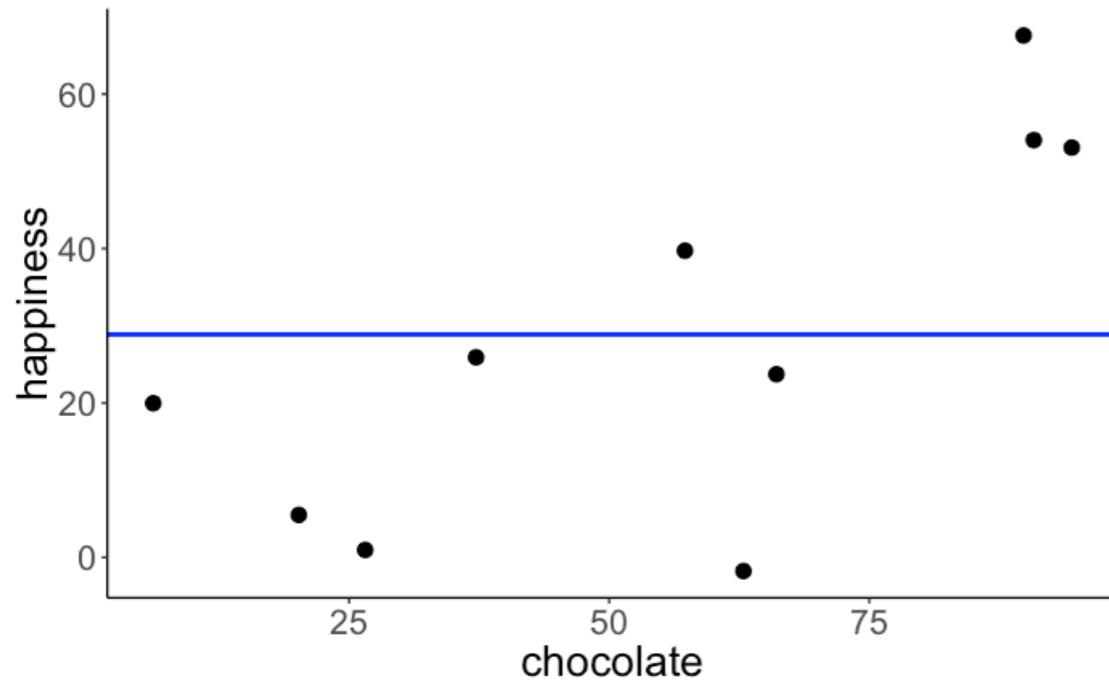
1. Define  $H_0$  as Model C (compact) and  $H_1$  as Model A (augmented)
- 2. Fit model parameters to the data**
3. Calculate the proportional reduction of error (PRE) in our sample
4. Decide whether the augmented model is **worth it** by comparing the observed PRE in our sample to the sampling distribution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

$$Y_i = 28.88 + e_i$$

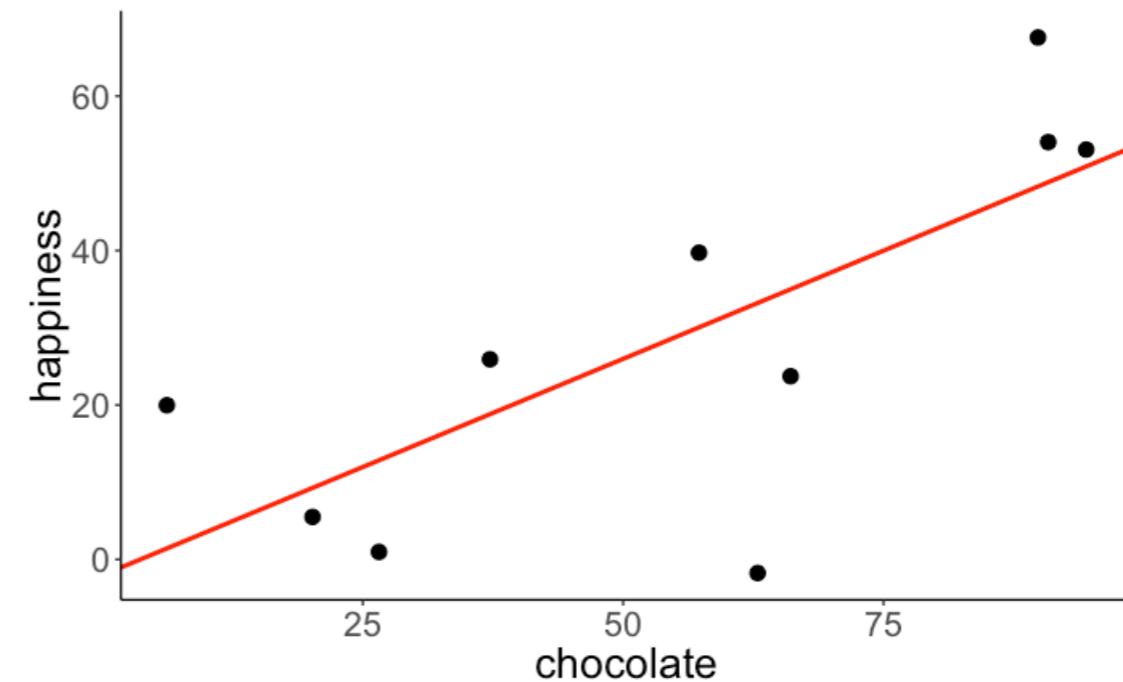
$H_1$ : Chocolate consumption and happiness are related.

### 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)

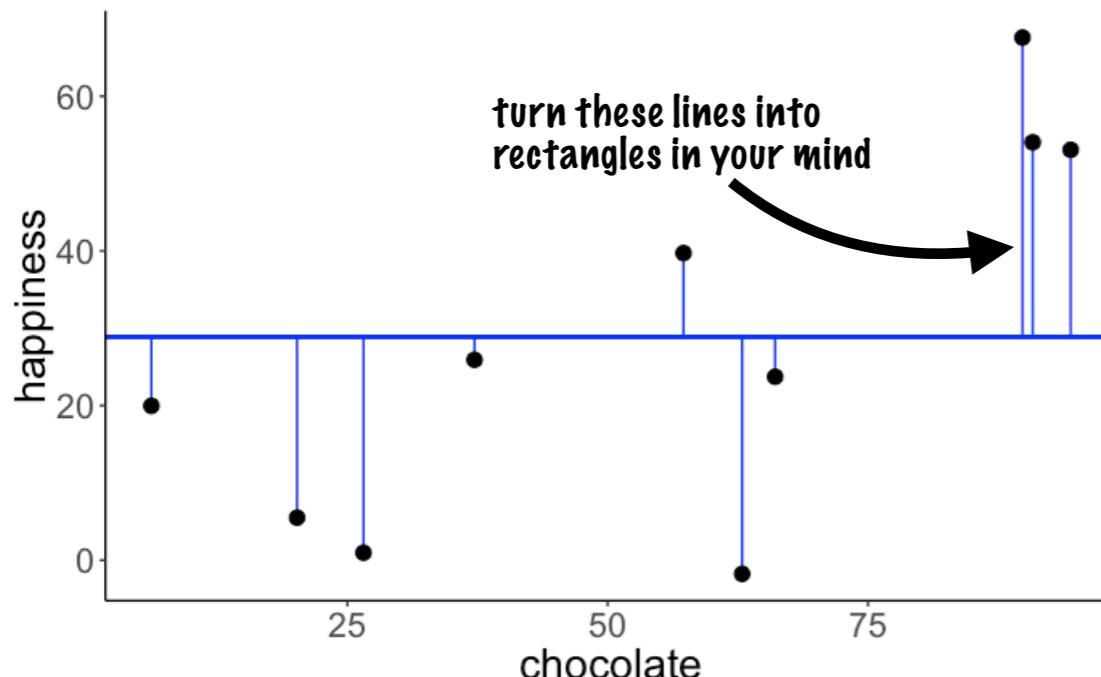
# Calculate PRE

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

Both models were fit to minimize the sum of squared errors

OLS = Ordinary **least squares** regression

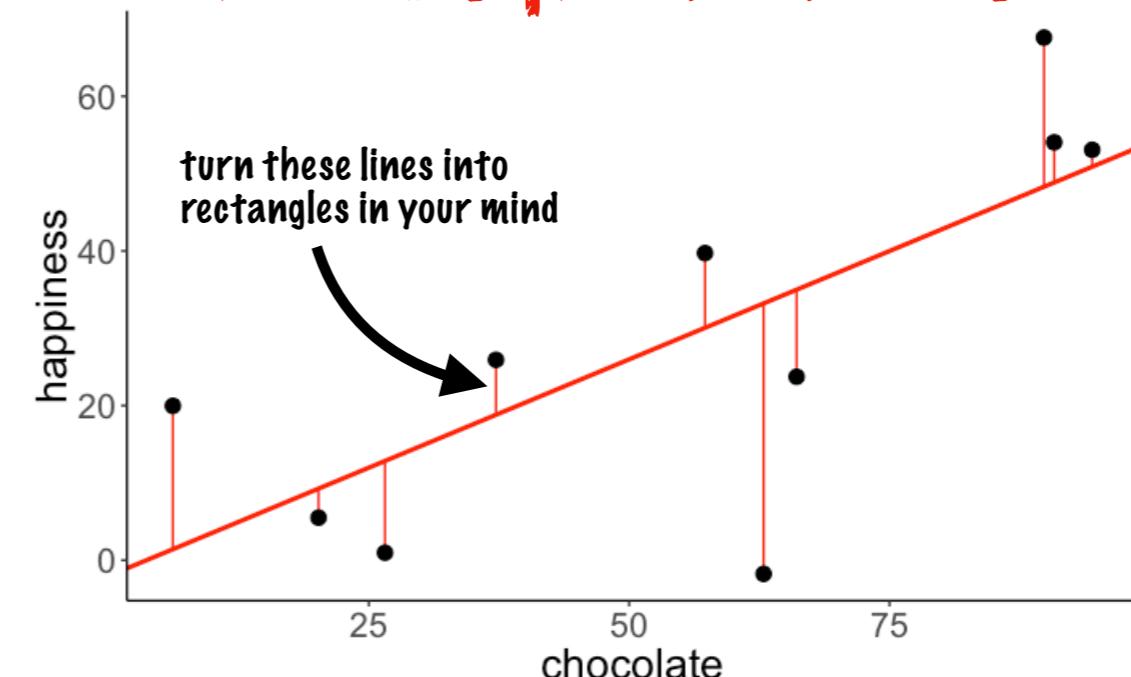
## Sum of squared errors



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

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

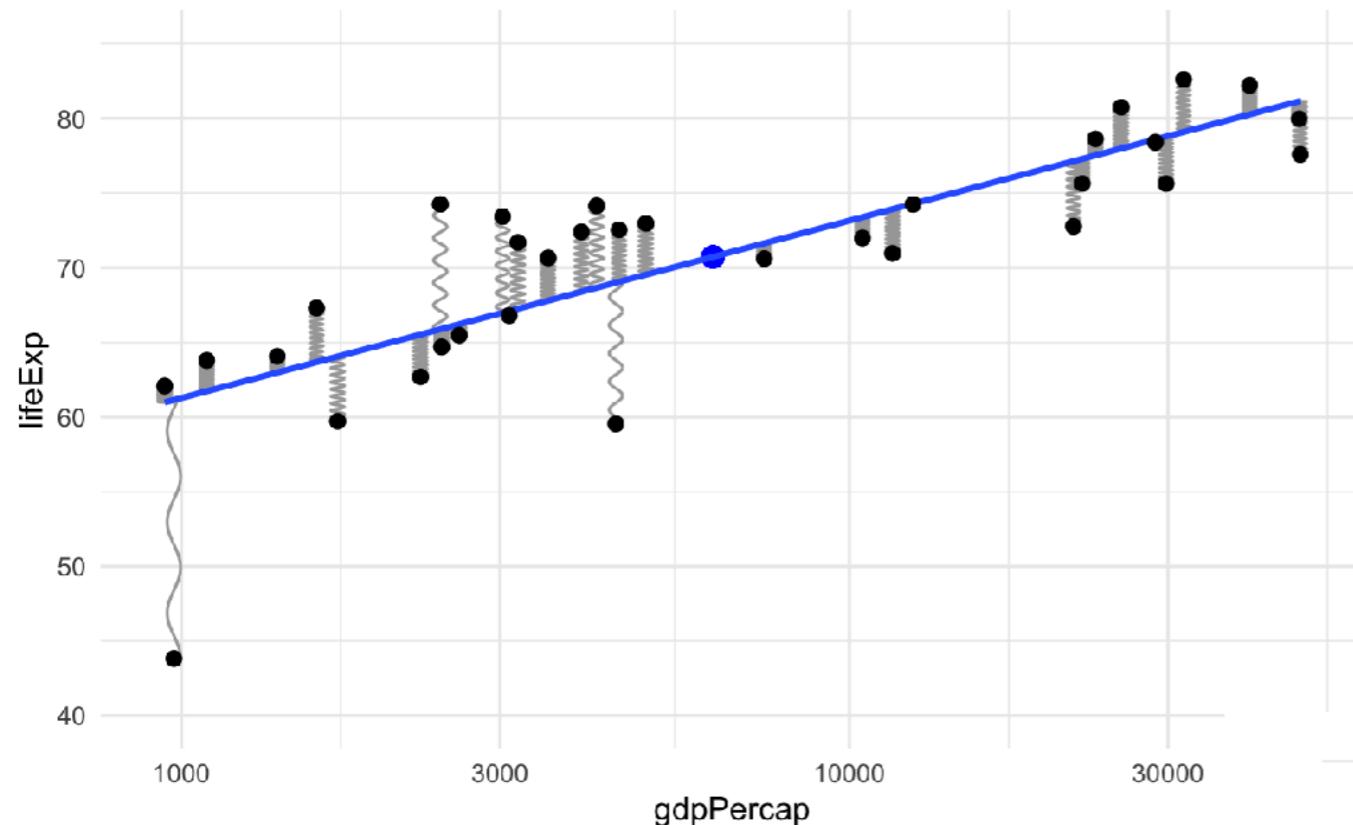
## Sum of squared errors



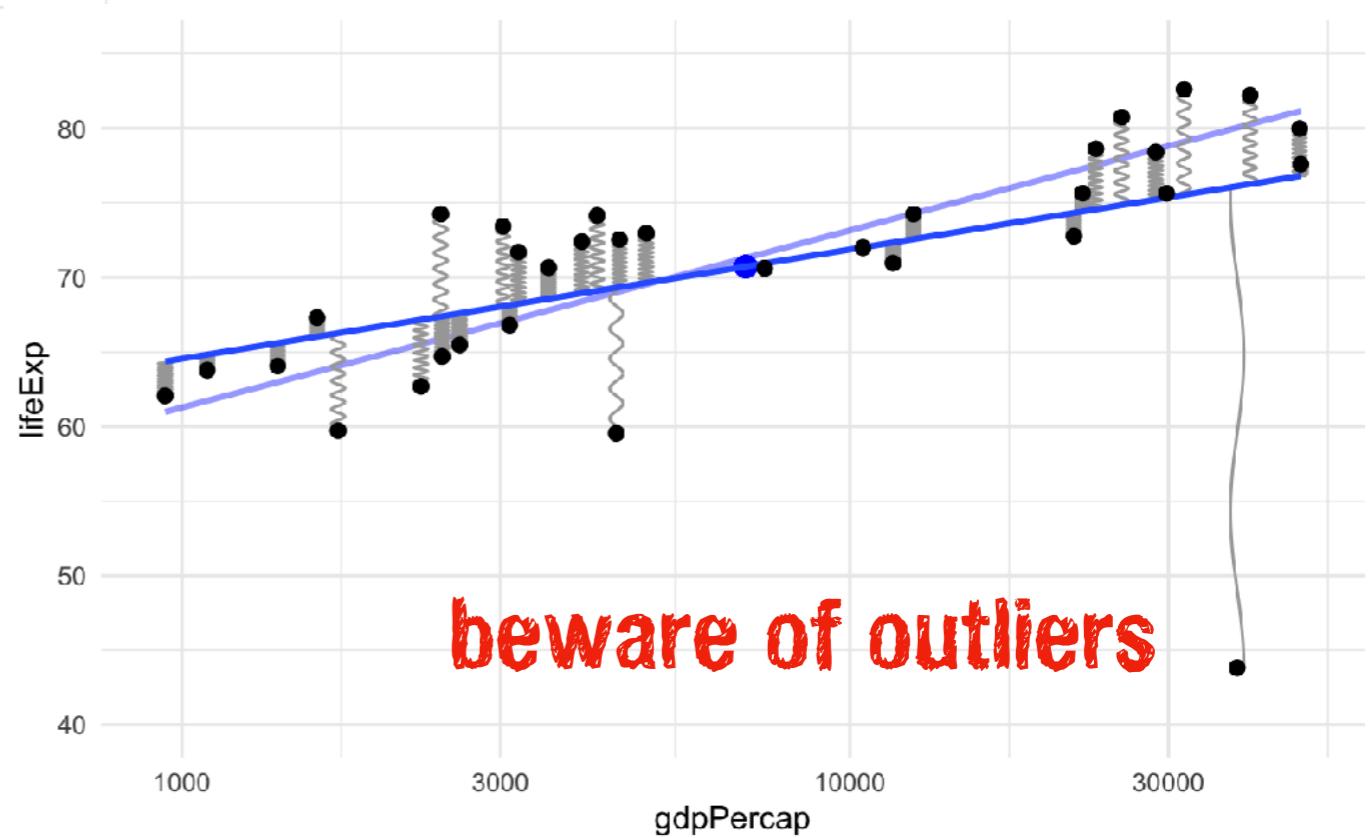
$$\text{SSE}(A) = 2396.946$$

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)

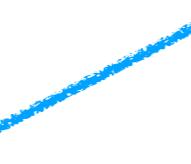
# Decide whether it's **worth it**

- 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  $\text{model}_C$  to create  $\text{model}_A$  but were not

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

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



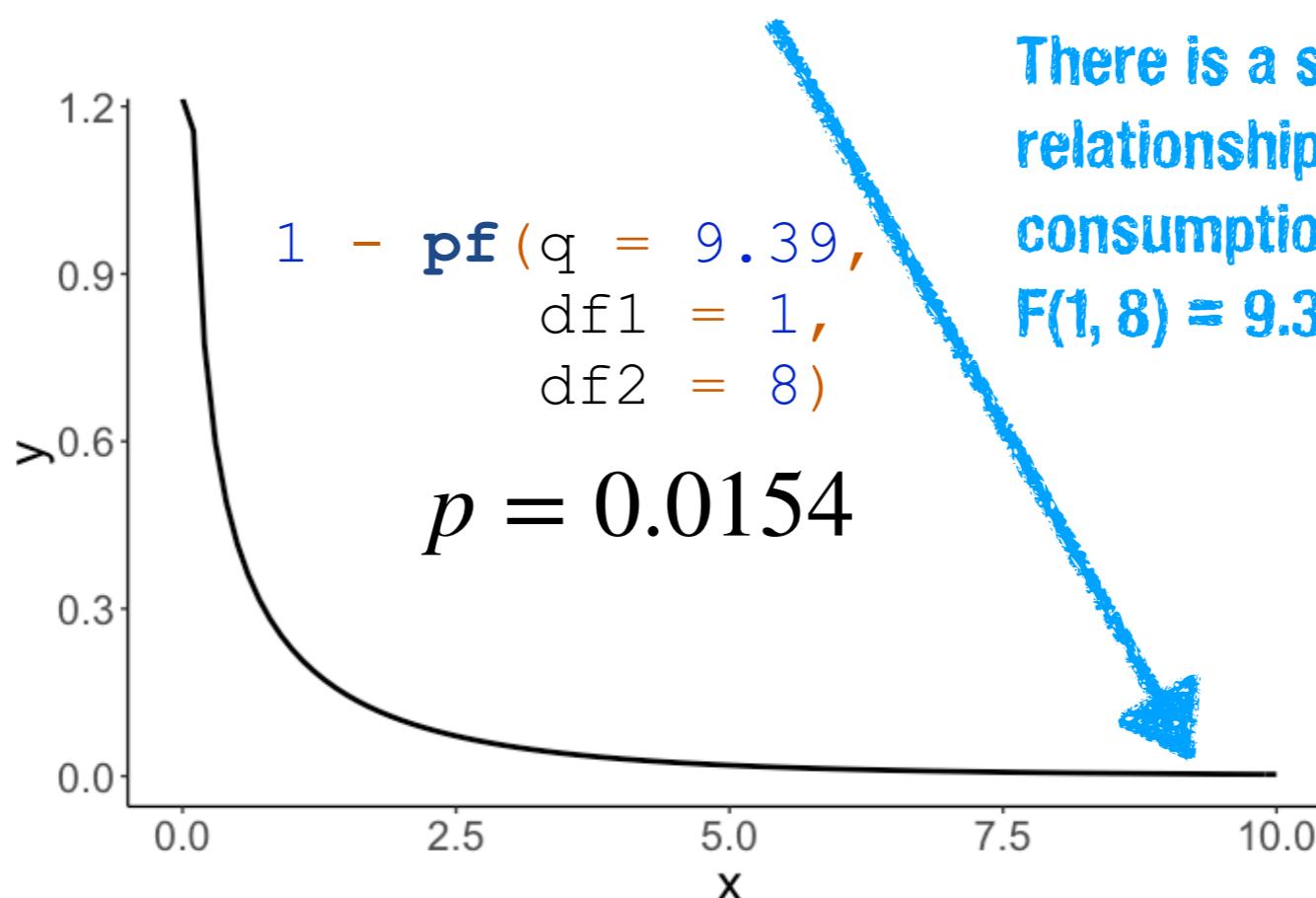
**number of observations  
vs. parameters in Model**

# Decide whether it's **worth it**

- To compute the  $F$  statistic, we need:

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

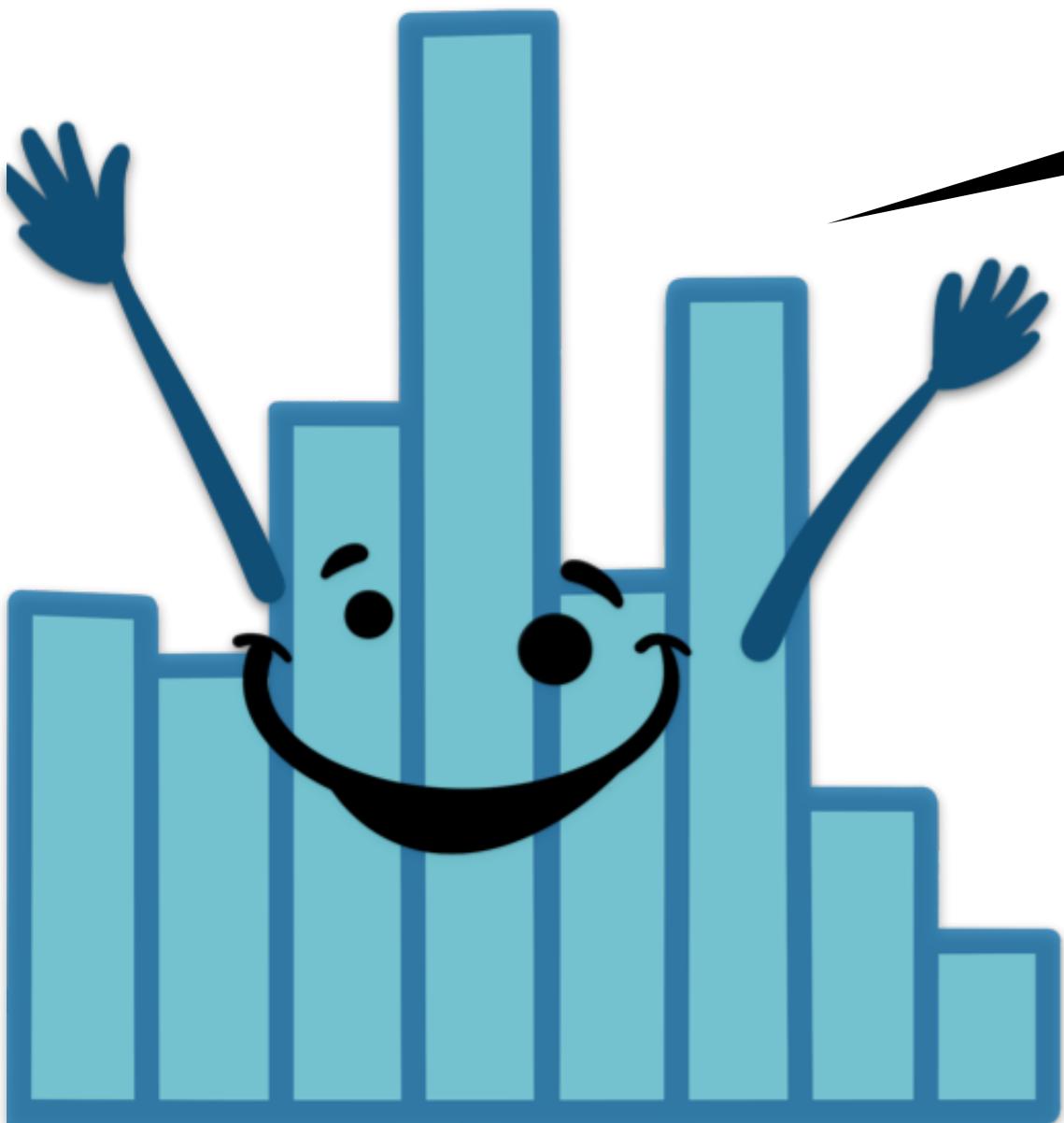
$$\begin{aligned} F &= \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})} \\ &= \frac{0.54/(2 - 1)}{(1 - 0.54)/(10 - 2)} \\ &= 9.39 \end{aligned}$$



We're listening to "Pata Pata" by "Miriam Makeba" submitted by Tobi

02:00

stretch break!



# 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

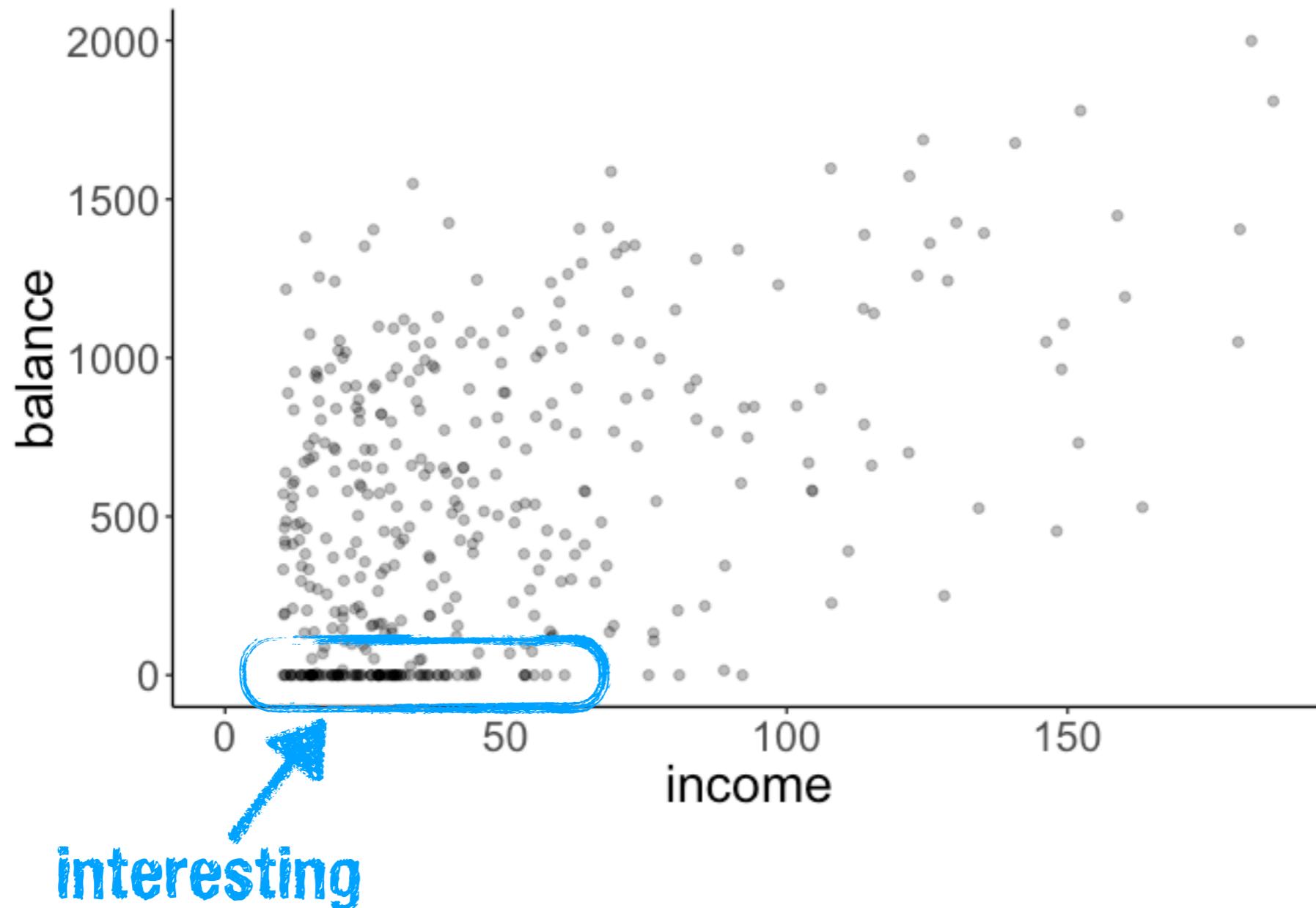
**nrow(df.credit) = 400**

**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

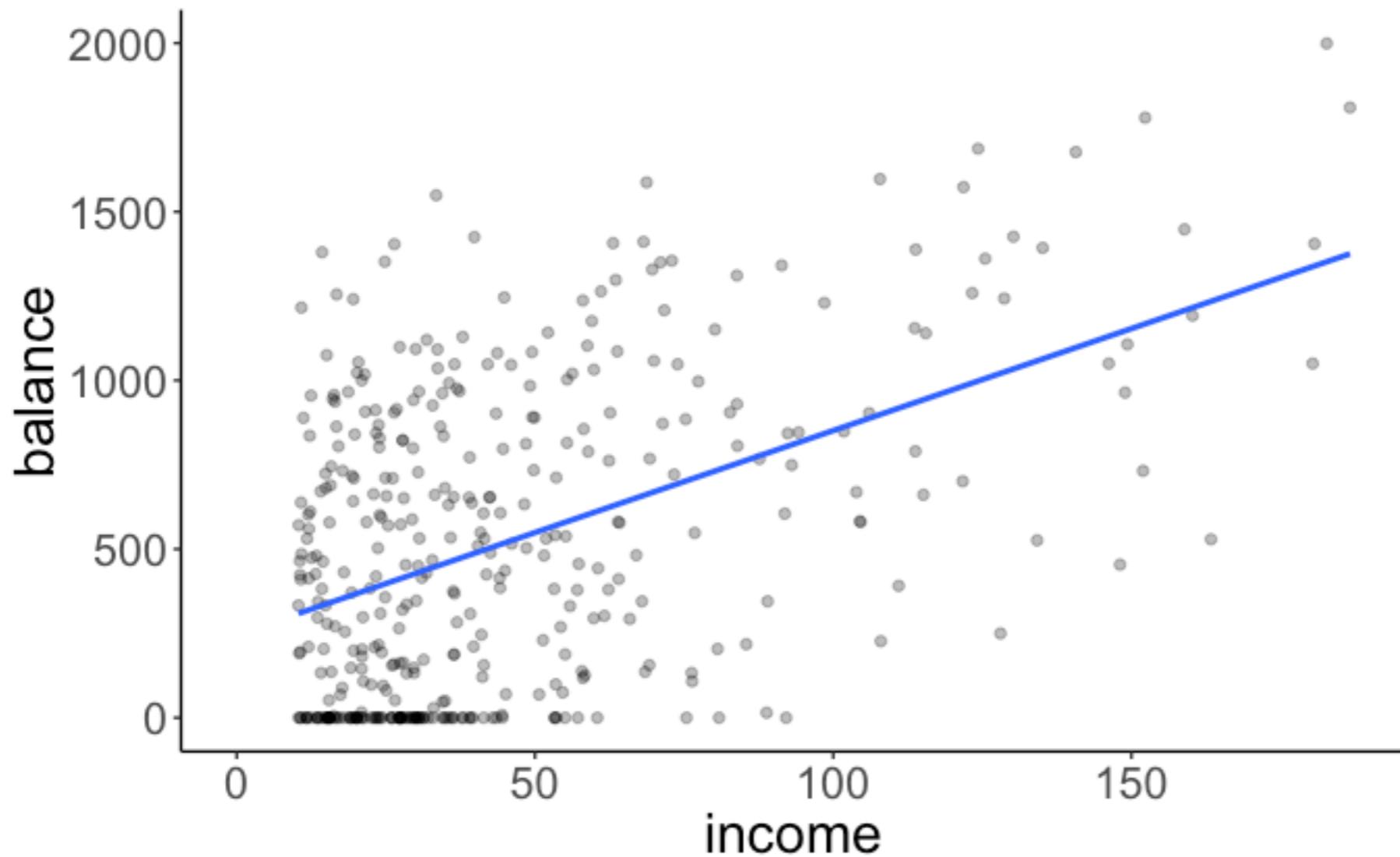
# Always plot the data first ...

```
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)
```

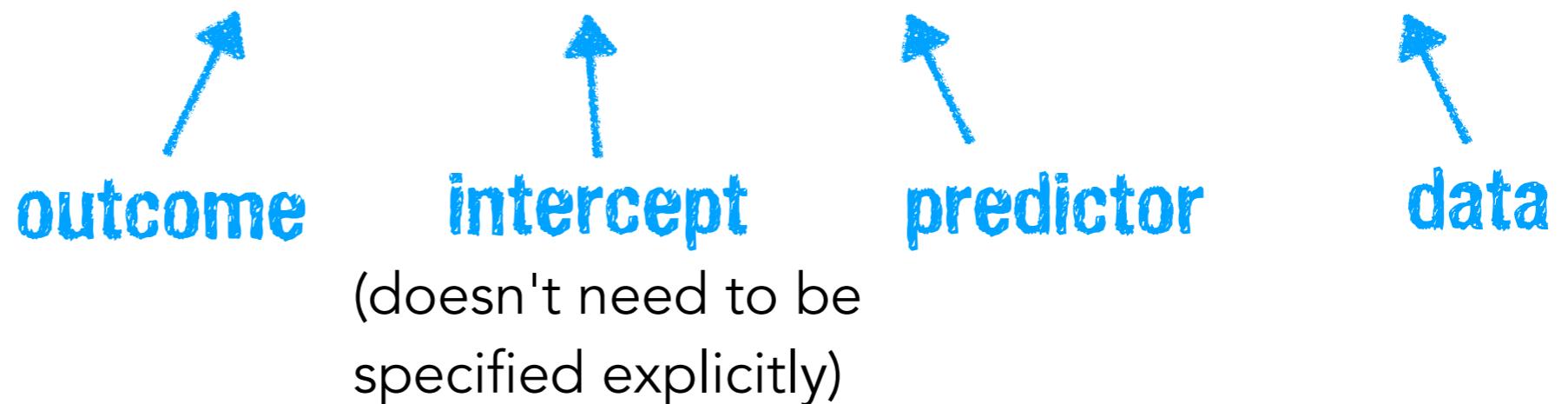


# Linear model

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

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

```
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:

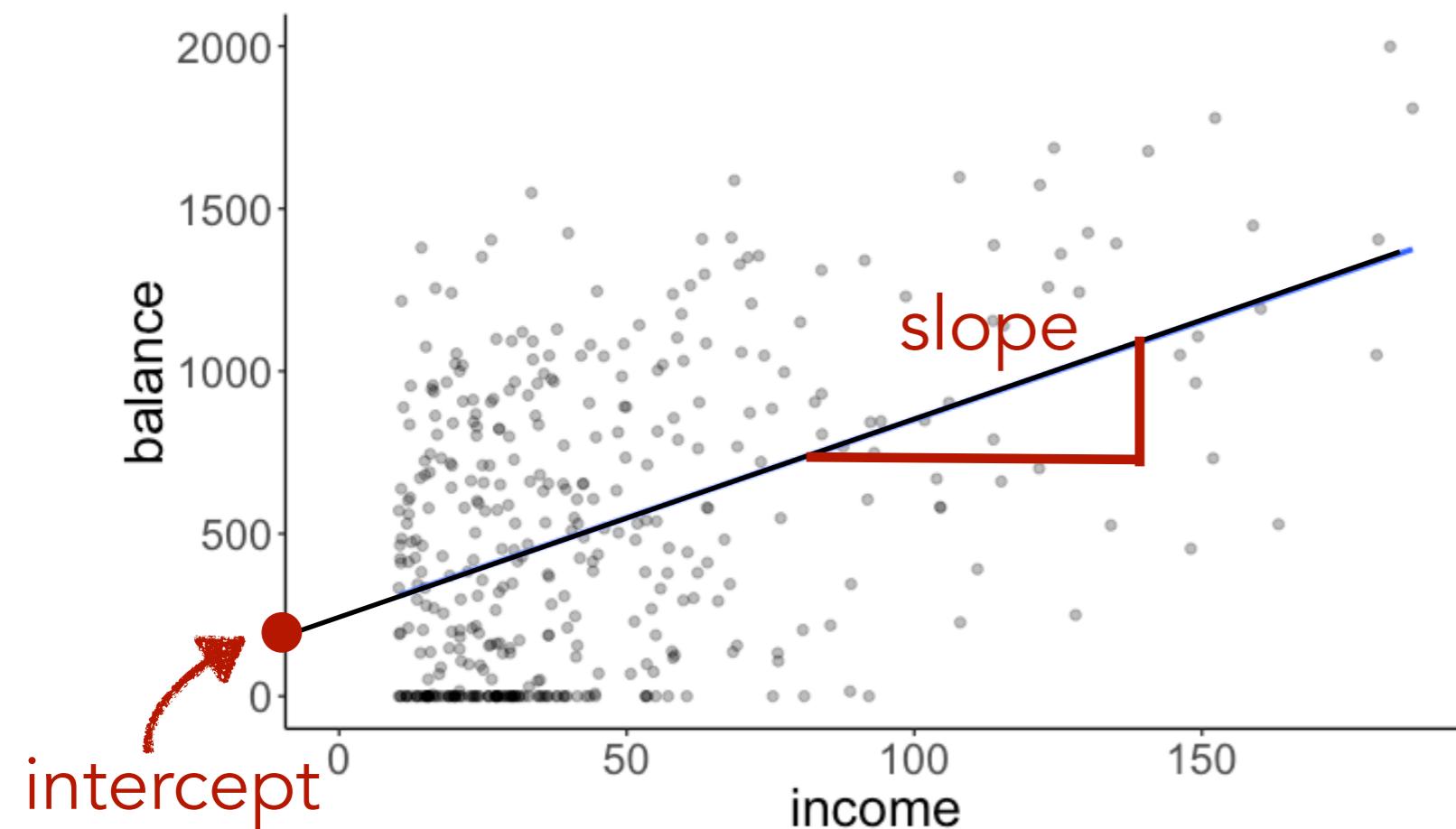
(Intercept) 246.515

income 6.048

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

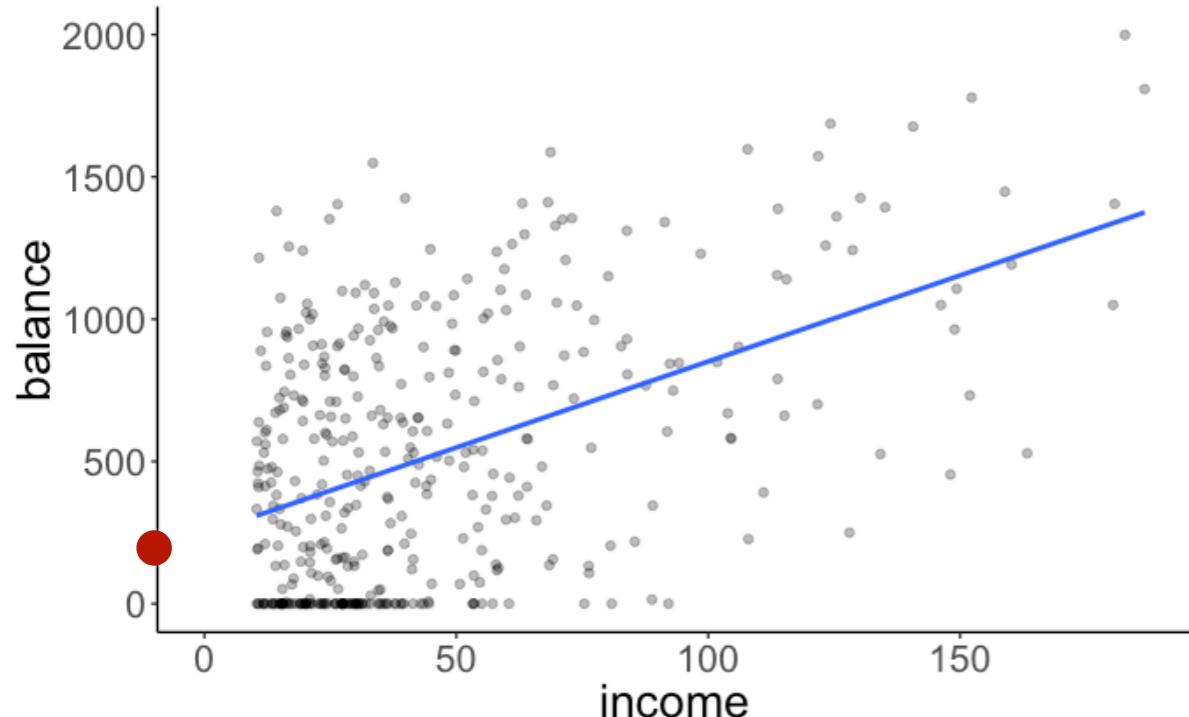
$$\text{balance}_i = b_0 + b_1 \cdot \text{income}_i + e_i$$

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

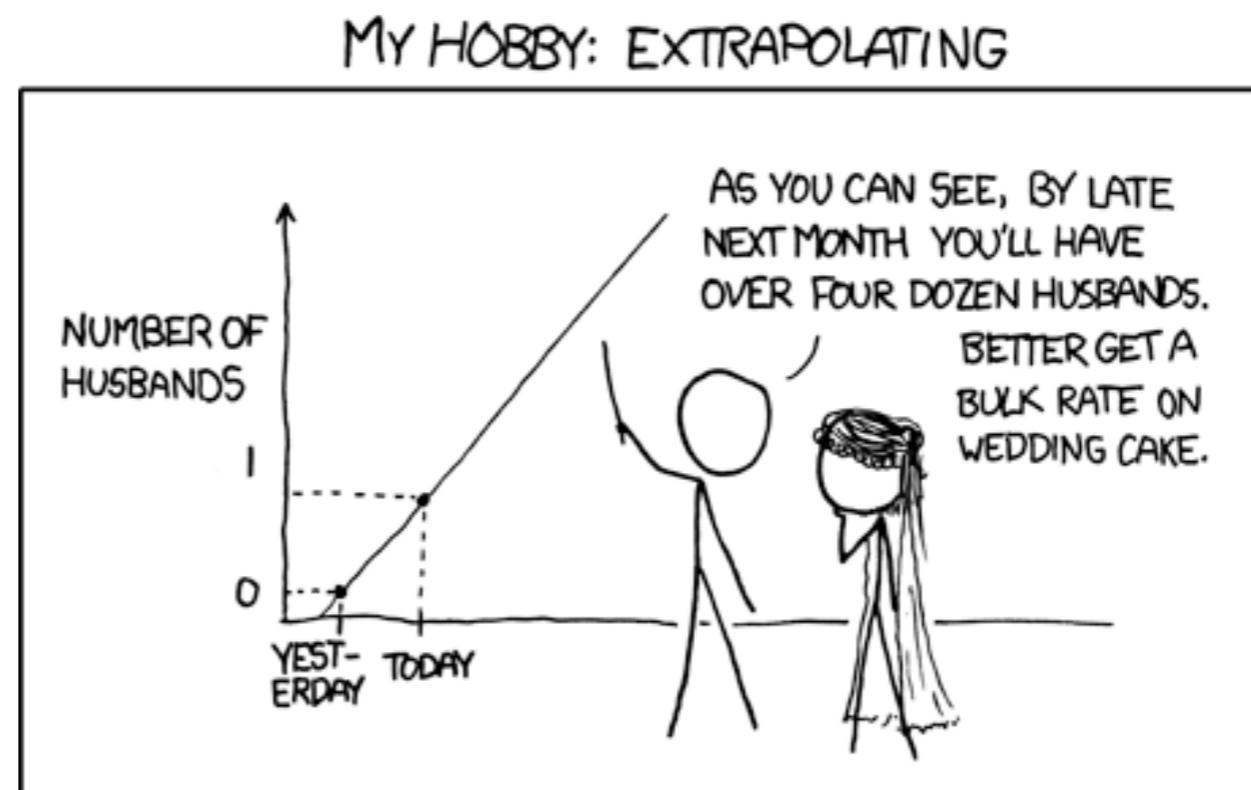


For each additional thousand dollars income, a person's average credit card is predicted to increase by \$6.05.

# 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, ...)



```
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

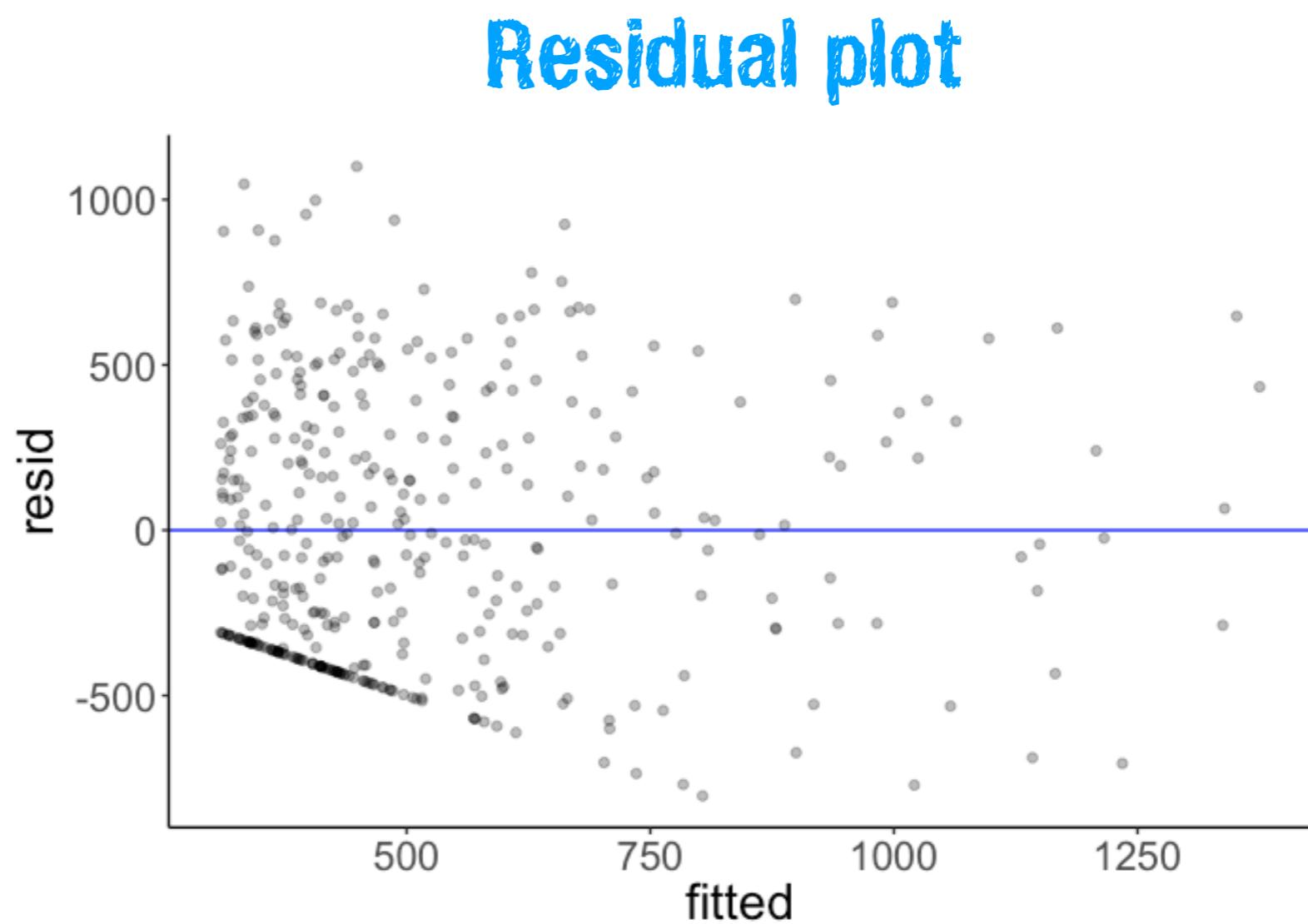
# summary()

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

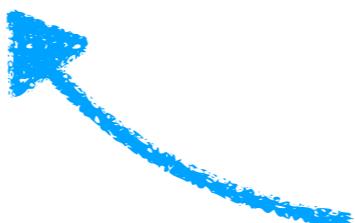
# summary()

```
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)

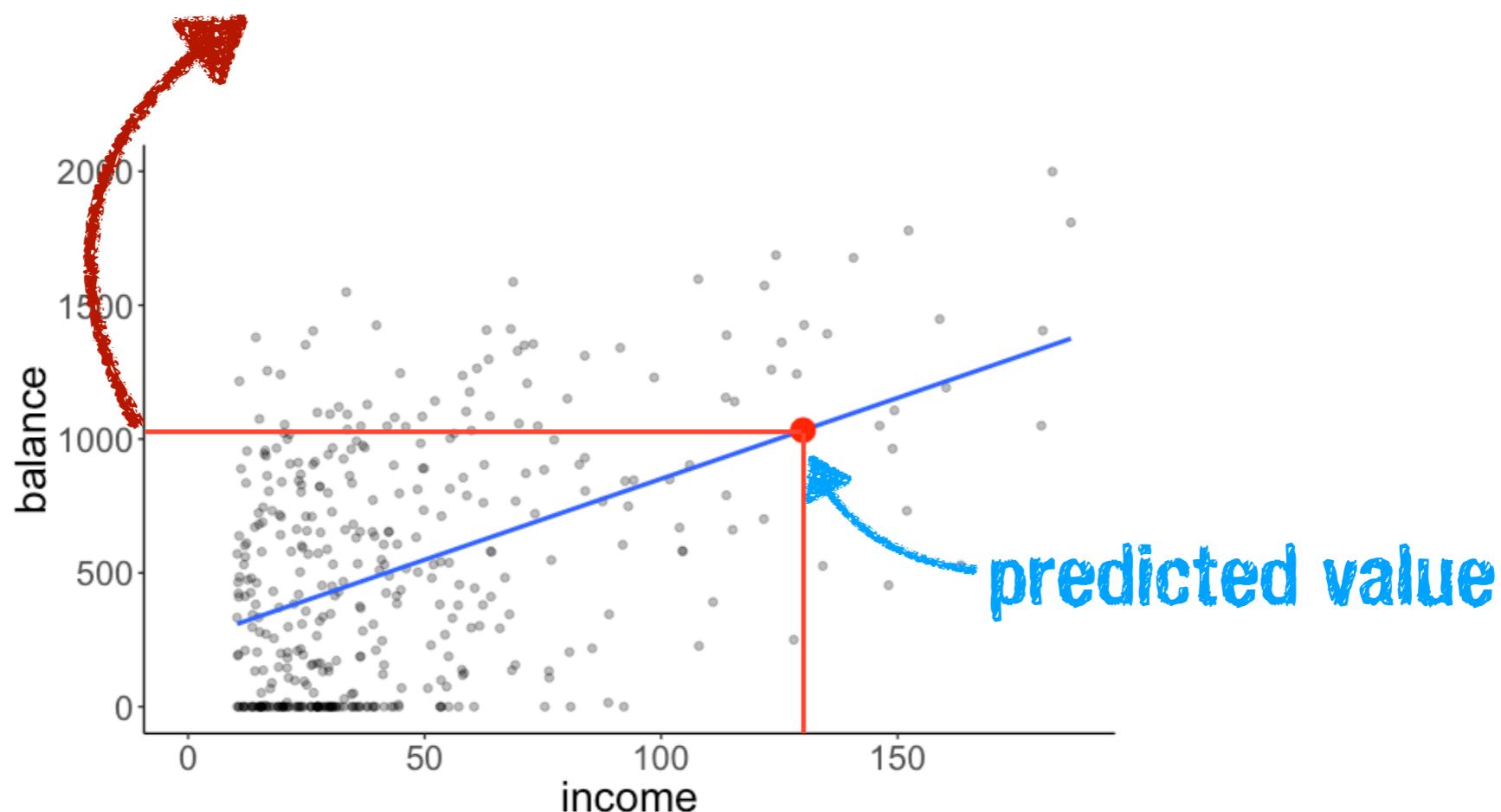
# Making predictions

```
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()
```

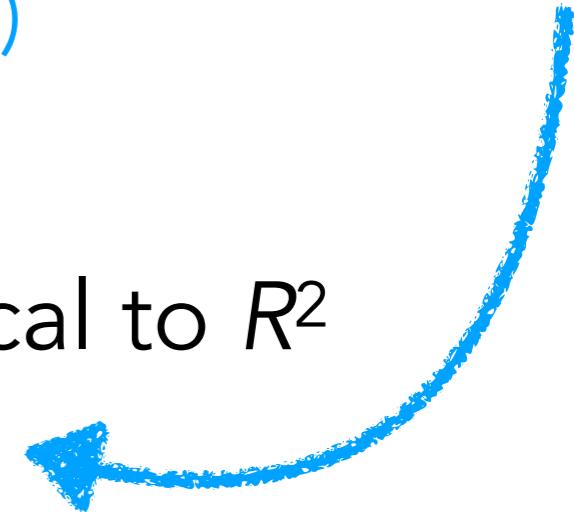
$R^2$

```
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  $r$



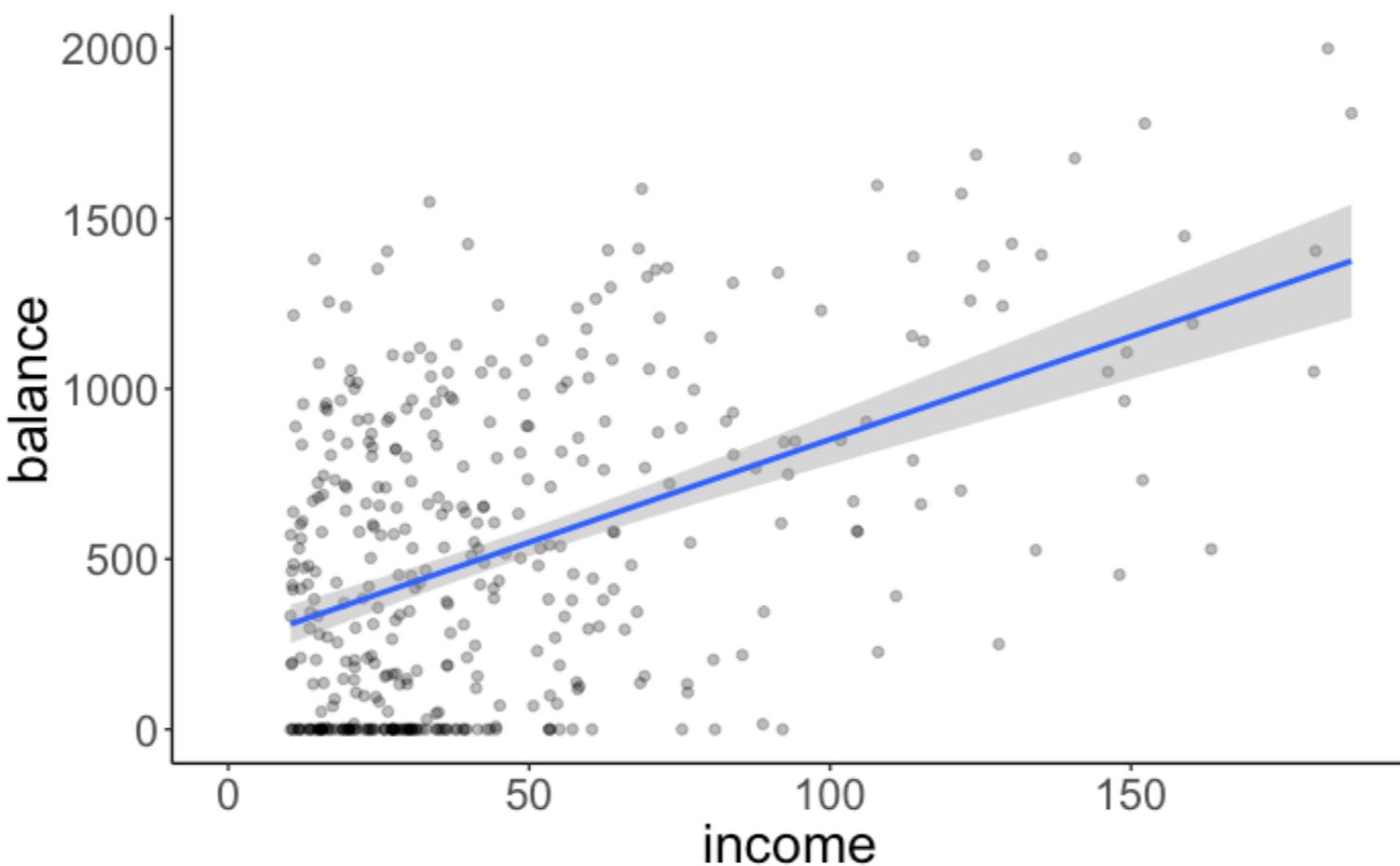
```
cor(df.credit$balance,  
df.credit$income)
```

$$R^2 = 0.215$$

$$r = .463$$

effect size measure

# Reporting the results

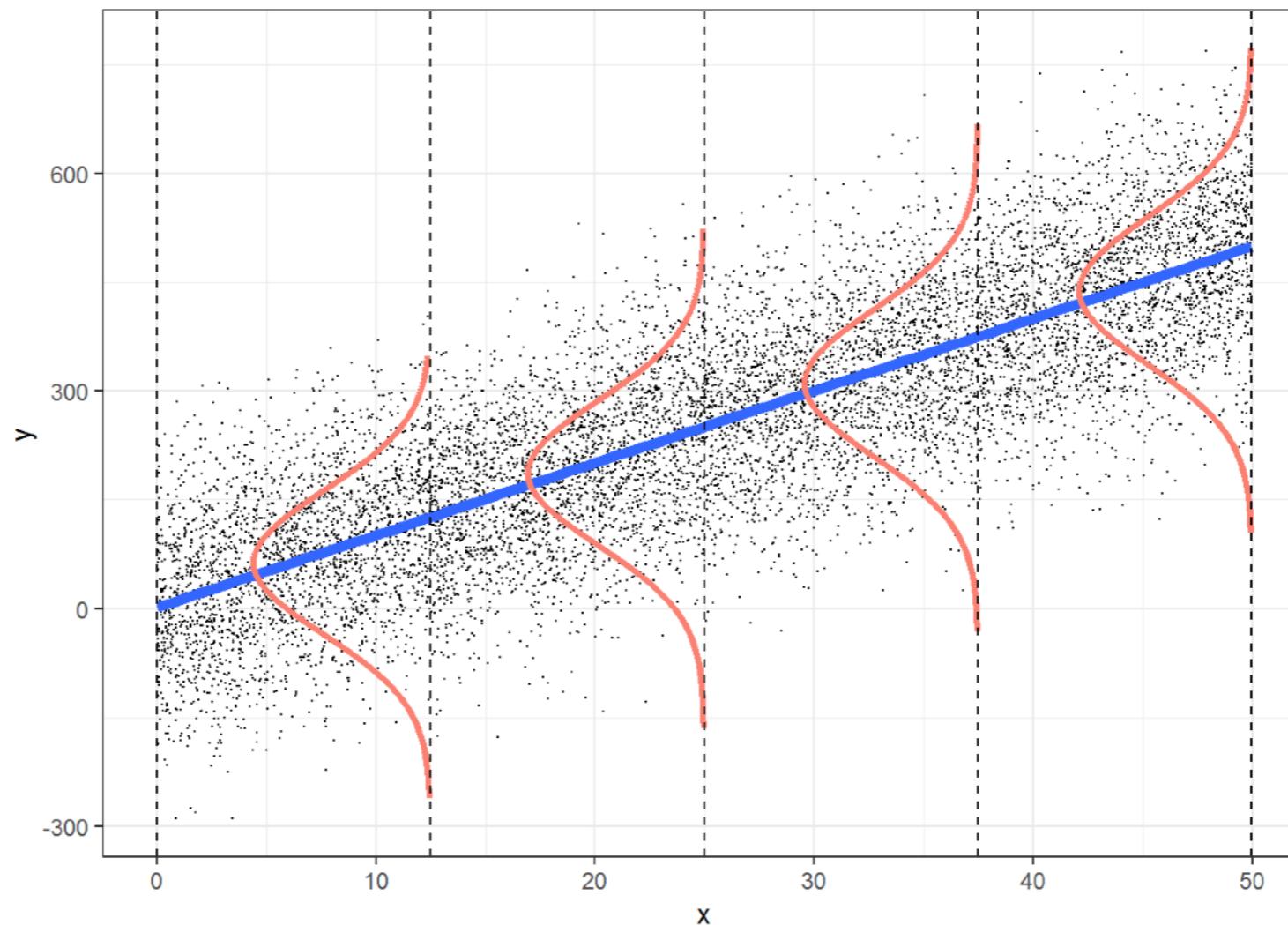


There is a significant relationship between a person's income and the average balance on their credit cards  
 $F(1, 389) = 108.99, p < .001, r = .463$ .

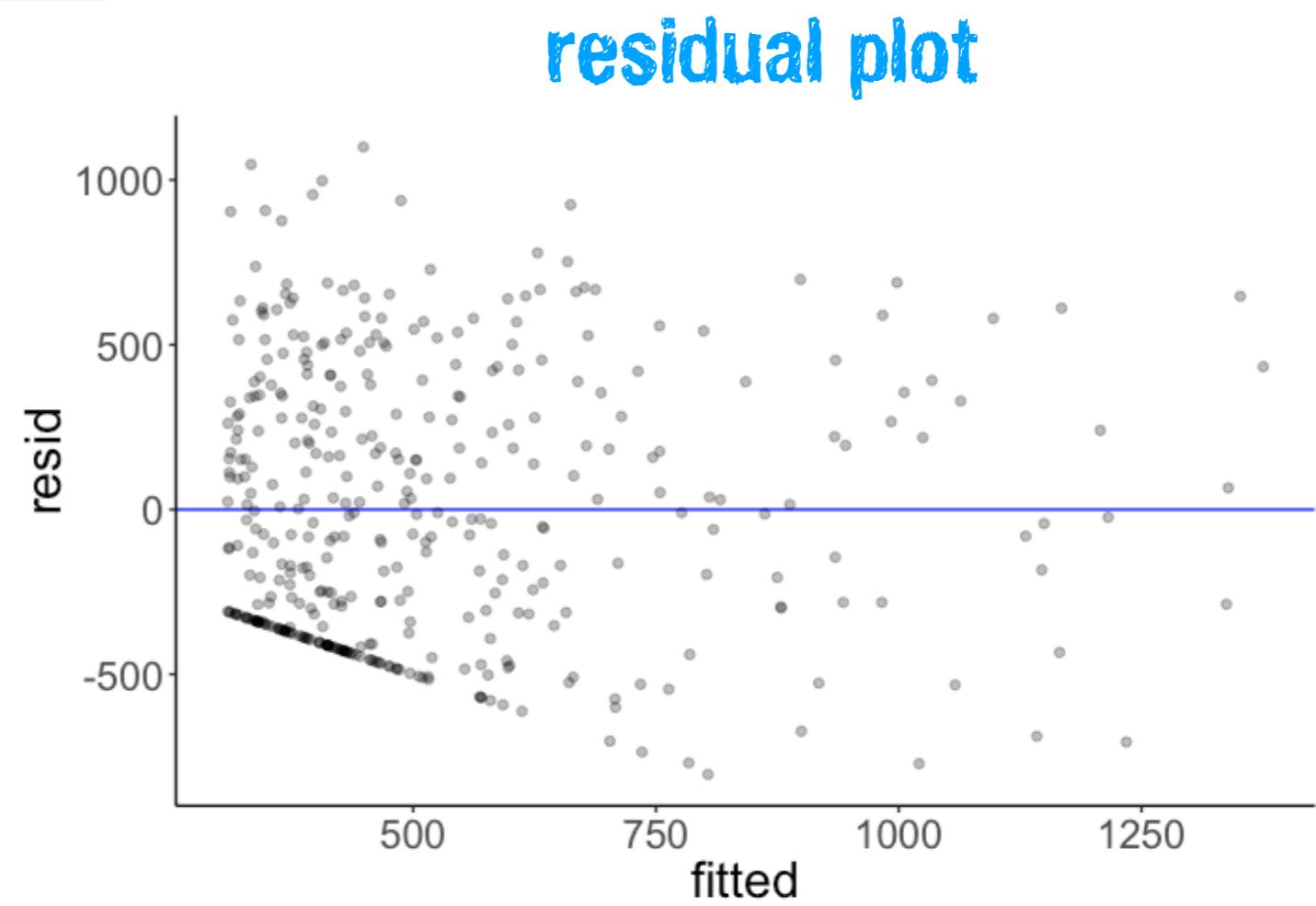
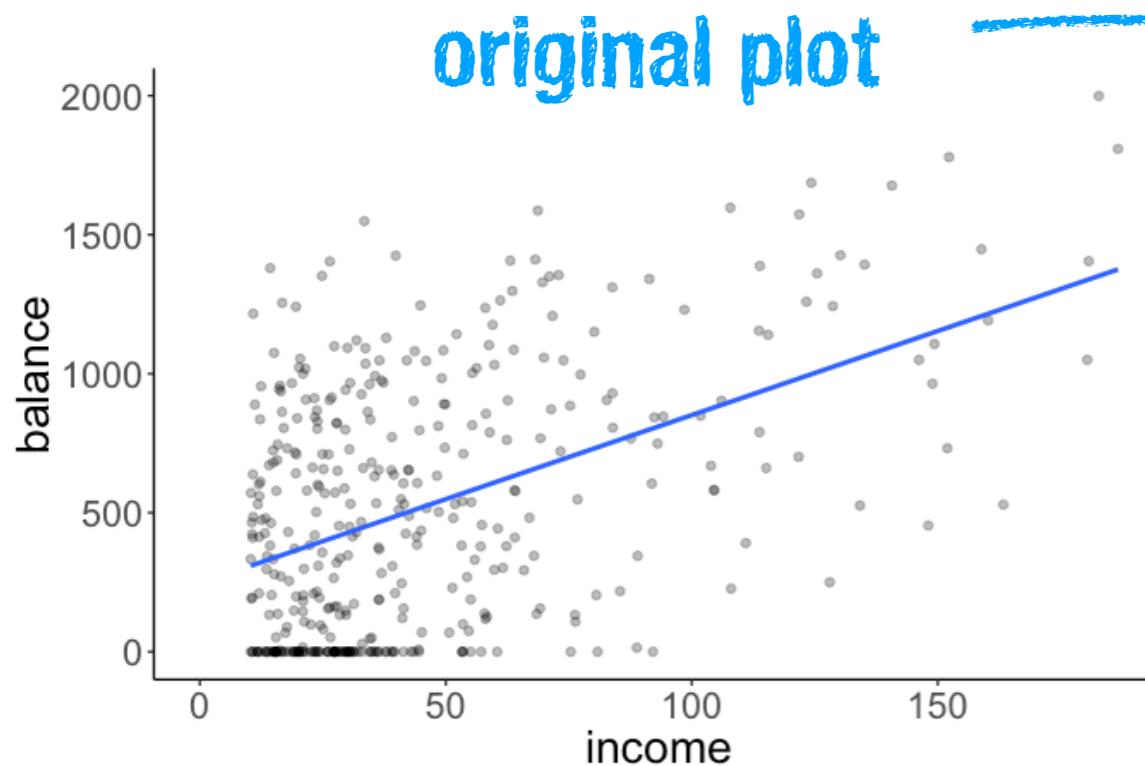
With each additional \$1000 of income, the average balance is predicted to increase by \$6.05 [4.91, 7.19] (95% CI).

# Model assumptions

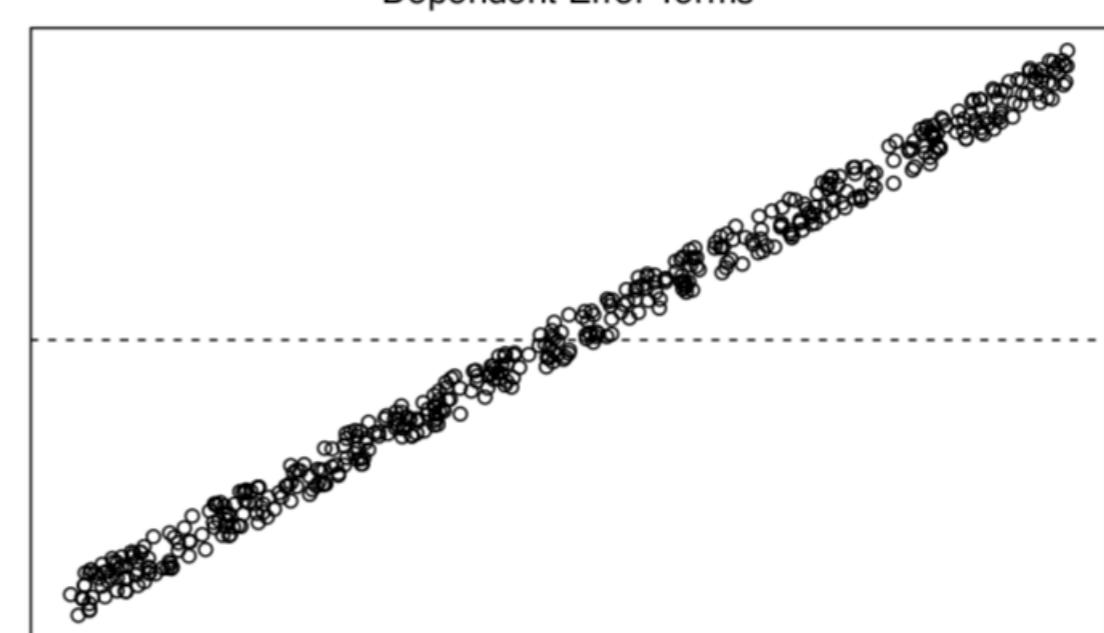
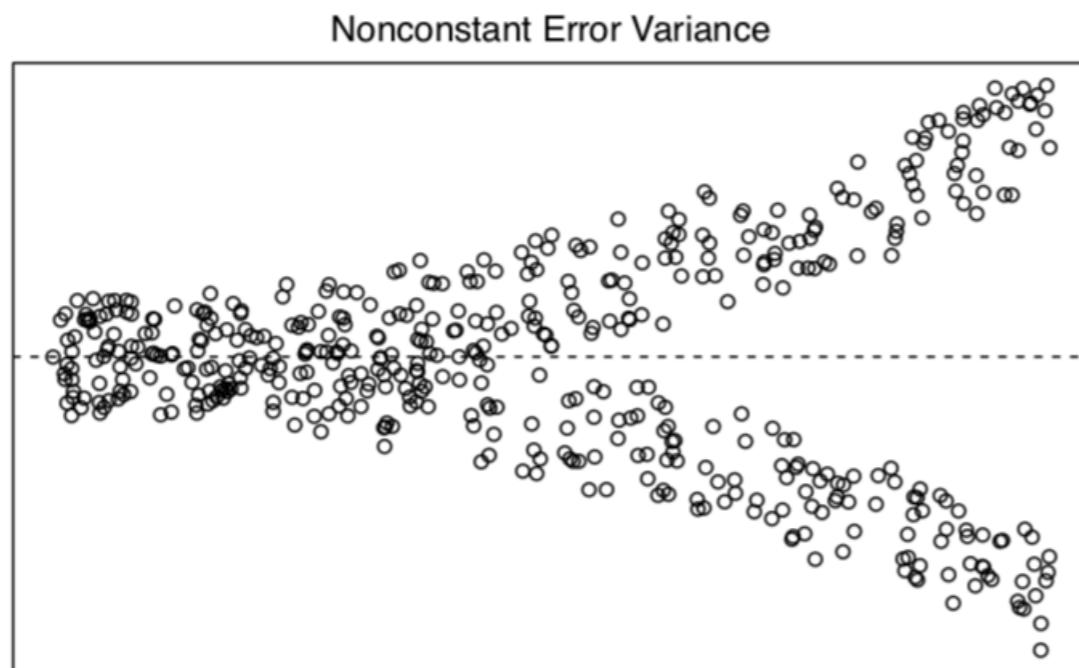
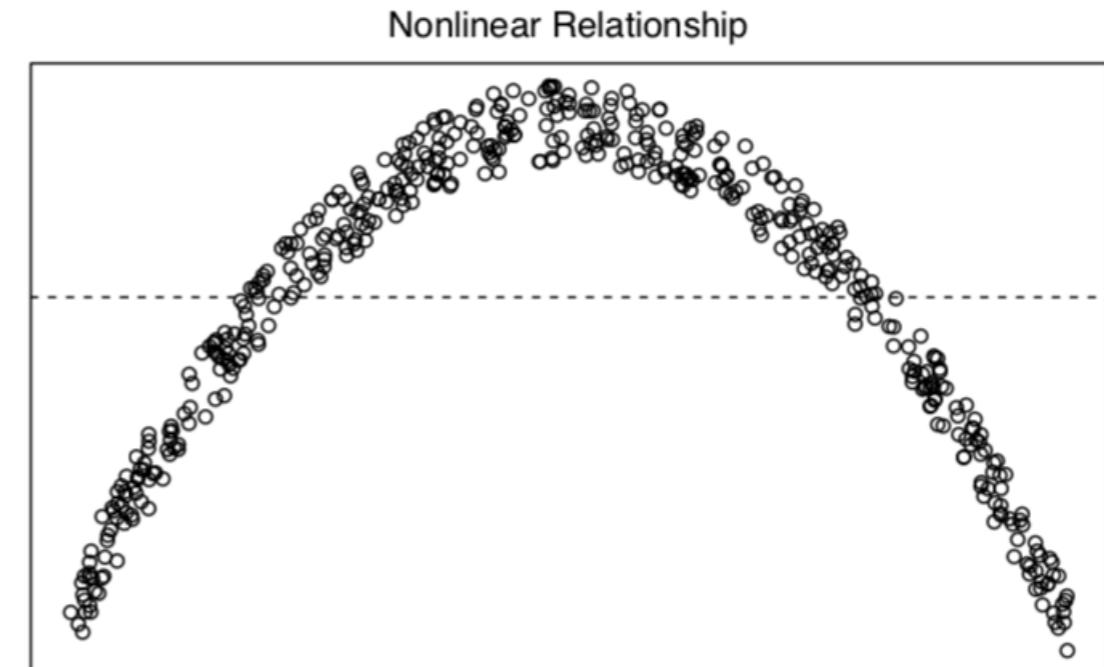
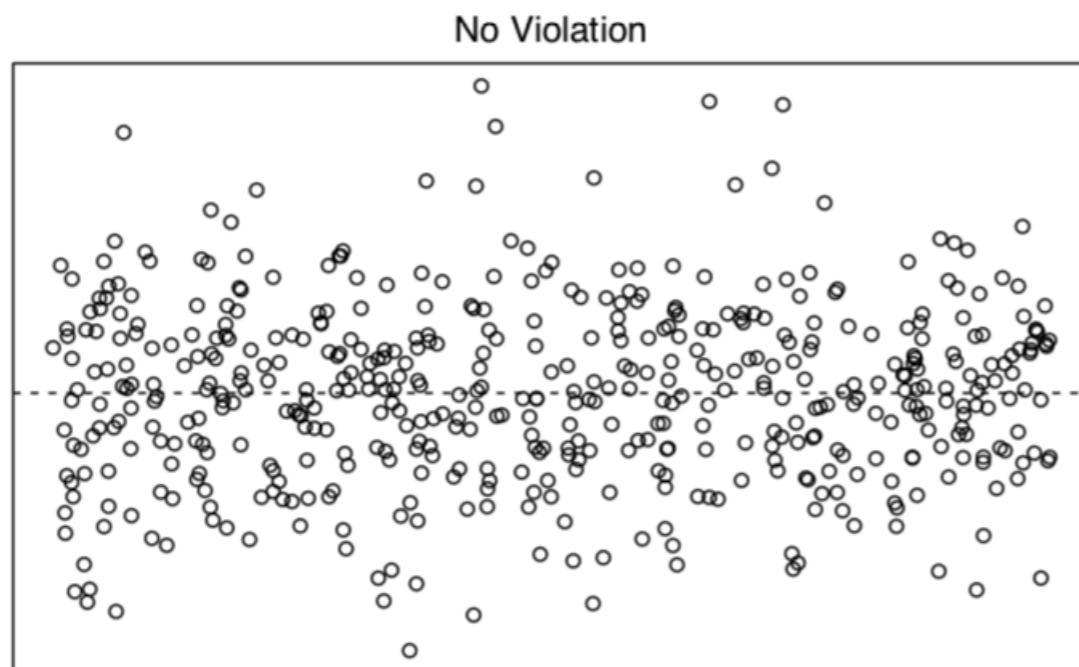
- independent observations
- $Y$  is continuous
- errors are normally distributed
- errors have constant variance
- error terms are uncorrelated



# Model assumptions



# Model assumptions



# Plan for today

- Quick recap
- Regression
  - The conceptual tour
  - The R route

# **Feedback**

# How was the pace of today's class?

much      a little      just      a little      much  
too      too      right      too      too  
slow      slow

# How happy were you with today's class overall?



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

**Thank you!**