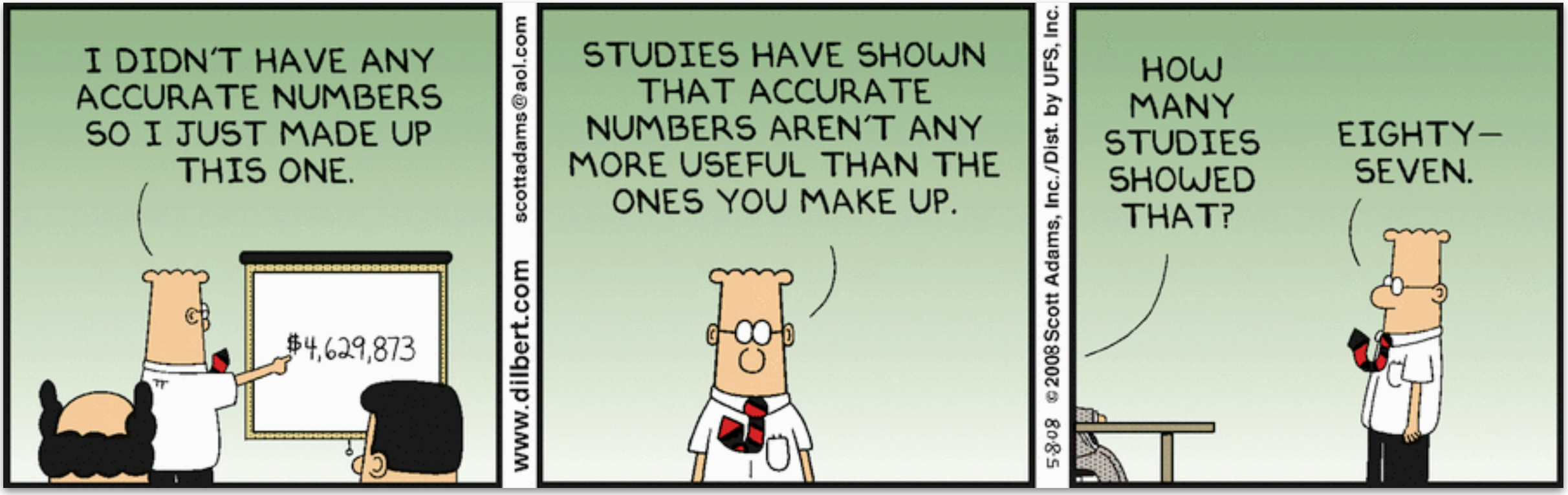
PSYCH 252: Statistical Methods for Behavioral and Social Sciences

Linear model 2



02/02/2024

Logistics

Class recordings

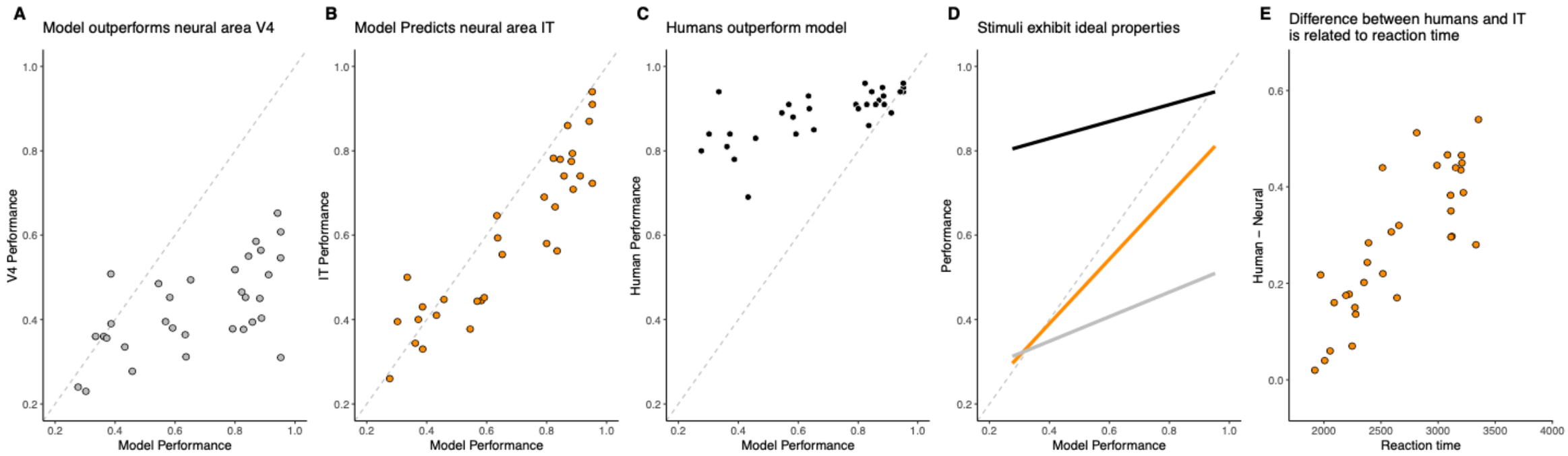
Lecture recordings 2022

- Introduction
- Visualization I
- Visualization II
- Data wrangling I
- Data wrangling II
- Probability
- Simulation I

Homework 2

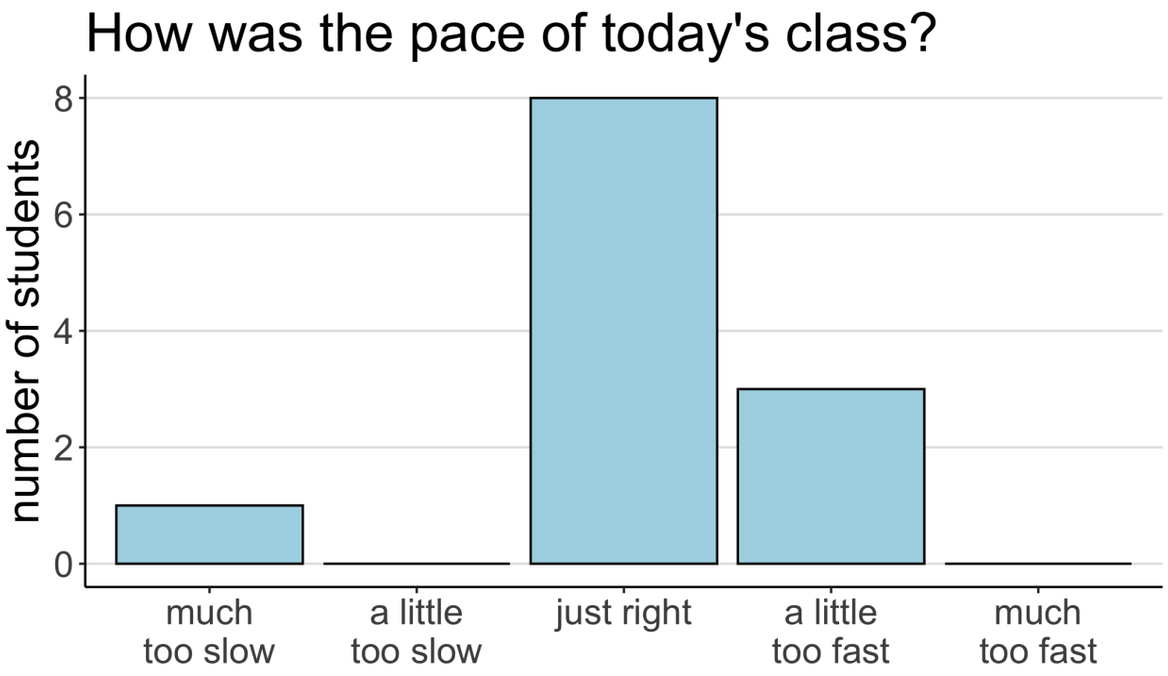
Homework 2

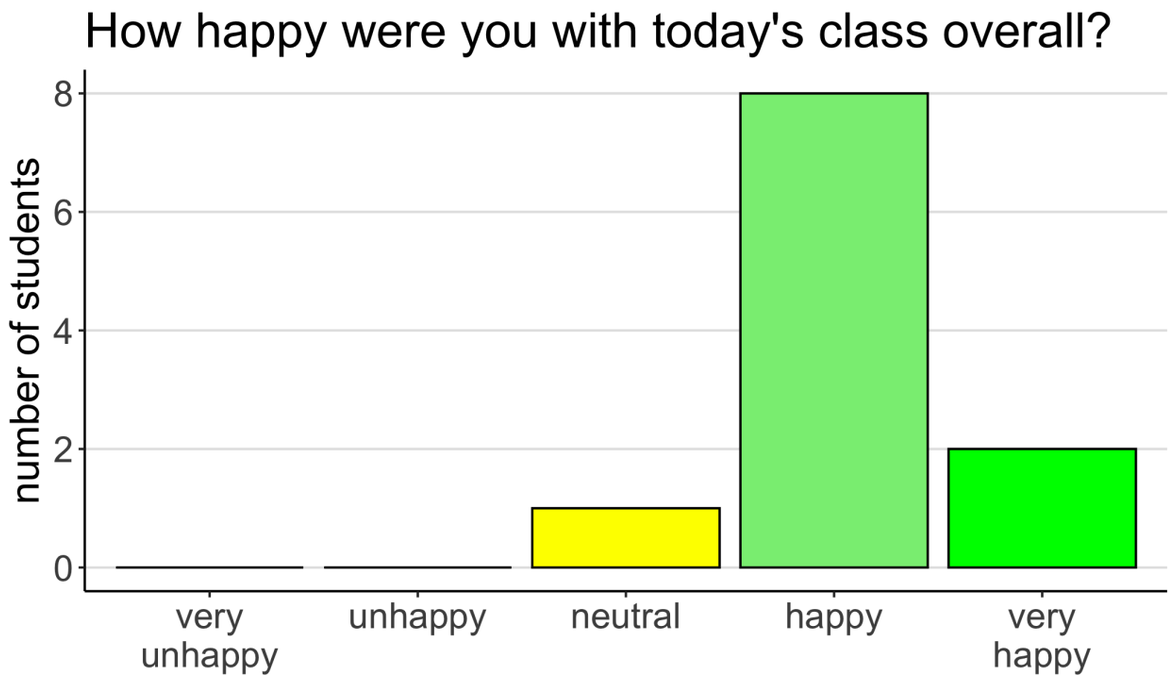
Grades are posted, and solutions are on Canvas.



Feedback

Your feedback





Just want to voice support for the recap at the beginning of class — please don't shorten it!

These concepts are just hard I'll have to self study

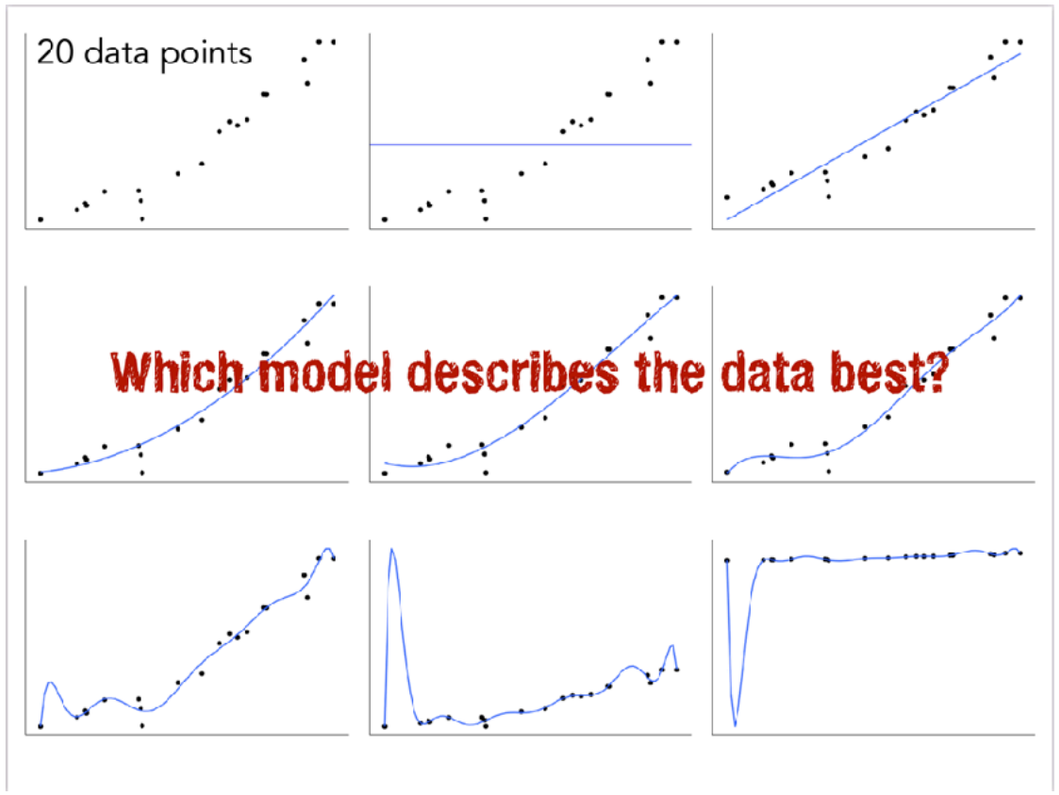
I would like to defend long recaps! Esp as we move into regressions!!!

Plan for today

- Quick recap
- Who is the correlation champ?
- Regression
  - The conceptual tour
  - The R route

Quick recap

Quick recap: Modeling data



- 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

Compact model | **Augmented model**

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

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

Proportional reduction in error (PRE)

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

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**

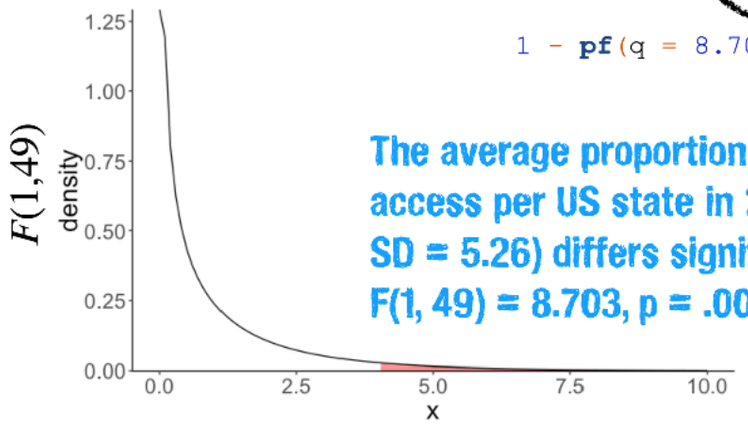
Decide whether it's **worth it**

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

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

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

```
1 - pf(q = 8.703, df1 = 1, df2 = 49)
```



The average proportion of internet access per US state in 2003 (M = 72.8, SD = 5.26) differs significantly from 75%, F(1, 49) = 8.703, p = .005.

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

$$Y_i = b_0 + e_i$$

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

```
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

C: $Y_i = 75 + e_i$ **A**: $Y_i = \overline{Y} + e_i$

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

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

$$= 1 - \frac{1355}{1595} \approx .15$$

Model A has 15% less error than Model C.

SSE(C) = 1595 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

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

Model C

$$Y_i = 75 + e_i$$

0 parameters

Model A

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

1 parameter

Sampling distribution of PRE

```r
# simulation parameters
n_samples = 1000
sample_size = 50
mu = 75 # true mean of the distribution
sigma = 5 # true standard deviation of the errors

# function to draw samples from the population distribution
fun.draw_sample = function(sample_size, mu, sigma){
  sample = mu + rnorm(sample_size, mean = 0, sd = sigma)
}

# draw samples
samples = n_samples %>%
  replicate(fun.draw_sample(sample_size, mu, sigma)) %>%
  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

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

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

Sampling distribution of PRE

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

calculate SSE for each model

calculate PRE

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

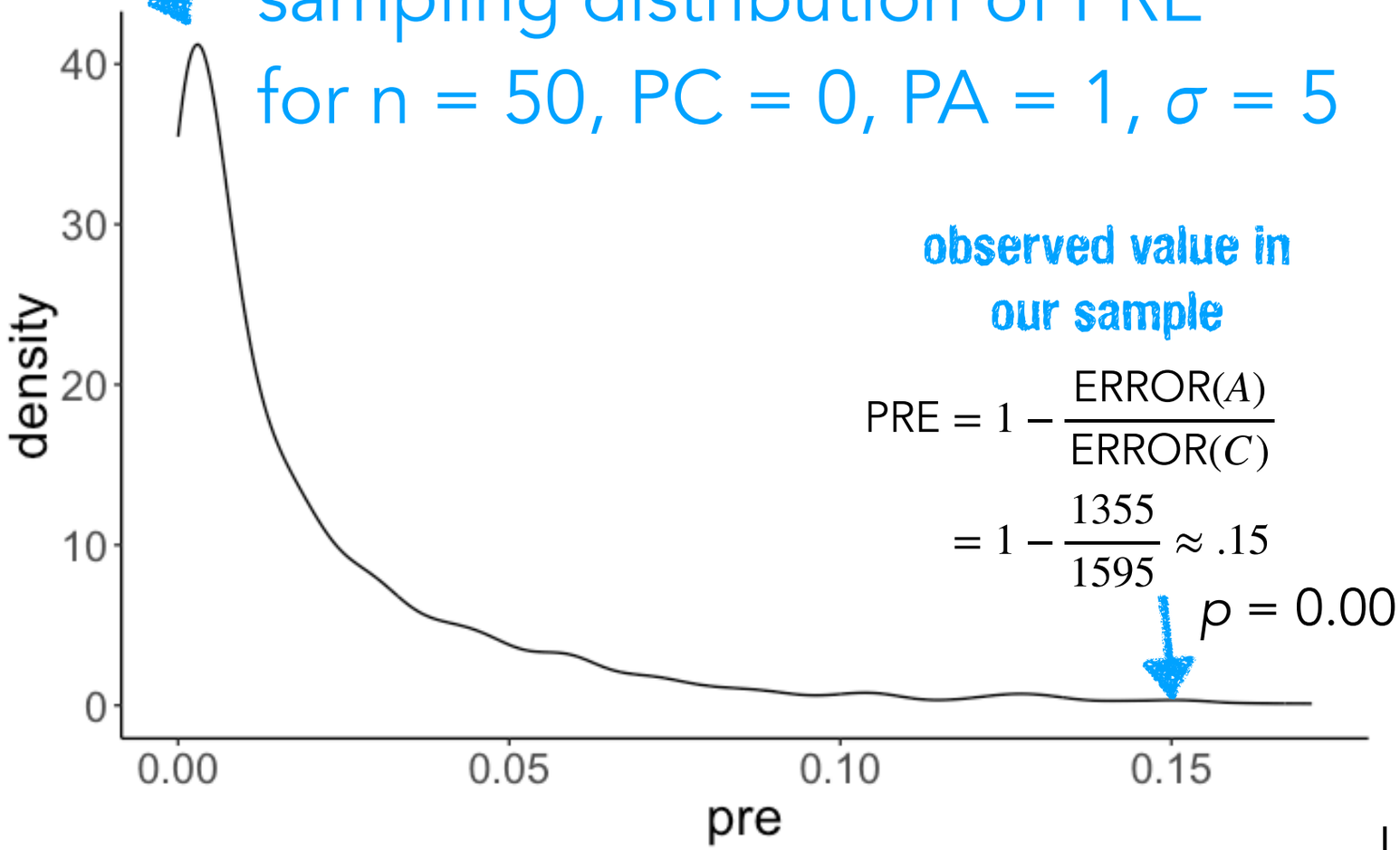
sample	sse_compact	sse_augmented	pre
1	1244.66	1241.52	0.00
2	953.25	894.95	0.06
3	1083.60	1083.32	0.00
4	888.06	798.19	0.10
5	1246.57	1233.25	0.01

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
for n = 50, PC = 0, PA = 1, $\sigma = 5$

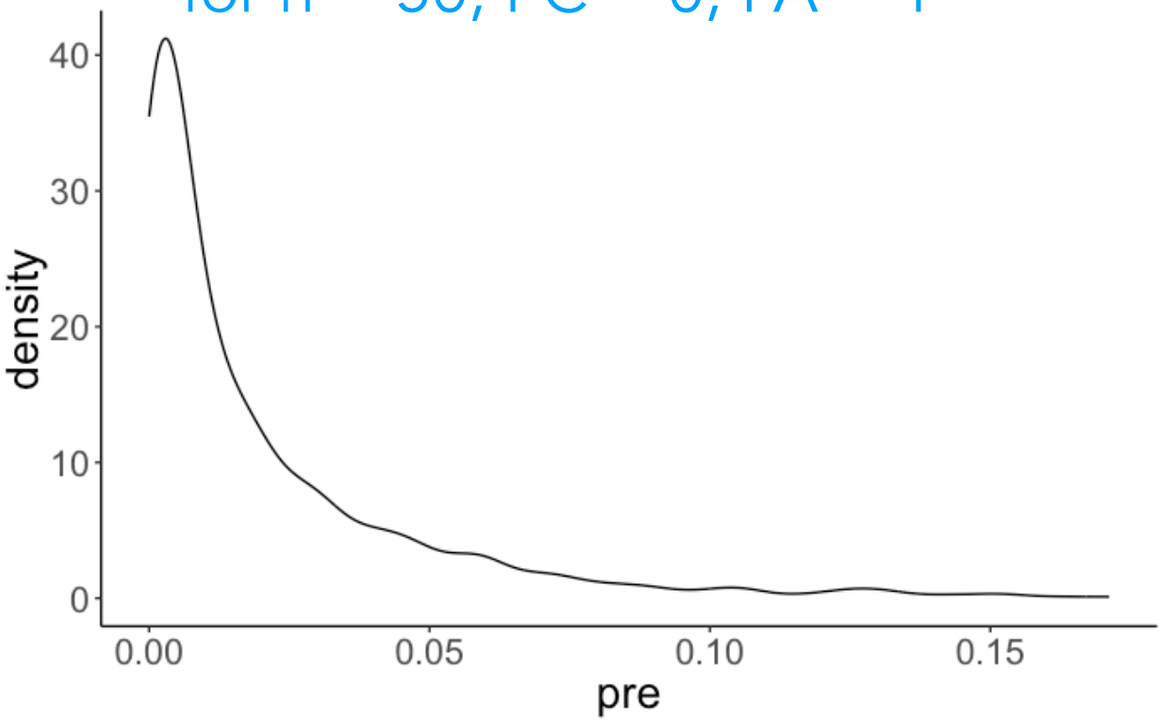


Sampling distribution of PRE

deterministic mapping

sampling distribution of PRE

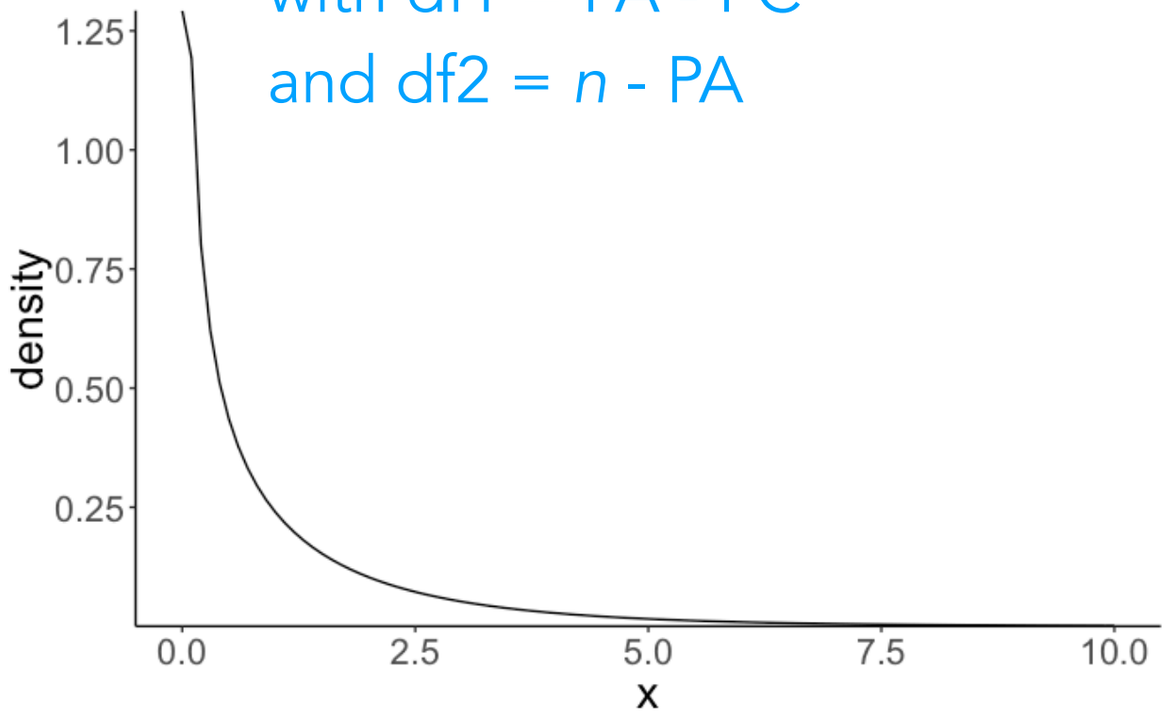
for n = 50, PC = 0, PA = 1



***F*(df1, df2) distribution**

with df1 = PA - PC

and df2 = *n* - PA

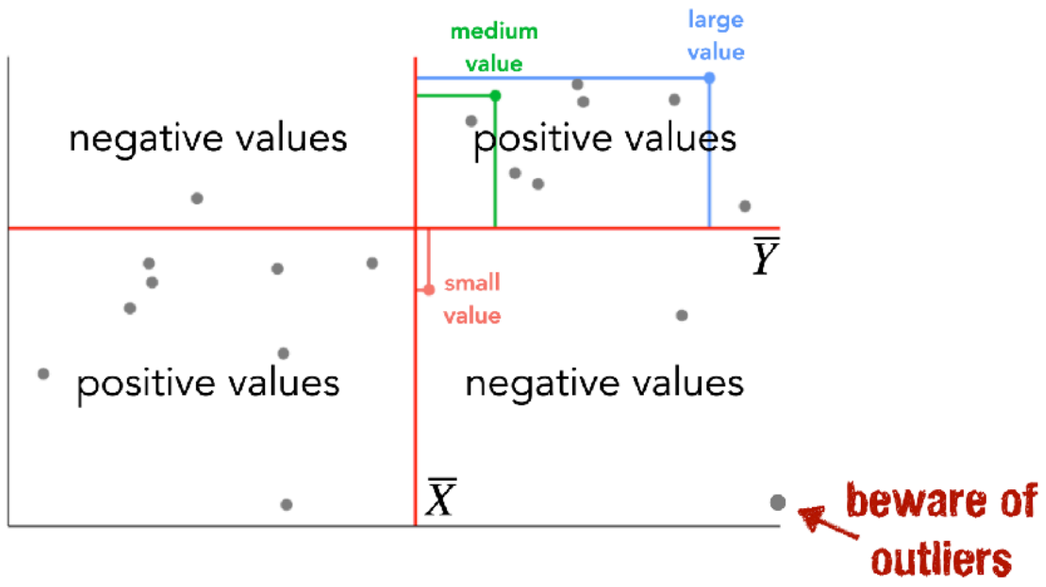


we use the F-distribution since it comes with R (and is the standard statistic to report)

Quick recap: Correlation

sample covariance

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



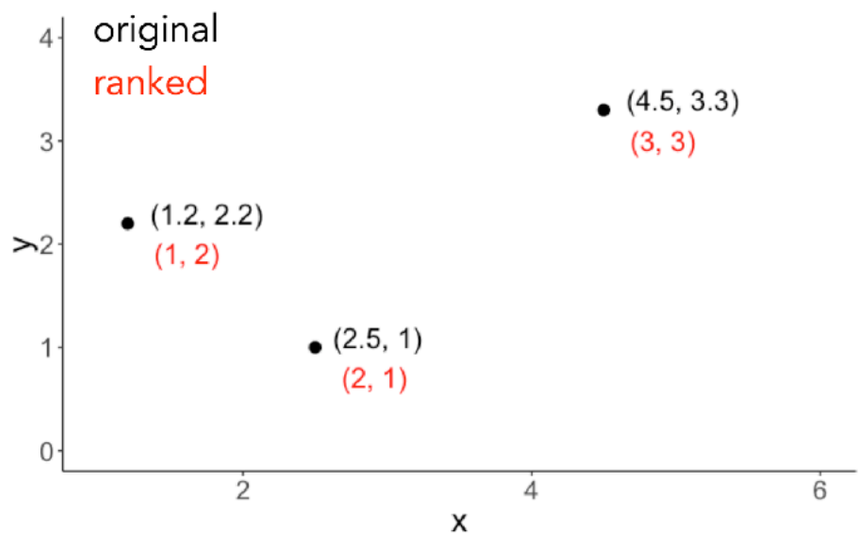
sample correlation coefficient

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

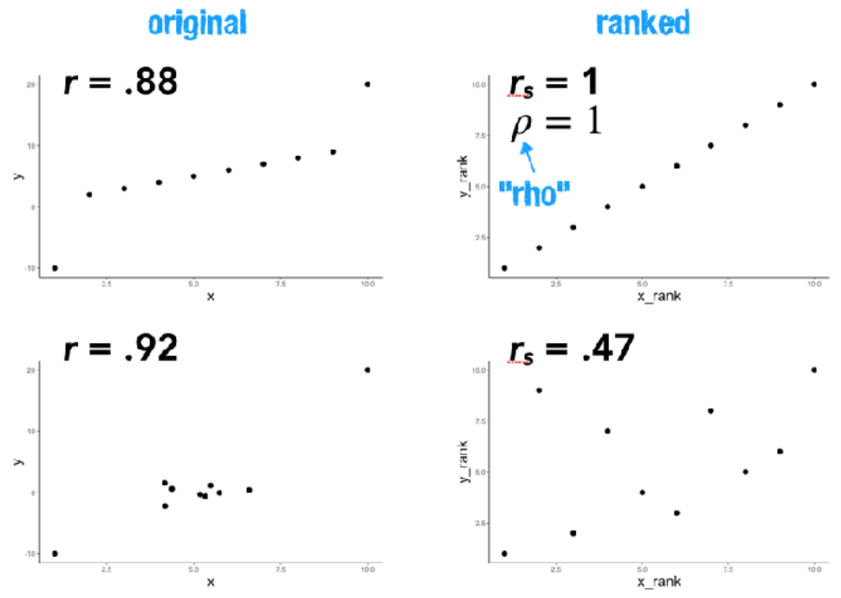
standardized covariation
(dividing by the standard deviations)

Spearman rank order correlation

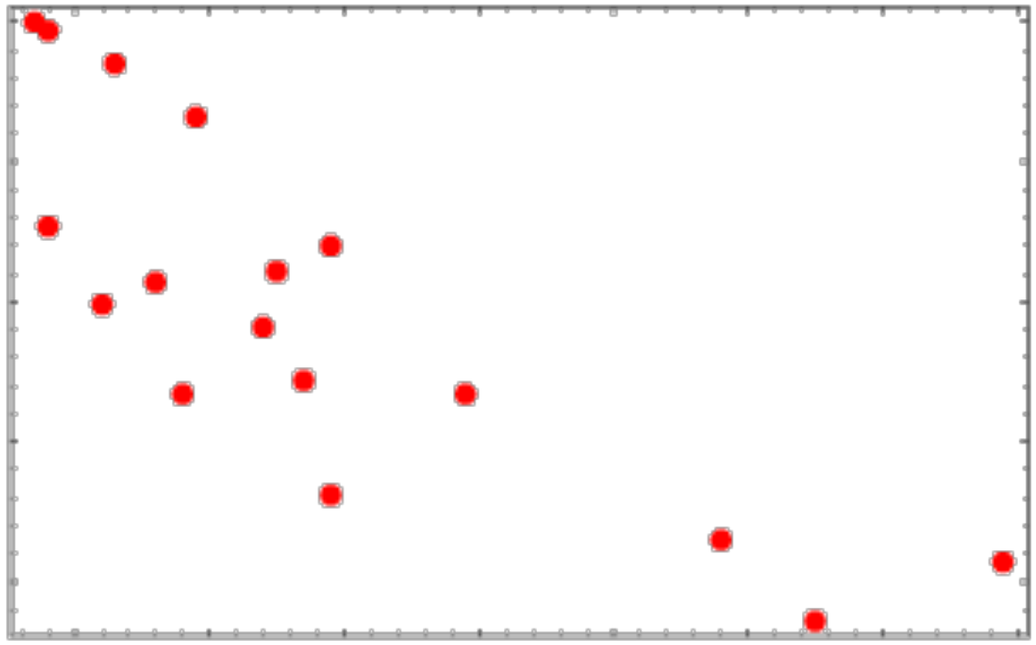
- transform original data into ranks
- calculate correlation on the ranked data



Spearman rank order correlation



Who is the correlation champion?



Winner gets chocolate!

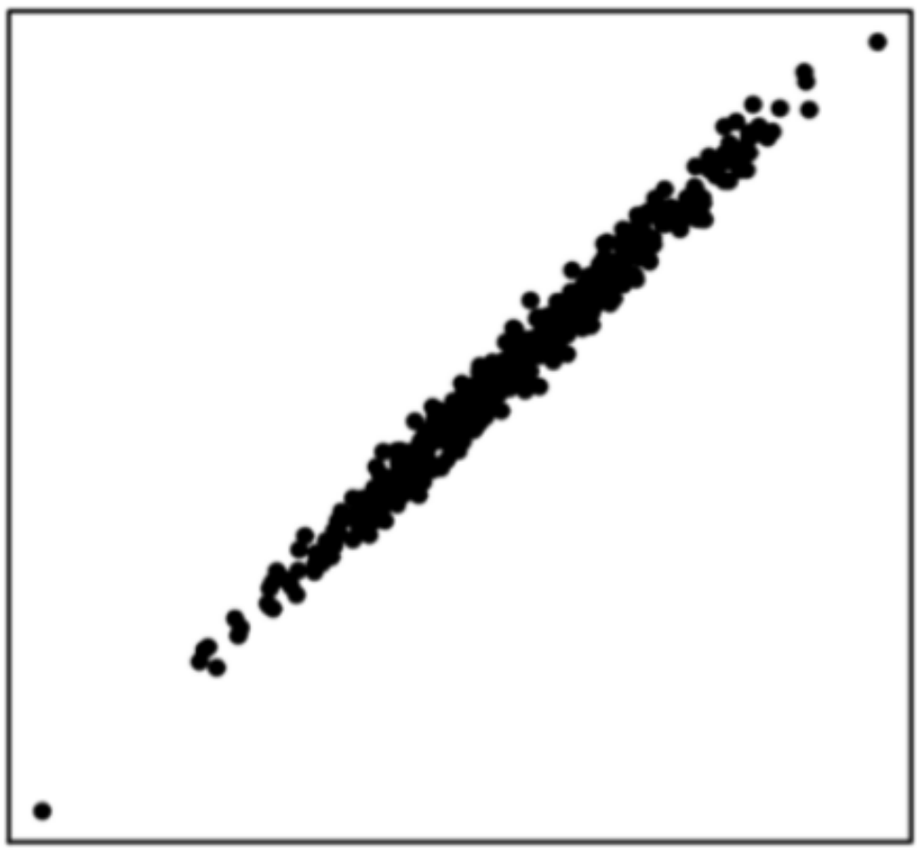
-1 : -0.75
-0.75 : -0.5
-0.5 : -0.25
-0.25 : 0
0 : 0.25
0.25 : 0.5
0.5 : 0.75
0.75 : 1

Who is the correlation champion?

Win up to 1,000 points per answer

In what range is the correlation coefficient?

0



-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

15

0

-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

-0.25 : 0

0 : 0.25

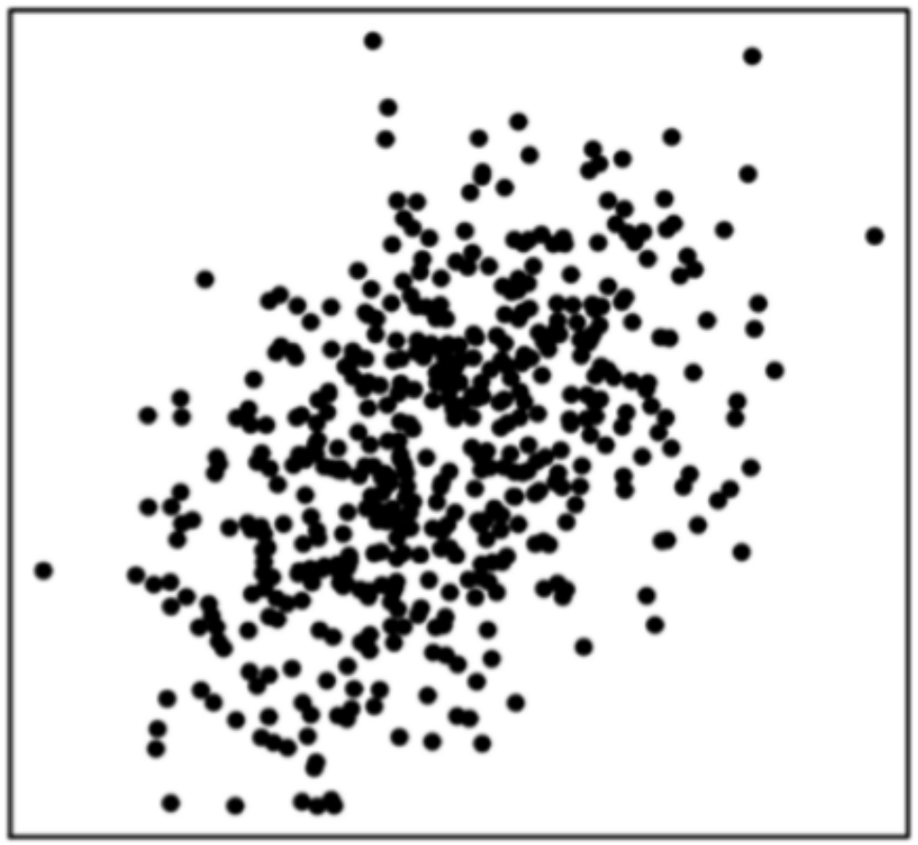
0.25 : 0.5

0.5 : 0.75

0.75 : 1

15

0



-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

-0.25 : 0

0 : 0.25

0.25 : 0.5

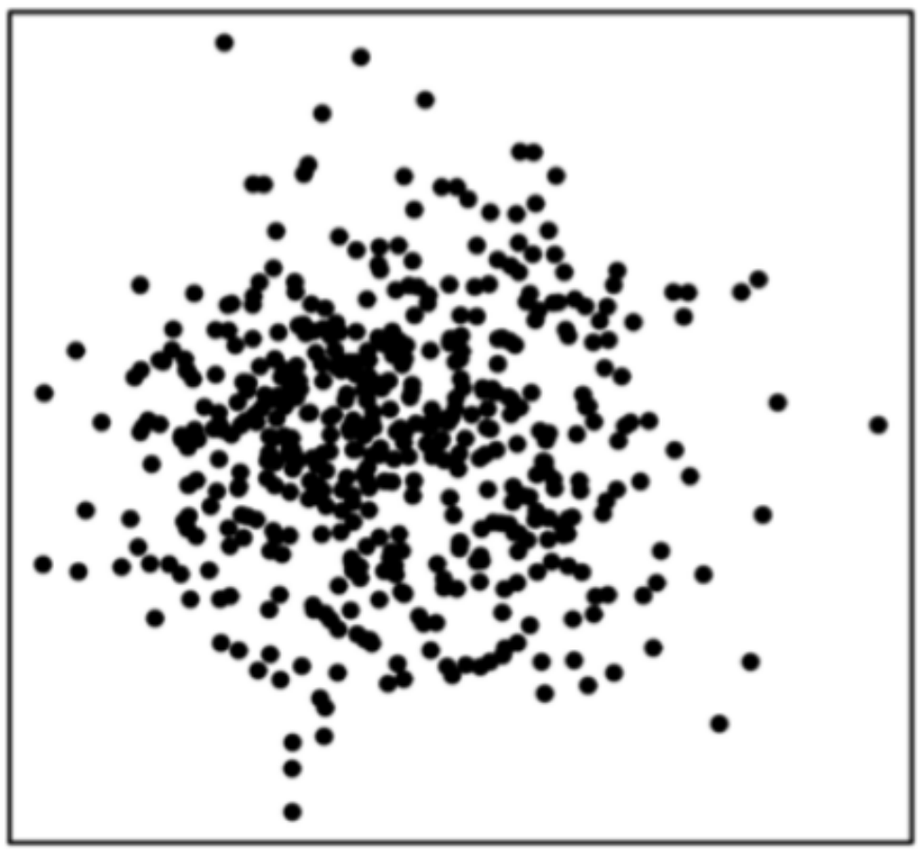
0.5 : 0.75

0.75 : 1

15

Leaderboard

0



-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

-0.25 : 0

0 : 0.25

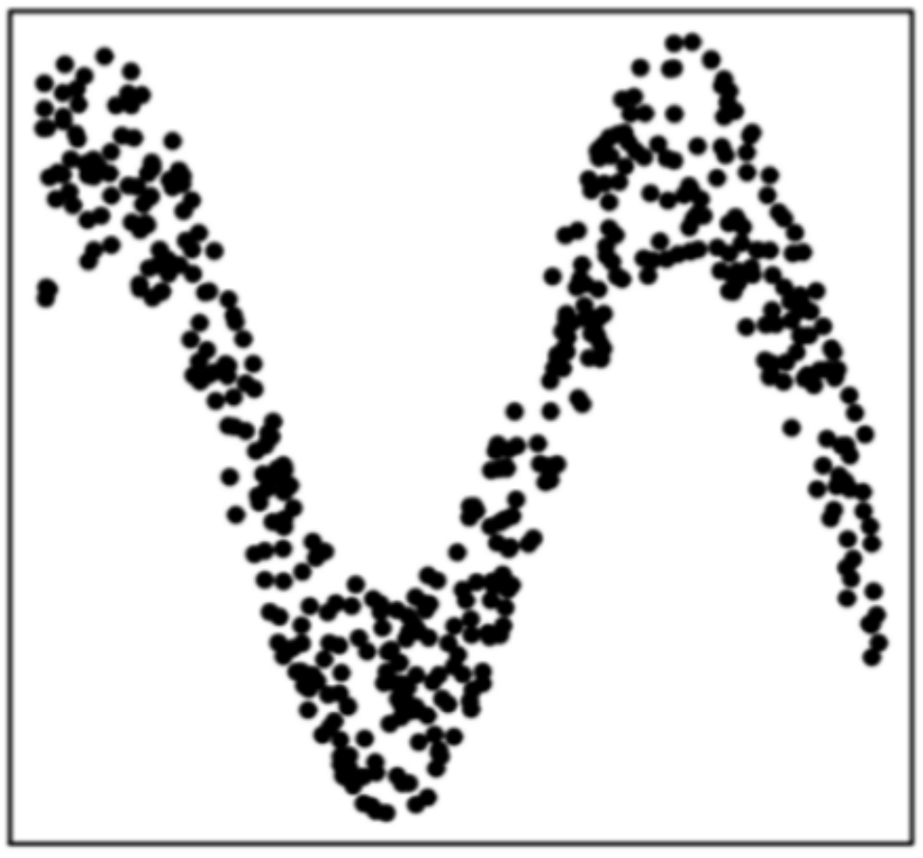
0.25 : 0.5

0.5 : 0.75

0.75 : 1

15

0



-1 : -0.75

-0.75 : -0.5

-0.5 : -0.25

-0.25 : 0

0 : 0.25

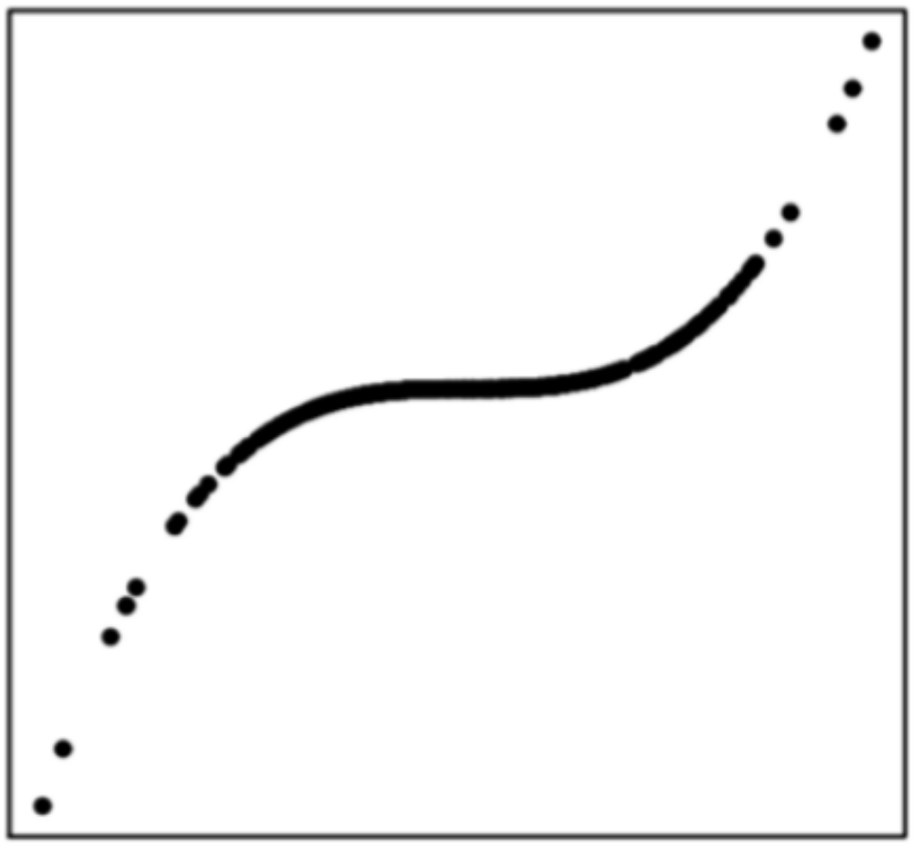
0.25 : 0.5

0.5 : 0.75

0.75 : 1

15

0



-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

15

Leaderboard

Nobody has responded yet.

Solution

XX

Regression

The conceptual tour

Linear model: Simple regression

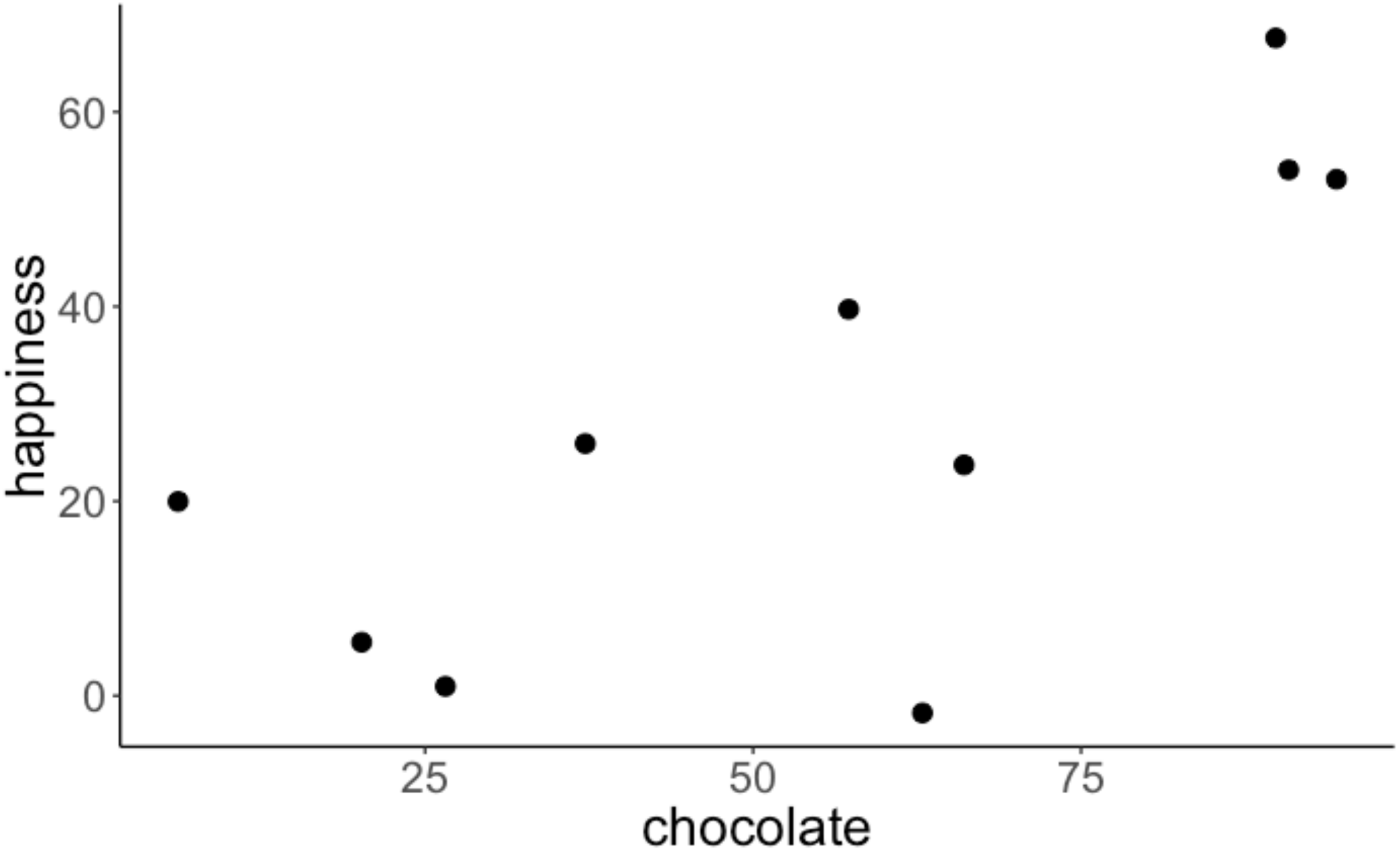
$$\texttt{Data} = \texttt{Model} + \texttt{Error}$$

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

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

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

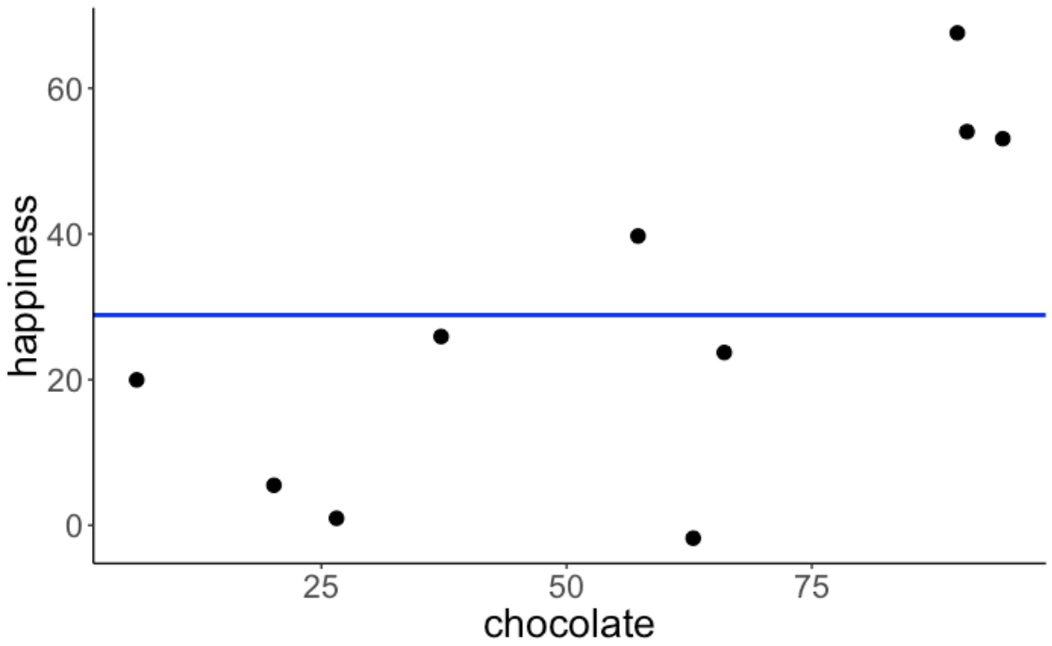
1. Define H_0 as Model C (compact) and H_1 as Model A (augmented)
2. **Fit model parameters to the data**
3. Calculate the proportional reduction of error (PRE) in our sample
4. Decide whether the augmented model is **worth it** by comparing the observed PRE in our sample to the sampling distribution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

$$Y_i = 28.88 + e_i$$

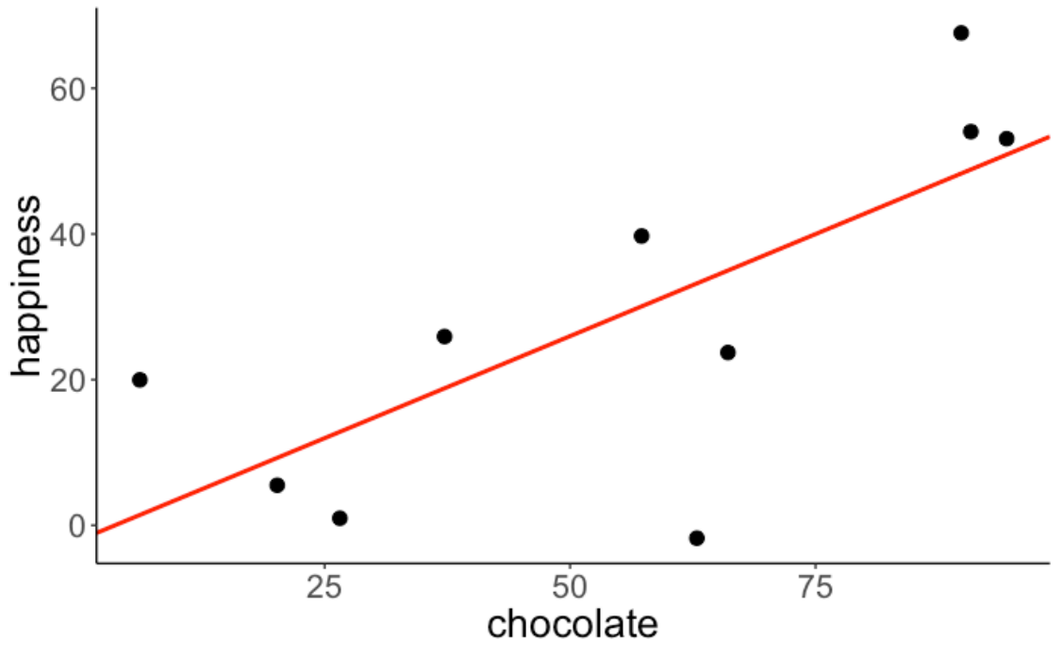
H_1: Chocolate consumption and happiness are related.

Model A

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

chocolate consumption

Model prediction



Fitted model

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

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)

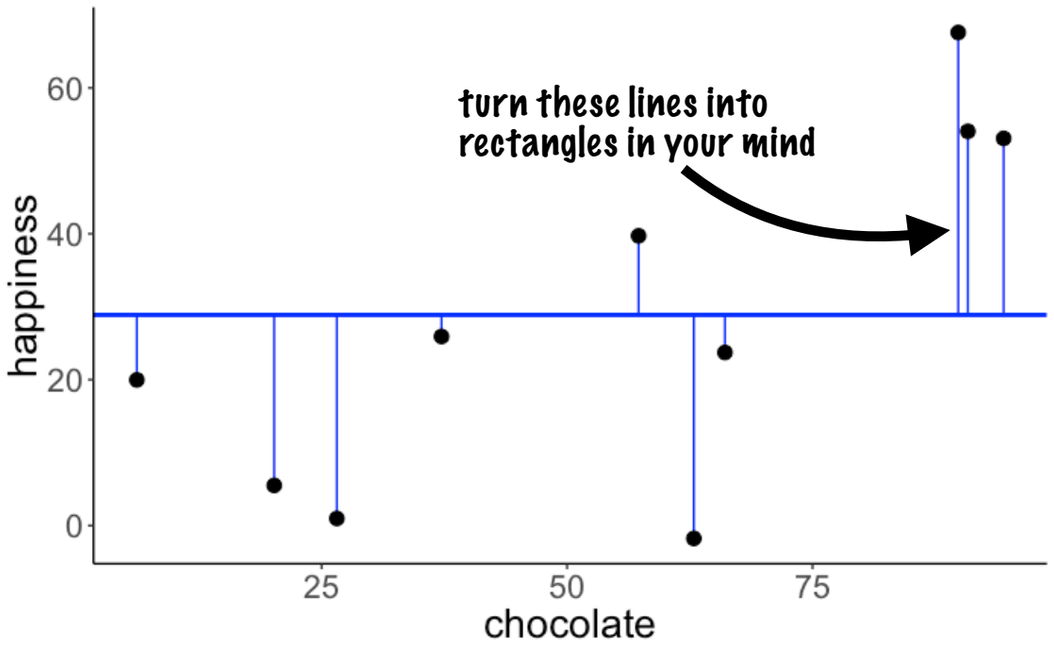
Calculate PRE

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

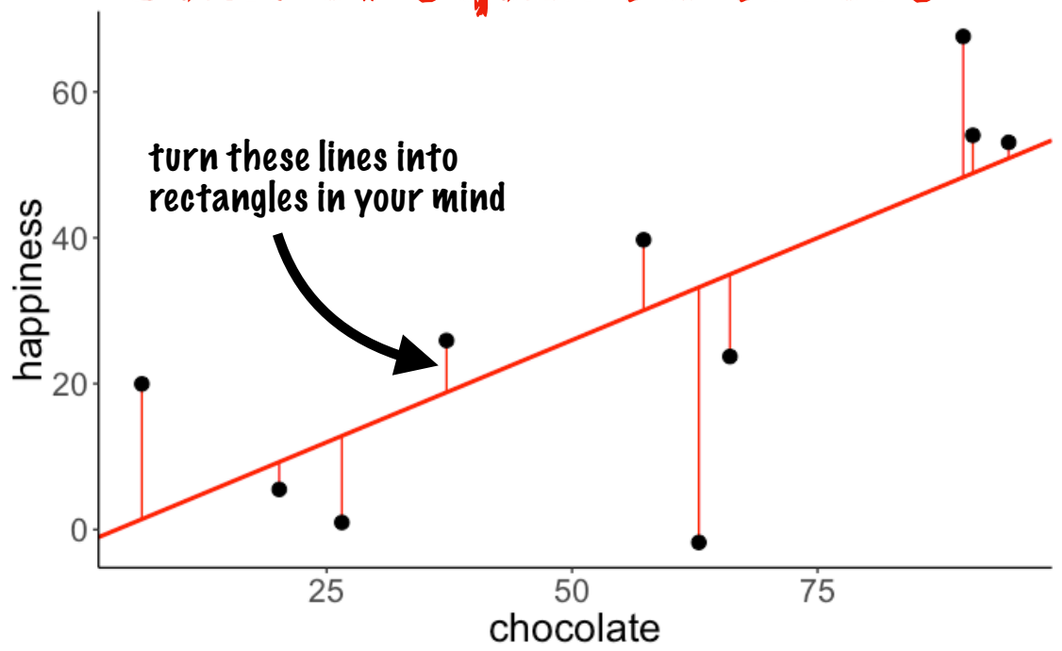
Both models were fit to minimize the sum of squared errors

OLS = Ordinary **least squares** regression

Sum of squared errors



Sum of squared errors



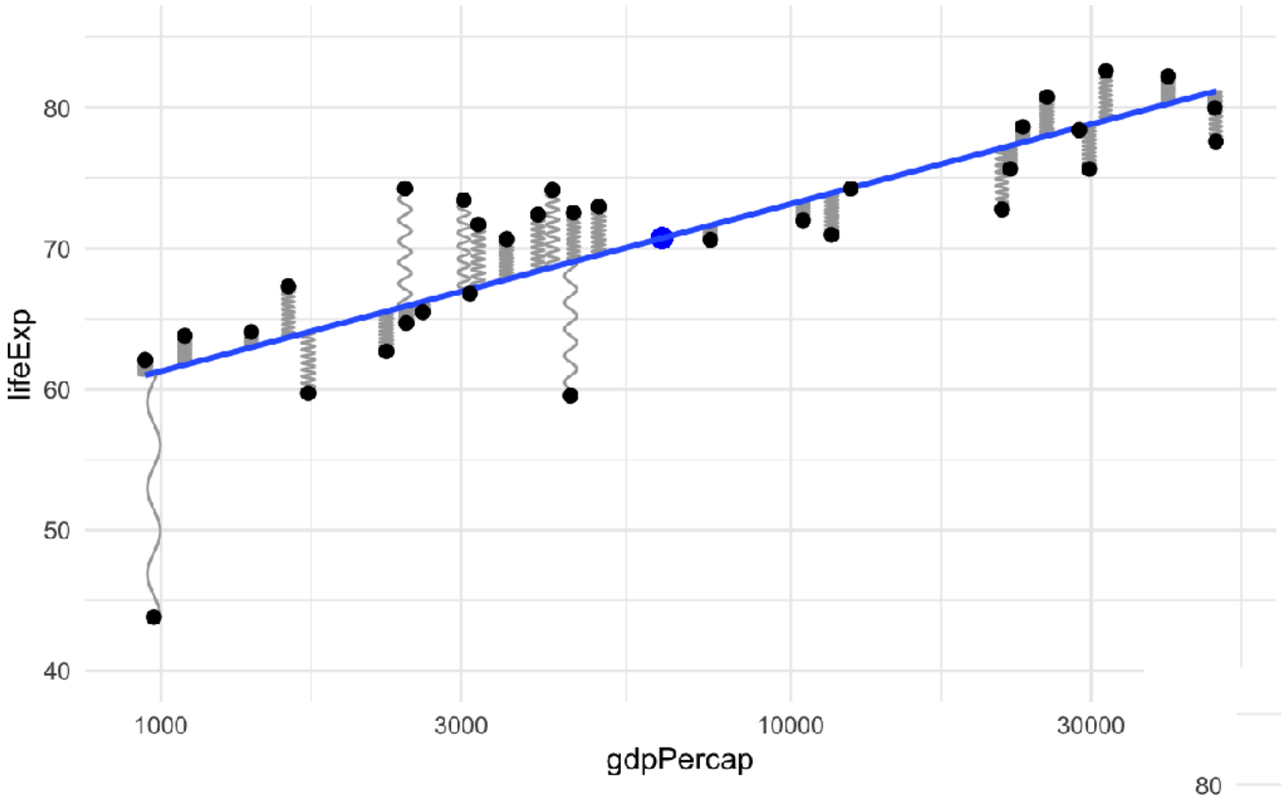
SSE(C) = 5215.016

SSE(A) = 2396.946

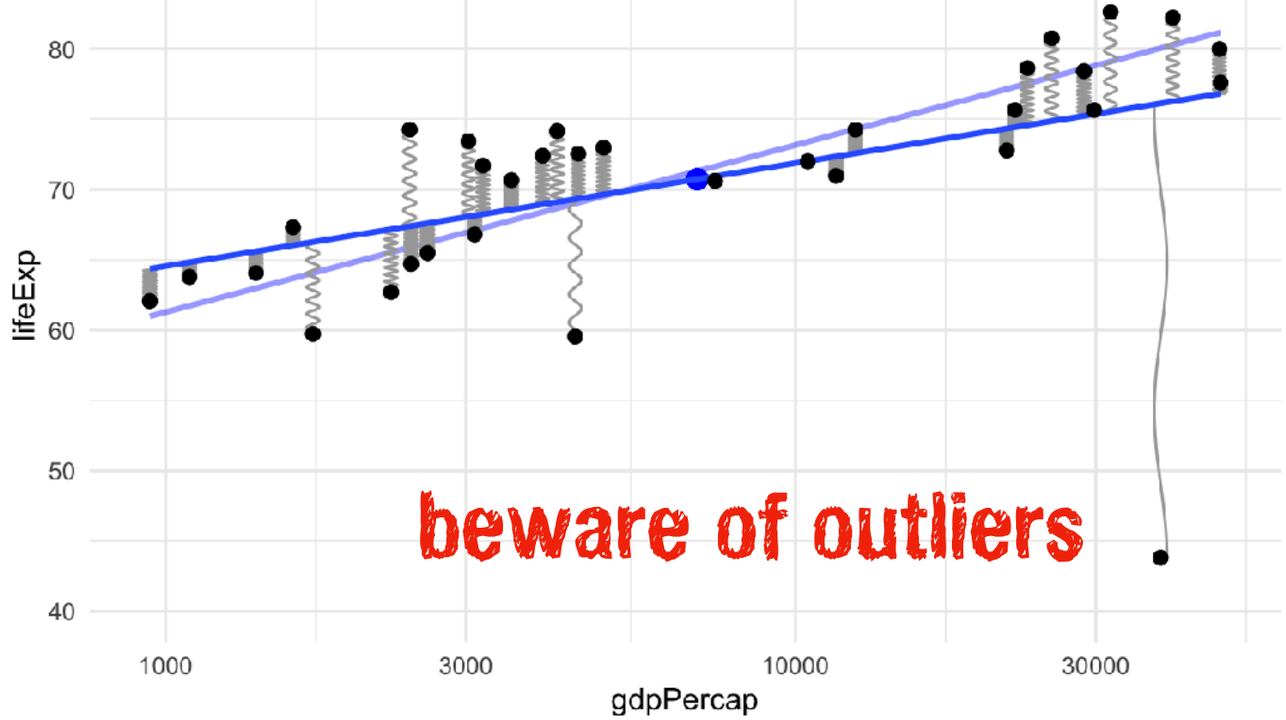
$$\text{PRE} = 1 - \frac{2396.946}{5215.016} \approx 0.54$$

The augmented model reduces the error by 54%.

Least squares as springs



each point is attached to the line with an identical spring



https://joshualoftus.com/posts/2020-11-23-least-squares-as-springs/

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

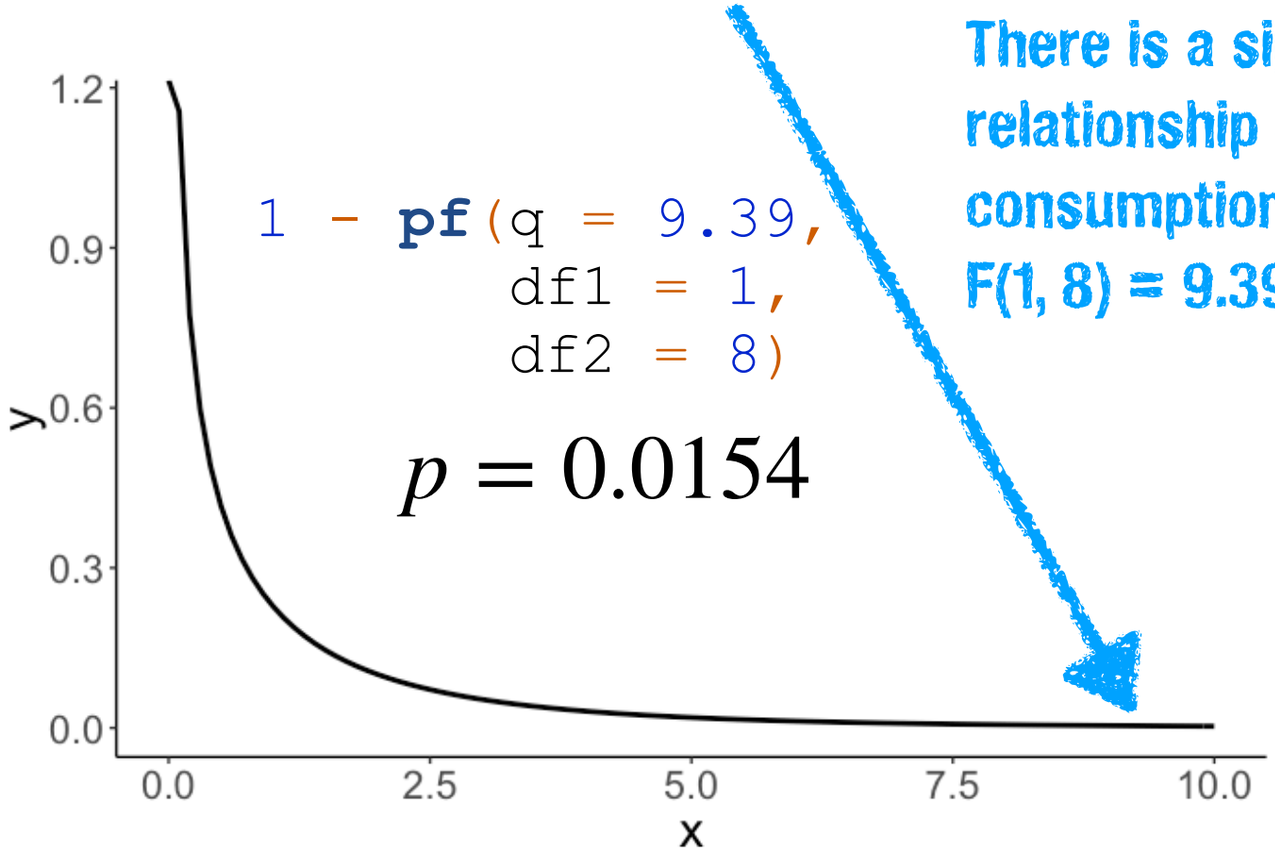
Decide whether it's **worth it**

- To compute the *F* statistic, we need:
  - PRE = 0.54
  - PC = 1
  - PA = 2
  - $n = 10$

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

$$= \frac{0.54/(2 - 1)}{(1 - 0.54)/(10 - 2)}$$

$$= 9.39$$

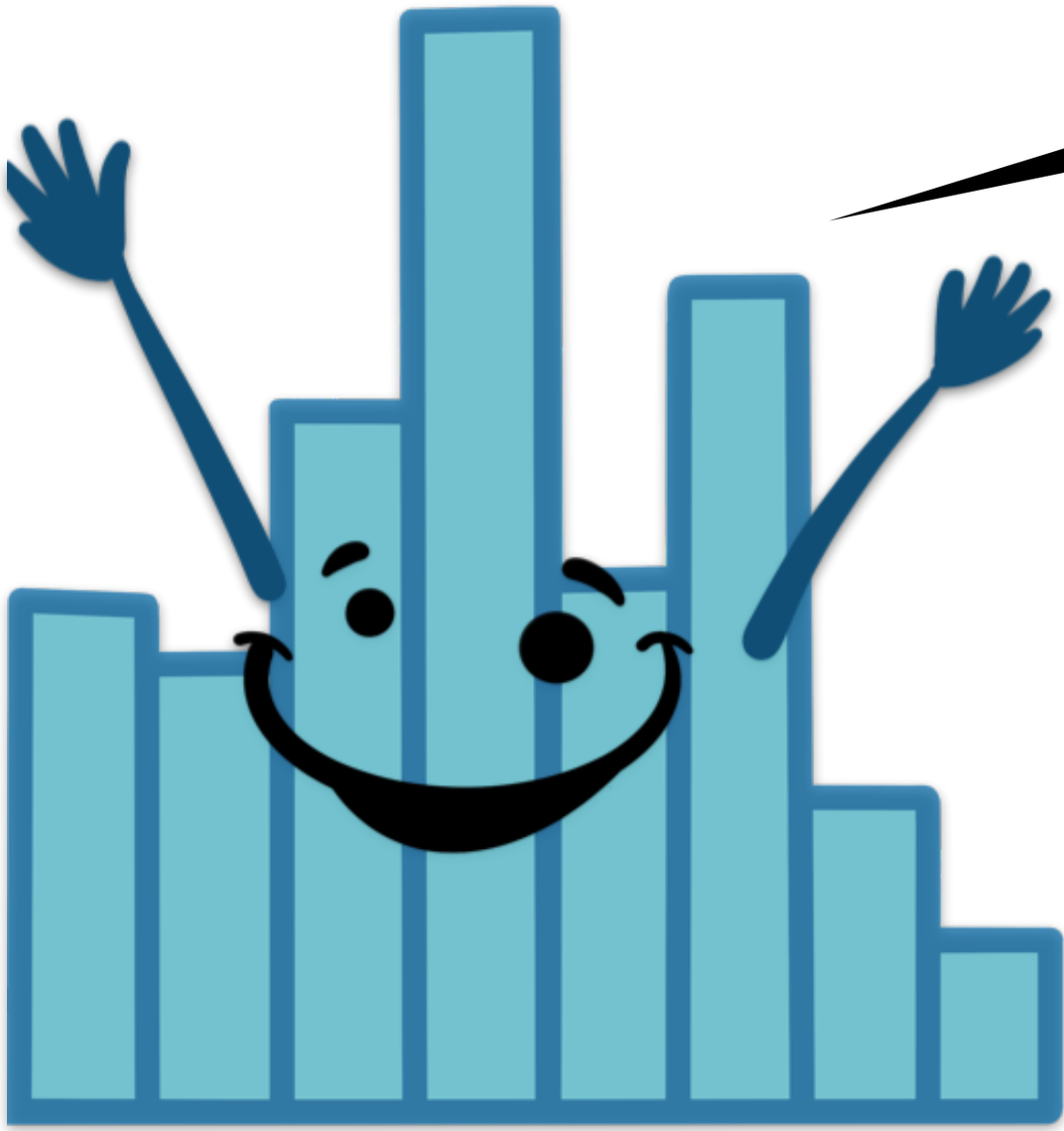


There is a significant relationship between chocolate consumption and happiness F(1, 8) = 9.39, p = .0154.

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

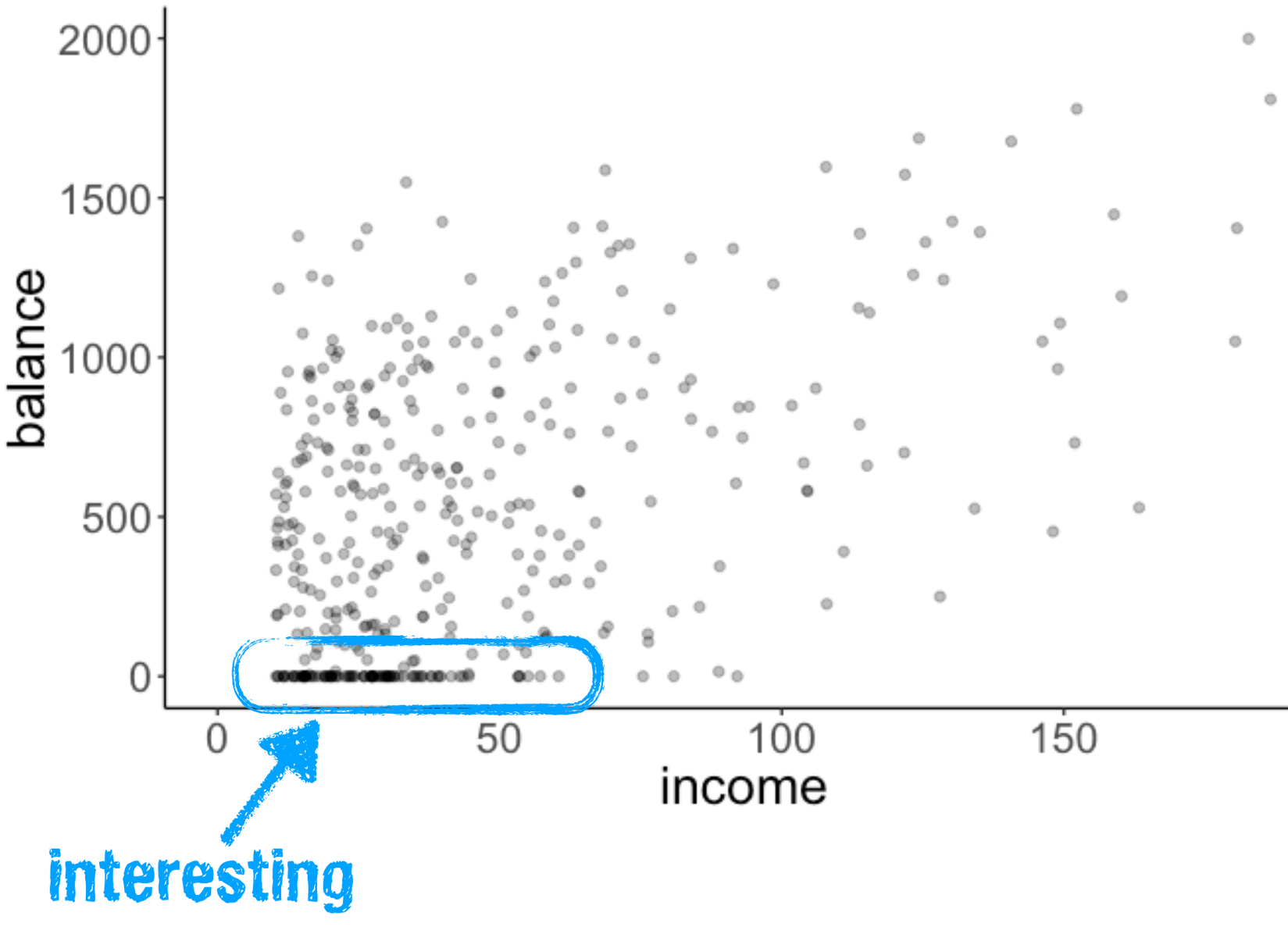
variable	description
income	in thousand dollars
limit	credit limit
rating	credit rating
cards	number of credit cards
age	in years
education	years of education
gender	male or female
student	student or not
married	married or not
ethnicity	African American, Asian, Caucasian
balance	average credit card debt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

`nrow(df.credit) = 400`

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

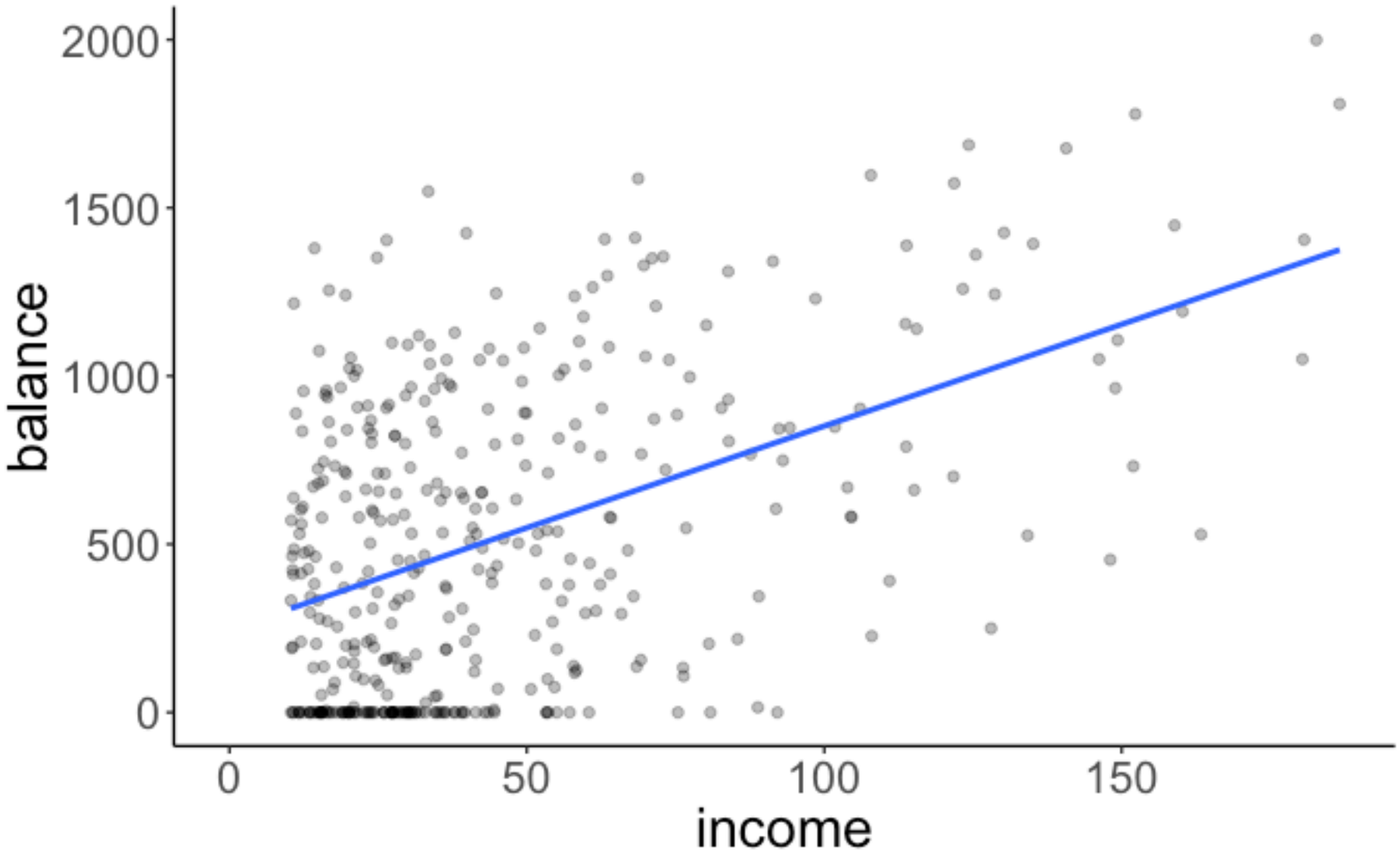
Always plot the data first ...

```
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)
```

outcome — intercept — predictor — data

(doesn't need to be specified explicitly)

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

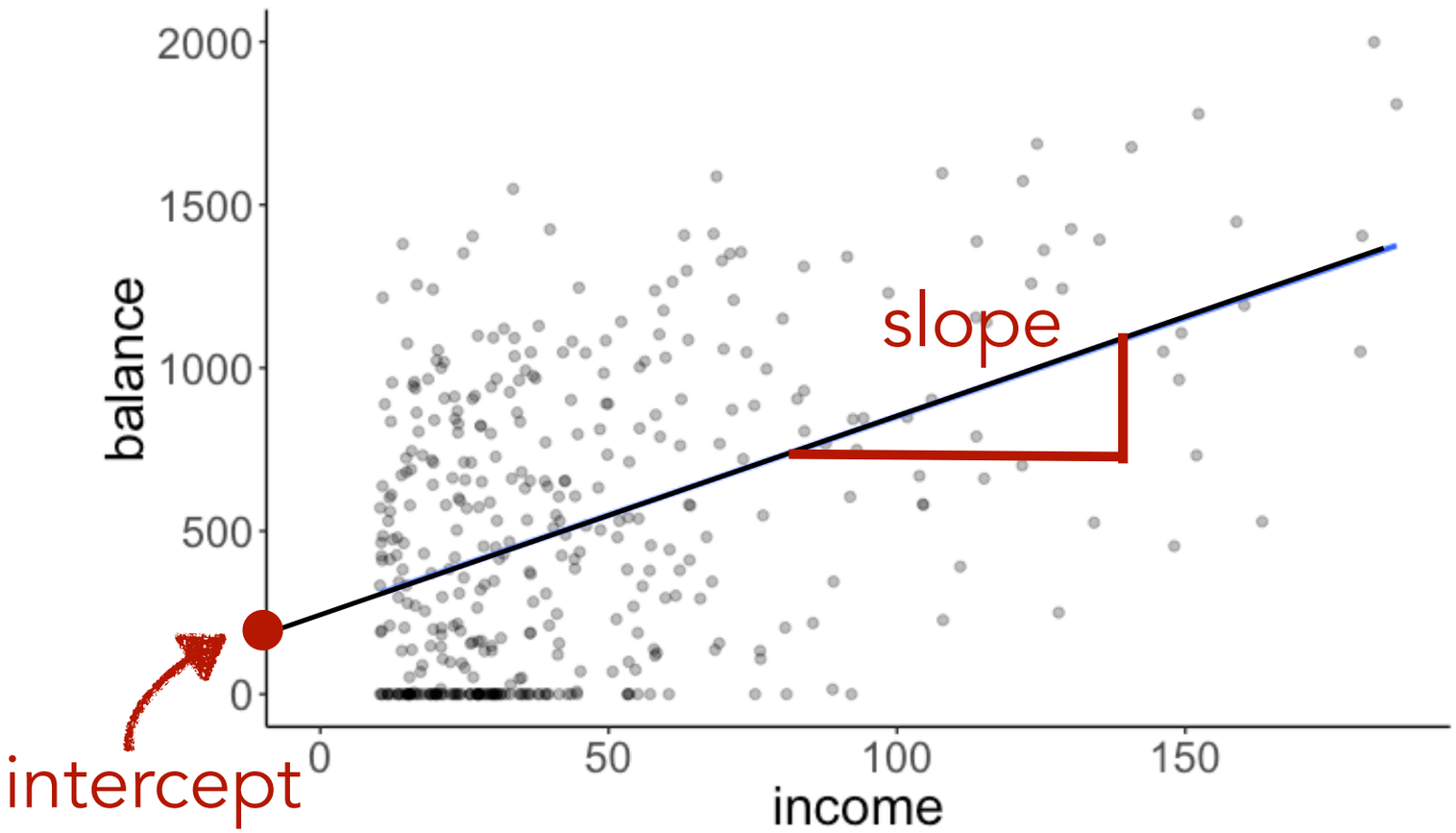
Interpreting regression parameters

```
Coefficients:
(Intercept)          income
    246.515           6.048
```

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

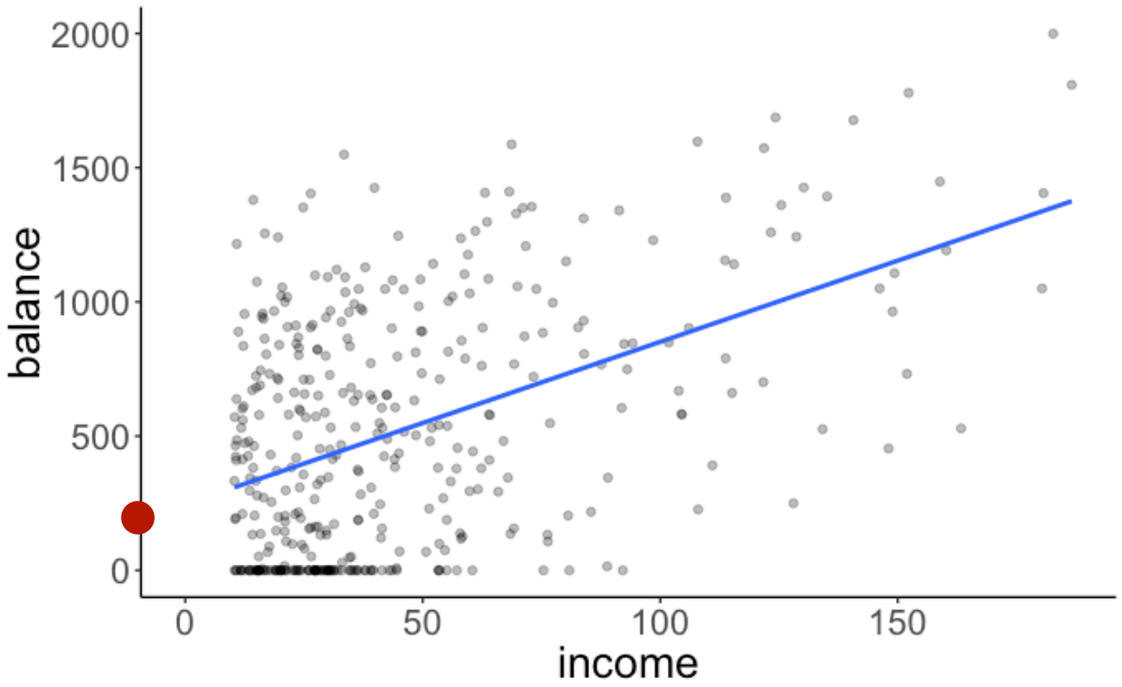
$$\text{balance}_i = b_0 + b_1 \cdot \text{income}_i + e_i$$

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

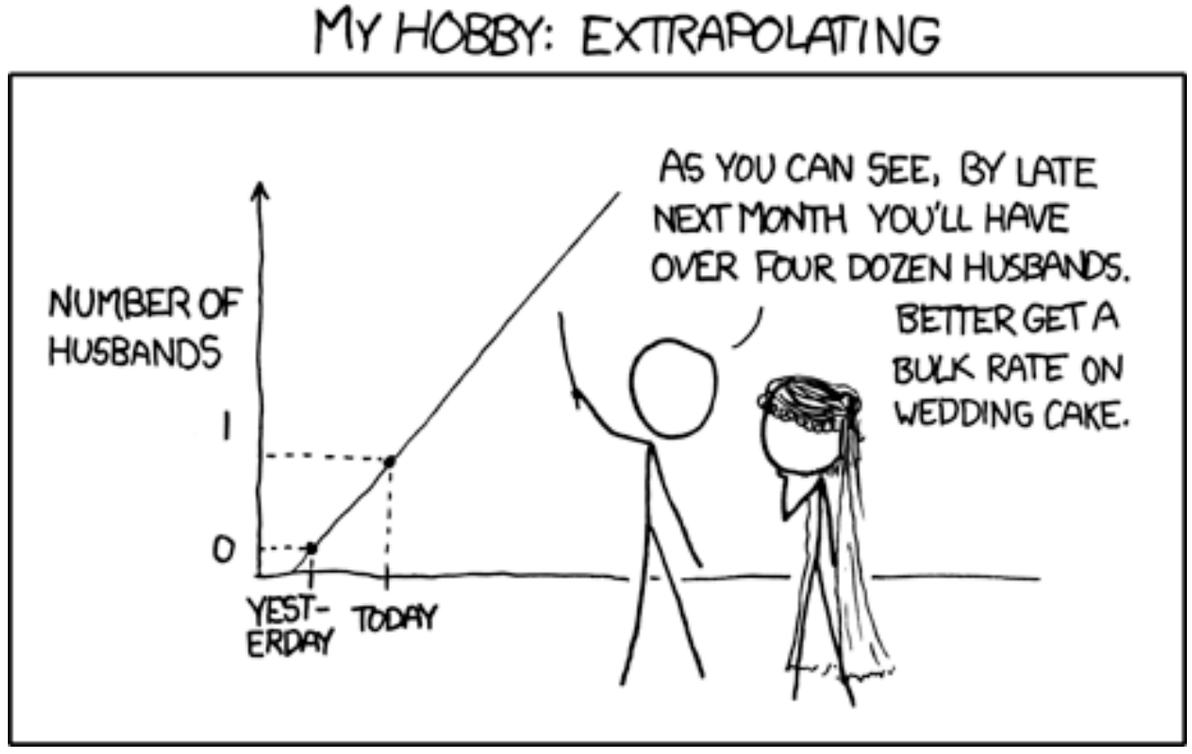


For each additional thousand dollars income, a person's average credit card is predicted to increase by $6.05.

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	.cooksd	.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

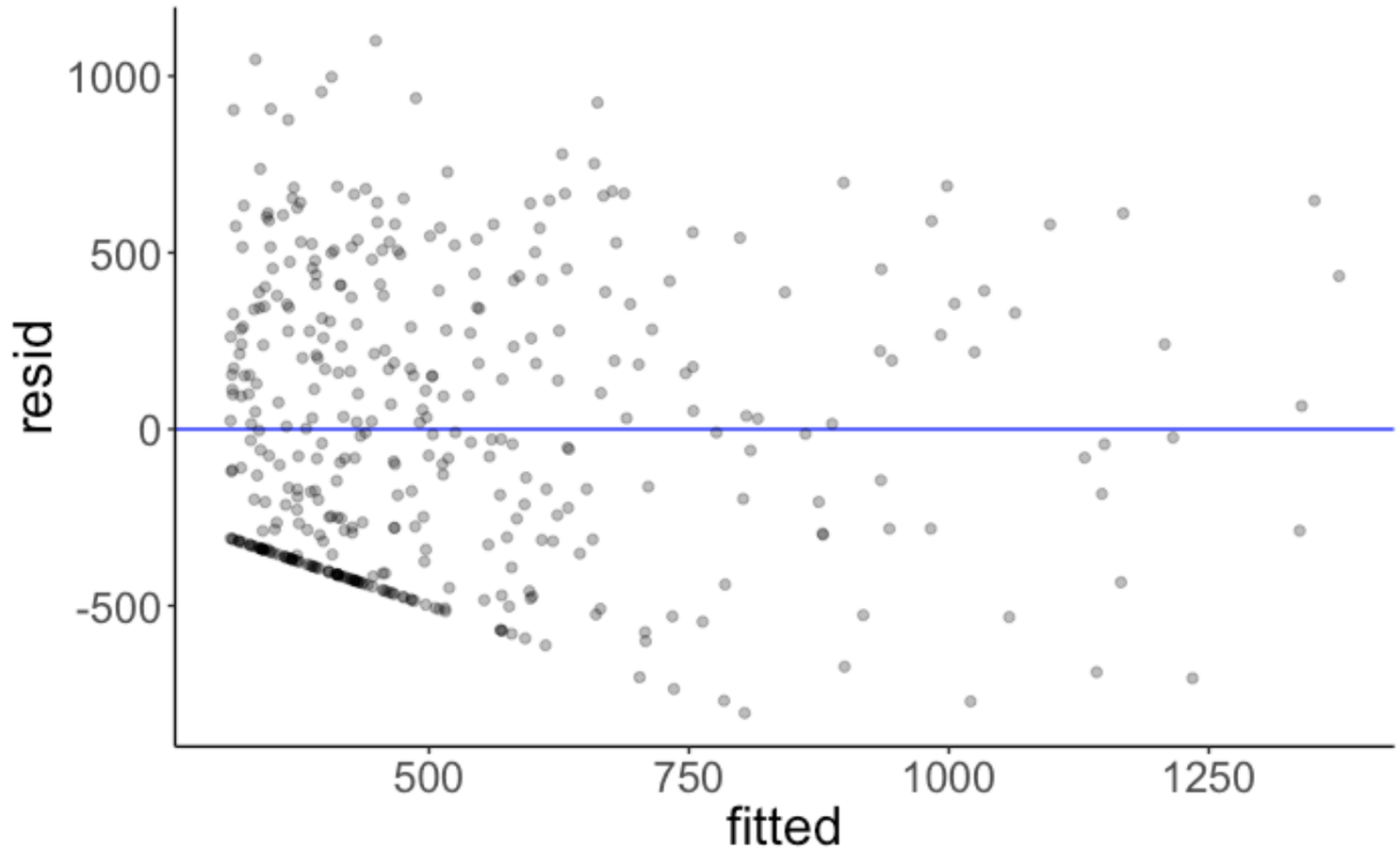
`summary()`

```
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

Residual plot



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

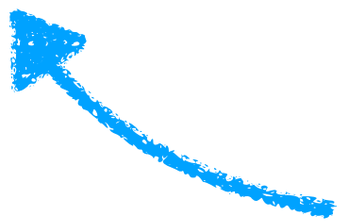
summary()

```
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)

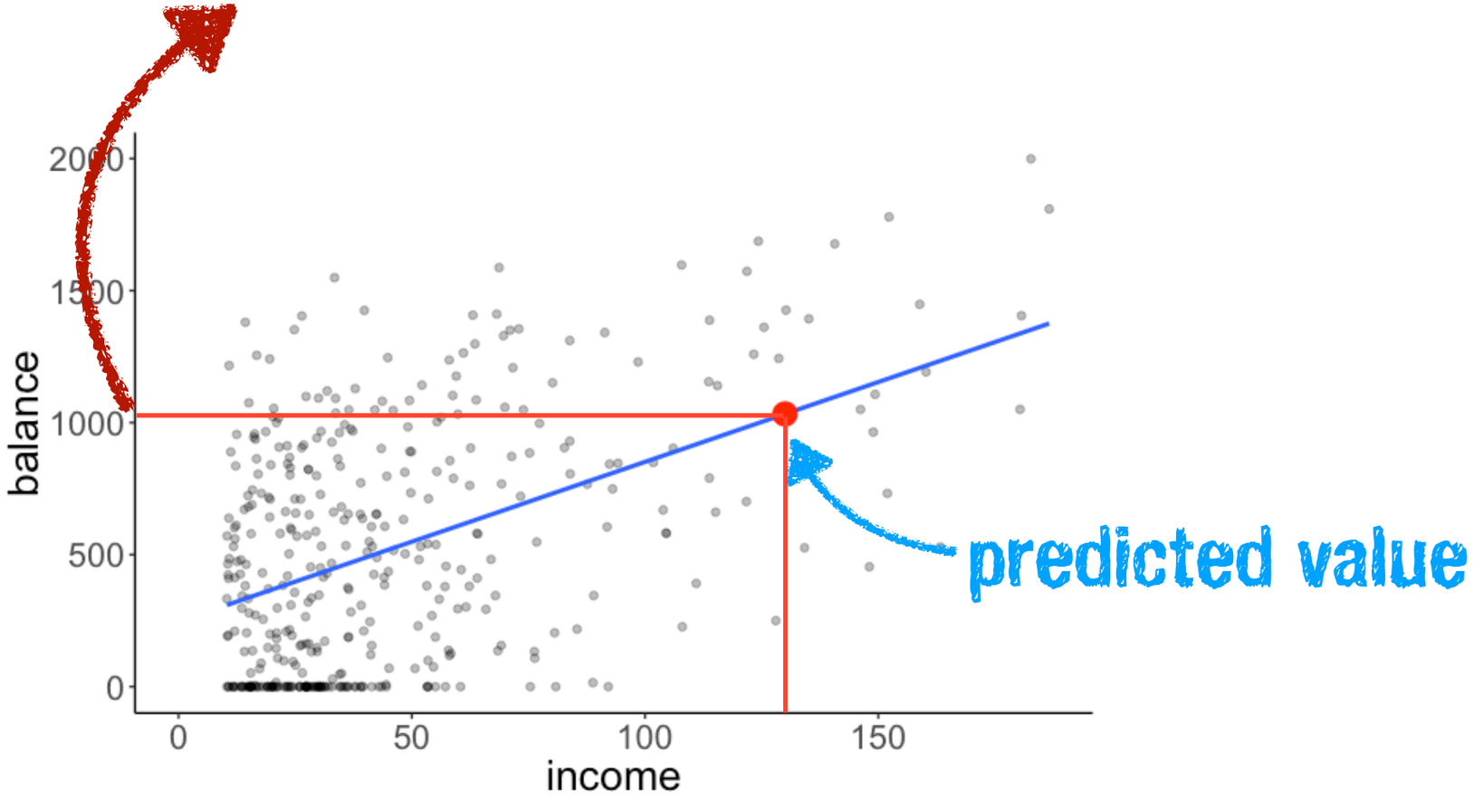
Making predictions

```
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

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

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

---

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)

```
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*

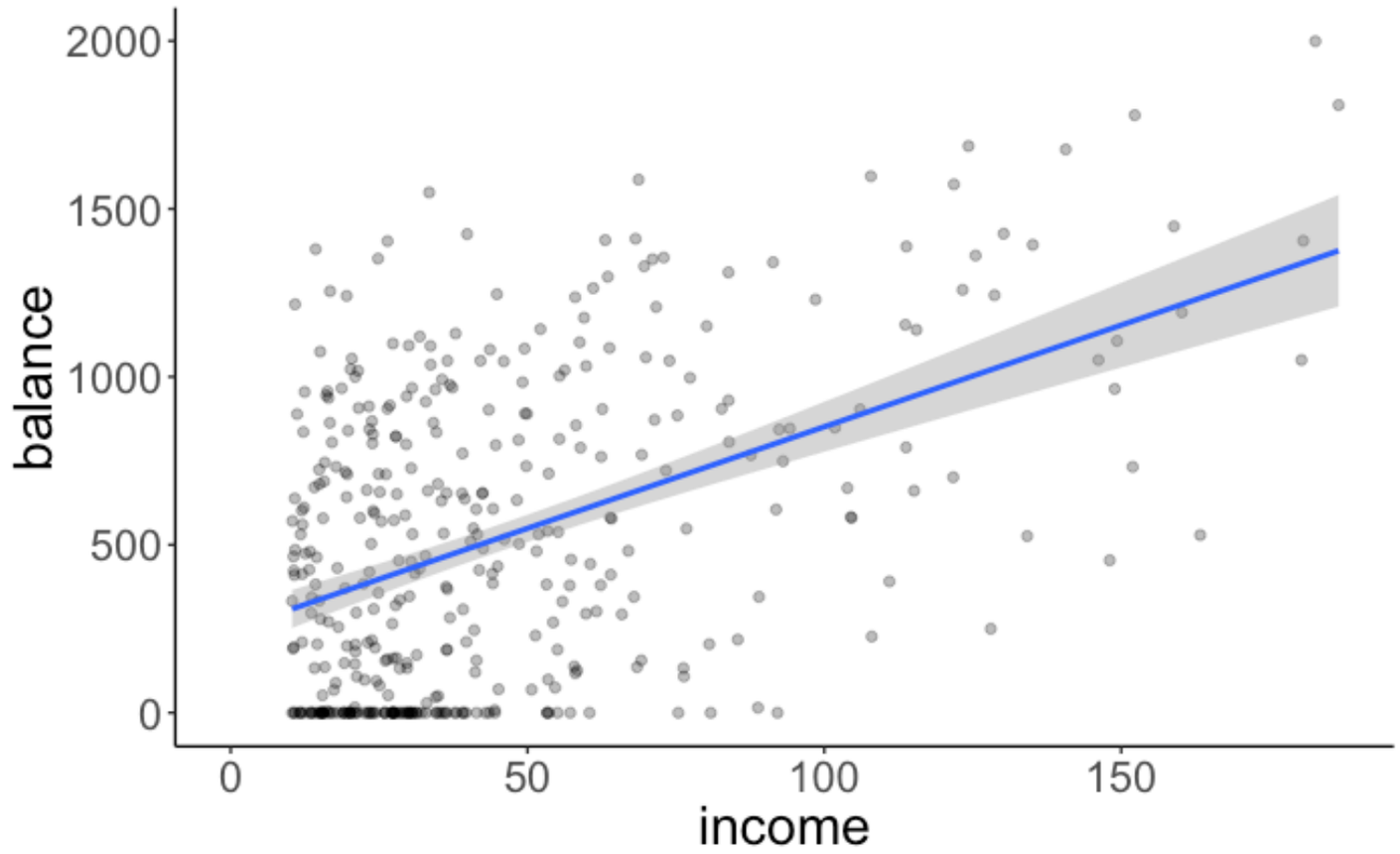
```
cor(df.credit$balance,
    df.credit$income)
```

$$R^2 = 0.215$$

$$r = .463$$

effect size measure

Reporting the results

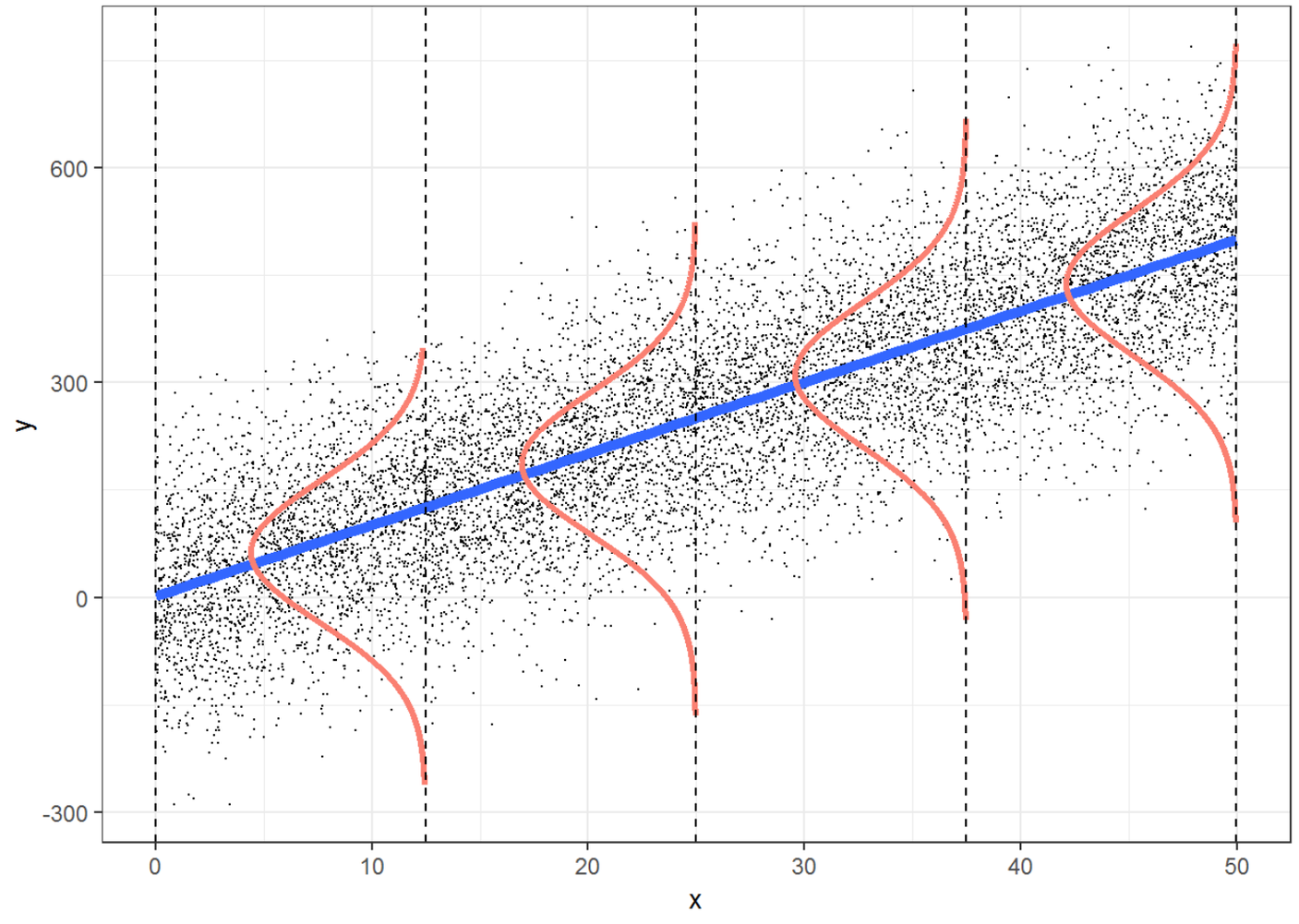


There is a significant relationship between a person's income and the average balance on their credit cards $F(1, 389) = 108.99$, $p < .001$, $r = .463$.

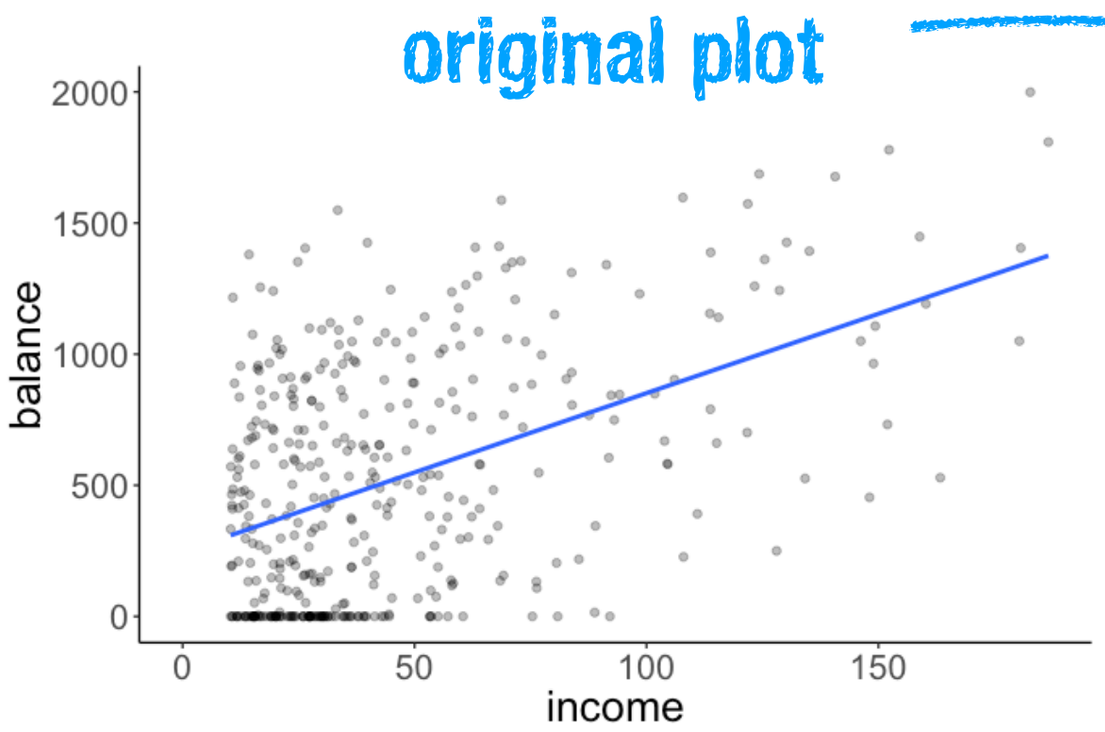
With each additional $1000 of income, the average balance is predicted to increase by $6.05 [4.91, 7.19] (95% CI).

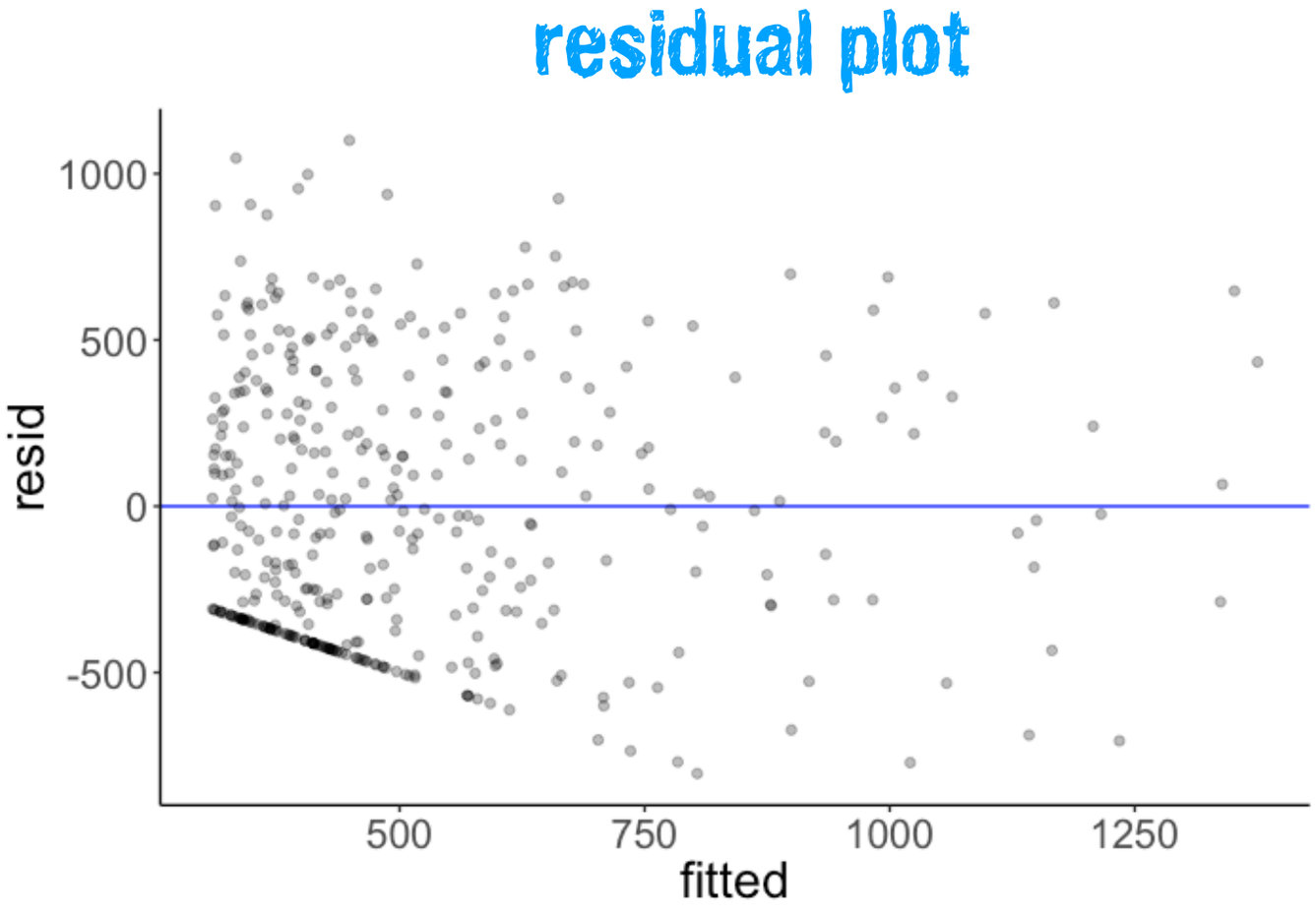
Model assumptions

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

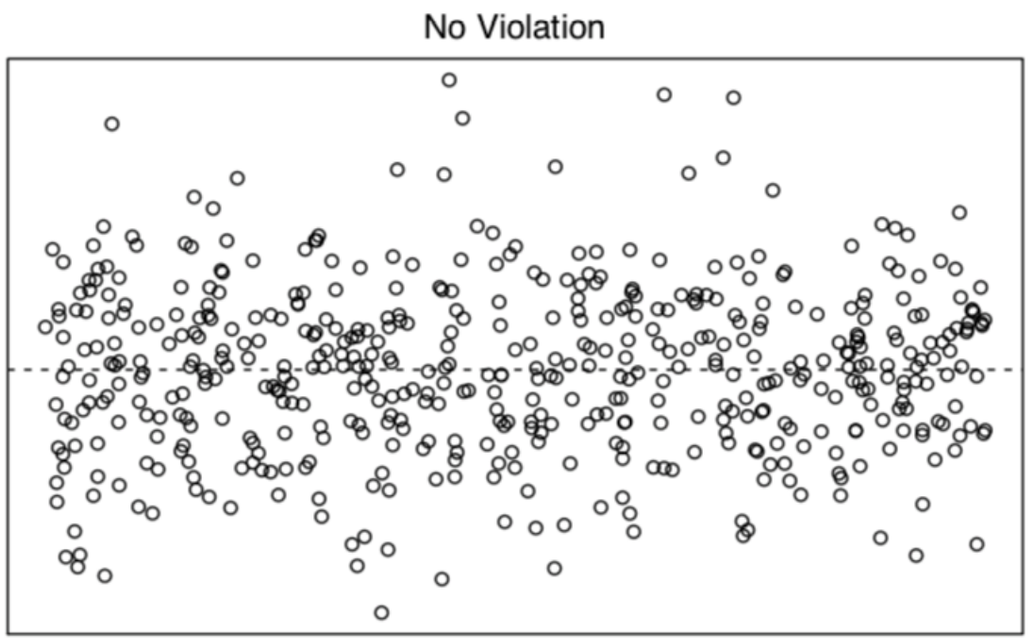


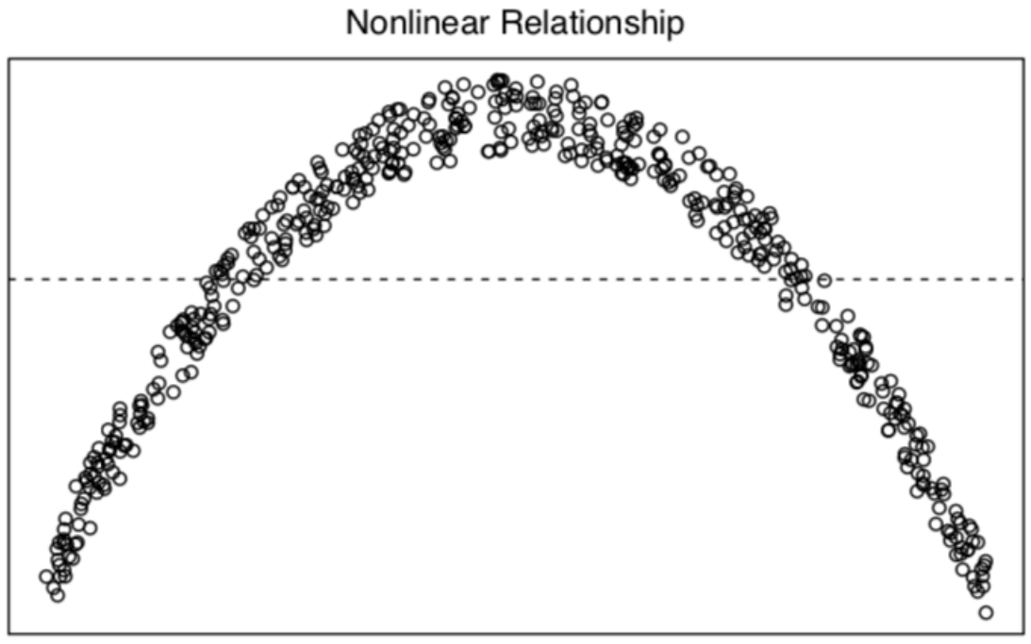
Model assumptions

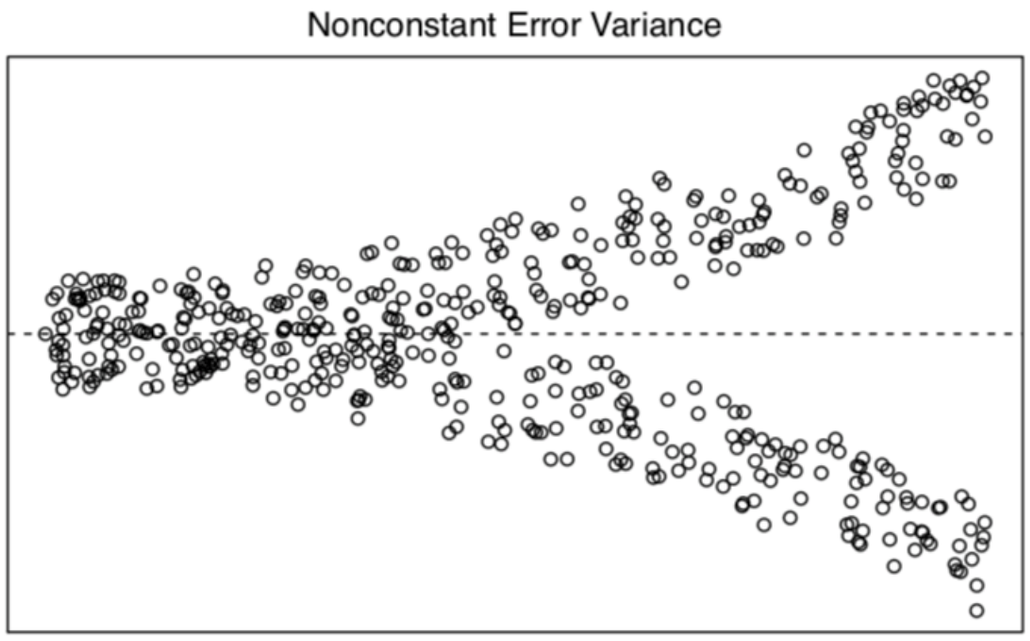


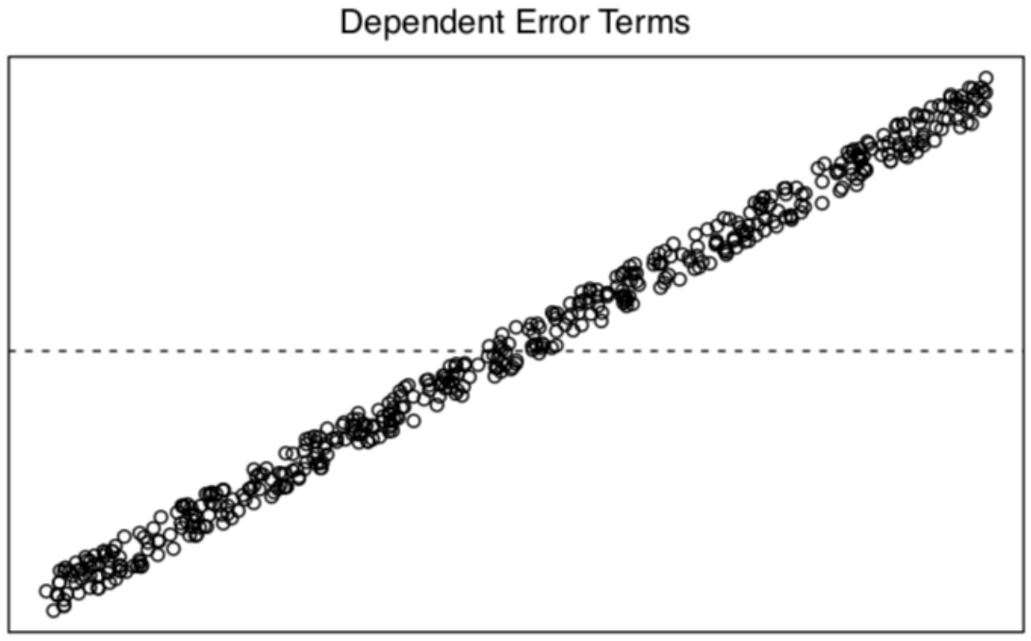


Model assumptions







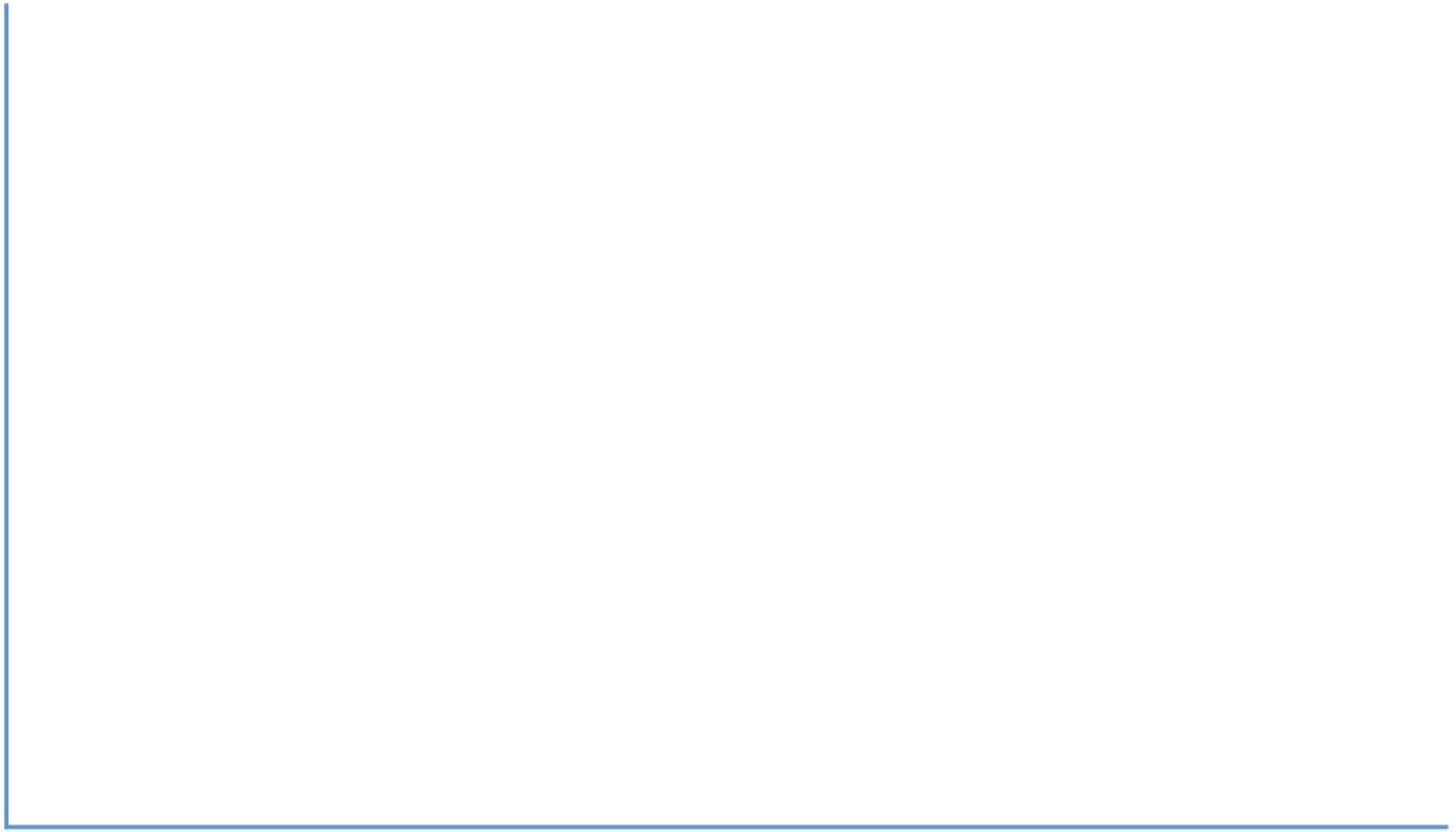


Plan for today

- Quick recap
- Who is the correlation champ?
- 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?

Thank you!