Multiple Linear Regression

“The height of sophistication is simplicity.”
— Clare Boothe Luce

👨‍🏫 Statistical Reasoning Lecture #2
Alexander Savi, 2025
📚 Mehmetoglu & Mittner Ch. (4, 5), 8, 10.1, .2, .4, .7, .8

Untitled by Koen Derks (aRtsy package) 🎨

News

INTERVIEW

‘Trump neemt risico’s op basis van gevoel, niet op basis van analyse. Het is allemaal improvisatie’

Nate Silver | statisticus, pokerspeler Nate Silver kijkt als cijferaar naar de Amerikaanse politiek en samenleving. Hij gelooft in risico’s nemen om te willen winnen. „De rijkste mensen van de wereld verdubbelen hun rijkdom elke tien jaar.”

Bas Blokker

1 september 2025 om 15:48 • Leestijd 7 minuten

— NRC (Sep. 1, 2025)

📢 Contact

How

Direct

- ❏ Lecture (WF/SR)
- ❏ Tutorial (WF)
- ❏ Consultation hour (SR)

Indirect (if not otherwise possible)

- ❏ Canvas or email

Who

Is it about WF?

- ❏ Tutorial teacher (initial point of contact)
- ❏ Jolien (substantive questions, during lecture)

Is it about SR?

- ❏ Substantive: Alexander
- ❏ Weekly assignments: Jonas

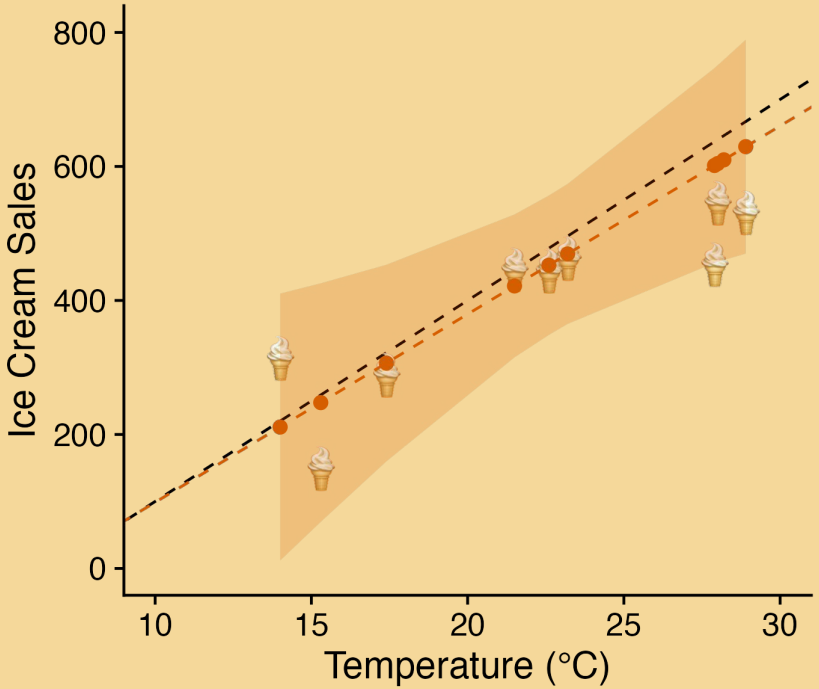
Is it cross-course?

- ❏ Alexander (course coordinator)

📢 Announcements

- ❏ SMASH
- ❏ Web lectures & attendance
- ❏ Personal course manual (for learning, can’t bring to exam)
- ❏ Online consultation with Jonas (restricted access to UvA accounts)

🔁 Recap



- ❏ RMSE
- ❏ Explained variance R^2
- ❏ *t* statistic
- ❏ Effect size f^2
- ❏ Confidence interval

🔮 Today

Topics

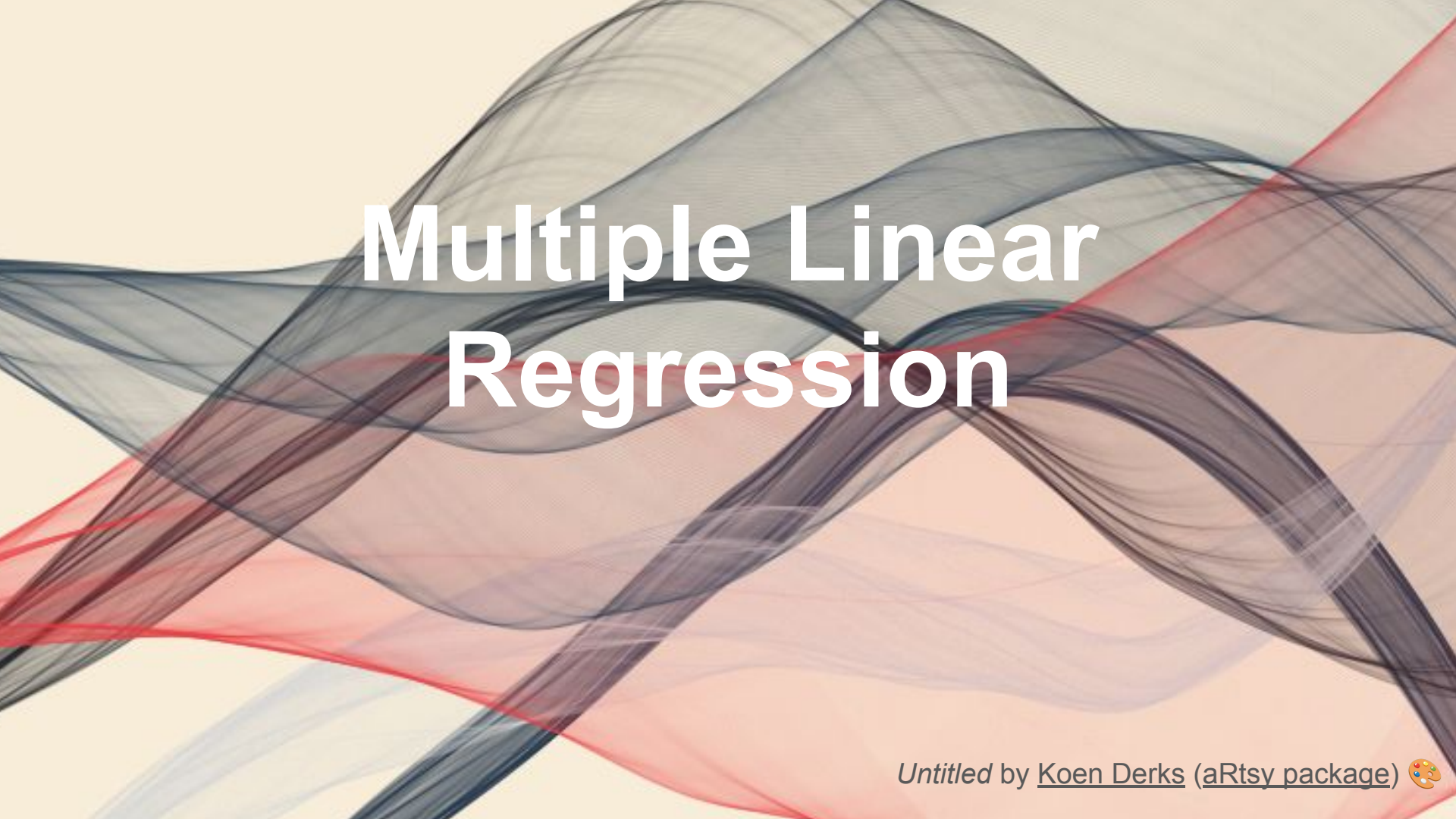
Statistical reasoning with GLM
Multiple linear regression
| Multiple linear regression
| Moderation/interaction analysis
Dummy-variable regression
Logistic regression
Multilevel and longitudinal analysis
Bayesian statistics

Learning goals

Estimate the relationships between more than two continuous variables.

Determine whether the relationship between two continuous variables depends on a third continuous variable.

Multiple Linear Regression



Iris

Q. Are the dimensions of the petals and sepals of the iris flower related?

H. The length of a petal is related to the length and the width of a sepal.

E. […]

💡 A data set made famous by Ronald Fisher and with its very own Wikipedia page.

More than one (continuous) independent variables

 Willem the Supervillain

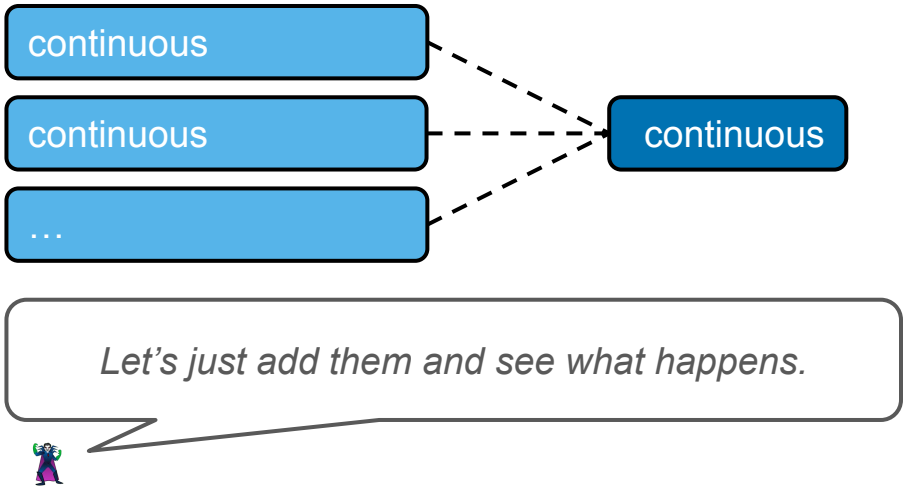
I'm on it, let's move fast and break things.

 Miró the Superhero

Sure, let's move carefully and fix things.

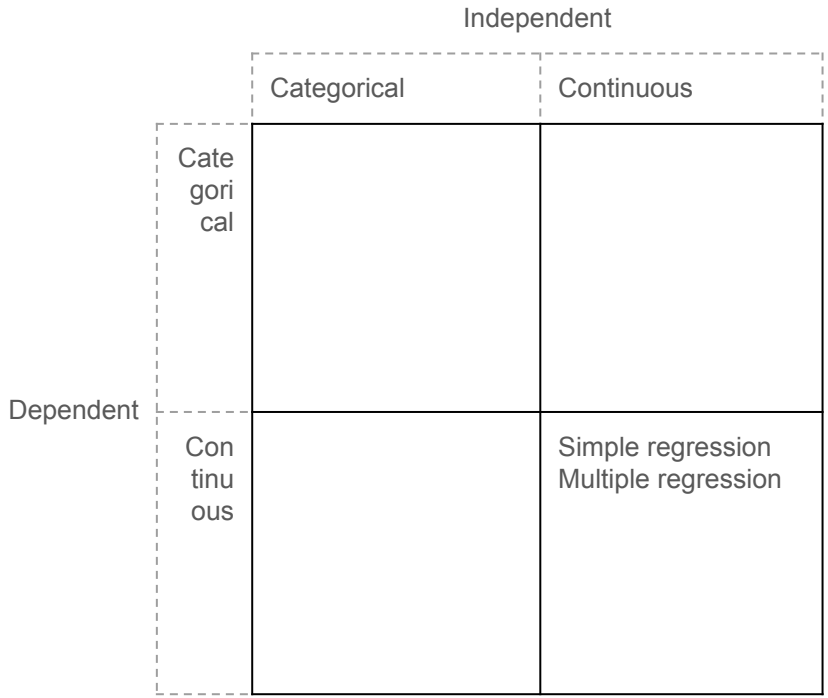
I have never tried that before, so I think I should definitely be able to do that.

Model thinking



$$\text{Petal. Length} = \beta_0 + \beta_1(\text{Sepal. Length}) + \beta_2(\text{Sepal. Width}) + \epsilon$$

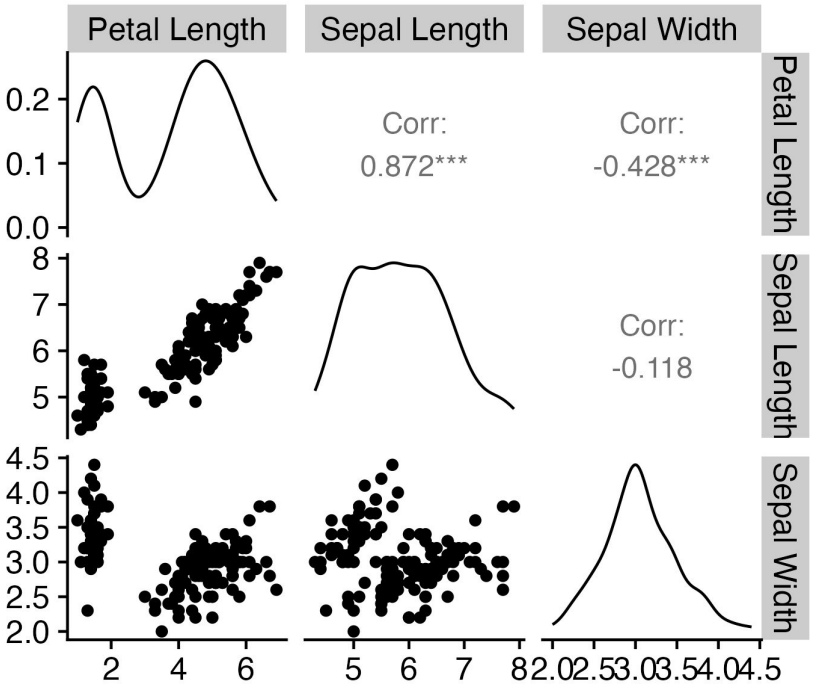
```
mod <- Petal.Length ~ Sepal.Length +
Sepal.Width
```



Data & transformations

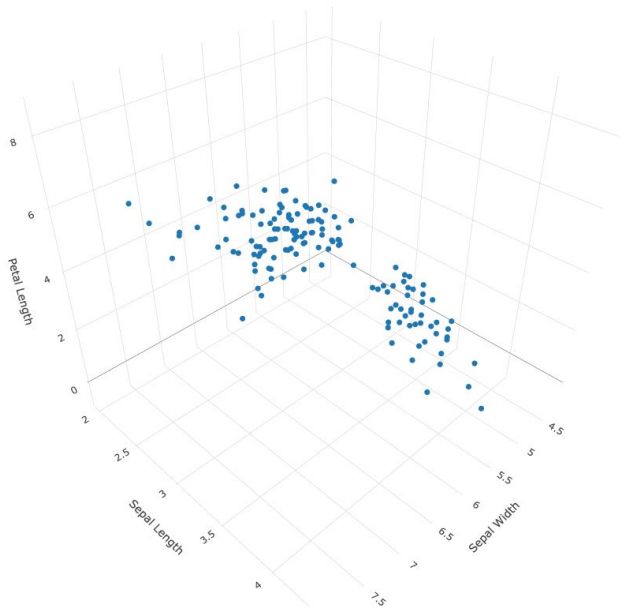
```
data(iris)
help(iris)
head(iris)
str(iris)
pairs(iris)
```

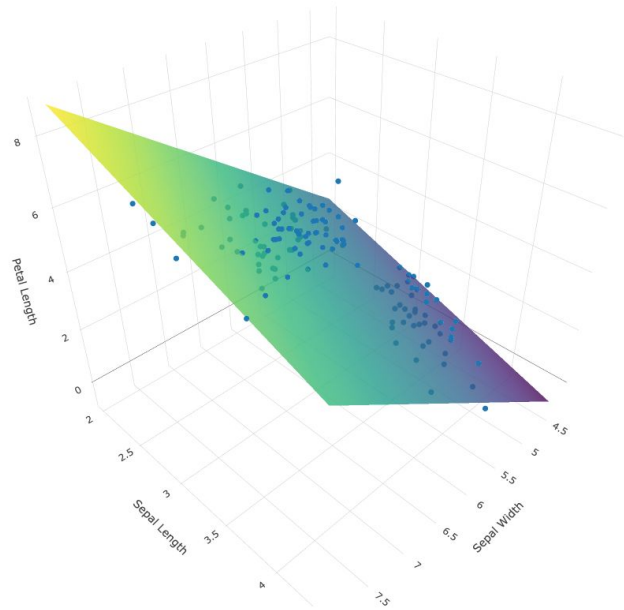
```
'data.frame':  150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```



🎲 Guess the correlation

Data & transformations





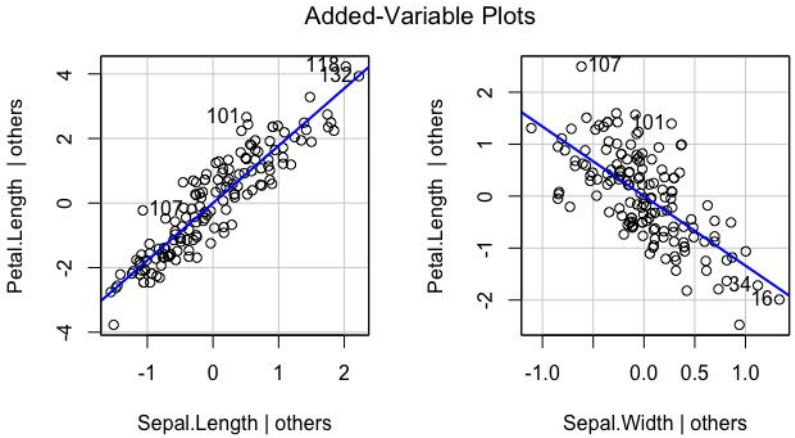
Results

```
fit <- lm(formula = mod, data = iris)
summary(fit); resid(fit); confint(fit)
```

```
Call:
lm(formula = mod, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-1.25582 -0.46922 -0.05741  0.45530  1.75599

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.52476    0.56344  -4.481 1.48e-05 ***
Sepal.Length  1.77559    0.06441  27.569  < 2e-16 ***
Sepal.Width  -1.33862    0.12236 -10.940  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6465 on 147 degrees of freedom
Multiple R-squared:  0.8677,Adjusted R-squared:  0.8659
F-statistic:   482 on 2 and 147 DF,  p-value: < 2.2e-16
```

$$\widehat{\text{Petal. Length}} = -2.52 + 1.78(\text{Sepal. Length}) - 1.34(\text{Sepal. Width})$$

Partial slope coefficient



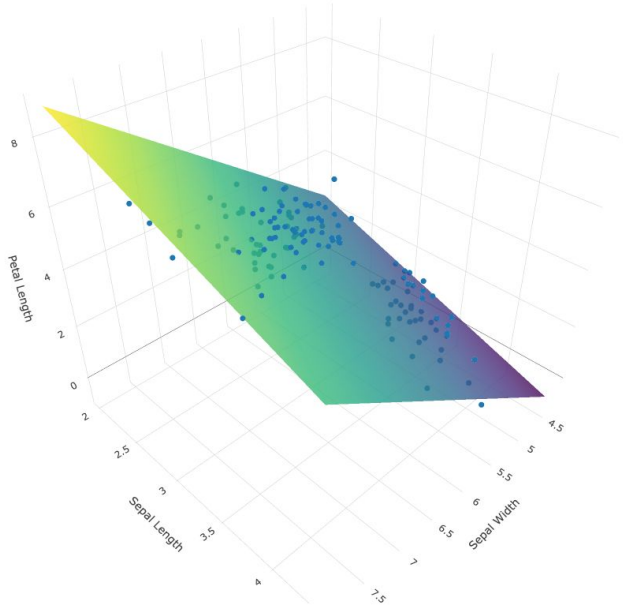
"| others" = *holding the other variables constant*

```
library("car")
car::avPlots(fit)
```

Hey, Willem the Supervillain, the meanings of your slope coefficients have changed!

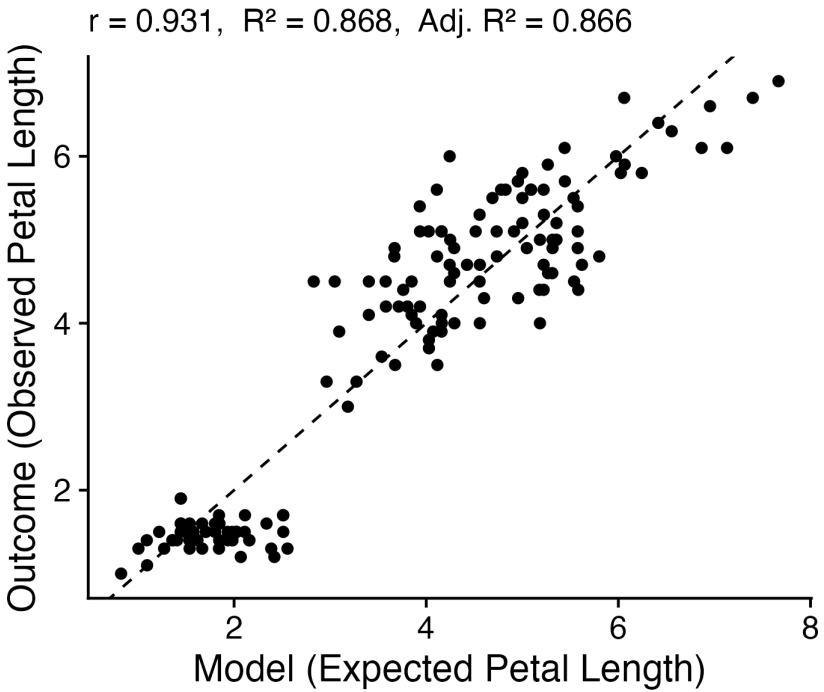
"The predicted mean Petal Length (Y) changes by 1.78 (predicted β_1) for every unit increase in Sepal Length (X_1), *having controlled for Sepal Width (X_2)*."

Ms. Signal 👫 Mr. Noise



Evaluation | (Adjusted) R^2

Hey, Willem the Supervillain, be careful, your explained variance is inflated!



Penalty for increasing the number of independent variables, more so for smaller n.

```
observed <- iris$Petal_Length
expected <- fitted(fit)
cor(observed, expected)^2
```

Evaluation | Model comparison

Statistical significance

Explained variance (multiple R^2)

Predictive validity

Effect size

```
# model comparison
mod_0 <- Petal.Length ~ Sepal.Length
fit_0 <- lm(formula = mod_0, data = iris)
anova(fit, fit_0)
```

Standardization

- For comparison of slope coefficients (if expressed in different or unintuitive measurements)
- What about iris data set?

```
iris_std <- iris |>
  mutate(across(where(is.numeric), scale))
fit_std <- lm(formula = mod, data = iris_std)
summary(fit_std)
```

Unstandardized

The mean of petal length changes by 1.78 centimeters for every centimeter increase in sepal length, having controlled for sepal width.

$$\widehat{\text{Petal. Length}} = -2.52 + 1.78(\text{Sepal. Length}) - 1.34(\text{Sepal. Width})$$

Standardized

The mean of petal length changes by .83 standard deviations for every standard deviation increase in sepal length, having controlled for sepal width.

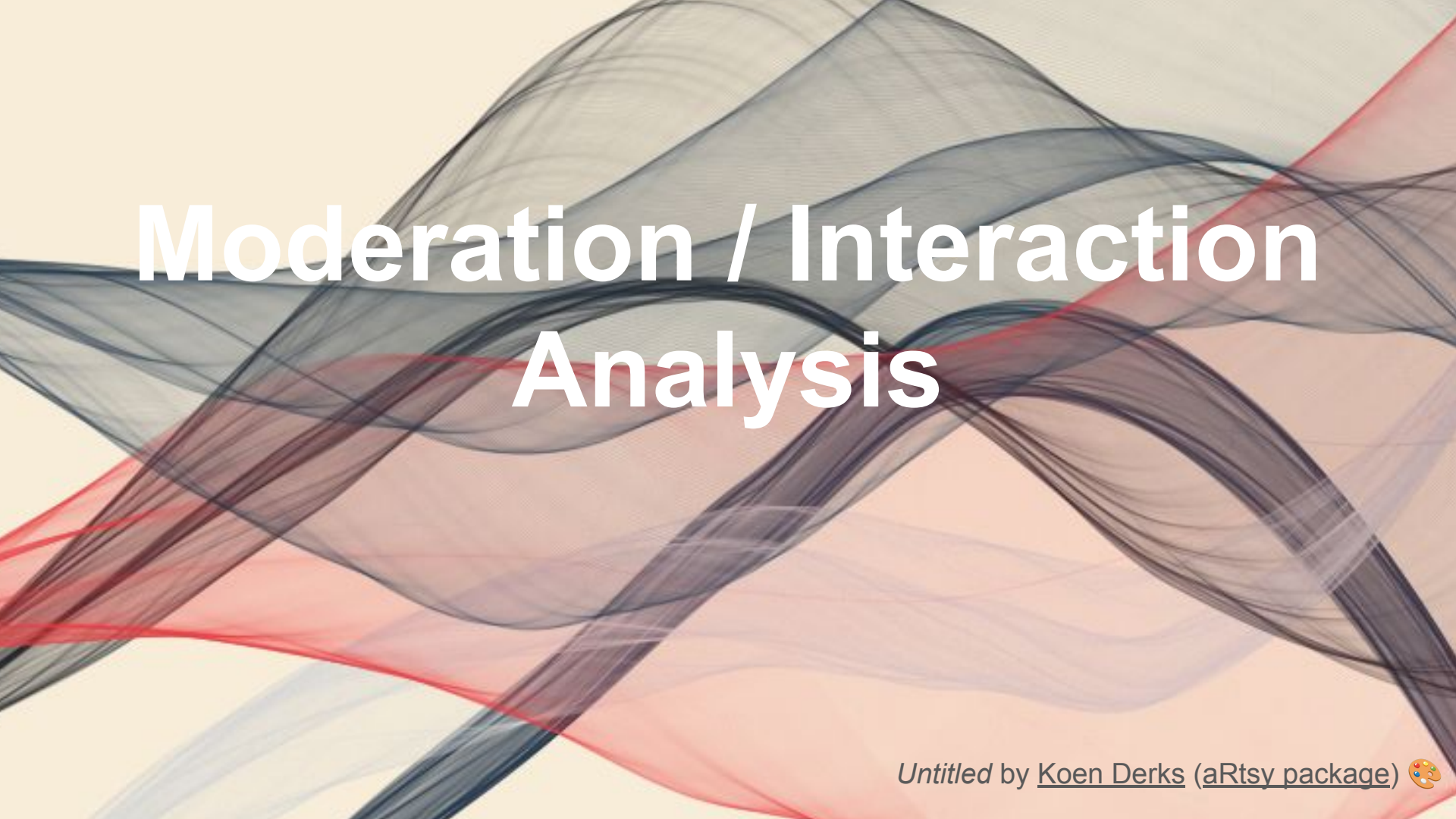
$$\widehat{\text{Petal. Length}} = 0 + 0.83(\text{Sepal. Length}) - 0.33(\text{Sepal. Width})$$

💩 R is the shit

```
library("report")
report(fit, alpha = .001)
```

```
We fitted a linear model (estimated using OLS) to predict
Petal.Length with Sepal.Length and Sepal.Width (formula:
Petal.Length ~ Sepal.Length + Sepal.Width). The model
explains a statistically significant and substantial
proportion of variance (R2 = 0.87, F(2, 147) = 482.00, p <
.001, adj. R2 = 0.87). The model's intercept, corresponding
to Sepal.Length = 0 and Sepal.Width = 0, is at -2.52 (95%
CI [-3.64, -1.41], t(147) = -4.48, p < .001). Within this
model:

  - The effect of Sepal Length is statistically significant
and positive (beta = 1.78, 95% CI [1.65, 1.90], t(147) =
27.57, p < .001; Std. beta = 0.83, 95% CI [0.77, 0.89])
  - The effect of Sepal Width is statistically significant
and negative (beta = -1.34, 95% CI [-1.58, -1.10], t(147) =
-10.94, p < .001; Std. beta = -0.33, 95% CI [-0.39, -0.27])

Standardized parameters were obtained by fitting the model
on a standardized version of the dataset. 95% Confidence
Intervals (CIs) and p-values were computed using a Wald
t-distribution approximation.
```

Moderation / Interaction Analysis



Untitled by Koen Derks (aRtsy package) 🎨

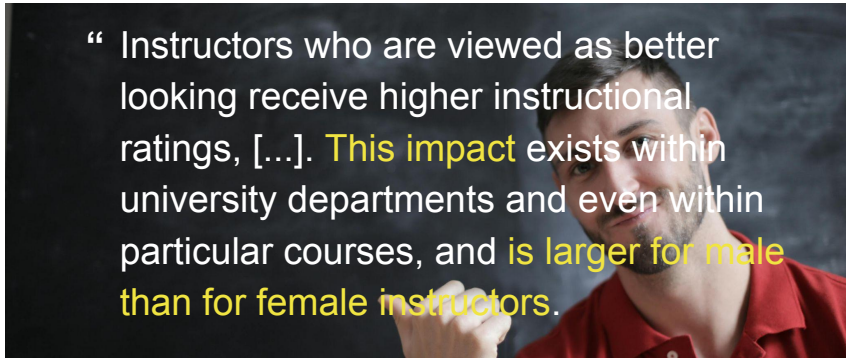
Student evaluations

De student als consument maakt vrouwelijke docenten extra kwetsbaar

Nieuws | door Frans van Heest

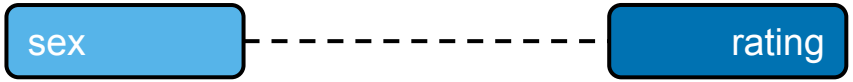
13 september 2023 | Vrouwelijke docenten worden aantoonbaar gediscrimineerd door studentenevaluaties, maar toch blijft het instrument voor veel universiteiten belangrijk om medewerkers te beoordelen. Cursusevaluaties moedigen echter middelmatig onderwijs aan en zijn extra nadelig voor vrouwen.

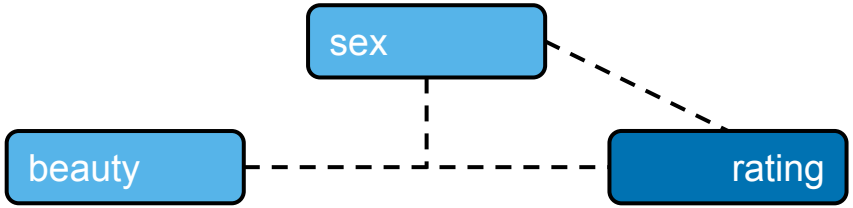
— ScienceGuide 🔒 (Sep. 13, 2023)



— Hamermesh & Parker (2005) 🔒; NBER

— Photo by Andrea Piacquadio





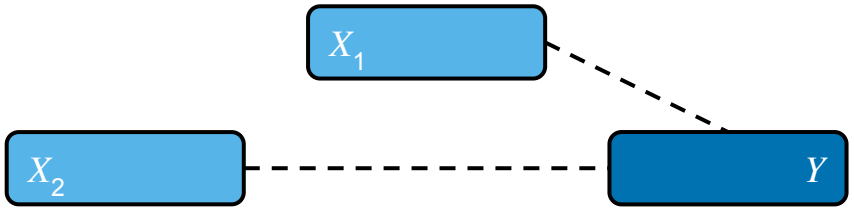
Main effects, interaction effects

Willem, did you know we've been analyzing main effects, using 'linear additive models'?

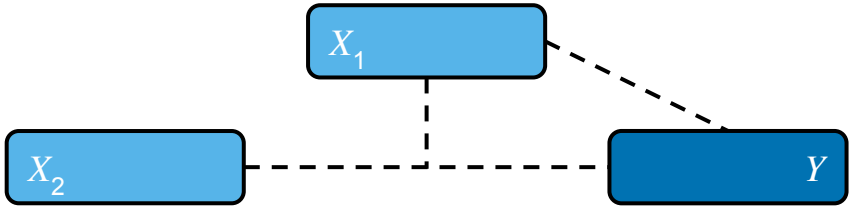
Sure, whatever, let's move on. I'm going to do interactions and start multiplying terms.

If this is what you call an interaction, we're definitely not going to multiply.

Additive



Nonadditive



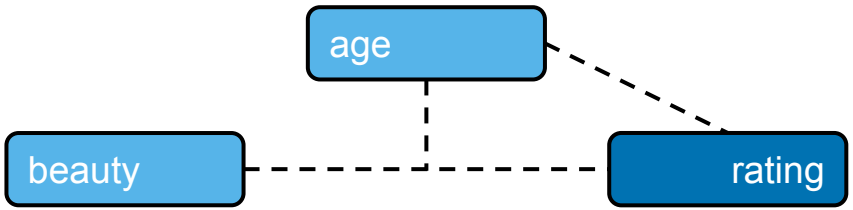
📖 10.1, 10.2

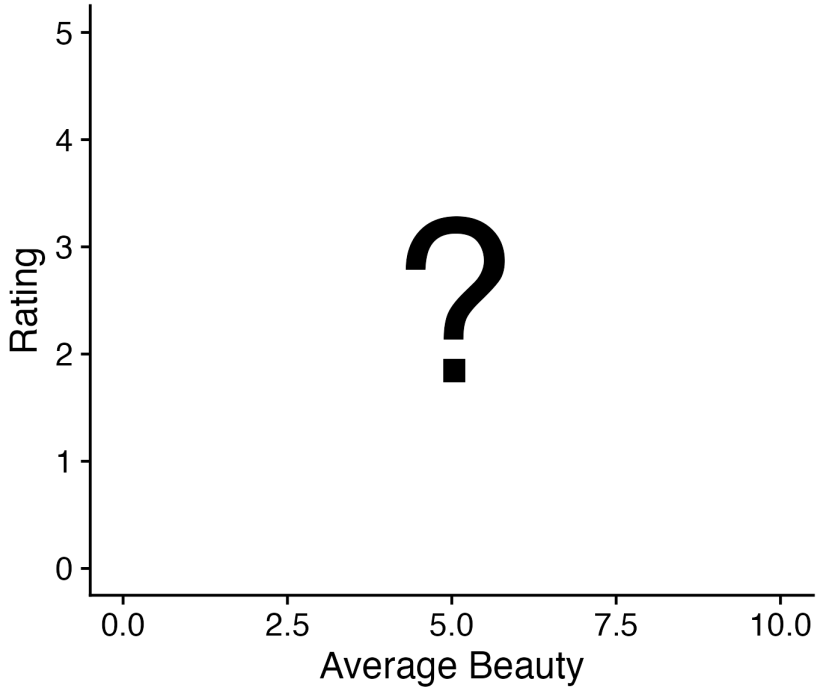
Interactions

Q. Is the effect of beauty on instructional rating modified by age?

H. The effect of beauty on rating is larger for [younger/older] teachers than for [younger/older] teachers.

E. …

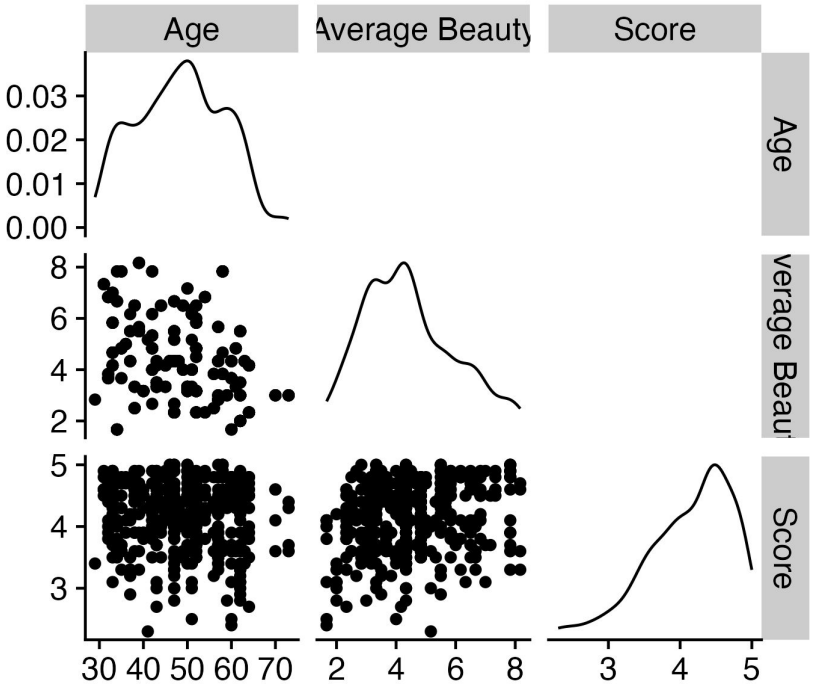




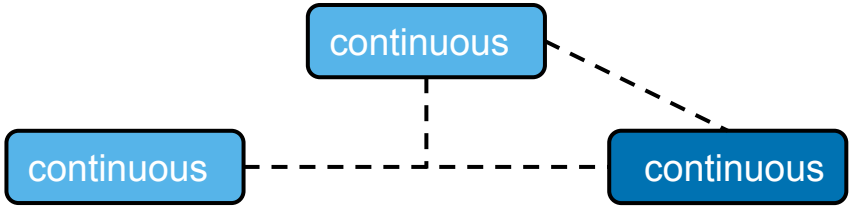
Data

```
library("moderndive")
data(package = "moderndive")
data(evals)
help(evals)
pairs(evals)
str(evals)
```

```
tibble [463 × 14] (S3: tbl_df/tbl/data.frame)
 $ ID          : int [1:463] 1 2 3 4 5 6 7 8 9 10 ...
 $ prof_ID     : int [1:463] 1 1 1 1 2 2 2 3 3 4 ...
 $ score       : num [1:463] 4.7 4.1 3.9 4.8 4.6 4.3 2.8 4.1 3.4 4.5 ...
 $ age         : int [1:463] 36 36 36 36 59 59 59 51 51 40 ...
 $ bty_avg     : num [1:463] 5 5 5 5 3 ...
 $ gender      : Factor w/ 2 levels "female","male": 1 1 1 1 2 2 2 2 2 1 ...
 $ ethnicity   : Factor w/ 2 levels "minority","not minority": 1 1 1 1 2 2 2 2 2 2 ...
 $ language    : Factor w/ 2 levels "english","non-english": 1 1 1 1 1 1 1 1 1 1 ...
 $ rank        : Factor w/ 3 levels "teaching","tenure track",..: 2 2 2 2 3 3 3 3 3 3 ...
 $ pic_outfit  : Factor w/ 2 levels "formal","not formal": 2 2 2 2 2 2 2 2 2 2 ...
 $ pic_color   : Factor w/ 2 levels "black&white",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ cls_did_eval: int [1:463] 24 86 76 77 17 35 39 55 111 40 ...
 $ cls_students: int [1:463] 43 125 125 123 20 40 44 55 195 46 ...
 $ cls_level   : Factor w/ 2 levels "lower","upper": 2 2 2 2 2 2 2 2 2 2 ...
```



Model thinking



Let's just multiply those terms.

$$\text{score} = \beta_0 + \beta_1(\text{age}) + \beta_2(\text{bty\_avg}) + \beta_3(\text{age} \times \text{bty\_avg}) + \epsilon$$

```
mod <- score ~ age + bty_avg + age:bty_avg
mod <- score ~ age * bty_avg
```

Dependent \ Independent	Categorical	Continuous
Categorical		
Continuous		Simple regression Multiple regression *(additive / nonadditive)*

Results

```
dat <- evals
fit <- lm(formula = mod, data = dat)
summary(fit)
```

```
Call:
lm(formula = mod, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9410 -0.3517  0.1231  0.4040  1.0066

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.156077   0.367797  14.019  < 2e-16 ***
age         -0.026128   0.007342  -3.559 0.000412 ***
bty_avg     -0.187800   0.075724  -2.480 0.013494 *
age:bty_avg  0.005318   0.001580   3.366 0.000827 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5287 on 459 degrees of freedom
Multiple R-squared:  0.06096,    Adjusted R-squared:  0.05482
F-statistic: 9.933 on 3 and 459 DF,  p-value: 2.349e-06
```

$$\widehat{\text{score}} = 5.16 - 0.03(\text{age}) - 0.19(\text{bty\_avg}) + 0.01(\text{age} \times \text{bty\_avg})$$

*So the **simple main effect** of average beauty on score is −0.19, **when age is 0**. Well done, Willempie, that’s illuminating. You could’ve centered your predictors.*

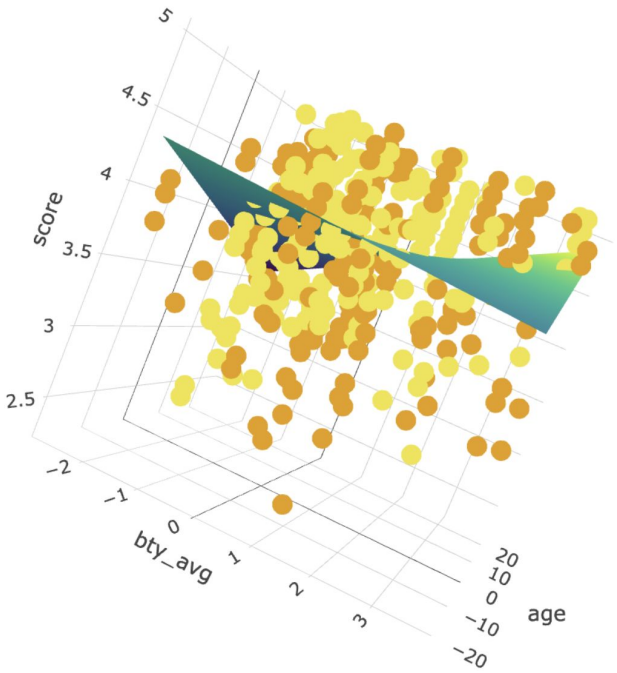
Talk to the hand, I’m only interested in the interaction effect, anyway.

```
dat$age <- dat$age - mean(dat$age) # 48.4
```

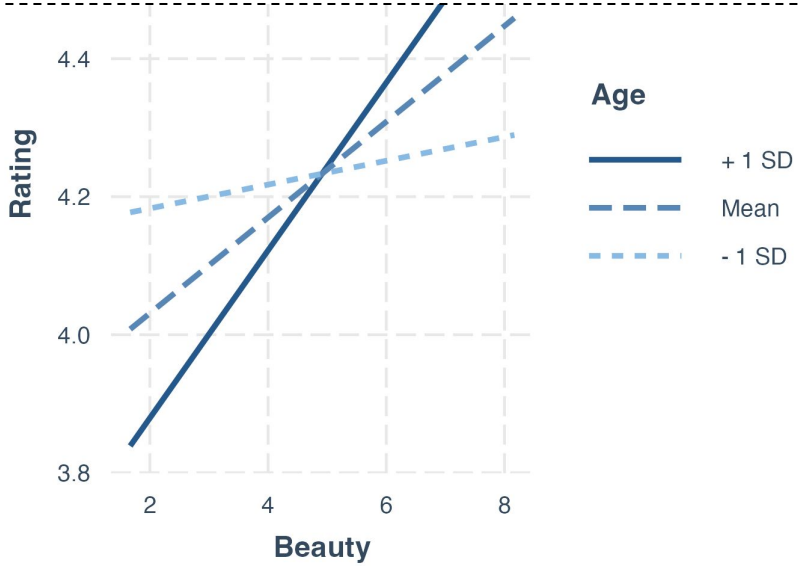
💡 To mean center? Practical advice on last paragraph of the Discussion section.

📖 10.4

Visualize interaction



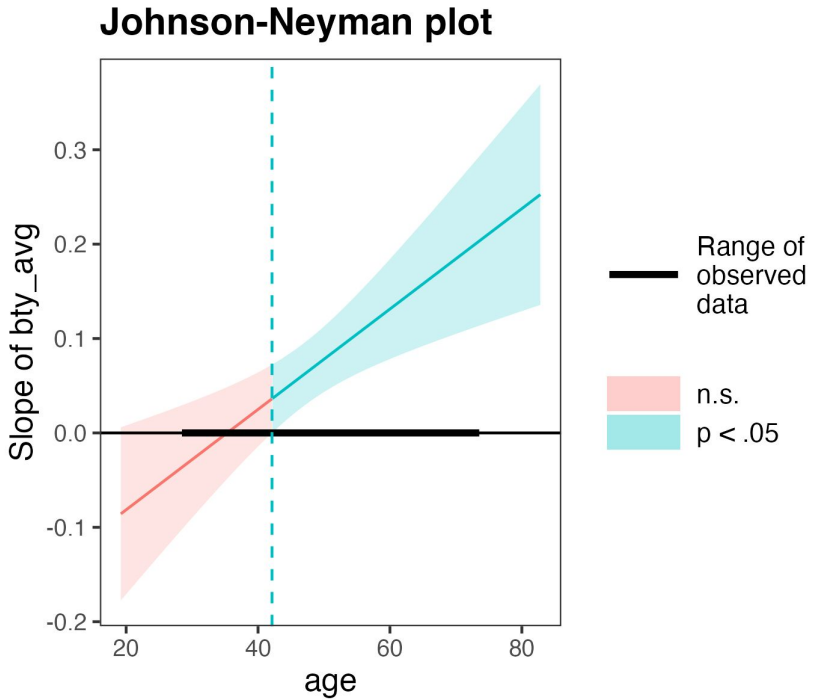

```
library("interactions")
interactions::interact_plot(model = fit, pred =
bty_avg, modx = age, data = dat)
```



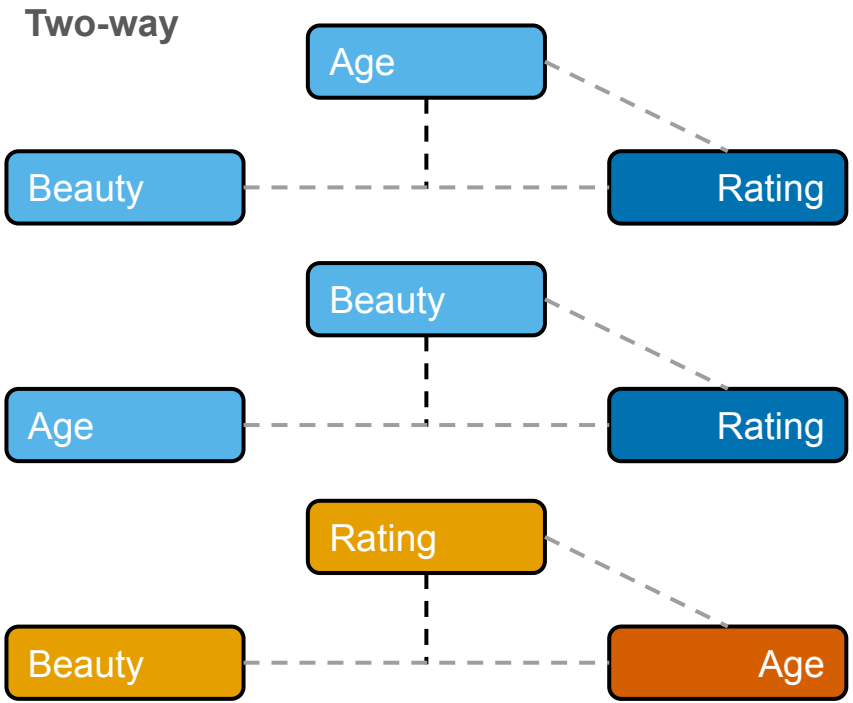
Simple slopes analysis & Johnson–Neyman interval

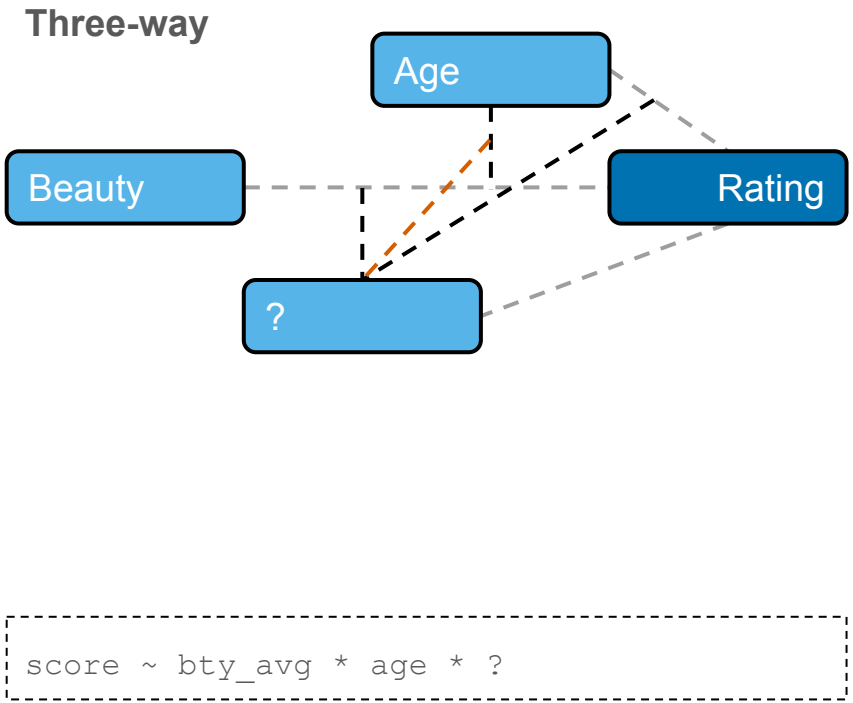
```
library("sandwich")
interactions::sim_slopes(fit, pred = bty_avg, modx
= age)
interactions::johnson_neyman(fit, pred = bty_avg,
modx = age, alpha = .05)
```

```
JOHNSON-NEYMAN INTERVAL

When age is OUTSIDE the interval [16.72, 42.12], the slope of bty_avg is p < .05.

Note: The range of observed values of age is [29.00, 73.00]

SIMPLE SLOPES ANALYSIS

Slope of bty_avg when age = 38.56227 (- 1 SD):

  Est.   S.E.   t val.      p
------ ------ -------- ------
  0.02   0.02     0.81   0.42

Slope of bty_avg when age = 48.36501 (Mean):

  Est.   S.E.   t val.      p
------ ------ -------- ------
  0.07   0.02     4.06   0.00

Slope of bty_avg when age = 58.16775 (+ 1 SD):

  Est.   S.E.   t val.      p
------ ------ -------- ------
  0.12   0.02     4.91   0.00
```

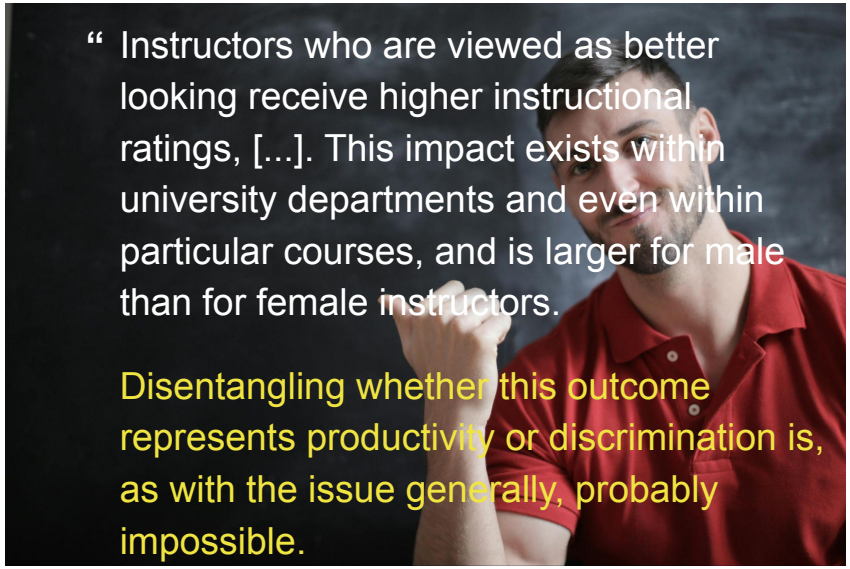


Higher-order interactions





Causality



— Hamermesh & Parker (2005) 🔒; NBER
— Photo by Andrea Piacquadio

🧠 Students don’t know what’s best for their own learning (The Conversation)

Model formulae in R

```
y ~ x   # with intercept
y ~ 1 + x   # with intercept
y ~ 0 + x   # without intercept

y ~ x + z   # add a term
y ~ x - z   # remove a term
y ~ I(x + z)   # sum two terms
y ~ x:z   # create an interaction term
y ~ x * z   # create crossed terms (x + z + x:z)
y ~ x %in% z)   # create nested terms (x + x:z)
```

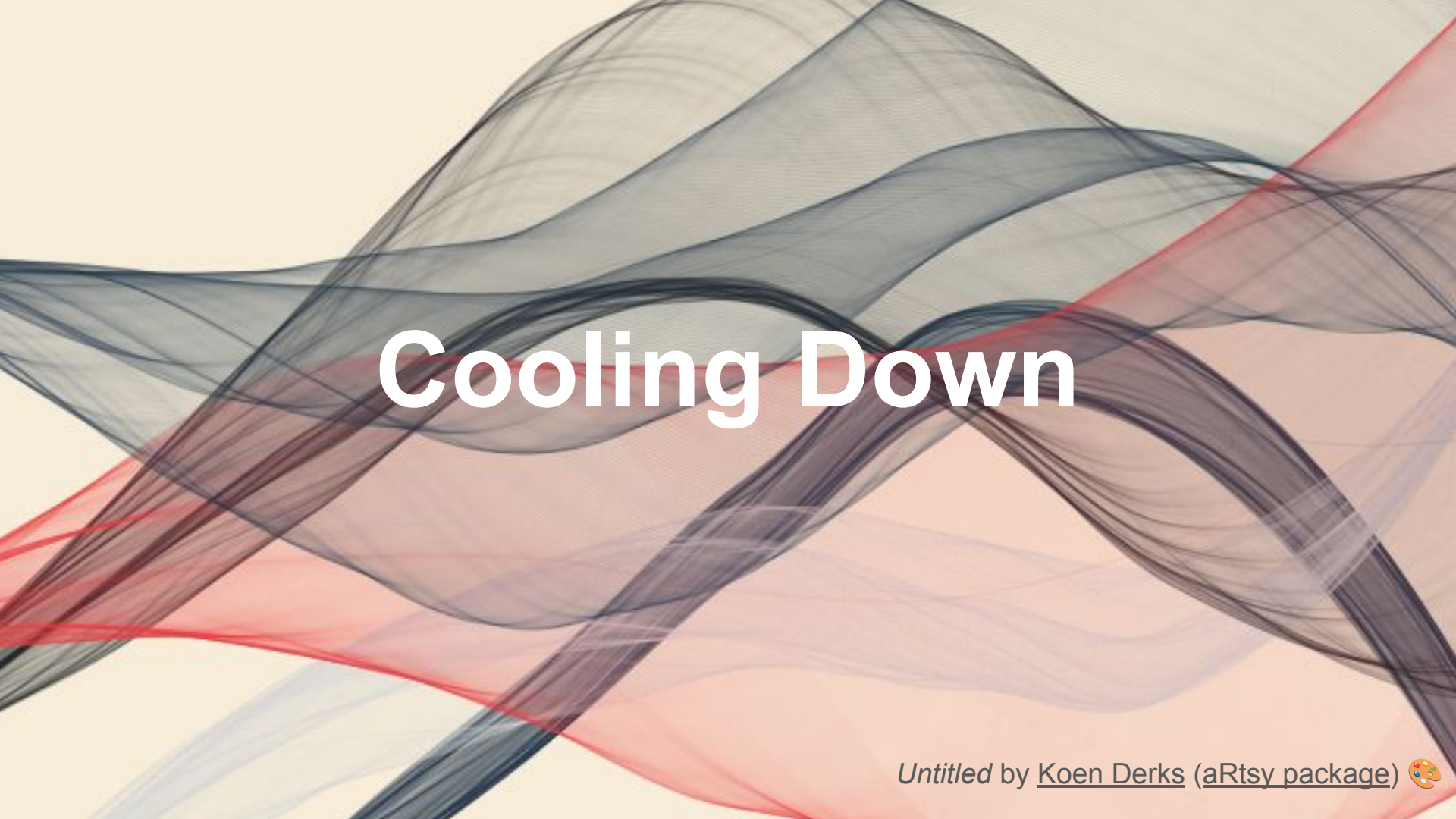
and there's more…

Traditional name	Model formula	R code
Bivariate regression	Y ~ X1 (continuous)	`lm(Y ~ X)`
One-way ANOVA	Y ~ X1 (categorical)	`lm(Y ~ X)`
Two-way ANOVA	Y ~ X1 (cat) + X2(cat)	`lm(Y ~ X1 + X2)`
ANCOVA	Y ~ X1 (cat) + X2(cont)	`lm(Y ~ X1 + X2)`
Multiple regression	Y ~ X1 (cont) + X2(cont)	`lm(Y ~ X1 + X2)`
Factorial ANOVA	Y ~ X1 (cat) * X2(cat)	`lm(Y ~ X1 * X2)` or `lm(Y ~ X1 + X2 + X1:X2)`

Table from An Introduction to R

💡 Nearly anything can be described with a (generalized linear) regression model. A cheat sheet for model formulae. Understand the *t*-test and ANOVA as a linear model (cheat sheet).

Cooling Down



Untitled by Koen Derks (aRtsy package) 🎨

Takeaways

There is no such thing as too much power*

* always determine the amount of participants you need to detect a particular effect size *before* the experiment

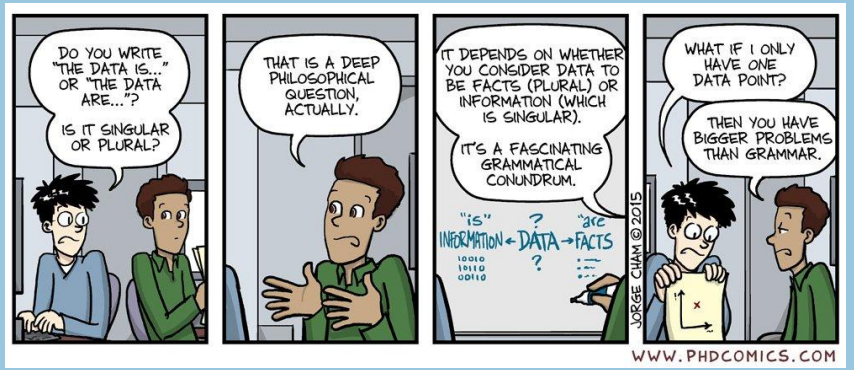


Illustration by Jorge Cham

Takeaways

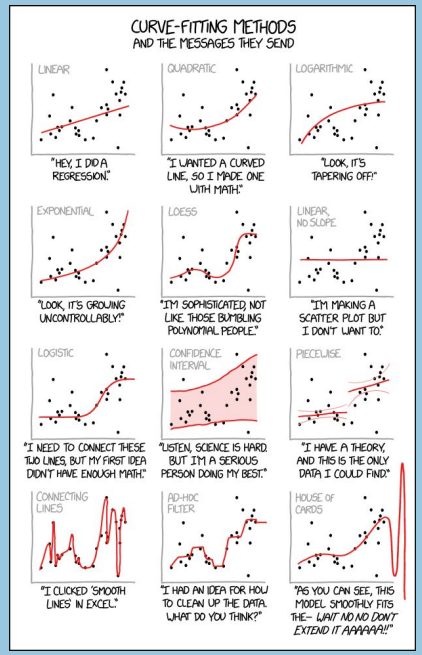


Illustration by Randall Munroe (wtf)

Takeaways

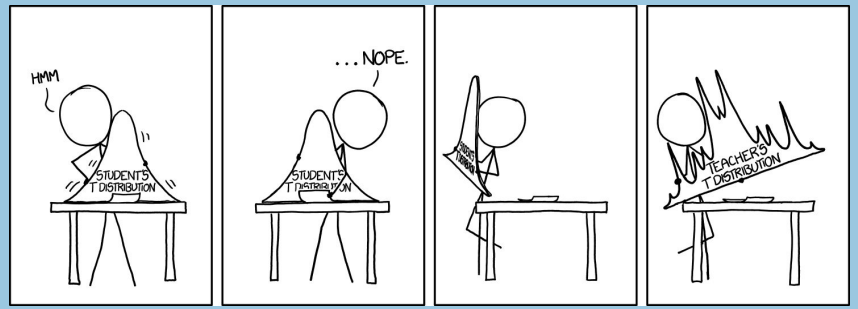


Illustration by Randall Munroe (wtf)

Takeaways

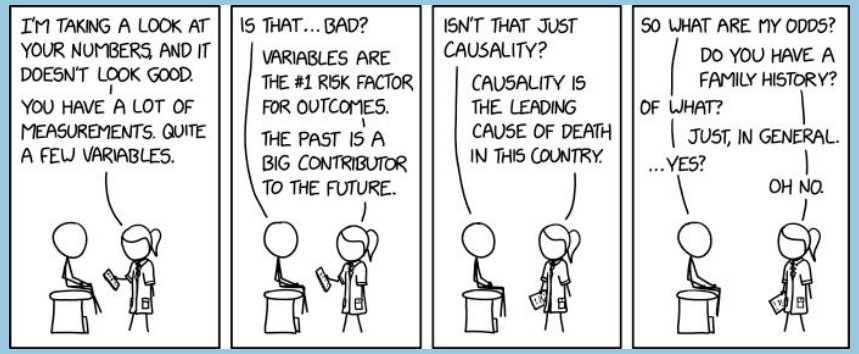


Illustration by Randall Munroe (wtf)

🥇 Exam(ple) question

Je wilt een auto gebruiken, maar maakt je ook zorgen om het milieu. Je gebruikt de `mtcars` dataset in R om er achter te komen wat de eigenschappen van een zuinige auto zijn. Je onderzoekt hoe de relatie tussen het verbruik (`mpg`) en de paardekracht (`hp`) wordt gemodereerd door het gewicht (`wt`).

A. Rond af op twee decimalen en rapporteer de beta-coefficient van de significante interactie.
B. Rond af op twee decimalen en rapporteer tot welk gewicht (`wt`) de paardekracht (`hp`) een negatieve relatie heeft met het verbruik (`mpg`).

💡 This R data set is frequently used in tutorials, help files, and question-and-answer websites like Stack Overflow and the Posit Forum.

🔮 Next week

Topics

Statistical reasoning with GLM
Multiple linear regression
| Dummy-variable regression
Logistic regression
Multilevel and longitudinal analysis
Bayesian statistics

Illustration by Jennifer Cheuk

✍️ Colophon

Slides
alexandersavi.nl/teaching/

License