Multiple Linear Regression

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

👨‍🏫 Statistical Reasoning Lecture #3
Alexander Savi, 2024
📚 Whitlock/Schluter, Ch. 17; Agresti/Franklin, Ch. 13

by Koen Derks (aRtsy package) 🎨

📰 News

NOS Nieuws • Vandaag, 05:11

Nederlander wint Ig Nobelprijs met 350.757 keer kop of munt gooien

💡 The 34th First Annual Ig Nobel Ceremony (2024)

— NOS Nieuws (Sep. 13, 2024)

📢 Announcements

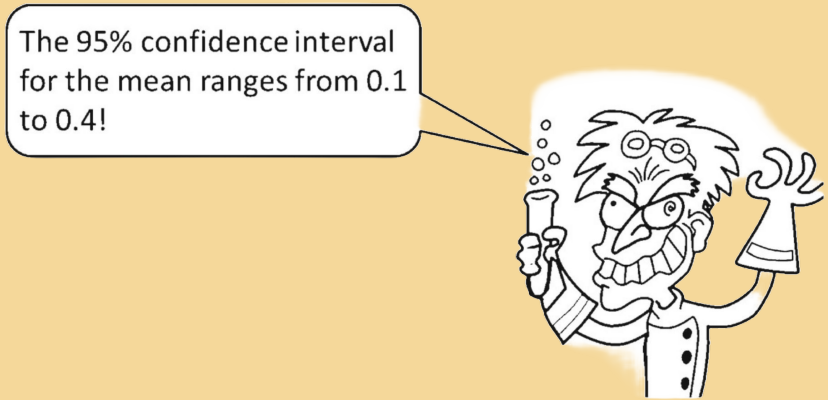
- Personal course manual (for learning, can’t bring to exam)
- Web lectures & attendance

Warming Up

by Koen Derks (aRtsy package) 🎨

🔁 Recap

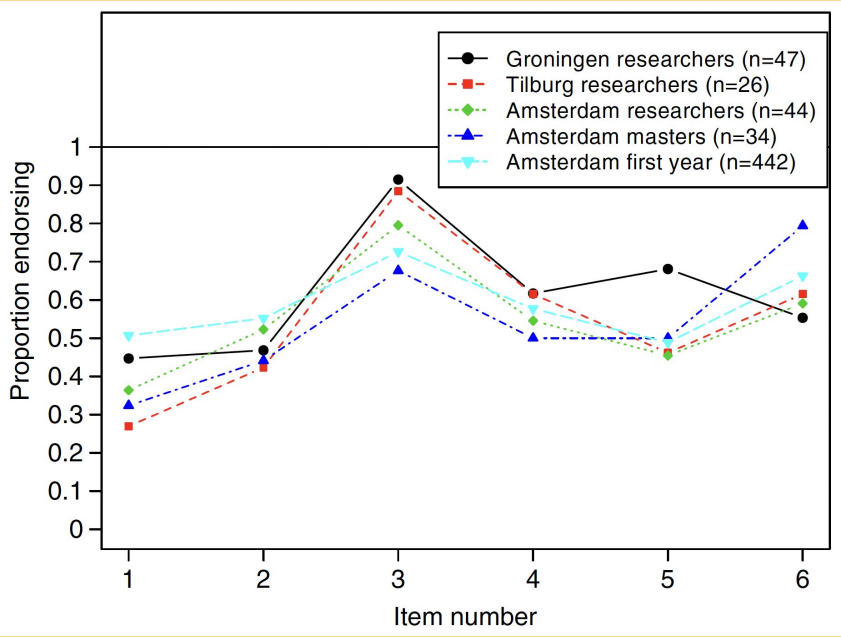
Professor Bumbledorf conducts an experiment, analyzes the data, and reports:



True or false?

1. The probability that the true mean is greater than 0 is at least 95%.
2. The probability that the true mean equals 0 is smaller than 5%.
3. The “null hypothesis” that the true mean equals 0 is likely to be incorrect.
4. There is a 95% probability that the true mean lies between 0.1 and 0.4.
5. We can be 95% confident that the true mean lies between 0.1 and 0.4.
6. If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4.

🔁 Recap



💡 Hoekstra et al., 2014

True or false?

1. The probability that the true mean is greater than 0 is at least 95%.
2. The probability that the true mean equals 0 is smaller than 5%.
3. The “null hypothesis” that the true mean equals 0 is likely to be incorrect.
4. There is a 95% probability that the true mean lies between 0.1 and 0.4.
5. We can be 95% confident that the true mean lies between 0.1 and 0.4.
6. If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4.

🔮 Overview

Topics

Probabilities & distributions
Frequentist inference
| Multiple linear regression
| Moderation
| 🔁 ~~*F* statistic and distribution~~
Factorial ANOVA
Nonparametric inference
Bayesian inference

Learning goals

Estimate the relationships between more than two variables.

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

~~Test complex models with the *F* distribution.~~

Multiple Linear Regression

Estimating Relationships Between Variables

by Koen Derks (aRtsy package) 🎨

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.

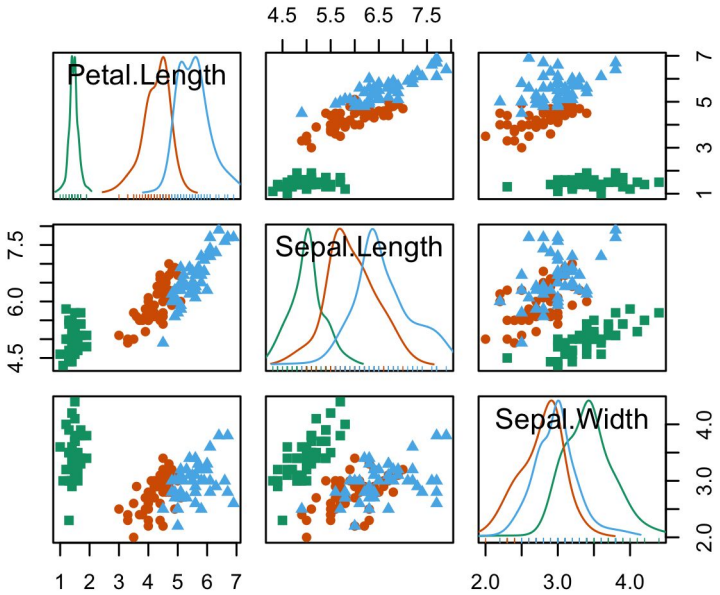
E. […]

Illustration by [Diego Mariano]

Data

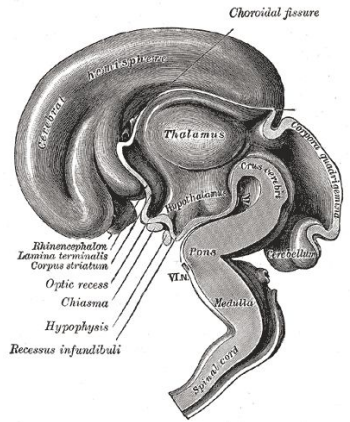
```
data(iris)
?"iris"
head(iris)
str(iris)
plot(iris)
```

```
> str(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 ...
```



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

Statistical model



Outcome = Model + Error

- Perseverance = Student Population + Error
- Petal Length = Sepal Length + Sepal Width + Error

[Model formulae](#) in R:

y ~ model

- y : dependent variable
- ~ : “is modeled by”
- model : independent variable(s)

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

Linear model

Outcome = Model + Error

Y_i = *Model* + e_i

$\hat{Y}_i$ = *Model*

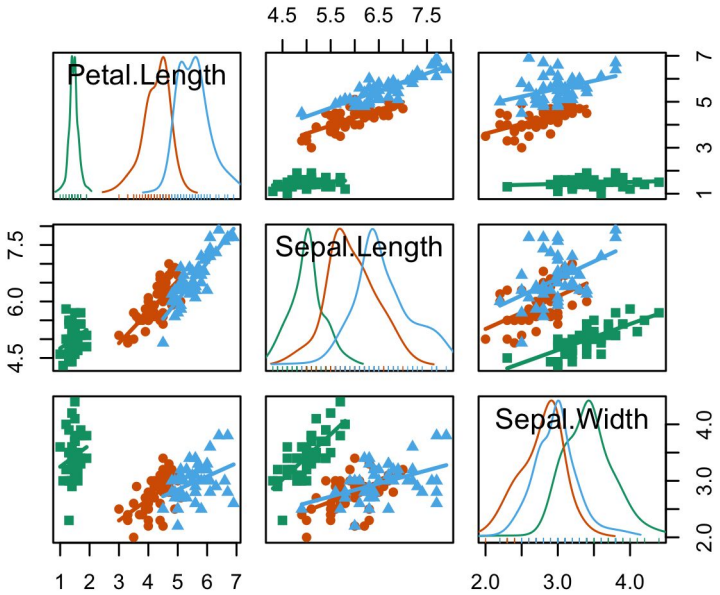
Linear equation: $y = a\,x + b$

$\beta_o + \beta_1 X_i$ (simple lin. reg.)

$\beta_o + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}$ (multiple lin. reg.)

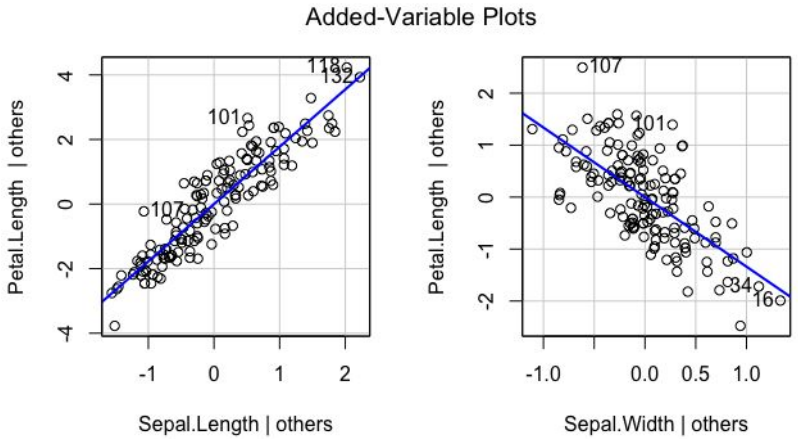
Petal Length$_i$ = β_o + β_1 *Sepal Length*$_i$ + β_2 *Sepal Width*$_i$ + e_i

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



Results

```
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
```

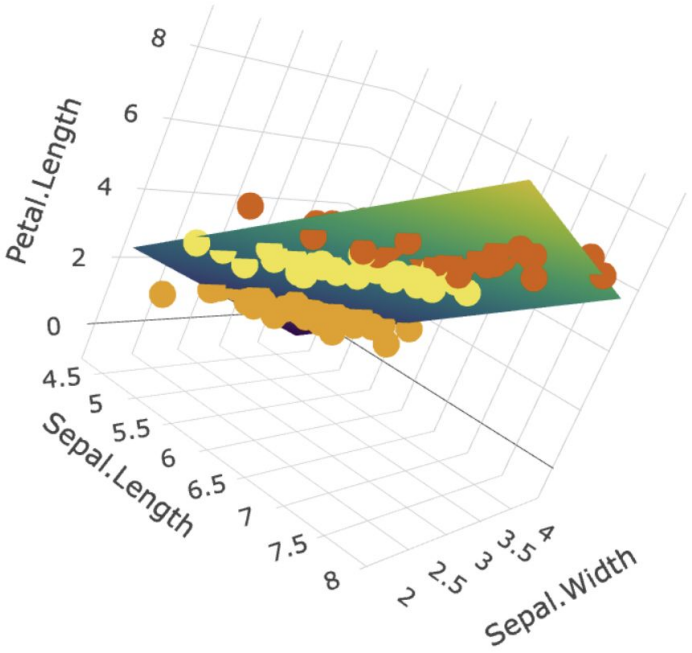


“| others” = *holding the other variables constant*

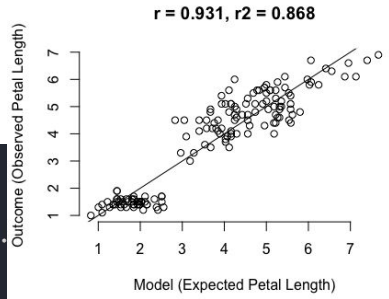
$$\textit{Petal Length}_i = -2.52 + 1.78 \times \textit{Sepal Length}_i + -1.34 \times \textit{Sepal Width}_i + e_i$$

💡Compute t-statistic for β_1 (same procedure as for the mean): $t = (1.776 - 0) / 0.064 = 27.569$

Results


Petal.Length
8
6
4
2
0
Sepal.Length
4.5
5
5.5
6
6.5
7
7.5
8
Sepal.Width
2
2.5
3
3.5
4

Model evaluation




```
Residuals:
     Min       1Q   Median
-1.25582 -0.46922 -0.05741  0.

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
```

```
# statistical significance of predictors
summary(fit)

# multiple R2 (explained variance)
observed <- iris$Petal_Length
expected <- fitted(fit)
cor(observed, expected)^2

# F statistic

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

# predictive validity
predict(fit, new_data)
```

Statistical model II

```
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

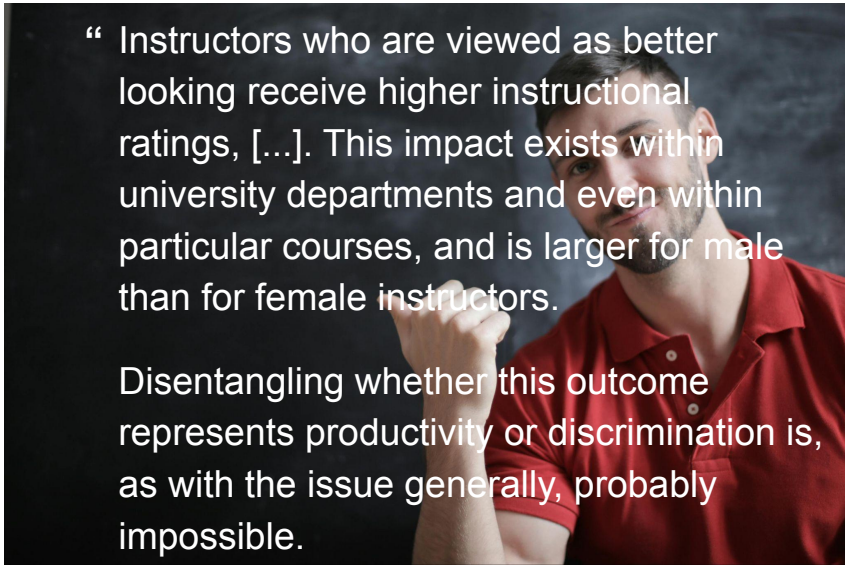
💡 Nearly anything can be described with a (generalized linear) regression model. A cheat sheet for model formulae. Understand the *t*-test and ANOVA as a linear model (cheat sheet).

Multiple Linear Regression

Moderation / Interaction

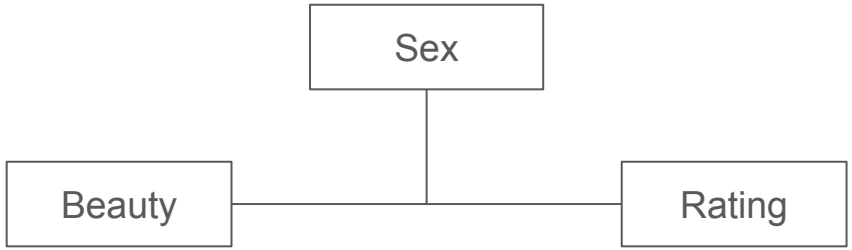
by Koen Derks (aRtsy package) 🎨

Moderation / interaction



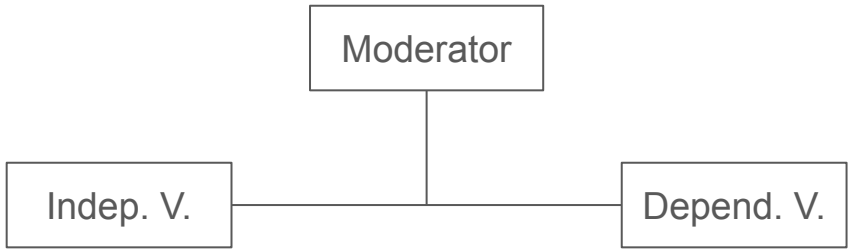
— Hamermesh & Parker, 2005 🔒 ; NBER

— Photo by Andrea Piacquadio



Moderation / interaction

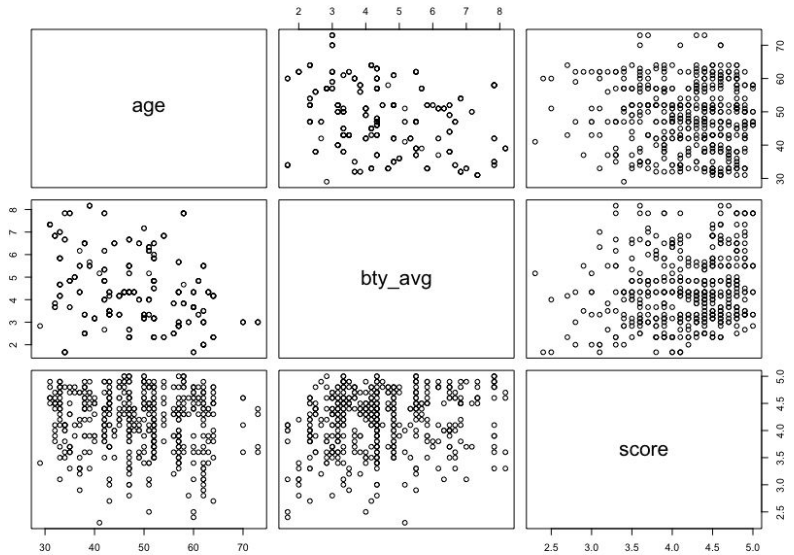
> In statistics and regression analysis, moderation (also known as effect modification) occurs when the relationship between two variables depends on a third variable. The third variable is referred to as the moderator variable [...]. — Wikipedia



Data

```
library("moderndive")
help("evals")
```

```
> str(evals)
tibble [463 × 16] (S3: tbl_df/tbl/data.frame)
 $ ID          : int [1:463] 117 227 409 116 120 250 111 124 125 92 ...
 $ prof_ID     : int [1:463] 20 42 83 20 20 48 20 21 21 17 ...
 $ score       : num [1:463] 3.3 3.3 3.3 3.4 3.4 3.4 3.5 3.5 3.5 3.6 ...
 $ age         : int [1:463] 57 39 47 57 57 50 57 52 52 56 ...
 $ bty_avg     : num [1:463] 4.33 8.17 6.67 4.33 4.33 ...
 $ gender      : Factor w/ 2 levels "female","male": 1 1 1 1 1 1 1 1 1 1 ...
 $ ethnicity   : Factor w/ 2 levels "minority","not minority": 2 2 2 2 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",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ 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 1 2 2 2 2 2 2 2 ...
 $ cls_did_eval: int [1:463] 8 22 16 14 12 18 17 31 17 34 ...
 $ cls_students: int [1:463] 19 24 21 20 15 28 28 36 19 49 ...
 $ cls_level   : Factor w/ 2 levels "lower","upper": 2 1 1 2 2 2 2 2 2 2 ...
 $ mean_gender : num [1:463] 4.09 4.09 4.09 4.09 4.09 ...
 $ mean_rank   : num [1:463] 4.28 4.28 4.28 4.28 4.28 ...
```



Moderation

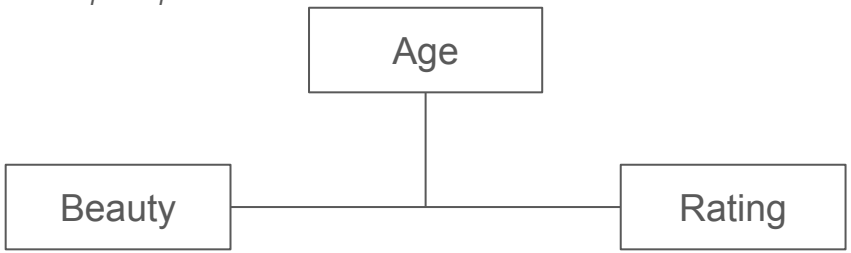
```
mod <- score ~ bty_avg * age
```

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

H. What's your hypothesis? 🤔

E. (*Your hypothesis in terms of your operationalization.*)

$Rating_i = \beta_o + \beta_1 Beauty_i + \beta_2 Age_i + \beta_3 Beauty_i \times Age_i + e_i$



Linear regression w/ interaction term

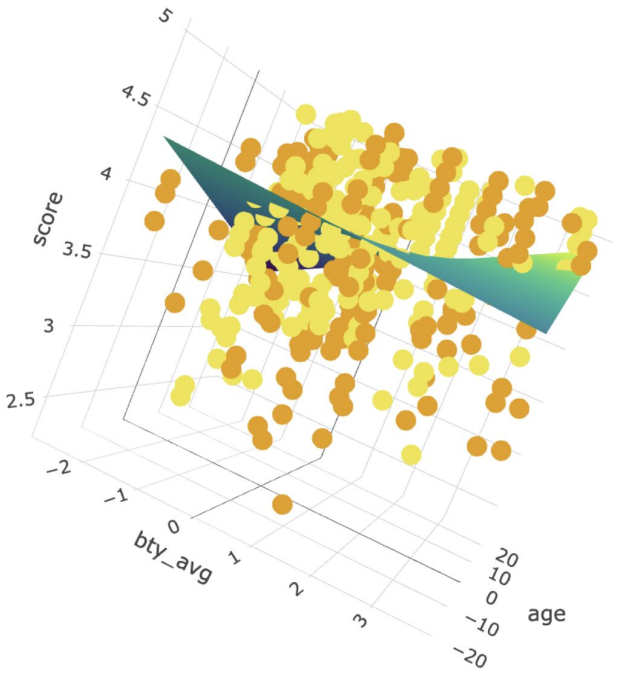
```
# mean centering
dat <- evals
dat$bty_avg <- dat$bty_avg -
mean(dat$bty_avg) # 4.4
dat$age <- dat$age - mean(dat$age) # 48.4

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

```
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)  4.198930   0.025603 164.001  < 2e-16 ***
age         -0.002636   0.002638  -0.999 0.318201
bty_avg      0.069389   0.017107   4.056 5.86e-05 ***
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
```

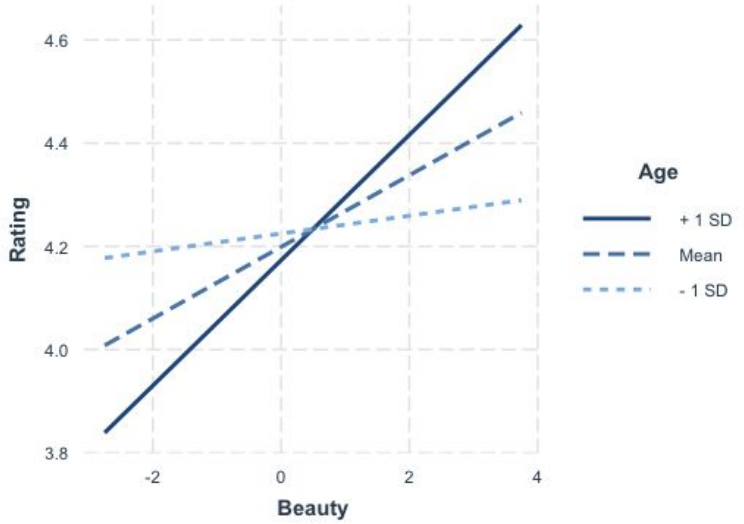
💡 To mean center or not to mean center? See last paragraph of the Discussion section for practical advice.

$$Rating_i = 4.20 + 0.07\ Beauty_i - 0.00\ Age_i + 0.01\ Beauty_i \times Age_i + e_i$$

Visualize interaction




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

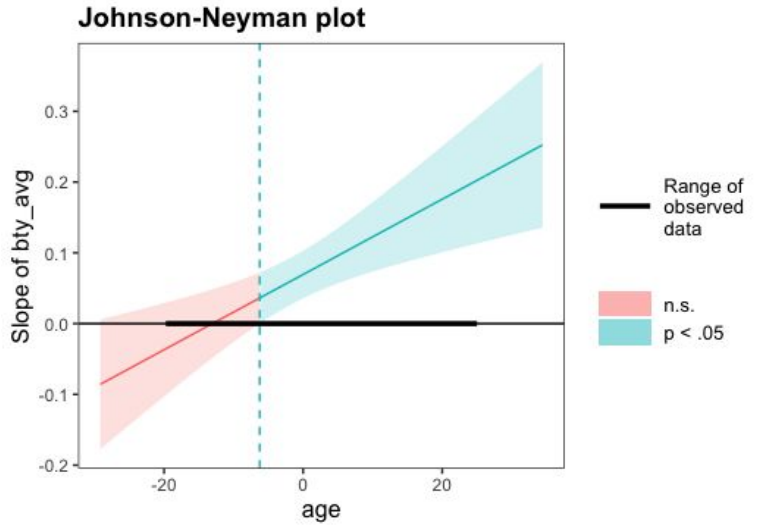


Simple slopes analysis & Johnson–Neyman interval

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

```
SIMPLE SLOPES ANALYSIS

Slope of bty_avg when age = -9.802742e+00 (- 1 SD):

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

Slope of bty_avg when age =  1.930589e-14 (Mean):

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

Slope of bty_avg when age =  9.802742e+00 (+ 1 SD):

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

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



Model evaluation

See previous lecture

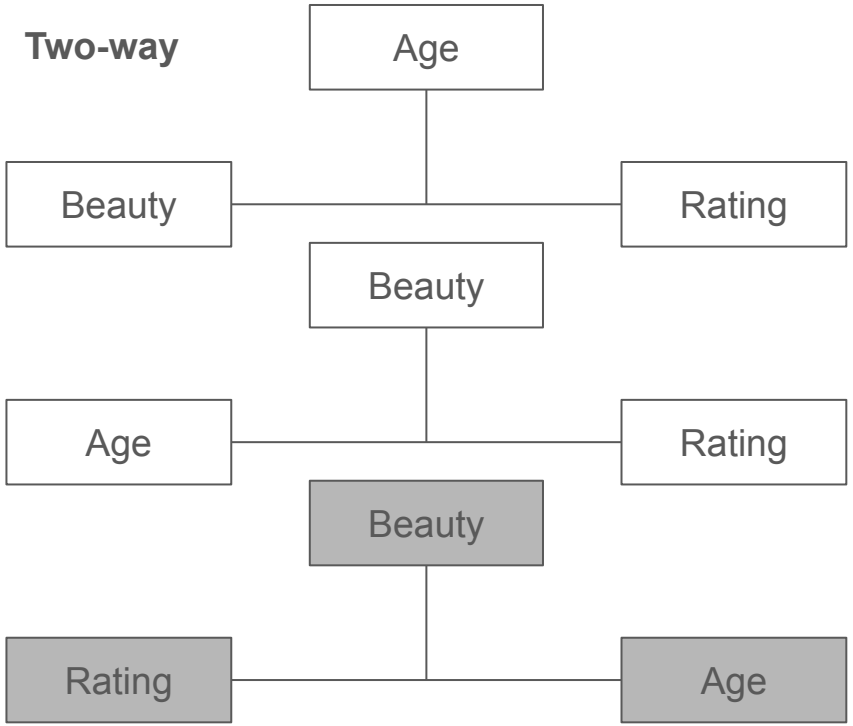
p-values; R^2

```
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)  4.198930   0.025603 164.001  < 2e-16 ***
age         -0.002636   0.002638  -0.999 0.318201
bty_avg      0.069389   0.017107   4.056 5.86e-05 ***
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
```

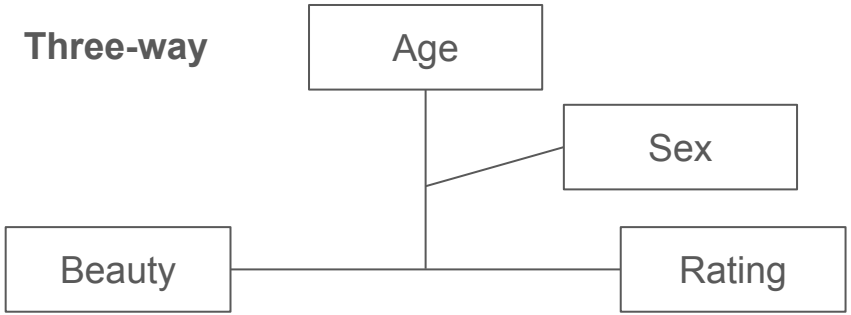
🧠 Students don’t know what’s best for their own learning (The Conversation)

Higher-order interactions

Two-way



Three-way



$$Rating_i = \beta_o + \beta_1\, Beauty_i + \beta_2\, Age_i + \beta_3\, Sex + \beta_4\, Beauty_i \times Age_i + \beta_5\, Beauty_i \times Sex_i + \beta_6\, Age_i \times Sex_i + \beta_7\, Beauty_i \times Age_i \times Sex_i + e_i$$

```
score ~ bty_avg * age * gender
```

Cooling Down

by Koen Derks (aRtsy package) 🎨

Takeaways

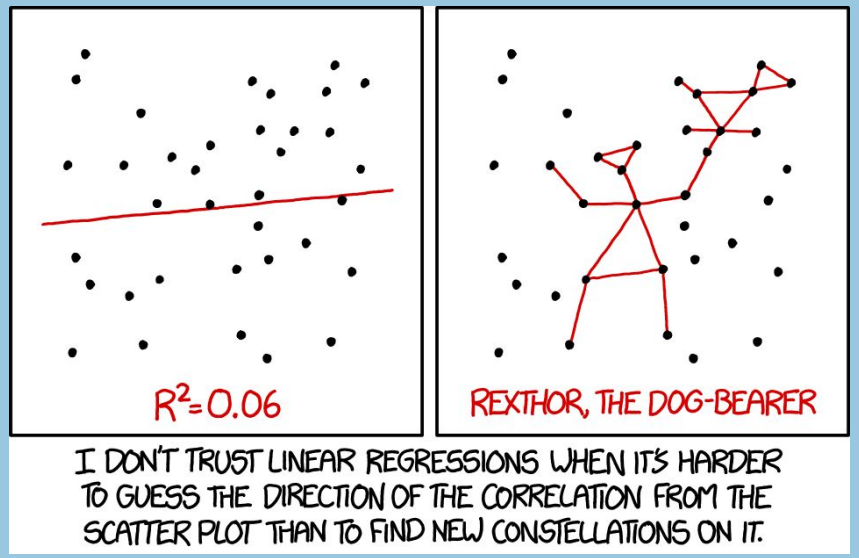


Illustration by Randall Munroe (wtf)

Takeaways

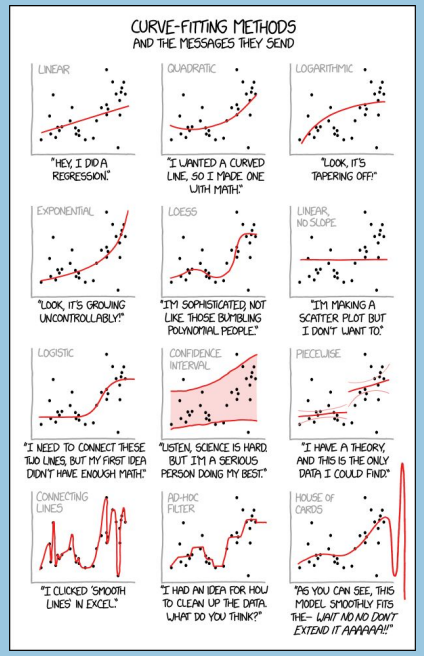


Illustration by Randall Munroe (wtf)

Takeaways

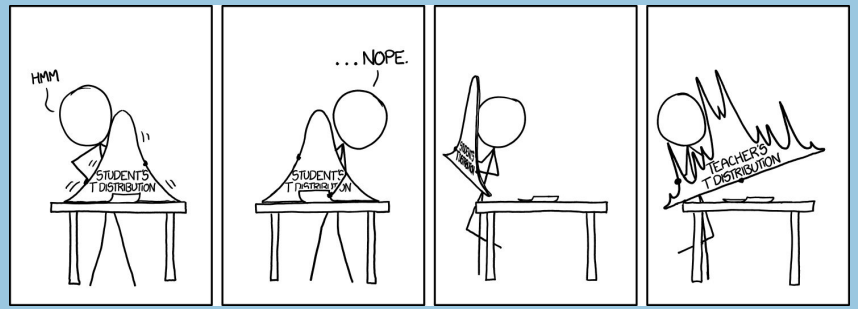
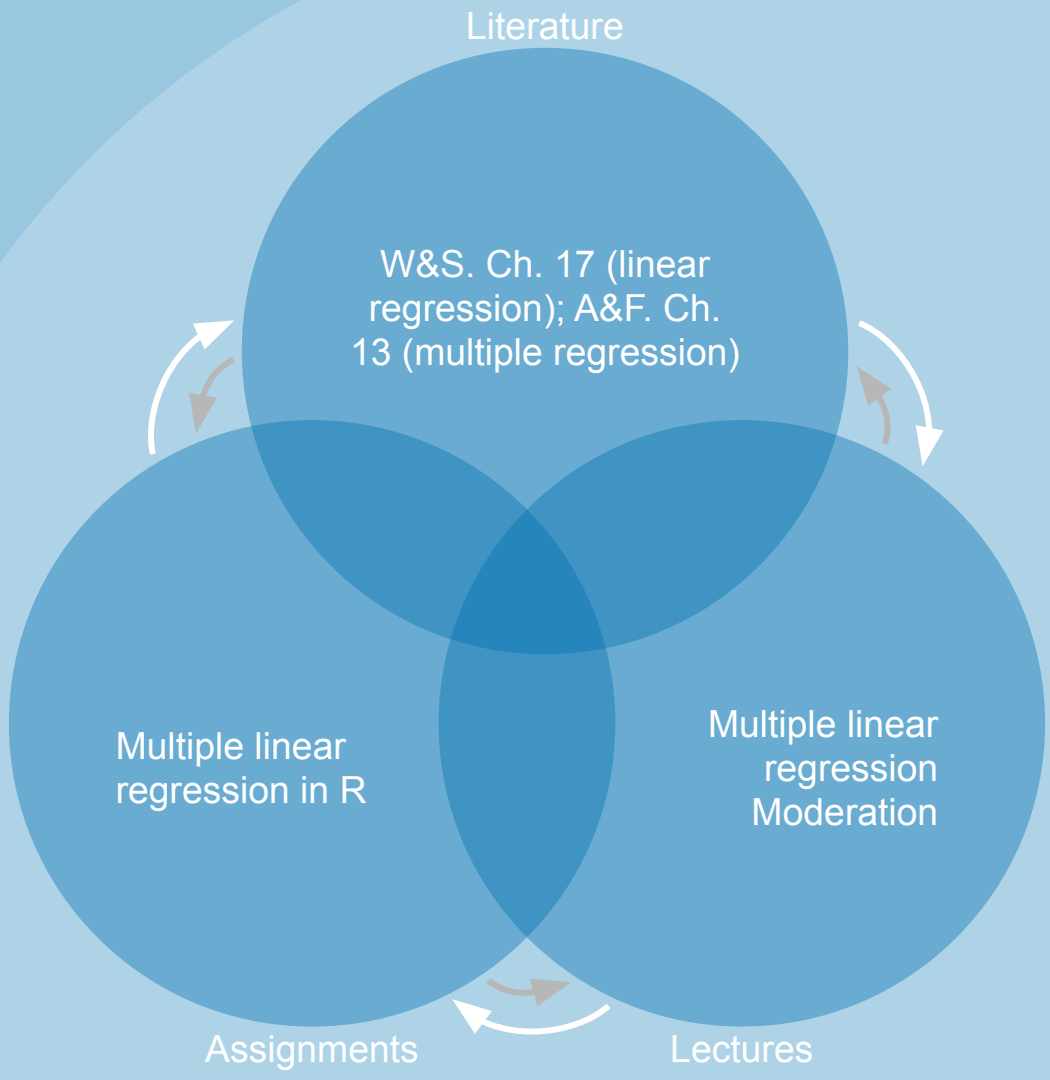


Illustration by Randall Munroe (wtf)

🥇 Nail it


Literature
W&S. Ch. 17 (linear regression); A&F. Ch. 13 (multiple regression)
Multiple linear regression in R
Multiple linear regression Moderation
Assignments
Lectures

🥇 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.

👩‍💻 Take-home assignments

📆 Weekly assignment

🍻 Pub quiz

Create an *informative* four-choice question about the content of today's lecture.

An informative question has a large spread in responses across answer options.

Clarify answer options (which are (in)correct and why).

Illustration adapted from Snippets.com

🔮 Overview

Topics

Probabilities & distributions
Frequentist inference
Multiple linear regression
| Factorial ANOVA
Nonparametric inference
Bayesian inference

Illustration by Jennifer Cheuk

⚠️ Don’t look here!

Show that a *t*-test and linear regression analysis return the same results.

Share your attempt (and tell whether you needed hints)!

Hints (select and copy/paste the invisible text below to reveal it)

0.

1.

2.

3.

✍️ Colophon

Slides
alexandersavi.nl/teaching/

License