Regression modelling strategies

Stefan Hartmann
HHU Düsseldorf

Course overview

Heinrich Heine
Universität Düsseldorf

- From data to straight lines
- Implementation in R
- Outlook: Other types of regression modelling (esp. count data)

Before we start...

Heinrich Heine
Universität Düsseldorf

Debunking some myths about statistical modelling

- **Myth 1: There is one right model for every dataset.**

→ No: "All models are wrong but some are useful" (George Box)

- **Myth 2: You can find the "right" model using a flowchart**

→ No: Each dataset presents a challenge on its own right, and modelling involves clearly stating your prior assumptions and expectations and "translating" them into a model.

- **Myth 3: Statistical modelling is objective.**

→ No: Modelling always involves the researcher's choices and expectations.

Basics of frequentist modeling

Heinrich Heine
Universität Düsseldorf

- 1. The goal is to estimate **parameter values** (e.g., the mean voice onset time for stops produced by American English speakers).
- 2. These parameters have **true** (**population**) **values**, which are approximated by taking a sample.
- 3. The **population** from which the sample is taken is **infinitely large**.
- 4. Samples drawn from the population are **representative** and **random** (e.g., samples are from all American English speakers, randomly).

prerequisite: **independence** of data points!

Basics of statistical modeling

Heinrich Heine
Universität Düsseldorf

What is a statistical model?

- A statistical model is the explanation of a dependent variable with the help of independent variables:

$$y = f(x)$$

- A model is never "true": "All models are wrong but some are useful" (George Box)

$$y = f(x) + \varepsilon$$

Basics of statistical modeling

Heinrich Heine
Universität Düsseldorf

Models and distributions

- Most statistical models assume that the observed data has been "generated" by a process following a certain distribution

- as such, both theoretical and empirical distributions play a role in modeling:

  - the distribution that we observe in the actual data,

  - a theoretical probability distribution that helps modelling the observed data.

Statistical distributions

Heinrich Heine
Universität Düsseldorf

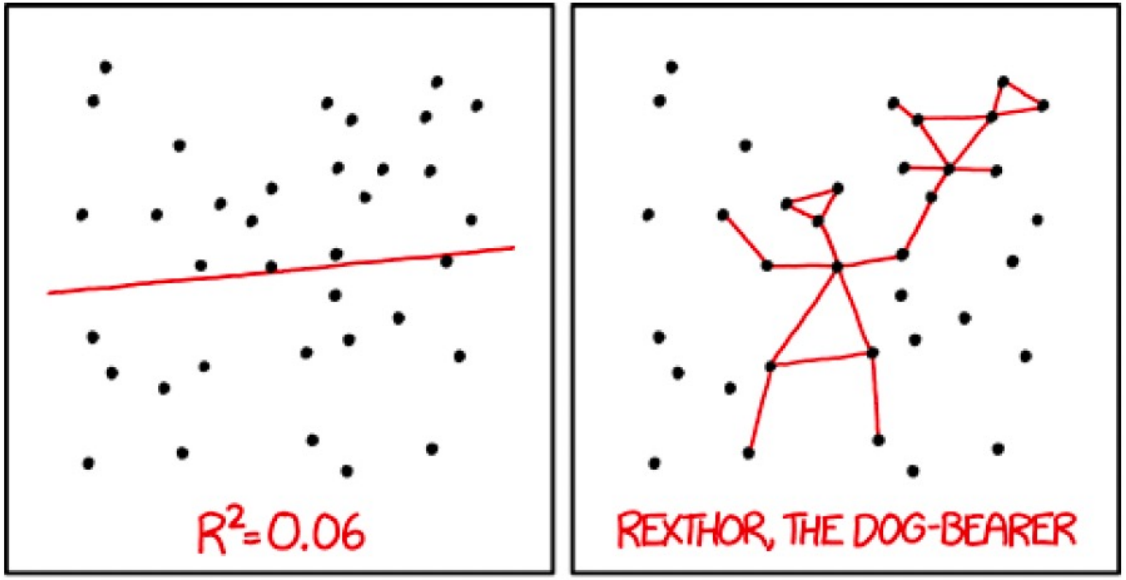
- Statistical models model **distributions**
- In statistical modeling, we use theoretical distributions to model observed data

Statistical distributions

- Shinyapp to play around with probability distributions:

https://tanguylefort.shinyapps.io/probas/

Visual inspection is crucial!



Types of variables

- Dependent variable
- Indipendent variable
- Confounding variable

Types of variables

Heinrich Heine
Universität Düsseldorf

dependent variable

- the variable that is the object of study
- also called *outcome variable* or *response variable*

Types of variables

Heinrich Heine
Universität Düsseldorf

Independent variable

- assumed influential variable
- can be manipulated by the researcher directly or indirectly
- also called *predictor variable*

Types of variables

Heinrich Heine
Universität Düsseldorf

Confounding variables

- all further variables that influence the dependent variable apart from the independent variable(s)

independent variable 1

independent variable 2

independent variable 3

...

dependent variable

confounding variable 1

confounding variable 2

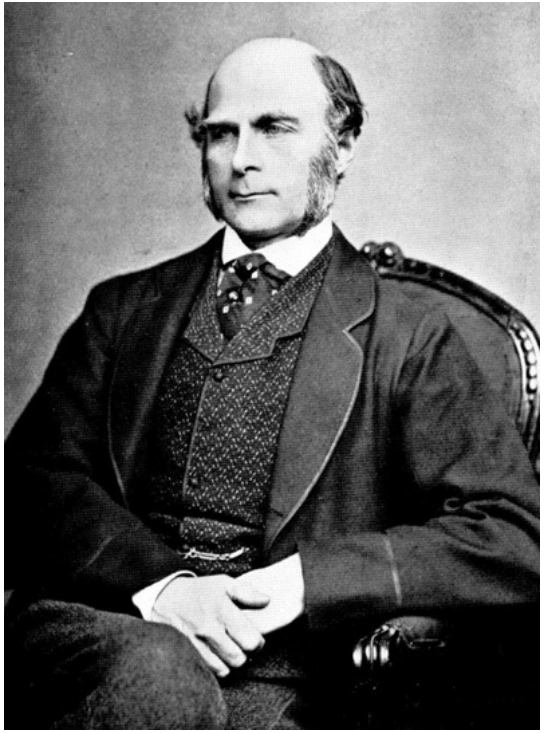
...

Regression models

Why "regression"?

- Francis Galton (1822-1911): Sons of very tall and very short men tend to be more similar to the population mean → "regression to the mean"

- predicting the height of each son from the father's height would be error-prone: better to use the population of fathers!



(McElreath 2020: 94f.; Picture: public domain / Wikimedia Commons)

The linear model

Simple linear regression

- We can predict all kinds of data using the following general equation:

$$Outcome_i \sim (model) + error_i$$

- In regression, the model we fit is linear → we summarize the data with a straight line.

The linear model

Heinrich Heine
Universität Düsseldorf

Simple linear regression

- Simple linear regression boils down to fitting **a straight line** to our data.
- In mathematical terms, a straight line can be defined by two parameters:
  - the **intercept (b_0)**
  - the **slope (b_1)**

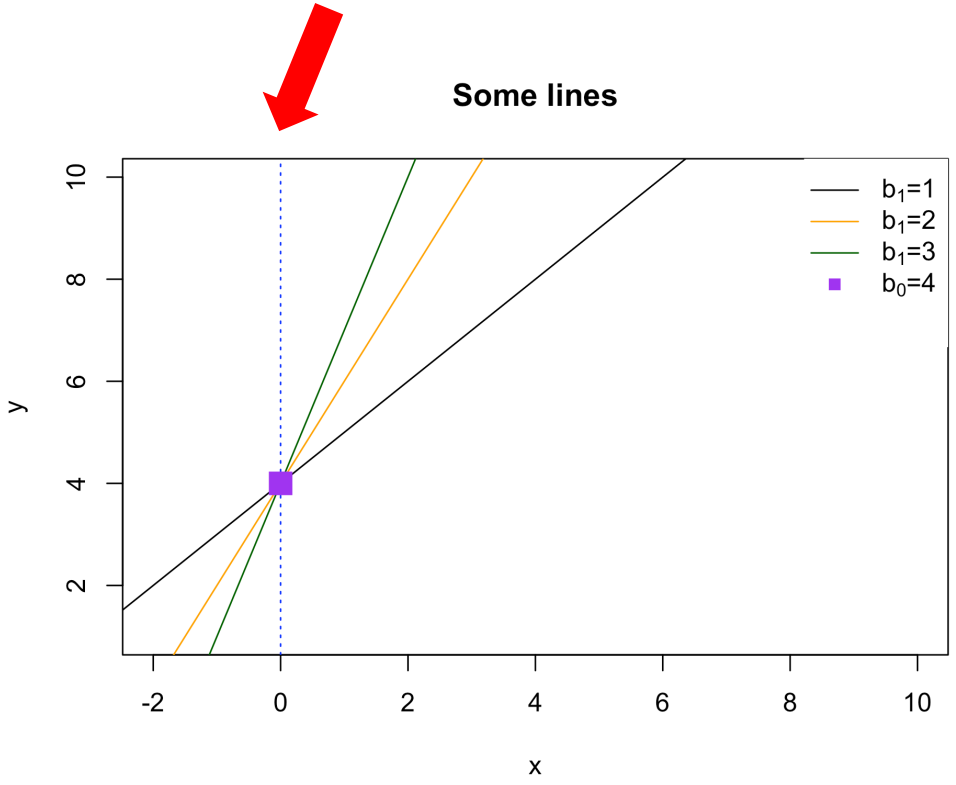
The linear model

Intercept and slope

- The **intercept (b_0)** is the point where x = 0 (or, where the regression line crosses the vertical axis of the graph)

- The **slope** (or gradient, **(b_1)** of the line is defined as change in y over change in x:

$$slope = \frac{\Delta y}{\Delta x}$$



The linear model

Simple linear regression

- Simple linear regression boils down to fitting **a straight line** to our data.
- In mathematical terms, a straight line can be defined by two parameters:
  - the **intercept (b_0)**
  - the **slope (b_1)**
- Hence, the linear model has the following basic form:

$$y = b_0 + b_1 * x$$

intercept

slope

datapoints on the x axis for which we generate the prediction

The linear model

Excursus: Notation conventions

- $\hat{Y}$:: Model estimate
- Y: actual values in the population

from $\hat{Y}$ to Y:

$\hat{Y} = b_0 + b_1 * x$

$Y = b_0 + b_1 * x + \varepsilon$

- (sometimes also lowercase; conventions differ wildly)

The linear model

Simple linear regression

- (Non-linguistic) example: Correlation of height and weight
- Dataset: Heights and weights of !Kung San individuals (an indigeneous people living in the Kalahari desert), from work by Nancy Howell
- Data: **https://t1p.de/howell_raw**

The linear model

Simple linear regression

```
# read data
hw <- read_csv("data/howell.csv")

# only adults
hw <- subset(hw, age >= 18)

# fit a model
m01 <- lm(height ~ weight, data = hw)
summary(m01)
```

```
> summary(m01)
```

```
Call:
lm(formula = height ~ weight, data = hw)
```

the formula, as we entered it

```
Residuals:
     Min       1Q   Median       3Q      Max
-19.7464  -2.8835   0.0222   3.1424  14.7744
```

distribution of residuals

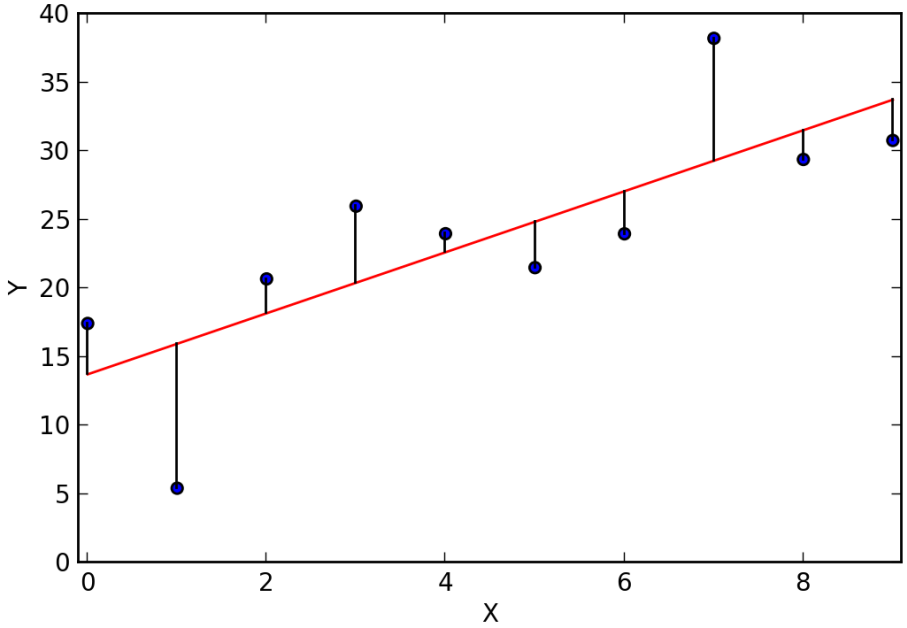
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 113.87939    1.91107   59.59   <2e-16 ***
weight        0.90503    0.04205   21.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

coefficients

```
Residual standard error: 5.086 on 350 degrees of freedom
Multiple R-squared:  0.5696,    Adjusted R-squared:  0.5684
F-statistic: 463.3 on 1 and 350 DF,  p-value: < 2.2e-16
```

summary statistics

Residuals



Residuals: vertical distances between the predicted and the observed values.
(Figure just for illustration; data are not from our model ☺)

```
Call:
lm(formula = height ~ weight, data = hw)

Residuals:
     Min       1Q   Median       3Q      Max
-19.7464  -2.8835   0.0222   3.1424  14.7744

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 113.87939    1.91107   59.59   <2e-16 ***
weight        0.90503    0.04205   21.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.086 on 350 degrees of freedom
Multiple R-squared:  0.5696,    Adjusted R-squared:  0.5684
F-statistic: 463.3 on 1 and 350 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = height ~ weight, data = hw)

Residuals:
     Min       1Q   Median       3Q      Max
-19.7464  -2.8835   0.0222   3.1424  14.7744

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 113.87939    1.91107   59.59   <2e-16 ***
weight        0.90503    0.04205   21.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.086 on 350 degrees of freedom
Multiple R-squared:  0.5696,    Adjusted R-squared:  0.5684
F-statistic: 463.3 on 1 and 350 DF,  p-value: < 2.2e-16
```

A note on p-values

Heinrich Heine
Universität Düsseldorf

Frequent misinterpretations

- Myth 1: The *p*-value represents the probability of the null hypothesis being true.

→ Wrong: The null hypothesis is an assumption, its truth cannot be known. Instead, the p-value gives the probability of getting a distribution like the one observed in our data in a population for which the null hypothesis is true.

- Myth 2: The *p*-value represents the strength of an effect.

→ Wrong: *p*-values usually depend on sample size. This is why measures of effect size should always be given alongside p-values.

- Myth 3: if $p<0.05$, one is justified to believe more strongly on one's alternative hypothesis.

→ Wrong: p-value only measures **in**compatibility with **null** hypothesis.

```
Call:
lm(formula = height ~ weight, data = hw)

Residuals:
     Min       1Q   Median       3Q      Max
-19.7464  -2.8835   0.0222   3.1424  14.7744

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 113.87939    1.91107   59.59   <2e-16 ***
weight        0.90503    0.04205   21.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.086 on 350 degrees of freedom
Multiple R-squared:  0.5696,    Adjusted R-squared:  0.5684
F-statistic: 463.3 on 1 and 350 DF,  p-value: < 2.2e-16
```

The F value

- F = effect (or "treatment") variance / error variance
- in other words: F tells us how much a model can explain relative to how much it can't explain

- operationalized as:

$$\frac{mean\ model\ sum\ of\ squares}{mean\ resiudal\ sum\ of\ squares}$$

- The more random variation (→ higher value in the denominator → lower F value), the more likely it is that any patterns found in the data are due to chance.

(Field et al. 2015: 260)

```
Call:
lm(formula = height ~ weight, data = hw)

Residuals:
     Min       1Q   Median       3Q      Max
-19.7464  -2.8835   0.0222   3.1424  14.7744

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 113.87939    1.91107   59.59   <2e-16 ***
weight        0.90503    0.04205   21.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.086 on 350 degrees of freedom
Multiple R-squared:  0.5696,    Adjusted R-squared:  0.5684
F-statistic: 463.3 on 1 and 350 DF,  p-value: < 2.2e-16
```

R^2

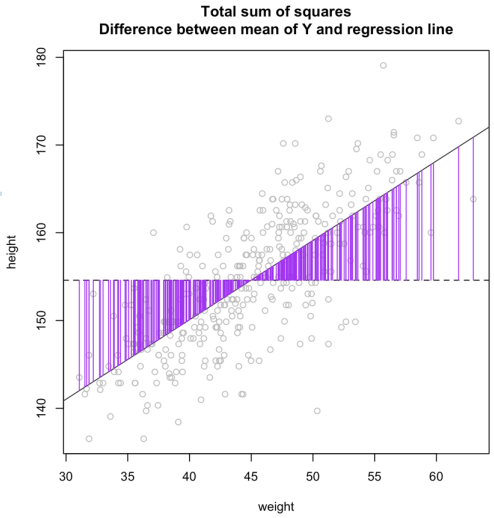
- R^2 represents the proportion of the variance "explained" (or, more neutrally: described) by the predictors in a regression model.
- Basic idea of R^2 (also called "coefficient of determination"): compare the model with the simplest model, i.e. the overall mean.
- R^2 and F are closely related:

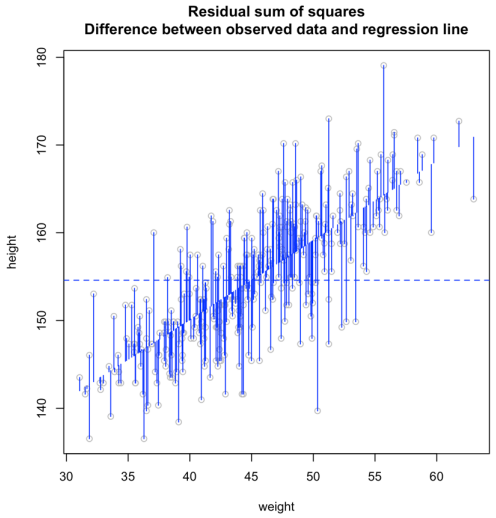
$$R^2 = 1 - (1 + F \cdot \frac{p-1}{n-p})^{-1}$$

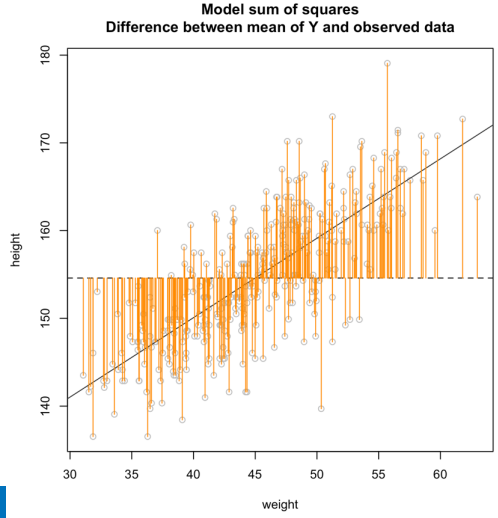
- with p = number of parameters and n = number of observations

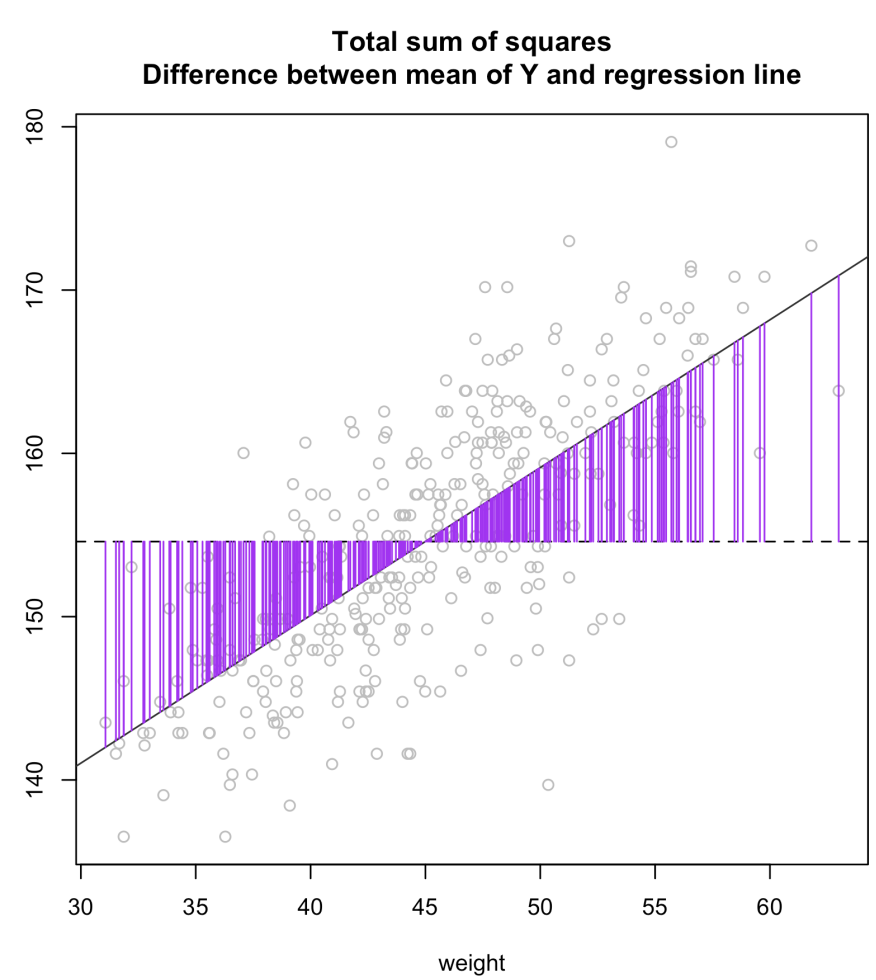
R^2

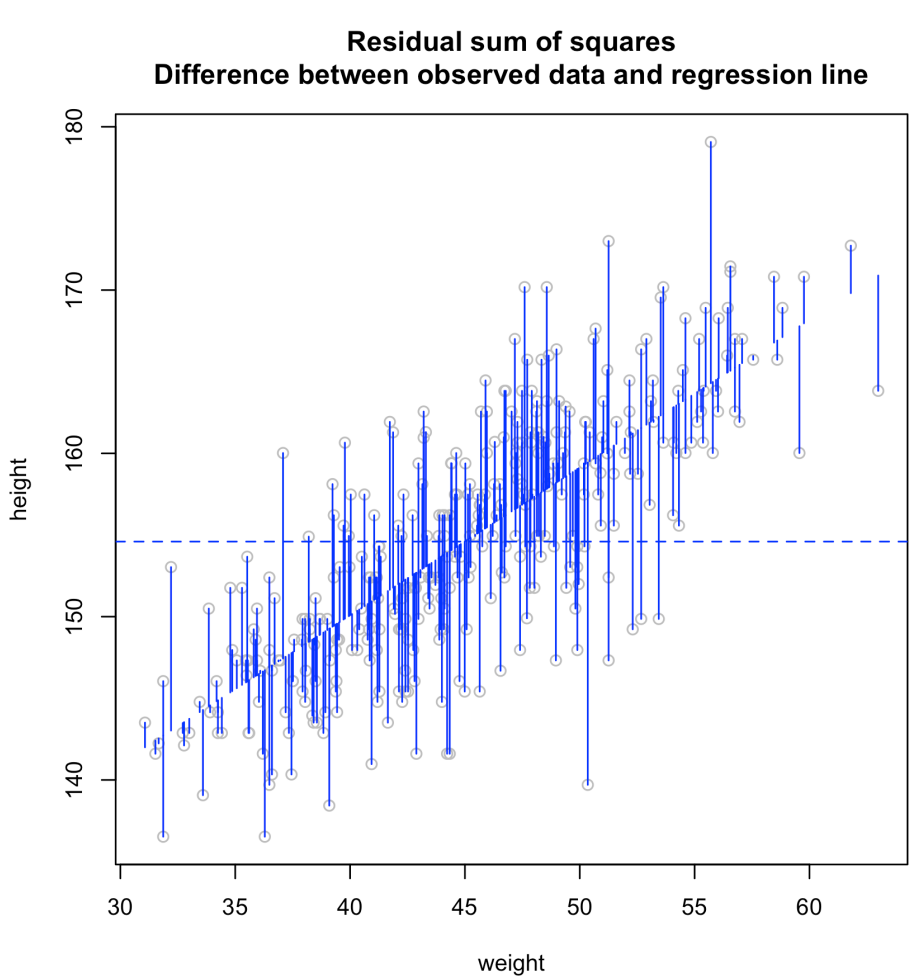
$$R^2 = \frac{SS_M}{SS_T}$$



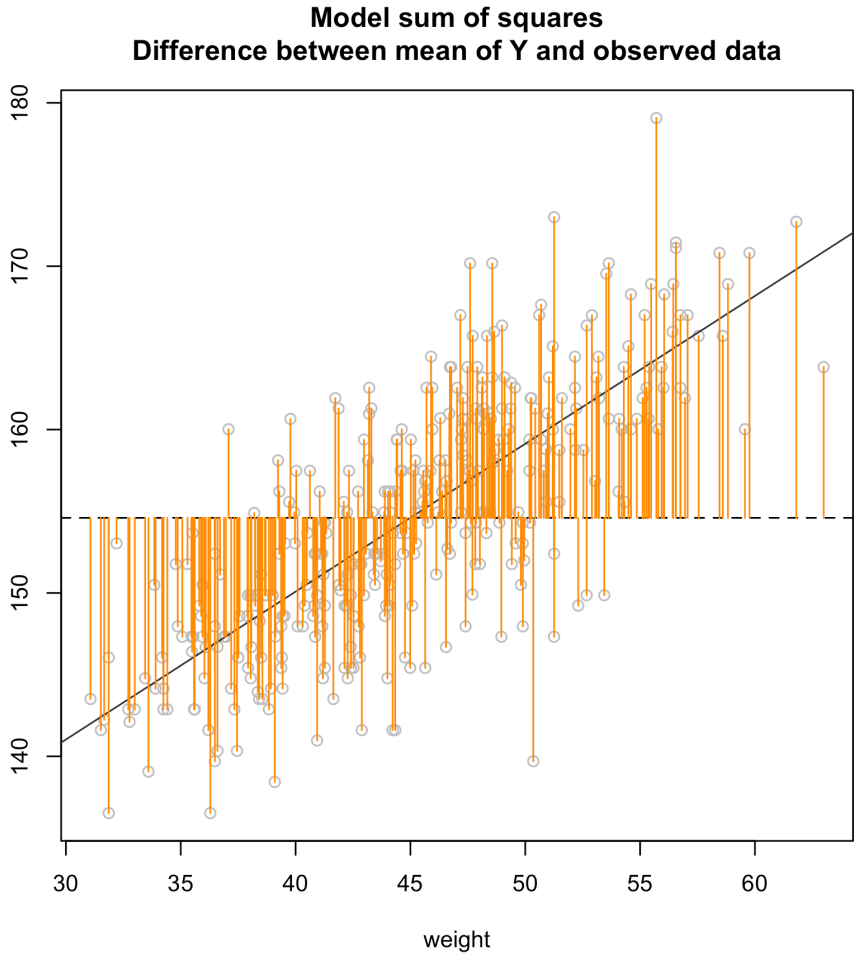








Model sum of squares
Difference between mean of Y and observed data


Model sum of squares
Difference between mean of Y and observed data
180
170
160
150
140
30
35
40
45
50
55
60
weight

R^2

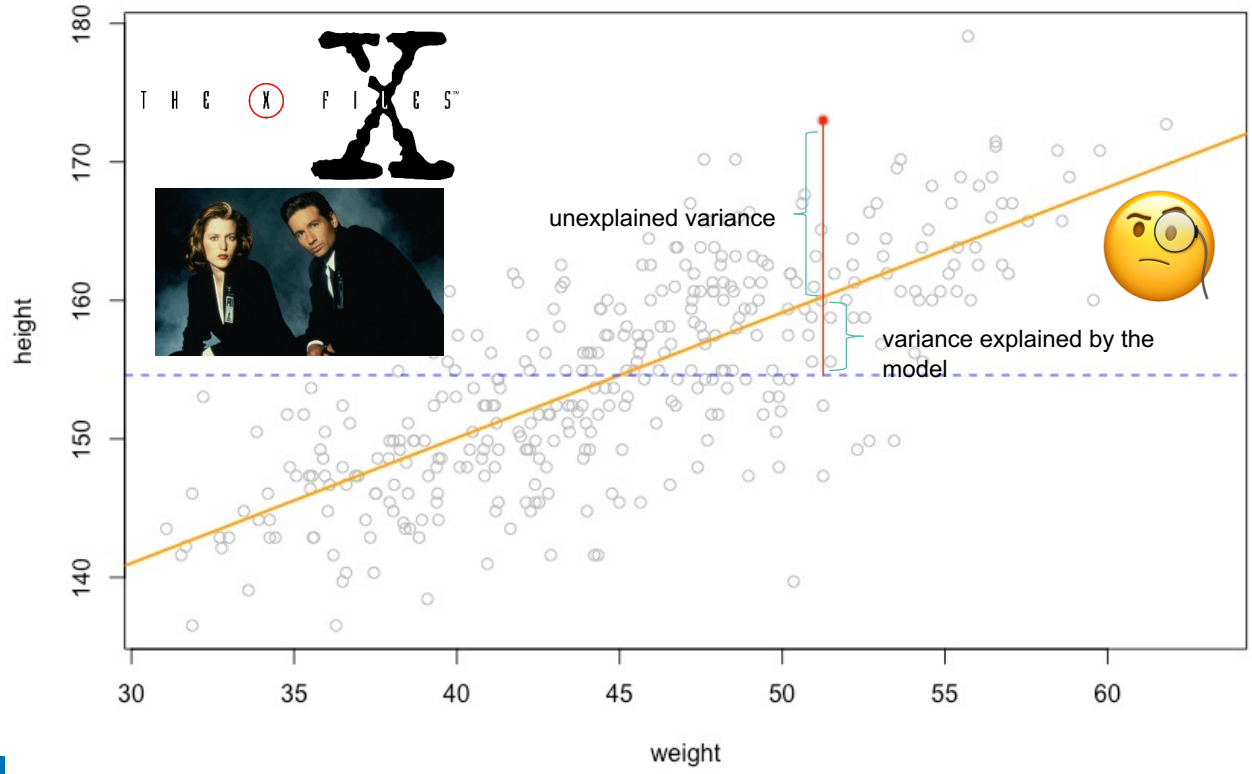
$$R^2 = \frac{SS_M}{SS_T}$$

or:

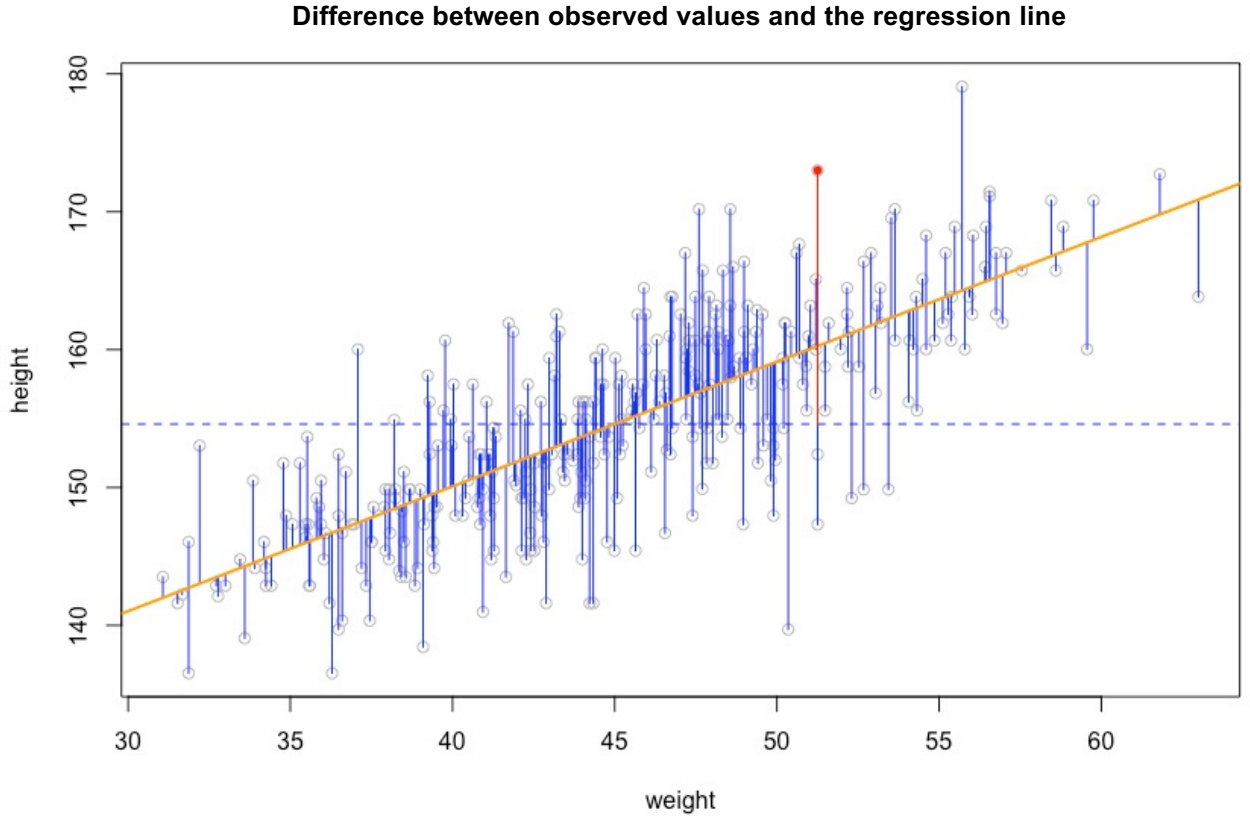
$$R^2 = \frac{SSE_{model}}{SSE_{null}}$$

... because SS_T is the model sum of squares (SS_M) of the null model (i.e. the model with just the intercept that basically just models the overall mean)!

F-Ratio


height
weight
180
170
160
150
140
30
35
40
45
50
55
60
unexplained variance
variance explained by the model

F-Ratio


Difference between observed values and the regression line
height
weight
140
150
160
170
180
30
35
40
45
50
55
60

Beyond R^2

Heinrich Heine
Universität Düsseldorf

- Problem of R^2: as soon as we add more predictors, the model fit automatically becomes better.
- Hence, the output of R's modelling functions always reports an **adjusted** R2 value that penalizes models for involving more predictors

Adjusted R^2

- $R^2_{adj} = 1 - \frac{(1-R^2)(N-1)}{N-l-1}$

- N: Number of datapoints/observations
- k: Number of parameters

Akaike Information Criterion

Heinrich Heine
Universität Düsseldorf

$$AIC = 2k - 2\ln(L)$$

- introduces penalty for additional parameters
- larger AIC: worse model fit
- smaller AIC: better model fit
- however, there are no clear rules-of-thumn when AIC is significantly smaller or "better" than in a null model (sometimes difference of 6 mentioned as a point of orientation)
- alternative formula for least-square regression types:

$$AIC = n \ln\left(\frac{SSM}{n}\right) + 2k$$

```
Call:
lm(formula = height ~ weight, data = hw)

Residuals:
     Min       1Q   Median       3Q      Max
-19.7464  -2.8835   0.0222   3.1424  14.7744

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 113.87939    1.91107   59.59   <2e-16 ***
weight        0.90503    0.04205   21.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.086 on 350 degrees of freedom
Multiple R-squared:  0.5696,    Adjusted R-squared:  0.5684
F-statistic: 463.3 on 1 and 350 DF,  p-value: < 2.2e-16
```

Challenge

Heinrich Heine
Universität Düsseldorf

- Which height does our model predict for a person who weights 150 kg?
- Which height does it predict for a person who weighs 2 kg?
- Which height does it predict for a person who weighs -73kg?
- What does this tell us about the validity of the model?

Challenge

- Which height does our model predict for a person who weights 150 kg?

→ We can plug in the value in our default formula:

$y = b_0 + b_1 * x$

outcome = intercept + slope * x

```
> summary(m01)

Call:
lm(formula = height ~ weight, data = hw)

Residuals:
     Min       1Q   Median       3Q      Max
-19.7464  -2.8835   0.0222   3.1424  14.7744

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 113.87939    1.91107   59.59   <2e-16 ***
weight        0.90503    0.04205   21.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.086 on 350 degrees of freedom
Multiple R-squared:  0.5696,    Adjusted R-squared:  0.5684
F-statistic: 463.3 on 1 and 350 DF,  p-value: < 2.2e-16
```

intercept

slope

"Check list" for regression models

1. inspect data
   - are there missing values (NAs)?
   - Are there outliers?
   - visualize!
   - consider centering numeric variables
     - (often scaling and log-transform also make sense)
2. fit model
3. interpret coefficients
4. check assumptions!!!
5. evaluate model, compare with null models

Meme from https://github.com/rmcelreath/stat_rethinking_2024/

A linguistic example

Heinrich Heine
Universität Düsseldorf

- example dataset of reaction time data:

https://t1p.de/elp_freq_raw

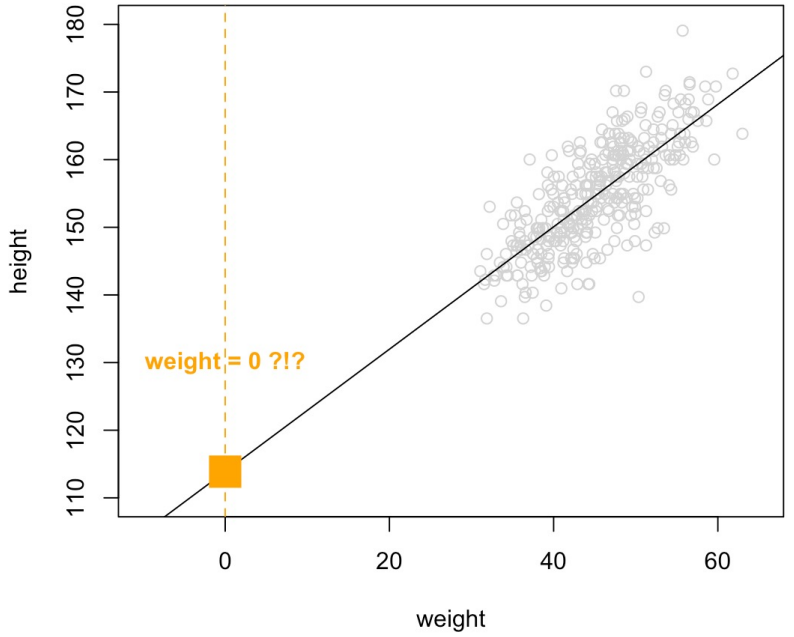
Data transformation

Data transformation

Linear transformation: Centering data

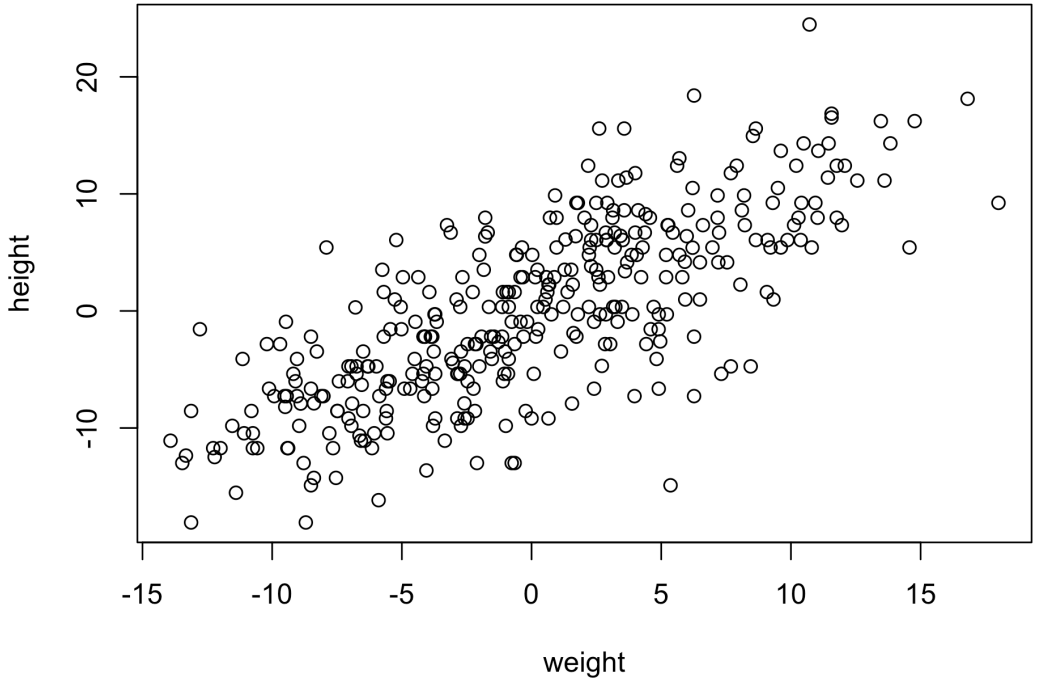
- Centering means to subtract the mean of the variable from each data point

- Centering can be helpful because it makes the intercept more interpretable: e.g. nobody weighs 0kg



Data transformation

Linear transformation: Centering data



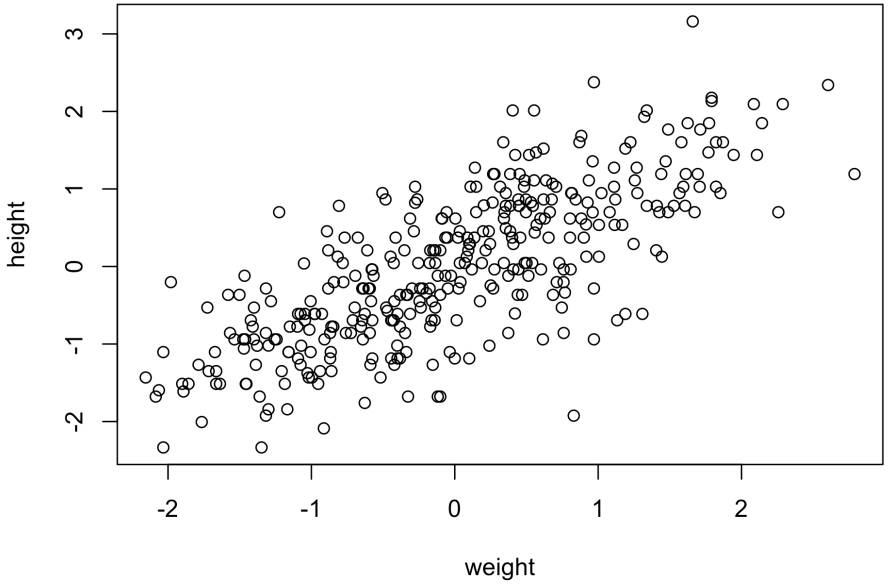
Data transformation

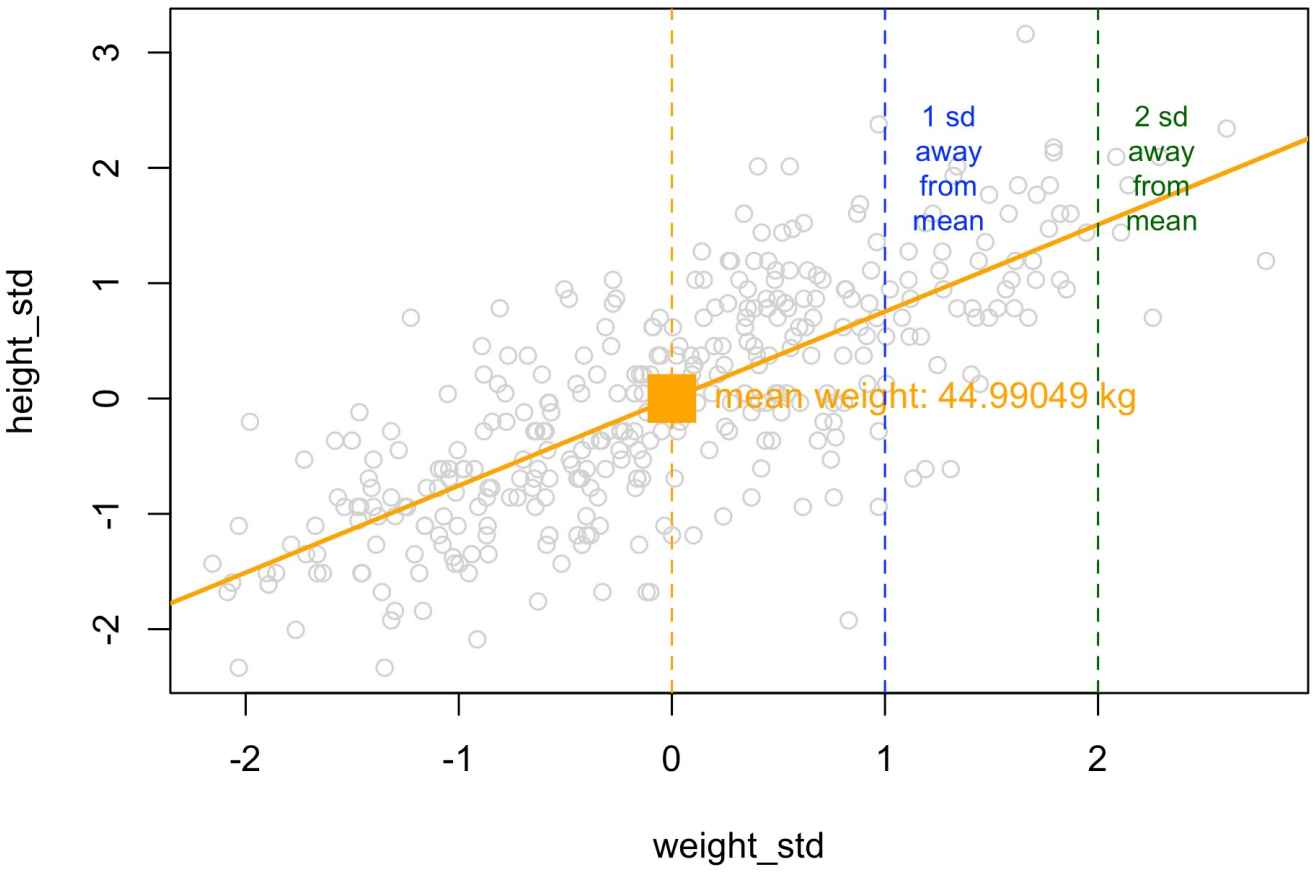
Linear transformation: Standardizing data

- Standardizing (or z-scoring) means that the centered variable is divided by the standard deviation of the sample

- This means that we get rid of the variable's metric:

Uncentered – centered – scaled




height_std
3
2
1
0
-1
-2
-2
-1
0
1
2
weight_std
1 sd
away
from
mean
2 sd
away
from
mean
mean weight: 44.99049 kg

Nonlinear transformation

Using logarithms

- Logarithmization is often used when analyzing linguistic data because it makes data less skewed
- The logarithm is the inverse of the exponential function:

Logarithms	Exponentiation
$log_{10}(1) = 0$	$10^0 = 1$
$log_{10}(10) = 1$	$10^1 = 10$
$log_{10}(100) = 2$	$10^2 = 100$
$log_{10}(1000) = 3$	$10^3 = 1000$

Nonlinear transformation

Using logarithms

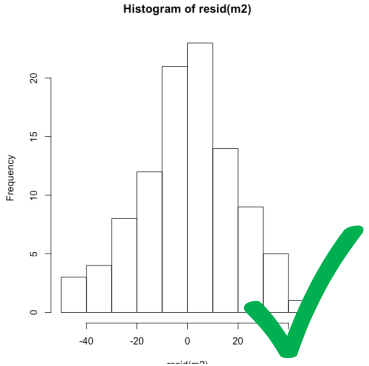
- many cognitive and linguistic phenomena are scaled logarithmically – e.g. there is evidence that processing times and and phonetic variables like loudness and pitch are scaled logarithmically
- Logarithms can be taken to different bases – the **natural logarithm** is the logarithm to the base of *e* (Euler's number, ~2.72)
- R's `log()` function gives you the natural logarithm by default, log10 the logarithm to the base of 10
- Note that the logarithm is only defined for positive numbers, hence `log(0)` is undefined! When working with datasets that contain zero values, a common trick is to add 1 to all datapoints. R's function `log1p()` automatically calculates `log(x+1)` for you.

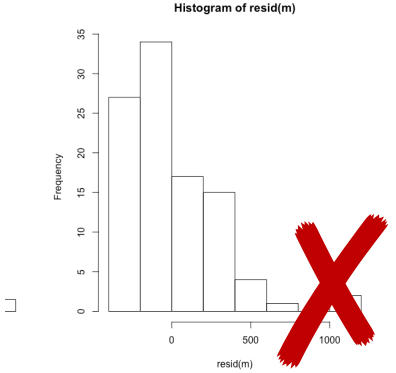
- https://t1p.de/example_maedchen

Assumptions

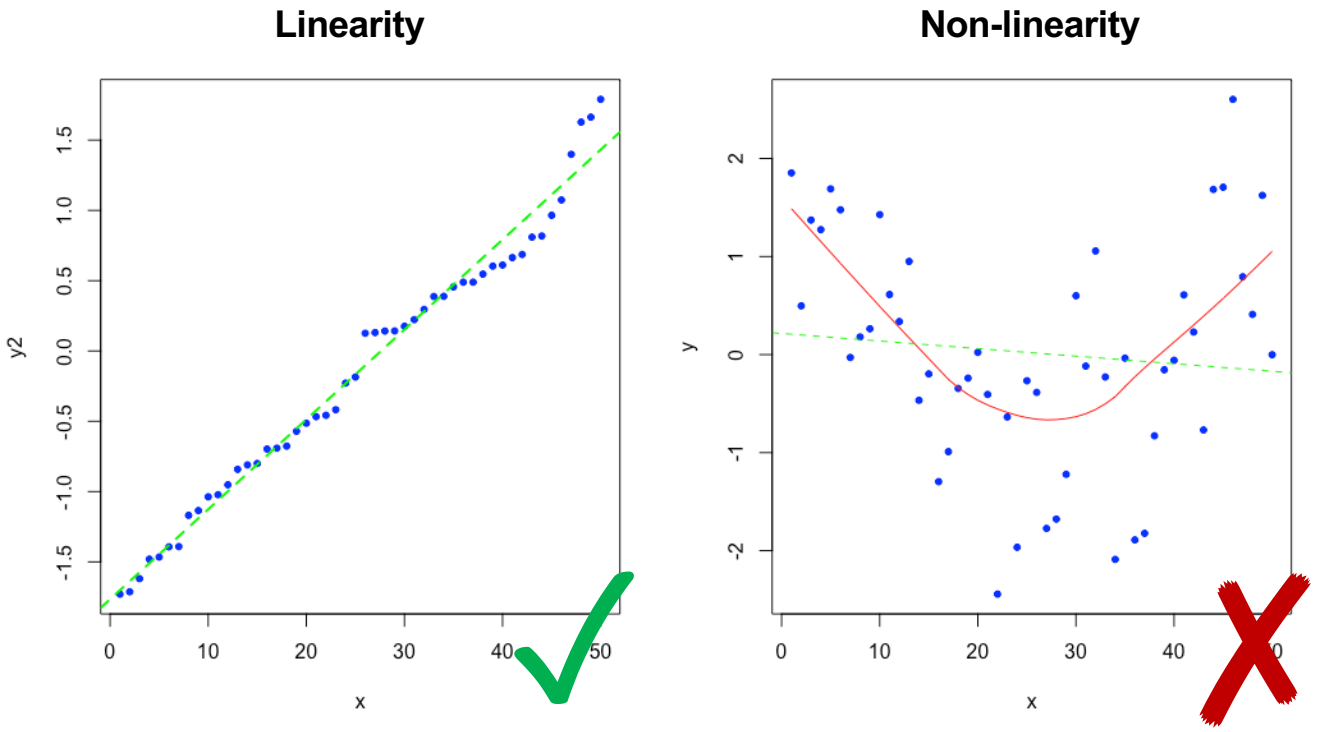
Normal distribution of residuals

- An important assumption is that the residuals are normally distributed.
- This can be checked visually by plotting a histogram of residual or, a Q-Q (quantile-quantile) plot of residuals, or by plotting residuals against fitted values.

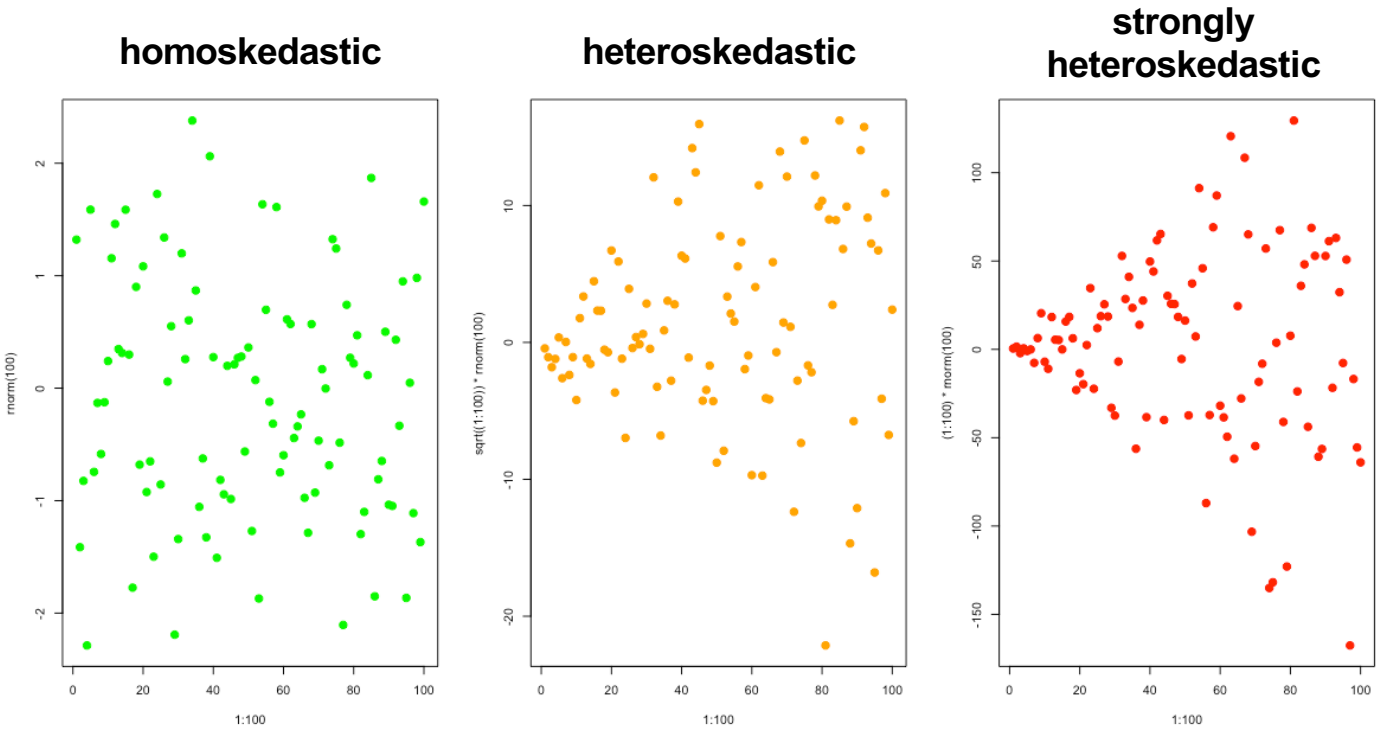




Linearity



Homoskedasticity of residuals



No multicollinearity

- Multicollinearity refers to a strong linear dependency between predictors
- (Only relevant for models with more than one predictors of course!)
- can be tested using **variance inflation factors** (VIFs)

No multicollinearity

What is (multi)collinearity?

- strong association between two or more predictor variables
- Example from McElreath (2021): predicting an individual's height using the length of their legs
- The predictors "left leg" and "right leg" contribute basically the same information – hence, the coefficients change drastically when both predictors are included, compared to when just one of them is included!
- sample size interacts with collinearity: all other things being equal, more data means that regression coefficients can be estimated more precisely.

(Wiinter 2020: 114f.; McElreath 2021: 160ff.)

No multicollinearity

Variance inflation factors

- Variance Inflation Factors measure the degree of (multi)collinearity
- related to R^2:

$$VIF_i = \frac{1}{R_i^2}$$

- implemented in R package car: car::vif
- rules of thumb proposed in the literature vary widely – some see VIFs < 10 as unproblematic, others < 5, yet others are even stricter (see Levshina 2015: 160 and https://www.reneshbedre.com/blog/variance-inflation-factor.html)

Multiple regression

Adding more predictors

- Most phenomena we investigate are not monocausal – hence, we usually want to include more than one predictor in our model.
- This is easily possible by adding more predictors in a linear model.
- But how do we interpret the results?

Example: age and gender as predictors of height

```
Call:
lm(formula = height ~ age + gender, data = hw)

Residuals:
     Min       1Q   Median       3Q      Max
-23.2225  -4.9538  -0.5237   5.5549  28.4786

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.6859     1.2167  58.920  < 2e-16 ***
age           4.5278     0.1134  39.932  < 2e-16 ***
gendermale    3.4857     1.2148   2.869  0.00458 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.409 on 189 degrees of freedom
Multiple R-squared:  0.8944,    Adjusted R-squared:  0.8933
F-statistic: 800.7 on 2 and 189 DF,  p-value: < 2.2e-16
```

influence of *gender* when we already know the influence of *age*

Categorical predictors

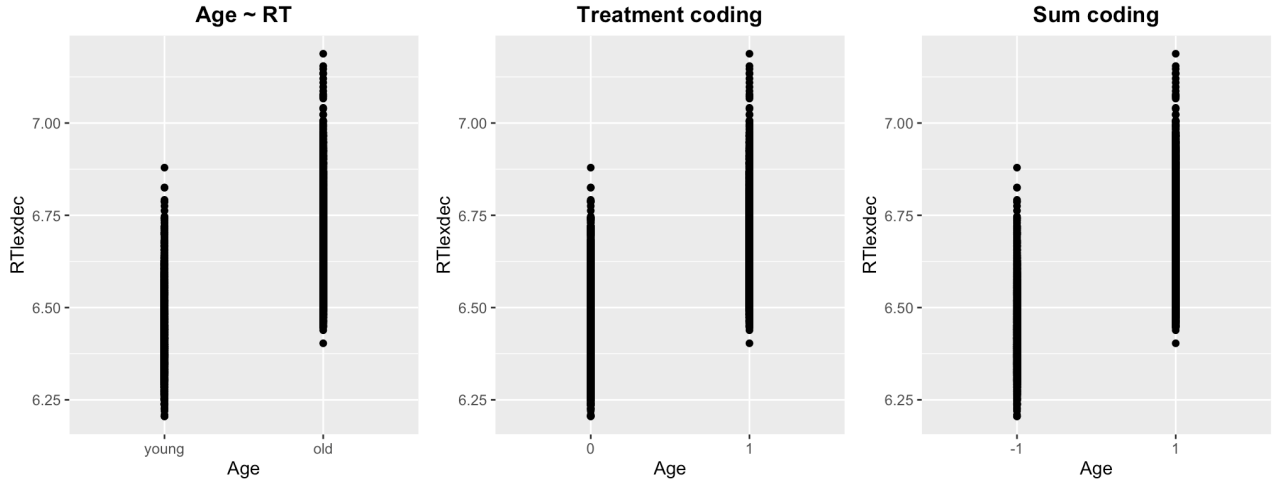
The challenge

- So far, we have worked with **numeric** predictors – but what if a predictor is **categorical**?
- e.g. experimental condition vs. control condition; gender; text type; ...
- to convert these categories into numeric values, regression uses **coding schemes**
- e.g. category A = 0, category B = 1
- two types of dummy coding are widespread:
  - treatment coding (also called dummy coding)
  - sum coding (also called contrast coding)

Categorical predictors

Sum coding and treatment coding

- treatment coding: category at $x = 0$ serves as the reference (base) level
- sum coding: the intercept is in the middle of the two categories (conceptually analogous to centering numeric variables)



Categorical predictors

Sum coding and treatment coding

- Treatment coding takes the **grand mean** as the intercept. Each level of a factor is changed to be compared to a reference (base) level.
- Sum coding compares the mean of a response variable at a given level to the overall mean of the response variable.
- sum coding can have advantages for the interpretation of variables in some cases (see Winter 2020: 125)

Interactions

Understanding interactions

- When we can assume that two or more factors **jointly** influence the response variable, it can make sense to model an interaction...
- ...i.e. to allow parameters to be conditional on further aspects on the data (McElreath 2021: 238)
- to incorporate an interaction, the two predictors are multiplied by each other:

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 (x_1 * x_2)$$

McElreath 2020: 238; Winter 2020: 133f.)

Interactions

Understanding interactions

- Example from McElreath (2020):
  - Response variable: Bloom size of tulips
  - predictor variables: water and shade
  - Tulips need both water and light to grow:
    - Without water, the plant will dry out
    - Without light, the plant will also die at some point.

Interactions

Interpreting interactions

- Important: Interactions change the meanings of coefficients!

Model without interaction

Heinrich Heine
Universität Düsseldorf

```
> summary(m00)

Call:
lm(formula = blooms ~ water.c + shade.c, data = t)

Residuals:
     Min       1Q   Median       3Q      Max
-121.025  -28.253   -5.502   37.646  115.262

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   128.99      11.70  11.021 7.12e-11 ***
water.c        75.80      14.33   5.288 2.01e-05 ***
shade.c       -41.60      14.33  -2.902  0.00782 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.82 on 24 degrees of freedom
Multiple R-squared:  0.6026,    Adjusted R-squared:  0.5694
F-statistic: 18.19 on 2 and 24 DF,  p-value: 1.553e-05
```

additional influence of *water* when we already know the influence of *shade*

additional influence of *shade* when we already know the influence of *water*

Model with interaction

```
Call:
lm(formula = blooms ~ water.c + shade.c + water.c:shade.c, data = t)

Residuals:
    Min      1Q  Median      3Q     Max
-121.03  -26.92    5.27   35.46   75.40

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      128.994      9.432  13.676 1.56e-12 ***
water.c           75.802     11.552   6.562 1.07e-06 ***
shade.c          -41.603     11.552  -3.601  0.00151 **
water.c:shade.c  -52.852     14.148  -3.736  0.00108 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.01 on 23 degrees of freedom
Multiple R-squared:  0.7526,    Adjusted R-squared:  0.7204
F-statistic: 23.33 on 3 and 23 DF,  p-value: 3.663e-07
```

expected change in the response variable when *water* increases by one unit and *shade* is at its mean (here: 0)

expected change in the response variable when *shade* increases by one unit and *water* is at its mean (here: 0)

Interaction as "slope adjustment term": a) influence of *water* on response variable when increasing *shade* by one unit; b) influence of *shade* on response variable when increasing *water* by one unit.

References

- Baayen, R. H. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.
- Brehm, Laurel & Phillip M. Alday. 2022. Contrast coding choices in a decade of mixed models. *Journal of Memory and Language* 125. 104334. https://doi.org/10.1016/j.jml.2022.104334.
- Winter, Bodo. 2019. *Statistics for linguists: an introduction using R*. New York: Routledge.

Course overview

Heinrich Heine
Universität Düsseldorf

- Part 1: Basics of regression modelling
- **Part 2: Generalized linear models**
- Part 3: Mixed-effects regression models

Straight lines are (not) enough

Heinrich Heine
Universität Düsseldorf

Why do we need GLMs?

- Linear models are well-suited for data with a numeric/continuous response variable – e.g. reaction times
- (Caveat: in many cases, straight lines are not the best fit even for those data types; check out e.g. polynomial regression modeling for "curving" straight lines)
- But in many cases, we are dealing with **categorical** response variables – e.g., variant A vs. variant B; error vs. no error etc.

Straight lines are (not) enough

Logistic regression

- In the context of logistic regression, we are usually interested in modeling *p* as a function of one or more predictors.
- Remember our regression equation:

$$y_i = \beta_0 + \beta_1 x_i$$

- Ideally, we want different probabilities for different values of *x*.
- The regression equation $\beta_0 + \beta_1 x_i$ can predict any continuous variable – but **probabilities** have to be between 0 and 1!
- As such, we have to 'squeeze' the output in the interval [0,1] – and that's what the logistic function does!

Straight lines are (not) enough

The challenge

- We have to 'squeeze' [–∞, +∞] in the interval [0,1] in order to model probabilities.
- We can do this using the logistic function
- in logistic regression, the logit function serves as the so-called link function

Straight lines are (not) enough

Basic idea of GLMs

- when fitting a simple linear model, the underlying assumption is that the process that generates the response variable follows a normal distribution
- GLMs generalize the linear model framework to incorporate data-generated processes that follow any distribution.

Simple linear regression: $y_i = Normal(\mu_i, \sigma)$

Binomial logistic regression: $y_i = binomial(N = 1, p)$

$= bernoulli(p)$

Bernoulli distribution is a special case of the binomial distribution with N = 1!

Straight lines are (not) enough

Heinrich Heine
Universität Düsseldorf

Link functions

Linear regression: $I(\beta_0 + b_1 * x_i)$

$$\downarrow$$

$$y_i = Normal(\mu_i, \sigma)$$

Logistic regression: $logistic(\beta_0 + b_1 * x_i)$

$$\downarrow$$

$$y_i = Bernoulli(p)$$

Logistic regression

Example: Dative alternation (from Baayen 2008)

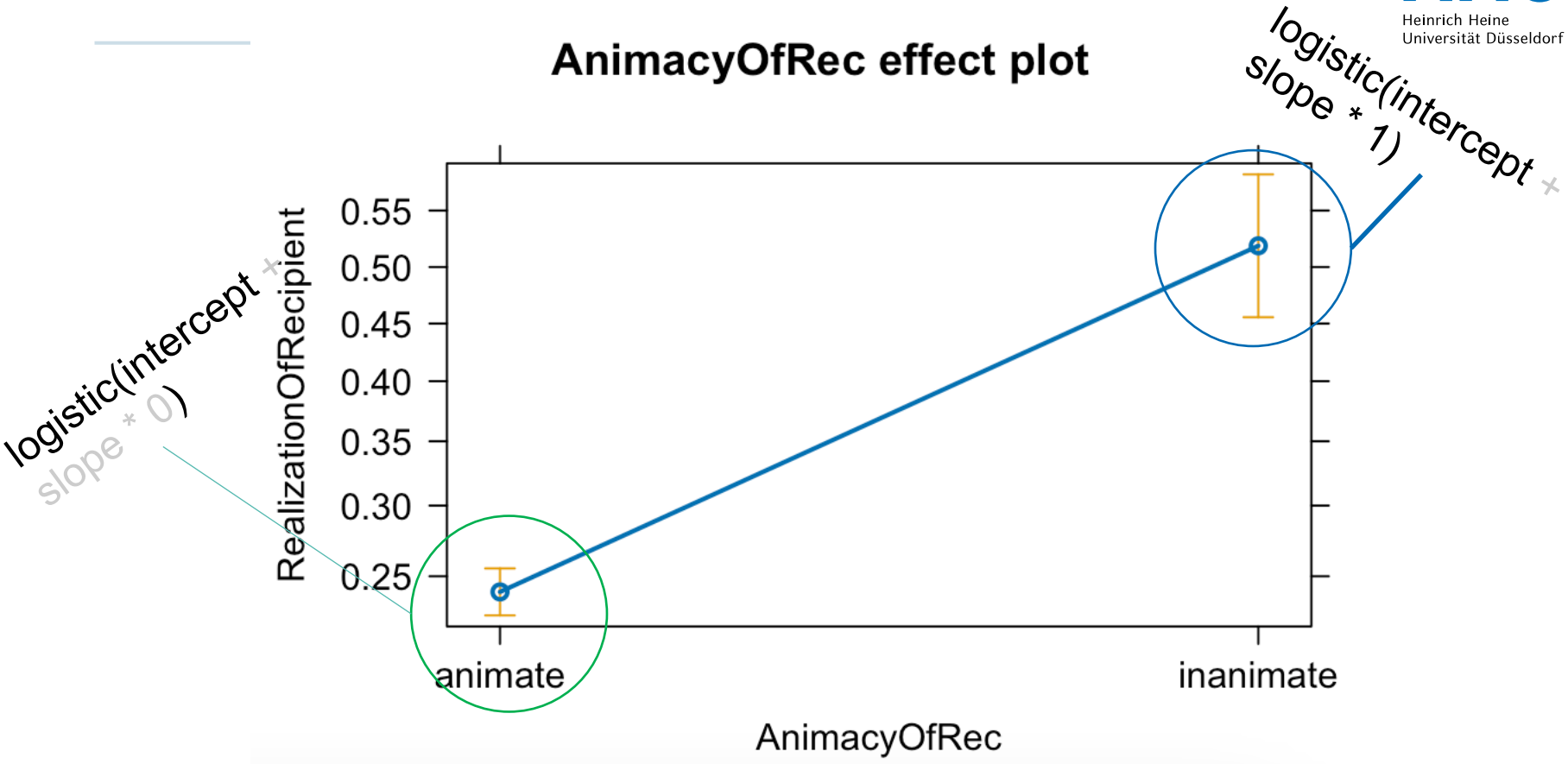
- e.g. *I gave her the book* (double-object construction) vs. *I gave the book to her* (PP construction)
- Response variable: RealizationOfRecipient – NP vs. PP
- Predictor variable: AnimacyOfRecipient – animate vs. inanimate

Interpreting coefficients

```
glm(formula = RealizationOfRecipient ~ AnimacyOfRec, family = "binomial",
    data = d)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.15406    0.04259 -27.094   <2e-16 ***
AnimacyOfRecinanimate  1.22941    0.13629   9.021   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3741.1  on 3262  degrees of freedom
Residual deviance: 3662.2  on 3261  degrees of freedom
AIC: 3666.2

Number of Fisher Scoring iterations: 4
```

log odds of observing PP when AnimacyOfRec is at its base level (animate)

logistic(-1.15406) = 0.2397 (in R. `plogis()`) → prob. of observing prep. dative when AnimacyOfRec is at its base level ≈ 25%


AnimacyOfRec effect plot
logistic(intercept + slope * 0)
logistic(intercept + slope * 1)
RealizationOfRecipient
0.55
0.50
0.45
0.40
0.35
0.30
0.25
animate
inanimate
AnimacyOfRec

Other models to look into

Heinrich Heine
Universität Düsseldorf

- Logistic regression models have become an important tool in linguistics, but for some data types, other link functions work better
- we work a lot with count data – as such, it can be useful to take a look at model types that are ideal for working with count data, especially Poisson regression and negative binomial regression.
- Generalized Additive Mixed Models (GAMM) as another valuable way of investigating non-linear relationships between predictors and response variables (see e.g. Fabian Tomaschek's introduction here https://osf.io/k98c4/)

Course overview

Heinrich Heine
Universität Düsseldorf

- Part 1: Basics of regression modelling
- Part 2: Generalized linear models
- **Part 3: Mixed-effects regression models**


GLM
GLMM

From GLM to GLMM

Motivation

- A key assumption in (G)LMs – and in most statistical tests – is that the datapoints are **independent**

- Common violations of the assumption of independence include:

  - multiple datapoints from the same participant in experimental studies

  - multiple datapoints from the same author and/or text in corpus studies

From GLM to GLMM

Motivation

- violating the independence assumption massively increases the danger of type I errors (rejecting a null hypothesis that is actually true in the population)

- Possible solutions for dealing with non-independence:
  - using averages (so that each participant only contributes one datapoint) – but this means losing information!
  - mixed models

Fixed vs. random effects

Heinrich Heine
Universität Düsseldorf

fixed effects	random effects
repeatable (in principle)	not repeatable
systematic influence	unsystematic influence
all relevant factor levels taken into account	sample of the population, perhaps not all factor levels taken into account (e.g. we cannot "sample" all inhabitants of Tilburg)
of interest	often not of interest

Fixed vs. random effects

Heinrich Heine
Universität Düsseldorf

- Examples for random effects:
  - participants (in repeated-measure designs)
  - items
  - corpus texts or their authors
  - language families (in typological studies)
- The line between fixed and random effects cannot always be drawn clearly.
- Random effects are **necessarily categorical**: "the whole point of fitting a mixed model is to account for dependent clusters of dara points that somehow group together" (Winter 2020: 236)

Random intercepts and random slopes

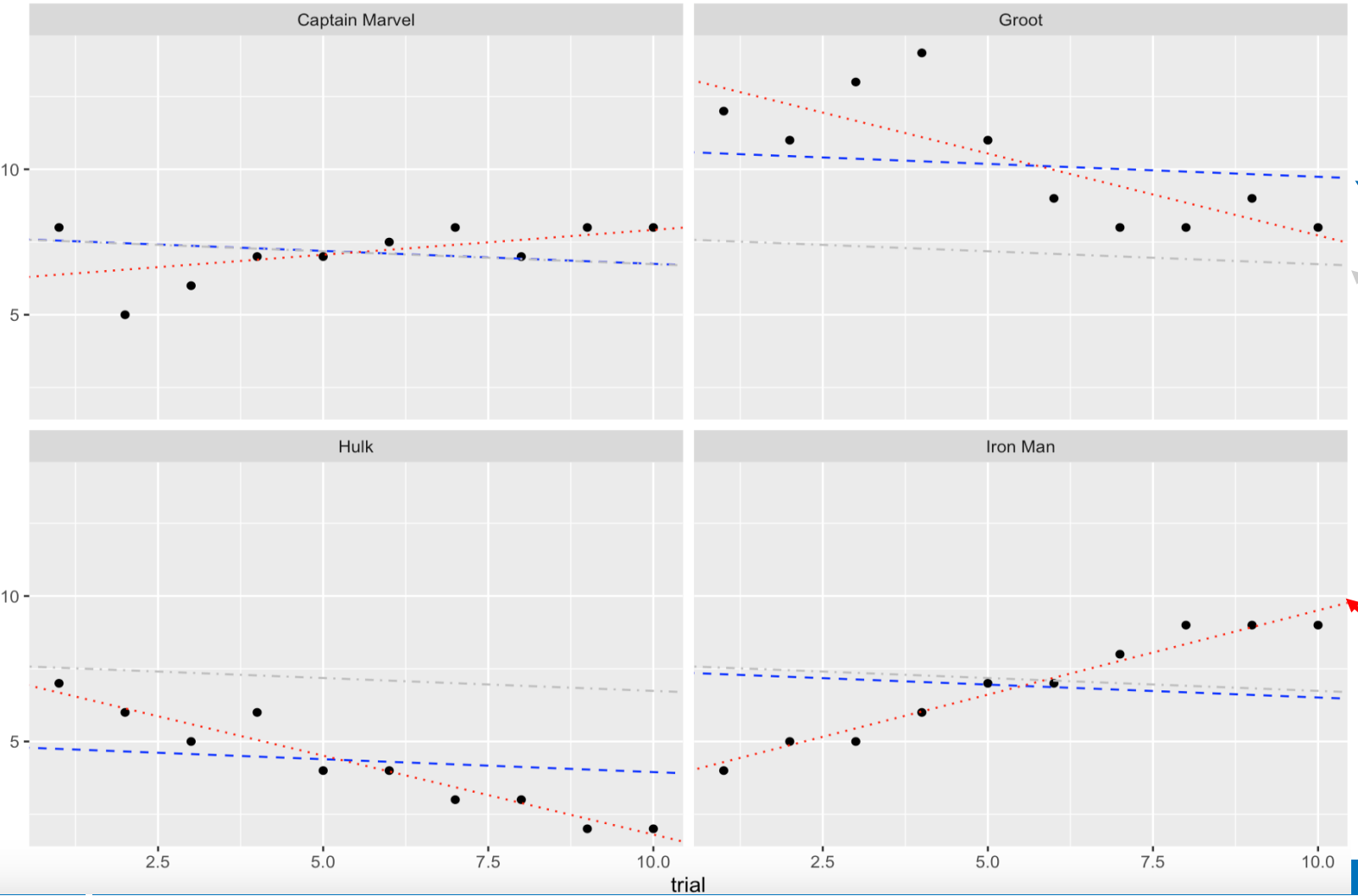
- Random effects can be added to a model via varying intercepts and/or varying slopes

- Varying intercept model:

$$y = \beta_{0j} + \beta_i * trial + \varepsilon$$

- Varying intercept and varying slopes model:

$$y = \beta_{0j} + \beta_{ij} * trial + \varepsilon$$



Heinrich Heine
Universität Düsseldorf

random intercepts

no random intercepts / slopes

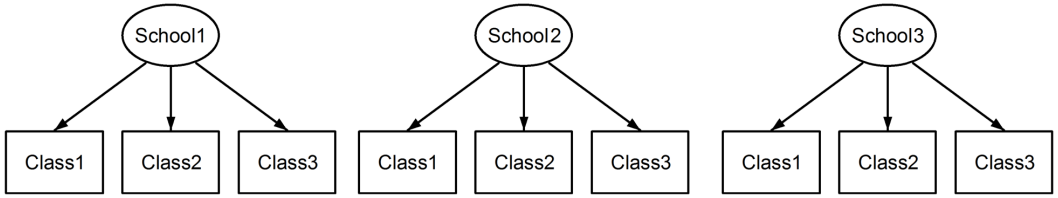
random intercepts & random slopes

Random effects

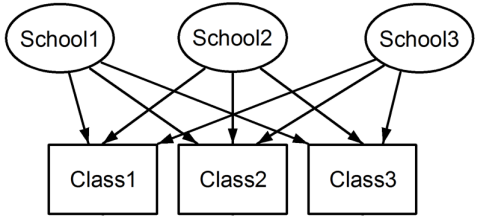
Nested and crossed random effects

- data points share multiple taxomomically or hierarchically organized characteristics → **nested** random effects

lme4: `(1 | class) + (1 | class : school)`



- factor levels appear in more than one level of the upper factor → **crossed** random effects



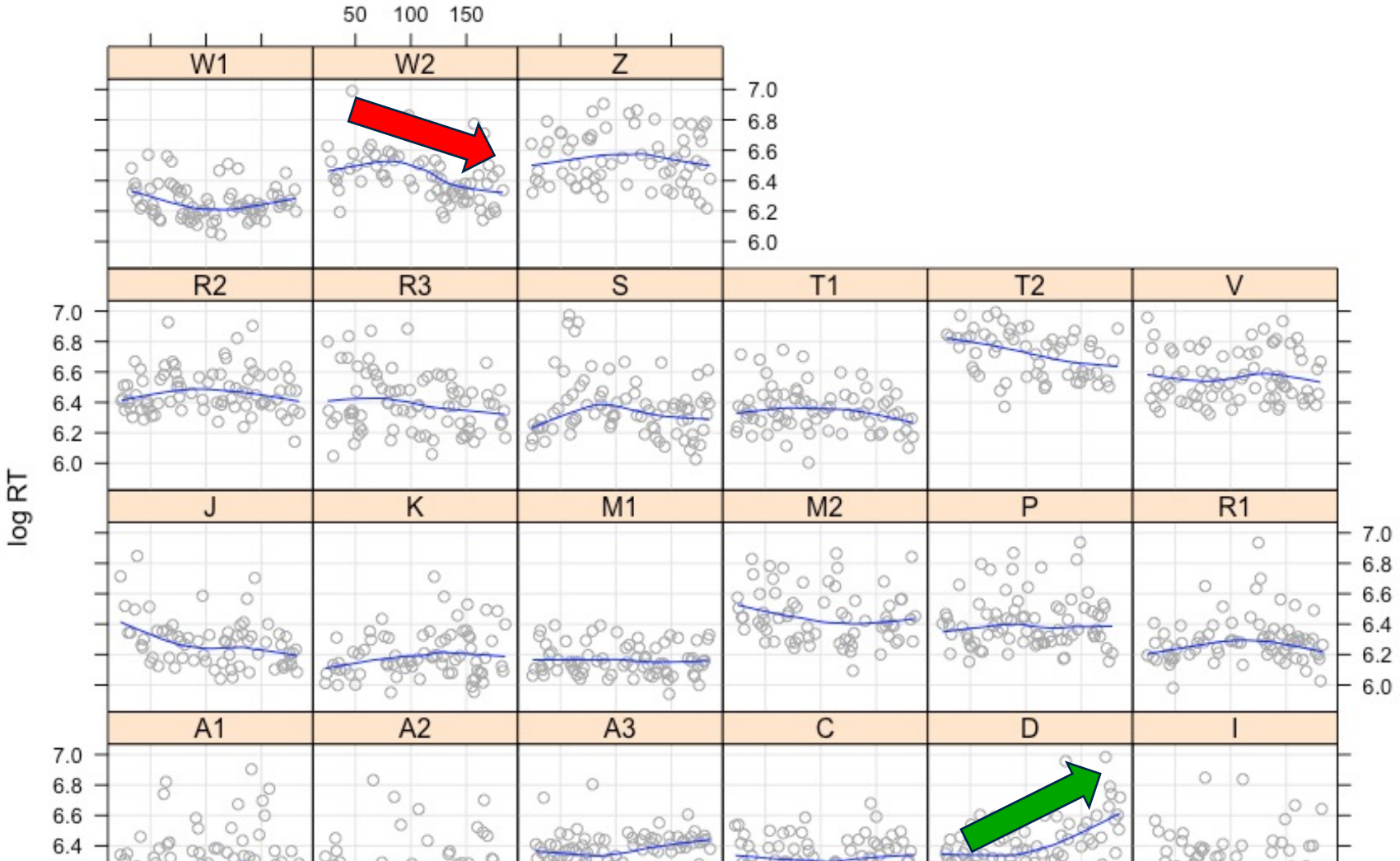
lme4: `(1 | class) + (1 | school)`

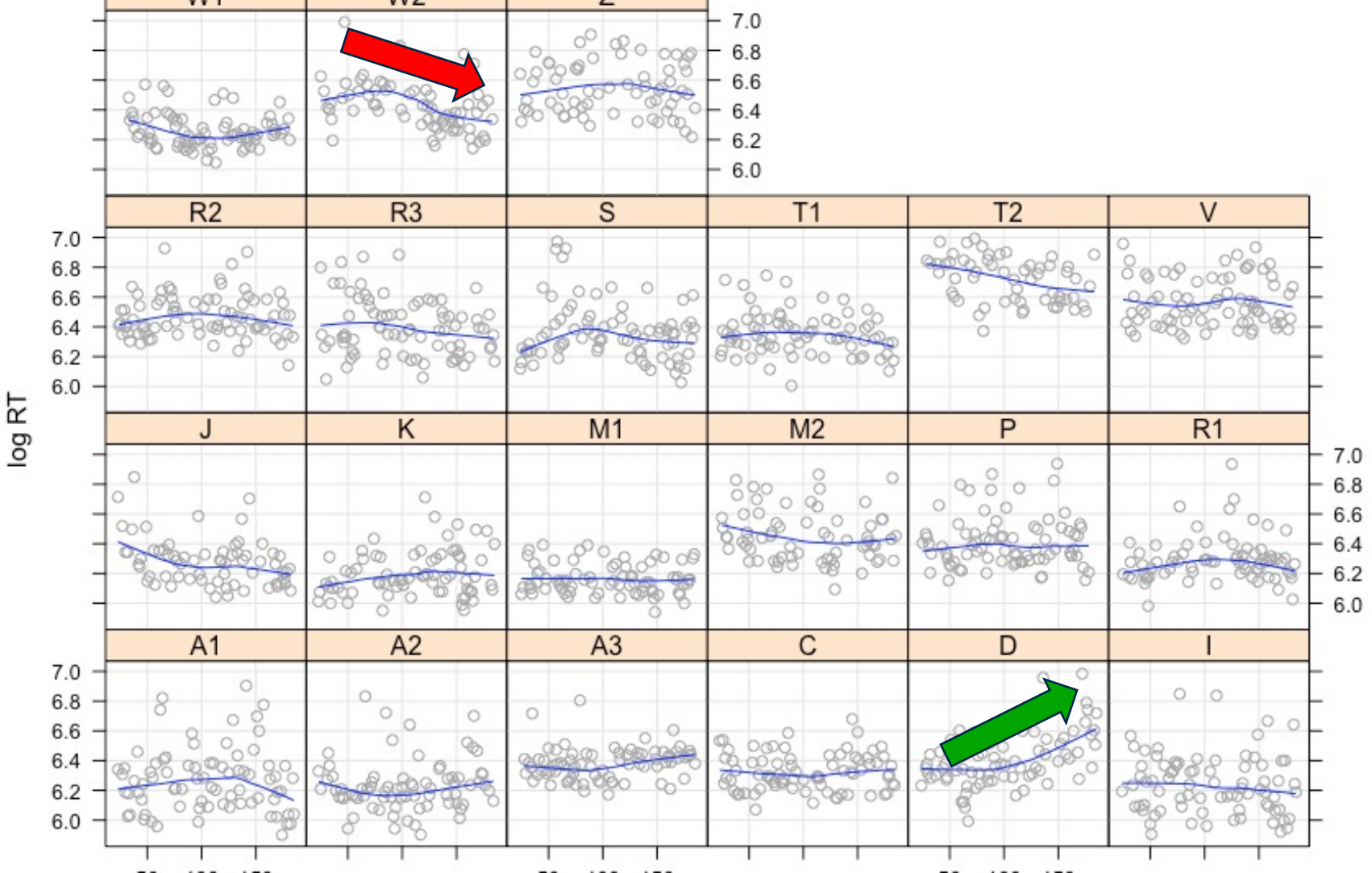
(Gries 2021; https://stats.stackexchange.com/questions/228800/crossed-vs-nested-random-effects-how-do-they-differ-and-how-are-they-specified)

Random intercepts and random slopes

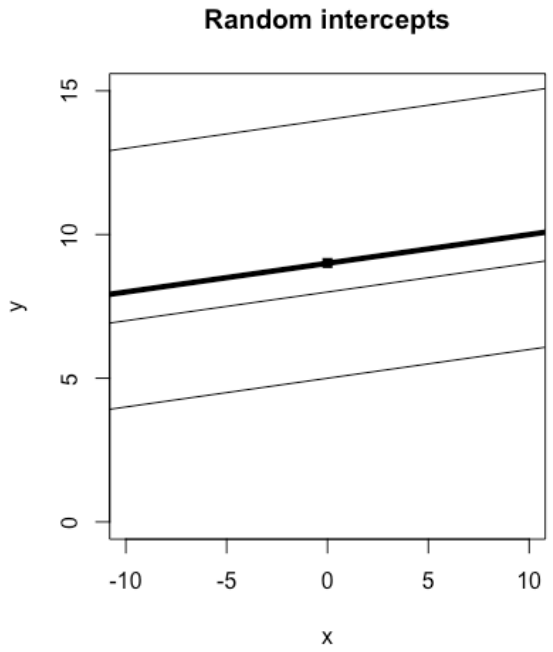
- **Random intercept model:** the fixed effect is evaluated against an error term that captures subject- or item-specific variability in the response

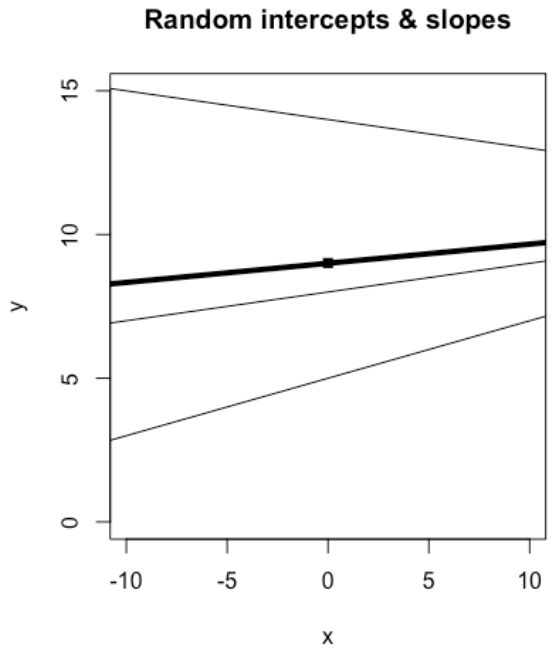
- **Random slope model:** the fixed effect is evaluated against an error term that captures subject- or item-specific variability in how the fixed effect affects the response


50 100 150
W1
W2
Z
R2
R3
S
T1
T2
V
J
K
M1
M2
P
R1
A1
A2
A3
C
D
I
7.0
6.8
6.6
6.4
6.2
6.0
log RT


W1
W2
Z
R2
R3
S
T1
T2
V
J
K
M1
M2
P
R1
A1
A2
A3
C
D
I
log RT
7.0
6.8
6.6
6.4
6.2
6.0

Random intercepts vs. random slopes





Example analysis

- from Baayen (2008): Reaction times ~ Word frequencies

```
lme4::lmer(RT ~ Frequency.c +
             (1 | Subject) +
             (1 | Word),
           data = lexdec3,
           REML = F)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod
Formula: RT ~ Frequency.c + (1 | Subject) + (1 | Word)
   Data: lexdec3

     AIC      BIC   logLik deviance df.resid
 -1300.1  -1273.3    655.0  -1310.1     1552

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.3061 -0.6712 -0.1020  0.5295  4.4149

Random effects:
 Groups   Name        Variance Std.Dev.
 Word     (Intercept) 0.002229 0.04721
 Subject  (Intercept) 0.017771 0.13331
 Residual             0.022636 0.15045
Number of obs: 1557, groups:  Word, 79; Subject, 21
```

general measures of model fit

Random effects structure

```
Random effects:
 Groups   Name        Variance Std.Dev.
 Word     (Intercept) 0.002229 0.04721
 Subject  (Intercept) 0.017771 0.13331
 Residual             0.022636 0.15045
Number of obs: 1557, groups:  Word, 79; Subject, 21
```

Fixed effects (as in typical linear models)

```
Fixed effects:
             Estimate Std. Error t value
(Intercept)  6.372703   0.029819  213.71
Frequency.c -0.038017   0.005158   -7.37

Correlation of Fixed Effects:
            (Intr)
Frequency.c 0.005
```

R^2 for mixed models

Heinrich Heine
Universität Düsseldorf

- package MuMIn offers options for calculating R^2 for mixed models based on indices proposed by Nakagawa & Schielzeth (2013)
  - marginal R^2: characterizes the variance described by the fixed effects
  - conditional R^2: describes the variance described by both fixed and random effects

Problems and solutions

Convergence issues

- Mixed models sometimes fail to converge
- This can happen when trying to fit an overly complex model to a fairly sparse dataset
- Centering and/or scaling continuous variables sometimes helps avoid convergence issues
- using a different optimizer can also help, or using the all_fit() function in the afex() package to try out different optimization algorithms.
- For "small *n* large *p* problems" it can also be helpful to look into Classification and Regression Trees (CART) and Random Forests (see e.g. Tagliamonte & Baayen 2012, Levshina 2015)

Problems and solutions

How to choose the "right" model

- As mentioned before, there is no such thing as a single right model for a given dataset.
- This is a big advantage because it gives us a lot of freedom in toying with different models...
- At the same time, it is a challenge because we need convince people (incl. reviewers) that our modeling choices make sense ☺

Problems and solutions

How to choose the "right" model

- Different strategies in finding the "best" model, e.g.
  - backward selection: putting all potentially relevant/influential variables into the model, checking which ones significantly improve the model fit, and only keeping the influential ones in the final model
  - forward selection: starting with an empty model and iteratively adding predictors

- Arguably, however, it is better to come up with a clear "data story" first (e.g. using DAGs), then fitting one model and computing model diagnostics (incl. comparing it with a null model).

References

- Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.
- Field, Andy, Jeremy Miles, & Zoë Field. 2012. *Discovering Statistics Using R*. Los Angeles: Sage.
- Levshina, Natalia. 2015. *How to do linguistics with R. Data exploration and statistical analysis*. Amsterdam, Philadelphia: John Benjamins.
- McElreath, Richard. 2020. *Statistical Rethinking. A Bayesian Course with R and Stan*. 2nd edn. Boca Raton: CRC Press.
- Tagliamonte, Sali A. & R. Harald Baayen. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24(02). 135–178. https://doi.org/10.1017/S0954394512000129.
- Winter, Bodo. 2020. *Statistics for linguists: an introduction using R*. New York: Routledge.