DRS Spring School – week 2

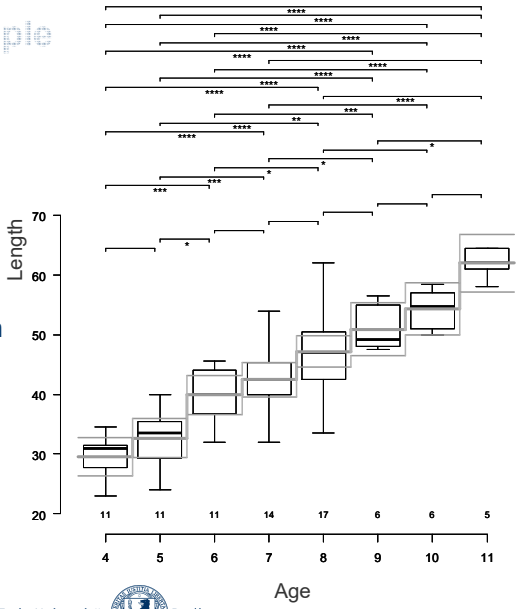
Regression

PD Dr. Lorenz Gygax (HU Berlin)

A motivating example

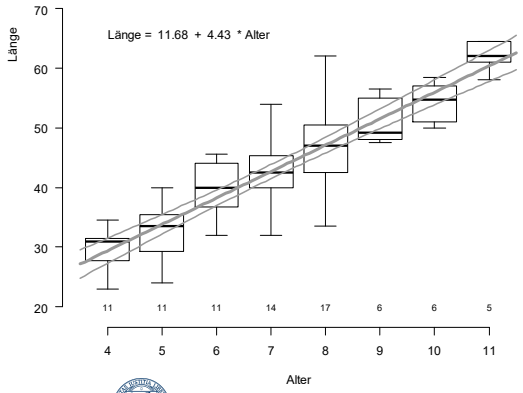
- Eels:
  Body length in dependence of age
- Pair-wise comparisons?
  (Tukey pair-wise post-hoc comparison)
- Estimate?

→ Some different approach might be useful



Regression analysis

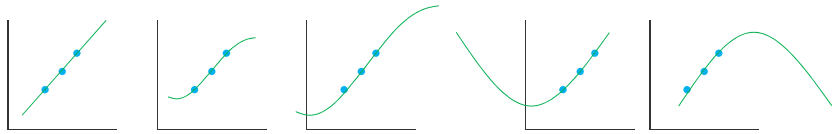
- Dependence of an outcome variable on one (or several) continuous explanatory variables
- Pragmatically:
  Slope ≠ 0 is more easily supported than a difference between groups
- Fit?

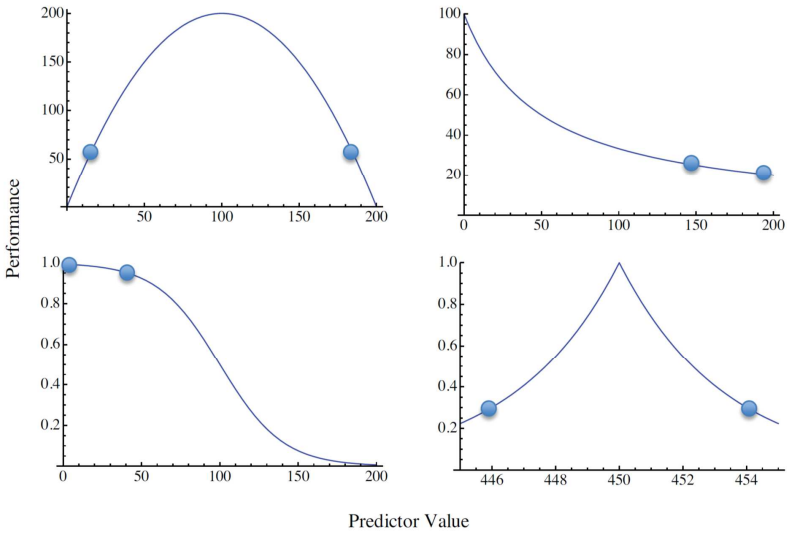


«(non-)categorical» thinking

- Very young eels: length = 0, length < 0?
- Very old eels: length → ∞?

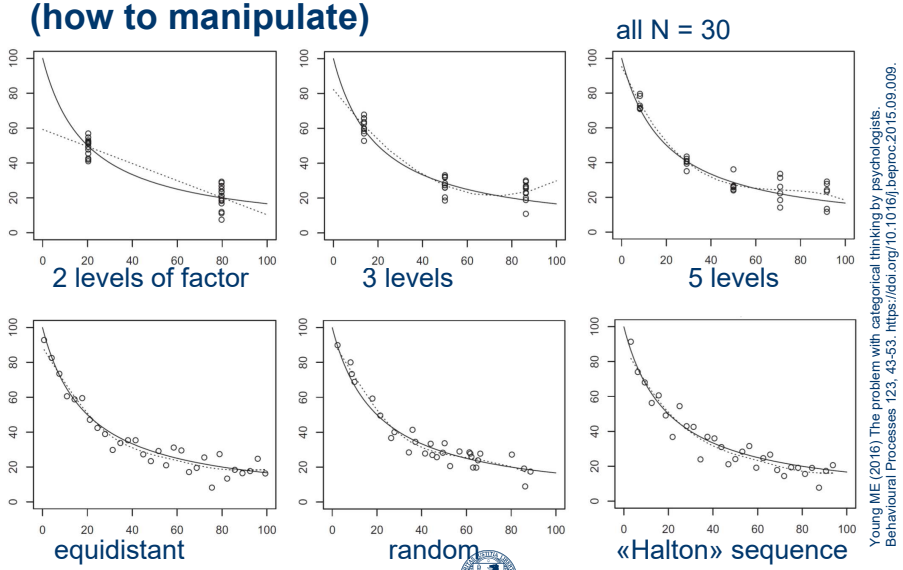
- Is there some flattening, what is the shape of the relationship?





Young ME (2016) The problem with categorical thinking by psychologists. Behavioural Processes 123, 43-53. https://doi.org/10.1016/j.beproc.2015.09.009.

Choice of values for explanatory variable (how to manipulate)



Young ME (2016) The problem with categorical thinking by psychologists. Behavioural Processes 123, 43-53. https://doi.org/10.1016/j.beproc.2015.09.009.

Choice of values for explanatory variable (how to manipulate)

- Works for all «ordinal» scales with the same principle
- Works with repeated measurements, too, e.g.:
  - Each subject receives a low, moderate and a high dosage of treatment substance (at different points in time)
  - Each subject receives different dosages

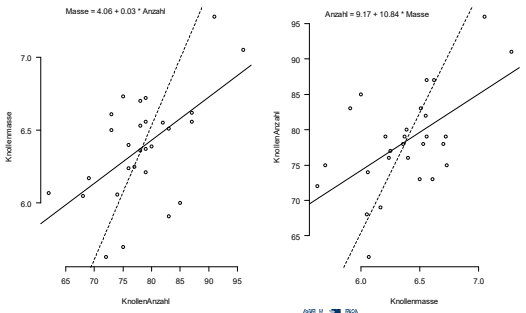
→ Whenever possible choose a continuous explanatory variable with varying values

→ Model this data without bias

↔ What is common in a research field

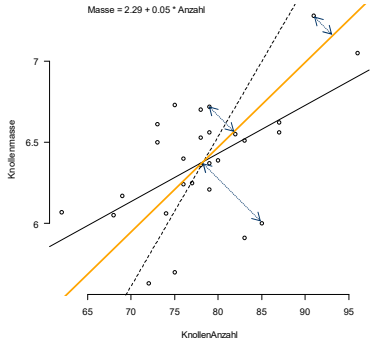
Correlation versus regression

- No differentiation of outcome and explanatory variables: the relationship is symmetrical
- With a correlation, do not draw a an estimated regression line
- Asymmetry in regression:



Correlation versus regression

- Exception: Special cast of
  orthogonal regression (total least square)
  (rarely used)



Correlation versus regression

→ For a regression (in common usage) you need to define:
- one (or several) explanatory variables («causes»)
- an outcome variable («effect»)

- BUT take care: this only sounds like causality
- Note:
  The experimental design and not the statistical method allows (at best) a causal interpretation
- Even a regression shows only a relationship in a non-experimental study (cp. Correlation) and at best a presumed causality

Simple regression

- Summary;
  you have seen everything on the slides before already!

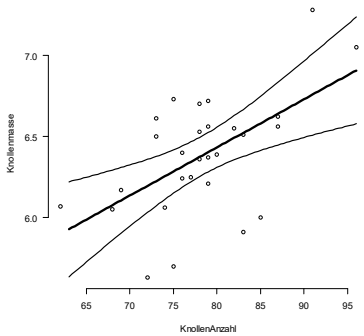
- In principle, straight lines and their slopes are estimated
  (↔ below / transformations)
- Is the pattern / relationship linear in at least the observed interval?
  (see also checks of assumptions)

- The explanatory variables need (should be) measured without error (cp. «errors in variables»)

Simple regression

- Model contains (usually):
  H_0: slope = 0
  H_A: slope ≠ 0

- Test often based on an
  F-test, here:
  slope = 0.030 (relevant?)
  $F_{1,26} = 12.31$, $p = 0.002$



- 95% confidence interval: in 95 of 100 repetitions of this experiment, the slope is expected to be within this interval.

Multiple regression

- (Hyper-)surfaces are estimated, e.g.

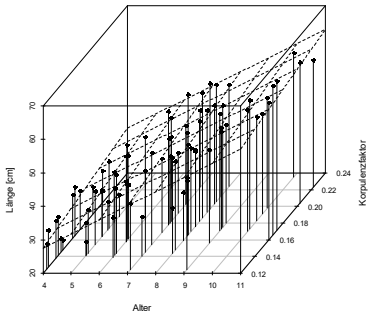
```
aal.2way.lm <- lm (Laenge_cm ~ Alter_Otholiten + Korpulenzfaktor)
```

- Estimate:

```
summary (aal.2way.lm)
[…]
Coefficients:
                  Estimate Std. Error …
(Intercept)         4.8040     4.0994 …
Alter_Otholiten     4.2207     0.3068 …
Korpulenzfaktor    48.4295    24.8047 …
[…]
```

- Model equation:

```
Länge = 4.80 + 4.22 * Alter + 48.42 * KF
```



Multiple regression

- 1. Global test
  Do all the explanatory variables together explain the outcome?

```
summary (aal.2way.lm)
[…]
Residual standard error: 5.223 on 78 degrees of freedom
Multiple R-squared:  0.7549,    Adjusted R-squared:  0.7486
F-statistic: 120.1 on 2 and 78 DF,  p-value: < 2.2e-16
[…]
```

  $F_{2,78} = 120.1$, $p < 0.0001$

- 2. specific tests: the influence of which single explanatory variable can be statistically supported?
  - complementing an existing model
  - when omitted from a «maximum» model

Multiple regression

- Complementing an existing model (Typ-I Sums-of-squares):
  Do additional variables generate additional information?

```
anova (aal.2way.lm)
[…]
Response: Laenge_cm
                     Df Sum Sq Mean Sq F value  Pr(>F)
aal.df$Alter_Otholiten  1 6448.9  6448.9 236.387 < 2e-16 ***
aal.df$Korpulenzfaktor  1  104.0   104.0   3.812 0.05448 .
Residuals              78 2127.9    27.3
[…]
```

- Omitting from a «maximum» model (Typ-III Sums-of-squares):
  Is information lost without a variable?

```
drop1 (aal.2way.lm)
[…]
Model: Laenge_cm ~ Alter_Otholiten + Korpulenzfaktor
                     Df Sum of Sq    RSS    AIC F value  Pr(>F)
<none>                           2127.9 270.74
aal.df$Alter_Otholiten  1      5163 7290.9 368.49 189.250 < 2e-16 ***
aal.df$Korpulenzfaktor  1       104 2231.9 272.61   3.812 0.05448 .
[…]
```


Raw versus standardized slope

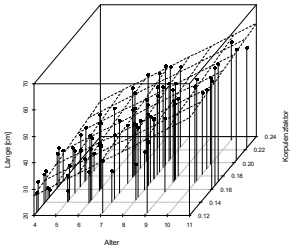
- Model equation:

```
Länge = 4.80 + 4.22 * Alter + 48.42 * KF
```

- Which influence is more important (relevance for the subject field)?

→ Normalise the explanatory variablen
$(X_i - \bar{x})/sd(x)$

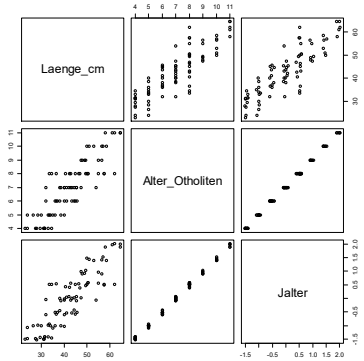
→ All variables have a comparable spread:
a standard deviation of 1 (and a mean of 0)




```
summary (aal.2wayNorm.lm)
[…]
Coefficients:
                  Estimate Std. Error …
(Intercept)        42.7235     0.5803 …
Alter_OtholitenN    8.5592     0.6222 …
KorpulenzfaktorN    1.2148     0.6222 …
[…]
```

Collinearity

- Assumption:
  Alternative determination of age
  Which way to determine age is better?



Collinearity

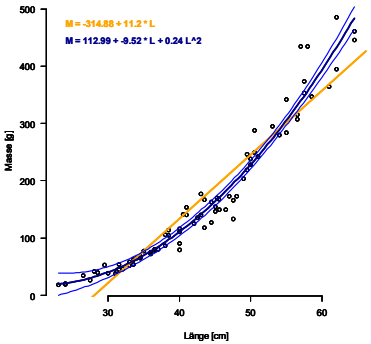
	lm (Laenge_cm ~ Alter_Otholiten + Jalter)
Slope Otholitenalter	-0.86 ± 5.52
Slope alternative method	10.77 ± 11.24
P-value Otholitenalter (drop1)	$F_{1,78}$ = 0.0240 p = 0.88
P-Wert alternative method (drop1)	$F_{1,78}$ = 0.9178 p = 0.34

Collinearity

- Problem:
  This is true for all «correlated» explanatroy variables
  (even if they have different data types)
- Occurs more often in observational (epidemiological) studies
- Importance that all combinations of values on the explanatory variables are observed (planning!): balanced data-set without confounding variables

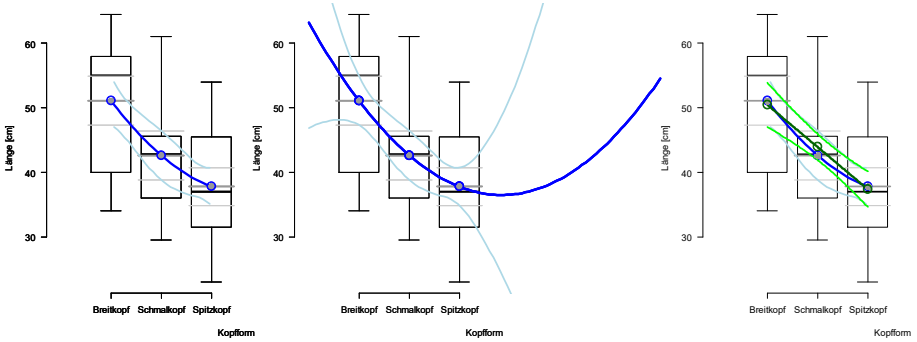
Non-linearity, polynomials

- No straight line
  (in observed range)



Polynomials & and ordered factors

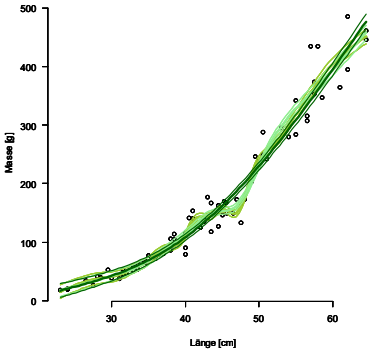
- ordered factor with k levels ≙ polynomial with degree (k-1)
- here: $Y = a + b_1 * X1 + b_2 * (X1)^2$



- No extrapolation while using polynomials!
- Polynomials can be simplified: omit higher degrees

Non-linearity: splines

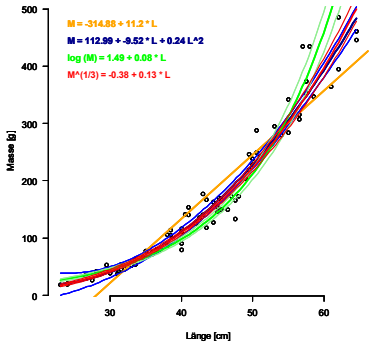
- Smooth, but «unrestricted» curve
- Decision on number of «turning points»
- 15, 9, 3



- Approach implemented in a generalised way in so called GAMs (Generalized additive models)

Non-linearity: transformations

- Logarithm?
- Allometry?



Extensions:

- ***Robust*** approaches
  - when data-/residuals not fully «normal»
  - e.g. long tails in distribution
  - some (rare) outliers
- ***Non-parametric*** approaches
  - similar, distribution is estimated from the data
  - e.g. re-sampling procedures (bootstrap, jack-knife)

→ (at the moment:) laboriuous
→ Limited complexity of the models

Collinearity

	`lm (Laenge_cm ~ Alter_Otholiten + Jalter)`	`lm (Laenge_cm ~ Alter_Otholiten)`	`lm (Laenge_cm ~ Jalter)`
Slope Otholitenalter	-0.86 ± 5.52	4.43 ± 0.29	
Slope alternative method	10.77 ± 11.24		9.03 ± 0.59
P-value Otholitenalter (drop1)	$F_{1,78} = 0.0240$ $p = 0.88$	$F_{1,79} = 228.26$ $p < 0.0001$	
P-Wert alternative method (drop1)	$F_{1,78} = 0.9178$ $p = 0.34$		$F_{1,79} = 231.78$ $p < 0.0001$