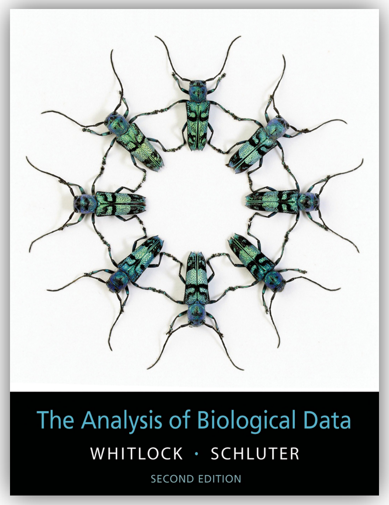
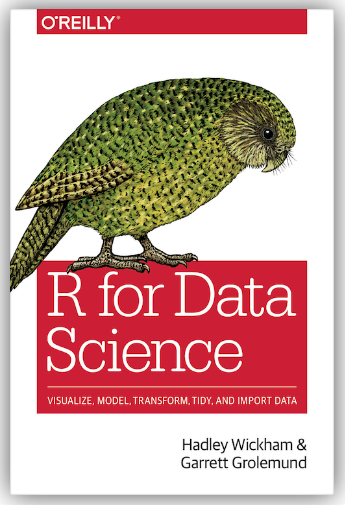
Data Science in Bioinformatics

Palle Villesen & Thomas Bataillon





Outline for week 12

- Chapter 18
  - Mole rats & ANCOVA
  - Exercises wrap up dn/ds and & expression &…
- Thursday session is CANCELLED
- The final assignment is released

Body mass in lazy Mole rats

- Two castes

  Workers

  Infrequent workers (LAZY)

Regression

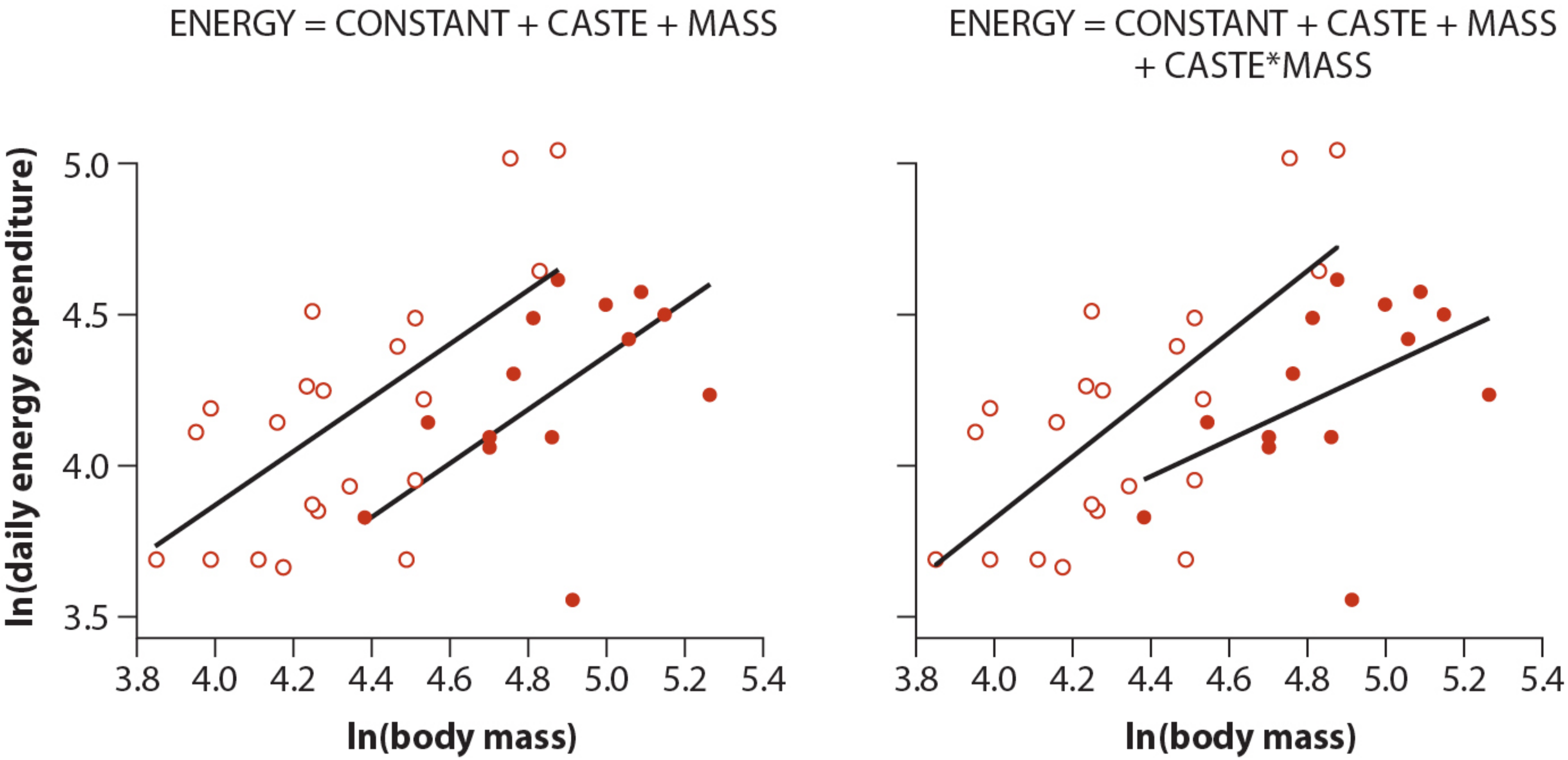
Y ~ Covariate

ANOVA

Y ~ Factor

Ancova model

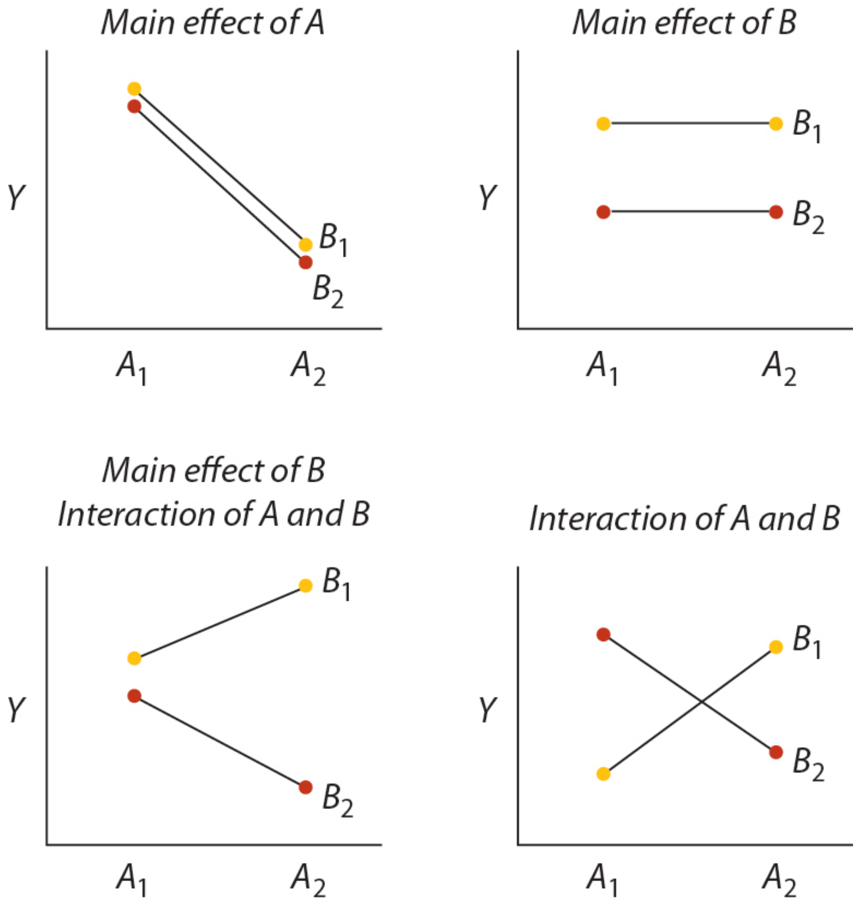
Y ~ Covariate + Factor



ENERGY = CONSTANT + CASTE + MASS

ENERGY = CONSTANT + CASTE + MASS + CASTE*MASS

Models with/ without interaction

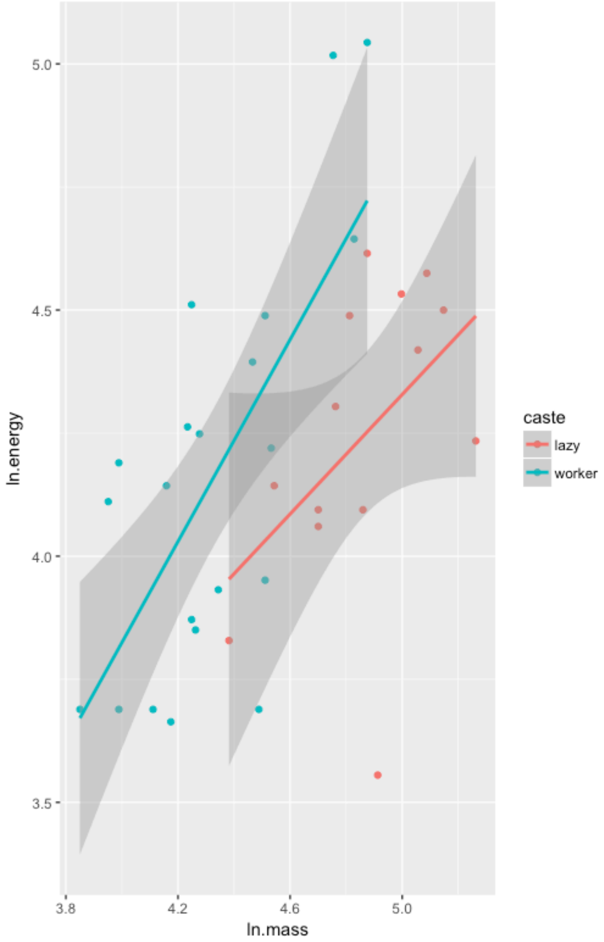


Main effect of A

Main effect of B

Main effect of B
Interaction of A and B

Interaction of A and B



Issues when fitting models

Simple models

Easy to test

Easy to interpret

Easy to fool

(underlying extra covariates not included in the model)

Multiple factors models

How to test effects?

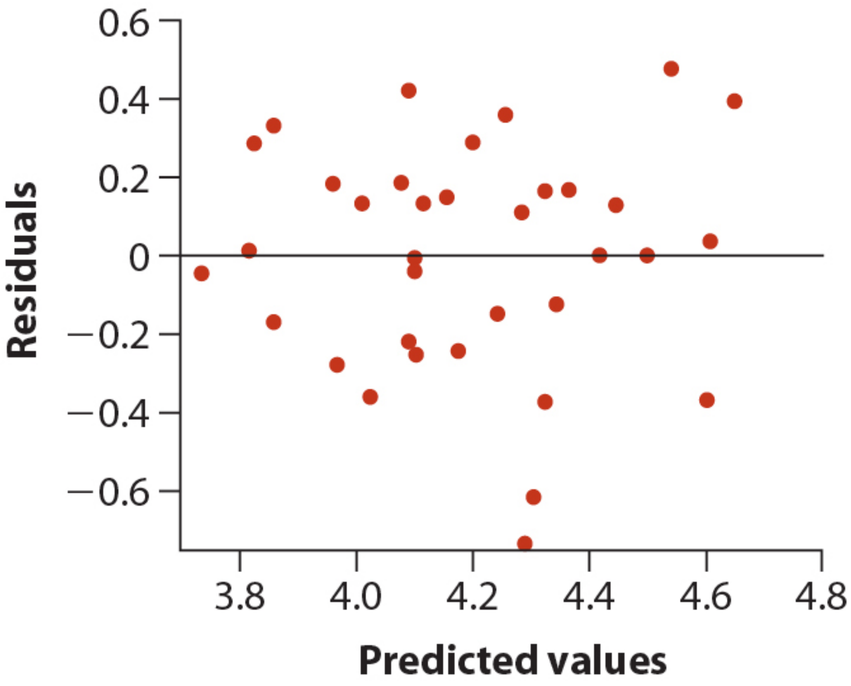
- type I, type II, type III

What model should I choose?

- Sequential
- Search all models
- Model selection
- Model Averaging

→ See Machine Learning in Bioinformatics

After fitting your model …



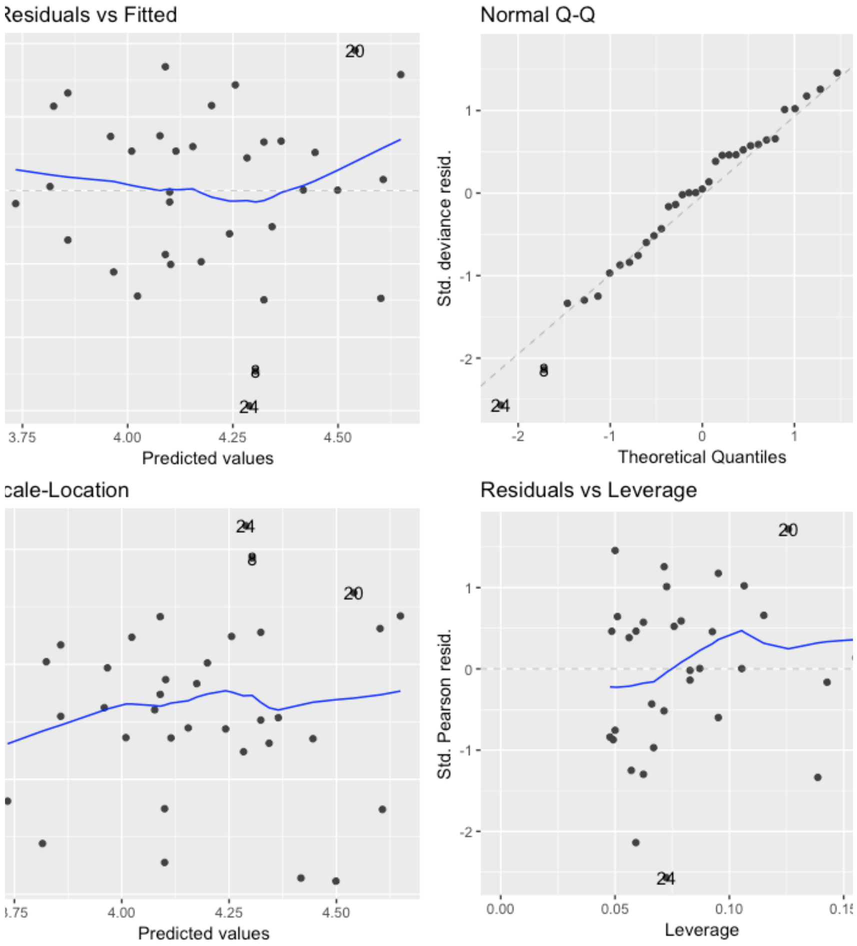
Check your model

fit1 = lm()

plot(fit1)

ggplot2::autoplot(fit)

… refit … repeat



esiduals vs Fitted

Normal Q-Q

cale-Location

Residuals vs Leverage

The last word ?

“The need for statisticians to reject the role of 'guardian of proven truth', and to resist attempts to provide once-for-all solutions and tidy over-unifications of the subject”

https://en.wikipedia.org/wiki/John_Tukey