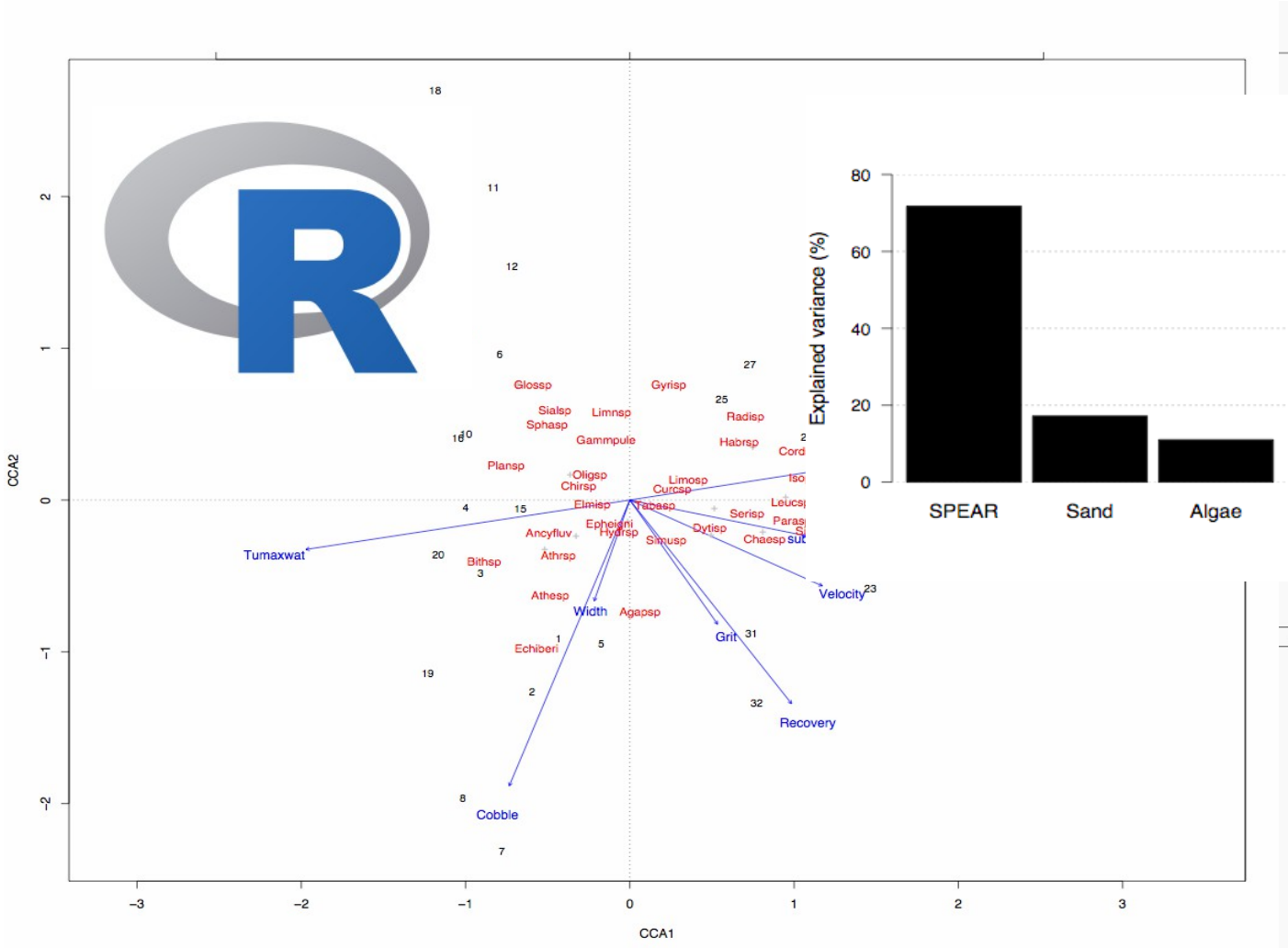
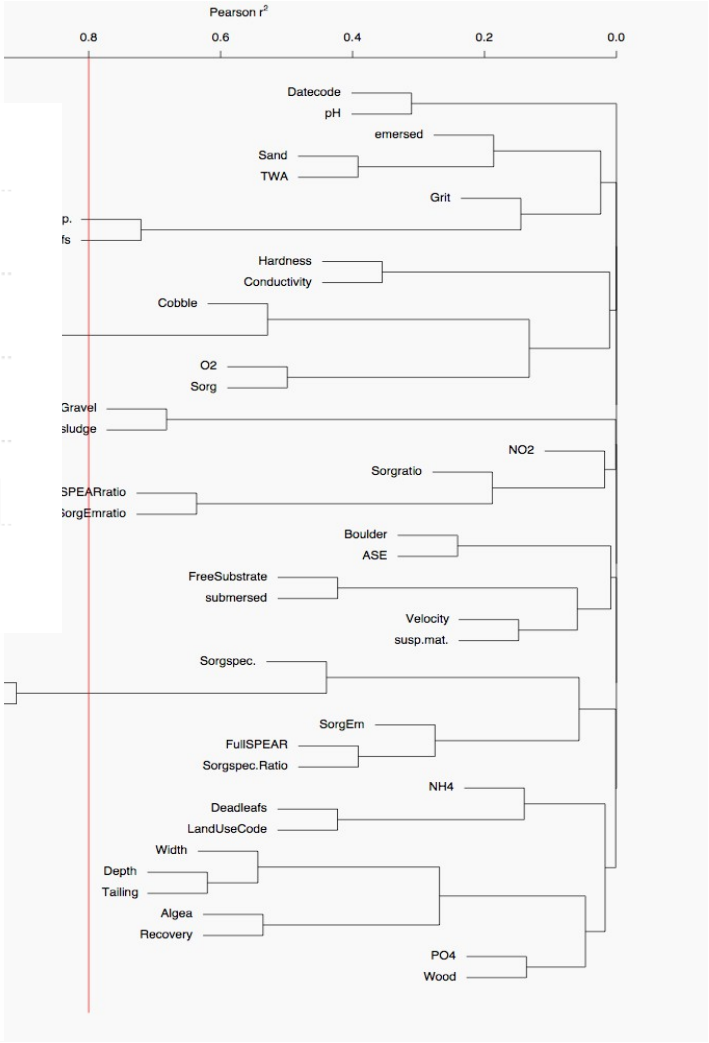
Applied Multivariate Statistics

University of Koblenz-Landau 2017/18





Ralf B. Schäfer

Short introduction

- Professor for Quantitative Landscape Ecology
- Current teaching: Statistics (M.Sc.); GIS (B.Sc./M.Sc.); Environmental Modelling (B.Sc./M.Sc.); Aquatic Ecotoxicology (M.Sc.); Environmental Philosophy (B.Sc.)
- Research focus:
  - Community ecology of freshwater invertebrates and microorganisms
  - Response of freshwater ecosystems to different (anthropogenic) stressors (e.g. pollution)
  - Trophic linkages between aquatic & terrestrial systems
- Primarily field studies/experiments and data analyses/ modelling

www.landscapecology.uni-landau.de

Organisation

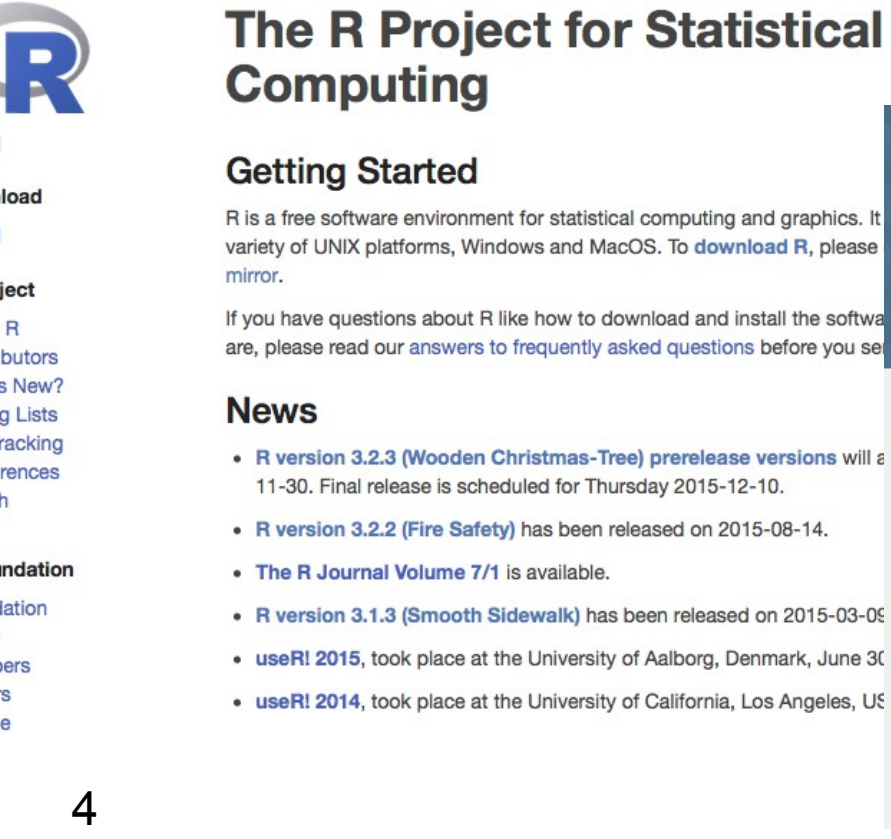
- Lecture material (including course schedule and literature list) can be found on github and website:
  https://github.com/rbslandau/statistics_multi
  https://goo.gl/EhPVFG

- Inverted classroom: Self study of lecture and demonstration, Q&A and exercises in class room

- Contact time: 2 hours per week; Own study time: approximately 1 day per week

Using your own notebook

- feel free to you use your own WLAN-enabled notebook!
- install R (http://mirrors.softliste.de/cran/) oder RStudio (recommended for beginners - http://www.rstudio.com/)
- Run “0_Install_packgs.R”, provided on github
- for installation of additional packages run install.packages(“package to be installed”)



Course objectives: Learning outcomes

- Classify, explain and interpret the different types of (multivariate) statistical approaches
- Select and apply the appropriate statistical method for the research goal
- Demonstrate moderate level of statistical modelling skills, including scripting in R

Two incorrect ways of thinking about stats

1. **Overconfidence:** Statistics is like mathematics and provides a single, correct answer
   But statistical thinking differs from mathematical thinking

2. **Disbelief:** Anything goes – statistics cannot be trusted
   But: statistics provide quantitative support of the complete research process

Tintle (2015) *Amer. Statist.* 69: 362

Statistical modelling, simulation and the linear model

Contents

Learning targets

- Explain the data analysis cycle and apply tools for exploratory data analysis
- Explain approaches to statistical modelling and simulation and apply simulation-based methods
- Diagnosing and interpreting the linear model

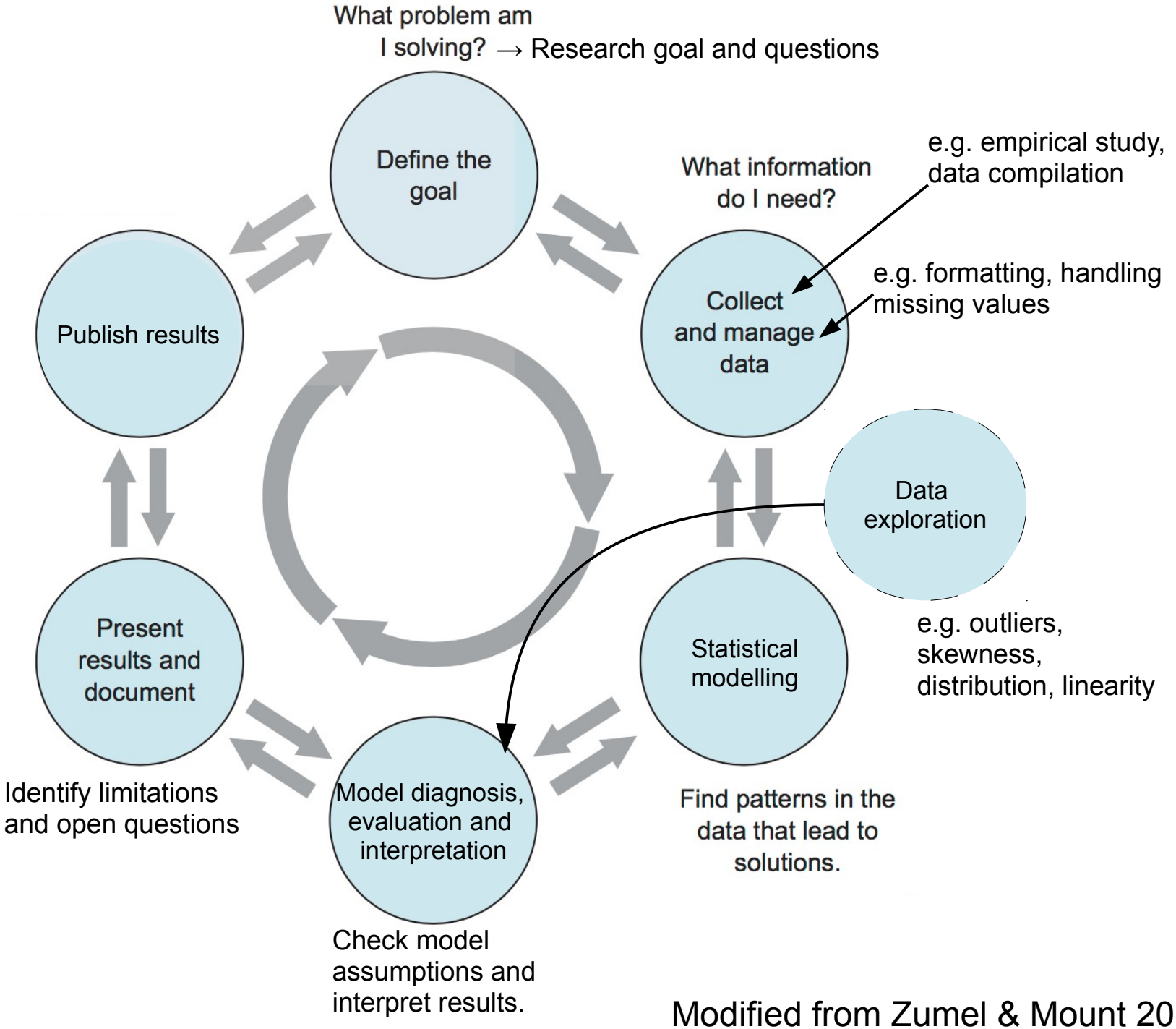
Learning targets and study questions

- Explain the data analysis cycle and apply tools for exploratory data analysis
  - Explain the steps of the data analysis cycle.
  - Summarise the elements of exploratory analysis. Which graphical tools are essential?
- Explain approaches to statistical modelling and simulation and apply simulation-based methods
  - Discuss the two different approaches to statistical modelling and links through simulation-based approaches.
  - Explain the purpose and critically discuss permutation tests.
  - Explain the purpose and critically discuss bootstrapping.
  - Explain the main idea of cross-validation and discuss the selection of *k* with respect to the bias-variance trade-off.

Learning targets and study questions

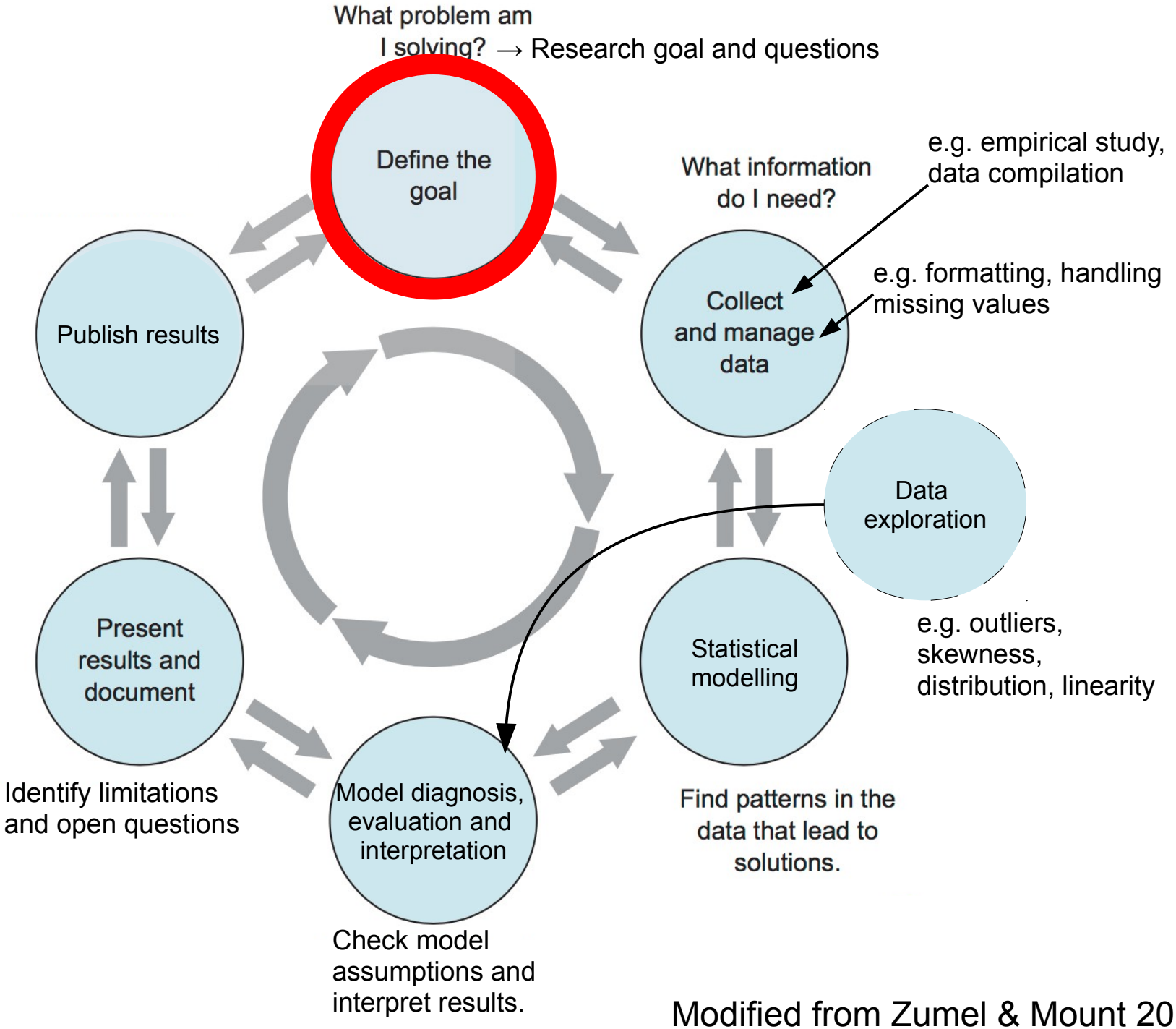
- Diagnosing and interpreting the linear model
  - Describe the assumptions of the linear regression and explain how they can be checked.
  - Which types of outliers exist? When is an outlier important?
  - Discuss the application of bootstrapping and cross-validation for the linear model.

Data analysis cycle



Modified from Zumel & Mount 2014: 6

Data analysis cycle



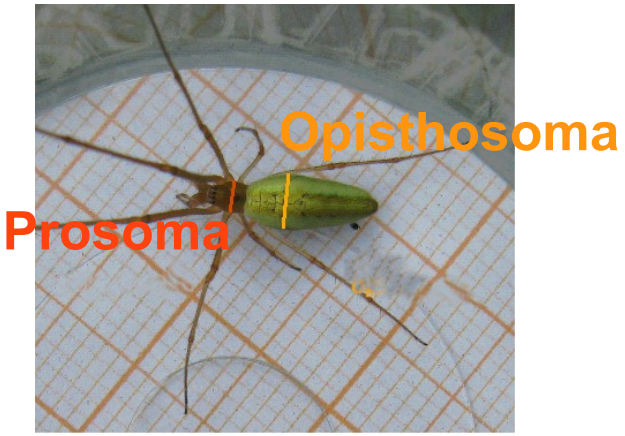
Modified from Zumel & Mount 2014: 6

Define research goal and question

- Research goals (e.g. prediction, estimation, inference) and questions should inform study design and methods
- Aim: Test scientific hypothesis → Formulate testable hypothesis

Example

Question: Does the body condition of riparian spiders differ between restored and non-restored stream stretches?

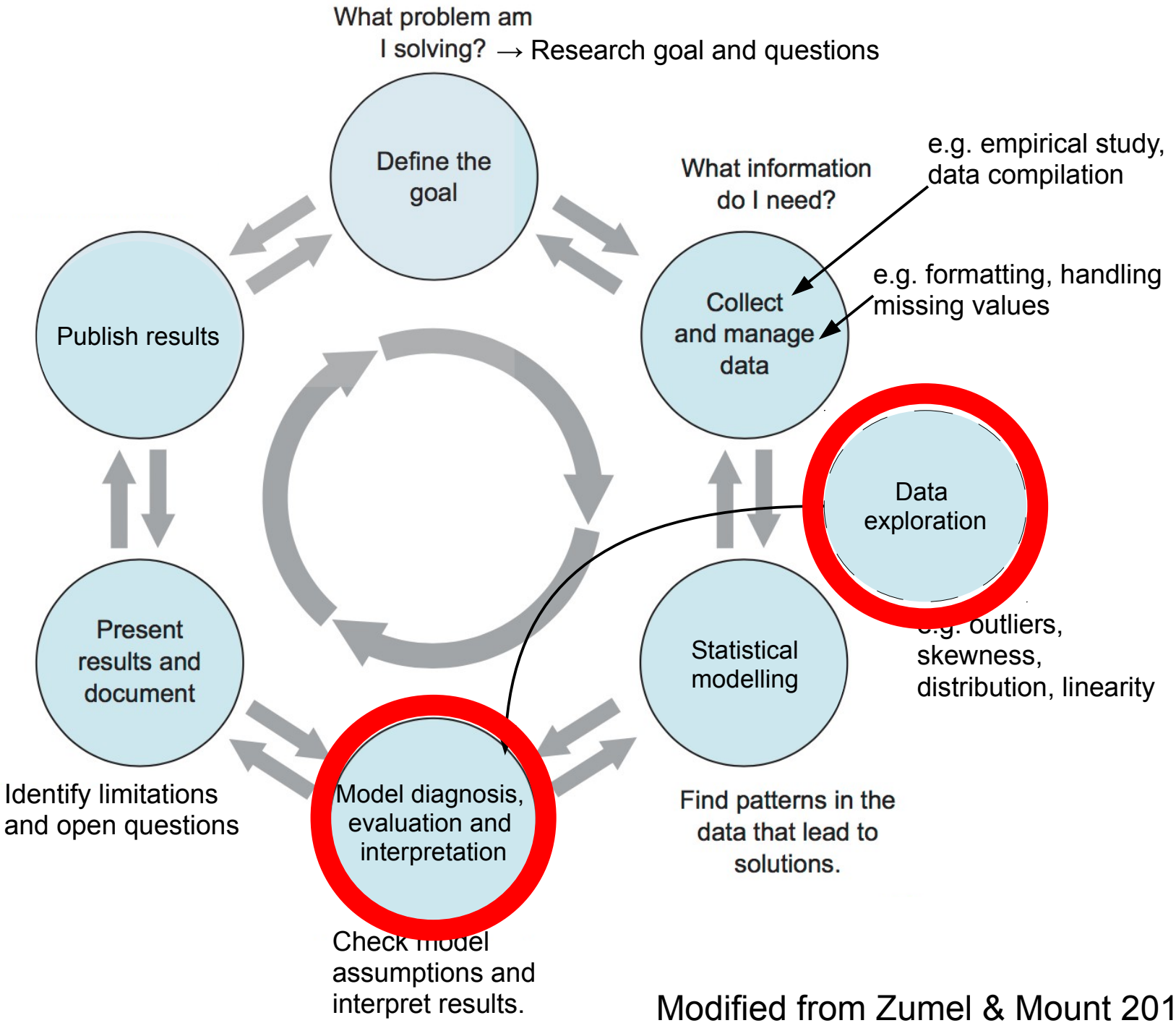


Scientific hypothesis: Restoring stream stretches alters aquatic communities, resulting in different emerging insects on which riparian spiders prey. This affects the spiders' body condition derived from prosomal (pr.) and opisthosomal (op.) width.

- Testable hypothesis: The sample means for the body condition are drawn from populations with the same μ:

$$H_0: \ \mu_{restored} = \mu_{non-restored} \qquad H_1: \ \mu_{restored} \neq \mu_{non-restored}$$

Data analysis cycle



Modified from Zumel & Mount 2014: 6

Tools for data exploration

- Useful for inspecting data before the modelling but also for model diagnosis
- Zuur et al. (2009) urge data inspection before modelling

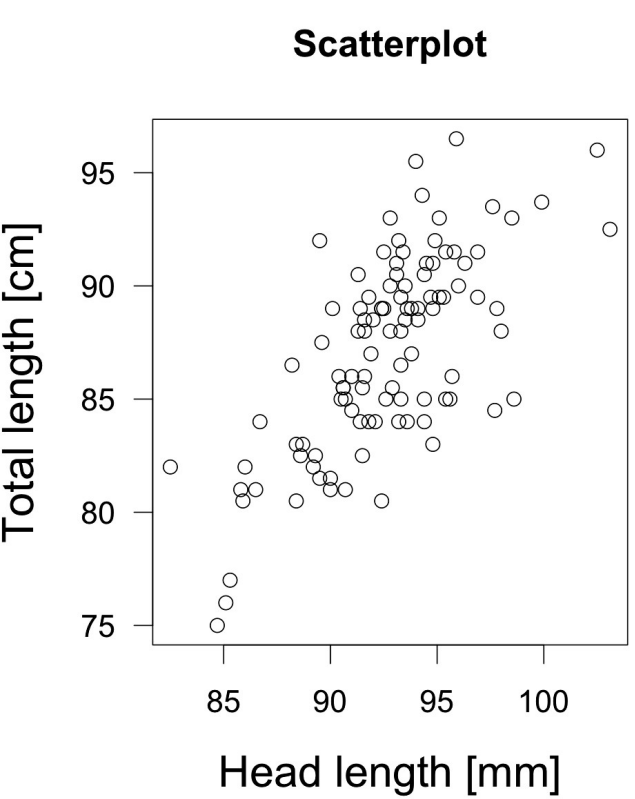
➡ **GIGA: Garbage in – Garbage out**

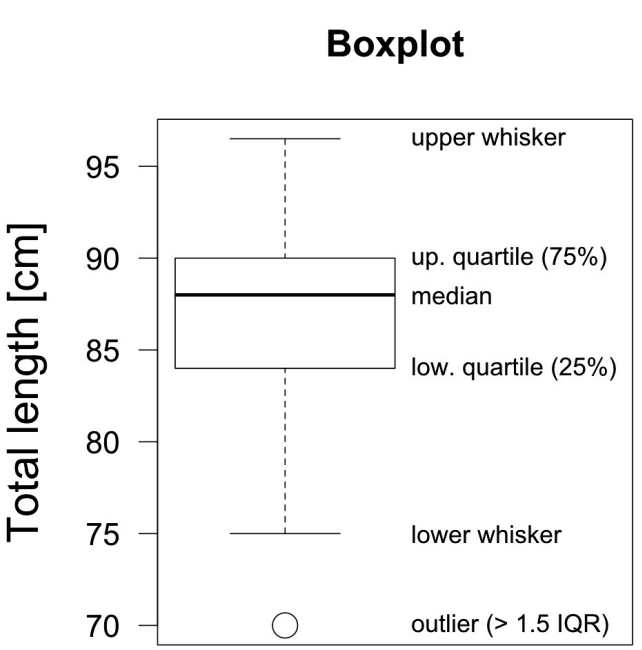
<u>Elements of data exploration – Checking for:</u>

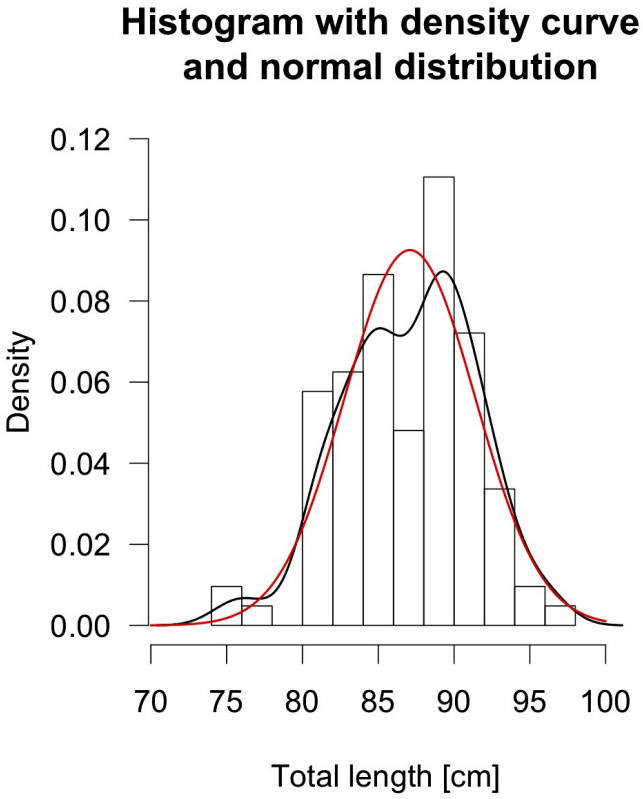
1. Outliers (e.g. boxplot)
2. Variance homogeneity (e.g. conditional boxplot)
3. Normal distribution (e.g. QQ-plot)
4. (Double) zeros (e.g. frequency plot)
5. Collinearity (e.g. pairwise scatterplots)
6. Relationship explanatory and response variable (e.g. scatterplots)
7. Spatial- or temporal autocorrelation (e.g. variograms)

Data exploration

Common plots for looking at the data







Linearity?
Collinearity?

Outliers?

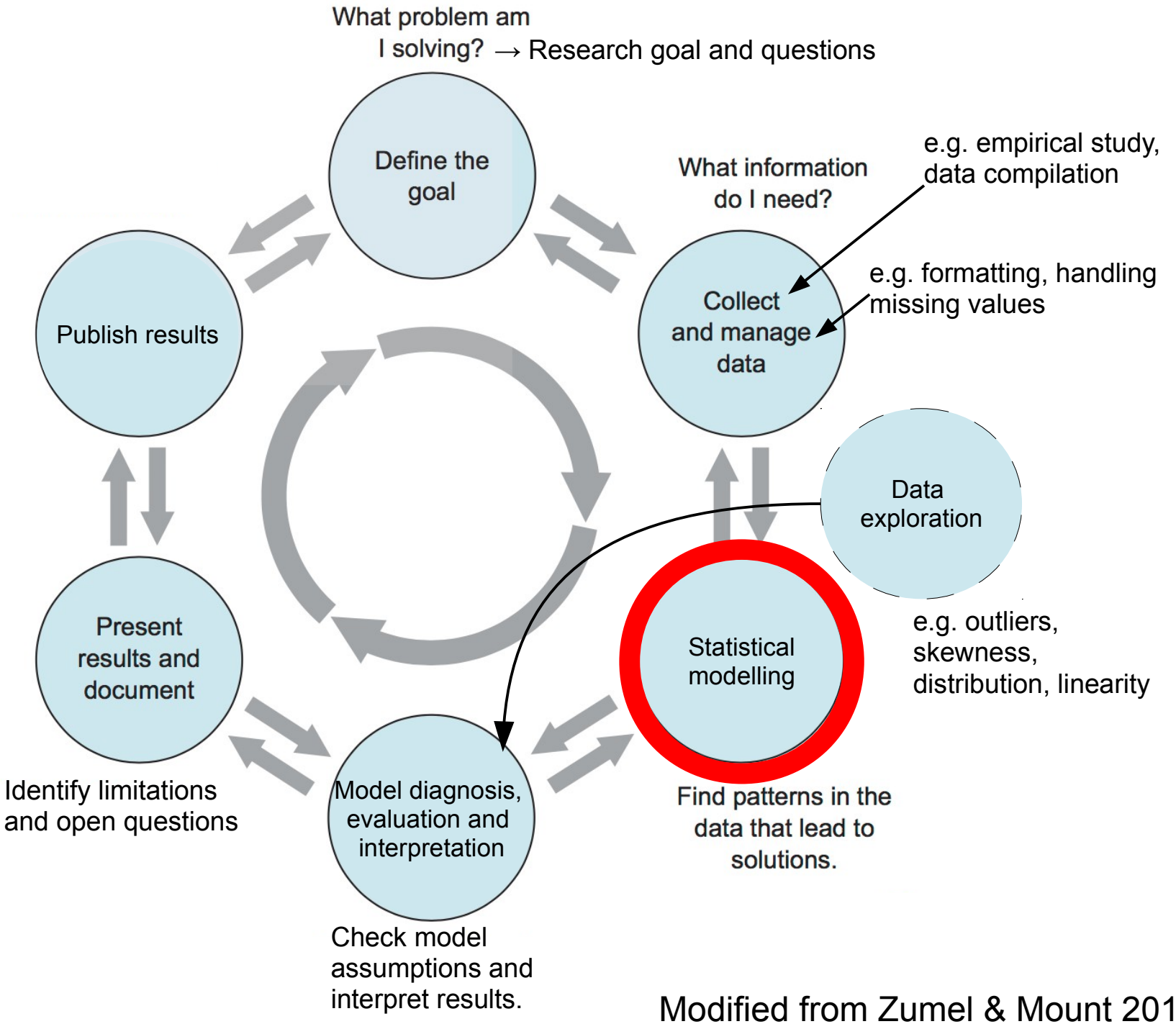
Asymmetry of distribution?
Normality?

Statistical modelling, simulation and the linear model

Contents

1. Framework for data analysis and tools for data exploration
2. **Statistical modelling and simulation-based tools**
3. Permutation and Monte Carlo simulation
4. Bootstrapping
5. Cross-Validation and Bias-variance trade-off
6. Revisiting the linear model

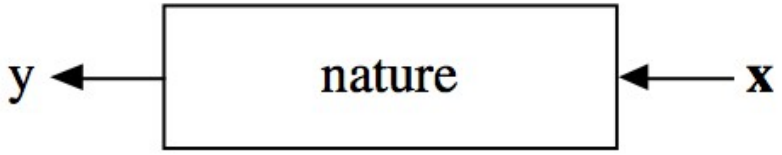
Data analysis cycle



Modified from Zumel & Mount 2014: 6

Statistical modelling: The two cultures

Real world: Processes lead to association between **x** and **y**



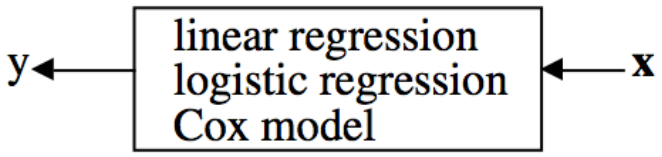
Examples for goals of statistical modelling: predict unknown **y** from **x**, estimate how **x** is related to **y**

Data modelling culture (classical statistics)

Common data model

response variables = f(predictor variables, random noise, parameters)

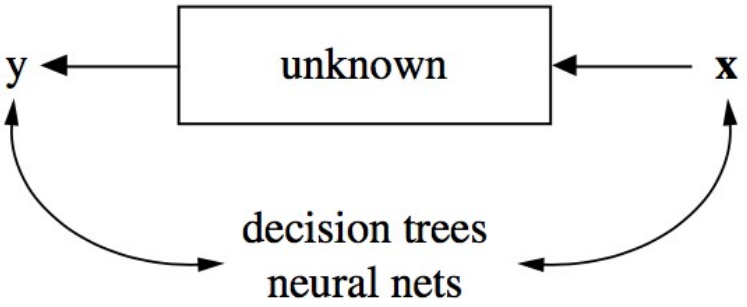
Estimate parameters from data



Model validation: Check residuals

Algorithmic modeling culture (machine learning)

Find algorithm that operates on **x** to predict **y**



Model validation: Predictive accuracy

Breiman 2001 *Statistical Science* 16: 199

Statistical modelling: the classical view

- Fit model to data to inform estimation, inference or prediction (e.g. estimate point or interval, test hypothesis)
  - Example: The arithmetic mean $\bar{x}$ is an estimate of the true population mean μ and s^2 is an estimate of the true variance σ^2
- Most models incorporate a deterministic (fixed effect) and a stochastic component (random effect)
  - Example: $y_i = b_0 + b_1 x_i + \epsilon_i \quad \text{with} \quad \epsilon \sim N(0, \sigma^2)$
- All models rely on assumptions → Model diagnosis
  - e.g. normal distribution, independence of observations
- Goodness of fit measures aid to choose between multiple models that fit the data
  - e.g. AIC, R^2, RMSE

Simulation-based approaches in data analysis

- Compatible with both cultures
- Infuses algorithm-based thinking into classical statistics
- Examples for simulation-based approaches for estimation, inference or model diagnosis in classical statistics:

1. **Permutation test** → Permuting (shuffling) the data to derive null distribution. Mainly used for inference
2. **Bootstrapping** → Randomly sampling subsets from the data with replacement. Mainly used for estimation of parameter distribution
3. **Cross-validation** (CV) → Splitting data into sets (i.e. sampling without replacement). Mainly used for validation of predictive models

Statistical modelling, simulation and the linear model

Contents

1. Framework for data analysis and tools for data exploration
2. Statistical modelling and simulation-based tools
3. **Permutation and Monte Carlo simulation**
4. Bootstrapping
5. Cross-Validation and Bias-variance trade-off
6. Revisiting the linear model

Permutation test: Algorithm

Repeat *k* times

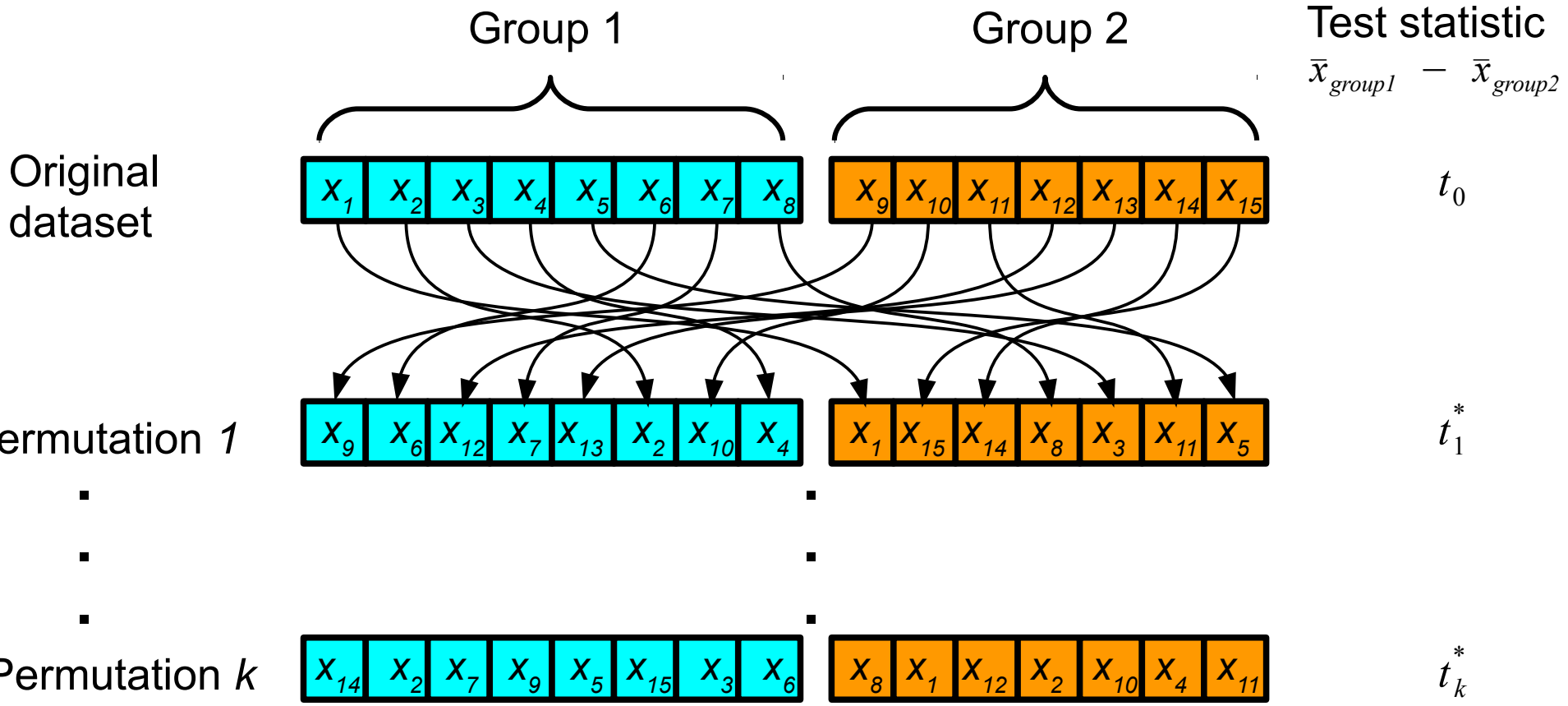
1) Permute values in data set
2) Compute test statistic t^* for permuted data

3) Compare test statistic t_0 to generated null distribution

Permutation test: Algorithm

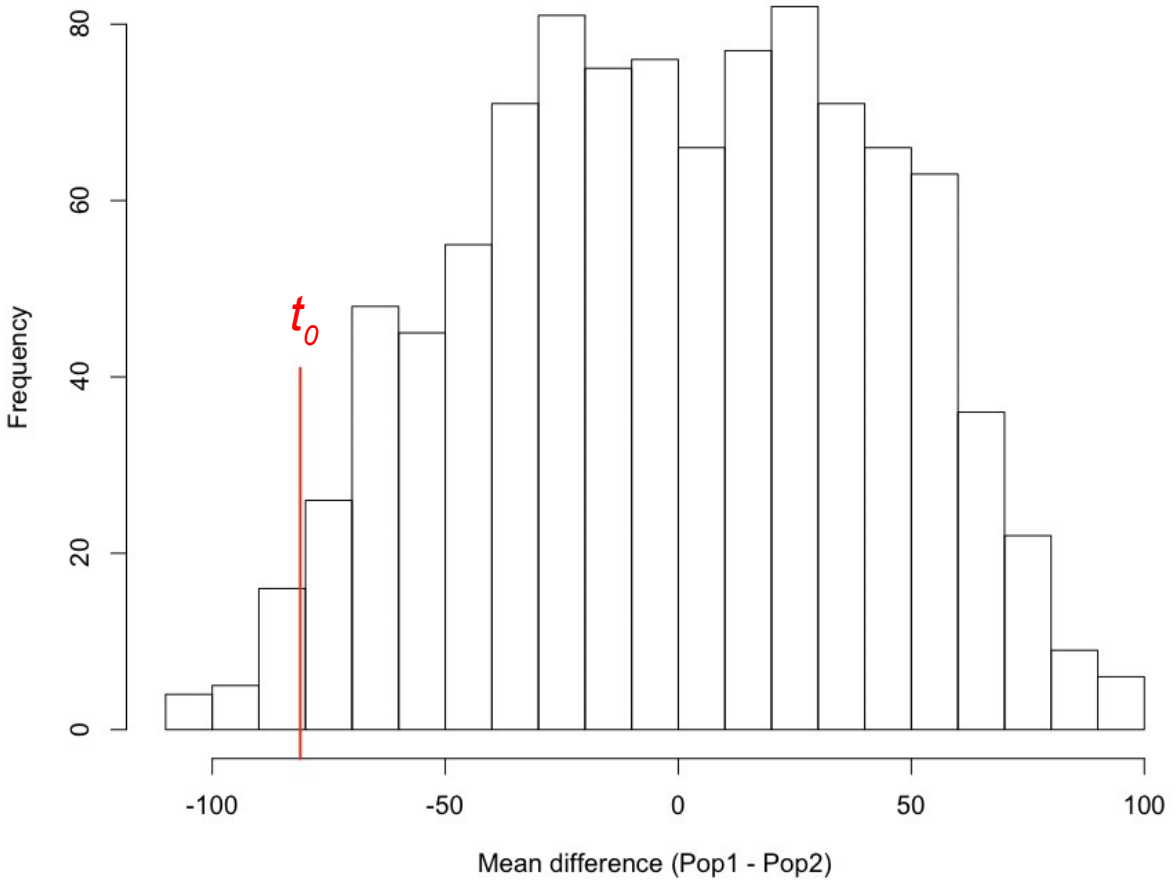
Repeat *k* times {
1) Permute values in data set
2) Compute test statistic *t** for permuted data
3) Compare test statistic t_0 to generated null distribution

<u>Example:</u> Permutation test of difference in group mean



Permutation test: Generated distribution

$$p = \frac{\sum_{i=1}^{k} 1 \text{ if } t_i^* \leq t_0, else\, 0}{k+1}$$



- Test informs whether pattern in data is due to chance
- Inference regarding statistical population only valid if distribution of sample data matches actual distribution of statistical population → particularly problematic for small *n*

Permutation test: Advantages and limitations

- Advantages
  - Free from distributional assumptions
  - Applicable to complex designs through restricting permutations
- Limitations
  - Generalisation to statistical population requires matching distribution
  - Statistical hypothesis testing can imply distributional assumptions that apply to the permutation test, if aiming to infer to the statistical population (e.g. testing for mean differences affected by variance)
  - Computationally intensive: Number of all possible permutations for a dataset is factorial *n*, i.e. *n*! (e.g. $35! \approx 10^{40}$) → Monte Carlo simulation

Legendre & Legendre 2012: 25ff

Monte-Carlo simulation

- Uses repeated random sampling to solve problems probabilistically (even though they can be deterministic in reality)
- Permutation tests use random numbers to randomly permute data → approximate with MC simulation
- Legendre & Legendre (2012): use at least 10,000 permutations for inference

Entrance of casino in Monte Carlo, Monaco

Edvard Munch - At the Roulette Table in Monte Carlo

Statistical modelling, simulation and the linear model

Contents

1. Framework for data analysis and tools for data exploration
2. Statistical modelling and simulation-based tools
3. Permutation and Monte Carlo simulation
4. **Bootstrapping**
5. Cross-Validation and Bias-variance trade-off
6. Revisiting the linear model

Bootstrapping: Idea and algorithm

- Inference on statistic *t* is based on sampling distribution
  - Ideally: Draw all or many samples from statistical population
  - Reality: Most frequently only one sample available
  - → **Idea:** Draw samples from an estimate of the statistical population (i.e. the sample) and use these to estimate property (e.g. variance) of the statistic *t*
- Algorithm:

  1) Draw random sample with replacement from data
  2) Compute statistic *t** for bootstrap sample

  } Repeat *k* times

  3) Use the *k* estimates to derive property of statistic
- Exhaustive bootstrapping ($k = n^n$) computationally demanding → approximate with Monte Carlo simulation
- Given todays computer power 10^4-10^5 simulations viable

Bootstrapping: Example

Example: Bootstrap to the mean (to derive variance)

t (here: mean)

Original dataset

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

$\bar{x} = 8$

Sampling with replacement

BS sample 1

15	7	8	4	15	1	11	9	1	3	6	14	2	11	12

$\bar{x}^* = 7.93$

BS sample 2

6	13	2	10	12	5	7	10	1	13	8	8	15	3	10

$\bar{x}^* = 8.2$

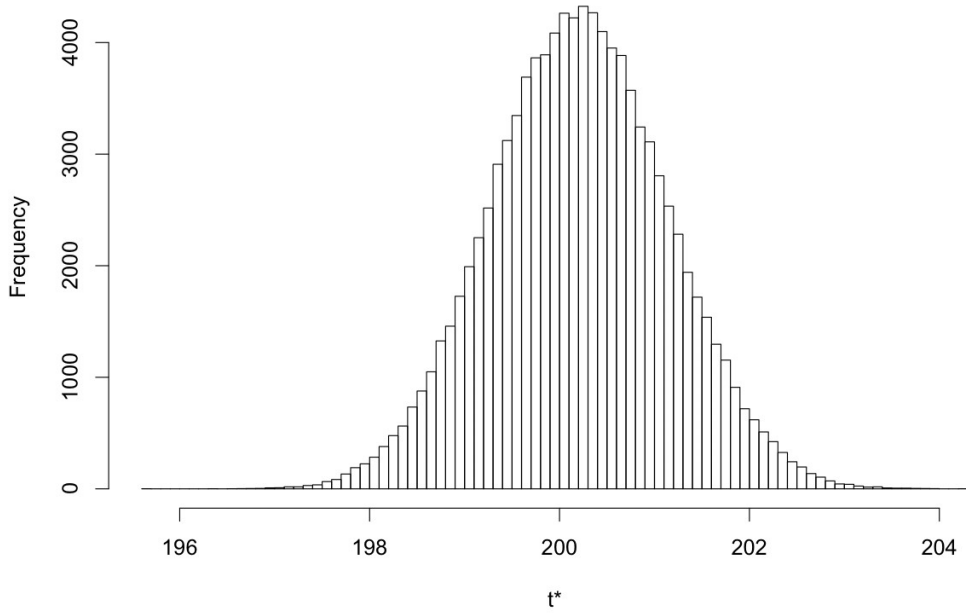
⋮

BS sample *k*

12	7	7	15	5	10	8	13	6	11	8	1	14	2	12

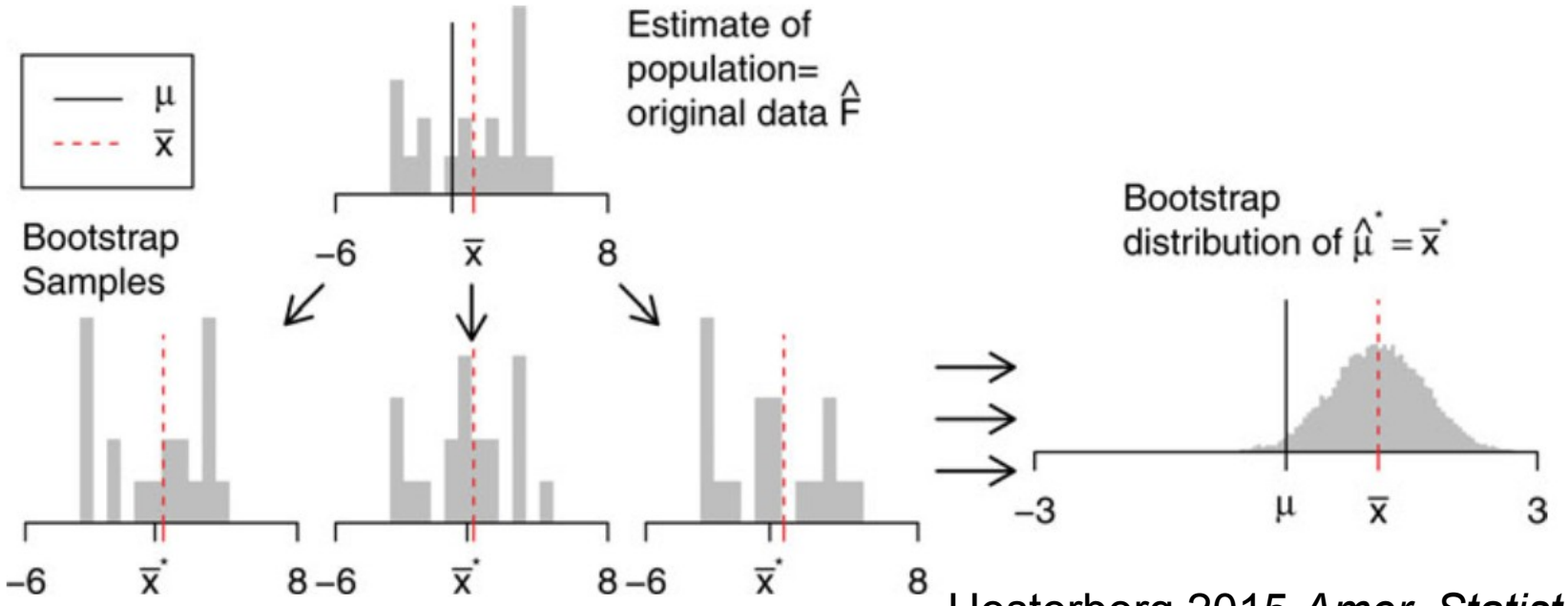
$\bar{x}^* = 8.73$

Distribution of statistic *t*



Bootstrapping: Limitations

- Do not use for hypothesis testing
- No distributional assumptions implied, but not reliable for all distributions, particularly at small *n* (see Hesterberg 2015)
- Small *n*: use adjusted bootstrap percentiles (Bca) or switch to parametric statistics (allow for additional assumptions)
- Bootstrap does not improve estimate of population parameter μ, centred at $\bar{x}$



Hesterberg 2015 *Amer. Statist.* 69:371

Statistical modelling, simulation and the linear model

Contents

1. Framework for data analysis and tools for data exploration
2. Statistical modelling and simulation-based tools
3. Permutation and Monte Carlo simulation
4. Bootstrapping
5. **Cross-Validation and Bias-variance trade-off**
6. Revisiting the linear model

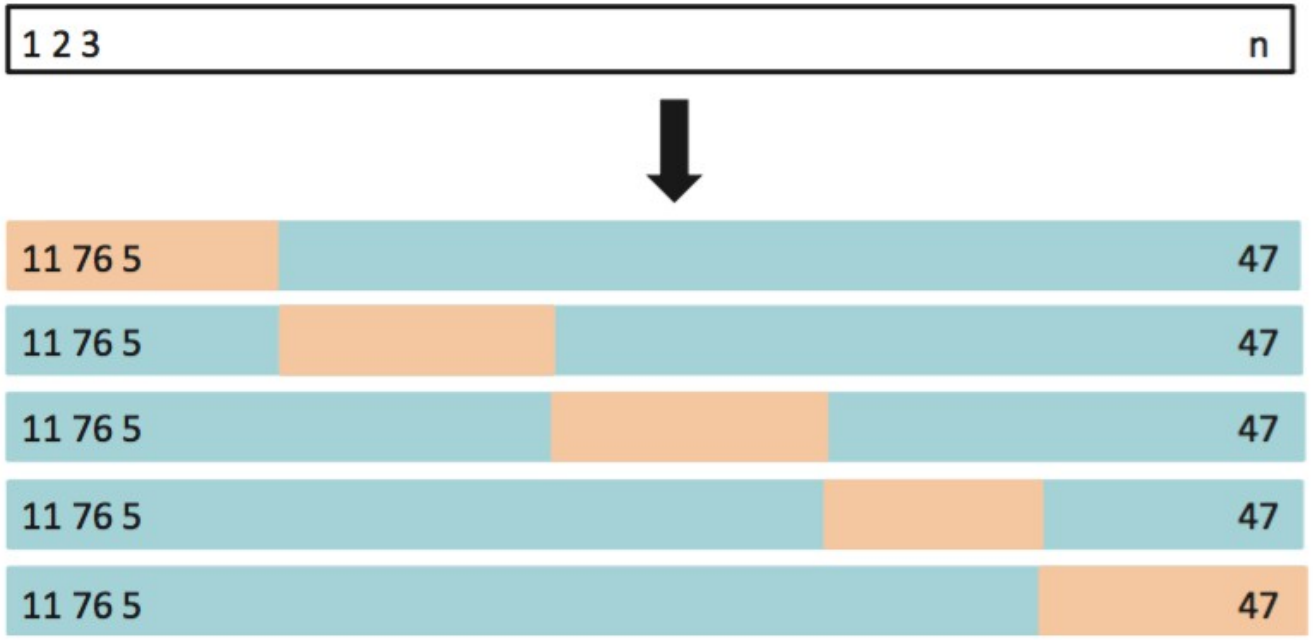
Cross-validation (CV)

- Objective: Evaluate predictive accuracy of a fitted model
- Can be checked if independent training data (used to fit model) and test data (new data) are available → Rare case
- **Idea:** Split the available data into training and test set and predict the (known) observations in the test set from a model fitted with the training data
- Algorithm:

1. Draw *k* random samples without replacement from data
2. For each *k:*
    1. Fit the model to the other *k*-1 parts
    2. Predict *k* from model and calculate the prediction error
3. Calculate prediction error as average over the *k* estimates

Cross-validation (CV)

Example: $k = 5$



- Problem of choosing *k*:
  - *k* = n (Leave-one-out CV predicts each observation from all others) → low bias, but high variance
  - *k* = 2 (split data into half) → low variance, but high bias
- *k* typically set to 5 or 10

Taken from James *et al.* 2013: 181

Bias-variance trade-off

Definitions in context of model validation:

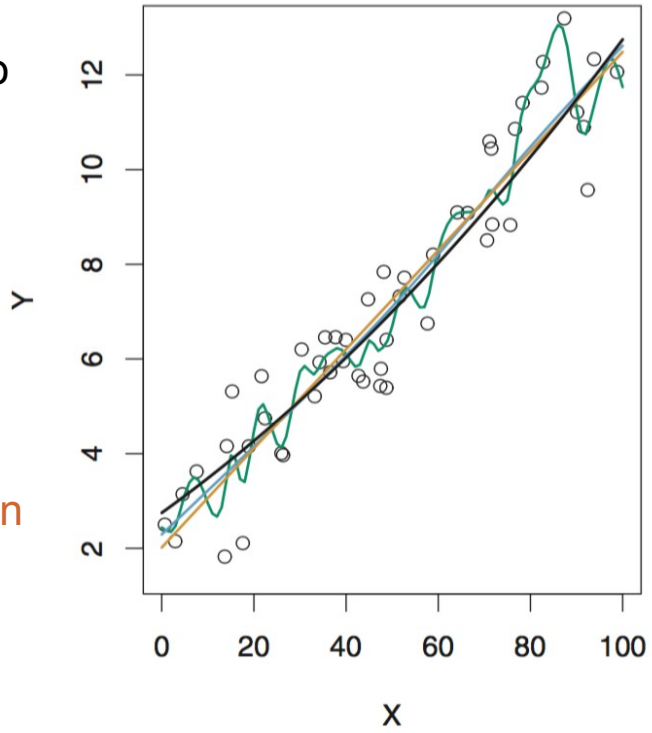
- **Bias:** error when approximating training data
- **Variance:** variability in error when approximating test data

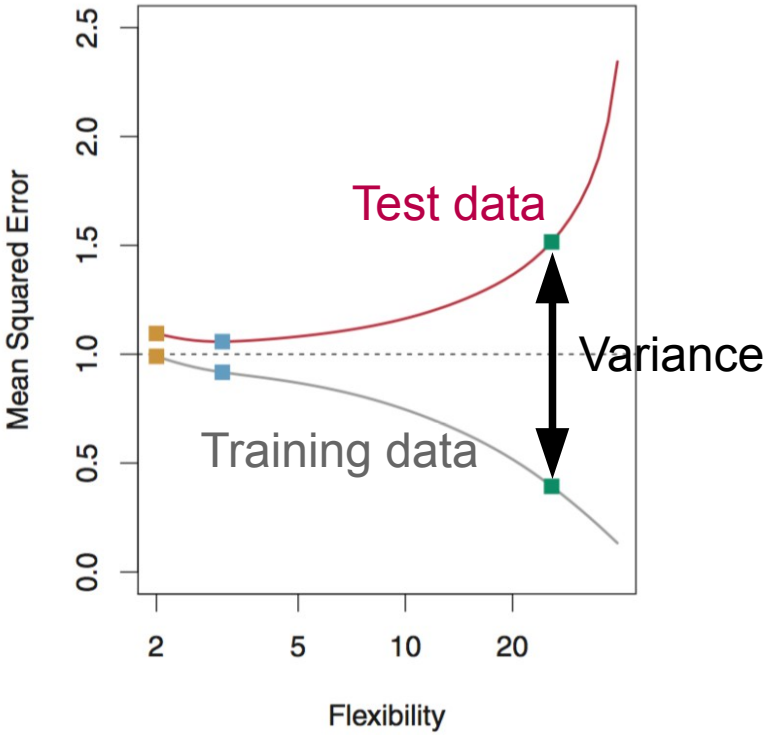
function used to simulate data

highly flexible smoother

little flexible smoother

linear regression



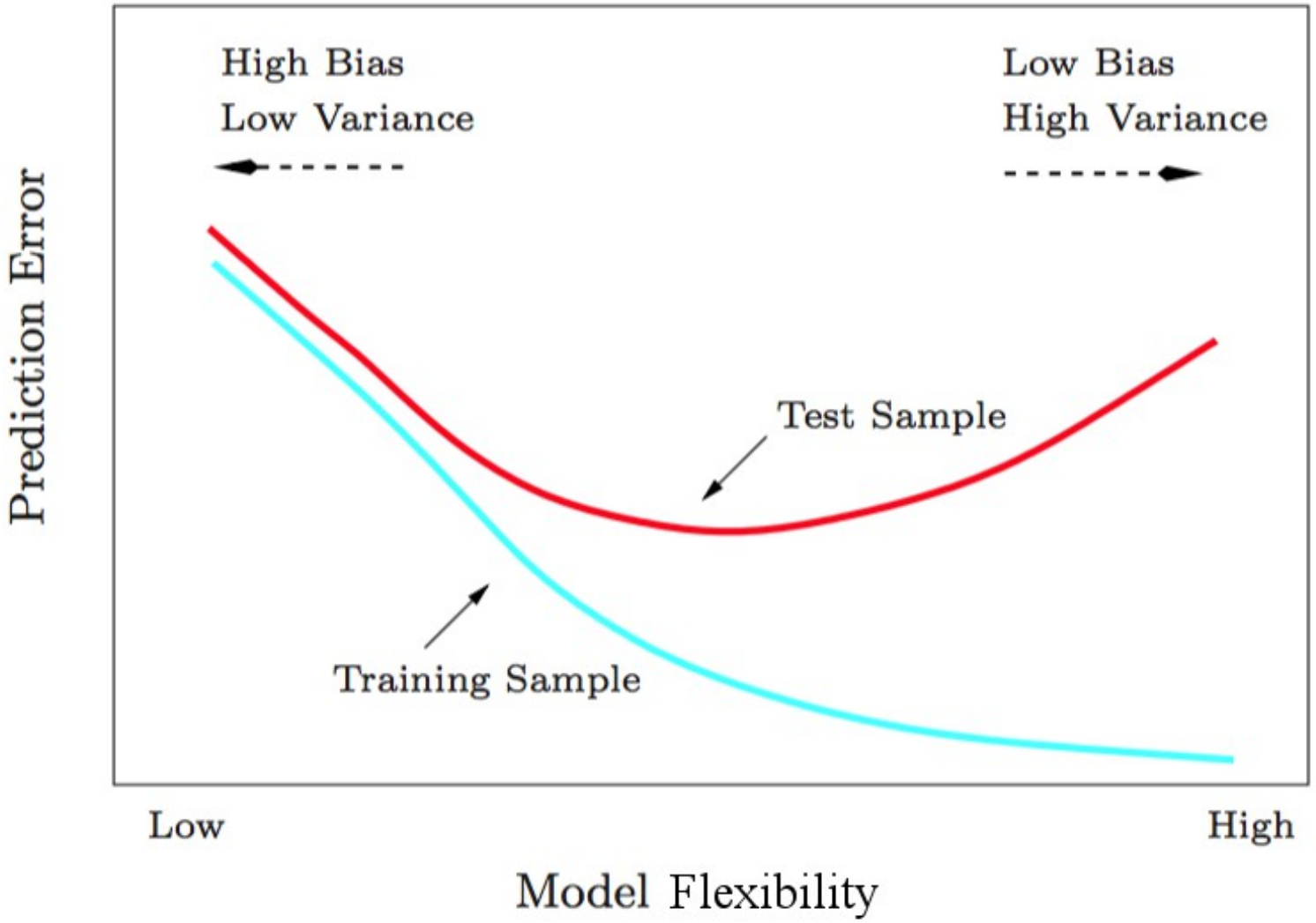


Higher flexibility (higher *k* in CV) → lower error for training data (i.e. lower bias), but variance will increase from some point

Taken from James *et al.* 2013: 33

Bias-variance trade-off

Higher flexibility (higher *k* in CV) → lower error for training data (i.e. lower bias), but variance will increase from some point → Optimise combined error



Taken from Hastie, Tibshirani and Friedman 2011: 38

Statistical modelling, simulation and the linear model

Contents

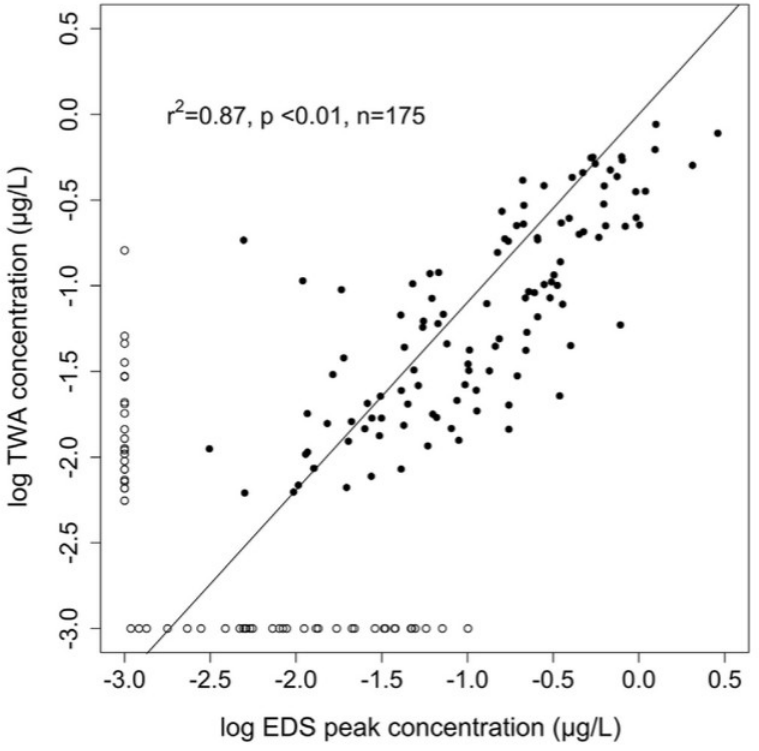
1. Framework for data analysis and tools for data exploration
2. Statistical modelling and simulation-based tools
3. Permutation and Monte Carlo simulation
4. Bootstrapping
5. Cross-Validation and Bias-variance trade-off
6. **Revisiting the linear model**

Relationship between two continuous variables: linear regression model

- Bivariate relationship between an explanatory variable and a response variable with:

$$y_i = b_0 + b_1 x_i + \epsilon_i \qquad \text{with} \quad \epsilon \sim N(0, \sigma^2)$$

- Example: Can we approximate pesticide runoff concentrations with passive sampling?



Fernandez et al. 2014

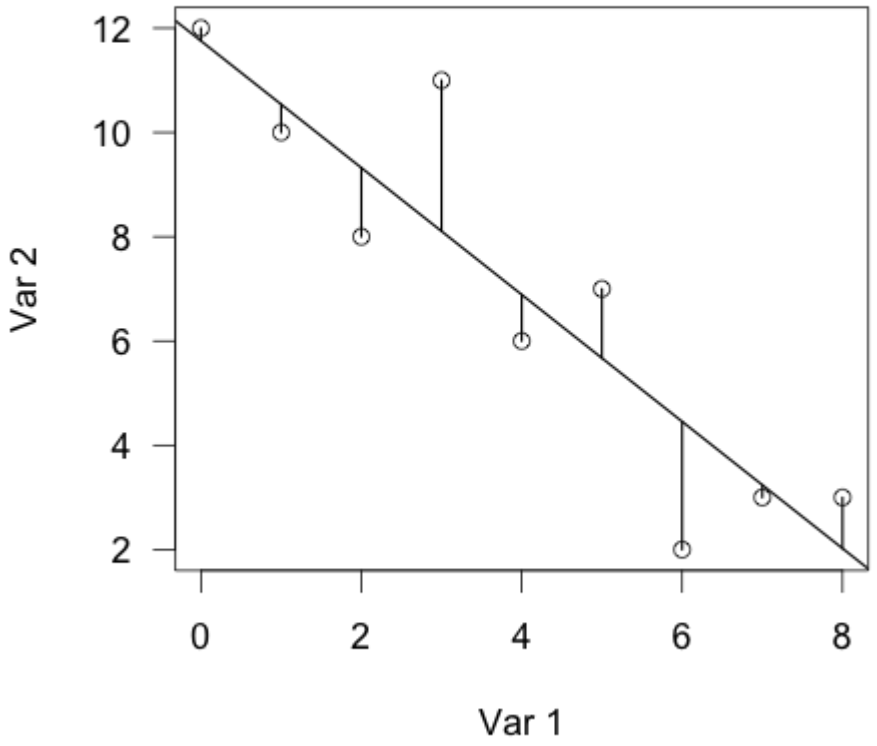
Relationship between two continuous variables: linear regression model

- Bivariate relationship between an explanatory variable and a response variable with:

$$y_i = b_0 + b_1 x_i + \epsilon_i \qquad \text{with} \quad \epsilon \sim N(0, \sigma^2)$$

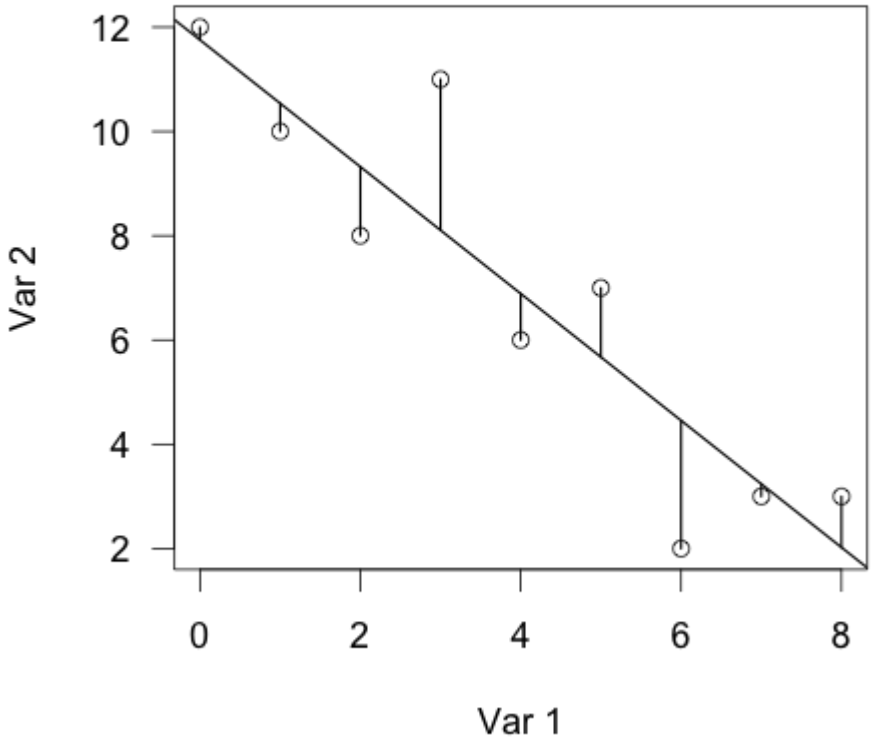
- Aim: minimise ε (also called error sum of squares: SSE)

$$\text{SSE} = \sum_{1}^{n} (y_i - b_0 - b_1 x_i)^2$$

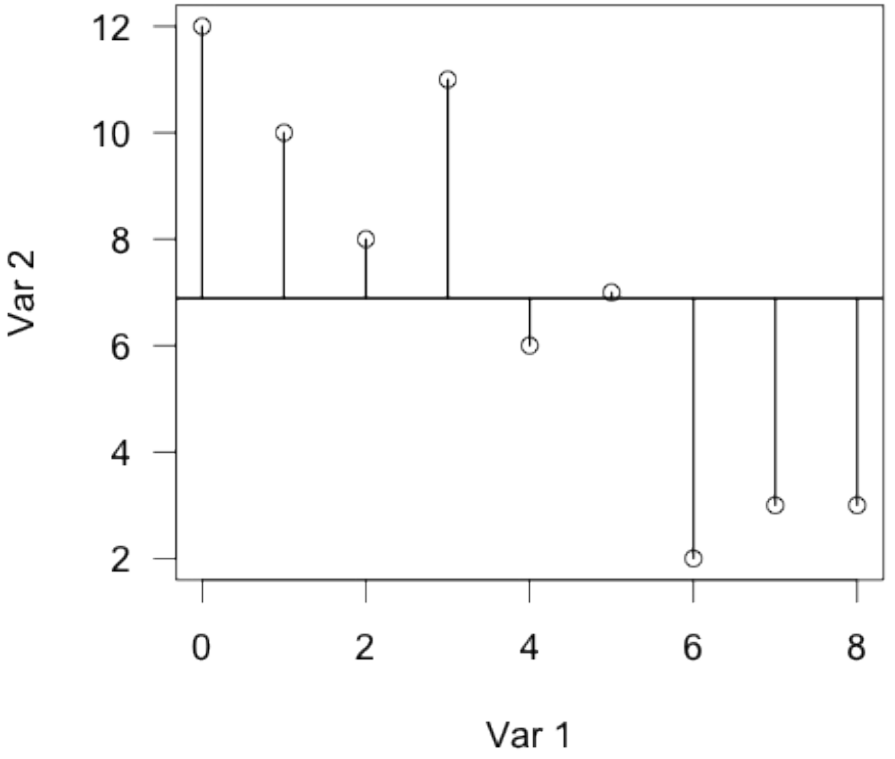


Linear regression model

$$SSE = \sum_{1}^{n} (y_i - b_0 - b_1 x_i)^2$$



$$SSY = \sum (y - \bar{y})^2$$



SSY = SSR + SSE

Total variation

Explained variation

Unexplained variation

% of explained variance:

$$R^2 = \frac{SSR}{SSY}$$

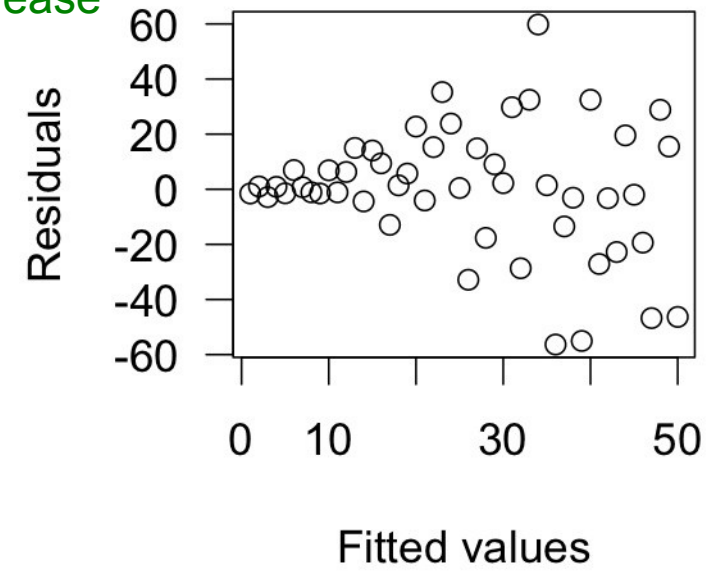
$$adj.\ R^2 = 1-(1-R^2)\frac{n-1}{n-p-1}$$

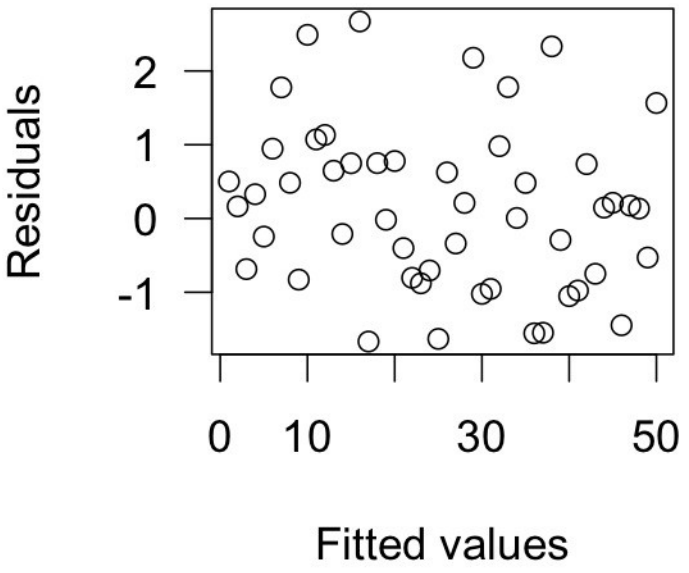
Linear regression model

- Assumptions:
  - Linear relationship (graphical diagnostics)
  - Normal distribution of error (graphical diagnostics)
  - Variance homogeneity (graphical diagnostics)
  - Independence of errors (graphical diagnostics)
- If one or more assumptions not met, alternatives include:
  - Generalised linear model, Generalised least squares, Mixed models
  - Variable transformation (but using an appropriate model such as a Generalised linear model is usually the better option)

Model diagnostics: Variance homogeneity

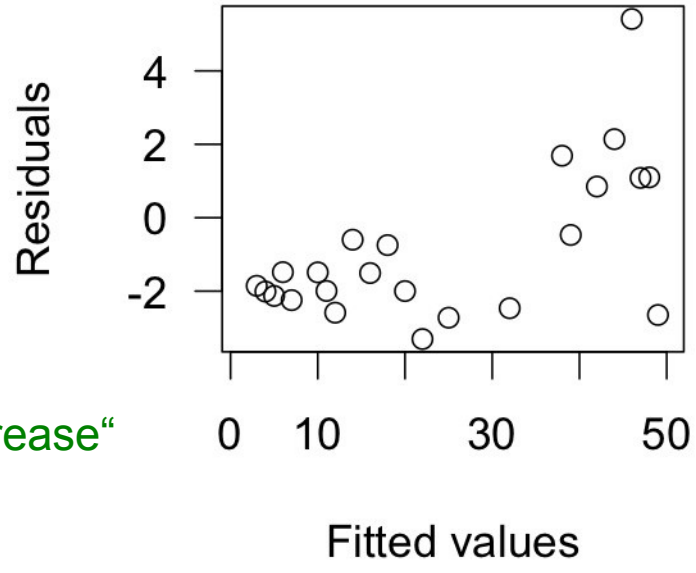
„strong increase“



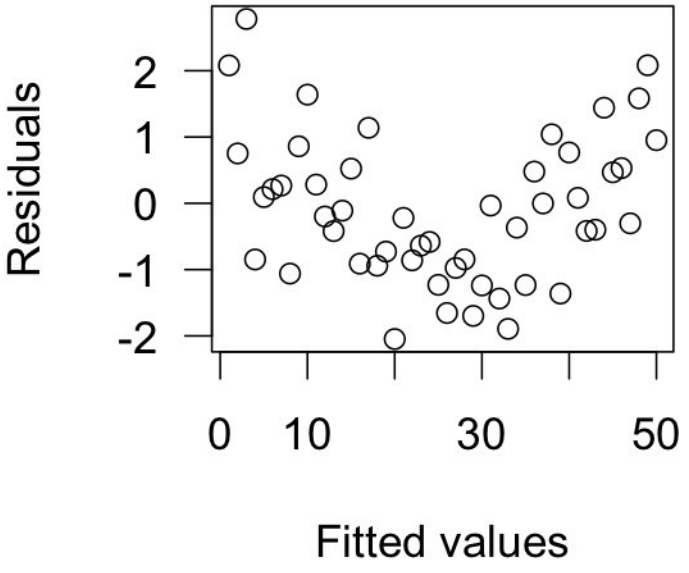


„normal“

Residuals vs. fitted values plots



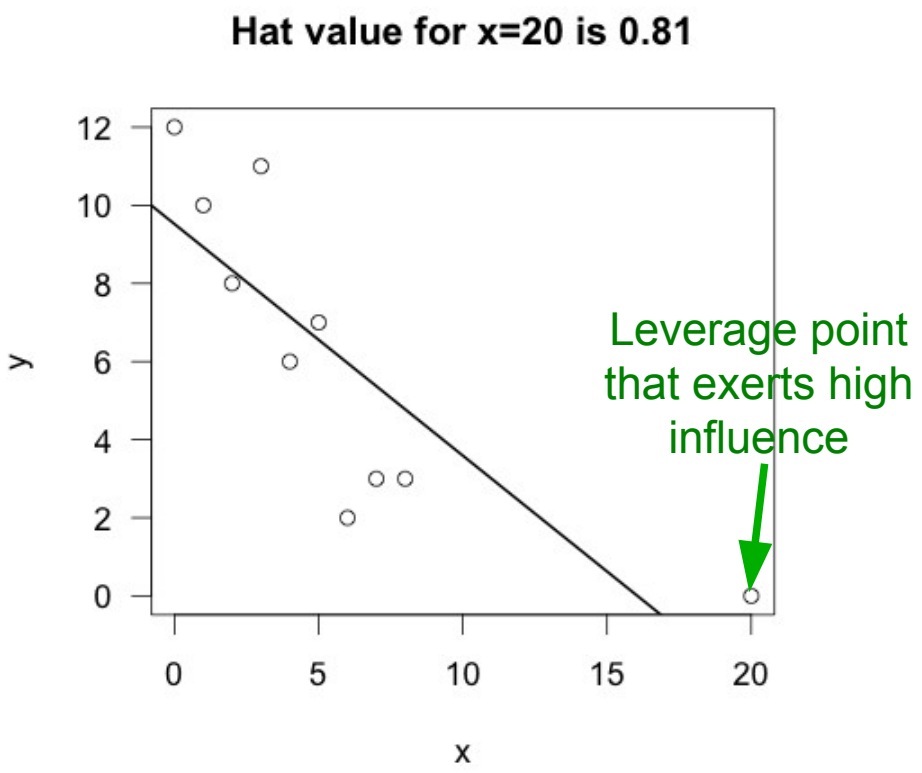
„slight increase“

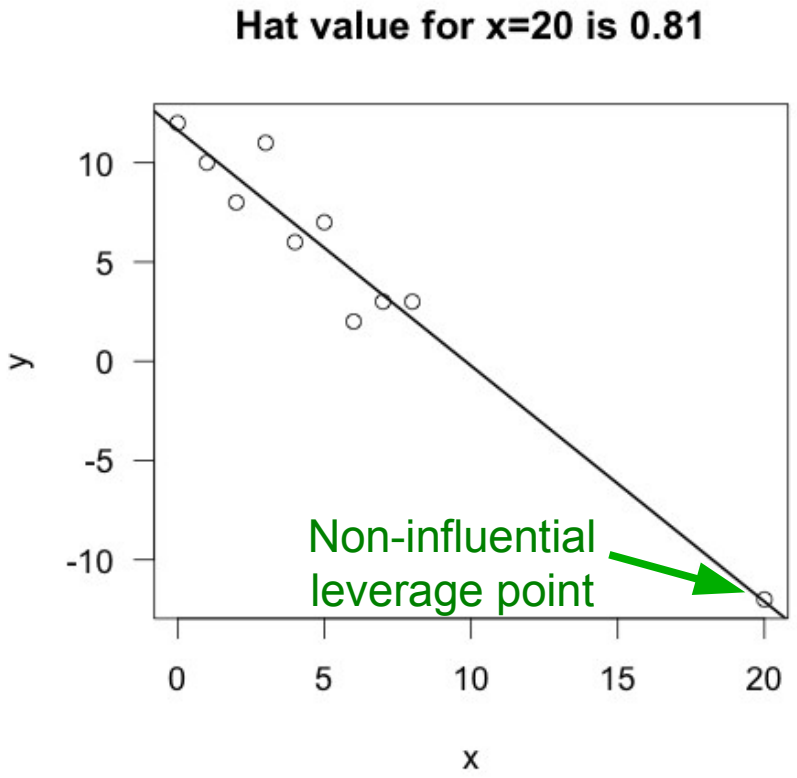


„non-linear“

Further model diagnostics

Leverage points (predictor outlier)

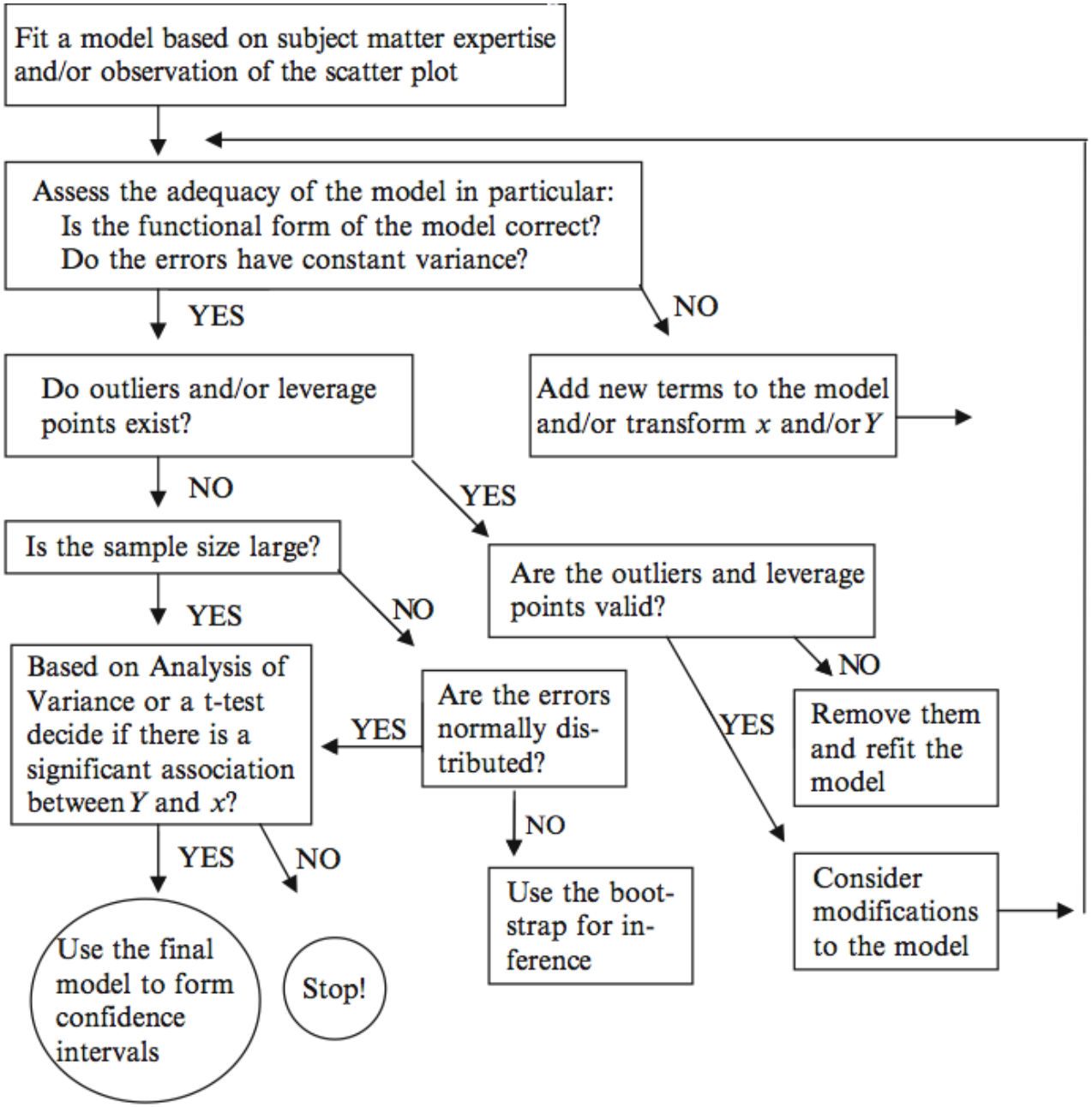




How to deal with leverage points/outliers?

- Check whether values are plausible
- Check robustness of model results when removing observations
- Fit different statistical model or transform data

Flowchart for simple linear regression



Taken from Sheather 2009: p.103

Simulation-based approaches to simple linear regression

- Predictive accuracy measured with Mean square prediction error (MSPE):

$$\mathrm{MSPE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$ for the new observations 1 to *m*

- Cross-validation (CV): Calculate CV-MSPE and CV-R^2
- Bootstrapping (BS) in regression analysis:
  - of residuals: BS residuals, add to $\hat{y}$ to generate new y^* and calculate regression coefficients → ***x*** fixed
  - of cases: BS complete cases and calculate regression coefficients → ***x*** random
  - If ***x*** and ***y*** random sample (e.g. *x* not fixed in experiment), residuals correlated or exhibit non-constant variance → BS cases

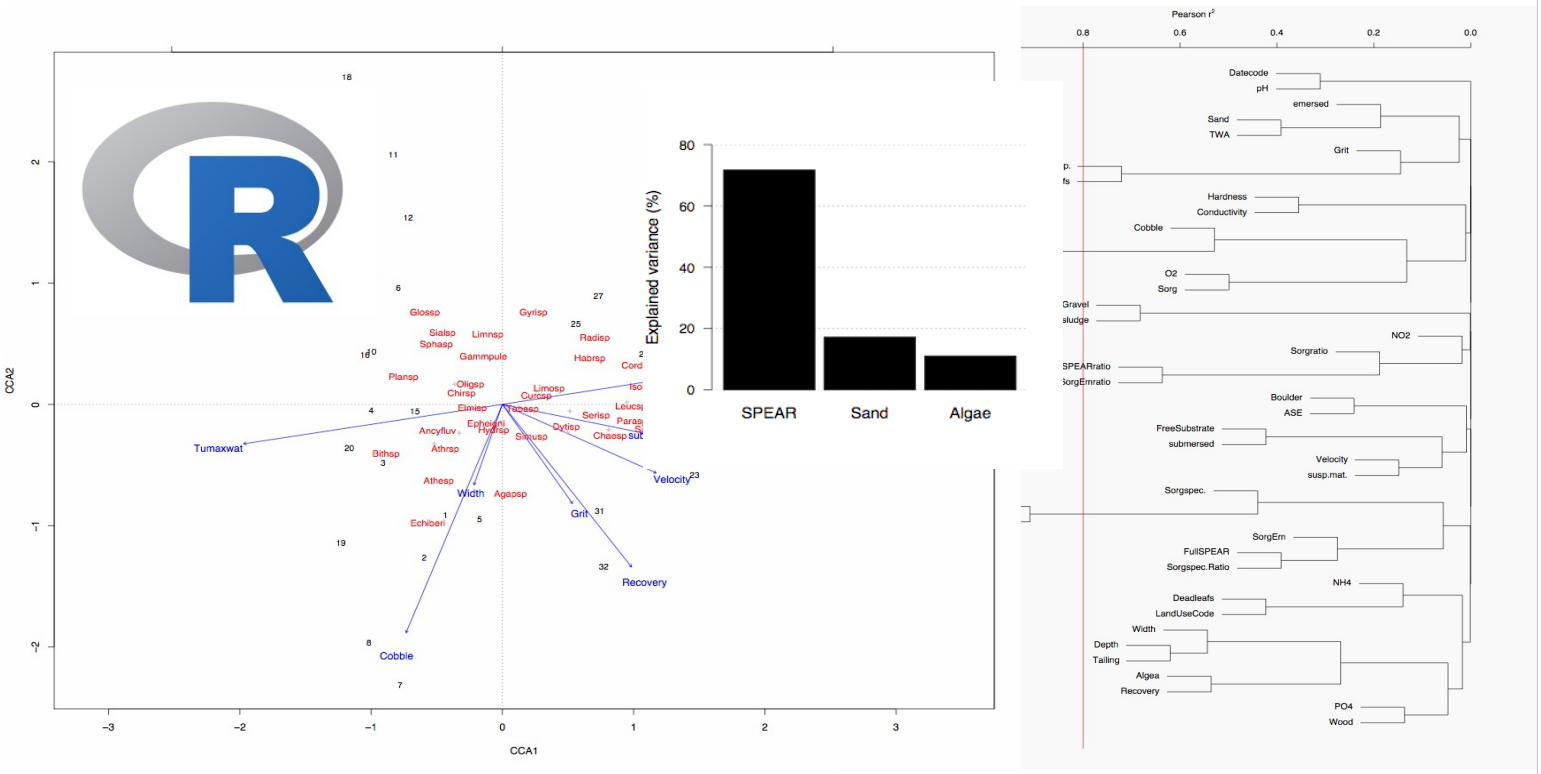
Exercise

We will work with the data set “possum” that includes biometric measurements of possums in Victoria, Australia.

Conduct a linear regression analysis, diagnose and interpret the model and apply simulation-based approaches.

Applied Multivariate Statistics

University of Koblenz-Landau 2017/18



Ralf B. Schäfer

These slides and notes complement the lecture with exercises "Applied multivariate statistics for environmental scientists". Do not hesitate to contact me if you have any comments or you found any errors (text or slides): schaefer-ralf@uni-landau.de

While I made notes below the slides, some aspects are only mentioned in the R scripts associated with the lecture.

Short introduction

- Professor for Quantitative Landscape Ecology
- Current teaching: Statistics (M.Sc.); GIS (B.Sc./M.Sc.); Environmental Modelling (B.Sc./M.Sc.); Aquatic Ecotoxicology (M.Sc.); Environmental Philosophy (B.Sc.)
- Research focus:
  - Community ecology of freshwater invertebrates and microorganisms
  - Response of freshwater ecosystems to different (anthropogenic) stressors (e.g. pollution)
  - Trophic linkages between aquatic & terrestrial systems
- Primarily field studies/experiments and data analyses/ modelling

www.landscapecology.uni-landau.de

2

Organisation

- Lecture material (including course schedule and literature list) can be found on github and website:
  https://github.com/rbslandau/statistics_multi
  https://goo.gl/EhPVFG

- Inverted classroom: Self study of lecture and demonstration, Q&A and exercises in class room

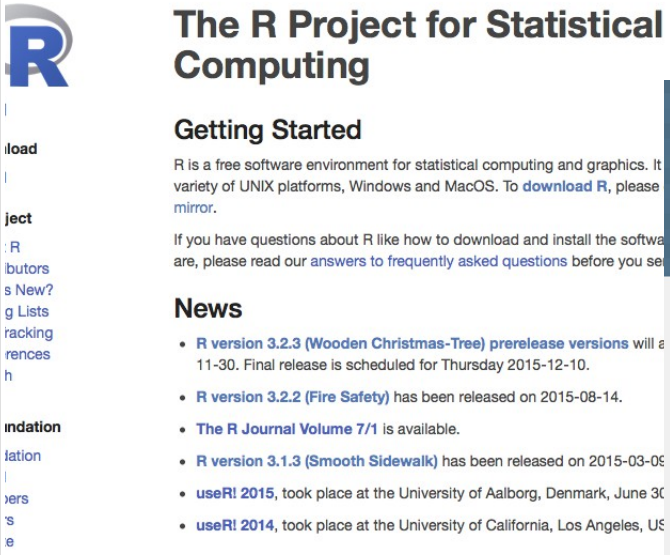
- Contact time: 2 hours per week; Own study time: approximately 1 day per week

3

Literature references that are listed in the literature list are cited in short form on slides. For literature not contained in the literature list, I give the complete reference on the slide or in the notes for the respective slide.

Using your own notebook

- feel free to you use your own WLAN-enabled notebook!
- install R (http://mirrors.softliste.de/cran/) oder RStudio (recommended for beginners - http://www.rstudio.com/)
- Run “0_Install_packgs.R”, provided on github
- for installation of additional packages run install.packages(“package to be installed”)



4

Course objectives: Learning outcomes

- Classify, explain and interpret the different types of (multivariate) statistical approaches
- Select and apply the appropriate statistical method for the research goal
- Demonstrate moderate level of statistical modelling skills, including scripting in R

5

Two incorrect ways of thinking about stats

1. **Overconfidence:** Statistics is like mathematics and provides a single, correct answer
   But statistical thinking differs from mathematical thinking

2. **Disbelief:** Anything goes – statistics cannot be trusted
   But: statistics provide quantitative support of the complete research process

Tintle (2015) *Amer. Statist.* 69: 362

Tintle N., Chance B., Cobb G., Roy S., Swanson T. & VanderStoep J. (2015) Combating Anti-Statistical Thinking Using Simulation-Based Methods Throughout the Undergraduate Curriculum. *The American Statistician* 69, 362–370.

Statistical modelling, simulation and the linear model

Contents

Learning targets

- Explain the data analysis cycle and apply tools for exploratory data analysis
- Explain approaches to statistical modelling and simulation and apply simulation-based methods
- Diagnosing and interpreting the linear model

8

Learning targets and study questions

- Explain the data analysis cycle and apply tools for exploratory data analysis
  - Explain the steps of the data analysis cycle.
  - Summarise the elements of exploratory analysis. Which graphical tools are essential?
- Explain approaches to statistical modelling and simulation and apply simulation-based methods
  - Discuss the two different approaches to statistical modelling and links through simulation-based approaches.
  - Explain the purpose and critically discuss permutation tests.
  - Explain the purpose and critically discuss bootstrapping.
  - Explain the main idea of cross-validation and discuss the selection of *k* with respect to the bias-variance trade-off.

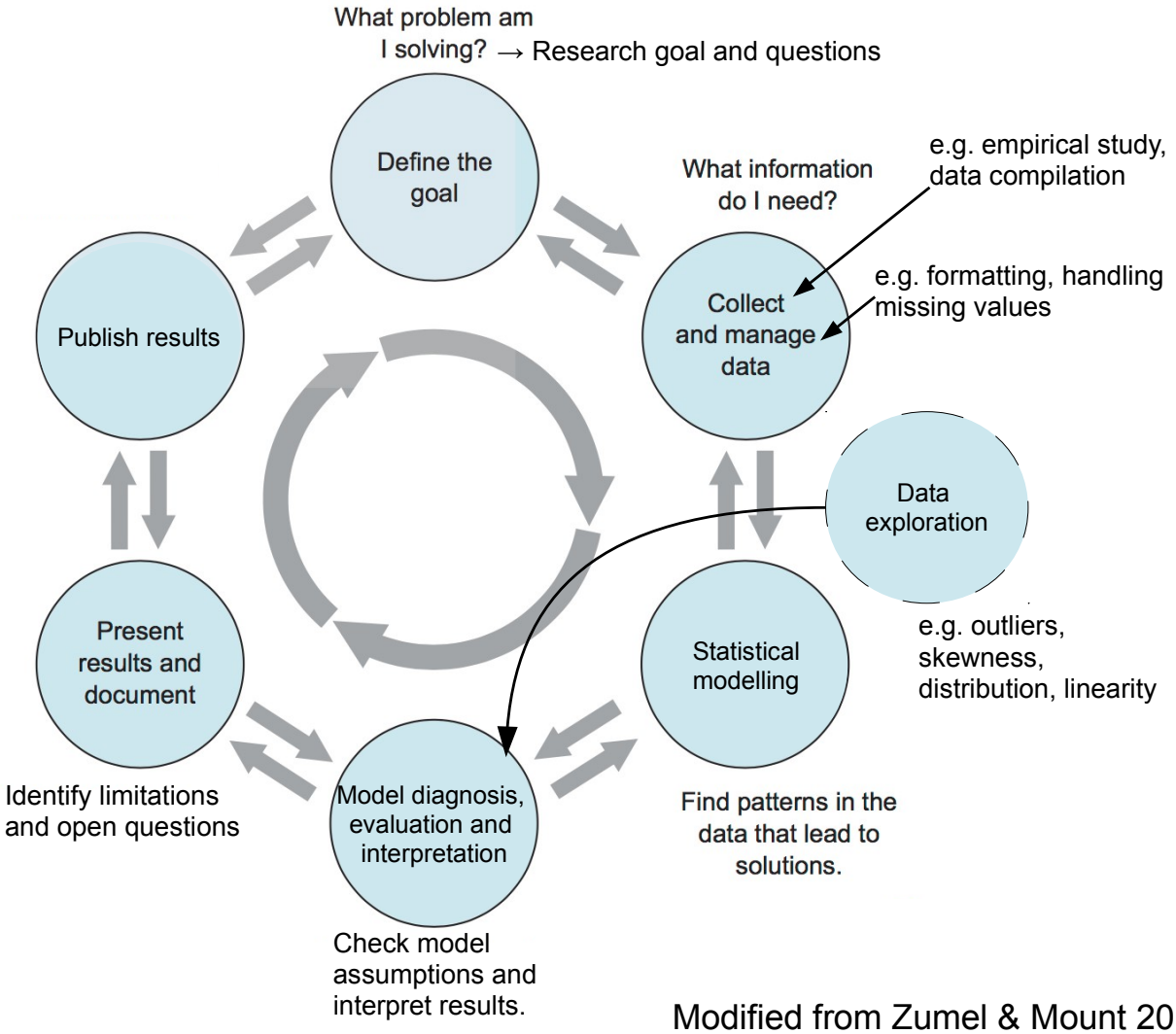
9

Learning targets and study questions

- Diagnosing and interpreting the linear model
  - Describe the assumptions of the linear regression and explain how they can be checked.
  - Which types of outliers exist? When is an outlier important?
  - Discuss the application of bootstrapping and cross-validation for the linear model.

10

Data analysis cycle



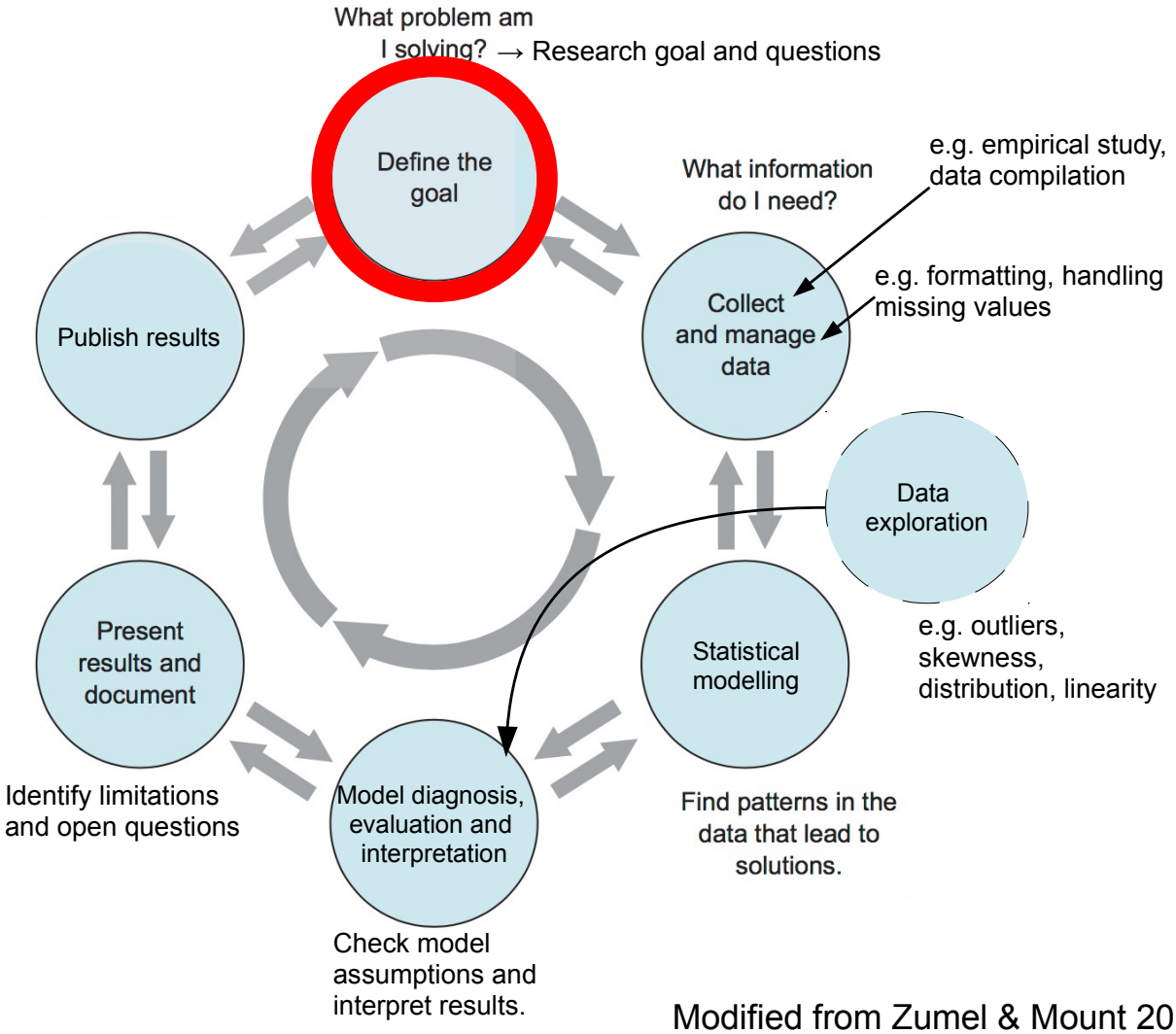
11

Modified from Zumel & Mount 2014: 6

Zumel N. & Mount J. (2014) Practical data science with R. Manning Publications Co, Shelter Island, NY.

Data exploration visualised with dashed line as it will depend on the research context if and when data exploration is conducted. However, most frequently data exploration (e.g. descriptive statistics such as data summaries) is employed before statistical modelling and the characteristics of the data set are explored to aid model selection. In some studies and disciplines, eventually no statistical modelling is done and only descriptive statistics is reported. Nevertheless, in case that a clear research hypothesis has been established before data collection, data exploration may not be required before statistical modelling. Still, the techniques related to data exploration will be needed to check model assumptions. Note that you must not establish research or statistical hypotheses after data exploration.

Data analysis cycle



Modified from Zumel & Mount 2014: 6

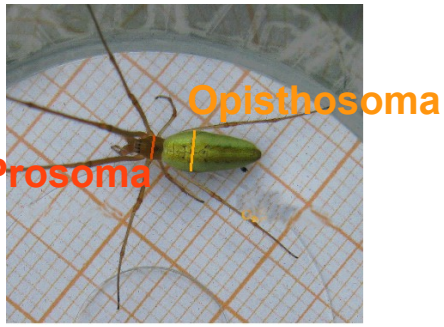
12

Define research goal and question

- Research goals (e.g. prediction, estimation, inference) and questions should inform study design and methods
- Aim: Test scientific hypothesis → Formulate testable hypothesis

Example

Question: Does the body condition of riparian spiders differ between restored and non-restored stream stretches?



Scientific hypothesis: Restoring stream stretches alters aquatic communities, resulting in different emerging insects on which riparian spiders prey. This affects the spiders' body condition derived from prosomal (pr.) and opisthosomal (op.) width.

- Testable hypothesis: The sample means for the body condition are drawn from populations with the same μ:

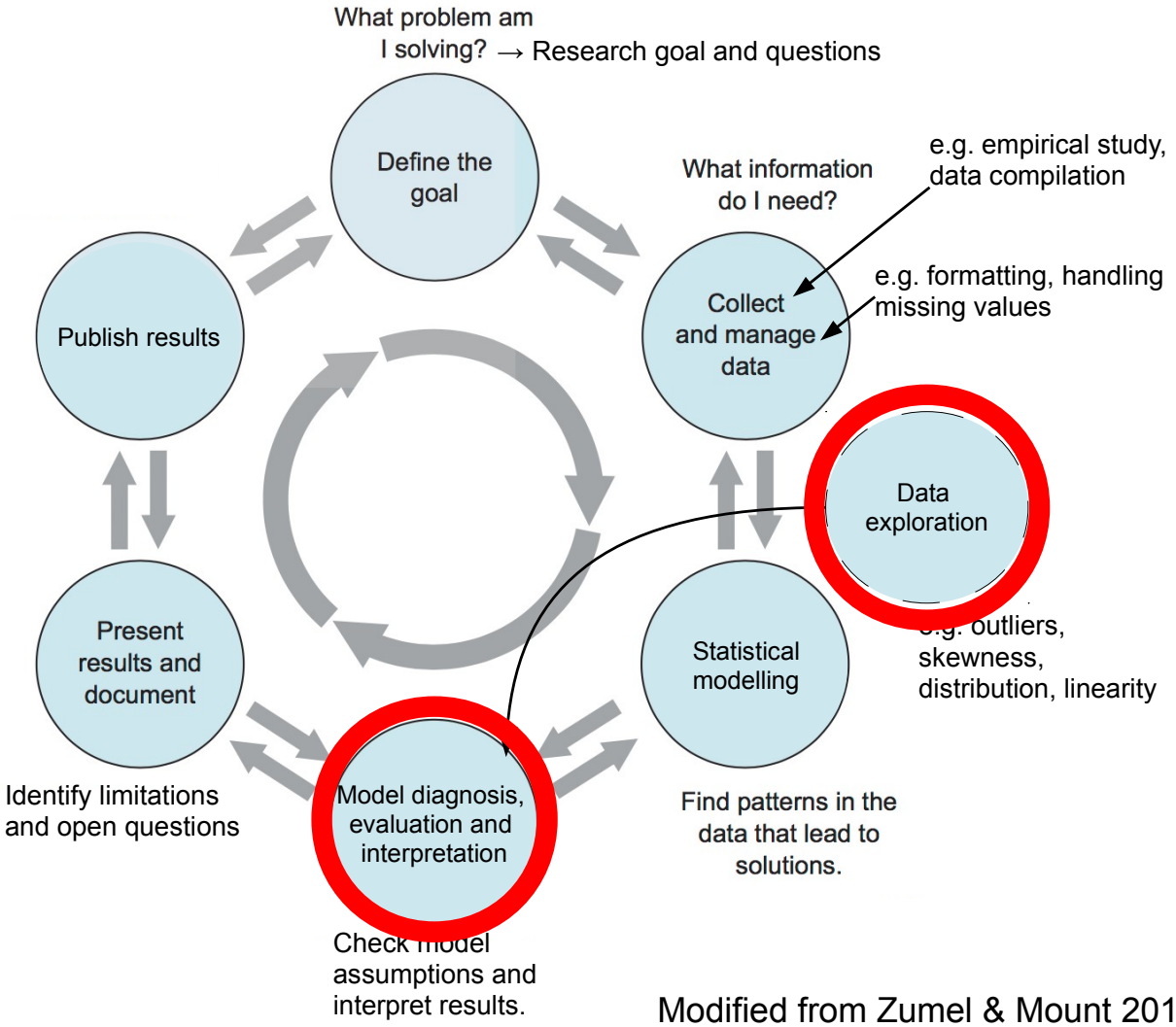
$$H_0: \ \mu_{restored} = \mu_{non-restored} \qquad H_1: \ \mu_{restored} \neq \mu_{non-restored}$$

13

River restoration may lead to improvements such as increased species richness of the aquatic invertebrate community. Terrestrial predators in the riparian zone such as spiders, in turn, may benefit from an increase in the biomass and diversity of aquatic emergent prey. In a study we therefore compared the body condition, using a proxy based on prosomal and opisthosomal width, between non-restored and restored stream reaches.

Statistical hypothesis testing consisted of comparing the central tendencies (sample means) using a paired t-test (each line corresponds to a different stream).

Data analysis cycle



Modified from Zumel & Mount 2014: 6

14

Tools for data exploration

- Useful for inspecting data before the modelling but also for model diagnosis
- Zuur et al. (2009) urge data inspection before modelling

➡ **GIGA: Garbage in – Garbage out**

Elements of data exploration – Checking for:

1. Outliers (e.g. boxplot)
2. Variance homogeneity (e.g. conditional boxplot)
3. Normal distribution (e.g. QQ-plot)
4. (Double) zeros (e.g. frequency plot)
5. Collinearity (e.g. pairwise scatterplots)
6. Relationship explanatory and response variable (e.g. scatterplots)
7. Spatial- or temporal autocorrelation (e.g. variograms)

A recommended read is:

Zuur, A.F; Ieno, E.N; Elphick, C.S (2009): A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution* 1: 3–14.

http://onlinelibrary.wiley.com/wol1/doi/10.1111/j.2041-210X.2009.00001.x/full

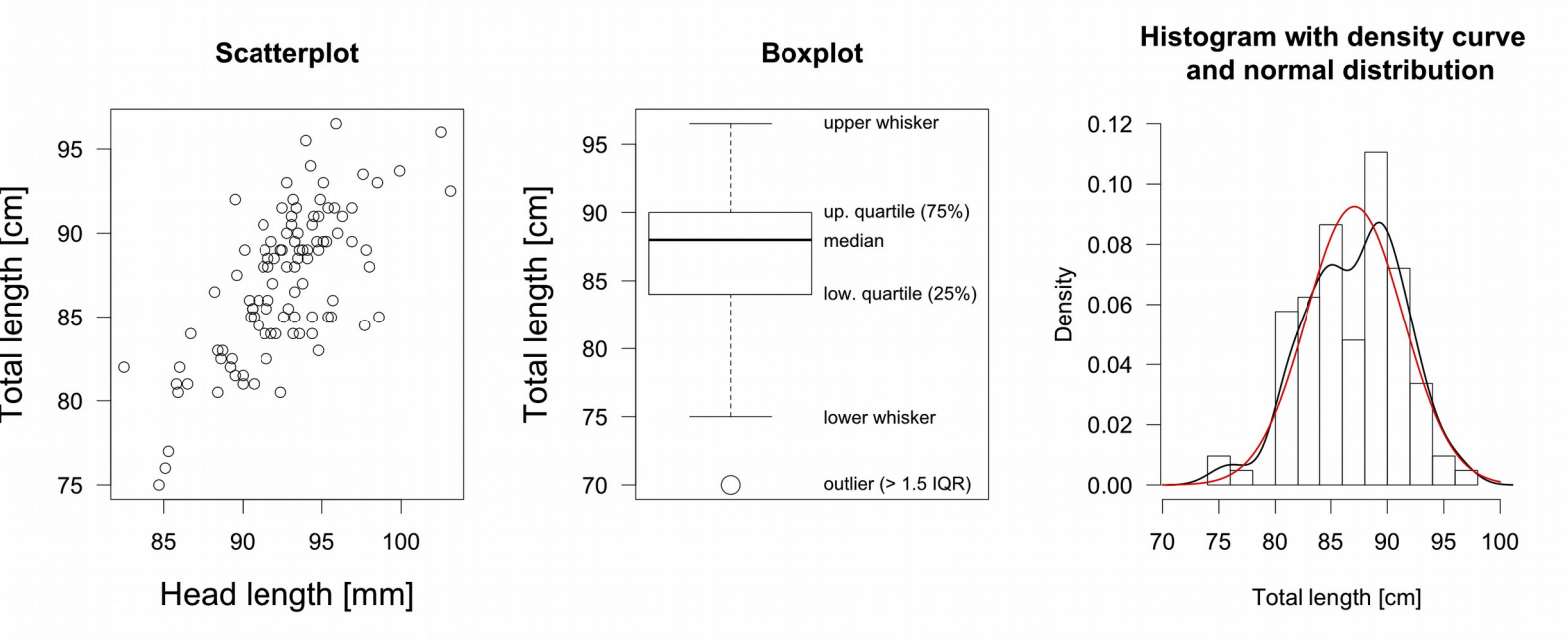
You have already encountered several of the elements of data exploration in the course and you will meet them later again.

Double zeros are often occurring for species data (i.e. absence of a species in pairs of sites) and may complicate interpretation. In addition, several zeros in the response variable can lead to biased parameter estimates and in such a situation models tailored for zero-inflated data should be used. For zero-inflated models see: Zuur, A.F & Ieno, E.N. (2016): Beginner's Guide to Zero-Inflated Models with R. Highland statistics.

http://highstat.com/index.php/beginner-s-guide-to-zero-inflated-models

Data exploration

Common plots for looking at the data



Linearity?
Collinearity?

Outliers?

Asymmetry of distribution?
Normality?

16

There are several rules of thumb as to what can be regarded as an outlier – but it remains more or less a subjective decision. John Tukey suggested to define Y as an outlier if: Y < (Q1 − 1.5 IQR) or Y > (Q3 + 1.5 IQR), where Q1 denotes the lower quartile, Q3 denotes the upper quartile, and IQR = (Q3 − Q1) denotes the interquartile range. In practice, the type of data, number of observations and knowledge about the data should be taken into account when deciding whether an observation is classified as outlier.

A beanplot represents an alternative to a boxplot that has several advantages. Beanplots have been introduced by Peter Kampstra: Kampstra P. 2008: Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. Journal of Statistical Software, Code Snippets. 28 (1): 1-9. Freely available at http://www.jstatsoft.org/v28/c01/
We will quickly look at a beanplot in the practical part.

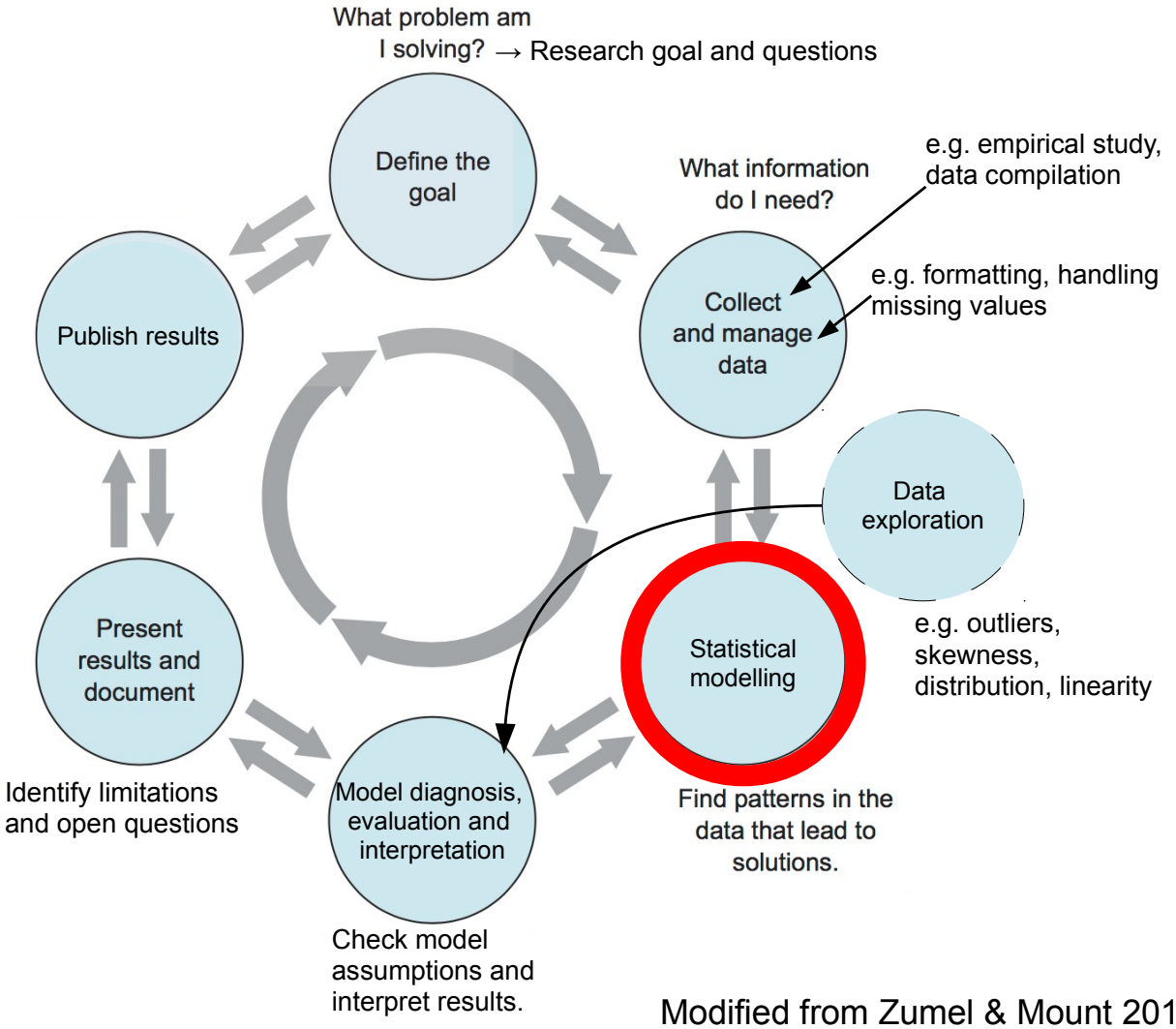
Statistical modelling, simulation and the linear model

Contents

1. Framework for data analysis and tools for data exploration
2. **Statistical modelling and simulation-based tools**
3. Permutation and Monte Carlo simulation
4. Bootstrapping
5. Cross-Validation and Bias-variance trade-off
6. Revisiting the linear model

17

Data analysis cycle

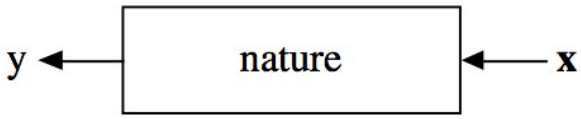


Modified from Zumel & Mount 2014: 6

18

Statistical modelling: The two cultures

Real world: Processes lead to association between **x** and **y**



Examples for goals of statistical modelling: predict unknown **y** from **x**, estimate how **x** is related to **y**

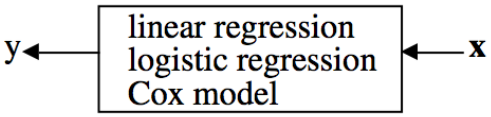
Data modelling culture
(classical statistics)

Algorithmic modeling culture
(machine learning)

Common data model

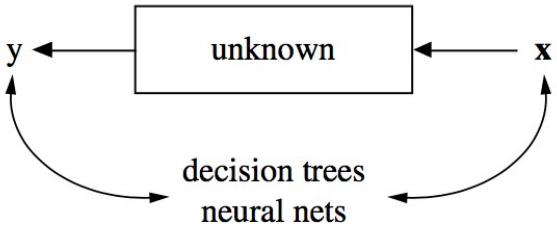
response variables = f(predictor variables, random noise, parameters)

Estimate parameters from data



Model validation: Check residuals

Find algorithm that operates on **x** to predict **y**



Model validation: Predictive accuracy

Breiman 2001 *Statistical Science* 16: 199

Breiman L. (2001) Statistical modeling: The two cultures. Statistical Science 16, 199–215.

The very readable debate is available here:
https://projecteuclid.org/download/pdf_1/euclid.ss/1009213726

Statistical modelling: the classical view

- Fit model to data to inform estimation, inference or prediction (e.g. estimate point or interval, test hypothesis)
  - Example: The arithmetic mean $\bar{x}$ is an estimate of the true population mean μ and s^2 is an estimate of the true variance σ^2
- Most models incorporate a deterministic (fixed effect) and a stochastic component (random effect)
  - Example: $y_i = b_0 + b_1 x_i + \epsilon_i \quad \text{with} \quad \epsilon \sim N(0, \sigma^2)$
- All models rely on assumptions → Model diagnosis
  - e.g. normal distribution, independence of observations
- Goodness of fit measures aid to choose between multiple models that fit the data
  - e.g. AIC, R^2, RMSE

Any observation contains signal and noise. In a statistical model, this relates to the fitted value and the residual.

Simulation-based approaches in data analysis

- Compatible with both cultures
- Infuses algorithm-based thinking into classical statistics
- Examples for simulation-based approaches for estimation, inference or model diagnosis in classical statistics:
  1. **Permutation test** → Permuting (shuffling) the data to derive null distribution. Mainly used for inference
  2. **Bootstrapping** → Randomly sampling subsets from the data with replacement. Mainly used for estimation of parameter distribution
  3. **Cross-validation** (CV) → Splitting data into sets (i.e. sampling without replacement). Mainly used for validation of predictive models

21

Statistical modelling, simulation and the linear model

Contents

1. Framework for data analysis and tools for data exploration
2. Statistical modelling and simulation-based tools
3. **Permutation and Monte Carlo simulation**
4. Bootstrapping
5. Cross-Validation and Bias-variance trade-off
6. Revisiting the linear model

22

Permutation test: Algorithm

Repeat *k* times:

1) Permute values in data set
2) Compute test statistic t^* for permuted data

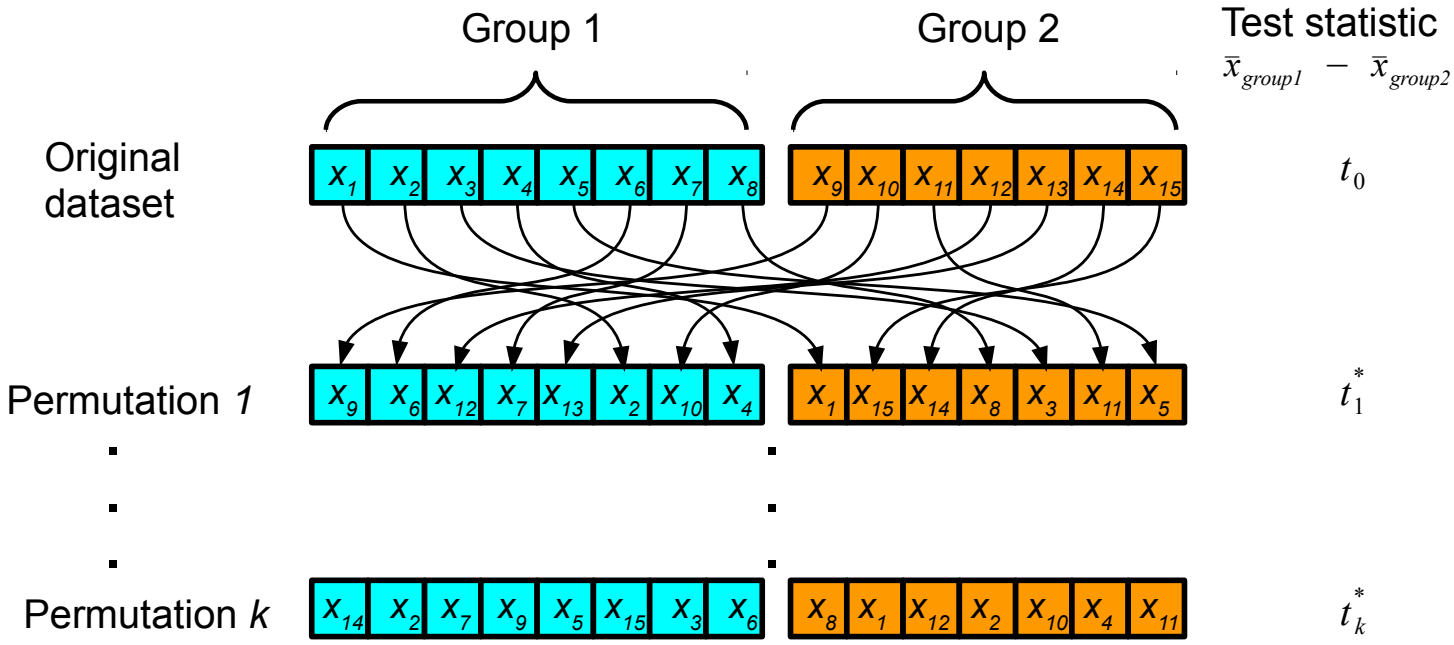
3) Compare test statistic t_0 to generated null distribution

23

Permutation test: Algorithm

Repeat *k* times
1) Permute values in data set
2) Compute test statistic *t** for permuted data
3) Compare test statistic t_0 to generated null distribution

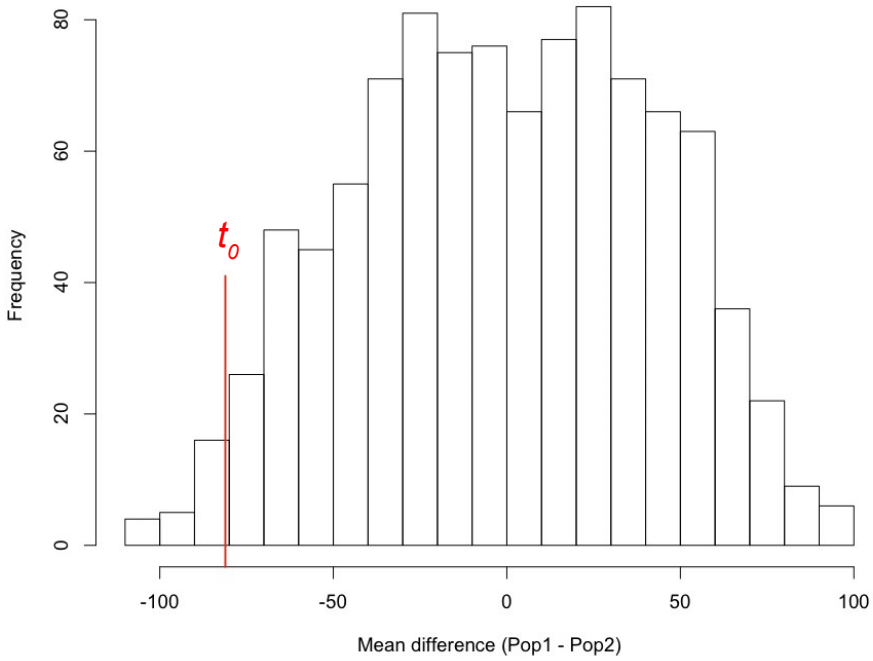
Example: Permutation test of difference in group mean



24

Permutation test: Generated distribution

$$p = \frac{\sum_{i=1}^{k} 1 \text{ if } t_i^* \leq t_0, else\, 0}{k+1}$$



- Test informs whether pattern in data is due to chance
- Inference regarding statistical population only valid if distribution of sample data matches actual distribution of statistical population → particularly problematic for small *n*

The p-value is computed as the fraction of test statistics *t**, which are based on permutated data, that are more extreme (lower or higher depending on the hypothesis) than the non-permuted test statistic.

If the sample distribution deviates from the actual distribution of the statistical population, the permutation test only allows to infer conclusions that apply to the data at hand. These may not be very interesting. For the example of the mean comparison, this would translate to being unable to test the null hypothesis:

$$H_0: \quad \mu_{group1} = \mu_{group2}$$

What a small sample size *n* is, depends on the context and no single number applies to all situations. For example, it will depend on the statistical distribution, statistical test etc. However, as a rule of thumb, sample sizes < 30 for a population are small. Still, much larger sample sizes can be required to reliably generalize from the permutation test to the statistical population.

Permutation test: Advantages and limitations

- Advantages
  - Free from distributional assumptions
  - Applicable to complex designs through restricting permutations
- Limitations
  - Generalisation to statistical population requires matching distribution
  - Statistical hypothesis testing can imply distributional assumptions that apply to the permutation test, if aiming to infer to the statistical population (e.g. testing for mean differences affected by variance)
  - Computationally intensive: Number of all possible permutations for a dataset is factorial *n*, i.e. *n*! (e.g. $35! \approx 10^{40}$) → Monte Carlo simulation

26 Legendre & Legendre 2012: 25ff

For instance, two null hypotheses are tested simultaneously (1. equality of mean, 2. equality of variance) when testing for differences among sample means. This dual aspect of classical tests such as analysis of variance or t-test also applies to the related permutation test and prohibits to draw unequivocal conclusions regarding the mean difference without consideration of variance equality.

Monte-Carlo simulation

- Uses repeated random sampling to solve problems probabilistically (even though they can be deterministic in reality)
- Permutation tests use random numbers to randomly permute data → approximate with MC simulation
- Legendre & Legendre (2012): use at least 10,000 permutations for inference

Entrance of casino in Monte Carlo, Monaco

Edvard Munch - At the Roulette Table in Monte Carlo

27

Name refers to the city, it was chosen as code name for a secret project in the context of nuclear weapon research in Los Alamos, USA.

The larger the number of MC-based permutations, the lower is the error when approximating the distribution of all possible permutations with the MC-based permutation.

Picture sources:
Photo of Casino
https://pixabay.com/de/spielbank-casino-monte-carlo-monaco-188882/

Picture of Munch:
https://upload.wikimedia.org/wikipedia/commons/1/1f/Edvard_Munch_-_At_the_Roulette_Table_in_Monte_Carlo_-_Google_Art_Project.jpg

Statistical modelling, simulation and the linear model

Contents

1. Framework for data analysis and tools for data exploration
2. Statistical modelling and simulation-based tools
3. Permutation and Monte Carlo simulation
4. **Bootstrapping**
5. Cross-Validation and Bias-variance trade-off
6. Revisiting the linear model

28

Bootstrapping: Idea and algorithm

- Inference on statistic *t* is based on sampling distribution
  - Ideally: Draw all or many samples from statistical population
  - Reality: Most frequently only one sample available
  - → **Idea:** Draw samples from an estimate of the statistical population (i.e. the sample) and use these to estimate property (e.g. variance) of the statistic *t*
- Algorithm:
  1) Draw random sample with replacement from data
  2) Compute statistic *t** for bootstrap sample

  } Repeat *k* times
  3) Use the *k* estimates to derive property of statistic
- Exhaustive bootstrapping ($k = n^n$) computationally demanding → approximate with Monte Carlo simulation
- Given todays computer power 10^4-10^5 simulations viable

29

The name bootstrapping alludes to the phrase “pulling oneself up by one’s bootstraps,” which has been voiced by the fictional character Baron Münchhausen.

In analogy to the permutation tests, the following applies to bootstrapping: The larger the number of MC-based bootstrap samples, the lower is the error when approximating the bootstrap distribution with the MC-based samples.

Bootstrapping: Example

Example: Bootstrap to the mean (to derive variance)

t (here: mean)

Original dataset

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

$\bar{x} = 8$

Sampling with replacement

BS sample 1

15	7	8	4	15	1	11	9	1	3	6	14	2	11	12

$\bar{x}^* = 7.93$

BS sample 2

6	13	2	10	12	5	7	10	1	13	8	8	15	3	10

$\bar{x}^* = 8.2$

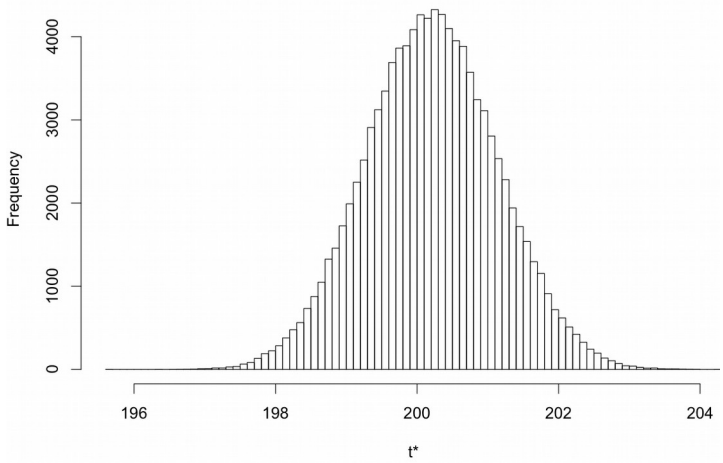
⋮

BS sample *k*

12	7	7	15	5	10	8	13	6	11	8	1	14	2	12

$\bar{x}^* = 8.73$

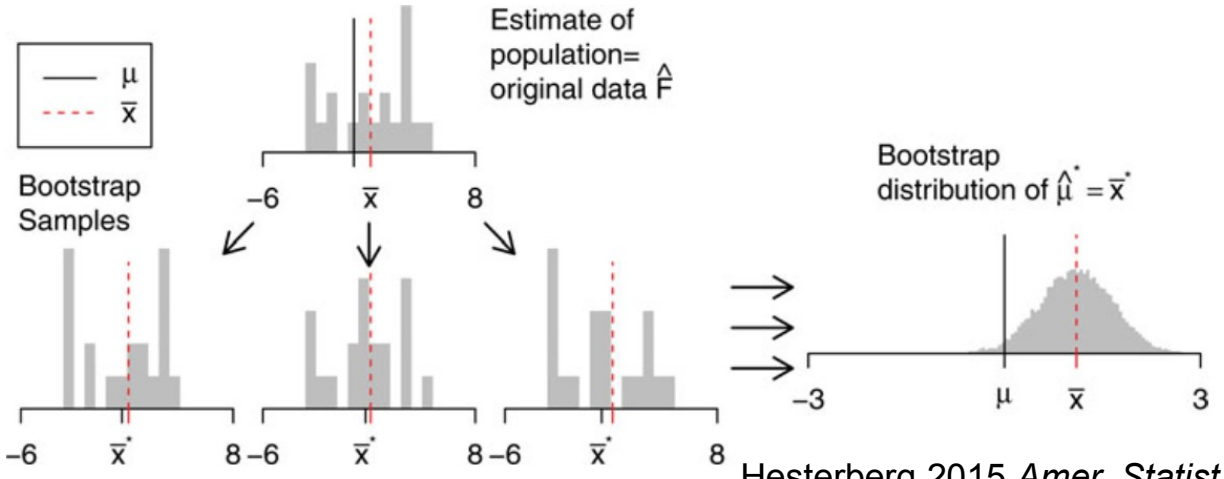
Distribution of statistic *t*



30

Bootstrapping: Limitations

- Do not use for hypothesis testing
- No distributional assumptions implied, but not reliable for all distributions, particularly at small *n* (see Hesterberg 2015)
- Small *n*: use adjusted bootstrap percentiles (Bca) or switch to parametric statistics (allow for additional assumptions)
- Bootstrap does not improve estimate of population parameter μ, centred at $\bar{x}$



Hesterberg 2015 *Amer. Statist.* 69:371

Bootstrapping is generally less accurate than permutation tests for hypothesis testing.

BCa corrects for bias and skewness in the distribution of bootstrap estimates.

A very nice introduction and overview on bootstrapping is provided by:

Hesterberg T.C. (2015) What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum. The American Statistician 69, 371–386.

Freely available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4784504/pdf/utas-69-371.pdf

Statistical modelling, simulation and the linear model

Contents

32

Cross-validation (CV)

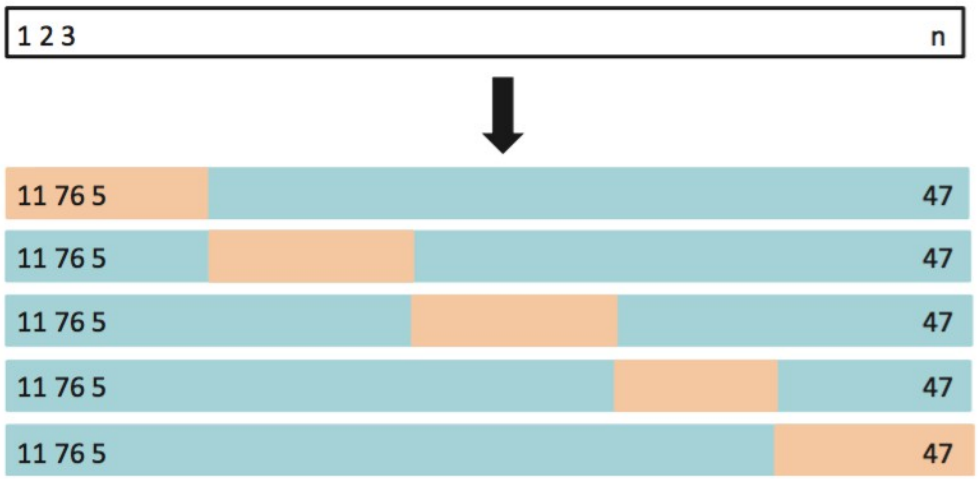
- Objective: Evaluate predictive accuracy of a fitted model
- Can be checked if independent training data (used to fit model) and test data (new data) are available → Rare case
- **Idea:** Split the available data into training and test set and predict the (known) observations in the test set from a model fitted with the training data
- Algorithm:
  1. Draw *k* random samples without replacement from data
  2. For each *k:*
     1. Fit the model to the other *k*-1 parts
     2. Predict *k* from model and calculate the prediction error
  3. Calculate prediction error as average over the *k* estimates

Predictive accuracy measures the accuracy of predictions for new data.

CV is typically used in validation, but can also be used as goodness-of-fit measure to guide parameter estimation (see shrinkage methods later).

Cross-validation (CV)

Example: k = 5



- Problem of choosing k:
  - k = n (Leave-one-out CV predicts each observation from all others) → low bias, but high variance
  - k = 2 (split data into half) → low variance, but high bias
- k typically set to 5 or 10

Taken from James *et al.* 2013: 181

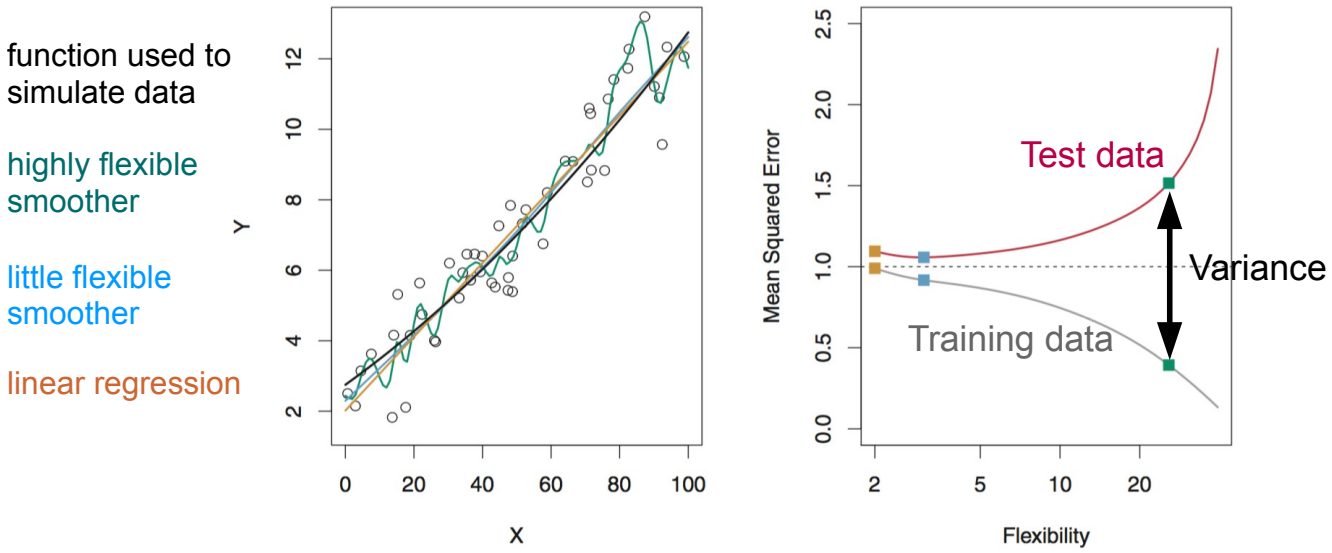
The bias-variance trade-off will be discussed in detail on the next slides. In brief, there is a trade-off between bias (error when estimating the 'true' prediction accuracy of the sample data) and variance (variability of the error when estimating new (test) data). If we use a major fraction of the data (extreme case: $k = n$, where we use n-1 observations) in model fitting, the error of estimating the prediction accuracy of the full data is probably very low (low bias). However, the variability of the error when predicting a few (or only one for $k = n$) observations from different training sets is most likely high, which translates to a high variance. Conversely, if we use only half of the data (k = 2) in model fitting, we decrease the variance. In other words, the error when predicting the test set is most likely similar for the two training sets. But this comes at the cost of bias. In the case of k = 2, we are estimating the predictive accuracy from only a fraction of the data, whereas in practice all observations will be used in prediction. The prediction accuracy estimated from the fraction of the data is likely to differ (i.e. lower or higher) from that of the complete data set, i.e. exhibit bias. Thus, the bias increases when the relative size of the training set in CV decreases.

k is typically set to 5 or 10, i.e. the data is partitioned in 5 or 10 groups during CV as a compromise between bias and variance. Leave-one-out CV is considered less reliable than 5- or 10-fold CV (see Harrell 2015: 172).

Bias-variance trade-off

Definitions in context of model validation:

- **Bias:** error when approximating training data
- **Variance:** variability in error when approximating test data



Higher flexibility (higher *k* in CV) → lower error for training data (i.e. lower bias), but variance will increase from some point

35

Taken from James *et al.* 2013: 33

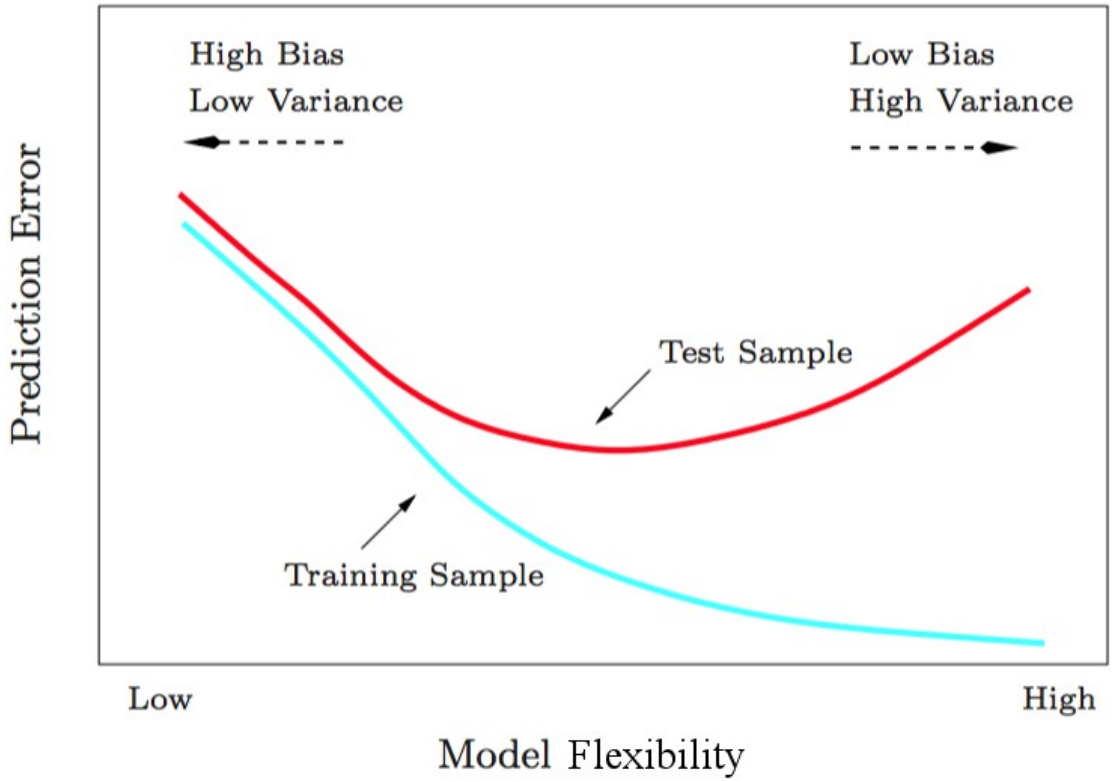
The left figure displays the fit of different models to data originating from the function plotted in black.

The models rank regarding bias: linear regression > little flexible smoother > highly flexible smoother.

Regarding variance (see right figure), the ranking is: highly flexible smoother > little flexible smoother > linear regression.

Bias-variance trade-off

Higher flexibility (higher *k* in CV) → lower error for training data (i.e. lower bias), but variance will increase from some point → Optimise combined error



Taken from Hastie, Tibshirani and Friedman 2011: 38

For a mathematical derivation of the bias-variance trade-off see Matloff(2017): 48f.

Statistical modelling, simulation and the linear model

Contents

1. Framework for data analysis and tools for data exploration
2. Statistical modelling and simulation-based tools
3. Permutation and Monte Carlo simulation
4. Bootstrapping
5. Cross-Validation and Bias-variance trade-off
6. **Revisiting the linear model**

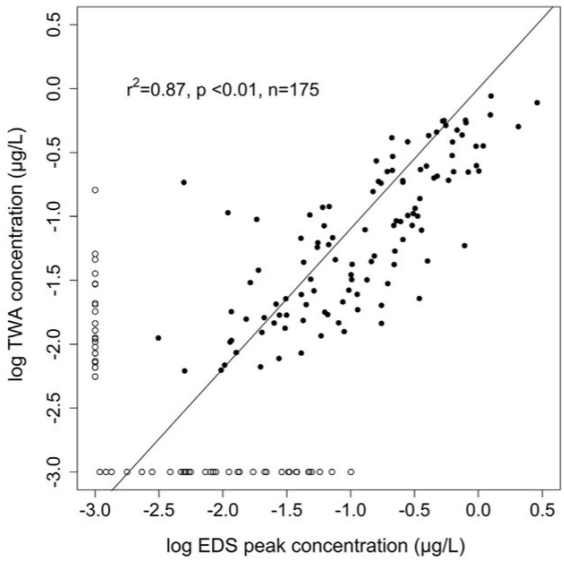
37

Relationship between two continuous variables: linear regression model

- Bivariate relationship between an explanatory variable and a response variable with:

$$y_i = b_0 + b_1 x_i + \epsilon_i \qquad \text{with} \quad \epsilon \sim N(0, \sigma^2)$$

- Example: Can we approximate pesticide runoff concentrations with passive sampling?



Fernandez et al. 2014

The figure shows the concentrations of pesticides measured with passive samplers (TWA concentrations) and event-driven samplers (EDS peak concentrations). The concentrations are relatively similar for pesticides that were quantified in samples from both sampling devices, i.e. follow almost a 1:1 relationship, which means that passive sampling is a suitable technique to approximate peak concentrations. The non-filled points indicate cases where a compound was only quantified in samples of one of the sampling devices. Further details can be found in:

Fernández D., Vermeirssen E.L.M., Bandow N., Muñoz K. & Schäfer R.B. (2014) Calibration and field application of passive sampling for episodic exposure to polar organic pesticides in streams. Environmental Pollution 194, 196–202.

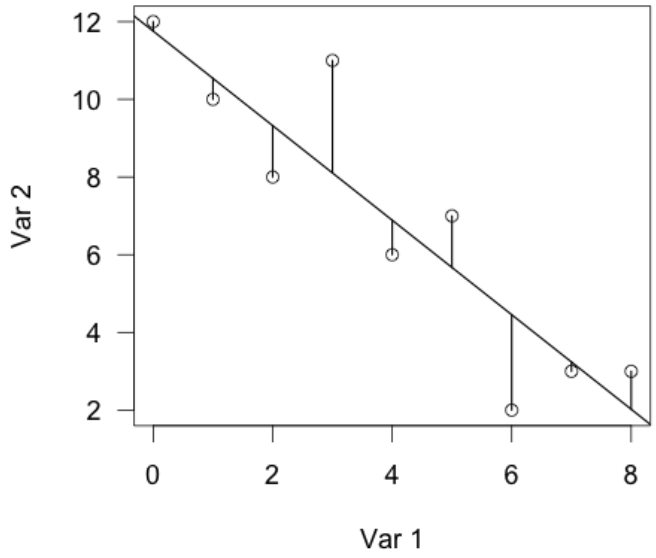
Relationship between two continuous variables: linear regression model

- Bivariate relationship between an explanatory variable and a response variable with:

$$y_i = b_0 + b_1 x_i + \epsilon_i \qquad \text{with} \quad \epsilon \sim N(0, \sigma^2)$$

- Aim: minimise ε (also called error sum of squares: SSE)

$$\text{SSE} = \sum_1^n (y_i - b_0 - b_1 x_i)^2$$



39

The fitted values for the regression model, i.e. the estimates for *y* are given as:

$$\hat{y}_i = b_0 + b_1 x_i$$

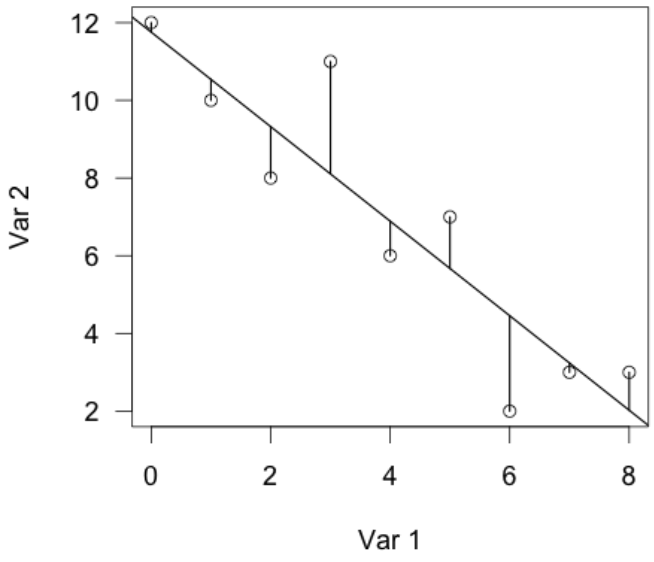
A measure that is similar to the SSE is the Mean Squared Error (MSE), which is given as:

$$MSE = \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

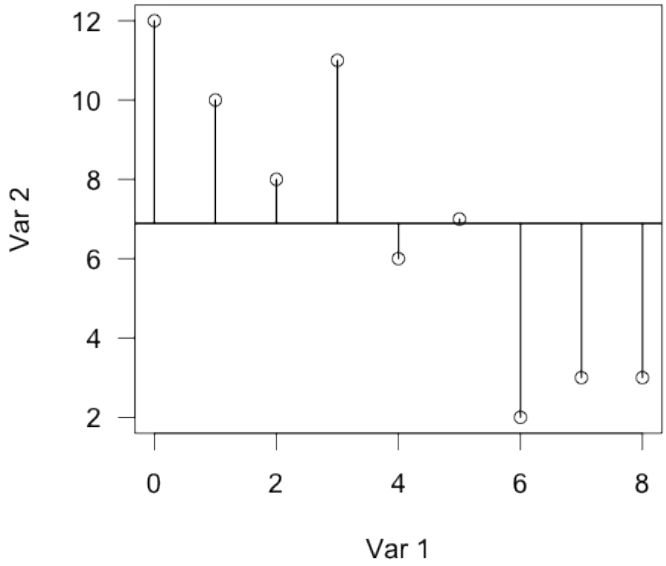
The denominator accounts for the number of explanatory variables *p* and the intercept and requires adjustment in case of no-intercept models (i.e. the denominator would turn into *n-p*). MSE is typically used when assessing the quality of the estimation. In case that the predictive accuracy is assessed, the Mean Squared Prediction Error (MSPE) is used for new observations y_{n+1} to y_m.

Linear regression model

$$SSE = \sum_{1}^{n} (y_i - b_0 - b_1 x_i)^2$$



$$SSY = \sum (y - \bar{y})^2$$



SSY = SSR + SSE

Total variation

Explained variation

Unexplained variation

% of explained variance:

$$R^2 = \frac{SSR}{SSY}$$

$$adj.\ R^2 = 1-(1-R^2)\frac{n-1}{n-p-1}$$

SSR refers to regression sum of squares and can be calculated as the summed quadratic differences between the fitted values and the mean for the response variable. SSE and SSY are defined as for the analysis of variance (ANOVA). Indeed, both ANOVA and linear regression are linear models and in R most functions apply to either of them.

The square root of the R^2 has the same absolute value as the Pearson correlation coefficient. The R^2 is typically used to measure the goodness of fit of a regression model.

The adjusted R^2 should be preferred over the normal R^2 as it takes the number of explanatory variables *p* into account (*n* is sample size). The denominator is *n*-*p*-1 accounting for the number of *p* and the intercept. However, this is more important in the case of multiple linear regression.

Linear regression model

- Assumptions:
  - Linear relationship (graphical diagnostics)
  - Normal distribution of error (graphical diagnostics)
  - Variance homogeneity (graphical diagnostics)
  - Independence of errors (graphical diagnostics)
- If one or more assumptions not met, alternatives include:
  - Generalised linear model, Generalised least squares, Mixed models
  - Variable transformation (but using an appropriate model such as a Generalised linear model is usually the better option)

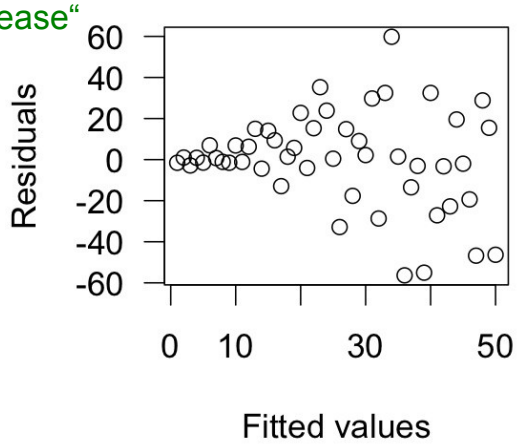
Although hypothesis tests for checking the assumptions exist, most textbooks recommend graphical diagnostics. For data that is spatially or temporally structured or data that is nested/ hierarchically structured, the independence of errors assumption is typically violated. Since time series and spatial data are beyond the scope of this course (and is discussed in the "Advanced GIS" course), I refer to Faraway (2015): Linear models in R. p.81-83 for diagnostic tools to spot serial correlation or see Plant (2012): Spatial data analysis in ecology and agriculture using R. Spatial and temporal structure can be incorporated into the model using generalised least squares (see chapter 4 in Zuur, A. F. et al. 2009: Mixed effects models and extensions in ecology with R. Springer: New York). Nested/hierarchically structured data can be modelled with mixed effect models, which are discussed in the first part of this course. We will also discuss generalised linear models later in this course. Variable transformation and robust regression are discussed in many textbooks (e.g. Maindonald & Braun 2010, Quinn & Keough 2002) and are beyond the scope of this course (but variable transformation has been extensively discussed in the preceding course of univariate statistics).

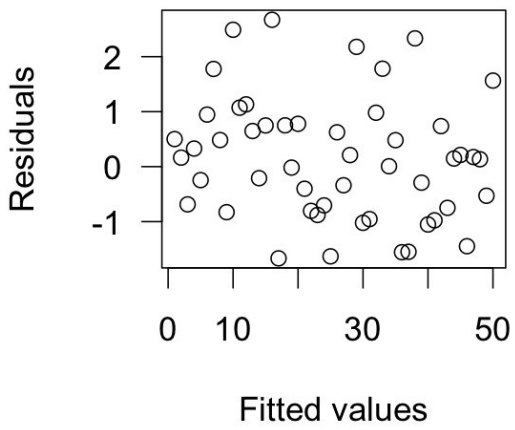
In linear regression analysis, we usually do not take the measurement error in *x* into account. This is discussed in detail in Warton, D.I., Wright, I.J., Falster, D.S., and Westoby, M. (2006). Bivariate line-fitting methods for allometry. Biological Reviews 81, 259-291.

Warton et al. (2006) also provide information on alternatives to linear regression that should be used if the measurement error is relevant.

Model diagnostics: Variance homogeneity

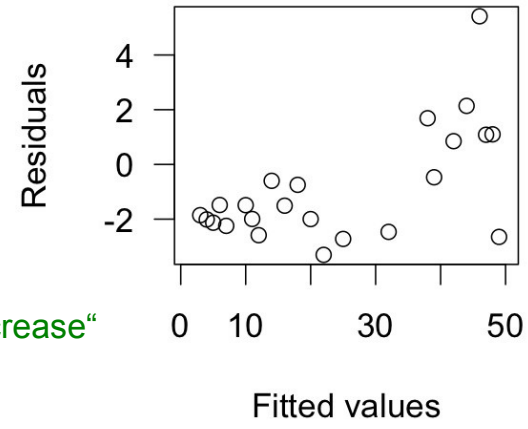
„strong increase“



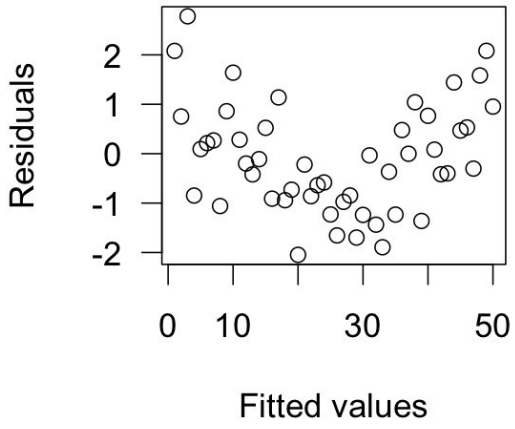


„normal“

Residuals vs. fitted values plots



„slight increase“



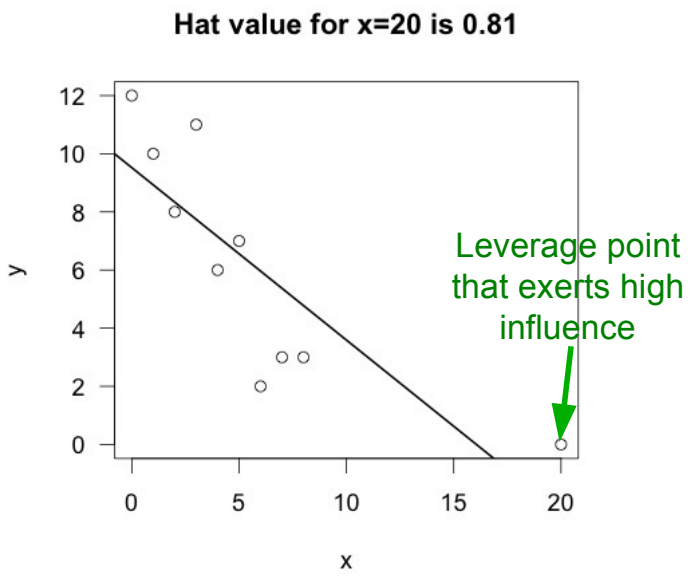
„non-linear“

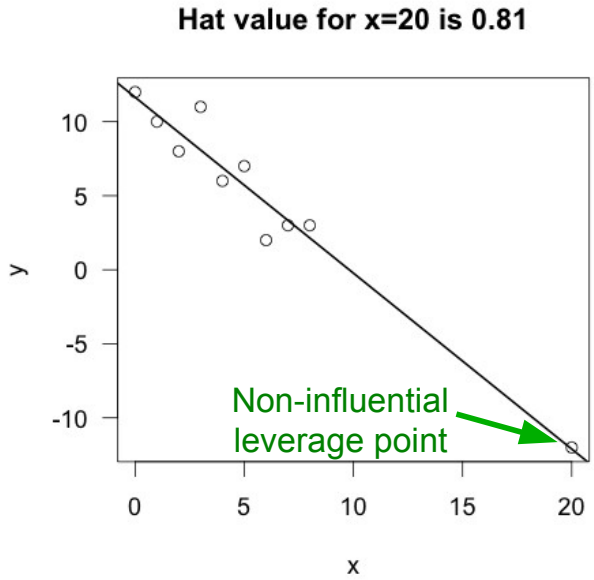
The graphical diagnostics for checking variance homogeneity are the same for linear regression and ANOVA (and other linear models), but the *x*-axis of ANOVA (and *t*-test) would display the factor levels and, consequently, the plots would not describe a continuous pattern.

The displayed residuals-fitted values plots can be used to check whether the assumption of variance homogeneity, also termed homoscedasticity, (and the assumption of linearity in the case of regression) holds. If the residuals are not randomly distributed (upper right) but display patterns, this may indicate variance heterogeneity, also termed heteroscedasticity, (bottom and top left) or non-linearity (bottom right). In case of departures from the assumption of homoscedasticity, generalized least squares, generalized linear or additive models can represent a suitable alternative for continuous data. Depending on the data, data transformation or weighting observations may also be used to alleviate the issue, though transformation should only be considered if the data cannot be modelled non-transformed.

Further model diagnostics

Leverage points (predictor outlier)





How to deal with leverage points/outliers?

- Check whether values are plausible
- Check robustness of model results when removing observations
- Fit different statistical model or transform data

Beside checking for assumptions, model diagnostics are used to detect influential points, leverage points (predictor outliers) and model outliers (outlier in response variable indicating model failure). Influential points exercise high influence on the model fit, but may not be outliers. Leverage points and outliers do not fit the model, but are not necessarily influential.

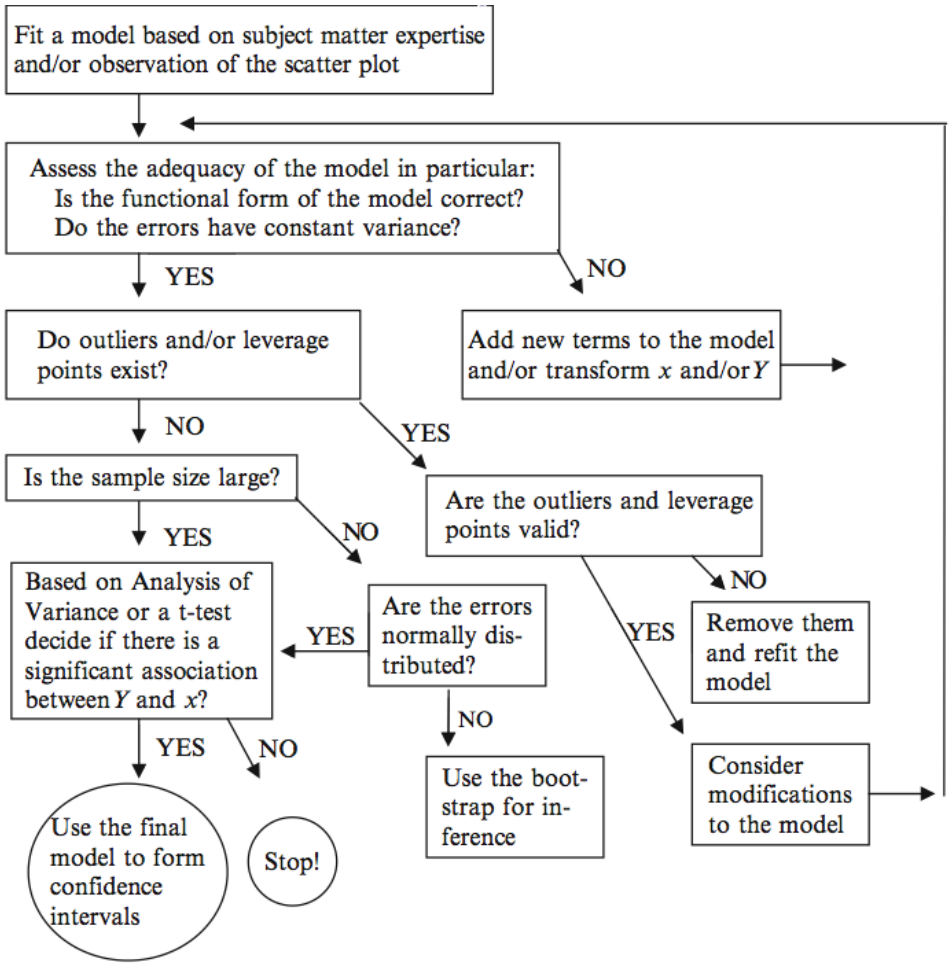
Leverage points (1) exercise high influence on the fitted *y* (but not necessarily on the model fit) and (2) are distant from the other *x*-values. The leverage is calculated in terms of so-called hat values, which will be explained later in the course, and the average hat value is *p*/*n*, where *p* is the number of parameters in the model (including the intercept) and *n* is the number of observations. Faraway (2015):83 and Sheater (2009) suggest to look at points with hat values > 2 *p*/*n* more closely. However, hat values are independent of the response variable *y* and graphical inspection is most suitable to check whether a high leverage point is really problematic. A nice illustration of leverage can be found here: http://www.rob-mcculloch.org/teachingApplets/Leverage/index.html.

Outliers in the response variable can be identified with studentized residuals. Here, points that deviate more than 2 standard deviations from the regression line may be considered as outlier (see Sheater 2009: p.60). There are of course different rules of thumb as to what can be regarded as an outlier – but it remains more or less a subjective decision. John Tukey suggested to define Y as an outlier if: Y < (Q1 − 1.5 IQR) or Y > (Q3 + 1.5 IQR), where Q1 denotes the lower quartile, Q3 denotes the upper quartile, and IQR = (Q3 − Q1) denotes the interquartile range. Hence, you could use a boxplot to identify an outlier.

Another important measure in diagnostics plots represents Cooks distance. Cooks distance measures the influence of observations on the model fit by calculating the combined effect of leverage and of the magnitude of the residual. The higher Cooks distance the larger the change in model fit when the point is removed from the model. A point with a high Cooks distance tends to be either an outlier or a leverage point or both. There are different rules of thumb as to when consider a point as influential (e.g. Cooks *D* > 1 or Cooks *D* > 4/*n*-2), but in practice it is important to look for gaps in the values of Cooks distance (Sheater 2009: p.68).

Methods such as robust regression and quantile regression have been developed to reduce the influence of influential points. However, they are outside the scope of this course.

Flowchart for simple linear regression



Taken from Sheather 2009: p.103

Note that this flowchart only serves the purpose of giving orientation, whereas the suggestions may differ from the suggestions presented in this lecture. For example, if the errors do not have constant variance, the flowchart suggests the addition of new terms to the model and/or variable transformation. However, we have discussed in the lecture that other model types such as the Generalized linear model can be more appropriate for the data. Hence, before transformation of data, you should check whether the data can be directly modelled using a Generalized linear model or others (cf. Szöcs & Schäfer 2015). Note also that the bootstrap may not be reliable for small sample sizes, see the part on bootstrapping.

Szöcs E. & Schäfer R. (2015) Ecotoxicology is not normal. Environmental Science and Pollution Research 22, 13990–13999.

Simulation-based approaches to simple linear regression

- Predictive accuracy measured with Mean square prediction error (MSPE):

$$\text{MSPE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$ for the new observations 1 to *m*

- Cross-validation (CV): Calculate CV-MSPE and CV-R^2
- Bootstrapping (BS) in regression analysis:
  - of residuals: BS residuals, add to $\hat{\boldsymbol{y}}$ to generate new $\boldsymbol{y}^*$ and calculate regression coefficients → ***x*** fixed
  - of cases: BS complete cases and calculate regression coefficients → ***x*** random
  - If ***x*** and ***y*** random sample (e.g. *x* not fixed in experiment), residuals correlated or exhibit non-constant variance → BS cases

Bold variables indicate vectors.

The mean square prediction error measures how well a new observation is predicted. For the relationship with other error measures see here.

The algorithm for residual bootstrapping is easiest understood, when reformulating the equation for the ordinary linear regression model (see here for details) to:

$$\epsilon_i = y_i - b_0 + b_1 x_i \Leftrightarrow \epsilon_i = y_i - \hat{y}_i$$

The bootstrap samples (samples with replacement) are drawn from the the *n* residuals $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ yielding to a bootstrap sample of residuals $\epsilon_1^*, \epsilon_2^*, \ldots, \epsilon_n^*$

These bootstrapped residuals are added to the vector of fitted responses ($\hat{\boldsymbol{y}}$) to obtain a vector of new responses ***y****: $y_i^* = \hat{y}_i + \epsilon_i^*$

These new responses are used to calculate new bootstrapped regression coefficients (i.e. b_0^*, b_1^*). The procedure is repeated 1,000 to 10,000 times and as usual in bootstrapping delivers the distribution for a test statistic *t** (here for b_0 and b_1).

For bootstrapping cases, pairs of *x* and *y* are bootstrapped and then the regression model is fitted, also providing bootstrapped regression coefficients (i.e. b_0^*, b_1^*).

Now when to use what? In case that the residuals exhibit non-constant variance or are correlated, the bootstrap sample does not preserve the properties of the population sample and bootstrapping of cases should be preferred. However, if the observations for the predictors (***x***) have not been drawn randomly, but are fixed (for example, fixed concentration levels in an experiment), bootstrapping residuals should be preferred as it preserves these original x. For further details see Fox (2015): 658-660 and Hesterberg (2015) *Americ. Statist.* 69: 371–386.

45

Exercise

We will work with the data set “possum” that includes biometric measurements of possums in Victoria, Australia.

Conduct a linear regression analysis, diagnose and interpret the model and apply simulation-based approaches.

46