Multiple regression analysis

Contents

1. **Model purpose, learning targets and mathematical basics**
2. Model definition, case study and modelling scheme
3. Goodness of fit and model selection
4. Stepwise model selection
5. The LASSO
6. Data preparation & Multicollinearity
7. Model diagnosis and analysis, small sample sizes and general tutorial

Multiple linear regression model

- Used to establish a (linear) relationship between multiple explanatory (predictor) and one response variable
- Research goals:
  - Prediction (predict response to explanatory variables)
  - Explanation (identify important explanatory variables)
  - Estimation (determination of effect size)
  - Hypothesis testing.

Multiple linear regression model

- Example: Which variable(s) do best explain the response of different groups of organisms?

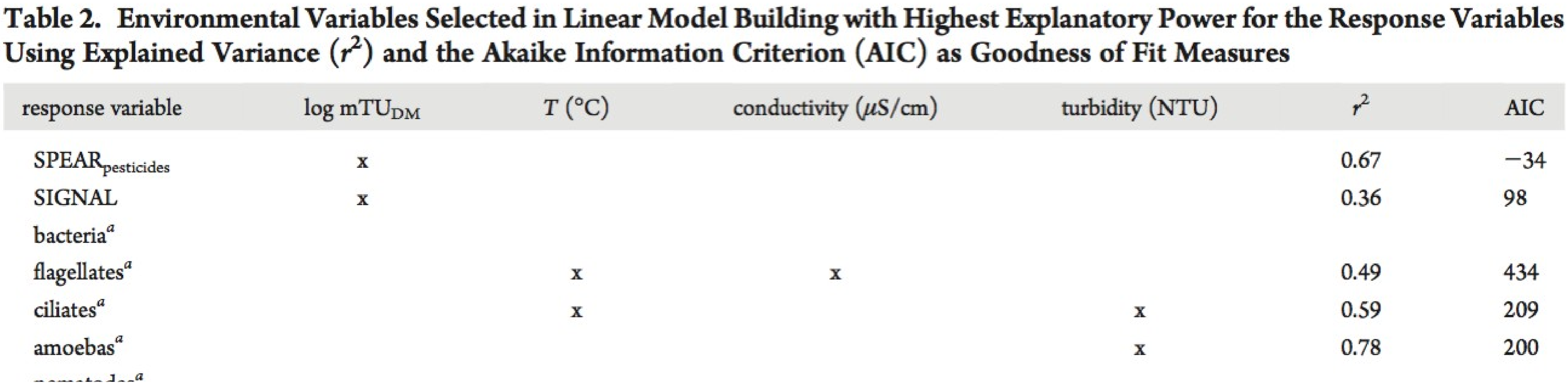


Table 2. Environmental Variables Selected in Linear Model Building with Highest Explanatory Power for the Response Variables Using Explained Variance (r^2) and the Akaike Information Criterion (AIC) as Goodness of Fit Measures

response variable	log mTU_{DM}	T (°C)	conductivity (μS/cm)	turbidity (NTU)	r^2	AIC
$SPEAR_{pesticides}$	x				0.67	−34
SIGNAL	x				0.36	98
bacteria[a]						
flagellates[a]		x	x		0.49	434
ciliates[a]		x		x	0.59	209
amoebas[a]				x	0.78	200
nematodes[a]						

Schäfer et al. 2011

Learning targets

- Describe the mathematical basis of multiple linear regression
- Explain and apply strategies to identify the best-fit model
- Interpret models and apply variable-importance measures
- Describe the modelling steps

Learning targets and study questions

- Describe the mathematical basis of multiple linear regression
  - What is the Hat matrix?
- Explain and apply strategies to identify the best-fit model
  - How does the research question influence the selection of the best-fit model?
  - Which goodness of fit measures can be used to compare models?
  - List the model selection strategies and criticise step-wise model selection.
  - Categorise and justify approaches to improve and replace stepwise model selection.

Learning targets and study questions

- Interpret models and apply variable-importance measures
  - Which types of model diagnostics are required for multiple regression models?
  - Outline methods to diagnose and to deal with collinearity.
  - Compare methods to check the relative importance of variables.
- Describe the modelling steps

Mathematical basics I

Matrix algebra represents the basis of multivariate statistics

Matrix: $\boldsymbol{A}_{m,n} = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,n} \\ a_{2,1} & a_{2,2} & a_{2,n} \\ a_{m,1} & a_{m,2} & a_{m,n} \end{pmatrix}$ with *m* rows and *n* columns

`matrix()`

`nrow(), ncol()`

Transpose matrix: $\boldsymbol{A}_{2,2} = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix}$ yields $\boldsymbol{A}^{t}_{2,2} = \begin{pmatrix} a_{1,1} & a_{2,1} \\ a_{1,2} & a_{2,2} \end{pmatrix}$

`t()`

Addition and substraction of $\boldsymbol{a}_{mn}$ and $\boldsymbol{b}_{mn}$

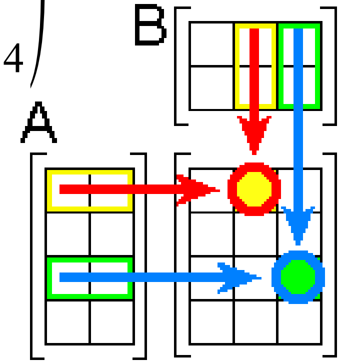
`+,-`

$$\begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} + \begin{pmatrix} b_1 & b_2 \\ b_3 & b_4 \end{pmatrix} = \begin{pmatrix} a_1 + b_1 & a_2 + b_2 \\ a_3 + b_3 & a_4 + b_4 \end{pmatrix}$$

Mathematical basics I

Matrix multiplication %*%

$$\begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} * \begin{pmatrix} b_1 & b_2 \\ b_3 & b_4 \end{pmatrix} = \begin{pmatrix} a_1 * b_1 + a_2 * b_3 & a_1 * b_2 + a_2 * b_4 \\ a_3 * b_1 + a_4 * b_3 & a_3 * b_2 + a_4 * b_4 \end{pmatrix}$$



$\boldsymbol{A} * \boldsymbol{B}^t$ a%*%t(b) ; tcrossprod(a,b)

$\boldsymbol{A}^t * \boldsymbol{B}$ t(a)%*%b ; crossprod(a,b)

Identity matrix for 2x2 matrix: $\boldsymbol{E} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

For $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ the inverse is: $A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

solve()

with the following characteristics:

$$\boldsymbol{A}^{-1} * \boldsymbol{A} = \boldsymbol{E}$$
$$\boldsymbol{A} * \boldsymbol{A}^{-1} = \boldsymbol{E}$$

Multiple regression analysis

Contents

1. Model purpose, learning targets and mathematical basics
2. **Model definition, case study and modelling scheme**
3. Goodness of fit and model selection
4. Stepwise model selection
5. The LASSO
6. Data preparation & Multicollinearity
7. Model diagnosis and analysis, small sample sizes and general tutorial

Defining the multiple linear regression model

- Relationship between several explanatory variables and a response variable with:

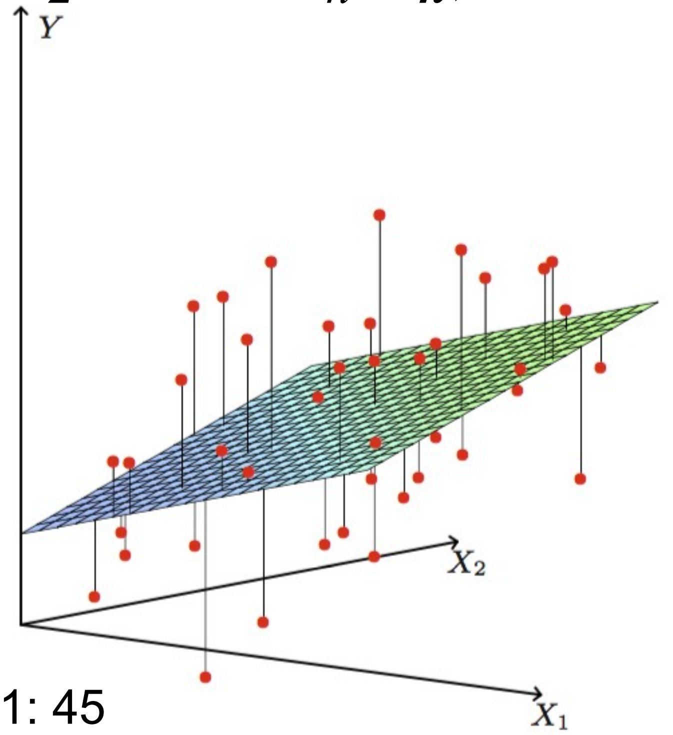
$$\boldsymbol{y} \;=\; b_0 + b_1\boldsymbol{x_1} + b_2\boldsymbol{x_2} + \ldots + b_m\boldsymbol{x_m} + \epsilon_i \qquad \text{with} \;\; \epsilon \sim N(0, \sigma^2)$$

- Aim is to minimise ε, which is SSE in the linear regression model:

$$\text{SSE} = \sum (\boldsymbol{y} - b_0 - b_1\boldsymbol{x_1} - b_2\boldsymbol{x_2} - \ldots - b_n\boldsymbol{x_n})^2$$

<u>Example</u>

Regression plane for two predictors:



Taken from Hastie, Tibshirani and Friedman 2011: 45

Defining the multiple linear regression model

Model in matrix form

$$\boldsymbol{y} = b_0 + b_1\boldsymbol{x_1} + b_2\boldsymbol{x_2} + \ldots + b_m\boldsymbol{x_m} + \epsilon \quad \Leftrightarrow \quad \hat{\boldsymbol{y}} = b_0 + b_1\boldsymbol{x_1} + b_2\boldsymbol{x_2} + \ldots + b_m\boldsymbol{x_m}$$

Full notation for the number of observations (*n*):

$$\begin{aligned} \hat{y}_1 &= b_0 + b_1 x_{11} + b_2 x_{12} + \ldots + b_m x_{1m} \\ \hat{y}_2 &= b_0 + b_1 x_{21} + b_2 x_{22} + \ldots + b_m x_{2m} \\ &\vdots \\ \hat{y}_n &= b_0 + b_1 x_{n1} + b_2 x_{n2} + \ldots + b_m x_{nm} \end{aligned} \quad \xrightarrow{\text{matrix}} \quad \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1\mathrm{m}} \\ 1 & x_{21} & x_{22} & \cdots & x_{2\mathrm{m}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{pmatrix}$$

$$\Big\downarrow$$

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b}$$

b can be estimated as (see Legendre & Legendre 2012: 88):

$$\boldsymbol{b} = (\boldsymbol{X^t}\boldsymbol{X})^{-1}(\boldsymbol{X^t}\boldsymbol{y})$$

Substitution yields:

$$\hat{\boldsymbol{y}} = \underbrace{\boldsymbol{X}(\boldsymbol{X^t}\boldsymbol{X})^{-1}(\boldsymbol{X^t}}_{\text{Hat matrix}}\boldsymbol{y})$$

Case study: Ostracods

Which patterns and factors control the diversity of marine arctic ostracods?

136 ostracod samples from different regions
10 explanatory variables

Aim: Identify most important explanatory variables for diversity of marine ostracods.
→ For explanation search for most parsimonious model

"It is futile to do with more things that which can be done with fewer"

http://www.phdcomics.com/comics/archive.php?comicid=1237

Modelling scheme (mainly for explanation)

Which variables should be included in the multiple regression model?

Full: Model 1: *Diversity ~ Var a, Var b, Var c, Var d, Var e*

Reduced: Model 2: *Diversity ~ Var a, Var b, Var c*

Model 3: *Diversity ~ Var a, Var b, Var c, Var d*

⋮

Model n: *Diversity ~ Var b, Var d, Var e*

Strategies

- Compare pre-specified models
- Best subset model selection
- Stepwise model selection
- Shrinkage methods

↓

Quantitative model comparison via goodness of fit measures

Best-fit model

- model diagnostics
- model validation

Multiple regression analysis

Contents

1. Model purpose, learning targets and mathematical basics
2. Model definition, case study and modelling scheme
3. **Goodness of fit and model selection**
4. Stepwise model selection
5. The LASSO
6. Data preparation & Multicollinearity
7. Model diagnosis and analysis, small sample sizes and general tutorial

Goodness of fit (GOF) measures

- R^2 or adj. R^2
  - R^2 increases with each additional variable in model (also noise)
  - adj. R^2 should be preferred for model comparison, because it penalises for additional variables
- Information theoretic goodness of fit measures:

$$\mathrm{AIC} = n \log\left(\frac{SSE}{n}\right) + 2p + const.$$

n = sample size

p = parameters in model

$$\mathrm{AIC_c} = \mathrm{AIC} + \frac{2p(p+1)}{n-p-1}$$

$$\mathrm{BIC} = n \log\left(\frac{SSE}{n}\right) + \ln(n)\, p + const.$$

  - The lower the value, the better the model

Model selection strategies

How to identify the best-fit model?

- Ideally: Comparison of a limited number of *a priori* specified models
- Traditionally used: (1) best subset and (2) stepwise selection
- (1) Best subset: 2^m (m = number of variables) w/o interactions → computationally demanding
- (2) Stepwise model selection requires start model, computes all models for next step (inclusion or exclusion of variable) and selects best model. Algorithm is repeated until change of included variables would reduce model fit.
- Stepward selection procedures: backward (variable elimination), forward (variable inclusion), both (combined)

Multiple regression analysis

Contents

1. Model purpose, learning targets and mathematical basics
2. Model definition, case study and modelling scheme
3. Goodness of fit and model selection
4. **Stepwise model selection**
5. The LASSO
6. Data preparation & Multicollinearity
7. Model diagnosis and analysis, small sample sizes and general tutorial

Stepwise model selection

- Coupled to hypothesis testing:
  - Partial *F*-test for differences in explained variance:

$$\frac{(SSE_{reduced\,model} - SSE_{full\,model}) / (df_{reduced\,model} - df_{full\,model})}{SSE_{full\,model} / df_{full\,model}}$$

  - *t*-test for predictor: $H_0:\ \beta_i = 0 \qquad H_1:\ \beta_i \neq 0$
    Predictors rejected where null hypothesis cannot be rejected
  - Multiple testing (e.g. a series of tests on same data) leads to inflation of *p*-values (computed *p*-values biased low)
    see: Taylor & Tibshirani (2015) PNAS 112: 7629
  - should only be considered for data sets with few variables (< 5) and a high *n:p* ratio (> 20)
- Coupled to information-theoretic criteria (AIC, BIC)

Problems of stepwise model selection

Problems include (see Harrell 2015: 68):

- *R^2* values biased high
- Standard errors and confidence intervals too low/narrow
- Regression coefficients biased high, require shrinkage
- Collinearity renders variable selection arbitrary
- Allows to not think about the problem

“Let the computer find out” is a poor strategy and usually reflects the fact that the researcher did not bother to think clearly about the problem of interest and its scientific setting”
(Burnham and Anderson, 2002)

Problems generally apply to the stepwise modelling strategy, irrespective of GOF
(Murtaugh 2014 *Ecology* 95: 611; Harrell 2015: 69)

(Partial) fixes

- Modify stepwise approach or related results:
  - correction of *p*-values for sequential testing (Fithian 2015 *ArXiv e-prints*)
  - employ bootstrapping or cross-validation on all steps of model selection
    (but see Harrell 2015: 70f, Austin 2008 *J Clin Epidem*)
  - apply shrinkage factor(s) *c* to regression coefficients, which is/are estimated via CV:

<u>Global shrinkage factor</u>

$$\hat{b}_0^s = (1-\hat{c})\bar{y}+\hat{c}\,\hat{b}_0$$

$$\hat{b}_j^s = \hat{c}\,\hat{b}_j,\; j = 1,\ldots,p$$

<u>Parameterwise shrinkage factor</u>

$$\hat{b}_0^s = (1-\hat{c}_0)\bar{y}+\hat{c}_0\hat{b}_0$$

$$\hat{b}_j^s = \hat{c}_j\hat{b}_j,\; j = 1,\ldots,p$$

- Use shrinkage method such as the LASSO (Least Absolute Shrinkage and Selection Operator)

Multiple regression analysis

Contents

1. Model purpose, learning targets and mathematical basics
2. Model definition, case study and modelling scheme
3. Goodness of fit and model selection
4. Stepwise model selection
5. **The LASSO**
6. Data preparation & Multicollinearity
7. Model diagnosis and analysis, small sample sizes and general tutorial

Shrinkage method: LASSO

- Ordinary least square regression:

$$\underset{b_0,\, b}{minimize} \left\{ \sum \left(\boldsymbol{y} - b_0 - \sum_{j=1}^{p} b_j \boldsymbol{x_j} \right)^2 \right\}$$

- Linear regression with LASSO:

$$\underset{b_0,\, b}{minimize} \left\{ \sum \left(\boldsymbol{y} - b_0 - \sum_{j=1}^{p} b_j \boldsymbol{x_j} \right)^2 + \lambda \sum_{j=1}^{p} |b_j| \right\}$$

Other formulation:

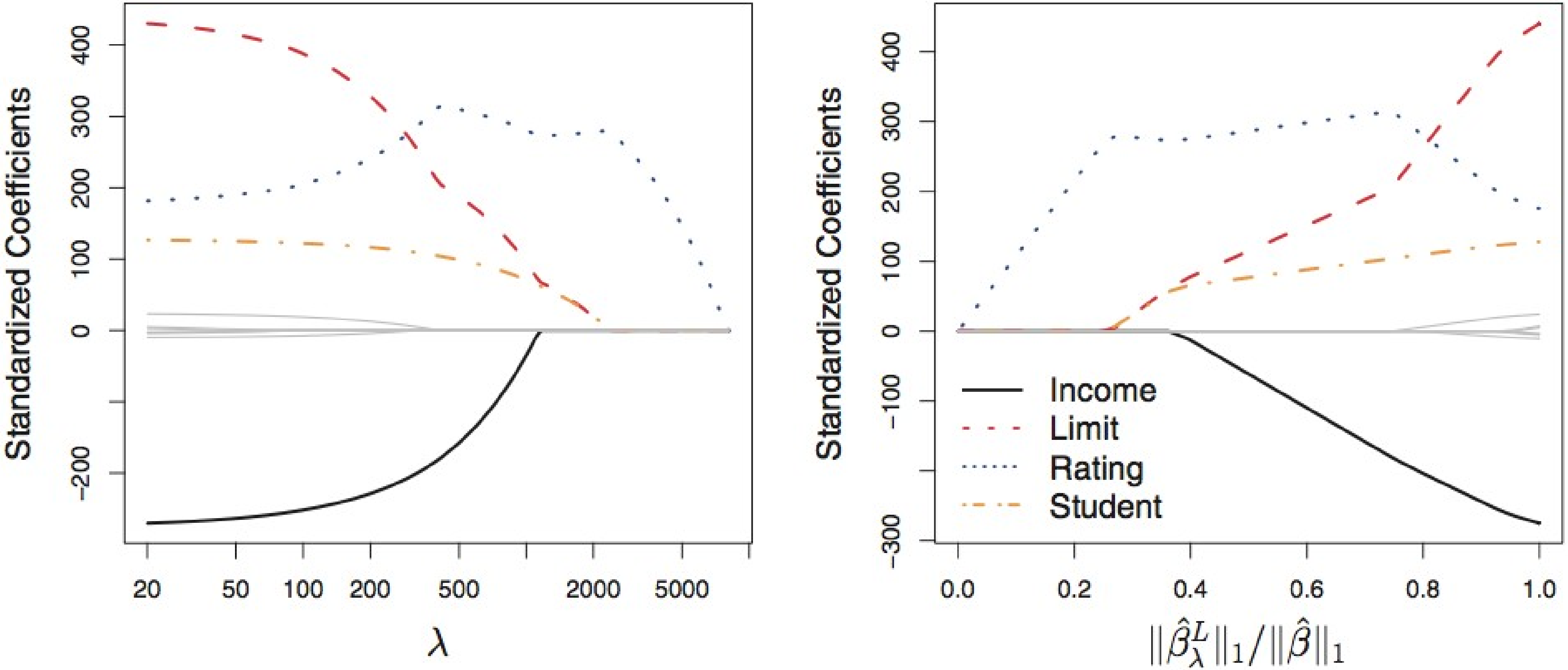
$$\underset{b_0,\, b}{minimize} \left\{ \sum \left(\boldsymbol{y} - b_0 - \sum_{j=1}^{p} b_j \boldsymbol{x_j} \right)^2 \right\} \ \text{subject to} \sum_{j=1}^{p} |b_j| \leq s$$

- Simultaneous selection of variables and estimation of (shrinked) regression coefficients

Shrinkage method: LASSO

$$\underset{b_0, b}{minimize} \left\{ \sum (\boldsymbol{y} - b_0 - \sum_{j=1}^{p} b_j \boldsymbol{x_j})^2 + \lambda \sum_{j=1}^{p} |b_j| \right\}$$

<u>Example plots</u>



- How do we identify the optimal λ? → Cross-validation

Taken from James *et al.* 2013: 220

Multiple regression analysis

Contents

1. Model purpose, learning targets and mathematical basics
2. Model definition, case study and modelling scheme
3. Goodness of fit and model selection
4. Stepwise model selection
5. The LASSO
6. **Data preparation & Multicollinearity**
7. Model diagnosis and analysis, small sample sizes and general tutorial

Data preparation

- Check distribution of predictors and transform if strongly skewed and spanning orders of magnitude
- Check for multicollinearity:
  - Strong correlation between explanatory variables
  - Can lead to incorrect estimates of the regression coefficient and non-significance of relevant predictors in the model
  - Inspect visually and using correlation analysis or variance inflation factors (VIF):

$$\mathrm{VIF} = \frac{1}{1 - R_j^2}$$

R_j is the explained variance for the linear model where the (explanatory) variable *x_j* is explained by all other variables in the model

Dealing with multicollinearity

- Select explanatory variables based on scientific knowledge
- Scatterplots and VIFs can aid in identifying variables with high multicollinearity, but can not suggest what to do
- Do not automatically remove the variable with the highest VIF! Check relevance of variables based on current scientific understanding
- Approaches to deal with multicollinearity:
  - Omit variables from model
  - Select alternative model (e.g. ridge regression, elastic net, principal component regression).

Multiple regression analysis

Contents

1. Model purpose, learning targets and mathematical basics
2. Model definition, case study and modelling scheme
3. Goodness of fit and model selection
4. Stepwise model selection
5. The LASSO
6. Data preparation & Multicollinearity
7. **Model diagnosis and analysis, small sample sizes and general tutorial**

Model diagnostics and analysis

1. Check assumptions of simple regression model (normality and independence of residuals, homogeneity of residual variance, linearity)
2. Check for leverage points, outliers and influential points
3. Use cross-validation to determine prediction accuracy (unless used in model selection)

Measures for relative importance of variables

- Standardized betas, explained variance or both
- Standardized betas are scaled regression coefficients:

$$\hat{b}_{k\,,\,standardized} = \hat{b}_k \frac{s_k}{s_y}$$

s_k = standard variation of predictor *k*
s_y = standard variation of response *y*

- Hierarchical partitioning (Chevan & Sutherland 1991) and PMVD (Feldman 2005) most suitable

Dealing with small sample sizes

- *n*/*p* ratio << 10, in extreme cases $n < p$

- OLS regression and LASSO unreliable, several modelling approaches not applicable for $n < p$ (e.g. backward elimination)

- Approaches to deal with small sample sizes:

  - Remove variables manually based on scientific understanding, very low variability or narrow distribution, and missing values

  - Apply redundancy techniques before modelling that reduce number of variables through statistical algorithms, e.g. variable clustering, principal component analysis (PCA)

  - Select alternative model: Elastic net

Brief tutorial for multiple regression

1. Transform variables if necessary (check range, distribution)

2. Check for multicollinearity, if present, omit variables or adjust regression method

Data preparation

3. Choose modelling strategy (e.g. specify models *a priori*, LASSO) in line with research question

4. Identify best-fit model by applying modelling strategy

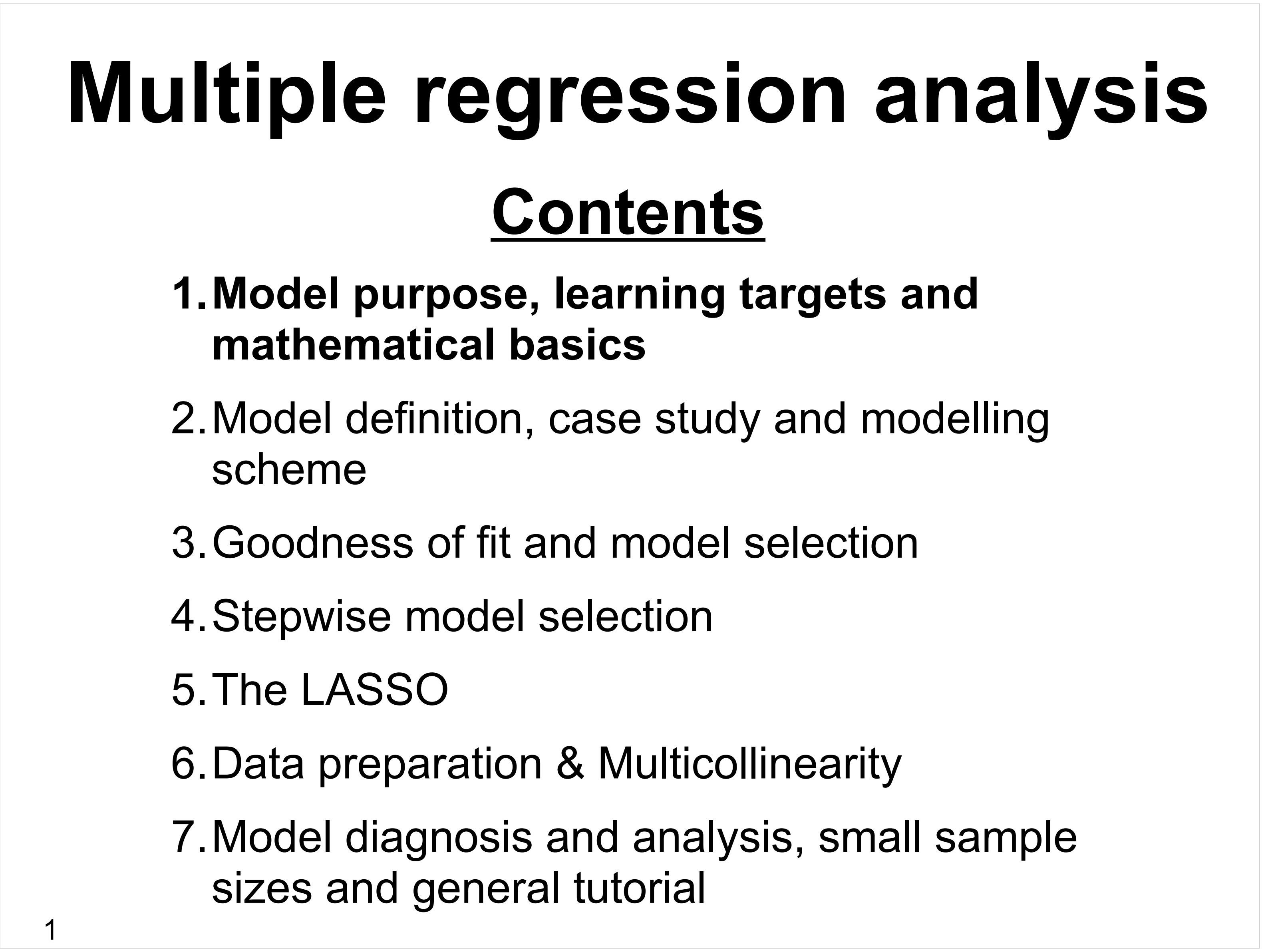
Modelling

5. Run diagnostics for best-fit model

6. Validate model using cross-validation or validation sample

7. Determine variable importance

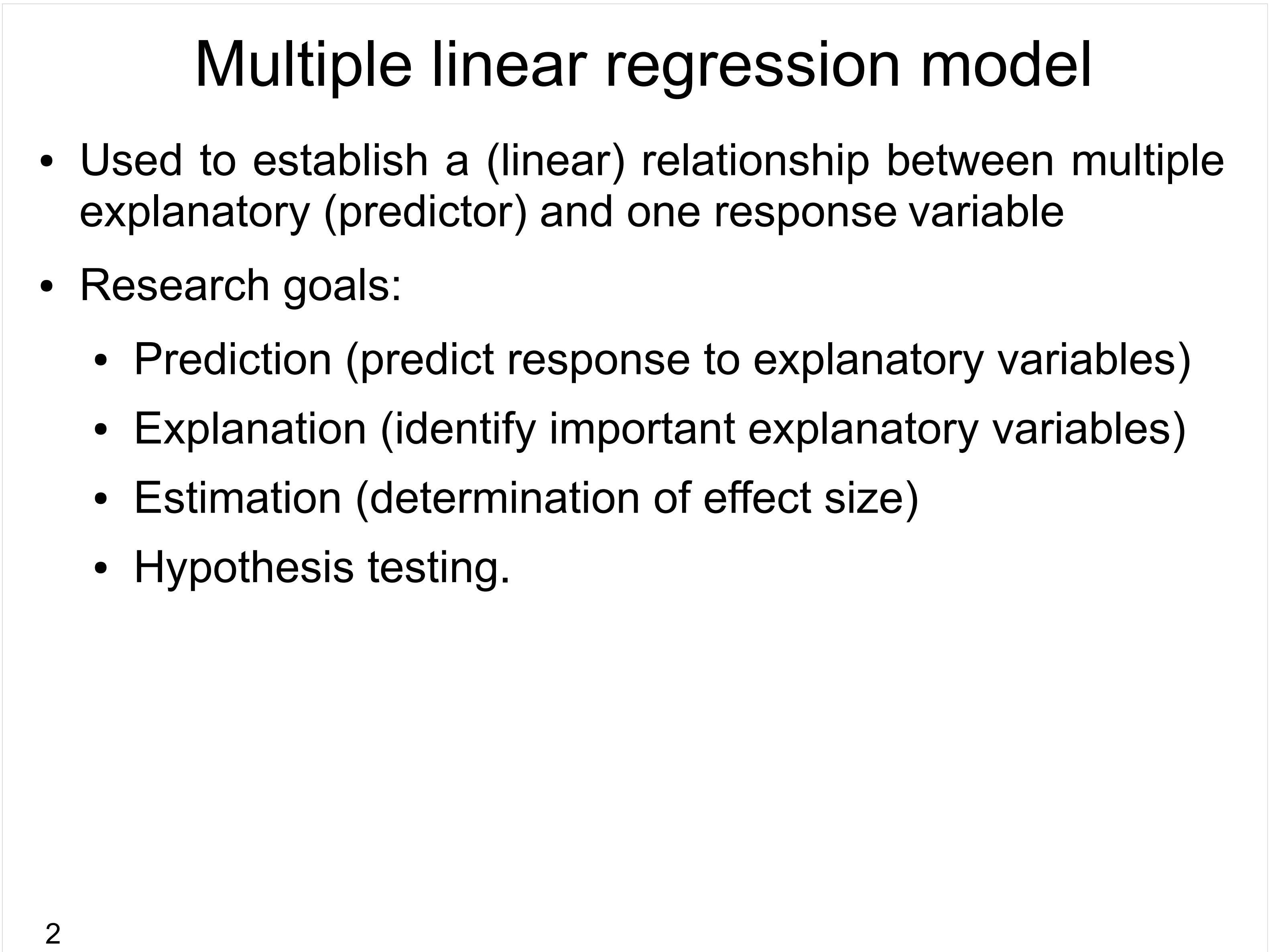
Model diagnosis and analysis



Multiple regression analysis

Contents

1. **Model purpose, learning targets and mathematical basics**
2. Model definition, case study and modelling scheme
3. Goodness of fit and model selection
4. Stepwise model selection
5. The LASSO
6. Data preparation & Multicollinearity
7. Model diagnosis and analysis, small sample sizes and general tutorial



Multiple linear regression model

- Used to establish a (linear) relationship between multiple explanatory (predictor) and one response variable
- Research goals:
  - Prediction (predict response to explanatory variables)
  - Explanation (identify important explanatory variables)
  - Estimation (determination of effect size)
  - Hypothesis testing.

Example for prediction: Predicting the distribution of a species based on climatic and environmental variables. Example for effect estimation: quantify the increase of plant growth under increasing levels of nutrients after accounting for the effects of precipitation and climatic variables. Example for hypothesis testing: Test the hypothesis that a regression coefficient deviates from a given value.

Multiple linear regression model

- <u>Example:</u> Which variable(s) do best explain the response of different groups of organisms?

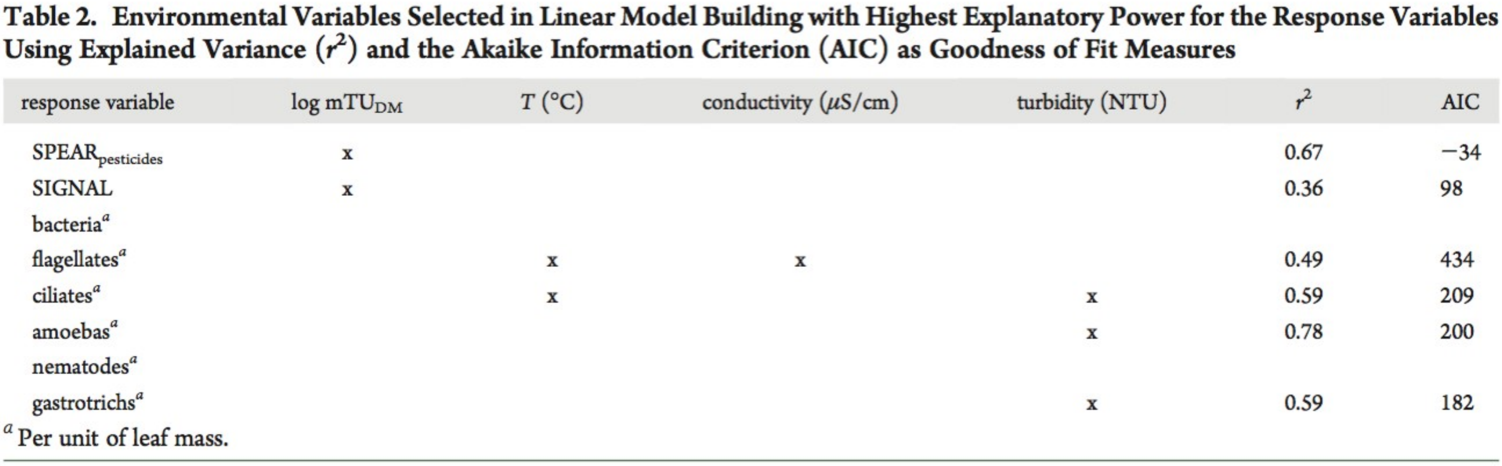


Table 2. Environmental Variables Selected in Linear Model Building with Highest Explanatory Power for the Response Variables Using Explained Variance (r^2) and the Akaike Information Criterion (AIC) as Goodness of Fit Measures

response variable	log $\mathrm{mTU_{DM}}$	T (°C)	conductivity (μS/cm)	turbidity (NTU)	r^2	AIC
$\mathrm{SPEAR_{pesticides}}$	x				0.67	−34
SIGNAL	x				0.36	98
bacteria[a]						
flagellates[a]		x	x		0.49	434
ciliates[a]		x		x	0.59	209
amoebas[a]				x	0.78	200
nematodes[a]						
gastrotrichs[a]				x	0.59	182

[a] Per unit of leaf mass.

Schäfer et al. 2011

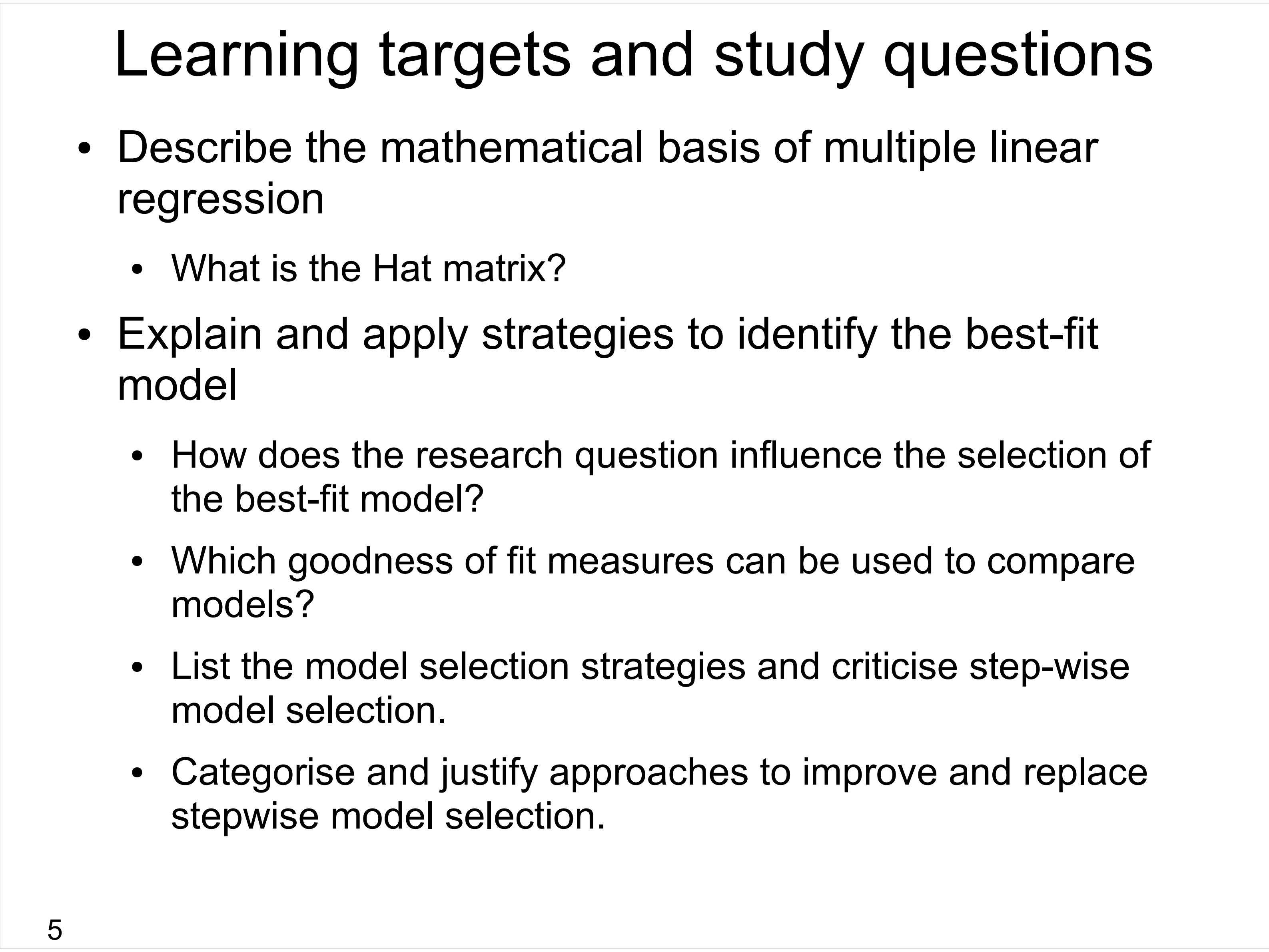
The example relates to a field study in South-East Australia on the relationship between pesticides and invertebrates as well as microorganisms. The SPEAR index has been developed to indicate pesticide effects in invertebrate communities, whereas the SIGNAL index indicates general ecological degradation.

Log mTU is the the maximum Toxic Unit in a sampling site, calculated for pesticides in 24 South-East Australian streams.

Schäfer R.B., Pettigrove V., Rose G., Allinson G., Wightwick A., von der Ohe P.C., et al. (2011) Effects of pesticides monitored with three sampling methods in 24 sites on macroinvertebrates and microorganisms. Environmental Science & Technology 45, 1665–1672.

Learning targets

- Describe the mathematical basis of multiple linear regression
- Explain and apply strategies to identify the best-fit model
- Interpret models and apply variable-importance measures
- Describe the modelling steps

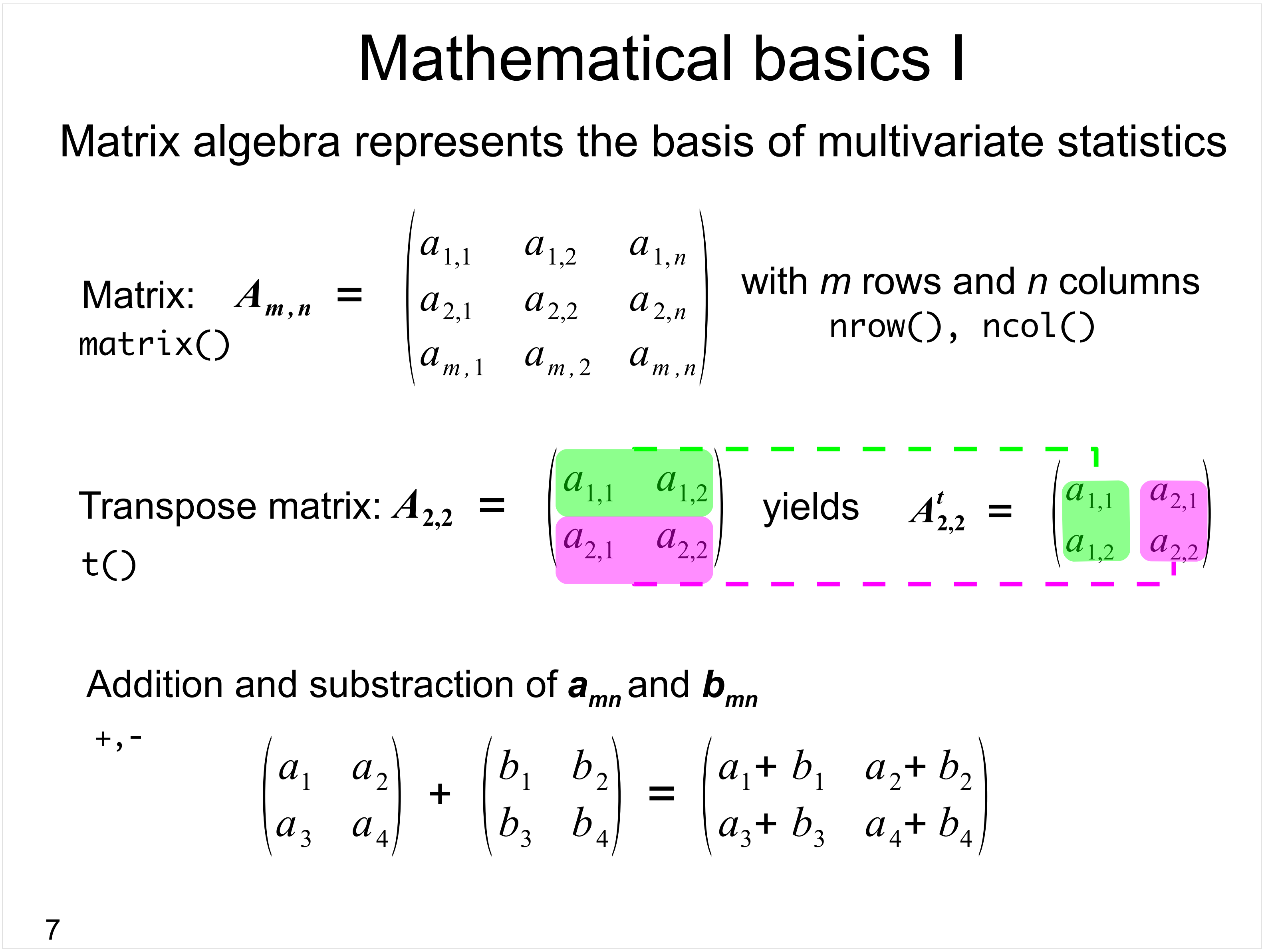


Learning targets and study questions

- Describe the mathematical basis of multiple linear regression
  - What is the Hat matrix?
- Explain and apply strategies to identify the best-fit model
  - How does the research question influence the selection of the best-fit model?
  - Which goodness of fit measures can be used to compare models?
  - List the model selection strategies and criticise step-wise model selection.
  - Categorise and justify approaches to improve and replace stepwise model selection.

Learning targets and study questions

- Interpret models and apply variable-importance measures
  - Which types of model diagnostics are required for multiple regression models?
  - Outline methods to diagnose and to deal with collinearity.
  - Compare methods to check the relative importance of variables.
- Describe the modelling steps



Mathematical basics I

Matrix algebra represents the basis of multivariate statistics

Matrix: $\boldsymbol{A}_{\boldsymbol{m},\boldsymbol{n}} = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,n} \\ a_{2,1} & a_{2,2} & a_{2,n} \\ a_{m,1} & a_{m,2} & a_{m,n} \end{pmatrix}$ with *m* rows and *n* columns

`matrix()`

`nrow(), ncol()`

Transpose matrix: $\boldsymbol{A}_{\boldsymbol{2,2}} = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix}$ yields $\boldsymbol{A}^{t}_{\boldsymbol{2,2}} = \begin{pmatrix} a_{1,1} & a_{2,1} \\ a_{1,2} & a_{2,2} \end{pmatrix}$

`t()`

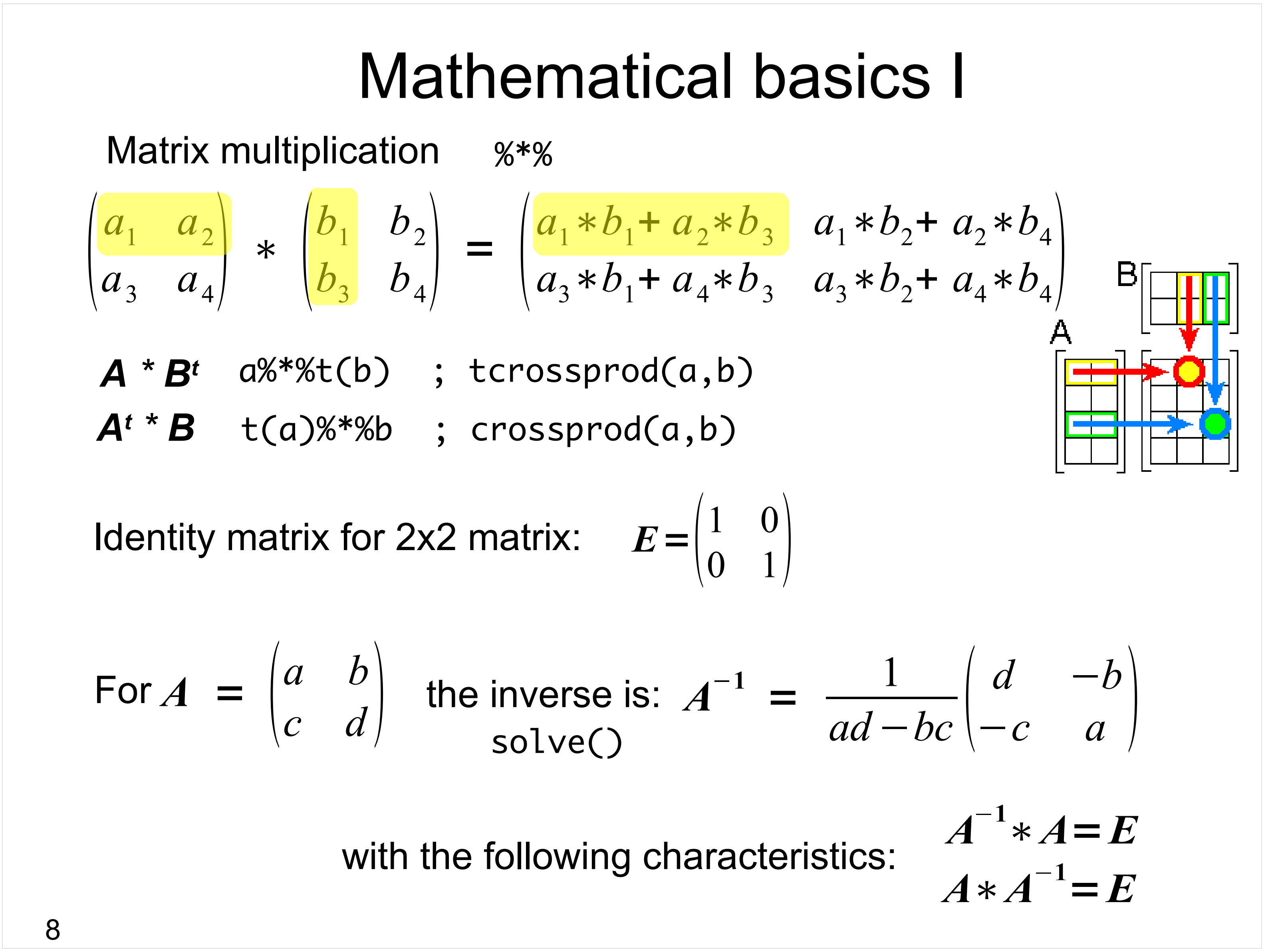
Addition and substraction of $\boldsymbol{a}_{mn}$ and $\boldsymbol{b}_{mn}$

`+,-`

$$\begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} + \begin{pmatrix} b_1 & b_2 \\ b_3 & b_4 \end{pmatrix} = \begin{pmatrix} a_1 + b_1 & a_2 + b_2 \\ a_3 + b_3 & a_4 + b_4 \end{pmatrix}$$

The background for many multivariate techniques, including multiple linear regression, is matrix algebra. Although we do not discuss the mathematical background in detail and mainly use existing R functions, a brief overview on the mathematical operations that constitute the basis of most methods and their implementation in R is given. You can downloaded a related script on the course website.

Matrices and vectors in bold.



Mathematical basics I

Matrix multiplication %*%

$$\begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} * \begin{pmatrix} b_1 & b_2 \\ b_3 & b_4 \end{pmatrix} = \begin{pmatrix} a_1 * b_1 + a_2 * b_3 & a_1 * b_2 + a_2 * b_4 \\ a_3 * b_1 + a_4 * b_3 & a_3 * b_2 + a_4 * b_4 \end{pmatrix}$$

$\boldsymbol{A} * \boldsymbol{B}^t$ `a%*%t(b) ; tcrossprod(a,b)`

$\boldsymbol{A}^t * \boldsymbol{B}$ `t(a)%*%b ; crossprod(a,b)`

Identity matrix for 2x2 matrix: $\boldsymbol{E} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

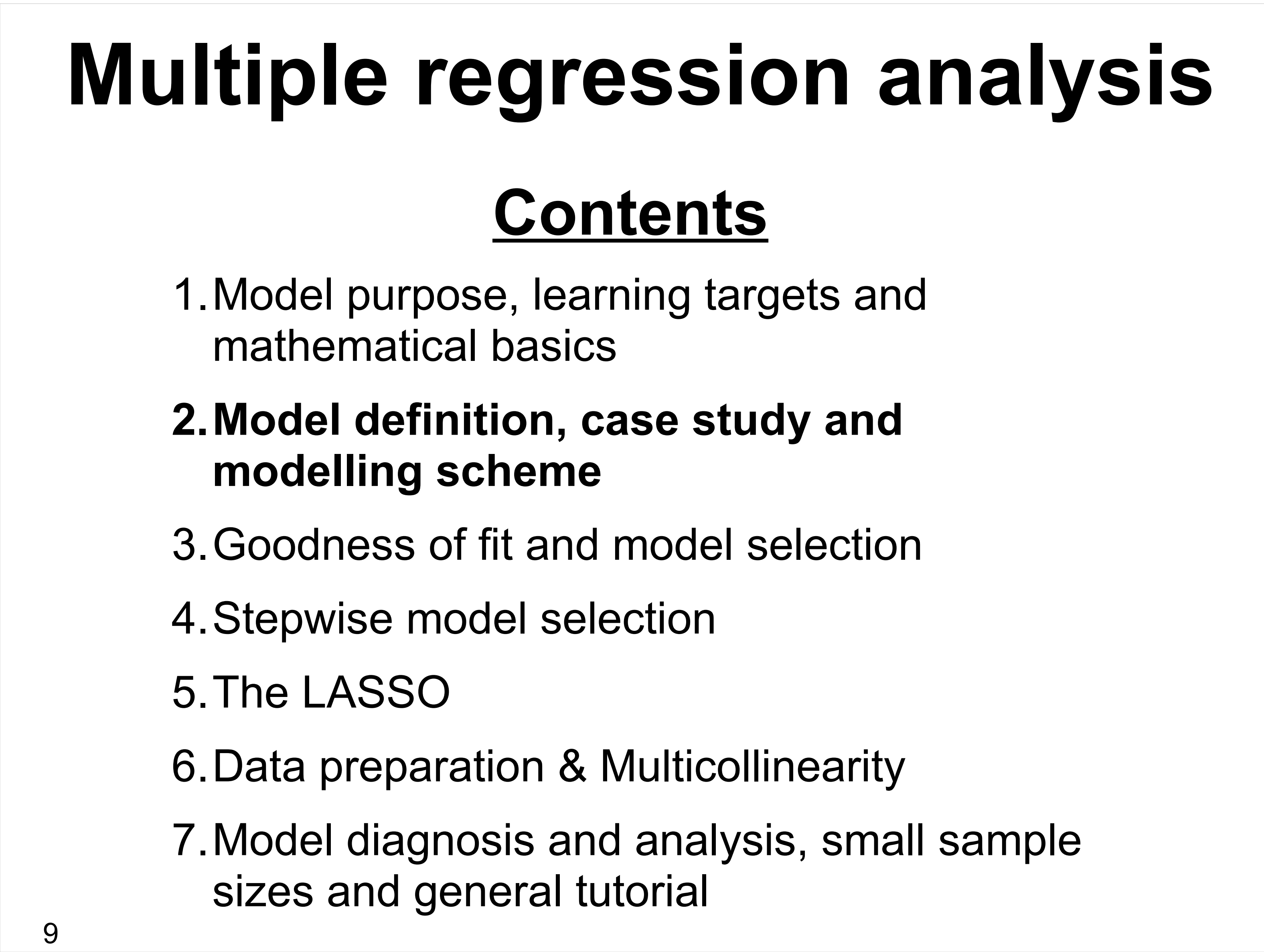
For $\boldsymbol{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ the inverse is: $\boldsymbol{A}^{-1} = \dfrac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

`solve()`

with the following characteristics: $$\boldsymbol{A}^{-1} * \boldsymbol{A} = \boldsymbol{E}$$ $$\boldsymbol{A} * \boldsymbol{A}^{-1} = \boldsymbol{E}$$

Division is not defined for a matrix, but can be achieved by multiplication with the inverse of a matrix.

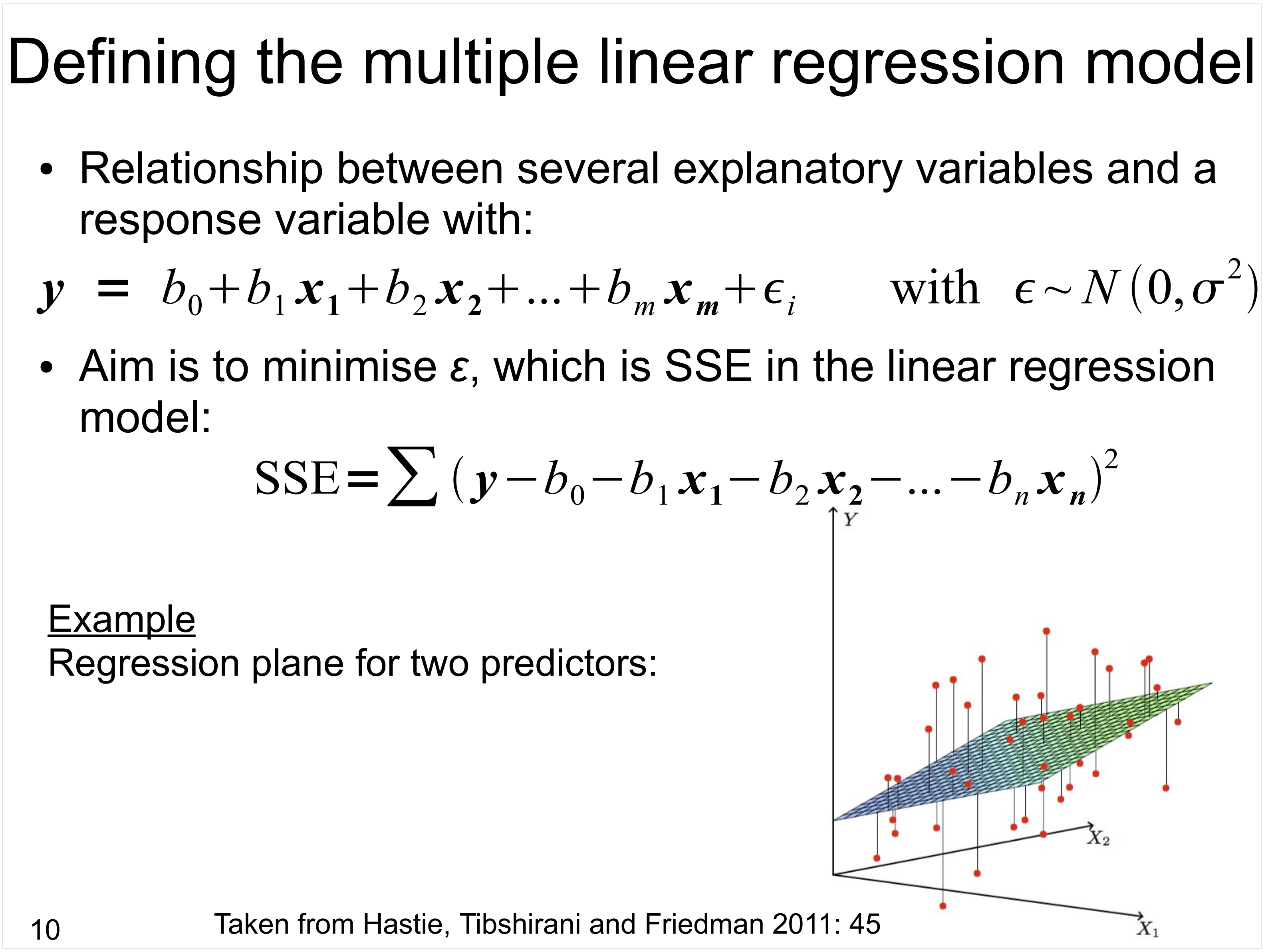
The described methods and functions are not applicable for calculation of the inverse for all matrices, but only for *n*n* matrices where the matrix rank equals the number of rows (or columns). This condition is equivalent to the existence of *n* independent linear combinations of the row or column vectors. The calculation of the rank of a matrix is described in Meyer 2000: Matrix Analyses and Applied Matrix Algebra: 44-45 (see: http://www.matrixanalysis.com/). In R this can be done with *qr*(), but note that this function does not yield a result for all matrices due to the abovementioned conditions.



Multiple regression analysis

Contents

1. Model purpose, learning targets and mathematical basics
2. **Model definition, case study and modelling scheme**
3. Goodness of fit and model selection
4. Stepwise model selection
5. The LASSO
6. Data preparation & Multicollinearity
7. Model diagnosis and analysis, small sample sizes and general tutorial



Defining the multiple linear regression model

- Relationship between several explanatory variables and a response variable with:

$$\boldsymbol{y} \;=\; b_0+b_1\boldsymbol{x_1}+b_2\boldsymbol{x_2}+\ldots+b_m\boldsymbol{x_m}+\epsilon_i \qquad \text{with} \quad \epsilon \sim N(0,\sigma^2)$$

- Aim is to minimise ε, which is SSE in the linear regression model:

$$\text{SSE}=\sum(\boldsymbol{y}-b_0-b_1\boldsymbol{x_1}-b_2\boldsymbol{x_2}-\ldots-b_n\boldsymbol{x_n})^2$$

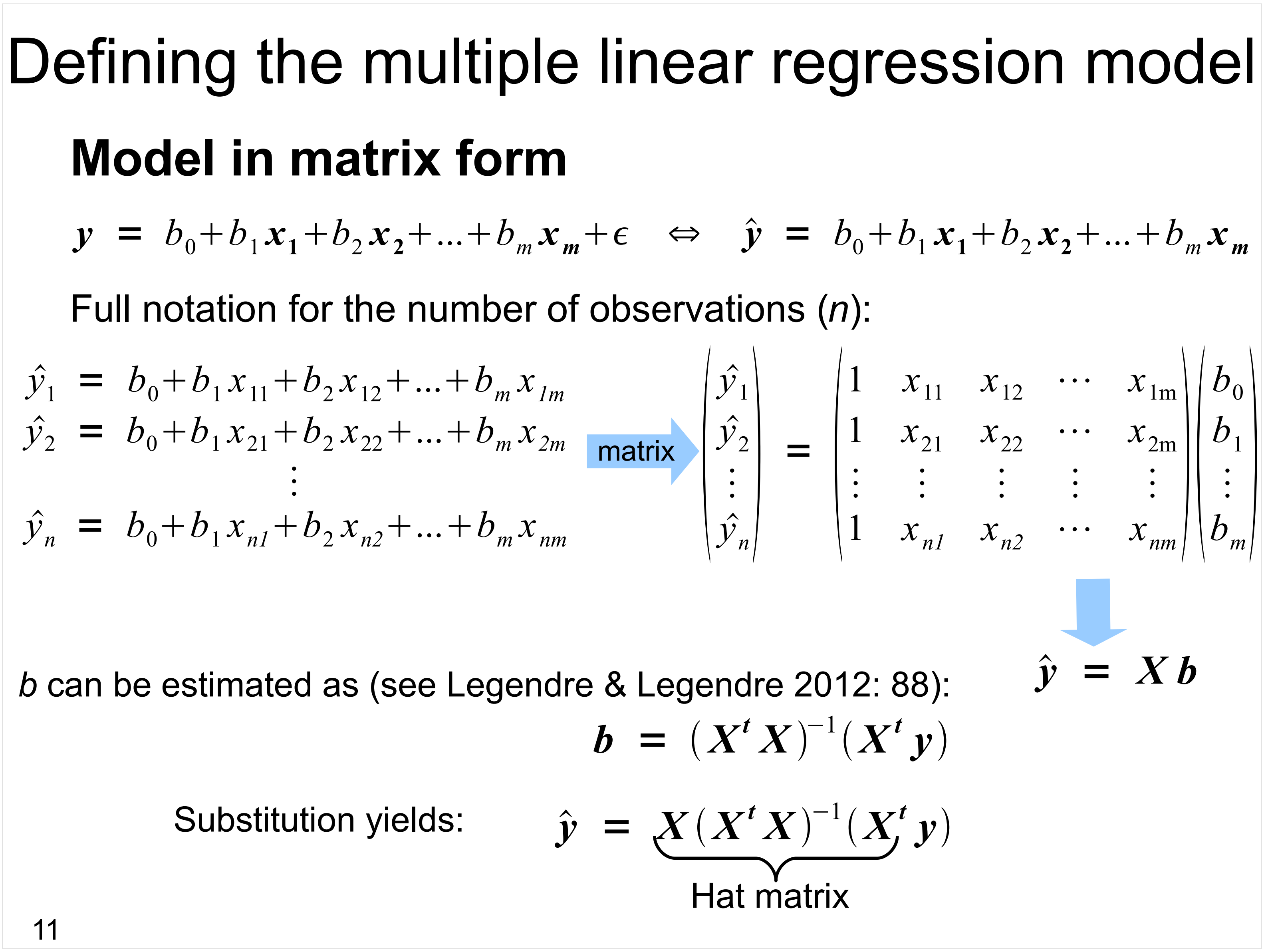
Example

Regression plane for two predictors:

10

Taken from Hastie, Tibshirani and Friedman 2011: 45

The linear regression that aims to minimize SSE is also called ordinary least square (OLS) regression.



Defining the multiple linear regression model

Model in matrix form

$$\boldsymbol{y} = b_0+b_1\boldsymbol{x_1}+b_2\boldsymbol{x_2}+...+b_m\boldsymbol{x_m}+\epsilon \quad \Leftrightarrow \quad \hat{\boldsymbol{y}} = b_0+b_1\boldsymbol{x_1}+b_2\boldsymbol{x_2}+...+b_m\boldsymbol{x_m}$$

Full notation for the number of observations (*n*):

$$\begin{aligned}\hat{y}_1 &= b_0+b_1x_{11}+b_2x_{12}+...+b_mx_{1m}\\ \hat{y}_2 &= b_0+b_1x_{21}+b_2x_{22}+...+b_mx_{2m}\\ &\vdots\\ \hat{y}_n &= b_0+b_1x_{n1}+b_2x_{n2}+...+b_mx_{nm}\end{aligned} \xrightarrow{\text{matrix}} \begin{pmatrix}\hat{y}_1\\ \hat{y}_2\\ \vdots\\ \hat{y}_n\end{pmatrix} = \begin{pmatrix}1 & x_{11} & x_{12} & \cdots & x_{1\mathrm{m}}\\ 1 & x_{21} & x_{22} & \cdots & x_{2\mathrm{m}}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm}\end{pmatrix}\begin{pmatrix}b_0\\ b_1\\ \vdots\\ b_m\end{pmatrix}$$

$$\Downarrow$$

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b}$$

b can be estimated as (see Legendre & Legendre 2012: 88):

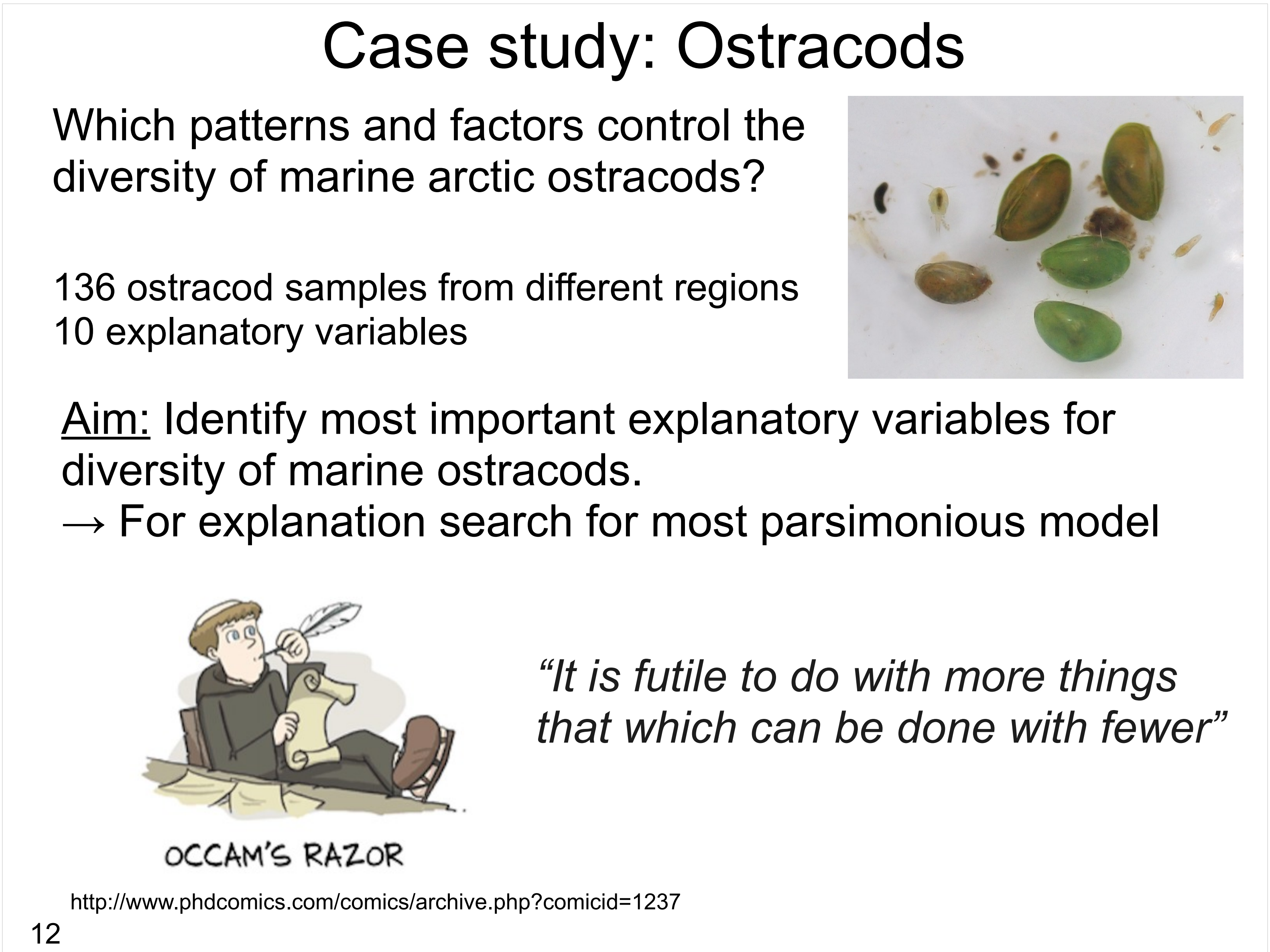
$$\boldsymbol{b} = (\boldsymbol{X}^{\boldsymbol{t}}\boldsymbol{X})^{-1}(\boldsymbol{X}^{\boldsymbol{t}}\boldsymbol{y})$$

Substitution yields:

$$\hat{\boldsymbol{y}} = \underbrace{\boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{t}}\boldsymbol{X})^{-1}(\boldsymbol{X}^{\boldsymbol{t}}}_{\text{Hat matrix}}\boldsymbol{y})$$

11

The hat matrix inherits its name from the fact that it is used to compute $\hat{\boldsymbol{y}}$ from $\boldsymbol{y}$



Case study: Ostracods

Which patterns and factors control the diversity of marine arctic ostracods?

136 ostracod samples from different regions
10 explanatory variables

Aim: Identify most important explanatory variables for diversity of marine ostracods.
→ For explanation search for most parsimonious model

“It is futile to do with more things that which can be done with fewer”

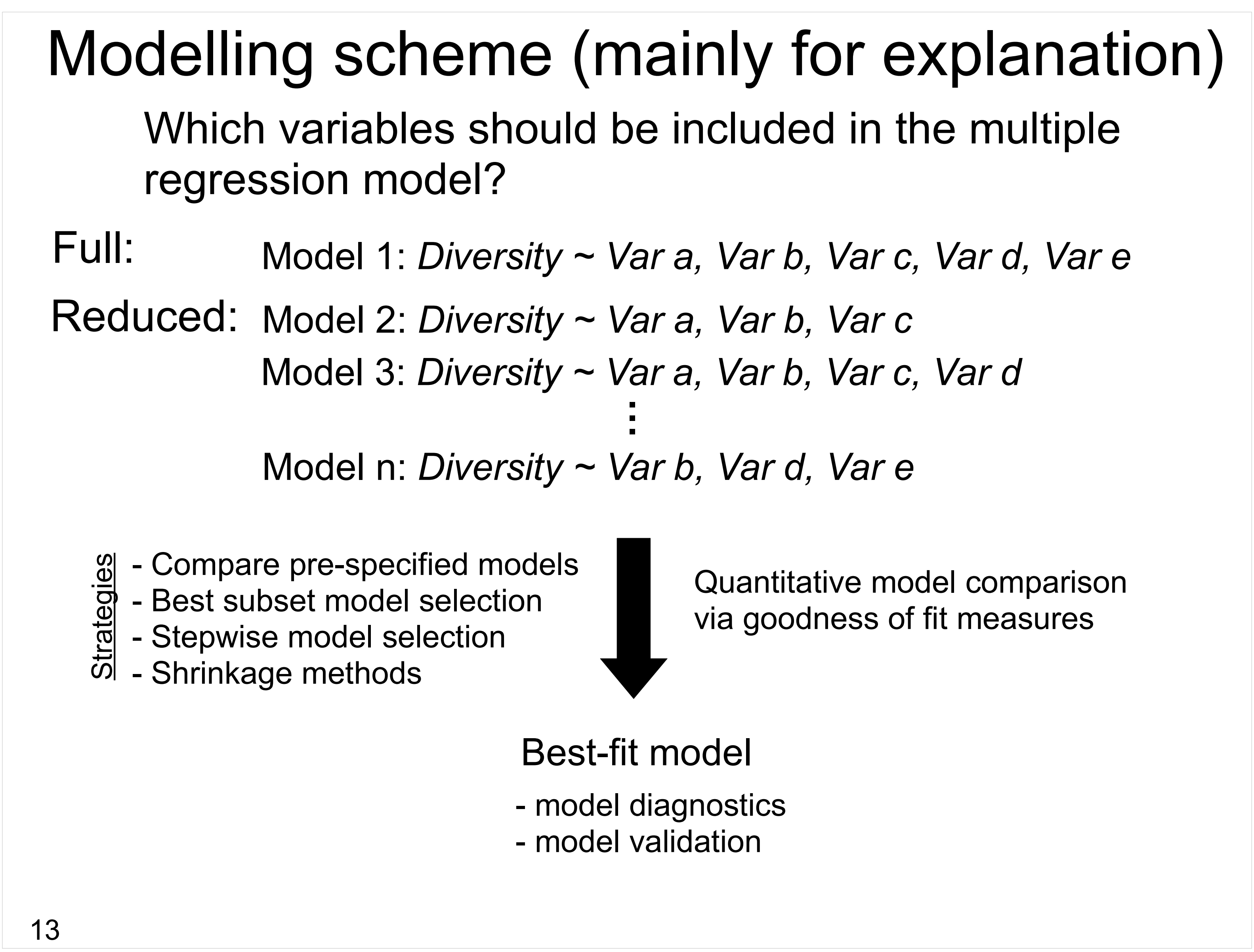
http://www.phdcomics.com/comics/archive.php?comicid=1237

The ostracod picture has been taken from: https://www4.uwm.edu/fieldstation/naturalhistory/bugoftheweek/images/ostracod12

We assume that the relationship between explanatory variables and the species richness is largely linear (or quadratic) in the case study – see for details:
Yasuhara M., Hunt G., van Dijken G., Arrigo K.R., Cronin T.M. & Wollenburg J.E. (2012) Patterns and controlling factors of species diversity in the Arctic Ocean. Journal of Biogeography 39, 2081–2088.

Occam’s razor may be translated to our situation as: Given a similar predictive or explanatory power, models with fewer variables are generally better than those with more variables.

Harrell (2015: 70, 95ff) argues that the full model typically provides meaningful *p*-values, confidence intervals and parameter estimates and has the highest predictive power. Thus, model parsimony is primarily relevant when we aim to identify the most important variables. Notwithstanding, when building models for prediction, we also prefer the model with fewer variables to one with more variables for a similar predictive power.



Modelling scheme (mainly for explanation)

Which variables should be included in the multiple regression model?

Full: Model 1: *Diversity ~ Var a, Var b, Var c, Var d, Var e*

Reduced: Model 2: *Diversity ~ Var a, Var b, Var c*

Model 3: *Diversity ~ Var a, Var b, Var c, Var d*

⋮

Model n: *Diversity ~ Var b, Var d, Var e*

Strategies

- Compare pre-specified models
- Best subset model selection
- Stepwise model selection
- Shrinkage methods

Quantitative model comparison via goodness of fit measures

Best-fit model

- model diagnostics
- model validation

13

For prediction, we often can use the full model and do not need to select a modelling strategy (see previous slide). If we aim to determine an effect size or test a specific hypothesis, we should have a pre-specified model and other strategies are largely irrelevant.

Why do we not test the importance of variables using multiple bivariate regressions?

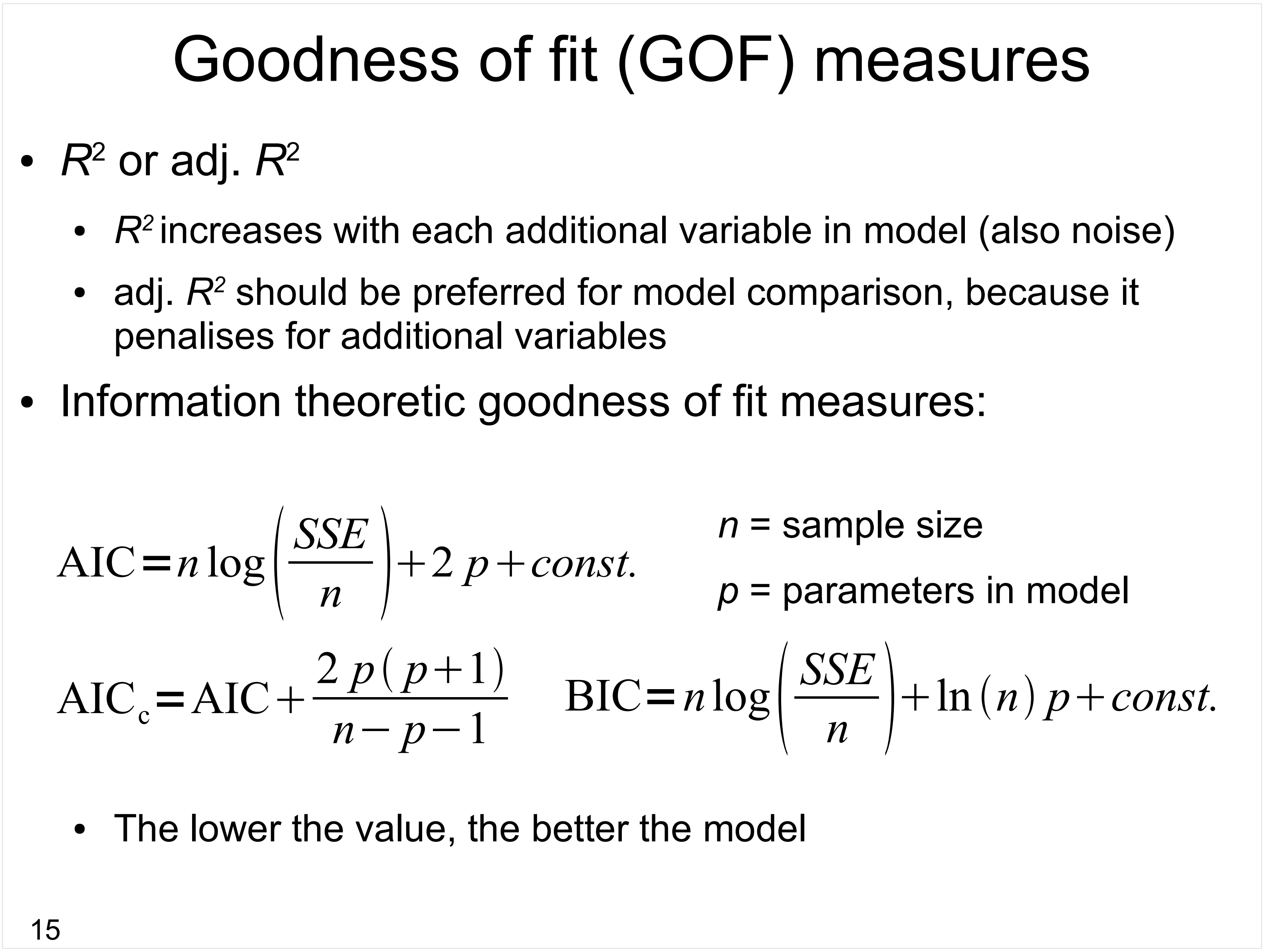
This is because in bivariate regressions, important variables may be ignored that exert a high explanatory power after other variables have been included in the model. For further explanation see:

Sun G.-W., Shook T.L. & Kay G.L. (1996) Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. Journal of Clinical Epidemiology 49, 907 – 916.

Multiple regression analysis

Contents

1. Model purpose, learning targets and mathematical basics
2. Model definition, case study and modelling scheme
3. **Goodness of fit and model selection**
4. Stepwise model selection
5. The LASSO
6. Data preparation & Multicollinearity
7. Model diagnosis and analysis, small sample sizes and general tutorial



Goodness of fit (GOF) measures

- R^2 or adj. R^2
  - R^2 increases with each additional variable in model (also noise)
  - adj. R^2 should be preferred for model comparison, because it penalises for additional variables
- Information theoretic goodness of fit measures:

$$\mathrm{AIC} = n\log\left(\frac{SSE}{n}\right) + 2p + const.$$

n = sample size

p = parameters in model

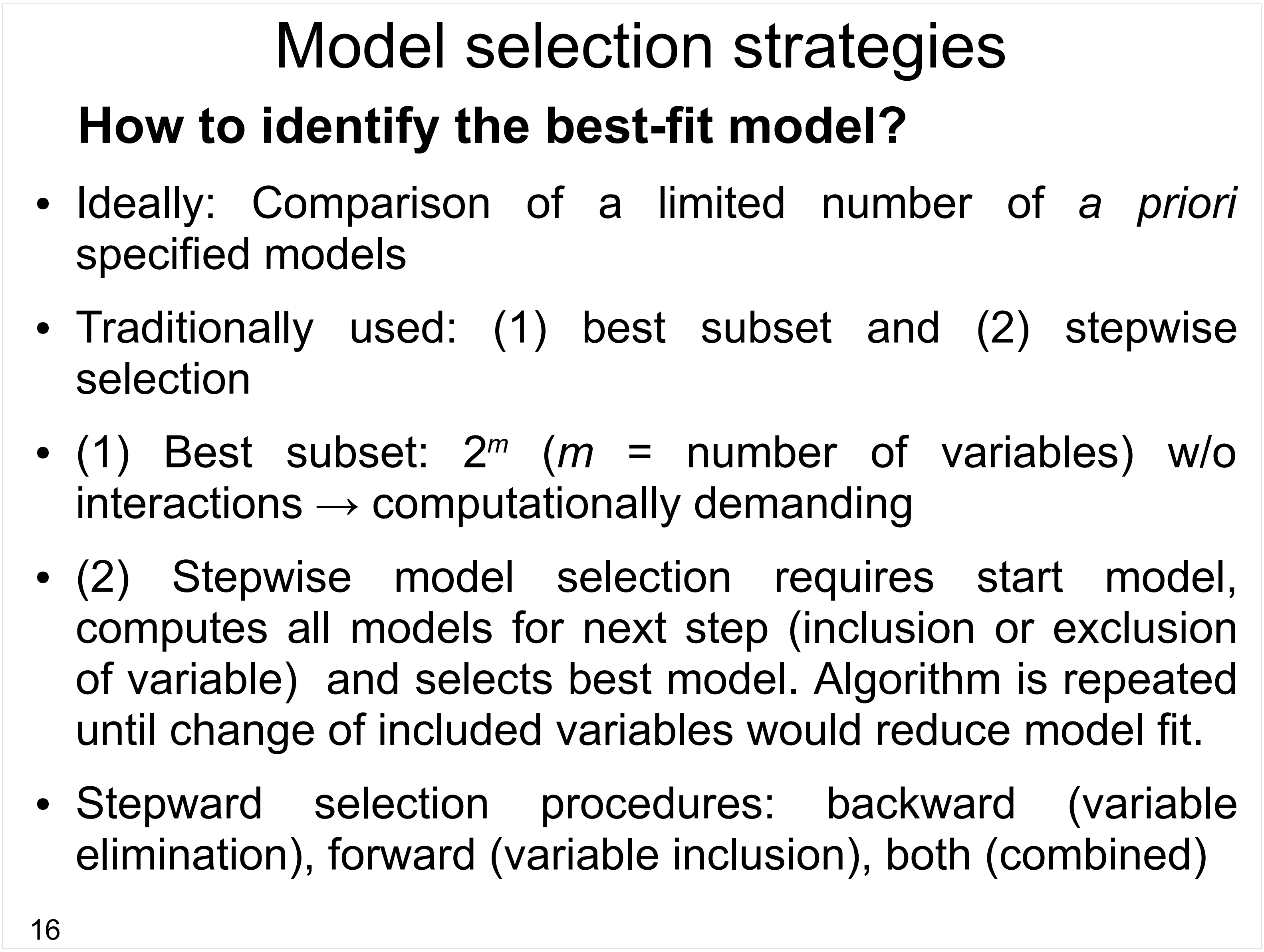
$$\mathrm{AIC_c} = \mathrm{AIC} + \frac{2p(p+1)}{n-p-1} \qquad \mathrm{BIC} = n\log\left(\frac{SSE}{n}\right) + \ln(n)\,p + const.$$

  - The lower the value, the better the model

n = sample size; p = parameters in model

Here, the information criteria are expressed for models subject to least square fitting. Generally, the AIC is given as: AIC = - 2log(L) + 2p, where L is the likelihood function for the parameters in the model. Similarly, the BIC is given as: BIC = -2log(L) + $p$log(n)

Note that the R function AIC() a simplified version of AIC without the constant term is calculated. This is justified because the constant term is redundant when comparing models i.e. the addition of a constant does not affect the ranking of the absolute values of the models. The BIC gives higher penalty to more complex models than the AIC for $n \geq 8$ and may thus aid in selection of more parsimonious models. The AIC tends towards over-fitting especially for smaller data sets (e.g. $n < 50$). The corrected AIC ($\mathrm{AIC_C}$) is the recommended alternative. In fact, the corrected AIC could always be used as it converges with the AIC for larger sample sizes.



Model selection strategies

How to identify the best-fit model?

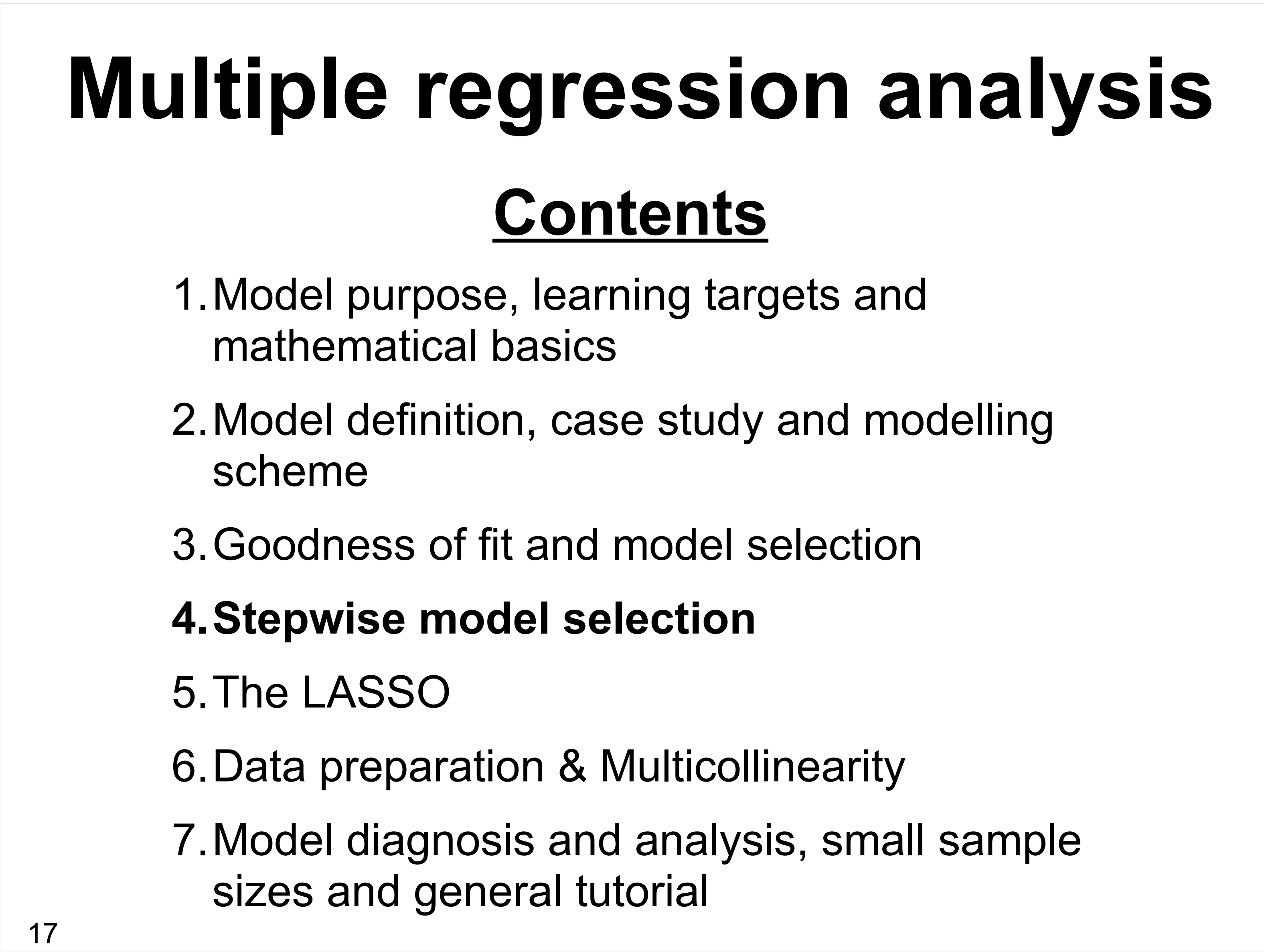
- Ideally: Comparison of a limited number of *a priori* specified models
- Traditionally used: (1) best subset and (2) stepwise selection
- (1) Best subset: 2^m (m = number of variables) w/o interactions → computationally demanding
- (2) Stepwise model selection requires start model, computes all models for next step (inclusion or exclusion of variable) and selects best model. Algorithm is repeated until change of included variables would reduce model fit.
- Stepward selection procedures: backward (variable elimination), forward (variable inclusion), both (combined)

If our aim is to estimate effect sizes or to test hypotheses, we should pre-specify a model (or a few models).

Computing all possible models represents an exhaustive search for the best regression model. This procedure is most useful when no initial hypotheses on the ranking of the scientific relevance of variables is available. If the number of possible models is large, different models may have a similar goodness of fit so that the selection of only one model is problematic. Model averaging of all models up to a certain threshold of a GOF measure can be applied in this case (see R demonstration). A review by Grueber et al. (2011) discusses several issues associated with model selection and averaging: http://onlinelibrary.wiley.com/doi/10.1111/j.1420-9101.2010.02210.x/abstract. Note the criticism of Cade (2015) regarding model averaging: http://www.esajournals.org/doi/pdf/10.1890/14-1639.1 and that part of the criticism on stepwise model selection (see next slides) also applies to best subset selection.

Best subset selection quickly becomes computationally demanding, especially if model validation would be applied to the whole process. For example, for ten-fold CV and 10 variables, 1024 * 10 models would need to be fitted.

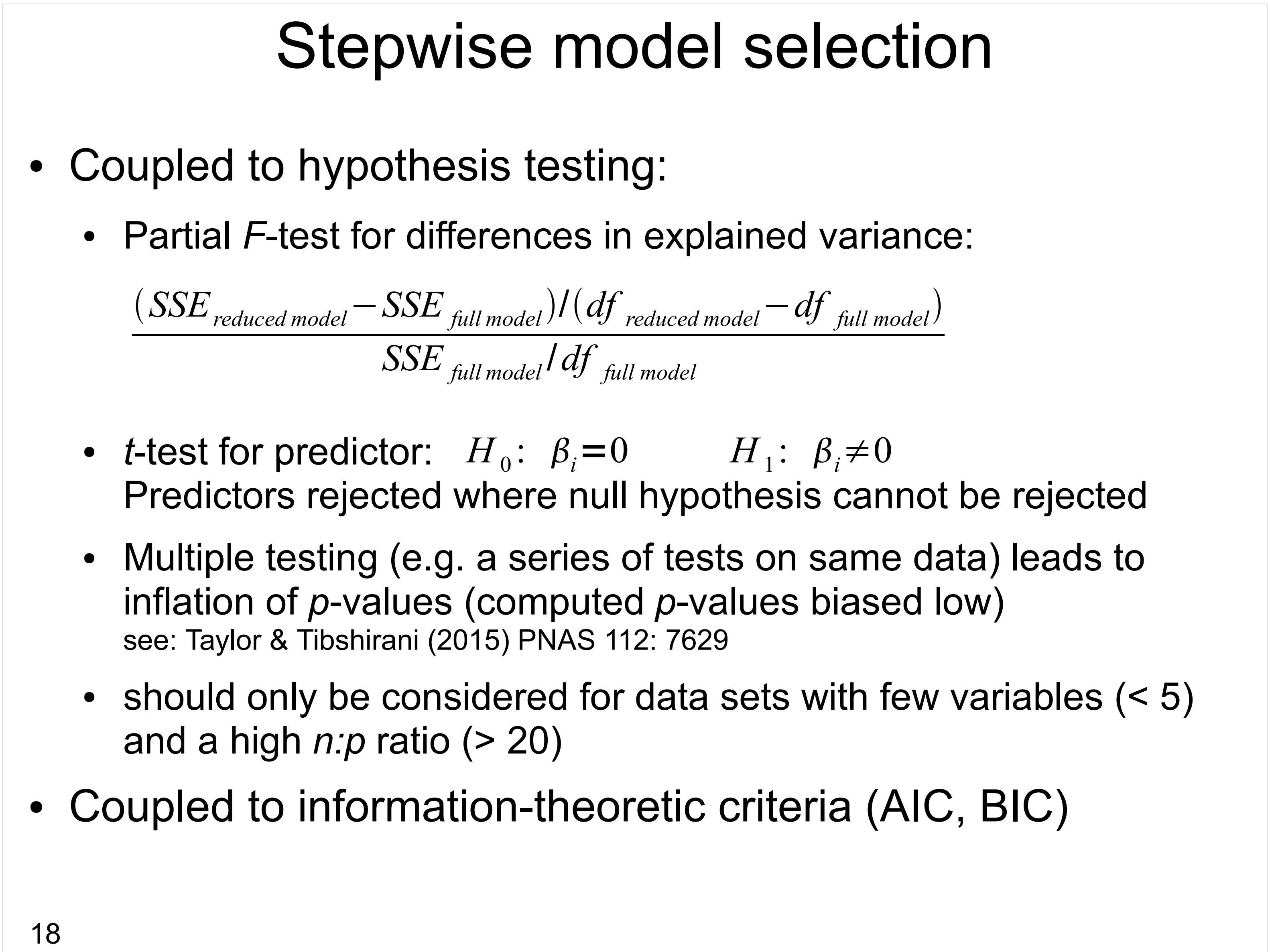
Moreover, compared to a more constrained search, best subset selection is likely to yield to a model with high variance (see Hastie, Tibshirani and Friedman 2011: 59).



Multiple regression analysis

Contents

1. Model purpose, learning targets and mathematical basics
2. Model definition, case study and modelling scheme
3. Goodness of fit and model selection
4. **Stepwise model selection**
5. The LASSO
6. Data preparation & Multicollinearity
7. Model diagnosis and analysis, small sample sizes and general tutorial



Stepwise model selection

- Coupled to hypothesis testing:
  - Partial *F*-test for differences in explained variance:

$$\frac{(SSE_{reduced\ model} - SSE_{full\ model}) / (df_{reduced\ model} - df_{full\ model})}{SSE_{full\ model} / df_{full\ model}}$$

  - *t*-test for predictor: $H_0\colon\ \beta_i = 0$ $H_1\colon\ \beta_i \neq 0$
    Predictors rejected where null hypothesis cannot be rejected
  - Multiple testing (e.g. a series of tests on same data) leads to inflation of *p*-values (computed *p*-values biased low)
    see: Taylor & Tibshirani (2015) PNAS 112: 7629
  - should only be considered for data sets with few variables (< 5) and a high *n:p* ratio (> 20)
- Coupled to information-theoretic criteria (AIC, BIC)

18

n = sample size

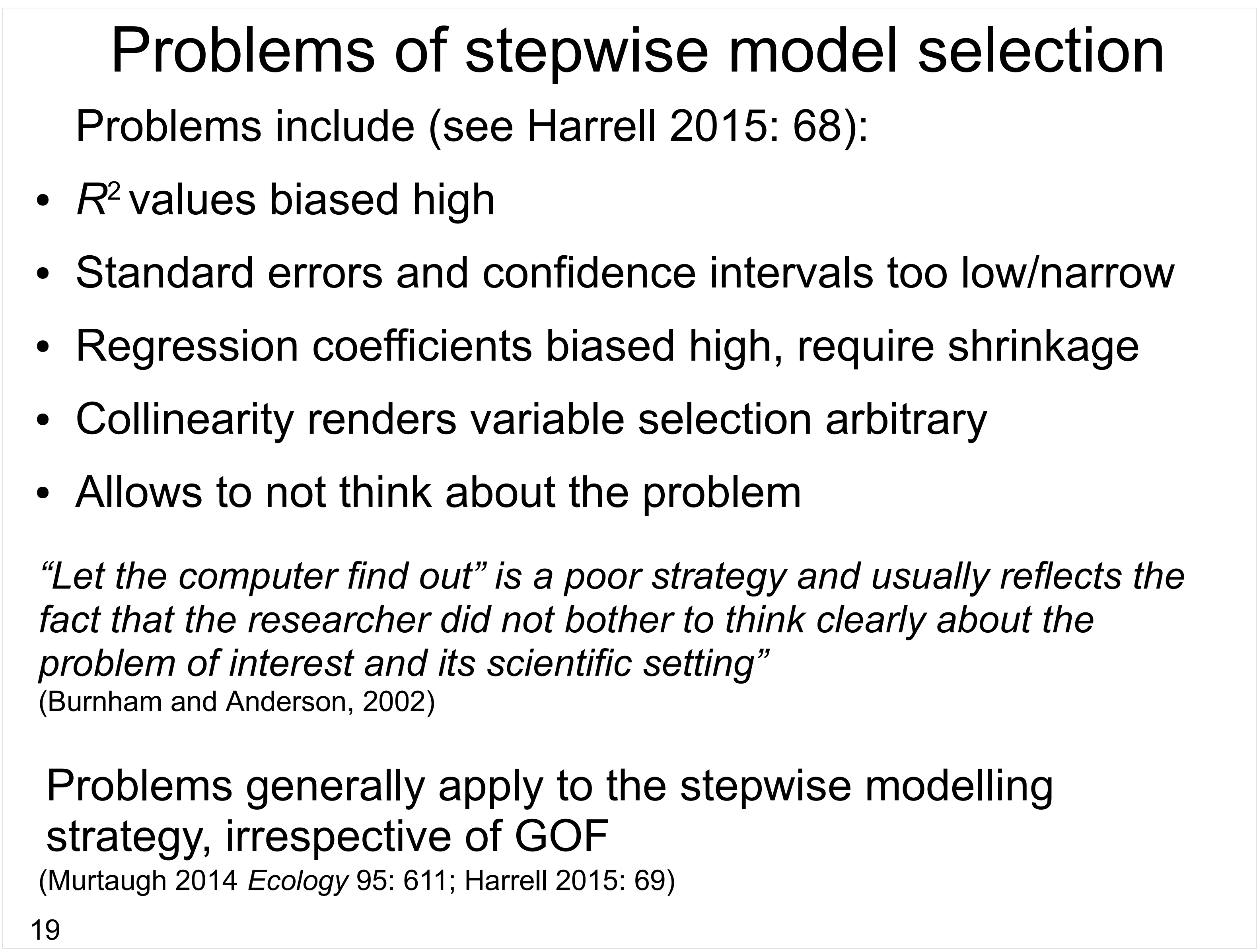
p = parameters in model

Note that the SSE is equivalent to RSS (residual sum of squares), which you also find in text books and journal articles.

Murtaugh (2014; Ecology 95: 611-617) pointed out that the different approaches (i.e. hypothesis testing using p-values and information-theoretic ones) are intimately linked. Thus, they face similar problems in model comparison and selection.

The paper by Taylor & Tibshirani (2015) can be freely accessed:

http://www.pnas.org/content/112/25/7629.abstract



Problems of stepwise model selection

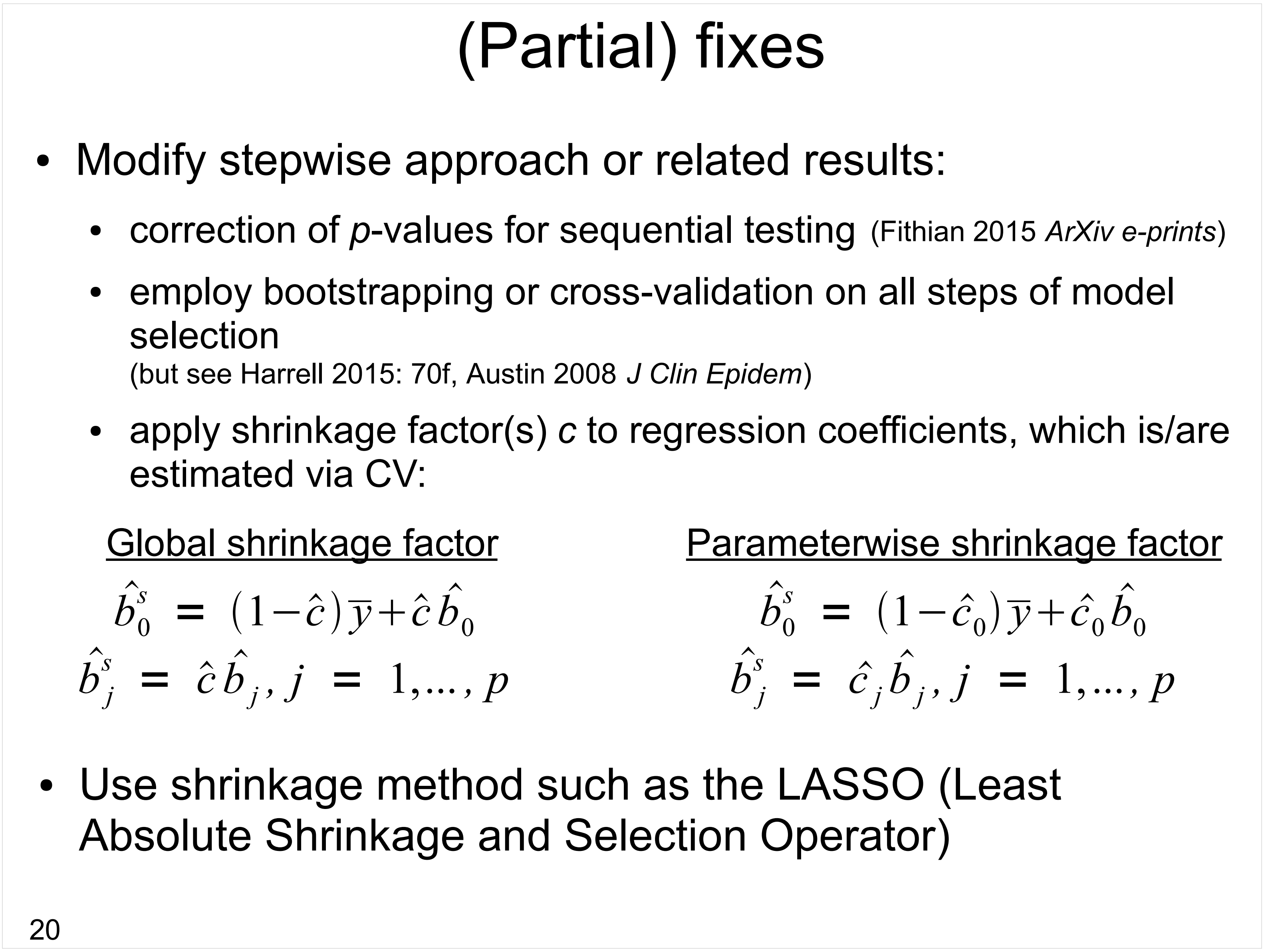
Problems include (see Harrell 2015: 68):

- *R*² values biased high
- Standard errors and confidence intervals too low/narrow
- Regression coefficients biased high, require shrinkage
- Collinearity renders variable selection arbitrary
- Allows to not think about the problem

“Let the computer find out” is a poor strategy and usually reflects the fact that the researcher did not bother to think clearly about the problem of interest and its scientific setting”
(Burnham and Anderson, 2002)

Problems generally apply to the stepwise modelling strategy, irrespective of GOF
(Murtaugh 2014 *Ecology* 95: 611; Harrell 2015: 69)

If stepwise model selection is used, Harrell (2015: p. 70) suggests to use backward selection, because this would perform better in the presence of collinear variables and starts with the full model, which is the only model providing accurate p-values, standard errors etc. However, backward selection cannot be used, if $n < p$ (this case is discussed in detail later).



(Partial) fixes

- Modify stepwise approach or related results:
  - correction of *p*-values for sequential testing (Fithian 2015 *ArXiv e-prints*)
  - employ bootstrapping or cross-validation on all steps of model selection
    (but see Harrell 2015: 70f, Austin 2008 *J Clin Epidem*)
  - apply shrinkage factor(s) *c* to regression coefficients, which is/are estimated via CV:

Global shrinkage factor

$$\hat{b}_0^s = (1-\hat{c})\bar{y}+\hat{c}\,\hat{b}_0$$

$$\hat{b}_j^s = \hat{c}\,\hat{b}_j, j = 1,\dots,p$$

Parameterwise shrinkage factor

$$\hat{b}_0^s = (1-\hat{c}_0)\bar{y}+\hat{c}_0\hat{b}_0$$

$$\hat{b}_j^s = \hat{c}_j\hat{b}_j, j = 1,\dots,p$$

- Use shrinkage method such as the LASSO (Least Absolute Shrinkage and Selection Operator)

Austin (2008) found no improved performance of bootstrapping model selection compared to backward stepwise selection. Harrell (2015: 70f) discusses several drawbacks of the bootstrap approach.

CV is similarly likely to underestimate the true variance.

The application of shrinkage factors after the model selection is called post-selection shrinkage.

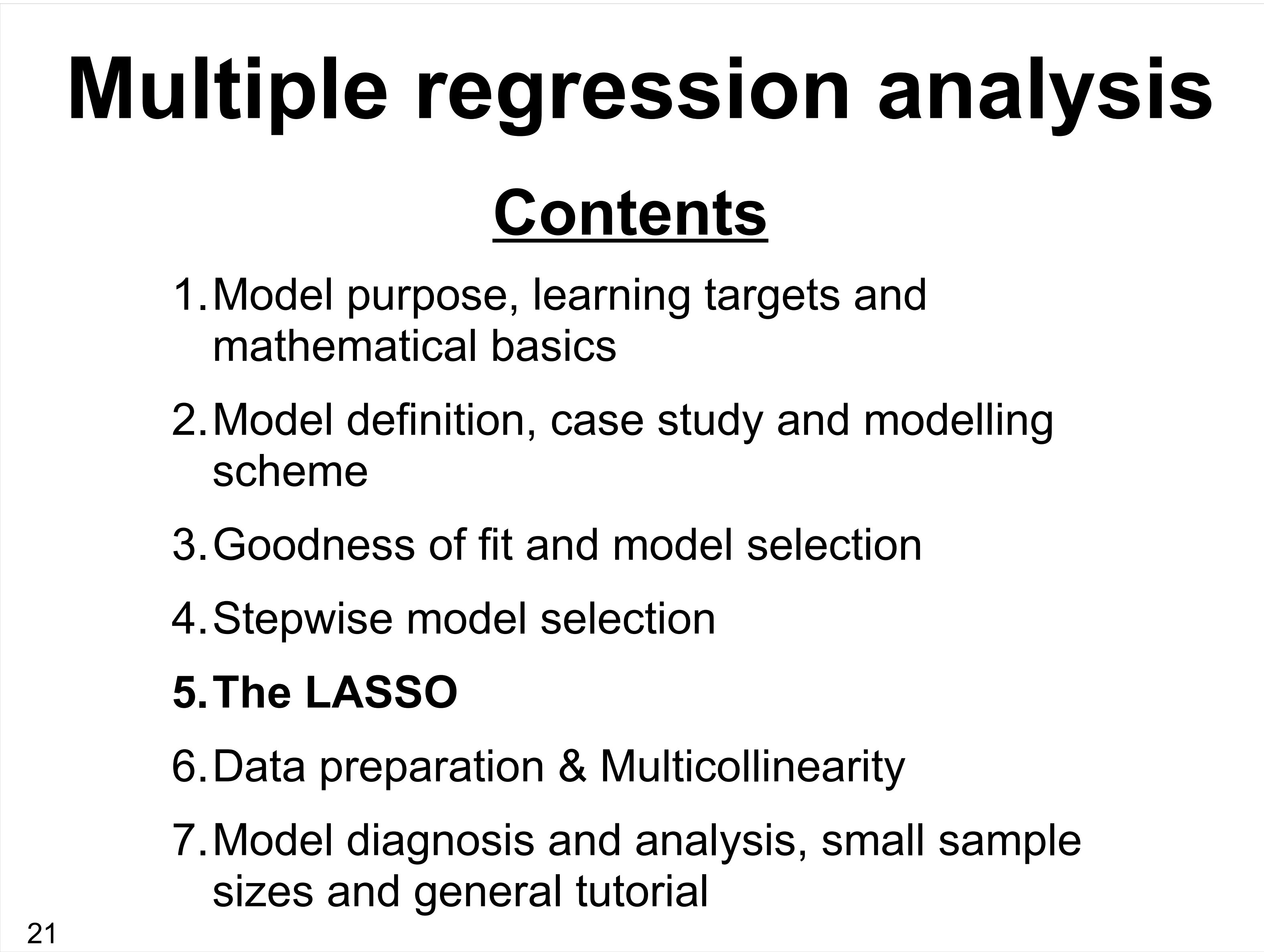
A simulation study found that backward stepwise elimination performed equally well as the LASSO in the identification of true predictors, particularly in conjunction with parameter-wise shrinkage (Houwelingen & Sauerbrei 2013). However, no approach performed best in all scenarios. Interestingly, backward stepwise elimination yielded often to more parsimonious (sparser) models than the LASSO (see next slides).

References

Austin P.C. (2008) Bootstrap model selection had similar performance for selecting authentic and noise variables compared to backward variable elimination: a simulation study. Journal of Clinical Epidemiology 61, 1009 – 1017.e1.

Fithian W., Taylor J., Tibshirani R. & Tibshirani R. (2015) Selective Sequential Model Selection. ArXiv e-prints, 1–36. http://adsabs.harvard.edu/abs/2015arXiv151202565F

Houwelingen H.C. van & Sauerbrei W. (2013) Cross-Validation, Shrinkage and Variable Selection in Linear Regression Revisited. Open Journal of Statistics 03, 79–102.

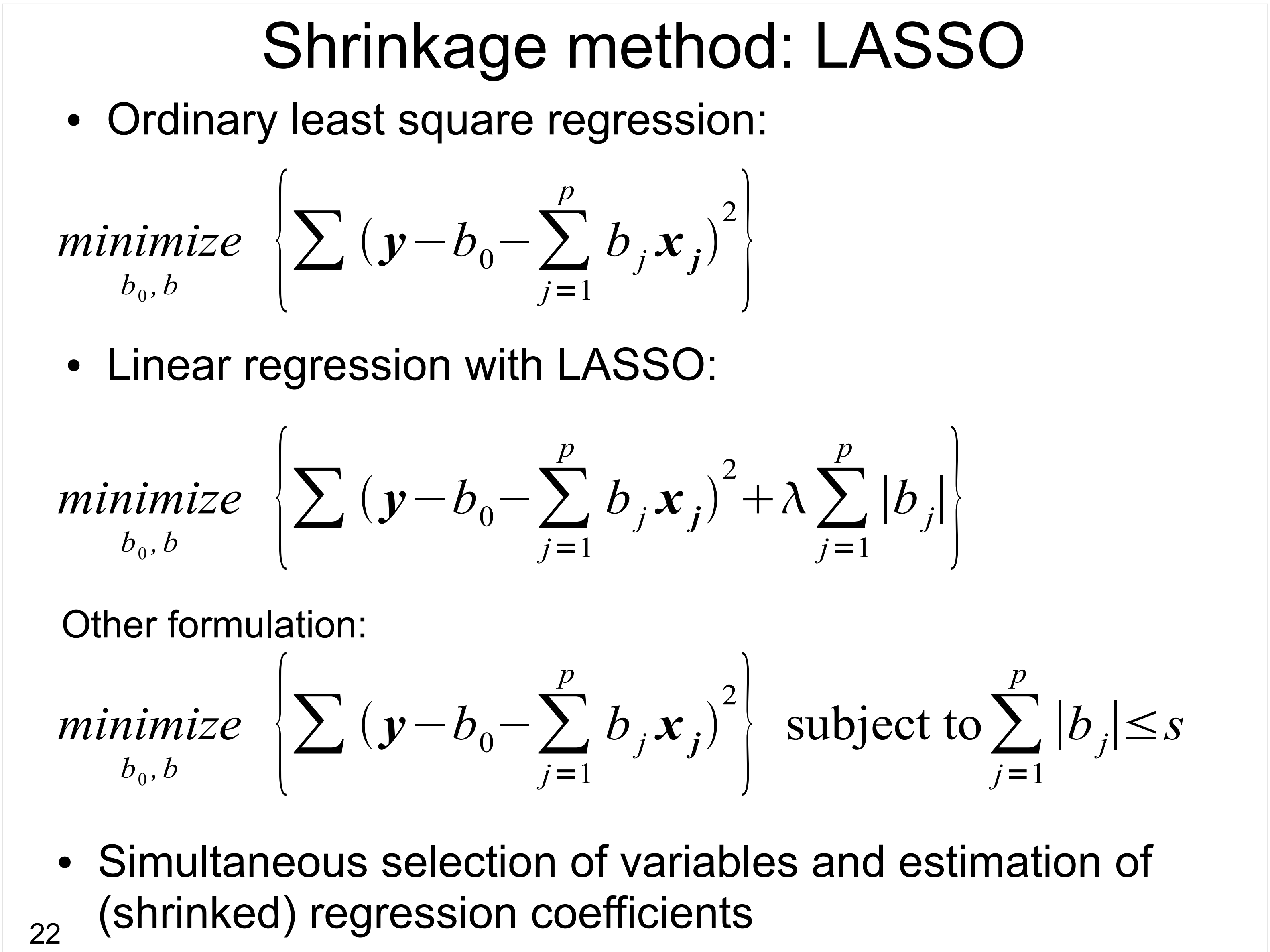


Multiple regression analysis

Contents

1. Model purpose, learning targets and mathematical basics
2. Model definition, case study and modelling scheme
3. Goodness of fit and model selection
4. Stepwise model selection
5. **The LASSO**
6. Data preparation & Multicollinearity
7. Model diagnosis and analysis, small sample sizes and general tutorial

21



Shrinkage method: LASSO

- Ordinary least square regression:

$$\underset{b_0,\, b}{minimize} \left\{ \sum \left(\boldsymbol{y} - b_0 - \sum_{j=1}^{p} b_j \boldsymbol{x_j} \right)^2 \right\}$$

- Linear regression with LASSO:

$$\underset{b_0,\, b}{minimize} \left\{ \sum \left(\boldsymbol{y} - b_0 - \sum_{j=1}^{p} b_j \boldsymbol{x_j} \right)^2 + \lambda \sum_{j=1}^{p} |b_j| \right\}$$

Other formulation:

$$\underset{b_0,\, b}{minimize} \left\{ \sum \left(\boldsymbol{y} - b_0 - \sum_{j=1}^{p} b_j \boldsymbol{x_j} \right)^2 \right\} \text{ subject to} \sum_{j=1}^{p} |b_j| \leq s$$

- Simultaneous selection of variables and estimation of (shrinked) regression coefficients

The equation for ordinary least square regression is identical to the one introduced earlier (see this slide), where we discussed that the aim in ordinary least square regression (OLS) is to minimize the SSE.

Using the LASSO is motivated by two problems of OLS regression:

1. Regression coefficients are biased high, leading to high variance in prediction. Shrinking of regression coefficients increases the bias (remember the bias-variance tradeoff) but reduces the variance.
2. For a large number of predictors, interpretation of the OLS result becomes tricky. Focusing on a smaller subset improves interpretation.

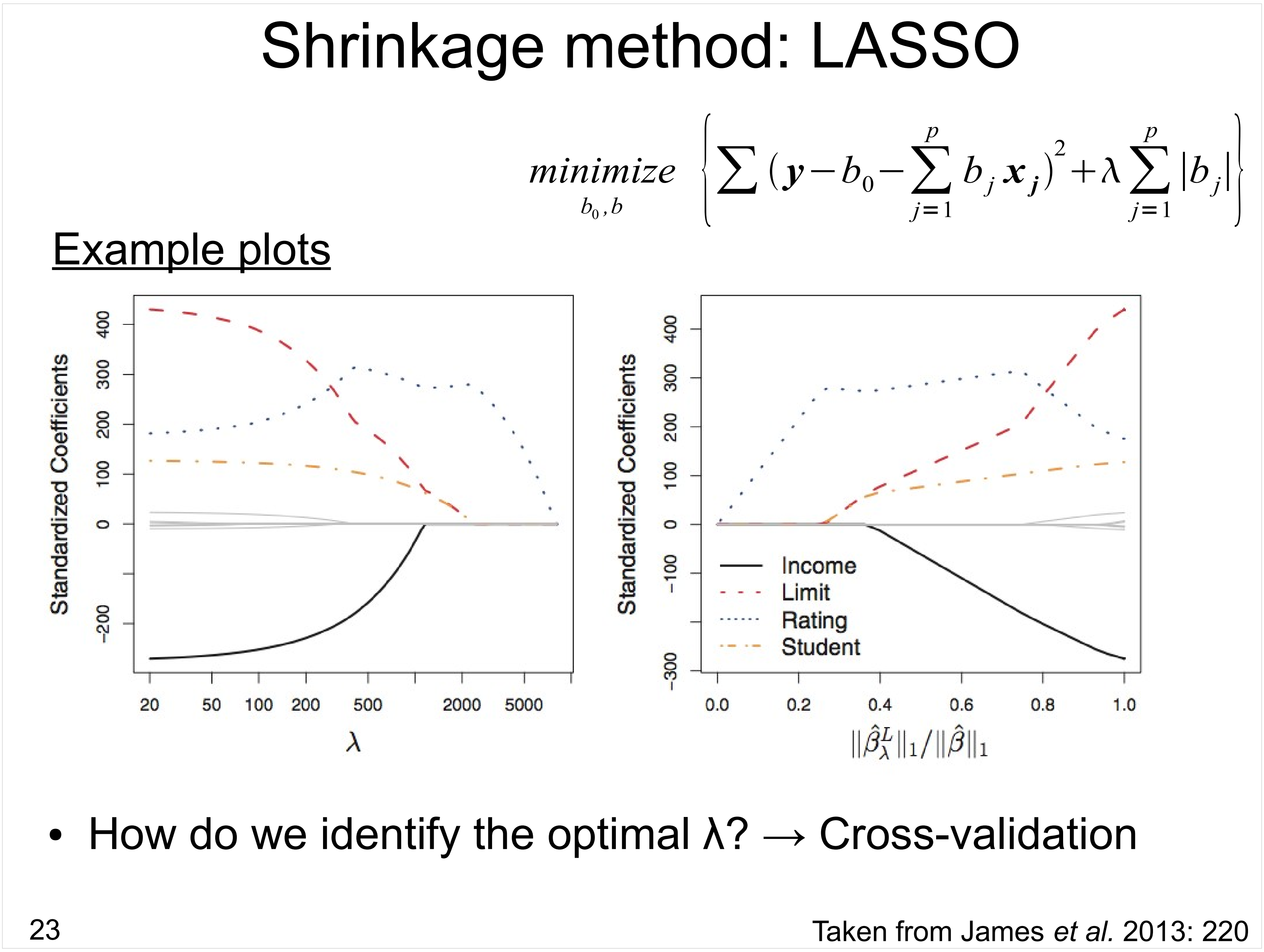
Through introduction of the penalty term in LASSO, the regression coefficients are shrunk. With increasing penalty (i.e. larger λ and smaller *s*, respectively) some regression coefficients are shrunk to 0, which means that the LASSO results in variable selection.

The penalty term is called *ℓ¹*-norm in statistical terms, where norm is a function that assigns a value to a vector (in our case to the vector of regression coefficients) and *ℓ¹* defines a specific mathematical space.

Note that the LASSO assumes sparsity, i.e. that in case of a large number of predictors only a small subset is relevant. Moreover, the relevant predictors should not have a high intercorrelation (i.e. collinearity) with non-relevant variables. Collinearity is discussed in more detail later.

A comparison of regression using LASSO and several other techniques can be found in Hastie, Tibshirani & Friedman (2011: 57ff). The LASSO is typically among the methods with the lowest prediction error.

For application in R see James et al. (2013) and for further developments of shrinkage (or more precise: sparse) methods see Hastie, Tibshirani & Wainwright (2015). All these books are freely downloadable, the URLs are provided with the literature list.



Shrinkage method: LASSO

$$\underset{b_0,b}{minimize}\left\{\sum\left(\boldsymbol{y}-b_0-\sum_{j=1}^{p}b_j\boldsymbol{x}_j\right)^2+\lambda\sum_{j=1}^{p}|b_j|\right\}$$

Example plots

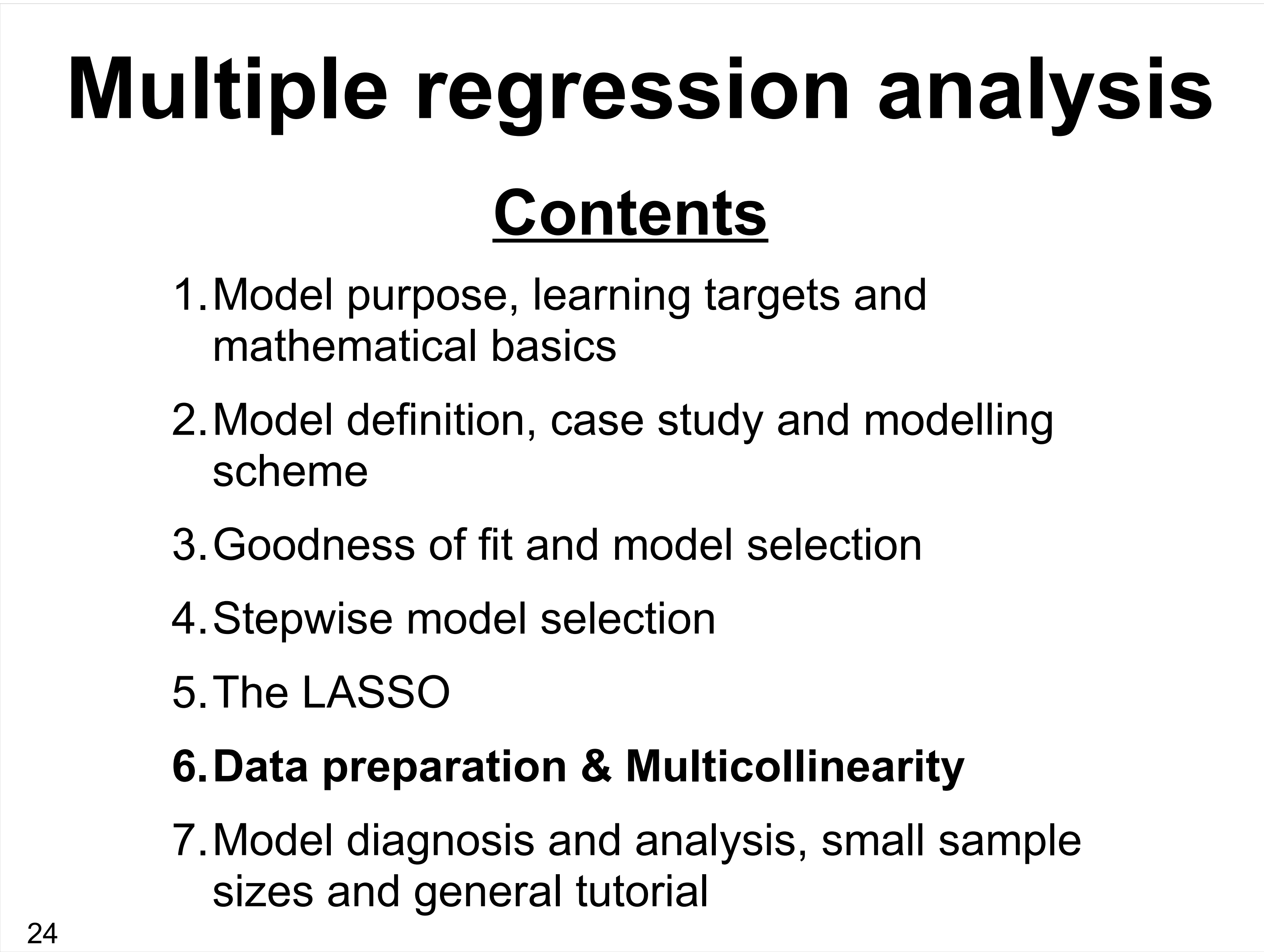
- How do we identify the optimal λ? → Cross-validation

Taken from James *et al.* 2013: 220

For the LASSO analysis, variables are typically standardized to mean of zero and standard deviation of one. Note that this leads to a zero intercept, i.e. b_0 can be removed in computation.

The left hand plot shows the standardised regression coefficients along increasing λ on the x axis. For very low values of λ, the regression coefficients are the same as for OLS regression. As λ becomes very high, all regression coefficients are shrunk towards zero and eventually we obtain the null model.

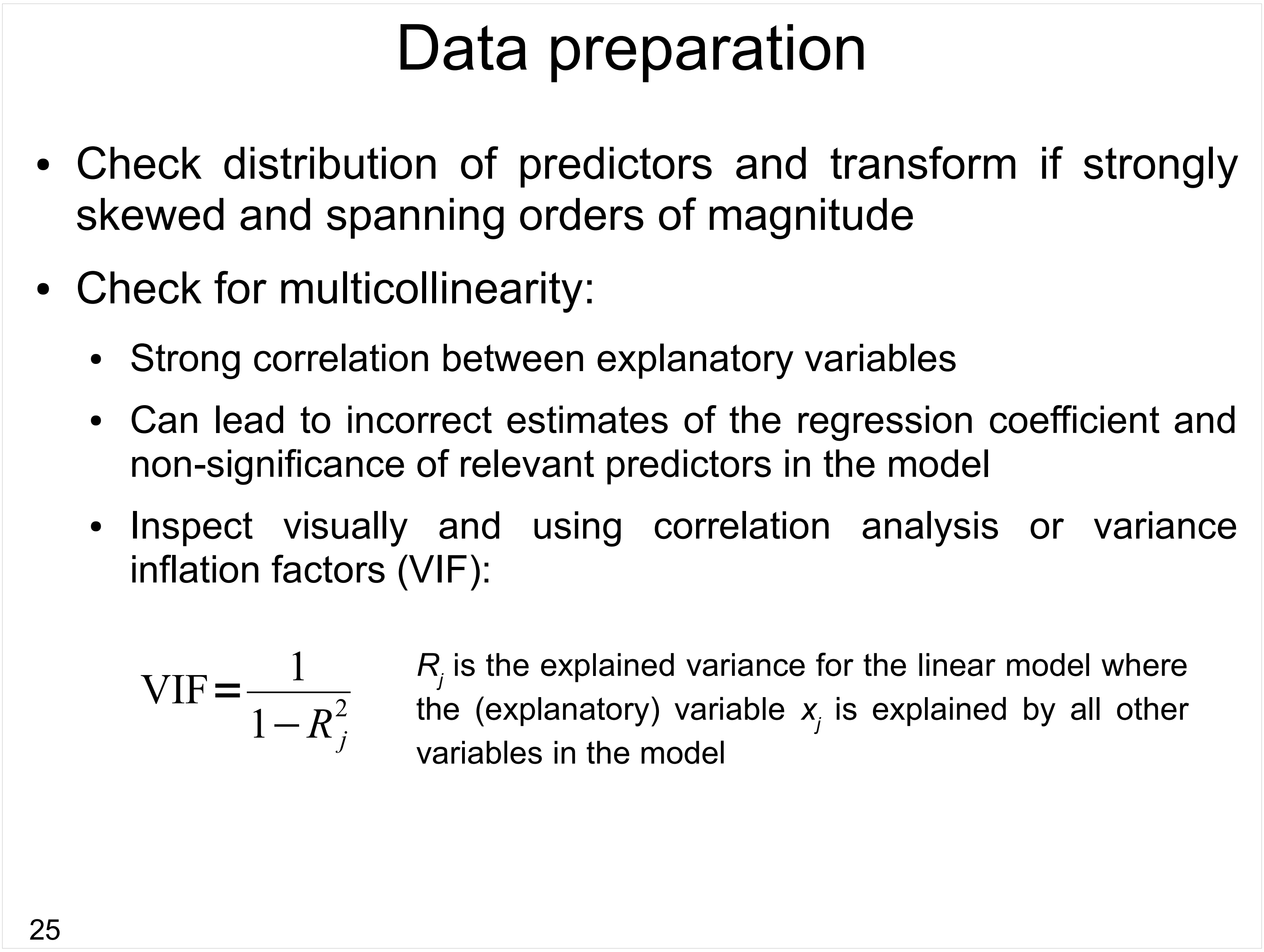
The right hand plot displays the ratio of the absolute sum of the standardized regression coefficients for the LASSO (i.e. ℓ^1-norm) and the absolute sum of the standardized regression coefficients from OLS. How to choose an optimal λ using cross validation is shown in the R demonstration.



Multiple regression analysis

Contents

1. Model purpose, learning targets and mathematical basics
2. Model definition, case study and modelling scheme
3. Goodness of fit and model selection
4. Stepwise model selection
5. The LASSO
6. **Data preparation & Multicollinearity**
7. Model diagnosis and analysis, small sample sizes and general tutorial



Data preparation

- Check distribution of predictors and transform if strongly skewed and spanning orders of magnitude
- Check for multicollinearity:
  - Strong correlation between explanatory variables
  - Can lead to incorrect estimates of the regression coefficient and non-significance of relevant predictors in the model
  - Inspect visually and using correlation analysis or variance inflation factors (VIF):

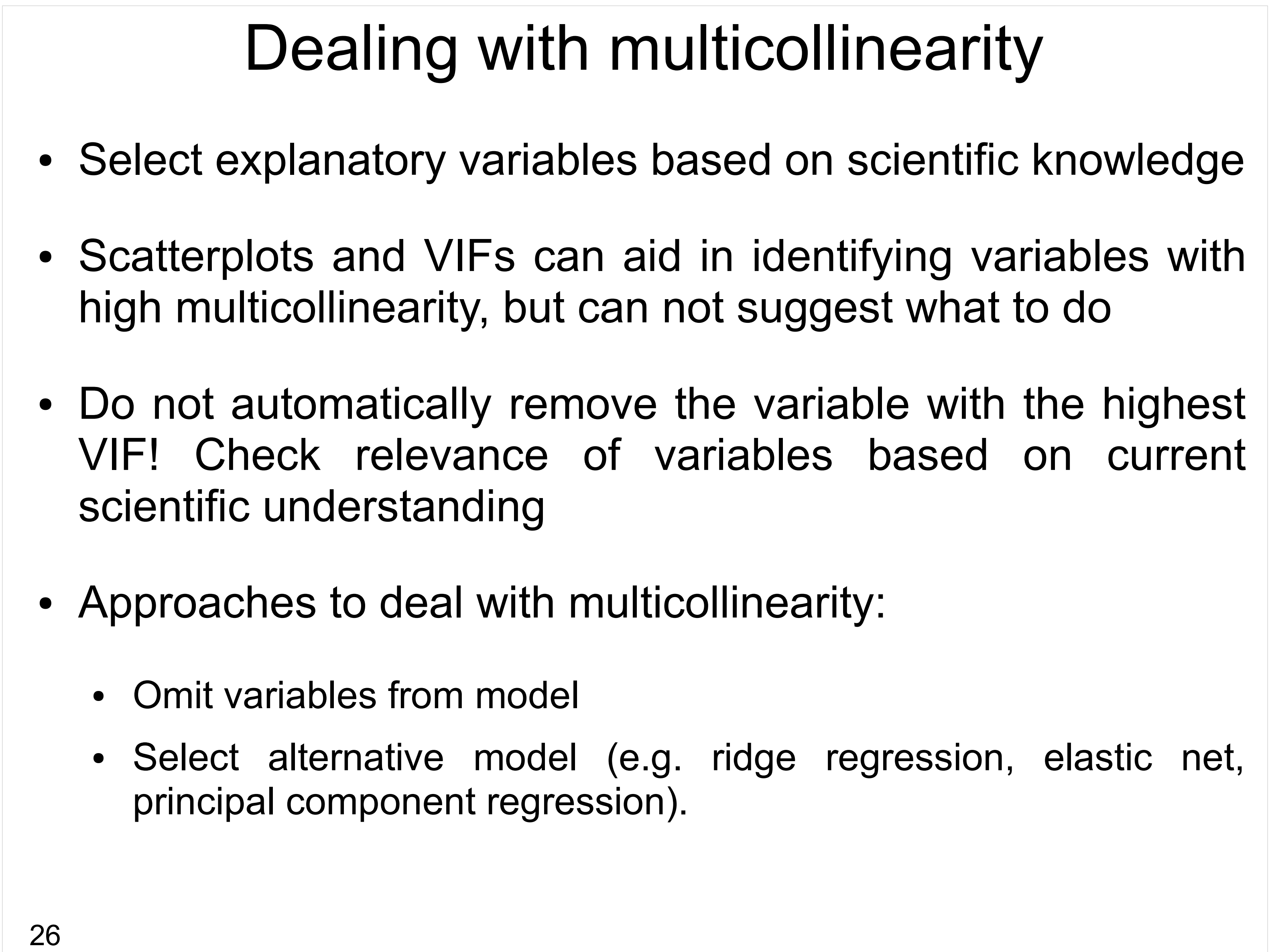
$$\mathrm{VIF} = \frac{1}{1 - R_j^2}$$

R_j is the explained variance for the linear model where the (explanatory) variable x_j is explained by all other variables in the model

Generally, transformation of the explanatory variables is necessary, if they are highly skewed and the data span several orders of magnitude. Especially chemical data often exhibit a skewed distribution (due to detection limits) that should be transformed before multiple regression analysis.

Multicollinearity does not affect predictions made on the same data set, or new data sets, where variables exhibit a similar collinearity structure as for the original data.

Regarding the VIF there are different rules of thumb as to when one should worry about collinearity. Most textbooks suggest that for VIF values >4 (Fox 2008:309, Kabacoff 2011: 200) or >5 (Sheather 2009:203) collinearity may represent a problem. However, drawing a sharp line within a continuum of values (e.g. VIFs) remains to some extent arbitrary.



Dealing with multicollinearity

- Select explanatory variables based on scientific knowledge
- Scatterplots and VIFs can aid in identifying variables with high multicollinearity, but can not suggest what to do
- Do not automatically remove the variable with the highest VIF! Check relevance of variables based on current scientific understanding
- Approaches to deal with multicollinearity:
  - Omit variables from model
  - Select alternative model (e.g. ridge regression, elastic net, principal component regression).

For a comparison of methods how to deal with collinearity see Dormann et al. (2012) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 35: 001-020.

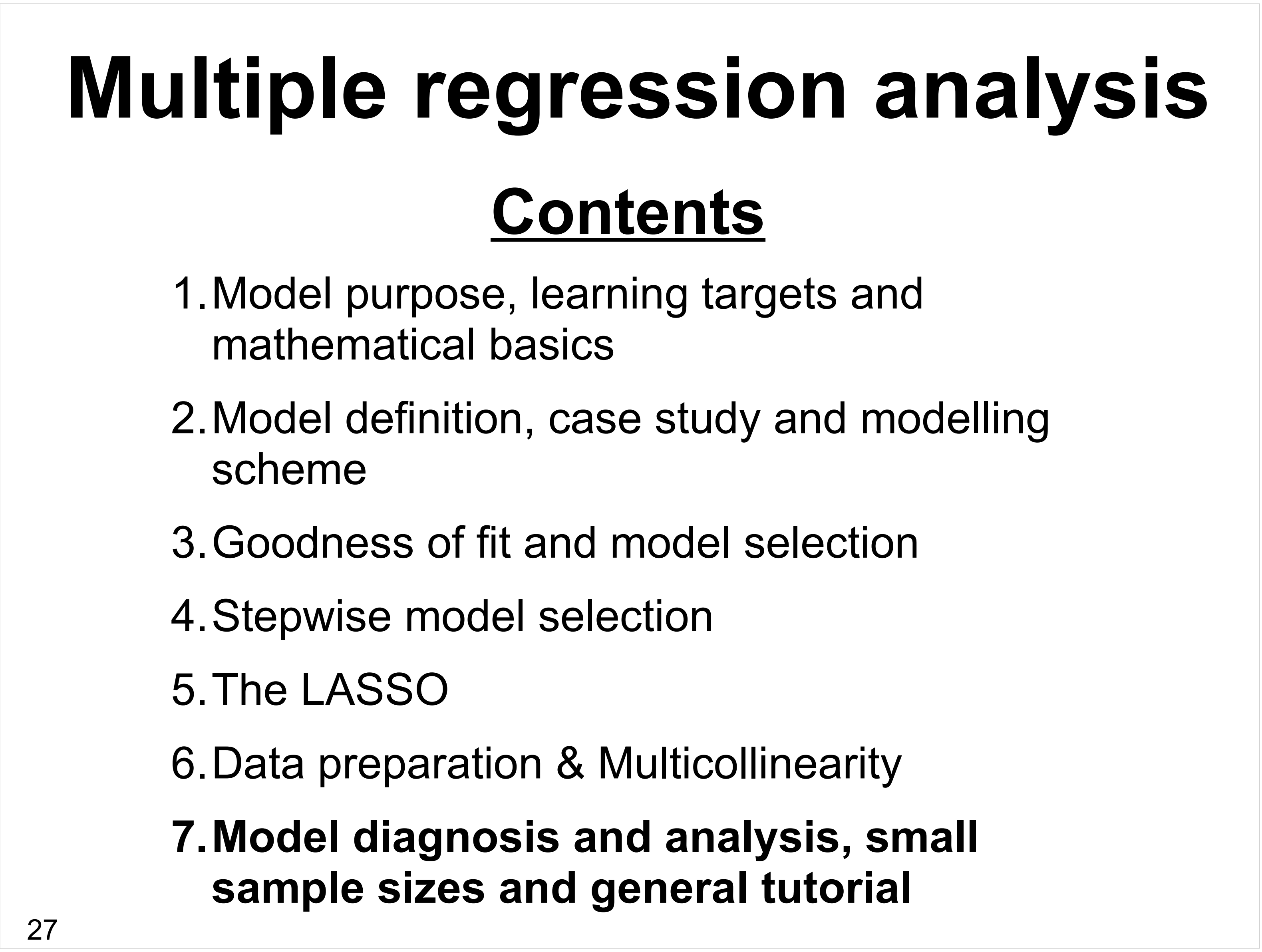
Interestingly, threshold-based pre-selection (i.e. omission of one of the variables if pairwise correlation exceeds a threshold, e.g. Pearson correlation coefficient $|r| > 0.7$.) was not outperformed by more sophisticated methods.

Ridge regression represents a shrinkage method similar to the LASSO, but uses a quadratic penalty term called *ℓ²*-norm (see James, Witten, Hastie and Tibshirani 2013: 61ff):

$$\underset{b_0,\, b}{minimize} \left\{ \sum \left(\boldsymbol{y} - b_0 - \sum_{j=1}^{p} b_j \boldsymbol{x_j} \right)^2 + \lambda \sum_{j=1}^{p} b_j^2 \right\}$$

Compared to the LASSO, ridge regression can better deal with collinear variables, but does not perform variable selection (i.e. regression coefficients are not shrunk to zero). Both represent special cases of the elastic net, which combines both *ℓ¹ and ℓ²* penalties and is briefly discussed later.

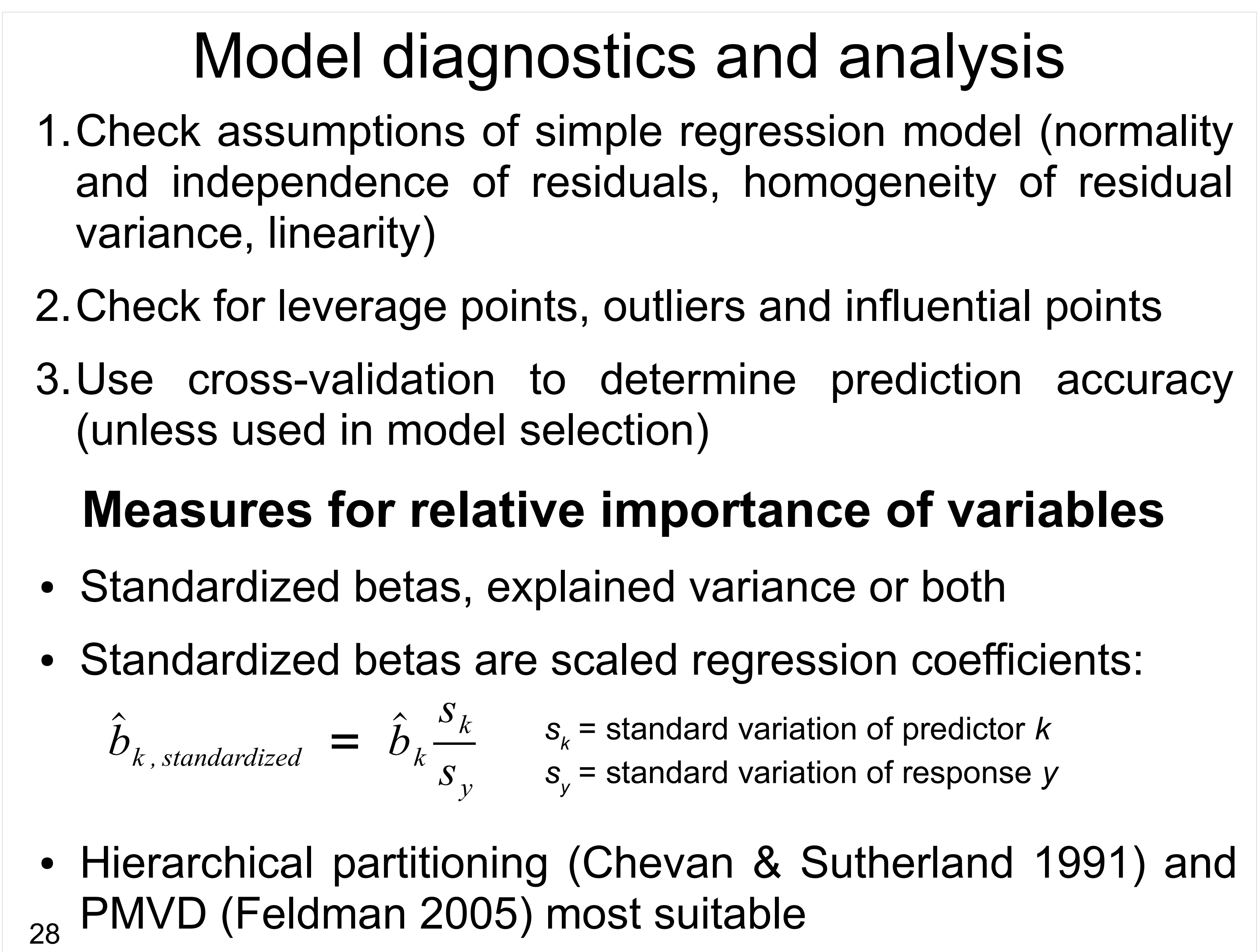
Principal component regression is covered later in the course. A comparison of methods including the LASSO, ridge regression and principle component regression can be found in Hastie, Tibshirani & Friedman (2011: 57ff).



Multiple regression analysis

Contents

1. Model purpose, learning targets and mathematical basics
2. Model definition, case study and modelling scheme
3. Goodness of fit and model selection
4. Stepwise model selection
5. The LASSO
6. Data preparation & Multicollinearity
7. **Model diagnosis and analysis, small sample sizes and general tutorial**



Model diagnostics and analysis

1. Check assumptions of simple regression model (normality and independence of residuals, homogeneity of residual variance, linearity)
2. Check for leverage points, outliers and influential points
3. Use cross-validation to determine prediction accuracy (unless used in model selection)

Measures for relative importance of variables

- Standardized betas, explained variance or both
- Standardized betas are scaled regression coefficients:

$$\hat{b}_{k\,,\,standardized} = \hat{b}_k \frac{s_k}{s_y}$$

s_k = standard variation of predictor *k*
s_y = standard variation of response *y*

- Hierarchical partitioning (Chevan & Sutherland 1991) and PMVD (Feldman 2005) most suitable

As mentioned before, the independence of residuals may not hold true for time series or spatial data. If temporal or spatial autocorrelation exists, this should be included in the model structure for example using a generalised least squares model.

One possibility to cope with non-linearity or non-normality is to transform the response variable (see Sheather 2009: 167 ff). The box-cox transformation is a widely used technique that can transform most variables to normal distribution. Note that if the data can be modelled with a generalised linear model, then this is preferable to transformation (cf. Warton and Hui 2011; Szöcs & Schäfer 2015).

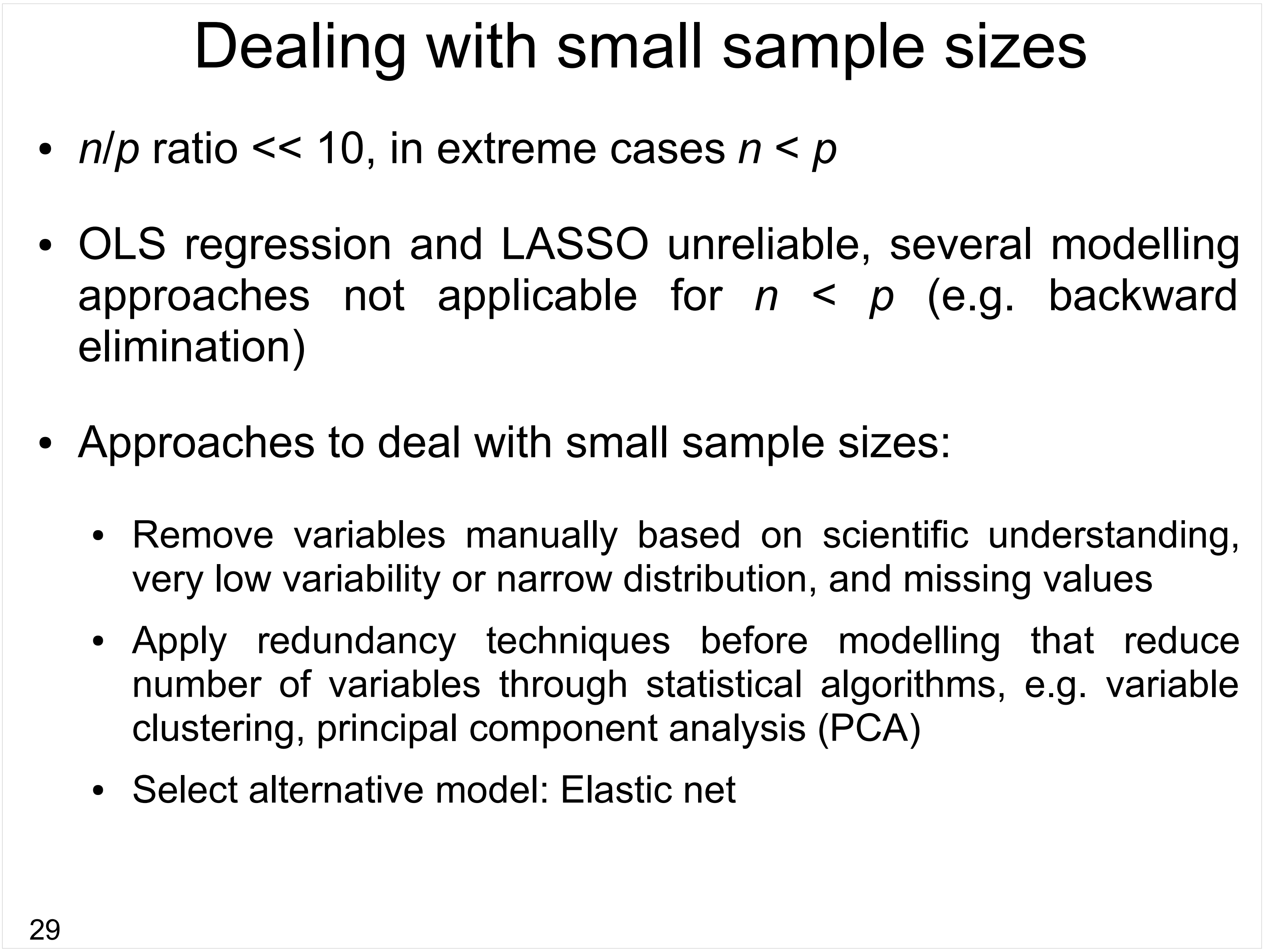
Standardized betas are contested as they are not related to the partitioning of R^2 (unless the explanatory variables are non-correlated) and do not account for the direct effect of a variable in the model (for example: high direct effect of predictor may be assigned to correlating predictors and result in low beta). Nevertheless, they inform on the change of the response variable for one unit change in the predictor.

Note that hierarchical partitioning is a general variable importance measure that can be used with any model that provides a goodness of fit metric. For linear models and R^2 as metric, hierarchical partitioning is equivalent to the LMG method that is often mentioned in journal articles.

For a general discussion of variable importance measures refer to: Grömping U. (2015) Variable importance in regression models. *WIREs Comput Stat* 7, 137–152, Johnson & Lebreton (2004). History and Use of Relative Importance Indices in Organizational Research. *Organizat Res Meth*, 7, 238–257 or with a special emphasis on R: Grömping, U. (2006) Relative importance for linear regression in R: The package relaimpo. *JSS,* **17**.

Warton, D.I., and Hui, F.K.C. (2011). The arcsine is asinine: the analysis of proportions in ecology. Ecology 92, 3–10.

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



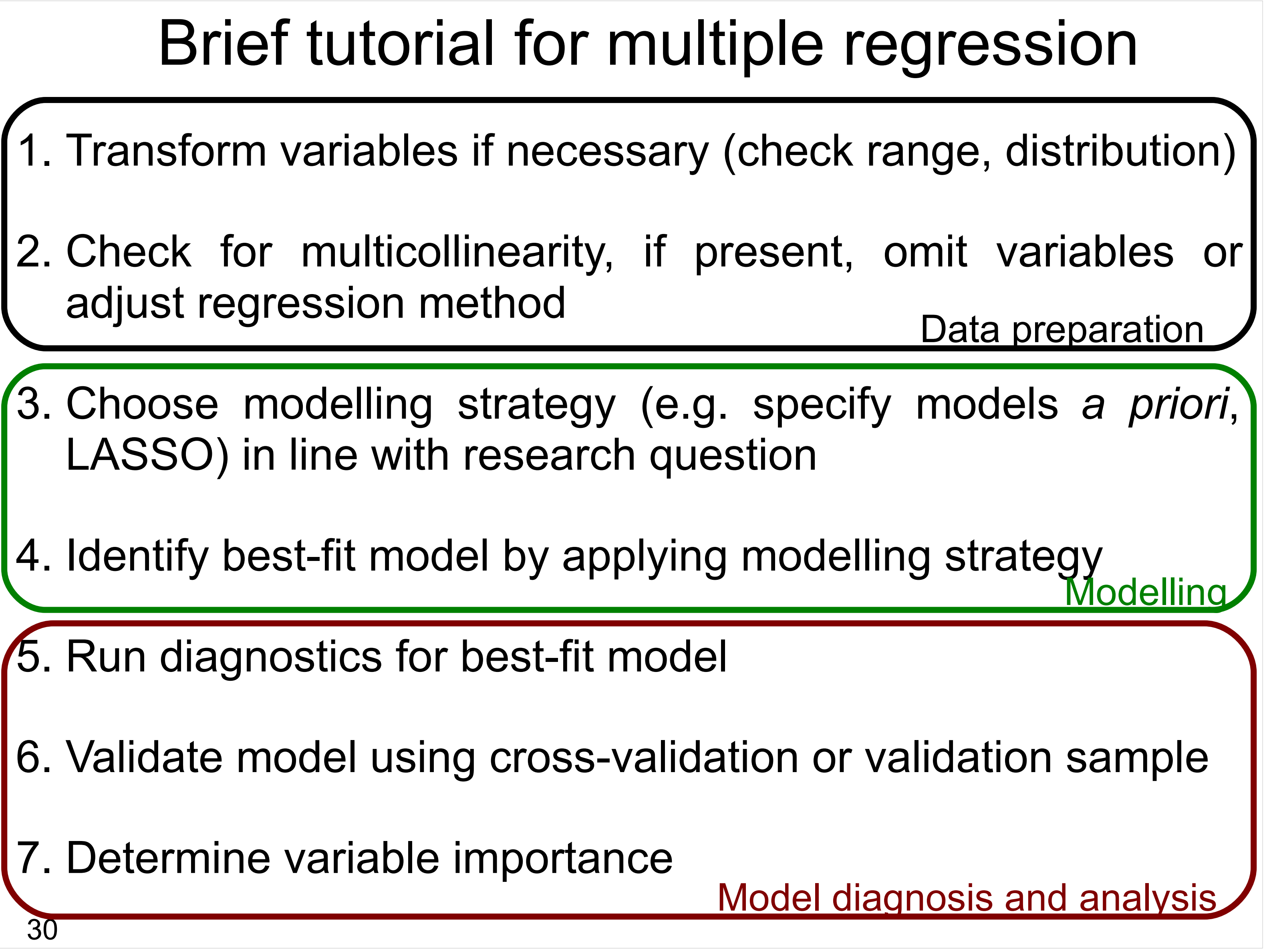
Dealing with small sample sizes

- *n*/*p* ratio << 10, in extreme cases *n* < *p*
- OLS regression and LASSO unreliable, several modelling approaches not applicable for *n* < *p* (e.g. backward elimination)
- Approaches to deal with small sample sizes:
  - Remove variables manually based on scientific understanding, very low variability or narrow distribution, and missing values
  - Apply redundancy techniques before modelling that reduce number of variables through statistical algorithms, e.g. variable clustering, principal component analysis (PCA)
  - Select alternative model: Elastic net

n = sample size; *p* = parameters in model

Running a PCA before regression analysis is also called principal component regression. We will comprehensively discuss PCA and cluster analysis later in the course. Briefly, PCA constructs new, non-correlated gradients from a data set and in case of collinearity this can help to reduce the number of variables. Similarly, cluster analysis identifies similar groups of variables, where group representatives could be subsequently selected for modeling. See Harrell (2015: p. 79ff) for further details and techniques. The approach is particularly powerful, if the predictors are strongly correlated, for example, in the case of bioclimatic or water quality variables. For an application see: Bhowmik & Schäfer (2015) Large Scale Relationship between Aquatic Insect Traits and Climate. PLoS ONE 10, e0130025. Freely accessible at: http://dx.doi.org/10.1371%2Fjournal.pone.0130025

The elastic net combines both ℓ^1 *and* ℓ^2 penalties of the LASSO and ridge regression. It can deal with *n* < *p* situations and with collinear variables. For details see: Zou H. & Hastie T. (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320.



Brief tutorial for multiple regression

1. Transform variables if necessary (check range, distribution)

2. Check for multicollinearity, if present, omit variables or adjust regression method

Data preparation

3. Choose modelling strategy (e.g. specify models *a priori*, LASSO) in line with research question

4. Identify best-fit model by applying modelling strategy

Modelling

5. Run diagnostics for best-fit model

6. Validate model using cross-validation or validation sample

7. Determine variable importance

Model diagnosis and analysis

30

For further details on regression modeling strategies see Harrell (2015) p. 63ff, with more detailed check lists for different modeling objectives (e.g. prediction or effect estimation) on the pages 94ff.

You can find an overview on the implementation of standard techniques for multiple regression here: http://www.statmethods.net/stats/regression.html.