Kurs Bio144:
Datenanalyse in der Biologie

Stefanie Muff & Owen L. Petchey

Lecture 7: Model selection

19./20. April 2018

Overview

- Model selection and model checking.
- Automatic model selection and its caveats.
- Model selection bias.
- Selection criteria: AIC, AIC_c, BIC
- Collinearity of covariates
- Explanation vs prediction.
- Occam's razor principle.

Course material covered today

The lecture material of today is based on the following literature:

- "Lineare regression" chapters 5.1-5.4
- Chapter 27.1 and 27.2 by Clayton and Hills "Choice and Interpretation of Models" (pdf provided)

Optional reading:

- Paper by Freedman (1983): "A Note on Screening Regression Equations" (Sections 1 and 2 are sufficient to get the point)

Developing a model

So far, our regression models "fell from heaven": The model family and the terms in the model were almost always given.

However, it is often not clear *a priori* which terms are relevant for a model.

The two **extreme situations** are

1. It is clear/known that $\mathbf{y}$ depends on a set of regressors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}$.
2. The study has the aim to find connections between the outcome $\mathbf{y}$ and the regressors. It is not known *how* or *if* each regressor influences $\mathbf{y}$.

The reality lies often in between:
Interest centers around one predictor (e.g., a new medication), but the effect of other potential influence factors must be taken into account.

Why is finding a model so hard?

Remember from week 1:

> Ein Modell ist eine Annäherung an die Realität. Das Ziel der Statistik und Datenanalyse ist es immer, dank Vereinfachungen der wahren Welt gewisse Zusammenhänge zu erkennen.

Box (1979): "All models are wrong, but some are useful."

$\rightarrow$ There is often not a "right" or a "wrong" model – but there are more and less useful ones.

$\rightarrow$ Finding a model with good properties is sometimes an art...

→ Even among statisticians there is no real consensus about how (or if!) to do model selection:

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7**, 679–692 doi: 10.1111/2041-210X.12541

SPECIAL FEATURE: 5TH ANNIVERSARY OF *METHODS IN ECOLOGY AND EVOLUTION*

The relative performance of AIC, AIC_C and BIC in the presence of unobserved heterogeneity

Mark J. Brewer[1,*], Adam Butler[2] and Susan L. Cooksley[3]

[1]Biomathematics and Statistics Scotland, Craigiebuckler, Aberdeen, AB15 8QH,UK; [2]Biomathematics and Statistics Scotland, JCMB, The King's Buildings, Edinburgh, EH9 3JZ,UK; and [3]The James Hutton Institute, Craigiebuckler, Aberdeen, AB15 8QH, UK

Summary

1. Model selection is difficult. Even in the apparently straightforward case of choosing between standard linear regression models, there does not yet appear to be consensus in the statistical ecology literature as to the right approach.

Note: The first sentence of a paper in *Methods in Ecology and Evolution* from 2016 is: "Model selection is difficult."

Mercury example

Let us look at the mercury example. The **research question** was:
"Gibt es einen Zusammenhang zwischen Quecksilber(Hg)-Bodenwerten von Wohnhäusern und der Hg-Belastung im Körper (Urin, Haar) der Bewohner?"

- *Hg concentration in urine* (Hg_{urine}) is the **response**.
- *Hg concentration in the soil* (Hg_{soil}) is the **predictor of interest**.

In addition, the following variables were monitored for each person, because they might influence the mercury level in a person's body:

smoking status; number of amalgam fillings; age; number of monthly fish meals; indicator if fish was eaten in the last 3 days; mother vs child; indicator if vegetables from garden are eaten; migration background; height; weight; BMI; sex; education level.

Thus: In total additional 13 variables!

How many variables can I include in my model?

Rule of thumb:

> Include no more than $n/10$ (10% of n) variables into your linear regression model, where n is the number of data points.

In the mercury example there are 156 individuals, so a **maximum of 15 variables** should be included in the model.

Remarks:

- Categorical variables with k levels already require $k - 1$ dummy variables. For example, if 'education level' has three categories, 2 variables are used up.
- Whenever possible, the model should **not be blown up** unneccessarily. Even if there are many data points, the use of too many variables may lead to an **overfitted** model.

  $\longrightarrow$ See https://en.wikipedia.org/wiki/Overfitting.

In the mercury study, the following variables were included using *a priori* knowledge:

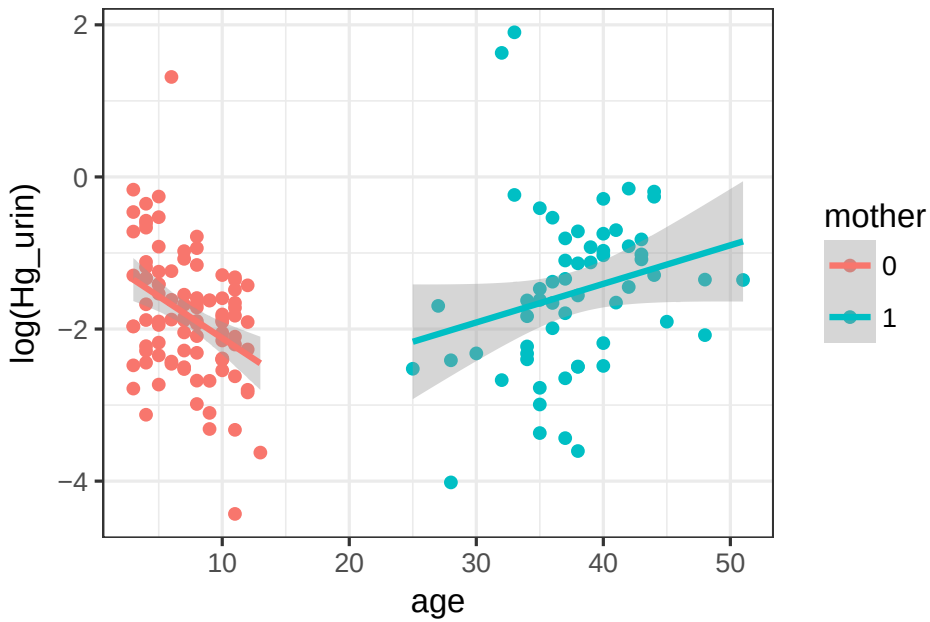
Variable	Meaning	type	transformation
Hg_urin	Hg conc. in urine (response)	continuous	log
Hg_soil	Hg conc. in the soil	continuous	log
vegetables	Eats vegetables from garden?	binary	
migration	Migration background	binary	
smoking	Smoking status	binary	
amalgam	No. of amalgam fillings	count	$\sqrt{.}$
age	Age of participant	continuous	
fish	Number of fish meals/month	count	$\sqrt{.}$
last_fish	Fish eaten in last 3 days?	binary	
mother	Mother or child?	binary	

Let us now fit the full model (including all covariates) in R:

	Coefficent	95%-confidence interval	p-value
Intercept	-0.94	from -1.10 to -0.79	< 0.0001
log10(Hg_soil)	0.03	from -0.05 to 0.11	0.47
vegetables	0.079	from -0.03 to 0.19	0.15
migration	-0.048	from -0.21 to 0.12	0.57
smoking	0.23	from 0.01 to 0.45	0.039
sqrt(amalgam)	0.36	from 0.27 to 0.45	< 0.0001
age	-0.0073	from -0.02 to 0.01	0.32
mother	-0.04	from -0.50 to 0.42	0.86
sqrt(fish)	0.087	from 0.03 to 0.14	0.003
last_fish	0.29	from 0.13 to 0.44	0.0003

- There is no evidence that mercury in soil influences mercury in urine ($p = 0.47$).
- We find $R^2 = 0.40$. Is this "good"?
- Are there additional terms that might be important?

Remember from slides 7-10 of week 4 that the dependency of mercury in urine on age is different for mothers and children:



→ We need to include an interaction term *mother·age*.

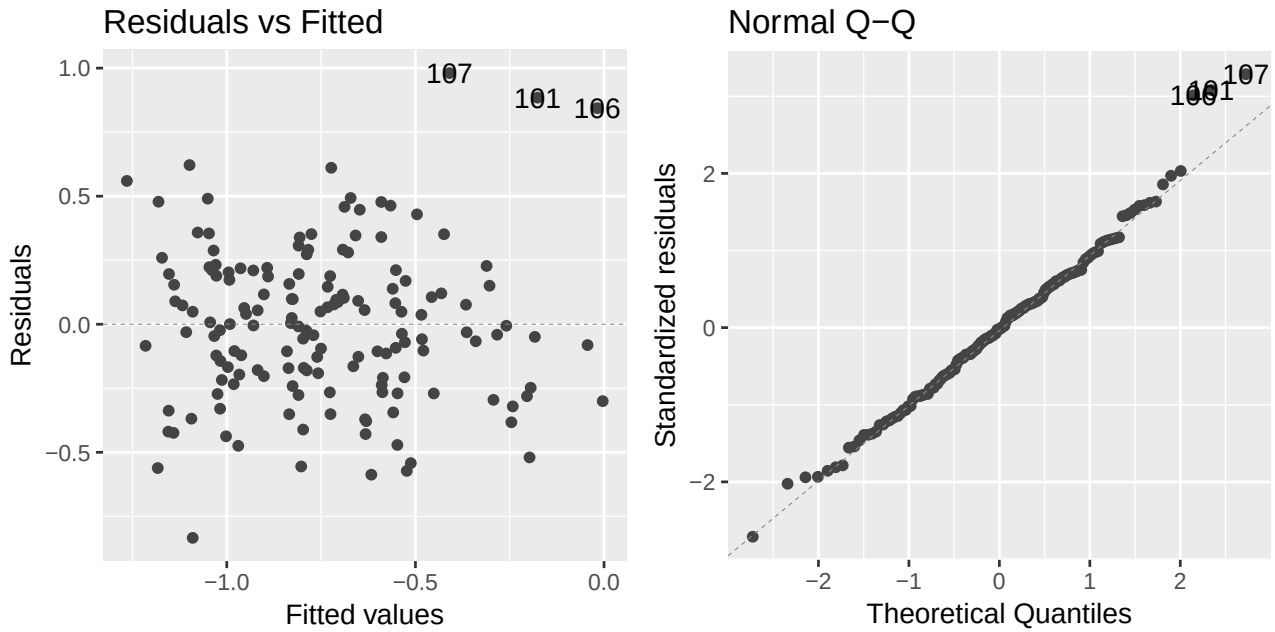
Fitting the model again with the additional term:

	Coefficent	95%-confidence interval	p-value
Intercept	-0.68	from -0.88 to -0.47	< 0.0001
log10(Hg_soil)	0.033	from -0.05 to 0.11	0.42
vegetables	0.07	from -0.03 to 0.17	0.18
migration	-0.036	from -0.19 to 0.12	0.65
smoking	0.27	from 0.06 to 0.48	0.012
sqrt(amalgam)	0.33	from 0.24 to 0.42	< 0.0001
age	-0.042	from -0.06 to -0.02	0.0004
mother1	-1.03	from -1.70 to -0.35	0.003
sqrt(fish)	0.079	from 0.03 to 0.13	0.004
last_fish	0.30	from 0.15 to 0.45	< 0.0001
age:mother1	0.055	from 0.03 to 0.08	0.0002

- There is evidence that the interaction is relevant ($p < 0.001$).
- R^2 has now clearly increased to 0.45.

$\rightarrow$ The interaction term apparently improved the model.

A model checking step (always needed, but we did it already in weeks 4):



This looks ok, no need to improve the model from this point of view.

Even if the model checking step revealed no violations of the assumptions (the model seems to be fine), we sometimes want to know:

- Which of the terms are important/relevant?
- Are there additional terms that might be important?
- Would it be possible to further "improve" the model?

Often, the desire is to find a model that is in some sense "optimal" or "best".

Is importance reflected by p-values?

A widely used practice to determine the "importance" of a term is to look at the p value from the t or F-test and check if it falls below a certain threshold (usually $p < 0.05$).

However, there are a few problems with this approach:

- **A small p-value does not necessarily mean that a term is (biologically, medically) important – and vice versa!**
- When carrying out the tests with $H_0 : \beta_j = 0$ for all variables sequentially, one runs into a multiple testing problem.
  (Remember the ANOVA lecture, week 5, slide 23).
- The respective tests depend crucially on the correctness of the normality assumption.
- Covariates are sometimes collinear, which leads to more uncertainty in the estimation of the respective regression parameters, and thus to larger p-values.

For all these reasons, we **strongly disagree** with the remark in Stahel's script 5.2, second part in paragraph d.

The first part is ok:

> Man kann also nicht behaupten, dass ein Term mit signifikantem Test-Wert einen „statistisch gesicherten" Einfluss auf die Zielgrösse habe.

But we disagree with this:

> Statt die Tests für strikte statistische Schlüsse zu verwenden, begnügen wir uns damit, die P-Werte der t-Tests für die Koeffizienten (oder direkt die t-Werte) zu benützen, um die *relative* Wichtigkeit der entsprechenden Regressoren anzugeben, insbesondere um die „wichtigste" oder die „unwichtigste" zu ermitteln.

Automatic model selection procedures

It would be very convenient if there were **objective** or even **automatic** procedures to select the "best" model. Wouldn't it?

In fact, such procedures have been proposed in the past. For example:

- Forward selection:
  Start with a large/full model. In each step, remove the variable with the largest p-value. Do this until only variables with low p-values remain in the model.

- Backward selection:
  Start with an empty model. In each step, add the predictor with the highest importance (lowest p-value). Do this until none of the missing coefficients has a low p-value when adding it.

Model selection bias

Let us reproduce an illustrative example that was described by Freedman (1983).

Aim of the example:

To illustrate how model selection purely based on p-values can lead to biased parameters and overestimated effects.

Procedure:

1. Randomly generate 100 data points for 50 covariables $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(50)}$ and a response $\mathbf{y}$:

```
> set.seed(9856)
> data <- data.frame(matrix(rnorm(51*100),ncol=51))
> names(data)[51] <- "Y"
```

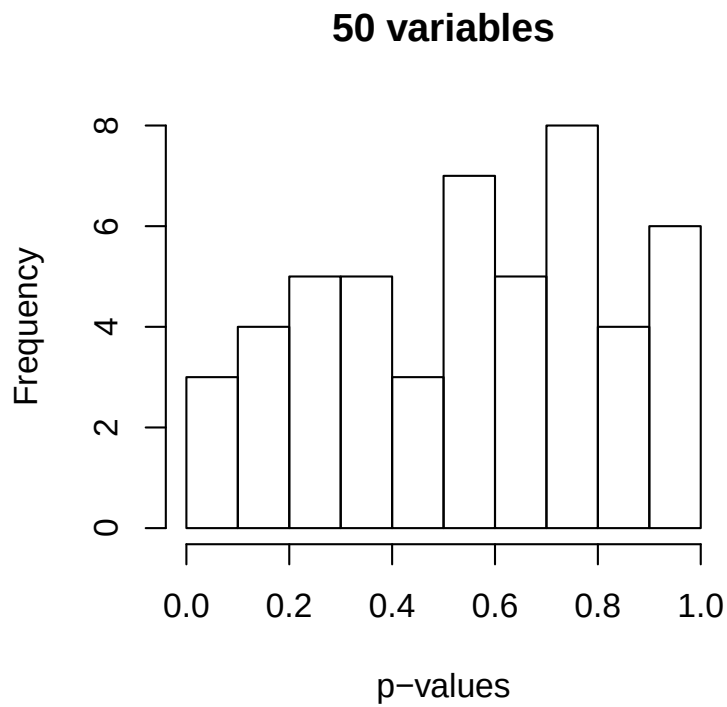
   data is a 100×51 matrix, where the last column is the response. The data were generated completely independent, the covariates do not have any explanatory power for the response!

2. Fit a linear regression model of **y** against all the 50 variables

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_{50} x_i^{(50)} + e_i \ .$$

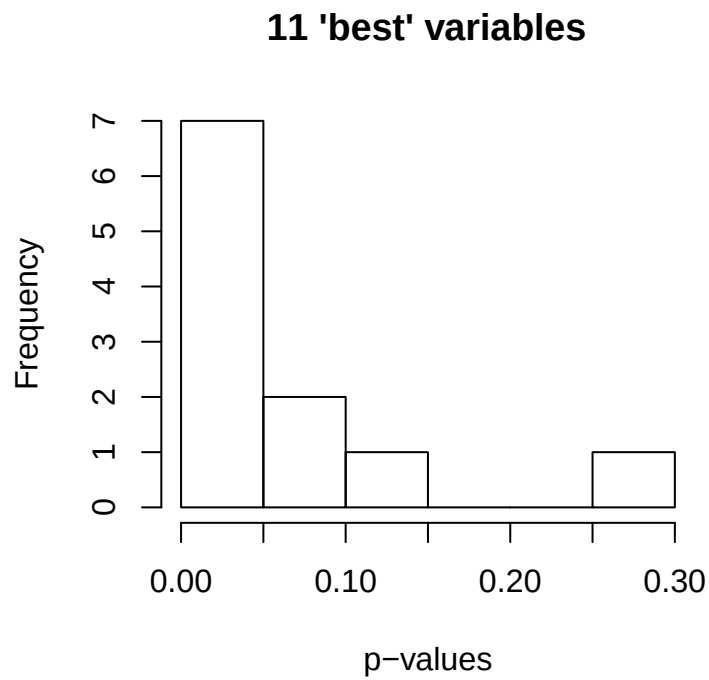
```
> r.lm <- lm(Y~.,data)
```

As expected, the distribution of the p-value is (more or less) uniform between 0 and 1:



3. Then pick all variables that have a $p < 0.25$ and re-fit the model using only these. 11 variables fulfil this criterion:

```
> # select all variables with p<0.25
> r.lm.red <- lm(Y ~ X1 + X2 + X4 + X5 + X14 + X18 + X20 + X27 + X36 + X38 + X49,data)
```



The distribution of the p-values is now skewed: many of them reach rather small values (7 have $p < 0.05$). This happened although none of the variables has any explanatory power!

Important note

Automatic model selection procedures may lead to biased parameter estimates and wrong conclusions!

See, e.g., Freedman (1983); Copas (1983).

Please note that **we therefore strongly discourage the use of automated model selection procedures**. So please ignore large parts of chapter 5.3 in the Stahel script!!

(Or read it to see how you should **not** do it.)

More modern ways to do model selection

Remember: R^2 is not suitable for model selection, because it *always* increases (improves) when a new variable is included.

In 2002, Burnham and Anderson suggested the use of so-called information-criteria for model selection.

The idea is to find a balance between

Good model fit $\leftrightarrow$ **Low model complexity**

$\rightarrow$ Penalize models with more parameters.

AIC

The most prominent criterion is the AIC (Akaike Information Criterion), which measures the quality of a model.

The AIC of a model with likelihood L and p parameters is given as

$$AIC = -2\log(L) + 2p \ .$$

Important: The lower the AIC, the better the model!

The AIC is a compromise between

- a high likelihood L (good model fit)
- few model parameters p (low complexity)

AIC_c: The AIC for low sample sizes

When the number of data points n is small with respect to the number of parameters p in a model, the use of a corrected AIC, the AIC_c is recommended.

The **corrected AIC** of a model with n data points, likelihood L and p parameters is given as

$$AIC_c = -2\log(L) + 2p \cdot \frac{n}{n-p-1} \ .$$

Burnham and Anderson **recommend to use AIC_c in general, but for sure when the ratio $n/p < 40$.**

In the mercury example, we have 156 data points and 11 parameters (including the intercetp β_0), thus $n/p = 156/11 \approx 14 < 40 \Rightarrow AIC_c$ should be used!

BIC, the brother/sister of AIC

Other information criteria were suggested as well. Another prominent example is the BIC (Bayesian Information Criterion), which is similar in spirit to the AIC.

The BIC of a model for n data points with likelihood L and p parameters is given as

$$BIC = -2\log(L) + p \cdot \ln(n) \ .$$

Again: The lower the BIC, the better the model!

The only difference to AIC is the complexity penalization. The BIC criterion is often **claimed to estimate the predictive quality** of a model. More recent research indicates that AIC and BIC perform well under different data structures (Brewer et al., 2016).

Don't worry: No need to remember all these AIC and BIC formulas by heart!

What you should remember:

AIC, AIC_c and BIC all have the **aim to find a good quality model by penalizing model complexity**.

Example of AIC use

(Essentially the same as BIC use.)

Remember that we first fitted the mercury example **without** (`r.lm1`) and **with** (`r.lm2`) the interaction *mother* · *age*. The model improvement is also confirmed by the **reduction in AIC$_c$** (which we used because $n/p < 40$):

```
> library(AICcmodavg)
> AICc(r.lm1)

[1] 109.2942

> AICc(r.lm2)

[1] 96.66951
```

Interpretation: The AIC$_c$ of the model with interaction is clearly lower, thus the model with interaction is to be preferred.

We can further play around with AIC and, for instance, fit a model withouth the binary *migration* variable:

```
> r.lm3 <- lm(log10(Hg_urin) ~ log10(Hg_soil) + vegetables  + smoking +
+              sqrt(amalgam) + age * mother + sqrt(fish) + last_fish,d.hg)
> AICc(r.lm3)
```

```
[1] 94.53696
```

Interpretation: We observe a further reduction of the AIC.

This "success" brings us to another idea:

Could we do model selection simply by minimizing the AIC?
Without actually "thinking"?

Model selection with AIC?

Given m potential variables to be included in a model.

- In principle it is possible to minimize the AIC over all 2^m possible models. Simply fit all models and take the "best" one (lowest AIC).
- Sometimes several models have a very similar, low AIC. It is then possible to do **model averaging**, e.g. over the 3 models with lowest AIC, by weighting them according to the AIC (not done in this course).

However:

- As always when one tries to circumvent "thinking", this may lead to unreasonable models.
- AIC is in some sense equivalent to p-values-based statistics for **nested models** (i.e., when the smaller model contains a subset of the larger model).

Another complication: Collinearity of covariates

(See Stahel chapter 5.4)

Given a set of covariates $\mathbf{x^{(1)}}, \mathbf{x^{(2)}}, \mathbf{x^{(3)}}, ..., \mathbf{x^{(p)}}$. If it is possible for one of the covariates to be written as a linear combination of the others

$$x_i^{(j)} = \sum_{k \neq j} c_k x_i^{(k)} \quad \text{for all} \quad i = 1, ..., n$$

the set of covariates is said to be collinear.

Examples:

- Three vectors in a 2D-plane are always collinear.
- A covariate can be written a a linear combination of two others: $\mathbf{x^{(j)}} = c_1 \cdot \mathbf{x^{(1)}} + c_2 \cdot \mathbf{x^{(2)}}$, then the three variables are collinear.

In statistics, the expression "collinearity" is also used when such a collinearity relationship is *approximately* true. For example, when two variables $\mathbf{x^{(1)}}$ and $\mathbf{x^{(2)}}$ have a high correlation.

What is the problem with collinearity?

I like to do simple (and extreme) examples to understand the point. An extreme form of collinearity is when two covariates are identical $\mathbf{x^{(1)}} = \mathbf{x^{(2)}}$. In the regression model

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + e_i \ ,$$

the slope coefficients β_1 and β_2 cannot be uniquely determined (there are many equally "optimal" solutions to the equation)!

When the variables are collinear, this problem is less severe, but the β coefficients can be estimated less precisely $\rightarrow$ standard errors too high $\rightarrow$ p-values too large.

Detecting collinearity

The variance inflation factor (VIF) is a measure for collinearity. It is defined for each covariate $\mathbf{x}^{(j)}$ as

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R^2 of the regression of $\mathbf{x}^{(j)}$ against all other covariates. (Note: if R_j^2 is large, this means large collinearity and thus a large VIF).

Examples:

- $R_j^2 = 0 \rightarrow$ no collinearity $\rightarrow$ VIF=$1/1 = 1$.
- $R_j^2 = 0.5 \rightarrow$ some collinearity $\rightarrow$ VIF=$1/(1\text{-}0.5) = 2$.
- $R_j^2 = 0.9 \rightarrow$ high collinearity $\rightarrow$ VIF=$1/(1\text{-}0.9) = 10$.

What do do against collinearity?

- Avoid it, e.g. in experiments.
- Remove the variable with an inacceptably high R_j^2 or VIF_j. The tolerance of VIFs are different in the literature and range from 4 to 10 as a maximum tolerable VIF.
- Be aware of it and interpret your results with the respective care.
- See also Stahel 5.4(i) for a "recipe".

Predictive and explanatory models

Before we continue to discuss model selection, let us introduce an important discrimination between models that aim at explanation and those that aim at prediction:

- **Explanatory models**: These are models that aim at understanding the (causal) relationship between covariates and the response.
  Example: The mercury study aims to understand if Hg-concentrations in the soil (covariable) influence the Hg-concentrations in humans (response).
- **Predictive models**: These are models that aim to predict the outcome of future subjects.
  Example: In the bodyfat example the aim is to predict people's bodyfat from factors that are easy to measure (age, BMI, weight,..).

Please note: The model selection strategy depends on this distinction.

Prediction vs explanation

When the aim is *prediction*, the best model is the one that best predicts the fate of a future subject. This is a well defined task and automatic strategies to find the model which is best in this sense are potentially useful.

However, when used for *explanation* the best model will depend on the scientific question being asked, **and automatic selection strategies have no place**.

(Clayton and Hills, 1993, chapters 27.1 and 27.2)

Your aim is prediction?

Ideally, the predictive ability of a model is tested by a cross-validation (CV) approach, but AIC, AIC_c, BIC are useful approximations to CV.

Automatic model selection procedures using criteria like AIC, AIC_c, BIC may threfore be useful.

▸ Find a description of the CV idea here.

Your aim is explanation?

Then please select your covariates according to a priori knowledge.

> It will rarely be necessary to include a large number of variables in the analysis, because only a few exposures are of genuine scientific interest in any one study, and there are usually very few variables of sufficient *a priori* importance for their potential confounding effect to be controlled for. Most scientists are aware of the dangers of analyses which search a long list of potentially relevant exposures. These are known as *data dredging* or *blind fishing* and carry a considerable danger of false positive findings. Such analyses are as likely to impede scientific progress as to advance it. There are similar dangers if a long list of potential confounders is searched, either with a view to explaining the observed relationship between disease and exposure or to enhancing it — findings will inevitably be biased. Confounders should be chosen *a priori* and not on the basis of statistical significance. In particular, variables which have been used in the design, such as matching variables, must be included in the analysis.

(Clayton and Hills, 1993, chapters 27.1 and 27.2)

Correct way of model selection for explanatory models

Automatic model selection strategies are thus **not recommended for explanatory models** (example: mercury study). We therefore recommend the following steps:

- **Think about a suitable model before you start**. This includes the model family (e.g., linear model), but also potential variables that are relevant using *a priori* knowledge.
- After fitting the model, check if modelling assumptions are met.
- If modelling assumptions are not met, **try to find out why and adapt the model**.
  - Is a transformation of variables needed?
  - Wrong model family (e.g., linear regression is not suitable)?
  - Are terms missing in the model?
  - Outliers?
  - ...
- Interpret the model coefficients (effect sizes) and the p-values properly (see next week).

Occam's Razor principle

The principle essentially says that an **explanatory model** should not be made more complicated than necessary.

This is also known as the principle of **parsimony** (Prinzip der Sparsamkeit):

> Systematic effects should be included in a model **only** if there is convincing evidence for the need of them.

▸ See Wikipedia for "Ockham's Rasiermesser"

Final comments on the mercury example

Given that the mercury model is an **explanatory model**, we should not remove a variable (e.g., migration) simply because it reduces AIC (see slide 28).

Therefore, given the a priori selection of variables and the model validation results, the model from slide 12 was used in the final publication (Imo et al., 2017).

A predictive model: The bodyfat example

The bodyfat study is a typical example for a **predictive model**.

There are 12 potential predictors (plus the response). Let's fit the full model:

```
> r.bodyfat <- lm(bodyfat ~ ., d.bodyfat)
```

	Coefficent	95%-confidence interval	p-value
Intercept	-115.96	from -228.65 to -3.26	0.044
age	0.02	from -0.04 to 0.08	0.52
gewicht	-0.76	from -1.46 to -0.07	0.032
hoehe	0.58	from -0.04 to 1.21	0.068
bmi	2.48	from 0.26 to 4.70	0.029
neck	-0.60	from -1.04 to -0.16	0.008
chest	-0.14	from -0.37 to 0.08	0.20
abdomen	0.92	from 0.74 to 1.11	< 0.0001
hip	-0.31	from -0.61 to -0.01	0.046
thigh	0.25	from -0.05 to 0.55	0.11
knee	0.073	from -0.43 to 0.58	0.78
ankle	-0.49	from -1.17 to 0.19	0.15
biceps	0.17	from -0.16 to 0.49	0.32

Given the predictive nature of the model, we do some automatic model selection for this dataset, for instance using the `stepAIC()` function from the `MASS` package:

```
> library(MASS)
> # model with optimal AIC:
> r.AIC <- stepAIC(r.bodyfat, direction = c("backward"), trace = FALSE,AICc=TRUE)
> AICc(r.bodyfat)

[1] 1413.99

> AICc(r.AIC)

[1] 1408.469
```

$\rightarrow$ The AICc for the optimal model is 1408.5, compared to the full model with an AICc of 1414.0. This is not a very dramatic improvement, though...

But at least the model was now reduced, and only 8 of the 12 variables retain:

	Coefficent	95%-confidence interval	p-value
Intercept	-112.87	from -222.74 to -2.99	0.044
gewicht	-0.75	from -1.42 to -0.07	0.031
hoehe	0.53	from -0.09 to 1.15	0.091
bmi	2.17	from 0.01 to 4.33	0.049
neck	-0.54	from -0.97 to -0.11	0.014
abdomen	0.91	from 0.76 to 1.06	< 0.0001
hip	-0.29	from -0.58 to 0.00	0.05
thigh	0.30	from 0.04 to 0.56	0.023
ankle	-0.45	from -1.09 to 0.18	0.16

Note: AICc minimization may lead to a model that may retain variables with relatively large p-values (e.g., ankle).

Again: Such a procedure should not be applied for an explanatory model. Moreover, the coefficients should not be (over-)interpreted.

Summary

- Model selection is difficult and controversial.
- Different approaches for predictive or explanatory models.
- Automatic model selection procedures are inappropriate for explanatory models.
- *P*-values should not be used for model selection.
- AIC, AIC_c, BIC: balance between model fit and model complexity.

Reading for the self-study week

You can find all reading tasks in the first "Self study week" on OpenEdX. The following pdfs are provided there:

- Statistical significanc vs. biological importance (Interleaf from Whitlock and Schluter)
- Correlation vs. causation (Interleaf from Whitlock and Schluter)
- P-Werte in der NZZ (3. April 2016)
- S. Goodman (2016): Aligning statistical and scientific reasoning (a really thoughtful article in one of the world's leading scientific journals)

And **optionally**:

- Paper by Freedman (1983): "A Note on Screening Regression Equations"

References:

Brewer, M. J., A. Butler, and S. L. Cooksley (2016). The relative performance of AIC, AIC_c and *bic* in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution 7*, 679–692.

Clayton, D. and M. Hills (1993). *Statistical Models in Epidemiology*. Oxford: Oxford University Press.

Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 45*, 311–354.

Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician 37*, 152–155.

Imo, D., S. Muff, R. Schierl, K. Byber, C. Hitzke, M. Bopp, M. Maggi, S. Bose-O'Reilly, L. Held, and H. Dressel (2017). Risk assessment for children and mothers in a mercury-contaminated area using human-biomonitoring and individual soil measurements: A cross-sectional study. Technical report, University of Zurich.