Lecture 8: Model/variable selection

BIO144 Data Analysis in Biology

Stephanie Muff & Owen Petchey

University of Zurich

22 April, 2021

Overview

- Predictive vs explanatory models.
- Selection criteria: AIC, AIC_c, BIC.
- Automatic model selection and its caveats.
- Model selection bias.
- Collinearity of explanatory variables
- Occam's razor principle.

Course material covered today

The lecture material of today is partially 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 freedman1983: "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 immediately obvious which terms are relevant to include in a model.

Importantly, the approach to find a model **heavily depends on the aim** for which the model is built.

The following distinction is important:

- The aim is to predict future values of **y** from known regressors.
- The aim is to explain **y** using known regressors. Ultimately, the aim is to find causal relationships.

→ Even among statisticians there is no real consensus about how, if, or when to select a model:

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."

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...

Predictive and explanatory models

Before we continue to discuss model/variable selection, we need to be clear about the scope of the model:

- **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,..).
- **Explanatory models**: These are models that aim at understanding the (causal) relationship between explanatory variables and the response.
  Example: The mercury study aims to understand if Hg-concentrations in the soil (explanatory) influence the Hg-concentrations in humans (response).

→ 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 "objective" variable selection 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 variable selection strategies have no place**.

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

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 (without interactions):

	Coefficient	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

Model selection for predictive models

- Remember: R^2 is not suitable for model selection, because it *always* increases (improves) when a new variable is included.
- Ideally, the predictive ability of a model is tested by a cross-validation (CV) approach. ▸ Find a description of the CV idea here.
- CV can be a bit cumbersome, and sometimes would require additional coding.
- Approximations to CV: So-called information-criteria like AIC, AIC_c, BIC.
- The idea is that the "best" model is the one with the smallest value of the information criterion (where the criterion is selected in advance).

Information-criteria

Information-criteria for model selection were made popular by

The idea is to find a balance between

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

$\rightarrow$ Reward models with better model fit.

$\rightarrow$ 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 bodyfat example, we have 243 data points and 13 parameters (including the intercept β_0), thus $n/p = 143/13 \approx 19 < 40 \Rightarrow AIC_c$ should be used for model selection!

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

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**.

Model selection with AIC/AICc/BIC

Given m potential variables to be included in a model.

- In principle it is possible to minimize the AIC/AICc/BIC over all 2^m possible models. Simply fit all models and take the "best'' one (lowest AIC).
- This is cumbersome to to "by hand." Useful to rely on implemented procedures in R, which search for the model with minimal AIC/AICc/BIC.
- Backward selection: **Start with a large/full model.** In each step, **remove** the variable that leads to the smallest deterioration in model fit (smallest R^2). Do this until no further variables can be deleted without loss of fit.
- Forward selection: **Start with an empty model** In each step, **add** the predictor that leads to the largest improvement (largest R^2). Do this until no further improvement is possible.

"Best" predictive model for bodyfat

Given the predictive nature of the bodyfat model, we search for the model with minimal AICc, for instance using the `stepAIC()` function from the `MASS` package:

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(AICcmodavg)
r.AIC <- stepAIC(r.bodyfat, direction = c("both"),
                 trace = FALSE,AICc=TRUE)
AICc(r.bodyfat)
```

```
## [1] 1413.99
```

The model was reduced, and only 8 of the 12 variables retain:

	Coefficient	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 1: AICc minimization may lead to a model that retains variables with relatively large p-values (e.g., ankle).

Note 2: We could continue here and for example include interactions, transformations of variables etc.

Cautionary note about the "best" predictive model

It is tempting to look at the coefficients and try to interpret what you see, in the sense of "Increasing the weight by 1kg will cause a bodyfat reduction by -0.75 percentage points."

However, the coefficients of such an optimized "best" model should **not be interpreted** like that!

$\rightarrow$ Model selection may lead to biased parameter estimates, thus do not draw (biological, medical,..) conclusions from models that were optimized for prediction, for example by AIC/AICc/BIC minimization!

See, e.g., freedman1983, copas1983.

Your aim is explanation?

"Explanation" means that you will want to interpret the regression coefficients, 95% CIs and p-values. It is then often assumed that some sort of causality ($x \rightarrow y$) exists.

In such a situation, you should formulate a confirmatory model:

- **Start with a clear hypothesis**
- **Select your explanatory variables according to a priori knowledge.**
- Ideally formulate **only one** or a few model(s) **before you start analysing your data.**

Confirmatory models have a long tradition medicine. In fact, the main conclusions in a study are only allowed to be drawn from the main model (which needs to be specified even before data are collected):

> 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.

(chapters 27.1 and 27.2, clayton.hills1993)

Confirmatory vs exploratory

Any **additional analyses** that you potentially do with your data have the character of **exploratory models**.

→ Two sorts of **explanatory models/analyses**:

- Confirmatory:
  - Clear hypothesis and **a priori** selection of regressors for y.
  - **No variable selection!**
  - Allowed to interpret the results and draw quantitative conclusions.
- Explanatory:
  - Build whatever model you want, but the results should only be used to generate new hypotheses, a.k.a. "speculations."
  - Clearly label the results as "exploratory."

Interpretation of exploratory models?

Results from exploratory models can be used to generate new hypotheses, but it is then not allowed to draw causal conclusions from them, or to over-interpret effect-sizes.

→ In biological publications it is (unfortunately) still common practice that exploratory models, which were optimized with model selection criteria (like AIC), are used to draw conclusions as if the models were confirmatory.

→ We illustrate why this is a problem on the next slides.

Model selection bias

Aim of the example:

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

Procedure:

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

```
set.seed(123456)
data <- data.frame(matrix(rnorm(51*100),ncol=51))
names(data)[51] <- "Y"
```

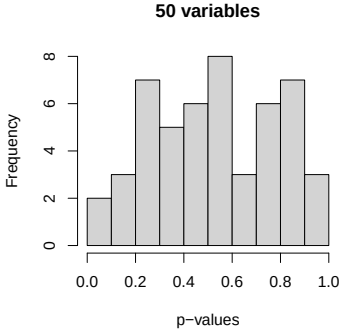
`data` is a 100×51 matrix, where the last column is the response. The **data were generated completely independently**, the explanatory variables do not have any explanatory power for the response!

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

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

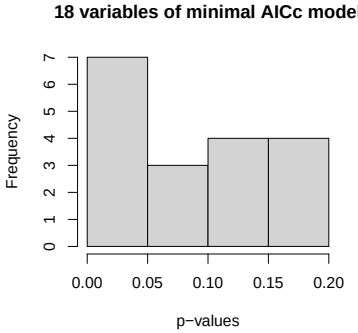
```
r.lm <- lm(Y~.,data)
```

As expected, the distribution of the p-values is (more or less) uniform between 0 and 1, with none below 0.05:



3. Then use AICc minimization to obtain the objectively "best" model:

```
r.AICmin <- stepAIC(r.lm, direction = c("both"),
                    trace = FALSE,AICc=TRUE)
```



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!

Main problem with model selection:

When model selection is carried out based on objective criteria, the effect sizes will to be too large and the uncertainty too small. So you end up being too sure about a too large effect.

$\rightarrow$ This is why such procedures should in general not be used when the aim is explanation.

Variable selection using p-values?

When you read publications, you might eventually see that people use p-values to do model selection. Also Stahel (Section 5.3) recommends such a procedure. However:

$\rightarrow$ Variable selection using p-values is an especially bad idea.

$\rightarrow$ Please NEVER do variable selection based on p-values$^{(\star)}$.

What is the problem?

$^{(\star)}$Even not when the aim is prediction.

Importance is not 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 of week 6, slide 28).
* The respective tests depend crucially on the correctness of the normality assumption.
* Explanatory variables 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.

> 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.

And we disagree with *p*-values based model selection suggested in Section 5.3 because

- It will also lead to model selection bias **freedman1983**.
- *P*-values are less suitable for model selection than AIC/AICc/BIC for the reasons mentioned on the previous slide.

An explanatory model: 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 $k=3$ categories, $k-1=2$ parameters 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.
  $\rightarrow$ See https://en.wikipedia.org/wiki/Overfitting.

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

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	$\sqrt{\cdot}$
age	Age of participant	continuous	
fish	Number of fish meals/month	count	$\sqrt{\cdot}$
last_fish	Fish eaten in last 3 days?	binary	
mother	Mother or child?	binary	
mother:age	Interaction term	binary:continuous	

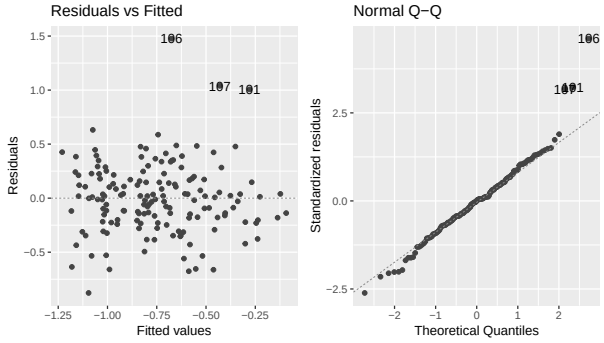
Let us now fit the full model (including all explanatory variables) in R:

	Coefficient	95%-confidence interval	p-value
Intercept	-0.68	from -0.90 to -0.46	< 0.0001
log10(Hg_soil)	0.03	from -0.06 to 0.11	0.49
vegetables	0.058	from -0.05 to 0.17	0.31
migration	-0.013	from -0.19 to 0.16	0.88
smoking	0.35	from 0.12 to 0.58	0.003
sqrt(amalgam)	0.29	from 0.19 to 0.39	< 0.0001
age	-0.041	from -0.07 to -0.02	0.001
mother	-1.09	from -1.82 to -0.35	0.004
sqrt(fish)	0.07	from 0.01 to 0.13	0.02
last_fish	0.31	from 0.15 to 0.47	0.0002
age:mother	0.057	from 0.03 to 0.09	0.0004

- The p-value for mercury in soil, $\log_{10}(Hg_{soil})$, is rather high: $p = 0.49$.
- **Question:** What is the answer to the main research question?

http://www.klicker.uzh.ch/bkx

A model checking step (always needed, but we did it already in lecture 5):



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?
- Can we find **additional patterns** in the data?

$\rightarrow$ We can go on from here and analyse the model in an exploratory manner. Such analyses can be useful to generate new hypotheses.

It would be tempting to check if there would be models with lower AICc.

For example, we can fit models where certain terms are omitted. Let's start with a model where the interaction *mother* · *age* is removed (denoted as `r.lm0`). The AICc then increases clearly, confirming that the term is relevant:

```
AICc(r.lm0)
```

```
## [1] 135.0549
```

```
AICc(r.lm1)
```

```
## [1] 123.8577
```

On the other hand, the model where we omit the binary *migration* variable would give a reduced AICc:

```
r.lm0 <- lm(log10(Hg_urin) ~ log10(Hg_soil) + vegetables  + smoking +
              sqrt(amalgam) + age * mother + sqrt(fish) + last_fish,d.hg)
AICc(r.lm0)
```

```
## [1] 121.5335
```

But: Given that the mercury model is an **explanatory, confirmatory model**, we should not remove a variable (e.g., migration) simply because it reduces AICc.

→ Therefore, given the a priori selection of variables and the model validation results, the model from slide was used in the final publication CITATION:imo.etal2017.

→ Any further analyses with other models would need to be labelled as exploratory.

Another complication: Collinearity of explanatory variables

(See Stahel chapter 5.4)

Given a set of explanatory variables $x^{(1)}, x^{(2)}, x^{(3)}, ..., x^{(p)}$. If it is possible for one of the explanatory variables 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 explanatory variables is said to be collinear.

Examples:

- Three vectors in a 2D-plane are always collinear.
- An explanatory variable can be written as a linear combination of two others: $x^{(j)} = c_1 \cdot x^{(1)} + c_2 \cdot x^{(2)}$, 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 $x^{(1)}$ and $x^{(2)}$ have a high correlation.

What is the problem with collinearity?

A simple (and extreme) example to understand the point: Assume two explanatory variables are identical $x^{(1)} = x^{(2)}$. In the regression model

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

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 explanatory variable $x^{(j)}$ as

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

where R_j^2 is the R^2 of the regression of $x^{(j)}$ against all other explanatory variables (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/(1-0.5) = 2.
- $R_j^2 = 0.9 \rightarrow$ high collinearity $\rightarrow$ VIF=1/(1-0.9) = 10.

What to do against collineartiy

- Avoid it, e.g. in experiments.
- Consider to not include a 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."

Important note: We would not care much about collinearity in a predictive model. If collinearity was a problem, AIC/AICc/BIC would probably anyway select a subset where some collinearity is eliminated (because model complexity is balanced against model fit).

Recommended procedure for explanatory models I

Before you start:

- **Think about a suitable model**. This includes the model family (e.g., linear model), but also potential variables that are relevant using **a priori** knowledge.
- Declare a strategy what you do if *e.g.* modelling assumptions are not met or in the presence of collinearity.
  - What kind of variable transformations would you try, in which order, and why?
  - What model simplifications will be considered it it is not possible to fit the intended model?
  - How will you deal with outliers?
  - How will you treat missing values in the data?
  - How will you treat collinear explanatory variables?
  - ...

It is advisable to write your strategy down as a "protocol" before doing any analyses.

Recommended procedure for explanatory models II

Analyze the data following your "protocol":

- Fit the model and check if modelling assumptions are met.
- If modelling assumptions are not met, **adapt the model** as outlined in your protocol.
- Interpret the model coefficients (effect sizes) and the p-values properly (see next week).

After the analysis that was specified in the "protocol":

- Any additional analyses, which you did not specify in advance, are purely exploratory.

One more thing: 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 knowledge or convincing evidence for the need of them.

▸ See Wikipedia for "Ockham's Rasiermesser"

Summary * Model/variable selection is difficult and controversial.
* Different approaches for predictive or explanatory models.
* Discriminate explanatory models into confirmatory and exploratory.
* AIC, AIC_c, BIC: balance between model fit and model complexity. * Automatic model selection leads to biased parameter estimates and *p*-values. * Threfore, automatic model selection procedures are inappropriate for explanatory models. * *P*-values should not be used for model selection, even not for predictive models. ## 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. Burnham, K.P. and D.R. Anderson (2002). *Model selection and multimodel inference: a practical*