Kurs Bio144:
Datenanalyse in der Biologie

Stefanie Muff & Owen L. Petchey

Week 7: Model selection

6./7. April 2017

Overview (todo: check)

- Model selection and model checking.
- Automatic model selection and its caveats.
- Selection criteria: AIC, BIC, Cp
- Model selection bias.
- Explanation v.s. prediction
- Relative importance of individual terms.

Course material covered today

- Stahel script chapters ...
- Chapter 27.1 and 27.2 by Clayton and Hills "Choice and Interpretation of Models" (pdf provided)

Optional reading:

- Paper by Freedman (1983)
- Check the Ellis book proposed by Owen

Recap of Last week

- **ANCOVA**
- **Linear alebra**

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* is the **response**.
- *Hg concentration in the 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:

Indicator if vegetables from garden are eaten; migration background; smoking status; number of amalgam fillings; age; number of monthly fish meals; indicator if fish was eaten in the last 3 days; mother; 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

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

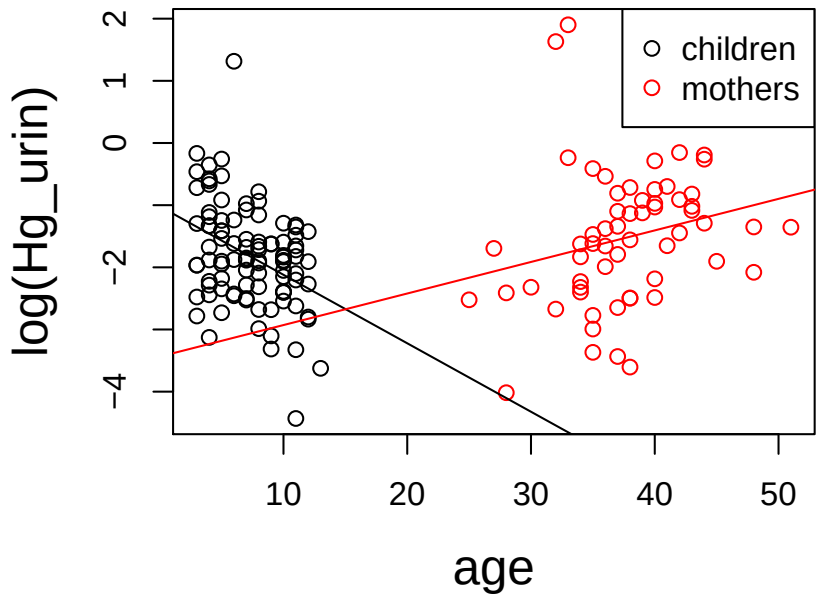
Variable	Meaning	type	transformation
Hg_urin	Hg conc. in urine (response)	continuous	log
Hg_soil	Hg conc. in the soil	continuous	log
vegetables	Eats vegetables from garden?	binary	
migration	Migration background	binary	
smoking	Smoking status	binary	
amalgam	No. of amalgam fillings	count	$\sqrt{.}$
age	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

- 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*.

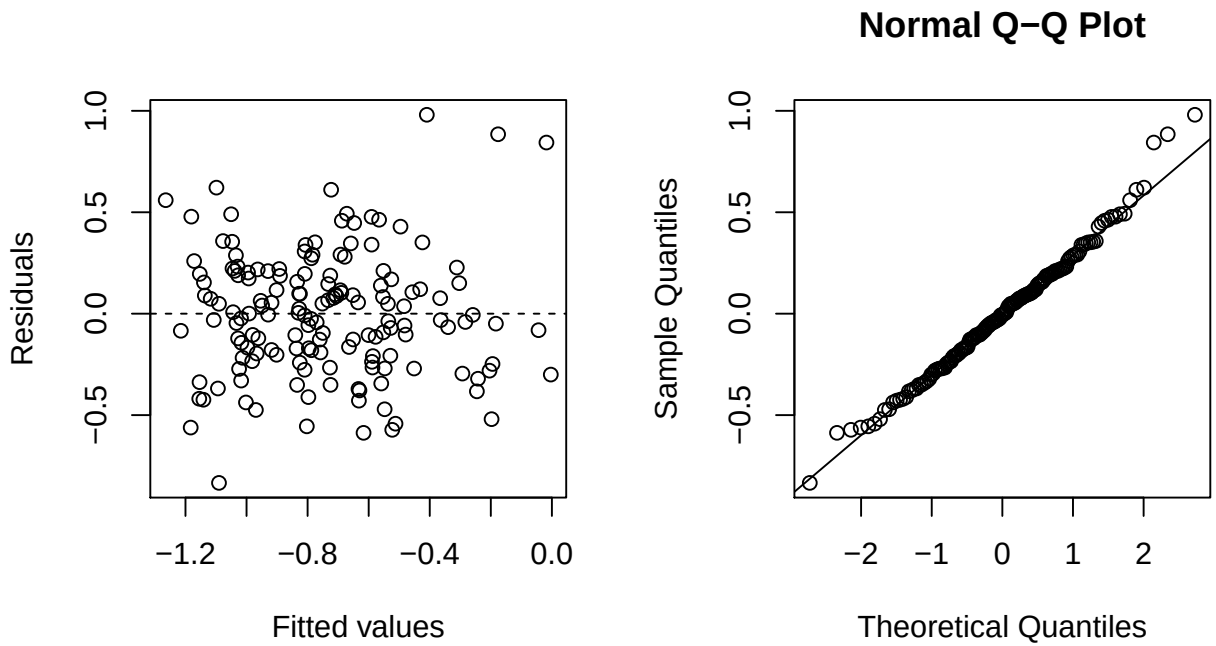
Fitting the model again with the additional term:

	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
mother	-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:mother	0.055	from 0.03 to 0.08	0.0002

- The p-value of the interaction is < 0.001, thus very small.
- 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 usually 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-test and check if it falls below a certain threshold (usually $p < 0.05$).

However, there are a few problems with this approach:

- When carrying out the t-tests with $H_0 : \beta_j = 0$ for all variables , one runs into a multiple testing problem.
  (Remember the ANOVA lecture, week 5, slide 25).
- 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.
- **A small p-value does not necessarily mean that a term is (biologically, medically) important – and vice versa!**

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.

Important note

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

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

Please note that **we 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 complexit**

$\rightarrow$ Penalize models with more parameters.

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)

Example of AIC use

Remember that we first fitted the mercury example **without** (`r.lm`) and **with** (`r.lm2`) the interaction *mother* · *age*. The model improvement is also confirmed by the **reduction in AIC**:

```
> AIC(r.lm)
[1] 107.4609
> AIC(r.lm2)
[1] 94.48769
```

Interpretation: The AIC 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)
> AIC(r.lm3)
```

```
[1] 92.70363
```

Interpretation: We observe a further reduction of AIC.

This success brings us to another idea:

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

- AIC for 'all subsets selection'
  For m variables there are 2^m possible models. Fit all models and take the "best" one (lowest AIC or BIC).
- Caveates and recommended practice

BIC, the brother of AIC

Definition and connection to AIC.

Model selection bias

Freedman (1983) example

The bodyfat example

Say that here we aim at prediction, not at explanation and that these are different situations in terms of model selection.

Summary

References:

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.