Lecture 8: Model/variable selection

BIO144 Data Analysis in Biology

Stephanie Muff & Owen Petchey

University of Zurich

05 November, 2020

Overview

- Predictive vs explanatory models.
- Selection criteria: AIC, AIC_c, BIC.
- Automatic model selection and its caveats.
- Model selection bias.
- Collinearity of covariates
- 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.

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

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 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).

→ 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 largest improvement (smallest AIC/AICc/BIC). Do this until no further improvement is possible.
- Forward selection: **Start with an empty model** In each step, **add** the predictor that leads to the largest improvement (smallest AIC/AICc/BIC). 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)
library(AICcmodavg)

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

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

```
AICc(r.AIC)
```

```
## [1] 1408.469
```

→ The AICc for the optimal model is 1, compared to the full model with an AICc of 2.

Note: Owen will also use `direction=c("forward")` and `direction=c("backward")` in the BC videos.