Processing and Analysis of Biological Data

Model selection

Øystein H. Opedal, Department of Biology, Lund University

2024-08-11

Model selection in confirmatory vs. exploratory analyses

As we have seen in many of the examples in previous chapters, statistical modelling is often used to estimate the parameters of a predefined model representing a theoretically expected relationship, or a biological hypothesis. In these cases, "model selection" is done by the researcher before performing the analysis, and the model structure is kept fixed whatever the parameter estimates and their uncertainty may be (they are in any case the results to be reported). Such analyses can be seen as "confirmatory", i.e. they are used to confirm the patterns in the data under the preselected statistical (and corresponding biological) model.

Confirmatory analyses can sometimes involve modification of model structure. A typical example is ANCOVA-type analyses, where one typically start from a full model allowing differences in both slopes and intercepts, and then test hypotheses about heterogeneity in slopes and intercepts by sequentially dropping first the interaction term and eventually the linear terms.

In more complex analyses with several to many candidate predictors, the model structure may not be well defined beforehand, and the analyses can be seen as "exploratory". Traditional statistical textbooks give detailed accounts of strategies for model selection of this kind, such as 'backward selection' with the aim of reducing the model to a 'minimum adequate model', where all terms are statistically significant (referred to as the principle of parsimony, Occam's razor etc.).

An intermediate strategy is to propose a set of competing models, one of which is often a null model representing the possibility of no effect of any predictor. The latter is represented by a model with only an intercept. To evaluate the support for the different models, we can use e.g. information criteria.

Information criteria

Due to the many problems with P-values, information criteria have emerged as an alternative approach for selecting among competing models. The philosophy behind these is to maximize the "information" carried by a model (or, strictly, minimizing the information lost), under the constraint of keeping the model as simple as possible. Thus, they are typically based on comparing the log likelihood of the alternative, nested models penalized by the number of parameters in each model. *Nested models* means that one model is a special case of the other, e.g. that a certain parameter of the more complex model is set to zero.

The most common information criterion in frequentist statistics is the AIC, the 'Akaike Information Criterion' defined as $AIC = -2ln(\hat{L}) + 2k$, where $ln(\hat{L})$ is the log likelihood of the model and k is the number of free parameters in the model. Recall that the likelihood represents the probability of the data given some parameters, and is what is maximized when we fit models with maximum likelihood. The lower the AIC value, the better the model.

Because the AIC value is directly based on the (log) likelihood, it is important that the candidate models are fitted to the same data. Thus, if we have missing values for some predictors, we should remove those observations completely before fitting the candidate models (using e.g. the `na.omit` function in `R`).

When we have relatively small data sets, it turns out that the standard AIC tends to select overly complex models. It is therefore recommended to introduce some extra penalty for extra parameters, yielding the finite-sample corrected AIC, $AICc$:

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

A rule of thumb is that the $AICc$ should be used when the ratio n/k is greater than 40, i.e. when we have less than 40 data points per estimated parameter.

The AIC values of several models are often summarized as differences, ΔAIC, from the highest ranked model. Another way to quantify the relative performance of a set of candidate model is to compute weights as

$w_i = \frac{exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^{R} exp(-\frac{1}{2}\Delta_r)}$

where the Δ values are the ΔAIC.

As an example, we simulate some data with two candidate predictors. Given a set of nested candidate models, we can build the AIC table as follows. Take some time to work through the code line by line, to make sure you understand what is happening at each step.

```
set.seed(12)
x1 = rnorm(200, 10, 3)
group = as.factor(sample(c("A", "B"), 200, replace=T))
y = 0.5*x1 + rnorm(200, 0, 4)
y[group=="A"] = y[group=="A"] + rnorm(length(y[group=="A"]), 2, 1)

m1 = lm(y ~ x1 * group)
m2 = lm(y ~ x1 + group)
m3 = lm(y ~ x1)
m4 = lm(y ~ group)
m5 = lm(y ~ 1)

mlist = list(m1, m2, m3, m4, m5)
AICTab = AIC(m1, m2, m3, m4, m5)
AICTab$logLik = unlist(lapply(mlist, logLik))
AICTab = AICTab[order(AICTab$AIC, decreasing=F),]
AICTab$delta = round(AICTab$AIC - min(AICTab$AIC), 2)
lh = exp(-0.5*AICTab$delta)
AICTab$w = round(lh/sum(lh), 2)
AICTab
```

```
##    df      AIC    logLik delta    w
## m2  4 1110.687 -551.3437  0.00 0.69
## m1  5 1112.549 -551.2745  1.86 0.27
## m3  3 1117.746 -555.8731  7.06 0.02
## m4  3 1119.242 -556.6211  8.55 0.01
## m5  2 1124.156 -560.0779 13.47 0.00
```

As expected, the likelihood is greater (less negative) for the more complex model ($m1$), but after penalizing for the number of parameters, $m2$ is the highest ranked and thus best supported. A traditional rule of thumb has been that 2 AIC units indicates strong support. A difference of 2 AIC units corresponds roughly to where a *likelihood ratio test* would be statistically significant (read P<0.05), so this practice in one way defeats the purpose of moving from significance testing to model comparison.

Data exercise: Model selection

Pick any of the datasets we have worked with in this course that includes more than one candidate predictor variable. Use AIC to perform model selection, produce a neat summary table with an informative legend, and interpret the results (biologically and statistically).

Model selection in linear mixed models

When we are working with mixed models, a couple of additional complications emerges in terms of model selection. First, model selection can in principle involve both fixed and random effects. Although random effects are often (or even usually) included because we specifically want to estimate the variance associated with it, or because of our knowledge of the data structure (e.g. avoiding pseudoreplication), there are cases where we want to compare models that differ in random-effect structure. The general recommendation in these cases is to perform model selection for the fixed and random effects separately, for example by first choosing random structure while keeping the fixed effects constant, and then choosing fixed effects while keeping the chosen random structure constant.

Second, when performing model selection (at least for fixed effects), we must compare models fitted with maximum likelihood (ML). This is the default in the `glmmTMB` package, but not e.g. in the `lme4` package. In `lme4` the default is to fit models with restricted maximum likelihood (REML), which is expected to give better (less biased) parameter estimates under some conditions. In either package, we can can choose the model-fitting method by changing the `REML` argument in the model call, with `REML=F` meaning that the model is fitted with maximum likelihood.

A general approach to fixed-effects AIC model selection in mixed models is thus to select the highest ranked model based on models fitted with ML, and then refit the highest ranked model with REML to obtain the parameter estimates.

For choosing random structures either ML or REML can be used (as long as the fixed effects are kept constant), with REML sometimes recommended.