Workshop 8: Model selection

Selecting among candidate models requires a **criterion** for evaluating and comparing models, and a **strategy** for searching the possibilities. In this workshop we will explore some of the tools available in R for model selection. If you are working from your own computer you may need to download and install the `leaps` package from the CRAN website to carry out all the exercises.

Scaling of basal metabolic rate in mammals

Savage et al. (2004, *Functional Ecology* 18: 257-282) used data to re-evaluate competing claims for the value of the allometric scaling parameter β relating whole-organism metabolic rate to body mass in endotherms:

$$\mathrm{BMR} = \alpha M^{\beta}$$

In this formula BMR is basal metabolic rate, M is body mass, and α is a constant. On a log scale this can be written as

$$\ln(\mathrm{BMR}) = \ln(\alpha) + \beta \ln(M)$$

where β is now a slope parameter of an ordinary linear regression - a linear model. Theory based on optimization of hydrodynamic flows through the circulation system predicts that the exponent should be $\beta = 3/4$, whereas we would expect $\beta = 2/3$ if metabolic rate scales with heat dissipation and therefore body surface area. These alternative scaling relationships represent distinct biophysical hypotheses. We will use them as candidate models and apply model selection procedures to compare their fits to data.

Savage et al. compiled data from 626 species of mammals. To simplify, and reduce possible effects of non-independence of species data points, they took the average of $\ln(BMR)$ among species in small intervals of $\ln(M)$. The resulting values of basal metabolic rate and mass can be downloaded here. Body mass is in grams, whereas basal metabolic rate is in watts.

1. Plot the data. Is the relationship between mass and metabolic rate linear on a log scale?

2. Fit a linear model to the log-transformed data (original data are not on the log scale). What is the estimate of slope?

3. Produce a 95% confidence interval for the estimate of slope. Does the interval include either of the candidate values for the scaling parameter β?

4. Add the best-fit regression line to the plot in (1).

5. Now compare the fits of the two candidate models to the data. To accomplish this you need to force a regression line having a specified slope through the (log-transformed) data. To fit a model with a particular slope, b, you do the following:

   ```
   z <- lm(y ~ 1 + offset(b*x), data=mydata)
   ```

6. Replot the data indicating the relationship between $\ln(M)$ and ln(BMR). Add to this plot the best-fit line having slope 3/4. Repeat this for the slope 2/3. By eye, which line appears to fit the data best?

7. Compare the residual sum of squares of the two models you fit in (5). Which has the smaller value? Do these values agree with your visual assessment of your plots in (6)?

8. Calculate the log-likelihood of each model fitted in (5). Which has the higher value? Hint: use the logLik function.

9. Calculate AIC for the two models, and the AIC difference. By this criterion, which model is best? How big is the AIC difference? Hint: Use the AIC function.

10. In general terms, what does AIC score attempt to measure?

11. Store your differences from the minimum AIC in a vector called delta. Using this vector, you can calculate the Akaike weights of the two models using the following:

    ```
    L <- exp(-0.5 * delta) # relative likelihoods of models
    w <- L/sum(L)          # Akaike weights
    ```

    Which has the higher weight of evidence in its favor? These weights would be used in Multimodel Inference (such as model averaging), which we won't go into in this course. The weights should sum to 1. They are sometimes interpreted as the posterior probability that the given model is the "best" model, assuming that the "best" model is one of the set of models being compared, but this interpretation makes assumptions that we won't go into right now.

12. Summarize the overall findings. Do both models have some support, according to standard criteria, or does one of the two models have essentially no support?

13. Why is it not possible to compare the two models using a conventional log-likelihood ratio test?

14. Optional: Both theories mentioned earlier predict that the relationship between basal metabolic rate and body mass will conform to a power law - in other words that the relationship between ln(BMR) and $\ln(M)$ will be linear. Is the relationship linear in mammals? Use AIC to compare the fit of a linear model fitted to the relationship between ln(BMR) and $\ln(M)$ with the fit of a quadratic regression of ln(BMR) on

$\ln(M)$ (a model in which both $\ln(M)$ and $(\ln(M))^2$ are included as terms). Don't force a slope of 2/3 or 3/4. Note that quadratic regression can be implemented using the following:

```
z <- lm(y ~ poly(x,2), data=mydata)
```

Plot both the linear and quadratic regression curves with the data. Which model has the most support? Which has the least? On the basis of this analysis, does the relationship between basal metabolic rate and body mass in mammals conform to a power law?

Bird abundance in forest fragments

In the current example we are going data dredging, unlike the previous example. There are no candidate models. Let's just try all possibilities and see what turns up. The data include a set of possible explanatory variables and we want to known which model, of all possible models, is the best. Sensibly, we wish to identify both the best model and those models that are near-best and should be kept under consideration (e.g., for use in planning, or subsequent multimodel inference).

The response variable is the abundance of forest birds in 56 forest fragment in southeastern Australia by Loyn (1987, cited in Quinn and Keough [2002] and analyzed in their Box 6.2). Abundance is measured as the number of birds encountered in a timed survey (units aren't explained). Six predictor variables were measured in each fragment:

- `area`: fragment area (ha)
- `dist`: distance to the nearest other fragment (km)
- `ldist`: distance to the nearest larger fragment (km)
- `graze`: grazing pressure (1 to 5, indicating light to heavy)
- `alt`: altitude (m)
- `yr.isol`: number of years since fragmentation.

The data can be downloaded here.

1. Using histograms, scatter plots, or the pairs command, explore the frequency distributions of the variables. Several of the variables are highly skewed, which will lead to outliers having excessive leverage. Transform the highly skewed variables to solve this problem. (I log-transformed `area`, `dist` and `ldist`. The results are not perfect.)

2. Use the `cor` command to estimate the correlation between pairs of explanatory variables. The results will be easier to read if you round to just a couple of decimals. Which are the most highly correlated variables?

3. Using the model selection tool dredge() in the MuMIn package, determine which linear model best predicts bird abundance (use AIC as the criterion). Ignore interactions. (You will need to install theMuMIn package if you haven't yet done so.) You'll need to set the na.action option before using the dredge command:

   ```
   options(na.action='na.fail')
   ```

   You'll also need to set the beta, evaluate, and rank arguments within the dredge function.

4. How many variables are included in the best model?

5. How many models in total have an AIC difference less than or equal to 7?

6. Calculate the Akaike weights of all the models retained. How much weight is given to the best model? Are there common features shared among the models having the highest weights?

7. How many models are in the "confidence set" whose cumulative weights reach 0.95?

8. Use a linear model to fit the "best" model to the data. Produce a summary of the results. Use visreg to visualize the relationship between bird abundance and each of the three variables in the "best" model one at a time. Which variable has the strongest relationship with bird abundance in this model?

Optional: Let's try analyzing the data using stepAIC(), which would also allow us to include interaction terms if we wished. Return to the data frame in which any variables requiring transformation have been replaced with the transformed variables.

1. Use stepAIC to find the "best" model (having no interaction terms). Review the results printed out on the screen.

2. Fit a linear model to the "best" model you found using stepAIC.

3. Inspect the results of the linear model fit. Use the drop1 command. Could step-wise regression used with null hypothesis significance testing have resulted in this model? How can you tell? Is it justifiable to keep terms in the "best" model that are not statistically significant according to basic significance testing?

4. Calculate AIC for the best model. Write this number down somewhere because we will compare it with another model fitted below. Note: the AIC value that you compute will differ from that printed out by stepAIC for this model. Not to worry: stepAIC uses the command extractAIC instead of AIC. The computations yield results that differ only by a constant, so AIC differences are unaffected as long as the same method is applied to models being compared.

5. Run stepAIC again, but this time use a model that includes all two-way interaction terms. This is already pushing the data to the limit, because there are only 56 data points. Scan the printed output on the screen to see the sequence of steps that stepAIC takes to find the best model.

6. Summarize the results of fitting a linear model to the best-fitting model from (5). Is the best model the same or different from the one picked out in (1) from the analysis of the additive model?

7. Calculate AIC for the best model analyzed in (6). How does it compare to the AIC value computed in (4) for the best additive model (the best model without interaction terms)? Considering the difference in AIC between the two models, which has more support? Do the two models have roughly equivalent support or does one have "essentially no support", as defined in lecture?