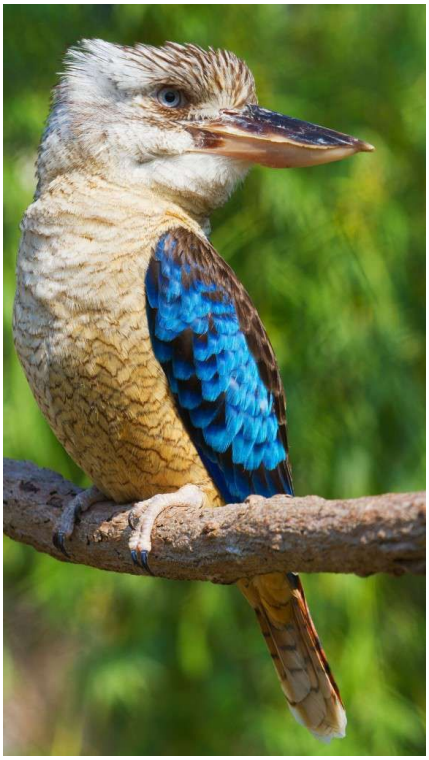
BIOL343 – Assignment #6

Model selection and apropos figures



Assigned Sunday 1 March 2020
Due Saturday 7 March 1159pm

For the last assignment, you were working with R.H. Loyn's data on bird abundance in Australian forest patches. You evaluated the distributions of the response variable and each predictor, analyzed univariate regressions and their assumptions, and put together a multiple regression with added-variable plots. In this assignment you'll use the same data to come up with the minimum adequate regression model (MAM) for bird abundance in Australian forest fragments.

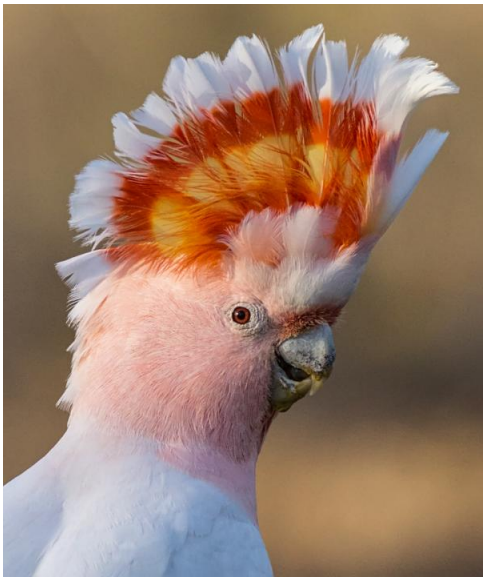
As usual, you will use **ggplot2** for graphs and **dplyr** functions whenever possible. To find the MAM, you will use the two different approaches to model selection that we discussed: stepwise selection using likelihood ratio tests and model dredging based on the corrected Akaike Information Criterion (AICc). The former uses functions in base R. The latter uses functions in the package, **MuMIn**. Dredging with MuMIn requires that you execute `options(na.action = "na.fail")` so that you are always comparing models built from the same data. In the Loyn dataset there are no missing values, but in a more realistic scenario you might have missing values of one predictor for some individuals and missing values of a second predictor for a different set of individuals, and so on. For sensible model comparison, you can only use complete cases.

Here's what Hana and I are expecting you do do.

(1) The data are in: BirdsInForestPatches.csv, and recall from the last assignment that you $\log_{10}$-transformed fragment area, distance to nearest forest patch and distance to largest forest patch. You'll those $\log_{10}$-transformed predictors again here.

(2) Perform backwards selection using the protocol we developed in lecture. Start with the full model and eliminate one predictor at a time chosen as the

predictor with the highest P value from an analysis using `car::Anova()`. Larger and smaller models should be compared using likelihood-ratio tests implemented with `anova()`.



(3) Once you've arrived at the MAM, see if you get the same model using forward selection implemented by the `add1()` function.

(4) Once you have the MAM, check assumptions using a quick 1-function graphical analysis. No figure caption required here.

(5) Determine the relative strength of the predictors in the model by rerunning it to calculate standardized partial regression coefficients.

(6) Make a formal graph that shows how well the data fit the MAM.

(7) Like backwards elimination, we begin model selection using AICc with a full model. Use the `dredge()` function to run all possible models and, for each, compute the log-likelihood, corrected AIC, ΔAIC, Akaike weight and r^2. Express the regression parameters as standardized partial regression coefficients.

(8) What predictors are included in the "top model" and does this set of predictors differ from the MAM that you arrived at using backwards and forwards selection?

(9) A common convention is to consider all models with ΔAIC ≤ 2 to be statistically indistinguishable. Let's consider how useful this criterion is in this case. Use **ggplot** to make a scatterplot with the rank of each model in terms of AICc on the x-axis (ranked from lowest AICc to highest) and ΔAIC on the y-axis. Plot a red horizontal line at ΔAIC = 2 to indicate how many other models you should consider along with the top model. This graph requires a figure caption, etc. In the text of your R notebook, interpret this graph. Do you see any obvious breakpoints in ΔAIC that separate models into likely vs. unlikely groups? Do any of these break points correspond to ΔAIC ≤ 2?

(10) Given the ΔAIC ≤ 2 criterion, how many other potential models should you consider along with the top model? Among this set of models, how frequently is each predictor included? Do the standardized partial regression coefficients of included predictors change much from model to model? Briefly describe the relevant trends.

(11) Calculate evidence ratios for all potential models and interpret these values along with Akaike weights. How likely is it that the top model is the best model?

(12) Run this top model and use **ggplot** to create “added-variable” plots that show the partial relation between the response and each predictor. If there is more than one predictor, arrange all the plots together using `plot_grid()`. This is a formal graph requiring a figure caption, etc.

(13) Make sure that all the formal graphs asked for above are beautifully rendered with complete and informative figure legends.

You should upload your file to the Assignment #6 OnQ dropbox by Saturday 7 March 2020 at 1159pm. As before, please submit a **PDF version** of your .html R notebook document called "StudentNumber_A6.pdf", where the file name starts with your student number.

Before you upload your assignment, please make sure that your PDF file is **complete** and as nicely formatted and organized as possible. Write concise code with no redundancy. Pay close attention to all the formatting guidelines and tips provided for the previous assignments.