**Solutions:**

ST2304 Exercises Week 11: Looing Basckwards and Ahead

**All answers to questions are written in red**

**Problem 1**

In 1898 Bumpus (or, probably, one of his technicians) collected some sparrows that had been blown out of their trees during a snow storm. The aim of this was to look at which survived (we will get to that part later). For now, we want to look at predictions of body mass using measurements of different aspects of body size.

**Figure 1.1**: Pairwise scatterplot between all the morphometric variables (Total.Length, Wingspread, Skull.Length, Humerus.Length, Femur.Length, Tarsus.Length, Sternum.Length, Skull.Width) in the house sparrow dataset.

We can see that almost everthying is positively correlated, to a greater or lesser extent. So first, try finding a model to explain weight using Sex and all of the morphometric measurements (Total.Length, Wingspread, Skull.Length, Humerus.Length, Femur.Length, Tarsus.Length, Sternum.Length, Skull.Width).

**Questions:**

1. Compare all of the subsets of models with these variables as main effects (see the last exercise for help), using AIC as your criterion. What is the best model?

Using an AIC based model approach the best model included the following variables; total length, skull length, humerus length, femur length, sternum length and skull width. Below is an example of summary table showing the best model output with the associated p values.

Only total length and humus length had significant p values although sternum length was close to signficiant.

The intercept (i.e. body mass at zero length and width) was -23.41 and total length was positively correlated with body mass (0.12) and with humerus length (20.12).

1. Compare all of the subsets of models with these variables as main effects, using BIC as your criterion. You can do this by giving bestglm(Xy=SomeData, IC="BIC").

The best model based on the BIC criterion was far simpler and only included total length and humerus length, below an the output of the summer from the best model based on BIC;

The use of BIC generally results in the selection of a simpler model due to the way it penalises the models with increasing numbers of parameters. Therefore, with an AIC approach you are more likely to overfit and with BIC you are more likely to underfit. You can see that when comparing the R squared for the BIC (43%) is not much lower than AIC (48%), even with many less parameters.

AIC = -2\*ln(likelihood) + 2\*k,

and

BIC = -2\*ln(likelihood) + ln(N)\*k,

where:

k = model degrees of freedom

N = number of observations

We can now look to see if the effects of any of these variables is different in the different sexes (i.e. if there is an interaction). We can write first order interactions of one variable with several like this: A\*(B + C). This can be expanded to A\*B + A\*C which becomes A + B + C + A:B + A:C.

1. Decide whether you want to use the best model from AIC or BIC, and explain briefly why you chose that model (there are arguments for both).

This is of course up to you, however in this case, where there are many variables that could potentially explaining the same variation in body mass. I would therefore avoid the risk of overfitting the model and go with a BIC approach and it has been argued that AIC is too liberal and can often prefer a more complex (and wrong model) over a simpler one.

Fit the model you chose with the interactions of these variables with Sex. Use anova() to compare the models with the interactions and without. use anova() on the model with the interactions to see if any should be included.

1. Would you include any interactions?

Using an ANOVA table comparing a model with the an interaction between sex and total length and humus length and one without any interactions, there was no significant improvement in the model fit (p > 0.01). Also the interactions themselves did not explain a significant amount of variance.



1. Can you suggest a better approach that doing this (bestglm can’t do the best approach, at least not properly)?

Rather than considering all parameters and all possible interactions and their combinations, as is done in bestglm() and such packages, a more reasonable approach could be to formulate a list of candidate models that you know are biologically relevant, based on your knowledge of the species.

1. Once you have decided on a final model, look at the fit of the model (residuals etc.). Is there any sign that the model is not linear? Are there outliers?

By looking at plots of the residuals, the model appears to fit relatively well, however a few potential outliers were identified and might want to be investigated. At higher fitted values of body mass residuals tended to be above the expected value, indicating that perhaps the data is not exactly linear.

**Problem 2**

1. Fit a model with just Total Length as a covariate, and survival as a response. How well does the model fit the data? How much of it does it explain, and then look at the residuals.

It is clear that the model does not fit well at all and explains approximately 6% of the variance in the data (R squared = 0.06).

1. Explain why the plot of the residuals against the fitted values (or agaisnt total length) looks like it does?

All the residual plots look strange since values can only be either 0 or 1 and certainely do not have a normal distribution as is assumed by the linear model. This is because we are trying to fit a linear model to categorical data, but we first need to convert this categorical variable into a normally distributed one, before we can analyse it with a linear regression and so in the next exercise we will learn about logistic regression models.