ST2304 Exercises Week 11: Looing Basckwards and Ahead

Problem 1: Trying to Remember what you did before Easter

The short version: we want to predict weights of English/house sparrows (*Passer domesticus*) from some morphological measurements, using data from 1898. the data are from here:

https://bioquest.org/numberscount/data-details/?product_id=31396

and a scan of the paper is here:

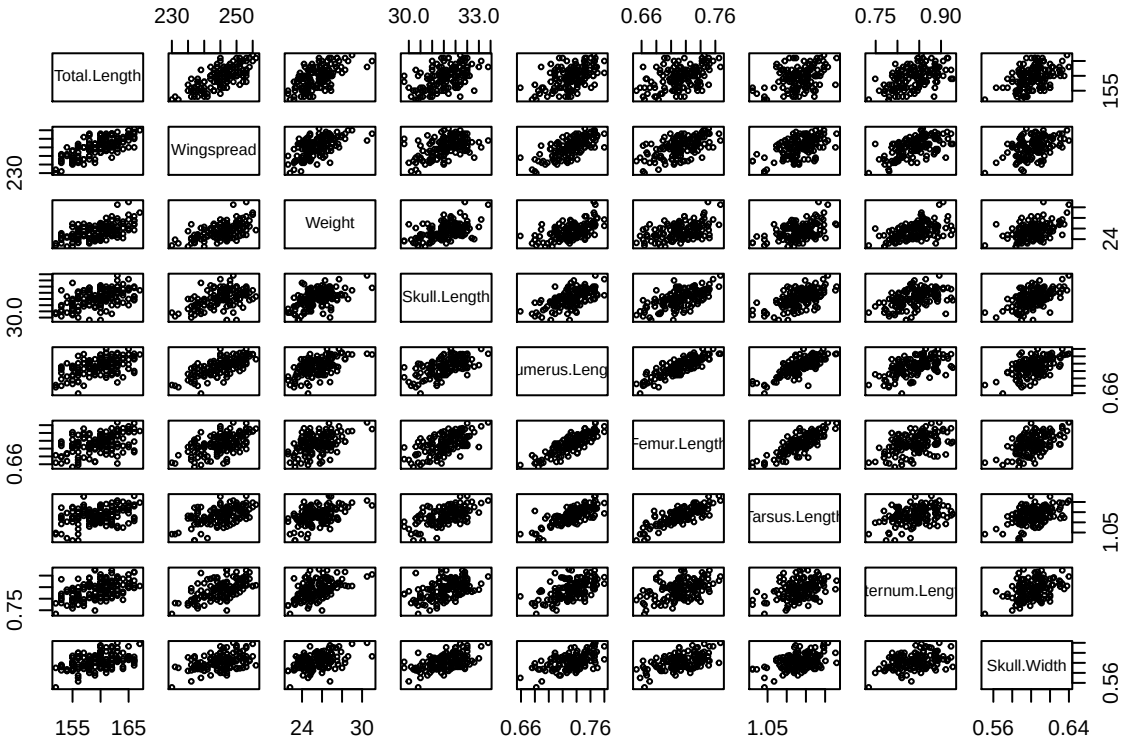
https://biodiversitylibrary.org/page/5071528

The long version: 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.

The data are on Blackboard, so can be read in like this, and then we can plot the data:

```
Bumpus <- read.csv("../Data/31396_Bumpus_English_Sparrow_Data.csv", na.strings = "Unknown")
# This line removes the units from the column names. It uses regular expressions
# Use ?plotmath to learn more if you want to disappear down that rabit hole
#  (anyone wanting to do bioinformatics should find themselves there at some point)
names(Bumpus) <- gsub("\\.+[[:alpha:]]*$", "", names(Bumpus))

pairs(Bumpus[,-(1:4)], cex=0.5)
```



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).

- 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?
- 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")`.

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`.

- 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).
- 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.
- Would you include any interactions?
- can you suggest a better approach that doing this (bestglm can't do the best approach, at least not properly)?

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?

Exam-type Questions

For a model, if $l(\theta|y)$ is the likelihood at the maximum likleihood estimates, $\hat{\theta}$, p is the number of parameters and n is the numbner of data points, then we can select models using AIC:

$$AIC = -2l(\hat{\theta}|y) + 2p$$

or BIC:

$$BIC = -2l(\hat{\theta}|y) + log(n)p$$

- if we are to compare different models with AIC, how would we know which model is best?
- for the Bumpus data, look at the summaries (these will be provided in the exam), the "best" model according to AIC and BIC are different. Why are they different?
- For the model selected by BIC, write out the model as a mathematical equation.
- one can argue endlessly about which model is best, nad this partly depends on what the model is used for. What factors and summaries of the model would you use to help decide which model is best?
- Look at the plots of the residuals against the fitted values, and the normal probability ølots (again, these will be provided in the exam). Are there any indications that the model is fitting poorly?

Problem 2: Finding out why what you did before Easter is not enough

The Bumpus data was collected and used to look at survival, and the effect that body size has on survival after a storm. Here we will fit a model using regression, and see what the model looks like (basically, horrible). Later we will fit a better model.

The `Status` variable shows whether the bird survived or not. We can convert that into a number (0 for died, 1 for survived) and analyse that using regression:

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