Biology 583 homework – learning to use R

If you haven't done so, install R and add the CAR and MASS libraries.

Remember you must hand in a set of R commands that executes all of the work in the questions below, and also hand in a copy of the output of these commands.

Download the dataset http://583.brianmcgill.org/birdsdiet.csv into a dataframe

What does this data seem to represent? Hint the command "names" could be useful.

There are two variables that are binary. Convert them to logical (True/False). There is one variable that is categorical. Is it ordered or unordered? Convert it to the appropriate variable type. Note you will need to consult a source on R for how to do this – we haven't discussed how to do this in class.

What is the mass of the 18^{th} family? What is the name of this family?

What is the skew of mass? What is the skew of log mass? Hint to look up how to calculate skew, you might want to use the help.search("skew") syntax to find the right function and the ?command syntax to see how it works.

What is the average AvgAbund of the families that eat insects? of the families that eat a mixture of plants and insects? of Passerines? Again help.search might help you find how to take an average of a vector of numbers.

Do a pairs plot to see how variables are related to each other.

Do a histogram of log(MaxAbund) using a method that gives stair-step histograms and one that gives smoothed histograms.

Do a boxplot of log(AvgAbund) versus Aquatic/non-aquatic. Now do a t-test is there a significant difference?

Can you find the command "box.cox.powers"? If not, connect to the library car (use "library(car)") and now try.

What is the optimal Box-Cox exponent for Mass (we will talk about what this means later).

Load data from the car library called Soils using "data(Soils)". What are the 25^{th}, 50^{th} and 75^{th} percentiles of Calcium levels in the data

Regression

Biological background

One of the big questions in ecology is to understand commonness and rarity of species. It has been understood for a long time that body size and diet have big effects on the abundance of a species. Big, meat-eating species are rare. Small, plant-eating species are common. However, there is enormous scatter in this data. I have recently done a study showing that these two variables are highly explanatory (r^2 goes up) if you do the analysis at a family level rather than a species level. You will analyze a dataset that explores this issue. The following variables are in the data:

Variable Name	**Description**	**Type**
Family	Common name of family	String
MaxAbund	The highest observed abundance at any site in North America	Continuous/numeric
AvgAbund	The average abundance across all sites where found in NA	Continuous/numeric
Mass	The body size in grams	Continuous/metric
Diet	Type of food consumed	Discrete – 5 levels (Plant; PlantInsect; Insect; InsectVert; Vertebrate)
Passerine	Is it a songbird/perching bird	Boolean (0/1)
Aquatic	Is it a bird that primarily lives in/on/next to the water	Boolean (0/1)

Getting started

Load a dataframe from http://583.brianmcill.org/birdsdiet.csv using read.table or read.csv. Look at the variables and match them to the table above.

You will load this data into a dataframe: R's way of holding statistical data. It is basically a fancy array that recognizes the row/column structure of statistical data and allows discrete as well as continuous data.

Try:

```
summary(birddietdata);
```

Can you tell from the last command what percentage of families are passerines and what percentage are aquatic? (NOTE: this is a trick question – think back to the coding – what does a 0 mean?). What is the average mass and average maximum and average abundances. How many in each diet category? Write a brief summary of this data for the start of your "Results" section.

Explore the basic relationships:

```
pairs(birddietdata) #note plot(birddietdata) does the same thing
```

Think about what relationships you see among variables but don't write this in the results as it is speculation.

Regression

Our main hypothesis is that abundance is a function of mass and diet. Lets analyze each of these independently. Do a regression of maximum abundance on mass:

```
m1<-lm(MaxAbund~Mass,data=birddietdata)
```

Run through the diagnostics (always do this before looking at p-values):

```
plot(m1)
```

What do the residual plots suggest? Are MaxAbund & Mass normal? Supplement the QQ plot with a histogram (check your class notes on how to do this). Do they need a transformation? Rerun the analysis with appropriate transformations (log10 gives a log-base-10 transformation). Do any other problems show up in the diagnostics (heteroscedasticity, non-independence, high leverage)? You can check out a high-leveraged point if you know the row # (shown in the plot of Cook's statistic for outliers) as:

```
birddietdata$Family[row]
```

Ignore the line about 53 levels – the part above tells you the family. What are the three most leveraged families? Be sure to include this in your results. Now you can look at the coefficients and p-values:

```
summary(m1)
```

You can also plot the regression

```
plot(log10(MaxAbund)~log10(Mass),data=birddietdata,
     xlab="log Mass (g)",ylab="Max Abund");
abline(reg=m1) #substitute the appropriate object for m1
```

Write up the p, r^2, coefficients, etc for this regression in your results.

Does the result improve if we analyze only terrestrial birds? The LM command has a subset capability. Lets get the results and put it into a new model object

```
m2<-lm(log10(MaxAbund)~log10(Mass),data=birddietdata,
    sub=!birddietdata$Aquatic)
```

Notice how we used ! (or "not") Aquatic to get terrestrial birds. Look at the r^2 for this subset. Is it higher? Write this up in the results. Do the same for just passerine birds (remember, think about the coding so you get passerines). Write this up in the results.