Section 1 - Multivariate regression

Biological background

This data set explores environmental variables that drive the abundance and presence/absence of a species. Here we look at one species, the Dickcissel. This is a grassland bird with peak abundances in Kansas. It was chosen because it has less noise than many species. The variables are:

abund – the number of individuals observed at each route
Present – presence/absence of the species
Land use variables (within 20 km radius of center of route, mixed=mixed forest)
broadleaf, conif, crop, grass, shrub, urban, wetland
NDVI – vegetation index – a measure of productivity
Climate (DD=degree days, FD=frost days, Tma=max temp, Tmi=min temp, P=precip) **clDD, clFD, clTma, clTmi, clP**

Exploration

Load the data (with read.table). It is in http://128.196.231.204/614/dickcissel.csv and is comma delimited with a header as usual. In the code below, I will just put a "?" where you should insert the actual name of your data object.

Do a histogram on the target variable (abund). Is a transformation needed? You will probably want to use this transformation over and over. So you can just replace the variable as follows:

```
?$abund=log10(?$abund+0.1)
```

Do a quick exploration of the data using the "pairs" command and the "cor" command. The cor command can only take numeric input, and this dataset has a categorical variable (Present). To remove this variable use subscripting:

```
cor(?[,-2])
```

This says to keep all rows (no subscripting before the comma) and to keep all columns except the 2nd. Identify two or three variables strongly (positive or negatively correlated with abundance) and include these in the results.

Scaled regression

Let's compare the relative importance of three (one climate, one productivity, one land cover). Run a multiple regression on abund~clTma+NDVI+grass. Now, we actually want to compare the strength of these three, so rerun the regression on scaled data:

```
m.scaled=lm(scale(abund)~scale(clTma)+ scale(NDVI)+
    scale(grass), data=data)
```

How do these results compare with the unscaled? Notice if the p-values for individual terms change. Does the F or p for the whole model change? Report the relative strength of effect of these three variables on abundance.

Polynomial regression

Now lets explore the possibility of a quadratic relationship between abundance and temperature (clDD). First plot the relationship:

```
plot(abund~clDD,data=?)
```

Does it appear curvilinear? Now lets run three nested models:

```
temp.linear=lm(abund~clDD,data=?)
temp.quad=lm(abund~clDD+I(clDD^2),data=?)
temp.cubic=lm(abund~clDD+I(clDD^2)+I(clDD^3),data=?)
```

Run a summary on the cubic – does the cubic term look significant. Find out by using the ANOVA command to compare these nested models:

```
anova(temp.linear,temp.quad,temp.cubic)
```

Which nested model should we accept? The anova command lists in increasing order of complexity. We should take the lowest line that has a significant p-value. Run a summary of this model. Report the regression formula, p and r^2 for this model.

Stepwise regression

Next we are going to do a stepwise regression. First run the model with everything except the Present variable:

```
full=lm(abund~.-Present, data=?) # note the period just
     after the tilde before the minus
```

Report the r^2 and p-value for the full model. Now run a stepwise regression

```
step=step(full)
```

Report which variables were dropped from the full model and what the new p-value and r^2 are. Did these change in the direction you expected? Mention why they changed in these directions briefly in the results.

Test for bouncing betas problem

We saw in the pairs anc cor commands that clDD, clFD and clTmn are all strongly correlated. Lets use just these variables to predict abundance. We might predict that since these are highly collinear, there will be a bouncing beta problem.

Run the regression once lm(abund~clDD+clFD+clTmn,…) Now run the regression on a random subset. First see how many rows there are in the dataframe:

```
dim(mydataframe)
```

This will tell you the number of rows, call it n. Now run the regression with 100 randomly chosen rows as:

```
summary(lm(sqrt(abund)~clDD+clFD+clTmn,subset=sample(1:n,10
0),…))
```

Rerun this 10 times using the up arrow to rerun. Comparing the coefficients do they change by a lot or a little? Do they change sign? Is there bouncing betas?

Borcard partition

Do a Borcard partition with all land-cover variables in one category and all climate variables in the other (leave out NDVI). What proportion of the variance in log abundance is explained by climate alone, by land cover alone, and by both combined? The formulas for this can be found in the lecture notes. To have R do the calculations for you, please note that the r2 is the 8th item reported in the summary. So if mod is the result returned by an lm command, then summary(mod)[[8]] returns the r2 (see Crawley pp 361-362).

Section 2 – ANOVA, ANCOVA

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://128.196.231.204/614/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.

ANOVA

OK, now lets run an ANOVA using diet. Run a linear model for MaxAbund as a function of diet. REMEMBER – you need to run a histogram or check residual plots to see if MaxAbund needs to be transformed! Also remember, ANOVA uses the lm command just like a regression. Any time you do an ANOVA you should verify the coding is sensible. To view the coding do:

```
model.matrix(m3) # or whatever your current model object is
```

What is the effects size of various levels of diet? (remember the summary command is a good way to get this). Note how R uses the diets in alphabetical order. It might be desirable to report effect sizes relative to diet=”Plant” as this is the baseline biologically. To do this issue:

```
birddietdata$Diet<-relevel(birddietdata$Diet,ref=”Plant”)
```

Note that this permanently changes your dataframe. Rerun the ANOVA (lm command). Write up this ANOVA in the results (note the p-value etc shouldn’t change but the treatment effects do).

Move on to a two-way ANOVA. Analyze how MaxAbund varies with both diet and aquatic/terrestrial. To get an interaction plot use:

```
interaction.plot(birddietdata$Diet,birddietdata$Aquatic,
     birddietdata$MaxAbund)
```

Note that the interaction.plot command doesn’t take formulas or the data= option. What a pain! What do the gaps in the line for the Aquatic group mean? Does the plot suggest an interaction? Find out using the lm command (remember a + separates dependent variables, a * gives interaction and individual terms – check your class notes). Look at the results of this analysis using the summary command on the model object. This gives significances for individual treatment levels. To get signifigance for whole terms, try:

```
anova(m6) #or whatever model object you have
#you can also do anova(m5,m6) to contrast nested models
```

Is the interaction term significant? Rerun without an interaction term. What is the treatment effect of diet and aquatic/terrestrial on abundance (when the interaction is omitted)? Hope you didn’t forget that abundance still needs a log transformation or figured this out when you ran through the diagnostic plots with plot(modelobject) which you do every time - right? Write this up in the results.

ANCOVA

Finally, lets do an ANCOVA with diet and mass. You should get the pattern of how to do this by now. Was diet significant? What is the treatment sizes of diet & mass. What is your r^2? How does this compare to the r^2 with just mass? You probably want to use the “ANOVA” command here also. What if you limit the analysis to terrestrial birds (using the “sub” option). What is this r^2? Write these results into your report. Did this analysis allow the slopes to vary between diets? (check out the coefficients using “summary”). To let the slopes vary by diet using interaction:

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

Use the ANOVA command on this result as well as summary, etc.

Unbalanced ANOVA and types of sums squares.

Now lets look at sums squares and the effect of order in unbalanced designs. The default anova command does not work. You need to use the “Anova” command (note upper case A) discussed in class that does Type II & Type III sum squares. It also has commands that give easy access to Cook’s distance (measure of influence) and other nice features. It is found in a package called ‘car’. Go ahead and download the car package (already downloaded on lab machines).

Let's see if type I & type III sum squares work as advertised. Reload the birdsdiet dataset. Recall that this is observational data (not experimental) and is unbalanced. In particular, the # of aquatic and non-aquatic families is unequal. And certainly the # of Aquatic plant eaters is not equal to the number of terrestrial plant eaters, etc. In short highly unbalanced. Use the table command to see how unbalanced it is.

Run the two way ANOVA with aquatic vs. diet with aquatic as the first factor (diet second), then with diet as the first factor. Save these both into model objects. Run anova on each model object and compare the output. Now run the Anova command (library car) with type III sum of squares on each model. Compare the output.

Section 3 – Path analysis

We will work with the Wilcox data set (you can read the brief paper given as an optional paper for the week on multivariate regression). He looked at islands in the Sea of California in an island biogeography context. We will limit ourselves to four variables here. Species richness, which we want to explain, area, elevation, and time since isolation. Area & elevation are strongly correlated (tall islands are larger). Moroever, area has a direct on richness (through # of individuals) and an indirect effect shared with elevation based on habitat complexity.

Step 1 – load the Wilcox data http://128.196.231.204/614/wilcox1978.csv

Step 2- do a pairs plot on the data, identify correlations

Step 3 – subset the data to just have the 4 variables of interest (e.g. d2<-d[,c("S","Area_km2",…. (you finish it)

Step 4 – calculate and store a correlation matrix (different from in class where we used a covariance matrix) – the cor command will do this

Step 5 – create a text file that species the following paths:
Elevation→Area (recall there is a path, a path name, and an NA for each path)
Area→Richness
Elevation→Richness
Time→Richness
four error terms (e.g. Elevation<->Elevation)

Step 6 – load the text file (m<-specify.model(file="") in package sem should help)

Step 7 – run the model, get a summary. (hint sem(mod,correlation,N=…) should help)

Step 8 – create an alternative model with a latent variable "Habitat", but drop the Elevation→area link, so:
Area→Habitat
Elevation→Habitat
Habitat→Richness
Time→Richness
four error terms (e.g. Elevation<->Elevation)

Step 9 – Run this model

Step 10 – compare the summaries of the two models visually. Which looks better? Does time or habitat have a stronger impact on richness.

Step 11 – formally compare the two models using the anova command

What to turn in

For this homework I want you to turn in the commands and output like last time. For the multivariate regression portion & the unbalanced sum of squares sections & the path analysis – just answer questions in the results box. For the ANOVA/ANCOVA part, please write a prose (i.e. like you would use in a journal) methods and results paragraph. This is to give you practice on turning numbers and computer code into prose, something you have to do in every paper you write.