What to turn in

For section 1a – please write a prose paragraph of methods and of results and put these into the appropriate boxes. For Sections 1b-4 just answer the questions in bold and place in the results section (no prose required). As always submit the code and output in the appropriate boxes.

#1a Blocking – Whether to treat blocking factors

We will work with a made up dataset on the effects of precipitation and grazing on species richness in the Southwest. A mixture of 30 species of annual plant seeds is well-mixed. And scattered in common gardens at four sites across the Southwest. The experimental design contains four levels of moisture and three levels of grazing (disturbance). A fully crossed, balanced, no replicate design is used. For reasons you don't have to worry about (it made it easier to generate the data!) the species richness is decimal in nature and you should analyze this using GLM instead of GLIM. You can load the dataset at http://128.196.231.204/614/rich.csv .

Field Cod

Explore this dataset and decide what linear model to use. Note: str(d) where d is any object in R (e.g. data.frame or model) is a quick way to explore something. Think about interaction terms, blocking, and nonlinear responses. You may want to try several models and see which one is best. Analyze the results.

For this section please write describe your process and findings by writing one paragraph suitable to go under the heading "Statistical methods" (and submit it in the methods box) and one paragraph suitable for results (and submitting it in the results box).

#1b Blocking – How to treat blocking factors (GLM, GLS, GEE, LMM, GLMM)

In the last section we saw that using a blocking factor can be beneficial.In this section we will use the same dataset as last time. However, we will now convert richness to an integer by rounding Use the function round to update the dataframe to make richness integer.

Now run the model "rich~moist+disturb" with the blocking factor added in appropriately in six separate statistical models. For the generalized models use Poisson regression, otherwise just use linear regression assuming that richness is far enough from zero that we can treat it as normal. For GLM/GLIM, the block will be treated as a fixed factor. For GLS/GEE place the block in the error structure. For LMM/GLMM treat the block as a random effect. The six models are: GLM (lm in R), GLIM (glm in R), GLS (gls in library nlme, use the corCompSymm to create the correlation structure), GEE (geeglm in geepack use corstr="exchangeable"), LMM (lme in nlme), and GLMM (glmer in lme4). **1) Obtain the AIC score for the five models that have AIC (one model doesn't – which one?) and determine the best fit. How do the effect sizes and significances vary between models (you can't get p-values for GLMM)?**

#2 - Phylogenetic data

Here we will use data that have phylogenies. Prepare by loading the "ape" library (you may need to download it first. Then type:

```
library(ape)
library(help=ape)
```

Now load the phylogeny:

```
tree=read.tree('http://128.196.231.204/614/primate.tree')
plot(tree) #display it
```

Now load a dataframe that has maximum lifespan and weight for the same primates. We have to use a new feature of dataframes where each row has a name:

```
d=read.csv('http://128.196.231.204/614/primatelh.csv',h=T,row.names=1)
```

Now look at d. What fields does it have. Do you see the rownames (just type "d<enter>").

Estimate the ancestral values for weight. Plot the tree with the PICs and tip values for weight indicated by circle size (see class notes or APE slides on course home page**) 2) Do the ancestral values make sense? Which node has the biggest estimate value?**

You may recall in the Garland and Ives paper PICs and GLS were claimed to be the same. Let us test this:

```
 w=pic(d$Weight,tree) #calculate Phylogenetic Independent
Contrasts on weight
 l=pic(d$MaxLife,tree)
 m.pic=lm(l~w-1) #note PICS go through zero so remove the
constant
 summary(m.pic)
```

Now try a GLS (hint: use the gls command from library nlme combined with a corBrownian correlation structure from library ape). **3) Which parts of the output of a PIC regression can we use? Slope, intercept, p-value? Which parts of this regression can we use? 4) How do the slopes estimated by these two methods compare? 3) Now what if we didn't do a phylogenetic method? Run a straight GLM using the lm command? how does it compare? Did you gain or lose power from PICs to standard regression?**

#3 - Type II regression

Lets do a model II regression. You will need to use the lmodel2 command in the lmodel2 library. Load our old friend, the birdsdiet2.csv file. We will regress log(MaxAbund) vs. log(Mass). Now run the Model II regression (look up the details in the help for the function lmodel2). Be sure to specify 1000 permutations so you get a p-value based on reshuffling. **5) How do the different slopes compare? Are the intercept and slope for SMA significantly different than for OLS?**

Now e are going to visually compare the 3 slopes. Type:

```
beta<-m$regression.results
beta
abline(beta[1,2],beta[1,3],lty=1)
abline(beta[2,2],beta[2,3],lty=2)
abline(beta[3,2],beta[3,3],lty=3)
legend(2,0,c("OLS","MA","SMA"),lty=c(1,2,3))
```

6) Visually which line seems like the best fit. Why?

#4a Mixed – ANOVAR/longitudinal/growth curve

Now we will look at an ANOVAR (repeated measure). Load the dataset sitka. To do this, type

```
library(MASS) #the Sitka data is in the MASS packag
library(nlme)  # & we will need the lme command in a minute anyway
```

```
data(Sitka) #load the dataset
names(Sitka) # for some reason the makers of R like commands with
lower case but data that starts Upper
summary(Sitka)
```

This dataset has four variables. One treatment was applied ozone/no-ozone. The logsize of the trees (height+2*diameter – a forestry measurement) was measured as the dependent variable. These measurements were repeated five times through the year and recorded in the variable "time" as # of days since Jan, 1 1988 (hint: treat Time as an ordered variable). The field "tree" gives the ID of the different trees this study was repeated on. The hypothesis is that the treatment and growth curve over time interact (hint: you need to look at an interaction effect). Develop the appropriate formulas for an lme command and run it. Run the anova command. **7) Report the significance of the treatment effect and the interaction of treatment on time (change of growth curve with treatment).** Just for comparison (don't report the results as it is statistically invalid) run this same analysis using the lm command (ignoring the repeated measure on tree). **8) How does this affect the significance of the interaction term? How did it effect the estimate of effect sizes (β)? What does this imply about the best model to use?**

#4b Mixed - Split plot

Finally we will work with a classic nested, split-plot design. Load the dataset Wheat (using the "data" command) from the library nlme (use the library command before the data command if you haven't already used it in a session). Wheat is grown in trays in a greenhouse. Each tray receives one of 4-moisture levels. Moisture levels cannot be done differently within a tray due to the watering mechanism. However, different wells within a tray can receive a different fertilizer level. At the end of growing, the plants are harvested and dry matter yield is weighed. Thus there are three replicate trays per each of four moisture levels and four fertilizer levels per tray. Load the data, identify the appropriate LME command and run it. Remember that tray is like another factor. You have to decide which things should be fixed or random, and for the random, specify the full nesting structure. Then run the summary and anova commands on the resulting object. **9) Report the significance and effect size of the treatment and interaction terms.**

#4c Mixed – Nested data and HLM

Now we will work with nested data. This dataset is related to the birdsdiet2.csv dataset we have used repeatedly, but it has data at the species level. Download http://128.196.231.204/614/birds.csv Each species has a total abundance (in North America), maximum observed abundance, number of routes observed on (proxy for range size), a body mass, and of course a genus, family and order. At the family level we have trophic level/diet. Instead of a log transform for abundance and number of routes, you will want to use the sqrt transform since there are some zeros (log should still work for mass, and no transformation for diet)

Start by finding out where the variation in the key variables resides. To do this use:

```
m<lme(sqrt(MaxAbund)~1,random=~1|Order/Family/Genus,data=?)
varcomp(m,scale=1) #scale makes add up to 100%
```

10) How much variance is at the species/genus/family/order level in max abundance? How about range size, body mass and diet? Are these labile or conserved traits? Where is the

variance for diet – careful – this is coded at the family level so don't include species or genus in the analysis (or you will get an error) and remember the lowest level is implied rather than explicit.

OK, now we will conduct an HLM. We want to regress MaxAbund vs. Mass. We want to explore the desirability of estimating a slope and/or an intercept on a family by family basis. A good first step is always to plot the data. Use xyplot (or coplot or groupedData) to plot MaxAbund vs. Mass for all families. **11) Do they look like there are different slopes in different families? Can you identify one family that looks clearly positive? clearly negative?** Using lm run a) a flat (species-level only) regression, then using lme run hierarchical models at the species/family levels for b) a fixed slope, variable intercept model, c) a variable slope, variable intercept model, and d) use model c) and include the level II effect of diet at the family level (no interaction of diet with mass). **12) Use model selection to pick the best model. What are the effect sizes and significances in this model?**