Modes of statistics & GLIM homework #3

This home work comes in three parts. The first deals with bootstrapping. The second with GLIM and logistic and Poisson regression (and zero inflated) . The final unit deals with survivorship analysis.

IMPORTANT

You will be doing a different way of handing in homework compared to the last two weeks. I will not ask you to write a methods section. LEAVE THE METHOD SECTION BLANK when you hand in the homework. Similarly, for the results, I will just ask for a list of specific numbers (p-values, etc) and occasionally a 1-2 sentence comment You do not need to write this into a prose section. Just hand in the question #'s and their answers. Everything you need to report is found below in bold with a number in front of it. So a highly hypothetical homework would look like (numbers made up):

no methods
results:

1) p=0.3. AIClinear=334, AICquadratic=330, delta=4
2) clP approximately 400
3) variables X & Z are significant
4) ….

#1 - Bootstrapping

We will use a dataset that is very close to the "Dickcissel" dataset we already have used. Instead of containing the 700+ sites where the Dickcissel is found, this data contains environmental variables for all sites where the BBS occurs (all of continental US and southern Canada). Load this datset from http://128.196.231.204/614/dickcissel.csv and briefly explore it.

A very commonly used measure of aridity/humidity is to take the ratio PET/AET (PET=Potential Evapotranspiration, AET=Actual Evapotransportation). While we don't have either PET or AET in this dataset we have clDD which is roughly proportional to PET and we have clP which is pretty close to AET. Thus a ratio of clDD/clP will give us a measure of aridity (high numbers mean much more could evaporate than does – i.e. dry or arid), which we will call AI (Aridity Index).

Calculate AI from the data. Explore the distribution of AI with hist and density as in HW #1. **1) Briefly describe in words the distribution of AI.** Calculate the mean of AI. Since this is a ratio, it does not immediately follow that the mean is just the mean of clDD/mean of clP. Although we have many datapoints and can calculate the mean of the ratio (rather than the ratio of the means) **(2) Report the mean of AI for our sample)**, our data is still a sample of a larger population. Better to use bootstrapping.. Use the boot command (library boot) to resample our data 1000 times and generate a mean and 95% confidence interval for the mean. **3) Report the bootstrap mean and 95% confidence interval.** Compare the 500th site (use the AI[500] notation to access this row) to the bootstrap mean. **4) Is the 500th site significantly different than the mean for the range of the Dickcissel region (i.e. all points)?**

Unit #2a – Logistic regression

This unit also uses the dickcissel dataset. In the past we did regressions using log10(abund) as the dependent variable. For logistic regression we will use the variable Present which indicates presence (abund>0) or absence (abund=0). Note that this is a habitat suitability model similar to the two paper by Zabel. We want Present to be coded as a 0 for absent and a 1 for present. Newer versions of R don't do this, so type the following:

```
d$Present=as.numeric(d$Present)-1
```

Lets look at how preciptation affects presence/absence

```
plot(Present~clP,data=?) # doesn’t give what you expect
plot(clP~Present,data=?) # little more useful
attach(?)# where ? is your dataframe – this means we won’t have to put
    ?$clP just clP
plot(clP,Present) # uses plot(x,y) format (could do
    plot(?$clP,?$Present) w/o attach command)
plot(clP+clP^2,Present) # looks like there might be some separation in
    quadratic
```

Now confirm the intuition from plots using the glm command. Create two model objects using the glm command. Have one using Present~clP and one with Present~clP+I(clP^2). When you use the glm command, don’t forget you have to ALWAYS SPECIFY THE FAMILY. What family should you use for logistic? Compare these two objects using the “anova” command. Is this significant? Do you remember what the distribution of this difference is? now try

```
    anova(m1,m2,test=”Chisq”) # m1,m2 are your two model objects
```

5) Report the p-value for a quadratic relation of precipitation vs. linear. Also report the two AIC values and the Δ. How would Burnham and Anderson describe the support for the linear model relative to the quadratic (look up Δ in the table given in the class notes)

Plot the linear case:

```
    plot(clP,Present)
    x=seq(0,200,0.5) #creates a vector of 0,0.5,1,1.5, ….
    lines(x,predict(m1,data.frame(clP=x),type=”response”))
    #m1 is your model object
```

Looking at the graph, **6) report at approximately what value of clP do we get a 50% probability of being present?**

Unit #2b – Poisson regression for log linear models

Start by loading the dataset http://128.196.231.204/614//birdsdiet2.csv . Let’s printout a contingency table of the count of cells in each category of Passerine/Aquatic/Diet:

Fiel

```
    attach(?)
    table(Diet,Suborder,Habit)#prints out contingency table
```

To do Poisson regression we need a column that has counts. To get it do this:

```
    d2=as.data.frame(table(Diet,Suborder,Habit))
```

Look at your new dataframe. What did we do? What is the variable/column with counts? Run a poisson regression using glm on your new dataframe. Don’t forget to specify a family! Analyze your results using the Anova command (library car). Alternatively, for GLIM, using anova(m,test=”Chisq”) will give the same result. **3) Report what significant interactions are there?**

Look for overdispersion: **3a) is the residual deviance close to the degrees of freedom?**. If it is much larger it is overdispersed. Run a model assuming overdispersion using the family=quasipoisson. Look at the output. **3b) What changed?**

Unit # 2c – Poisson (etc) regression on count data

Load the dataset “dickcissel.csv”. Look at the fields again (you’ve seen this before). **7) How many abundances are zero? What percentage of all records is this?** (hint: sum(d$abund==0) and length(d$abund). Note that the abundances are decimal (averaged over 5 years). Multiply abundance by 5 and store this back in the data frame for the remaining analyses in this section 2c.

Fit a Poisson regression of d$abund vs. clDD, clFD, clP, NDVI.

Check for overdispersion. **8) What is the deviance? How does this compare to the degrees of freedom?**

Now run a quasipoisson, a negative binomial (hint: glm.nb in library MASS), a zero-inflated negative binomial, and a hurdle negative binomial (hint library pscl for the last two – see class lecture notes). **9) what are the AICs for each model? Can you get AIC for every model? Which model appears the best?**

Unit #3 – Survivorship

We will briefly explore survivorship analysis using a dataset from Pollock et al 1989 (it is an optional paper on the website if you want to downloaded it). They radiocollared birds, and then tracked how long they lived. In some cases deaths were observed. In other cases the radio collar failed while the bird was still alive (this is censored data). Two possible explanatory variables are juvenile (1st year) vs. adult and condition (weight/wing length).). Load this datset from http://128.196.231.204/614/radio_surv.csv . Briefly explore the dataset and make sure you can find the four variables mentioned. All the functions you need are in the library "survival" so be sure to load this with a library statement now.

Fiel

The key step in using the library survival is to have the dependent variable on the left of the ~ be a Kaplan-Meier survivorship object by using Surv(time) for uncensored data and Surv(time,uncensored) for censored data (where the 2nd variable is a 1 if the "death" event was observed, and a 0 is if the observation stopped).

To get a quick plot do:

```
 m<-survfit(Surv(Days,Uncensored)~Adult,data=d)
plot(m,lty=c(1,3))
legend(10,0.3,c("Juv","Adult"),lty=c(1,3))
```

Note that in the plot command we have given a "lty=" command to specify two line types (dashed, solid) – see the help on the plot command for more details. We then add a legend at the x coordinate 10 & y-coordinate 0.3. **10) Do juveniles or adults die faster initially? Longer term (60 days) do there appear to be big differences in the rates?**

Now we want to fit a regression-type model exploring condition & age. We have two choices. To do a non-parametric model called Cox proportional hazards we use:

```
mnp<- coxph(Surv(Days,Uncensored)~Adult+Condition,data=d)
```

We can now look at the model object in a very familiar way. **11) Does age have a positive or negative effect on survivorship? How about condition? Are either of these statistically significant?**

To do a parametric model we specify a specific curve shape,then we know exactly how to interpret the coefficients as they are parameters in the curve. Specifically the default curve shape, which we will use is Weibull hazard. Other possibilities include things like exponential (constant hazard rates).

```
mex<-survreg((Surv(Days,Uncensored)~Adult+Condition),data=d)
```

12) How does this summary differ from the non-parametric (both in terms of what is or is not output and in terms of biological interpretation in this specific case)?