**Carter Adamson – Fish 558 Lab 13 HW**

*4/23/24*

*Guidelines:*

* *Include your Name, course, lab number, and date at the top of the document*
* *Number and label the questions and answers clearly! (We should easily be able to find your answers!)*
* *Include all of the requested output (e.g., values, data tables, and plots), not just the code for them. (We will not copy your code into R to see if it works).*
* *Include informative captions for figures and tables. See research articles for examples. [We will take points off this time if these are not included!].*
* *Submit a Word document unless directed otherwise (no r files or pdfs please).*
* *Include all your code used for the problems.*
* *Answer ALL questions using complete sentences that are clear and informative.*

**FOR 458 UNDERGRADUATES ONLY (26 pts) [grad student extra credit]**

1. In Lab we used statistical models to calculate standardized annual indices for Atlantic Croaker based on catch-per-unit-effort (CPUE) data from a trawl survey. Now, you will be using similar methodology to develop a standardized annual index for pinfish from Tampa Bay. The data can be found in “**pinfish1992-2006.csv**”. Standardized seine pulls were conducted from 1992-2006, and data were obtained on the number of pinfish caught (*number*), the *year*, vegetation (*veg*: none vs. seagrass), bottom type (*bot*: Sand vs. Mud), *depth*, temperature (*temp*), and salinity (*sal*). Make sure that you treat *year*, *veg*, and *bot* as factors. I recommend naming the data “AC1” to match the lab script (for ease of coding). Also, you can think of the *number* column as CPUE (where the unit of effort is one standard seine pull), and you can create a column named *CPUE* for easier coding (to match the lab script). All of this data setup can be accomplished with this code:

**AC1=read.csv("pinfish1992-2006.csv")**

**AC1$CPUE = AC1$number**

**AC1$year = as.factor(AC1$year)**

**AC1$veg = as.factor(AC1$veg)**

**AC1$bot = as.factor(AC1$bot)**

Use the dataset to do the following:

* 1. **Generate some exploratory plots to examine the data. For example, create a boxplot of log(CPUE+1) for each categorical variable, and a scatter plot of log(CPUE+1) for each continuous variable. What patterns do you see in the data that may be important to account for when developing an index using a statistical model? (4 pts)**

A graph of different sizes and numbers

Description automatically generated with medium confidence

Figure 1: Histograms of catch-per-unit effort (CPUE, top) and only positive CPUE values (bottom) for pinfish in Tampa Bay. In this case, CPUE refers to the raw number of fish pulled with each pull. Since we assume that pulls are standardized for effort (i.e. the same area is fished each time), fish counts can be interpreted directly as CPUE.

The first step in exploring these data was to generate histograms of the response variable(s) to get an idea of the distribution and shape of catch (assumed to be CPUE). Figure 1 shows histogram for all CPUE values above, as well as only positive values below. We can see that this dataset is dominated by zeros; the first bin contains more than 700 records with zeros included, but closer to 400 without zeros. We can clearly see that low catches are more common than high catches. Though zero-inflated models were not called for in this assignment, they may be appropriate to use here.

A close-up of several graphs

Description automatically generated

Figure 2: Box and scatter plots for the effects of various categorical (boxplots) and continuous (scatter plots) covariates on pinfish CPUE, log-transformed as log(CPUE+1).

The plots in Figure 2 use log(CPUE+1) as a response variable to compress outlying data and get an easier-to-visually-interpret look at the relationships between the covariates and CPUE. The boxes show the general distribution of log(CPUE+1) for the categorical variables of year, vegetation, and substrate type. All three of these covariates exhibit some variation, which is most obvious with vegetation (segrass appears to be associated with higher catch) and year (years like 1993 and 2006 have conspicuously lower medians and lower quartiles may be compressed on zero. Bottom substrate is less conclusive. Once again, August stands out as having a distinctly different distribution than the other months, with The scatter plots show the effects of increasing the values of the three continuous variables: depth, salinity, and temperature. There is a lot of noisy data here that makes it difficult to pick out a concrete pattern just from these simple visualizations. Modeling will help illuminate these effects. It’s worth noting, however, that all three of these variables have a large number of zero or near-zero values, which can be seen as a horizontal “line” of data points on the x-axis. Again, accounting for the extra zeros while modeling might help combat this pattern, but the directions for this assignment don’t call for modeling extra zeros so I didn’t do this.

* 1. Modify the lab script to calculate the nominal annual mean CPUE and generate standardized indices for annual pinfish trends using these statistical models:
     1. 0) nominal means (i.e., no standardization; use this for comparison)
     2. 1) generalized linear model (GLM) with a normal distribution,
     3. 2) a lognormal GLM (with bias-corrected estimates),
     4. 3) a gamma GLM
     5. 4) a poisson GLM
     6. 5) a negative binomial GLM

Include all available explanatory variables (year, veg, bot, depth, sal, temp) in each of models 1-5 (variables are not needed for the “nominal means”). For each of the 6 methods above, generate standardized annual indices with a 95% confidence interval (CI) for each year.

Some notes/tips for modifying the lab script to accomplish this task:

* To make predictions for your models, you will need to create a different “p.data” object than what we did in lab. This will have to be a data frame that includes all of the covariates in your models (year, veg, bot, depth, sal, temp). For categorical variables, you will need to pick one of the factor levels (e.g., veg=”Seagrass”), except for year which should have all factor levels represented (e.g., year=levels(AC1$year) ). For continuous quantitative variables you will need to pick a specific value like the mean (e.g., temp = mean(AC1$temp) ).
* Be sure to bias correct your lognormal GLM estimates; do NOT use the non-bias corrected code.

For the two count models (poisson and negative binomial (NB) GLM):

* To use the rootogram() function, you will need to install the “countreg” and “topmodels” packages, and that can be done using the following code:

install.packages("countreg", repos="http://R-Forge.R-project.org")

install.packages("topmodels", repos = "https://R-Forge.R-project.org")

* You should REMOVE the “offset(..)” term in your Poisson and NB model, because we are treating all of the seine hauls as identical (i.e., we are not accounting for different amounts of area swept).
* When generating predictions using the predict() function, your “newdata=” can refer to “p.data” because we aren’t using the offset of log.area. Your predict code would look like this:
  + out<-predict(mod4, newdata=p.data, type="response", se.fit=T)

**Include all of your code at the end of your homework to receive points for this question** (3 pts).

* 1. Create a multipanel plot with all 6 indices. Include a title on each panel to easily distinguish the models, and include the 95% CI. (4 pts)

A group of graphs showing the number of glm

Description automatically generated

Figure 3: Time series plot of pinfish indices for Tampa Bay from 1992 to 2006. Panels represent the time series for indices calculated in various ways: catch nominal means, or by a generalized linear models using five different distributions. Dotted vertical lines represent the 95% confidence interval.

* 1. Create a single plot with all indices overlaid, but standardize each index to its overall mean such that each index is centered on 1. How do the different indices compare? (4 pts)

A graph with colorful lines

Description automatically generated

Figure 4: Time series plot of standardized pinfish indices for Tampa Bay from 1992 to 2006. Different colors correspond to various methods of calculating the indices.

Overall, the different indices generally show similar patterns. For example, all of them show peaks in 1998, 2001, and 2004. However, the magnitude of these peaks varies, most obviously in 2001, where the Negative Binomial GLM predicts a much higher index than the others. The Negative Binomial isn’t consistently on top, though. In the 2004 peak, it’s the lognormal GLM predicting the highest index.

* 1. Determine which model you think is most appropriate for generating the annual index. Justify your answer based on the type of data you are modeling, model diagnostics (e.g., residual plots, rootogram), and AIC. Rootograms are only used for the count models (poisson and negative binomial). Recall that AIC can only be used to compare models that use the same response variable, so in this case, AIC can only be used to compare the normal, poisson, and negative binomial GLMs. You should also address how much Percent Deviance is explained by each model. (3 pts)

I decided to limit my model selection to just the Poisson and Negative Binomial, because we are dealing with count data. Since we didn’t divide by seine area, the data is limited to integers, so the count models are most appropriate. To assess which of these models to use, I first generate rootograms.

A graph of a graph

Description automatically generated A graph of a function

Description automatically generated

Poisson

Negative Binomial

Figure 5: Rootograms for the poisson GLM (left) and the negative binomial GLM (right).

The rootogram looks slightly better for the negative binomial, since more of the bars hang close to zero. There isn’t a major difference, but this is some evidence in favor of the negative binomial being the best model.

Table 1: Summary of model comparison between the poisson and negative binomial GLMs.

|  |  |  |
| --- | --- | --- |
| **Model** | **AIC** | **% Deviance Explained** |
| Poisson | 103477.79 | 0.516 |
| Negative Binomial | 5878.73 | 0.353 |

Table 1 shows that AIC was orders of magnitude lower for the negative binomial model, indicating major support for that being the better model. Between that evidence and the better rootogram, I decided to choose the **negative binomial** model as my beset model, despite it explaining less deviance than the poisson model.

* 1. Using your best model, generate a multipanel plot to visualize the effect that each covariate or explanatory factor has on the pinfish catches (use the **termplot()** or the **visreg()** function). Describe what trends and patterns you see for the pinfish data (make sure to address all modeled variables). For example, what are the patterns by year? Are catches higher for a specific bottom type? Etc. (5 pts)

A group of graphs showing different types of numbers

Description automatically generated with medium confidence

Figure 6: Termplot diagrams for the effects of six covariates on an index of pinfish catch, using the negative binomial candidate model.

Figure 6 shows the relative effects of the various covariates on the pinfish index. The index fluctuates throughout the years, but no major pattern sticks out there. However, we can see clearly that pinfish are caught more in areas with seagrass and muddy bottoms. Pinfish catch increases with increasing depth and salinity, and decreases as temperatures increase.

* 1. In your own words, describe what is meant when we say we are calculating a “standardized CPUE”, and use this exercise for reference. For example, how did the annual index from your best model compare to the nominal index (i.e., annual means with no standardization)? What does this comparison tell you about the effects of using standardized CPUE vs. nominal CPUE? (3 pts)

Since the purpose of the index of abundance is to accurately reflect abundance trends, it’s important to address any effects that don’t have to do with the population’s actual abundance. The purpose of standardizing CPUE is to do just that. For example, the best model I used here was able to account for catch variation due to seagrass presence, variation due to substrate, due to depth, and so on. With that information taken into account, we can be confident that the index more accurately reflects abundance trends.

1. Answer these questions (+1 pt):
   1. How many hours did you spend on this assignment as a whole?

6.5 hours (+ more for the grad project component)

* 1. I continue to encourage you to work with classmates collaboratively. Did you work with anyone else or at least consult with someone? Who? How did you communicate?

Solo.

* 1. Were there any particular things you struggled with in this lab and how did you overcome them?

Choosing the “best” model was a little tricky because deviance explained was higher for a model with worse AIC and diagnostics. It was a matter of weighing the evidence and prioritizing.

**FOR 558 GRADUATE STUDENTS ONLY (5 pts)**

Work on your class projects this week. Include a short synopsis of what you worked on and where you are with the analysis. Address any obstacles or challenges you may be facing (analytical or conceptual). (5 pts)

What I worked on this week was getting the data into the right format for a capture history. I had been having issues due to the use of two different tagging system, but combining those into one ID field did the trick nicely. Assembling capture histories involved creating an entirely new data frame iteratively, which was time consuming to code (and takes some time to run!). At this point I am comfortable with the capture histories and I’m preparing to run the CJS analysis with RMark. The last thing I need to investigate before doing that is how to code the robust design of the survey, with secondary occasions (surveys) nested within primary occasions (seasons). I’ll poke around and try to work this challenge out, and my plan B in case I can’t make that work will be to do a simpler CJS analysis without accounting for the robust design.

# Appendix: **CODE**

library(here)

library(ggplot2)

library(gridExtra)

library(countreg)

library(topmodels)

library(tidyverse)

# 458 SECTION (extra credit) ####----

pinfish <- read.csv(here("Lab 13 - CPUE Standardization, model-based", "pinfish1992-2006.csv"))

pinfish$CPUE <- pinfish$number

pinfish$year <- as.factor(pinfish$year)

pinfish$veg <- as.factor(pinfish$veg)

pinfish$bot <- as.factor(pinfish$bot)

pinfish$logCPUE <- log(pinfish$CPUE+1)

pinfish$CPUE1 <- pinfish$CPUE+1

pinfish.pos <- subset(pinfish, CPUE>0)

## a - exploratory plots ####----

#CPUE histograms

par(mfrow=c(2,1))

hist(pinfish$CPUE, main="CPUE")

hist(pinfish.pos$CPUE, main ="CPUE>0")

#covariate effects on CPUE

p1 <- pinfish %>% ggplot(aes(x=year, y=logCPUE))+ggtitle("Year")+geom\_boxplot()+labs(y="log(CPUE+1)")

p2 <- pinfish %>% ggplot(aes(x=veg, y=logCPUE))+ggtitle("Vegetation")+geom\_boxplot()+labs(y="log(CPUE+1)")

p3 <- pinfish %>% ggplot(aes(x=bot, y=logCPUE))+ggtitle("Bottom Substrate")+geom\_boxplot()+labs(y="log(CPUE+1)")

p4 <- pinfish %>% ggplot(aes(x=depth, y=logCPUE))+ggtitle("Depth (m)")+geom\_point()+labs(y="log(CPUE+1)") #units not specified. Guessing m, since most frequen in 16 to 32 inch of water according to a source

p5 <- pinfish %>% ggplot(aes(x=sal, y=logCPUE))+ggtitle("Salinity (ppt)")+geom\_point()+labs(y="log(CPUE+1)")

p6 <- pinfish %>% ggplot(aes(x=temp, y=logCPUE))+ggtitle("Temperature (C)")+geom\_point()+labs(y="log(CPUE+1)")

grid.arrange(p1, p2, p3, p4, p5, p6) #combine all covariate effects onto one grid of plots

## b - mean CPUE & indices ####----

###0:nominal----

#summary stats by year

nominal = pinfish %>% group\_by(year) %>%

summarize(mean = mean(CPUE),

var = var(CPUE),

cv = sqrt(var)/mean, #coefficient of variation

n = length(CPUE),

se = sqrt(var/n),

LCI = mean-1.96\*se,

UCI = mean+1.96\*se)

###1: normal GLM----

mod.norm=glm(CPUE~year+veg+bot+depth+sal+temp, data=pinfish, family=gaussian(link='identity'))

#standardized predictions

pred.data<-data.frame(year=levels(pinfish$year),veg="Seagrass", bot="Sand",

depth=mean(pinfish$depth), sal=mean(pinfish$sal), temp=mean(pinfish$temp))

out.norm<-predict(mod.norm, newdata=pred.data, type="response",se.fit=T) #bc covariates are held constant, predictions are standardized yearly indices

#data frame with yearly predictions and summary stats

summ.norm<-as.data.frame(cbind(as.numeric(levels(pinfish$year)),out.norm$fit,out.norm$se.fit))

names(summ.norm)<-c("year","mean","SE")

summ.norm$CV<-summ.norm$SE/summ.norm$mean

summ.norm$LCI<-summ.norm[,2]-1.96\*summ.norm[,3] #Lower 95% Confidence Interval (LCI)

summ.norm$UCI<-summ.norm[,2]+1.96\*summ.norm[,3] #Upper 95% Confidence Interval (UCI)

###2: lognormal GLM----

mod.lognorm=glm(logCPUE~year+veg+bot+depth+sal+temp, data=pinfish,family=gaussian(link='identity'))

#import and run bias correction function, then add summary stats

lnorm.bias.cor=dget(here("Lab 13 - CPUE Standardization, model-based", "lnormBC.r"))

summ.cor=lnorm.bias.cor(mod.lognorm)

summ.lognorm <- cbind(as.numeric(levels(pinfish$year)), summ.cor)

names(summ.lognorm) <- c("year", "mean", "SE")

summ.lognorm$CV<-summ.cor[,2]/summ.cor[,1]

summ.lognorm$LCI<-summ.cor[,1]-1.96\*summ.cor[,2]

summ.lognorm$UCI<-summ.cor[,1]+1.96\*summ.cor[,2]

###3: gamma GLM----

mod.gamma=glm(CPUE1~year+veg+bot+depth+sal+temp, data=pinfish,family=Gamma(link='inverse'))

#predictions

out.gamma<-predict(mod.gamma, newdata=pred.data, type="response",se.fit=T)

#summary data frame

summ.gamma<-as.data.frame(cbind(as.numeric(levels(pinfish$year)),out.gamma$fit,out.gamma$se.fit))

names(summ.gamma)<-c("year","mean","SE")

summ.gamma$CV<-summ.gamma$SE/summ.gamma$mean

summ.gamma$LCI<-summ.gamma[,2]-1.96\*summ.gamma[,3]

summ.gamma$UCI<-summ.gamma[,2]+1.96\*summ.gamma[,3]

###4: poisson GLM----

mod.poisson=glm(CPUE~year+veg+bot+depth+sal+temp, data=pinfish, family=poisson)

#predictions

out.poisson<-predict(mod.poisson, newdata=pred.data, type="response", se.fit=T)

#summary data frame

summ.poisson <- as.data.frame(cbind(as.numeric(levels(pinfish$year)),out.poisson$fit, out.poisson$se.fit))

names(summ.poisson) <- c("year", "mean", "SE")

summ.poisson$CV<-summ.poisson$SE/summ.poisson$mean

summ.poisson$LCI<-summ.poisson[,2]-1.96\*summ.poisson[,3]

summ.poisson$UCI<-summ.poisson[,2]+1.96\*summ.poisson[,3]

#see section e for rootogram

###5: nb GLM----

mod.nb=glm.nb(CPUE~year+veg+bot+depth+sal+temp, data=pinfish, link="log")

#predictions

out.nb<-predict(mod.nb, newdata=pred.data, type="response", se.fit=T)

#summary data frame

summ.nb <- as.data.frame(cbind(as.numeric(levels(pinfish$year)),out.nb$fit, out.nb$se.fit))

names(summ.nb) <- c("year", "mean", "SE")

summ.nb$CV<-summ.nb$SE/summ.nb$mean

summ.nb$LCI<-summ.nb[,2]-1.96\*summ.nb[,3]

summ.nb$UCI<-summ.nb[,2]+1.96\*summ.nb[,3]

#see section e for

##c - multipanel plot ####----

plot.nom <- ggplot(data=nominal, aes(x=year, y=mean, group=1))+geom\_line()+

geom\_errorbar(aes(ymin=LCI, ymax=UCI, width=0.1), linetype=2)+

theme\_bw()+ggtitle("Nominal")

plot.norm <- ggplot(data=summ.norm, aes(x=year, y=mean, group=1))+geom\_line()+

geom\_errorbar(aes(ymin=LCI, ymax=UCI, width=0.1), linetype=2)+

theme\_bw()+ggtitle("Normal GLM")

plot.lognorm <- ggplot(data=summ.lognorm, aes(x=year, y=mean, group=1))+geom\_line()+

geom\_errorbar(aes(ymin=LCI, ymax=UCI, width=0.1), linetype=2)+

theme\_bw()+ggtitle("Lognormal GLM")

plot.gamma <- ggplot(data=summ.gamma, aes(x=year, y=mean, group=1))+geom\_line()+

geom\_errorbar(aes(ymin=LCI, ymax=UCI, width=0.1), linetype=2)+

theme\_bw()+ggtitle("Gamma GLM")

plot.poisson <- ggplot(data=summ.poisson, aes(x=year, y=mean, group=1))+geom\_line()+

geom\_errorbar(aes(ymin=LCI, ymax=UCI, width=0.1), linetype=2)+

theme\_bw()+ggtitle("Poisson GLM")

plot.nb <- ggplot(data=summ.nb, aes(x=year, y=mean, group=1))+geom\_line()+

geom\_errorbar(aes(ymin=LCI, ymax=UCI, width=0.1), linetype=2)+

theme\_bw()+ggtitle("NB GLM")

grid.arrange(plot.nom, plot.norm, plot.lognorm, plot.gamma, plot.poisson, plot.nb)

##d - overlaid plot ####----

#assemble data for easier plotting

#standardizing by dividing each index by its mean

plot.d.data <- as.data.frame(cbind(as.numeric(levels(pinfish$year)), nominal$mean/mean(nominal$mean),

summ.norm$mean/mean(summ.norm$mean),summ.lognorm$mean/mean(summ.lognorm$mean),

summ.gamma$mean/mean(summ.gamma$mean),summ.poisson$mean/mean(summ.poisson$mean),

summ.nb$mean/mean(summ.nb$mean)))

names(plot.d.data) <- c("year","nom", "norm", "lognorm","gamma","poisson", "nb")

palette <- c( "#FF0033", "#33CC33", "#FFCC00", "#66CCCC", "#FF99FF", "#9900FF")

plot.d.data %>% ggplot(aes(x=year, y=nom, color="Nominal Means"))+geom\_line()+

geom\_line(aes(y=norm, color="Normal GLM"))+

geom\_line(aes(y=lognorm, color="Lognormal GLM"))+

geom\_line(aes(y=gamma, color="Gamma GLM"))+

geom\_line(aes(y=poisson, color="Poisson GLM"))+

geom\_line(aes(y=nb, color="Negative Binomial GLM"))+

labs(x="Year", y="Pinfish Index", color="Index Type")+

theme\_bw()+scale\_color\_manual(values=palette)

##e - determine best model ####----

#diagnostic plots

#normal

par(mfrow=c(1,2))

plot(mod.norm, which=c(1:2))

#lognormal

plot(mod.lognorm, which=c(1,2)) #the straight line here represents CPUE = 0 (log 0 is 1)

#gamma (limited usefulness)

par(mfrow=c(1,1))

plot(mod.gamma, which=c(1))

#poisson: rootogram

breaks = seq(0, 300, by=5)

rootogram(mod.poisson, breaks=breaks)

#nb: rootogram

rootogram(mod.nb, breaks=breaks)

#AIC

AIC(mod.norm, mod.poisson, mod.nb)

#deviance explained

dev.norm = (mod.norm$null.deviance - mod.norm$deviance) / mod.norm$null.deviance

dev.lognorm = (mod.lognorm$null.deviance - mod.lognorm$deviance) / mod.lognorm$null.deviance

dev.gamma = (mod.gamma$null.deviance - mod.gamma$deviance) / mod.gamma$null.deviance

dev.poisson = (mod.poisson$null.deviance - mod.poisson$deviance) / mod.poisson$null.deviance

dev.nb = (mod.nb$null.deviance - mod.nb$deviance) / mod.nb$null.deviance

##f - plot covariate effects ####----

par(mfrow=c(3,2))

termplot(mod.nb, partial.resid=F, se=T, col.se="black", col.term="black")