NES8010 Data Analaysis and Modelling

General linear model example using fish count data

February 2021

This is an example of how to analyse count data using linear regression.

Count data need to be handled differently by glm than a continuous variable as we are dealing with whole numbers only (integers) and we cannot have negative numbers. The glm framework allows us to do this by selecting the model 'family' based on the expected error structures. For the case of count data we use the 'poisson' family.

# Dataset

*Western Indian Ocean damselfish.*

Coral reefs in the Indian ocean are suffering from multiple threats; climate change, polluntion and overfishing. To address this, in particular the latter, some areas have been closed to fishing. To investigate the effect of closures on the abundance of fish surveys were carried out at sites within closed areas and outwith. To determine that any differences were due to the presence/ absence of fish and not through observer bias (an individual surveyors ability to detect a species) these surveys were repeated by multiple observers.

The analysis focusses on the abundance of 7 of the small damselfish at the sites within and outsite areas protected from fishing.

We are testing the hypothesis that areas protected from fishing have a greater abundance of fish than areas that are fished. We will look at only one species in this example - allowing you to explore the relationships of the other fish by copying and amending the code given.

First, read in the data. Ensure you are either using RProj or have the correct filepath to where your data are stored using the Session> set working >directory> choose working directory from the Rstudio menus.

The code in the notes uses the here package and the data are in the Data directory and checks that the packages we need for the exercise are loaded to our workspace.

wants <- c("ggfortify", "here", "ggplot2")  
has <- wants %in% rownames(installed.packages())  
if(any(!has)) install.packages(wants[!has])  
  
  
library(here)

## here() starts at C:/Users/nacm4/OneDrive - Newcastle University/Documents/MSc/NES8010 Quantitative Skills/NES8010 Quantitative Skills

fish<- read.csv(here("Data", "fish.csv"))  
head(fish)

## Site Protection Observer Chrysiptera.unimaculata Dascyllus.trimaculatus  
## 1 1 Protected A 12 0  
## 2 1 Protected A 8 0  
## 3 1 Protected A 9 0  
## 4 1 Protected B 6 0  
## 5 1 Protected B 6 0  
## 6 1 Protected B 10 0  
## Plectroglyphydodon.lacrymatus Pomacentrus.caeruleus Pomacentrus.pavo  
## 1 128 0 0  
## 2 153 0 0  
## 3 138 0 0  
## 4 194 0 0  
## 5 182 0 0  
## 6 237 0 0  
## Pomacentrus.sulfureus Stegastes.fasciolatus  
## 1 0 4  
## 2 2 6  
## 3 1 5  
## 4 3 3  
## 5 1 3  
## 6 2 3

The data file has very long heading names - the full species name for each of the damselfish oberved. You can search www.fishbase.org to see what these species look like but for modelling in R long column names are a pain as we have to keep typing them out, so we will shorten them to a genus species code.

You can rename the columns of an object very easy by assigning a character vector (that is the length of the number of columns of the object you are renaming).

names(fish)<-c("site", "protection", "observer", "Chr.unim", "Das.trim", "Ple.lacr", "Pom.caer",   
 "Pom.pavo", "Pom.sulf", "Ste.fasc")

First of all, data exploration.

summary(fish)

## site protection observer Chr.unim   
## Min. :1.0 Length:90 Length:90 Min. : 2.00   
## 1st Qu.:2.0 Class :character Class :character 1st Qu.:16.00   
## Median :3.5 Mode :character Mode :character Median :29.50   
## Mean :3.5 Mean :31.54   
## 3rd Qu.:5.0 3rd Qu.:45.00   
## Max. :6.0 Max. :88.00   
## Das.trim Ple.lacr Pom.caer Pom.pavo   
## Min. : 0.000 Min. : 7.00 Min. : 0.000 Min. :0.00000   
## 1st Qu.: 0.000 1st Qu.: 17.00 1st Qu.: 1.000 1st Qu.:0.00000   
## Median : 1.000 Median : 32.50 Median : 3.000 Median :0.00000   
## Mean : 3.956 Mean : 45.97 Mean : 2.911 Mean :0.08889   
## 3rd Qu.: 6.750 3rd Qu.: 59.50 3rd Qu.: 4.000 3rd Qu.:0.00000   
## Max. :22.000 Max. :237.00 Max. :11.000 Max. :3.00000   
## Pom.sulf Ste.fasc   
## Min. :0.0000 Min. :0.000   
## 1st Qu.:0.0000 1st Qu.:0.000   
## Median :0.0000 Median :1.000   
## Mean :0.3556 Mean :1.478   
## 3rd Qu.:1.0000 3rd Qu.:3.000   
## Max. :3.0000 Max. :6.000

table (fish$observer, fish$protection)

##   
## Fished Protected  
## A 9 9  
## B 9 9  
## C 9 9  
## D 9 9  
## E 9 9

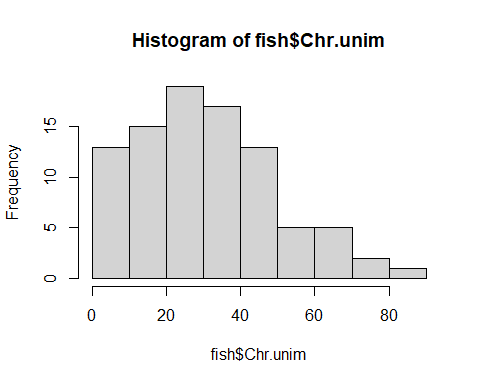
table (fish$observer, fish$site)

##   
## 1 2 3 4 5 6  
## A 3 3 3 3 3 3  
## B 3 3 3 3 3 3  
## C 3 3 3 3 3 3  
## D 3 3 3 3 3 3  
## E 3 3 3 3 3 3

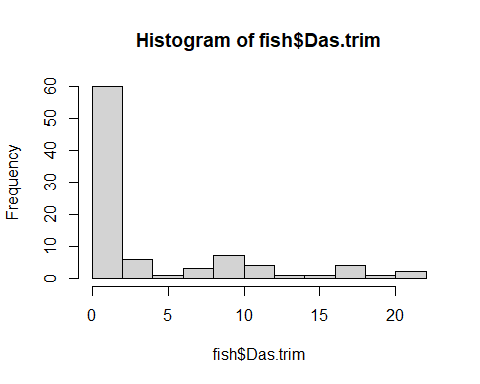
There are three sites (1-3) that are protected and three (4-6) that are fished. There are 5 observers and they all surveyed all sites 3 times.

As the data are counts, lets plot a frequency distribution of each species.

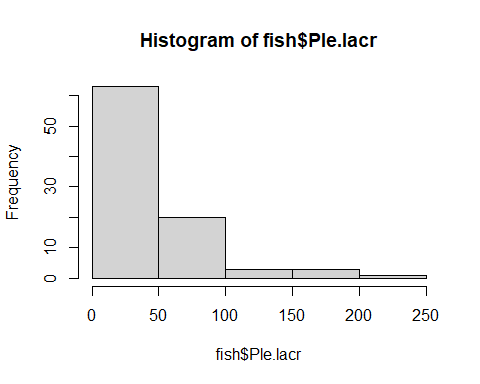
hist(fish$Chr.unim)



hist(fish$Das.trim)



hist(fish$Ple.lacr)



We can see (as we might expect with count data) that there is a skew to the distributions. We will focus on the first fish *Chr.unim* for the rest of the analysis

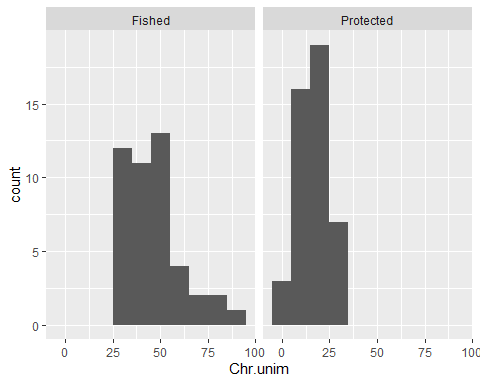
We can also tabulate the actual counts to see how they are dispersed.

table(fish$Chr.unim)

##   
## 2 3 4 6 8 9 10 11 12 13 14 16 18 20 22 23 24 27 29 30 32 33 34 36 37 38   
## 1 1 1 3 1 4 2 1 3 1 1 5 2 2 2 5 3 3 4 2 2 5 3 2 1 2   
## 39 41 45 46 47 48 49 50 51 52 55 58 61 67 69 78 88   
## 2 1 3 2 3 1 1 2 2 1 1 1 3 1 1 2 1

It is also useful to plot the distribution separately for the protected and fished site. We can do this easily with the hist command so we will use the graphics package *ggplot2* to make these histograms quickly and side by side.

library(ggplot2)  
m<- ggplot(fish, aes(x=Chr.unim))  
m <- m + geom\_histogram(binwidth = 10)  
m + facet\_grid( ~ protection)



What do you notice about the number of fish counted at the different sites?

You can repeat this code but substitute the species code for another species. If you are interested in this graphics package there are really useful help files with examples online <http://docs.ggplot2.org/current/> for every graph imaginable.

We can model this relationship (and test for an effect of site and protection) using a poisson glm model. Remember as the data are counts of fish the response variable has a particular structure (you can't count part of fish, so the numbers will all be positive integers) we expect the residual errors to be structured - which is why we use a poisson model structure.

M1.C.unim<- glm(Chr.unim~protection+as.factor(site), data=fish, family= "poisson")  
summary(M1.C.unim)

##   
## Call:  
## glm(formula = Chr.unim ~ protection + as.factor(site), family = "poisson",   
## data = fish)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.3775 -1.0149 -0.0099 0.7989 3.7042   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 4.05468 0.03400 119.252 < 2e-16 \*\*\*  
## protectionProtected -2.01780 0.09926 -20.329 < 2e-16 \*\*\*  
## as.factor(site)2 0.88269 0.11087 7.961 1.70e-15 \*\*\*  
## as.factor(site)3 1.14117 0.10711 10.654 < 2e-16 \*\*\*  
## as.factor(site)4 -0.46378 0.05472 -8.475 < 2e-16 \*\*\*  
## as.factor(site)5 -0.24506 0.05131 -4.776 1.79e-06 \*\*\*  
## as.factor(site)6 NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 1016.1 on 89 degrees of freedom  
## Residual deviance: 153.5 on 84 degrees of freedom  
## AIC: 622.17  
##   
## Number of Fisher Scoring iterations: 4

The model output tells use there is significant effect of protection and site of course these are confounded as the site is nested in protection, so let's only look at protection but add in the variable to test for an observer effect. We can then compare the inclusion of protection in the model versus that of the site variable.

M2.C.unim<- glm(Chr.unim~protection+observer, data=fish, family= "poisson")  
summary(M2.C.unim)

##   
## Call:  
## glm(formula = Chr.unim ~ protection + observer, family = "poisson",   
## data = fish)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.011 -1.700 -0.065 1.249 4.413   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.78429 0.04452 84.993 < 2e-16 \*\*\*  
## protectionProtected -1.01894 0.04251 -23.968 < 2e-16 \*\*\*  
## observerB -0.07112 0.06203 -1.147 0.25157   
## observerC 0.18263 0.05832 3.132 0.00174 \*\*   
## observerD -0.02252 0.06126 -0.368 0.71322   
## observerE 0.14643 0.05880 2.490 0.01277 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 1016.15 on 89 degrees of freedom  
## Residual deviance: 336.58 on 84 degrees of freedom  
## AIC: 805.25  
##   
## Number of Fisher Scoring iterations: 4

There is a large effect for protection (exp(1.63)= nearly 5 times as many fish oberved at site 5 than 1-3) and slight but significant differences in the number of fish observered by each observer.

M3.C.unim<- glm(Chr.unim~as.factor(site)+observer, data=fish, family= "poisson")  
summary(M3.C.unim)

##   
## Call:  
## glm(formula = Chr.unim ~ as.factor(site) + observer, family = "poisson",   
## data = fish)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.7523 -0.9308 -0.2188 0.8185 2.9252   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.98483 0.10099 19.654 < 2e-16 \*\*\*  
## as.factor(site)2 0.88269 0.11087 7.961 1.7e-15 \*\*\*  
## as.factor(site)3 1.14117 0.10711 10.654 < 2e-16 \*\*\*  
## as.factor(site)4 1.55402 0.10263 15.141 < 2e-16 \*\*\*  
## as.factor(site)5 1.77274 0.10086 17.576 < 2e-16 \*\*\*  
## as.factor(site)6 2.01780 0.09926 20.329 < 2e-16 \*\*\*  
## observerB -0.07112 0.06203 -1.147 0.25158   
## observerC 0.18263 0.05832 3.132 0.00174 \*\*   
## observerD -0.02252 0.06126 -0.368 0.71322   
## observerE 0.14643 0.05880 2.490 0.01277 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 1016.15 on 89 degrees of freedom  
## Residual deviance: 125.12 on 80 degrees of freedom  
## AIC: 601.79  
##   
## Number of Fisher Scoring iterations: 4

We still have a significant effect of site and also an effect of observer. Remember the levels of a factor are compared with the baseline (first level is the first alphabetically, here site 1 for the site and HB for observer).

Let's try a few more candidate models, first with ony one variable and then with an interaction of two variables (allowing for the fact that different observers may have significantly differnt counts of fish at specific sites)

M4.C.unim<- glm(Chr.unim~observer, data=fish, family= "poisson")  
summary(M4.C.unim)

##   
## Call:  
## glm(formula = Chr.unim ~ observer, family = "poisson", data = fish)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -7.5056 -2.7987 -0.2854 2.3503 7.3125   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.39934 0.04307 78.920 < 2e-16 \*\*\*  
## observerB -0.07112 0.06203 -1.147 0.25158   
## observerC 0.18263 0.05832 3.132 0.00174 \*\*   
## observerD -0.02252 0.06126 -0.368 0.71322   
## observerE 0.14643 0.05880 2.490 0.01277 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 1016.15 on 89 degrees of freedom  
## Residual deviance: 987.76 on 85 degrees of freedom  
## AIC: 1454.4  
##   
## Number of Fisher Scoring iterations: 5

M5.C.unim<- glm(Chr.unim~as.factor(site), data=fish, family= "poisson")  
summary(M5.C.unim)

##   
## Call:  
## glm(formula = Chr.unim ~ as.factor(site), family = "poisson",   
## data = fish)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.3775 -1.0149 -0.0099 0.7989 3.7042   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.03688 0.09325 21.843 < 2e-16 \*\*\*  
## as.factor(site)2 0.88269 0.11087 7.961 1.7e-15 \*\*\*  
## as.factor(site)3 1.14117 0.10711 10.654 < 2e-16 \*\*\*  
## as.factor(site)4 1.55402 0.10263 15.141 < 2e-16 \*\*\*  
## as.factor(site)5 1.77274 0.10086 17.576 < 2e-16 \*\*\*  
## as.factor(site)6 2.01780 0.09926 20.329 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 1016.1 on 89 degrees of freedom  
## Residual deviance: 153.5 on 84 degrees of freedom  
## AIC: 622.17  
##   
## Number of Fisher Scoring iterations: 4

M6.C.unim<- glm(Chr.unim~as.factor(site)\*observer, data=fish, family= "poisson")  
summary(M6.C.unim)

##   
## Call:  
## glm(formula = Chr.unim ~ as.factor(site) \* observer, family = "poisson",   
## data = fish)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.26950 -0.54586 -0.08245 0.55532 2.37941   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.26868 0.18570 12.217 < 2e-16 \*\*\*  
## as.factor(site)2 0.46135 0.23711 1.946 0.051692 .   
## as.factor(site)3 1.00215 0.21712 4.616 3.92e-06 \*\*\*  
## as.factor(site)4 1.31484 0.20915 6.287 3.24e-10 \*\*\*  
## as.factor(site)5 1.36890 0.20798 6.582 4.64e-11 \*\*\*  
## as.factor(site)6 1.72645 0.20154 8.566 < 2e-16 \*\*\*  
## observerB -0.27625 0.28273 -0.977 0.328527   
## observerC -1.17007 0.38157 -3.066 0.002166 \*\*   
## observerD -0.18924 0.27595 -0.686 0.492852   
## observerE 0.06669 0.25834 0.258 0.796290   
## as.factor(site)2:observerB 0.55849 0.34361 1.625 0.104091   
## as.factor(site)3:observerB -0.03275 0.33142 -0.099 0.921279   
## as.factor(site)4:observerB 0.44615 0.31146 1.432 0.152017   
## as.factor(site)5:observerB 0.29365 0.31198 0.941 0.346583   
## as.factor(site)6:observerB -0.05532 0.30760 -0.180 0.857269   
## as.factor(site)2:observerC 1.33041 0.43110 3.086 0.002028 \*\*   
## as.factor(site)3:observerC 1.11811 0.41423 2.699 0.006949 \*\*   
## as.factor(site)4:observerC 1.13233 0.40555 2.792 0.005237 \*\*   
## as.factor(site)5:observerC 1.56967 0.40031 3.921 8.81e-05 \*\*\*  
## as.factor(site)6:observerC 1.53591 0.39495 3.889 0.000101 \*\*\*  
## as.factor(site)2:observerD 0.34958 0.34118 1.025 0.305542   
## as.factor(site)3:observerD 0.03930 0.32174 0.122 0.902776   
## as.factor(site)4:observerD 0.01797 0.31048 0.058 0.953845   
## as.factor(site)5:observerD 0.43665 0.30294 1.441 0.149474   
## as.factor(site)6:observerD 0.06531 0.29871 0.219 0.826944   
## as.factor(site)2:observerE 0.24780 0.32304 0.767 0.443022   
## as.factor(site)3:observerE -0.05411 0.30315 -0.179 0.858329   
## as.factor(site)4:observerE -0.02143 0.29129 -0.074 0.941340   
## as.factor(site)5:observerE 0.07231 0.28835 0.251 0.801998   
## as.factor(site)6:observerE 0.16743 0.27880 0.601 0.548139   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 1016.147 on 89 degrees of freedom  
## Residual deviance: 55.849 on 60 degrees of freedom  
## AIC: 572.52  
##   
## Number of Fisher Scoring iterations: 4

AIC(M1.C.unim,M2.C.unim,M3.C.unim,M4.C.unim,M5.C.unim,M6.C.unim)

## df AIC  
## M1.C.unim 6 622.1736  
## M2.C.unim 6 805.2498  
## M3.C.unim 10 601.7903  
## M4.C.unim 5 1454.4387  
## M5.C.unim 6 622.1736  
## M6.C.unim 30 572.5232

The model with site and observer (and the interaction between the two) is the best fit (assessed by comparison of AIC values)

If we check the residual plots we find the model is a good fit.

#plot(M6.C.unim)

So we have found there is a significant difference in the abundance of damselfish found on the fished sites compared to the protected sites. The damselfish are more abundant at the fished sites. This is because there are no larger fish (they have been fished) to prey on them allowing the small fish to increase in number (a phenomenon known as 'prey release') at the protected site the damselfish numbers are kept in check by larger fish.

Is this the same for all the fish species? Does the same observer have trouble detecting the other species too?

How else might we structure an analysis of all the fish species at once?