How to deal with non-linear count data? Macro-invertebrates in wetlands

In this session we will recognize the advantages of making an effort to better identify the proper error distribution of data and choose the most informative type of model to assess our hypotheses. We use data collected by Bohlen *et al.* (2014) aimed at understanding the effect of hydrology on species abundance to evaluate government policies encouraging water retention. They used a stratified random sampling method to gather data on abundance of several organisms in wetlands within four ranches in Highlands and Okeechobee Counties in Florida, USA (Figure 1). Here, we focus on the abundance of macro-invertebrates. We do not use their whole wealth of data, and instead focus on a subset of relevant variables.



Figure 1. Sampling design within a wetland and example macro-invertebrates

Bohlen *et al.* (2014) proposed hypotheses on the shape of the responses of organisms to wetland water depth. In particular they expected that macro-invertebrate abundance may increase with water depth to a maximum and then decline (Figure 2). They also predicted that macro-invertebrate abundance may covariate with that of other organisms, and that this variation may vary among ranches because of management history and local attributes.



Figure 2. Hypothesis of change of macro-invertebrate abundance with water depth

Model validation: identification of the proper error distribution.

In this demo we evaluate an additive model including water depth, abundance of fish, and ranch as explanatory variables of macro-invertebrate abundance. The first variable is continuous, the second corresponds to counts and the third is categorical including four ranches. We start with a linear model with Gaussian errors; followed by a generalized linear model with negative binomial errors and then a general additive model with negative binomial errors. These models represent a sample of possible ways to assess the data. We emphasize validation of these models as a protocol to compare the quality and amount of information that can be obtained from them. You may wonder how we chose this model; please consider previous sessions with model selection approaches that addresses this question. Here we focus on the importance of proper error distribution. For simplicity, in this analysis we ignore that data was nested within wetlands by ranch.



Figure 3. Histogram of macro-invertebrate and fish abundances

The numbers of macro-invertebrate caught were extremely variable. Most frequently none were caught but once there were 18 captured (Figure 3). Fish were even more variable (Figure 3). For a moment, we ignore the distribution of the data and proceed to evaluate a linear model with Gaussian error. We will see that this is not a good decision but it is, sadly, a relatively common procedure among many ecologists.

We start by subsetting the data to remove missing values. We identify macro-invertebrates as the dependent variable y, and depth, fish and ranch as independent variables. We include the variable depth² to characterize the non-linear nature of the response as hypothesized. We transform fish abundance to its natural logarithm to reduce the leverage of extreme values. We change the variable that identifies the ranches from strings of names to strings of numbers to facilitate the reading of the output. We also create some plots to visualize the data (code in the R script).

```
dataforstats <- read.table("dataforstats_final_BB121613.txt", header=T)
sub1 <-subset(dataforstats,!is.na(upland_elev_m ))

y <- as.numeric(subl$macroct)
depth <- subl$depth
depth2 <- depth^2
fish <- log(subl$fishct)
rancho <- rep(1,length(subl$ranch))
rancho[subl$ranch=="bir"] <- 2
rancho[subl$ranch=="pal"] <- 3
rancho[subl$ranch=="wil"] <- 4
rancho <- factor(rancho)</pre>
```

Macro-invertebrate abundance did appear to be higher in intermediate depths, to co-vary with fish and be variable among ranches (Figure 4).



Figure 4. Observed macro-invertebrate abundance with water depth and fish abundance in four ranches in South Florida.

We use the function *1m* to obtain a linear model with Gaussian errors (default error distribution).

```
lm(formula = y \sim depth + depth2 + rancho + fish)
Residuals:
    Min
             10 Median
                             30
                                    Max
-3.7797 -1.7483 -0.9956
                         0.7213 16.4276
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.1577935 0.3775342
                                    3.067
                                           0.00225 **
                                           0.00529 **
depth
            0.0955440
                        0.0341426
                                    2.798
depth2
            -0.0017283
                        0.0006599
                                   -2.619
                                           0.00903 **
rancho2
            -0.4303603
                        0.2725263
                                   -1.579
                                           0.11479
rancho3
            -0.3762514
                        0.3586106
                                   -1.049
                                           0.29448
rancho4
            -0.8806389
                        0.3227926
                                   -2.728
                                           0.00654 **
                                    3.728 0.00021 ***
fish
             0.4980347
                        0.1336034
_ _ _
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
Residual standard error: 2.807 on 651 degrees of freedom
                               Adjusted R-squared: 0.03491
Multiple R-squared: 0.04372,
F-statistic: 4.961 on 6 and 651 DF, p-value: 5.592e-05
```

PCB 6468 - Methods in Experimental Ecology II Pedro F. Quintana-Ascencio, David G. Jenkins & Lina M. Sánchez-Clavijo



Figure 5. Macro-invertebrate abundance with **water depth** in four ranches in South Florida. The size of the dots relates to the amount of samples with a given combination of macro-invertebrates and depth. Samples with no fish are in blue, those with fish in black. The models were fit for fish abundance = 5. Linear model with Gaussian errors in green, generalized linear model with negative binomial errors in red and general additive model in black.



Figure 6. Macro-invertebrate abundance with **fish abundance** in four ranches in South Florida. The size of the dots relates to the amount of samples with a given combination of macro-invertebrates and fish. The models were fit for depth = 10. Linear model with Gaussian errors in green, generalized linear model with negative binomial errors in red and general additive model in black.

The shape of our linear model with Gaussian errors, represented with a green line in Figures 5-6 is consistent with the hypothesis of Bohlen et al. (2014). The output of the model allows the rejection of the null hypothesis of no evidence of an association with depth and that of no covariance with fish. It also indicates some differences among ranches. However, it explains less than 3% of the variance, and its residuals have a strong pattern suggesting increasing positive deviations from predicted values as abundance of macro-invertebrates increases (Figure 7). This model also indicates increasing number of macro-invertebrates with fish that appear not to be consistent with the data. We should not trust this model.

Now we recognize that nature of the data as counts and the nonlinear relationship with fish. We evaluate a generalized linear model with negative binomial errors and allow for non-linearity in the covariance with fish adding the variable fish². We call the function *glmmadmb* from the package glmmADMB. Notice that we do not include zero-inflation even though the nature of the data may require this procedure.

```
glmmadmb(formula = y \sim depth + depth2 + fish + fish2 + rancho, data = sub1, family =
"nbinom", zeroInflation = FALSE)
AIC: 2446.1
Coefficients:
         Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.028111 0.204410 0.14 0.8906
                           0.018151
depth 0.055513
depth2 -0.000949
                                           3.06
                                                    0.0022 **
                                          -2.73
              -0.000949 0.000348
0.770460 0.187250
                                                   0.0064 **
depth2
                                          4.11 3.9e-05 ***
diplini0.7704600.1872504.113.9e-05**fish-0.2075370.071678-2.900.0038**rancho2-0.1962970.139040-1.410.1580rancho3-0.0859090.182590-0.470.6380rancho4-0.4795170.168690-2.840.0045**
___
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
Number of observations: total=658
Negative binomial dispersion parameter: 0.66195 (std. err.: 0.059911)
Log-likelihood: -1214.04
```

The generalized linear model with negative binomial errors and quadratic response to fish abundance is more consistent with the data (Figures 6 and 7) and has better patterns in the residuals (Figure 7; less marked departures of the predicted values with increasing abundance of macro-invertebrates and fish). However, this departure from the assumptions still does not give strong support to this model.

Finally, we asses a general additive model allowing for more flexible association patterns for the response to depth and fish abundance. This approach allows evaluating nonlinear patterns using a piece by piece fitting across the range of the data. We call the function *gam* from the package mgcv.

PCB 6468 - Methods in Experimental Ecology II Pedro F. Quintana-Ascencio, David G. Jenkins & Lina M. Sánchez-Clavijo Spring 2016 04/06/2016

edf Ref.df Chi.sq p-value s(depth) 3.161 3.954 11.52 0.0207 * 2.817 3.361 25.31 2.68e-05 *** s(fish) _ _ _ `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1 0 Signif. codes: R-sq.(adj) = 0.0361Deviance explained = 6.64% UBRE score = 0.047136Scale est. = 1n = 658



Figure 7. Residuals for the linear model with Gaussian errors (left), generalized linear model with negative binomial errors (center) and general additive model (right).

The general additive model with negative binomial errors was consistent with the data and has almost removed any pattern in the residuals (Figure 7). This model has the smallest AIC value (Im (8 df) AIC = 3234.416; GLM (9 df) AIC = 2446.080; GAM (edf = 9.978163) AIC = 2443.411) and the best association between observed and predicted values (Figure 8). We concur with Bohlen *et al.* (2014) supporting a non-linear association among water depth and micro-invertebrate abundance which, in the studied wetlands, reached maximum abundances around 20 cm. We also found that fish and macro-invertebrates co-varied, and identified significant

variation among ranches (Figure 9). The causes of ecological patterns are diverse and rarely follow linear relationships. We can benefit from assessing models with more realistic and flexible assumptions allowing us to better describe these patterns.



Figure 8. Association among observed and predicted values for the linear model with Gaussian errors (left), generalized linear model with negative binomial errors (center)and general additive model(right).



Figure 9. Predicted number of macro-invertebrates with depth and fish (natural logarithmic transformed) after the GAM model with negative binomial errors, range: white= 4 organisms, black 0 organisms.

Patrick J. Bohlen, Elizabeth Boughton, John E. Fauth, David Jenkins, Greg Kiker, Pedro F. Quintana-Ascencio, Sanjay Shukla, and Hilary M. Swain. 2014. Assessing Trade-Offs among Ecosystem Services in a Payment-for-Water Services Program on Florida Ranchlands Final Report. USA Environmental Protection Agency.