*Using Generalized Linear Models to assess factors influencing oyster counts*

Generalized linear models (GLM) are a very robust “class” of statistical models that are basically extensions of simple linear regression model but allow the dependent variable (i.e., counts of oysters) to be non-normal. We used GLM models to make a preliminary assessment of factors including Year, Locality (i.e., Horseshoe, Lone Cabbage etc.), Site (Inshore, Nearshore, Offshore), and Treatment (Rock, No Rock) influence oyster counts. We assumed that oyster counts were likely to increase as more oyster bar area was assessed so we included effort (total transect length from the collapsed transects) as an offset. We also conducted an exploratory analyses of how Suwannee River discharge (USGS gauge 02323500), as a proxy for salinity, nutrients, and other factors, influenced counts on oyster reefs. We assessed how river discharge in year of sampling as well as a 1 year lag of river discharge influenced oyster counts.

We defined the analyses as part of three epochs similar to described elsewhere. Epoch 1 years = 2010-2013 and included all localities and sites. No restoration was done in Epoch 1. Epoch 2 years included all years (2010-2018) but only the Locality = Lone Cabbage and restoration effects were assessed from the “pilot project” reefs. The Epoch 3 analyses included all years 2010-2018 and all localities and sites, including restoration effects. We built a small subset of models that examined basic questions for each epoch. Epoch 1 focused on whether differences in oyster counts were apparent between localities and sites. Epoch 2 examined the influence of the pilot restoration project on Lone Cabbage reef, and Epoch 3 assessed over all years and treatments what factors had the largest influence on counts.

*Assessing distribution of data*

We chose to work with counts of oysters instead of oyster density to allow us to incorporate sampling effort as an offset in our model. To assess the distribution these data come from we considered the following: count data are discrete, the variance of counts exceeds the mean (see earlier tables), and we graphically examined fit of a negative binomial model a histogram of oyster counts from Epoch 3 (all data). The negative binomial appears to describe these count data and was assumed for all GLM models presented.

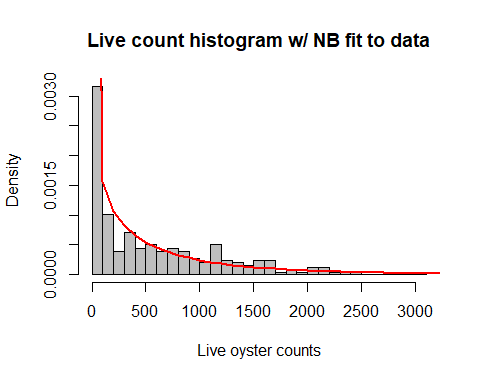


Figure. Histogram of density (y-axis) of live oysters counted (x-axis) from collapsed transects across all sites and years (Epoch 3 in the GLM discussion). The red line represents the predicted density of oyster counts if these data follow a negative binomial distribution. Visual assessment of model fit appears good.

*Candidate models*

We fit a series of single factor models using locality as a random effect and effort as an offset. Factors considered included Site, Season, mean annual daily discharge, and a 1 year time lag on mean and monthly annual discharge (i.e., discharge in the previous year or month influenced oyster counts in the present year). A limited set of models were fit of biological or project interest. We used a forward selection process where we fit each parameter individually and then retained statistically significant factors (p<0.1). Final model comparison was then made with AIC when appropriate.

*All data all years combined, no random effects*

The full dataset included all years, effort, locations, and treatment. We fit a negative binomial GLM to these data and examined the relationship between oyster counts and Year, Locality, Site, Treatment, 1 year lag on annual discharge, and log(effort) as an offset

> summary(full.mod.all)

Call:

glm.nb(formula = count\_live ~ year + locality + site + treatment +

ann\_dis\_lag\_sc + offset(log(tran\_length)) - 1, data = d,

init.theta = 0.8250328677, link = log)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.9843 -1.0918 -0.3747 0.2851 2.2437

Coefficients:

Estimate Std. Error z value Pr(>|z|)

year -0.03374 0.05609 -0.602 0.5475

localityCK 70.56687 112.82327 0.625 0.5317

localityCR 70.71074 112.82576 0.627 0.5308

localityHB 70.39406 112.82728 0.624 0.5327

localityLC 69.68484 112.87681 0.617 0.5370

localityLT 69.43443 113.26755 0.613 0.5399

siteN -1.52363 0.17748 -8.585 < 2e-16 \*\*\*

siteO -1.79954 0.18473 -9.742 < 2e-16 \*\*\*

treatmentrocks 0.78981 0.43616 1.811 0.0702 .

ann\_dis\_lag\_sc 2.65379 0.52772 5.029 4.94e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.825) family taken to be 1)

Null deviance: 8340.5 on 256 degrees of freedom

Residual deviance: 304.2 on 246 degrees of freedom

AIC: 3628.7

Number of Fisher Scoring iterations: 1

Theta: 0.8250

Std. Err.: 0.0654

2 x log-likelihood: -3606.7220

We found no significant effect between Year and Locality but did find significant effects by Site, Treatment, and a 1 year lag on annual discharge. Largest effects were by Site (inshore, nearshore, or offshore). So the largest effect on mean oyster counts per transect come from the site (inshore is the highest density, then nearshore, offshore) and river discharge the year before has a positive benefit on mean oyster counts. Model comparison using AIC between models with and without the annual discharge lag suggested improvement in model fit when information on discharge was included (delta AIC ~ 22).

*Epoch 1 only*

We fit a model to the Epic 1 data (all data prior to 2014) to assess the relationship between year, locality, siate, and 1 year lag on annual discharge on counts of live oysters. We used log(effort) as an offset.

> summary(full.mod.e1)

Call:

glm.nb(formula = count\_live ~ year + locality + site + ann\_dis\_lag\_sc +

offset(log(tran\_length)) - 1, data = d1, init.theta = 0.7804103593,

link = log)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.9368 -1.0873 -0.3953 0.2730 2.1187

Coefficients:

Estimate Std. Error z value Pr(>|z|)

year -0.3155 0.1581 -1.996 0.0459 \*

localityCK 637.9586 318.2812 2.004 0.0450 \*

localityCR 638.1299 318.3059 2.005 0.0450 \*

localityHB 637.8407 318.3063 2.004 0.0451 \*

localityLC 637.0968 318.3360 2.001 0.0454 \*

siteN -1.5615 0.1825 -8.556 <2e-16 \*\*\*

siteO -1.8014 0.1910 -9.430 <2e-16 \*\*\*

ann\_dis\_lag\_sc 0.9114 1.1383 0.801 0.4233

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.7804) family taken to be 1)

Null deviance: 7483.77 on 230 degrees of freedom

Residual deviance: 274.82 on 222 degrees of freedom

AIC: 3222.6

Number of Fisher Scoring iterations: 1

Theta: 0.7804

Std. Err.: 0.0650

2 x log-likelihood: -3204.5920

We found significant effects of locality and site but not lag of annual discharge in this model. This may be because contrast in river discharge was not high in these years.

We then fit the same model but treated site as a random effect. This model did not converge. Perhaps because there is a lot of information in the site variable because of the strong contrast. This is an area for further work.

Interestingly, if you drop year from this same model, then the model will converge

> summary(localityd1.2)

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [glmerMod]

Family: Negative Binomial(0.7645) ( log )

Formula: count\_live ~ locality + ann\_dis\_lag\_sc + (1 | site) + offset(log(tran\_length))

Data: d1

AIC BIC logLik deviance df.resid

3235.4 3259.5 -1610.7 3221.4 223

Scaled residuals:

Min 1Q Median 3Q Max

-0.8725 -0.6883 -0.3198 0.2672 4.1704

Random effects:

Groups Name Variance Std.Dev.

site (Intercept) 0.6169 0.7854

Number of obs: 230, groups: site, 3

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.6130 0.6039 2.671 0.007564 \*\*

localityCR 0.1534 0.2500 0.613 0.539549

localityHB -0.1592 0.2537 -0.627 0.530346

localityLC -0.8970 0.2510 -3.574 0.000352 \*\*\*

ann\_dis\_lag\_sc 2.6531 0.6464 4.105 4.05e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr) lcltCR lcltHB lcltLC

localityCR -0.229

localityHB -0.173 0.640

localityLC -0.205 0.686 0.665

ann\_ds\_lg\_s -0.569 -0.075 -0.157 -0.132

and Locality = Cedar Key and Locality = Lone Cabbage are significant as is the 1 year lag on annual discharge.

Overall it seems that site is important and that there may be a positive relationship between river discharge the previous year and oyster counts. Need to spend more time thinking about whether Site should be used as factor or random effect.

*Epoch 2 the pilot project*

Using all years but for the Lone Cabbage offshore site only, we fit a model that included year, treatment, and the 1 year lag on annual discharge.

> summary(fullmod.e2)

Call:

glm.nb(formula = count\_live ~ year + treatment + ann\_dis\_lag\_sc +

offset(log(tran\_length)) - 1, data = d2.1, init.theta = 1.07135215,

link = log)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.8232 -0.9940 -0.3429 0.4075 1.5338

Coefficients:

Estimate Std. Error z value Pr(>|z|)

year -0.15518 0.06341 -2.447 0.0144 \*

treatmentcontrol 311.05752 127.51265 2.439 0.0147 \*

treatmentrocks 312.14893 127.67712 2.445 0.0145 \*

ann\_dis\_lag\_sc 4.51355 0.74513 6.057 1.38e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.0714) family taken to be 1)

Null deviance: 785.96 on 54 degrees of freedom

Residual deviance: 62.53 on 50 degrees of freedom

AIC: 723.96

Number of Fisher Scoring iterations: 1

Theta: 1.071

Std. Err.: 0.190

2 x log-likelihood: -713.961

Results of this model suggest significant treatment and river discharge effects. Note the results are different if you include all of the LC control data and not just the offshore. This is because of the influence of site. Again, these data are only the Lone Cabbage offshore data.

If we add a station as a random effect to the model above, treatment is no longer significant

> summary(m1d2)

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [glmerMod]

Family: Negative Binomial(1.1153) ( log )

Formula: count\_live ~ treatment + ann\_dis\_lag\_sc + (1 | station) + offset(log(tran\_length))

Data: d2

AIC BIC logLik deviance df.resid

1348.6 1361.7 -669.3 1338.6 96

Scaled residuals:

Min 1Q Median 3Q Max

-1.0454 -0.5828 -0.1985 0.3573 3.3241

Random effects:

Groups Name Variance Std.Dev.

station (Intercept) 1.119 1.058

Number of obs: 101, groups: station, 19

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.5928 0.5513 -1.075 0.282

treatmentrocks 0.7617 0.5312 1.434 0.152

ann\_dis\_lag\_sc 3.7421 0.7326 5.108 3.25e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr) trtmnt

tretmntrcks -0.068

ann\_ds\_lg\_s -0.863 -0.099

This is really pretty interesting because if you look at the stddev of the random effect (station) you see that it is larger than the effect size for treatment. Need to think about this more but this may mean that the variability with the reefs is greater than the treatment effect, but, as the treatment is on the reefs then this may be a confounded result. Either way the magnitude of the treatment effect needs to be examined more closely as it may be biologically significant but not statistically significant at times.

*Key take away points*

Site is very important and site effects can dwarf any other effects. May need to move into a “ratio” approach of treatment:control if control sites are going to be from a different site. There also seems to be a potential relationship between the river discharge the year prior and oyster counts. But this warrants further work.

*More to do related to GLMs*

There is a lot more to do related to the GLM models. Chapter 6 in the Kery (2010) book is a nice deliberate way to walk through these types of models to present the models graphically. Need to take a closer look at the discharge data. Need to standardize the covariates relative to each other. Need to better understand when to consider a parameter as a fixed or random effect and if there are ways to compare these models (Bolker mentions this somewhere..a pseudo r-square?). These are good things to go over with Jose-Miguel Ponciano.