Homework 4

In this assignment we will again analyze NOAA reef fish survey data. The goal is to learn how to model count data, and to get a sense for the challenges in trying to deduce processes from observed patterns.

Imagine that we are interested in the habitat characteristics that favor *Ctenochaetus strigosus*, known as kole in Hawaiian, and with the English common name spotted surgeonfish in this dataset. We will focus on the role of different types of benthic substrate, and the most common substrate types in this dataset are turf algae (column 'ta'), hard coral (column 'hard_coral'), sand (column 'sand'), and crustose coralline algae (column 'cca'). I have included a dataset that only includes this species, and only the sampling period during which turf algae percent cover was measured.

(1) Create a model where the count of kole is a function of turf algae, hard coral, sand, and crustose coralline algae. For now, include all these terms in the model, because we are interested in all of them as potential drivers of abundance. Use a poisson distribution to model the count distribution. Using likelihood ratio tests for the model terms, which appear to be significantly related to kole abundance?

Plot the fitted model effects (it may be easiest to see the relationships with the y-axis on a log scale for this model). Based on these plots, which substrate types are predicted to have the largest effect on abundance? You can also compare the slope coefficients for the predictors, because all the predictors are measured on the same scale (percent cover).

(2) In the previous model we did not consider *overdispersion*, which means that the counts may have more variation than predicted by the poisson distribution. Now use a quasipoisson model instead. How much overdispersion is there, based on this model? How does accounting for overdispersion change the results? Which aspects of the model have changed, and which have not?

(3) Another option for overdispersed counts is to use a negative binomial distribution instead of a poisson distribution. How do the results from a negative binomial model compare to the quasipoisson approach?

(4) One of the main challenges in trying to decipher patterns from survey data is that 'correlation is not causation'. Make two new models, one where sand is the only predictor, and one where turf algae is the only predictor. For both models use the negative binomial distribution. How do the results of these models differ from the model in #3, where all four predictors were included together? What do you think could explain why the results have changed? It may help to look at a correlogram of the predictors. The function ggpairs() in the package GGally is particularly nice for this (note: I am not requiring you to make a correlogram here, but you may find it helpful).

Finally, what are your overall conclusions about substrate associations of kole, from this look at the data?