Homework 5 – Modeling kole counts

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) Start by exploring the distributions of the four predictor variables, and how they are correlated with one another. The function ggpairs() in the package GGally is particularly nice for this.

(2) For each of the four predictors, make single-predictor models where the response variable is kole counts. Consider which of the predictors could be transformed to reduce skew along the x-axis. For educational purposes, start by using a poisson distribution for the counts.

Quantify the degree of overdispersion in these models, and describe what it means. Based on likelihood ratio tests and effects plots, what are the relationships between kole and the four substrate types?

(3) Now make new single-predictor models that account for overdispersion (we discussed two different ways to do this in lecture, you can do one or both).

How have the results changed compared to #2, and why?

(4) Although the single-predictor models are useful, we may get a more accurate picture of the role of each substrate type if we include all the substrate types in one model. Create such a model. Based on likelihood ratio tests and effects plots, how do the results change when all predictors are included simultaneously? Why do you think the results have changed?

Don't forget to make some residual diagnostic plots as well.

Finally, what are your overall conclusions about the substrate associations of kole, from this look at the data? Are there any additional analyses you would want to do that we have not done in this assignment?