Capstone 1 - Inferential Statistics

Burns, Echelle

Dec 2019

In order to see how the predictor variables (receiver density, year, sea surface temperature, sea surface salinity, chlorophyll-a, depth gradient, lunar phase, zone) influenced the response variable (shark density), I ran a series of simple pairwise comparisons. In order to reduce biases that may be introduced by zero-inflated data, I removed all rows in which shark density values were equal to 0.

Receiver density (the number of receivers within a particular zone) was converted into a categorical variable. Results from an ordinary least squares ANOVA indicate a significant relationship between shark density and receiver density. Having two receivers in the water in the same zone resulted in significantly higher shark density values, compared to results from when one or three receivers were in the water ($p < 0.001$). It makes sense that having two receivers in the water at the same time would increase the ability to detect a shark in a particular zone; two receivers cover a larger proportion of the entire zone. However, under this logic, one would expect that three receivers would result in even higher shark densities, which is not supported by the model ($p = 0.625$). This can be explained by looking at the distribution of receiver density data. There are few instances in which three receivers are present in a single zone, which could explain why we do not see higher shark density values when three receivers are present.

Years were also converted into categorical variables and an ordinary least squares ANOVA was run. Again, there was a significant relationship between year and shark density. Years 2015, 2017, and 2018 had significantly higher shark densities when compared to 2014 ($p < 0.001$). However, 2017 exhibited statistically similar shark densities when compared to 2014 ($p = 0.153$) and 2019 had significantly lower shark densities compared to 2014 ($p = 0.028$). Perhaps years with higher densities corresponded to years with more receivers in the water or more ideal environmental conditions. Because 2019 is not yet over, seeing significantly smaller shark densities is not surprising.

When using all data values, it is difficult to determine a clear relationship between sea surface temperature and shark density; the variation in shark density at different temperature values is quite large. Such variation can be described partially by looking at the relationship between sea surface temperature and shark density across years; in 2014, 2015, and 2016, shark densities are higher at ~ 16 degrees C, but in 2017 shark densities are higher around 18-22 degrees C, and in 2018 shark densities are higher at 12-14 degrees C. However, even when data are split up by year, the variation at each temperature value remains high. A second approach converted the data into a series of binomial points, where the response variable would be simply shark presence or absence. However, a quick plot of these data show that there are sharks present at nearly all temperatures; this would not help to determine which temperatures are more appealing to juvenile white sharks. Therefore, the final approach included rounding temperatures into 0.1 degree increments and keeping only the maximum shark density that was observed in that temperature range. With this approach, we can still determine whether which temperatures may be more appealing to more than one individual. The data do not appear to have a linear relationship, which was supported by the results of a correlation test (correlation coefficient = -0.1938, p = 0.030). It did appear that there were two peaks in the dataset, where shark densities were higher around 12-14 degrees C and 18-22 degrees C. Therefore, I used bootstrapping (sampling with replacement) techniques to fit a fourth degree polynomial to the dataset. I ran the bootstrap 100 times to get an idea of how missing data values could influence the resulting curve. I also ran the polynomial fit on the entire dataset. The resulting formula from the model that used all data values was:

$$y = -0.013x^4 + 0.93x^3 - 24.5x^2 + 280x - 1169$$

I used a similar approach to calculate the relationship between maximum shark density and salinity (0.1 increments). Once again, there was no linear relationship between salinity and maximum shark density (correlation coefficient = -0.2736, p = 0.012) and there appeared to be a peak in shark density at 33-35 psu. These data were fit to a third degree polynomial and models were run for 100 bootstrapped samples and for the entire dataset. The resulting formula from the model that used all data values was:

$$y = 0.14x^3 - 15x^2 + 531x - 6245$$

For chlorophyll-a, there was only one instance in which shark density exceeded 0 at high chlorophyll levels (> 50 mg/m^3). Therefore, it appears that as long as chlorophyll levels are low, sharks are present. I also calculated the maximum shark density at each 0.1 value of chlorophyll-a. However, there appears to be no correlation between these two values (correlation coefficient = -0.1458, p = 0.203). After conducting a log transformation on shark density and chlorophyll-a values, however, a strong, negative correlation was present (correlation coefficient = -0.6036, p = 0.0). This suggests that as chlorophyll-a values increase logarithmically, maximum shark density values decrease logarithmically.

Shark density was also highly variable across all values of depth gradient (seafloor slope), so the maximum shark density per 0.001 was calculated. However, because depth gradient values occurred between a narrow range (0 m to 0.025 m), there was no linear relationship between maximum shark density and depth gradient (correlation coefficient = 0.359, p = 0.157). These results did not change when maximum shark density values were log-transformed (correlation coefficient = 0.282, p = 0.273).

Lunar phase was compared to shark density using an ordinary least squares ANOVA, but the results indicated that there was not a significant difference in shark density across different lunar phases (p > 0.05). Zone IDs were also compared to shark density using an ordinary least squares ANOVA. At least 8 zones had significantly higher shark densities than the other 42 zones that were present in the model. These zones may represent nearshore ‘hotspot’ areas in which juvenile white sharks spend their time.