Kendall Model

Kendall Fitzgerald

2025-04-04

Invertebrates to include in analysis: purple urchin, sunflower sea star, northern kelp crab

Setup

```
#load necessary packages
library(here)
```

```
## here() starts at /home/guest/ENV710/group_project
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(naniar)
library(gtsummary)
library(lme4)
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:gtsummary':
##
##     select
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
#load dataset
inverts_kelp <- read_csv("PISCO_kelpforest_swath.1.2.csv")
```

```
## Rows: 222227 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): campus, method, site, zone, classcode, disease, observer, notes, si...
## dbl (8): survey_year, year, month, day, transect, count, size, depth
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
#tidy data, the addition of recruit and juvenile classcodes does not make a difference so will move for
inverts_kelp <- inverts_kelp %>%
  filter(campus == "UCSC",
         classcode %in% c("PUGPRO", "STRPURAD", "PYCHEL", "MACPYRAD")) %>%
  #only include relevant columns for further analysis
  dplyr::select(campus, survey_year, site, zone, classcode, count, size, disease, depth)

#further tidy data based on relevant rows and sum count data
inverts_summarized <- inverts_kelp %>%
  group_by(campus, site, survey_year, classcode, depth) %>%
  summarise(
    total_count = sum(count, na.rm = TRUE),
    zone = first(zone),
    .groups = "drop"
  )

#group by species (classcode), year (survey_year), and zone
inverts_wide_grouped <- inverts_summarized %>%
  group_by(classcode, survey_year, zone) %>%
  summarise(across(c(total_count), sum, na.rm = TRUE), .groups = "drop")
```

```
## Warning: There was 1 warning in `summarise()`.
## i In argument: `across(c(total_count), sum, na.rm = TRUE)`.
## i In group 1: `classcode = "MACPYRAD"`, `survey_year = 1999`, `zone = "INNER"`.
## Caused by warning:
## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.
## Supply arguments directly to `.fns` through an anonymous function instead.
```

```
##
##   # Previously
##   across(a:b, mean, na.rm = TRUE)
##
##   # Now
##   across(a:b, \(x) mean(x, na.rm = TRUE))
```

```
#calculate the sum of all observations
sum(inverts_wide_grouped$total_count)
```

```
## [1] 825666
```

```
#pivot to wide format to separate out species into separate rows
inverts_wide <- inverts_wide_grouped %>%
  pivot_wider(names_from = classcode, values_from = total_count, values_fill = list(total_count = 0))

#calculate the sum of giant kelp observations
sum(inverts_wide$MACPYRAD)
```

```
## [1] 35873
```

```
#calculate the sum of northern kelp crab observations
sum(inverts_wide$PUGPRO)
```

```
## [1] 213
```

```
#calculate the sum of sunflower sea star observations
sum(inverts_wide$PYCHEL)
```

```
## [1] 4825
```

```
#calculate the sum of purple urchin observations
sum(inverts_wide$STRPURAD)
```

```
## [1] 784755
```

Step 1. Define Research Question

How does the presence of different invertebrates and habitat zone correlate with kelp abundance?

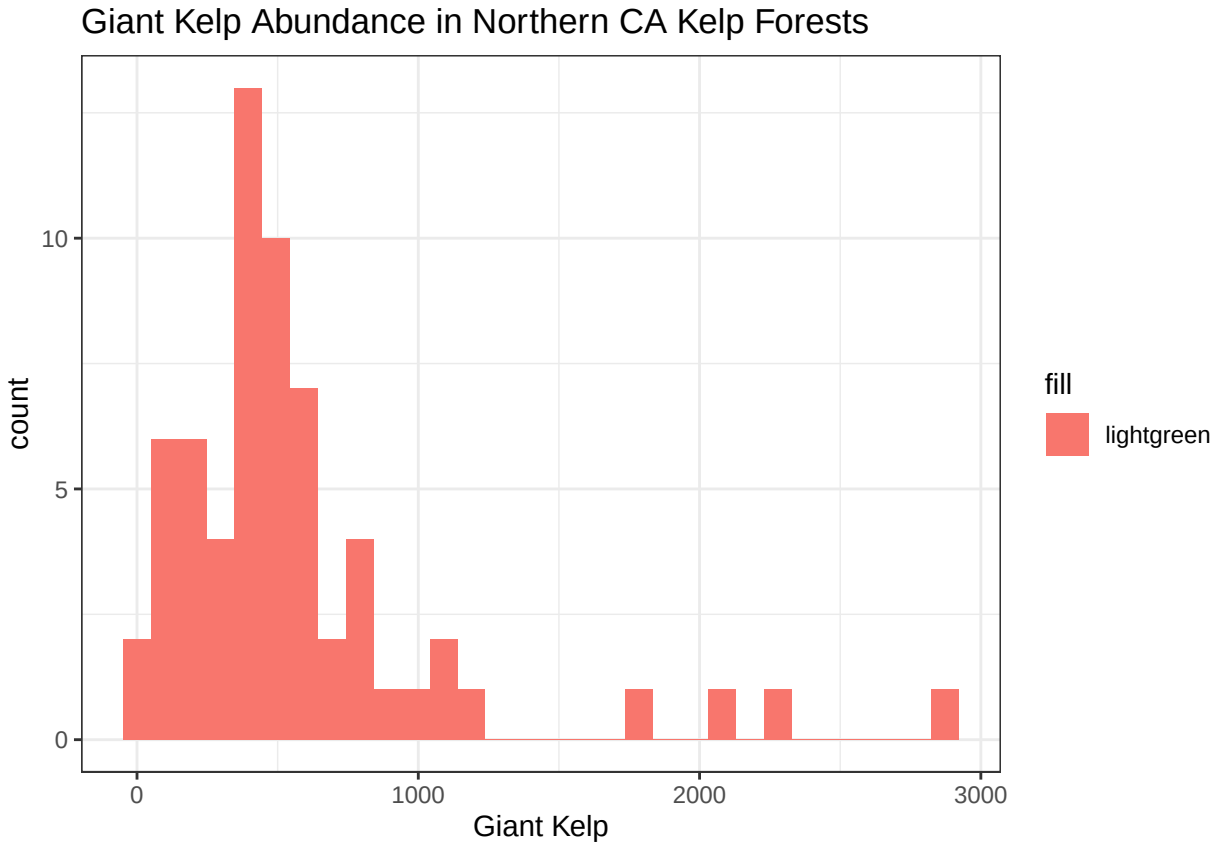
Null Hypothesis: Invertebrate presence and habitat zone do not have any correlation with kelp abundance.

Alternative Hypothesis: Invertebrate presence and habitat zone do have a correlation with kelp abundance.

Step 2. Examine data and possible correlations

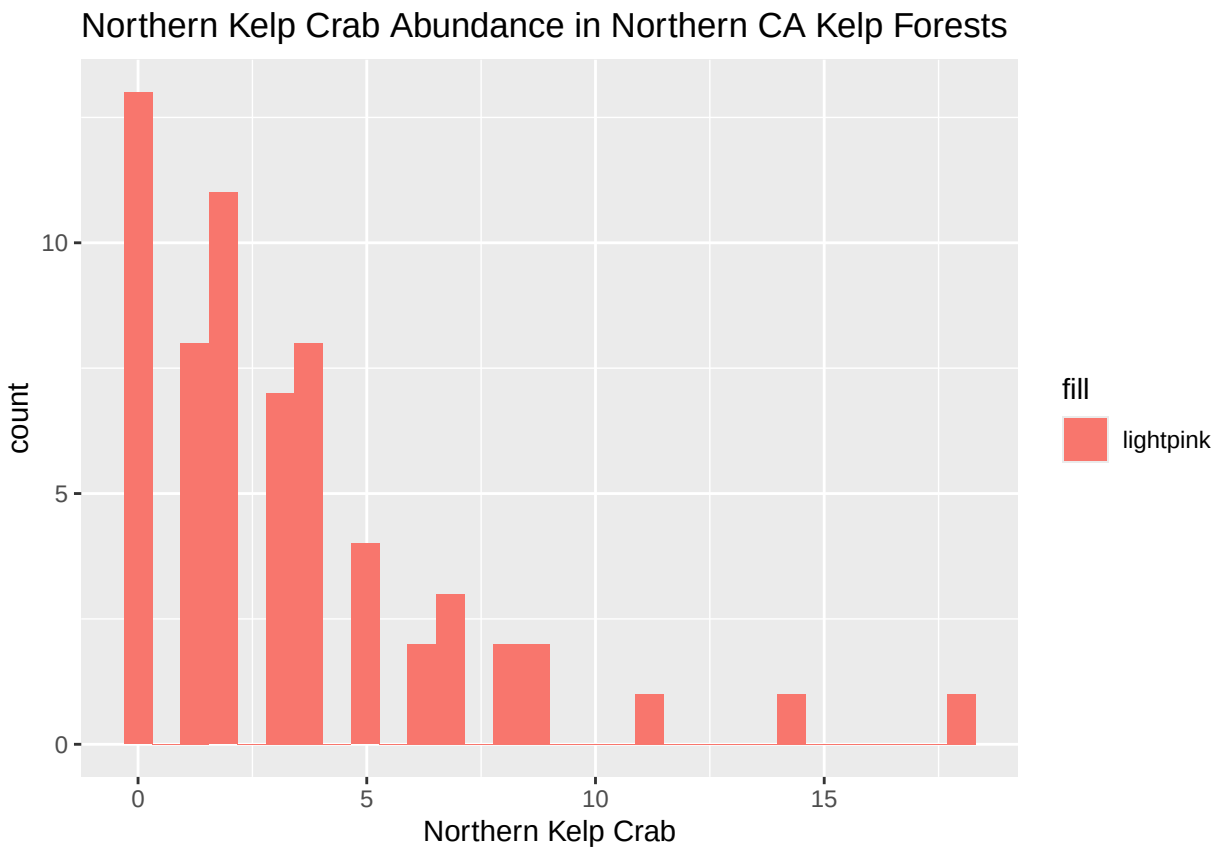
```r
#create and examine figures of raw values

#histograms of species count

#histogram of giant kelp
kelp_fig <- ggplot(inverts_wide, aes(x = MACPYRAD,
                                     fill = "lightgreen")) +
  geom_histogram() +
   labs(x = "Giant Kelp",
        title = "Giant Kelp Abundance in Northern CA Kelp Forests") +
  theme_bw()

kelp_fig
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



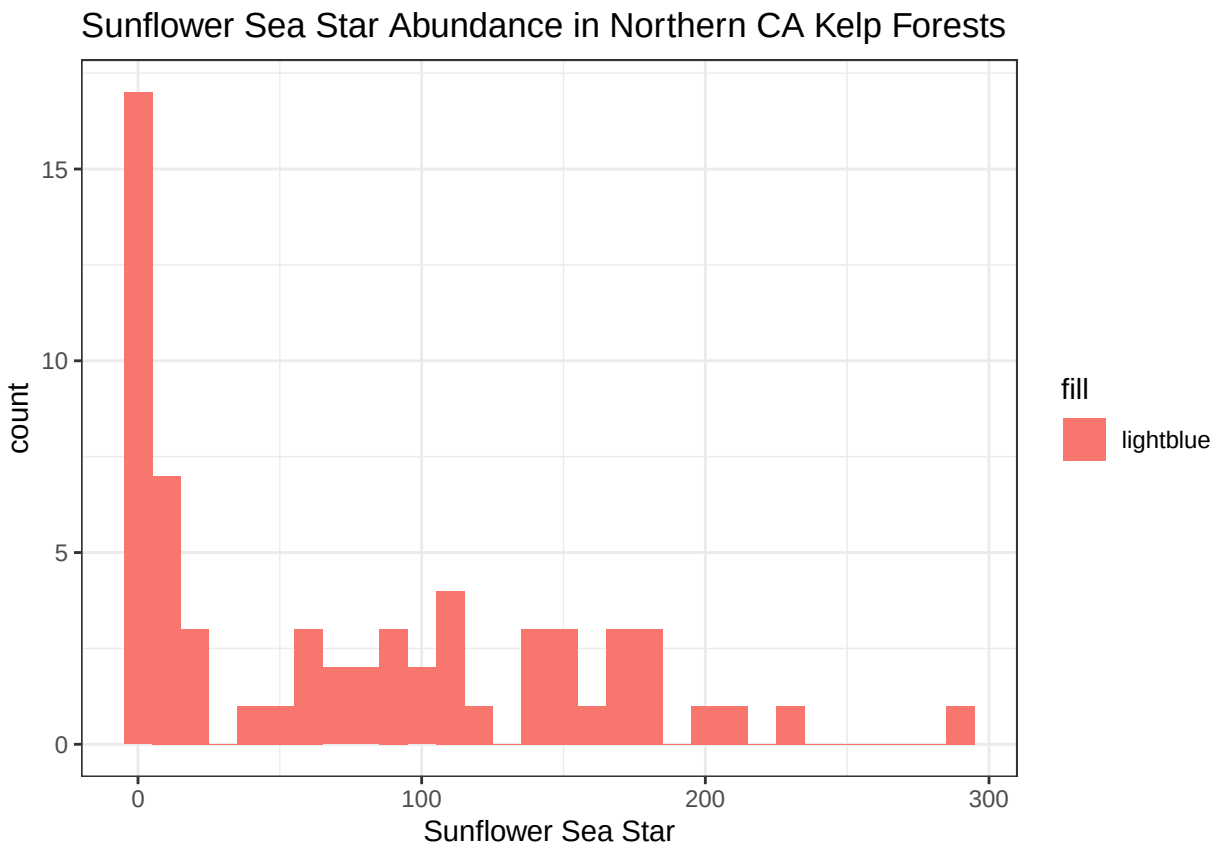

```r
#histogram of Northern kelp crab
crab_fig <- ggplot(inverts_wide, aes(x = PUGPRO,
                                     fill = "lightpink")) +
  geom_histogram() +
   labs(x = "Northern Kelp Crab",
        title = "Northern Kelp Crab Abundance in Northern CA Kelp Forests")
crab_fig
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



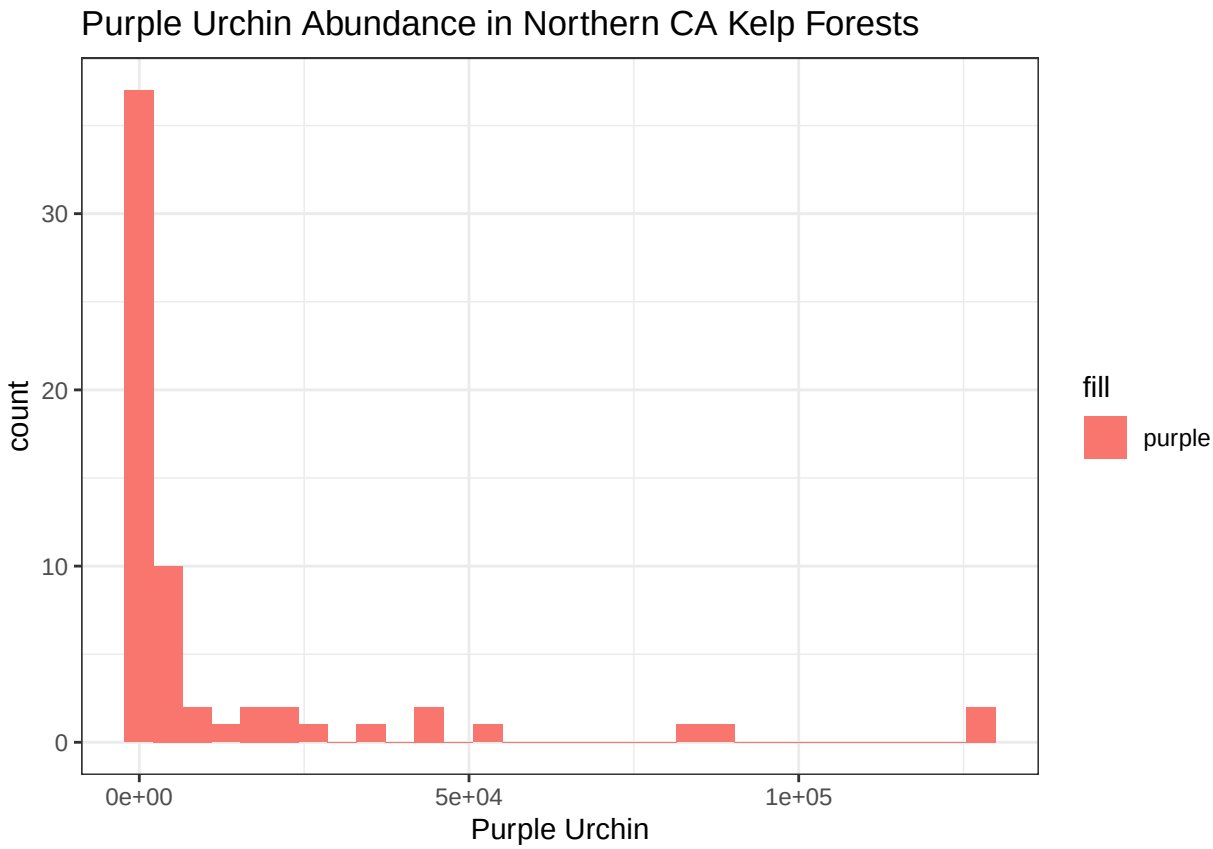

```
#histogram of Sunflower Sea Star
seastar_fig <- ggplot(inverts_wide, aes(x = PYCHEL,
                                       fill = "lightblue")) +
  geom_histogram() +
   labs(x = "Sunflower Sea Star",
        title = "Sunflower Sea Star Abundance in Northern CA Kelp Forests") +
  theme_bw()

seastar_fig
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

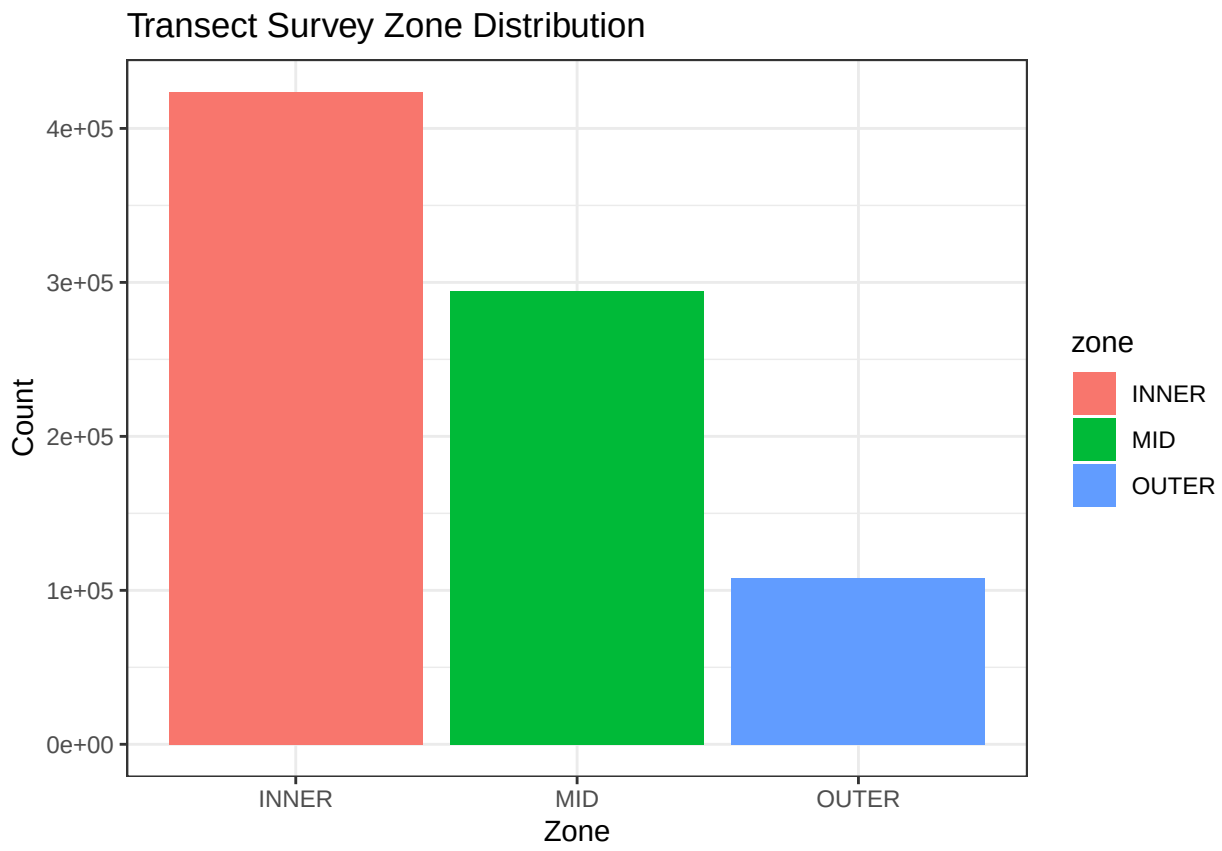



```
#histogram of purple urchin
urchin_fig <- ggplot(inverts_wide, aes(x = STRPURAD,
                                       fill = "purple")) +
  geom_histogram() +
   labs(x = "Purple Urchin",
        title = "Purple Urchin Abundance in Northern CA Kelp Forests") +
  theme_bw()

urchin_fig
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#bar graph of zone distribution (categorical variable)
zone_fig <- ggplot(inverts_kelp, aes(x = zone,
                                     y = count,
                                     fill = zone)) +
  geom_bar(stat = "identity") +
     labs(y = "Count",
       x = "Zone",
       title = "Transect Survey Zone Distribution") +
  theme_bw()

zone_fig
```



Relationship figures of species through time

```
#first tidy data so that it can be more easily understood by ggplot; summarize the counts by species an
species_counts <- inverts_summarized %>%
  group_by(classcode, survey_year) %>%
  summarise(total_count = sum(total_count, na.rm = TRUE), .groups = "drop")

#create labels for graph
labels <- c("MACPYRAD"="Giant Kelp",
                              "PUGPRO"="Northern Kelp Crab",
                              "PYCHEL"="Sunflower Sea Star",
                              "STRPURAD" = "Purple Urchin")

#create facet-wrapped figure to show species counts by year
species_separate <- ggplot(species_counts, aes(x = survey_year, y = total_count)) +
  geom_point(aes(color = classcode), size = 2, alpha = 0.7) +
  geom_smooth(aes(color = classcode), method = "loess", se = FALSE) +
  geom_vline(xintercept = 2013, linetype = "longdash", color = "black")  +
  facet_wrap(~ classcode, scales = "free_y", labeller = labeller(classcode = labels))+
  labs(title = "Distribution of Kelp Forest Species Abundance Used in Model Over Time",
       x = "Survey Year",
       y = "Total Count") +
  theme_bw() +
  theme(strip.text = element_text(face = "bold"),
```

```
        axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5, face = "bold", size = 12),
        legend.position = "none")

species_separate
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

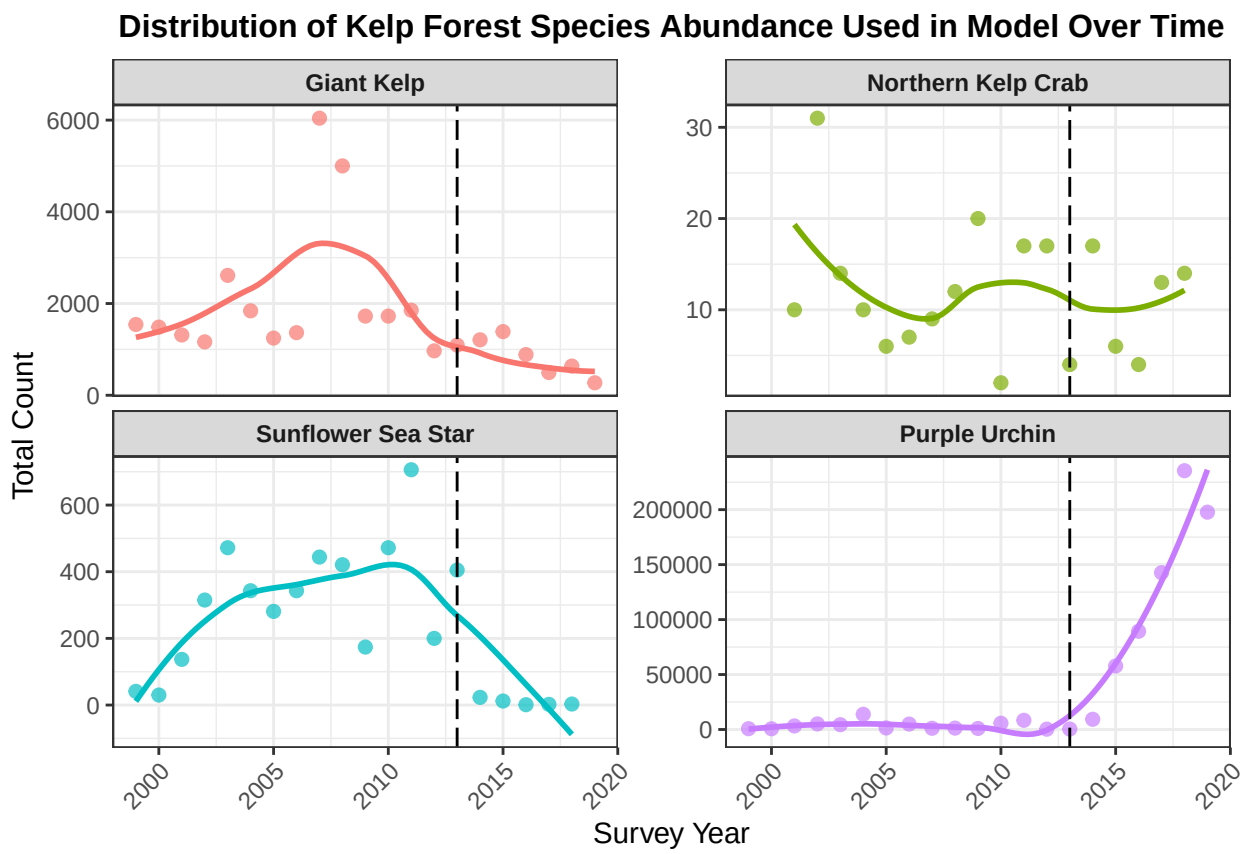


Figure 1: This figure shows the abundance of species used in the model over time. The black dotted line sitting at the year 2013 indicates the first large marine heatwave during that time period. This heatwave significantly contributed to the alteration of the kelp forest ecosystem.

Performing correlation tests between independent variables

```
#make dataframe with continuous predictor variables only
predictors <- inverts_wide %>%
  dplyr::select(PUGPRO, STRPURAD, PYCHEL)

#calculate the correlation matrix
cor_matrix <- cor(predictors, use = "complete.obs")

# View the correlation matrix
cor_matrix
```

```
##                PUGPRO    STRPURAD     PYCHEL
## PUGPRO    1.00000000 -0.02495791  0.2427995
## STRPURAD -0.02495791  1.00000000 -0.4112285
## PYCHEL    0.24279953 -0.41122852  1.0000000
```

Because correlation does not exceed 0.6 between all variables, the test showed little to moderate correlation between the continuous predictor variables. Since this is the case, we can move forward with step 3.

Step 3 - Fit regression model

```
#check the mean and variance of the count variable
mean(inverts_wide$MACPYRAD)
```

```
## [1] 569.4127
```

```
var(inverts_wide$MACPYRAD)
```

```
## [1] 274078.1
```

```
#because the variance is significantly higher than the mean, we cannot use Poisson regression so we wil

#fit a negative binomial regression model
kelp_glm_nb <- glm.nb(MACPYRAD ~ PUGPRO + STRPURAD + PYCHEL + zone,
                      data = inverts_wide)
```

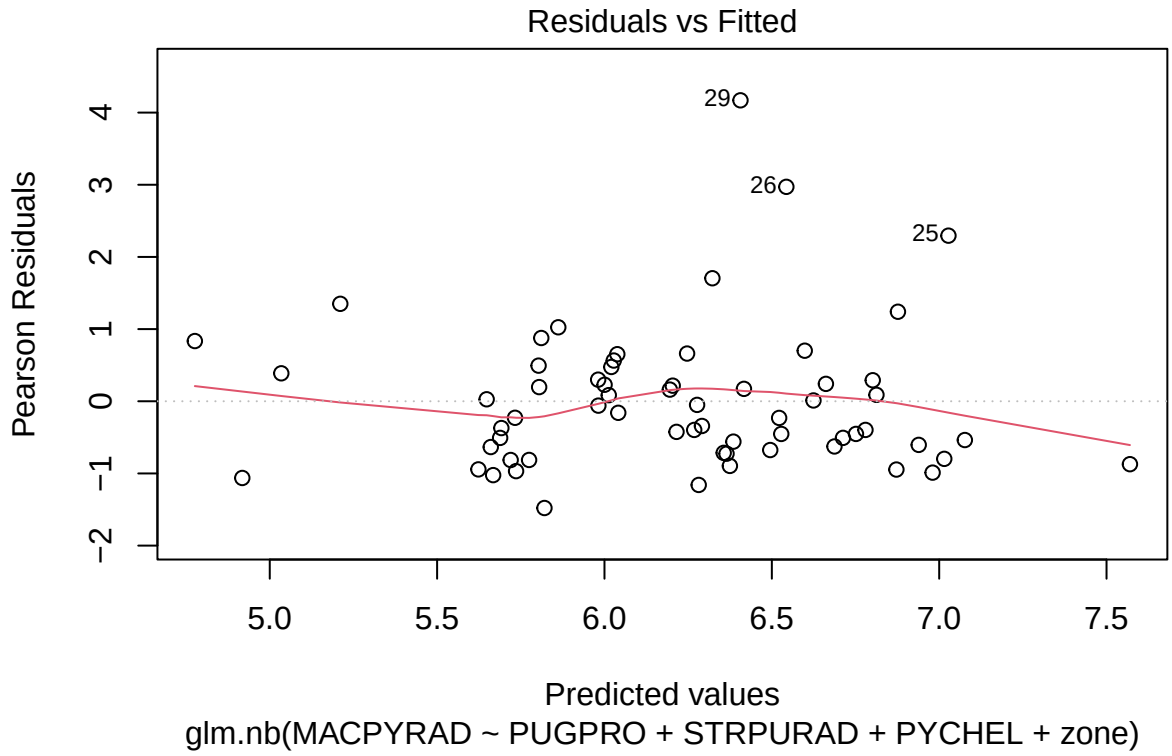
Step 4 - Evaluate Model Diagnostics

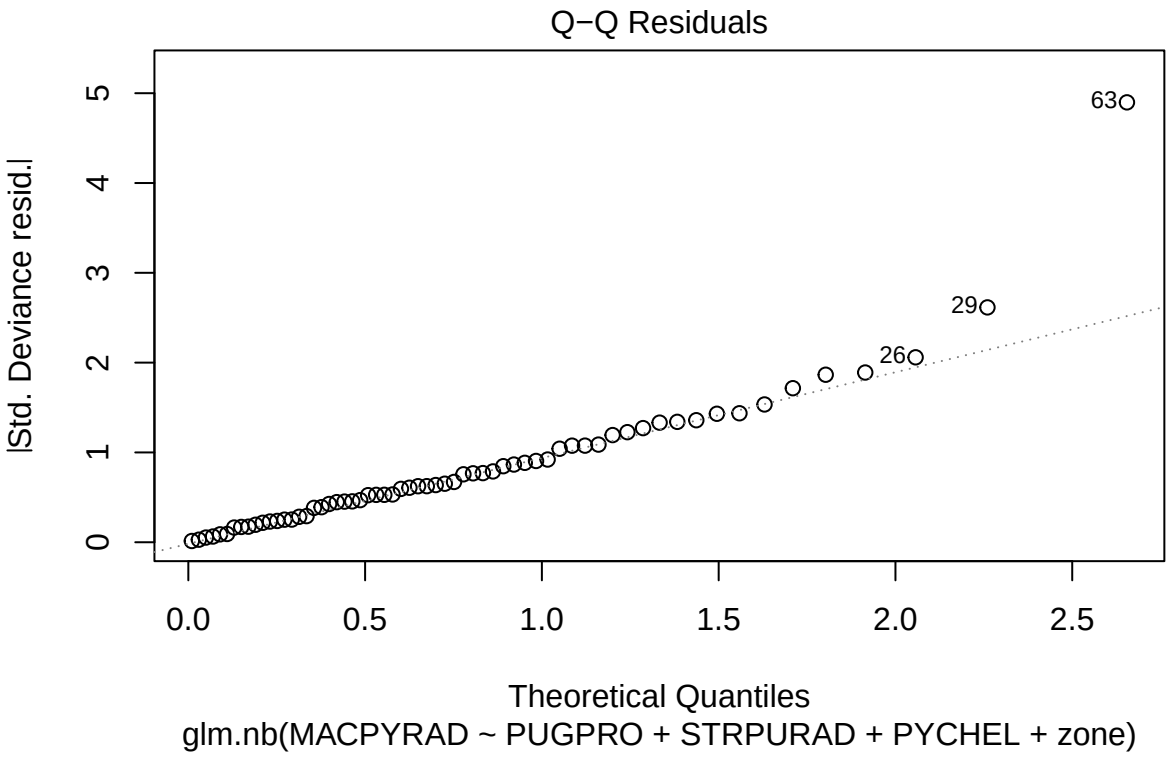
```
#Examine overall model output
summary(kelp_glm_nb)
```

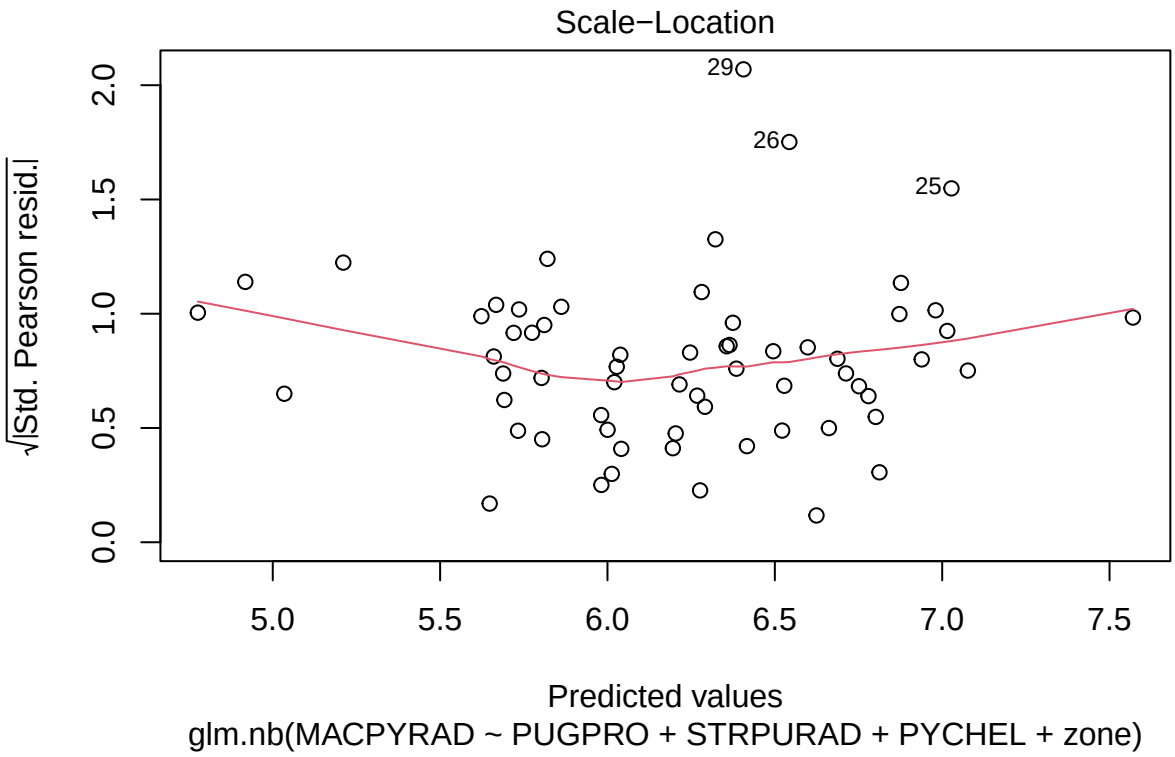
```
##
## Call:
## glm.nb(formula = MACPYRAD ~ PUGPRO + STRPURAD + PYCHEL + zone,
##     data = inverts_wide, init.theta = 2.201056427, link = log)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.975e+00  2.240e-01  26.667  < 2e-16 ***
## PUGPRO      -3.216e-02  2.574e-02  -1.250    0.211
## STRPURAD    -8.380e-06  3.573e-06  -2.345    0.019 *
## PYCHEL       5.521e-03  1.321e-03   4.178 2.94e-05 ***
## zoneMID      1.507e-01  2.112e-01   0.713    0.476
## zoneOUTER   -5.740e-02  2.227e-01  -0.258    0.797
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(2.2011) family taken to be 1)
##
```

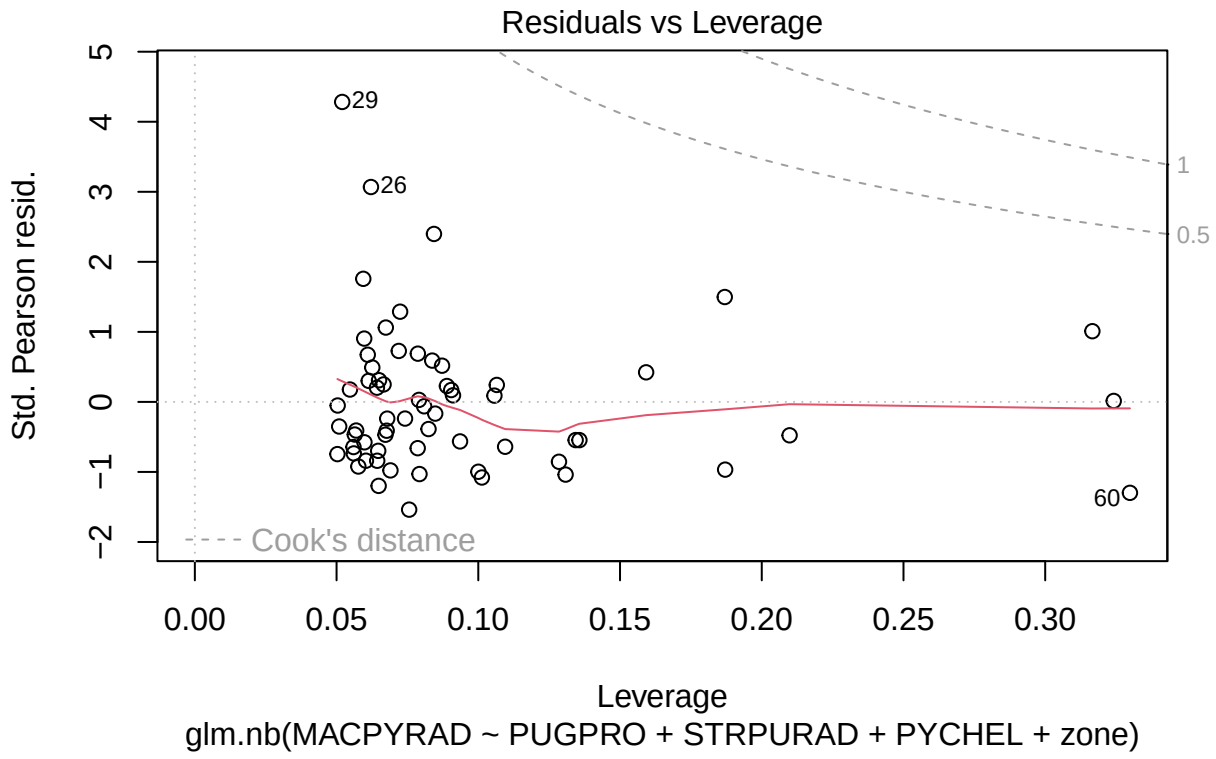
```
##     Null deviance: 106.503  on 62  degrees of freedom
## Residual deviance:  70.837  on 57  degrees of freedom
## AIC: 906.23
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  2.201
##           Std. Err.:  0.383
##
##  2 x log-likelihood:  -892.233
```

```
#plot model output residuals
plot(kelp_glm_nb)
```




Q−Q Residuals
|Std. Deviance resid.|
0
1
2
3
4
5
63
29
26
0.0
0.5
1.0
1.5
2.0
2.5
Theoretical Quantiles
glm.nb(MACPYRAD ~ PUGPRO + STRPURAD + PYCHEL + zone)


Scale−Location
29
26
25
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
2.0
5.0
5.5
6.0
6.5
7.0
7.5
Predicted values
glm.nb(MACPYRAD ~ PUGPRO + STRPURAD + PYCHEL + zone)




```
#take out relevant outliers based on residual plots
inverts_final <- inverts_wide[-c(26, 29, 63), ]

#calculate the new sum of giant kelp observations
sum(inverts_final$MACPYRAD)
```

```
## [1] 31474
```

```
#calculate the new sum of northern kelp crab observations
sum(inverts_final$PUGPRO)
```

```
## [1] 205
```

```
#calculate the new sum of sunflower sea star observations
sum(inverts_final$PYCHEL)
```

```
## [1] 4591
```

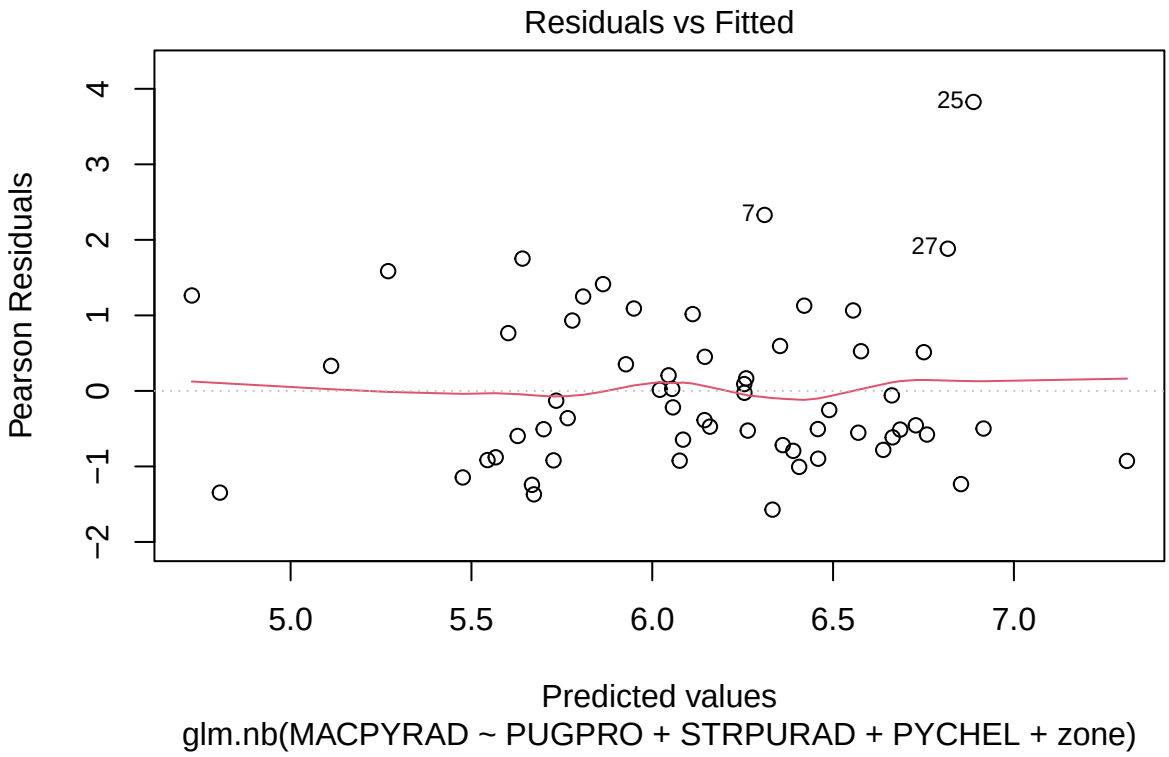
```
#calculate the new sum of purple urchin observations
sum(inverts_final$STRPURAD)
```

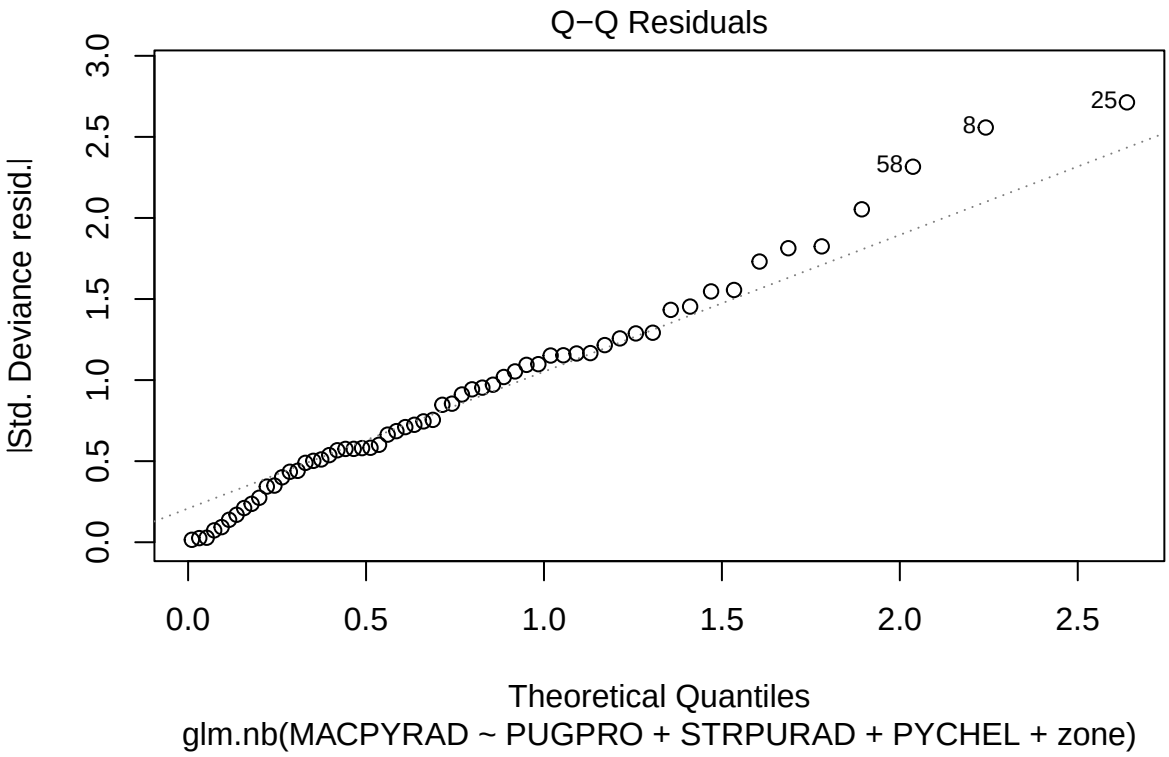
```
## [1] 782875
```

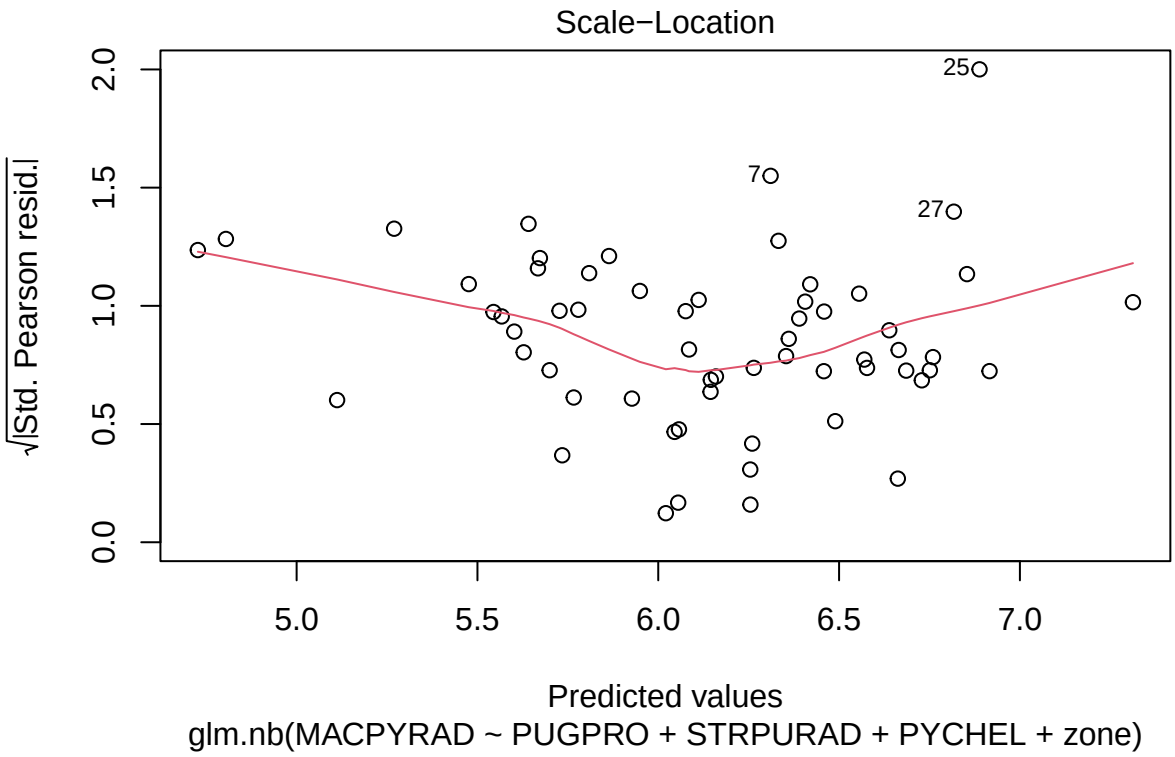
```r
#refit model without outliers
kelp_glm_refit <- glm.nb(MACPYRAD ~ PUGPRO + STRPURAD + PYCHEL + zone,
                         data = inverts_final)

#re-examine model output
summary(kelp_glm_refit)
```

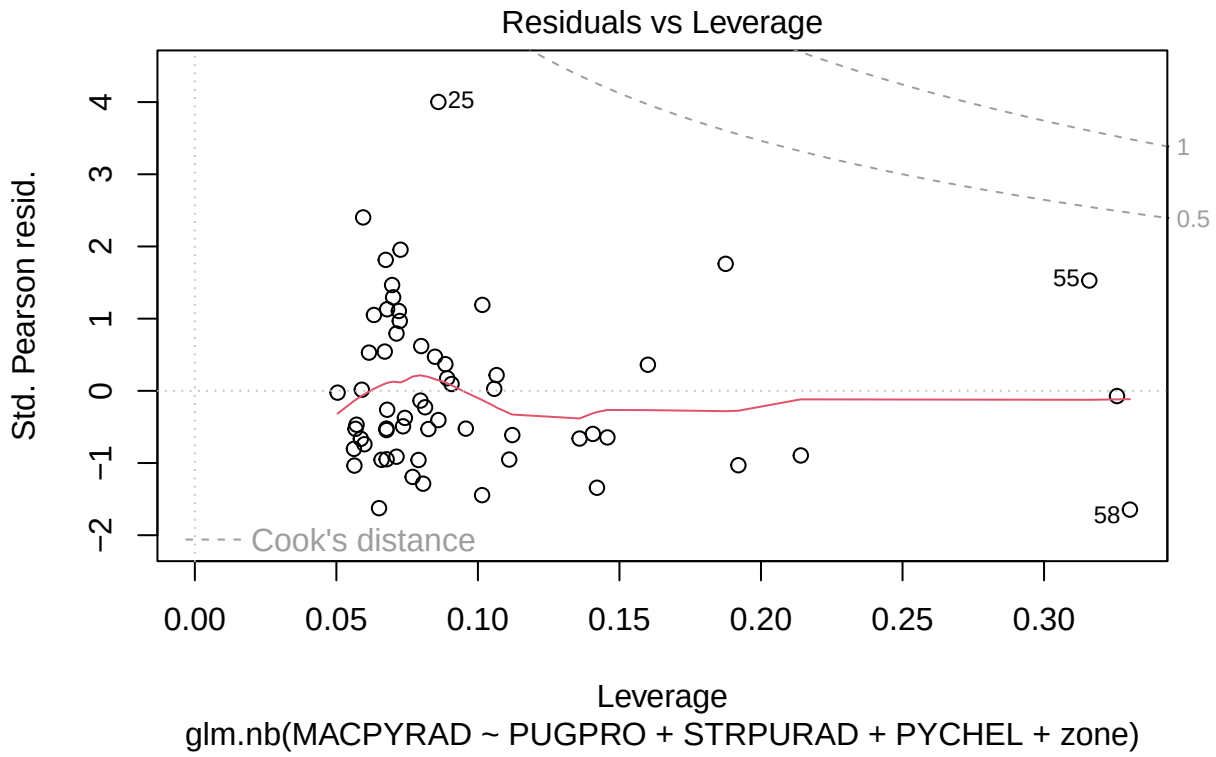
```
##
## Call:
## glm.nb(formula = MACPYRAD ~ PUGPRO + STRPURAD + PYCHEL + zone,
##     data = inverts_final, init.theta = 3.948276584, link = log)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.026e+00  1.677e-01  35.936  < 2e-16 ***
## PUGPRO      -1.561e-02  1.947e-02  -0.802 0.422712
## STRPURAD    -9.691e-06  2.689e-06  -3.604 0.000314 ***
## PYCHEL       4.155e-03  1.006e-03   4.131 3.62e-05 ***
## zoneMID      1.757e-01  1.581e-01   1.112 0.266152
## zoneOUTER   -2.947e-01  1.727e-01  -1.707 0.087836 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(3.9483) family taken to be 1)
##
##     Null deviance: 121.172  on 59  degrees of freedom
## Residual deviance:  62.554  on 54  degrees of freedom
## AIC: 828.22
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  3.948
##           Std. Err.:  0.701
##
##  2 x log-likelihood:  -814.223
```

```r
#re-examine model residuals
plot(kelp_glm_refit)
```


Residuals vs Fitted
Pearson Residuals
4
3
2
1
0
−1
−2
25
7
27
5.0
5.5
6.0
6.5
7.0
Predicted values
glm.nb(MACPYRAD ~ PUGPRO + STRPURAD + PYCHEL + zone)


Q−Q Residuals
|Std. Deviance resid.|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
25
8
58
Theoretical Quantiles
glm.nb(MACPYRAD ~ PUGPRO + STRPURAD + PYCHEL + zone)


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
2.0
25
7
27
5.0
5.5
6.0
6.5
7.0
Predicted values
glm.nb(MACPYRAD ~ PUGPRO + STRPURAD + PYCHEL + zone)




```
#perform Akaike Information Criterion test to compare the two models
AIC(kelp_glm_nb, kelp_glm_refit)
```

```
##                df      AIC
## kelp_glm_nb     7 906.2329
## kelp_glm_refit  7 828.2233
```

The results of the Akaike Information Criterion (AIC) test showed that the re-fit model had a lower AIC score and, therefore, was a better fit of the data.

Step 5 - Interpret the model and communicate the results

```
library(broom)
library(knitr)
library(gt)

#create table from model
model_table <- tidy(kelp_glm_refit)

#tidy table
model_table_simple <- broom::tidy(kelp_glm_refit) %>%
  dplyr::select(term, estimate, p.value) %>%
  dplyr::mutate(
```

```
    term = dplyr::case_when(
      term == "(Intercept)" ~ "Intercept",
      term == "PUGPRO" ~ "Northern Kelp Crab (PUGPRO)",
      term == "STRPURAD" ~ "Purple Urchin (STRPURAD)",
      term == "PYCHEL" ~ "Sunflower Sea Star (PYCHEL)",
      term == "zoneMID" ~ "Zone: Mid",
      term == "zoneOUTER" ~ "Zone: Outer",
      TRUE ~ term),
    Significance = dplyr::case_when(
      p.value < 0.001 ~ "***",
      p.value < 0.01 ~ "**",
      p.value < 0.05 ~ "*",
      p.value < 0.1 ~ ".",
      TRUE ~ ""),
    estimate = round(estimate, 3),
    p.value = signif(p.value, 3)) %>%
  dplyr::rename(
    Predictor = term,
    Coefficient = estimate,
    `P-value` = p.value)

#use kable package to make it a more attractive output
nbmodel_table <- kable(model_table_simple, caption = "Negative Binomial Regression Model Coefficients")

nbmodel_table
```

Table 1: Negative Binomial Regression Model Coefficients

Predictor	Coefficient	P-value	Significance
Intercept	6.026	0.00e+00	***
Northern Kelp Crab (PUGPRO)	-0.016	4.23e-01	
Purple Urchin (STRPURAD)	0.000	3.14e-04	***
Sunflower Sea Star (PYCHEL)	0.004	3.62e-05	***
Zone: Mid	0.176	2.66e-01	
Zone: Outer	-0.295	8.78e-02	.

The Negative Binomial regression model suggests that the abundance of Giant Kelp in study areas of Northern California kelp forests is significantly influenced by the presence of Sunflower Sea Stars (B = 0.004, p = 3.62e-05) and Purple Urchins (B = -9.691e-06, p = 0.0003). Northern kelp crabs (B = -0.0156, p = 0.42), Mid Zone (B = 0.176, p = 0.266), and Outer Zone (B = -0.295, p = 0.0878) by contrast are not significant predictors for Giant Kelp abundance. The significant negative relationship between Giant Kelp and Purple Urchins shows that as urchin populations increase, kelp abundance decreases which makes logical sense as urchins predate on kelp. When left unchecked, they can turn entire kelp forests into "Urchin Barrens" with little to no kelp population. The strong positive relationship between Sunflower Sea Star abundance and kelp abundance also supports this narrative, as Sea Stars primarily predate on urchins and keep their population in check, which in turn helps the kelp population.