Diversity Indices

Katie Lankowicz

2023-07-05

Motivation

This document serves as an example for how to reshape field data to quantify biodiversity in each net haul for CBASS. We will calculate two indices: Shannon and Simpson. The Shannon index is defined as S-WI = -sum(p_i_) * ln(p_i_), or the sum of the proportion of the entire community made up of species i multiplied by the natural log of the proportion of the entire community made up of species i. High Shannon index values indicate high biodiversity. An index value of 0 indicates that a community only has one species. The index is sensitive to rare species, and is generally better at providing quantitative measures of species richness. Shannon assumes random sampling and that all species are represented in each net haul, which is likely a poor assumption for shore-based seines that occur in different habitats of Casco Bay.

The second diversity index is the Simpson index. This index is defined as 1 - sum(p_i_)^2, or 1 minus the sum of the proportion of the entire community made up of species i squared. The index will range between 0 and 1, with 0 indicating no diversity and 1 indicating infinite diversity. The Simpson index is not as sensitive to rare species, and does a better job of quantifying relative abundance.

Data loading

Before calculating the indices, the data need to be loaded. We also need to determine which species were caught in at least one seine haul of the 509 seine hauls that caught fish.

```
##
##                alewife           american eel          american shad
##                    105                      1                      4
##           atlantic cod       atlantic herring      atlantic menhaden
##                      2                     95                      4
##    atlantic silverside       banded killifish       blueback herring
##                    304                      3                      6
##               bluefish             butterfish             common dab
##                     16                      3                      1
##          crevalle jack         emerald shiner          golden shiner
##                      1                      1                      2
##             green crab         grubby sculpin                   hake
##                    356                     26                      2
##         horseshoe crab        largemouth bass       longhorn sculpin
##                      5                      4                      6
##               lumpfish                 mullet              mummichog
##                      2                      5                    162
##  ninespine stickleback      northern pipefish        northern puffer
##                      3                     28                      2
##             periwinkle                 permit                pollock
##                      1                      2                      5
##               red hake          river herring            rock gunnel
##                      1                      1                      7
##              sandlance         shortfin squid      shorthorn sculpin 
##                     63                      1                     20 
##     shortnose sturgeon          slimy sculpin        smallmouth bass 
##                      1                      1                      4 
##                  smelt           spotted hake           striped bass 
##                      3                      1                      6 
##        striped sculpin threespine stickleback                 tomcod 
##                      1                     12                     59 
##            unID gunnel           unID sculpin            unID shiner 
##                      1                     17                      3 
##          unID sturgeon             white hake           white mullet 
##                      4                      2                      1 
##           white sucker        winter flounder 
##                      5                    209
```

Data reshaping

Now the data need to be merged and reshaped. Currently, the abundance dataframe has a row for each species caught in each seine haul. We want to reshape this so we have one row for each seine haul, and each species is represented as its own column.

Diversity index calculation

That was the hard part. Now that the data have been reshaped, we can calculate our diversity indices. There is a built-in function for this within the `vegan` package.

Models

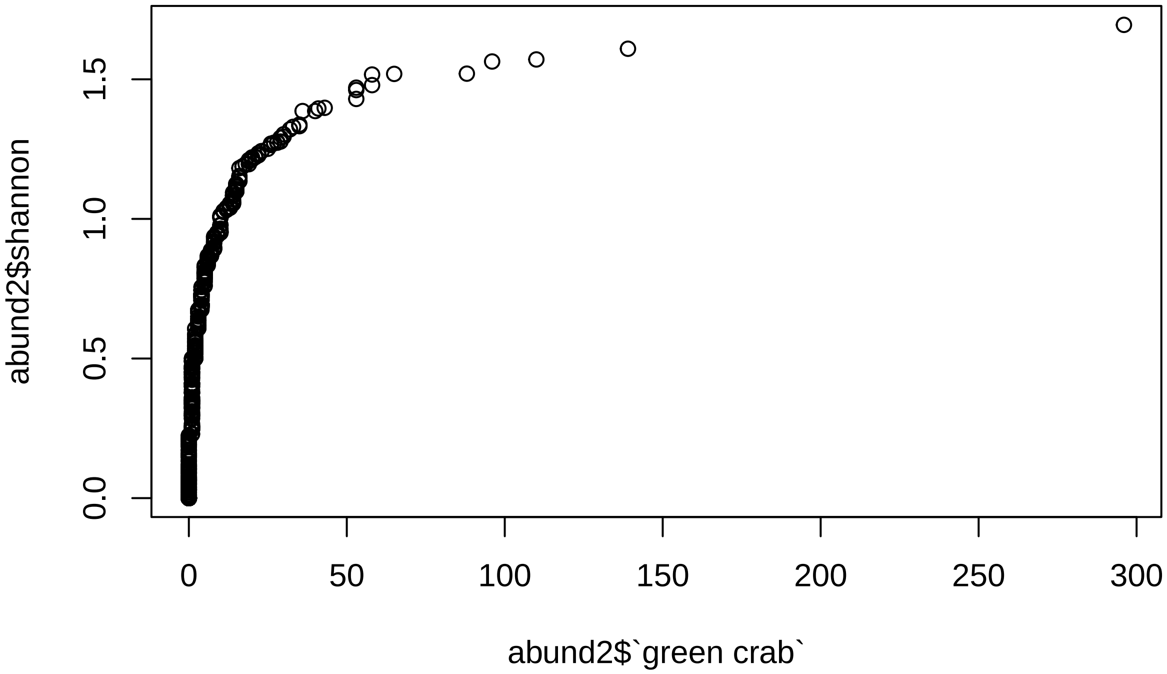
Eventually, we will use this information to build a Generalized Additive Model to test the relationship between green crab abundance and species diversity. For now, we will check some model assumptions and do some less-complicated test models to confirm that we have enough data to address our research questions.

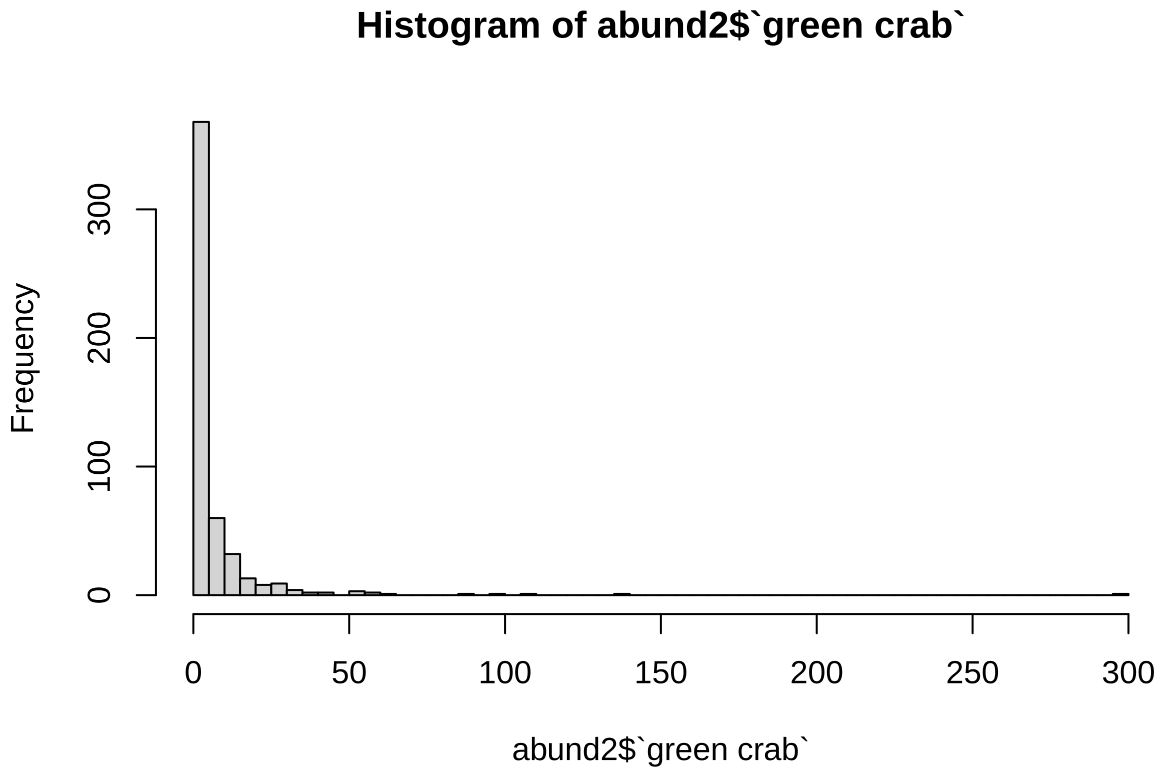
Assumptions

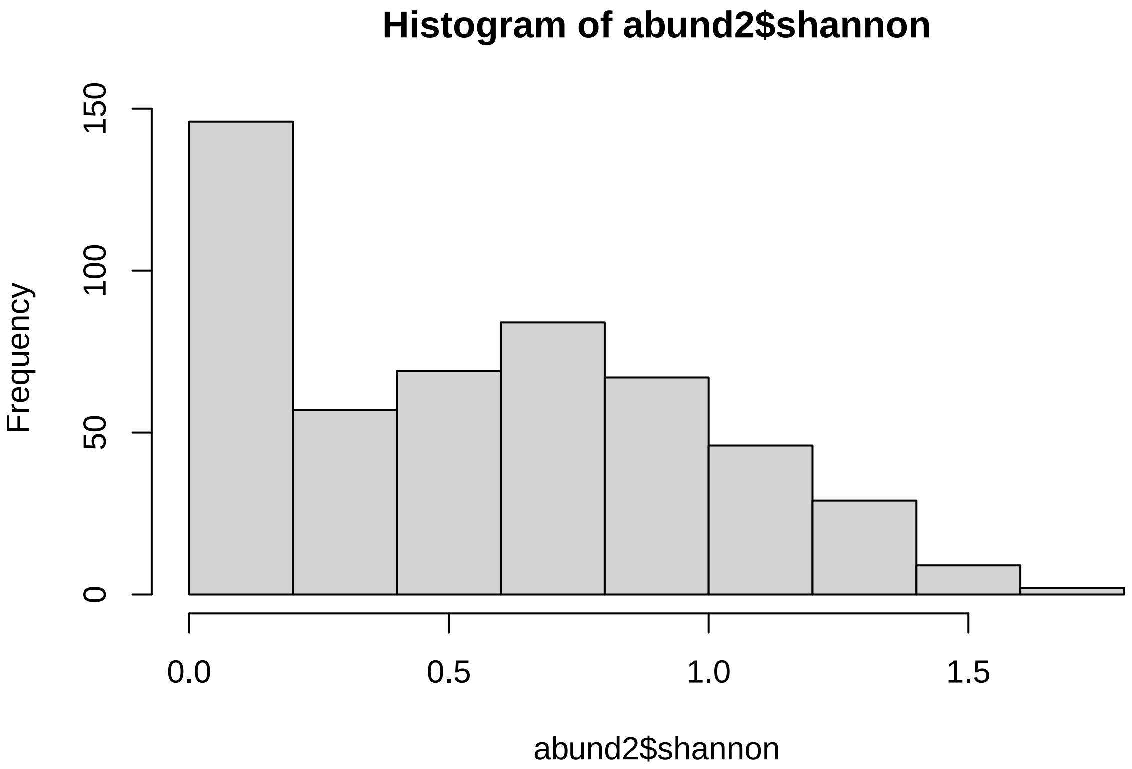
For a linear model, there are four major assumptions that must be met. Call your variables X (indepenent variable) and Y (dependent variable). These assumptions are:

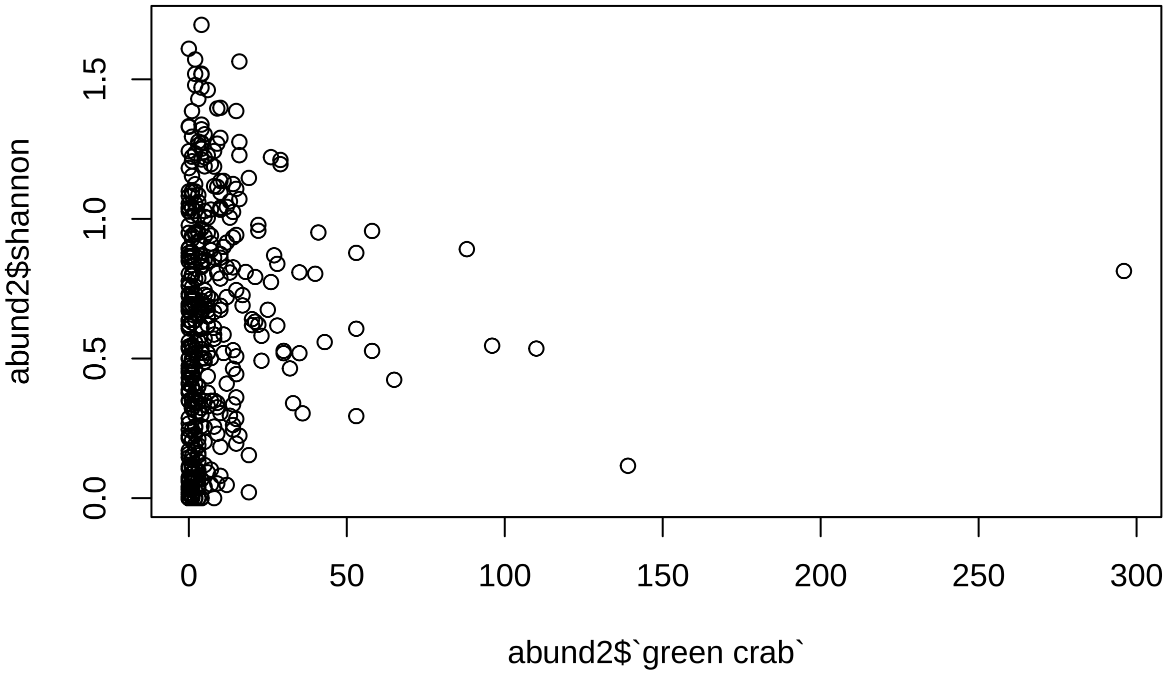
Linearity: The relationship between X and the mean of Y is linear. Homoscedasticity: The variance of residuals is the same for any value of X. Independence: Observations are independent of each other. Normality: For any fixed value of X, Y is normally distributed.

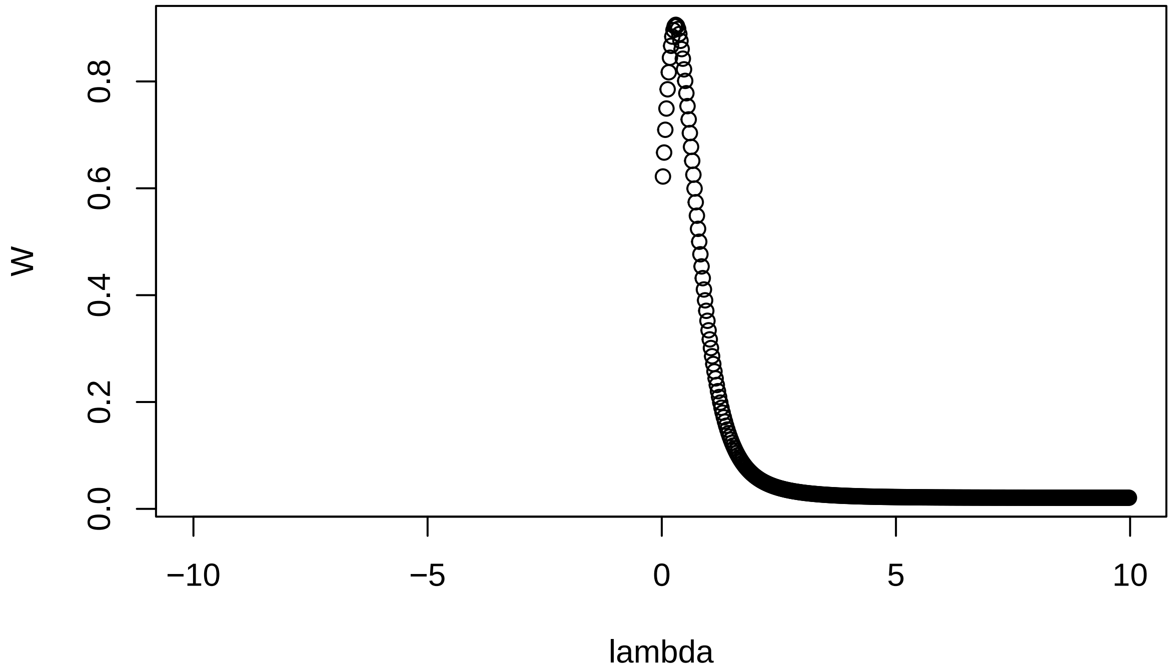
Let's check this now. The X variable will be the catch of green crabs and the Y variable will be the Shannon index.


abund2$shannon
0.0
0.5
1.0
1.5
0
50
100
150
200
250
300
abund2$`green crab`

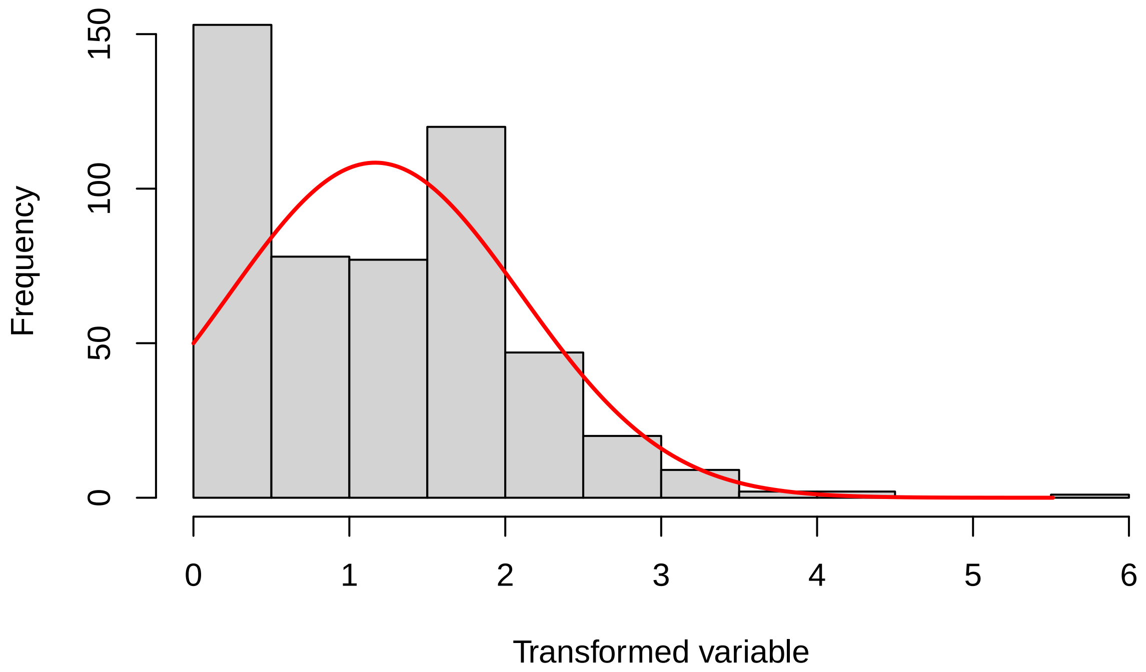

Histogram of abund2$`green crab`
Frequency
0
100
200
300
0
50
100
150
200
250
300
abund2$`green crab`

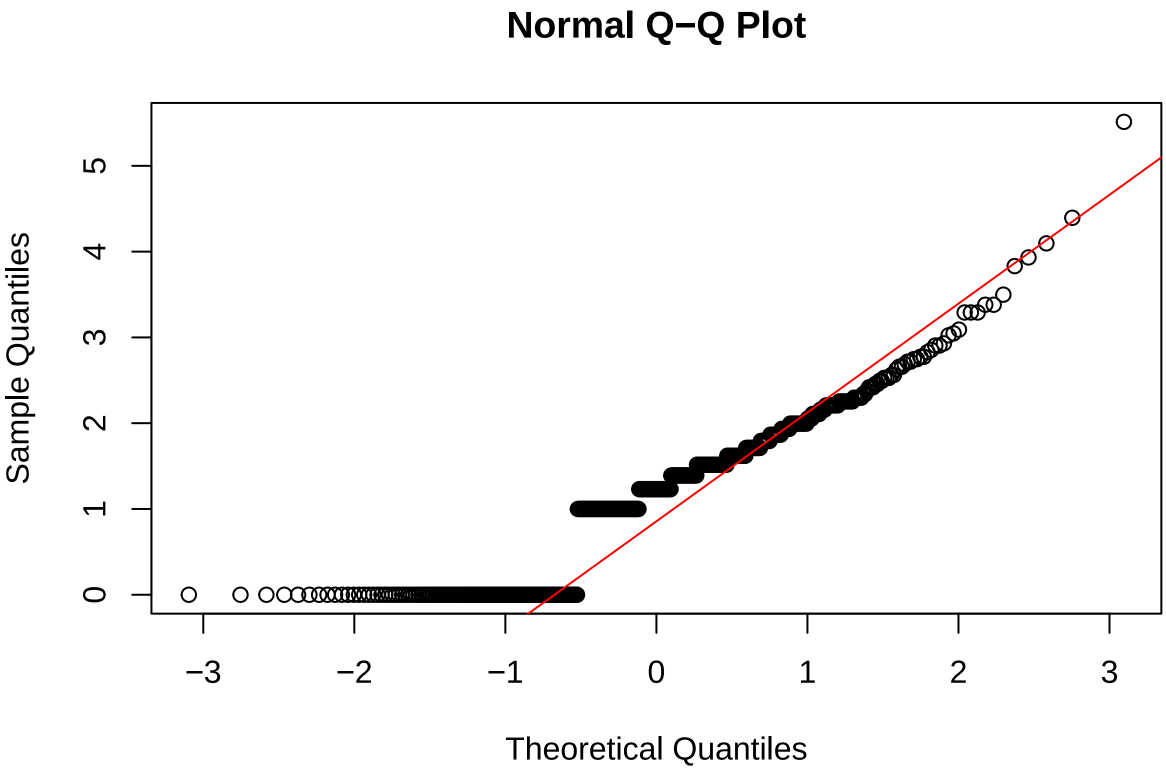

Histogram of abund2$shannon
Frequency
0
50
100
150
0.0
0.5
1.0
1.5
abund2$shannon

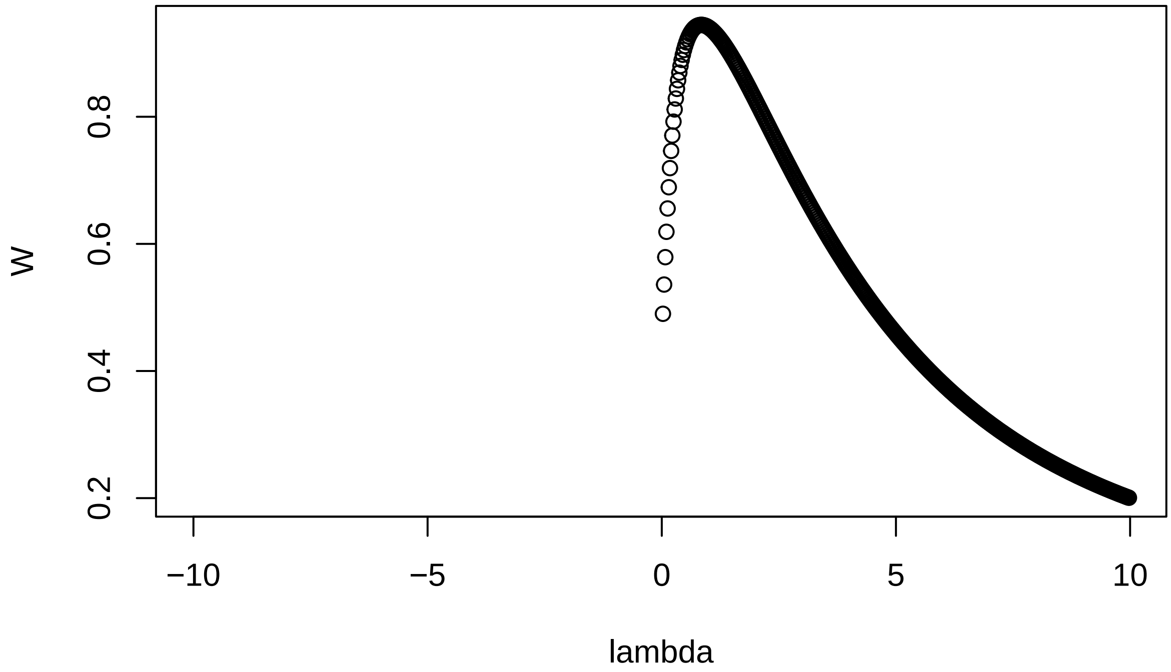

abund2$shannon
0.0
0.5
1.0
1.5
0
50
100
150
200
250
300
abund2$`green crab`



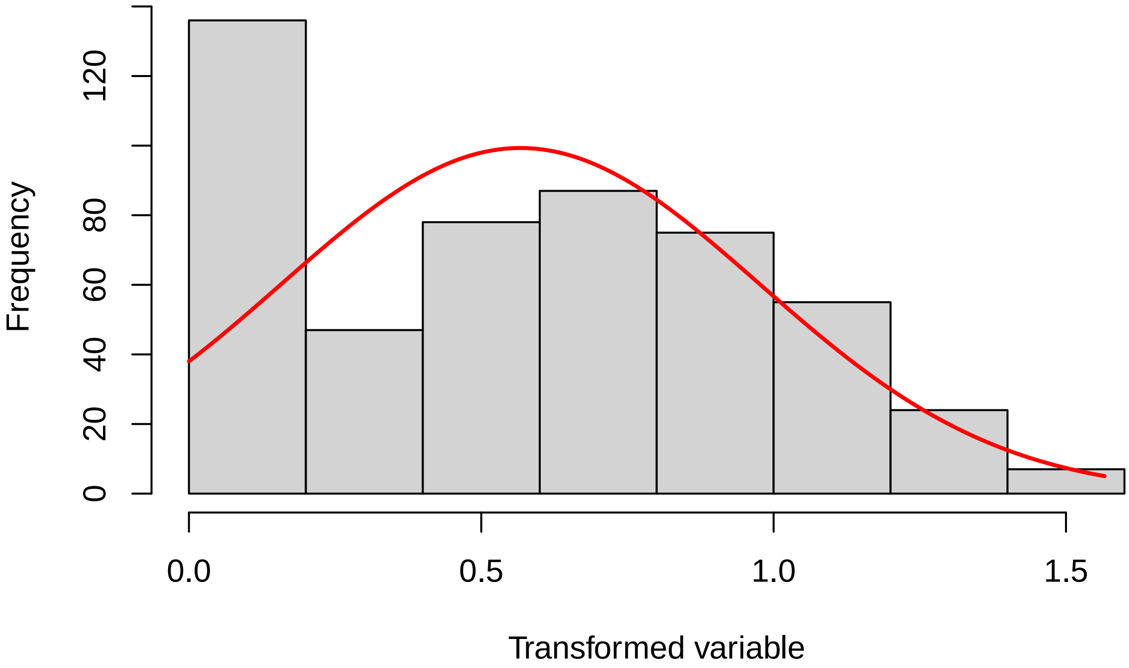
```
##
##     lambda      W Shapiro.p.value
## 413    0.3 0.9059       3.379e-17
##
## if (lambda >  0){TRANS = x ^ lambda}
## if (lambda == 0){TRANS = log(x)}
## if (lambda <  0){TRANS = -1 * x ^ lambda}
```

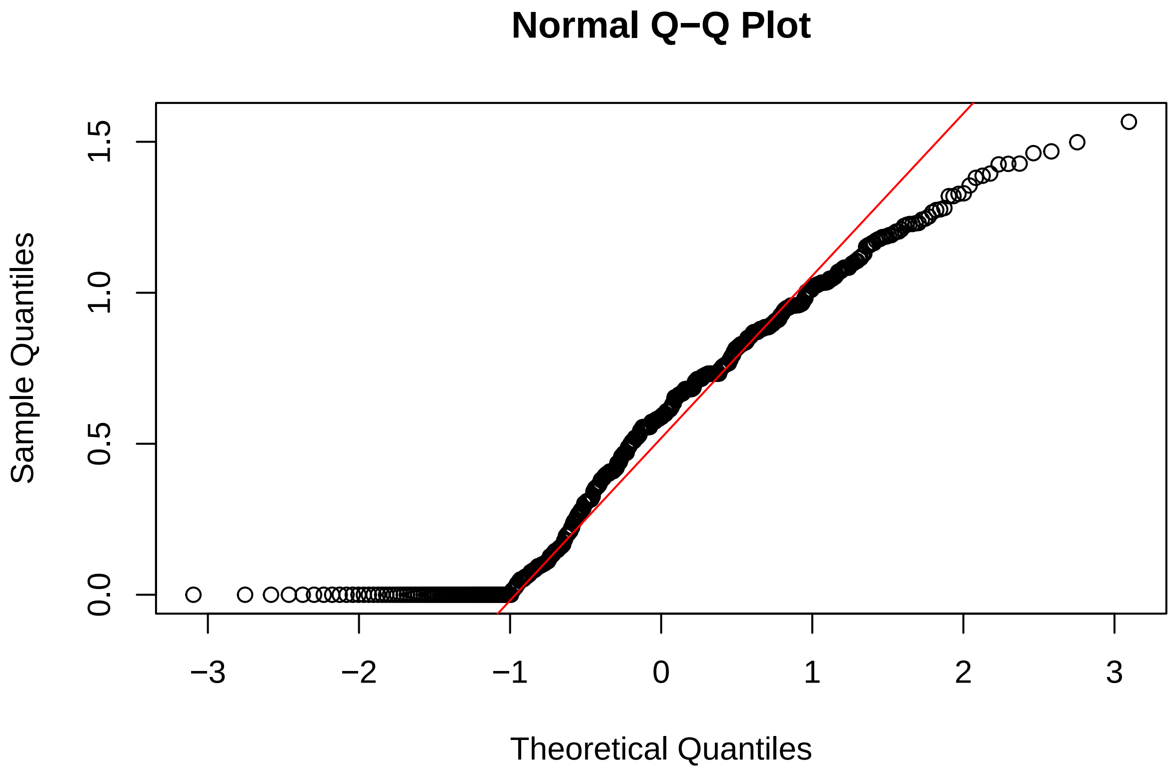

150
100
50
0
Frequency
0
1
2
3
4
5
6
Transformed variable

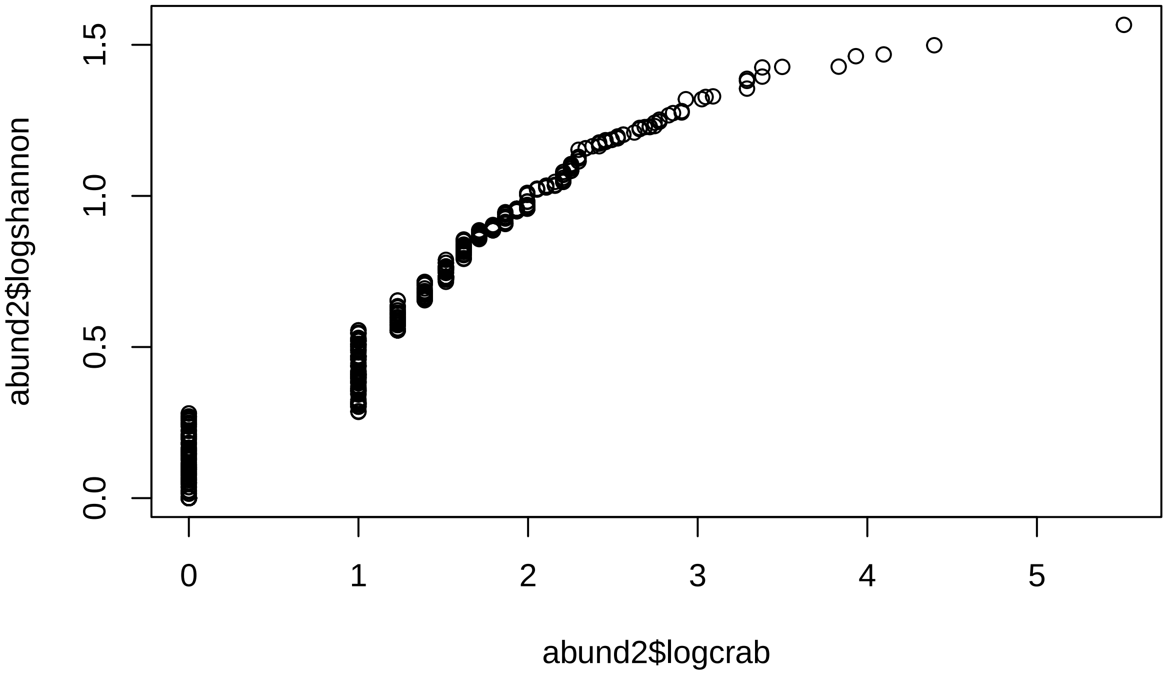

Normal Q−Q Plot
Sample Quantiles
0
1
2
3
4
5
Theoretical Quantiles
−3
−2
−1
0
1
2
3

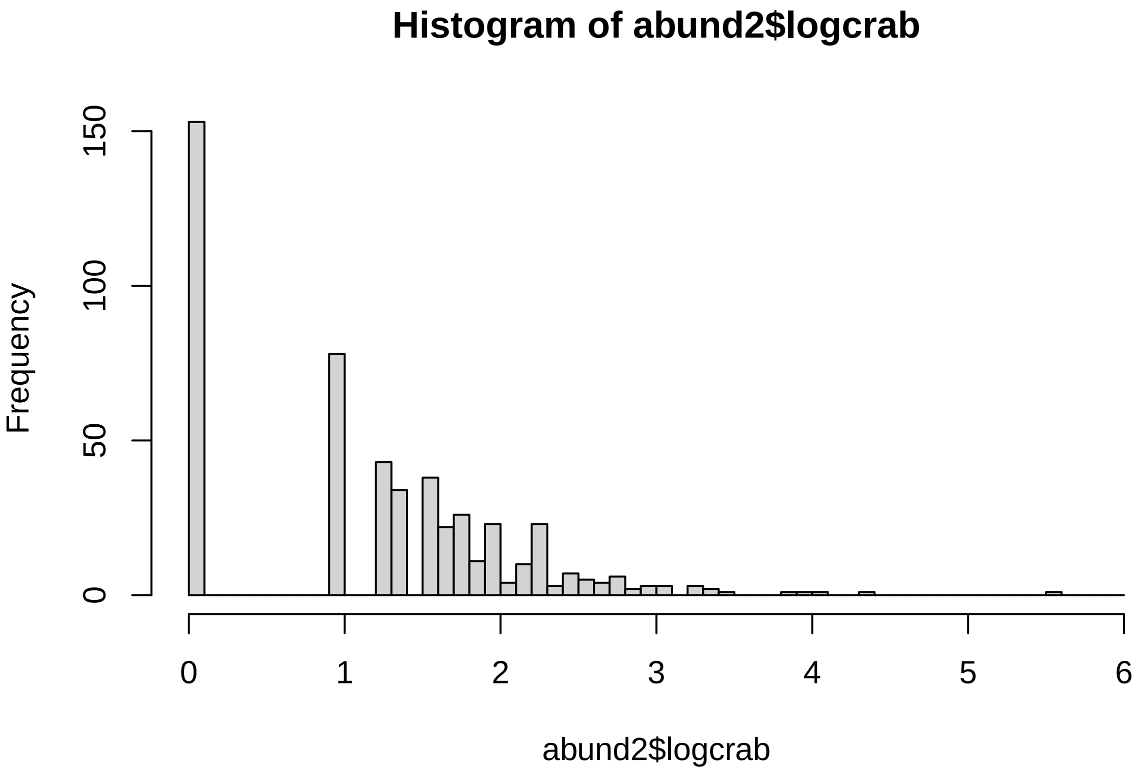


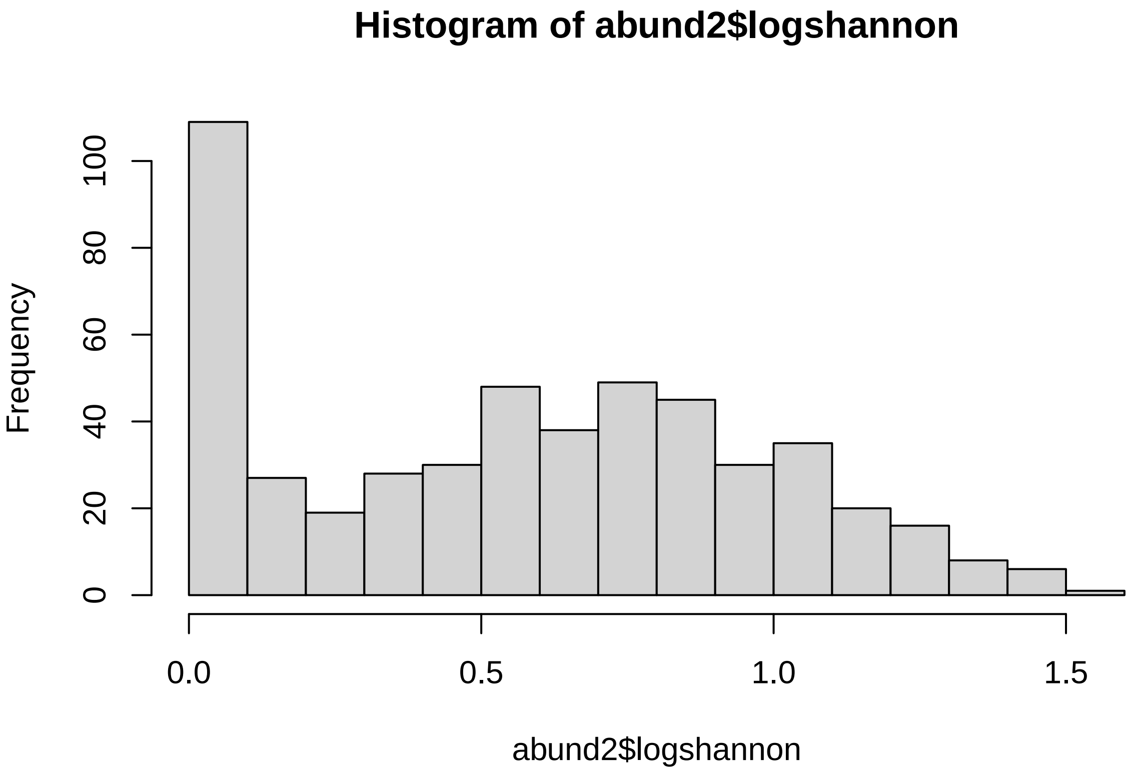
```
##
##     lambda      W Shapiro.p.value
## 435   0.85 0.9446       7.419e-13
##
## if (lambda >  0){TRANS = x ^ lambda}
## if (lambda == 0){TRANS = log(x)}
## if (lambda <  0){TRANS = -1 * x ^ lambda}
```

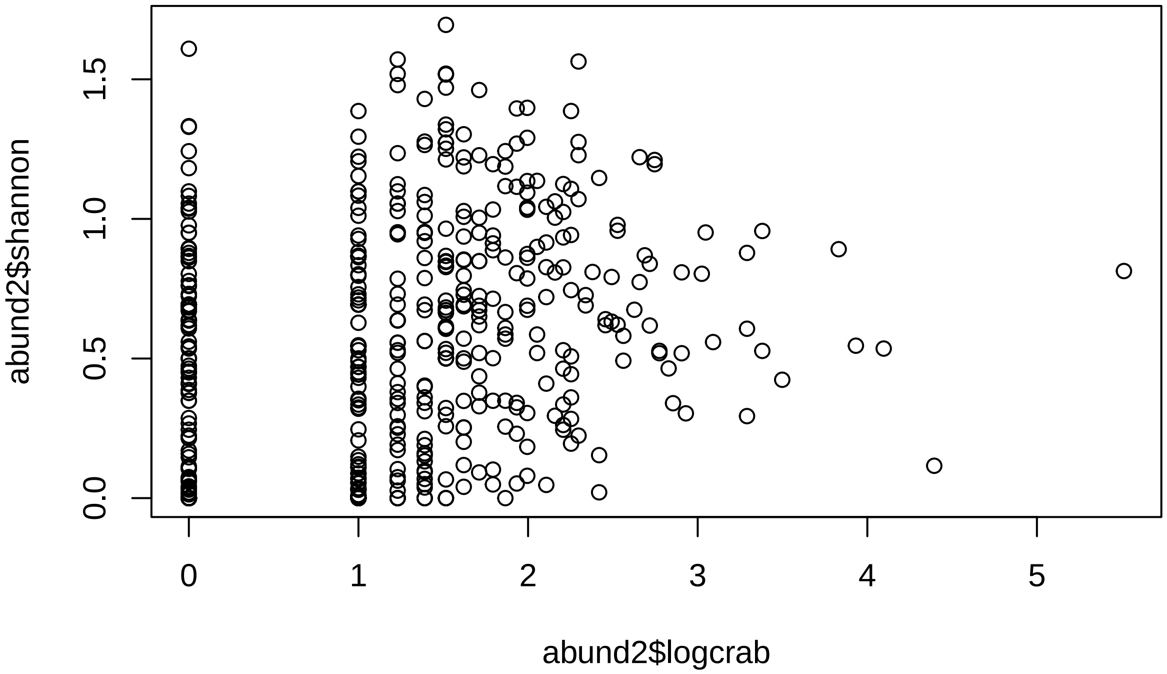

Frequency
0
20
40
60
80
120
0.0
0.5
1.0
1.5
Transformed variable


Normal Q−Q Plot
Sample Quantiles
0.0
0.5
1.0
1.5
Theoretical Quantiles
−3
−2
−1
0
1
2
3


abund2$logshannon
0.0
0.5
1.0
1.5
0
1
2
3
4
5
abund2$logcrab


Histogram of abund2$logcrab
Frequency
0
50
100
150
0
1
2
3
4
5
6
abund2$logcrab


Histogram of abund2$logshannon
Frequency
0
20
40
60
80
100
0.0
0.5
1.0
1.5
abund2$logshannon



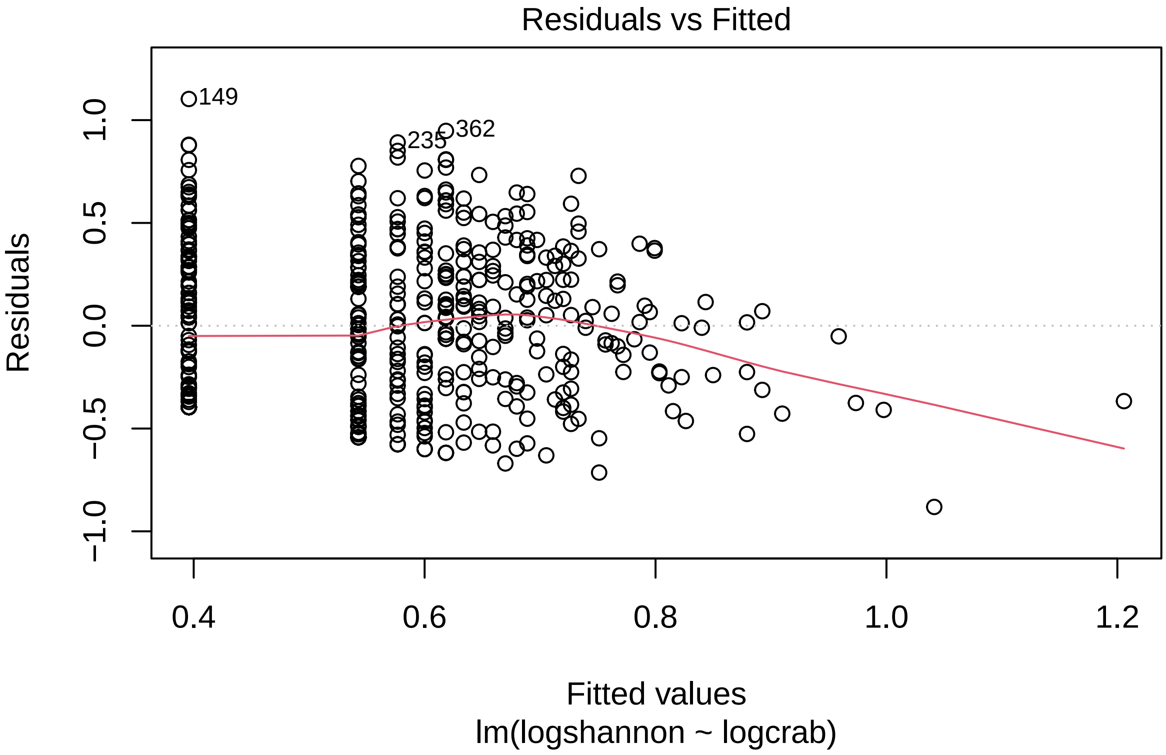
There are clearly violations of assumptions here, but we'll continue with the exercise anyway.

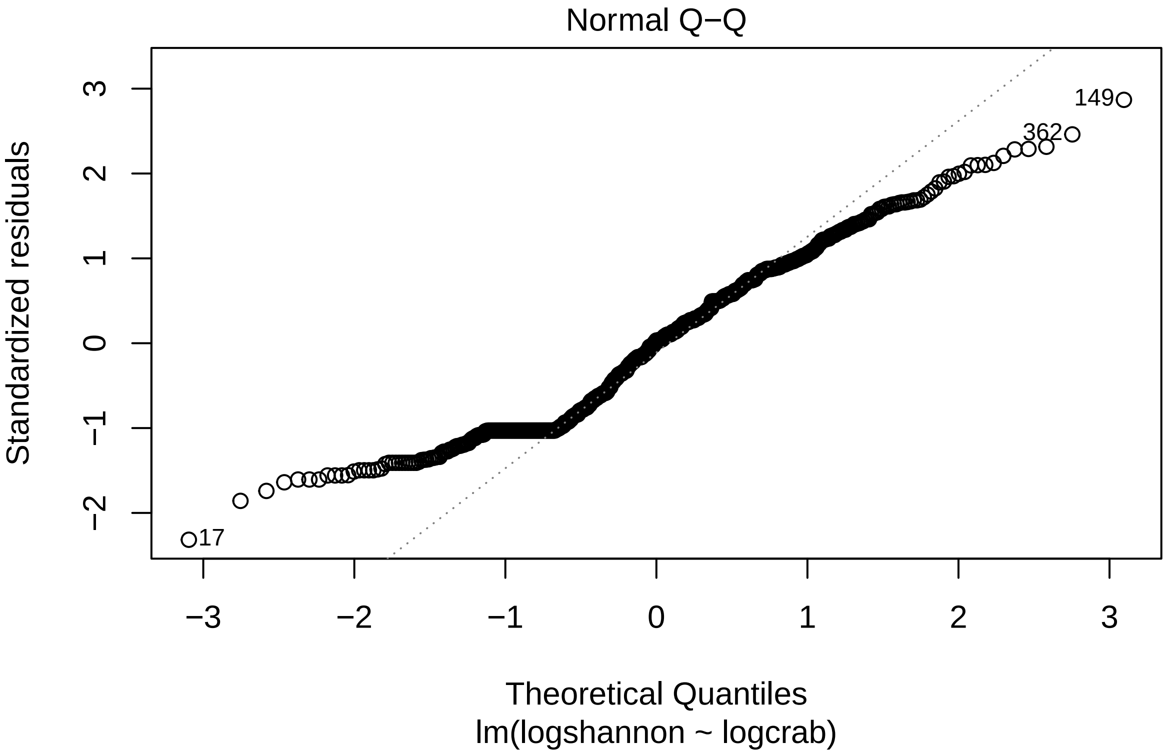
Linear models

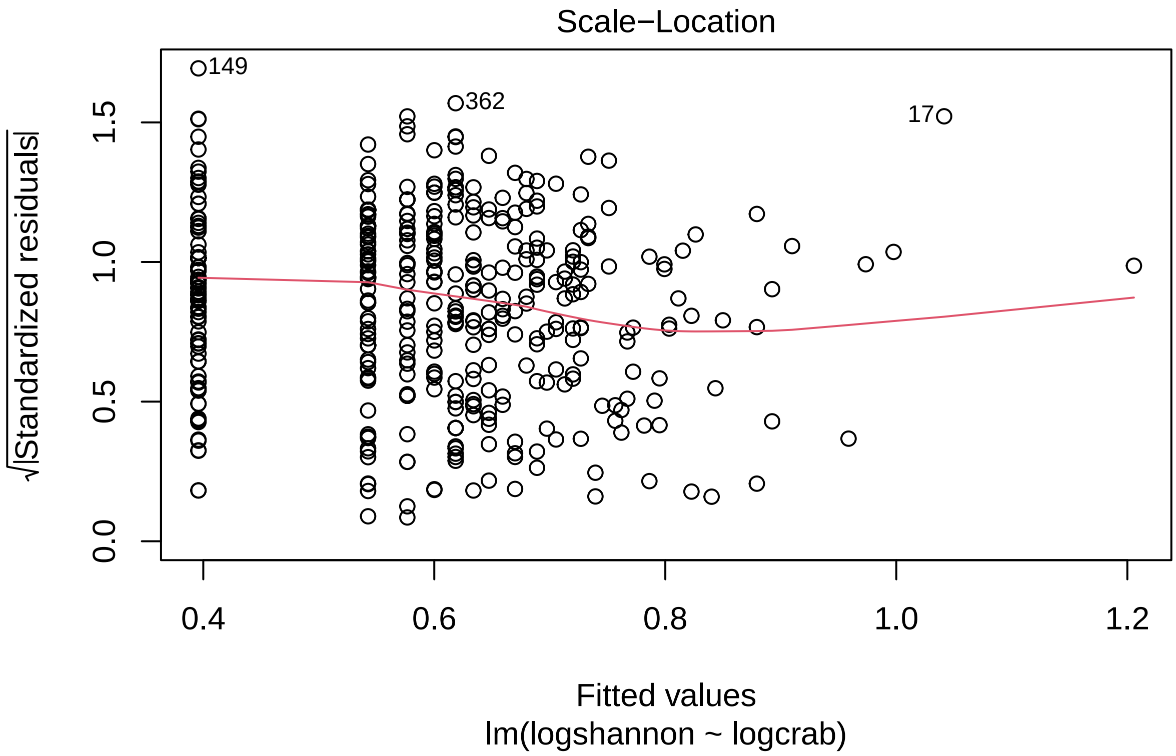
We'll use our transformed data to build a linear model that relates green crab abundance to Shannon diversity indices at each of our sites.

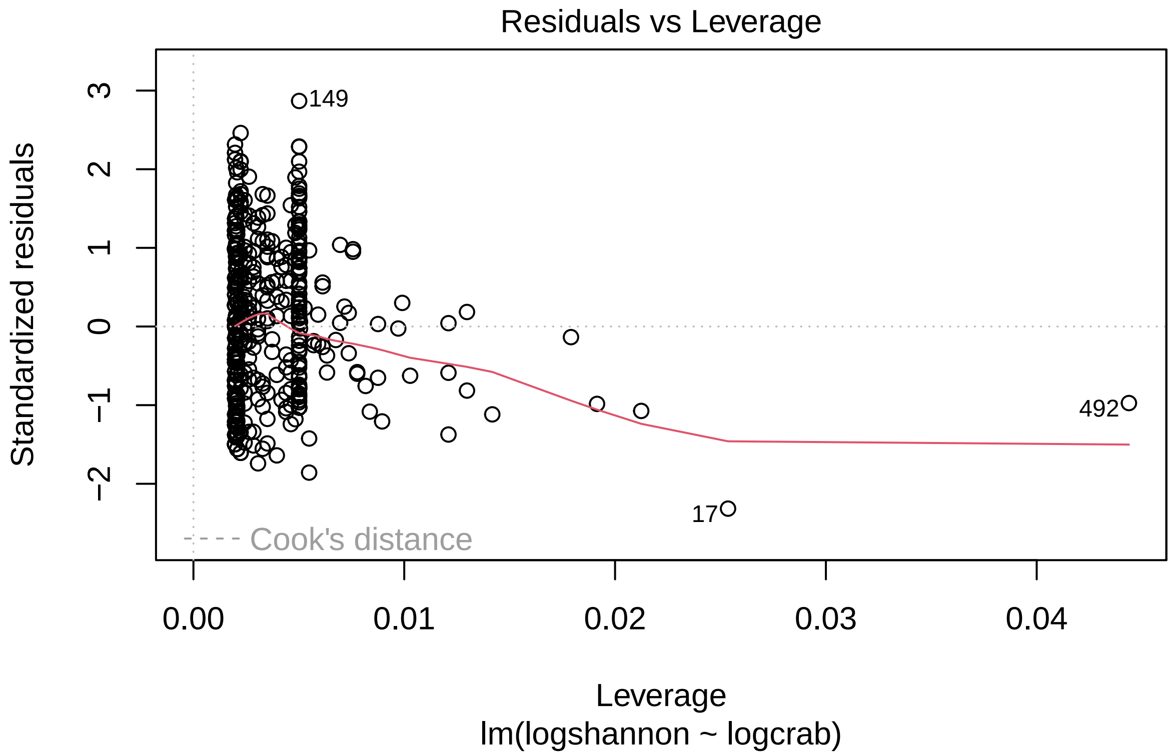
```
##
## Call:
## lm(formula = logshannon ~ logcrab, data = abund2)
##
## Coefficients:
## (Intercept)      logcrab
##      0.3958       0.1469
```

```
##
## Call:
## lm(formula = logshannon ~ logcrab, data = abund2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88132 -0.39579  0.01213  0.31221  1.10277
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.39579    0.02729  14.501  < 2e-16 ***
## logcrab      0.14691    0.01826   8.045 6.13e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3855 on 507 degrees of freedom
## Multiple R-squared:  0.1132, Adjusted R-squared:  0.1115
## F-statistic: 64.73 on 1 and 507 DF,  p-value: 6.132e-15
```




Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
149
362
17
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(logshannon ~ logcrab)

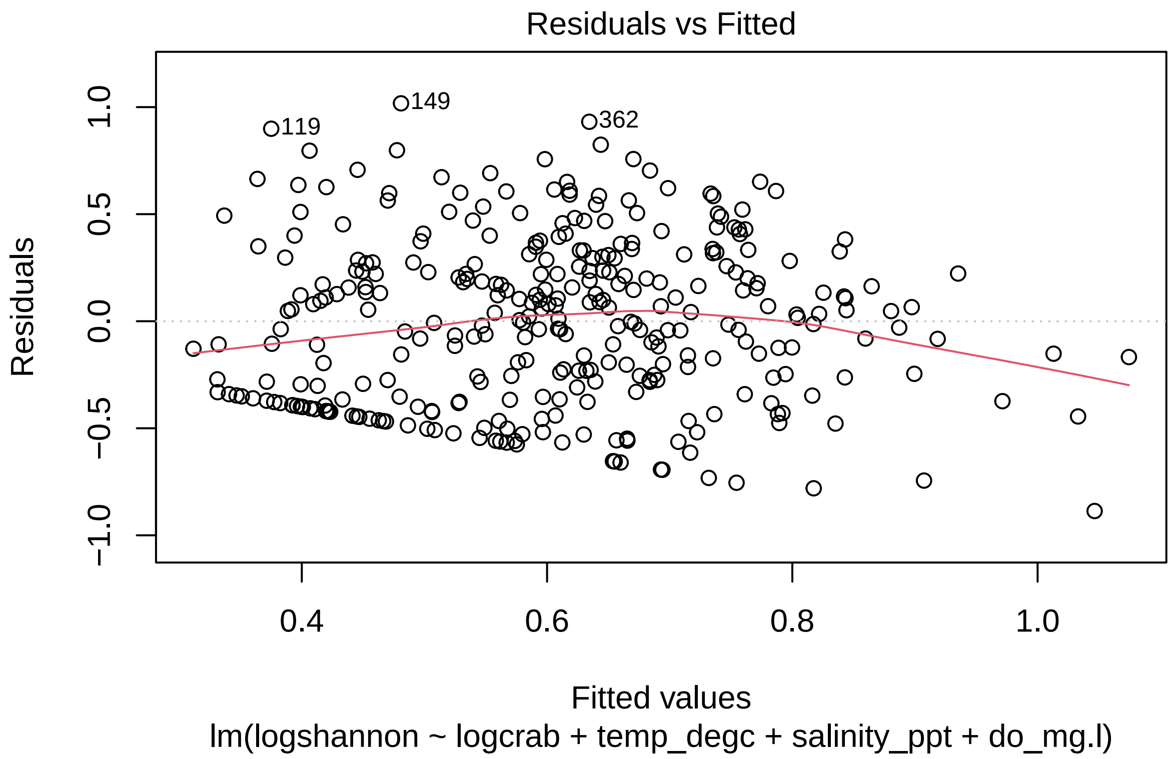

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
0.4
0.6
0.8
1.0
1.2
149
362
17
Fitted values
lm(logshannon ~ logcrab)

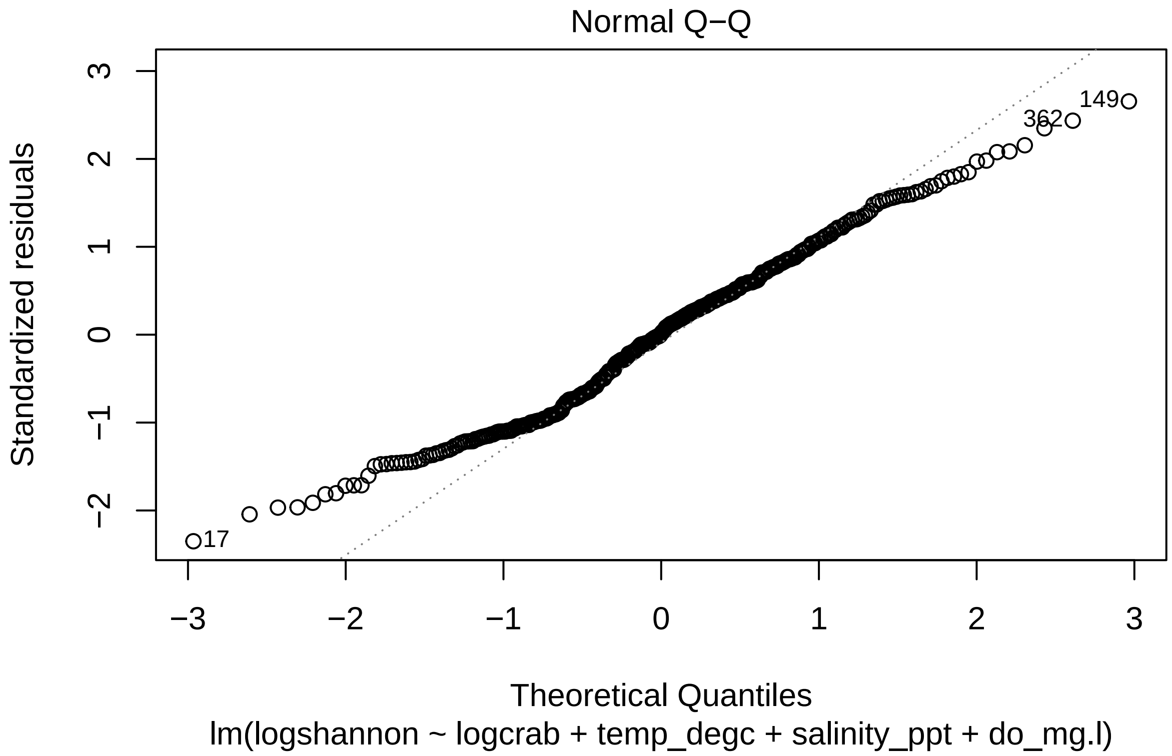


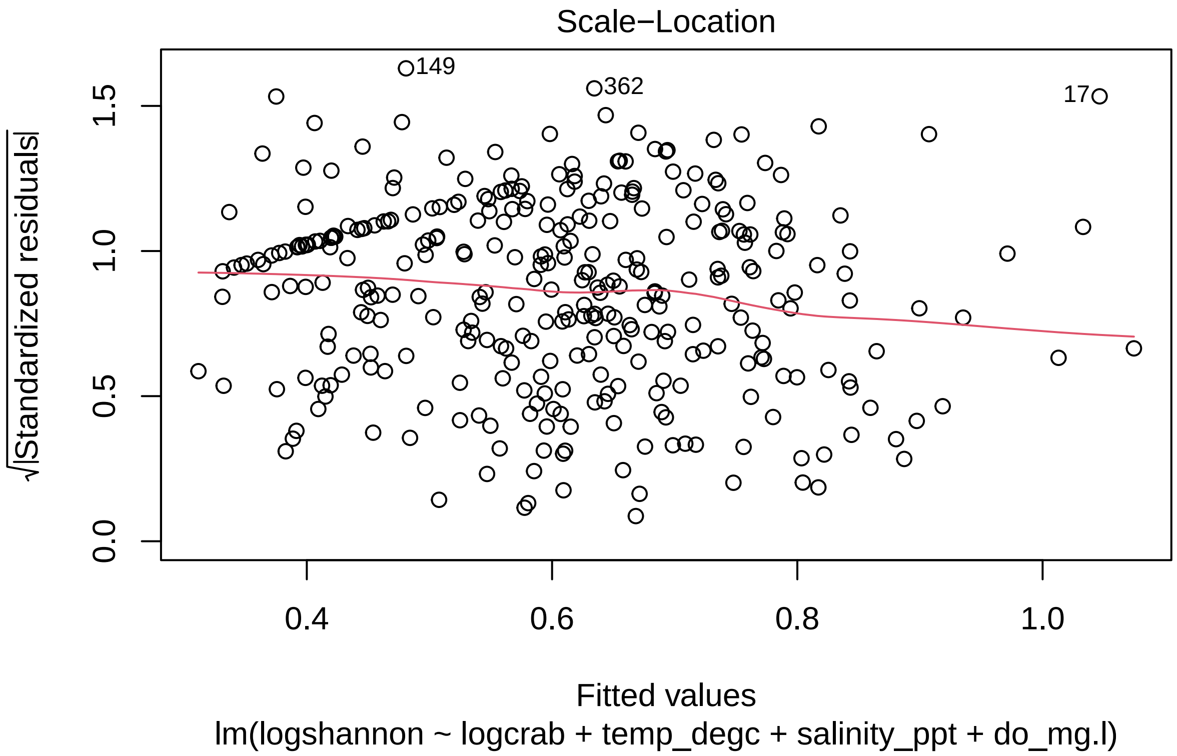
This is a bad model! We already knew the assumptions were likely to be violated due to the weird distributions of both the green crab catch and the Shannon diversity index. The model results indicate that green crab abundance explains only 11% of the variance in Shannon diversity index (this is the R-squared value). The plots show the residuals have unequal variance along the range of X values. Let's try adding a few more terms to see what happens.

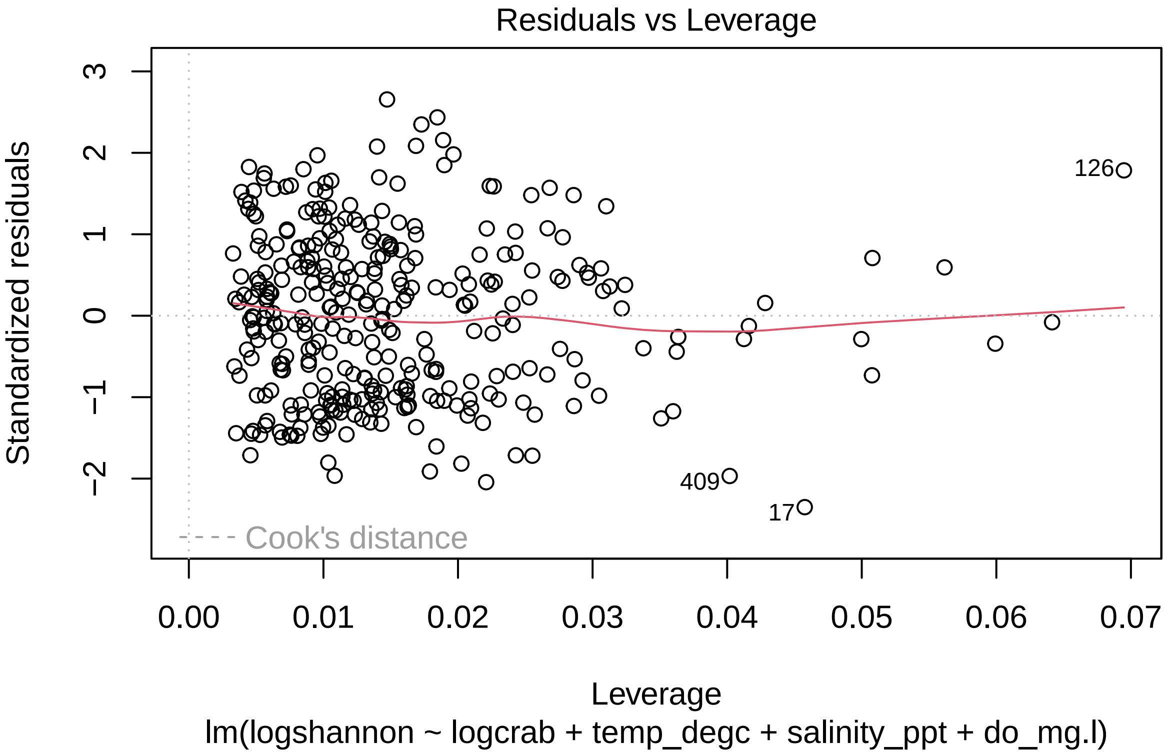
```
## 
## Call:
## lm(formula = logshannon ~ logcrab + temp_degc + salinity_ppt + 
##     do_mg.l, data = abund2)
## 
## Coefficients:
##  (Intercept)       logcrab     temp_degc  salinity_ppt       do_mg.l  
##    -0.014727      0.142658      0.008435     -0.001148      0.032394
```

```
## 
## Call:
## lm(formula = logshannon ~ logcrab + temp_degc + salinity_ppt + 
##     do_mg.l, data = abund2)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.88645 -0.34714  0.00511  0.27807  1.01766 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -0.014727   0.263207  -0.056 0.955415
## logcrab       0.142658   0.023949   5.957 6.67e-09 ***
## temp_degc     0.008435   0.009945   0.848 0.396982
## salinity_ppt -0.001148   0.002949  -0.389 0.697408
## do_mg.l       0.032394   0.009723   3.332 0.000962 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3861 on 326 degrees of freedom
##   (178 observations deleted due to missingness)
## Multiple R-squared:  0.1277, Adjusted R-squared:  0.117
## F-statistic: 11.93 on 4 and 326 DF,  p-value: 4.684e-09
```




Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(logshannon ~ logcrab + temp_degc + salinity_ppt + do_mg.l)
149
362
17


Scale−Location
149
362
17
√|Standardized residuals|
0.0
0.5
1.0
1.5
0.4
0.6
0.8
1.0
Fitted values
lm(logshannon ~ logcrab + temp_degc + salinity_ppt + do_mg.l)



That didn't help very much. In fact, the model summary indicates that temperature and salinity do not contribute anything of value to the model (note that the Pr(>|t|) column gives them both high t-scores, so they aren't significant). We can remove them. However, we didn't add much explanatory power to the model by using dissolved oxygen as another independent variable. Maybe this is because we do not have DO values for 176 of our 509 seine hauls.

At any rate, this is the rough process we will follow later on. We will use a more flexible model (Generalized Additive Model) that will relax some of the strict assumptions of a linear model. We will also test other potential covariates for inclusion.