Univariate Assignment

Read in tree data

```
# read in directly from website:
trees <- read.csv('https://raw.githubusercontent.com/dmcglinn/quant_methods/gh-pages/data/treedata_subset.csv')
```

1. **Carry out an exploratory analysis using the tree dataset. Metadata for the tree study can be found here. Specifically, I would like you to develop and compare models for species cover for a habitat generalist *Acer rubrum* (Red maple) and a habitat specialist *Abies fraseri* (Frasier fir).**

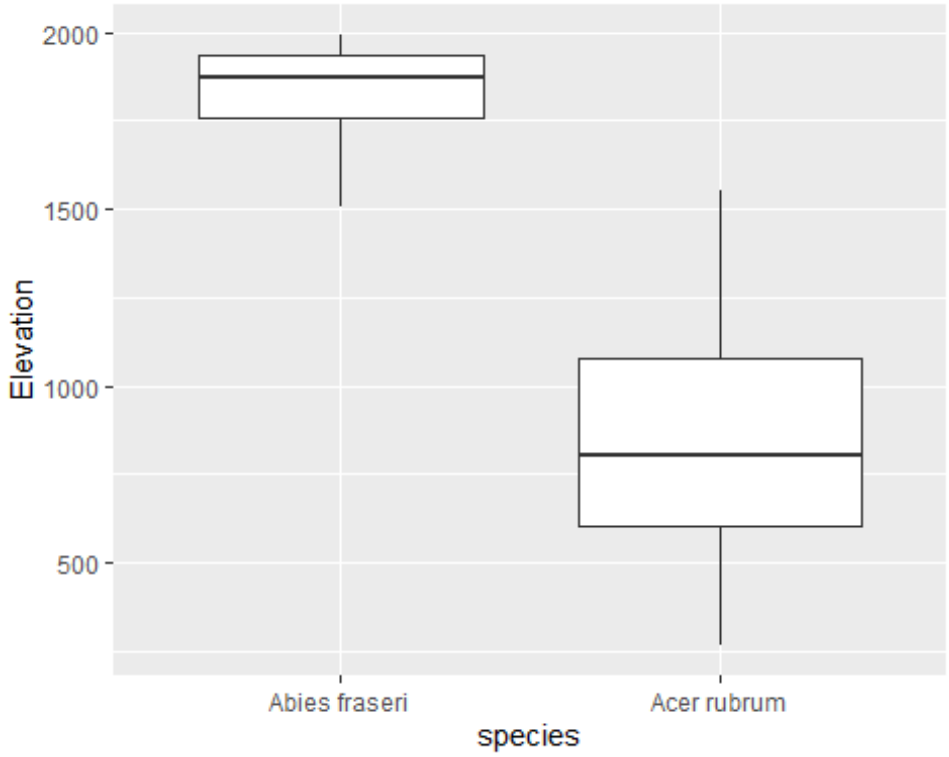
Restructure and subset the data using the following R code:

```
# we wish to model species cover across all sampled plots
# create site x sp matrix for two species
sp_cov <- with(trees, tapply(cover, list(plotID, spcode),
                             function(x) round(mean(x))))
sp_cov <- ifelse(is.na(sp_cov), 0, sp_cov)
sp_cov <- data.frame(plotID = row.names(sp_cov), sp_cov)
# create environmental matrix
cols_to_select <- c('elev', 'tci', 'streamdist', 'disturb', 'beers')
env <- aggregate(trees[ , cols_to_select], by = list(trees$plotID),
                 function(x) x[1])
names(env)[1] = 'plotID'
# merge species and enviornmental matrices
site_dat <- merge(sp_cov, env, by='plotID')
# subset species of interest
abies <- site_dat[ , c('ABIEFRA', cols_to_select)]
acer  <- site_dat[ , c('ACERRUB', cols_to_select)]
names(abies)[1] <- 'cover'
names(acer)[1] <- 'cover'
```

Visualize the data:

```
library(ggplot2)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(gridExtra)
```

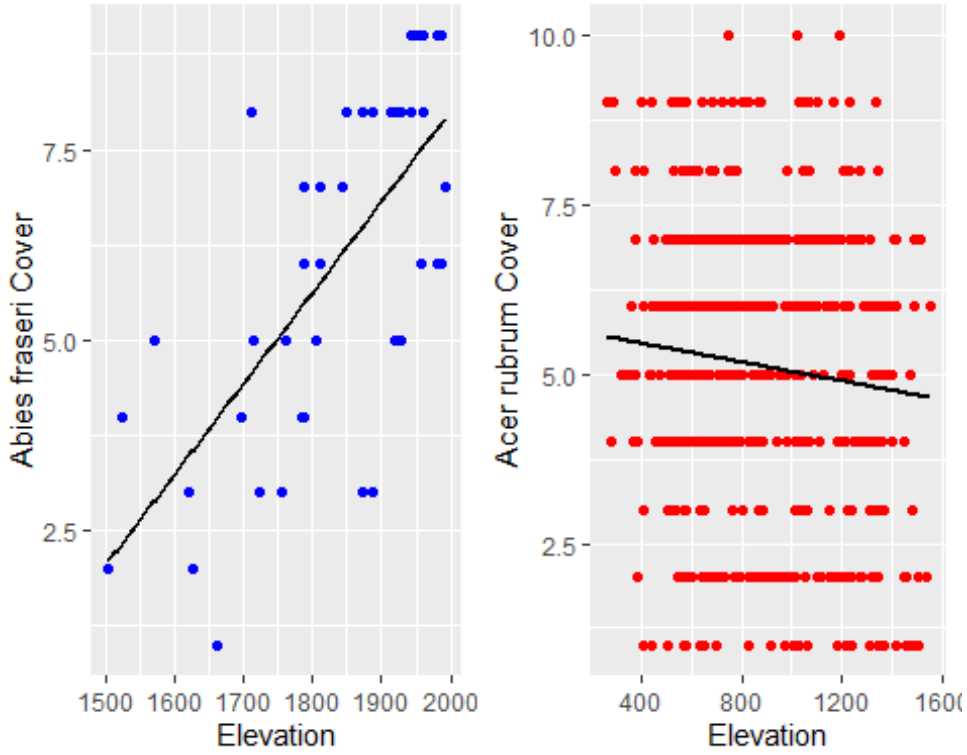
```
## Warning: package 'gridExtra' was built under R version 4.1.2

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

AbiesFraseri<-trees %>% filter(species=='Abies fraseri')
AcerRubrum<-trees %>% filter(species=='Acer rubrum')
BothSpecies<-rbind(AbiesFraseri,AcerRubrum)
#ggplot(BothSpecies, aes(x=cover, y=elev))+
  #geom_point(data=BothSpecies, aes(x=cover, y=elev, color=species))+
  #geom_smooth(data=BothSpecies, aes(linetype=species, color=species), method
=lm, fullrange=TRUE)
AbiesFraseri_plot<-ggplot(AbiesFraseri, aes(x=elev, y=cover))+
  geom_point(data=AbiesFraseri, aes(x=elev, y=cover), color='blue')+
  geom_smooth(data=AbiesFraseri, aes(linetype=species), method=lm, fullrange=
TRUE, se=FALSE, color='black')+
  theme(legend.position="none")+
  xlab("Elevation")+
  ylab("Abies fraseri Cover")
AcerRubrum_plot<-ggplot(AcerRubrum, aes(x=elev, y=cover))+
  geom_point(data=AcerRubrum, aes(x=elev, y=cover), color='red')+
  geom_smooth(data=AcerRubrum, aes(linetype=species), method=lm, fullrange=TR
UE, se=FALSE, color='black')+
  theme(legend.position="none")+
  xlab("Elevation")+
  ylab("Acer rubrum Cover")
SpecElev_plot<-ggplot(BothSpecies, aes(x=species,y=elev))+
  geom_boxplot()+
  ylab("Elevation")
print(SpecElev_plot)
```




```
grid.arrange(AbiesFraseri_plot,AcerRubrum_plot, ncol=2)
```

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

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



The slopes for each species appear different in the above graph, so and Interaction effect model will be used (the species being evaluated seems to influence the variation in cover with elevation)

Create and evaluate model:

```
trees_int_mod <- lm(cover ~ species + elev + species:elev,
              data = BothSpecies)
summary(trees_int_mod)

##
## Call:
## lm(formula = cover ~ species + elev + species:elev, data = BothSpecies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4433 -1.3303  0.5126  1.2497  5.1009
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -15.814670   4.252825  -3.719 0.000215 ***
## speciesAcer rubrum        21.546652   4.258800   5.059 5.28e-07 ***
## elev                       0.011914   0.002314   5.148 3.36e-07 ***
## speciesAcer rubrum:elev   -0.012613   0.002328  -5.419 8.06e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.998 on 763 degrees of freedom
## Multiple R-squared:  0.05294,    Adjusted R-squared:  0.04922
## F-statistic: 14.22 on 3 and 763 DF,  p-value: 5.053e-09
```

```
summary(aov(trees_int_mod))
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## species        1   32.8   32.85   8.231  0.00423 **
## elev           1   20.2   20.19   5.058  0.02479 *
## species:elev   1  117.2  117.17  29.360 8.06e-08 ***
## Residuals    763 3044.9    3.99
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.1.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.1.2
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
Anova(trees_int_mod, type=3)
```

```
## Anova Table (Type III tests)
##
## Response: cover
##               Sum Sq  Df F value    Pr(>F)
## (Intercept)    55.18   1  13.828  0.000215 ***
## species       102.15   1  25.597 5.276e-07 ***
## elev          105.75   1  26.499 3.358e-07 ***
## species:elev  117.17   1  29.360 8.063e-08 ***
## Residuals    3044.88 763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Compare the p-values you observe using the function `Anova` to those generated using `summary`.

The p-values generated by the Anova function and the summary function are the same. They both show significant variation between species (p=5.3* 10^-7), with elevation (p=3.4* 10^-7), and between species and elevation (p=8.1* 10^-8).

For each species address the following additional questions:

How well does the exploratory model appear to explain cover?

```
abies_int_mod <- lm(cover ~ elev, data=abies)
Anova(abies_int_mod, type=3)
```

```
## Anova Table (Type III tests)
##
## Response: cover
##             Sum Sq  Df F value    Pr(>F)
## (Intercept) 105.64   1  118.53 < 2.2e-16 ***
## elev        165.04   1  185.19 < 2.2e-16 ***
## Residuals   652.38 732
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
acer_int_mod <- lm(cover ~ elev, data=acer)
Anova(acer_int_mod, type=3)
```

```
## Anova Table (Type III tests)
##
## Response: cover
##             Sum Sq  Df F value    Pr(>F)
## (Intercept) 4278.1   1  664.95 < 2.2e-16 ***
## elev         832.8   1  129.44 < 2.2e-16 ***
## Residuals   4709.5 732
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Each model produces a significant result (p = <2.2*10^-16), indicating that within each species, elevation has a significant effect on cover.

Which explanatory variables are the most important?

```
abies_int_mod <- lm(cover ~ elev + tci + streamdist + disturb + beers, data=a
bies)
Anova(abies_int_mod, type=3)
```

```
## Anova Table (Type III tests)
##
## Response: cover
##             Sum Sq  Df  F value    Pr(>F)
## (Intercept)  59.64   1  70.4167 2.501e-16 ***
## elev         98.13   1 115.8739 < 2.2e-16 ***
## tci           2.11   1   2.4895   0.11505
## streamdist    3.98   1   4.6951   0.03057 *
## disturb      28.40   3  11.1771 3.545e-07 ***
## beers         1.50   1   1.7679   0.18406
## Residuals   614.85 726
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

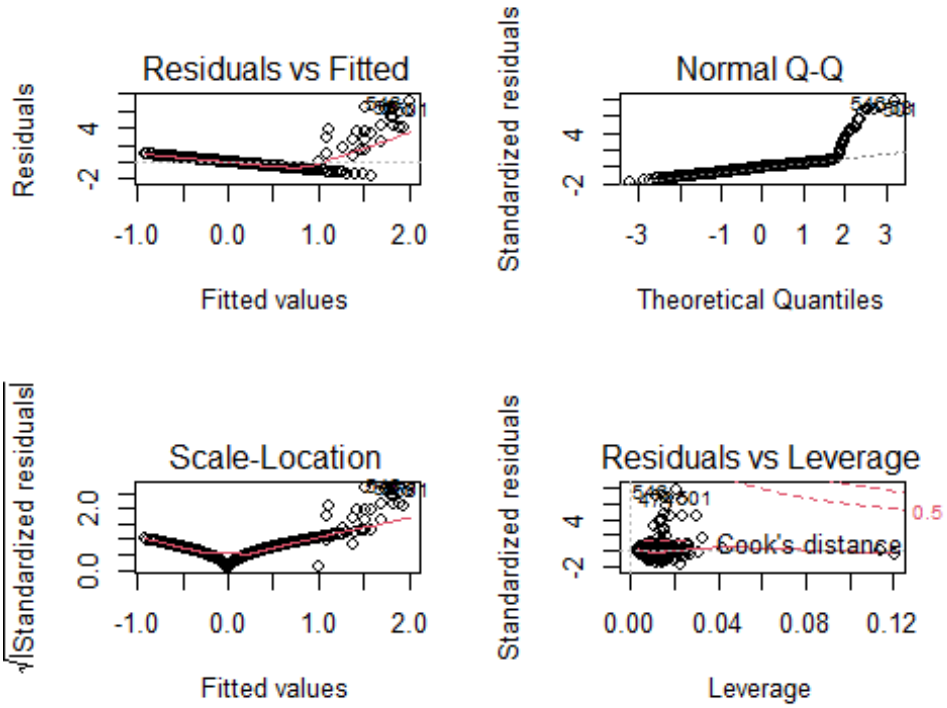
```
acer_int_mod <- lm(cover ~ elev + tci + streamdist + disturb + beers, data=ac
er)
Anova(acer_int_mod, type=3)
```

```
## Anova Table (Type III tests)
##
## Response: cover
##              Sum Sq  Df  F value    Pr(>F)
## (Intercept) 1845.7   1 295.0456 < 2.2e-16 ***
## elev         664.1   1 106.1624 < 2.2e-16 ***
## tci           55.8   1   8.9257  0.002907 **
## streamdist    10.8   1   1.7340  0.188316
## disturb       44.1   3   2.3479  0.071433 .
## beers         55.1   1   8.8144  0.003087 **
## Residuals   4541.7 726
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

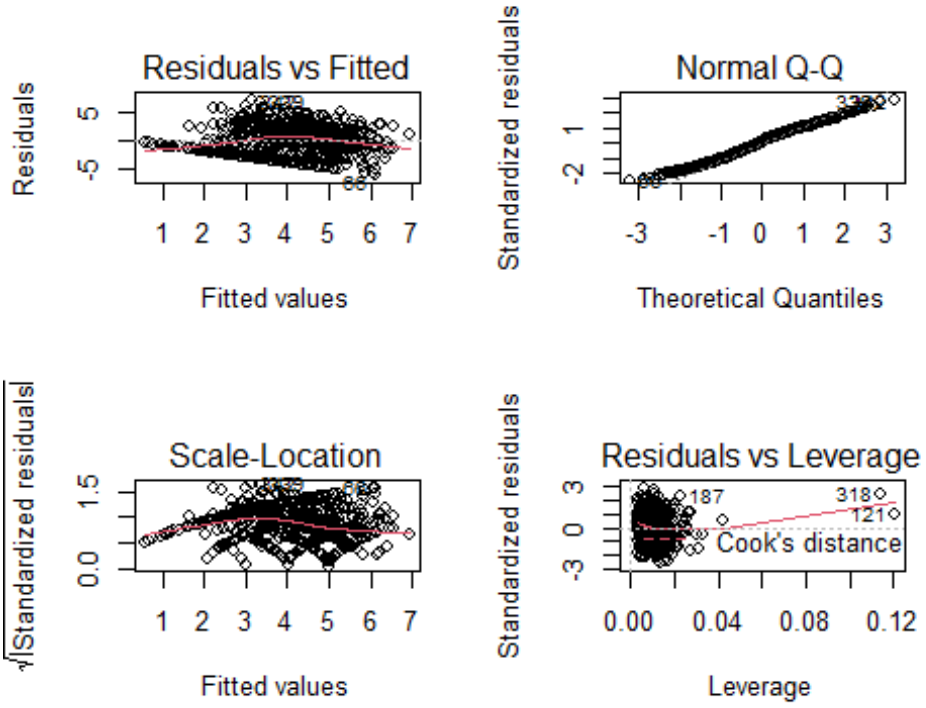
For Abies (habitat specialist), elevation has the greatest effect on cover (p = <2.2* 10^-16), followed by disturbance level (p = 3.5* 10^-7) and distance from streams (p = 0.03). For Acer (habitat generalist), elevation also has the greatest effect on cover (p = <2.2*10^-16), followed by site water potential (tci, p = 0.003) and heat load index (beers, p = 0.003).

Do model diagnostics indicate any problems with violations of OLS assumptions?

```
par(mfrow = c(2,2))
plot(abies_int_mod)
```




```
par(mfrow = c(2,2))
plot(acer_int_mod)
```



Based on the above plots, the Abies model violates assumptions of normality (top right graph deviates from linear arrangement) and homoscedasticity (top left and bottom left graphs have noticeable patterns). The bottom right plot also suggests influence of outliers. For the Acer model, assumption of normality holds and no large effect due to outliers. The homoscedasticity graphs are more spread out as compared to the Abies plots, but patters are still visible, so assumption of homoscedasticity is questionable. Overall, the Acer model holds to the assumptions more than the Abies model (there are far more acer samples in the dataset than abies samples, which may explain some of this).

Are you able to explain variance in one species better than another, why might this be the case?

```
nrow(subset(trees, species=="Abies fraseri"))

## [1] 44

nrow(subset(trees, species=="Acer rubrum"))

## [1] 723
```

I believe the discrepancy in sample size is the most likely culprit. There are only 44 samples for Abies fraseri (the model for which violated most assumptions), while there are 723 samples for Acer rubrum.

2. **You may have noticed that the variable cover is defined as positive integers between 1 and 10. and is therefore better treated as a discrete rather than continuous variable. Re-examine your solutions to the questions above but from**

the perspective of a General Linear Model (GLM) with a Poisson error term (rather than a Gaussian one as in OLS). The Poisson distribution generates integers 0 to positive infinity so this may provide a good first approximation. Your new model calls will look as follows:

```
acer_poi <- glm(cover ~ elev + tci + streamdist + disturb + beers, data = ace
r,
           family='poisson')
Anova(acer_poi, type=3)
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: cover
##            LR Chisq Df Pr(>Chisq)
## elev        171.549  1  < 2.2e-16 ***
## tci          14.411  1  0.0001469 ***
## streamdist    1.451  1  0.2283874
## disturb      12.521  3  0.0057946 **
## beers        12.175  1  0.0004844 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
abies_poi <- glm(cover ~ elev + tci + streamdist + disturb + beers, data = ab
ies,
           family='poisson')
Anova(abies_poi, type=3)
```

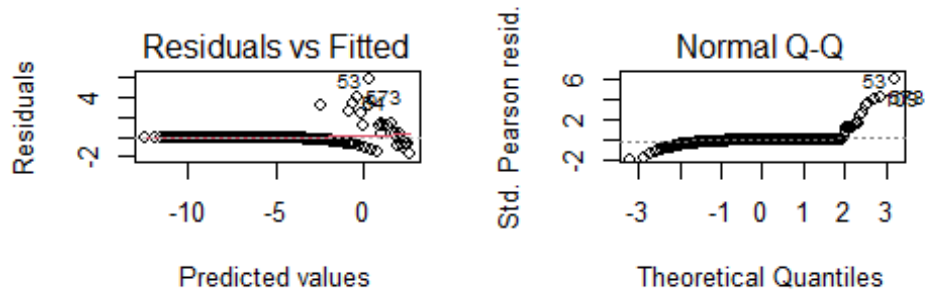
```
## Analysis of Deviance Table (Type III tests)
##
## Response: cover
##            LR Chisq Df Pr(>Chisq)
## elev         420.83  1  < 2.2e-16 ***
## tci            7.34  1   0.006742 **
## streamdist     7.63  1   0.005748 **
## disturb       21.89  3  6.863e-05 ***
## beers          0.01  1   0.904161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

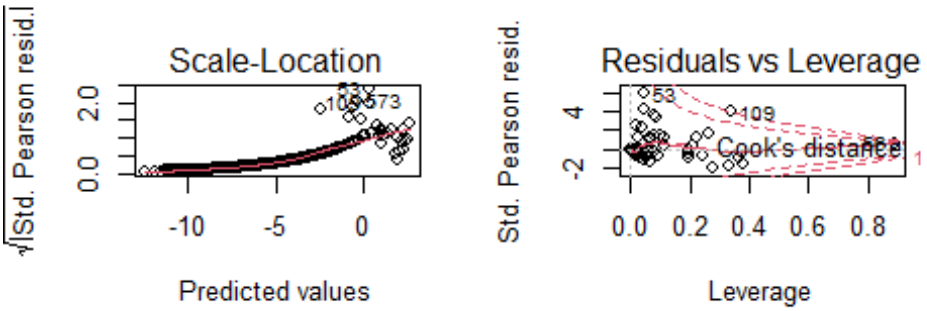
Which explanatory variables are the most important?

For Acer (habitat generalist), elevation has the greatest effect on cover (p = <2.2* 10^-16), followed by site water potential (tci, p = 0.0001), heat load index (beers, p = 0.0005), and disturbance level (p = 0.005). For Abies (habitat specialist), elevation also has the greatest effect on cover (p = <2.2* 10^-16), followed by disturbance level (p = 6.9* 10^-5), distance from stream (p = 0.006), and site water potential (tci, p = 0.007).

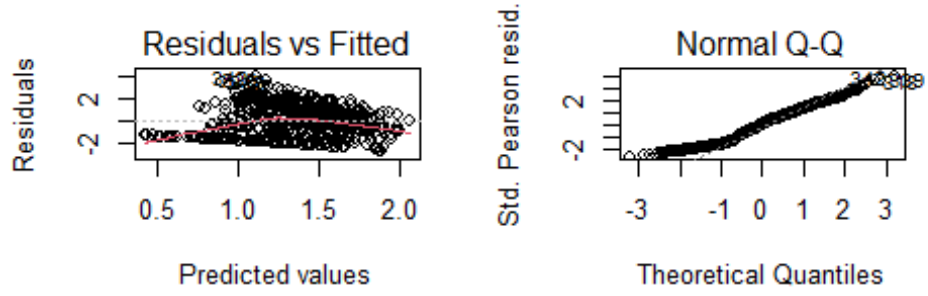
Do model diagnostics indicate any problems with violations of OLS assumptions?

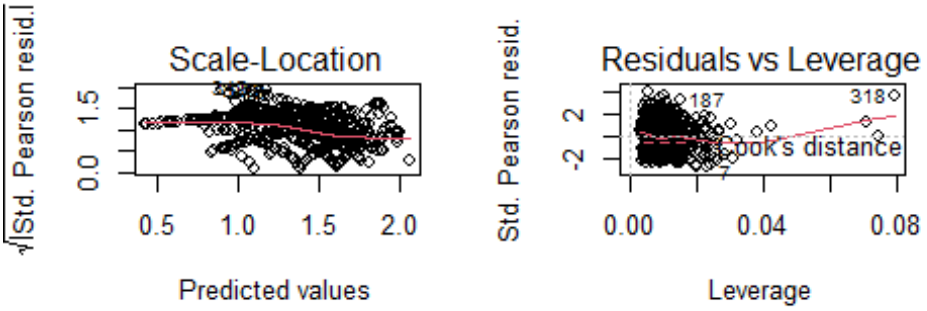
```
par(mfrow = c(2,2))
plot(abies_poi)
```






```
par(mfrow = c(2,2))
plot(acer_poi)
```





The assumption violations are similar to those seen in the original models, with Abies breaking most assumptions, while Acer holds true to most assumptions.

Are you able to explain variance in one species better than another, why might this be the case?

The number of Abies samples is still rather low in this new model, so I would still say this is the cause for some of the assumption issues.

For assessing the degree of variation explained you can use a pseudo-R-squared statistic (note this is just one of many possible)

```
pseudo_r2 <- function(glm_mod) {
                1 -  glm_mod$deviance / glm_mod$null.deviance
}
pseudo_r2(acer_poi)

## [1] 0.1342074

pseudo_r2(abies_poi)

## [1] 0.8951796
```

Compare your qualatitive assessment of which variables were most important in each model. Does it appear that changing the error distribution changed the results much? In what ways?

The levels of significance are more extreme in the glm models (significance is greater for variables in common, and some variables became significant in the glam model). This follows the difference between a gaussian (lm) and poisson (glm) distribution, where the poisson distribution is slightly more skewed right.

3. **Provide a plain English summary (i.e., no statistics) of what you have found and what conclusions we can take away from your analysis?**

Based on this data, elevation has a significant influence on cover for both species. In addition to elevation, Acer (habitat generalist) cover is influenced by site water potential, heat load index, and disturbance level, while Abies (habitat specialiest) cover is influenced by disturbance level, distance from stream, and site water potential. However, it is important to note that the Abies models did not meet the necessary assumptions. This is most likely due to the small number of Abies samples in the data set, as compared to the number of Acer samples.