R Univariate Models Homework 3

Caroline Tribble

```
#install.packages("ggplot2")       # for ggplot
#install.packages("gridExtra")     # for grid.arrange to arrange ggplots
#install.packages("scatterplot3d") # for scatterplot3d to make 3d graphic
#install.packages("MASS")          # for stepAIC to automate model selection

library(ggplot2)
library(gridExtra)
library(scatterplot3d)
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 3.6.2
```

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')
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec =
## dec, : embedded nul(s) found in input
```

Examine this dataset and see how the data is structured, see function str

```
str(trees)
```

```
## 'data.frame':    8025 obs. of  9 variables:
##  $ plotID    : Factor w/ 734 levels "ATBN-01-0303",..: 20 53 54 56 109 188 452 471 471 471 ...
##  $ spcode    : Factor w/ 52 levels "ABIEFRA","ACERNEG",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ species   : Factor w/ 52 levels "Abies fraseri",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ cover     : int  1 8 3 3 5 2 4 8 8 5 ...
##  $ elev      : num  1660 1712 1722 1754 1570 ...
##  $ tci       : num  5.7 3.82 3.89 3.15 11.85 ...
##  $ streamdist: num  491 454 453 492 0 ...
##  $ disturb   : Factor w/ 4 levels "CORPLOG","LT-SEL",..: 1 4 2 3 2 4 4 4 4 4 ...
##  $ beers     : num  0.224 0.834 1.333 1.471 0.496 ...
```

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). Because this dataset includes both continuous and discrete explanatory variables use the function Anova in the package car. This will estimate partial effect sizes, variance explained, and p-values for each explanatory variable included in the model.

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

For each species address the following additional questions:

a. How well does the exploratory model appear to explain cover?

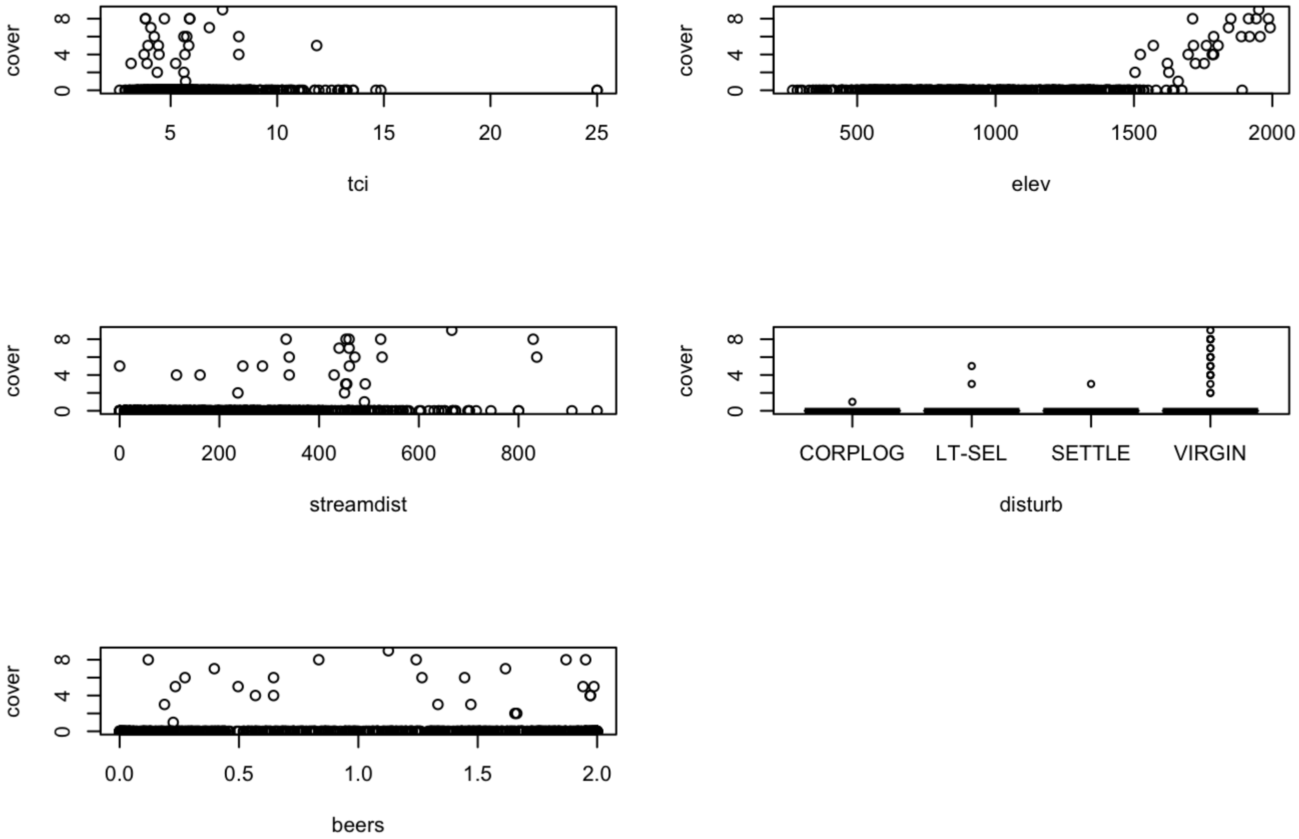
b. Which explanatory variables are the most important?

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

```
# 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'
```

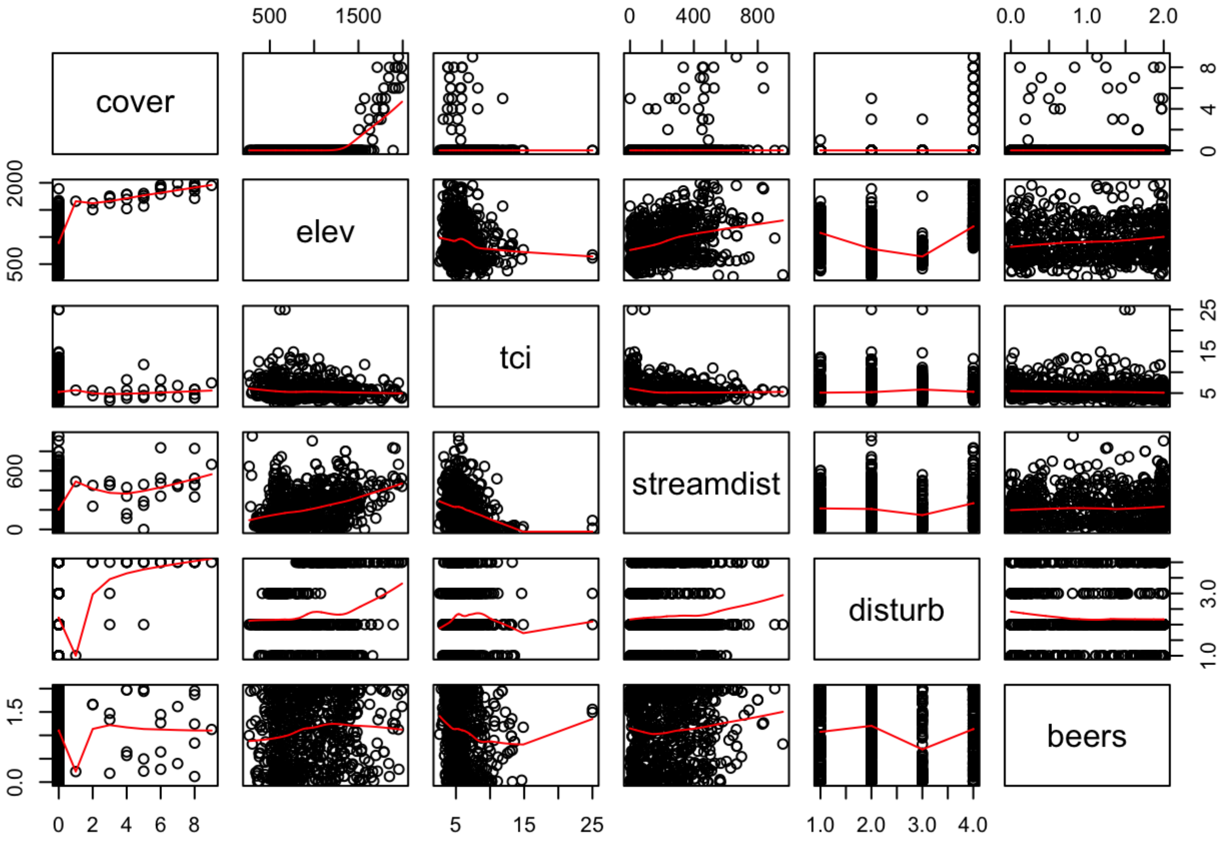
Frasier fir (Abies fraseri)

```
par(mfrow=c(3,2))
plot(cover ~ tci, data = abies)
plot(cover ~ elev, data = abies)
plot(cover ~ streamdist, data = abies)
plot(cover ~ disturb, data = abies)
plot(cover ~ beers, data = abies)
```



Looking at these plots showing the effect of site “water potential” (tci), elevation in meters (elev), distance in meters from nearest stream (streamdist), plot disturbance history (disturb), or heat load index (beers) on the local abundance as a percentage of horizontal cover (cover) of Frasier fir (Abies fraseri). Looking at these graphs it can be seen tci and elevation will be important factors to include in the model because cover shows a positive relationship with increasing elevation past >1500 meters and a negative relationship with increasing water site potential. Across all explanatory variables the response variable, cover of Frasier fir, is highly zero-inflated. This would make sense since we know this is a specialist species that appears to only occur at elevations >1500 meters and in areas with 0 - 15 tci. It is difficult to see any notable patterns in the effect of stream distance, disturbance, or heat load index on species cover and will be included in the model testing to be conservative.

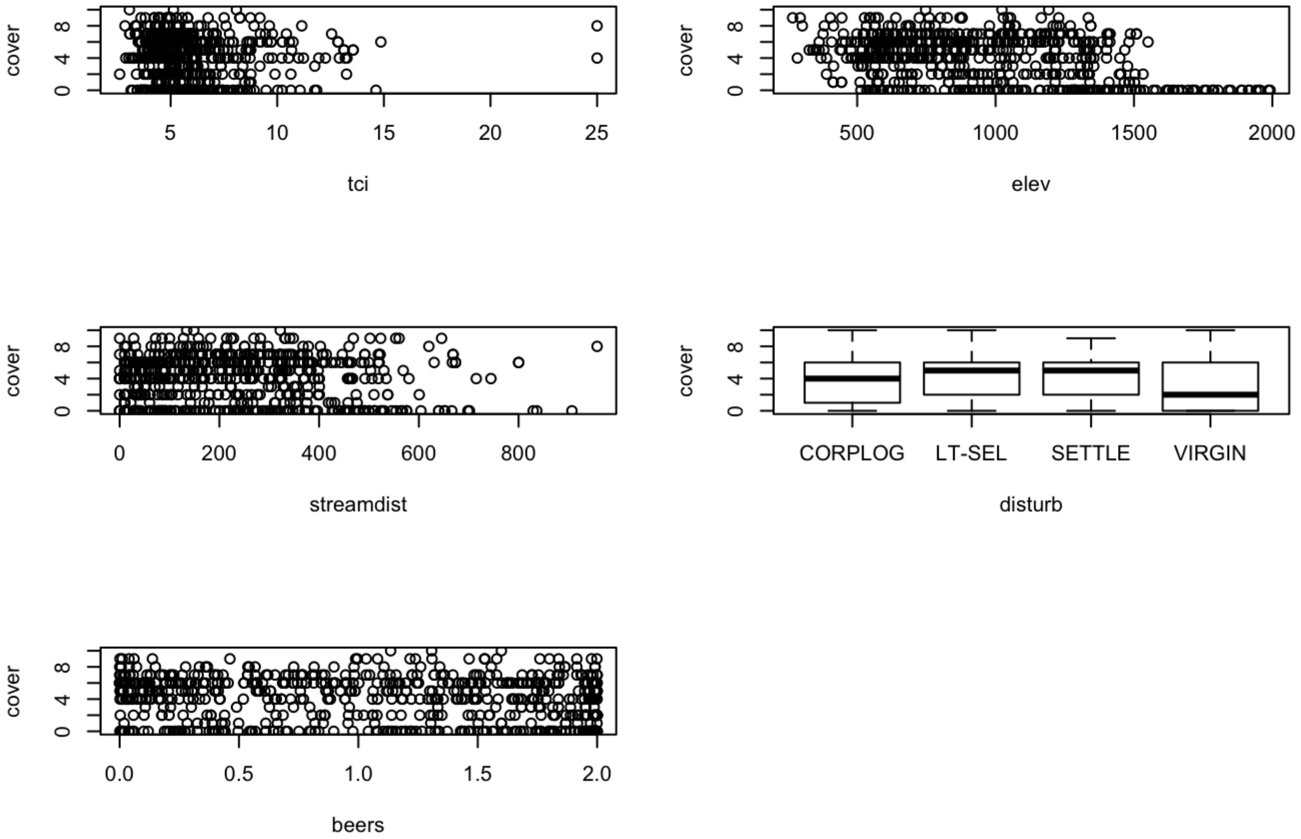
```
pairs(abies, panel= panel.smooth)
```



The pairs plot plots all the variables against one another. The first row of plots with the label ‘cover’ displays the effect of elev, tci, streamdist, disturb, and beers on species cover respectively. A line has also been fitted to each plot and it appears elevation has the strongest linear signal. The other plots show no linear relationship due to the large number of zeros in cover of frasier fir.

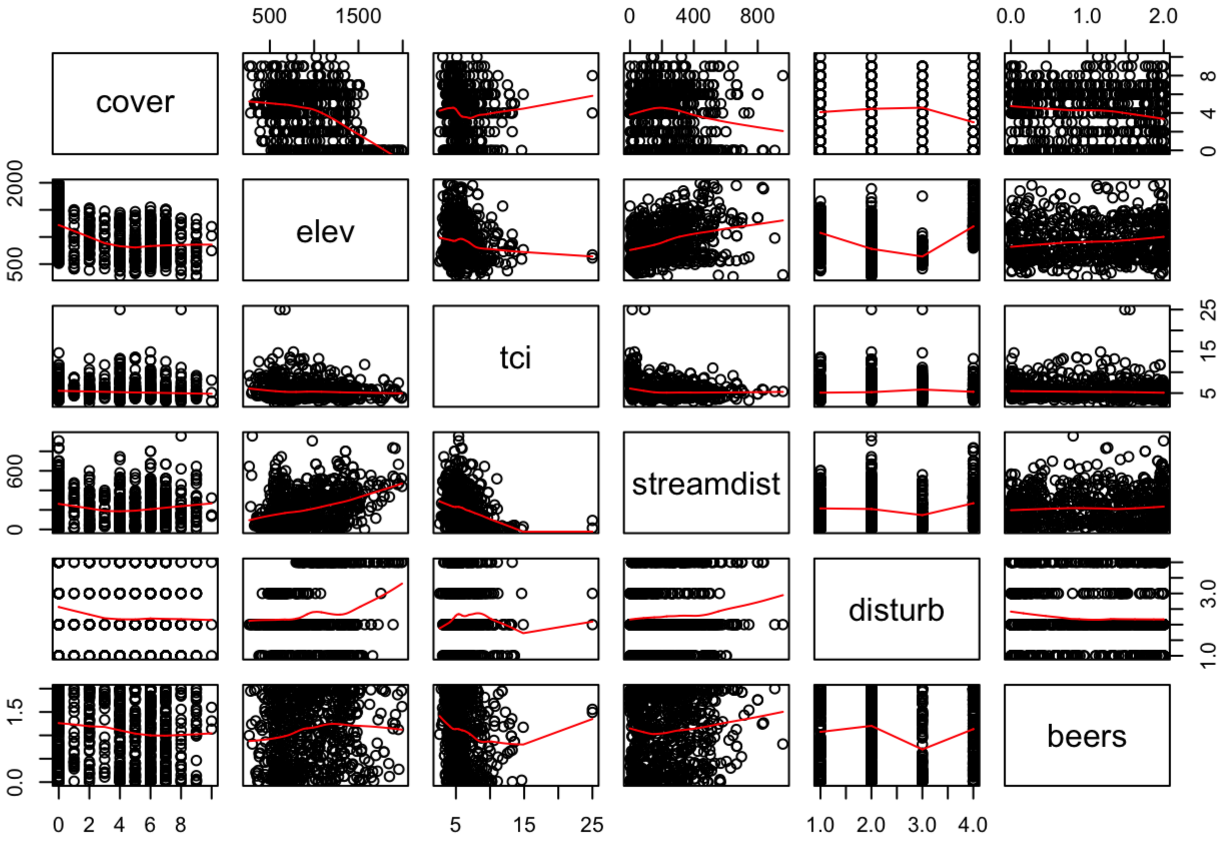
Red Maple (Acer Rubrum)

```
par(mfrow=c(3,2))
plot(cover ~ tci, data = acer)
plot(cover ~ elev, data = acer)
plot(cover ~ streamdist, data = acer)
plot(cover ~ disturb, data = acer)
plot(cover ~ beers, data = acer)
#lines(lowess(acer$beers,acer$cover), col ="red")
```



Similar to the plots above, we are observing the same effects on species cover for the red maple (Acer rubrum). Looking at these graphs it can be seen tci and elevation again will be important factors to include in the model because cover shows a negative relationship with increasing elevation only past > 1550 meters and a negative relationship with increasing water site potential. Unlike frasier fir, cover of the red maple is much more widely distributed across each explanatory variable. This would make sense since we know this is a generalist species. It is difficult to see any notable patterns in the effect of disturbanc or heat load index on species cover and will be included in the model testing to be conservative.

```
pairs(acer, panel= panel.smooth)
```



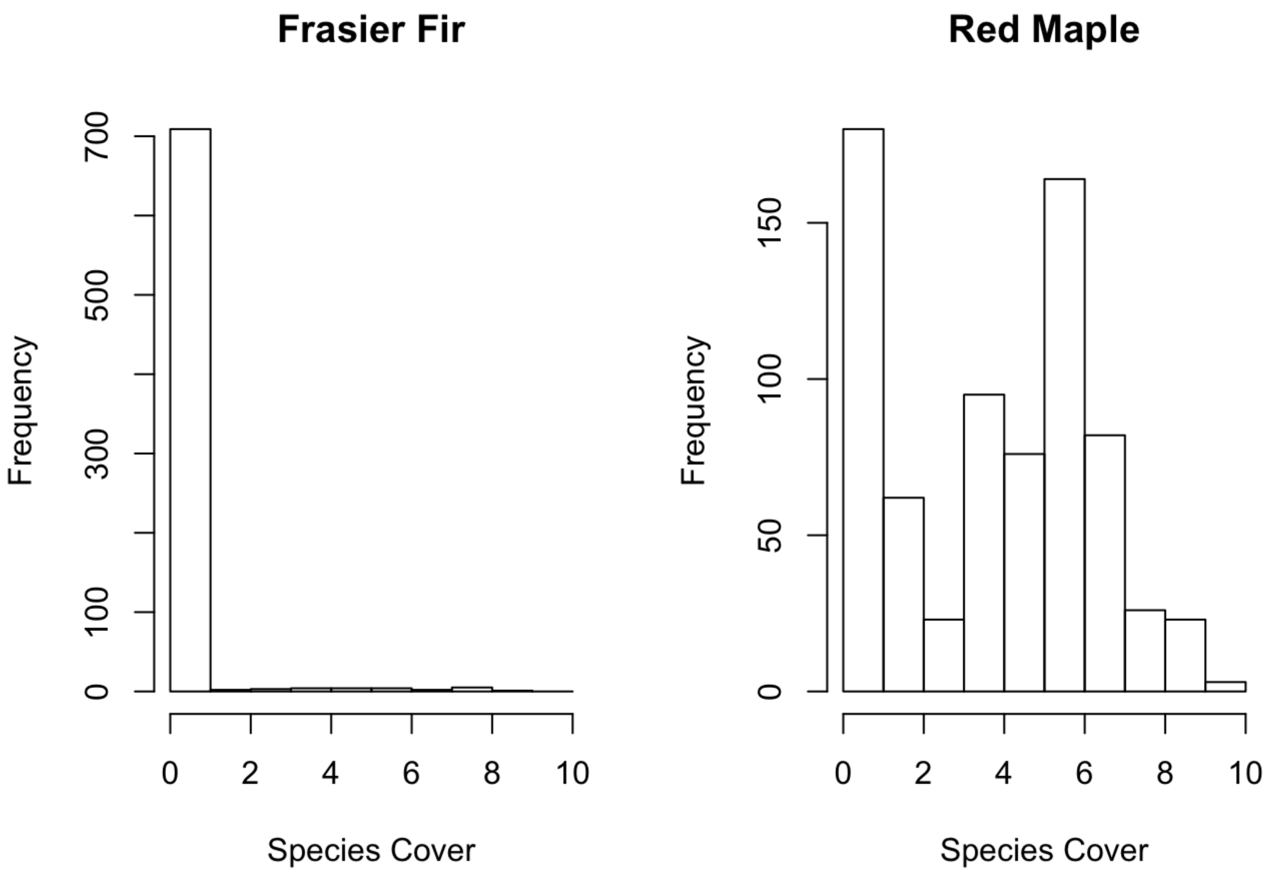
Similarly looking at the first row with the response variable cover it can be see elevation has the strongest signal and the other explanatory variables are a bit more difficult to interpret just looking at them.

```
sapply(abies, class)
```

```
##       cover        elev         tci  streamdist     disturb       beers
##   "numeric"   "numeric"   "numeric"   "numeric"    "factor"   "numeric"
```

Histograms to see how the frequency of our response variable, cover, is distributed for each species.

```
par(mfrow= c(1,2))
hist(abies$cover, breaks = seq(0,10, by =1), main = "Frasier Fir", xlab = "Species Cover")
hist(acer$cover, breaks = seq(0,10, by =1), main = "Red Maple", xlab = "Species Cover")
```



These histograms show again that species cover of the specalist tree species, frasier fir, is highly zero-inflated, which makes sense since it appears this species is endemic to a certain level of high elevation (> 1500 meters). The generalist species, red maple, follows a more bimodal distribution in cover type displaying much more variation than the specialist species.

```
library(car)
```

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

```
abies_mod <- aov(cover ~ elev + tci + streamdist + as.factor(disturb) + beers, data = abies)
Anova(abies_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 *  
## as.factor(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
```

```
summary(abies_mod)
```

```
##                     Df Sum Sq Mean Sq F value   Pr(>F)    
## elev                 1  165.0  165.04 194.881  < 2e-16 ***
## tci                  1    1.7    1.73   2.044   0.1533    
## streamdist           1    5.2    5.19   6.130   0.0135 *  
## as.factor(disturb)   3   29.1    9.71  11.460 2.39e-07 ***
## beers                1    1.5    1.50   1.768   0.1841    
## Residuals          726  614.8    0.85                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
abies2_mod <- aov(cover ~ elev + tci + streamdist + as.factor(disturb), data = abies)
Anova(abies2_mod, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: cover
##                     Sum Sq  Df  F value    Pr(>F)
## (Intercept)          68.82   1  81.1767 < 2.2e-16 ***
## elev                 96.64   1 113.9949 < 2.2e-16 ***
## tci                   2.40   1   2.8309   0.09290 .
## streamdist            4.03   1   4.7582   0.02948 *
## as.factor(disturb)   29.12   3  11.4479  2.43e-07 ***
## Residuals           616.35 727
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(abies2_mod)
```

```
##                     Df Sum Sq Mean Sq F value   Pr(>F)
## elev                 1  165.0  165.04 194.676  < 2e-16 ***
## tci                  1    1.7    1.73   2.042   0.1535
## streamdist           1    5.2    5.19   6.124   0.0136 *
## as.factor(disturb)   3   29.1    9.71  11.448 2.43e-07 ***
## Residuals          727  616.3    0.85
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
abies3_mod <- aov(cover ~ elev + streamdist + as.factor(disturb), data = abies)
Anova(abies3_mod, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: cover
##                     Sum Sq  Df  F value    Pr(>F)    
## (Intercept)          85.74   1 100.8844 < 2.2e-16 ***
## elev                 95.18   1 111.9841 < 2.2e-16 ***
## streamdist            2.97   1   3.4968   0.06189 .  
## as.factor(disturb)   29.85   3  11.7057 1.696e-07 ***
## Residuals           618.75 728                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(abies3_mod)
```

```
##                     Df Sum Sq Mean Sq F value  Pr(>F)    
## elev                 1  165.0  165.04 194.187 < 2e-16 ***
## streamdist           1    3.8    3.79   4.461   0.035 *  
## as.factor(disturb)   3   29.8    9.95  11.706 1.7e-07 ***
## Residuals          728  618.7    0.85                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
abies4_mod <- aov(cover ~ elev + as.factor(disturb), data = abies)
Anova(abies4_mod, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: cover
##                     Sum Sq  Df F value    Pr(>F)    
## (Intercept)          84.44   1  99.012 < 2.2e-16 ***
## elev                116.12   1 136.160 < 2.2e-16 ***
## as.factor(disturb)   30.67   3  11.986 1.147e-07 ***
## Residuals           621.72 729                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

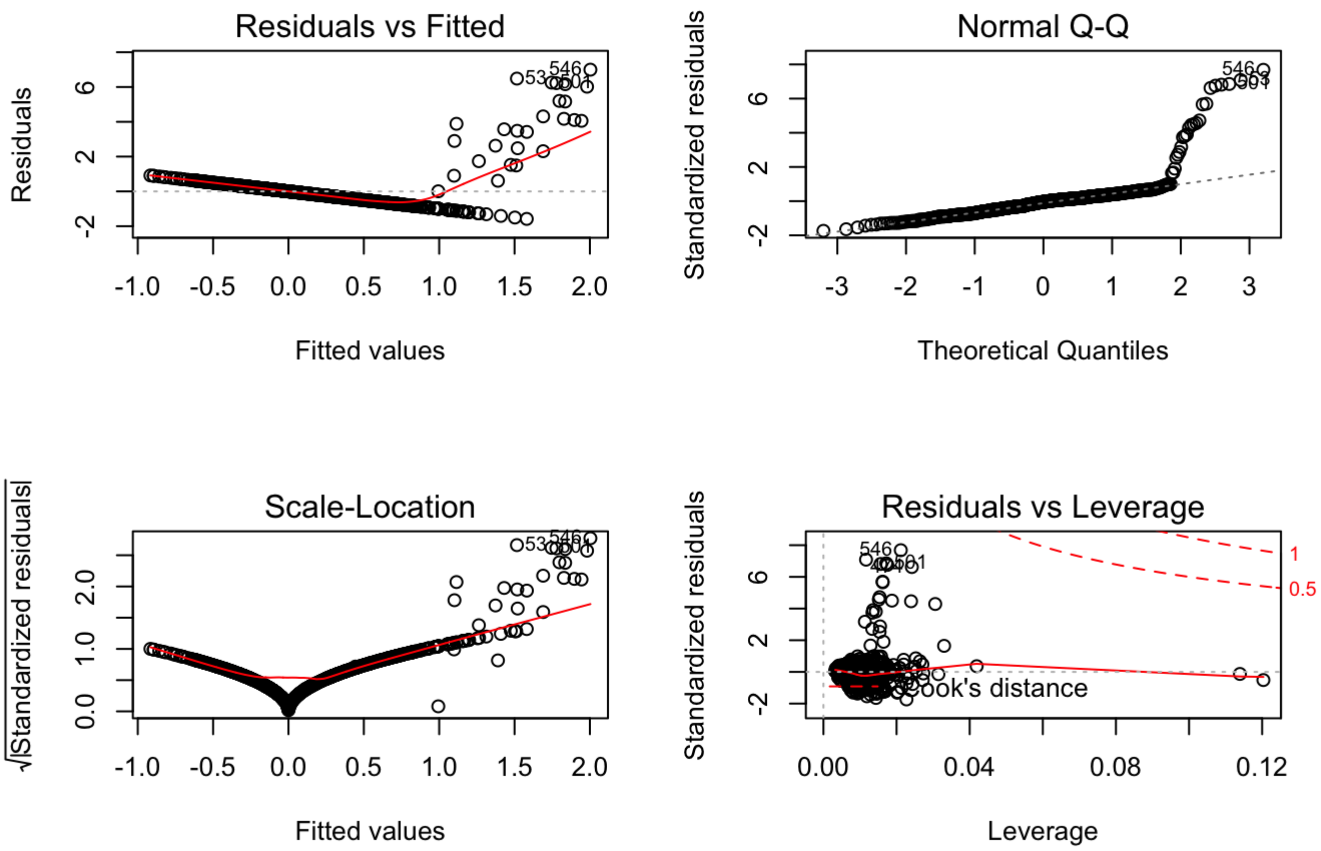
```
summary(abies4_mod)
```

```
##                      Df Sum Sq Mean Sq F value   Pr(>F)    
## elev                  1  165.0  165.04  193.52  < 2e-16 ***
## as.factor(disturb)    3   30.7   10.22   11.99 1.15e-07 ***
## Residuals           729  621.7    0.85                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

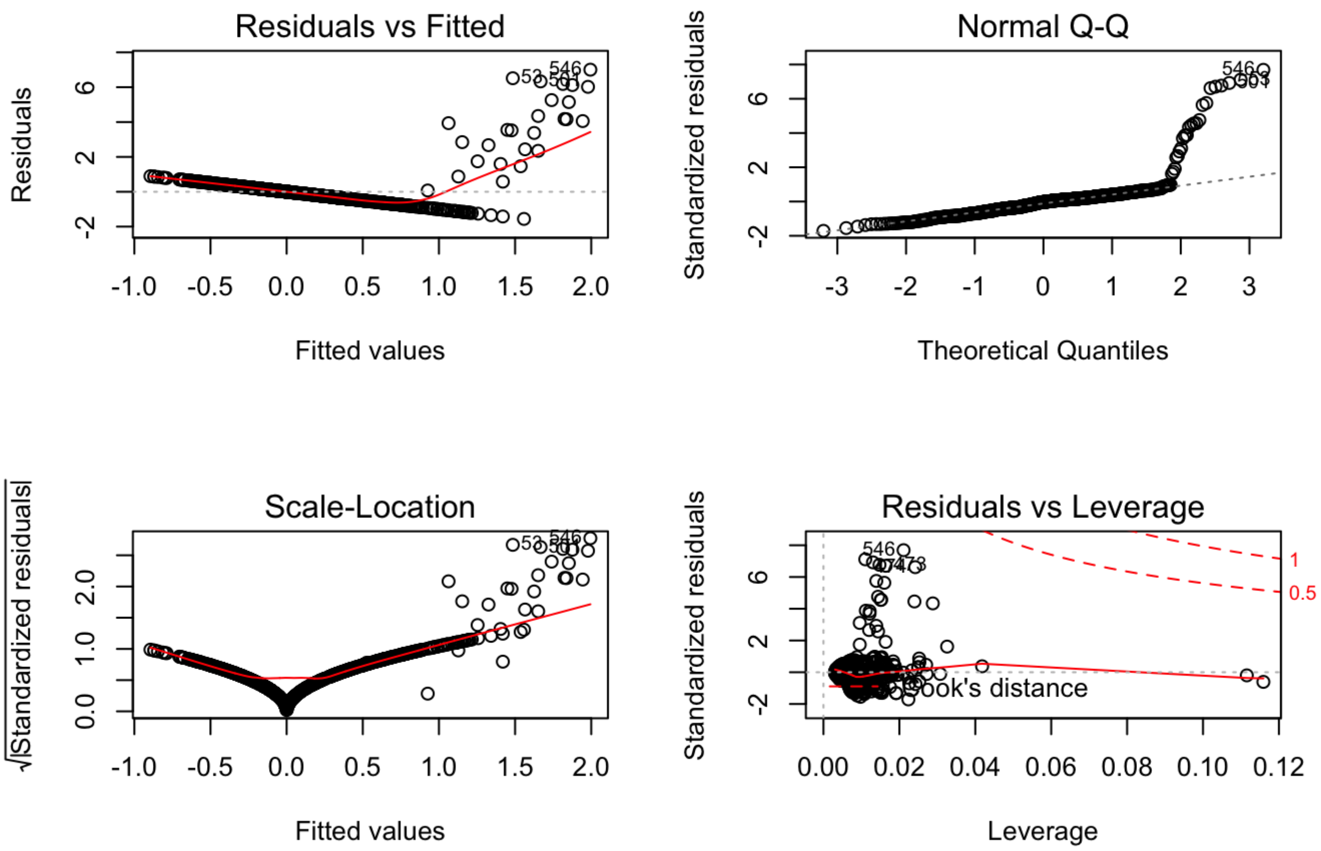
The steps above followed the model simplifcation process by trying to fit the maximal model and then simplifying by removing least significant terms and seeing how removal affects the variance and residual deviance of the model.

The Anova tables provides the analysis of variance table including the degrees of freedom, sum of squares, mean squares, f value, and p-value. The summary function provides the coefficients table of linear models and reports the same p-values and f values as the anova table.

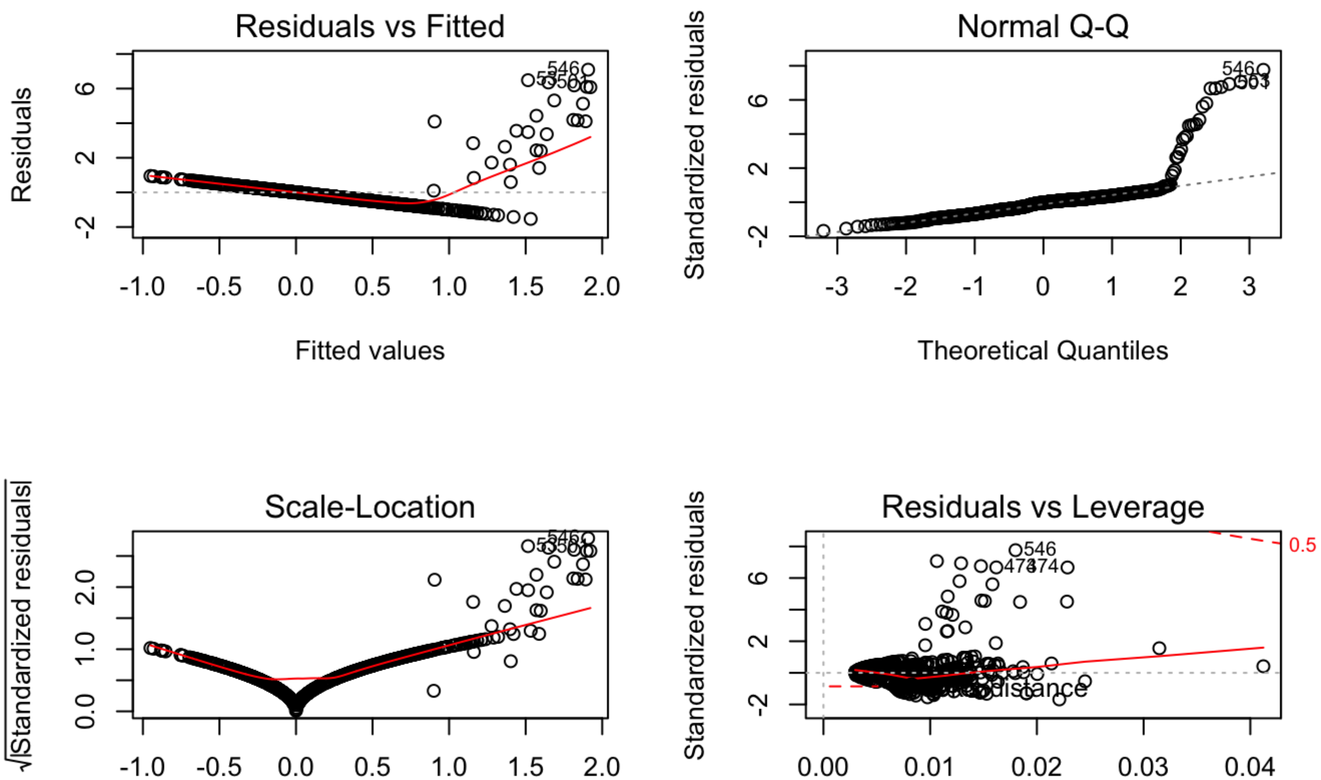
```
par(mfrow=c(2,2))
plot(abies_mod)
```



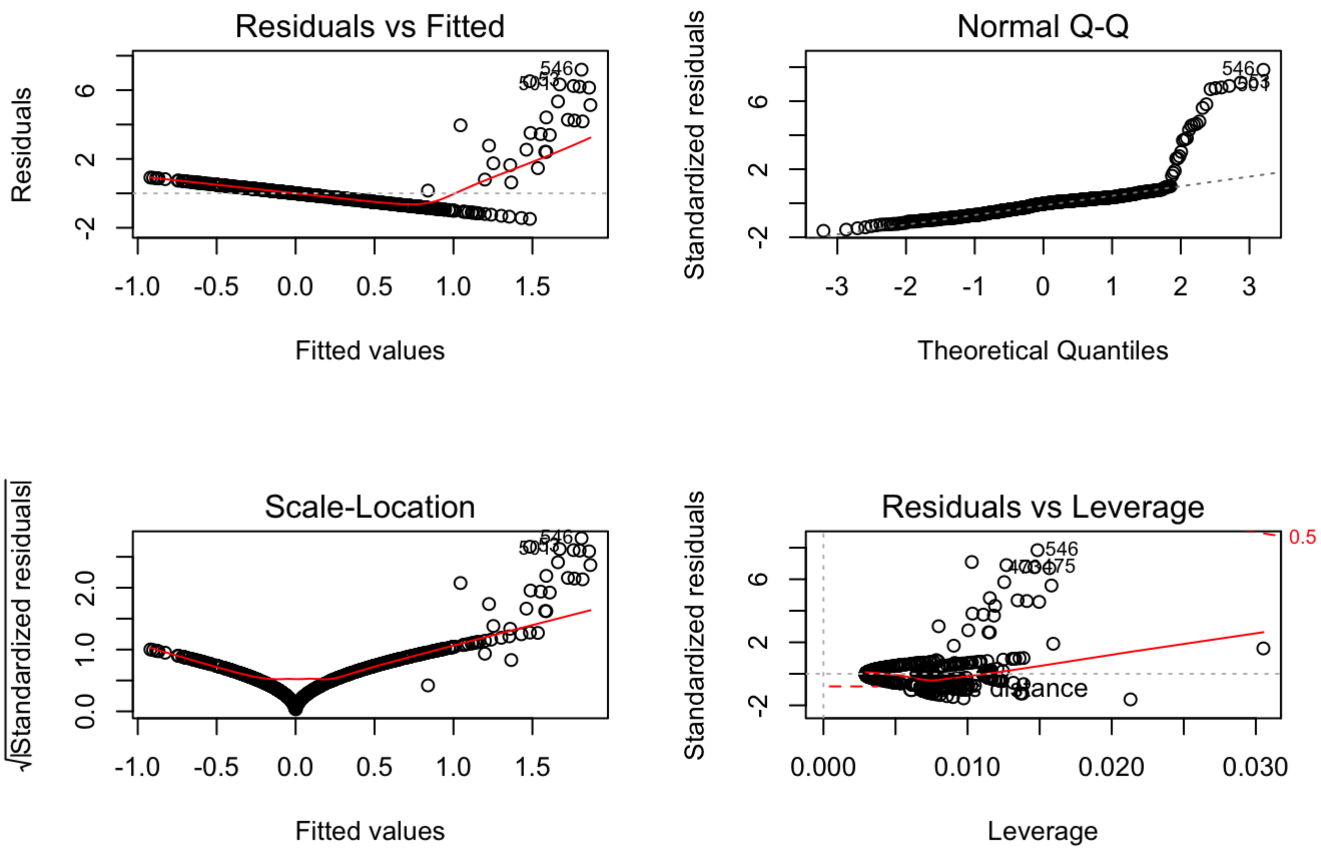
```
par(mfrow=c(2,2))
plot(abies2_mod)
```



```
par(mfrow=c(2,2))
plot(abies3_mod)
```



```
par(mfrow=c(2,2))
plot(abies4_mod)
```



In all the above plots if you focus on the residuals vs fitted and normal Q-Q plots, it can be seen that these data do not fit a normal distribution. For a normal distribution, the mean of the residuals vs fitted values is zero for homogeneity of variance and in the Q-Q plot the residuals would fit the line of linear fit well. Instead we see right skewness in both these graphs, likely due to the large number of zeros in cover. Thus the data does not adhere to OLS assumptions of a normal distribution and must be fitted to a different distribution or transformed to best understand the influence of each explanatory variable on frasier fir cover. Therefore, we cannot trust the p-values of significance in the above anovas run since the models disobey OLS assumptions. Since cover is zero-inflated can try a Poisson distribution to fit the model.

```
abiesglm_mod <- glm(cover ~ elev + tci + streamdist + as.factor(disturb) + beers, family = 'poisson', data = abie
s)
Anova(abiesglm_mod, 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 ** 
## as.factor(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
```

```
summary(abiesglm_mod)
```

```
## 
## Call:
## glm(formula = cover ~ elev + tci + streamdist + as.factor(disturb) + 
##     beers, family = "poisson", data = abies)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.0927  -0.1507  -0.0548  -0.0229   3.8859  
## 
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)    
## (Intercept)              -1.564e+01  1.359e+00 -11.509  < 2e-16 ***
## elev                      7.850e-03  5.883e-04  13.343  < 2e-16 ***
## tci                       1.688e-01  5.927e-02   2.848  0.00440 ** 
## streamdist               -1.692e-03  6.368e-04  -2.658  0.00787 ** 
## as.factor(disturb)LT-SEL  1.622e+00  1.068e+00   1.518  0.12904    
## as.factor(disturb)SETTLE  3.174e+00  1.161e+00   2.733  0.00628 ** 
## as.factor(disturb)VIRGIN  2.649e+00  1.025e+00   2.584  0.00976 ** 
## beers                    -1.826e-02  1.515e-01  -0.120  0.90409    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 940.37  on 733  degrees of freedom
## Residual deviance:  98.57  on 726  degrees of freedom
## AIC: 203.75
## 
## Number of Fisher Scoring iterations: 7
```

In this model the explanatory variables elevation, water potential, stream distance, and disturbance history all have a significant effect on frasier fir cover. Elevation has the strongest effect on frasier fir cover folllowed by disturbance history (corporate logging being most significant), stream distance, and tci. This can be observed in the p-values ($p<0.05$ and the lower the more significant) and Chisq values (higher the more significant) for each explantory variable.

Heat load index (beers) does not appear to contribute much to the variance and will be taken out to test how it impacts the model.If removal of beers does not significantly impact the deviance of the model then it will remain removed.

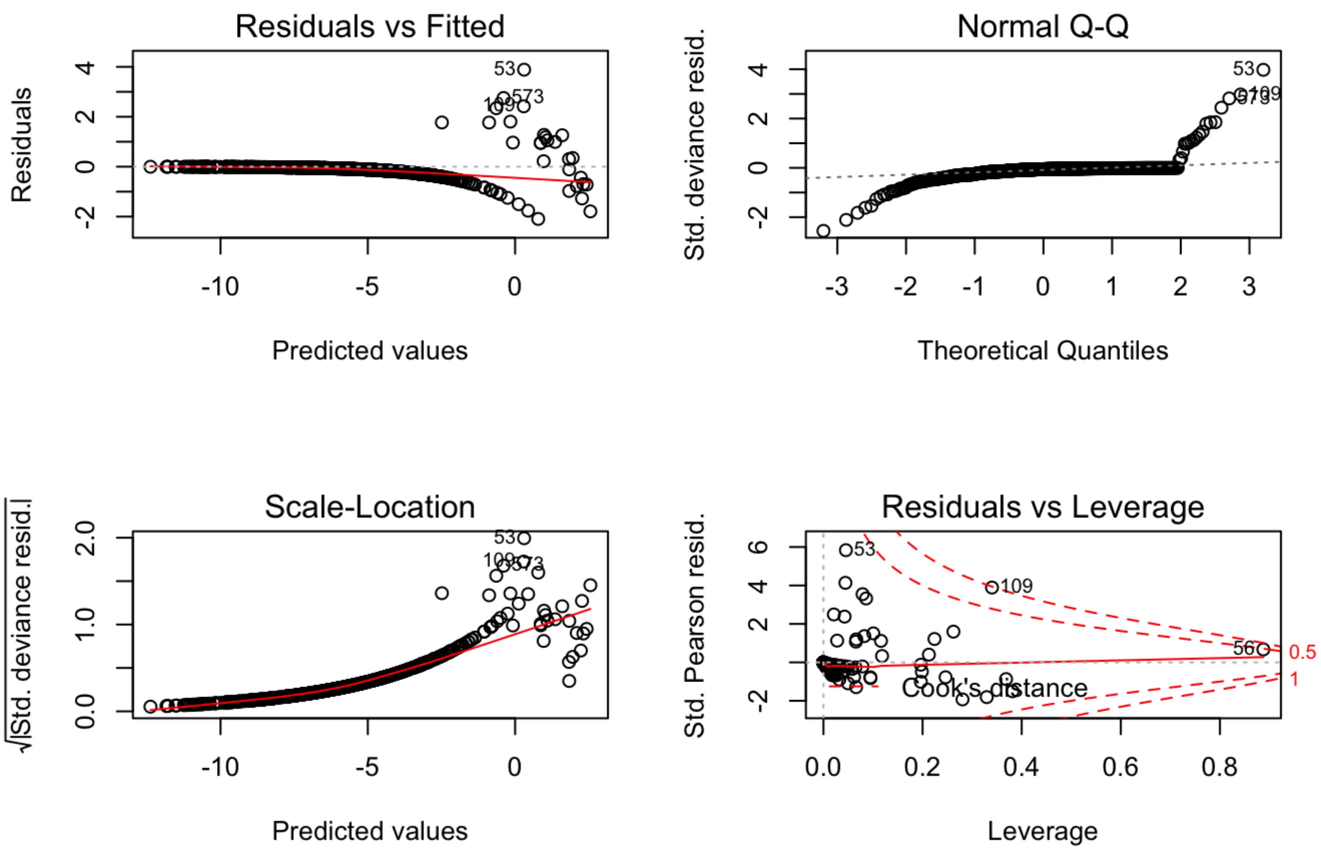
```
abies2glm_mod  <- glm(cover ~ elev + tci + streamdist + as.factor(disturb), family = 'poisson', data = abies)
Anova(abiesglm_mod, type = 3)
```

```
## 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 ** 
## as.factor(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
```

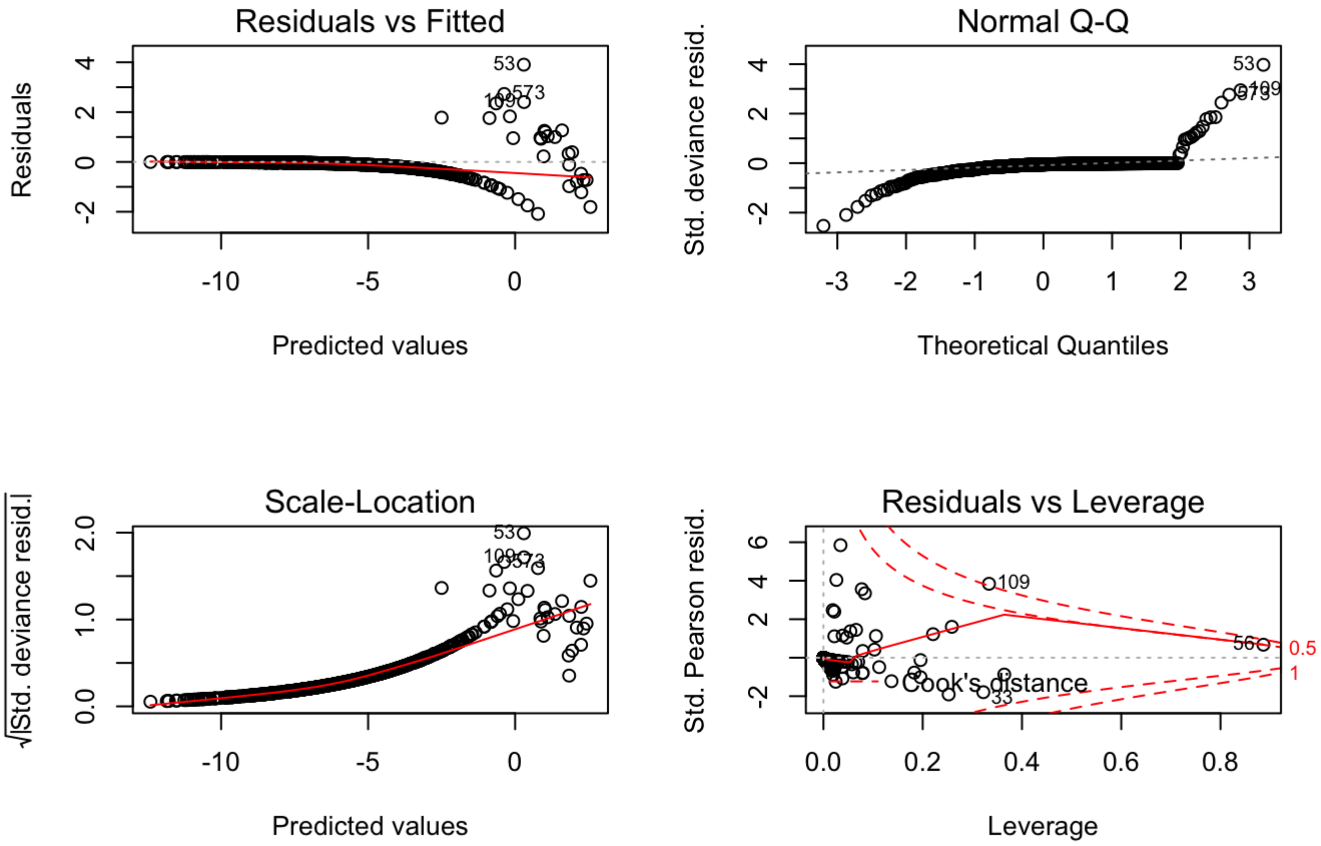
```
summary(abiesglm_mod)
```

```
## 
## Call:
## glm(formula = cover ~ elev + tci + streamdist + as.factor(disturb) + 
##     beers, family = "poisson", data = abies)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.0927  -0.1507  -0.0548  -0.0229   3.8859  
## 
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)    
## (Intercept)              -1.564e+01  1.359e+00 -11.509  < 2e-16 ***
## elev                      7.850e-03  5.883e-04  13.343  < 2e-16 ***
## tci                       1.688e-01  5.927e-02   2.848  0.00440 ** 
## streamdist               -1.692e-03  6.368e-04  -2.658  0.00787 ** 
## as.factor(disturb)LT-SEL  1.622e+00  1.068e+00   1.518  0.12904    
## as.factor(disturb)SETTLE  3.174e+00  1.161e+00   2.733  0.00628 ** 
## as.factor(disturb)VIRGIN  2.649e+00  1.025e+00   2.584  0.00976 ** 
## beers                    -1.826e-02  1.515e-01  -0.120  0.90409    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 940.37  on 733  degrees of freedom
## Residual deviance:  98.57  on 726  degrees of freedom
## AIC: 203.75
## 
## Number of Fisher Scoring iterations: 7
```

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



```
par(mfrow=c(2,2))
plot(abies2glm_mod)
```



When fitted to a poisson distribution using a general linear model, we see a great improvement in our residuals vs fitted values with the mean much closer to zero and the Q-Q plot closer to a linear relationship but still showing some leptokurtosis. The Poisson distribution models processes that produce few rare events and events must be independent of one another and generates integers 0 to positive infinity.

Step Function: AIC Comparison

```
step(abiesglm_mod)
```

```
## Start:  AIC=203.75
## cover ~ elev + tci + streamdist + as.factor(disturb) + beers
## 
##                       Df Deviance    AIC
## - beers                1    98.58 201.76
## <none>                      98.57 203.75
## - tci                  1   105.91 209.09
## - streamdist           1   106.20 209.38
## - as.factor(disturb)  3   120.46 219.64
## - elev                 1   519.40 622.58
## 
## Step:  AIC=201.76
## cover ~ elev + tci + streamdist + as.factor(disturb)
## 
##                       Df Deviance    AIC
## <none>                      98.58 201.76
## - streamdist           1   106.84 208.02
## - tci                  1   107.65 208.83
## - as.factor(disturb)  3   120.50 217.68
## - elev                 1   521.97 623.15
```

```
## 
## Call:  glm(formula = cover ~ elev + tci + streamdist + as.factor(disturb), 
##     family = "poisson", data = abies)
## 
## Coefficients:
##              (Intercept)                      elev  
##               -15.676816                  0.007853  
##                      tci                streamdist  
##                 0.171748                 -0.001710  
## as.factor(disturb)LT-SEL  as.factor(disturb)SETTLE  
##                 1.614142                  3.171190  
## as.factor(disturb)VIRGIN  
##                 2.646177  
## 
## Degrees of Freedom: 733 Total (i.e. Null);  727 Residual
## Null Deviance:       940.4 
## Residual Deviance: 98.58     AIC: 201.8
```

The step function confirms variables of significance using AIC comparison. If it finds removing a variable significantly increases AIC then it should remain in the model, but if it has little effect and lowers the AIC value then it can be removed. Here we can see that heat load index did not significantly describe variance in species cover and when removed lowered the AIC value 2 points so it should remain out of the model.

```
acer_mod <- aov(cover ~ tci + elev + streamdist + as.factor(disturb) + beers, data = acer)
Anova(acer_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 ***
## tci                   55.8   1   8.9257  0.002907 **
## elev                 664.1   1 106.1624 < 2.2e-16 ***
## streamdist            10.8   1   1.7340  0.188316
## as.factor(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
```

```
summary(acer_mod)
```

```
##                     Df Sum Sq Mean Sq F value  Pr(>F)
## tci                  1     13    12.8   2.043 0.15336
## elev                 1    883   883.4 141.219 < 2e-16 ***
## streamdist           1      9     9.3   1.494 0.22204
## as.factor(disturb)   3     40    13.3   2.123 0.09602 .
## beers                1     55    55.1   8.814 0.00309 **
## Residuals          726   4542     6.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the p-values (low) in the Anova table and the F value (high) it appears streamdist has the least significant effect and can be removed from the model. Disturbance is not significant but 0.07 is pretty close to the alpha value of 0.05 and thus may be significant with the right model.

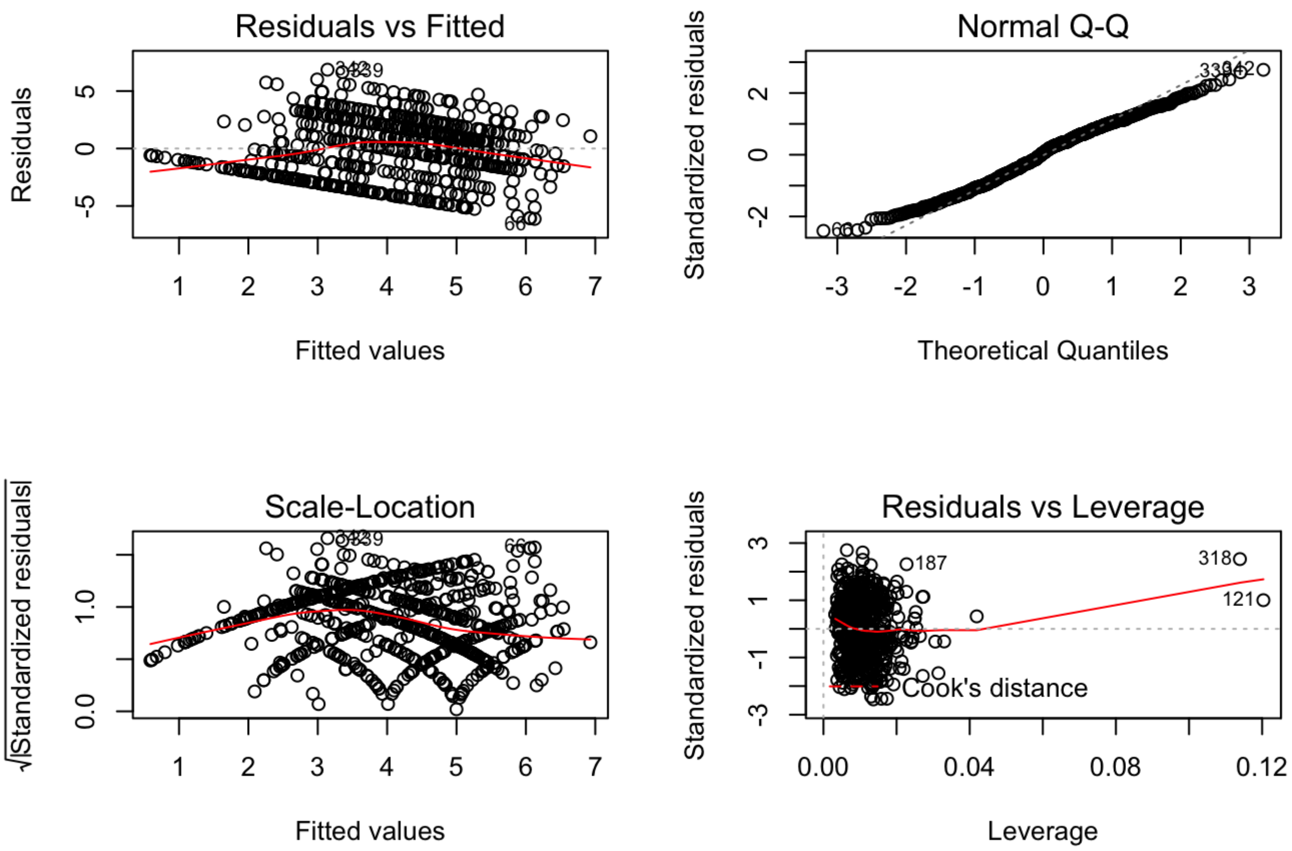
```
acer2_mod <- aov(cover ~ tci + elev + beers + as.factor(disturb), data = acer)
Anova(acer2_mod, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: cover
##                     Sum Sq  Df  F value    Pr(>F)
## (Intercept)         1941.4   1 310.0249 < 2.2e-16 ***
## tci                   69.6   1  11.1110 0.0009016 ***
## elev                 670.3   1 107.0353 < 2.2e-16 ***
## beers                 55.7   1   8.8982 0.0029500 **
## as.factor(disturb)    42.7   3   2.2751 0.0786159 .
## Residuals           4552.6 727
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

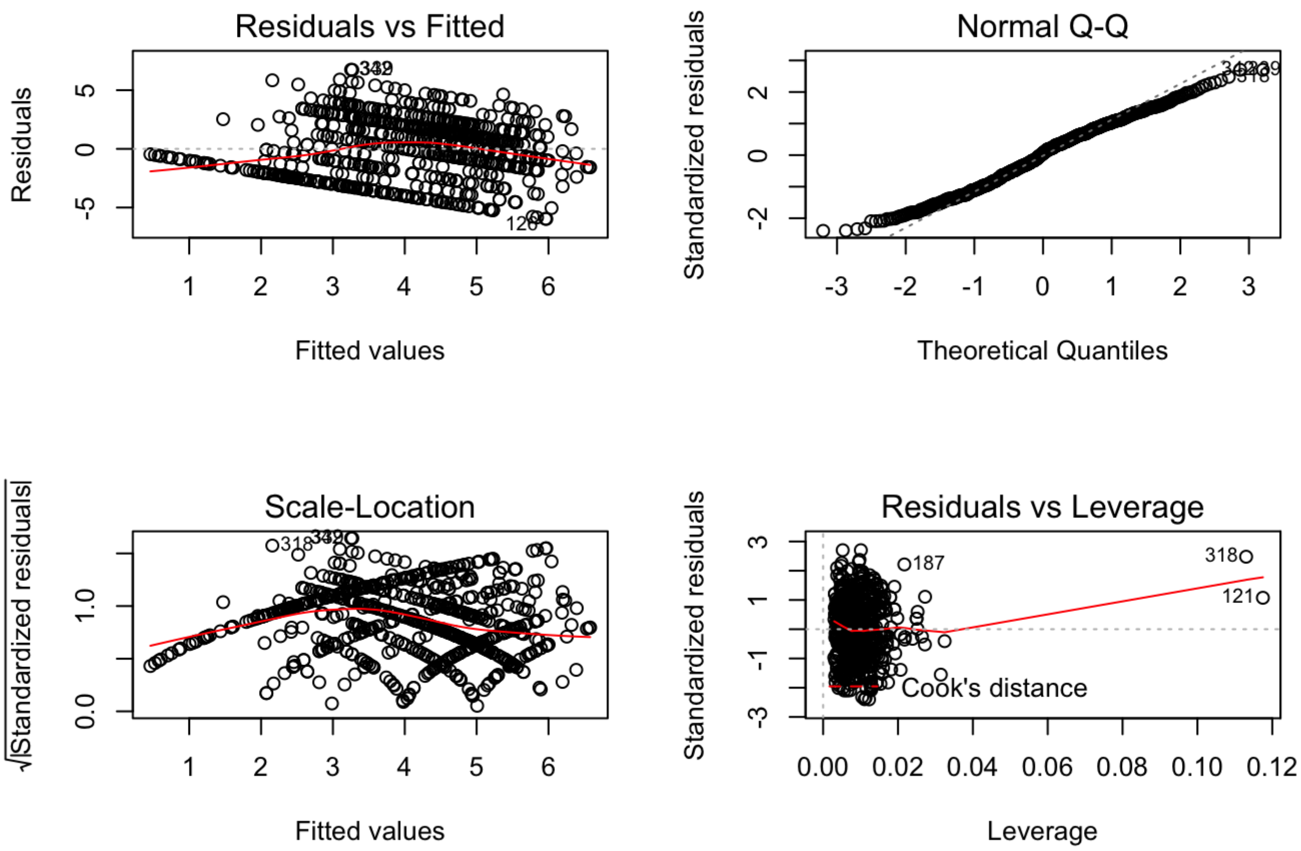
```
summary(acer2_mod)
```

```
##                     Df Sum Sq Mean Sq F value  Pr(>F)
## tci                  1     13    12.8   2.041 0.15357
## elev                 1    883   883.4 141.076 < 2e-16 ***
## beers                1     51    50.7   8.101 0.00455 **
## as.factor(disturb)   3     43    14.2   2.275 0.07862 .
## Residuals          727   4553     6.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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



```
par(mfrow=c(2,2))
plot(acer2_mod)
```



Fits a normal distribution well and OLS assumptions looking at these plots but we are treating cover as a continuous variable when it is discrete and we also are losing the different levels of disturbance history in understanding this model. Next will try to fit to a general linear model with a gaussian distribution.

```
acerglm_mod <- glm(cover ~ elev + tci + streamdist + as.factor(disturb) + beers, family = 'gaussian', data = ace
r)
Anova(acerglm_mod,type = 3)
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: cover
##                    LR Chisq Df Pr(>Chisq)
## elev                106.162  1  < 2.2e-16 ***
## tci                   8.926  1   0.002812 **
## streamdist            1.734  1   0.187900
## as.factor(disturb)    7.044  3   0.070516 .
## beers                 8.814  1   0.002989 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

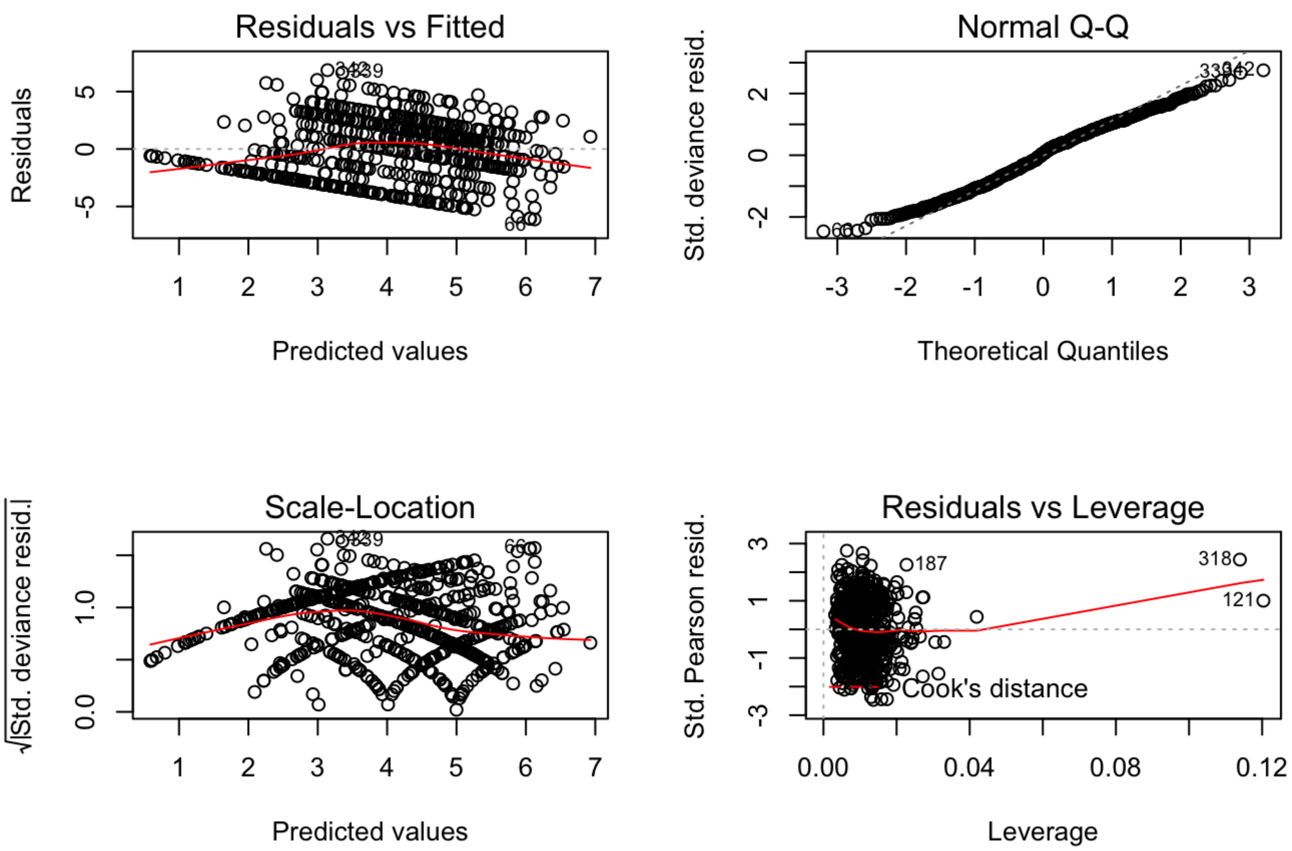
```
summary(acerglm_mod)
```

```
##
## Call:
## glm(formula = cover ~ elev + tci + streamdist + as.factor(disturb) +
##     beers, family = "gaussian", data = acer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.1258  -1.9460   0.1577   1.8624   6.8596
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               8.7372607  0.5086637  17.177  < 2e-16 ***
## elev                     -0.0034639  0.0003362 -10.304  < 2e-16 ***
## tci                      -0.1317294  0.0440921  -2.988  0.00291 **
## streamdist                0.0007520  0.0005711   1.317  0.18832
## as.factor(disturb)LT-SEL -0.4379126  0.2559816  -1.711  0.08756 .
## as.factor(disturb)SETTLE -0.9309789  0.3564239  -2.612  0.00919 **
## as.factor(disturb)VIRGIN -0.3601527  0.2941812  -1.224  0.22125
## beers                    -0.4101716  0.1381555  -2.969  0.00309 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 6.255795)
##
##     Null deviance: 5542.2  on 733  degrees of freedom
## Residual deviance: 4541.7  on 726  degrees of freedom
## AIC: 3438.8
##
## Number of Fisher Scoring iterations: 2
```

In this model the explanatory variables elevation, tci, beers, and disturbance history all have a significant effect on red maple cover. Elevation has the strongest effect on red maple cover folllowed by tci, heat load index (beers), and disturbance history (corporate logging being most significant). This can be observed in the p-values ($p<0.05$ and the lower the more significant) and Likelihood ratio (LR) Chisquare values (higher the more significant) for each explantory variable.

Stream distance does not appear to contribute much to the variance and will be taken out to test how it impacts the model. If removal of stream distance does not significantly impact the deviance of the model then it will remain removed.

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



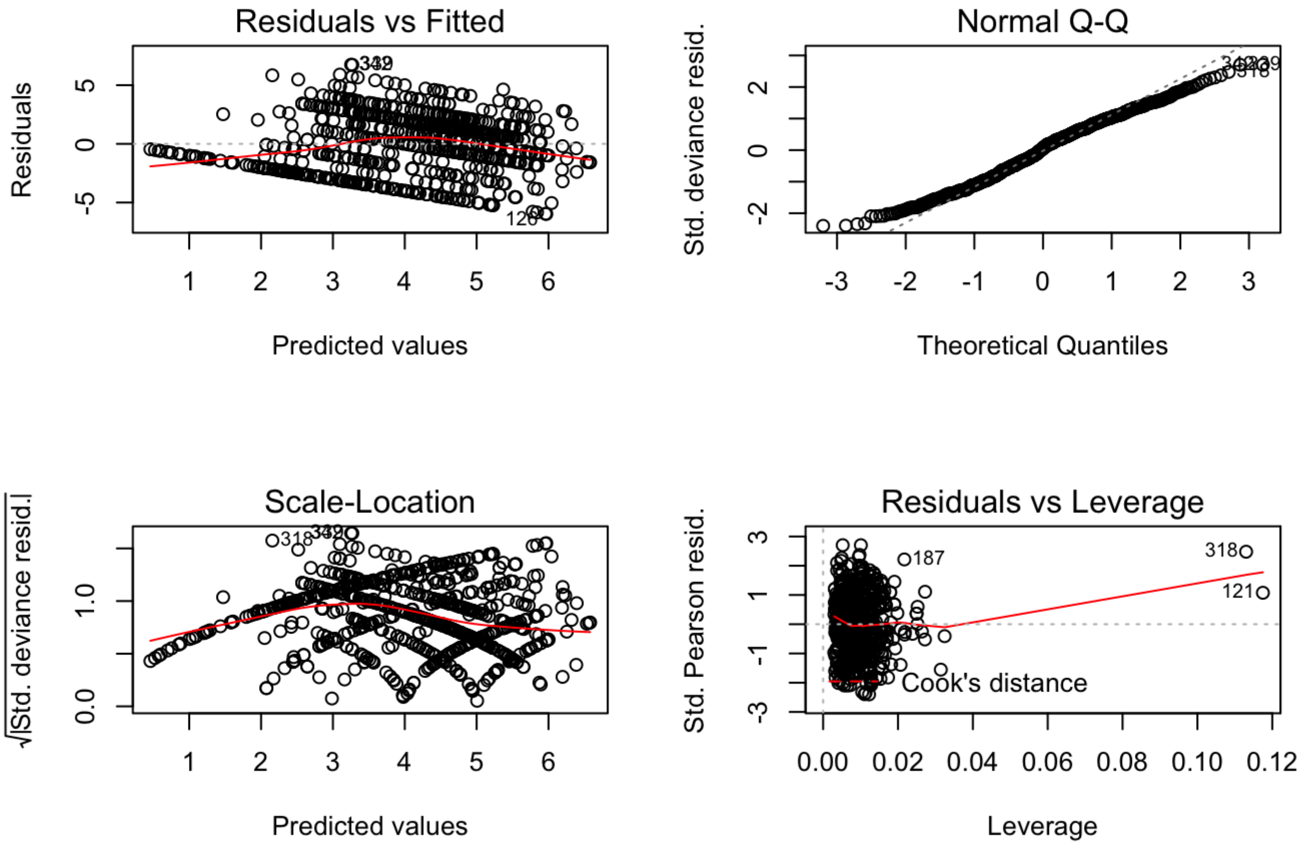
```
acer2glm_mod <- glm(cover ~ tci + elev + as.factor(disturb) + beers, family = 'gaussian', data = acer)
Anova(acer2glm_mod,type = 3)
```

```
## Analysis of Deviance Table (Type III tests)
## 
## Response: cover
##                    LR Chisq Df Pr(>Chisq)    
## tci                  11.111  1  0.0008582 ***
## elev                107.035  1  < 2.2e-16 ***
## as.factor(disturb)    6.825  3  0.0776772 .  
## beers                 8.898  1  0.0028545 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(acer2glm_mod)
```

```
## 
## Call:
## glm(formula = cover ~ tci + elev + as.factor(disturb) + beers, 
##     family = "gaussian", data = acer)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -5.9699  -1.9645   0.1384   1.8665   6.7468  
## 
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               8.8447322  0.5023269  17.608  < 2e-16 ***
## tci                      -0.1438223  0.0431469  -3.333 0.000902 ***
## elev                     -0.0033394  0.0003228 -10.346  < 2e-16 ***
## as.factor(disturb)LT-SEL -0.4046811  0.2548632  -1.588 0.112758    
## as.factor(disturb)SETTLE -0.9227220  0.3565486  -2.588 0.009848 ** 
## as.factor(disturb)VIRGIN -0.3360115  0.2937576  -1.144 0.253067    
## beers                    -0.4122964  0.1382157  -2.983 0.002950 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 6.262111)
## 
##     Null deviance: 5542.2  on 733  degrees of freedom
## Residual deviance: 4552.6  on 727  degrees of freedom
## AIC: 3438.5
## 
## Number of Fisher Scoring iterations: 2
```

```
par(mfrow=c(2,2))
plot(acer2glm_mod)
```



We can see that Red Maple fits a normal glm distribution well since in the residuals vs fitted values the residual mean is close to zero representing homoegeneity of variance and in the Q-Q plot the residuals fit the linear regression very well. Thus the data adheres to OLS assumptions of a normal distribution.

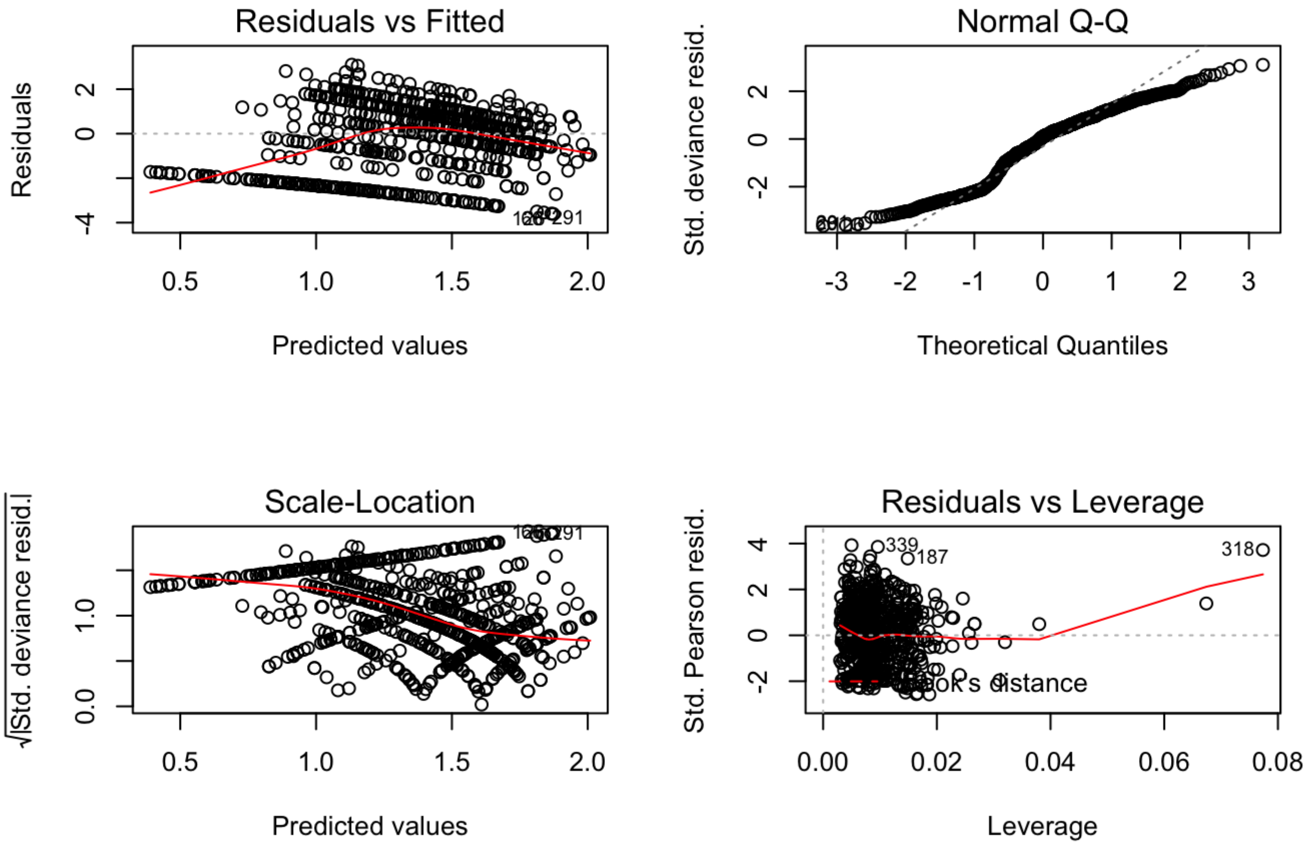
```
acer3glm_mod <- glm(cover ~ tci + elev + beers + as.factor(disturb), family = 'poisson', data = acer)
Anova(acer2glm_mod,type = 3)
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: cover
##                     LR Chisq Df Pr(>Chisq)    
## tci                   11.111  1  0.0008582 ***
## elev                 107.035  1  < 2.2e-16 ***
## as.factor(disturb)     6.825  3  0.0776772 .  
## beers                  8.898  1  0.0028545 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(acer3glm_mod)
```

```
## 
## Call:
## glm(formula = cover ~ tci + elev + beers + as.factor(disturb), 
##     family = "poisson", data = acer)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -3.6078  -1.5023   0.0954   0.8883   3.1017  
## 
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)    
## (Intercept)               2.618e+00  1.008e-01  25.967  < 2e-16 ***
## tci                      -3.636e-02  9.094e-03  -3.999 6.36e-05 ***
## elev                     -8.972e-04  6.983e-05 -12.848  < 2e-16 ***
## beers                    -9.601e-02  2.722e-02  -3.527 0.000420 ***
## as.factor(disturb)LT-SEL -1.216e-01  5.020e-02  -2.423 0.015392 *  
## as.factor(disturb)SETTLE -2.359e-01  6.804e-02  -3.467 0.000526 ***
## as.factor(disturb)VIRGIN -9.931e-02  6.294e-02  -1.578 0.114600    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 1917.4  on 733  degrees of freedom
## Residual deviance: 1661.6  on 727  degrees of freedom
## AIC: 3651.3
## 
## Number of Fisher Scoring iterations: 5
```

```
par(mfrow=c(2,2))
plot(acer3glm_mod)
```



Shows Poisson distribution does not fit species cover for the Red Maple well and a Gaussian (normal) distribution is more appropriate.

```
step(acerglm_mod)
```

```
## Start:  AIC=3438.75
## cover ~ elev + tci + streamdist + as.factor(disturb) + beers
## 
##                       Df Deviance    AIC
## - streamdist           1   4552.6 3438.5
## <none>                     4541.7 3438.8
## - as.factor(disturb)  3   4585.8 3439.8
## - beers                1   4596.8 3445.6
## - tci                  1   4597.5 3445.7
## - elev                 1   5205.8 3536.9
## 
## Step:  AIC=3438.5
## cover ~ elev + tci + as.factor(disturb) + beers
## 
##                       Df Deviance    AIC
## <none>                     4552.6 3438.5
## - as.factor(disturb)  3   4595.3 3439.4
## - beers                1   4608.3 3445.4
## - tci                  1   4622.1 3447.6
## - elev                 1   5222.8 3537.3
```

```
## 
## Call:  glm(formula = cover ~ elev + tci + as.factor(disturb) + beers, 
##     family = "gaussian", data = acer)
## 
## Coefficients:
##            (Intercept)                    elev  
##               8.844732               -0.003339  
##                    tci  as.factor(disturb)LT-SEL  
##              -0.143822               -0.404681  
## as.factor(disturb)SETTLE  as.factor(disturb)VIRGIN  
##              -0.922722               -0.336012  
##                  beers  
##              -0.412296  
## 
## Degrees of Freedom: 733 Total (i.e. Null);  727 Residual
## Null Deviance:       5542 
## Residual Deviance: 4553  AIC: 3439
```

The step function confirms variables of significance using AIC comparison. If it finds removing a variable significantly increases AIC then it should remain in the model, but if it has little effect and lowers the AIC value then it can be removed. Here we can see that stream distance did not significantly describe deviance in species cover and when removed lowered the AIC value 0.3 so cover is best described without it.

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

```
#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 question 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 approximatio
n.
#Your new model calls will look as follows:

#acer_poi = glm(cover ~ tci + elev + ... , data = my_data,
           #family='poisson')
```

^^Did this step above in model testing for each species^^

2. 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?

For both species elevation had the most significant effect on their cover and was most important in the model. Both species cover were also significantly effected by site water potential (tci) and plot disturbance history with corporate logging sites the most significant type. Dissimilarly, the frasier fir model was best explained with stream distance and without heat load index while red maple was the opposite. This could be due to these species differing preffered climates where heat load index at very high elevations that are cooler may be less important while stream distance is due to the harsher climate and limited natural resources. For the generalist species, stream distance may be less important than heat load index because this particular species may be better adapted to drought conditions than large drops or peaks in temperature.

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?

Frasier fir is a specialist species that occurs at elevations greater than 1500 meters. In addition to elevation, the distribution of this species is also influenced by a site's water potential (tci), the disturbance history of the plot, and the distance of the plot from the closest stream. Heat load index of a plot does not influence the cover of frasier fir. The red maple is a generalist species that occurs at elevtaions below 1500 meters and the species cover is influenced by tci, the disturbance history, and the heat load index of the plot. The distance to the closest stream does not have an impact on the cover of red maples in an area.