quant_methods_univariate_assignment

lloyd hill

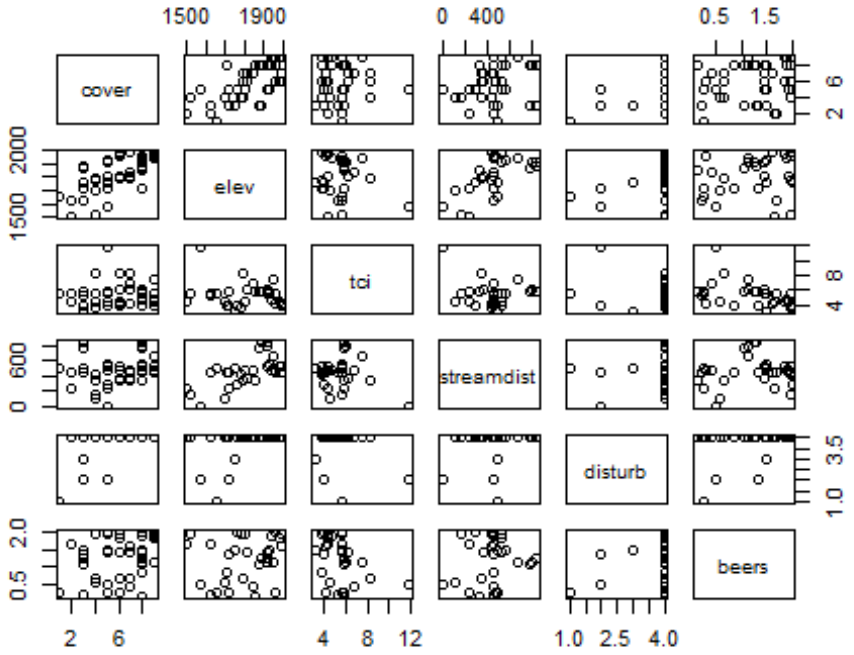
February 4, 2019

```
#import trees data
trees <-
read.csv("https://raw.githubusercontent.com/dmcglinn/quant_methods/gh-
pages/data/treedata_subset.csv")
head(trees)
```

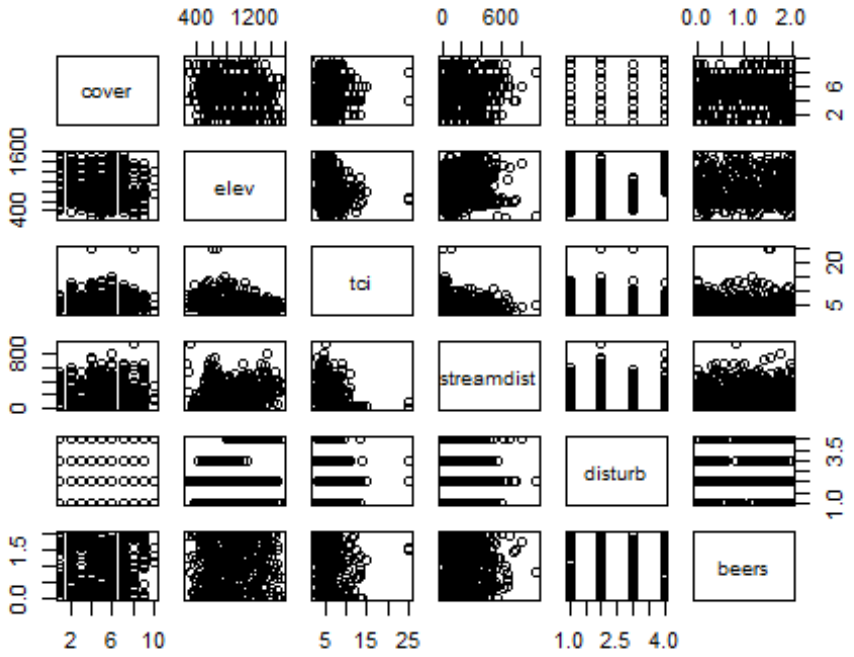
```
##         plotID  spcode       species cover elev       tci streamdist
## 1 ATBN-01-0403 ABIEFRA Abies fraseri     1 1660  5.701460      490.9
## 2 ATBN-01-0532 ABIEFRA Abies fraseri     8 1712  3.823586      454.0
## 3 ATBN-01-0533 ABIEFRA Abies fraseri     3 1722  3.893762      453.4
## 4 ATBN-01-0536 ABIEFRA Abies fraseri     3 1754  3.145527      492.5
## 5 FRID-01-0003 ABIEFRA Abies fraseri     5 1570 11.850000        0.0
## 6 PITT-01-0045 ABIEFRA Abies fraseri     2 1504  4.373741      237.1
##   disturb     beers
## 1 CORPLOG 0.2244286
## 2  VIRGIN 0.8340878
## 3  LT-SEL 1.3332586
## 4  SETTLE 1.4712484
## 5  LT-SEL 0.4961189
## 6  VIRGIN 1.6558421
```

```
#subset data into two species of interest, frasier fir and red maple
trees_abi <- subset(trees, trees$spcode == 'ABIEFRA')
trees_ace <- subset(trees, trees$spcode == 'ACERRUB')

#make pairs plots to explore potential relationships among variables of
interest (cover, elev, tci, streamdist, and beers)
pairs(trees_abi[,c(4:9)])
```




```
pairs(trees_ace[,c(4:9)])
```



```r
#after viewing these initial pairs plots, I can see some trends between elevation and cover as well as some possible interactions (eg, a positive relationship between elevation and stream distance, and a negative relationship between tci and elevation).

#I wanted to first build linear models to explore potential relationships between cover (response variable) and any habitat parameters (explanatory variables).

abi_lm1 <- lm(cover ~ elev + tci + streamdist + disturb + beers, data = trees_abi)
Anova(abi_lm1, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: cover
##             Sum Sq Df F value    Pr(>F)
## (Intercept) 59.401  1 23.1710 2.652e-05 ***
## elev        61.618  1 24.0358 2.022e-05 ***
## tci          5.667  1  2.2105    0.1458
## streamdist   1.636  1  0.6382    0.4296
## disturb     10.089  3  1.3118    0.2855
## beers        0.014  1  0.0056    0.9406
## Residuals   92.289 36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
ace_lm1 <- lm(cover ~ elev + tci + streamdist + disturb + beers, data = trees_ace)
Anova(ace_lm1, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: cover
##              Sum Sq  Df  F value    Pr(>F)
## (Intercept)  765.43   1 193.5096 < 2.2e-16 ***
## elev          40.44   1  10.2233  0.001448 **
## tci           12.58   1   3.1805  0.074947 .
## streamdist    29.09   1   7.3531  0.006856 **
## disturb        9.45   3   0.7962  0.496166
## beers         35.61   1   9.0034  0.002789 **
## Residuals   2828.21 715
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#I compared the Anova() function from the 'car' package to the R's basic 'summary' outputs. While the p-values are very similar, summary handles the categorical variable (disturb) differently. The summary table also includes R-squared values.

summary(abi_lm1)
```

```
##
## Call:
## lm(formula = cover ~ elev + tci + streamdist + disturb + beers,
##     data = trees_abi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4630 -0.6472  0.0788  1.0872  3.8017
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -20.561173   4.271449  -4.814 2.65e-05 ***
## elev            0.012370   0.002523   4.903 2.02e-05 ***
## tci             0.287641   0.193467   1.487   0.1458
## streamdist     -0.001266   0.001585  -0.799   0.4296
## disturbLT-SEL   2.188367   2.097905   1.043   0.3038
## disturbSETTLE   1.527604   2.341471   0.652   0.5183
## disturbVIRGIN   3.025596   1.735921   1.743   0.0899 .
## beers           0.037551   0.500269   0.075   0.9406
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.601 on 36 degrees of freedom
## Multiple R-squared:  0.5824, Adjusted R-squared:  0.5011
## F-statistic: 7.171 on 7 and 36 DF,  p-value: 2.215e-05
```

```
summary(ace_lm1)
```

```
##
## Call:
## lm(formula = cover ~ elev + tci + streamdist + disturb + beers,
##     data = trees_ace)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7073 -1.2446  0.3409  1.3575  5.2732
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.3502303  0.4564973  13.911  < 2e-16 ***
## elev          -0.0010108  0.0003161  -3.197  0.00145 **
## tci           -0.0627613  0.0351922  -1.783  0.07495 .
## streamdist     0.0012895  0.0004756   2.712  0.00686 **
## disturbLT-SEL  0.0829610  0.2166747   0.383  0.70192
## disturbSETTLE -0.1044556  0.2804213  -0.372  0.70963
## disturbVIRGIN  0.3088364  0.2518161   1.226  0.22044
## beers         -0.3269597  0.1089662  -3.001  0.00279 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.989 on 715 degrees of freedom
## Multiple R-squared:  0.04493,    Adjusted R-squared:  0.03558
## F-statistic: 4.805 on 7 and 715 DF,  p-value: 2.669e-05
```

#With these models, elevation is the only significant (p<0.05) variable for fir cover; whereas, elevation, stream distance and beers are all significant explanatory variables for red maple. These models also return an adjusted R-squared value of 0.5 for fir cover, and 0.04 for maple cover.

#While my first model includes all variables in this dataset, it may not be the best model for describing cover of these two species. There may in fact create a better model to describe the growth habits of both species. Since stream distance, elevation, and topographical convergence index are all indicators of topography, it is likely our data contain interactions. These models include interactions between stream distance, tci, and elevation. (I didn't include beers as a interaction because slope aspect should be independent of elevation).

```
abi_lm2 <- lm(cover ~ elev + tci + streamdist + elev:tci + elev:streamdist +
beers, data = trees_abi)
Anova(abi_lm2, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: cover
##                 Sum Sq Df F value  Pr(>F)
## (Intercept)      0.935  1  0.3692 0.54717
## elev             0.262  1  0.1033 0.74970
## tci              3.995  1  1.5770 0.21706
## streamdist       7.822  1  3.0882 0.08714 .
## beers            0.004  1  0.0015 0.96938
## elev:tci         4.631  1  1.8284 0.18452
## elev:streamdist  7.134  1  2.8164 0.10174
## Residuals       93.719 37
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(abi_lm2)
```

```
##
## Call:
## lm(formula = cover ~ elev + tci + streamdist + elev:tci + elev:streamdist
+
##     beers, data = trees_abi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3191 -0.7693  0.0858  0.8376  4.4619
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)      1.083e+01  1.782e+01   0.608   0.5472
## elev            -3.121e-03  9.709e-03  -0.321   0.7497
## tci             -2.829e+00  2.252e+00  -1.256   0.2171
## streamdist      -4.495e-02  2.558e-02  -1.757   0.0871 .
## beers            1.970e-02  5.099e-01   0.039   0.9694
## elev:tci         1.732e-03  1.281e-03   1.352   0.1845
## elev:streamdist  2.306e-05  1.374e-05   1.678   0.1017
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.592 on 37 degrees of freedom
## Multiple R-squared:  0.5759, Adjusted R-squared:  0.5071
## F-statistic: 8.373 on 6 and 37 DF,  p-value: 9.174e-06
```

```
ace_lm2 <- lm(cover ~ elev + tci + streamdist + elev:tci + elev:streamdist +
beers, data = trees_ace)
Anova(ace_lm2, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: cover
##                  Sum Sq  Df F value   Pr(>F)
## (Intercept)       94.33   1 24.0369 1.17e-06 ***
## elev              11.32   1  2.8854 0.089821 .
## tci               15.59   1  3.9722 0.046635 *
## streamdist        15.21   1  3.8771 0.049333 *
## beers             31.18   1  7.9449 0.004956 **
## elev:tci          27.38   1  6.9776 0.008434 **
## elev:streamdist    4.51   1  1.1502 0.283866
## Residuals       2809.74 716
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(ace_lm2)
```

```
##
## Call:
## lm(formula = cover ~ elev + tci + streamdist + elev:tci + elev:streamdist
+
##     beers, data = trees_ace)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9699 -1.3135  0.3179  1.3789  5.5621
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.268e+00  8.705e-01   4.903 1.17e-06 ***
## elev             1.811e-03  1.066e-03   1.699  0.08982 .
## tci              2.450e-01  1.229e-01   1.993  0.04664 *
## streamdist       2.544e-03  1.292e-03   1.969  0.04933 *
```

```
## beers           -3.014e-01  1.069e-01  -2.819  0.00496 **
## elev:tci        -4.216e-04  1.596e-04  -2.642  0.00843 **
## elev:streamdist -1.558e-06  1.453e-06  -1.072  0.28387
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.981 on 716 degrees of freedom
## Multiple R-squared:  0.05117,    Adjusted R-squared:  0.04321
## F-statistic: 6.435 on 6 and 716 DF,  p-value: 1.28e-06
```

```
#These models produce similar r-squared values as my first set of models, but the p-values have changed. For example, all variables in the fir cover lost their signifcance, and the interaction between elevation and tci showed siginificance for maple cover.
```

For each species address the following additional questions:

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

FIR The r-squared values (>0.5) from both of my models suggest that these habitat parameters do a pretty good job at explaining cover.

MAPLE The low r-sqaured values (<0.01) from both of my models suggest that these habitat parameters do not explain cover very well. There may be patterns (as explained by low P-values) but they do not make for a good linear regression.

1.2. Which explanatory variables are the most important?

FIR In my first model, elevation was the only explanatory variable with a significant p-value. Once I introduced interactions into the second model, elevation lost its significant P-value.

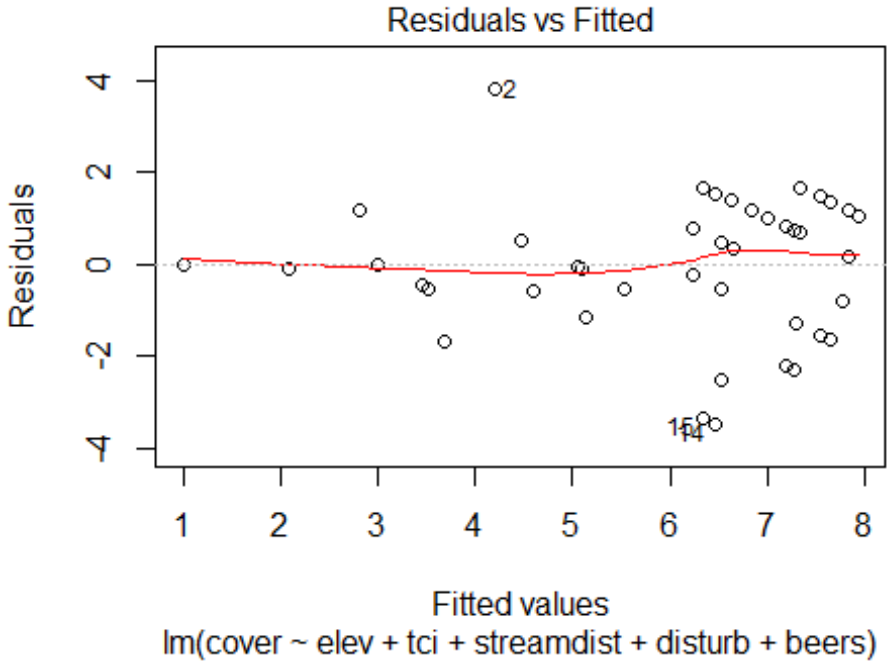
MAPLE In my first model, several explanatory variables demonstrated a siginificant p-value: elevation, stream distance, and beers. When I included interactions, these variables maintained significance and the interaction between elevation and tci showed significance.

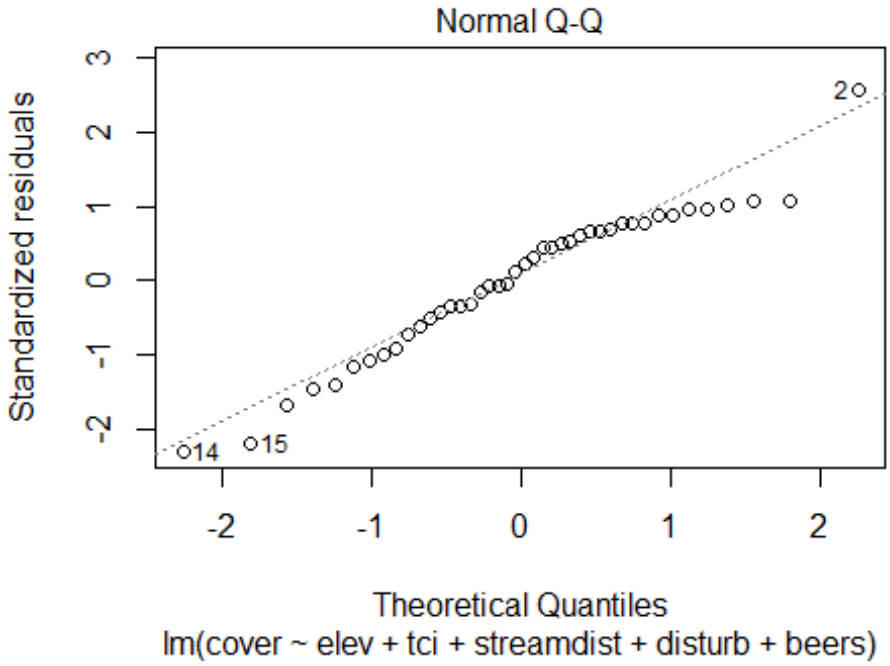
1.3. Do model diagnostics indicate any problems with violations of OLS assumptions?

I used the plot function to see how well these data matched the assumptions of our analysis. In both species, the residuals appear to be normally distributed and varaince looks homogenous. The study design also suggests samples were independent.

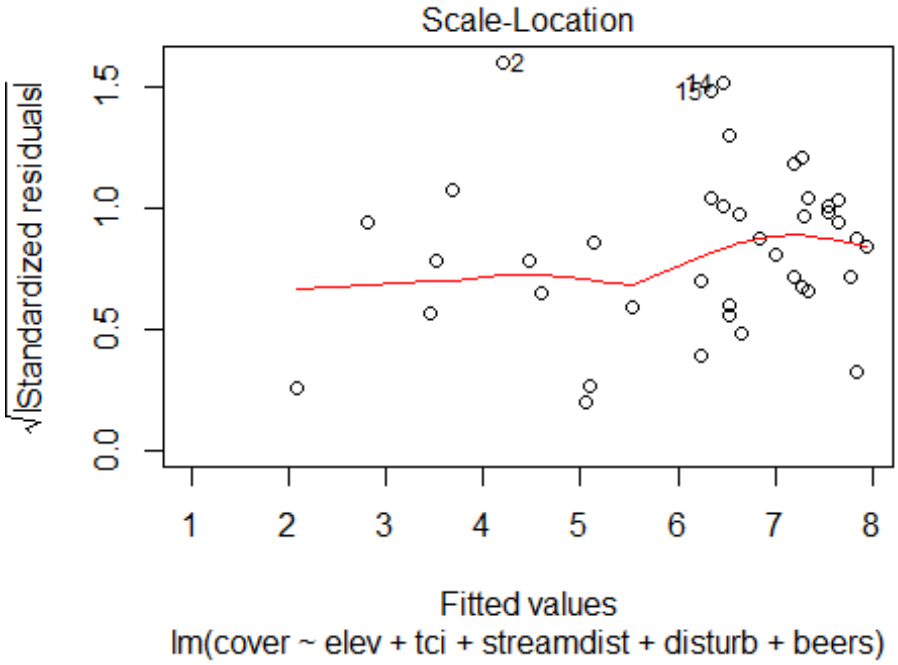
```
#FIR
plot(abi_lm1)
```

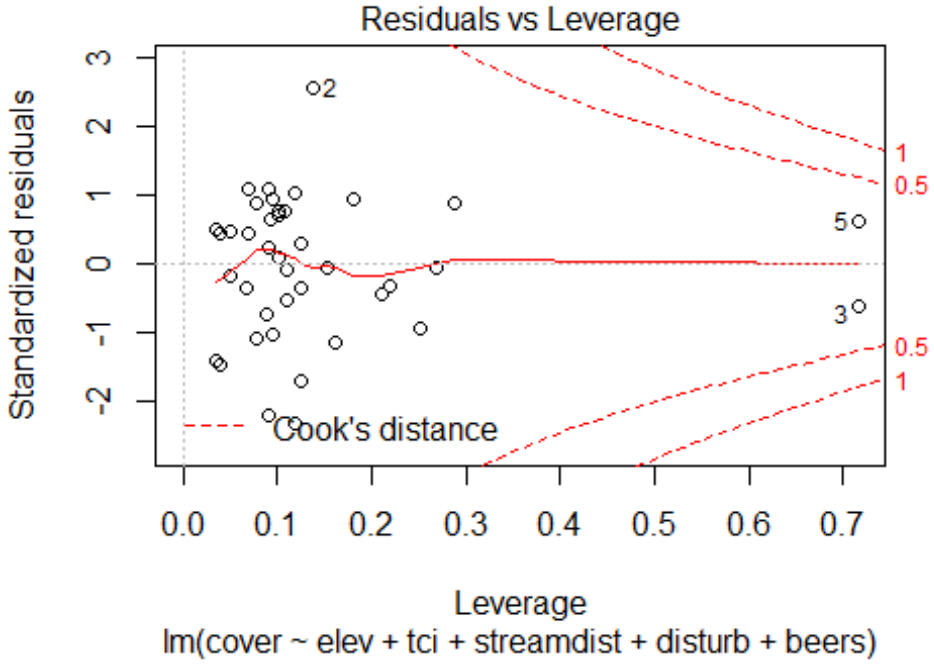
```
## Warning: not plotting observations with leverage one:
##   1, 4
```



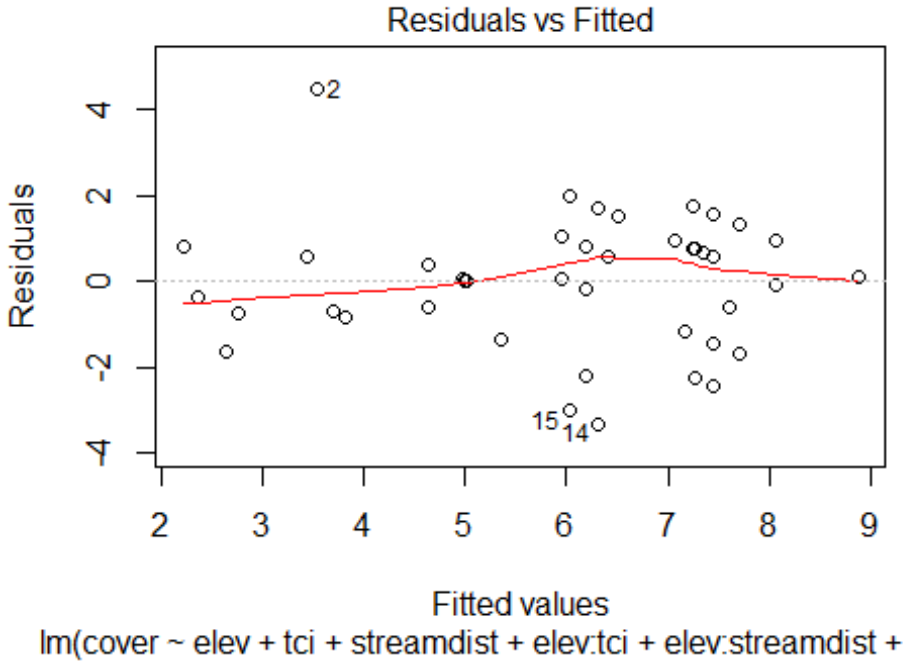


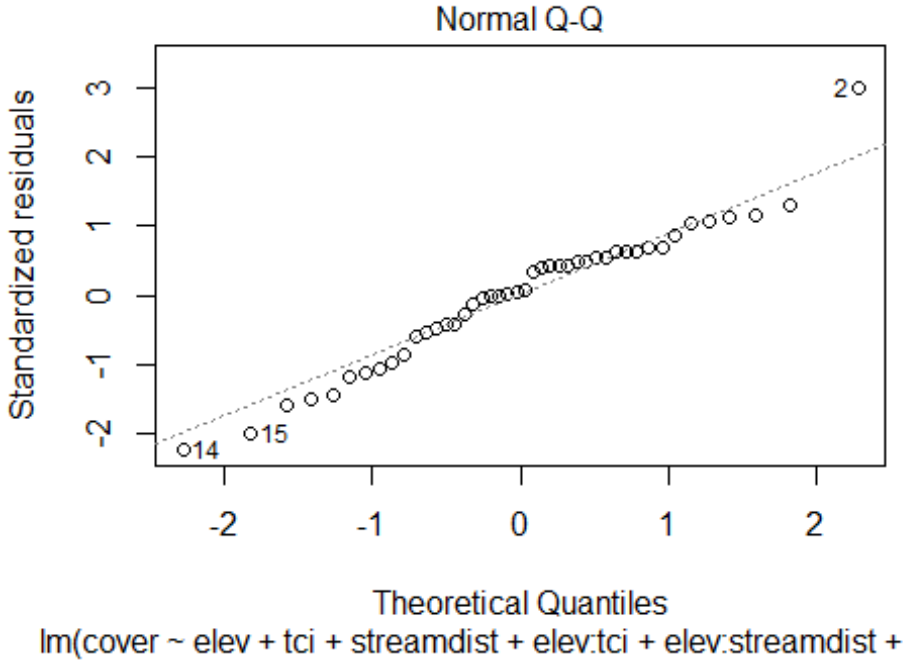

```
## Warning: not plotting observations with leverage one:
##   1, 4
```

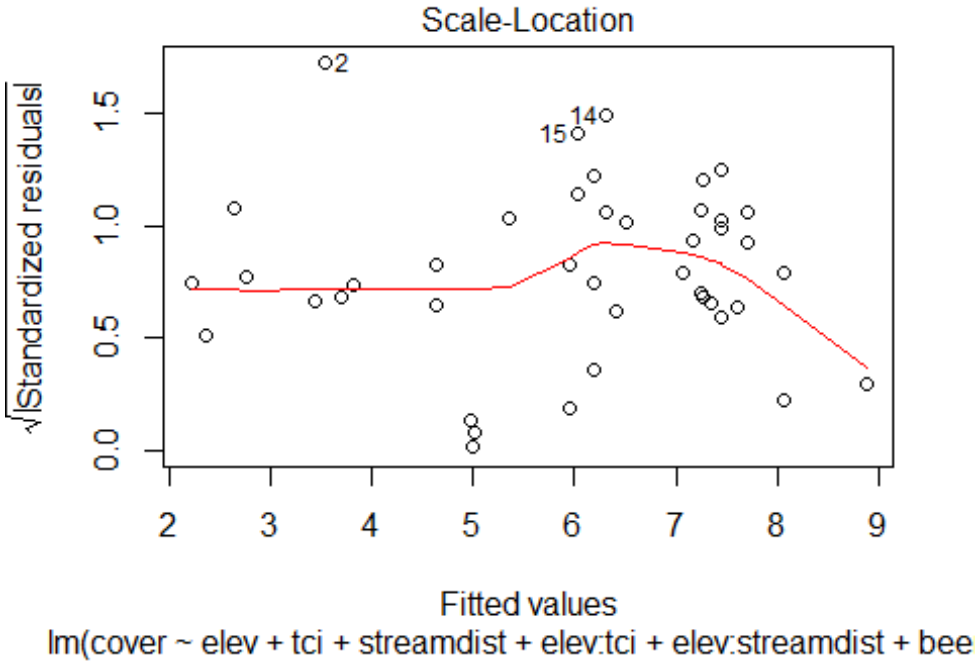

Scale-Location
√|Standardized residuals|
Fitted values
lm(cover ~ elev + tci + streamdist + disturb + beers)

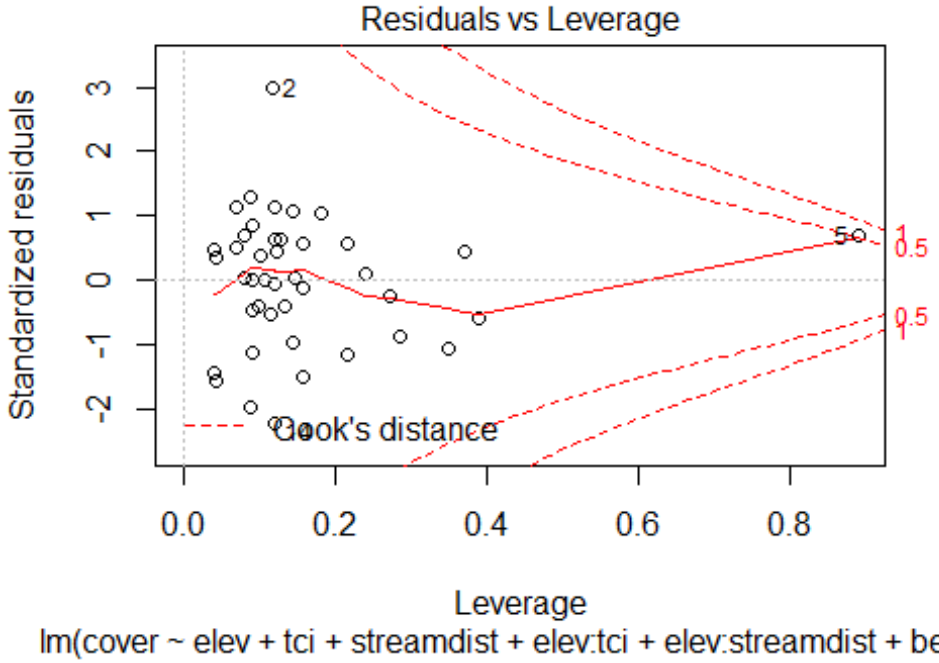

Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
lm(cover ~ elev + tci + streamdist + disturb + beers)

```
plot(abi_lm2)
```

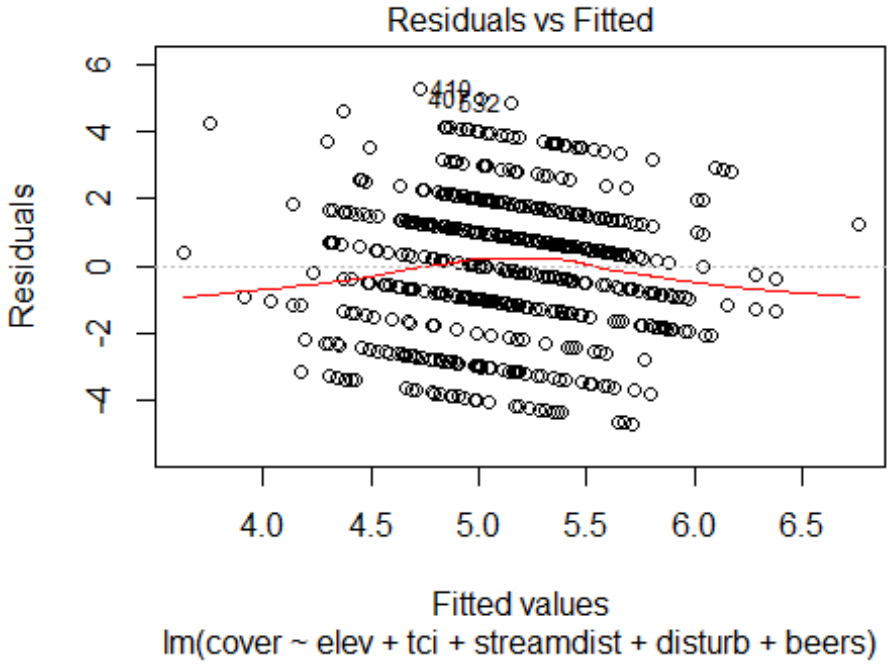

Residuals vs Fitted
Residuals
Fitted values
lm(cover ~ elev + tci + streamdist + elev:tci + elev:streamdist + bee
2
15
14

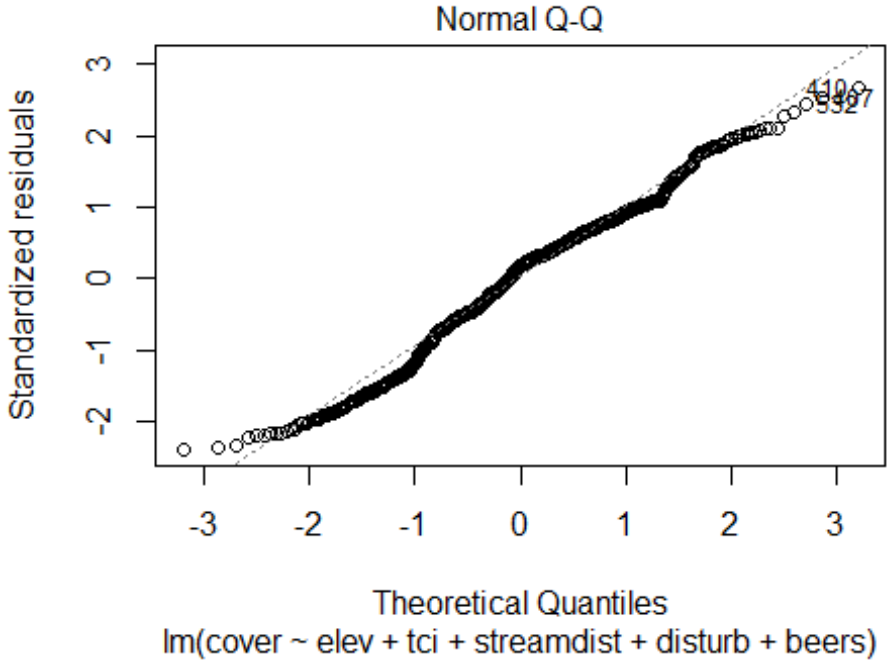

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(cover ~ elev + tci + streamdist + elev:tci + elev:streamdist + bee
2
15
14

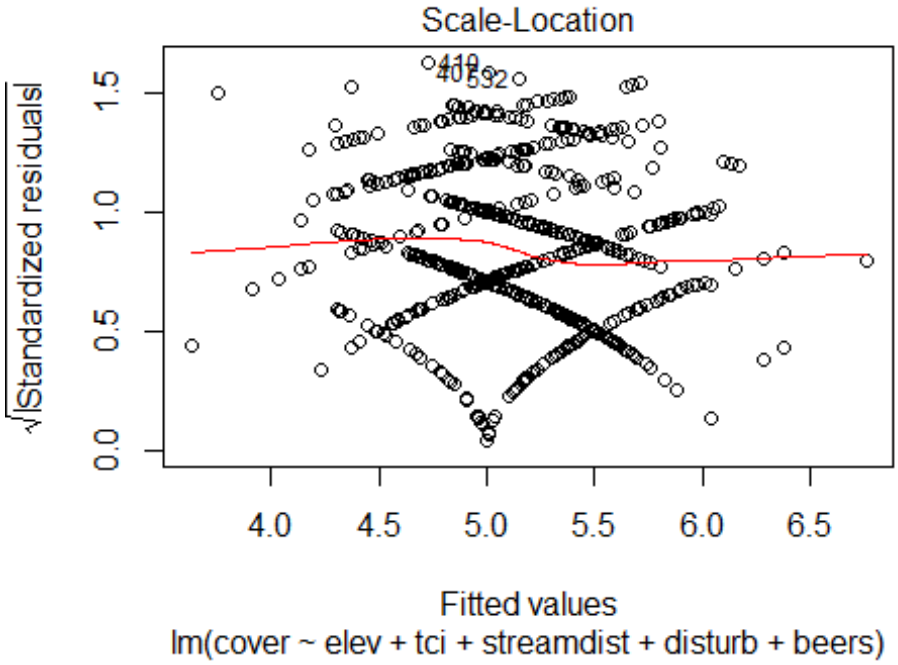


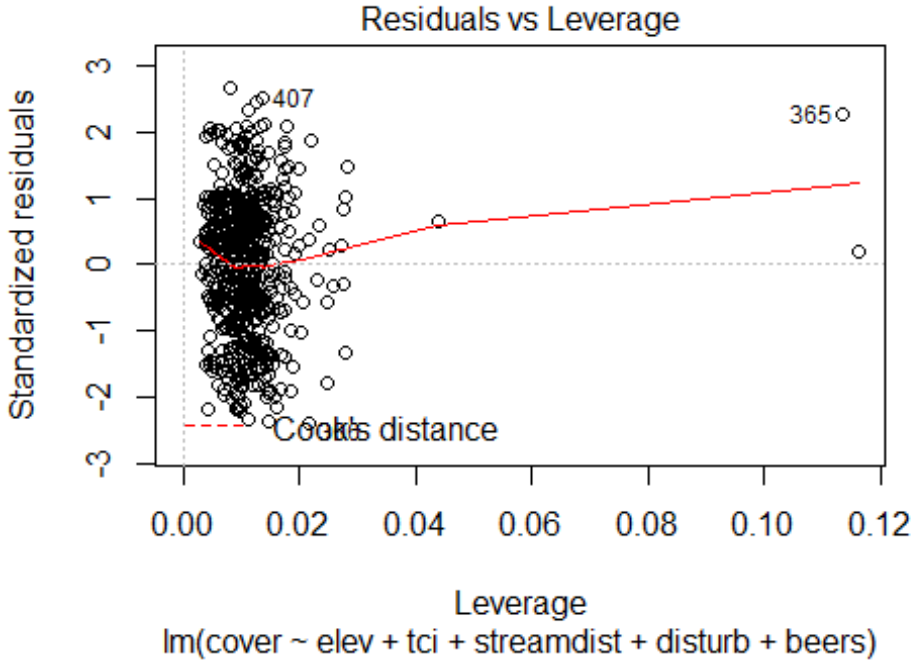



```
#MAPLE
plot(ace_lm1)
```

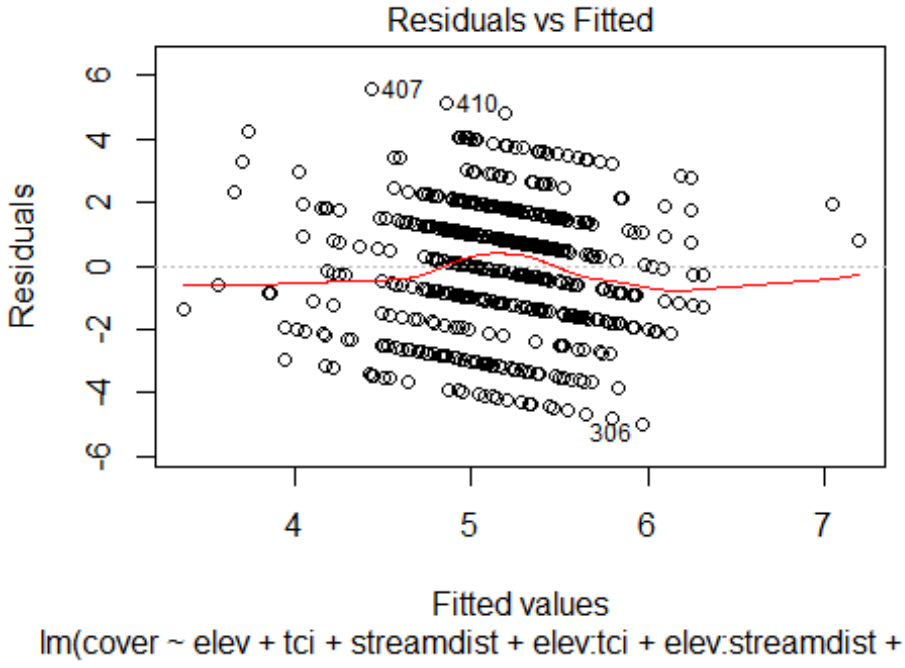

Residuals vs Fitted
Residuals
Fitted values
lm(cover ~ elev + tci + streamdist + disturb + beers)

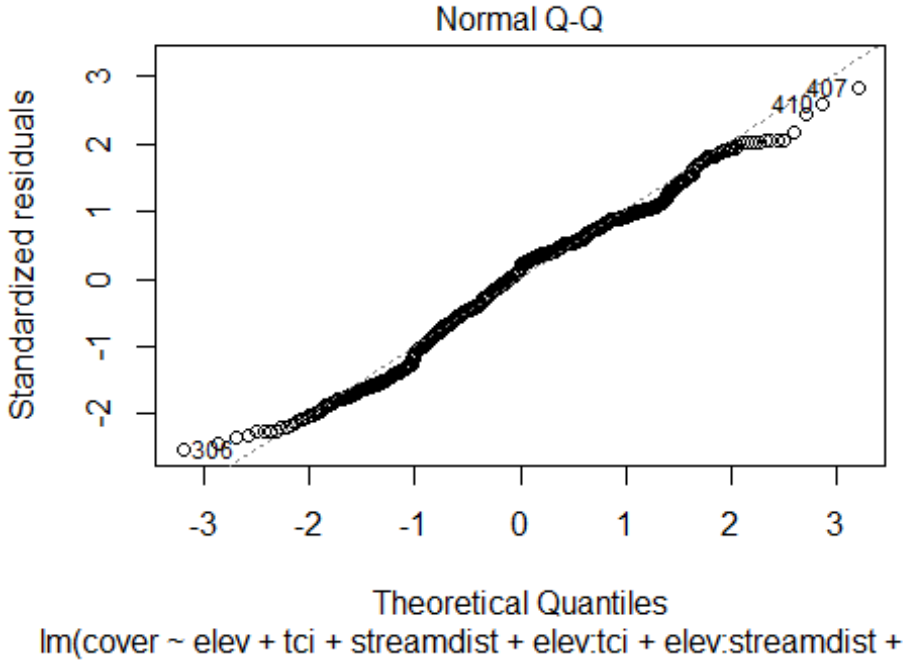

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(cover ~ elev + tci + streamdist + disturb + beers)

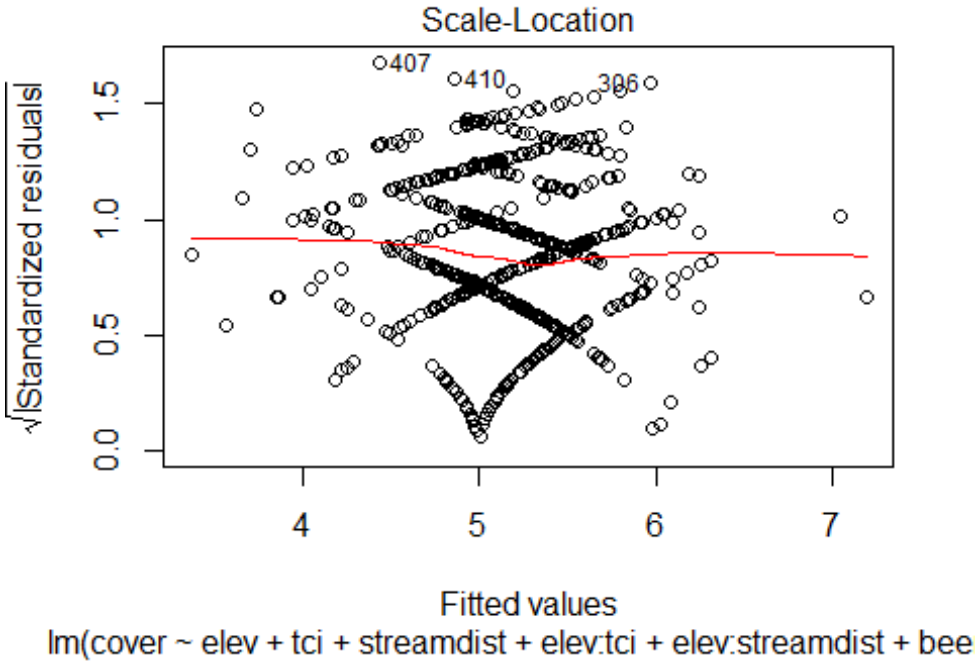

Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
4.0
4.5
5.0
5.5
6.0
6.5
Fitted values
lm(cover ~ elev + tci + streamdist + disturb + beers)

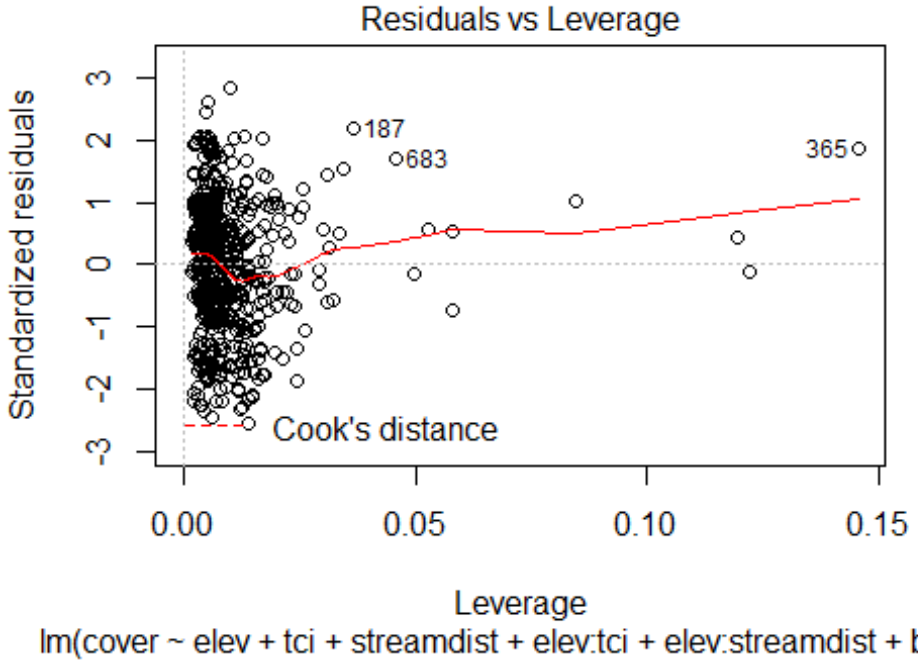

Residuals vs Leverage
Standardized residuals
-3
-2
-1
0
1
2
3
407
365
Cook's distance
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Leverage
lm(cover ~ elev + tci + streamdist + disturb + beers)

```
plot(ace_lm2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(cover ~ elev + tci + streamdist + elev:tci + elev:streamdist + bee
407
410
306


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(cover ~ elev + tci + streamdist + elev:tci + elev:streamdist + bee
407
410
306





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

The different tree species returned very different R-squared values. For fir cover r-squared = 0.5, which means 50% of the variance can be explained by our model. This is pretty good for ecological data. Conversely, maple returned a very low R-squared value of 0.04. Frasier fir is a habitat specialist, only growing on the tallest peaks of the Southern Appalachians; therefore, elevation explains cover very well. Red maple is a habitat generalist, and can tolerate a wide range of elevations and growing conditions accounting for the large amount of unexplained variance in the model.

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 approximation. Your new model calls will look as follows:

```
#Create glm model with poisson distrubution using same parameters as my
second lm for each of the species.
abi_poi <- glm(cover ~ elev + tci + streamdist + elev:tci + elev:streamdist +
beers, data = trees_abi, family = 'poisson')
ace_poi <- glm(cover ~ elev + tci + streamdist + elev:tci + elev:streamdist +
beers, data = trees_ace, family = 'poisson')

#write function to calculate pseudo r squared value from deviance/null
deviance.
pseudo_r2 <- function(my_glm) {
    (1 - my_glm$deviance/my_glm$null.deviance)
}

#compare the two methods
paste("Pseudo R-squared from GLM, Fir")

## [1] "Pseudo R-squared from GLM, Fir"

pseudo_r2(abi_poi)

## [1] 0.5572911

paste("Adjusted R-squared from LM, Fir")

## [1] "Adjusted R-squared from LM, Fir"

abi_lm2_summ <- summary(abi_lm2)
abi_lm2_summ$adj.r.squared

## [1] 0.5071111

paste("Pseudo R-squared from GLM, Maple")

## [1] "Pseudo R-squared from GLM, Maple"

pseudo_r2(ace_poi)
```

```
## [1] 0.04586806

paste("Adjusted R-squared from LM, Maple")

## [1] "Adjusted R-squared from LM, Maple"

ace_lm2_summ <- summary(ace_lm2)
ace_lm2_summ$adj.r.squared

## [1] 0.04321497
```

2.1. 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 fir and maple models, using the pseudo r-squared calculation and poisson distributed glm function increased the r-value. This suggests that this technique was able to explain more variation/error within the models. However, While the fir model maintained a decent r-squared, the maple models still show a weak r-squared.

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?

The results of my analyses support my initial hypotheses – that frasier fir cover will be strongly coorelated with elevation whereas red maple cover will be hard to explain by any one variable. Frasier fir is a habitat specialist restricted to the tallest peaks of the Southern Appalachians. Red maple, is a habitat generalist and tolerates a variety of growing conditions. While red maple cover is influenced by its habitat, it will be harder to detect any linear trend with the given data.

4. (optional) Examine the behavior of the function stepAIC() using the exploratory models developed above. This is a very simple and not very robust machine learning stepwise algorithm that uses AIC to select a best model. By default it does a backward selection routine.

```
#create models with all combinations of variables to run with stepAIC()
abi_lm_all <- lm(cover ~ elev * tci * streamdist * beers * disturb, data =
trees_abi)
ace_lm_all <- lm(cover ~ elev * tci * streamdist * beers * disturb, data =
trees_ace)

#run stepAIC() on these models to chose a model with the lowest AIC
#stepAIC(abi_lm_all)
#stepAIC(ace_lm_all)

#The lowest AIC produced for the fir data was for the model: 'AIC=46.26
(cover ~ elev + tci + streamdist + beers + elev:tci + elev:streamdist +
tci:streamdist + elev:beers + tci:beers + streamdist:beers + elev:tci:beers +
tci:streamdist:beers)"
```

```
#The lowest AIC produced for the maple data was for the model: 'AIC=952.1
(cover ~ elev + tci + streamdist + beers + disturb + elev:tci +
elev:streamdist + tci:streamdist + elev:beers + tci:beers + streamdist:beers
+ elev:disturb + tci:disturb + streamdist:disturb + beers:disturb +
elev:tci:streamdist + tci:streamdist:beers + elev:streamdist:disturb +
tci:streamdist:disturb + tci:beers:disturb + streamdist:beers:disturb +
tci:streamdist:beers:disturb)'


#While this technique may have dropped the AIC a few points, it failed to
produce a simple model. This shows the limitations of this type of model
selection.
```

5. (optional) Develop a model for the number of species in each site (i.e., unique plotID). This variable will also be discrete so the Poisson may be a good starting approximation. Side note: the Poisson distribution converges asymptotically on the Gaussian distribution as the mean of the distribution increases. Thus Poisson regression does not differ much from traditional OLS when means are large.

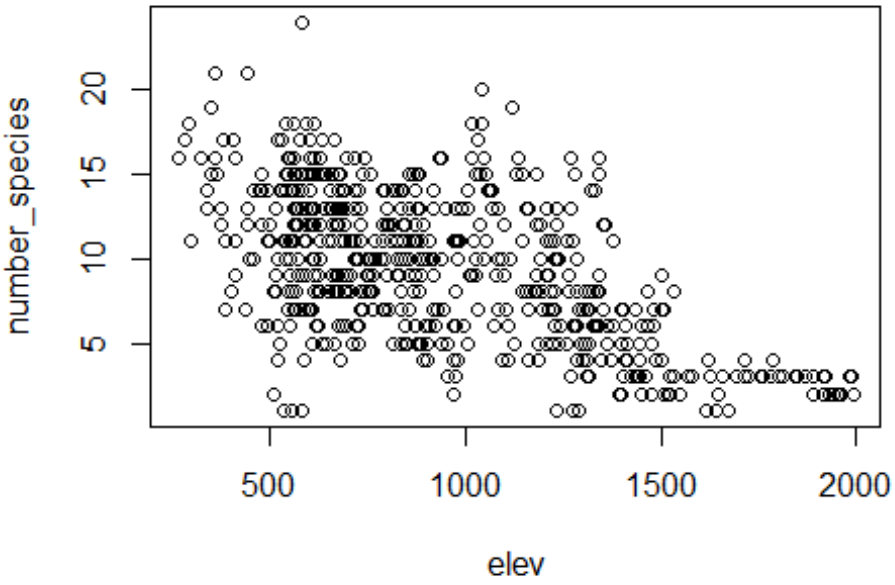
```
#I wasn't sure how to approach this, so I looked online and found the 'dplyr'
package. I wanted to model the species per plot as it relates to elevation.
My hypothesis is that species per plot decrease with elevation. My plot
appears to confirm this, but the resulting r-squared (using glm function from
previous question) was only 0.3. It appears that this relationship is not
linear, if I had more time I'd explore other regression types.

elev_spp <-   trees %>%
              group_by(elev) %>%
              summarize(number_species = n_distinct(species))

head(elev_spp)

## # A tibble: 6 x 2
##    elev number_species
##   <dbl>          <int>
## 1  267.             16
## 2  284.             17
## 3  294              18
## 4  302.             11
## 5  326.             16
## 6  340.             13

plot(elev_spp)
```




```
elev_spp_glm <- glm(number_species ~ elev, data = elev_spp, family =
'poisson')

pseudo_r2(elev_spp_glm)

## [1] 0.3125775
```