02/01/2022

trees <- read.csv('https://raw.githubusercontent.com/dmcglinn/quant\_methods/gh-pages/data/treedata\_subset.csv')

## Univariate Assignment

install.packages(“ggplot2”) install.packages(“gridExtra”) install.packages(“scatterplot3d”) install.packages(“MASS”) install.packages(“ggplot”) install.packages(“reshape2”)

# 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'

* plotID: unique code for each spatial unit (note some sampled more than once)
* date: when species occurrence recorded
* plotsize: size of quadrat in m2
* spcode: unique 7-letter code for each species
* species: species name
* cover: local abundance measured as estimated horizontal cover (ie, relative area of shadow if sun is directly above) classes 1-10 are: 1=trace, 2=0-1%, 3=1-2%, 4=2-5%, 5=5-10%, 6=10-25%, 7=25-50%, 8=50-75%, 9=75-95%, 10=95-100%
* utme: plot UTM Easting, zone 17 (NAD27 Datum)
* utmn: plot UTM Northing, zone 17 (NAD27 Datum)
* elev: elevation in meters from a digital elevation model (10 m res)
* tci: topographic convergence index, or site “water potential”; measured as the upslope contributing area divided by the tangent of the slope angle (Beven and Kirkby 1979)
* streamdist: distance of plot from the nearest permanent stream (meters)
* disturb: plot disturbance history (from a Park report); CORPLOG=corporate logging; SETTLE=concentrated settlement, VIRGIN=“high in virgin attributes”, LT-SEL=light or selective logging
* beers: transformed slope aspect (‘heat load index’); 0 is SW (hottest), 2 is NE (coolest)

Above shows a map of the regional and local location of the elevational transects included in the dataset (from [Fridley 2009](http://plantecology.syr.edu/fridley/Fridley2009_jamc.pdf)).

1. Carry out an exploratory analysis using the tree dataset. Metadata for the tree study can be found [here](../data/tree_metadata.txt). Specifically, I would like you to develop and compare models for species cover for a habitat generalist [*Acer rubrum* (Red maple)](http://www.durhamtownship.com/blog-archives/pix/November1407.jpg) and a habitat specialist [*Abies fraseri* (Frasier fir)](https://upload.wikimedia.org/wikipedia/commons/d/d0/Abies_fraseri_Mitchell.jpg). Because this dataset includes both continuous and discrete explanatory variables use the function Anova in the packages car as such

library(car)

## Loading required package: carData

acer\_mod = lm(acer$cover ~ acer$elev + acer$tci + acer$streamdist + acer$disturb + acer$beers, data = trees)  
Anova(acer\_mod, type = 3)

## Anova Table (Type III tests)  
##   
## Response: acer$cover  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 1845.7 1 295.0456 < 2.2e-16 \*\*\*  
## acer$elev 664.1 1 106.1624 < 2.2e-16 \*\*\*  
## acer$tci 55.8 1 8.9257 0.002907 \*\*   
## acer$streamdist 10.8 1 1.7340 0.188316   
## acer$disturb 44.1 3 2.3479 0.071433 .   
## acer$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)

##   
## Call:  
## lm(formula = acer$cover ~ acer$elev + acer$tci + acer$streamdist +   
## acer$disturb + acer$beers, data = trees)  
##   
## 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 \*\*\*  
## acer$elev -0.0034639 0.0003362 -10.304 < 2e-16 \*\*\*  
## acer$tci -0.1317294 0.0440921 -2.988 0.00291 \*\*   
## acer$streamdist 0.0007520 0.0005711 1.317 0.18832   
## acer$disturbLT-SEL -0.4379126 0.2559816 -1.711 0.08756 .   
## acer$disturbSETTLE -0.9309789 0.3564239 -2.612 0.00919 \*\*   
## acer$disturbVIRGIN -0.3601527 0.2941812 -1.224 0.22125   
## acer$beers -0.4101716 0.1381555 -2.969 0.00309 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.501 on 726 degrees of freedom  
## Multiple R-squared: 0.1805, Adjusted R-squared: 0.1726   
## F-statistic: 22.85 on 7 and 726 DF, p-value: < 2.2e-16

abies\_mod = lm(abies$cover ~ abies$elev + abies$tci + abies$streamdist + abies$disturb + abies$beers, data = trees)  
Anova(abies\_mod, type = 3)

## Anova Table (Type III tests)  
##   
## Response: abies$cover  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 59.64 1 70.4167 2.501e-16 \*\*\*  
## abies$elev 98.13 1 115.8739 < 2.2e-16 \*\*\*  
## abies$tci 2.11 1 2.4895 0.11505   
## abies$streamdist 3.98 1 4.6951 0.03057 \*   
## abies$disturb 28.40 3 11.1771 3.545e-07 \*\*\*  
## abies$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)

##   
## Call:  
## lm(formula = abies$cover ~ abies$elev + abies$tci + abies$streamdist +   
## abies$disturb + abies$beers, data = trees)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.5809 -0.4502 -0.0420 0.2346 6.9968   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.5705179 0.1871566 -8.391 2.50e-16 \*\*\*  
## abies$elev 0.0013315 0.0001237 10.764 < 2e-16 \*\*\*  
## abies$tci 0.0255969 0.0162231 1.578 0.115046   
## abies$streamdist 0.0004553 0.0002101 2.167 0.030574 \*   
## abies$disturbLT-SEL 0.3248142 0.0941853 3.449 0.000596 \*\*\*  
## abies$disturbSETTLE 0.5408814 0.1311418 4.124 4.15e-05 \*\*\*  
## abies$disturbVIRGIN 0.5584644 0.1082404 5.159 3.20e-07 \*\*\*  
## abies$beers -0.0675883 0.0508326 -1.330 0.184059   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9203 on 726 degrees of freedom  
## Multiple R-squared: 0.2478, Adjusted R-squared: 0.2406   
## F-statistic: 34.17 on 7 and 726 DF, p-value: < 2.2e-16

library(car)  
Anova(my\_mod, type=3)

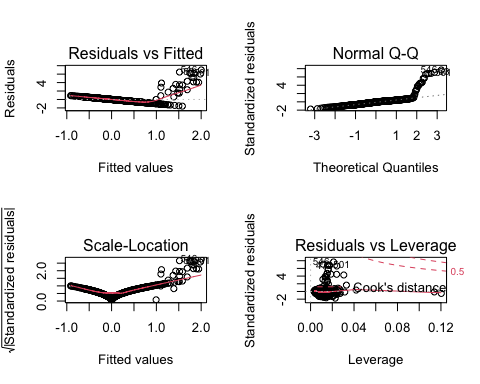
This will estimate partial effect sizes, variance explained, and p-values for each explanatory variable included in the model.

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

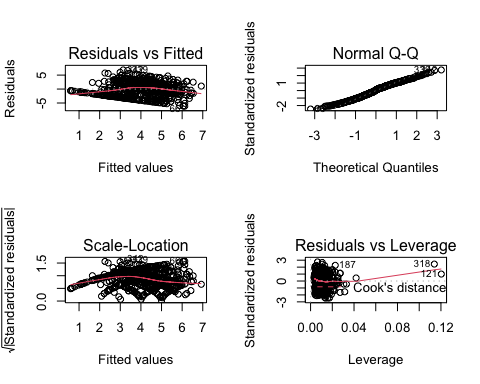
For each species address the following additional questions:

* how well does the exploratory model appear to explain cover? #The model explains the cover rather well in relation to the two species. The model shows that the total coverage for acer is a widespread dataframe, while abies (as expected from all the visual 0 values in the data) does not span far on the graphs at all. The summary function of course is much more straightforward and likely understandable by the average person who isn’t much of a computer wiz (me).

par(mfrow=c(2,2))  
plot(abies\_mod)



plot(acer\_mod)

 \* which explanatory variables are the most important? - the cover variables would be most important as it determines the total shade coverage of the area on the land. This further can help one understand the remaining variables as they all branch off of cover (such as tci, elev, etc…).

* do model diagnostics indicate any problems with violations of OLS assumptions?
* #yes, due to the curved trend lines the linear regression is not moving at a constant state which would more or less be required to fit the OLS which it doesn’t.
* are you able to explain variance in one species better than another, why might this be the case?
* #Clearly the abies variable had an average of near zero due to the large amount of zero covers in its data set. However, acer had much more of a variety therefor making it much easier to even view the data dispersal let alone determine variance.

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:  
  
  
```r  
acer\_poi = glm(acer$cover ~ acer$elev + acer$tci + acer$streamdist + acer$disturb + acer$beers, data = trees,  
 family = 'poisson')  
   
abies\_poi = glm(abies$cover ~ abies$elev + abies$tci + abies$streamdist + abies$disturb + abies$beers, data = trees,  
 family = 'poisson')   
summary(abies)

## cover elev tci streamdist   
## Min. :0.0000 Min. : 266.7 Min. : 2.610 Min. : 0.0   
## 1st Qu.:0.0000 1st Qu.: 651.0 1st Qu.: 4.528 1st Qu.: 78.1   
## Median :0.0000 Median : 874.5 Median : 5.317 Median :205.7   
## Mean :0.1866 Mean : 936.5 Mean : 5.813 Mean :229.8   
## 3rd Qu.:0.0000 3rd Qu.:1206.8 3rd Qu.: 6.277 3rd Qu.:340.1   
## Max. :9.0000 Max. :1992.0 Max. :25.000 Max. :957.5   
## disturb beers   
## Length:734 Min. :0.000106   
## Class :character 1st Qu.:0.410923   
## Mode :character Median :1.176209   
## Mean :1.092815   
## 3rd Qu.:1.758685   
## Max. :1.999999

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: 1660.1 on 726 degrees of freedom AIC: 3651.9

Number of Fisher Scoring iterations: 5

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

acer\_poi\_r2 <- (1 - acer\_poi$deviance/acer\_poi$null.deviance)  
print(acer\_poi\_r2)

## [1] 0.1342074

[1] 0.1342074

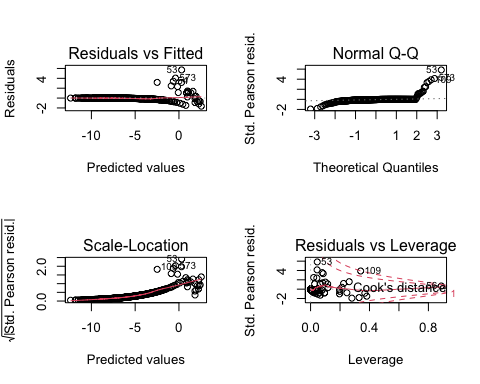
abies\_poi\_r2 <- (1 - abies\_poi$deviance/abies\_poi$null.deviance)  
print(abies\_poi\_r2)

## [1] 0.8951796

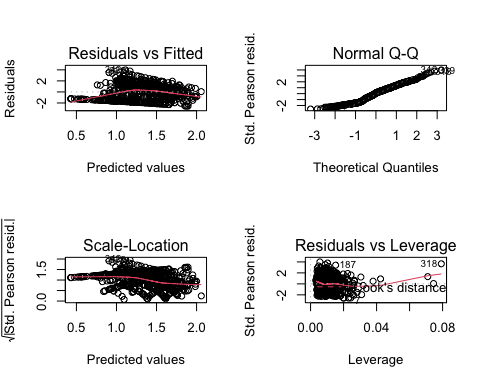
[1] 0.8951642

Compare your qualatitive assessment of which variables were most important in each model.

par(mfrow=c(2,2))  
plot(abies\_poi)



par(mfrow=c(2,2))  
plot(acer\_poi)



Does it appear that changing the error distribution changed the results much? In what ways?

#no it did not. The error did not change the overall results. Changing the numerical values to a smaller amount allowed the graph to shrink slightly. This however did not change overall data results of the experiment as it changed the physical representation on the graph.

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 analysis led to the conclusion that the fraser fir tree covers less area than the red maple did. This was discovered by measuring shadows from the trees as well as comparing the elevations, distribution, topogrophy, and slopes. Once the data was all measured it was easily concluded that the red maple tree had far greater cover in the environment than the fraser fir did. Showing anybody the graphs they would be able to tell a drastic difference in some sort of way regarldess of knowing the context.