Univariate Homework

Morgan Lattomus

2/1/2000

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). Because this dataset includes both continuous and discrete explanatory variables use the function Anova in the packages car as such

#install.packages('car')  
library(car)

## Loading required package: carData

#install.packages(ggplot2)  
library(ggplot2)  
#install.packages("tidyverse")  
#install.packages("funModeling")  
#install.packages("Hmisc")  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble 3.1.6 v dplyr 1.0.7  
## v tidyr 1.1.4 v stringr 1.4.0  
## v readr 2.1.1 v forcats 0.5.1  
## v purrr 0.3.4

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## x dplyr::recode() masks car::recode()  
## x purrr::some() masks car::some()

library(funModeling)

## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':  
##   
## src, summarize

## The following objects are masked from 'package:base':  
##   
## format.pval, units

## funModeling v.1.9.4 :)  
## Examples and tutorials at livebook.datascienceheroes.com  
## / Now in Spanish: librovivodecienciadedatos.ai

library(Hmisc)

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)

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'

site\_dat = merge(sp\_cov, env, by='plotID')

abies = site\_dat[ , c('ABIEFRA', cols\_to\_select)]  
acer = site\_dat[ , c('ACERRUB', cols\_to\_select)]  
names(abies)[1] = 'cover'  
names(acer)[1] = 'cover'

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

## Anova Table (Type III tests)  
##   
## Response: abies$cover  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 59.64 1 70.4167 2.501e-16 \*\*\*  
## abies$tci 2.11 1 2.4895 0.11505   
## abies$elev 98.13 1 115.8739 < 2.2e-16 \*\*\*  
## abies$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$cover ~ abies$tci + abies$elev + abies$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$tci 0.0255969 0.0162231 1.578 0.115046   
## abies$elev 0.0013315 0.0001237 10.764 < 2e-16 \*\*\*  
## 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

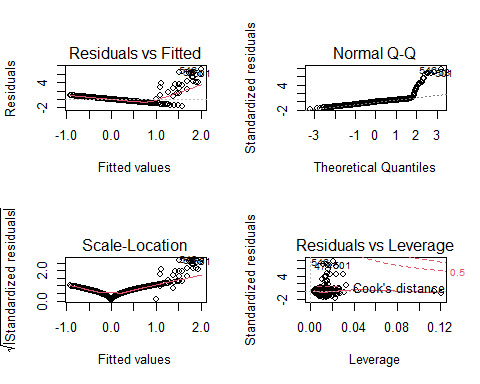
acer\_mod <- lm(acer$cover ~ acer$tci + acer$elev + acer$streamdist + acer$disturb + acer$beers, data = trees)  
Anova(acer\_mod, type = 3)

## Anova Table (Type III tests)  
##   
## Response: acer$cover  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 1845.7 1 295.0456 < 2.2e-16 \*\*\*  
## acer$tci 55.8 1 8.9257 0.002907 \*\*   
## acer$elev 664.1 1 106.1624 < 2.2e-16 \*\*\*  
## acer$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

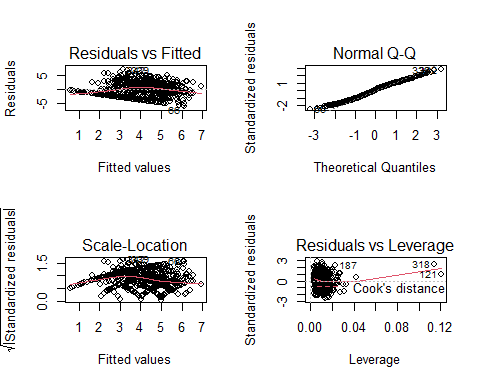
summary(acer\_mod)

##   
## Call:  
## lm(formula = acer$cover ~ acer$tci + acer$elev + acer$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$tci -0.1317294 0.0440921 -2.988 0.00291 \*\*   
## acer$elev -0.0034639 0.0003362 -10.304 < 2e-16 \*\*\*  
## 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

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



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



#Additional Questions #- How well does the exploratory model appear to explain cover?

# The fir tree has drastically less cluttered graphs, likely due to the fact that they have a lower abundance as a habitat specialist. The maple, however, has more data to include and work through on the graphs as they are more prominent due to their ability to function as a habitat generalist. Based off this data, we can conclude that the maple tree has more cover, and therefore a higher abundance than the fir tree.

#- Which explanatory variables are the most important?

# For this particular data set, the total tree cover over a particular area would be the most useful explanatory variable as the cover can be compared against over variables to determine the relationship between general abundance of trees versus the possible factor impacting a specialists’ or generalists’ ability to survive there.

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

# Yes, we can see that there is no constant linear relationship between any of the variables of the two species, which violates the first OLS assumption.

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

# The Fraser fir’s have little to no variance and the Red Maples having a wider range of variance. This variance can be explained better in the fir data, as the abundance would be expected to be lower as the tree would likely inhabit less area due to it being a habitat specialist.

1. Re-examine your solutions to the question above but from the perspective of a General Linear Model (GLM) with a Poisson error term.

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

##   
## Call:  
## glm(formula = acer$cover ~ acer$tci + acer$elev + acer$streamdist +   
## acer$disturb + acer$beers, family = "poisson", data = trees)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.6436 -1.4922 0.0961 0.8796 3.1435   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.593e+00 1.029e-01 25.211 < 2e-16 \*\*\*  
## acer$tci -3.415e-02 9.271e-03 -3.683 0.000230 \*\*\*  
## acer$elev -9.158e-04 7.139e-05 -12.828 < 2e-16 \*\*\*  
## acer$streamdist 1.351e-04 1.117e-04 1.209 0.226518   
## acer$disturbLT-SEL -1.254e-01 5.027e-02 -2.495 0.012603 \*   
## acer$disturbSETTLE -2.352e-01 6.801e-02 -3.458 0.000544 \*\*\*  
## acer$disturbVIRGIN -1.020e-01 6.301e-02 -1.619 0.105544   
## acer$beers -9.515e-02 2.725e-02 -3.491 0.000480 \*\*\*  
## ---  
## 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

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

##   
## Call:  
## glm(formula = abies$cover ~ abies$tci + abies$elev + abies$streamdist +   
## abies$disturb + abies$beers, family = "poisson", data = trees)  
##   
## 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 \*\*\*  
## abies$tci 1.688e-01 5.927e-02 2.848 0.00440 \*\*   
## abies$elev 7.850e-03 5.883e-04 13.343 < 2e-16 \*\*\*  
## abies$streamdist -1.692e-03 6.368e-04 -2.658 0.00787 \*\*   
## abies$disturbLT-SEL 1.622e+00 1.068e+00 1.518 0.12904   
## abies$disturbSETTLE 3.174e+00 1.161e+00 2.733 0.00628 \*\*   
## abies$disturbVIRGIN 2.649e+00 1.025e+00 2.584 0.00976 \*\*   
## abies$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

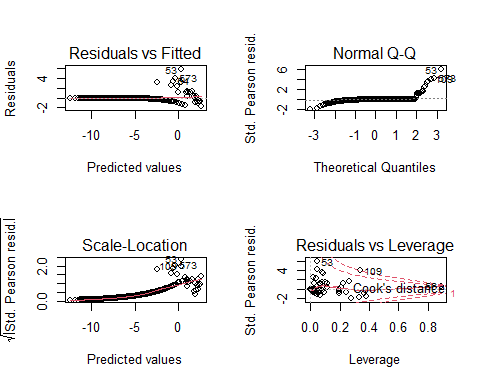
acer\_poi\_r2 = (1 - acer\_poi$deviance / acer\_poi$null.deviance)  
  
acer\_poi\_r2

## [1] 0.1342074

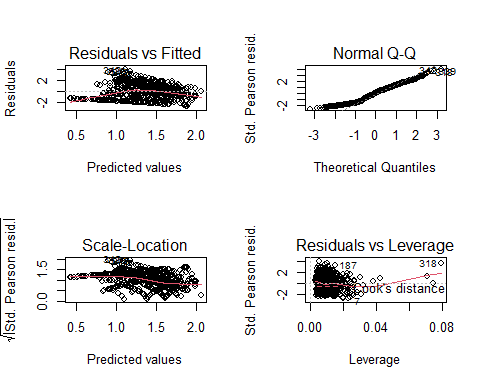
abies\_poi\_r2 = (1 - abies\_poi$deviance / abies\_poi$null.deviance)  
  
abies\_poi\_r2

## [1] 0.8951796

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?

# Changing the error distribution did not change much of the overall results, other than reducing the number of points on the graphical representations, making them marginally easier to read. The statistical analyses, however, did not changing along with the changing error distribution.

#3. Provide a plain English summary of what you have found and what conclusions we can take away from your analysis?

# Using the data provided, exploratory models were made which support the Fraser Fir’s role as a habitat specialist and the Red Maple’s role as a habitat generalist. The plots created using the predicted and fitted values show a higher amount of abundance in Red Maple and less abundance in the Fraser Fir. The total tree cover was determined to be the most important variable when analyzing this data set and its overall importance. Lastly, it was determined that changing the error distribution of the data did not drastically alter the representative data, either in the written statistical analyses or the graphical representation of the data sets.