Tree Model

Allison Conger

February 4, 2016

The goal of the assignment is to explore the data and create a model. Begin by importing the tree data:

```
treedata = read.csv('treedata_subset.csv')
```

Two subsets are created for the specific tree species Acer rubrum and Abies fraseri.

1. Creating a subset of data of the two tree species he wanted

```
Acer = subset(treedata, subset=species == 'Acer rubrum', select =
               c('cover', 'tci', 'beers', 'elev', 'streamdist','disturb'))

Abies = subset(treedata, subset=species == 'Abies fraseri', select =
               c('cover', 'tci', 'beers', 'elev', 'streamdist','disturb'))
```

Next a model is made comparing cover and all the other variables as independent variables. Including tci, beers, eleve, steamdist, and disturb.

```
#a model that combines all the other variables, for dataset acer
acermod = lm(cover ~ tci + beers + elev + streamdist + disturb, data=Acer)
#summarize the info
summary(acermod)
```

```
##
## Call:
## lm(formula = cover ~ tci + beers + elev + streamdist + disturb,
##     data = Acer)
##
## 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 ***
## tci           -0.0627613  0.0351922  -1.783  0.07495 .
## beers         -0.3269597  0.1089662  -3.001  0.00279 **
## elev          -0.0010108  0.0003161  -3.197  0.00145 **
## 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
## ---
## 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
```

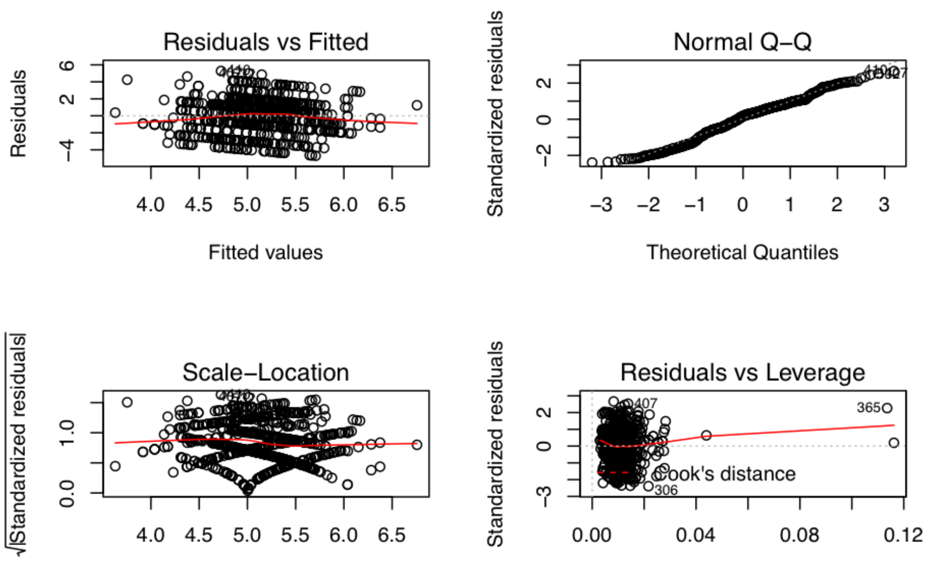
```r
#plot it
par(mfrow=c(2,2))
plot(acermod)

#abies model
abiesmod=lm(cover ~ tci + beers + elev + streamdist + disturb, data=Abies)

#info
summary(abiesmod)
```

```
##
## Call:
## lm(formula = cover ~ tci + beers + elev + streamdist + disturb,
##     data = Abies)
##
## 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 ***
## tci             0.287641   0.193467   1.487   0.1458
## beers           0.037551   0.500269   0.075   0.9406
## elev            0.012370   0.002523   4.903 2.02e-05 ***
## 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 .
## ---
## 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
```

```r
library(car)
```




```r
Anova(abiesmod, type=3)  #usefuller in some way
```

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

Our R value is fairly good at 0.5. big t values are the important ones.. f stat means somethings.The p-value is small but should be taken with a grain of salt as this is not a hard hypothesis testing.

For each species address the following additional questions:

- how well does the exploratory model appear to explain cover? Fairly well for the Abies model seeing as the R value is 0.5, however the Acer model is much lower.
- which explanatory variables are the most important? the variable with the highest t-value is elevatation for the Abies model, which is also true for the Acer model, however its a somewhat lower number.

- do model diagnostics indicate any problems with violations of OLS assumptions? The residuals seems higher for the Abies model (92.2)
- are you able to explain variance in one species better than another? the Acer model is a much larger sample set than the Abies model.

2.

Changing from a OLS model to a GLM model

```r
#acer glm model
acer_glm = glm(cover ~ tci + beers + elev + streamdist + disturb, data=Acer, family='poisson')
#summary
summary(acer_glm)
```

```
##
## Call:
## glm(formula = cover ~ tci + beers + elev + streamdist + disturb,
##     family = "poisson", data = Acer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4282  -0.5903   0.1391   0.5786   2.1038
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.873e+00  1.023e-01  18.315  < 2e-16 ***
## tci           -1.297e-02  8.159e-03  -1.589  0.11202
## beers         -6.391e-02  2.423e-02  -2.638  0.00834 **
## elev          -1.961e-04  7.047e-05  -2.783  0.00538 **
## streamdist     2.428e-04  1.030e-04   2.357  0.01843 *
## disturbLT-SEL  1.840e-02  4.880e-02   0.377  0.70619
## disturbSETTLE -1.739e-02  6.253e-02  -0.278  0.78099
## disturbVIRGIN  6.311e-02  5.638e-02   1.119  0.26293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 649.34  on 722  degrees of freedom
## Residual deviance: 623.38  on 715  degrees of freedom
## AIC: 3101.8
##
## Number of Fisher Scoring iterations: 4
```

```r
#abies glm model
abies_glm = glm(cover ~ tci + beers + elev + streamdist + disturb, data=Abies, family='poisson')
#summary
summary(abies_glm)
```

```
##
## Call:
## glm(formula = cover ~ tci + beers + elev + streamdist + disturb,
```

```
##     family = "poisson", data = Abies)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.47931  -0.35524   0.08027   0.36453   1.69535
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.1157009  1.5505526  -2.654  0.00795 **
## tci            0.0568868  0.0524222   1.085  0.27785
## beers         -0.0165548  0.1326724  -0.125  0.90070
## elev           0.0023508  0.0007292   3.224  0.00126 **
## streamdist    -0.0002186  0.0003969  -0.551  0.58176
## disturbLT-SEL  1.2440008  1.0827736   1.149  0.25060
## disturbSETTLE  1.0440232  1.1644892   0.897  0.36996
## disturbVIRGIN  1.4002993  1.0171140   1.377  0.16859
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 41.274  on 43  degrees of freedom
## Residual deviance: 16.126  on 36  degrees of freedom
## AIC: 189.3
##
## Number of Fisher Scoring iterations: 4
```

For assessing the degree of variation explained you can use a pseudo-R-squared statistic

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

Compare the residual sums of squares between the traditional OLS and glm models using anova (Note: not Anova) as such

```
anova(acermod, acer_glm)
```

```
## Analysis of Variance Table
##
## Model 1: cover ~ tci + beers + elev + streamdist + disturb
## Model 2: cover ~ tci + beers + elev + streamdist + disturb
##   Res.Df     RSS Df Sum of Sq F Pr(>F)
## 1    715 2828.21
## 2    715  623.38  0    2204.8
```

Does it appear that changing the error distribution changed the results much? In what ways? The RES. DF remainded the same while the RSS has lowered by a 10's place.

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 tree cover seems to be most influenced by the elevation of the trees. and The poisson model is more accurate than the OLS model. The abies model seems to be a more accurate model than the acer model.