Univariate Assignment

Caroline Oliver

1/30/2018

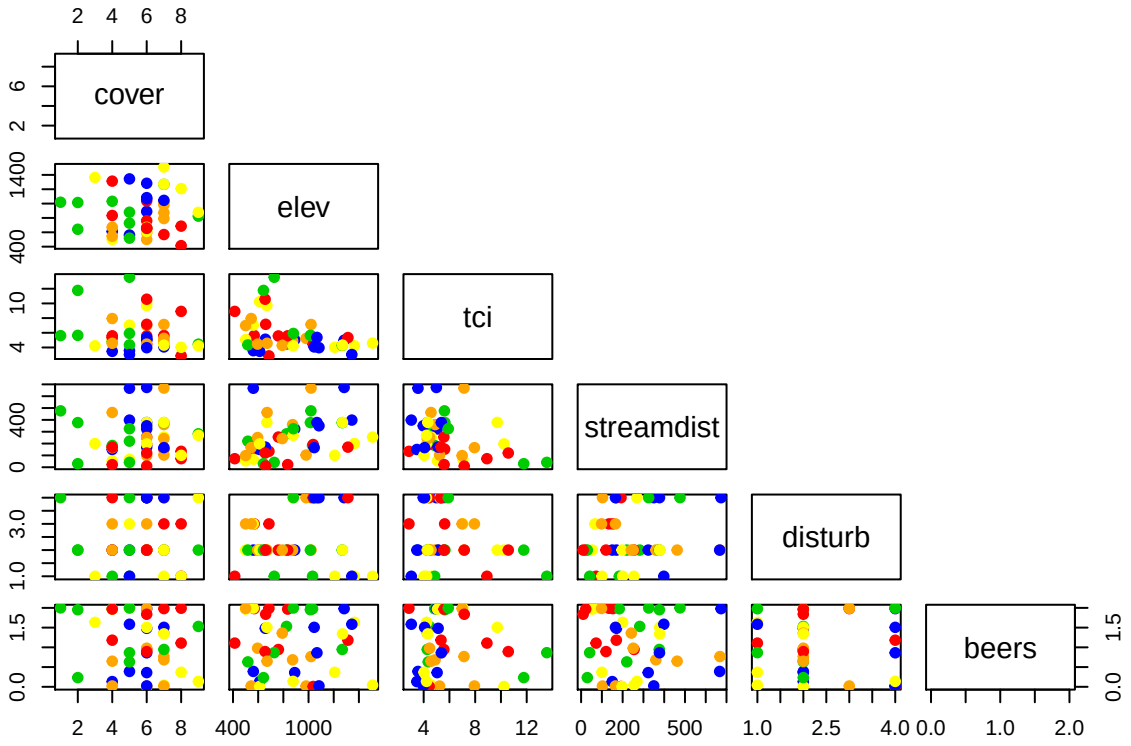
```
treeData = read.csv(file = "http://dmcglinn.github.io/quant_methods/data/treedata_subset.csv",
                    header = TRUE)

redMapleData = subset(treeData, species == "Acer rubrum")
frasierFirData = subset(treeData, species == "Abies fraseri")
```

Data Plots

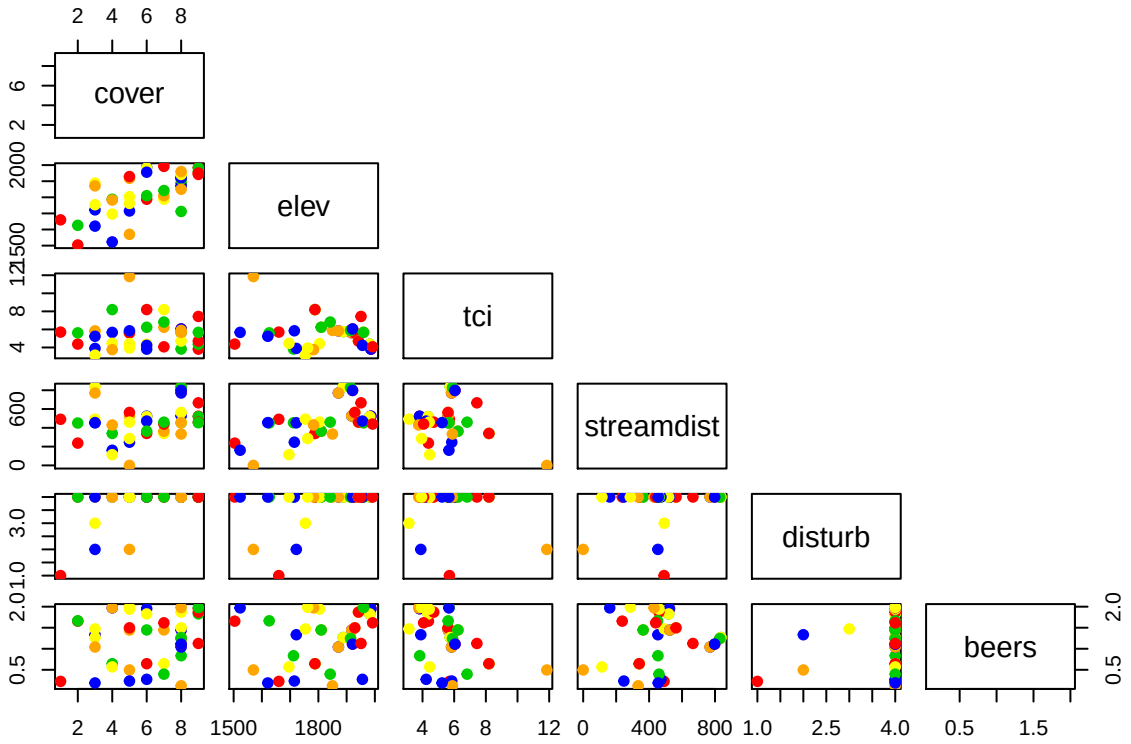
Pairs Plot for Red Maple Data

```
redSample <- sample(1:723, 44, replace=F)
pairs(redMapleData[redSample,4:9], pch = 19, col = c("red","green3","blue","yellow","orange"),
      upper.panel = NULL)
```



Pairs Plot for Fraiser Fir Data

```
pairs(frasierFirData[ ,4:9], pch = 19, col = c("red","green3","blue","yellow","orange"),
      upper.panel = NULL)
```



OLS Summary for Red Maple Data

```
redMapleOLS = lm(cover~elev + tci + streamdist + disturb + beers, data = redMapleData)
summary(redMapleOLS)
```

```
##
## Call:
## lm(formula = cover ~ elev + tci + streamdist + disturb + beers,
##     data = redMapleData)
##
## 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
```

OLS Summary for Fraiser Fir Data

```
frasierFirOLS = lm(cover~elev + tci + streamdist + disturb + beers, data = frasierFirData)
summary(frasierFirOLS)
```

```
##
## Call:
## lm(formula = cover ~ elev + tci + streamdist + disturb + beers,
##     data = frasierFirData)
##
## 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
```

Anova Summaires for all 3 Datasets

Red Maple Anova

```
anova(redMapleOLS)
```

```
## Analysis of Variance Table
##
## Response: cover
##             Df  Sum Sq Mean Sq F value   Pr(>F)
## elev         1   31.60  31.605  7.9900 0.004835 **
## tci          1   23.33  23.327  5.8972 0.015410 *
## streamdist   1   33.15  33.147  8.3800 0.003909 **
## disturb      3    9.35   3.117  0.7879 0.500831
## beers        1   35.61  35.613  9.0034 0.002789 **
## Residuals  715 2828.21   3.956
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Frasier Fir Anova

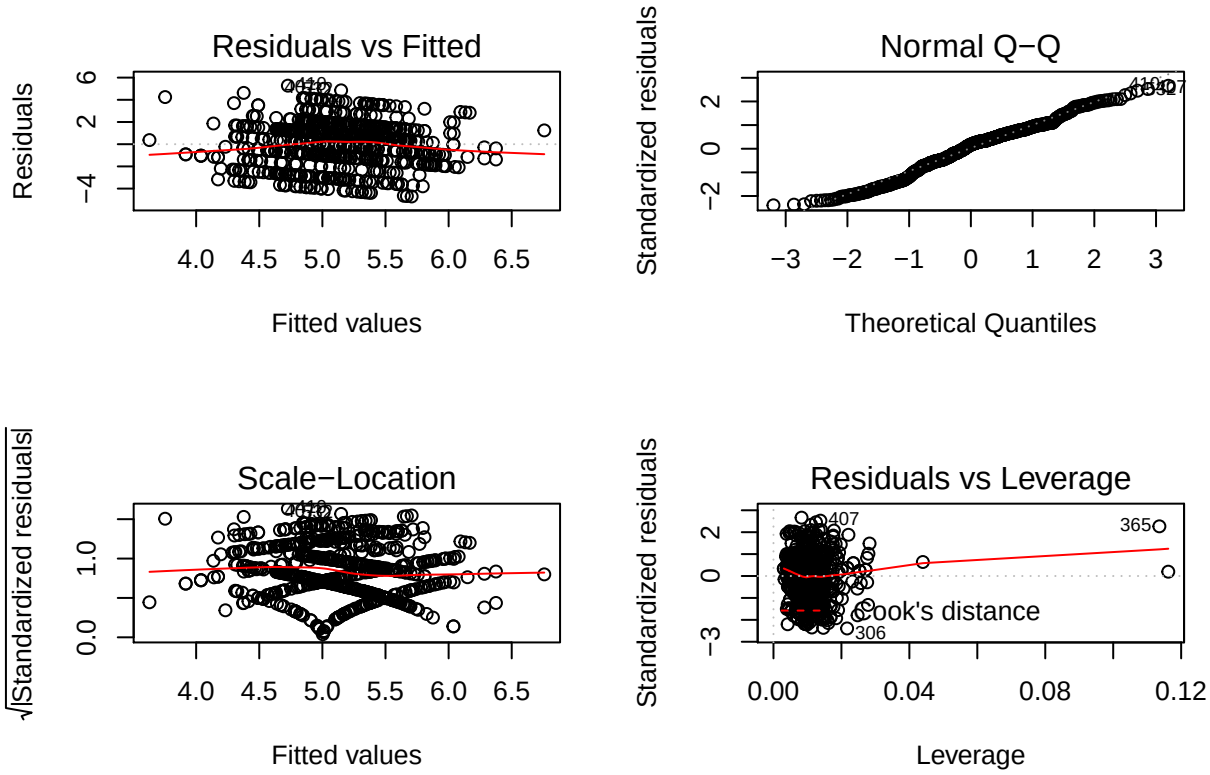
```
anova(frasierFirOLS)
```

```
## Analysis of Variance Table
##
## Response: cover
##            Df  Sum Sq Mean Sq F value  Pr(>F)
## elev        1 105.749 105.749 41.2502 1.9e-07 ***
## tci         1   9.239   9.239  3.6038 0.06569 .
## streamdist  1   3.007   3.007  1.1729 0.28601
## disturb     3  10.679   3.560  1.3886 0.26196
## beers       1   0.014   0.014  0.0056 0.94058
## Residuals  36  92.289   2.564
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear Model Plots

Red Maple Plots

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

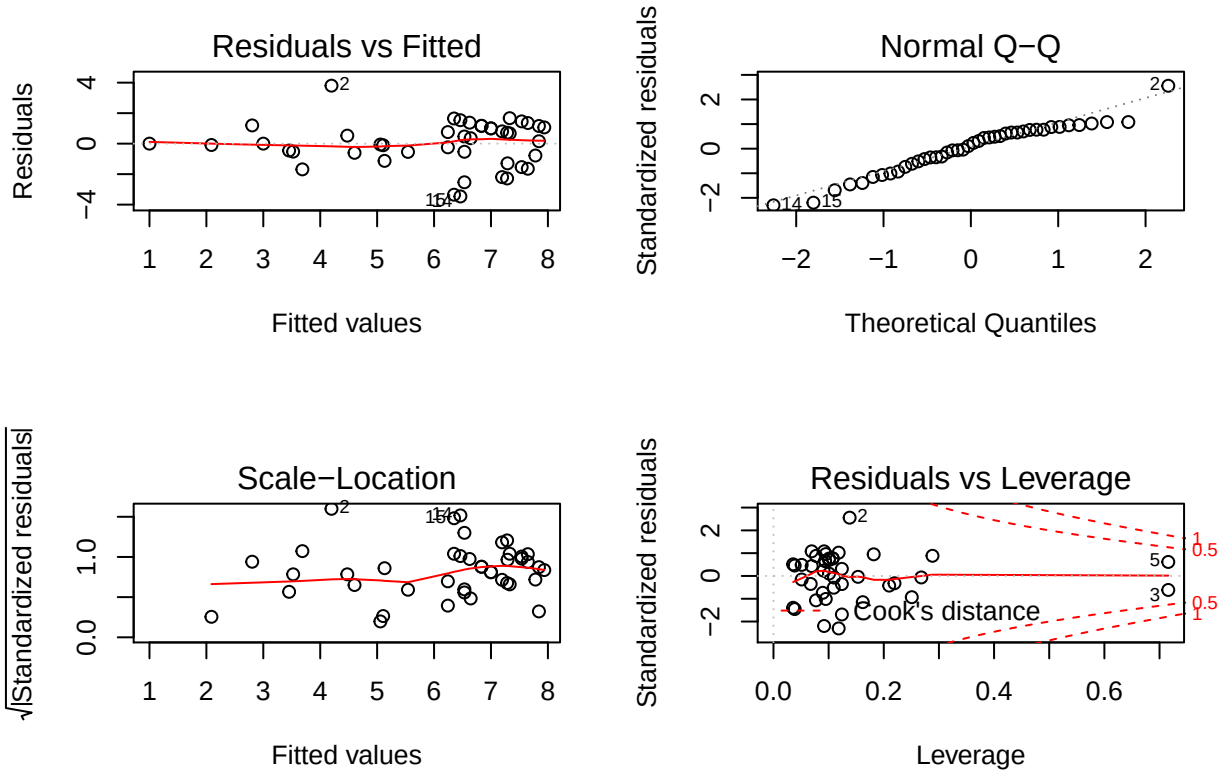


Frasier Fir Plots

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

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

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



Questions Set 1

How well does the exploratory model appear to explain cover?

To decide how welll the models explained cover we can look at the R squared values, F-statistic, and p-value. For the Red Maple Model the R squared values were much lower than that of the Frasier Fir Model which would seem to mean that the variance was better explained in the Red Maple Model than in the Frasier Fir Model. If you look at the F-statistics and p-values for the two models, the Frasier Fir Model had a better F-statistic at 7.171 over the Red Maple Model with an F-statistic value of 4.801. Similarly, the Frasier Fir Model had a better p-value score as well. However, it is necessary to note the difference in sample size between the two datasets. The Red Maple Model is modeled off of 723 values, where as the Frasier Fir Model is only modeled off of 44. Smaller models tend to require larger F-statistics than the larger models if the result is significant (i.e. there is a correlation between the preditor variables and the variable they are predicting) simply due to the smaller sample size. In the case of the Red Maple Model and the Frasier Fir Model, both seemed to do a relatively good job at explaining cover.

Which explanatory variables are the most important?

For the Red Maple Tree Data, the elevation stream distribution, and beers varaibles appear to be most important according to the anova performed on the Red Maple OLS model. All three variables had a significance code of 0.01. For the Frasier Fir Tree Data, elevation seems to be a very important varaible with a significance code of 0. The tci variable comes in as

second most important with a significance code of 0.1 which is not typically seen as statisically significant.

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

For both the Red Maple Tree Model and the Frasier Fir Tree Model, the linear model plots seem to conclude that the data does not violate the OLS assumptions of linearity which can be seen in the residuals vs fitted plot in the upper left hand corner, or mulitvariate normality which can be seen in the Normal Q-Q plot in the upper right hand corner. There is also little multicolinearity which can be seen in the original pairs plots for both the Red Maple and the Frasier Fir Datasets.

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

For variance we look at the adjusted R squared variable which for the Red Maple Data OLS is 0.03558 and for the Frasier Fir Data is 0.5011. This vast difference in adjusted R squared values can possibly be explained by the difference in sample size between the two datasets. The Red Maple Dataset has 723 values in it, while the Frasier Fir Dataset only has 44 values.

GLM Models for the 2 Datasets

Red Maple GLM

```
redMaple_glm = glm(cover~elev + tci + streamdist + disturb + beers, data= redMapleData,
                   family='poisson')
summary(redMaple_glm)
```

```
##
## Call:
## glm(formula = cover ~ elev + tci + streamdist + disturb + beers,
##     family = "poisson", data = redMapleData)
##
## 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 ***
## elev          -1.961e-04  7.047e-05  -2.783  0.00538 **
## tci           -1.297e-02  8.159e-03  -1.589  0.11202
## 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
## beers         -6.391e-02  2.423e-02  -2.638  0.00834 **
## ---
## 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
```

Fraiser Fir GLM

```
frasierFir_glm = glm(cover~elev + tci + streamdist + disturb + beers, data= frasierFirData,
                     family='poisson')
summary(frasierFir_glm)
```

```
##
## Call:
## glm(formula = cover ~ elev + tci + streamdist + disturb + beers,
##     family = "poisson", data = frasierFirData)
##
## 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 **
## elev           0.0023508  0.0007292   3.224  0.00126 **
## tci            0.0568868  0.0524222   1.085  0.27785
## 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
## beers         -0.0165548  0.1326724  -0.125  0.90070
## ---
## 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
```

Pseudo R Squared values for each GLM Model

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

Red Maple R Squared Value

```
pseudo_r2(redMaple_glm)
```

```
## [1] 0.03997917
```

Fraiser Fir R Squared Value

```
pseudo_r2(frasierFir_glm)
```

```
## [1] 0.60931
```

Anova Comparing Models

Anova Comparing Red Maple Models

```
anova(redMapleOLS, redMaple_glm)
```

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

Anova Comparing Fraiser Fir Models

```
anova(frasierFirOLS, frasierFir_glm)
```

```
## Analysis of Variance Table
##
## Model 1: cover ~ elev + tci + streamdist + disturb + beers
## Model 2: cover ~ elev + tci + streamdist + disturb + beers
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1     36 92.289
## 2     36 16.126  0    76.164
```

Question Set 2

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

Even though the Poisson GLM Model is more sophisticated, the results from the OLS Models and the GLM Models do not significantly differ. Even though the OLS Model is more robust, the significant vairables for both the Red Maple OLS Model and the Frasier Fir OLS Model are generally consistent with the significant varaibles for the Red Maple and Frasier Fir GLM Models with Poisson Distribution.

Question 3: Sumary & Conclusions

Similarly to what was said in Question Sets 1 & 2, the OLS Models for both the Red Maple Tree Data and the Frasier Fir Tree Data performed well given the size of the datasets. Even though the F-statistics and P-values differ between the two models and may appear to be better for one model than the other, both can be considered significant. Since the Frasier Fir Dataset only consisted of 44 data points, a larger F-statistic and smaller P-value, compared to the Red Maple Dataset, are needed to prove a significant relationship between the predictor variables and cover. On the other hand, the Red Maple Dataset represents a much larger sample size at 723 data points, so a slightly smaller F-statistic and P-value are still seen as significant and the model is deemed useful. Also to note, the more sophisticated GLM model with Poisson error distribution did not seem to provide any great difference in variable significance compared to the OLS Models for each of the datasets.