

```
trees <- read.csv('https://raw.githubusercontent.com/dmccglinn/quant_methods/gh-pages/data/treedata_subsets.csv')
str(trees)
```

```
## 'data.frame':      8038 obs. of  9 variables:
## $ plotID      : Factor w/ 734 levels "ATBN-01-0303",...: 20 53 54 56 109 188 452 471 471 471 ...
## $ spcode      : Factor w/ 52 levels "ABIEFRA","ACERNEG",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ species     : Factor w/ 51 levels "Abies fraseri",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ cover       : int   1 8 3 3 5 2 4 8 8 5 ...
## $ elev        : num   1660 1712 1722 1754 1570 ...
## $ tci         : num    5.7 3.82 3.89 3.15 11.85 ...
## $ streamdist  : num   491 454 453 492 0 ...
## $ disturb     : Factor w/ 4 levels "CORPLOG","LT-SEL",...: 1 4 2 3 2 4 4 4 4 4 ...
## $ beers       : num    0.224 0.834 1.333 1.471 0.496 ...
```

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* (Fraser fir).

```
summary(trees)
```

```
##           plotID           spcode           species
## UFRL-02-0160: 62  ACERRUB: 723  Acer rubrum      : 723
## UFRL-01-0090: 58  TSUGCAN: 653  Tsuga canadensis : 653
## UFRL-01-0091: 56  QUERRUB: 556  Quercus rubra   : 556
## UFRL-01-0093: 54  LIRITUL: 469  Liriodendron tulipifera: 469
## UFRL-02-0164: 54  NYSSSYL: 422  Nyssa sylvatica  : 422
## UFRL-02-0177: 54  MAGNFRA: 382  Magnolia fraseri : 382
## (Other)      :7700  (Other):4833  (Other)         :4833
##      cover      elev      tci      streamdist
## Min.   : 1.000   Min.   : 266.7   Min.   : 2.610   Min.   : 0.00
## 1st Qu.: 2.000   1st Qu.: 592.6   1st Qu.: 4.567   1st Qu.: 76.16
## Median : 4.000   Median : 791.4   Median : 5.254   Median :198.00
## Mean   : 3.959   Mean   : 849.8   Mean   : 5.840   Mean   :229.50
## 3rd Qu.: 6.000   3rd Qu.:1061.0   3rd Qu.: 6.418   3rd Qu.:340.60
## Max.   :10.000   Max.   :1992.0   Max.   :25.000   Max.   :957.50
##
##      disturb      beers
## CORPLOG:1501   Min.   :0.000106
## LT-SEL :3924   1st Qu.:0.317933
## SETTLE :1299   Median :1.089389
## VIRGIN :1314   Mean   :1.029020
##              3rd Qu.:1.685997
##              Max.   :1.999999
##
```

```
lm(trees$cover ~ trees$spcode)
```

```
##
## Call:
## lm(formula = trees$cover ~ trees$spcode)
##
## Coefficients:
##      (Intercept)  trees$spcodeACERNEG  trees$spcodeACERNIG
##              6.0227             -2.5227             -1.0227
## trees$spcodeACERRUB  trees$spcodeAILAALT  trees$spcodeBETUALL
##             -0.8899              2.9773             -1.1254
## trees$spcodeBETULEN  trees$spcodeBETUNIG  trees$spcodeCARYALB
##             -2.0817             -5.0227             -2.6727
## trees$spcodeCARYCOR  trees$spcodeCARYGLA  trees$spcodeCARYOVT
##             -3.4513             -2.6721             -2.3005
## trees$spcodeCASTDNT  trees$spcodeCATASPE  trees$spcodeCLADKEN
##             -3.6116             -5.0227             -2.8561
## trees$spcodeDIOSVIR  trees$spcodeFAGUGRA  trees$spcodeFRAXAME
##             -4.1656             -1.8824             -3.0755
## trees$spcodeFRAXPEN  trees$spcodeJUGLCIN  trees$spcodeJUGLNIG
##             -3.4602             -3.6894             -3.1532
## trees$spcodeJUNIVIR  trees$spcodeLIQUSTY  trees$spcodeLIRITUL
##             -4.2045             -2.0682             -1.9289
## trees$spcodeMAGNACU  trees$spcodeMAGNFRA  trees$spcodeMAGNTRI
##             -3.1111             -2.8630             -3.4672
## trees$spcodeNYSSSYL  trees$spcodePAULTOM  trees$spcodePICERUB
##             -2.2218             -4.6894             -1.6735
## trees$spcodePINUECH  trees$spcodePINUPUN  trees$spcodePINURIG
##             -2.5227             -1.1097             -1.7515
## trees$spcodePINUSTR  trees$spcodePINUVIR  trees$spcodePLATOCC
##             -2.3591             -1.7549             -2.0227
## trees$spcodeQUERALB  trees$spcodeQUERFLC  trees$spcodeQUERIMB
##             -2.1087             -2.2296             -2.5227
## trees$spcodeQUERMON  trees$spcodeQUERMUE  trees$spcodeQUERRUB
##             -0.9915             -3.7727             -2.0551
## trees$spcodeQUERSTE  trees$spcodeQUERVEL  trees$spcodeROBIPSE
##             -1.9227             -2.5874             -2.8824
## trees$spcodeSALINIG  trees$spcodeTILIAME  trees$spcodeTILIAMEH
##             -1.1894             -1.7627             -1.6290
## trees$spcodeTSUGCAN  trees$spcodeULMUALA  trees$spcodeULMUAME
##             -1.4148             -3.6894             -2.6894
## trees$spcodeULMURUB
##             -2.4513
```

#subset acer rubrum by selecting Acer rubrum then selecting columns of variables

```
acer <- subset(trees, subset=species=="Acer rubrum", select= c("cover", "tci", "elev", "beers", "disturb
```

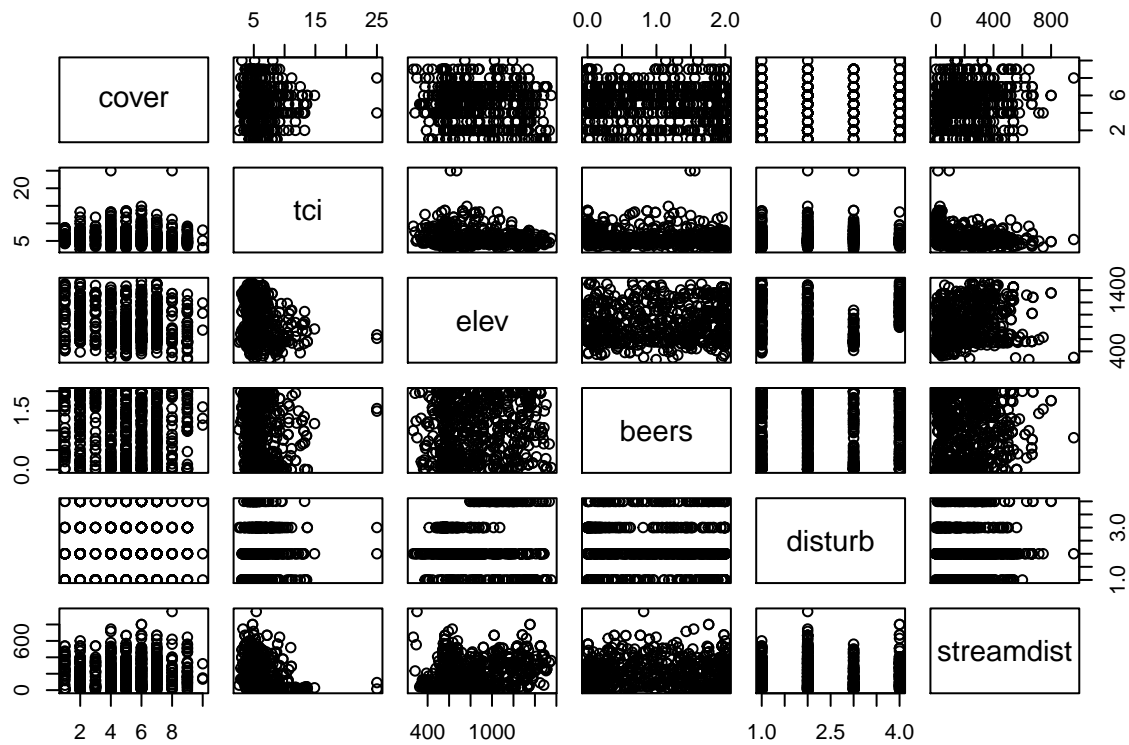
```
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor=3, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste0(prefix, txt)
  if(missing(cex.cor))
    cex.cor <- 0.8/strwidth(txt)
```

```

    text(0.5, 0.5, txt, cex = cex.cor)
}

plot(acer)

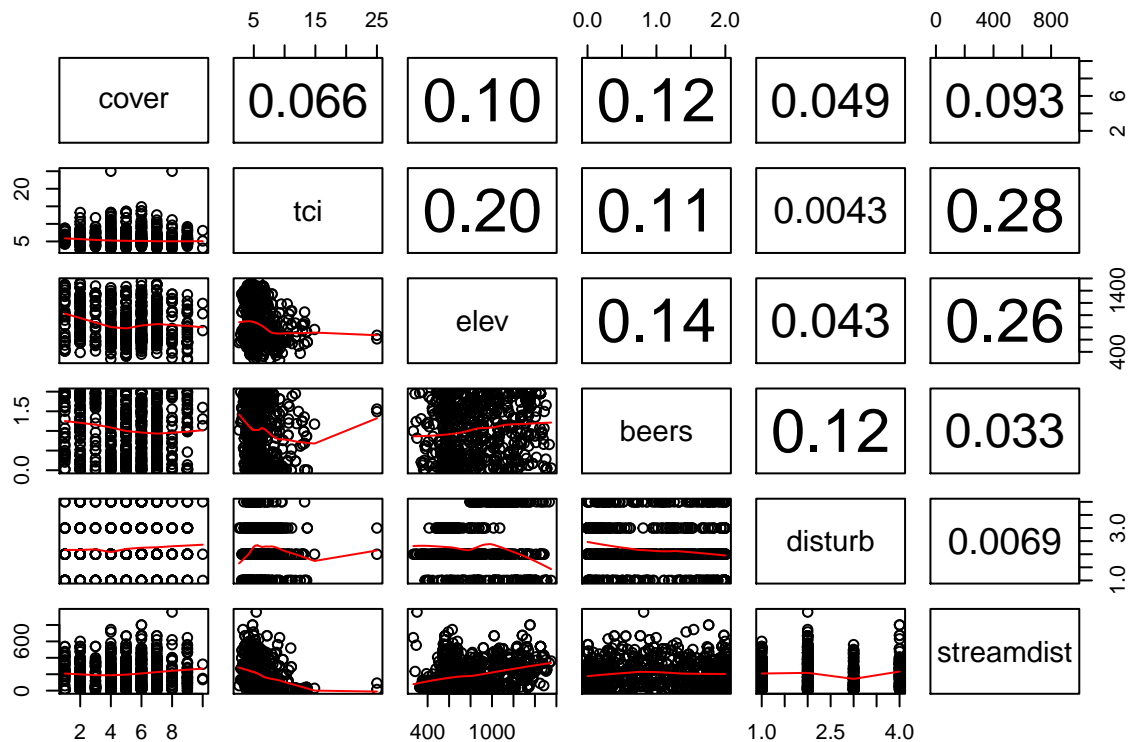
```



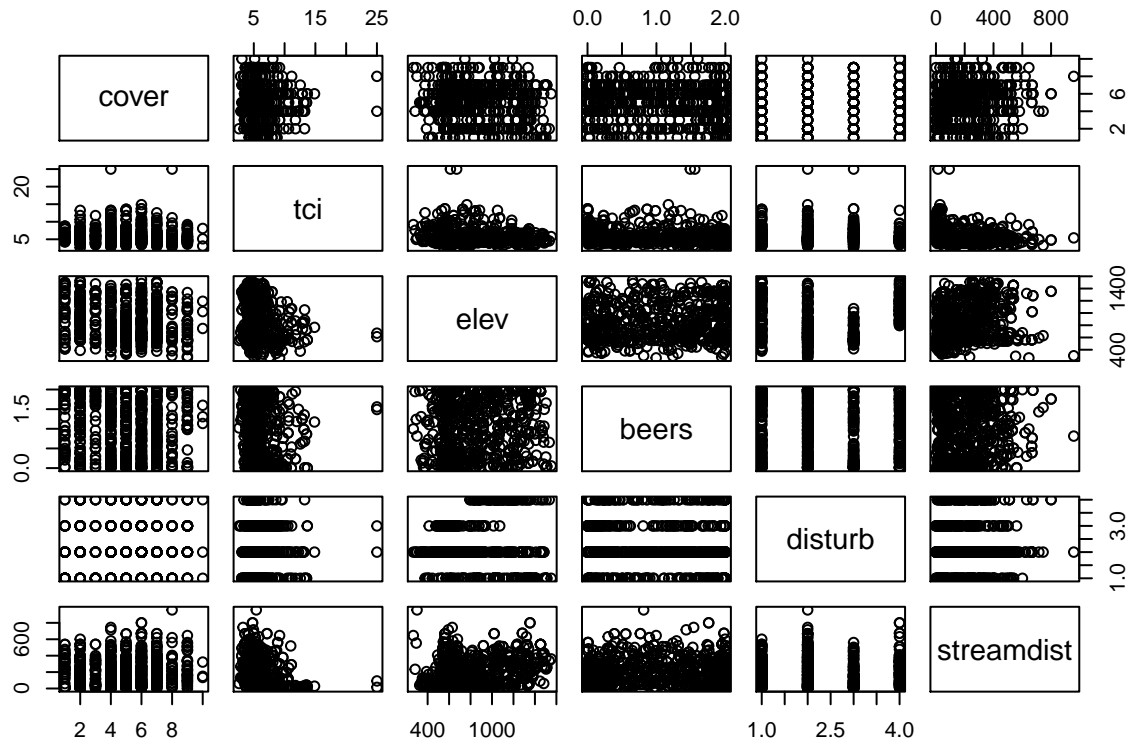
```

pairs(acer, lower.panel = panel.smooth, upper.panel = panel.cor)

```



```
plot(acer)
```



```
summary(acer)
```

```
##      cover      tci      elev      beers
##  Min.   : 1.000   Min.   : 2.610   Min.   : 266.7   Min.   :0.000106
## 1st Qu.: 4.000   1st Qu.: 4.564   1st Qu.: 601.9   1st Qu.:0.311609
## Median : 6.000   Median : 5.226   Median : 802.1   Median :1.068436
## Mean   : 5.133   Mean   : 5.837   Mean   : 857.6   Mean   :1.024309
## 3rd Qu.: 6.000   3rd Qu.: 6.408   3rd Qu.:1077.0   3rd Qu.:1.687648
## Max.   :10.000   Max.   :25.000   Max.   :1551.0   Max.   :1.999999
##      disturb      streamdist
## CORPLOG:147   Min.   : 0.00
## LT-SEL :343   1st Qu.: 68.54
## SETTLE :123   Median :194.20
## VIRGIN :110   Mean   :218.69
##           3rd Qu.:328.00
##           Max.   :957.50
```

#how well does the exploratory model appear to explain cover? It gives alot of information which is not
#which explanatory variables are the most important? The variables for Acer rubrum of importanace seem

```
acer.lm<-lm(cover ~ beers + elev + streamdist ,data= acer)
summary(acer.lm)
```

```
##
## Call:
## lm(formula = cover ~ beers + elev + streamdist, data = acer)
##
## Residuals:
```

```
##      Min      1Q  Median      3Q      Max
## -4.9339 -1.3307  0.2865  1.3691  5.2839
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.8119219  0.2420805  24.008  < 2e-16 ***
## beers       -0.3034710  0.1067192  -2.844  0.004587 **
## elev        -0.0008226  0.0002589  -3.178  0.001548 **
## streamdist   0.0015415  0.0004566   3.376  0.000774 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.992 on 719 degrees of freedom
## Multiple R-squared:  0.03688,    Adjusted R-squared:  0.03287
## F-statistic: 9.178 on 3 and 719 DF,  p-value: 5.775e-06
```

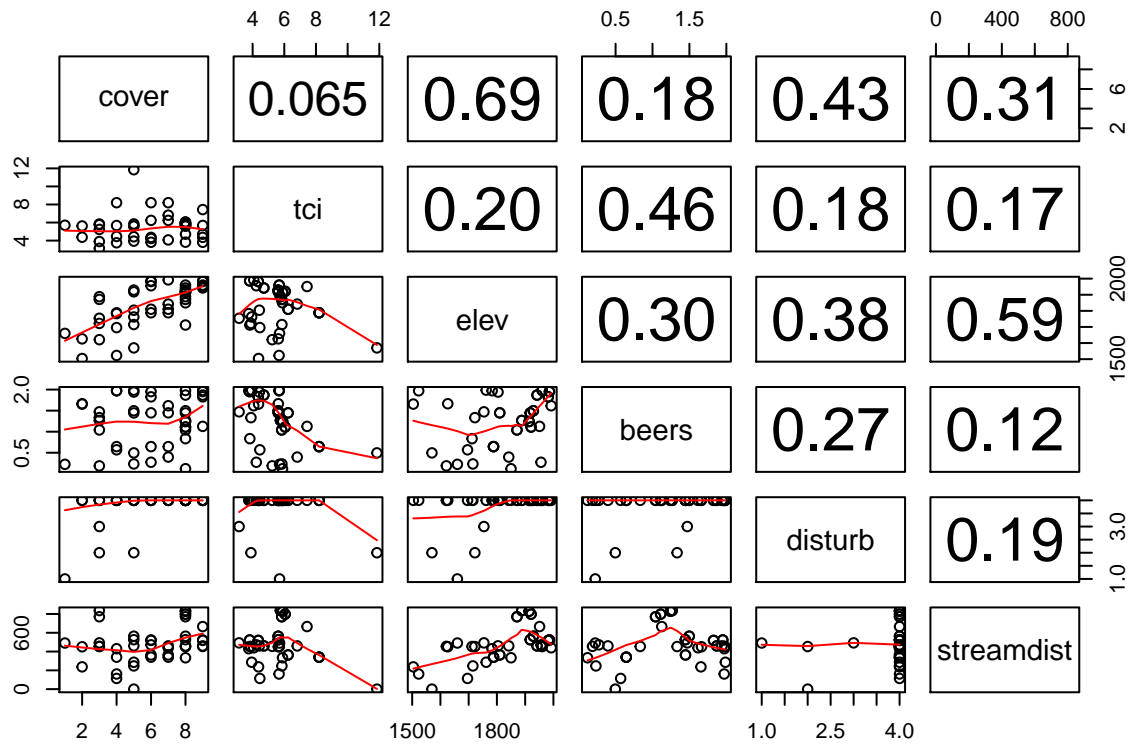
#The model explains 3.29% of the variance and has highly significant coefficients for the intercept and

#subset Abies fraseri by selecting Abies fraseri then selecting columns of variables

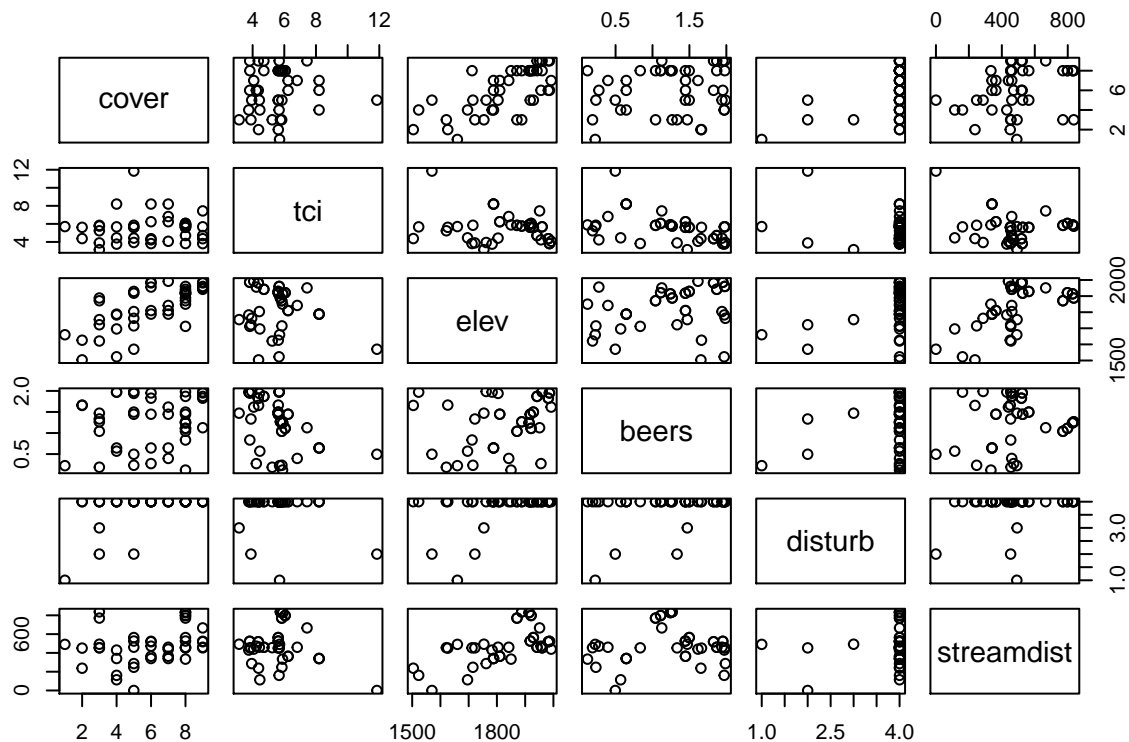
```
abies <- subset(trees,subset=species=="Abies fraseri", select= c("cover", "tci", "elev", "beers", "dist"))

panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor=3, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste0(prefix, txt)
  if(missing(cex.cor))
    cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor)
}

pairs(abies, lower.panel = panel.smooth, upper.panel = panel.cor)
```



```
plot(abies)
```



```
summary(abies)
```

```
##      cover      tci      elev      beers      disturb
##  Min.   :1.000   Min.   : 3.146   Min.   :1504   Min.   :0.1199   CORPLOG: 1
##  1st Qu.:4.000   1st Qu.: 4.370   1st Qu.:1760   1st Qu.:0.7868   LT-SEL : 2
```

```
## Median :6.000 Median : 5.646 Median :1871 Median :1.3891 SETTLE : 1
## Mean :6.023 Mean : 5.544 Mean :1833 Mean :1.2582 VIRGIN :40
## 3rd Qu.:8.000 3rd Qu.: 5.899 3rd Qu.:1932 3rd Qu.:1.8273
## Max. :9.000 Max. :11.850 Max. :1992 Max. :1.9867
## streamdist
## Min. : 0.0
## 1st Qu.:364.0
## Median :460.1
## Mean :486.7
## 3rd Qu.:535.6
## Max. :836.5
```

*#how well does the exploratory model appear to explain cover? It gives alot of information which is not
#which explanatory variables are the most important? The variables for Abies fraseri of importanace seen*

```
abies.lm<-lm(cover ~ disturb + elev + streamdist ,data= abies)
summary(abies.lm)
```

```
##
## Call:
## lm(formula = cover ~ disturb + elev + streamdist, data = abies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.3604 -1.2462  0.2371  1.1941  3.2542
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -18.274973   4.057937  -4.504 6.18e-05 ***
## disturbLT-SEL  2.824972   2.017886   1.400  0.1696
## disturbSETTLE  0.874518   2.294927   0.381  0.7053
## disturbVIRGIN  3.074096   1.697937   1.810  0.0781 .
## elev          0.011995   0.002478   4.841 2.18e-05 ***
## streamdist    -0.001298   0.001595  -0.814  0.4207
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.614 on 38 degrees of freedom
## Multiple R-squared:  0.5518, Adjusted R-squared:  0.4928
## F-statistic: 9.356 on 5 and 38 DF,  p-value: 7.133e-06
```

#The model explains 49% of the variance and has highly significant coefficients for the intercept and t

#do model diagnostics indicate any problems with violations of OLS assumptions? The Homoscedasticity se

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

```
#ANOVA TEST
anova(acer.lm)
```

```
## Analysis of Variance Table
```

```
##
## Response: cover
##           Df Sum Sq Mean Sq F value    Pr(>F)
## beers      1  41.40  41.403 10.4378 0.001291 **
## elev       1  22.60  22.600  5.6975 0.017246 *
## streamdist 1  45.22  45.220 11.4000 0.000774 ***
## Residuals 719 2852.03   3.967
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Stream distance and beers are the most signifcant varibales for acer, from the anova test

anova(abies.lm)
```

```
## Analysis of Variance Table
```

```
##
## Response: cover
##           Df Sum Sq Mean Sq F value    Pr(>F)
## disturb    3 46.202  15.401  5.9085 0.002061 **
## elev       1 73.999  73.999 28.3895 4.744e-06 ***
## streamdist 1  1.727   1.727  0.6626 0.420703
## Residuals 38 99.049   2.607
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Disturbance and elevation the most signifcant varibales for acer, from teh anova test
```

```
#2. You may have noticed that the variable cover is defined as positive integers between 1 and 10. and
```

```
#GLM
```

```
acer.glm = glm(cover ~ elev + streamdist + beers , data = acer,
               family='poisson')
summary(acer.glm)
```

```
##
## Call:
## glm(formula = cover ~ elev + streamdist + beers, family = "poisson",
##      data = acer)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.5225  -0.6144   0.1240   0.5825   2.1083
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.766e+00  5.327e-02  33.145  < 2e-16 ***
## elev        -1.586e-04  5.753e-05  -2.757  0.00583 **
## streamdist   2.919e-04  9.872e-05   2.956  0.00311 **
## beers       -5.910e-02  2.368e-02  -2.496  0.01258 *
## ---
## 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: 628.24  on 719  degrees of freedom
## AIC: 3098.6
##
## Number of Fisher Scoring iterations: 4
abies.glm = glm(cover ~ disturb + elev + streamdist ,data = abies,
               family='poisson')
summary(abies.glm)

##
## Call:
## glm(formula = cover ~ disturb + elev + streamdist, family = "poisson",
##      data = abies)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.4680  -0.6436   0.1216   0.4532   1.4139
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.4419365   1.4301035  -2.407   0.0161 *
## disturbLT-SEL  1.3625701   1.0646587   1.280   0.2006
## disturbSETTLE  0.8991409   1.1563757   0.778   0.4368
## disturbVIRGIN  1.4172746   1.0114793   1.401   0.1612
## elev           0.0021250   0.0006652   3.194   0.0014 **
## streamdist    -0.0001743   0.0003960  -0.440   0.6598
## ---
## 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: 17.750  on 38  degrees of freedom
## AIC: 186.93
##
## Number of Fisher Scoring iterations: 4
#Pseudo r-squared
pseudo_r2 = function(glm_mod) {
  1 - glm_mod$deviance / glm_mod$null.deviance
}

pseudo_r2(acer.glm)

## [1] 0.03249165
# The r-squared is now 3% for acer

pseudo_r2(abies.glm)

## [1] 0.5699441
# The r-squared is now 57% for abies

#Anova test

```

```
anova(acer.lm, acer.glm)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Model 1: cover ~ beers + elev + streamdist
```

```
## Model 2: cover ~ elev + streamdist + beers
```

```
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
```

```
## 1      719 2852.03
```

```
## 2      719  628.24  0    2223.8
```

```
#residual of squares is reduced with the acer.glm anova test
```

```
anova(abies.lm, abies.glm)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Model 1: cover ~ disturb + elev + streamdist
```

```
## Model 2: cover ~ disturb + elev + streamdist
```

```
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
```

```
## 1       38 99.049
```

```
## 2       38 17.750  0    81.299
```

```
#residual of squares is reduced with the abies.glm anova test
```

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?

From the analysis I've come to the following conclusions:

The *Acer rubrum*'s (*Acer*) cover can not be explained strictly by the variables relationship with cover. Cover may be the results of other significant relationships. Looking back at the plotted results, stream distance and other relationships have the most significance to better explain *Acer rubrum*'s abundance. The model I was able to generate only explained 3.29% *Acer rubrum*'s cover, which is not too surprising since this species is a generalist.

The *Abies fraseri* (*Abies*) cover is better explained by the models generated, I assume its easier to detect significant variables because the *Abies* is a habitat specialist. It is easier to identify significant influences for *Abies* abundance with the model and identify relationships. The model generated explained 49% of *Abies* cover.

NOTES for me:

The General Linear Model (GLM) is a useful framework for comparing how several variables affect different continuous variables. In it's simplest form, GLM is described as: $\text{Data} = \text{Model} + \text{Error}$ (Rutherford, 2001, p.3)

GLM is the foundation for several statistical tests, including ANOVA, ANCOVA and regression analysis. Despite their differences, each fits the definition of $\text{Data} = \text{Model} + \text{Error}$: • In ANOVA, "data" is the dependent variable scores, the "error" the model is the experimental conditions, and the "error" is the part of the model not explained by the data. • In regression analysis, the independent predictors make up the "model" and the residuals are the "error" component.