HW3_Margaret_Walker

Margaret Walker

February 6, 2016

1. 1. Carry out an exploratory analysis using the tree dataset. 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`.

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?
* which explanatory variables are the most important?
* do model diagnostics indicate any problems with violations of
  OLS assumptions?
* are you able to explain variance in one species better than         another?
```

I started by subsetting the data into two dataframes (one for each species). I only selected the variables needed for the analysis. See code below:

```
trees <- read.csv("../data/treedata_subset.csv")
acer <- subset(trees, species == "Acer rubrum", select = c("cover", "elev",
    "tci", "streamdist", "disturb", "beers"))
abies <- subset(trees, species == "Abies fraseri", select = c("cover", "elev",
    "tci", "streamdist", "disturb", "beers"))
```

Next I created models for both acer and abies with all of the variables (elevation, tci, streamdist, disturb, and beers). I used the summary() function to get information about each model. I also used the Anova function for both models. Finally, I plotted both models to check some of the assumptions. See work below:

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.1.3
```

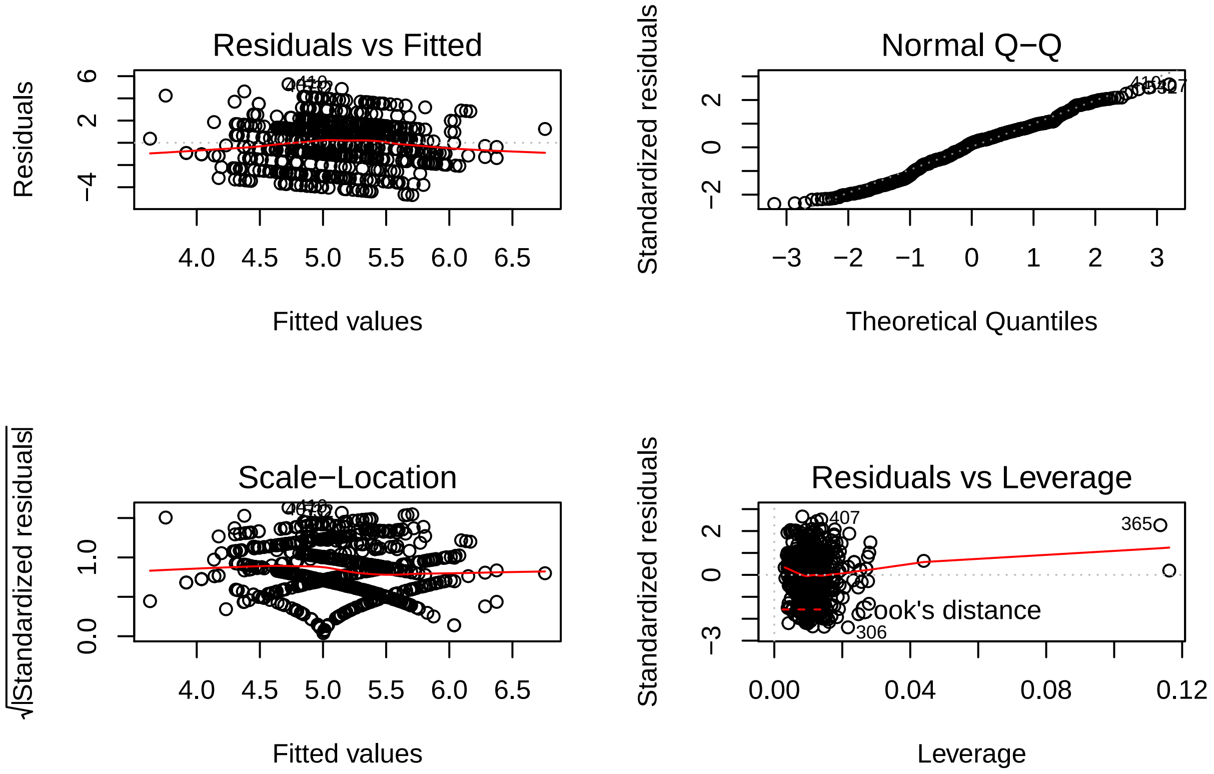
```
mod_acer <- lm(cover ~ . , data=acer)
summary(mod_acer)
```

```
##
## Call:
## lm(formula = cover ~ ., 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 ***
## 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
```

```
Anova(mod_acer)
```

```
## Anova Table (Type II tests)
##
## Response: cover
##              Sum Sq  Df F value   Pr(>F)
## elev          40.44   1 10.2233 0.001448 **
## tci           12.58   1  3.1805 0.074947 .
## streamdist    29.09   1  7.3531 0.006856 **
## disturb        9.45   3  0.7962 0.496166
## beers         35.61   1  9.0034 0.002789 **
## Residuals   2828.21 715
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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



```
mod_abies <- lm(cover ~ . , data=abies)
summary(mod_abies)
```

```
##
## Call:
## lm(formula = cover ~ ., 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 ***
## 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
```

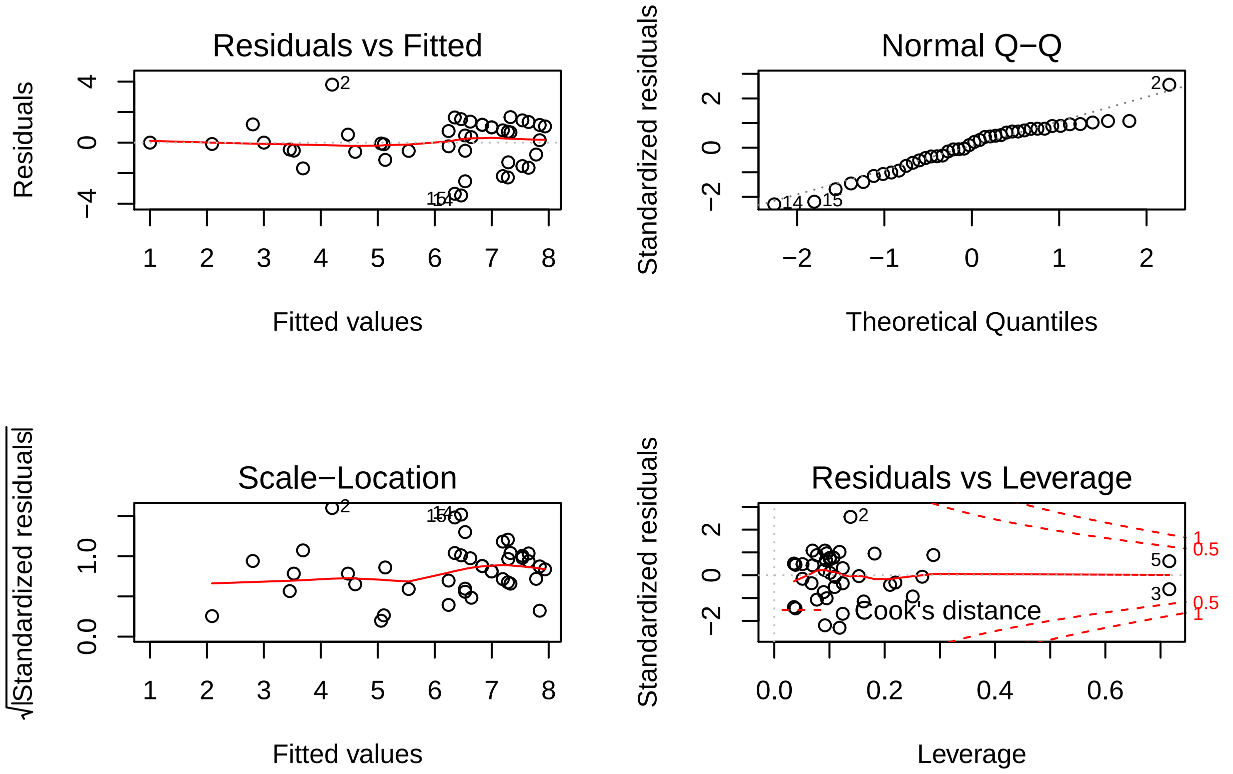
```
Anova(mod_abies, type=3)
```

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

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

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

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



First, for acer rubrum, when looking at the summary of the model it appears that elevation, stream distance, and beers are potentially important in determining cover. However, The adjusted R^2 is only 3.6%, which is very low. So only 3.6% of the variance in cover is explained by the model. When comparing the p-values from Anova() and summary() we get p-values almost equal with the exception of disturbance since summary separates out the levels of disturbance while Anova() doesn't. Also, when examining the plots of the acer model we see some issues. There appears to be some type of pattern in the residual vs. fitted plot which indicates that the variances are not equal. Also, on the qqplot there are a few outliers, which means that data are not necessarily normal. So, we may want to consider going another route for further analyses.

Next, for abies fraseri, when looking at the summary of the model it appears that elevation is the only variable important in determining the cover. Furthermore, the adjsuted R^2 is 50%, which means 50% of the variance in cover is explained by the model. Once again, the Anova() function gives us similar p-values except that disturbance is not separated by factor. Furthemore, when looking at the model plots we see some issues again. There are definitely a few outliers based on the qqplot. Also, based on the qqplot and the pattern on the residuals vs. fitted plot (more points on the right side) I am skeptical that the data are normal and the variances are equal. There also appears to be significant leverage for some of the points.

Based on the adjusted R^2 values we can describe more of the variance in cover for abies than for acer. Only one variable (elevation) significantly effects cover for abies, while several variables are important for acer. This makes since based on the fact that acer rubrum is a habitat generalist while abies fraseri is a habitat specialist.

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.

I created poisson glm's for both acer and abies and then I compared the glm to the ols models using aov. I also calculated the pseudo r^2 for both and plotted the models to look at the assumptions. See the code below:

First for acer rubrum:

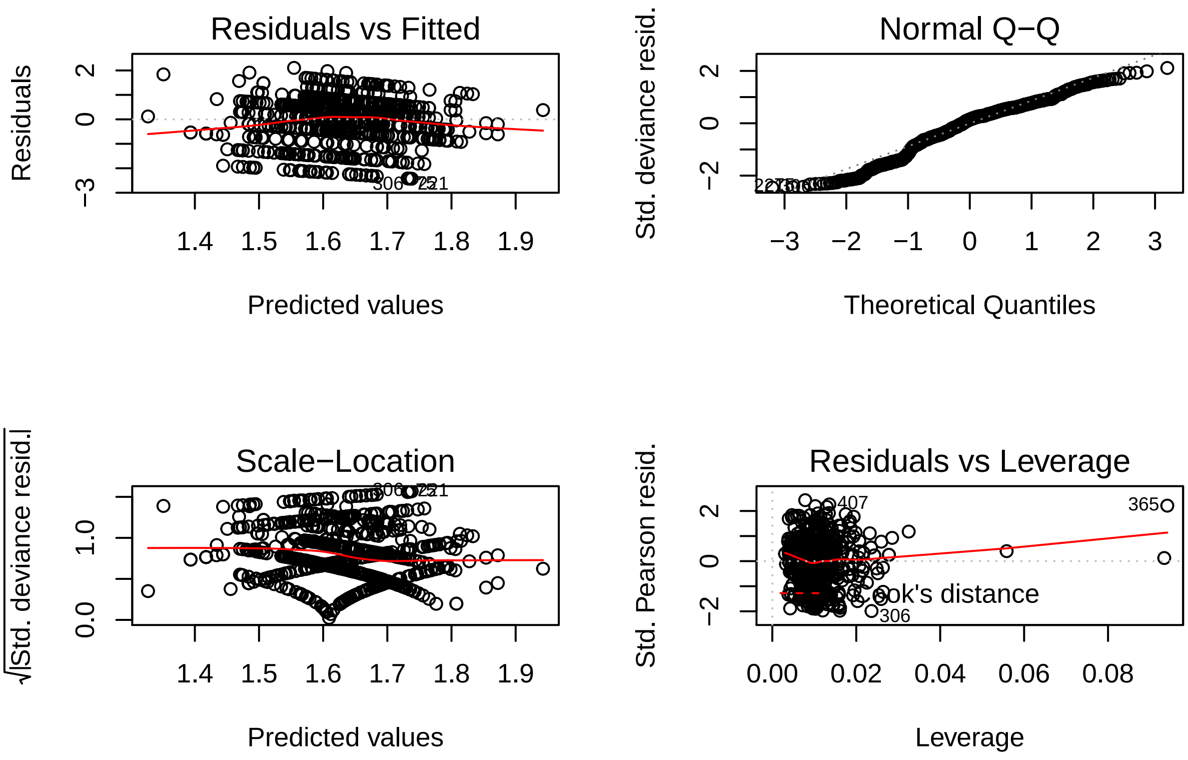
```r
glm_acer <- glm(cover ~ . , data=acer, family ="poisson")
summary(glm_acer)
```

```
##
## Call:
## glm(formula = cover ~ ., 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 ***
## 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
```

```
summary(mod_acer)
```

```
##
## Call:
## lm(formula = cover ~ ., 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 ***
## 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
```

```
psuedo_r2 <- function(glm_mod){
  1-glm_mod$deviance/glm_mod$null.deviance
}
par(mfrow=c(2,2))
plot(glm_acer)
```



```
psuedo_r2(glm_acer)
```

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

```
anova(mod_acer, glm_acer)
```

```
## 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
```

It appears that for the glm for acer elevation, streamdist, and beers are still important in determining cover. The pseudo R^2 is 3.99% which means the model explains about 4% of the variance in cover. This is slightly higher than the adjusted R^2 for the OLS model. Furthermore, when looking at the plots of the glm the qqplot looks a little better than the qqplot for the OLS model. This leads me to believe this error distribution is a better fit for the data. When using the anova() function comparing the OLS and glm model we also see how much of a better fit the poisson model is than the gaussian model. The residual sums of squares is significantly lower for the glm than the lm, which means much more error is explained by the glm model.

Next for abies fraser, see code below:

```
glm_abies <- glm(cover ~ . , data=abies, family="poisson")
summary(glm_abies)
```

```
##
```

```
## Call:
## glm(formula = cover ~ ., 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 **
## 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
```

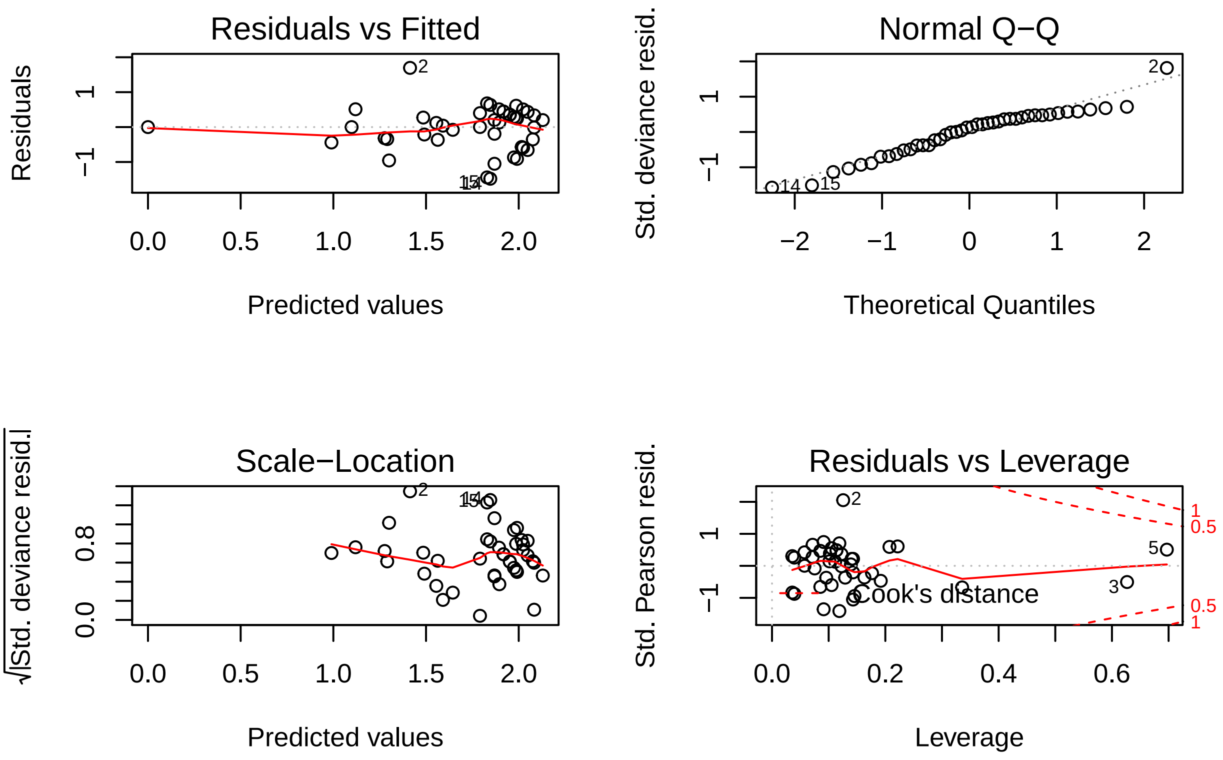
```
psuedo_r2(glm_abies)
```

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

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

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

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



```r
anova(mod_abies, glm_abies)
```

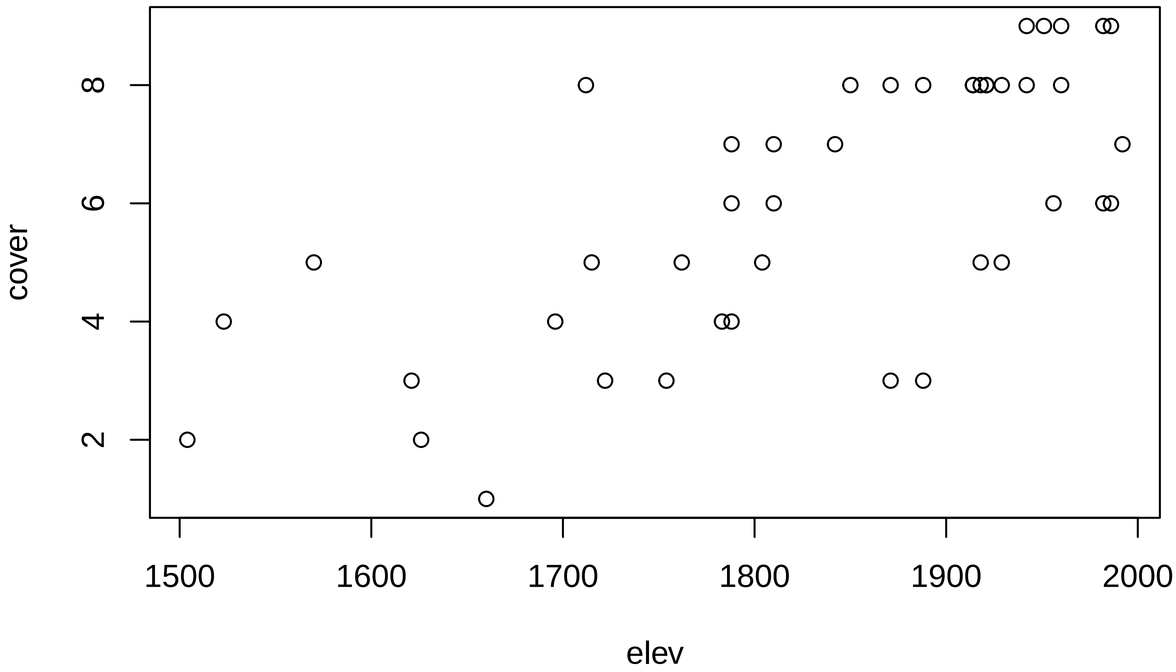
```
## Analysis of Variance Table
##
## Model 1: cover ~ elev + tci + streamdist + disturb + beers
## Model 2: cover ~ elev + tci + streamdist + disturb + beers
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1     36 92.289
## 2     36 16.126  0    76.164
```

Once again for abies it appears that only elevation is important in determining cover even for the glm. The psuedo R^2 for the glm is 60.9% which means almost 61% of the variance in cover is explained by this model. This pseudo R^2 is almost 10% higher than the adjusted R^2 for the ols model, which means it is a better fit. However, when looking at the plots there doesn't appear to be much different between those for the OLS model and the glm model. This could be due to the small sample size for abies fraseri compared to acer rubrum. Finally, when using the anova() function to compare the OLS and glm model we see that the residual sums of squares is much lower for the glm model, which indicates that it is a much better fit.

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?

Based on the analysis we can conclude some important things. First, the error distribution for the data is most likely a poisson distrubtion rather than a gaussian distribution due to the lower residual sums of squares for the poisson. Also, it appears that our model is not very good at describing the variance in cover for acer rubrum, but is a good predictor of cover for abies fraseri based on the pseudo R^2. In plain english, this tells us that none of the environmental variables are very good predictors in determining the cover of acer rubrum. This goes along with the fact that acer rubrum is a habitat generalist. That is, it can live in any habitat, which helps us understand why we couldn't determine any predictors of cover. On the other hand, abies fraseri cover appears to be predicted by elevation. That is, as elevation increases so does abies fraseri cover. This corresponds with the fact that abies is a habitat specialist and we should only see high cover under very specific environmental condities (i.e. high elevation). We can further see the relationship of abies cover and elevation in the following plot:

```
par(mfrow=c(1,1))
plot(cover~elev, data=abies)
```



4. (optional) Examine the behavior of the function `step()` using the exploratory models developed above. This is a very simple and not very robust machine learning stepwise algorithm that uses AIC to select a best model. By default it does a backward selection routine.

To use the step() function I start with the full glm poisson model for acer and abies. See work below:

```
step(glm_acer)
```

```
## Start:  AIC=3101.77
## cover ~ elev + tci + streamdist + disturb + beers
## 
##              Df Deviance    AIC
## - disturb     3   625.28 3097.7
## <none>            623.38 3101.8
## - tci         1   625.97 3102.4
## - streamdist  1   628.87 3105.2
## - beers       1   630.34 3106.7
## - elev        1   631.16 3107.5
## 
## Step:  AIC=3097.67
## cover ~ elev + tci + streamdist + beers
## 
##              Df Deviance    AIC
## <none>            625.28 3097.7
## - tci         1   628.24 3098.6
## - streamdist  1   631.22 3101.6
## - beers       1   632.24 3102.6
## - elev        1   634.11 3104.5
```

```
##
## Call:  glm(formula = cover ~ elev + tci + streamdist + beers, family = "poisson",
##     data = acer)
##
## Coefficients:
## (Intercept)         elev          tci   streamdist        beers
##   1.8700348   -0.0001719   -0.0138226    0.0002500   -0.0626543
##
## Degrees of Freedom: 722 Total (i.e. Null);  718 Residual
## Null Deviance:       649.3
## Residual Deviance: 625.3     AIC: 3098
```

```r
step(glm_abies)
```

```
## Start:  AIC=189.3
## cover ~ elev + tci + streamdist + disturb + beers
##
##              Df Deviance    AIC
## - disturb     3   19.521 186.70
## - beers       1   16.141 187.32
## - streamdist  1   16.431 187.61
## - tci         1   17.308 188.49
## <none>            16.125 189.30
## - elev        1   27.471 198.65
##
## Step:  AIC=186.7
## cover ~ elev + tci + streamdist + beers
##
##              Df Deviance    AIC
## - beers       1   19.533 184.71
## - streamdist  1   20.014 185.19
## - tci         1   21.459 186.64
## <none>            19.521 186.70
## - elev        1   35.334 200.51
##
## Step:  AIC=184.71
## cover ~ elev + tci + streamdist
##
##              Df Deviance    AIC
## - streamdist  1   20.055 183.23
## <none>            19.533 184.71
## - tci         1   21.731 184.91
## - elev        1   37.364 200.54
##
## Step:  AIC=183.23
## cover ~ elev + tci
##
```

```
##        Df Deviance    AIC
## <none>      20.055 183.23
## - tci   1   22.180 183.36
## - elev  1   41.120 202.30
```

```
##
## Call:  glm(formula = cover ~ elev + tci, family = "poisson", data = abies)
##
## Coefficients:
## (Intercept)         elev          tci
##   -3.137624     0.002469     0.065410
##
## Degrees of Freedom: 43 Total (i.e. Null);  41 Residual
## Null Deviance:       41.27
## Residual Deviance: 20.06     AIC: 183.2
```

The step function runs through the model removing variables until the lowest AIC value is obtained. Based on this function the best model to describe cover of acer rubrum contains the following variables: elevation, tci, streamdist, and beers. This is all of the variables except disturbance. This once again goes along with the fact that acer is a generalist. On the other hand, the step function tells us the best model to describe cover of abies contains only elevation and tci as variables.