

# Randich\_Univariate\_Assignment

Mikala

February 7, 2016

```
library(car)
trees <- read.csv('http://dmcglinn.github.io/quant_methods/data/treedat
a_subset.csv')
cols_of_interest = c('cover', 'elev', 'tci', 'streamdist', 'disturb', '
beers')
acer <- subset(trees, subset=species == 'Acer rubrum',
               select = cols_of_interest)
acermod = lm(cover ~ ., data = acer)
Anova(acermod, type = 3)

## Anova Table (Type III tests)
##
## Response: cover
##
```

	Sum Sq	Df	F value	Pr(>F)	
(Intercept)	765.43	1	193.5096	< 2.2e-16	***
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

summary(acermod)

##
## 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

abies <- subset(trees, subset=species == 'Abies fraseri',
                select = cols_of_interest)
abiesmod = lm(cover ~ ., data = abies)
Anova(abiesmod, 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

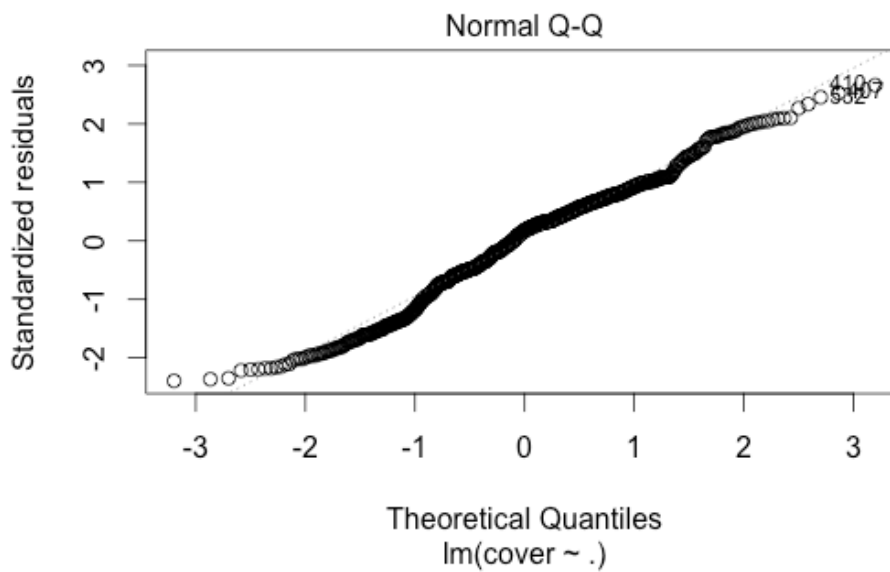
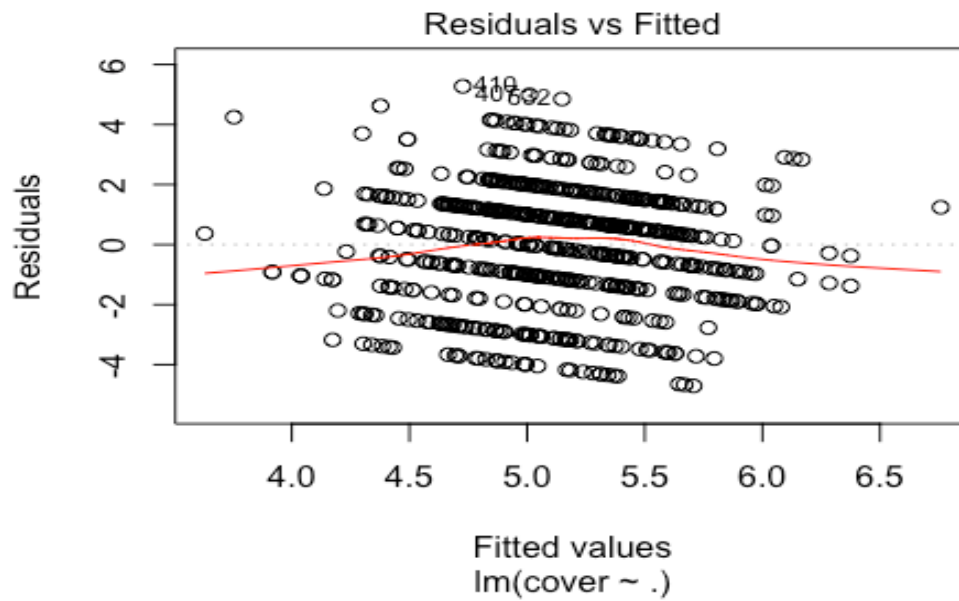
summary(abiesmod)

##
## 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
```

```
plot(acermod)
```



!!!

The exploratory model created for the *Abies fraseri* species to explain cover relatively well, with an R-squared adjusted value of 50.11%. The exploratory model created for the *Acer rubrum* species appears to explain cover less well, with an R-squared adjusted value of 3.558%.

For the species *Acer rubrum*, the significant factors in predicting tree cover determined by both a type 3 ANOVA and a summary in R were: elevation, streamdist (distance from nearest permanent stream), and beers (transformed slope aspect or 'heat load index'). For the species *Abies fraseri*, the only significant factor determined by the type 2 ANOVA and summary was elevation.

For the exploratory model developed to predict cover for the *Acer rubrum* species, the model assumption of normality seems to be reasonably satisfied, because the standard residuals generally fall along the line of theoretical quantities, with the exceptions of a couple outliers.

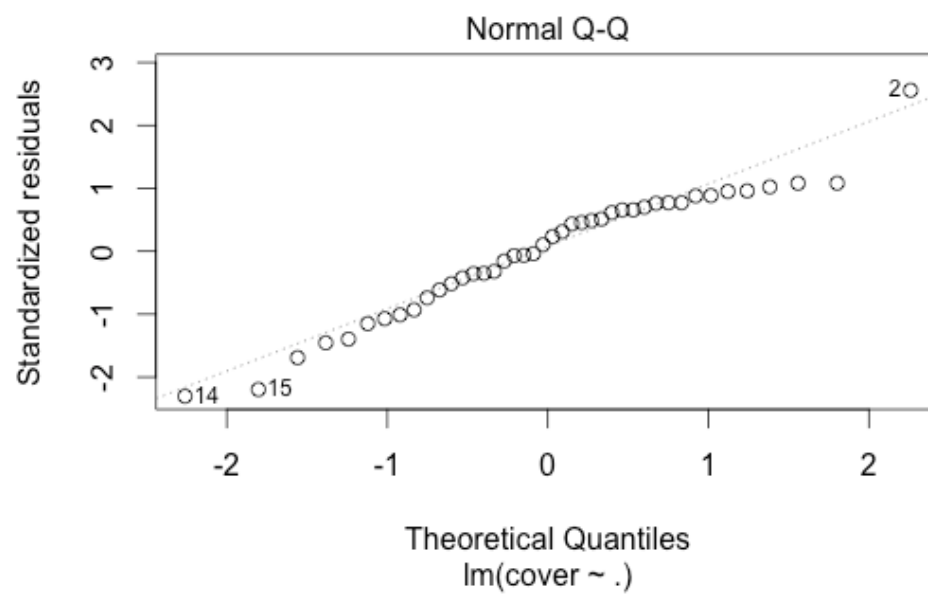
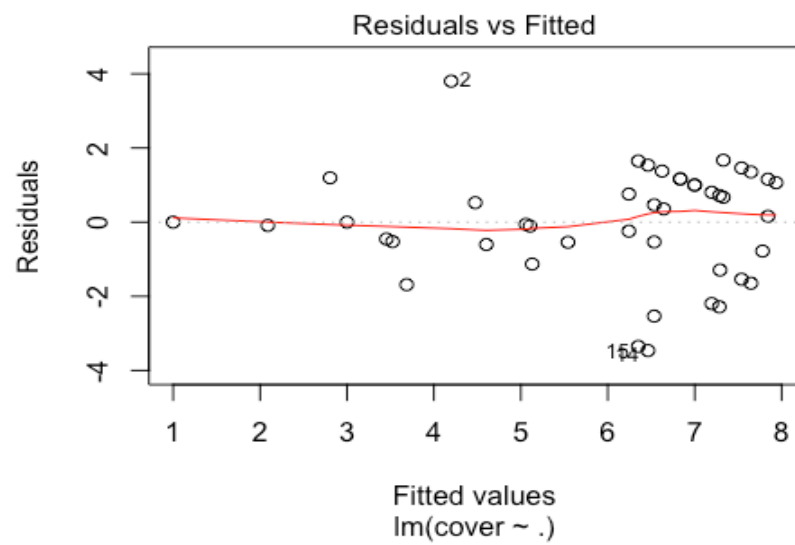
When examining the assumption of equal variance, there is clearly a pattern within the residual plots. This indicates the variables are affecting cover in a systematic way, and the model needs improvement. Additionally, there are a couple observations that have a disproportionate amount of leverage.

For the exploratory model developed to predict cover for the *Abies rubrum* species, the model assumptions generally appear to be satisfied.

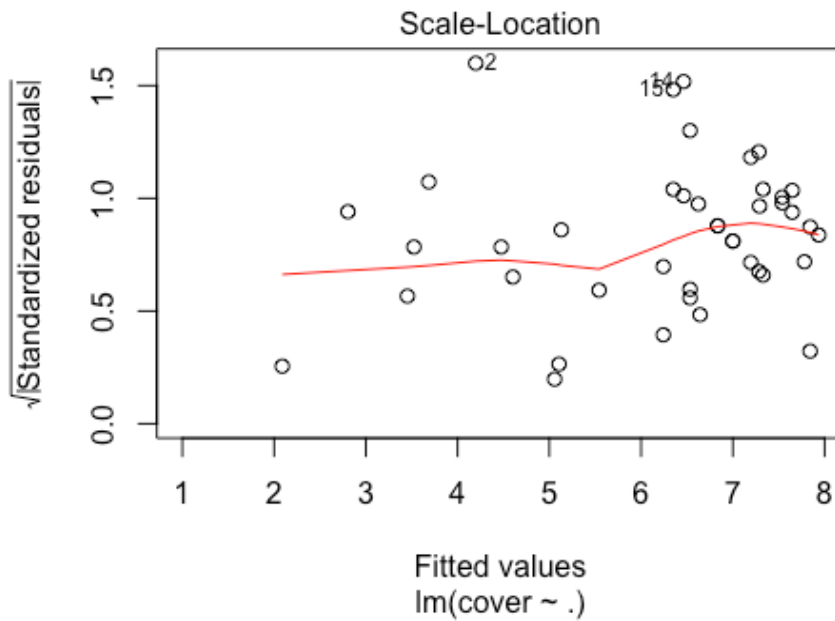
The data appears to be generally normally distributed, with the exceptions of a couple outliers and some tailing off at the end. The assumption of equal variance appears to be generally satisfied, although there is a left skew that may indicate further analysis is necessary. There are a couple observations that have a disproportionate amount of leverage.

```
plot(abiesmod)
```

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



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



Solely looking at the type 3 ANOVA models and summaries produced by R to explain cover for the *Acer rubrum* and *Abies fraseri* species, it would seem that we are better able to predict cover for the *Abies* species. The model for this species had a much higher R-squared adjusted value, and was the simpler model with only one significant explanatory variable. However, this species had a much smaller number of observations, which could have an impact on the model. The *Acer rubrum* model was more complicated with more explanatory variables produced and a low R-squared adjusted value. This exploratory model clearly had issues with the model assumption of equal variance, due to the striking pattern in the residual plots. However, the p-values for each explanatory factor produced were very low, and including only these variables could produce a better model.

```
acer_glm = glm(cover ~ . , data= acer, family='poisson')
acer_glm

##
## Call:  glm(formula = cover ~ ., family = "poisson", data = acer)
##
## Coefficients:
## (Intercept)          elev          tci      streamdist      disturbl
T-SEL
##  1.8730109      -0.0001961     -0.0129660       0.0002428       0.01
83973
## disturbSETTLE  disturbVIRGIN          beers
```

```
##      -0.0173856      0.0631125      -0.0639106
##
## Degrees of Freedom: 722 Total (i.e. Null);  715 Residual
## Null Deviance:      649.3
## Residual Deviance: 623.4      AIC: 3102

summary.glm(acer_glm)

##
## 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

abies_glm = glm(cover ~ ., data= abies, family='poisson')
summary.glm(abies_glm)

##
## 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

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

## [1] 0.03997917

pseudo_r2(abies_glm)

## [1] 0.60931

anova(acermod, acer_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(abiesmod, abies_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

```



For the general linear model (GLM) produced for the *Abies* species, again, elevation was determined to be the only significant explanatory variable for predicting cover. A different function was applied to the *Abies* GLM to determine an R-squared value of 60.93%. Because of this higher  $R^2$  value than the OLS model, the GLM appears to explain cover better than the OLS.

For the GLM produced for the *Acer* species, elevation, stream distance, and aspect were again determined to be the significant factors in predicting cover. The new  $R^2$  function produced a value of 3.99%, which is slightly higher than the adjusted  $R^2$  value produced by the OLS model for *Acer*. Similar to the OLS results, the GLM for the *Abies* species is a better predictor of cover than the GLM for the *Acer* species. The AIC values affirmed this, as *Abies* had a lower AIC value, indicating it may be the “better” model.

When examining the model assumptions for the GLMs, the residual plots were the same as those produced by the OLS models, raising the same questions about the assumption of equal variance being satisfied.

When examining the anova table produced by comparing the OLS to the GLM models for each *Acer* and *Abies* species, in both cases, the GLM model did a better job at predicting cover. The RSS (residual sum of squares) indicates how much the actual observations deviated from the predicted values. In both the *Acer* and *Abies* models, the GLM had a lower RSS than the OLS model.

From this analysis we can determine that when predicting tree cover for the species *Abies fraseri*, elevation is the key factor that is important. When predicting cover for the species *Acer rubrum*, key factors to consider are elevation, distance from a stream, and slope aspect or heat index. The exploratory models created to model cover based on other factors created better models for the *Abies* species than the *Acer* species, since it is the simpler model and more variance can be explained by the model. However, it is important to take into consideration that the assumptions required for these models may not have been adequately fulfilled, and further investigation and analysis may be required.