univariate

Julius Alipala

January 31, 2019

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.5.2
```

```
trees = read.csv('~/quant_methods/data/treedata_subset.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 ...
```

```
#subsets
red_maple = trees[trees$species == "Acer rubrum",]
str(red_maple)
```

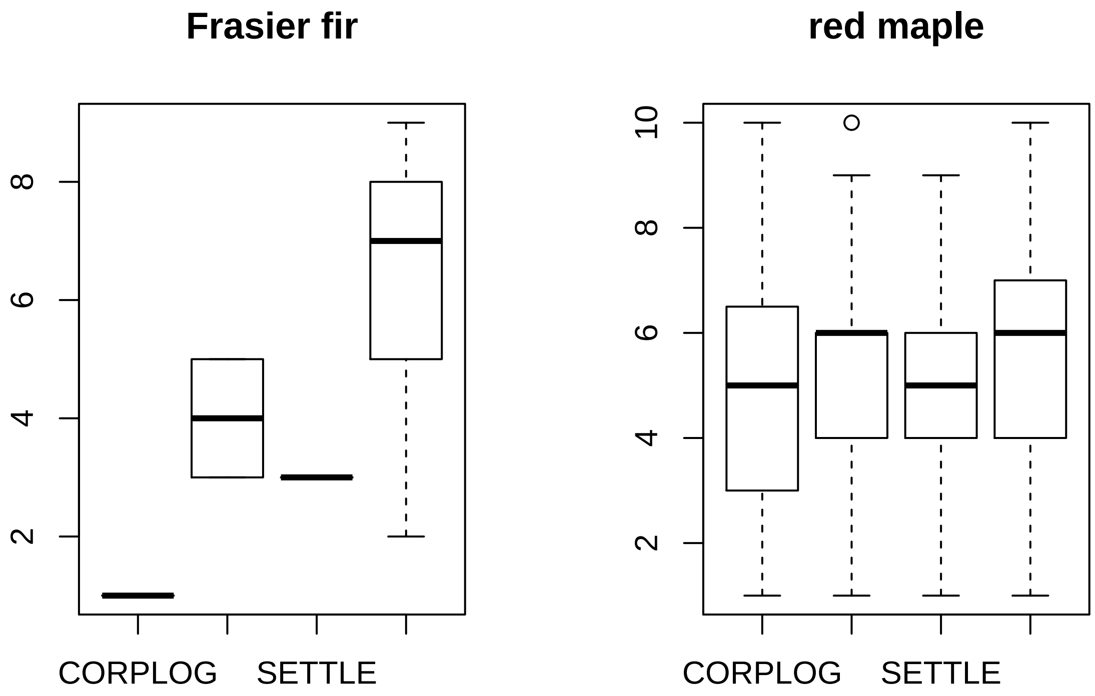
```
## 'data.frame':    723 obs. of  9 variables:
##  $ plotID    : Factor w/ 734 levels "ATBN-01-0303",..: 1 2 3 4 5 6 8 9 10 18 ...
##  $ spcode    : Factor w/ 52 levels "ABIEFRA","ACERNEG",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ species   : Factor w/ 51 levels "Abies fraseri",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ cover     : int  6 7 5 7 5 4 2 7 4 7 ...
##  $ elev      : num  896 947 1027 450 477 ...
##  $ tci       : num  4.71 4.45 6.15 4.13 5.59 ...
##  $ streamdist: num  197 125 175 202 134 ...
##  $ disturb   : Factor w/ 4 levels "CORPLOG","LT-SEL",..: 1 1 1 2 2 2 1 4 2 1 ...
##  $ beers     : num  1.991 0.817 0.586 0.86 0.101 ...
```

```
frasier_fir = trees[trees$species == "Abies fraseri",]
str(frasier_fir)
```

```
## 'data.frame':    44 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 ...
```

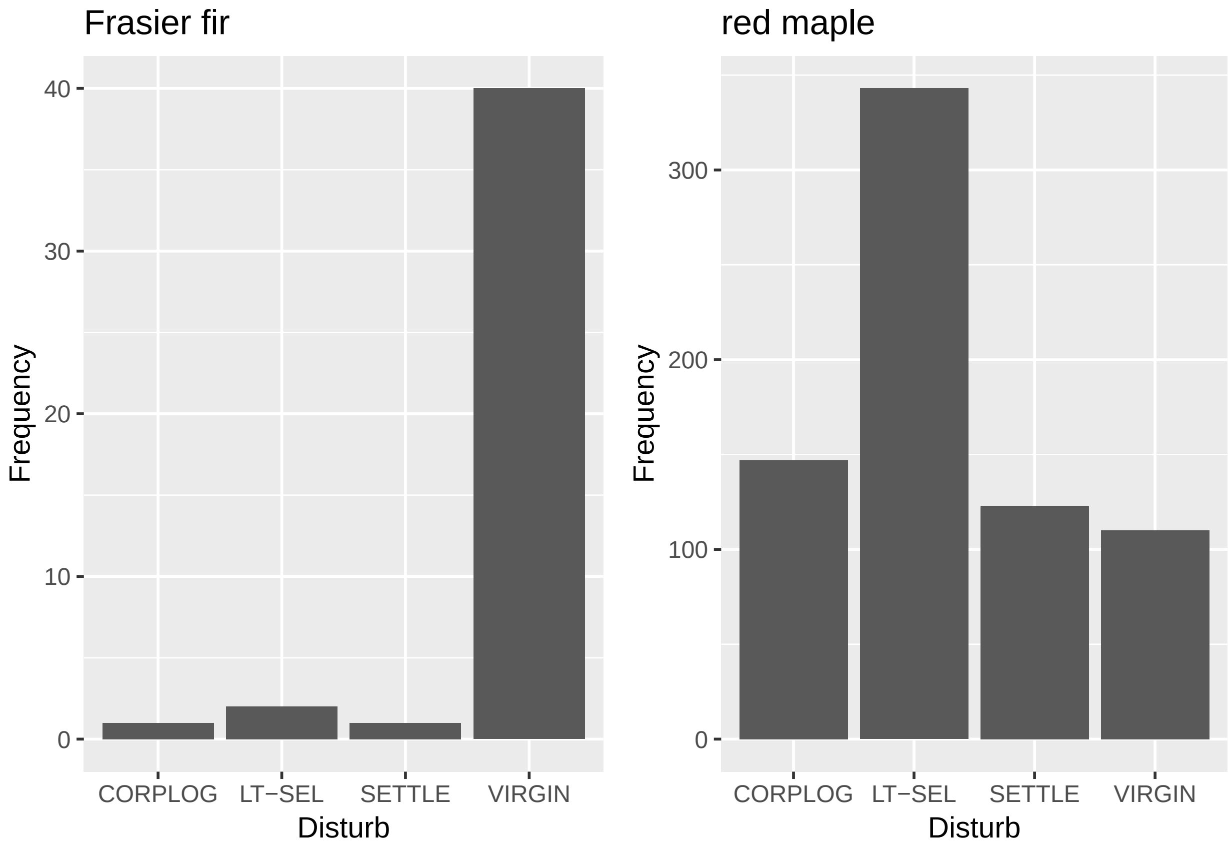
```
# red_maple - generalist
# frasier_fir - specialist

# more observations for generalist than specialist
```

```
par(mfrow=c(1,2))
boxplot(cover ~ disturb, frasier_fir)
title(main="Frasier fir")
boxplot(cover ~ disturb, red_maple)
title(main="red maple")
```



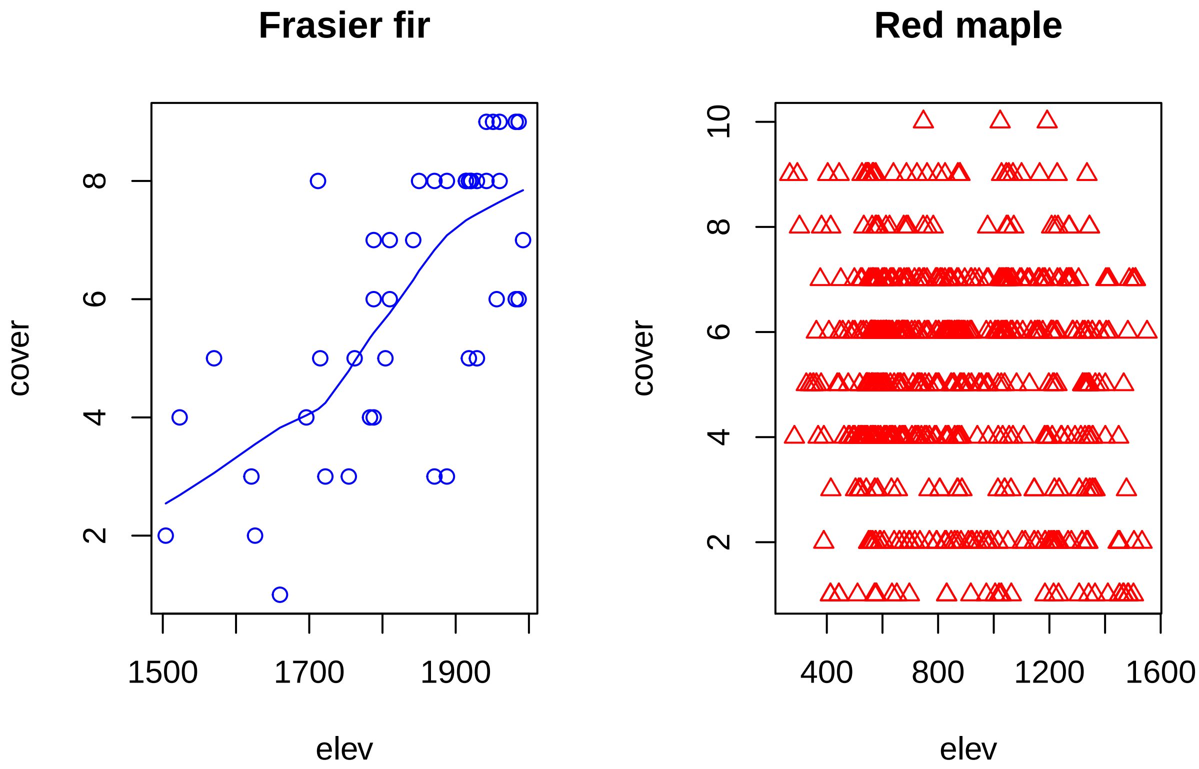
```
# outlier for red maple (lt-sel)
# more range in cover for red_maple
# only one instance of frasier_fir for CORPLOG and SETTLE
#
# plot suggests that being in a virgin environment affects cover for frasier fir, but there are very
# few observations of frasier fir in other environments
#

p1 = ggplot(data = frasier_fir,
            mapping = aes(x = disturb, group = 1)) +
      geom_bar() +
      labs(x = 'Disturb', y = 'Frequency') + ggtitle('Frasier fir')

p2 = ggplot(data = red_maple,
            mapping = aes(x = disturb, group = 1)) +
```

```r
      geom_bar() +
      labs(x = 'Disturb', y = 'Frequency') + ggtitle('red maple')

grid.arrange(p1, p2, nrow = 1)
```



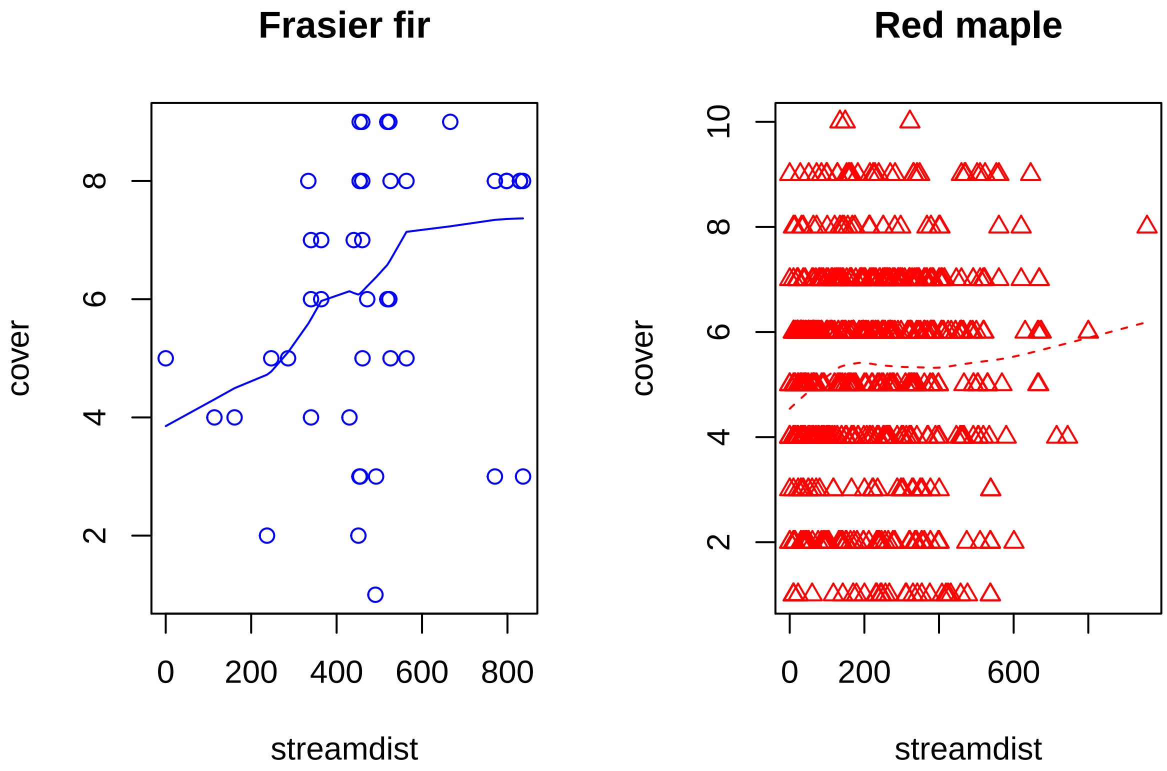
```r
# majority of observations of frasier firs are found in virgin environments
# frasier fir may be sensitive to disturbance; targeted for Christmas trees?
# majority of observations of blue maples found in environments with light/selective logging
```

```r
par(mfrow=c(1,2))
plot(cover ~ elev, frasier_fir, pch=1, col='blue')
title("Frasier fir")
lines(lowess(frasier_fir$elev, frasier_fir$cover), lt = 1, col='blue')
plot(cover ~ elev, red_maple, pch=2, col='red')
title("Red maple")
```



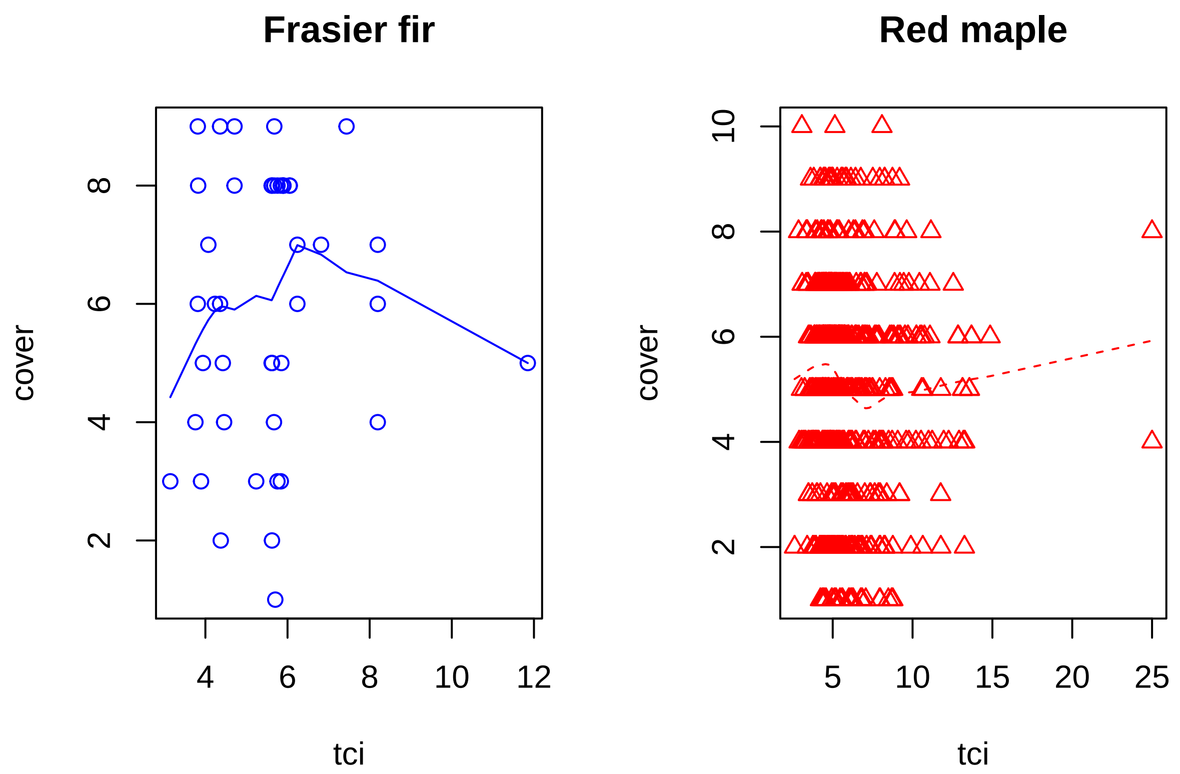
```
#lines(lowess(red_maple$elev, red_maple$cover), lty=2, col='red')

# elevation doesn't seem to effect cover for red maple
# higher elevation correlates with more cover for frasier firs
```

```
par(mfrow=c(1,2))
plot(cover ~ streamdist, frasier_fir, pch=1, col='blue')
title('Frasier fir')
lines(lowess(frasier_fir$streamdist, frasier_fir$cover), lt = 1, col='blue')
plot(cover ~ streamdist, red_maple, pch=2, col='red')
title('Red maple')
lines(lowess(red_maple$streamdist, red_maple$cover), lty=2, col='red')
```



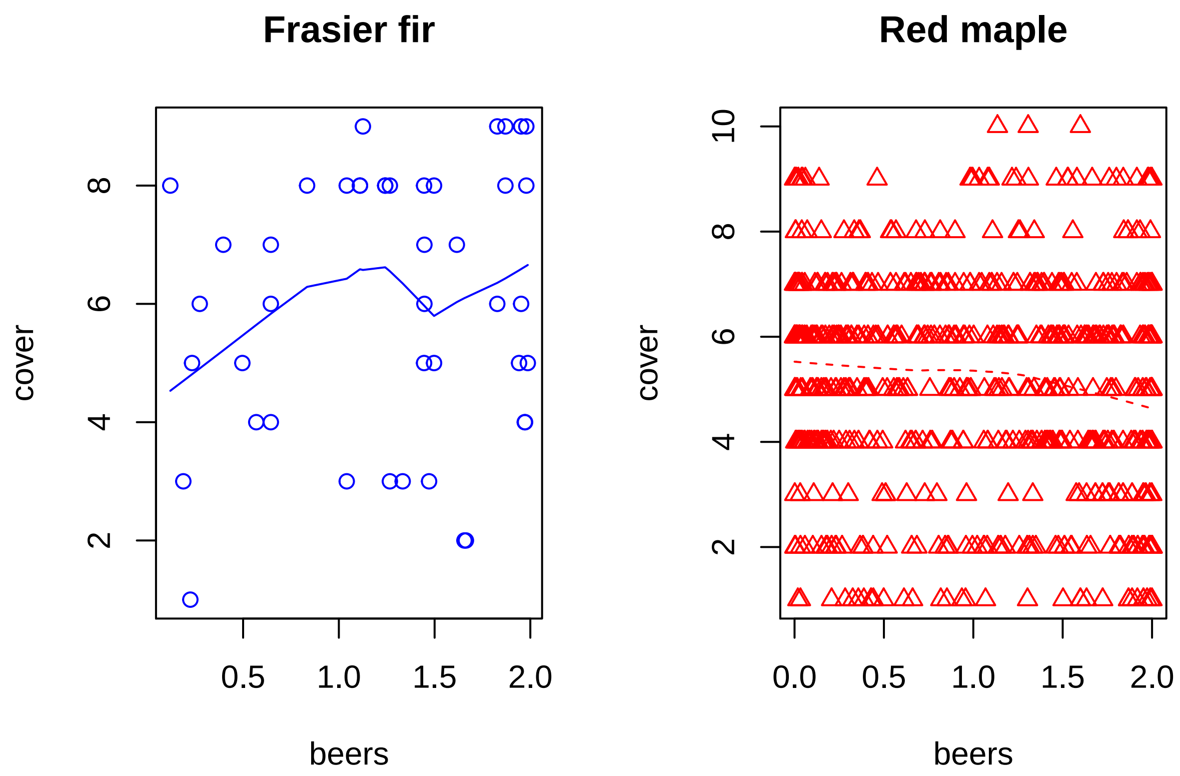
```
# stream distance may have an effect on cover for frasier fir
# red maple seems to thrive more when closer to a stream
# frasier fir have more observations between 200-600 meters from a stream
```

```
par(mfrow=c(1,2))
plot(cover ~ tci, frasier_fir, pch=1, col='blue')
title('Frasier fir')
lines(lowess(frasier_fir$tci, frasier_fir$cover), lt = 1, col='blue')
plot(cover ~ tci, red_maple, pch=2, col='red')
title('Red maple')
lines(lowess(red_maple$tci, red_maple$cover), lty=2, col='red')
```



```
# tci doesn't seem to have much effect on cover
# more observations when tci is lower
```

```
par(mfrow=c(1,2))
plot(cover ~ beers, frasier_fir, pch=1, col='blue')
title('Frasier fir')
lines(lowess(frasier_fir$beers, frasier_fir$cover), lt = 1, col='blue')
plot(cover ~ beers, red_maple, pch=2, col='red')
title('Red maple')
lines(lowess(red_maple$beers, red_maple$cover), lty=2, col='red')
```



```
# may have correlation between cover and beers for frasier fir
```

```
#null mod
null_mod = lm(cover ~ 1, data = frasier_fir)
null_mod
```

```
##
## Call:
## lm(formula = cover ~ 1, data = frasier_fir)
##
## Coefficients:
## (Intercept)
##       6.023
```

```
# mean cover for frasier fir is 6.023

par(mfrow=c(2,2))
plot(cover ~ 1, data = frasier_fir)
title("Frasier fir: null model")
abline(null_mod, lwd = 2)
abline(h = mean(frasier_fir$cover), col = 'red', lty = 2, lwd = 2)

#main effect model: elev
elev_mod = lm(cover ~ elev, data = frasier_fir)
elev_mod
```

```
##
## Call:
```

```
## lm(formula = cover ~ elev, data = frasier_fir)
##
## Coefficients:
## (Intercept)         elev
##   -15.81467      0.01191
```

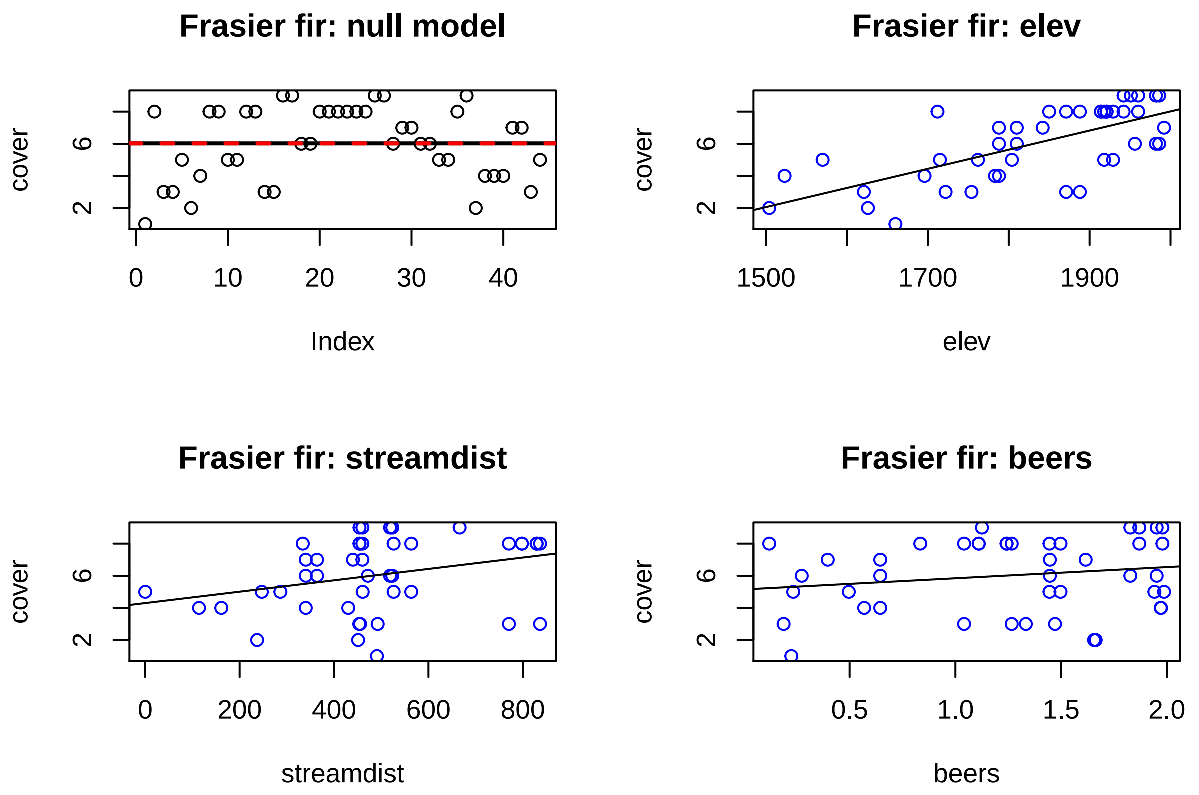
```
plot(cover ~ elev, frasier_fir, pch=1, col='blue')
title("Frasier fir: elev")
abline(elev_mod)

#streamdist
sd_mod = lm(cover ~ streamdist, data = frasier_fir)
sd_mod
```

```
##
## Call:
## lm(formula = cover ~ streamdist, data = frasier_fir)
##
## Coefficients:
## (Intercept)   streamdist
##    4.298319     0.003543
```

```
plot(cover ~ streamdist, frasier_fir, pch=1, col='blue')
title("Frasier fir: streamdist")
abline(sd_mod)

#beers
beers_mod = lm(cover ~ beers, data = frasier_fir)
beers_mod
```

```
##
## Call:
## lm(formula = cover ~ beers, data = frasier_fir)
##
## Coefficients:
## (Intercept)        beers
##      5.1474       0.6957
```

```
plot(cover ~ beers, frasier_fir, pch=1, col='blue')
title("Frasier fir: beers")
abline(beers_mod)
```



```
#null mod
null_rm = lm(cover ~ 1, data = red_maple)
null_rm
```

```
##
## Call:
## lm(formula = cover ~ 1, data = red_maple)
##
## Coefficients:
## (Intercept)
##       5.133
```

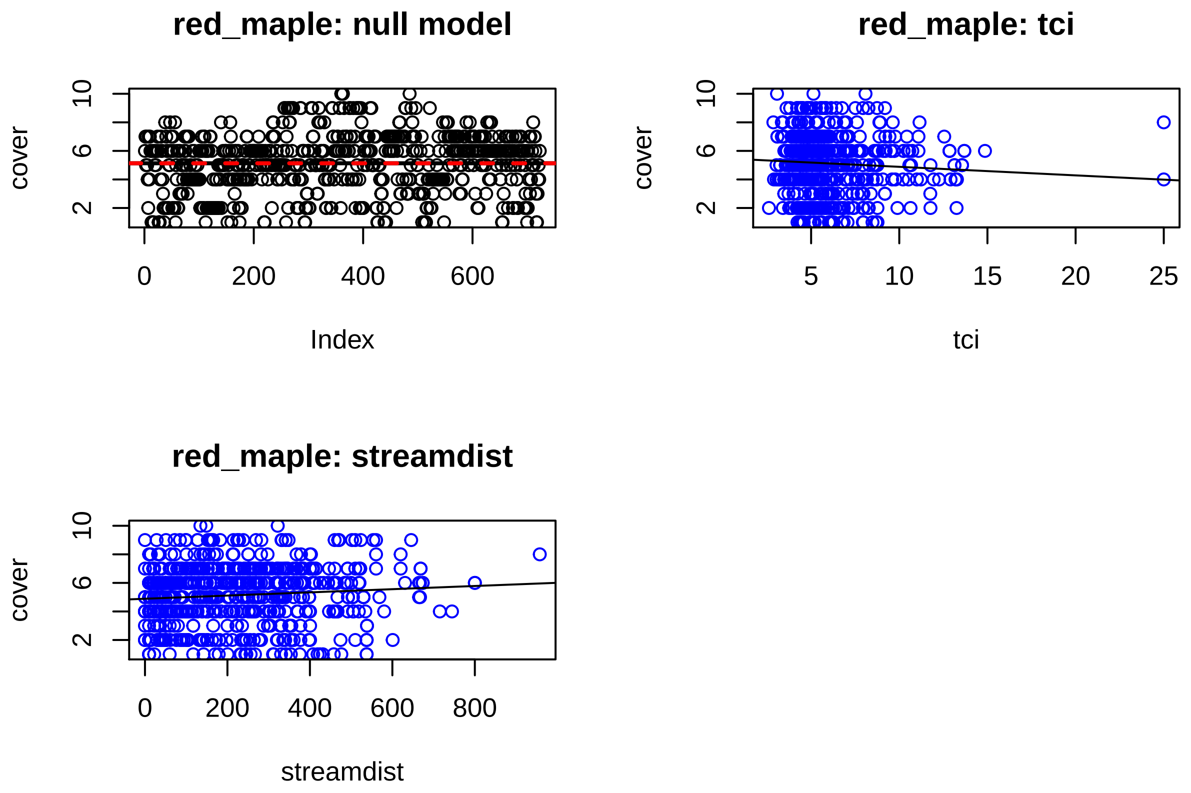
```
par(mfrow=c(2,2))
plot(cover ~ 1, data = red_maple)
title("red_maple: null model")
abline(null_rm, lwd = 2)
abline(h = mean(red_maple$cover), col = 'red', lty = 2, lwd = 2)

tci_rm = lm(cover ~ tci, data = red_maple)
tci_rm
```

```
##
## Call:
## lm(formula = cover ~ tci, data = red_maple)
##
## Coefficients:
## (Intercept)          tci
##     5.48197     -0.05983
```

```r
plot(cover ~ tci, red_maple, pch=1, col='blue')
title("red_maple: tci")
abline(tci_rm)

stream_rm = lm(cover ~ streamdist, data = red_maple)
stream_rm
```

```
##
## Call:
## lm(formula = cover ~ streamdist, data = red_maple)
##
## Coefficients:
## (Intercept)   streamdist
##    4.888126     0.001119
```

```r
plot(cover ~ streamdist, red_maple, pch=1, col='blue')
title("red_maple: streamdist")
abline(stream_rm)
```



```r
#anova(elev_mod)
#anova(sd_mod)
#anova(beers_mod)

#summary(elev_mod)
#summary(sd_mod)
#summary(beers_mod)
```

```r
#summary(tci_rm)
#summary(stream_rm)
```

```r
all_mod = lm(cover ~ elev + streamdist + beers, data=frasier_fir)
#summary(all_mod)

int_mod = lm(cover ~ elev * streamdist * beers, data=frasier_fir)

all_rm = lm(cover ~ elev + streamdist + beers, data=red_maple)
int_rm = lm(cover ~ elev * streamdist * beers, data=red_maple)
```

```r
AIC(null_mod)
```

```
## [1] 199.8769
```

```r
AIC(all_mod)
```

```
## [1] 175.6542
```

```r
AIC(elev_mod)
```

```
## [1] 173.2266
```

```r
AIC(beers_mod)
```

```
## [1] 200.4108
```

```r
AIC(sd_mod)
```

```
## [1] 197.5159
```

```r
AIC(int_mod)
```

```
## [1] 178.8721
```

```r
AIC(null_rm)
```

```
## [1] 3075.185
```

```r
AIC(stream_rm)
```

```
## [1] 3070.921
```

```r
AIC(tci_rm)
```

```
## [1] 3074.046
```

```r
AIC(all_rm)
```

```
## [1] 3054.014
```

```r
AIC(int_rm)
```

```
## [1] 3037.797
```

```r
summary(int_rm)
```

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

```
##     Min      1Q  Median      3Q     Max 
## -4.8056 -1.2068  0.2397  1.3522  5.2787 
## 
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           4.035e+00  6.306e-01   6.398 2.84e-10 ***
## elev                  1.578e-03  8.087e-04   1.952 0.051374 .  
## streamdist            2.740e-03  2.359e-03   1.161 0.245836    
## beers                 1.278e+00  4.983e-01   2.565 0.010506 *  
## elev:streamdist      -2.462e-06  2.860e-06  -0.861 0.389637    
## elev:beers           -2.114e-03  6.072e-04  -3.481 0.000531 ***
## streamdist:beers     -3.272e-04  1.986e-03  -0.165 0.869154    
## elev:streamdist:beers  1.483e-06  2.179e-06   0.681 0.496372    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.964 on 715 degrees of freedom
## Multiple R-squared:  0.06861,    Adjusted R-squared:  0.05949 
## F-statistic: 7.524 on 7 and 715 DF,  p-value: 9.205e-09
```

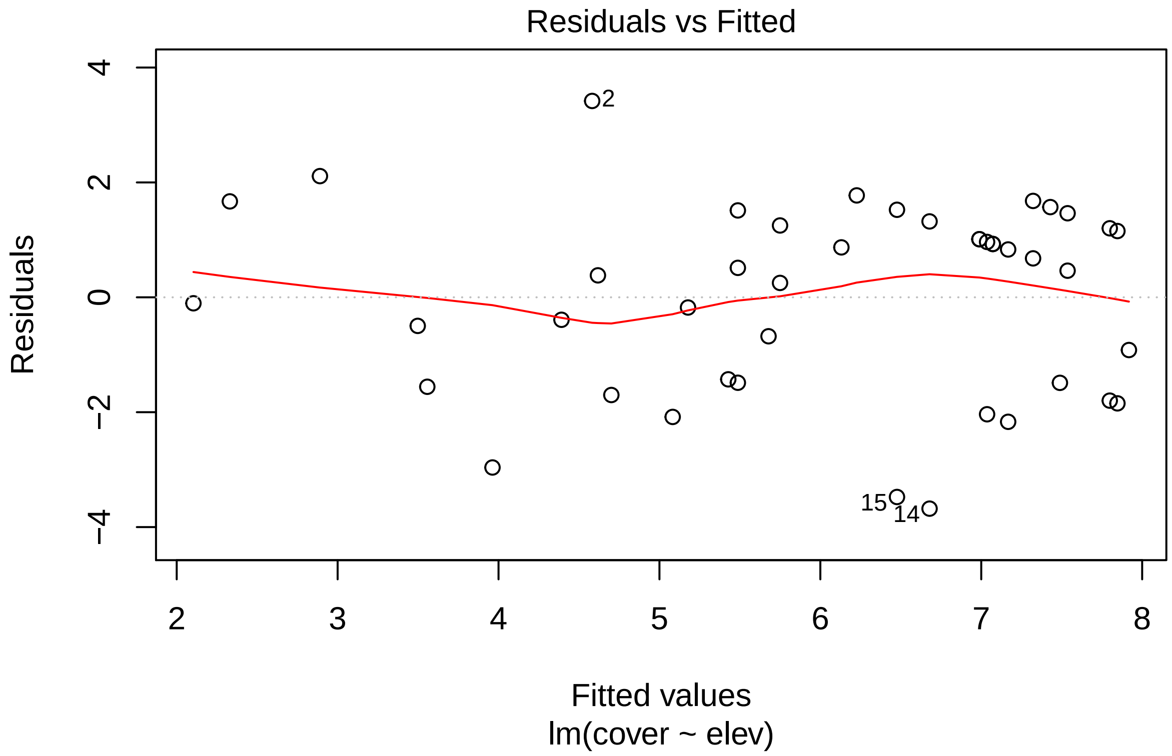
How well does the exploratory model appear to explain cover?

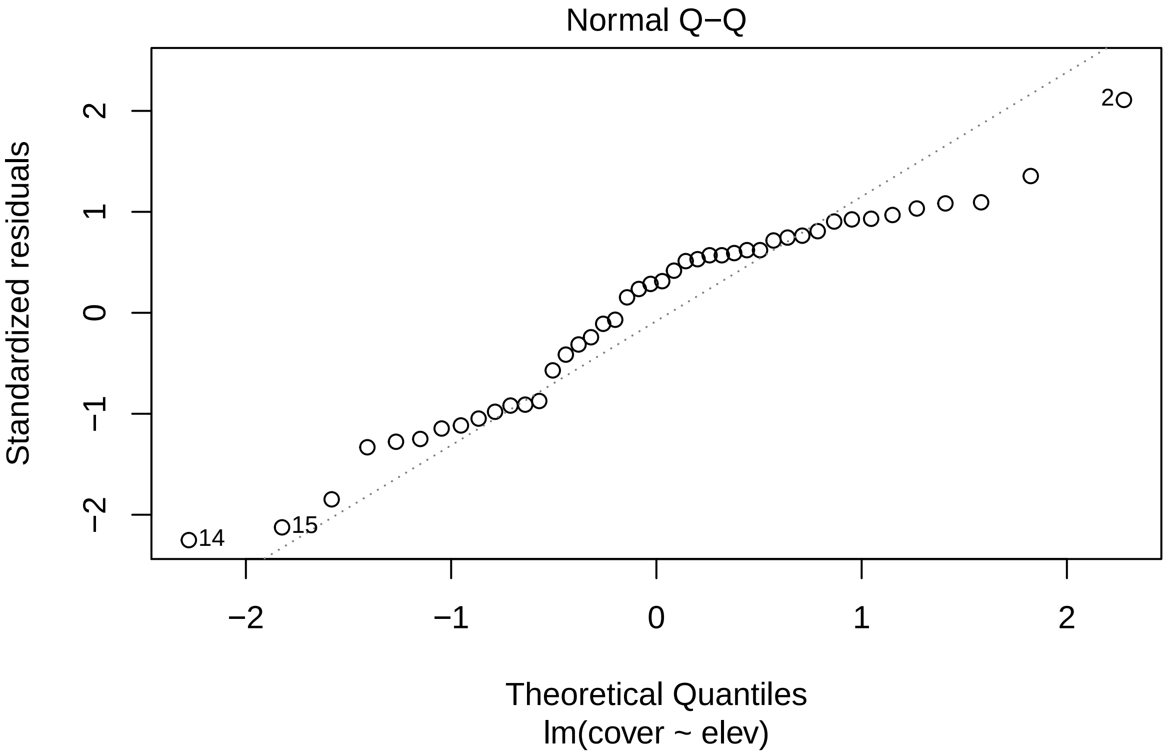
The cover~elev main effect model seems to be the best for frasier firs while the interaction effect model is the best for red maples. However, there doesn't seem to be a good model for the red maples. They do not differ very much from the null model.

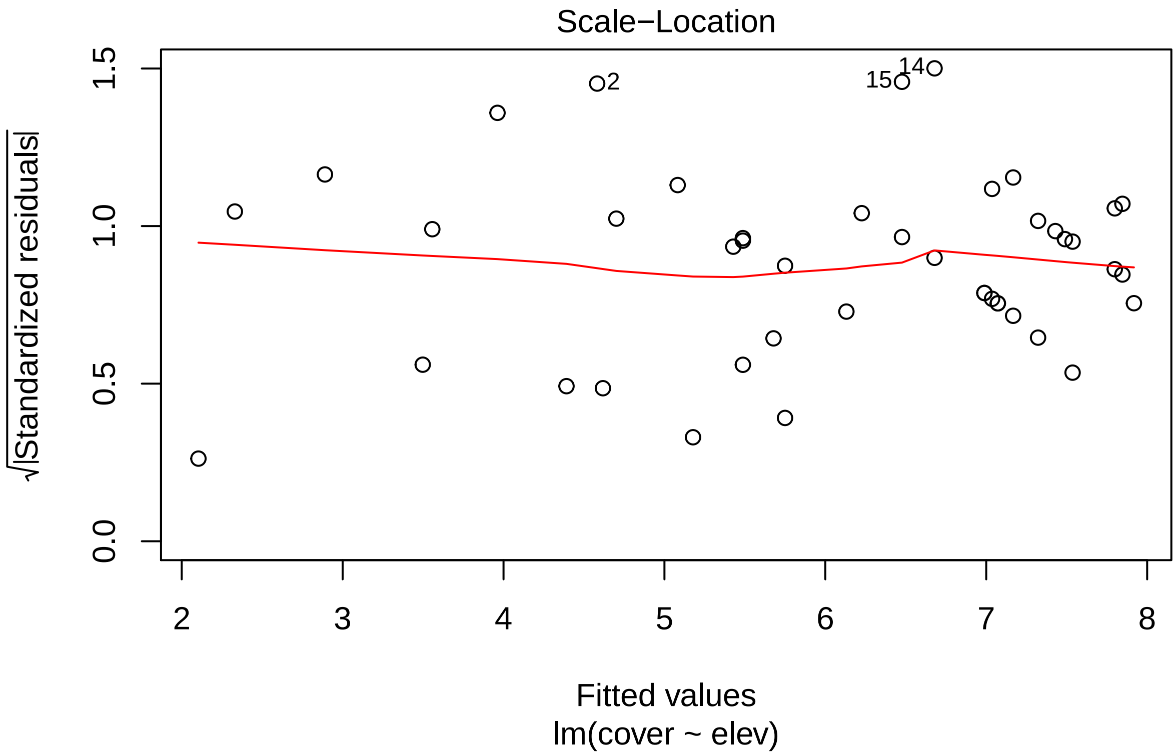
Which explanatory variables are the most important? For frasier firs, elev.

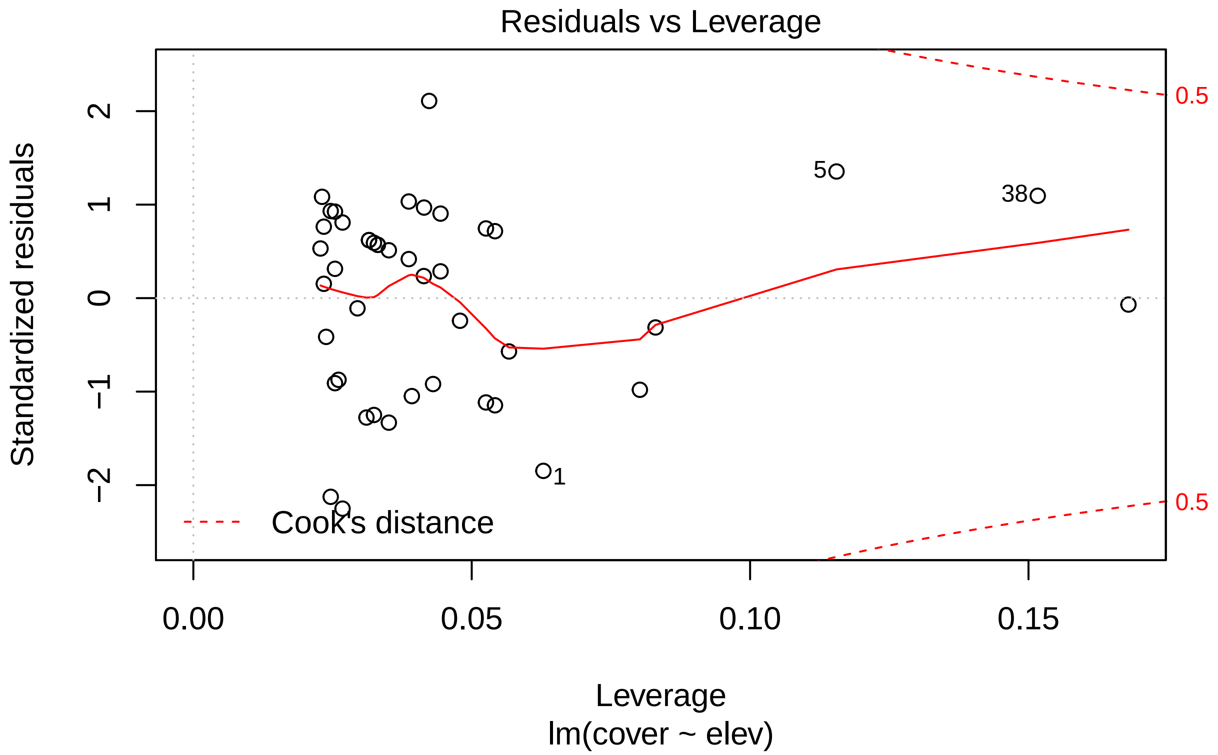
Do model diagnostics indicate any problems with violations of OLS assumptions? No

```
plot(elev_mod)
```


Residuals vs Fitted
Residuals
4
2
0
−2
−4
2
15
14
2
3
4
5
6
7
8
Fitted values
lm(cover ~ elev)


Normal Q−Q
Standardized residuals
−2
−1
0
1
2
14
15
2
−2
−1
0
1
2
Theoretical Quantiles
lm(cover ~ elev)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2
3
4
5
6
7
8
2
15
14
Fitted values
lm(cover ~ elev)



Are you able to explain variance in one species better than another, why might this be the case? The adjusted r-squared indicates that elevation explains almost 50% of the variance in cover. The data for red maple varies too much for a good model to be developed.

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

fras_poi = glm(cover ~ elev, data = frasier_fir,
           family='poisson')
summary(fras_poi)
```

```
##
## Call:
## glm(formula = cover ~ elev, family = "poisson", data = frasier_fir)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7627  -0.5757   0.2390   0.4400   1.5342
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.3784141  1.0143217  -2.345    0.019 *
## elev         0.0022556  0.0005425   4.158 3.21e-05 ***
## ---
## 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: 22.180  on 42  degrees of freedom
## AIC: 183.36
##
## Number of Fisher Scoring iterations: 4
```

```
summary(elev_mod)
```

```
##
## Call:
## lm(formula = cover ~ elev, data = frasier_fir)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.679 -1.488  0.488  1.214  3.418
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.814670   3.526227  -4.485 5.56e-05 ***
## elev          0.011914   0.001919   6.208 1.99e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.656 on 42 degrees of freedom
## Multiple R-squared:  0.4786, Adjusted R-squared:  0.4661
## F-statistic: 38.54 on 1 and 42 DF,  p-value: 1.991e-07
```

```
AIC(fras_poi)
```

```
## [1] 183.3563
```

```
pseudo_r2(fras_poi)
```

```
## [1] 0.4626207
```

```
acer_poi = glm(cover ~ elev * streamdist * beers, data = red_maple,
           family='poisson')
summary(acer_poi)
```

```
##
## Call:
## glm(formula = cover ~ elev * streamdist * beers, family = "poisson",
##     data = red_maple)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.47688  -0.56716   0.09887   0.57590   2.25635
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              1.429e+00  1.388e-01  10.291  < 2e-16 ***
## elev                     3.002e-04  1.758e-04   1.707  0.08782 .
## streamdist               4.884e-04  5.094e-04   0.959  0.33765
## beers                    2.638e-01  1.119e-01   2.358  0.01839 *
## elev:streamdist         -4.382e-07  6.177e-07  -0.709  0.47802
## elev:beers            -4.378e-04  1.373e-04  -3.188  0.00143 **
## streamdist:beers      -8.942e-05  4.347e-04  -0.206  0.83703
## elev:streamdist:beers  3.323e-07  4.826e-07   0.689  0.49113
## ---
## 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: 608.60  on 715  degrees of freedom
## AIC: 3087
##
## Number of Fisher Scoring iterations: 4
```

```
summary(int_rm)
```

```
##
## Call:
## lm(formula = cover ~ elev * streamdist * beers, data = red_maple)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8056 -1.2068  0.2397  1.3522  5.2787
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.035e+00  6.306e-01   6.398 2.84e-10 ***
## elev                   1.578e-03  8.087e-04   1.952 0.051374 .
## streamdist             2.740e-03  2.359e-03   1.161 0.245836
## beers                  1.278e+00  4.983e-01   2.565 0.010506 *
## elev:streamdist       -2.462e-06  2.860e-06  -0.861 0.389637
## elev:beers            -2.114e-03  6.072e-04  -3.481 0.000531 ***
## streamdist:beers      -3.272e-04  1.986e-03  -0.165 0.869154
## elev:streamdist:beers  1.483e-06  2.179e-06   0.681 0.496372
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.964 on 715 degrees of freedom
## Multiple R-squared:  0.06861,    Adjusted R-squared:  0.05949
## F-statistic: 7.524 on 7 and 715 DF,  p-value: 9.205e-09
```

```
AIC(acer_poi)
```

```
## [1] 3086.985
```

```
pseudo_r2(acer_poi)
```

```
## [1] 0.06274571
```

Compare your qualatitive assessment of which variables were most important in each model. Does it appear that changing the error distribution changed the results much? In what ways? This doesn't seem to change the results much and may be worse for red maple based on AIC.

Provide a plain English summary (i.e., no statistics) of what you have found and what conclusions we can take away from your analysis? For frasier firs, elevation, stream distance, and beers seem to have an effect on coverage. Out of these three variables, elevation seems to have the most effect. Much more observations were made for the red maple, but I could not determine a clear correlation between cover and any other variables.

The data seemed to be too evenly spread for a trend to be found.