Solutions to the Univariate Models assignment

Univariate Assignment

Read in tree data, metadata can be found here.

```
library(car)
```

```
## Loading required package: carData
```

```
library(MASS)
library(GGally)
```

```
## Loading required package: ggplot2

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
source('https://raw.githubusercontent.com/dmcglinn/quant_methods/gh-pages/scripts/utility_functions.R')

trees <- read.csv('https://raw.githubusercontent.com/dmcglinn/quant_methods/gh-pages/data/treedata_subs

trees$disturb <- as.factor(trees$disturb)


# each row provides a cover estimate for a different species
# in a different plot
```

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`. After loading `car` we can call the function like so:

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

   This will estimate partial effect sizes, variance explained, and p-values for each explanatory variable included in the model.

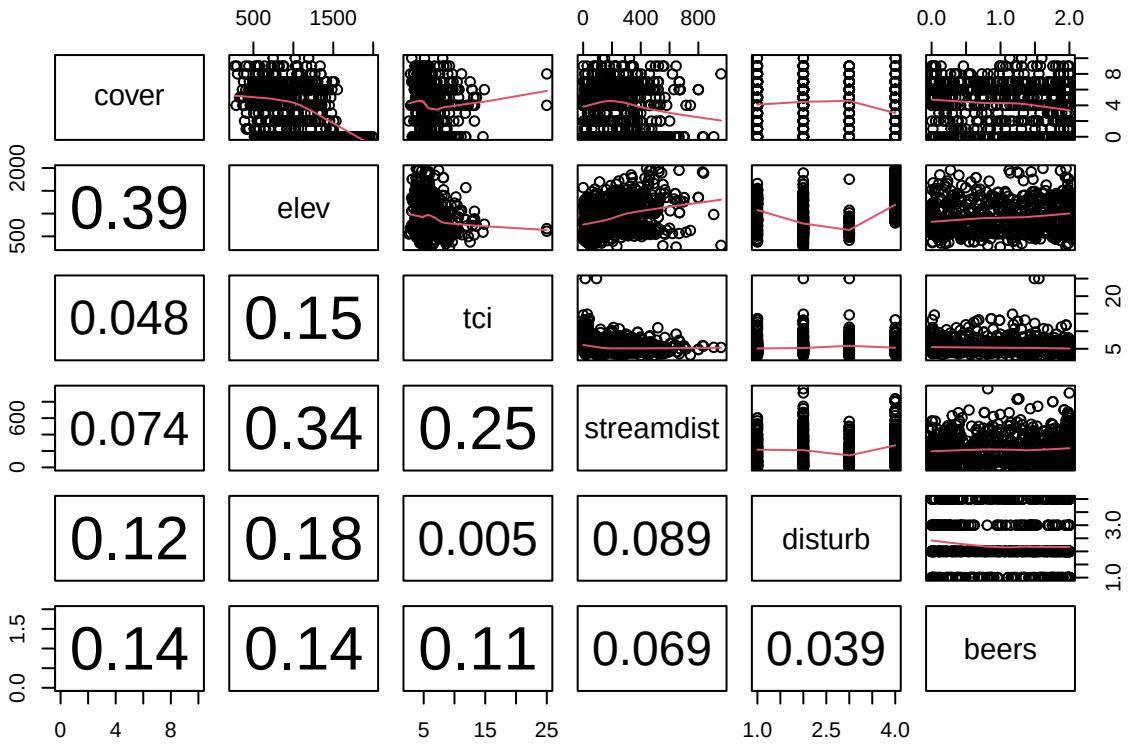
   Compare the p-values you observe using the function `Anova` to those generated using `summary`.

   For each species address the following additional questions:

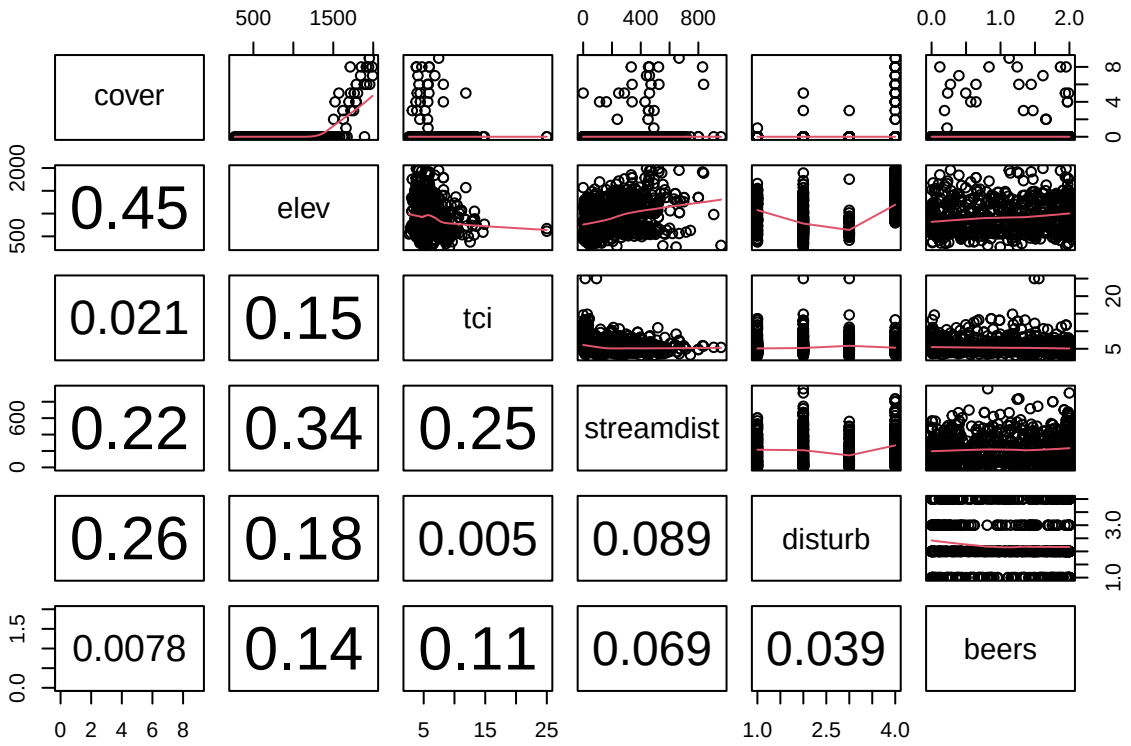
   ```
   * how well does the exploratory model appear to explain cover?
   * which explanatory variables are the most important?
   * do model diagnostics indicate any problems with violations of
     OLS assumptions?
   * are you able to explain variance in one species better than another?
   ```

```r
# we wish to model species cover across all sampled plots

# create site x sp matrix for two species
sp_cov <- with(trees, tapply(cover, list(plotID, spcode),
                             function(x) round(mean(x))))
sp_cov <- ifelse(is.na(sp_cov), 0, sp_cov)
sp_cov <- data.frame(plotID = row.names(sp_cov), sp_cov)

# create environmental matrix
cols_to_select <- c('elev', 'tci', 'streamdist',
                    'disturb', 'beers')
env <- aggregate(trees[ , cols_to_select], by = list(trees$plotID),
                 function(x) x[1])
names(env)[1] = 'plotID'

# merge species and enviornmental matrices
site_dat <- merge(sp_cov, env, by='plotID')

# subset species of interest
abies <- site_dat[ , c('ABIEFRA', cols_to_select)]
acer  = site_dat[ , c('ACERRUB', cols_to_select)]
names(abies)[1] = 'cover'
names(acer)[1] = 'cover'
```

```r
# prior to model fitting I will visually examine correlations with cover
pairs(acer, lower.panel = panel.cor, upper.panel = panel.smooth)
```



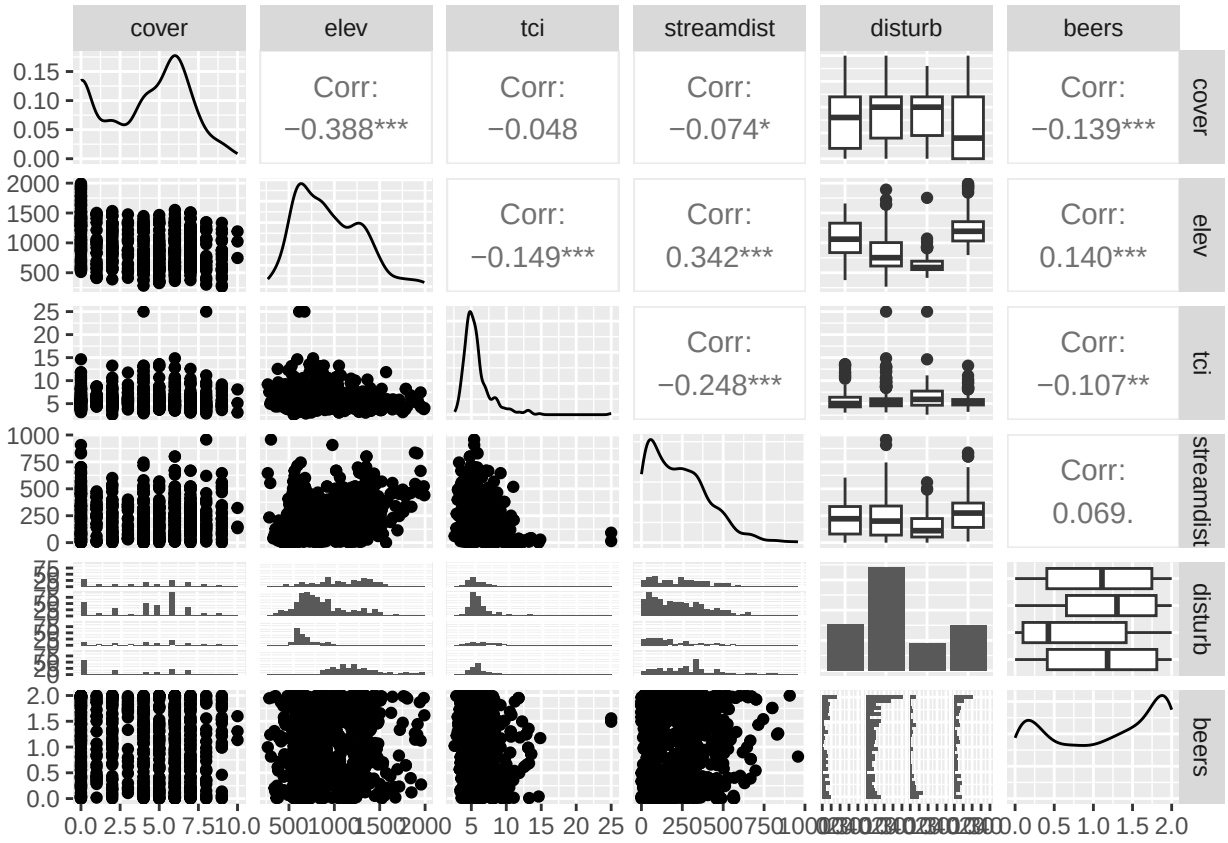

```
pairs(abies, lower.panel = panel.cor, upper.panel = panel.smooth)
```



The `GGally` package also has a slick ggplot option for pairs plots

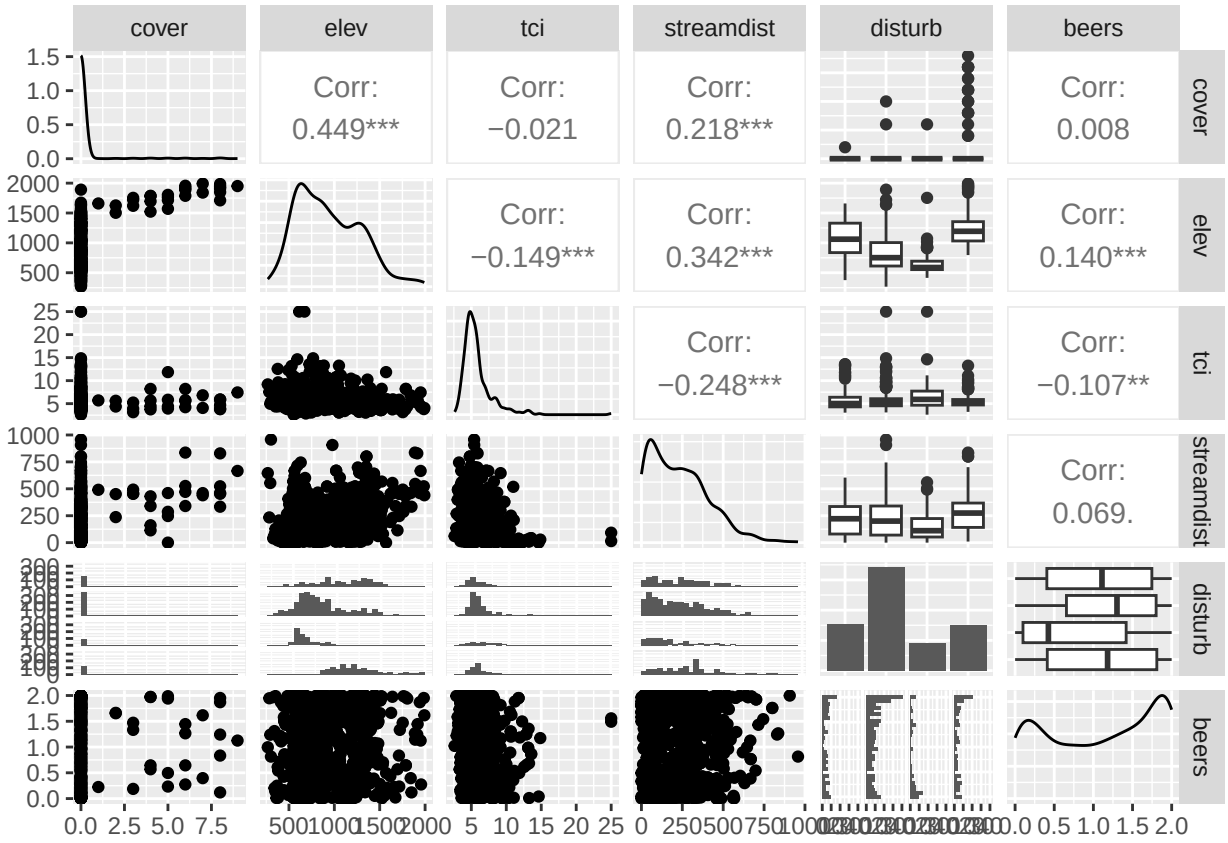
```
ggpairs(acer)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggpairs(abies)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



These plots already tell us the take home message in many ways which is that Red maple cover is not well explained by the measured enviornmental variables with the exception of elevation. Fraser fir (*Abies fraseri*) may correlate better with the enviornmental variables but it is difficult to tell because it has a cover of zero in so many plots.

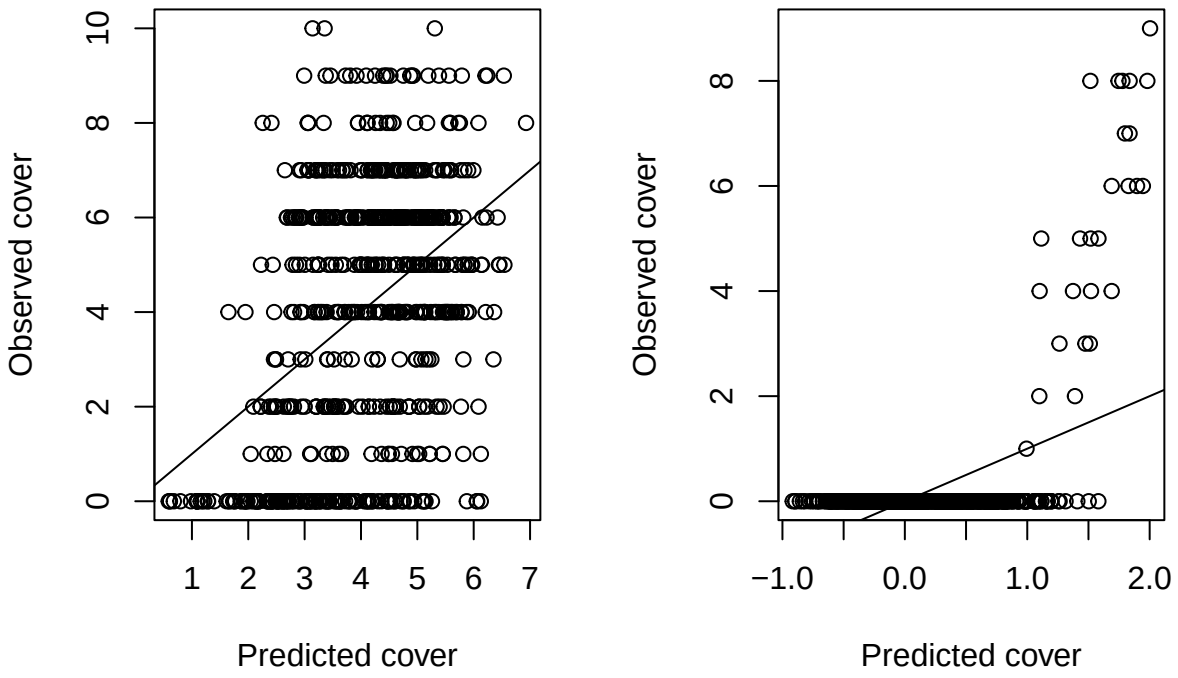
Now we will carry out OLS model fitting. As this is an explanatory modeling we will include all of the variables in the models. We do this for several reasons:

1. this is the best way to fairly compete the variables against one another
2. a partial correlation between the response and an explanatory variable may be stronger than its raw correlation due to interference from other variables.

```
# build OLS models using "." shorthand which is short for all variables in data
acer_lm <- lm(cover ~ . , data=acer)
abies_lm <- lm(cover ~ . , data=abies)
```

Let's first examine the overall model fit by examining the predicted-observed (PO) plot:

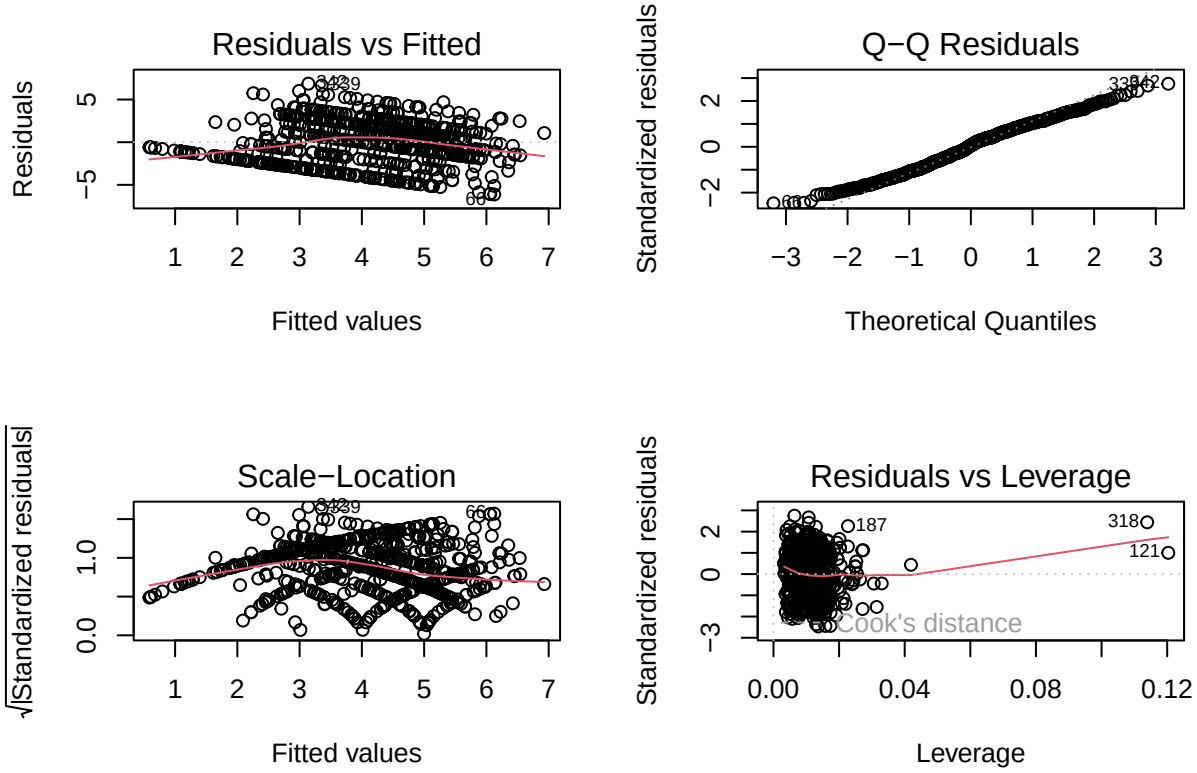
```
par(mfrow=c(1, 2))
plot(predict(acer_lm), acer$cover, xlab='Predicted cover', ylab='Observed cover')
abline(a=0, b=1)
plot(predict(abies_lm), abies$cover, xlab='Predicted cover', ylab='Observed cover')
abline(a=0, b=1)
```



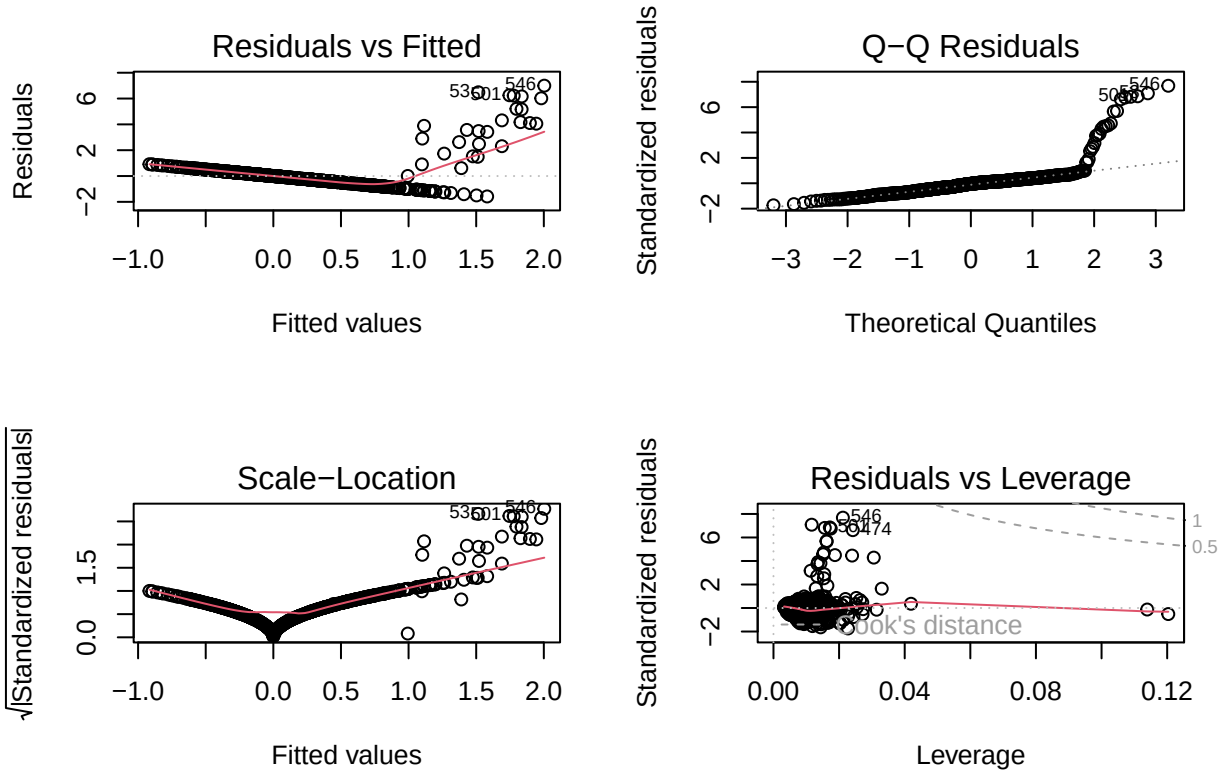

```
par(mfrow=c(1, 1))
```

There are some pretty large and systematic deviations from the one-to-one line which suggests that our model is performing poorly for both species. Additionally it is obvious that observed cover values are integers while the predicted values are continuous. Negative cover is predicted for *Abies fraseri* but this doesn't make any sense. We'll return to this idea later when we fit the Poisson GLM model which only predicts positive integer values.

```
# before diving into statistics we should check model diagonstistics
par(mfrow=c(2,2))
plot(acer_lm)
```




```
plot(abies_lm)
```




```
par(mfrow=c(1,1))
```

Those diagnostic plots are waving some red flags. In particular, the "Residuals vs Fitted" and "Scale-Location" plots indicate that their are some systematic deviations of the residuals. You should not see regular geometric patterns in residuals so that is a bit troubling. These systematic errors are occurring because the response `cover` is a discrete variable and the Gaussian error term in our OLS assumes a continuous response. So there is a mismatch between our error distribution and the variable we are modeling.

Now let's examine the output from the functions `summary` and `Anova` to examine 1. partial variable importance (t-value effect size and significance) 2. overall model performance (adjusted r-squared and significance)

```
summary(acer_lm)
```

```
##
## Call:
## lm(formula = cover ~ ., data = acer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1258 -1.9460  0.1577  1.8624  6.8596
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.7372607  0.5086637  17.177  < 2e-16 ***
## elev         -0.0034639  0.0003362 -10.304  < 2e-16 ***
```

```
## tci           -0.1317294  0.0440921  -2.988  0.00291 **
## streamdist     0.0007520  0.0005711   1.317  0.18832
## disturbLT-SEL -0.4379126  0.2559816  -1.711  0.08756 .
## disturbSETTLE -0.9309789  0.3564239  -2.612  0.00919 **
## disturbVIRGIN -0.3601527  0.2941812  -1.224  0.22125
## beers         -0.4101716  0.1381555  -2.969  0.00309 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.501 on 726 degrees of freedom
## Multiple R-squared:  0.1805, Adjusted R-squared:  0.1726
## F-statistic: 22.85 on 7 and 726 DF,  p-value: < 2.2e-16
```

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

```
## Warning in printHypothesis(L, rhs, names(b)): one or more coefficients in the hypothesis include
##      arithmetic operators in their names;
##   the printed representation of the hypothesis will be omitted

## Anova Table (Type III tests)
##
## Response: cover
##             Sum Sq  Df  F value    Pr(>F)
## (Intercept) 1845.7   1 295.0456 < 2.2e-16 ***
## elev         664.1   1 106.1624 < 2.2e-16 ***
## tci           55.8   1   8.9257  0.002907 **
## streamdist    10.8   1   1.7340  0.188316
## disturb       44.1   3   2.3479  0.071433 .
## beers         55.1   1   8.8144  0.003087 **
## Residuals   4541.7 726
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Overall the model for *Acer rubrum* explained 17% of the variance in cover based upon the computed adjusted R^2. This actually would not necessarily be considered a terrible R^2 value if it were not for the systematic deviations we already observed in the fitted and residual values of the model which indicate a fundamentally flawed model.

The strongest correlations for this species are negative with respect to `elev`, `tci` and `beers` as evidenced by their negative t-values which were significantly different than zero. This indicates that this species prefers lower elevations that are dryer and warmer.

```
summary(abies_lm)
```

```
##
## Call:
## lm(formula = cover ~ ., data = abies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5809 -0.4502 -0.0420  0.2346  6.9968
##
## Coefficients:
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.5705179  0.1871566  -8.391 2.50e-16 ***
## elev           0.0013315  0.0001237  10.764  < 2e-16 ***
## tci            0.0255969  0.0162231   1.578 0.115046
## streamdist     0.0004553  0.0002101   2.167 0.030574 *
## disturbLT-SEL  0.3248142  0.0941853   3.449 0.000596 ***
## disturbSETTLE  0.5408814  0.1311418   4.124 4.15e-05 ***
## disturbVIRGIN  0.5584644  0.1082404   5.159 3.20e-07 ***
## beers         -0.0675883  0.0508326  -1.330 0.184059
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9203 on 726 degrees of freedom
## Multiple R-squared:  0.2478, Adjusted R-squared:  0.2406
## F-statistic: 34.17 on 7 and 726 DF,  p-value: < 2.2e-16
```

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

```
## Warning in printHypothesis(L, rhs, names(b)): one or more coefficients in the hypothesis include
##      arithmetic operators in their names;
##   the printed representation of the hypothesis will be omitted

## Anova Table (Type III tests)
##
## Response: cover
##             Sum Sq  Df  F value    Pr(>F)
## (Intercept)  59.64   1  70.4167 2.501e-16 ***
## elev         98.13   1 115.8739 < 2.2e-16 ***
## tci           2.11   1   2.4895   0.11505
## streamdist    3.98   1   4.6951   0.03057 *
## disturb      28.40   3  11.1771 3.545e-07 ***
## beers         1.50   1   1.7679   0.18406
## Residuals   614.85 726
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For *Abies fraseri* the results are a bit different. This species is slightly better explained by the explanatory variables (adjusted $R^2 = 0.24$).

The ANOVA table indicates that this species is primarily driven by `elev` with a secondary role for `disturbance`. Note the F-statistics for those variables are much larger. The `summary` output which provides the coefficient estimates suggests that this species prefers high elevations (positive regression coefficient) and VIRGIN forests (the most positive β of the disturbance categories) - remember these β's are relative to the intercept which in this case is the disturbance category CORPLOG).

We were able to model the habitat specialist, *Abies fraseri*, about as well as the more generalist species *Acer rubrum*, but we have some serious concerns and doubts about our inferences because the model diagnostics and common sense suggest that a Gaussian error for this response variables is inappropriate.

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. Your new model calls will look as follows:

```
acer_poi <- glm(cover ~ tci + elev + ... , data= my_data,
                family='poisson')
```

For assessing the degree of variation explained you can use a pseudo-R-squared statistic (note this is just one of many possible)

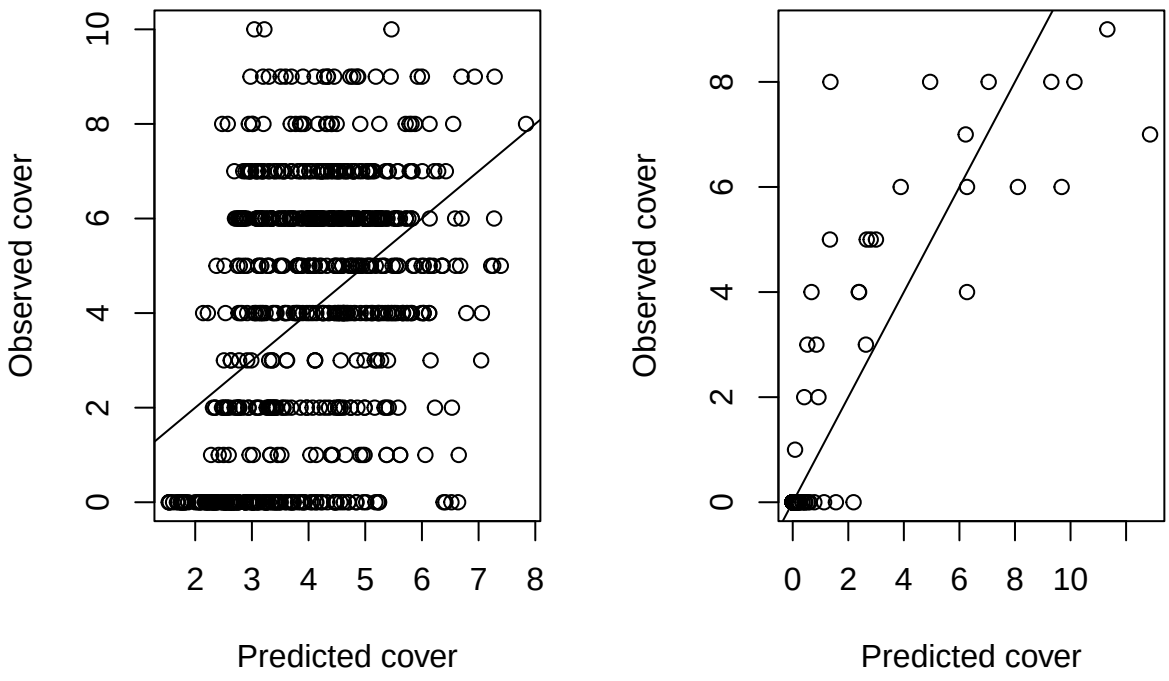
```
pseudo_r2 <- function(glm_mod) {
    1 -  glm_mod$deviance / glm_mod$null.deviance
}
```

Compare your qualative 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?

```
acer_poi <- glm(cover ~ . , data = acer, family = 'poisson')
abies_poi <- glm(cover ~ . , data = abies, family = 'poisson')
```

Let's examine the predicted-observed plots first to see if they systematic error is better handled.

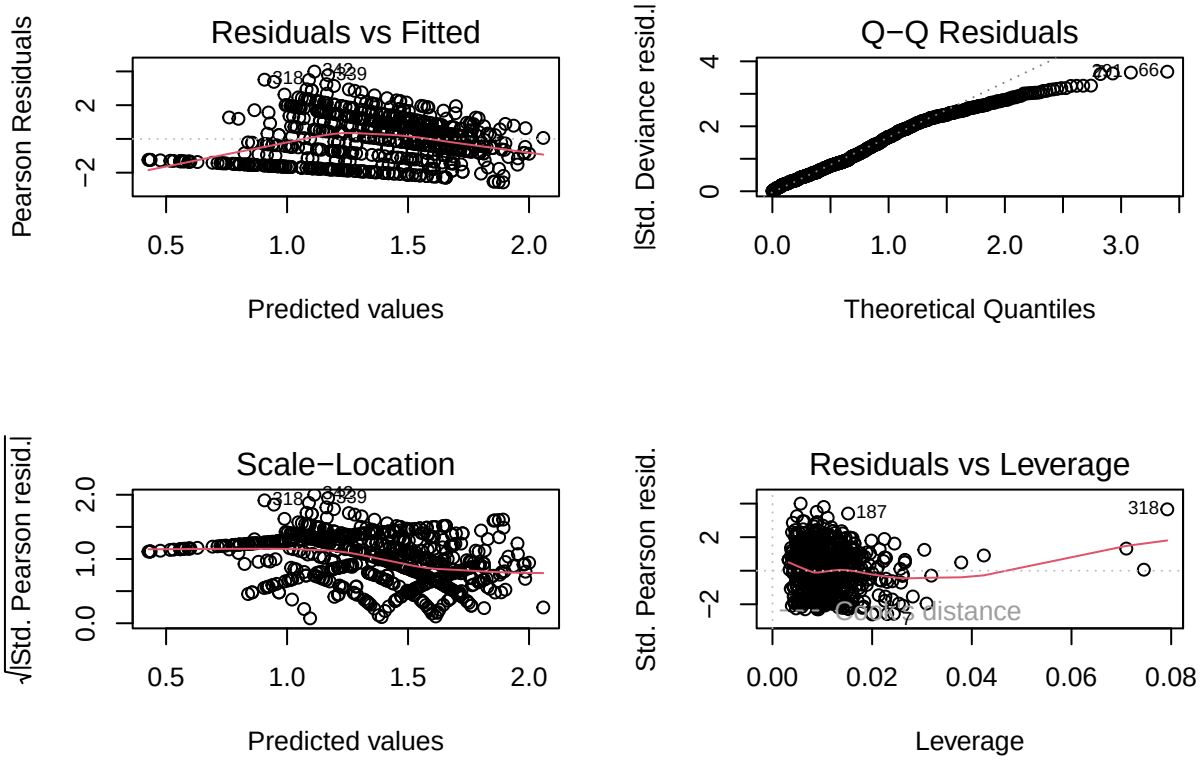
```
par(mfrow=c(1, 2))
plot(predict(acer_poi, type='response'), acer$cover,
     xlab='Predicted cover', ylab='Observed cover')
abline(a=0, b=1)
plot(predict(abies_poi, type='response'), abies$cover,
     xlab='Predicted cover', ylab='Observed cover')
abline(a=0, b=1)
```



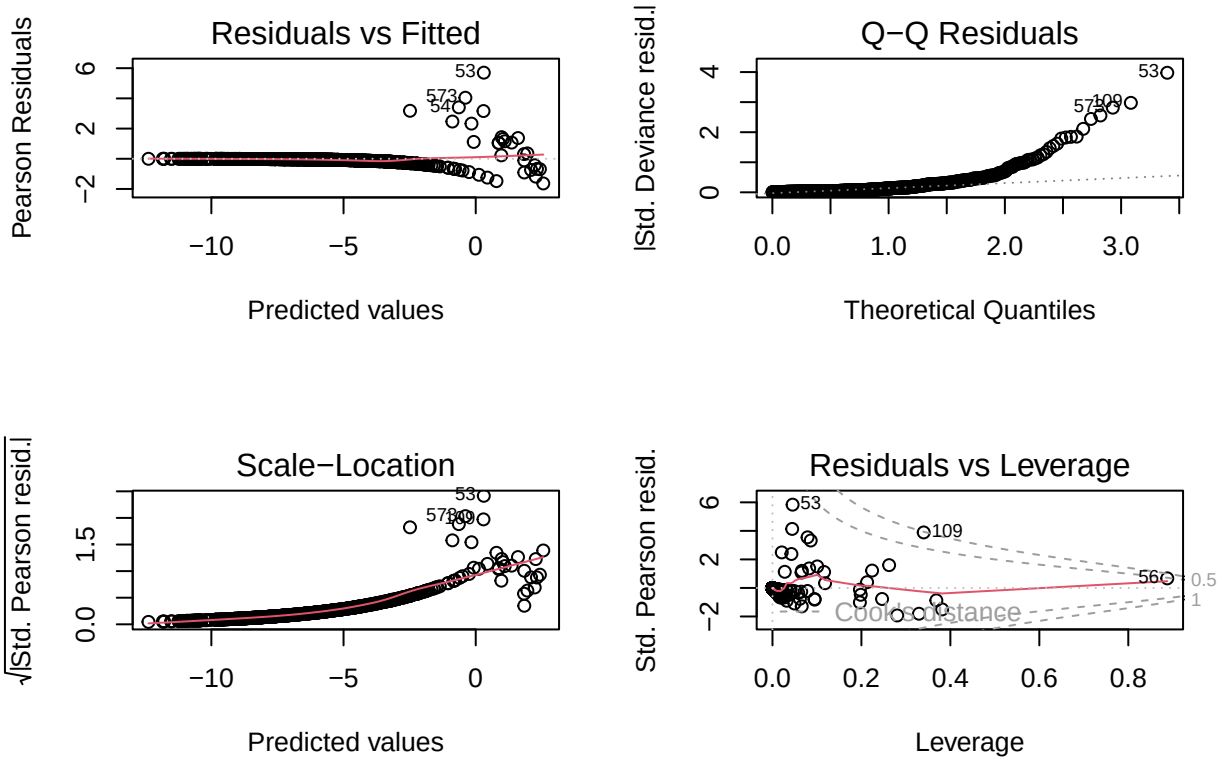
```
par(mfrow=c(1, 1))
```

These PO plots already look much more encouraging. There is still quite a bit of error but it is less systematic and their are no predicted negative cover values. Let's look at the model diagnostics:

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




```
plot(abies_poi)
```

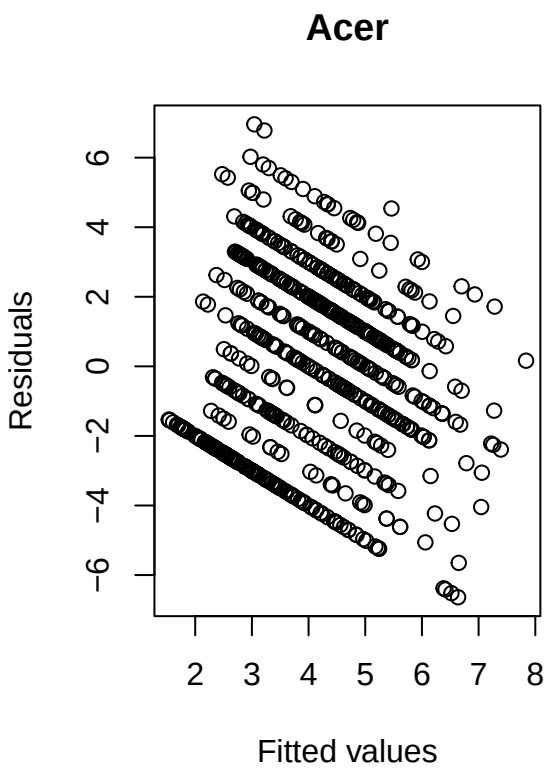


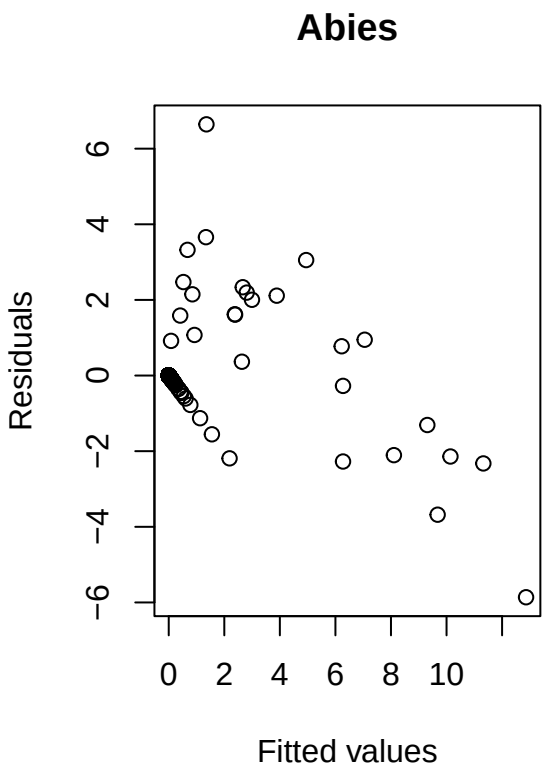

```
par(mfrow=c(1,1))
```

These model diagnostics look better but still not great. We can still observe some systematic patterns of error in the residuals plots. Also the Fraser fir residuals are still decidedly non-Normal. This seems to be due in part to the large number of samples with zero cover of *Abies fraserii*.

One important thing to note is that the predicted and residual values in the above plots are on the transformed log scale. It can be more intuitive and informative to plot them on the response scale as such

```
par(mfrow=c(1,2))
plot(predict(acer_poi, type='response'),
     residuals(acer_poi, type='response'),
     xlab='Fitted values', ylab='Residuals',
     main='Acer')
plot(predict(abies_poi, type='response'),
     residuals(abies_poi, type='response'),
     xlab='Fitted values', ylab='Residuals',
     main='Abies')
```






```
par(mfrow=c(1,1))
```

Before we examine differences in the model outputs, let's examine if using the Poisson error term improved model fit.

```
summary(acer_lm)$r.squared
```

```
## [1] 0.1805288
```

```
pseudo_r2(acer_poi)
```

```
## [1] 0.1342074
```

```
summary(abies_lm)$r.squared
```

```
## [1] 0.2478266
```

```
pseudo_r2(abies_poi)
```

```
## [1] 0.8951796
```

For Red maple the fit is worse but for Fraser fir we observe a big improvement in fit. This is in large part due to the fact that the Poisson distribution is truncated at zero and given that for Fraser fir there are a lot of sites with zero cover where the Gaussian model was doing a very poor job.

The PO plots For both Poisson models suggest that the models tend to under predict the occurrence of the zero category. This is most pronounced for Fraser fir. So even though we've explicitly recognized the positive discrete nature of our response variable it is still inflated with zeros relative to a Poisson distribution.

Given that the Poisson regression actually performed worse for the Red maple than the OLS regression. I'll not interpret that model further.

```
summary(abies_poi)
```

```
##
## Call:
## glm(formula = cover ~ ., family = "poisson", data = abies)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.564e+01  1.359e+00 -11.509  < 2e-16 ***
## elev           7.850e-03  5.883e-04  13.343  < 2e-16 ***
## tci            1.688e-01  5.927e-02   2.848  0.00440 **
## streamdist    -1.692e-03  6.368e-04  -2.658  0.00787 **
## disturbLT-SEL  1.622e+00  1.068e+00   1.518  0.12904
## disturbSETTLE  3.174e+00  1.161e+00   2.733  0.00628 **
## disturbVIRGIN  2.649e+00  1.025e+00   2.584  0.00976 **
## beers         -1.826e-02  1.515e-01  -0.120  0.90409
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 940.37  on 733  degrees of freedom
## Residual deviance:  98.57  on 726  degrees of freedom
## AIC: 203.75
##
## Number of Fisher Scoring iterations: 7
```

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

```
## Analysis of Deviance Table (Type III tests)
##
## Response: cover
##            LR Chisq Df Pr(>Chisq)
## elev         420.83  1  < 2.2e-16 ***
## tci            7.34  1   0.006742 **
## streamdist     7.63  1   0.005748 **
## disturb       21.89  3  6.863e-05 ***
## beers          0.01  1   0.904161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The take home messages we arrived at with the OLS modeling have not changed greatly for Fraser fir. More variables are statistically significant but the real story remains about the strong effect of elevation and disturbance.

At this point we may still not be very satisfied with our modeling though given the excess of zeros we observed above the Fraser firs. One option would be to try to use a negative binomial error term rather than Poisson which provides for greater aggregation of zeros due to the inclusion of an additional clumping parameter.

```
library(MASS)

abies_nb <- glm.nb(cover ~ . , data=abies,
                   control=glm.control(maxit=100))
AIC(abies_poi)
```

```
## [1] 203.7475
```

```
AIC(abies_nb)
```

```
## [1] 198.8835
```

That resulted in a modest decrease in the AIC. One last model to examine is a zero inflated Poisson model in which combines two models:

- a logistic regression for whether or not cover is zero or not, and
- a Poisson regression for variation in the size of cover.

```
library(pscl)
```

```
## Classes and Methods for R originally developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University (2002-2015),
## by and under the direction of Simon Jackman.
## hurdle and zeroinfl functions by Achim Zeileis.
```

```
# fit a model in which all the variables are included
# in the portion of the model with positive values, and
# only include elevation in the model for the zeros.

abies_zip <- zeroinfl(cover ~ . | elev, data=abies)

AIC(abies_poi)
```

```
## [1] 203.7475
```

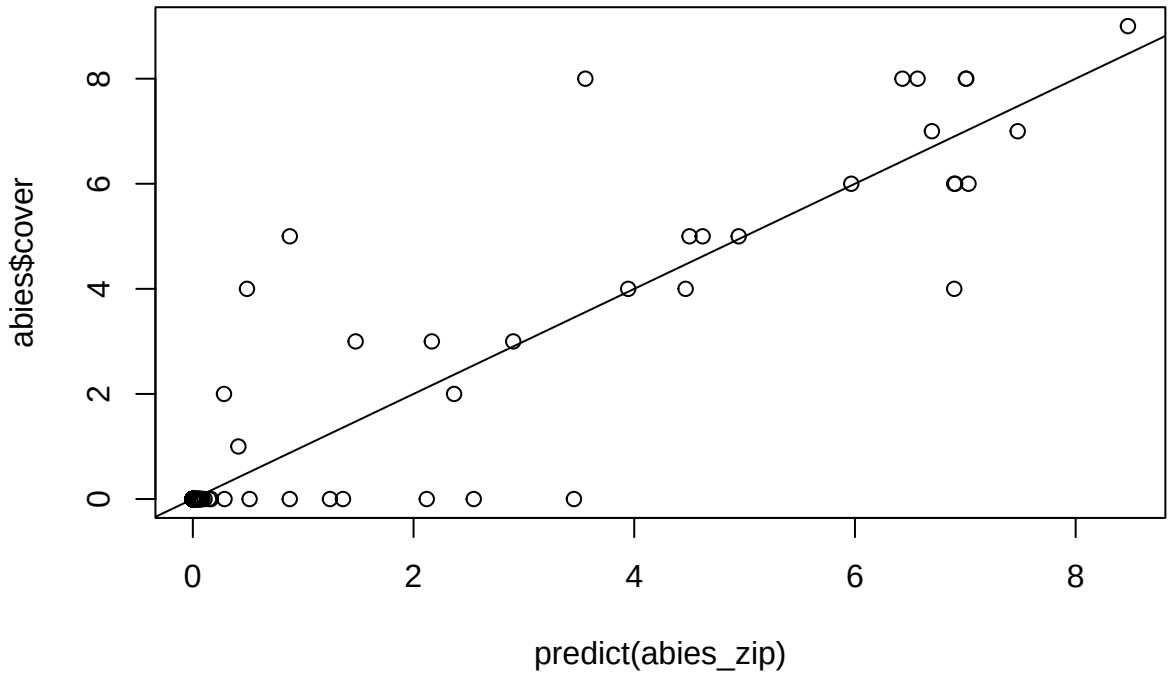
```
AIC(abies_nb)
```

```
## [1] 198.8835
```

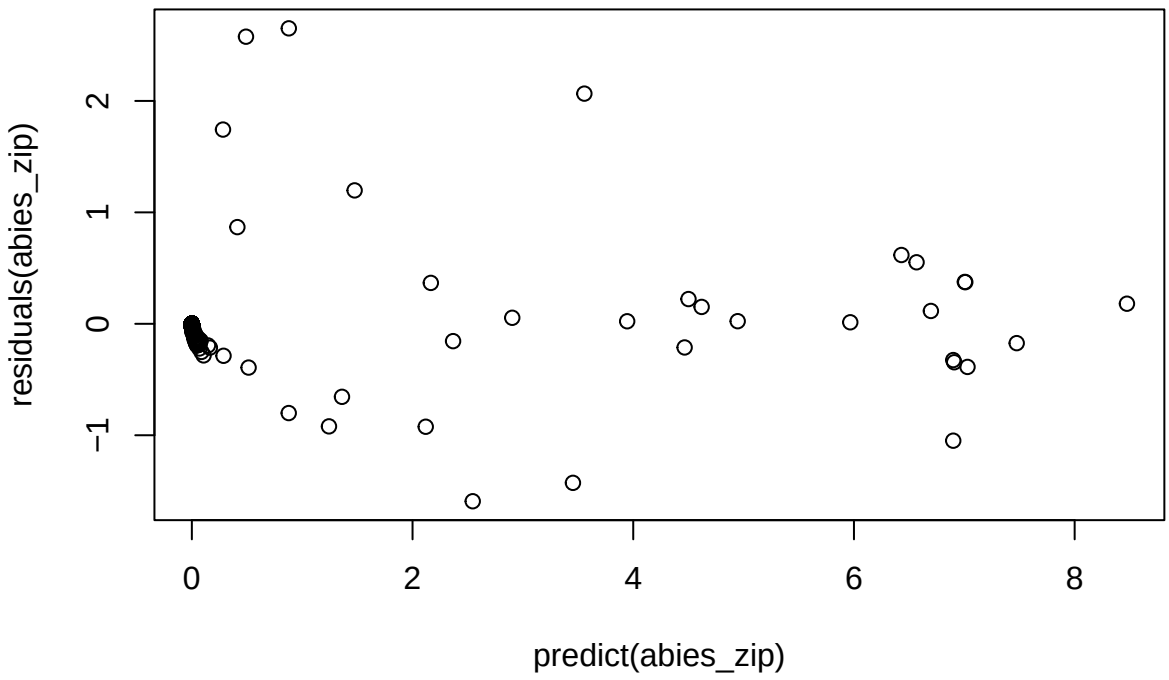
```
AIC(abies_zip)
```

```
## [1] 151.3478
```

```r
# PO plot
plot(predict(abies_zip), abies$cover)
abline(a=0, b=1)
```




```r
# diagonstic plot
plot(predict(abies_zip), residuals(abies_zip))
```



This model resulted in a substantial increase in the model adequacy as judged by the much lower AIC. The model diagnostic plot looks more reasonable and when we examine the predicted to observed plot we see that we are doing a better job predicting all the zeros in the dataset.

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

```
## Warning in printHypothesis(L, rhs, names(b)): one or more coefficients in the hypothesis include
##      arithmetic operators in their names;
##   the printed representation of the hypothesis will be omitted
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: cover
##             Df     Chisq Pr(>Chisq)
## elev         1 -241.7802    1.00000
## tci          1    1.9544    0.16211
## streamdist   1    0.5385    0.46304
## disturb      3    7.8730    0.04871 *
## beers        1    0.0000    0.99882
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above output from `Anova` isn't so helpful. We see that `disturb` is marginally important but in this case `summary` provides more useful information.

```
summary(abies_zip)
```

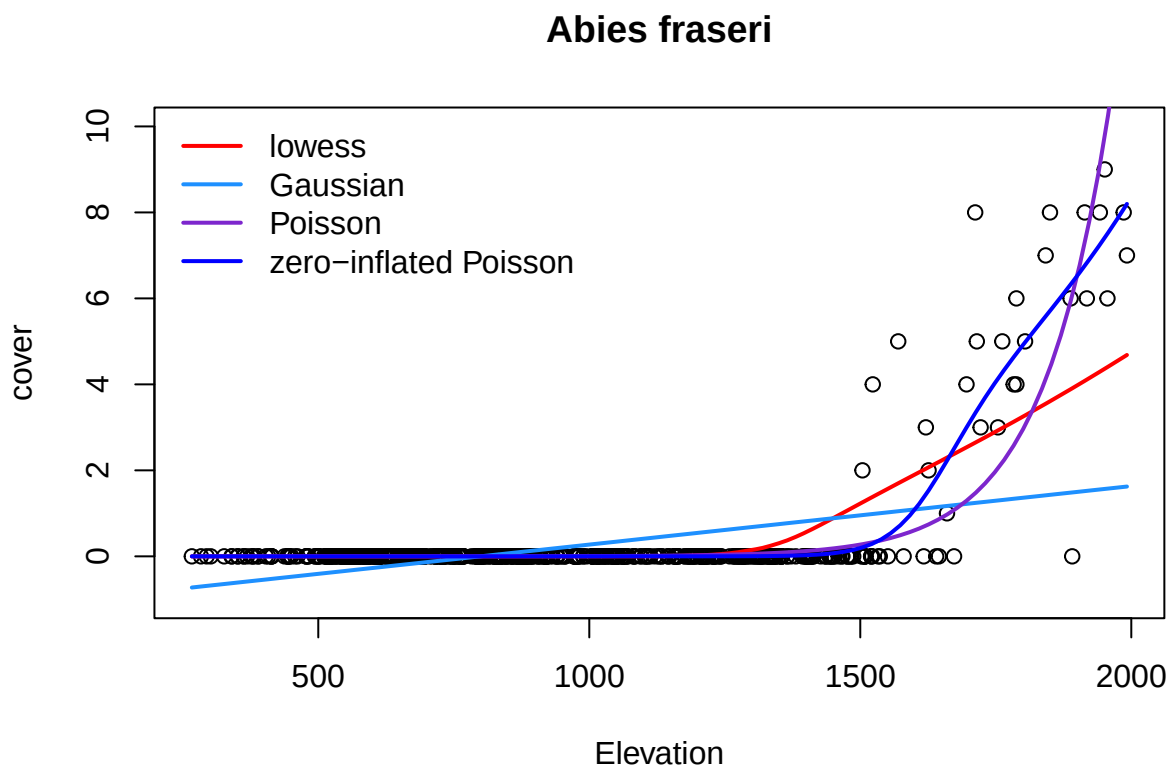
```
## Warning in sqrt(diag(object$vcov)): NaNs produced
```

```
##
## Call:
## zeroinfl(formula = cover ~ . | elev, data = abies)
##
## Pearson residuals:
##        Min         1Q     Median         3Q        Max
## -1.592e+00 -5.160e-03 -1.153e-04 -7.332e-06  2.651e+00
##
## Count model coefficients (poisson with log link):
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.7438707  1.2112540  -3.916 8.98e-05 ***
## elev           0.0022669        NaN     NaN      NaN
## tci            0.0881267  0.0630376   1.398   0.1621
## streamdist    -0.0003440  0.0004688  -0.734   0.4630
## disturbLT-SEL  1.1070129  1.1981011   0.924   0.3555
## disturbSETTLE  1.7581960  1.2934878   1.359   0.1741
## disturbVIRGIN  2.0320914  1.1474932   1.771   0.0766 .
## beers          0.0002380  0.1609456   0.001   0.9988
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 35.618310  11.505845   3.096  0.00196 **
## elev        -0.022248   0.007267  -3.062  0.00220 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 43
## Log-likelihood: -65.67 on 10 Df
```

Above we see the coefficient estimates for the Poisson portion of the model first and then below the coefficients for the logistic portion of the model. Elevation had a very important role on the performance of the logistic regression which isn't a big surprise based on the very first graph we made using `pairs`. Elevation likely also was relevant for the Poisson portion based on what we've learned with the other models but it is difficult to be sure because the estimate of standard error, the z-value, and the p-value are missing for this variable. This is because of the warning that was generated with summarizing the model output. I'm not exactly sure what happened but when you look at the variance-covariance matrix the diagonal element for elevation is very small and the sqrt of a very small number is NaN. This must be needed for downstream estimates of standard error and the like. To round out our impression of our ability to model the influence of elevation on *Abies frasier* cover we can plot our predictions against the data. To do so we'll make it a bit easier by only considering a zero-inflated Poisson model that only includes elevation.

```
abies_elev_gau <- glm(cover ~ elev, data=abies, family='gaussian')
abies_elev_poi <- glm(cover ~ elev, data=abies, family='poisson')
abies_elev_zip <- zeroinfl(cover ~ elev | elev, data=abies)

abies_newdata <- data.frame(elev = seq(min(abies$elev), max(abies$elev),
                                       length.out=100))
plot(cover ~ elev, data=abies, xlab='Elevation',
```

```
     ylab='cover', main='Abies fraseri',
     ylim=c(-1, 10))
lines(lowess(abies$elev, abies$cover), col='red', lwd=2)
lines(abies_newdata$elev,
      predict(abies_elev_gau, newdata=abies_newdata,
              type='response'), col='dodgerblue', lwd=2)
lines(abies_newdata$elev,
      predict(abies_elev_poi, newdata=abies_newdata,
              type='response'),  col='purple3', lwd=2)
lines(abies_newdata$elev,
      predict(abies_elev_zip, newdata=abies_newdata),
      col='blue', lwd=2)
legend('topleft', c('lowess', 'Gaussian', 'Poisson', 'zero-inflated Poisson'),
       col=c('red', 'dodgerblue', 'purple3', 'blue', 'green'),
       lwd=2, bty='n')
```



Let's examine if the zero-inflated Poisson improves our acer model as well.

```
acer_zip <- zeroinfl(cover ~ . | elev, data=acer)

AIC(acer_poi)
```

```
## [1] 3651.864
```

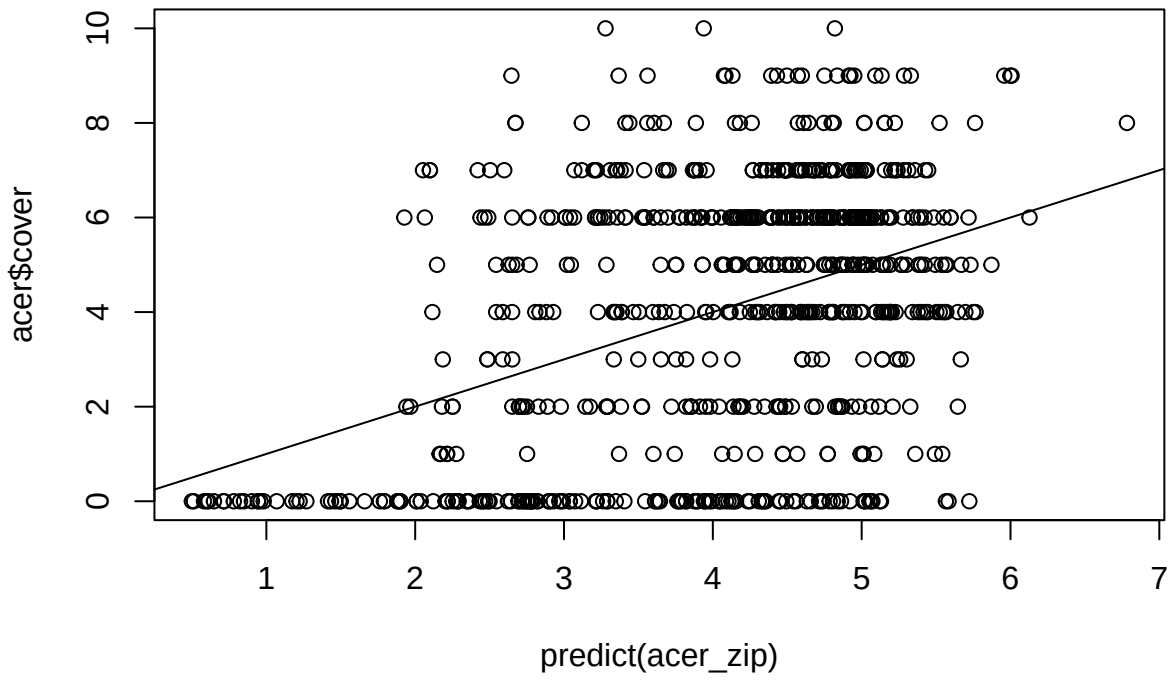
```r
AIC(acer_zip)
```

```
## [1] 3079.988
```

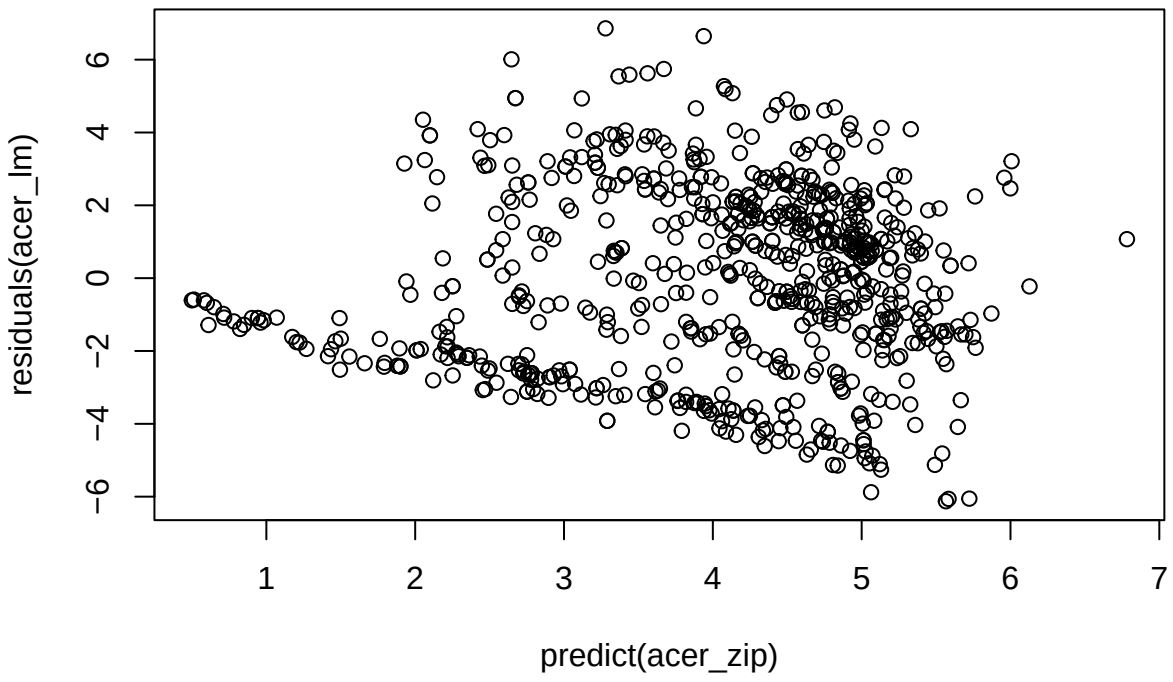
```r
summary(acer_zip)
```

```
##
## Call:
## zeroinfl(formula = cover ~ . | elev, data = acer)
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.11878 -0.71526  0.05087  0.67729  2.37672
##
## Count model coefficients (poisson with log link):
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.860e+00  1.100e-01  16.920   <2e-16 ***
## elev          -1.723e-04  6.836e-05  -2.521   0.0117 *
## tci           -1.205e-02  9.065e-03  -1.329   0.1838
## streamdist     2.623e-04  1.144e-04   2.294   0.0218 *
## disturbLT-SEL -1.728e-02  5.365e-02  -0.322   0.7474
## disturbSETTLE -2.848e-02  7.233e-02  -0.394   0.6937
## disturbVIRGIN  2.748e-03  6.273e-02   0.044   0.9651
## beers         -5.594e-02  2.769e-02  -2.020   0.0434 *
##
## Zero-inflation model coefficients (binomial with logit link):
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.1377535  0.4584546 -11.207   <2e-16 ***
## elev         0.0036193  0.0004052   8.932   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 17
## Log-likelihood: -1530 on 10 Df
```

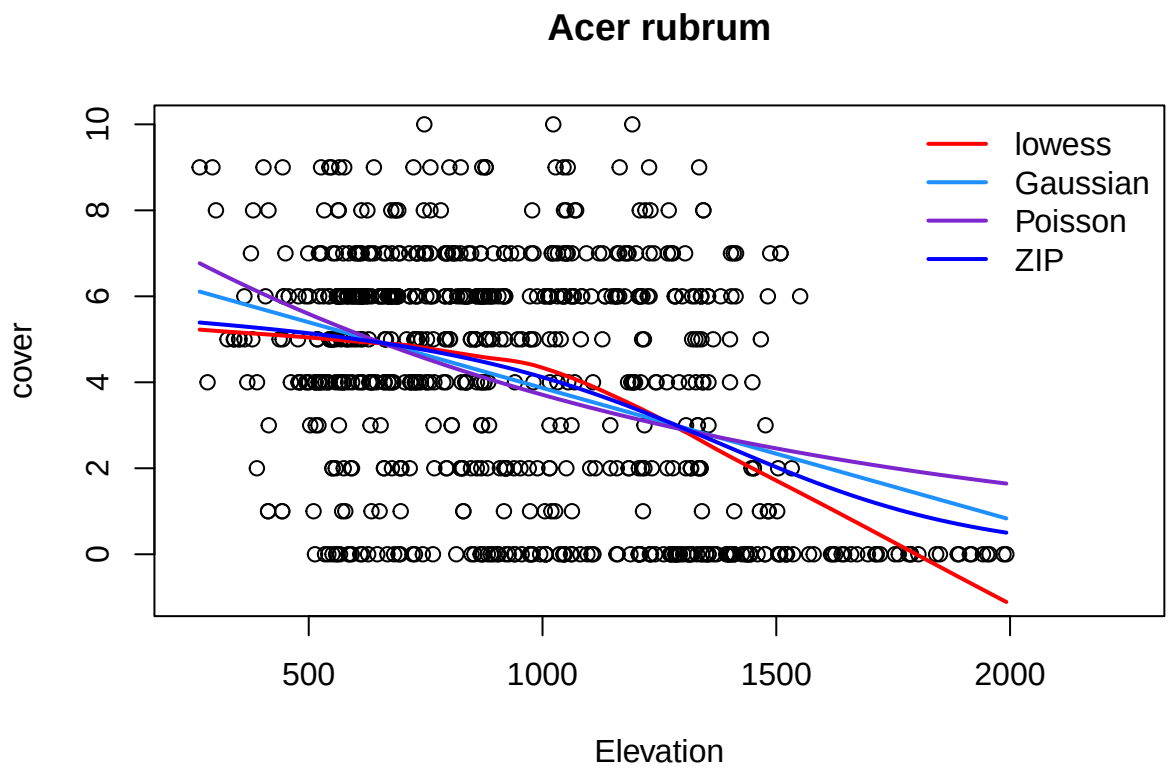
```r
plot(predict(acer_zip), acer$cover)
abline(a=0, b=1)
```




```
plot(predict(acer_zip), residuals(acer_lm))
```




```
acer_elev_gau <- glm(cover ~ elev, data=acer, family='gaussian')
acer_elev_poi <- glm(cover ~ elev, data=acer, family='poisson')
acer_elev_zip <- zeroinfl(cover ~ elev | elev, data=acer)

acer_newdata <- data.frame(elev = seq(min(acer$elev), max(acer$elev),
                                      length.out=100))

plot(cover ~ elev, data=acer, xlab='Elevation',
     ylab='cover', main='Acer rubrum',
     ylim=c(-1, 10), xlim=c(250, 2250))
lines(lowess(acer$elev, acer$cover), col='red', lwd=2)
lines(acer_newdata$elev,
      predict(acer_elev_gau, newdata=acer_newdata,
              type='response'), col='dodgerblue', lwd=2)
lines(acer_newdata$elev,
      predict(acer_elev_poi, newdata=acer_newdata,
              type='response'),  col='purple3', lwd=2)
lines(acer_newdata$elev,
      predict(acer_elev_zip, newdata=acer_newdata),
      col='blue', lwd=2)
legend('topright', c('lowess', 'Gaussian', 'Poisson', 'ZIP'),
       col=c('red', 'dodgerblue', 'purple3', 'blue', 'green'),
       lwd=2, bty='n')
```



The payoff is not as great for adopting a ZIP model for *Acer rubrum* but the AIC is still substantially lower and the diagnostic residual plot is much better behaved.

This last plot for *Acer rubrum* suggests a uni-modal response of this species with elevation which could either be best captured using a weighted averaging approach of including a quadratic elevation term into the model (i.e., `elev^2`).

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?

The take home messages from this analysis are that both species are responding to the environment although *Abies fraseri* which is more of a habitat specialist shows stronger correlations with the available enviornmental variables. Elevation was the most important variable in all the models we examined and ecologically this is not a big surprise either given its combined influence on moisture and temperature. From a modeling perspective we also gained some insight when working with discrete data. If there are a lot of zeros we observed that developing a separate model for the zeros was extremely beneficial for constructing more accurate and reasonable model predictions. Interestingly though the take home messages of which variables were important did not change greatly from our initial inference from the OLS models. This indicates OLS is fairly robust to substantial violations of its assumptions. Most importantly it is key to recognize that lot of the insight we gained from the model analysis was visually pretty obvious from our initial observation of the data patterns. This is a good reminder to always plot the data!

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.

5. (optional) Develop a model for the number of species in each site (i.e., unique plotID). This variable will also be discrete so the Poisson may be a good starting approximation. Side note: the Poisson distribution converges asymptotically on the Gaussian distribution as the mean of the distribution increases. Thus Poisson regression does not differ much from traditional OLS when means are large.