Forest Fires

PSTAT 126 Project Step 3

Sofia Ward, Henry McMahon, Pariya Akhiani, Syed Khaled

Northeast Region of Portugal 2-28-2008

Introduction This dataset is part of the UCI Machine Learning Repository. The data can be used to predict the burned area of a forest fire given information about weather conditions and fuel moisture. We are instead using the data to predict ISI, the initial spread index, for reasons which will be explained in our analysis. The data was taken from Montesinho Natural Park in northeastern Portugal in 2007-2008, so the results apply to a forest with similar foliage.

Variable Relationships We used a *ggpairs()* plot to determine the pairwise relationship between variables. There is a weak correlation between *Area* and the other predictors. There are large outliers in *Area*, and because most of the fires in our dataset have similar *Area*, it is not a good response variable. This is why we will predict *ISI.* Additionally, it is pretty clear that there is a second degree polynomial relationship between *FFMC* and *ISI*

Contextual variable selection is useful in this regard, but criteria based variable seleciton will still be carried out. From our domain knowledge, we know that *DMC*, *FFMC*, and *DC* are combinations of *temperature*, *RH* and *wind.* So, these predictors will need to be left out for meaningful analysis.

Initial Model Selection

Based on our domain knowledge, we select initial models. We had one model use all of the input terms and the other use the FFMC variable instead:

$$\text{Model 1:}\quad \text{ISI} \sim \text{temp} \ + \ \text{RH} \ + \ \text{wind} \ + \ \text{rain} \ + \ \text{seasons} \ + \ \text{day}$$
$$\text{Model 2:}\quad \text{ISI} \sim \ \text{wind} \ + \ \text{seasons} \ + \ \text{FFMC}^2 + \text{FFMC} + \ \text{day}$$

We compare the R^2 values, obtained from the summary of the *lm()* function used for each model:

```
## [1] "Model 1: 0.337021978677974"
```

```
## [1] "Model 2: 0.58704819254095"
```

The R^2 value for Model 2 is significantly higher than Model 1, indicating that Model 2 should be selected over Model 1.

Criterion Based Model Selection

We use stepwise search to find the ‘best’ model in terms of minimizing *AIC* and *BIC*:

```
## [1] "(Intercept), X:wind:seasonsSummer, Y:RH:rain, Y:wind:seasonsFall, temp:RH:wind, temp:wind:seasonsSprin
```

TO DO: Explain what this model means, write it nicely in LATEX.

Forward Selection on *Day* We use forward selection to justify that *day* is not a useful predictor for *ISI.*

TO DO: Figure out what model we were supposed to use for lm

The R^2 values of the model with and without the predictor *day* are:

```
## [1] "Without Day: 0.309358770631067"
```

```
## [1] "With Day: 0.337021978677974"
```

These values are very similar, so we can remove *day* from our model. Now, we know that *seasons* should be an interaction variable for *wind*, and *day* is not a useful predictor.

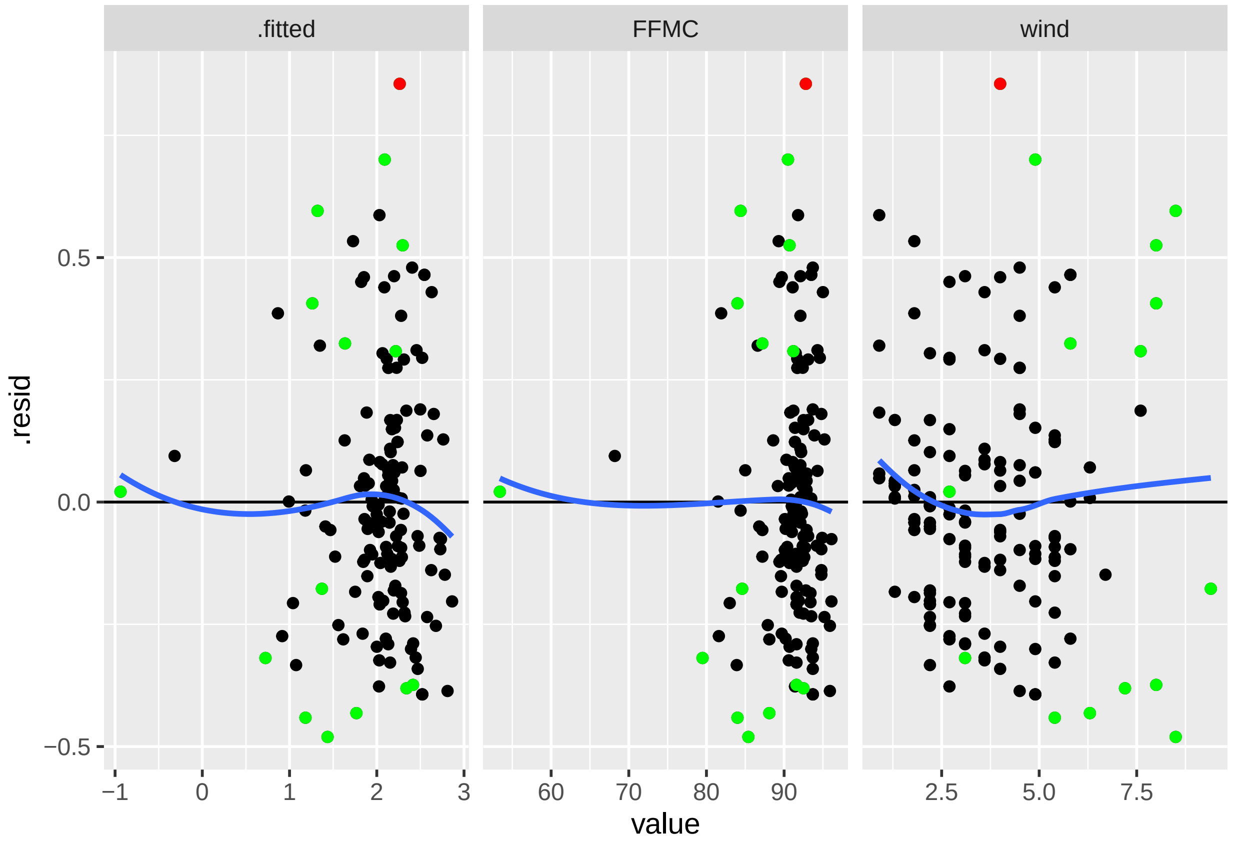
TO DO: fix mlm6. The criteria selection best models have very low R^2.

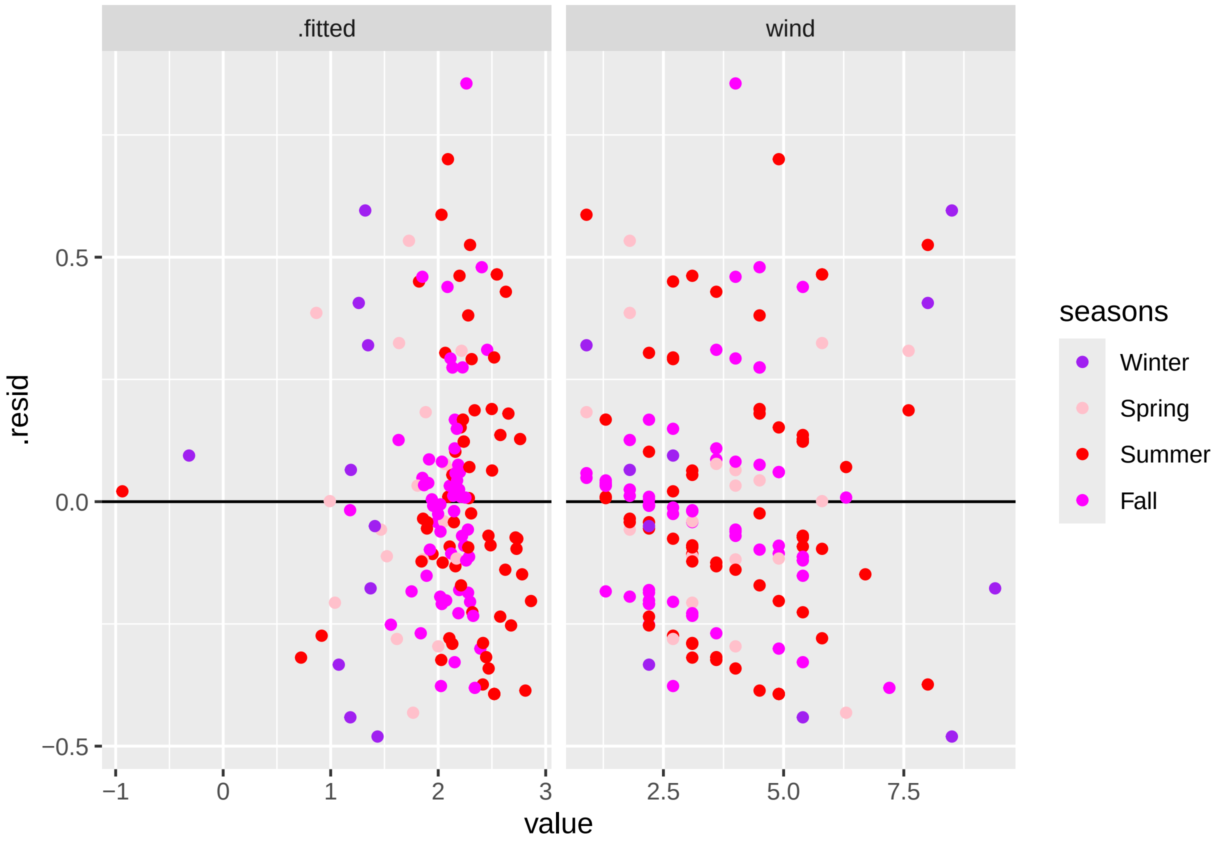
```
##
## Call:
## lm(formula = ISI ~ temp + RH + wind:seasons + rain, data = subset_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2154 -1.9553 -0.5217  1.8223 13.8776
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.67825    2.89902   0.924  0.35708
## temp                0.14399    0.09239   1.559  0.12125
## RH                  0.01413    0.02595   0.545  0.58676
## rain                3.32555    4.25905   0.781  0.43616
## wind:seasonsWinter -0.09451    0.25689  -0.368  0.71346
## wind:seasonsSpring  0.29278    0.28041   1.044  0.29814
## wind:seasonsSummer  1.01910    0.18586   5.483 1.77e-07 ***
## wind:seasonsFall    0.68202    0.21582   3.160  0.00192 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.458 on 147 degrees of freedom
## Multiple R-squared:  0.3081, Adjusted R-squared:  0.2751
## F-statistic:  9.35 on 7 and 147 DF,  p-value: 1.475e-09
```

```
##
## Call:
## lm(formula = log(ISI) ~ seasons:wind + poly(FFMC, 2), data = subset_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4803 -0.1821 -0.0194  0.1229  0.8554
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.86988    0.05236  35.715  < 2e-16 ***
## poly(FFMC, 2)1      5.90296    0.29344  20.116  < 2e-16 ***
## poly(FFMC, 2)2      1.25446    0.29566   4.243 3.88e-05 ***
## seasonsWinter:wind  0.03014    0.01711   1.762   0.0801 .
## seasonsSpring:wind  0.04190    0.01849   2.266   0.0249 *
## seasonsSummer:wind  0.05791    0.01409   4.109 6.56e-05 ***
## seasonsFall:wind    0.03723    0.01645   2.264   0.0250 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2595 on 148 degrees of freedom
## Multiple R-squared:  0.8005, Adjusted R-squared:  0.7924
## F-statistic: 98.99 on 6 and 148 DF,  p-value: < 2.2e-16
```

mlm6 is clearly a better model with a much higher R-squared, all variables are statistically significant and the model as a whole is significant.

Residual Checking & Diagnostics


.fitted
FFMC
wind
.resid
0.5
0.0
−0.5
−1
0
1
2
3
60
70
80
90
2.5
5.0
7.5
value


.fitted
wind
.resid
0.5
0.0
−0.5
−1
0
1
2
3
2.5
5.0
7.5
value
seasons
Winter
Spring
Summer
Fall

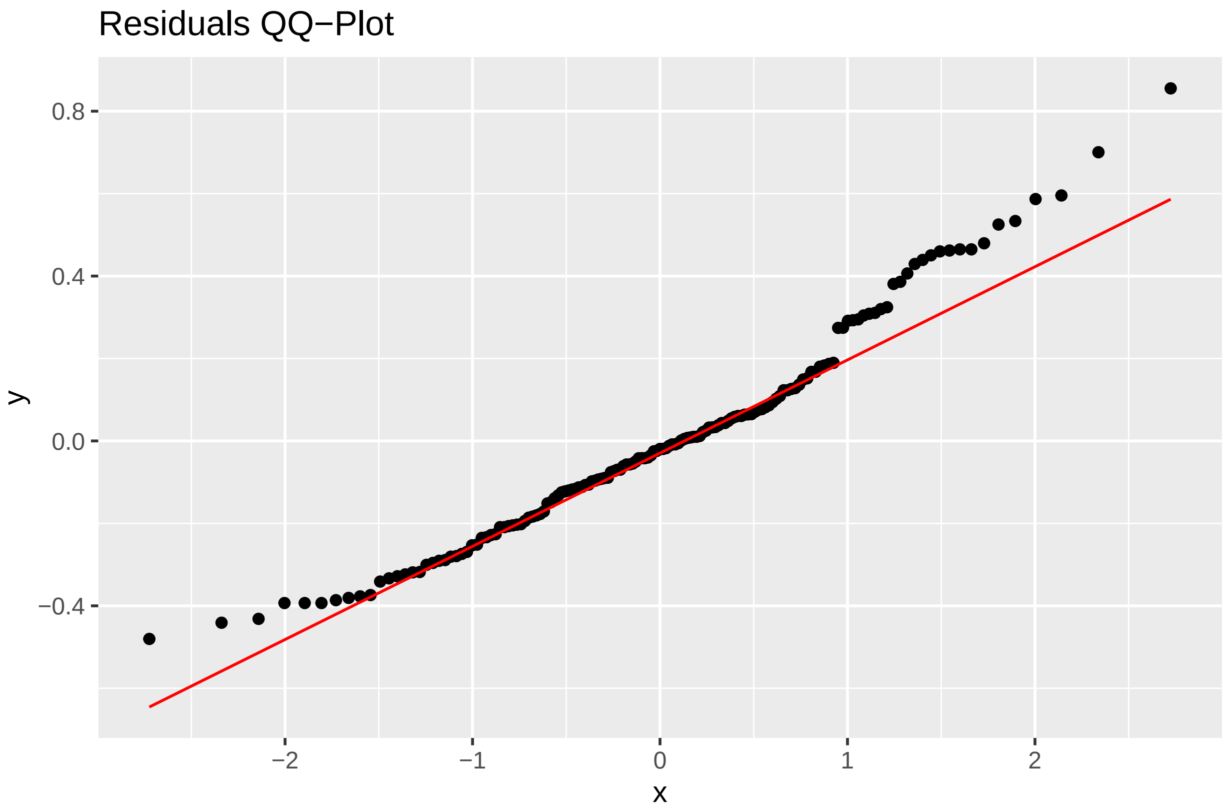
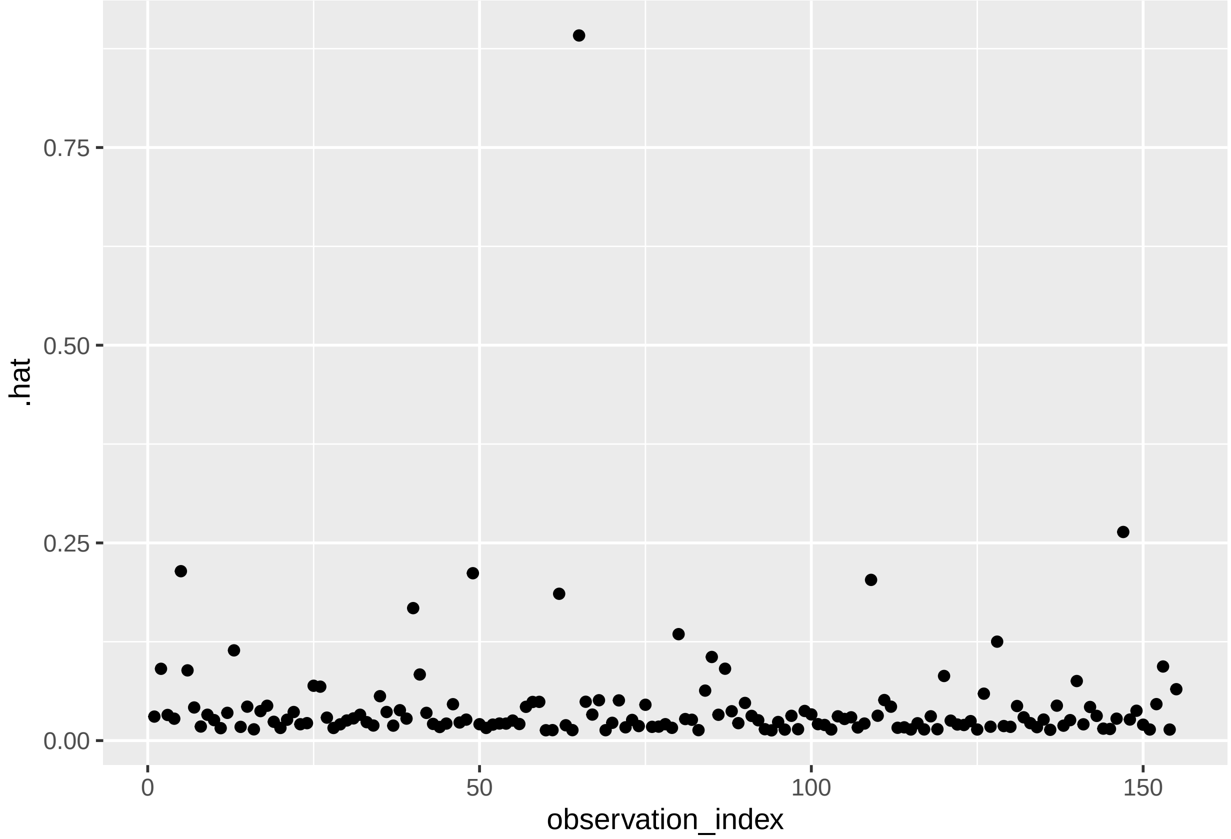
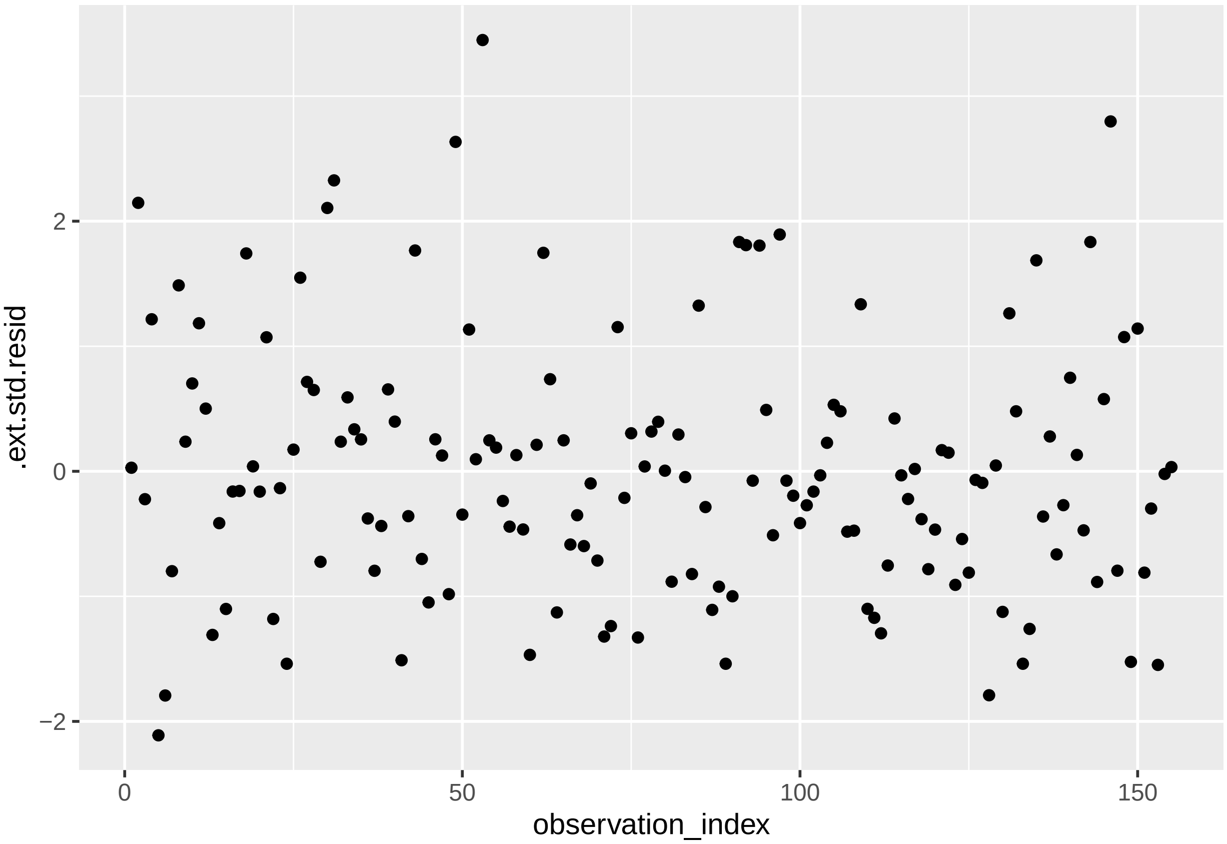


Figure 3: Residuals QQ−Plot

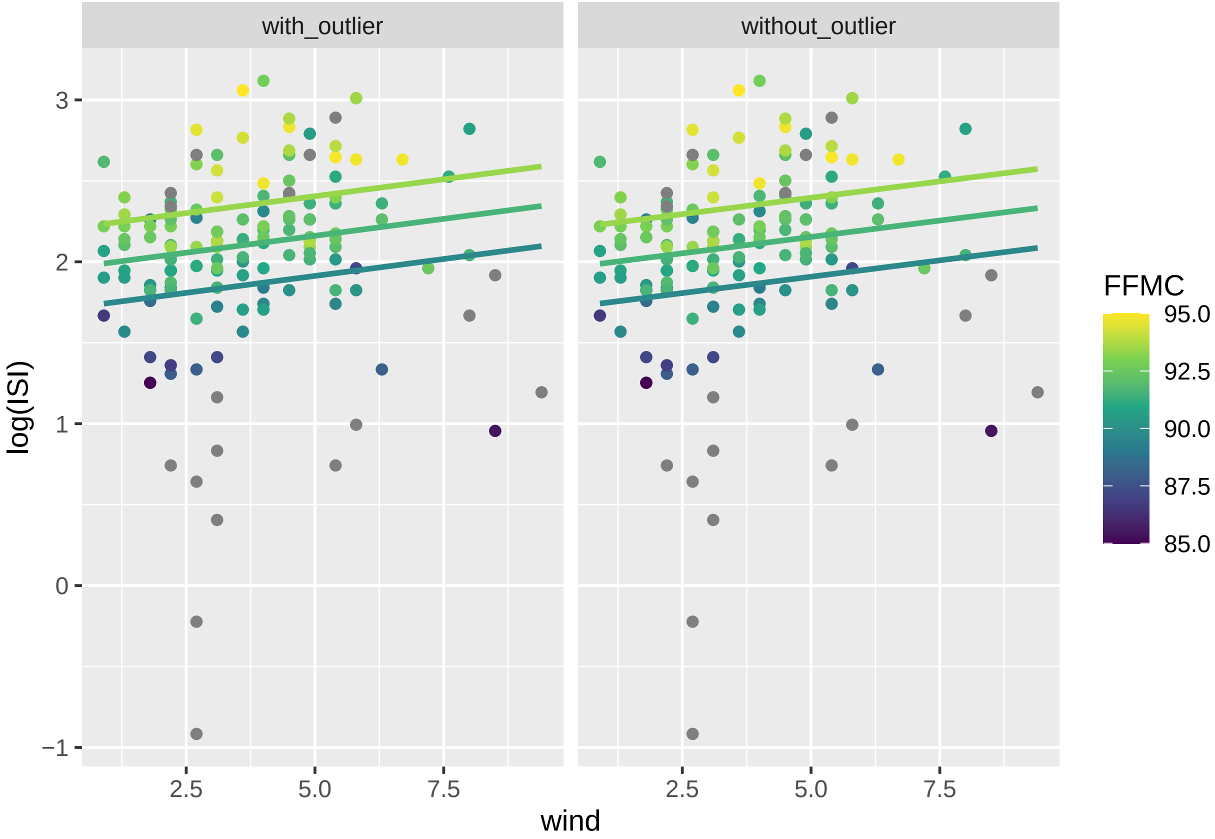


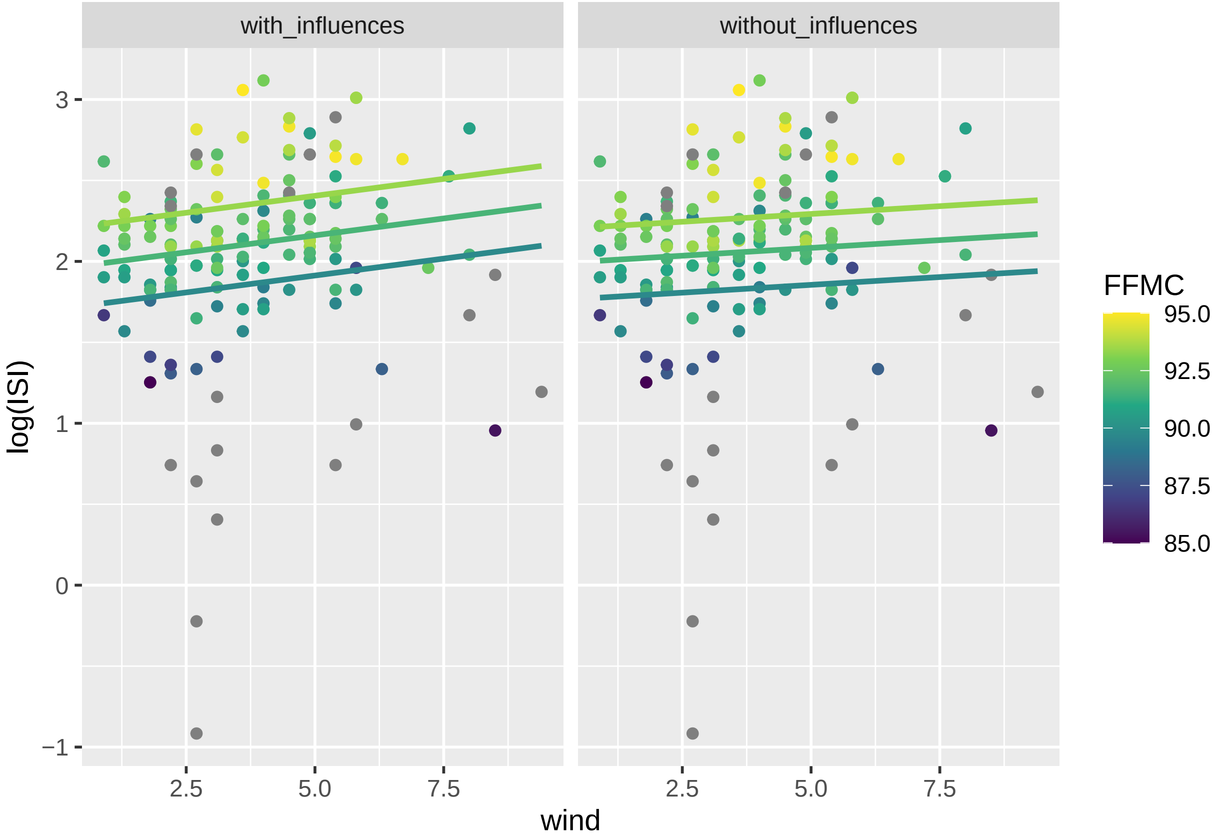


TO DO: Show plots

All of the residuals seem to be fine, but again you are able to see in the data that most of the fires fell within an extremely tight range. A square root transformation had to performed on the response variable, originally the fitted values were displaying non-constant variance through a somewhat pronounced fan pattern. However, there is a pretty clear outlier. Let's explore the fit without that data point.

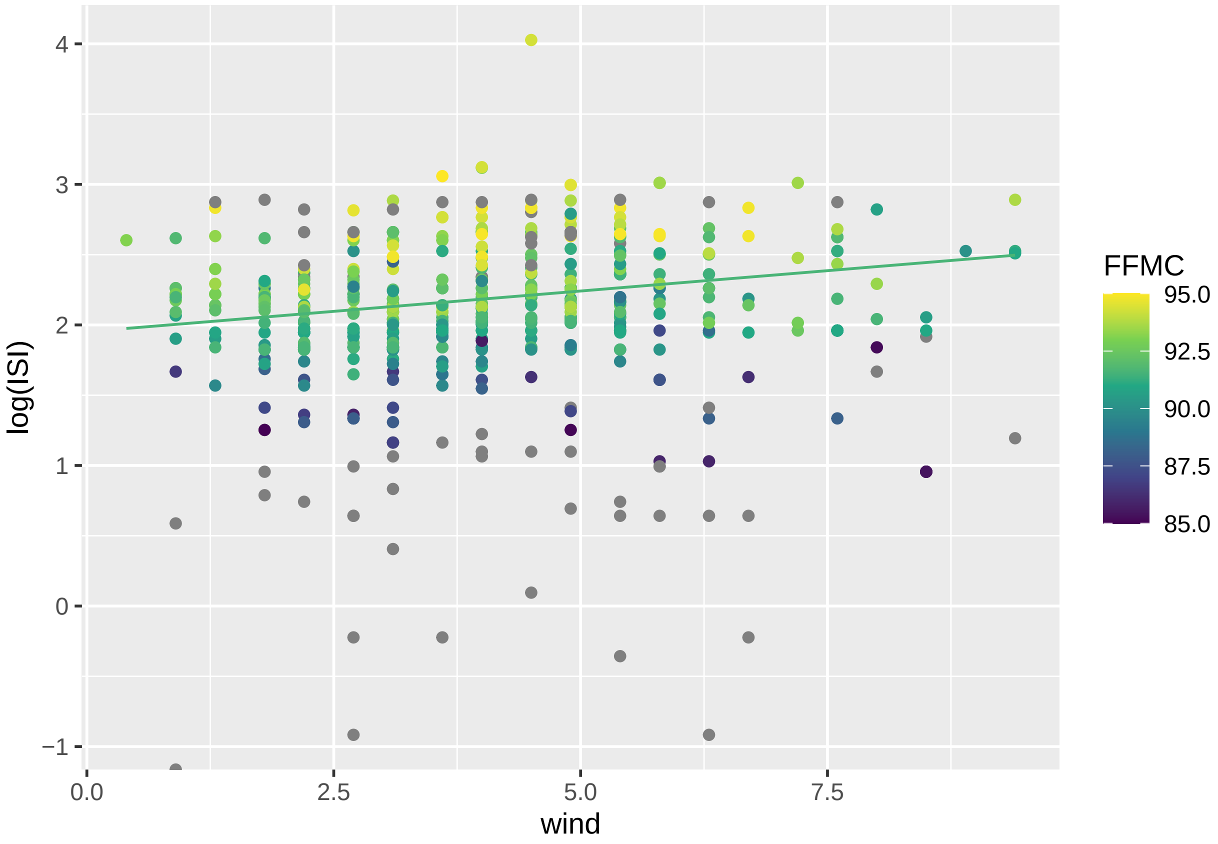
Now lets look at influence & outlier points.

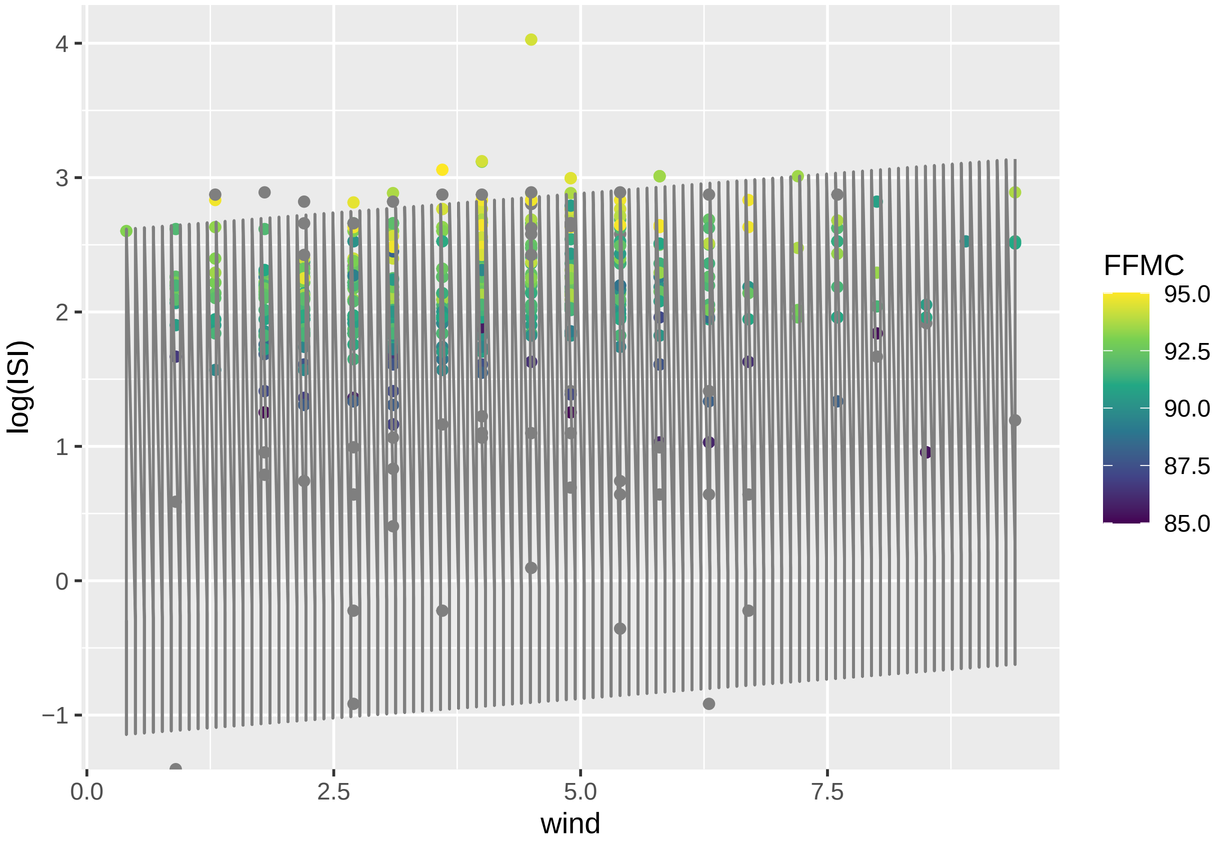




Visualizing the Fit

to discuss actually not sure if this is necessary. not in the directions anywhere...





gives estimated relationship when FFMC is at its median value # needs CI & PI added i think…

TO DO: Show plots with PI and CI

Summary:

TO DO: Add summary