151A Project - Dae Woong Ham

Introduction

In the first and main part of my project I will gear my whole analysis and decisions to the following purpose:

To propose not only a good predicting model but also a good explaining model for ONLY wildfires that seems to be significantly bigger than 0 (thus filtering out all the response variables >0)

This goal makes sense in the perspective of fireman. They don't necessarily care about fires that are weaker and over preparing is much better than underpreparing, thus this conservative estimate realistically and statistically makes sense. Moreover, statistically this will just make the analysis much easier because I can ignore all those clustered up 0 points.

I will organize this analysis in the following way. First I will do some very simple EDA on the response variable itself and see how it's distributed. Then I will really try to look at the possible explanatory variables (the other 12 variables) that could be helpful. After I explore my explanatory variable I will finally start proposing some models based on methodologies learned in class. I will then diagnose my model and see how it is performing. Lastly, there will be an evaluation of this model after all procedure and fixing is done to see how it generally does in a linear regression setting.

In the last part of my report, I will focus on my own question to create another model for the firemen. The above analysis seems great to the firemans to analyze only fires they should care about. However, it still begs the question of what about the fires they shouldn't care about? Ignoring all the 0 points takes away a lot of data. Wouldn't they want to predict that also? Therefore, that motivated this question of "trying to predict whether or not there will be a fire they care about". Therefore for this question I will shift my attention to changing the area into a dummy variable and predicting that with generalized linear model, with the main question being what covariates influences this dummy variable? The analysis and organization will almost mirror the above analysis.

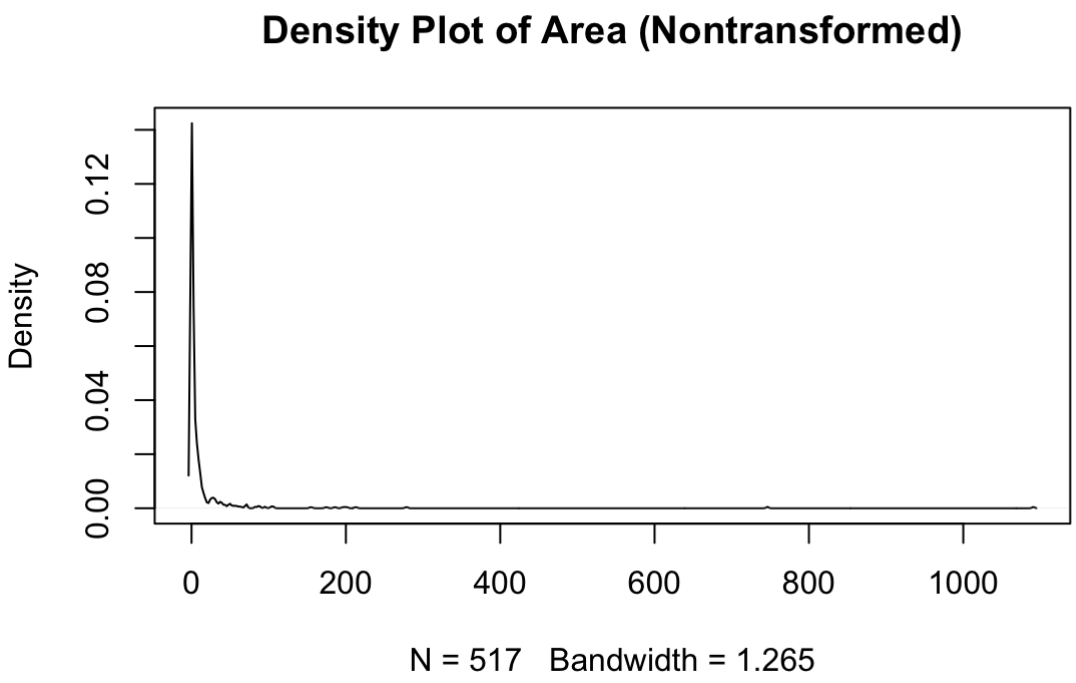
Description of Data

The very first EDA I will do is on the response variable itself. Let's look at the following results and plots.

```
## summary statistics of area and five biggest fires
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.52   12.85    6.57 1090.84
```

```
## [1] 1090.84  746.28  278.53  212.88  200.94
```



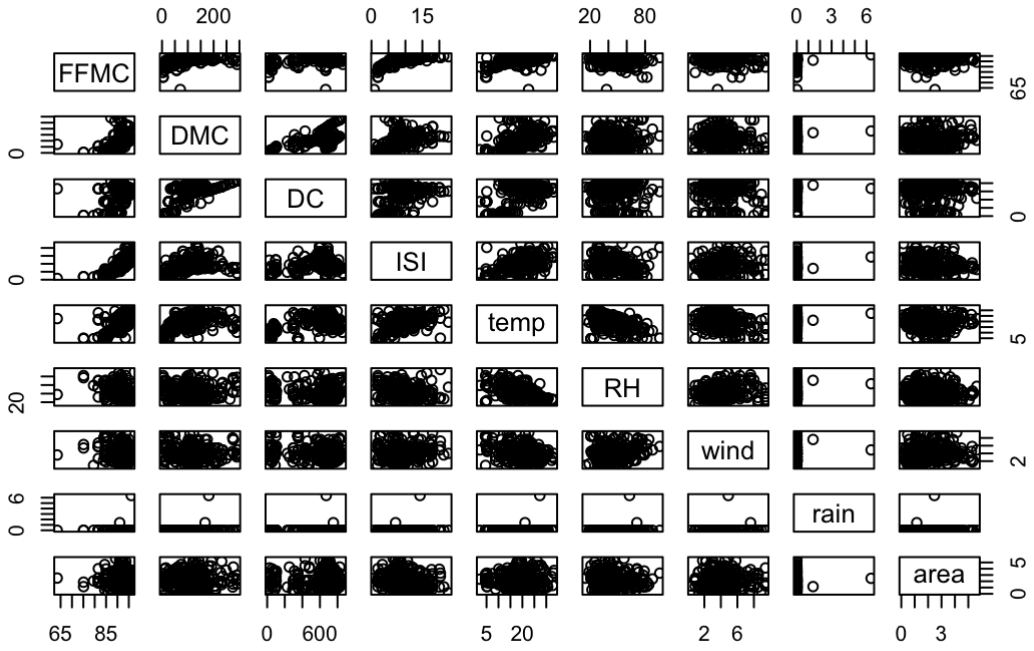
It seems that some of these wildfires were absolutely massive such as the one with the 1090 and 746 data points. For the purpose of this analysis I will not consider these necessary to predict simply due to it being unlikely. Moreover, I will do a shifted (1 + x) log transformation on this variable due to the very high skew.

Now as explained above because the response variable as shown in the plot seems to have a lot of values at 0 and this is not so important to the goal that I want to achieve (it will only obscure and make the linear model worse), I will perform the remainder of my analysis only on the response variables that matter, which is area > 0.

I will now do some EDA on the continuous variables on just the subset data with area values greater than 1 to see what sort of model I should expect for this part.

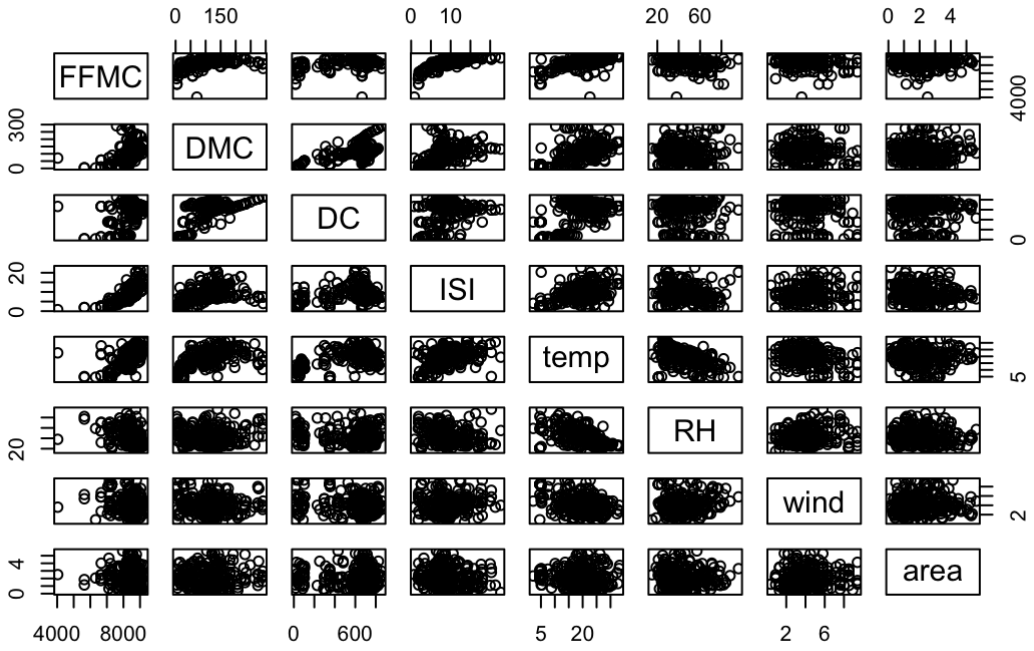
```
## correlation matrix of all continous variables
```

```
##              FFMC         DMC          DC         ISI        temp
## FFMC  1.00000000  0.47726134  0.40615160  0.70334427  0.55983872
## DMC   0.47726134  1.00000000  0.66958680  0.32495335  0.49792720
## DC    0.40615160  0.66958680  1.00000000  0.25556559  0.49428150
## ISI   0.70334427  0.32495335  0.25556559  1.00000000  0.46430017
## temp  0.55983872  0.49792720  0.49428150  0.46430017  1.00000000
## RH   -0.28205610  0.03590472 -0.07847560 -0.14567877 -0.49285044
## wind -0.16329673 -0.14144297 -0.23866800  0.07104590 -0.32413228
## rain  0.08201440  0.07614383  0.03668559  0.06766453  0.08200962
## area -0.05188504  0.01941425 -0.03229985 -0.10706557 -0.04250525
##                RH        wind        rain         area
## FFMC -0.28205610 -0.16329673 0.082014403 -0.051885044
## DMC   0.03590472 -0.14144297 0.076143835  0.019414253
## DC   -0.07847560 -0.23866800 0.036685587 -0.032299854
## ISI  -0.14567877  0.07104590 0.067664527 -0.107065570
## temp -0.49285044 -0.32413228 0.082009622 -0.042505254
## RH    1.00000000  0.14067307 0.099780520 -0.032330947
## wind  0.14067307  1.00000000 0.049226380  0.047750804
## rain  0.09978052  0.04922638 1.000000000  0.008627104
## area -0.03233095  0.04775080 0.008627104  1.000000000
```



Despite sub-setting the data for only area > 0, transforming, and removing big wildfires, this plot still looks very unpromising. Most of the covariates are not even close to a strong linear relationship and this is reflected in the correlation matrix in which the highest correlated variable is wind and temperature only at values of 0.047. I should expect the linear model to prove very weak.

Another thing that is apparent from the above plot is that the variables rain and FFMC seem extremely hard to visualize because rain has so little values greater than 0 and FFMC seems to have an outlier making it hard to visualize. I will remove FFMC's outlier and change rain into a categorical variable to basically 1 and 0 and not include it in my covariate EDA. Finally, because FFMC seems extremely negative skewed I will also transform this data with a square function. The following revised plot shows all these changes to the covariates for better visualization.



Despite doing all these small touches it still does not solve the non-linear problem. However, visually speaking it is definitely an improvement, and I can discern that area might have some potential to be explained by wind, RH, and temperature. Lastly, note that there is collinearity in some of continuous covariates also. Because of this EDA, I will try to do some PCA if I have some models with many covariates.

Analysis

Finding Model Candidates

Because I did not do any EDA on the categorical variables, now is the appropriate time to see if these categorical variables do make an impact on the response variable area. There are many ways to tackle this, but one of the most direct and logical ways is to just compare a full model with all the variables and a sub model with just the continuous variables. We will then see the summary statistics.

```
## Sumamry of Full Model with Everything
```

```
##
## Call:
## lm(formula = area ~ ., data = new_data_without_zero)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2026 -0.8707 -0.0744  0.5468  3.7265
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.068e+00  1.836e+00   1.126  0.26117
## X            4.165e-02  3.721e-02   1.119  0.26418
## Y           -9.951e-02  7.540e-02  -1.320  0.18815
## monthaug     7.844e-01  1.077e+00   0.728  0.46711
## monthdec     1.498e+00  9.152e-01   1.637  0.10294
## monthfeb    -3.824e-01  7.171e-01  -0.533  0.59441
## monthjul     1.814e-01  9.200e-01   0.197  0.84384
## monthjun    -2.736e-01  8.642e-01  -0.317  0.75182
## monthmar    -3.802e-01  6.703e-01  -0.567  0.57103
## monthmay     1.107e+00  1.365e+00   0.811  0.41804
## monthoct     2.737e+00  1.314e+00   2.082  0.03838 *
## monthsep     1.793e+00  1.213e+00   1.479  0.14046
## daymon      -1.035e-02  2.832e-01  -0.037  0.97089
## daysat       4.111e-01  2.703e-01   1.521  0.12961
## daysun       3.677e-01  2.662e-01   1.381  0.16854
## daythu       4.549e-04  2.975e-01   0.002  0.99878
## daytue       3.354e-01  2.749e-01   1.220  0.22367
## daywed       8.218e-02  2.913e-01   0.282  0.77812
## FFMC         4.106e-05  2.072e-04   0.198  0.84311
## DMC          7.197e-03  2.247e-03   3.203  0.00154 **
## DC          -4.325e-03  1.642e-03  -2.635  0.00897 **
## ISI         -2.652e-02  3.019e-02  -0.878  0.38058
## temp         2.603e-02  2.752e-02   0.946  0.34516
## RH          -1.811e-04  7.901e-03  -0.023  0.98174
## wind         3.228e-02  4.629e-02   0.697  0.48634
## rain1       -7.905e-02  8.950e-01  -0.088  0.92969
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.177 on 242 degrees of freedom
## Multiple R-squared:  0.121,  Adjusted R-squared:  0.03017
## F-statistic: 1.332 on 25 and 242 DF,  p-value: 0.1398
```

```
## Summary of Model with Only Continuous Variables
```

```
##
## Call:
## lm(formula = area ~ ., data = subset_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9511 -0.9179 -0.1741  0.6337  3.5092
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.207e+00  1.567e+00   1.409    0.160
## FFMC         5.877e-05  1.954e-04   0.301    0.764
## DMC          2.530e-03  1.812e-03   1.396    0.164
## DC          -3.242e-04  4.452e-04  -0.728    0.467
## ISI         -4.376e-02  2.763e-02  -1.584    0.114
## temp        -8.919e-03  1.936e-02  -0.461    0.645
## RH          -6.878e-03  6.195e-03  -1.110    0.268
## wind         4.082e-02  4.362e-02   0.936    0.350
##
## Residual standard error: 1.194 on 260 degrees of freedom
## Multiple R-squared:  0.02702,    Adjusted R-squared:  0.000821
## F-statistic: 1.031 on 7 and 260 DF,  p-value: 0.4095
```

From this information, it becomes blatantly obvious that the only continuous variable model lacks precision by many folds. It is also important to note that the full model despite all these transformations only could achieve an R^2 value of 0.121. This is not the best news for a linear modelling class because theory tells us that R^2 never decreases with more variables. In other words, this is the maximum R^2 we can get because my new proposed models will only drop variables not add them. This is a huge indication that perhaps the linear model is not a good fit.

Now it is also apparent that the full model seems to lack many statistical significances for many parameters as they all have high standard errors. In order to improve accuracy for not only prediction but for description purposes, I will now propose better candidate models.

My main way of choosing candidate models will be using the step() function. The step function is quite ideal because I can propose many candidate models by approaching the problem from different ways. First I can look at it in terms of AIC which is more prediction power based and not penalizing the amount of variables. I can also look at BIC which is less predictive power but more explanation based because it penalizes it for more variables. Lastly I could choose three potentially different candidate models by stepping forward, backwards or both ways for each information criteria, thus a total of potentially six candidate models. These will be my candidate models

So here are the following results the six different step method gave me:

1. Full model (forward AIC/BIC)
2. month + DMC + DC from (backwards and both AIC)
3. intercept, i.e. no variables (backwards and both BIC)

These results are very interesting because potentially there could be six different models but both the forward AIC and BIC methods gave me the only three models, one of which was the intercept model. Although this is consistent with the belief that the linear model might not be so powerful, it is still not so useful for my setting to give a powerful predictive model to the fireman who is using other data like temperature to predict. Therefore there is really only two candidate models. I will also introduce the following two models for possible consideration due to high correlation to area and to continue to test if the categorical variables are useful or not:

1. Full model
2. Continuous variables only model

3. month + DMC + DC
4. month + DMC + DC + temp + wind

Now I will begin my model selection process by using two measures: AIC/BIC and cross validation.

Model Comparison - Part 1

AIC/BIC comparisons

```
##                            df      AIC      BIC
## Month + DMC + DC           13 858.4584 905.1412
## Full Model                 27 874.3597 971.3163
## Continuous Subset Model     9 865.5764 897.8953
## Added temp and wind model  15 860.5188 914.3836
```

Here it seems clear that our two winners are the models with month DMC and DC and the model with all those three but with added temperature and wind. But I will also run all four of these models through cross validation.

The methodology I will be using is a four-fold cross validation. My cleaned up data with only >0 values of response variables has 268 rows so I can train on 201 data points and test on 67 data points every time. I will report the RMSE in a table and compare.

Cross Validation

```
##                             CV Error
## Full Model                  83.75193
## Just Continuous Variables 122.74696
## MonthDMCDC                111.53873
## Added temp wind           115.69681
```

Note when doing the cross validation because often times the month variable would not contain all the months in the training data to actually build the model properly I had to omit it from all models hoping it would not make a significant impact. However, as expected it seems like the CV error is significantly reduced when having the full model. However, in terms of explanation it might not be the best as indicated in our information criteria above because the full model is always problematic due to the multicollinearity problem. So I will see if PCA can actually help.

Analysis through PCA: Will it make the model better?

The motivation for doing PCA is because it seems from the cross validation result the full model does the best but there is going to be some bad multicollinearity problem. Theoretically speaking doing PCA to the categorical variables is not so clear and moreover doing PCA for the categorical variables seem unnecessary because I suspect close to no multicollinearity in those variables. Thus I will only do it to the continuous variable.

```
## Importance of components:
##                             PC1      PC2      PC3      PC4     PC5     PC6
## Standard deviation     649.0585 209.4644 43.18315 14.20583 4.05692 2.85322
## Proportion of Variance   0.9016   0.0939  0.00399  0.00043 0.00004 0.00002
## Cumulative Proportion    0.9016   0.9955  0.99951  0.99994 0.99997 0.99999
##                            PC7   PC8
## Standard deviation     1.60999 1.175
## Proportion of Variance 0.00001 0.000
## Cumulative Proportion  1.00000 1.000
```

It seems very obvious that by the fourth PCA, the remainder is not so important therefore I will take the first four PCA which essentially contains all the information about my continuous variable and concatenate it with the remaining categorical variables to redo a linear model with the PCAs and categorical variables and compare

summary statistics. Here is the following result

```
## PCA Model
```

```
##
## Call:
## lm(formula = area ~ ., data = improved_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1727 -0.8543 -0.1008  0.5573  3.8636
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.0791051  1.0165924   1.061  0.28951
## X            0.0390121  0.0368988   1.057  0.29143
## Y           -0.0915252  0.0741045  -1.235  0.21798
## monthaug     0.8187293  1.0491687   0.780  0.43593
## monthdec     1.3512368  0.8752391   1.544  0.12392
## monthfeb    -0.4005978  0.7102584  -0.564  0.57326
## monthjul     0.3140929  0.8796146   0.357  0.72134
## monthjun    -0.1124617  0.8226640  -0.137  0.89138
## monthmar    -0.3687855  0.6659491  -0.554  0.58024
## monthmay     1.2946291  1.3391397   0.967  0.33462
## monthoct     2.7208477  1.3063891   2.083  0.03831 *
## monthsep     1.7987089  1.1985062   1.501  0.13470
## rain1        0.2042859  0.8606548   0.237  0.81258
## daymon       0.0174324  0.2777874   0.063  0.95001
## daysat       0.4366958  0.2646469   1.650  0.10020
## daysun       0.4434835  0.2555810   1.735  0.08396 .
## daythu       0.0023837  0.2924907   0.008  0.99350
## daytue       0.3374219  0.2722068   1.240  0.21632
## daywed       0.0821895  0.2878857   0.285  0.77551
## PC1         -0.0003341  0.0002684  -1.245  0.21448
## PC2         -0.0029125  0.0014011  -2.079  0.03869 *
## PC3         -0.0077120  0.0023236  -3.319  0.00104 **
## PC4         -0.0060828  0.0053817  -1.130  0.25946
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.173 on 245 degrees of freedom
## Multiple R-squared:  0.1155, Adjusted R-squared:  0.03611
## F-statistic: 1.455 on 22 and 245 DF,  p-value: 0.09045
```

```
## Original Full Model
```

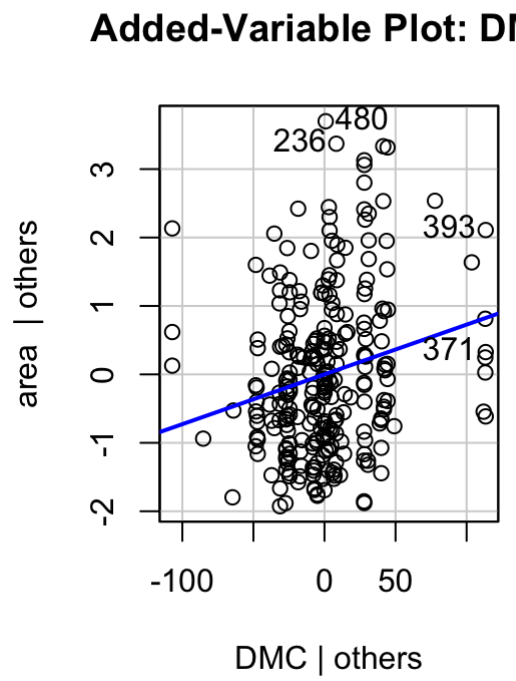
```
##
## Call:
## lm(formula = area ~ X + Y + month + day + FFMC + DMC + DC + ISI +
##     temp + RH + wind + rain, data = new_data_without_zero)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2026 -0.8707 -0.0744  0.5468  3.7265
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.068e+00  1.836e+00   1.126  0.26117
## X            4.165e-02  3.721e-02   1.119  0.26418
## Y           -9.951e-02  7.540e-02  -1.320  0.18815
## monthaug     7.844e-01  1.077e+00   0.728  0.46711
## monthdec     1.498e+00  9.152e-01   1.637  0.10294
## monthfeb    -3.824e-01  7.171e-01  -0.533  0.59441
## monthjul     1.814e-01  9.200e-01   0.197  0.84384
## monthjun    -2.736e-01  8.642e-01  -0.317  0.75182
## monthmar    -3.802e-01  6.703e-01  -0.567  0.57103
## monthmay     1.107e+00  1.365e+00   0.811  0.41804
## monthoct     2.737e+00  1.314e+00   2.082  0.03838 *
## monthsep     1.793e+00  1.213e+00   1.479  0.14046
## daymon      -1.035e-02  2.832e-01  -0.037  0.97089
## daysat       4.111e-01  2.703e-01   1.521  0.12961
## daysun       3.677e-01  2.662e-01   1.381  0.16854
## daythu       4.549e-04  2.975e-01   0.002  0.99878
## daytue       3.354e-01  2.749e-01   1.220  0.22367
## daywed       8.218e-02  2.913e-01   0.282  0.77812
## FFMC         4.106e-05  2.072e-04   0.198  0.84311
## DMC          7.197e-03  2.247e-03   3.203  0.00154 **
## DC          -4.325e-03  1.642e-03  -2.635  0.00897 **
## ISI         -2.652e-02  3.019e-02  -0.878  0.38058
## temp         2.603e-02  2.752e-02   0.946  0.34516
## RH          -1.811e-04  7.901e-03  -0.023  0.98174
## wind         3.228e-02  4.629e-02   0.697  0.48634
## rain1       -7.905e-02  8.950e-01  -0.088  0.92969
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.177 on 242 degrees of freedom
## Multiple R-squared:  0.121,  Adjusted R-squared:  0.03017
## F-statistic: 1.332 on 25 and 242 DF,  p-value: 0.1398
```

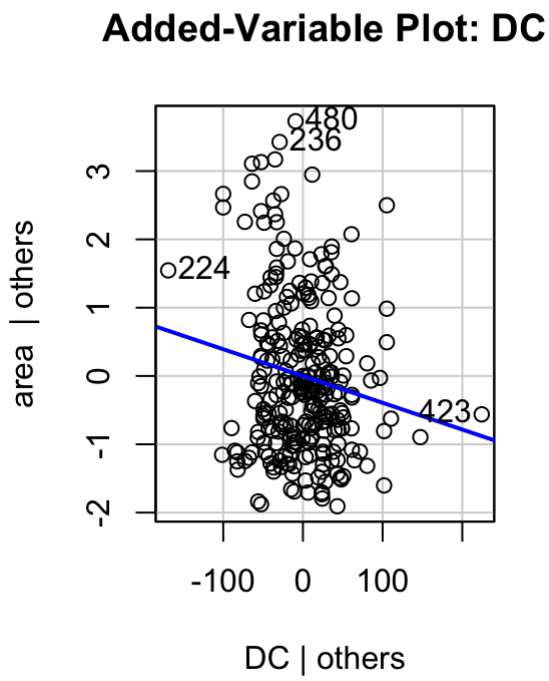
The PCA definitely helped increased the adjusted R^2 value without much loss of original R^2 value and most importantly made the omnibus test p-value 0.09 compared to 0.1398, giving much more robust coefficient estimates. This is expected due to the power of PCA. Therefore in conclusion I will propose two models to the fireman:

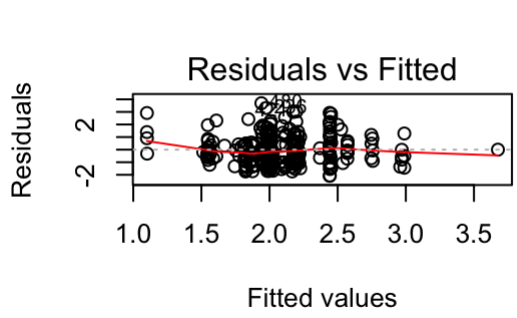
1. For Prediction: The Full Model with the PCAs (of course if the firemen want to predict they would have to change their explanatory variable to PCAs)
2. For Explanation: The Model with just month + DMC + DC

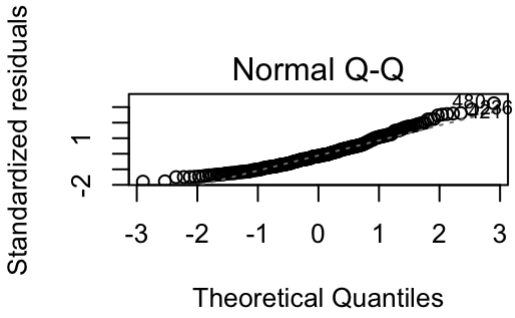
Last diagnostics with my explanation model

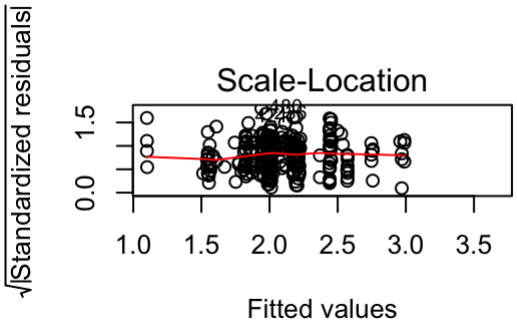
As explained above there is high evidence to believe linear regression is not a good idea. I will explore added-variable plots on my explanation model - specifically only the two variables, and general diagnostics plots.

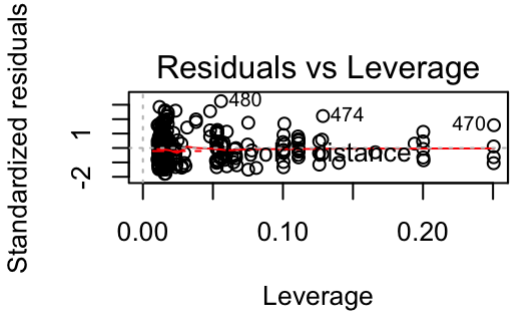












The first comment is that the added variable plots show quite unconvincing linear relationships, and these were the two best variables picked to model linearly area from AIC/BIC criterias! It is indicative and only expected from all the analysis above that the linear model would not be so great.

Surprisingly the normal error assumption and constant variance plots do not look as bad as expected. However, it is important to note that I am only diagnosing the explanation model, which has much less parameters. However, it is still worrying that the cook's distance really does not look very promising, as there seems to be a decent

amount of points that are influential and the normal QQ plot is itself not convincing enough. This is exactly what I expected and it is reflective that perhaps a linear model might not be best despite all these transformations. But this is the best I can do.

End of first part of Analysis for continuous variable response area

I will now begin to analysis my own question which is taking one step further to give the firemen another source of model they might be interested in - whether or not they should anticipate a big fire or not. This analysis requires making the area a dummy variable, in my case I made a decision to make anything bigger than 1 hectares count as something the fireman would care about.

The main purpose of this part of the analysis will be to show what variables will best explain this dummy variable. I will not go as into detail in my analysis as above. First, I will immediately go to modelling with a similar way I used above - the step() function to give me six potential candidate models.

Modelling dummy area response

It was fascinating to see that as similar to above the six different step methods only gave me again three candidate models. I will summarize those three models in the following along with where they originated from.

1. Full Model (forward AIC and BIC)
2. X, month, FFMC, rain (backwards and both AIC)
3. Intercept (backwards and both BIC)

For the same reason as above I do not want to consider the intercept model for the fireman to use to predict whether there will be a substantial fire or not. I will again add temperature and wind as another candidate model and only the continuous model again so here will be the four following candidate model. Moreover, note how it's very interested in the dummy variable setting of just predicting probabilities of being strong or negligible fires the best covariates changed from month + DMC + DC (in the above analysis) to the addition of X and FFMC and rain instead of DMC and DC. This, however, is not a huge surprise because I have added all the 0 points now and perhaps different covariates are stronger to predict this.

1. Full Model
2. X, month, FFMC, rain
3. X, month, FFMC, rain, temp, wind
4. Only Continuous variables

For model selection because we are looking at response variables as only 1 or 0 instead of cross validation I will just do AIC/BIC criterias due to the complications that arise when calculating root mean square errors needed to do cross validation. (Note for CV we have to figure out what exactly the errors/residuals are, it can't simply be yi - probabilities) Here is the following result

Model Selection using AIC/BIC

```
##                            df      AIC      BIC
## Y + Month + FFMC + Rain   15 711.6655 775.3280
## Full Model                28 731.7515 850.5882
## Continuous Subset Model    8 721.7215 755.6748
## Added temp and wind model 17 713.0326 785.1834
```

In this case the X + Month + FFMC + Rain had the lowest AIC but the model with only the continuous variables had a lower BIC. I, however, am more interested in giving a strong predicting power model to the firemans to indicate to them whether or not there wil be a big power so I will use AIC since they do not really care about

parsimonious models. Thus I will choose the X + Month + FFMC + Rain as my final model. Let's take a look at how this model is doing

Summary of our candidate model

```
##
## Call:
## glm(formula = area ~ X + month + FFMC + rain, family = binomial,
##     data = dummy_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3564  -1.1396  -0.8313   1.1728   1.7608
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.287e+00  1.374e+00  -1.664   0.0962 .
## X            7.198e-02  3.988e-02   1.805   0.0711 .
## monthaug    -1.069e-01  7.148e-01  -0.150   0.8811
## monthdec     1.591e+01  4.842e+02   0.033   0.9738
## monthfeb     3.758e-01  8.133e-01   0.462   0.6441
## monthjan    -1.493e+01  1.455e+03  -0.010   0.9918
## monthjul     7.972e-02  7.771e-01   0.103   0.9183
## monthjun    -2.208e-01  8.530e-01  -0.259   0.7957
## monthmar    -5.229e-01  7.408e-01  -0.706   0.4803
## monthmay     2.219e-01  1.567e+00   0.142   0.8874
## monthnov    -1.512e+01  1.455e+03  -0.010   0.9917
## monthoct    -6.592e-01  8.773e-01  -0.751   0.4524
## monthsep     2.006e-01  7.090e-01   0.283   0.7773
## FFMC         2.233e-04  1.587e-04   1.407   0.1593
## rain1       -1.127e+00  8.355e-01  -1.349   0.1773
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 712.31  on 514  degrees of freedom
## Residual deviance: 681.67  on 500  degrees of freedom
## AIC: 711.67
##
## Number of Fisher Scoring iterations: 14
```

```
## P-value for all the coef being 0
```

```
## [1] 0.00687173
```

Just like the above it makes sense that if the covariates won't explain the continuous areas greater than 0, there won't be much significance in a linear model for a dummy variable either. This is exactly what we see, most of the coefficients are simply not significant at all. However, there is comfort that the difference in deviances around 30+ gives a significant p-value of less than 1% (using the pchisq()). This at least gives comfort that these covariates are doing something useful for the prediction.

Discussion/Conclusion of Whole Project

In the first analysis despite doing all the necessary transformations, modelling and diagnosing, it did not change the fact that a linear model might not have been the best for the response variable area. This was further emphasized in my final analysis by looking at the diagnosis plots such as residuals vs. fitted values, etc. This dataset was quite exemplary in showing how often the response variable can have more noise than all the explanatory variable's power combined. However, because of the nature of this class and the projects

instructions, I still did come up with two candidate models for different purposes, predicting and explaining. The predicting model had PCA terms to solve multicollinearity. Of course that means if the fireman had to predict, they would have to put their X explanatory variable and change it to PCAs.

The latter part of this project was motivated by considering modelling area in terms of a different setting - the fireman simply caring of whether or not there will be a fire big enough to pay attention to. Since the first analysis just didn't consider all the unnecessary fires, thus giving a conservative estimate, I decided to explore exactly modelling those 0s also in a dummy variable setting. The linear setting, as expected, turned out to be equally as poor because even the final candidate model didn't seem so significant. However, this is really the best one can do sometimes in life.

One of the biggest lesson learned is that: one can clearly see through the analysis of these two questions that there are certain settings to do linear modelling and certain setting to not do linear modelling! People often forget that and just want to do linear modelling for everything, but sometimes the dataset is too noisy. This was a great reminder to the realistic limitations of linear models!