Contents

Abstract

Through this project, we are intended to discover environmental factors that are associated with burned area of forest using a dataset from UCI machine learning repository. Linear regression was employed as the statistical method, and the best fitted model was selected using a combination of forward selection, backward selection, and partial F-test. The final model was diagnosed and the associations between finally selected covariates and response were analysed, and suggestions on forest preservation were given based on the analysis results.

1 Introduction

1.1 Problem and Motivation

One major environmental concern is the severity of forest fires, which affects forest preservation, creates economical and ecological damage and causes human suffering. Such phenomenon is due to multiple causes such as human negligence and lightnings. Actions can be taken to control this disaster, prevending each year millions of forest hectares destroyed all around the world. In this study, we are intended to discover the most influential factors that cause larger forest burned, thus advise govenor to take advantage of this result.

1.2 Data

This study will consider forest fire data from the Montesinho natural park, from the Tras-os-Montes northeast region of Portuga. In the dataset, there were n=517 instances, with 13 variables:

- X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
- Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
- month - month of the year: 'jan' to 'dec'
- day - day of the week: 'mon' to 'sun'
- FFMC - FFMC index from the FWI system: 18.7 to 96.20
- DMC - DMC index from the FWI system: 1.1 to 291.3
- DC - DC index from the FWI system: 7.9 to 860.6
- ISI - ISI index from the FWI system: 0.0 to 56.10
- temp - temperature in Celsius degrees: 2.2 to 33.30
- RH - relative humidity in %: 15.0 to 100
- wind - wind speed in km/h: 0.40 to 9.40
- rain - outside rain in mm/m2 : 0.0 to 6.4
- area - the burned area of the forest (in ha): 0.00 to 1090.84

Data can be retrieved from UCI machine learning repository: https://archive.ics.uci.edu/ml/datasets/Forest+Fires

1.3 Questions of Interest

Without using geographic features, which are the spatial coordinates of the instances, we want to find the the factors that are statistically associated with the burned area of the forest, either positively, or negatively. By interpreting the estimation, finally we want to come up with suggestions for fast control of forest fire to minimize the loss.

2 Methods

Before fitting any linear regression models, we need to remove the outliers in the response and then we will perform a power transformation on the continuous variables, both predictors and response. Besides, the distributions of continuous variables and response by categorical variable, month, will be studied.

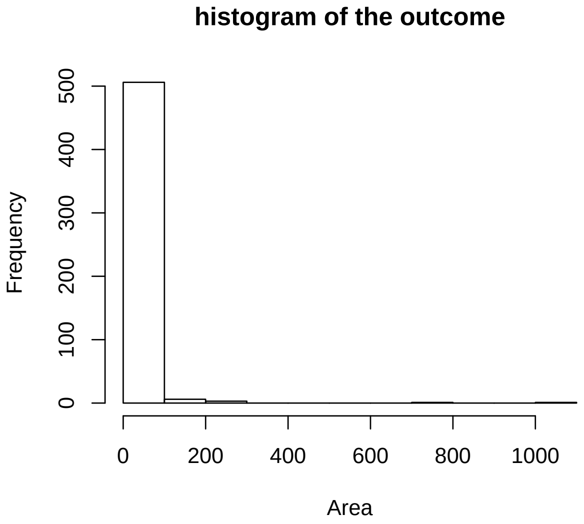
When fitting the linear regression models, we first include all potential covariates in the model, and then elimilate variables that have a multicolinearity problems (i.e. high variance inflation factor). Doing this will provide us with the full model. Next we will choose our base model only including temperature as the predictor. Forward selection and backward selection will be both performed using AIC, and the result will be compared using partial F-test to get our final model. Lastly, diagnostic check will be performed on the final model.

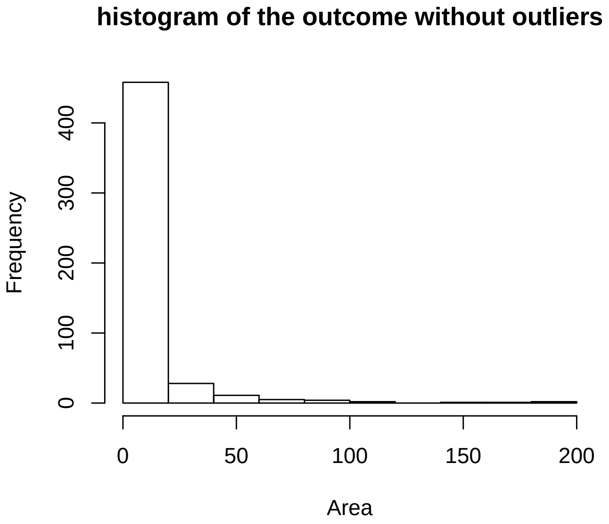
3 Results

3.1 Exploratory Data Analysis

3.1.1 Outcome Outliers

Before transforming the predicors and response, we removed the outliers in the response whose burned area is greater than 200 ha. And the distribution looks better:





3.1.2 Predictor Transformation

Before considering transformations for the response area, we will choose transformations for the continuous predictors. We can use a multivariate version of the Box-Cox method which will try to choose power transformations so that the predictors have approximately a multivariate normal distribution.

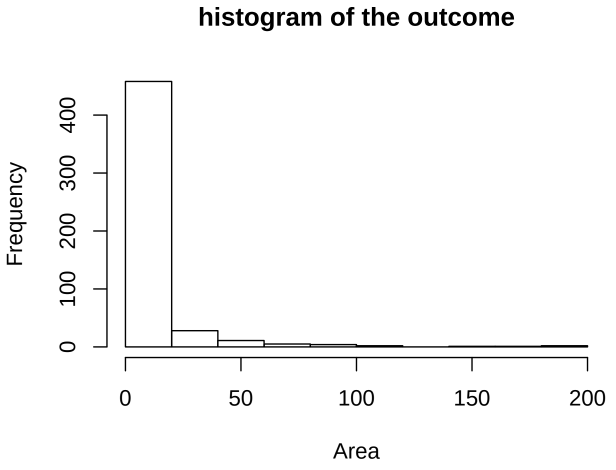
Seeing from the results (in Appendix), we should tranform variables RH by taking natural logarithm, as the $\lambda = 0$. In addition, we were supposed to power transform DMC and ISI by 0.33, but instead we take log transformation as the result is easier to interpret.

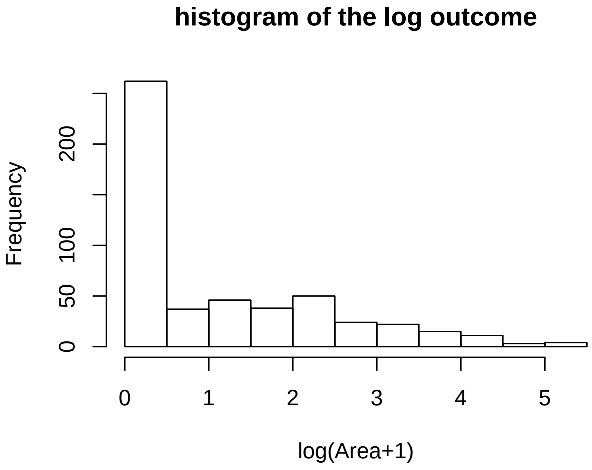
3.1.3 Outcome Transformation

Once the predictors are already transformed, we took a look at the distribution of the response, area, and we notice it is quite right skewed. Then we used the ordinary Box-Cox transformation for area using the already transformed predictors.

As shown in Appendix 2, the best transformation then is $\lambda = -0.55$. However, the log transformation is easier to interpret and is thus preferable if the fit is okay.

By taking log transformation for $area + 1$ (A simple solution for variables that can be 0 is to add a small constant before transforming), the histogram of the response looks better (much less right-screwed):

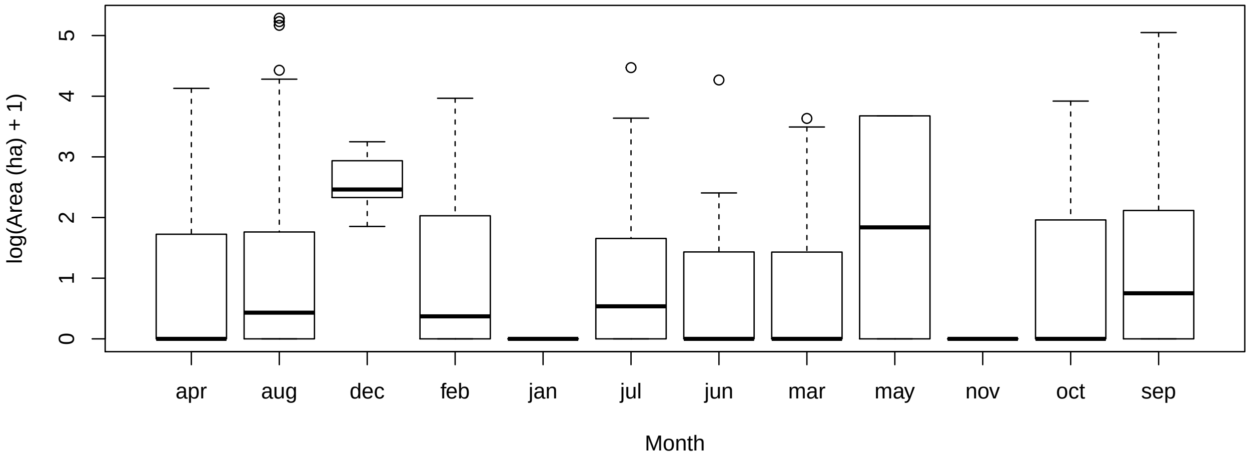




However, as there are many zero's in the response area, the distribution looks right skewed while the non-zero data looks normally distributed.

3.1.4 Categorical Variable

There are 12 numeric variables and 1 categorical variable, let's take a look at the distribution of transformed area by different month:



Apparently, the distributions of area are different by month, especially December has a higher IQR than other months, which may raise our concern in the later analysis that during December the burned area can be larger.

3.2 Regression Analysis

First of all, we considered the full model that includes all covariates that can be useful in the prediction:

$$log(area+1) \sim temp+log(RH)+wind+rain+FFMC+log(DMC)+DC+log(ISI+1)+month$$

On the other hand, we can argue that the response and log(area) will be positively correlated with the temperature, so we will only consider models including this predictor as our smallest model:

$$log(area+1) \sim temp$$

3.2.1 Multicollinearity

Before we proceed to build linear models, we need to check for multicollinearity among the covariates. The variables in this dataset can be correlated as FFMC, DMC, DC, and ISI were computed based on temp, RH, wind, and rain, referring to https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system.

As variable DC has a really large $GVIF^{(\frac{1}{2\times Df})}$ of 4.965 (see Appendix), we considered to exclude it from the full model. Thus, our full model becomes:

$$log(area+1) \sim temp+log(RH)+wind+rain+FFMC+log(DMC)+log(ISI+1)+month$$

3.2.2 Forward Model Selection

Beginning with the smallest model, the forward selection procedure will add variables one at a time until the chosen information criterion cannot be decreased anymore. Using AIC (by R default), we got our model as (see Appendix 2):

$$log(area + 1) = \beta_0 + \beta_1 temp + \beta_2 wind + \beta_{Jan} monthJan + \beta_{Feb} monthFeb + ... + \beta_{Dec} monthDec$$

3.2.3 Backward Model Selection

Beginning with the largest model, we remove one at a time until the information criterion cannot be decreased. Backward selection yields larger models than forward, and we had the same model as in the forward selection model (see Appendix 2)

Therefore, we will use this model as our preliminary model.

3.2.4 Partial F-test

Seeing from the summary table of our preliminary model (see Appendix 2), we notice that the p-value for wind is 0.1156, which means this predictor is not statistically significant. So, now we are using partial F-test to test if the reduced model without this covariate is as good as our preliminary model, and the null hypothsis and the alternative hypothesis for this test is:

$$H_0 : \beta_2 = 0$$
$$H_a : \beta_2 \neq 0$$

Since the test statistic is F=2.484 and p-value for the partial F test is 0.1156, we don't have evidence to reject the null hypothesis. Thus, the reduced model is as good as the preliminary model, and the reduced model is preferred as it is simpler. So, we will exclude wind from the our model.

Therefore, our final model is:

$$log(area + 1) = \beta_0 + \beta_1 temp + \beta_{Jan} monthJan + \beta_{Feb} monthFeb + ... + \beta_{Dec} monthDec$$

3.2.5 Diagnostic Check

By looking at the residual plot and normal QQ plot (see Appendix 2), we found the residuals were appoximately normally distributed with constant variance, while there was left tail in the QQ plot, which might be due to the right skewed distribution of $log(area + 1)$. Overall, the fit looks reasonable.

3.3 Interpretation

Seeing from the summary table of the final model in Appendix 2, we have a esmatimate for temp of $\beta_{temp} = 0.02501$ with standard error 0.01361, and p-value 0.06663. This means the effect of temperature is positive and moderately significant, which means 1 degree higher the temperature is, $e^{0.02501} = 1.025$ ha more forest will be burned on average, with 95% confidence interval of:

```
##          2.5 %   97.5 %
## temp 0.9982797 1.053111
```

Besides, the parameter of December is statistically significant (p=0.00752), with $\beta_{Dec} = 1.67060$, which means in December there will be $e^{1.67060} = 5.315$ ha more forest burned on average, with 95% confidence interval of:

```
##             2.5 %   97.5 %
## monthdec 1.564555 18.05834
```

4 Conclusion

Through the analysis, we found two significant factors that are associated with the area of forest burned, temperature and month. Unsurprisingly, the temperature is positively correlated with the burned area, and every 1 degree rising causes 1.025 ha more forest burned on average. In addition, during December there are 5.315 ha more forest burned than other months on average.

This result suggests us that global warming can be a main reason more forests being burned, and preventing global warming can help with forest preservation. On the other hand, governor should pay more attention and take actions such as increasing the number of forest rangers and fire fighers near foreset during the winter time, especially December.

Appendix 1: R code

```r
library(MASS)
library(car)
library(alr4)
fire <- read.csv("/Users/will/Desktop/PSTAT 126/forestfires.csv", header = T)

# remove the outliers in the response
par(mfrow=c(1,2))
hist(fire$area, main = "histogram of the outcome", xlab = "Area") # histogram of the outcome
fire <- fire[fire$area <= 200,]
hist(fire$area, main = "histogram of the outcome without outliers", xlab = "Area")
par(mfrow=c(1,1))

# Transforming Predictors
fire.pt = powerTransform(cbind(temp,RH,wind, DMC, DC, ISI+1) ~ 1, fire)
summary(fire.pt)
fire_trsf <- with(fire,
                  data.frame(area, temp, log(RH), wind, rain, FFMC, log(DMC), DC, log(ISI+1)))

# Transforming the outcome
fire.lm = lm(area+1~., data = fire_trsf)
bc <- boxCox(fire.lm)
bc$x[which.max(bc$y)]

par(mfrow=c(1,2))
# histogram of the outcome
hist(fire$area, main = "histogram of the outcome", xlab = "Area")
# histogram of the log outcome
hist(log(fire$area+1), main = "histogram of the log outcome", xlab = "log(Area+1)")
par(mfrow=c(1,1))

boxplot(log(fire$area+1) ~ fire$month, xlab = "Month", ylab = "log(Area (ha) + 1)")

full_mod <- lm(log(area+1) ~ temp + log(RH) + wind + rain + FFMC + log(DMC) + DC + log(ISI+1) +
               data = fire)
vif(full_mod) # remove DC because of high vif
f = ~ temp + log(RH) + wind + rain + FFMC + log(DMC) + log(ISI+1) + month

# The base model
m0 = lm(log(area+1) ~ temp, fire)
# Forward
m.forward = step(m0, f, direction = 'forward') # Uses AIC by default

m1 = update(m0, f)
# Backward
m.backward = step(m1, scope = c(lower = ~ temp), direction = 'backward')
```

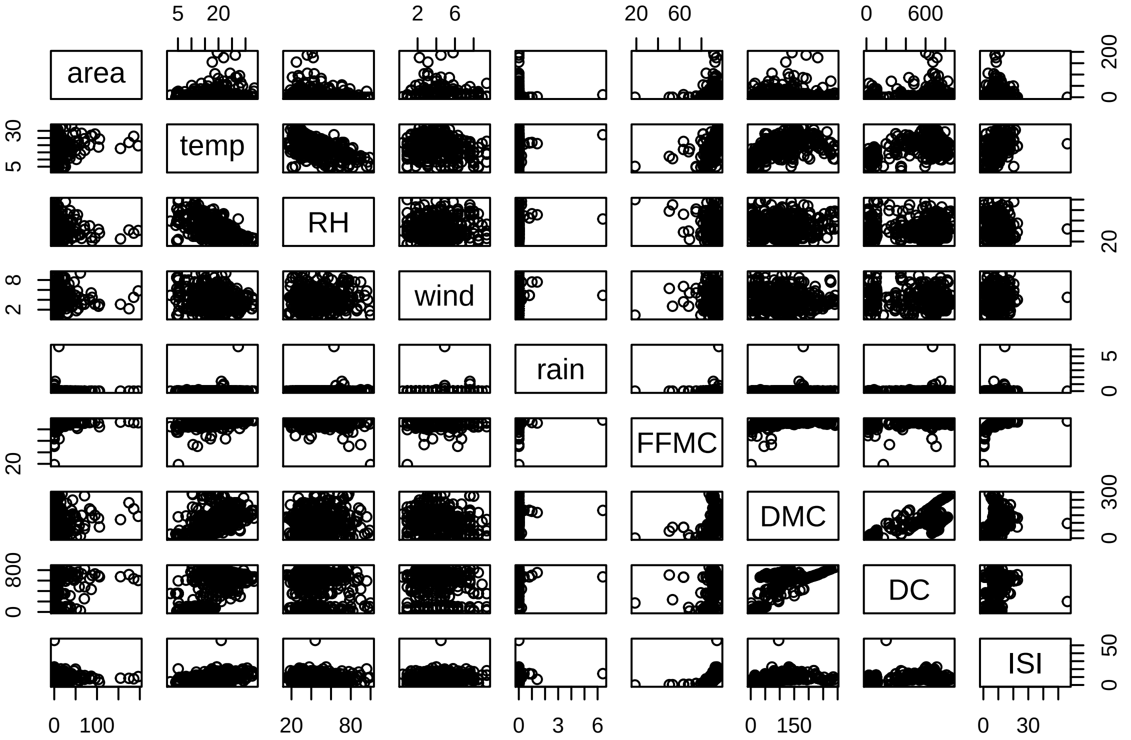
```r
summary(lm(log(area+1) ~ temp + wind + month, data=fire))

# partial F-test
anova(lm(log(area+1) ~ temp + month, data = fire),
      lm(log(area+1) ~ temp + wind + month, data = fire))

final_mod <- lm(log(area+1)~temp + month, data=fire)
# Diagnostic Check
par(mfrow=c(1,2))
plot(final_mod,1)
plot(final_mod,2)
par(mfrow=c(1,1))

summary(final_mod)

# confidence interval for parameters
exp(confint(final_mod, 'temp', level=0.95))
exp(confint(final_mod, 'monthdec', level=0.95))
```

Appendix 2: Plots and Outputs

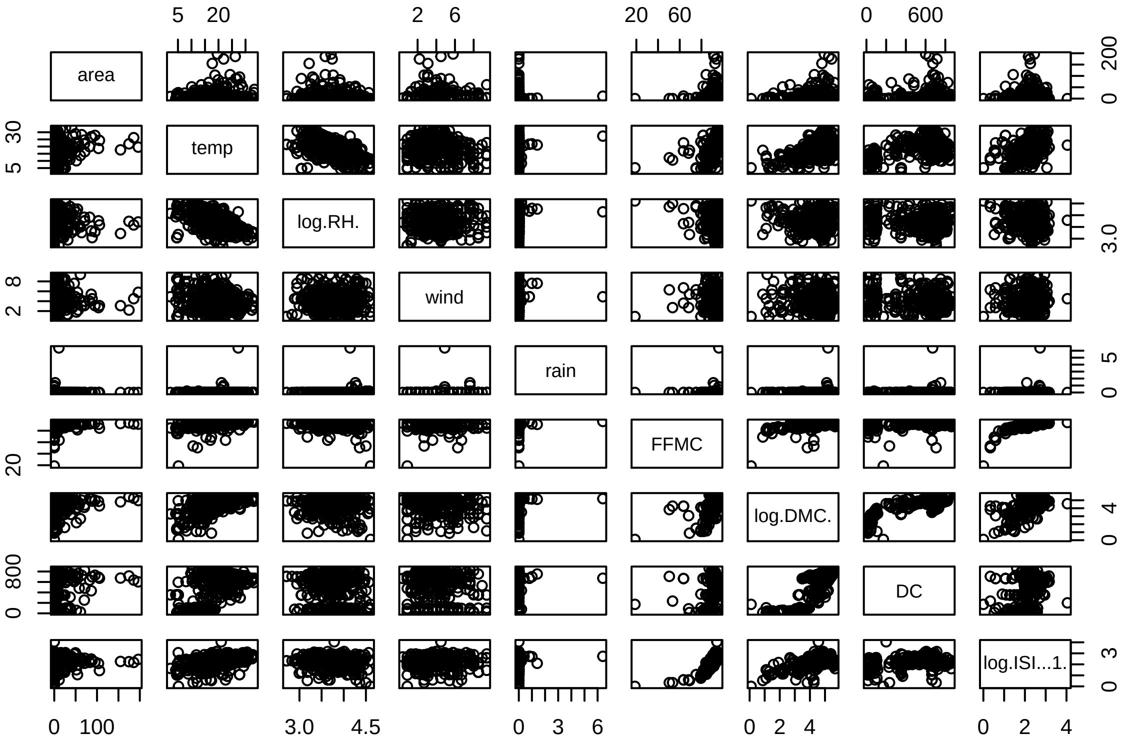
```r
# overview of the dataset
str(fire)
```

```
## 'data.frame':    512 obs. of  13 variables:
##  $ X    : int  7 7 7 8 8 8 8 8 8 7 ...
##  $ Y    : int  5 4 4 6 6 6 6 6 6 5 ...
##  $ month: Factor w/ 12 levels "apr","aug","dec",..: 8 11 11 8 8 2 2 2 12 12 ...
##  $ day  : Factor w/ 7 levels "fri","mon","sat",..: 1 6 3 1 4 4 2 2 6 3 ...
##  $ FFMC : num  86.2 90.6 90.6 91.7 89.3 92.3 92.3 91.5 91 92.5 ...
##  $ DMC  : num  26.2 35.4 43.7 33.3 51.3 ...
##  $ DC   : num  94.3 669.1 686.9 77.5 102.2 ...
##  $ ISI  : num  5.1 6.7 6.7 9 9.6 14.7 8.5 10.7 7 7.1 ...
##  $ temp : num  8.2 18 14.6 8.3 11.4 22.2 24.1 8 13.1 22.8 ...
##  $ RH   : int  51 33 33 97 99 29 27 86 63 40 ...
##  $ wind : num  6.7 0.9 1.3 4 1.8 5.4 3.1 2.2 5.4 4 ...
##  $ rain : num  0 0 0 0.2 0 0 0 0 0 0 ...
##  $ area : num  0 0 0 0 0 0 0 0 0 0 ...
```

```r
# correlation between variables
pairs(fire[c("area", "temp", "RH", "wind", "rain", "FFMC", "DMC", "DC", "ISI")])
```



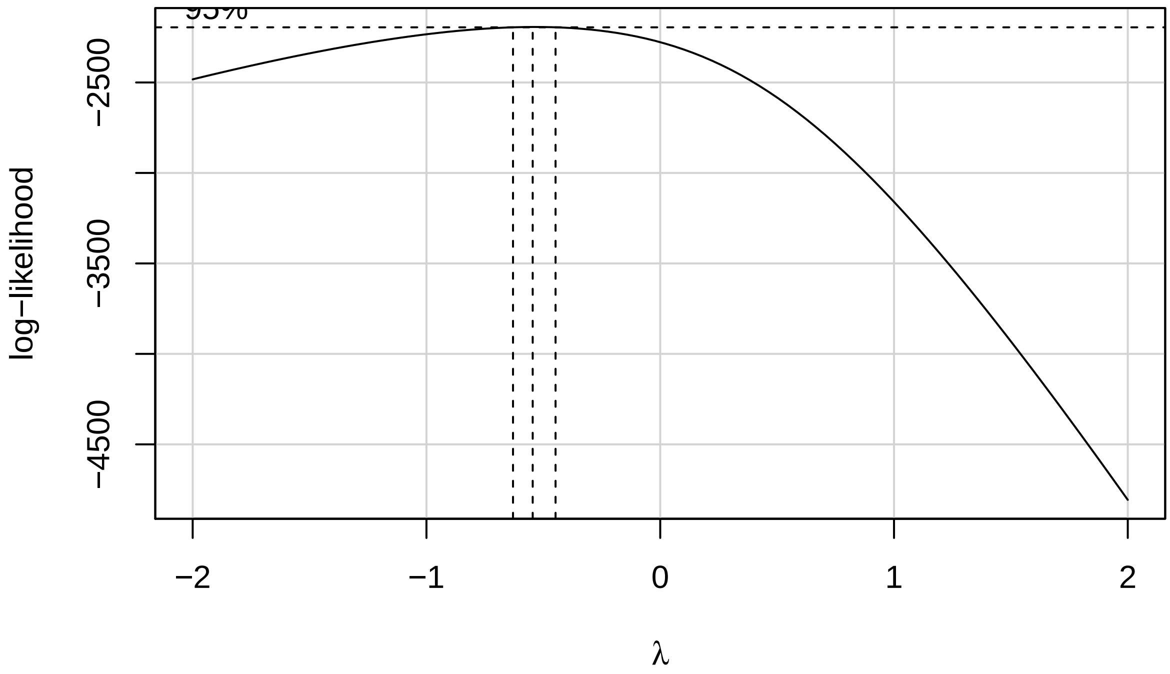
```
# correlation between trasnformed variables
pairs(fire_trsf)
```



```
# power transformation for predictors
summary(fire.pt)
```

```
## bcPower Transformations to Multinormality
##      Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## temp    1.0656        1.00       0.9076       1.2235
## RH      0.1645        0.00      -0.0272       0.3562
## wind    0.5578        0.50       0.4066       0.7090
## DMC     0.3543        0.33       0.2899       0.4187
## DC      1.2153        1.22       1.0755       1.3552
##         0.2950        0.33       0.1862       0.4038
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                     LRT df       pval
## LR test, lambda = (0 0 0 0 0 0) 810.4817  6 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                     LRT df       pval
## LR test, lambda = (1 1 1 1 1 1) 500.8152  6 < 2.22e-16
```

```
# Transforming the outcome
fire.lm = lm(area+1~., data = fire_trsf)
bc <- boxCox(fire.lm)
```



```
bc$x[which.max(bc$y)]
```

```
## [1] -0.5454545
```

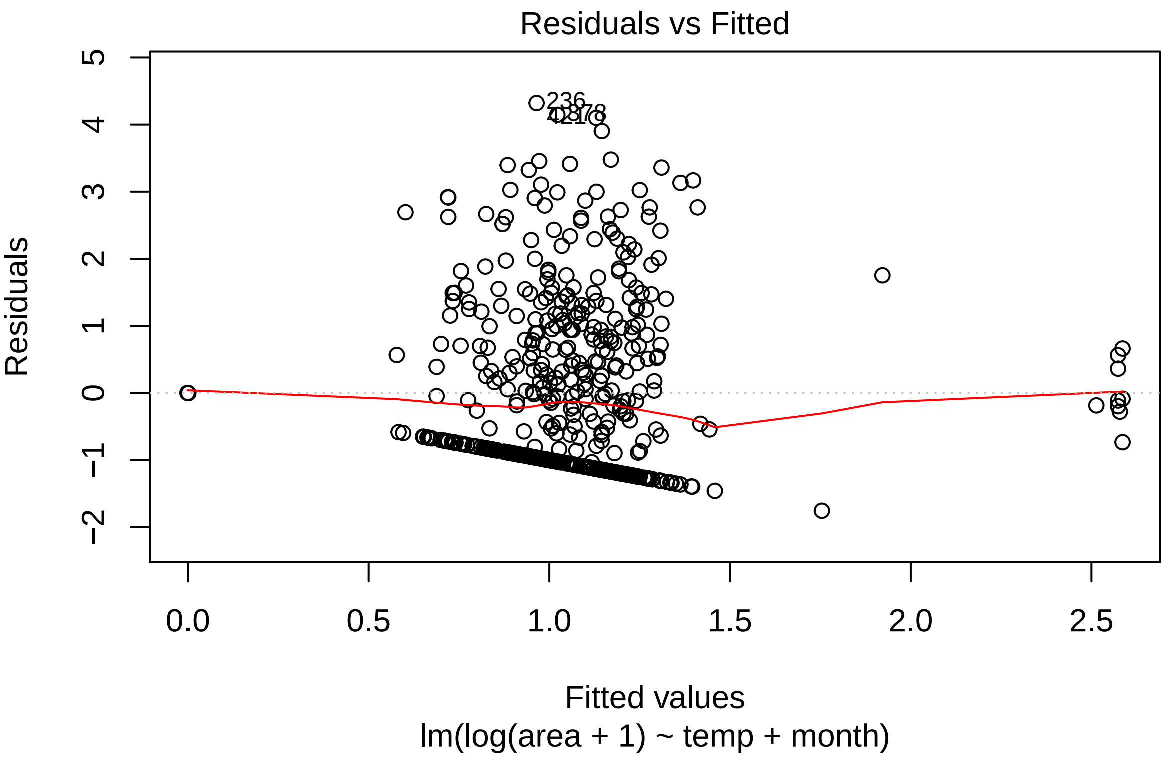
```
# remove DC because of high vif
vif(full_mod)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## temp            4.185693  1        2.045897
## log(RH)          2.435040  1        1.560462
## wind             1.257497  1        1.121382
## rain             1.058496  1        1.028832
## FFMC             3.682838  1        1.919072
## log(DMC)         9.445555  1        3.073362
## DC              24.655059  1        4.965386
## log(ISI + 1)     3.315722  1        1.820912
## month          138.359127 11        1.251176
```

```
# summary table of the second last model
summary(lm(log(area+1)~temp + wind + month, data=fire))
```

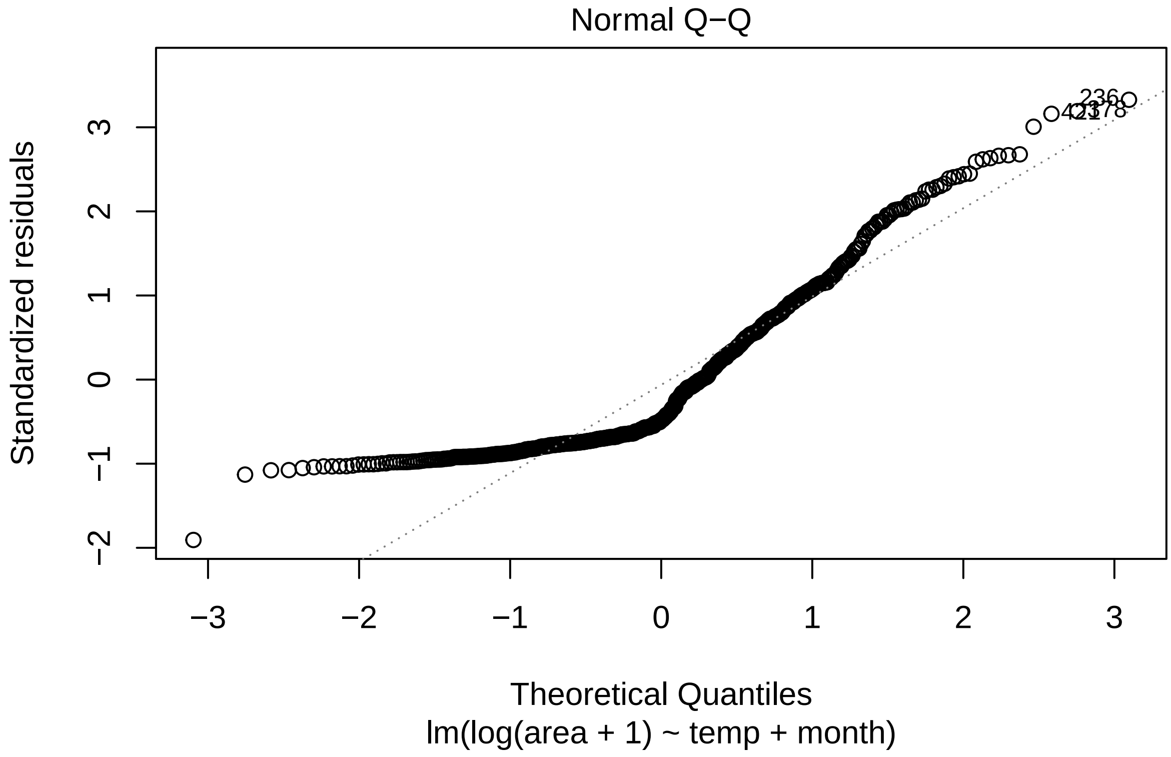
```
## 
## Call:
## lm(formula = log(area + 1) ~ temp + wind + month, data = fire)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.7690 -1.0030 -0.5808  0.8562  4.2484 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  0.49698    0.49881   0.996   0.3196  
## temp         0.02793    0.01371   2.037   0.0422 *
## wind         0.05483    0.03479   1.576   0.1156  
## monthaug    -0.30924    0.46265  -0.668   0.5042  
## monthdec     1.52923    0.62799   2.435   0.0152 *
## monthfeb     0.11589    0.52431   0.221   0.8252  
## monthjan    -0.75326    1.02630  -0.734   0.4633  
## monthjul    -0.37979    0.51125  -0.743   0.4579  
## monthjun    -0.45296    0.54830  -0.826   0.4091  
## monthmar    -0.36222    0.46864  -0.773   0.4399  
## monthmay     0.68779    1.01732   0.676   0.4993  
## monthnov    -1.07326    1.37094  -0.783   0.4341  
## monthoct    -0.24700    0.55361  -0.446   0.6557  
## monthsep    -0.04746    0.45722  -0.104   0.9174  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.301 on 498 degrees of freedom
## Multiple R-squared:  0.05016,    Adjusted R-squared:  0.02536 
## F-statistic: 2.023 on 13 and 498 DF,  p-value: 0.01749
```

```
# Diagnostic Check
plot(final_mod,1)
```



```r
plot(final_mod,2)
```

```
## Warning: not plotting observations with leverage one:
##   512
```



```
# summary of the final model
summary(final_mod)
```

```
##
## Call:
## lm(formula = log(area + 1) ~ temp + month, data = fire)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7541 -0.9929 -0.6225  0.8082  4.3209
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.78795    0.46407   1.698  0.09015 .
## temp         0.02501    0.01361   1.838  0.06663 .
## monthaug    -0.31348    0.46333  -0.677  0.49899
## monthdec     1.67060    0.62248   2.684  0.00752 **
## monthfeb     0.05888    0.52383   0.112  0.91055
## monthjan    -0.91927    1.02240  -0.899  0.36902
## monthjul    -0.40370    0.51179  -0.789  0.43060
## monthjun    -0.45749    0.54911  -0.833  0.40516
## monthmar    -0.34265    0.46917  -0.730  0.46553
## monthmay     0.68350    1.01883   0.671  0.50262
## monthnov    -1.08311    1.37296  -0.789  0.43055
## monthoct    -0.29846    0.55346  -0.539  0.58994
## monthsep    -0.08543    0.45726  -0.187  0.85188
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.303 on 499 degrees of freedom
## Multiple R-squared:  0.04542,    Adjusted R-squared:  0.02247
## F-statistic: 1.979 on 12 and 499 DF,  p-value: 0.02431
```