What Influences Forest Fires Area? (Lab 5)

Ye (Eric) Wang

February 27, 2016

Source: http://archive.ics.uci.edu/ml/datasets/Forest+Fires (http://archive.ics.uci.edu/ml/datasets/Forest+Fires) Reference: P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9.

Load the data

```
forestfire <- read.csv("http://archive.ics.uci.edu/ml/machine-learning-databases/forest-fires/forestfires.csv")
```

Preliminaries

Define the question of interest. Review the design of the study (for thinking about model assumptions). Correct erros in the data.

This is a **difficult** regression task, where the aim is to *understand how the burned area of forest fires, in the northeast region of Portugal, is related to the meteorological and other data*. And below gives some details of the variables considered in this study.

1. Spatial information
    - 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
2. Temporal information
    - month - month of the year: ‘jan’ to ‘dec’
    - day - day of the week: ‘mon’ to ‘sun’
3. FWI: The forest Fire Weather Index (FWI) is the Canadian system for rating fire danger
    - FFMC - FFMC index denotes the moisture content surface litter and influences ignition and fire spread: 18.7 to 96.20
    - DMC - DMC index represent the moisture content of shallow organic layers: 1.1 to 291.3
    - DC - DC index represent the moisture content of deep organic layers: 7.9 to 860.6
    - ISI - ISI index is a score that correlates with fire velocity spread: 0.0 to 56.10
4. Meteorological information
    - 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
5. area - the burned area of the forest (in ha): 0.00 to 1090.84

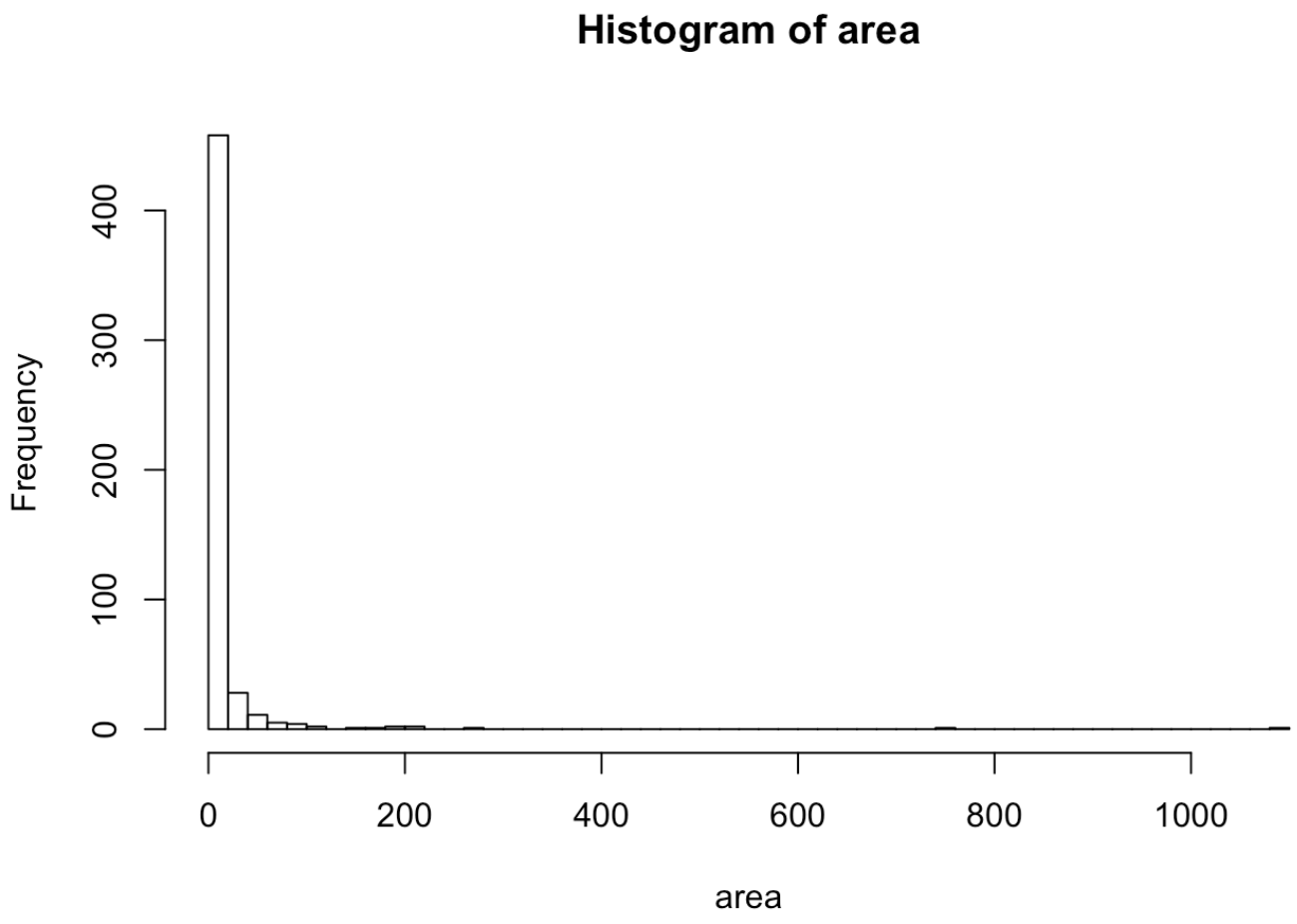
The first four rows denote the spatial and temporal attributes. Only two geographic features were included, the X and Y axis values where the fire occurred, since the type of vegetation presented a low quality (i.e. more than 80% of the values were missing). After consulting the Mon- tesinho fire inspector, we selected the month and day of the week temporal variables. Average monthly weather conditions are quite distinct, while the day of the week could also influence forest fires (e.g. work days vs weekend) since most fires have a human cause.

Explore the data

Look for initial answers to questions and for potential models.

Let us first explore our response variable `area` .

```
attach(forestfire)
hist( area, 40 )
```

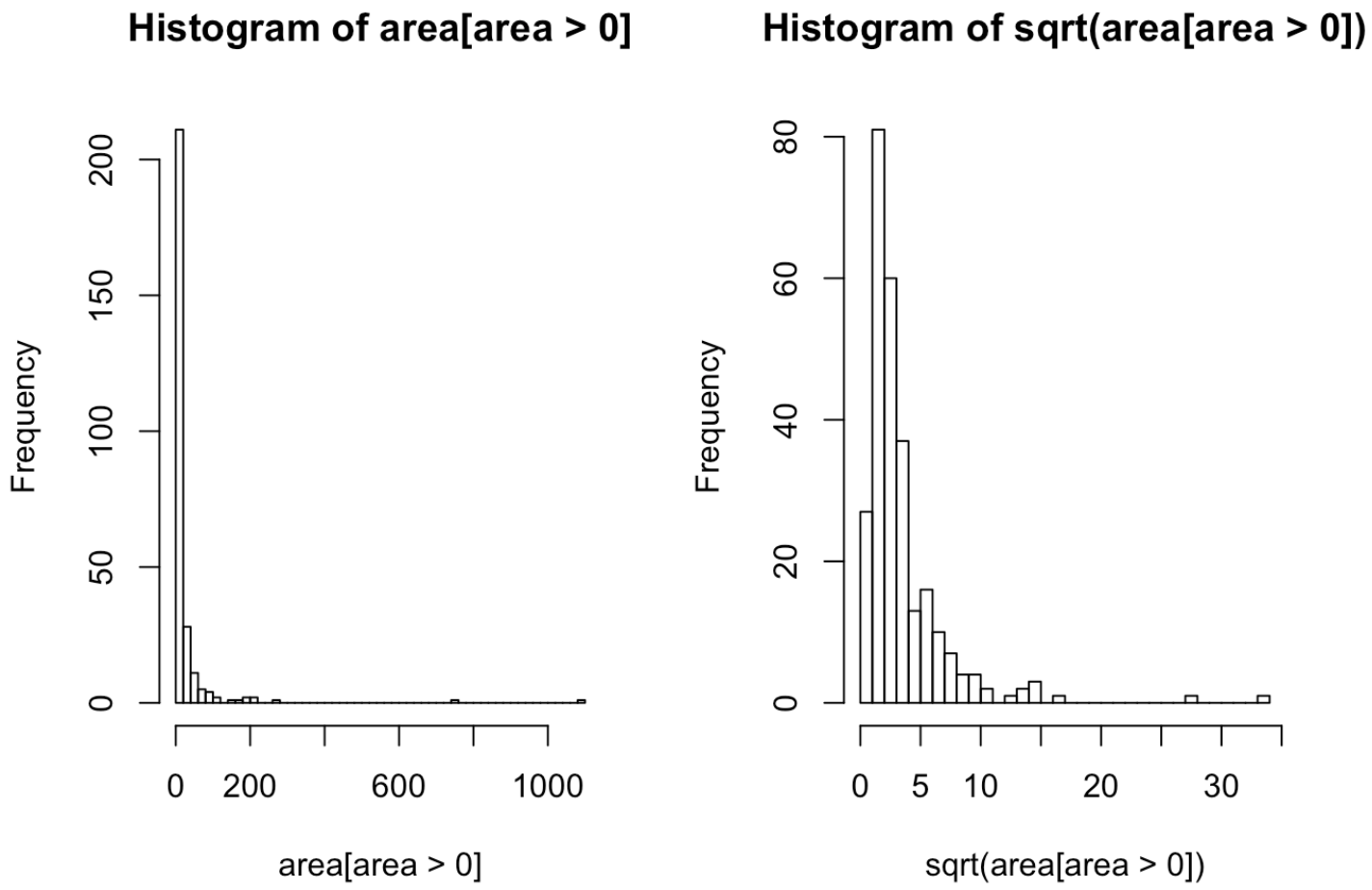



```
round(table(area==0)/nrow(forestfire),2)
```

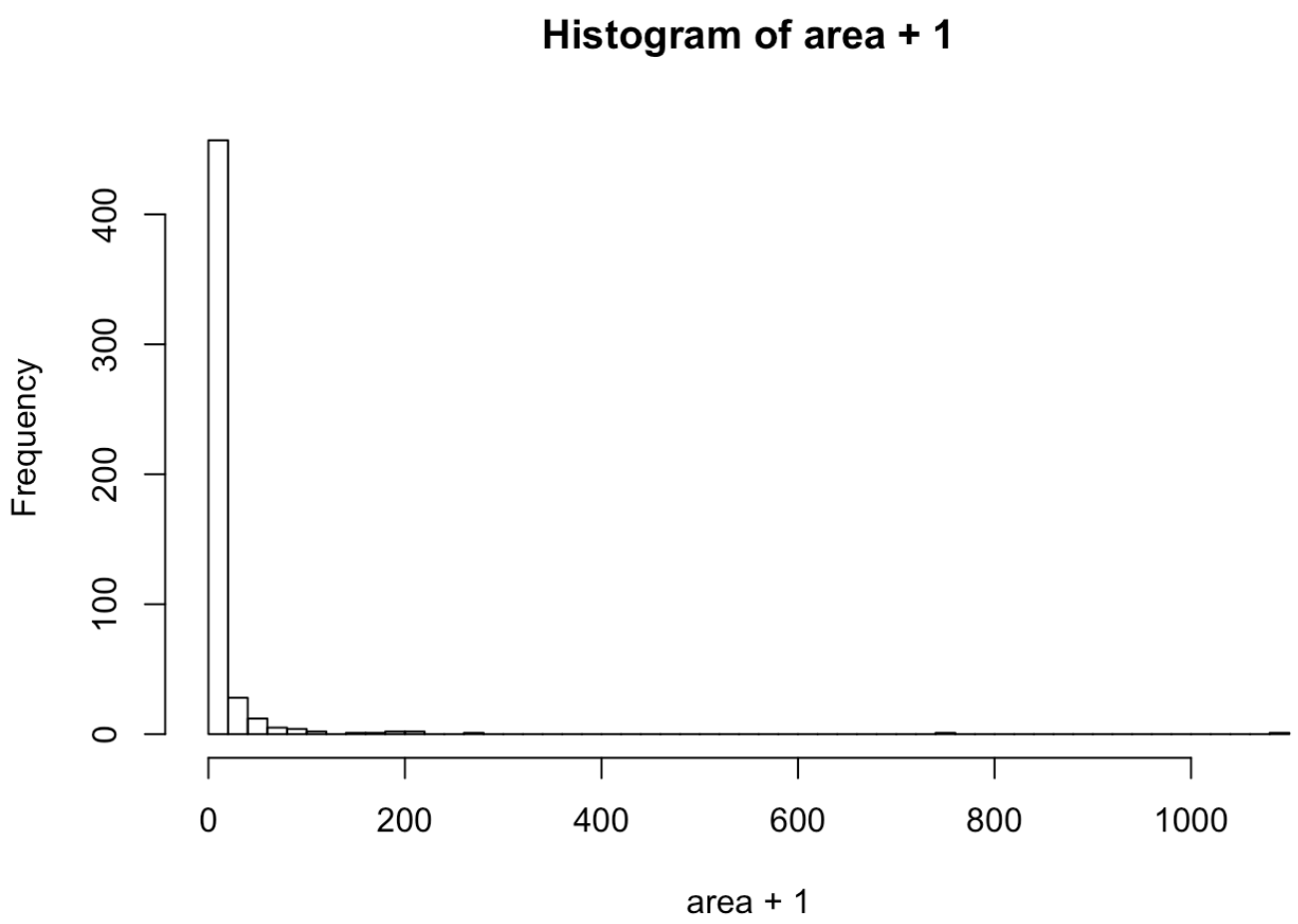
```
##
## FALSE  TRUE
##  0.52  0.48
```

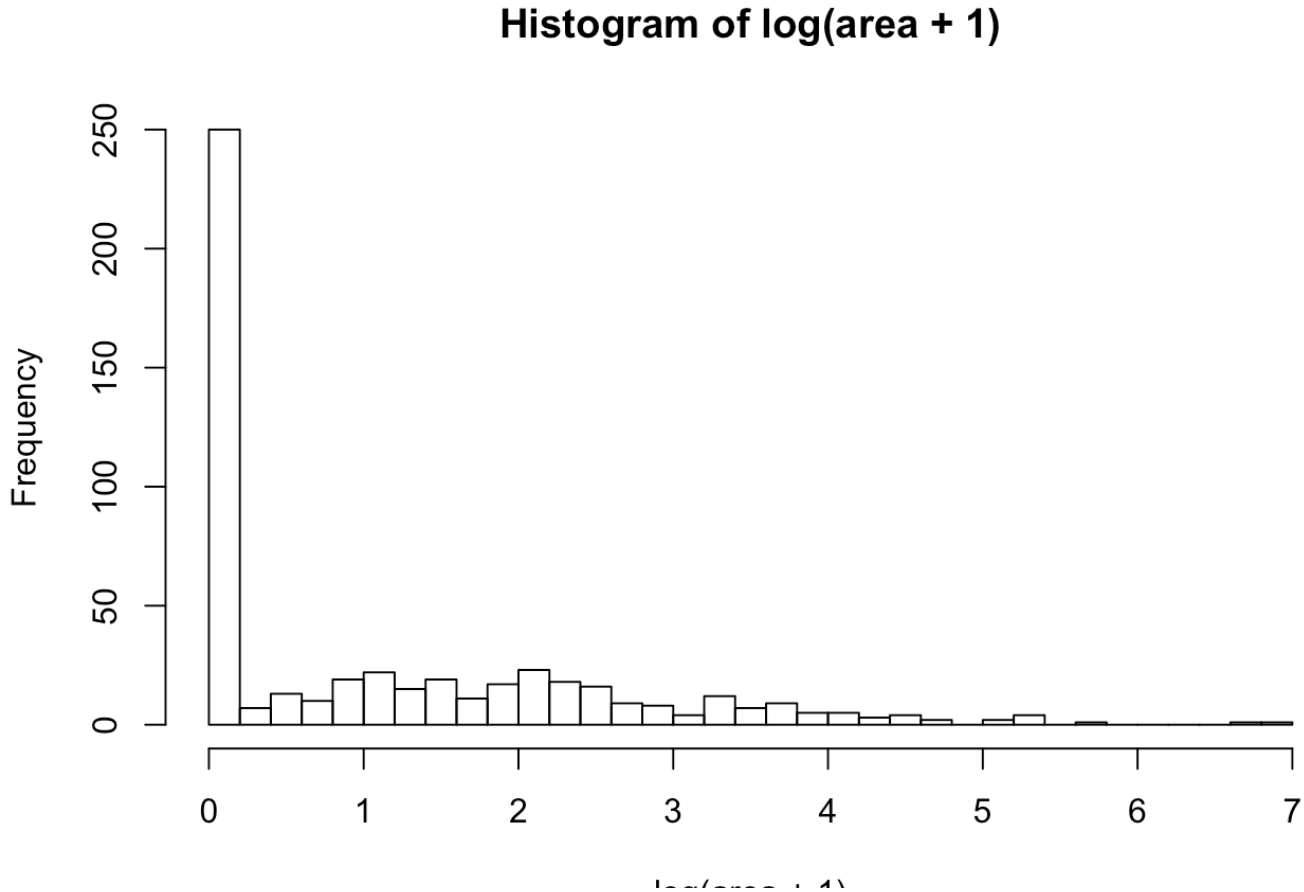
48% percent of the time there is no observation of a forest fire. This makes sense since there should be a positive probability that no forest fires are triggered at the time of observation.

Then let us take a look at the distribution of areas that are not trivial. There is a clear right skewness and hence we try a log tranformation.



Can we build a regression model that takes into account this excessive number of zeroes? In the original paper the authors proposed to transform the responses using `ln(y+1)`. But that does not solve the problem, as can be seen below.

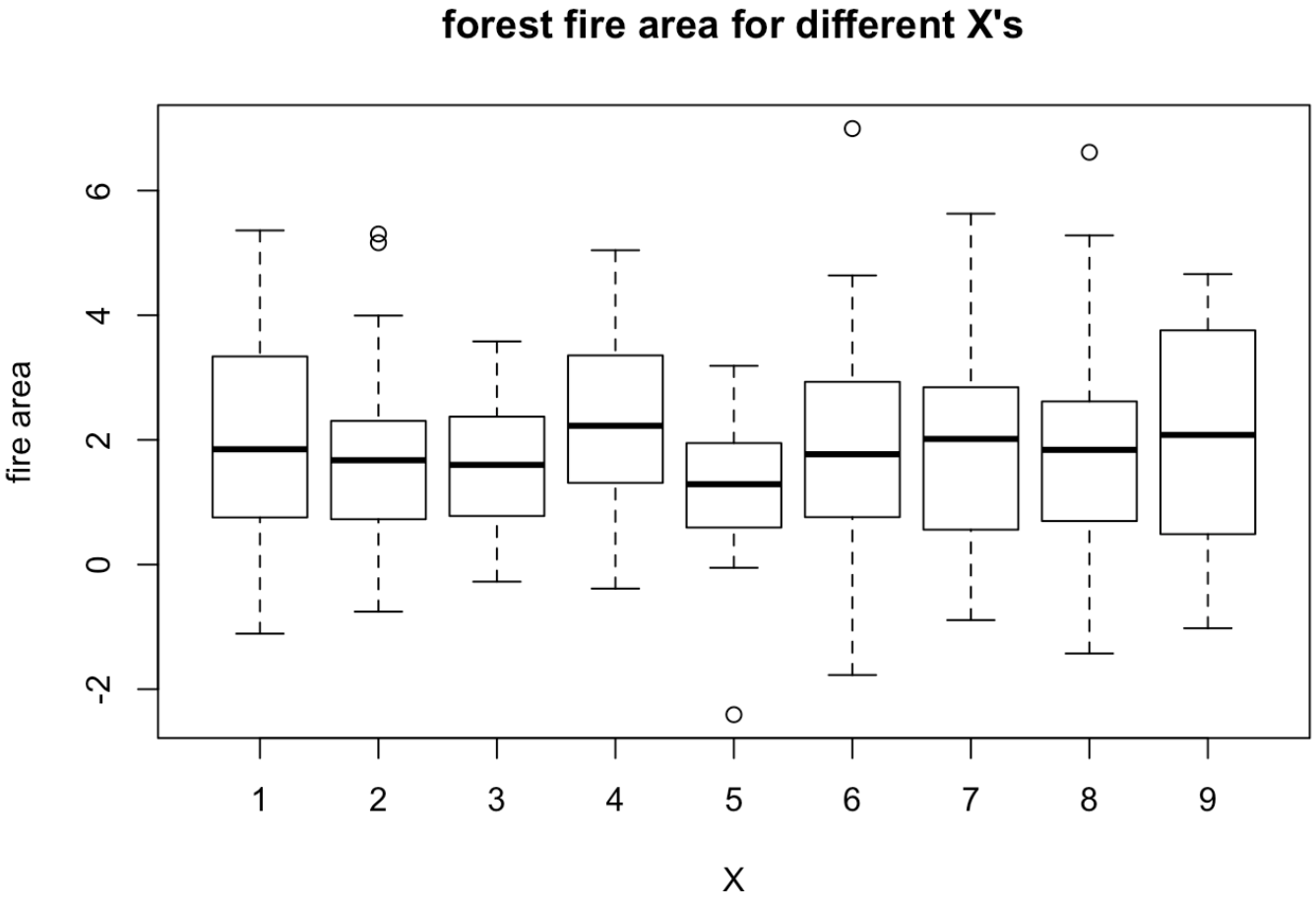

Histogram of area + 1
Frequency
0
100
200
300
400
0
200
400
600
800
1000
area + 1


Histogram of log(area + 1)
Frequency
0
50
100
150
200
250
0
1
2
3
4
5
6
7

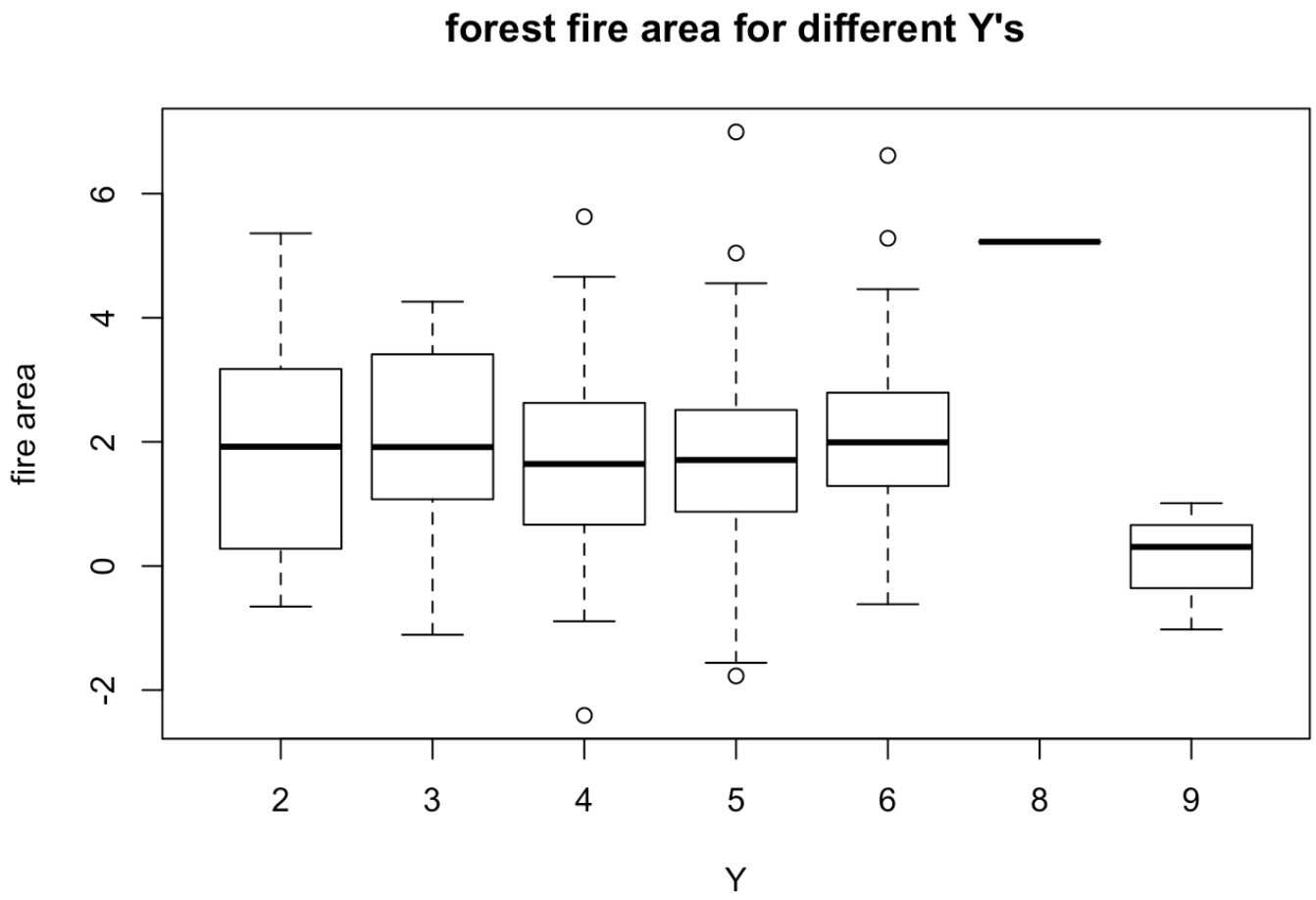
log(area + 1)

Moreover, I personally think the transformation itself makes no sense since under such transformation y can be smaller than 0, which is nonsense! I think the best model to solve this type of problems is the zero-inflation models, but these modles are beyond the scope of this lab so I will simply proceed with the multiple linear regression approach. But I will only focus on a subset of the data where areas are above zero. In other words, I will try to predict the area of the forest fires given that there is indeed a forest fire going on.

```
par(mfrow=c(1,1))
detach(forestfire)
forestfire <- forestfire[forestfire$area>0,]
# Let's explore the relationships between the response and the predictors
boxplot(log(area)~as.factor(X), data = forestfire, xlab = "X", ylab = "fire area",
main = "forest fire area for different X's")
```

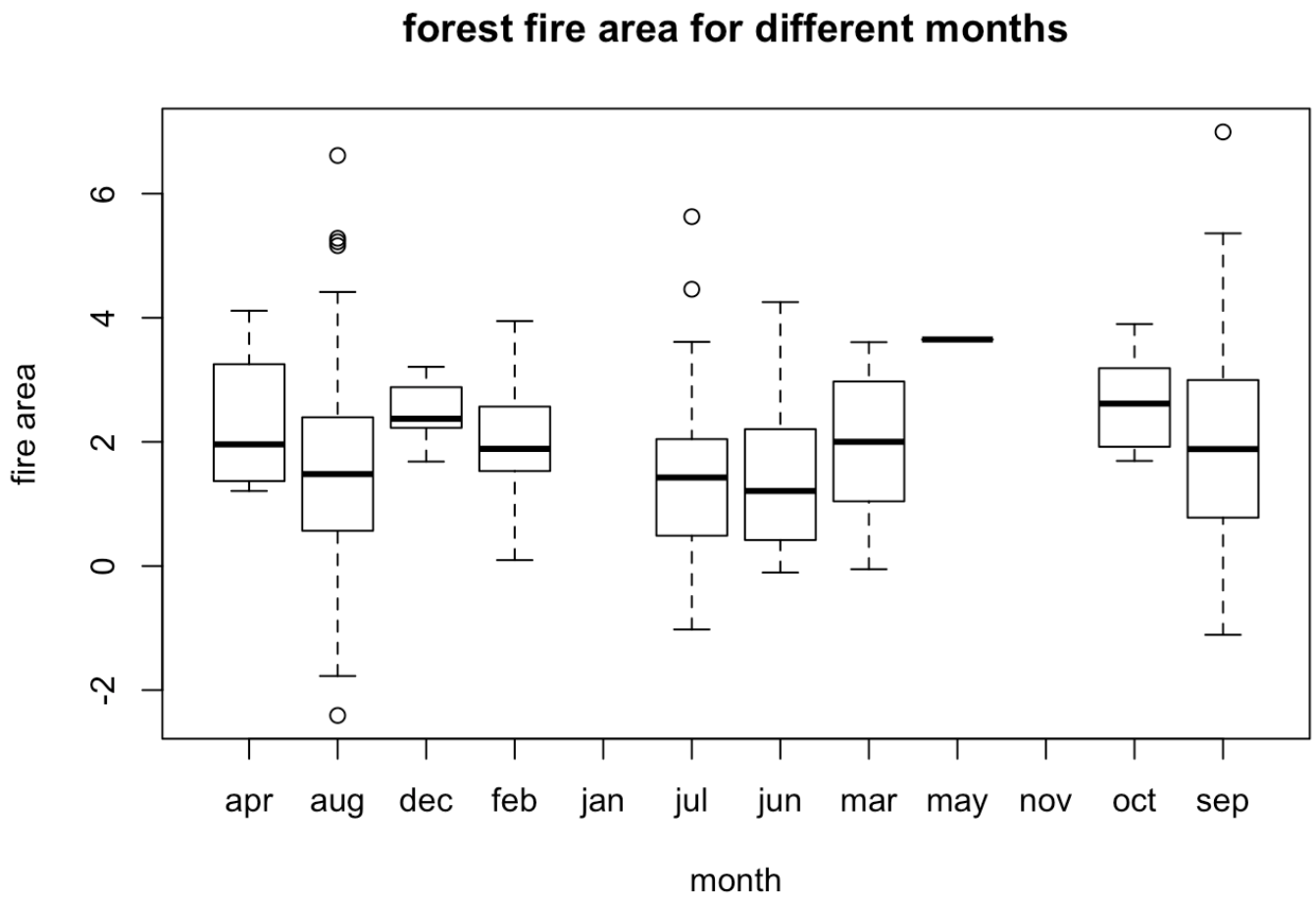



```
boxplot(log(area)~as.factor(Y), data = forestfire, xlab = "Y", ylab = "fire area",
main = "forest fire area for different Y's")
```



It seems that there is no obvious relationship between the spatial locations and the fire area, hence, for simplicity, I decide not to consider these two predictors.

```
boxplot(log(area)~month, data = forestfire, xlab = "month", ylab = "fire area", ma
in = "forest fire area for different months")
```

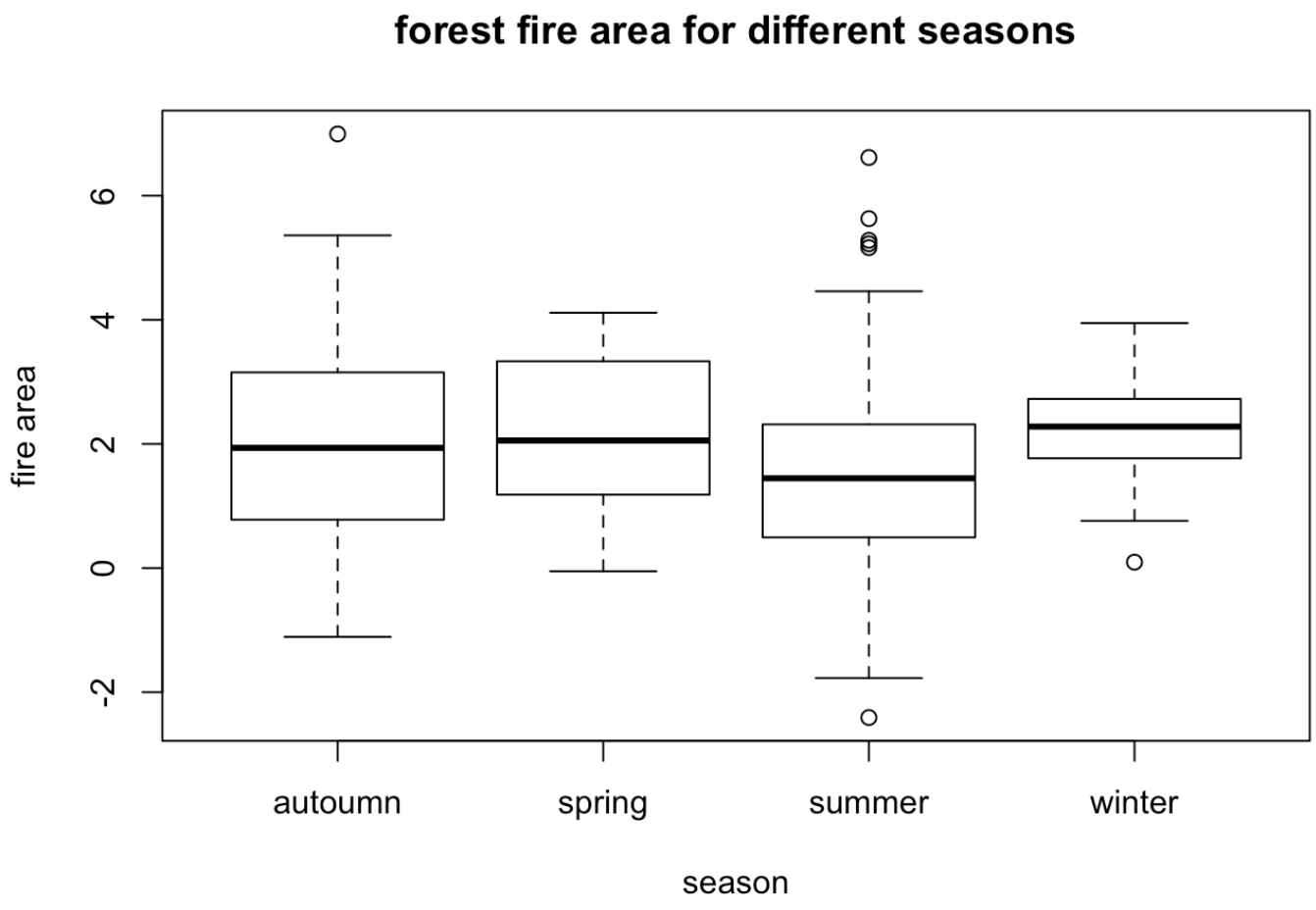



```
summary(forestfire$month)
```

```
## apr aug dec feb jan jul jun mar may nov oct sep
##   4  99   9  10   0  18   8  19   1   0   5  97
```

So the numbers of observations fall into every months are very unbalanced. In fact many months only have few observations. This introduce a great risk of overfitting. Hence I recode the month predictor into a categorical predictor of different seasons.

```
forestfire$season <- rep("spring", 270)
for (i in 1:270){
  if (forestfire$month[i] %in% c("feb","jan","dec")) forestfire$season[i] <- "wint
er"
  if (forestfire$month[i] %in% c("oct","nov","sep")) forestfire$season[i] <- "auto
umn"
  if (forestfire$month[i] %in% c("aug","jul","jun")) forestfire$season[i] <- "summ
er"
}
forestfire$season <- as.factor(forestfire$season)
forestfire$month <- NULL
boxplot(log(area)~season, data = forestfire, xlab = "season", ylab = "fire area",
main = "forest fire area for different seasons")
```

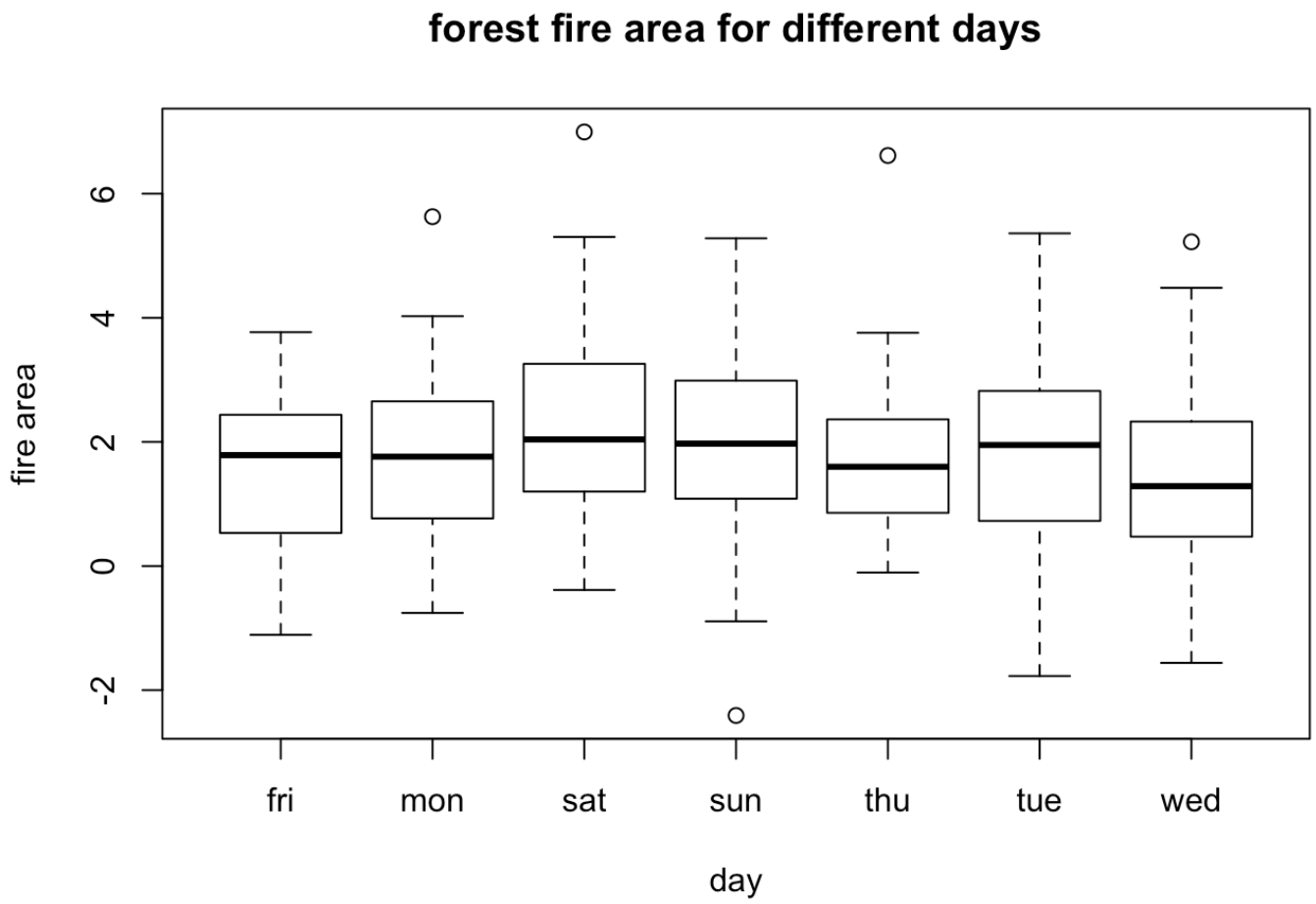



```
reg_season <- lm(log(area)~season,data = forestfire)
summary(reg_season)
```

```
##
## Call:
## lm(formula = log(area) ~ season, data = forestfire)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9829 -0.9652 -0.0766  0.8594  5.0401
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.0401     0.1499  13.612   <2e-16 ***
## seasonspring   0.0820     0.3434   0.239   0.8115
## seasonsummer  -0.4651     0.2020  -2.303   0.0221 *
## seasonwinter   0.1812     0.3782   0.479   0.6323
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.514 on 266 degrees of freedom
## Multiple R-squared:  0.02796,    Adjusted R-squared:  0.017
## F-statistic:  2.55 on 3 and 266 DF,  p-value: 0.0561
```

Now we seem to find some evididence that summer might tend to have less severe forest fires, probabliy due to the high humidity. This suggest that we should include season as a predictor in the model.

```
boxplot(log(area)~day, data = forestfire, xlab = "day", ylab = "fire area", main =
"forest fire area for different days")
```

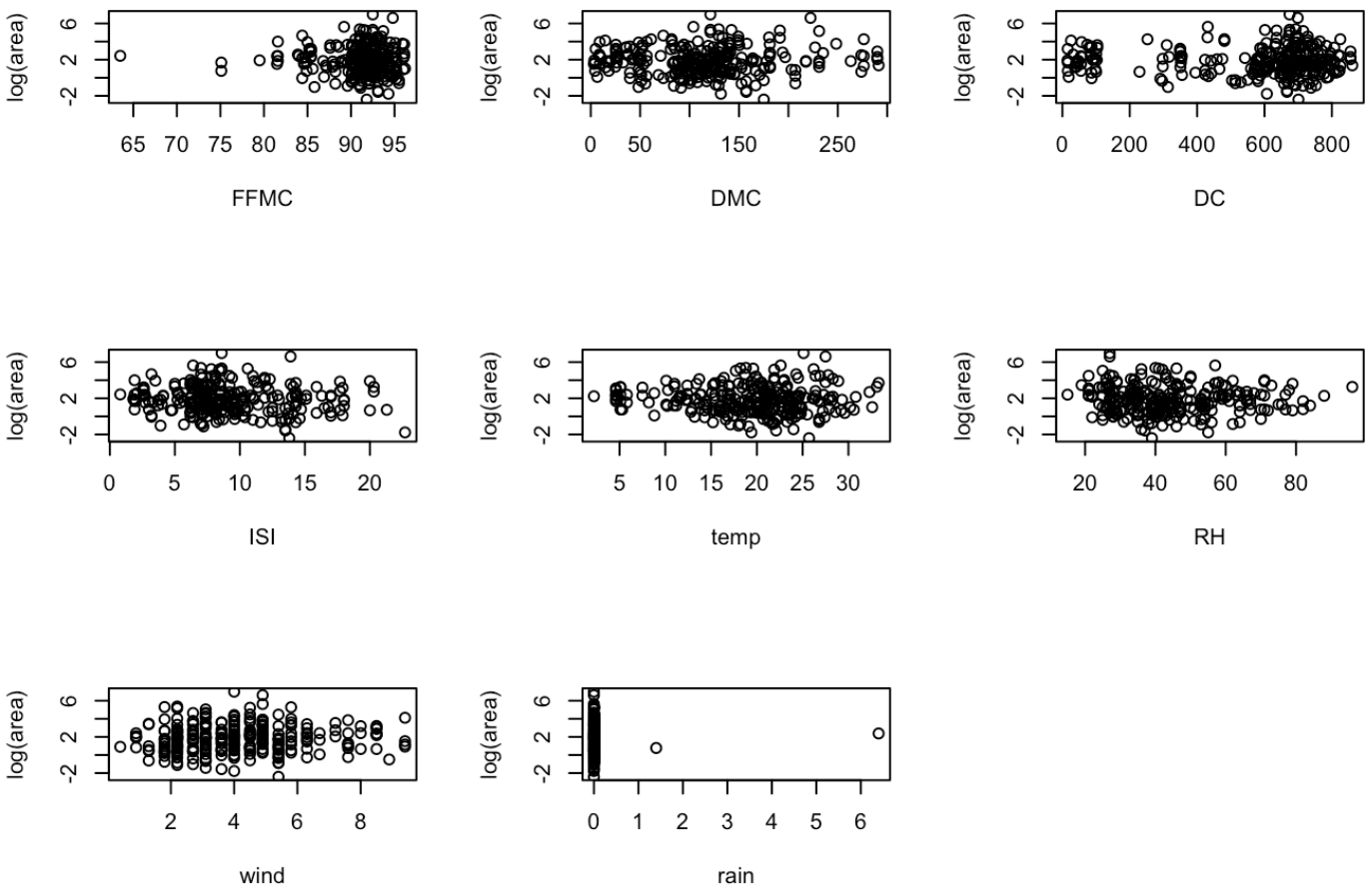



```
reg_day <- lm(log(area)~day,data = forestfire)
summary(reg_day)
```

```
##
## Call:
## lm(formula = log(area) ~ day, data = forestfire)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4189 -1.0289 -0.0829  0.9222  4.8357
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.6101     0.2327   6.918 3.48e-11 ***
## daymon        0.1985     0.3375   0.588   0.5570
## daysat        0.6266     0.3311   1.893   0.0595 .
## daysun        0.4008     0.3221   1.245   0.2144
## daythu        0.1693     0.3596   0.471   0.6381
## daytue        0.2564     0.3448   0.744   0.4577
## daywed       -0.1248     0.3563  -0.350   0.7264
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.526 on 263 degrees of freedom
## Multiple R-squared:  0.02305,    Adjusted R-squared:  0.0007647
## F-statistic: 1.034 on 6 and 263 DF,  p-value: 0.4033
```

We seem to have some very mild evidence that Saturday tend to have more severe forest fires probably because of the increased human activities during weekends. We should also include day as a predictor.

```
par(mfrow=c(3,3))
plot(log(area) ~ FFMC + DMC + DC + ISI + temp + RH + wind + rain,
     data = forestfire)
par(mfrow=c(1,1))
```



After looking at the scatterplots of all these predictors, I didn't find any obvious non-linear trend so I will first simply considering including the linear terms.

Last let us see if there is strong multi-colinearity in these continuous predictors.

```
which(cor(forestfire[,4:11])>0.8)
```

```
## [1]  1 10 19 28 37 46 55 64
```

Only diagonal elements exceed 0.8, seems that we do not need to worry about multi-colinearity.

Formulate an inferential model

We start from the simplest model, without any interaction and quadratic terms.

```
reg0 <- lm(log(area) ~season + day + FFMC + DMC + DC + ISI + temp + RH + wind + ra
in,
           data = forestfire)
summary(reg0)
```

```
##
## Call:
## lm(formula = log(area) ~ season + day + FFMC + DMC + DC + ISI +
##     temp + RH + wind + rain, data = forestfire)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2269 -0.9284  0.0104  0.8443  4.6198
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.114324   4.093447   0.272  0.78567
## seasonspring -0.500017   0.820988  -0.609  0.54304
## seasonsummer -0.892422   0.311260  -2.867  0.00449 **
## seasonwinter -0.025461   0.816610  -0.031  0.97515
## daymon        0.116749   0.348497   0.335  0.73790
## daysat        0.659230   0.335902   1.963  0.05080 .
## daysun        0.414535   0.332092   1.248  0.21310
## daythu        0.202118   0.368833   0.548  0.58418
## daytue        0.362894   0.349207   1.039  0.29971
## daywed        0.056571   0.366271   0.154  0.87738
## FFMC          0.011408   0.042759   0.267  0.78985
## DMC           0.006120   0.002578   2.374  0.01833 *
## DC           -0.001637   0.001181  -1.386  0.16692
## ISI          -0.027846   0.036041  -0.773  0.44047
## temp          0.016943   0.032874   0.515  0.60673
## RH           -0.005799   0.009560  -0.607  0.54470
## wind          0.071378   0.056706   1.259  0.20929
## rain          0.077246   0.242038   0.319  0.74988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.504 on 252 degrees of freedom
## Multiple R-squared:  0.09046,    Adjusted R-squared:  0.0291
## F-statistic: 1.474 on 17 and 252 DF,  p-value: 0.1042
```

As can be seen, the model fitting is terrible. This is not necessarily due to that we fit a bad model, it might simply because that the predictors we have do not have enough information to explain the responses. However, I decide to make one more try and hence I include the interaction terms between the four indices since I feel that the influece of these indices are not independent. I also include a quadratic term for wind and RH since I feel that the increase in fire area with respect to wind and the decrease in fire area with respect to humidity might be faster than linear.

```
reg1 <- lm(log(area) ~season + day + (FFMC + DMC + DC + ISI)^2 + temp + RH + wind
+ rain + I(wind^2) + I(RH^2),
           data = forestfire)
summary(reg1)
```

```
##
## Call:
## lm(formula = log(area) ~ season + day + (FFMC + DMC + DC + ISI)^2 +
##     temp + RH + wind + rain + I(wind^2) + I(RH^2), data = forestfire)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7567 -0.8923 -0.0498  0.8199  4.3699
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.813e+00  8.715e+00   0.208 0.835407
## seasonspring -3.403e-01  9.293e-01  -0.366 0.714505
## seasonsummer -1.295e+00  3.438e-01  -3.768 0.000206 ***
## seasonwinter  4.358e-01  1.009e+00   0.432 0.666291
## daymon       -4.883e-02  3.480e-01  -0.140 0.888519
## daysat        5.969e-01  3.313e-01   1.802 0.072817 .
## daysun        3.651e-01  3.259e-01   1.120 0.263791
## daythu        7.263e-02  3.711e-01   0.196 0.845014
## daytue        2.572e-01  3.513e-01   0.732 0.464700
## daywed        3.939e-03  3.639e-01   0.011 0.991373
## FFMC         -3.161e-03  1.050e-01  -0.030 0.976020
## DMC          -1.780e-01  1.150e-01  -1.548 0.122989
## DC            2.694e-02  1.987e-02   1.355 0.176533
## ISI           8.484e-02  9.511e-01   0.089 0.928991
## temp         -7.107e-03  3.417e-02  -0.208 0.835424
## RH           -5.319e-02  3.474e-02  -1.531 0.127048
## wind          4.317e-01  2.075e-01   2.081 0.038517 *
## rain          1.724e-02  2.391e-01   0.072 0.942601
## I(wind^2)    -4.274e-02  2.194e-02  -1.948 0.052593 .
## I(RH^2)       4.930e-04  3.399e-04   1.450 0.148222
## FFMC:DMC      2.238e-03  1.284e-03   1.743 0.082648 .
## FFMC:DC      -2.900e-04  2.386e-04  -1.215 0.225382
## FFMC:ISI     -9.532e-05  1.012e-02  -0.009 0.992496
## DMC:DC       -2.314e-05  1.406e-05  -1.646 0.101011
## DMC:ISI      -1.473e-04  8.419e-04  -0.175 0.861248
## DC:ISI       -1.759e-04  2.488e-04  -0.707 0.480089
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.472 on 244 degrees of freedom
## Multiple R-squared:  0.1566, Adjusted R-squared:  0.07013
## F-statistic: 1.812 on 25 and 244 DF,  p-value: 0.01247
```

This seems to be much better compared to the first model, at least now we have a significant F test. But this model involves too many predicitor and with most of them not significant. So let us then do a backward selection.

```
reg2 <- step(reg1, direction = "backward")
```

```
## Start:  AIC=233.51
## log(area) ~ season + day + (FFMC + DMC + DC + ISI)^2 + temp +
##     RH + wind + rain + I(wind^2) + I(RH^2)
##
##             Df Sum of Sq    RSS    AIC
## - day        6    12.831 541.67 227.99
## - FFMC:ISI   1     0.000 528.84 231.51
## - rain       1     0.011 528.85 231.52
## - DMC:ISI    1     0.066 528.91 231.55
## - temp       1     0.094 528.94 231.56
## - DC:ISI     1     1.084 529.93 232.06
## - FFMC:DC    1     3.202 532.04 233.14
## <none>                   528.84 233.51
## - I(RH^2)    1     4.560 533.40 233.83
## - RH         1     5.081 533.92 234.09
## - DMC:DC     1     5.874 534.72 234.49
## - FFMC:DMC   1     6.582 535.42 234.85
## - I(wind^2)  1     8.222 537.06 235.68
## - wind       1     9.382 538.22 236.26
## - season     3    35.198 564.04 244.91
##
## Step:  AIC=227.99
## log(area) ~ season + FFMC + DMC + DC + ISI + temp + RH + wind +
##     rain + I(wind^2) + I(RH^2) + FFMC:DMC + FFMC:DC + FFMC:ISI +
##     DMC:DC + DMC:ISI + DC:ISI
##
##             Df Sum of Sq    RSS    AIC
## - rain       1     0.000 541.67 225.99
## - temp       1     0.025 541.70 226.00
## - FFMC:ISI   1     0.113 541.79 226.04
## - DMC:ISI    1     0.267 541.94 226.12
## - DC:ISI     1     1.389 543.06 226.68
## - FFMC:DC    1     3.688 545.36 227.82
## <none>                   541.67 227.99
## - DMC:DC     1     4.892 546.57 228.41
## - I(RH^2)    1     5.103 546.78 228.52
## - RH         1     5.185 546.86 228.56
## - I(wind^2)  1     8.259 549.93 230.07
## - FFMC:DMC   1     8.325 550.00 230.10
## - wind       1     9.212 550.89 230.54
## - season     3    35.233 576.91 239.00
##
## Step:  AIC=225.99
## log(area) ~ season + FFMC + DMC + DC + ISI + temp + RH + wind +
##     I(wind^2) + I(RH^2) + FFMC:DMC + FFMC:DC + FFMC:ISI + DMC:DC +
##     DMC:ISI + DC:ISI
##
##             Df Sum of Sq    RSS    AIC
## - temp       1     0.026 541.70 224.00
## - FFMC:ISI   1     0.113 541.79 224.04
```

```
## - DMC:ISI    1     0.267 541.94 224.12
## - DC:ISI     1     1.389 543.06 224.68
## - FFMC:DC    1     3.693 545.37 225.82
## <none>                   541.67 225.99
## - DMC:DC     1     4.892 546.57 226.41
## - I(RH^2)    1     5.105 546.78 226.52
## - RH         1     5.214 546.89 226.57
## - I(wind^2)  1     8.259 549.93 228.07
## - FFMC:DMC   1     8.351 550.02 228.12
## - wind       1     9.214 550.89 228.54
## - season     3    35.289 576.96 237.03
##
## Step:  AIC=224
## log(area) ~ season + FFMC + DMC + DC + ISI + RH + wind + I(wind^2) +
##     I(RH^2) + FFMC:DMC + FFMC:DC + FFMC:ISI + DMC:DC + DMC:ISI +
##     DC:ISI
##
##             Df Sum of Sq    RSS    AIC
## - FFMC:ISI   1     0.102 541.80 222.05
## - DMC:ISI    1     0.269 541.97 222.13
## - DC:ISI     1     1.396 543.10 222.69
## - FFMC:DC    1     3.845 545.54 223.91
## <none>                   541.70 224.00
## - DMC:DC     1     4.867 546.57 224.41
## - I(RH^2)    1     5.079 546.78 224.52
## - RH         1     5.317 547.02 224.64
## - I(wind^2)  1     8.462 550.16 226.18
## - FFMC:DMC   1     8.719 550.42 226.31
## - wind       1     9.325 551.02 226.61
## - season     3    38.396 580.10 236.49
##
## Step:  AIC=222.05
## log(area) ~ season + FFMC + DMC + DC + ISI + RH + wind + I(wind^2) +
##     I(RH^2) + FFMC:DMC + FFMC:DC + DMC:DC + DMC:ISI + DC:ISI
##
##             Df Sum of Sq    RSS    AIC
## - DMC:ISI    1     0.236 542.04 220.17
## - DC:ISI     1     1.545 543.35 220.82
## - FFMC:DC    1     3.867 545.67 221.97
## <none>                   541.80 222.05
## - I(RH^2)    1     4.986 546.79 222.52
## - DMC:DC     1     5.124 546.93 222.59
## - RH         1     5.249 547.05 222.65
## - I(wind^2)  1     8.361 550.16 224.18
## - wind       1     9.230 551.03 224.61
## - FFMC:DMC   1    10.819 552.62 225.39
## - season     3    38.902 580.70 234.77
##
## Step:  AIC=220.17
## log(area) ~ season + FFMC + DMC + DC + ISI + RH + wind + I(wind^2) +
##     I(RH^2) + FFMC:DMC + FFMC:DC + DMC:DC + DC:ISI
```

```
##
##             Df Sum of Sq    RSS    AIC
## - DC:ISI     1     3.172 545.21 219.74
## - FFMC:DC    1     3.775 545.81 220.04
## <none>                   542.04 220.17
## - I(RH^2)    1     4.807 546.84 220.55
## - DMC:DC     1     4.942 546.98 220.62
## - RH         1     5.087 547.13 220.69
## - I(wind^2)  1     8.795 550.83 222.51
## - wind       1     9.733 551.77 222.97
## - FFMC:DMC   1    15.072 557.11 225.57
## - season     3    39.000 581.04 232.93
##
## Step:  AIC=219.74
## log(area) ~ season + FFMC + DMC + DC + ISI + RH + wind + I(wind^2) +
##     I(RH^2) + FFMC:DMC + FFMC:DC + DMC:DC
##
##             Df Sum of Sq    RSS    AIC
## <none>                   545.21 219.74
## - DMC:DC     1     4.806 550.02 220.11
## - I(RH^2)    1     4.874 550.08 220.15
## - ISI        1     5.045 550.25 220.23
## - RH         1     5.227 550.44 220.32
## - I(wind^2)  1     6.673 551.88 221.03
## - wind       1     8.016 553.23 221.68
## - FFMC:DC    1    10.908 556.12 223.09
## - FFMC:DMC   1    14.976 560.19 225.06
## - season     3    38.967 584.18 232.38
```

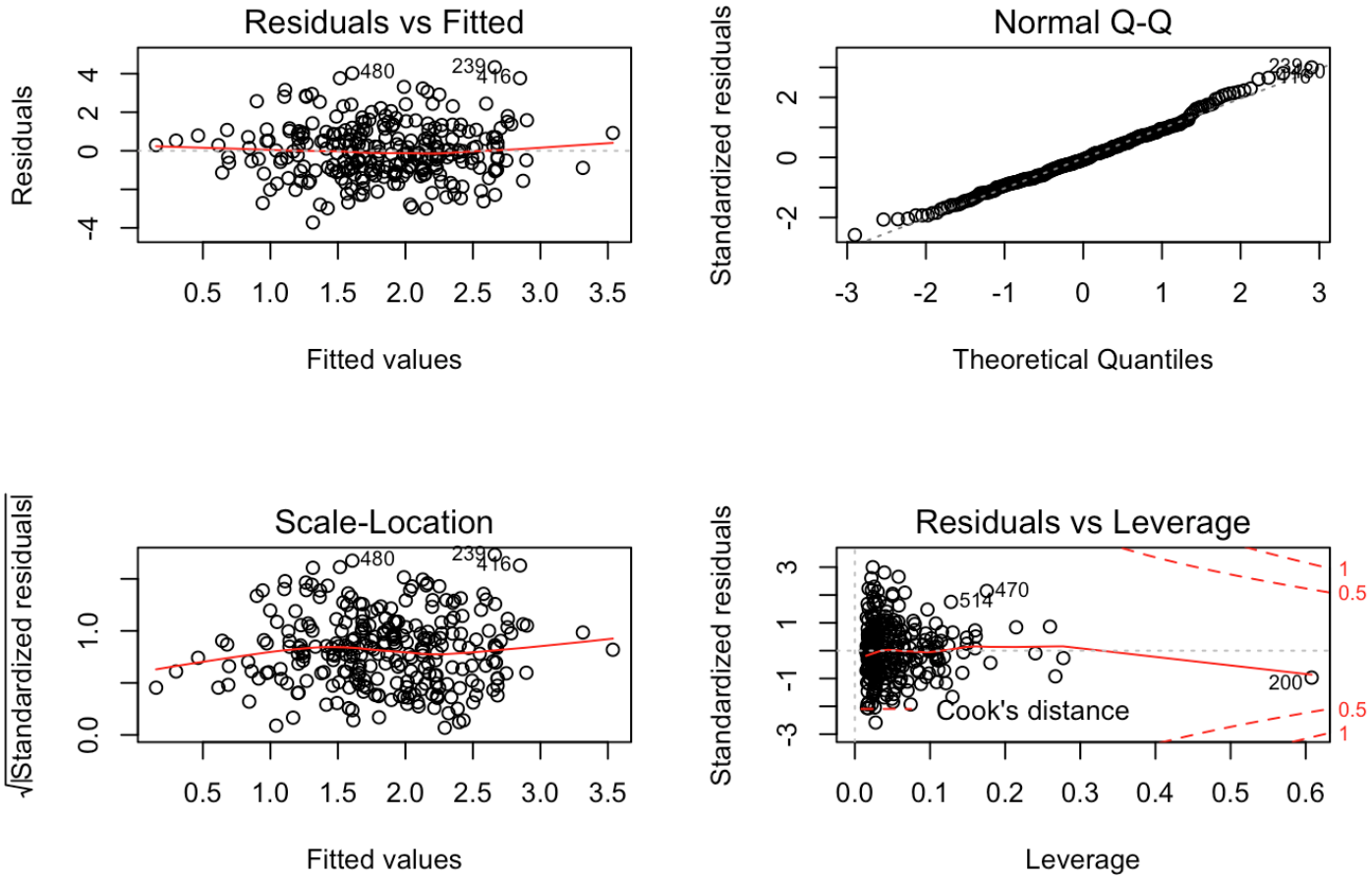
```
summary(reg2)
```

```
##
## Call:
## lm(formula = log(area) ~ season + FFMC + DMC + DC + ISI + RH +
##     wind + I(wind^2) + I(RH^2) + FFMC:DMC + FFMC:DC + DMC:DC,
##     data = forestfire)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7225 -1.0005 -0.1020  0.8899  4.3323
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.185e+00  6.971e+00  -0.457  0.64811
## seasonspring -5.612e-01  8.144e-01  -0.689  0.49142
## seasonsummer -1.312e+00  3.312e-01  -3.962 9.64e-05 ***
## seasonwinter  1.223e-01  8.195e-01   0.149  0.88151
## FFMC          6.721e-02  7.989e-02   0.841  0.40094
## DMC          -1.708e-01  7.568e-02  -2.257  0.02484 *
## DC            3.346e-02  1.481e-02   2.259  0.02470 *
## ISI          -5.370e-02  3.496e-02  -1.536  0.12574
## RH           -5.101e-02  3.263e-02  -1.563  0.11918
## wind          3.863e-01  1.995e-01   1.936  0.05393 .
## I(wind^2)    -3.641e-02  2.061e-02  -1.767  0.07848 .
## I(RH^2)       4.911e-04  3.252e-04   1.510  0.13230
## FFMC:DMC      2.127e-03  8.038e-04   2.647  0.00864 **
## FFMC:DC      -3.834e-04  1.697e-04  -2.259  0.02475 *
## DMC:DC       -2.006e-05  1.338e-05  -1.499  0.13505
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.462 on 255 degrees of freedom
## Multiple R-squared:  0.1304, Adjusted R-squared:  0.08271
## F-statistic: 2.732 on 14 and 255 DF,  p-value: 0.0008879
```

We obtained a “optiaml” model with the help of backward selection based on AIC. We want to proceed to find the significant predictors and understand how the values of these predictors are related to the areas of the fires (p-value, confidence intervals, t-statistics). But before doing that, let us first do some model checking.

Check the model.

```
par(mfrow=c(2,2))
plot(reg2)
```




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

There is no obvious violation of model assumptions so we are good to go.

Interpret the coefficients and commenting on the p-vals and the confidence interval. Will leave this part for you guys!