Statistical Learning Project: Forest Fires

Adriano Rasetta, Alessandro Manente, Giuliano Squarcina

June 13, 2020

1 Data

The dataset contains forest fire data from the Montesinho natural park, in the northeast region of Portugal. The data was collected from January 2000 to December 2003 and it was built using two sources. The first database was collected by the inspector that was responsible for the Montesinho fire occurrences. At a daily basis, every time a forest fire occurred, several features were registered. The second database was collected by the Braganca Polytechnic Institute, containing several weather observations that were recorded with a 30 minute period by a meteorological station located in the center of the Montesinho park.
This dataset is available at: http://www.dsi.uminho.pt/~pcortez/forestfires/.
The dataset contains:

- spatial location data (X, Y) within a $9x9$ grid
- temporal data (*month*, *day*)
- fire danger indexes (*FFMC*, *DMC*, *DC*, *ISI*)
- meteorological data

The following table explains each variable in detail:

Table 1: The preprocessed dataset attributes

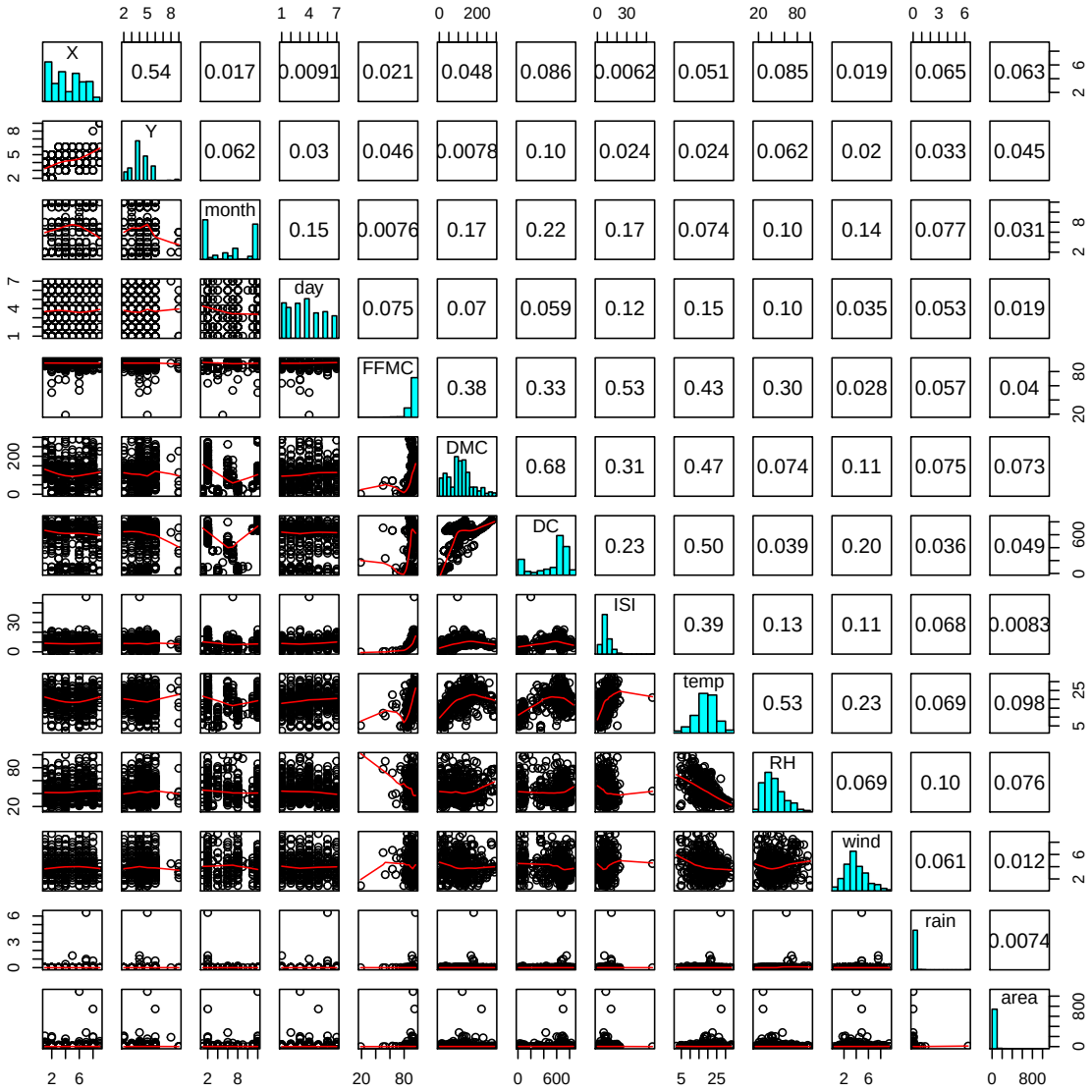
X	x-axis coordinate (from 1 to 9) of the $9x9$ grid in which the park area has been divided
Y	y-axis coordinate (from 1 to 9) of the $9x9$ grid in which the park area has been divided
month	Month when fire occurred: in some months fires are more frequent
day	Day of the week when fire occurred: in some days fires are more frequent due to human activites (e.g. work days vs weekend)
FFMC	Fine Fuel Moisture Code: denotes the moisture content surface litter and influences ignition and fire spread
DMC	Duff Moisture Code: the moisture content of shallow organic layers
DC	Drought Code: the moisture content of deep organic layers
ISI	Score indicating the fire velocity spread
temp	Outside temperature (in C°)
RH	Outside relative humidity (in %)
wind	Outside wind speed (in km/h)
rain	Outside rain (in mm/m^2)
area	Total burned area (in ha): the target variable we would like to predict

2 Data exploration

First of all we check the presence of `NA`:

```
##     X     Y month   day  FFMC   DMC    DC   ISI  temp    RH  wind  rain  area
##     0     0     0     0     0     0     0     0     0     0     0     0     0
```

The dataset is complete. To grasp a raw idea about the data we are dealing with, we use a *pair plot*, since the number of variables is still manageable:



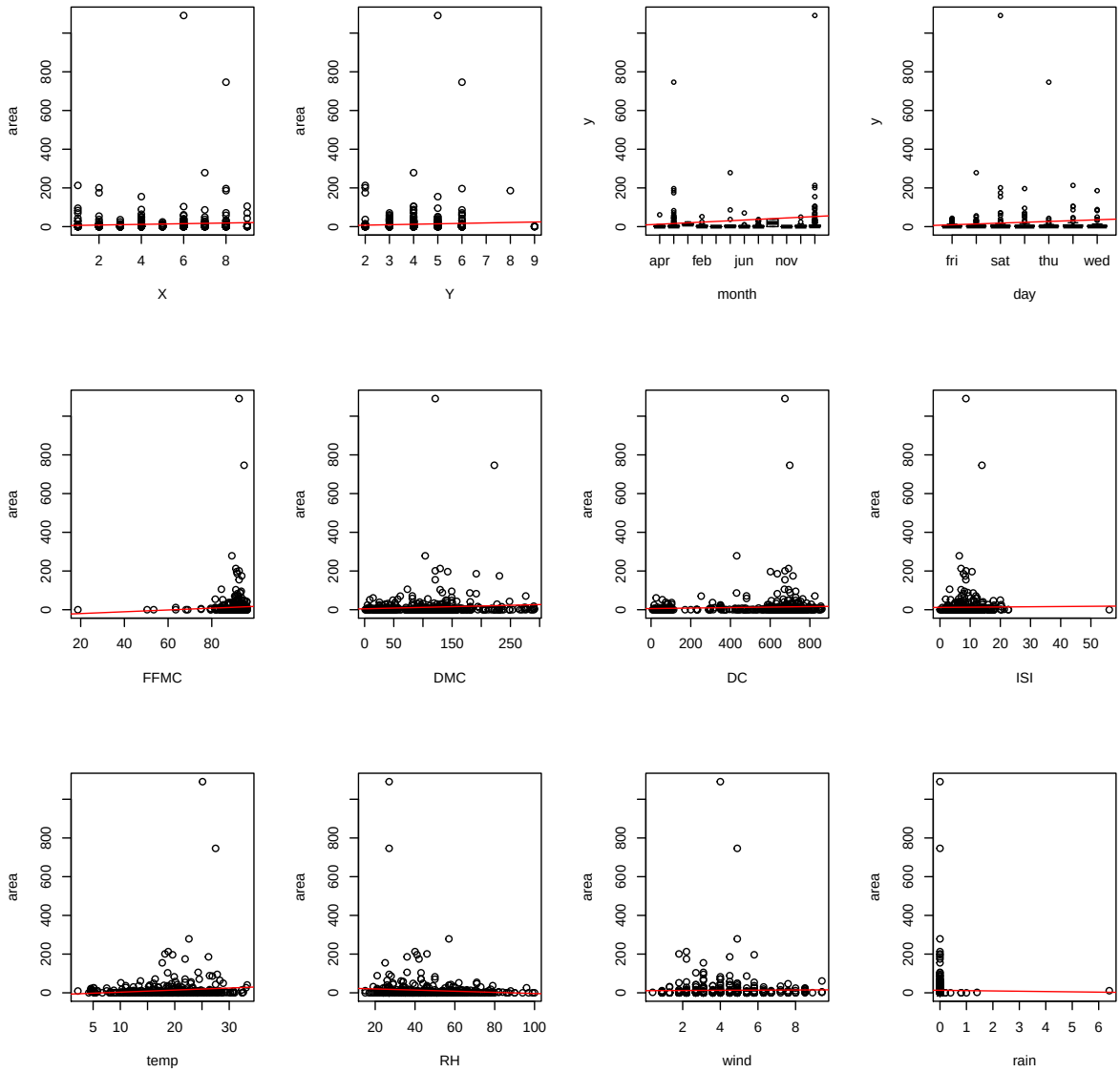
Analyzing the scatterplot we observe that:

- many scatter plots of (quantitative) regressors show a circular cluster of points, that implies a weak correlation between many pairs of regressors
- some others exhibit an almost 0 correlation (points are streched along an horizontal or vertical line); i.e. *wind* vs *rain*
- only few of them exhibit a high correlation, but anyway never above 0.68; i.e. *DMC* vs *DC*

The above observations can be quantized computing the correlation matrix:

```
##             X     Y  FFMC   DMC    DC   ISI  temp    RH  wind rain
## X        1.00  0.54 -0.02 -0.05 -0.09  0.01 -0.05  0.09  0.02 0.07
## Y        0.54  1.00 -0.05  0.01 -0.10 -0.02 -0.02  0.06 -0.02 0.03
## FFMC    -0.02 -0.05  1.00  0.38  0.33  0.53  0.43 -0.30 -0.03 0.06
## DMC     -0.05  0.01  0.38  1.00  0.68  0.31  0.47  0.07 -0.11 0.07
## DC      -0.09 -0.10  0.33  0.68  1.00  0.23  0.50 -0.04 -0.20 0.04
## ISI      0.01 -0.02  0.53  0.31  0.23  1.00  0.39 -0.13  0.11 0.07
## temp    -0.05 -0.02  0.43  0.47  0.50  0.39  1.00 -0.53 -0.23 0.07
## RH       0.09  0.06 -0.30  0.07 -0.04 -0.13 -0.53  1.00  0.07 0.10
## wind     0.02 -0.02 -0.03 -0.11 -0.20  0.11 -0.23  0.07  1.00 0.06
## rain     0.07  0.03  0.06  0.07  0.04  0.07  0.07  0.10  0.06 1.00
```

This excludes a multicollinearity problem. It's worth focus on the relation between regressors and the response *area* (last line of previous *pair plot*):



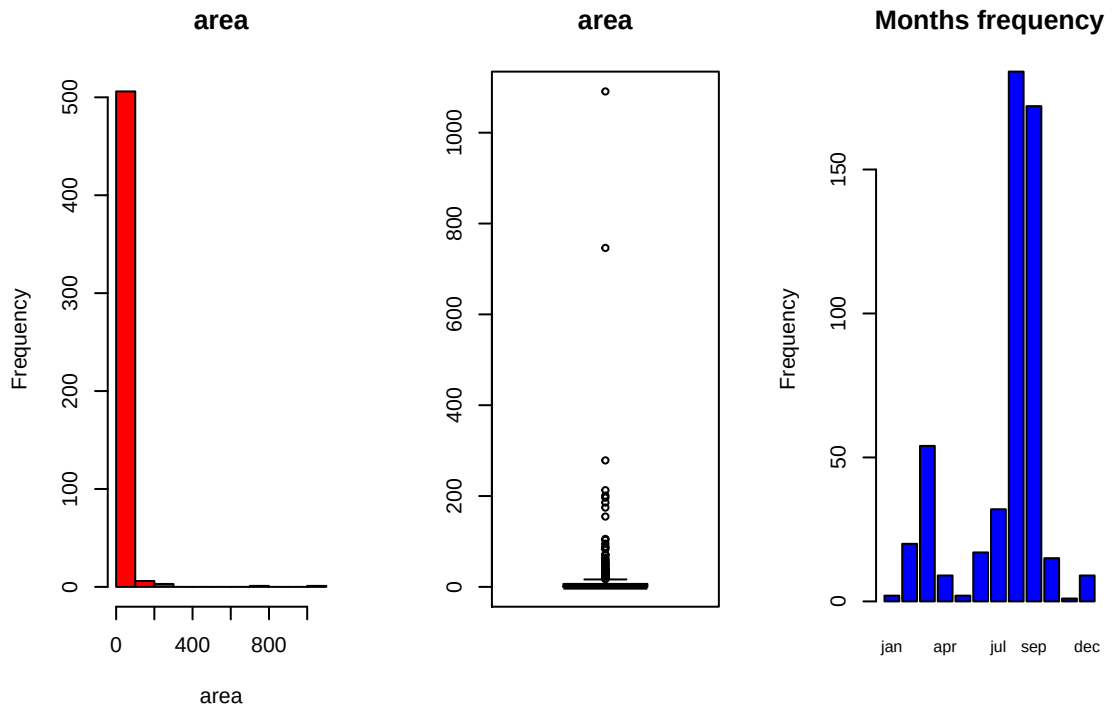
What appears is an almost flat cluster of points in each scatterplot, which means that as anyone of the regressor changes, the value of *area* is almost not affected. When all regressors are going to be considered simultaneously in the multiple regression setting, we expect almost all of them will remain not significant, defining a difficult regression problem.
Numerical summaries for each variable are also computed:

```
##        X               Y           month      day          FFMC
##  Min.   :1.000   Min.   :2.0   aug    :184   fri:85   Min.   :18.70
##  1st Qu.:3.000   1st Qu.:4.0   sep    :172   mon:74   1st Qu.:90.20
##  Median :4.000   Median :4.0   mar    : 54   sat:84   Median :91.60
##  Mean   :4.669   Mean   :4.3   jul    : 32   sun:95   Mean   :90.64
##  3rd Qu.:7.000   3rd Qu.:5.0   feb    : 20   thu:61   3rd Qu.:92.90
##  Max.   :9.000   Max.   :9.0   jun    : 17   tue:64   Max.   :96.20
##                                (Other): 38   wed:54
##       DMC              DC             ISI              temp
##  Min.   :  1.1   Min.   :  7.9   Min.   : 0.000   Min.   : 2.20
##  1st Qu.: 68.6   1st Qu.:437.7   1st Qu.: 6.500   1st Qu.:15.50
##  Median :108.3   Median :664.2   Median : 8.400   Median :19.30
##  Mean   :110.9   Mean   :547.9   Mean   : 9.022   Mean   :18.89
##  3rd Qu.:142.4   3rd Qu.:713.9   3rd Qu.:10.800   3rd Qu.:22.80
##  Max.   :291.3   Max.   :860.6   Max.   :56.100   Max.   :33.30
##
##        RH              wind            rain              area
##  Min.   : 15.00   Min.   :0.400   Min.   :0.00000   Min.   :   0.00
##  1st Qu.: 33.00   1st Qu.:2.700   1st Qu.:0.00000   1st Qu.:   0.00
##  Median : 42.00   Median :4.000   Median :0.00000   Median :   0.52
##  Mean   : 44.29   Mean   :4.018   Mean   :0.02166   Mean   :  12.85
##  3rd Qu.: 53.00   3rd Qu.:4.900   3rd Qu.:0.00000   3rd Qu.:   6.57
##  Max.   :100.00   Max.   :9.400   Max.   :6.40000   Max.   :1090.84
##
```

Note that:

- *area* has a very low mean compared with the $max(area)$. So its distribution appears right-skewed.
- *month* has categories with heterogeneuous frequencies

Analyzing these variables more in detail:

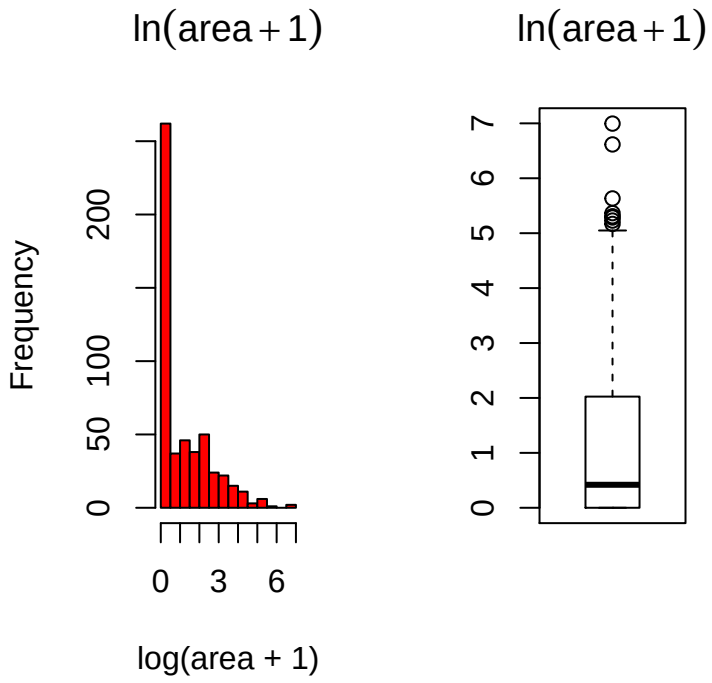


area is actually right-skewed and *month* has high heterogeneity in its class frequencies. To solve the first

issue, it's advisible to apply the following transformation to *area*:

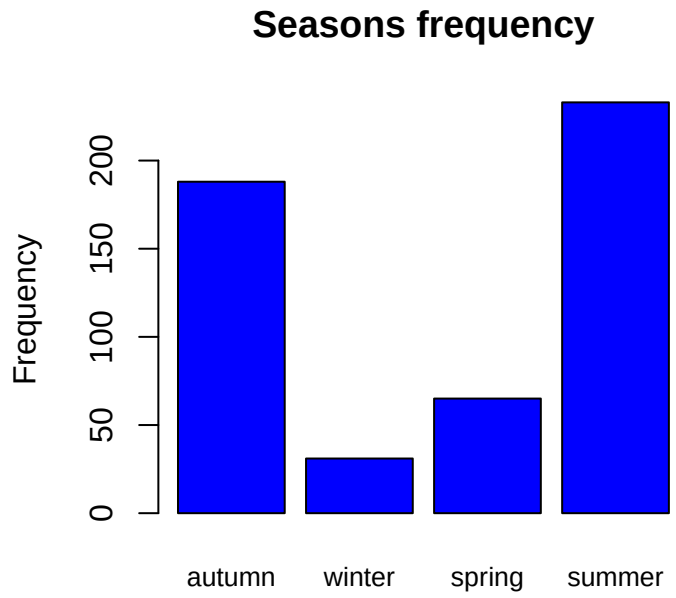
$$\ln(area + 1)$$

in this way the right skewed is reduced and the $+1$ prevents to have $\ln 0$, which is not defined.



To solve the second issue, the 12 categories in *month* are grouped in 4 categories, correspondig with the 4 seasons. The new categorical variable is called *season*.

```
## autumn winter spring summer
##    188     31     65    233
```



The 4 categories in *season* are less unbalanced than 12 categories in *month*: a regression model can learn more from few categories with higher frequency than many categories with low frequency. Moreover, in the perspective to use Cross-Validation to get more robust results, some categories could appear just in the training set and not in the validation set, creating a problem of unseen categories in the test set. It's also useful to reduce the number of of categories in *day* from 7 to 2, splitting the days of the week in "working day" and "week-end".

```
## working_day     weekend
##         338         179
```

Given tha data, the relevant question is:
Is it possible to predict the *ha* of forest burned on a given region in a given time interval, provided:

- **spatial location data**
- **temporal data**
- **fire danger indexes**
- **meteorological data?**

Such knowledge is particularly useful for improving firefighting resource management (e.g. prioritizing targets for air tankers and ground crews), reducing forest distruction, economical and ecological damage and preserving humans' life.
We are going to tackle such question looking for a model capable to provide a satisfactory answer.

3 Model Selection

We start estimating the full model, containing all regressors available. The aim is to check the significance of the full regression:

```
##
## Call:
## lm(formula = log(area + 1) ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8578 -1.0720 -0.5343  0.8908  5.3464
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.364e+00  1.558e+00  -0.876   0.3816
## X             4.337e-02  3.158e-02   1.373   0.1703
## Y             7.813e-03  5.961e-02   0.131   0.8958
## seasonwinter  7.329e-01  5.029e-01   1.457   0.1457
## seasonspring -1.360e-01  4.903e-01  -0.277   0.7816
## seasonsummer -3.633e-01  1.971e-01  -1.843   0.0659 .
## dayweekend    1.230e-01  1.307e-01   0.941   0.3469
## FFMC          1.592e-02  1.479e-02   1.076   0.2823
## DMC           2.462e-03  1.675e-03   1.470   0.1422
## DC            1.496e-05  7.428e-04   0.020   0.9839
## ISI          -1.667e-02  1.717e-02  -0.971   0.3322
## temp          2.662e-02  2.057e-02   1.294   0.1963
## RH           -1.812e-03  5.675e-03  -0.319   0.7497
## wind          8.370e-02  3.673e-02   2.279   0.0231 *
## rain          5.278e-02  2.125e-01   0.248   0.8039
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.383 on 502 degrees of freedom
## Multiple R-squared:  0.04799,Adjusted R-squared:  0.02144
## F-statistic: 1.808 on 14 and 502 DF,  p-value: 0.03473
```

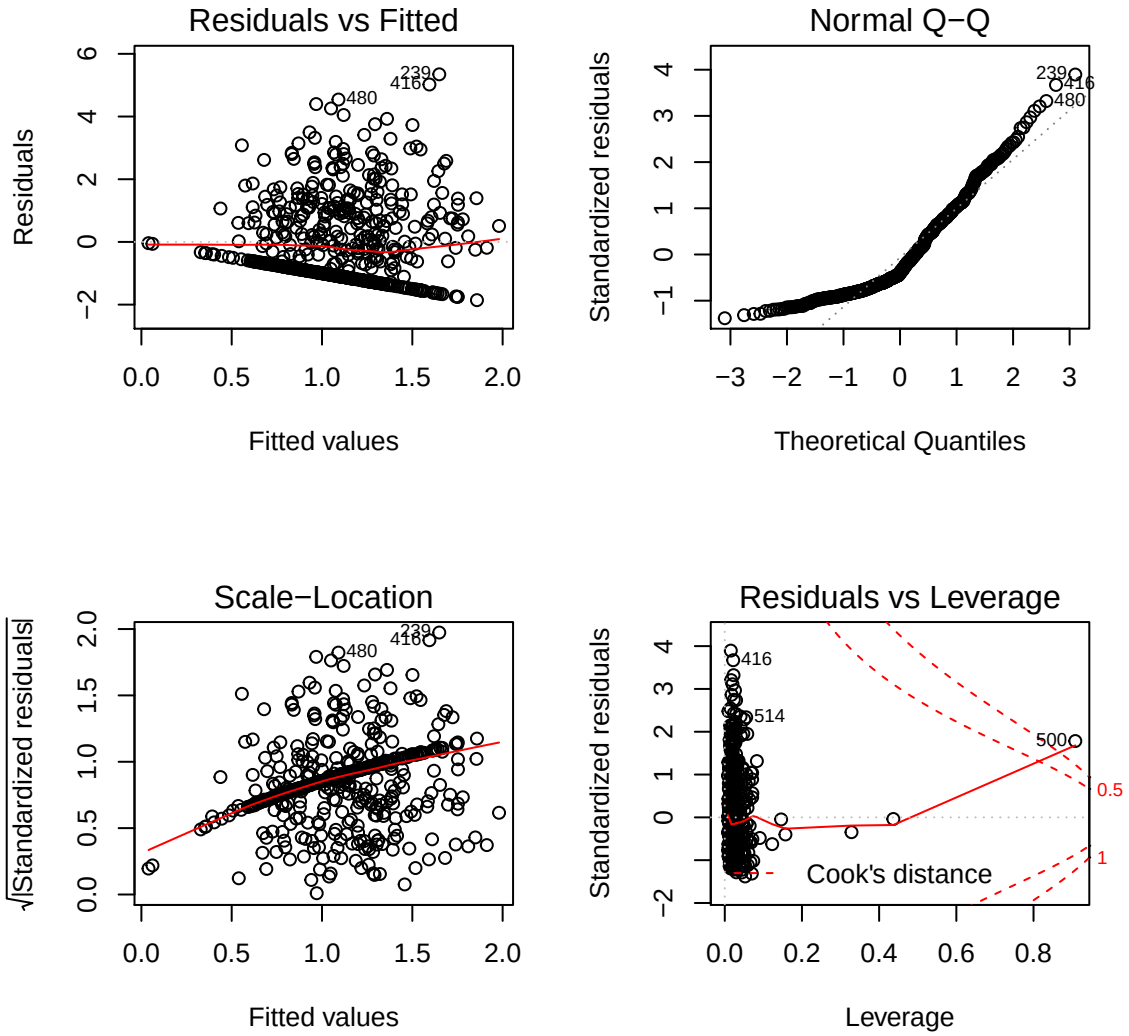
All coefficients but one are not significant at 5% significance level, as guessed in *Data Exploration* section. The unique significant coefficient at 5% significance level is *wind*. Also the *Adjusted R-squared* is low if compared to toy datasets, but is in line with many other regressions on real-world data. The *p-value* on *F-statistic* shows a mild evidence about the significance of the full regression.
Nevertheless, the regression model could be useful to make predicions. Such assumption can be tested in a simple way: compare the performance of the *naive prediction avg(area)* with the $\hat{y}$ estimated from the full model. If the prediction error is significantly smaller for $\hat{y}$, than it's meaningful to go on in the analysis of the regression model (i.e. variable selection, adding non-linearity, etc.). Using the *Root Mean Squared Error(RMSE)* the results are:

RMSE	64.34
RMSE naive	63.59

so it does not seem meaningful to carry on any further statistical analysis. Anyway the *RMSE* is sensitive to outliers, and our analysis highlighted their presence, due to the right-skewed distribution of *area*. Hence it is worth to use an *error* less sensitive to outliers: *Mean Absolute Error (MAE)*, which leads to:

MAE	12.87
MAE naive	18.57

so MAE is 30.66% smaller than $MAE\ naive$. The improvement is not negligible. It is meaningful to use the model for prediction: this implies that our first concern is to increase the *predictive* capability of the model, even at cost to reduce its interpretability.
We start analyzing the residuals of the *full model* estimated previously:



Residuals vs Fitted:

- there appears to be little pattern in the residuals, suggesting that the there is a straight-line relationship between the predictors and the response.
- there is a mild evidence of non-constant variances in the errors (heteroscedasticity) from the presence of a funnel shape in the residual plot: one possible solution is to transform the response Y using a concave function such as $\ln(Y)$ or $\sqrt{Y}$. Anyway we already *log-transform* the response variable *area*, and trying further concave transformation does not lead to sensitive improvements.

Normal Q-Q:

- residuals have a left-tail distribution lighter than Normal and a right-tail distribution heavier than Normal (right-skewed)

Scale-Location:

- the presence of heteroskedasticity can be seen also from how the *studentized residuals* spread along the ranges of $\hat{y}$. Ideally the red line should be horizontal.

Residuals vs Leverage:

- observations whose *studentized residuals* are greater than 3 in absolute value are possible outliers
- observation 500, whose *Leverage* is greater than *Cook's distance* (outside the red dashed line) is a *leverage point*. The regression results will be altered if we exclude this observation.

Excluding observation 500, the regression result is:

```
##
## Call:
## lm(formula = log(area + 1) ~ ., data = data[-500, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9128 -1.0542 -0.5016  0.8934  5.3553
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.740e+00  1.569e+00  -1.109   0.2678
## X             4.624e-02  3.156e-02   1.465   0.1434
## Y             6.433e-03  5.949e-02   0.108   0.9139
## seasonwinter  8.127e-01  5.038e-01   1.613   0.1073
## seasonspring -5.405e-02  4.914e-01  -0.110   0.9125
## seasonsummer -3.437e-01  1.970e-01  -1.744   0.0817 .
## dayweekend    1.132e-01  1.305e-01   0.867   0.3861
## FFMC          1.725e-02  1.478e-02   1.168   0.2435
## DMC           2.339e-03  1.673e-03   1.398   0.1626
## DC            1.203e-04  7.435e-04   0.162   0.8715
## ISI          -1.721e-02  1.714e-02  -1.004   0.3159
## temp          3.094e-02  2.067e-02   1.497   0.1350
## RH            6.651e-05  5.760e-03   0.012   0.9908
## wind          9.172e-02  3.692e-02   2.484   0.0133 *
## rain         -1.118e+00  6.890e-01  -1.623   0.1051
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.38 on 501 degrees of freedom
## Multiple R-squared:  0.05228,Adjusted R-squared:  0.0258
## F-statistic: 1.974 on 14 and 501 DF,  p-value: 0.018
```

now *rain* and *RH* have changed sign, but are still not signifcant. The *Adjusted R-square* is increased and also the significance of regression (lower *p-value* of *F-statistic*) is increased. We are interested in checking if this can increase the predictive power of the model:

MAE	12.87
$MAE_{-leveragePoint}$	12.88

The *MAEs* are the same, so no real improvement in performance from excluding the leverage point. Anyway is advisible to keep it out of dataset.

A possible way to increase the *predictive* power of the model is to extend it with some interactions. It's reasonable to expect that the effect of *DMC* on *area* also depends by *wind*. By words: the amount of bursted area as the moisture content of shallow organic layers increases, is not independent by the wind, which positively contributes to increase bursted area.

```
##
## Call:
```

```
## lm(formula = log(area + 1) ~ . + wind:DMC, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8153 -1.0676 -0.4924  0.9030  5.3370
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.065e+00  1.588e+00  -1.300   0.1942
## X             4.467e-02  3.156e-02   1.415   0.1576
## Y             8.173e-03  5.947e-02   0.137   0.8907
## seasonwinter  7.393e-01  5.068e-01   1.459   0.1452
## seasonspring -1.397e-01  4.957e-01  -0.282   0.7781
## seasonsummer -3.798e-01  1.989e-01  -1.909   0.0568 .
## dayweekend    1.385e-01  1.319e-01   1.050   0.2943
## FFMC          1.679e-02  1.477e-02   1.137   0.2562
## DMC           5.217e-03  2.813e-03   1.854   0.0643 .
## DC            2.145e-05  7.471e-04   0.029   0.9771
## ISI          -1.687e-02  1.713e-02  -0.985   0.3253
## temp          3.433e-02  2.083e-02   1.648   0.1000 .
## RH            1.666e-03  5.892e-03   0.283   0.7775
## wind          1.638e-01  6.761e-02   2.422   0.0158 *
## rain         -1.016e+00  6.932e-01  -1.466   0.1434
## DMC:wind     -7.018e-04  5.518e-04  -1.272   0.2040
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.379 on 500 degrees of freedom
## Multiple R-squared:  0.05534,Adjusted R-squared:  0.027
## F-statistic: 1.953 on 15 and 500 DF,  p-value: 0.0169
```

The *wind*DMC* coefficient is not significant and the $MAE_{interactionModel}$ =12.85, which is a not meaningful reduction. Further attempts to extend the model result to be not worthful, so we keep it as it is. Tested that is not possible to extend easily the model to get lower *MAE*, it is interesting to check if is possible to maintain a comparable *MAE* using fewer regressors, which could be beneficial for interpretability too.

3.1 Backward Selection

Using *backward selection* is possible to get the following model:

```
## Start:  AIC=347.43
## log(area + 1) ~ X + Y + season + day + FFMC + DMC + DC + ISI +
##     temp + RH + wind + rain
##
##          Df Sum of Sq    RSS    AIC
## - RH      1    0.0003 954.59 345.43
## - Y       1    0.0223 954.61 345.44
## - DC      1    0.0499 954.64 345.46
## - day     1    1.4339 956.02 346.20
## - ISI     1    1.9207 956.51 346.47
## - FFMC    1    2.5979 957.19 346.83
## <none>                954.59 347.43
## - DMC     1    3.7261 958.31 347.44
## - X       1    4.0914 958.68 347.64
## - temp    1    4.2695 958.86 347.73
```

```
## - rain    1    5.0213 959.61 348.14
## - wind    1   11.7598 966.35 351.75
## - season  3   21.2208 975.81 352.77
##
## Step:  AIC=345.43
## log(area + 1) ~ X + Y + season + day + FFMC + DMC + DC + ISI +
##     temp + wind + rain
##
##          Df Sum of Sq    RSS    AIC
## - Y       1    0.0224 954.61 343.44
## - DC      1    0.0498 954.64 343.46
## - day     1    1.4703 956.06 344.22
## - ISI     1    1.9268 956.51 344.47
## - FFMC    1    2.7553 957.34 344.92
## <none>                954.59 345.43
## - DMC     1    4.0358 958.62 345.61
## - X       1    4.0981 958.69 345.64
## - rain    1    5.2919 959.88 346.28
## - temp    1    8.2223 962.81 347.85
## - wind    1   11.8486 966.44 349.79
## - season  3   22.9629 977.55 351.69
##
## Step:  AIC=343.44
## log(area + 1) ~ X + season + day + FFMC + DMC + DC + ISI + temp +
##     wind + rain
##
##          Df Sum of Sq    RSS    AIC
## - DC      1    0.0464 954.66 341.47
## - day     1    1.4690 956.08 342.23
## - ISI     1    1.9346 956.54 342.49
## - FFMC    1    2.7430 957.35 342.92
## <none>                954.61 343.44
## - DMC     1    4.1421 958.75 343.68
## - rain    1    5.3000 959.91 344.30
## - X       1    6.2348 960.84 344.80
## - temp    1    8.2692 962.88 345.89
## - wind    1   11.8262 966.44 347.79
## - season  3   23.0493 977.66 349.75
##
## Step:  AIC=341.47
## log(area + 1) ~ X + season + day + FFMC + DMC + ISI + temp +
##     wind + rain
##
##          Df Sum of Sq    RSS    AIC
## - day     1    1.4497 956.11 340.25
## - ISI     1    2.0018 956.66 340.55
## - FFMC    1    2.7857 957.44 340.97
## <none>                954.66 341.47
## - rain    1    5.2568 959.91 342.30
## - X       1    6.1889 960.85 342.80
## - DMC     1    7.0208 961.68 343.25
## - temp    1    8.2705 962.93 343.92
## - wind    1   11.9800 966.64 345.90
## - season  3   24.4521 979.11 348.52
##
## Step:  AIC=340.25
```

```
## log(area + 1) ~ X + season + FFMC + DMC + ISI + temp + wind +
##     rain
##
##          Df Sum of Sq    RSS    AIC
## - ISI     1    2.0336 958.14 339.35
## - FFMC    1    2.4221 958.53 339.55
## <none>                956.11 340.25
## - rain    1    5.4331 961.54 341.17
## - X       1    6.1834 962.29 341.58
## - DMC     1    7.1907 963.30 342.12
## - temp    1    8.7143 964.82 342.93
## - wind    1   11.8687 967.98 344.62
## - season  3   24.0126 980.12 347.05
##
## Step:  AIC=339.35
## log(area + 1) ~ X + season + FFMC + DMC + temp + wind + rain
##
##          Df Sum of Sq    RSS    AIC
## - FFMC    1    1.1320 959.27 337.95
## <none>                958.14 339.35
## - rain    1    5.3965 963.54 340.24
## - X       1    6.0471 964.19 340.59
## - DMC     1    7.6587 965.80 341.45
## - temp    1    7.9426 966.08 341.61
## - wind    1   10.4423 968.58 342.94
## - season  3   27.4409 985.58 347.92
##
## Step:  AIC=337.95
## log(area + 1) ~ X + season + DMC + temp + wind + rain
##
##          Df Sum of Sq    RSS    AIC
## <none>                959.27 337.95
## - rain    1    5.3326 964.60 338.82
## - X       1    6.0684 965.34 339.21
## - DMC     1    9.0107 968.28 340.78
## - temp    1    9.8268 969.10 341.21
## - wind    1   11.0539 970.33 341.87
## - season  3   26.3976 985.67 345.96
```

Interpretation:
each row contains info about the model *without (-)* the regressor indicated at the beginning of the row. If the model without a given regressor has a particolar high RSS or AIC, then the excluded regressor is important, and its exclusion worsen a lot the model. By analogy, makes sense to exclude the regressor that is less useful in explaning the response, that is, the regressor which exclusion leads to the model with the lowest RSS (or AIC). The model selected is:

$$area = \beta_0 + \beta_1 rain + \beta_2 X + \beta_3 DMC + \beta_4 temp + \beta_5 wind + \beta_6 season$$

3.2 Forward Selection

Using *forward selection* is possible to get the following model:

```
## Start:  AIC=347.14
## log(area + 1) ~ 1
##
##          Df Sum of Sq     RSS    AIC
## + season  3   12.6125  994.64 346.64
```

```
## + wind    1    4.4036 1002.85 346.88
## + DC      1    4.3211 1002.93 346.92
## + DMC     1    4.2845 1002.97 346.94
## <none>                1007.25 347.14
## + X       1    3.6514 1003.60 347.27
## + RH      1    3.1529 1004.10 347.52
## + rain    1    3.0950 1004.16 347.55
## + temp    1    2.6100 1004.64 347.80
## + FFMC    1    2.0413 1005.21 348.09
## + Y       1    1.4400 1005.81 348.40
## + day     1    1.1124 1006.14 348.57
## + ISI     1    0.1589 1007.09 349.06
##
## Step:  AIC=346.64
## log(area + 1) ~ season
##
##        Df Sum of Sq    RSS    AIC
## + DMC   1    8.8364 985.80 344.03
## + temp  1    7.4545 987.18 344.76
## + wind  1    7.0305 987.61 344.98
## + FFMC  1    5.4167 989.22 345.82
## + X     1    4.1377 990.50 346.49
## + RH    1    3.9164 990.72 346.60
## <none>               994.64 346.64
## + DC    1    3.3763 991.26 346.88
## + rain  1    2.4675 992.17 347.36
## + Y     1    1.8886 992.75 347.66
## + day   1    1.3892 993.25 347.92
## + ISI   1    0.1103 994.53 348.58
##
## Step:  AIC=344.03
## log(area + 1) ~ season + DMC
##
##        Df Sum of Sq    RSS    AIC
## + wind  1    6.7833 979.02 342.47
## + temp  1    6.6556 979.15 342.54
## + RH    1    5.4267 980.38 343.18
## + X     1    4.6919 981.11 343.57
## <none>               985.80 344.03
## + FFMC  1    3.2670 982.54 344.32
## + rain  1    3.1344 982.67 344.39
## + Y     1    1.5255 984.28 345.23
## + day   1    1.3510 984.45 345.32
## + ISI   1    0.0558 985.75 346.00
## + DC    1    0.0095 985.79 346.03
##
## Step:  AIC=342.47
## log(area + 1) ~ season + DMC + wind
##
##        Df Sum of Sq    RSS    AIC
## + temp  1    9.1484 969.87 339.63
## + RH    1    6.1258 972.89 341.23
## + X     1    4.6396 974.38 342.02
## + rain  1    4.5296 974.49 342.08
## <none>               979.02 342.47
## + FFMC  1    2.8692 976.15 342.96
```

```
## + Y     1    1.8656 977.15 343.49
## + day   1    1.5704 977.45 343.64
## + ISI   1    0.0544 978.97 344.44
## + DC    1    0.0049 979.01 344.47
##
## Step:  AIC=339.63
## log(area + 1) ~ season + DMC + wind + temp
##
##        Df Sum of Sq    RSS    AIC
## + X     1    5.2665 964.60 338.82
## + rain  1    4.5307 965.34 339.21
## <none>              969.87 339.63
## + Y     1    1.8644 968.01 340.63
## + day   1    1.3279 968.54 340.92
## + FFMC  1    1.0923 968.78 341.04
## + ISI   1    0.6687 969.20 341.27
## + RH    1    0.2610 969.61 341.49
## + DC    1    0.0043 969.87 341.62
##
## Step:  AIC=338.82
## log(area + 1) ~ season + DMC + wind + temp + X
##
##        Df Sum of Sq    RSS    AIC
## + rain  1    5.3326 959.27 337.95
## <none>              964.60 338.82
## + day   1    1.3491 963.26 340.09
## + FFMC  1    1.0681 963.54 340.24
## + ISI   1    0.7448 963.86 340.42
## + RH    1    0.4698 964.13 340.56
## + DC    1    0.0211 964.58 340.80
## + Y     1    0.0209 964.58 340.80
##
## Step:  AIC=337.95
## log(area + 1) ~ season + DMC + wind + temp + X + rain
##
##        Df Sum of Sq    RSS    AIC
## <none>              959.27 337.95
## + day   1   1.18413 958.09 339.32
## + FFMC  1   1.13198 958.14 339.35
## + ISI   1   0.74342 958.53 339.55
## + DC    1   0.09549 959.18 339.90
## + RH    1   0.04570 959.23 339.93
## + Y     1   0.01214 959.26 339.95
```

Interpretation:
each row contains info about the model *with (+)* the regressor indicated at the beginning of the row. If the model with a given regressor has a particolar low RSS or AIC, then the included regressor is important, and its inclusion improve a lot the model. By analogy, makes sense to include the regressor that is more useful in explaning the response, that is, the regressor which inclusion leads to the model with the lowest RSS (or AIC). The model selected is:

$$area = \beta_0 + \beta_1 rain + \beta_2 X + \beta_3 DMC + \beta_4 temp + \beta_5 wind + \beta_6 season$$

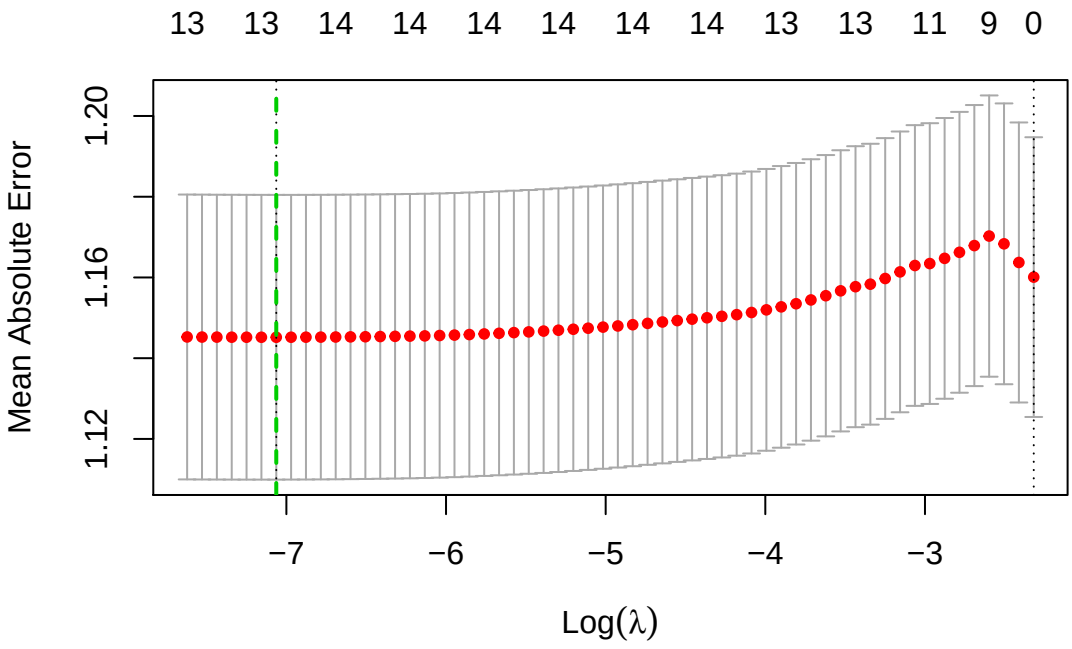
This is the same model selected by *backward selection*.

3.3 Parameters Shrinkage: Lasso Regression

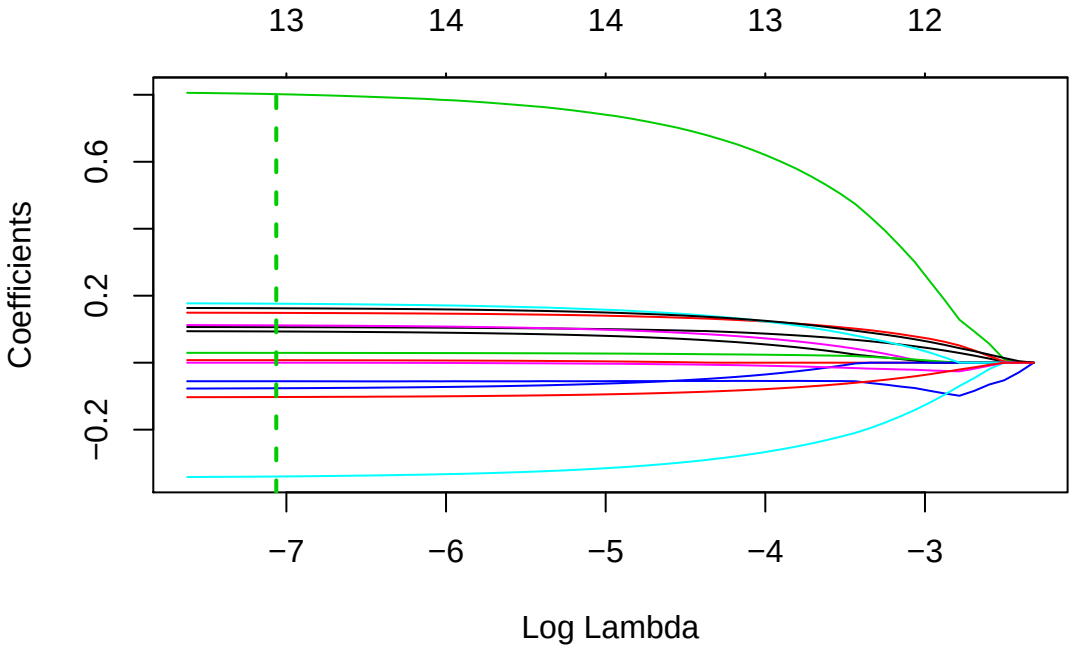
We try to perform variables selection using *Lasso Regression*. The *lasso* implicitly assumes that a number of the coefficients truly equal zero, so, in general, it should perform better in a setting where a relatively small number of predictors have substantial coefficients. Here we hope to identify a subset of regressors similar to the *backward selection*.
The regressors, except the categorical ones, are standardize, to neutralize the effect of different measurements unit on the shrinkage procedure. Then a *lasso* regression is performed, using *LOOCV* to find the optimal lambda (λ^*), which minimizes the error in the training set.
The estimated coefficents using λ^* are:

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                            1
## (Intercept)   1.181354960
## X             0.106121597
## Y             0.007427218
## seasonwinter  0.801939926
## seasonspring -0.055363676
## seasonsummer -0.339887048
## dayweekend    0.111417122
## FFMC          0.093192706
## DMC           0.148692274
## DC            0.029296260
## ISI          -0.076358375
## temp          0.176206797
## RH            .
## wind          0.162561367
## rain         -0.102532660
```

with λ^* =8.559 $\times 10^{-4}$. When $\lambda^* \approx 0$, the *lasso regression* is approximately equivalent to an *OLS regression*, so there is no particolar reason to use it. This explains why only *RH* variable is shrunk by *lasso*. So it does not seem to be an appropriate technique to be used in this dataset.
We used some standard plot to display better the situation:



As $\log(\lambda)$ increases, more regressors are pushed to zero, but the MAE in the *validation set* tends to increase monotonically until a peak is reached. Then there is a slightly decrease, but anyway the $min(MAE)$ is reached for λ^* found before.
The same information can be shown using the following plot:



So *lasso* regression keeps all variables but one (RH), which do not satisfy our aim. The *lasso model* is essentialy identic to the *full model*, impling similar performances:

MAE_{lasso}	12.88
MAE_{full}	12.87

Hence we use the model selected by *backward* and *forward* selection.

4 Results

Given the analysis of the previous section, this is the model we are going to (possibly) interpret and test:

$$area = \beta_0 + \beta_1 rain + \beta_2 X + \beta_3 DMC + \beta_4 temp + \beta_5 wind + \beta_6 season$$

The selected model leads to the following coefficients:

```
##
## Call:
## lm(formula = log(area + 1) ~ season + DMC + wind + temp + X +
##     rain, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8451 -1.0420 -0.5242  0.8792  5.4412
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.228914   0.386209  -0.593  0.55363
## seasonwinter  0.742239   0.334549   2.219  0.02695 *
## seasonspring -0.083895   0.242049  -0.347  0.72903
## seasonsummer -0.392401   0.145069  -2.705  0.00706 **
## DMC           0.002791   0.001279   2.182  0.02955 *
## wind          0.086356   0.035727   2.417  0.01600 *
## temp          0.032535   0.014276   2.279  0.02308 *
## X             0.047221   0.026367   1.791  0.07391 .
## rain         -1.117507   0.665651  -1.679  0.09380 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.376 on 507 degrees of freedom
## Multiple R-squared:  0.04763,Adjusted R-squared:  0.03261
## F-statistic:  3.17 on 8 and 507 DF,  p-value: 0.001622
```

of which only one is strongly not significant (*seasonspring*) and five are significant well below the 5% significance level. Compared to the initial *full model*, the reduction of categories in *month* and *day* and the exclusion of the *high leverage poing*, seem to be effective.
Even if our primary task is *prediction* and not *inference*, the result obtained allows us to try an interpretation of coefficients:

- $\beta_{\mathbf{X}}$: very mild evidence that as we move toward Eastern part of the park ($=X$ increases), *area* increases. This depends by the way ($9x9$ grid) used to divide the area of the park. This kind of information it's specific to the "Montesinho natural park", so not generalizable to other parks. For this reason we could think in a further refinement of the model, to drop it.

- $\beta_{\mathbf{seasonwinter}}$: the *constrast* category is "Autumn", so it represents the *ha* of burned forest due to the fact that we are in winter rather than in autumn. It is signficant, but its sign is not intuitive at first glace (why should be higher the burned area in Winter with respect to the Autumn, considering the in Winter there are less favourable conditions to fire propagations?). It could exhist a variable, like *tempestivity of intervention*, which is negatively related to *area* (higher *tempestivity* $\implies$ lower *area*), positively related to Summer (high amount of resources in terms of crews and equipments is invested in Summer because the risk of fires is high) and negatively related to Winter. So, on average, when a fire happens in Winter, it takes more time to be spotted and to be estinguished.

- $\beta_{\mathbf{seasonspring}}$: strongly not signifcant. No difference with respect to the constrast category Autumn.

- $\beta_{\mathbf{seasonsummer}}$: strongly significant. Specular interpretation provided for $\beta_{\mathbf{seasonwinter}}$

- $\beta_{\mathbf{DMC}}$: significant. As the moisture content of shallow organic layers increases, the *ha* of forest burned by a $\approx 0.3\%$. This is not the expected sign, but the magnitude is very small, maybe could have a scientific explanation.
- $\beta_{\mathbf{temp}}$: significant. Higher temperature foster fire propagation.
- $\beta_{\mathbf{wind}}$: significant. *wind* plays an important role: it increases by $\approx 9\%$ the *ha* of burned forest.
- $\beta_{\mathbf{rain}}$: mild signficant, due to a high Std. Error in its estimation. Anyway is out of dout that it play the major role in preventing fire propagation. An increase of $1mm/m^2$ in the amount of rain decreases the $1.11ha$ the area of burned forest by

To assess the robustness of the selected model, we use *Leave One Out Cross-Validation (LOOCV)*: data are shuffled and divided in n folds, with each one, in turn, used as *test set* and the remaning $n-1$ used as *training set*. Then the $mean(MAE)$ over the n folds is computed:

$MAE_{backward\,CV}$	12.95
$MAE_{full\,CV}$	12.98

The model selected appears to be robust to different training/test set splits and to perform better in the *test set* than the *full model*: the $MAE_{backward\,CV}$ is 0.2% lower than the $MAE_{full\,CV}$.
However, economical and practical considerations suggest to look for a variables selection approach based on data's cost and availability in cluster (i.e. the 4 *meteorological data* are available together and are cheaper than *fire danger indexes*). Hence we try to use only the four *meteorological data*, because:

- are cheap and easy to get
- *backward* selection already selected 3 of them (*rain*, *temp*, *wind*)

The estimated model is therefore:

$$area = \beta_0 + \beta_1 temp + \beta_2 RH + \beta_3 wind + \beta_4 rain$$

and we test it using $LOOCV$:

$MAE_{meteo\,CV}$	13
$MAE_{full\,CV}$	12.98

The $MAE_{meteo} = 13$, which is only 0.21% higher than the *full model* and slightly higher than the $MAE_{backward\,CV}$. So *meteorological data* are a good compromise between *prediction power* and *feasibility* of data.

Of major importance for the study, given the frequency of small fires ($< 1\ ha$) is to understand how our models perform with this kind of data. To understand it, to the *full model*, the *backward stepwise selection model* and the *meteo model* were asked to predict the *area* of samples with a value equal to 0 in the column *area*.
The results are the following:

$MAE0_{step\,CV}$	2.17
$MAE0_{meteo\,CV}$	2.07
$MAE0_{full\,CV}$	2.19

5 Conclusion

It is exceedingly difficult to manage a fire once it breaks out, so the main objective must be prevention. It requires the capacity to:

- spot fastly a new fire
- individuate areas with high probability to generate fires

To accomplish this task, there are four available options:

- *patrol the park*: can detect smaller fires but requires a lot of human resources
- *satellite infrared smoke scanners*: data are costly and with low spatial resolution
- *aerial monitoring through drones*: costly
- *local sensors*: data are cheap and with high spatial resolution

To achieve the goal we set in section 2, we chose an economical and efficient solution: *local sensors*. The results of our analysis produced two models: one derived from backward selection and the other only from meteorological data. Both cases provide a test error close to the complete model and indicate that error is on average 13 hectares: if we have a fire of 15 hectares, the model on average could foresee a burned area of 13 hectares larger or smaller of the real burned area. So, we have a range of $2 - 28$ hectares of surface.
A further point in favor of the two selected model is the excellent ability to identify small fires: by selecting the subset of the initial data that have a burned area below the hectare, both models provide very accurate predictions. The average error is only two hectares. This allows to calibrate the amount of resources (men and tools) proportionally to the fire's size:

- *small fires*: use park patrol, in fact if the exact position of the small fire is unknown, it's impossible to intervene
- *large fires*: use air tankers and ground crews

Possible future improvements to the current research could focus on the following question:
what is the role of vegetation? What are the fire prevention strategies currently in force? What are the times of reaction to emergencies? Could it be useful to add data on altitude, slope and exposure to wind and sun? Moreover, considering the main cause of the fires, the human being, it could be especially useful to know some information about the human incidence in the park: what are the areas with the most tourists and inhabitants? In conclusion, there are many ways to improve the efficiency of the model, but what matters most is that the approach to data can make a difference and allow better management of emergencies.

6 Technical Appendix

6.1 ERRORS

Two measures were used to compare the performances of the tested models: root-mean-square error (RMSE) and mean absolute error (MAE). In both metrics, lower values result in better predictive models.

$$RMSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x_i})^2 \tag{1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{x_i}| \tag{2}$$

In our particular case, we will take the MAE into greater consideration, as it is less sensitive to outliers, thus providing a more stable indicator.

6.2 METHOD FOR EVALUATING AND COMPARING MODELS

Akaike's information criterion (AIC) was used to compare models with different number of predictors. This method makes a mathematical adjustment to the training error rate in order to estimate the test error rate.

$$AIC = -2l(\hat{\theta}) + 2d \tag{3}$$

In the case of linear model with Gaussian errors:

$$AIC = -2l(\hat{\beta}, \hat{\theta}) + 2d = n\log\left(\frac{RSS}{n}\right) + 2d \tag{4}$$

We use it in the backward selection. We preferred the AIC to the BIC (Bayesian information criterion) to have a selection method capable of storing models with a large number of variables.

6.3 BACKWARD STEPWISE SELECTION

To select the model we used the Backward Stepwise Selection:

Input: A full model with p predictors, M_p. *Output:* Single model with best performance among the ones tested.

1. Let M_p be the full model with p predictors.
2. for $k = p, p-1, \ldots, 1$:
   - Consider all k models that contain all but one predictor in M_k.
   - Choose the one having smallest RSS or highest R^2, it is M_{k-1}.
3. Chose a model among the best models selected, $M_0, \ldots, M_p$, through a given criterion (cross-validated predicted error, C_p, AIC, BIC or $adjustedR^2$).

It is a heuristic research strategy that provides locally optimal solutions that approach an optimal solution globally in a reasonable amount of time. We use it as a cheaper strategy to find a satisfactory model.

6.4 FORWARD STEPWISE SELECTION

To select the model we used the Forward Stepwise Selection:

Input: A null model with 0 predictors, M_0.
Output: Single model with best performance among the ones tested.

1. Let M_0 be the null model with p predictors.
2. for $k = 0, 1, \ldots, p-1$:
   - Consider all $p-k$ models that augment by one predictor the model M_k.
   - Choose the one having smallest RSS or highest R^2, it is M_{k+1}.
3. Chose a model among the best models selected, $M_0, \ldots, M_p$, through a given criterion (cross-validated predicted error, C_p, AIC, BIC or $adjusted R^2$).

It is a heuristic research strategy that provides locally optimal solutions that approach an optimal solution globally in a reasonable amount of time. We use it as a cheaper strategy to find a satisfactory model.

6.5 ESTIMATE OF THE TEST ERROR

To estimate the test error associated with the model, we used the Leave-one-out cross validation (LOOCV):

1. Splitting the set of observations into two parts: a single observation (x_1, y_1) is used for the validation set, and the remaining observations make up the training set;
2. the statistical learning method is fitted on the $n-1$ training observations;
3. the prediction $\hat{y_i}$ is made for the excluded observation;
4. if Y is continuous, $MSE_i = (y_i - \hat{y_i})^2$ provides an estimate for the test error (approximately unbiased but poor);
5. repeat the procedure for every $i = 1, ..., n$ thereby obtaining n estimates of the test MSE.
6. The LOOCV estimate of the test MSE is the average of the n estimates,

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} RMSE_i \tag{5}$$

Although the LOOCV is computationally very expensive, the small size of the dataset has allowed to exploit its potential thus obtaining an estimate of the test error with a lower bias than the more common K-fold cross-validation.

6.6 LASSO

6.6.1 General Problem

The lasso is a shrinkage method that, thanks to the 1−norm on the coefficients β_j acts in a non linear manner, as can be seen below:

$$\min_{\beta} \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \tag{6a}$$

$$\text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s, \tag{6b}$$

It's equivalent Lagrangian Form is:

$$\min_{\beta} \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{7}$$

6.6.2 Pathwise Coordinate Descent (Multivariate Case)

Fixed the penalization λ we optimize w.r.t. each β_j, keeping the $\beta_k,\ k \neq j$ fixed, in this way:

- Assume all regressors are standardized with zero mean and unit norm
- Let $\tilde{\beta_k}$ be the estimate of β_k for the parameter λ
- rearrange the lagrangian form to isolate β_j

$$R(\tilde{\beta}(\lambda), \beta_j) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{k\neq j}\tilde{\beta_k}x_{ik} - \beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\tilde{\beta}(\lambda) + \lambda|\beta_j| \tag{8}$$

- It has an explicit solution, giving the update

$$\tilde{\beta}(\lambda) \longleftarrow S\left(\sum_{i=1}^{n} x_{ij}(y_i - \tilde{y_i}^j), \lambda\right) \tag{9}$$

  Where $S(t, \lambda) = sign(t)(|t| - \lambda)_+$ is the '$soft - thresholding$' operator

Cycling through each variable in turn yields the lasso estimate $\tilde{\beta}$.