STATISTICAL APPROACHES TO ANALYZE FOREST FIRES

FALL 2017

PURDUE UNIVERSITY

Table of Contents

Data Background

Forest fires affects forest preservation, create economical and ecological lost, and cause human suffering. Therefore, it is necessary to forecast the forest fire for a successful firefighting. Weather conditions such as temperature, rain, wind and air humidity are known to affect fire occurrence. There are several ways to detect fire such as satellite-based, smoke scanners and local sensors. However, the first two methods have high equipment and maintenance cost. The last one, local sensors such as meteorological stations, can easily record the environmental data in real-time with low costs. In our project, we downloaded the forest fire data from UCI machine learning repository (http://archive.ics.uci.edu/ml/datasets/Forest+Fires). This data provides the environmental conditions of forest fire in the northeast region of Portugal, by using meteorological and other data such as coordinates of the fire, dates and months of the fire. Our goal is to build a regression model to analyze the factors affects the forest fire. To achieve this goal, we set two separate goals for this analysis report: a) analyzing the factors influencing the occurrence of forest fires, and b) analyzing the factors influencing the amount of area burned in the forest fires.

Firstly, let's take a closer look at the data. The data set includes 517 instances without missing value. Our goal is to build a regression model to analyze the factors affects the forest fire. Thus, the factor *area* (in *ha*) is considered to be the response variable in this study. The possible attributes include location and date variables (*X, Y, month* and *day*), the Fire Index factors (*FFMC, DMC, DC* and *ISI*), and the weather factors (*temp, RH, wind* and *rain*).

Location factors, *X* and *Y*, are referring to the coordinates of the observation area on the map of the target forest. Specifically, the map is divided into 81 segments (see Figure 1), indicating 1-9 on the x-axis and 1-9 on the y-axis. However, since the coordinates are categorical factors and hard to interpret in this case, we decided to exclude these two factors from our models. Also, according to literature review, the number of forest fires occurred in different areas does not seem to depend on their coordinates (see Figure 2). The *month* factor indicates twelve different months all year round (12 levels: January to December), and the *day* factor indicates the day of the week observed (5 levels: Monday to Friday).

The forest Fire Weather Index (FWI) is the Canadian system for rating fire danger: the FFMC denotes the moisture content surface litter and influences ignition and fire spread, while the DMC and DC represent the moisture content of shallow and deep organic layers, which affect fire intensity. The ISI is a score that correlates with fire velocity spread, while BUI represents the amount of available fuel. The FWI index is an indicator of fire intensity and it combines the two previous components. Although different scales are used for each of the FWI elements, high values suggest more severe burning conditions. Also, the fuel moisture codes require a memory (time lag) of past weather conditions: 16 hours for FFMC, 12 days for DMC and 52 days for DC.

The weather factor *temp* recorded the average temperature (in °C) of that area according to the weather report. The factor relative humidity (RH) is the ratio (in %) of the partial pressure of water vapor to the equilibrium vapor pressure of water at a given temperature. *Wind* factor refers to the outside wind speed (in km/h), and the *rain* factor indicates the average amount of precipitation (in mm/m^2).

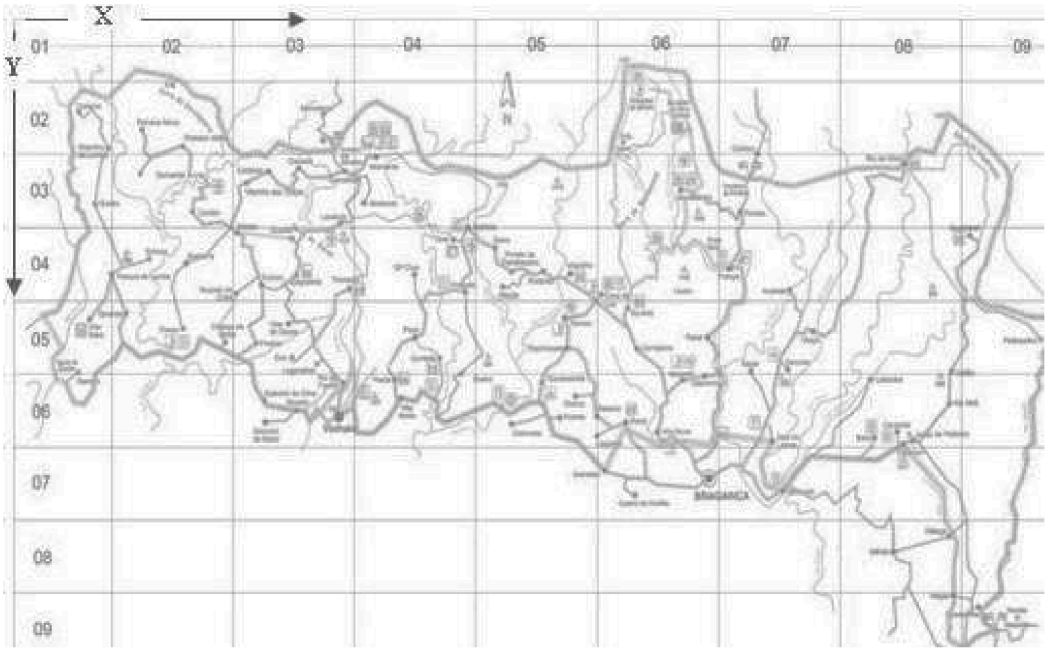


Figure 1: The map of the Montesinho natural park (Cortez and Anibal, 2007)

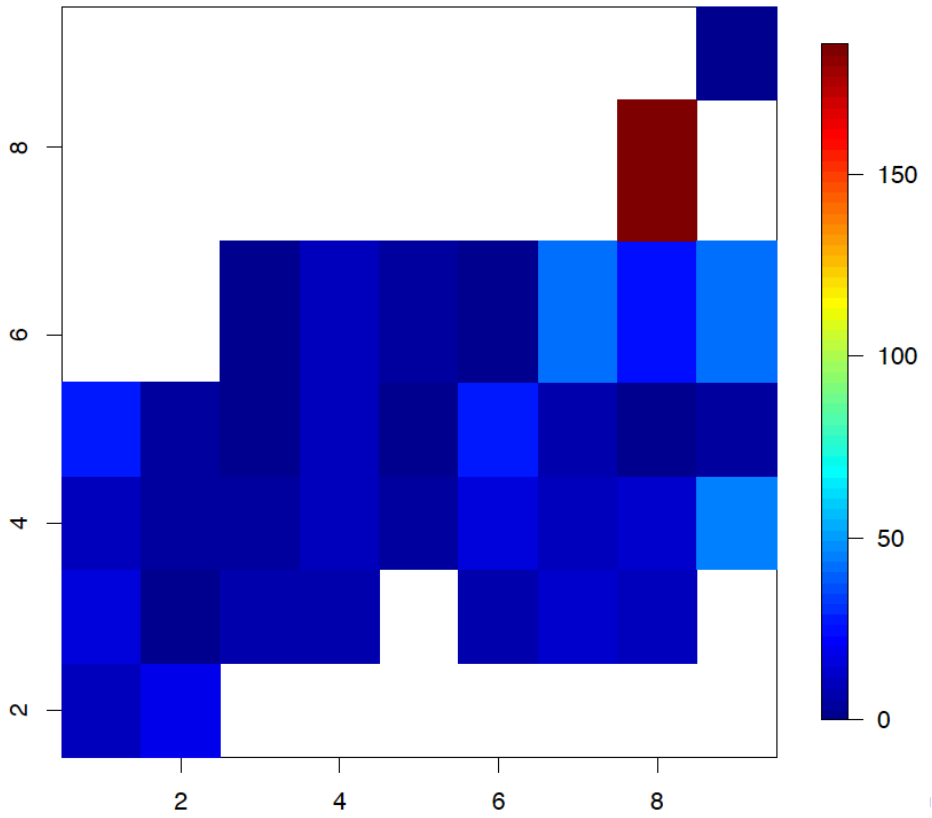
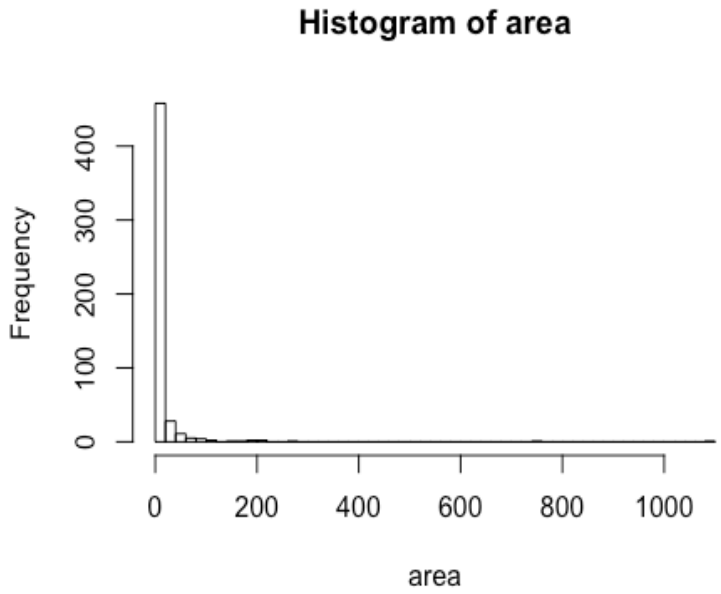


Figure 2: The intensity of fire occurrence based on coordinates (Cortez and Anibal, 2007)

We can see from the histogram below that over 48% of observations are with area equal to zero. This makes sense since there should be a positive probability that no forest fires are triggered at the time of observation. Also, we noticed that there is a clear right skewness and hence we might need to try transformation on response variable later.



Part 1: Factors Influencing Fire Occurrence

1. Logistic Regression Model

By creating a dummy variable *factor_area* with two levels (*factor_area*=1 means the area was burned while *factor_area*=0 means the area wasn't burned), we are able to analysis the probability of the occurrence of fire with a binomial model.

1.1 Model Selection

First, we include all the variables into model and fit it. Because there are data unavailable in some levels of categorical factors month and day, we exclude these two factors from the model. After stepwise selection, we fitted a logistic regression model with two random variables, *DC* and *wind*. However, the drop1 output shows that the factor *wind* can be removed from the model. Thus, the final logistic regression model was as follow:

$$logit(fire) = 0.0007844\, X_{DC} \quad (1)$$

To understand the model better, we can give an example: when DC=500, logit(fire) = 0.3922, and P(fire occurs when DC=500) = 0.5968122. However, based on the output below, we can see that this model's residual deviance is almost as much as the null deviance, which indicates that the model is poor fit based on this data. The Goodness of fit test below shows the same conclusion

that the model is a bad fit. In addition, since the estimate of the overdispersion parameter is close to 1, we ruled out the possibility of overdispersion.

```
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.3405883  0.2154146  -1.581   0.1139
## DC           0.0007844  0.0003582   2.190   0.0285 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 715.69  on 516  degrees of freedom
## Residual deviance: 710.85  on 515  degrees of freedom
## AIC: 714.85

#chi-square test fails to reject the null hypothesis and suggests that
two models are the same, and we can go with the one-factor model instea
d of the two-factor one.

pchisq(deviance(bmod11)-deviance(bmod1),df.residual(bmod11)-df.residual
(bmod1),lower=F)

## [1] 0.07797296

#goodness-of-fit test

pchisq(deviance(bmod11),df.residual(bmod11),lower=F)

## [1] 2.063469e-08

#estimate of the overdispersion parameter

sum(residuals(bmod11,type="pearson")^2)/df.residual(bmod11)

## [1] 1.003841
```

1.2 Model Transformation

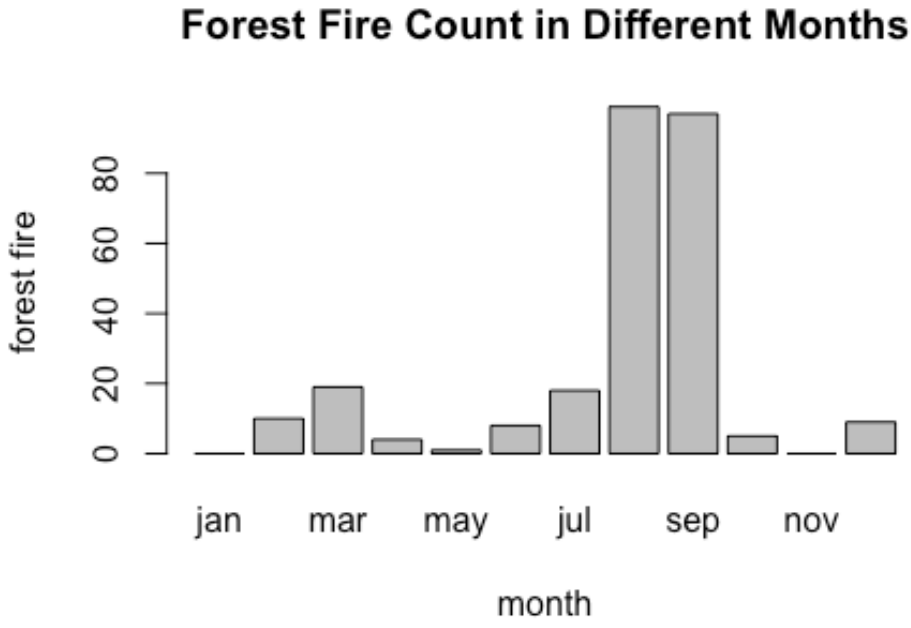
Since the last model doesn't perform well regarding explaining the deviance of the response variable, we performed model transformation to improve the model fit. By adding interaction and quadratic terms in model, we were able to increase the R-square from 0.01 to 0.06. After the stepwise model selection, the final logistic regression model with quadratic terms is as follow:

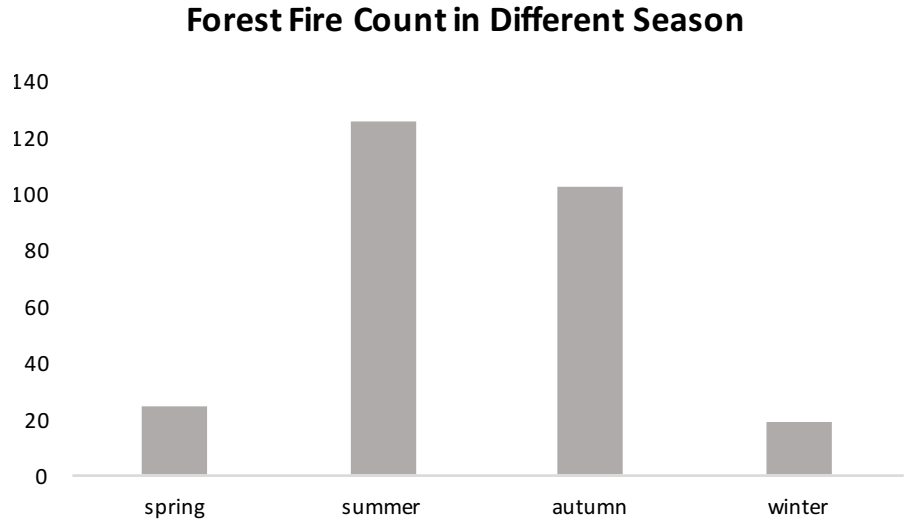
$$logit(fire) = 0.001X_{DC} - 0.297X_{temp} + 0.099X_{Rh} + 0.009X_{temp^2} - 0.001X_{RH^2} + 0.013X_{wind^2} \quad (2)$$

Using the chi-square test to compare model (1) and (2), we reject the null hypothesis and conclude that some of the extra terms in model (2) are significant, so we should choose model (2) instead of model (1). In addition, since the estimate of the overdispersion parameter is close to 1, we ruled out the possibility of overdispersion.

```
## glm(formula = factor_area ~ DC + temp + RH + I(temp^2) + I(RH^2) +
##     I(wind^2), family = binomial(link = logit), data = databi)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7952  -1.1905   0.7504   1.1151   1.6522
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.6785635  1.0204014  -0.665 0.506053
## DC           0.0010874  0.0004692   2.317 0.020477 *
## temp        -0.2968125  0.0882905  -3.362 0.000774 ***
## RH           0.0993346  0.0324213   3.064 0.002185 **
## I(temp^2)    0.0085481  0.0022870   3.738 0.000186 ***
## I(RH^2)     -0.0009940  0.0003135  -3.171 0.001520 **
## I(wind^2)    0.0132654  0.0058412   2.271 0.023146 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 715.69  on 516  degrees of freedom
## Residual deviance: 684.51  on 510  degrees of freedom
## AIC: 698.51
##
## Number of Fisher Scoring iterations: 4

#Chi-square test to compare model (1) and (2)

pchisq(deviance(bmod11)-deviance(bmod21),df.residual(bmod11)-df.residua
l(bmod21),lower=F)

## [1] 7.661332e-05

#estimate of the overdispersion parameter

sum(residuals(bmod21,type="pearson")^2)/df.residual(bmod21)

## [1] 1.010178
```

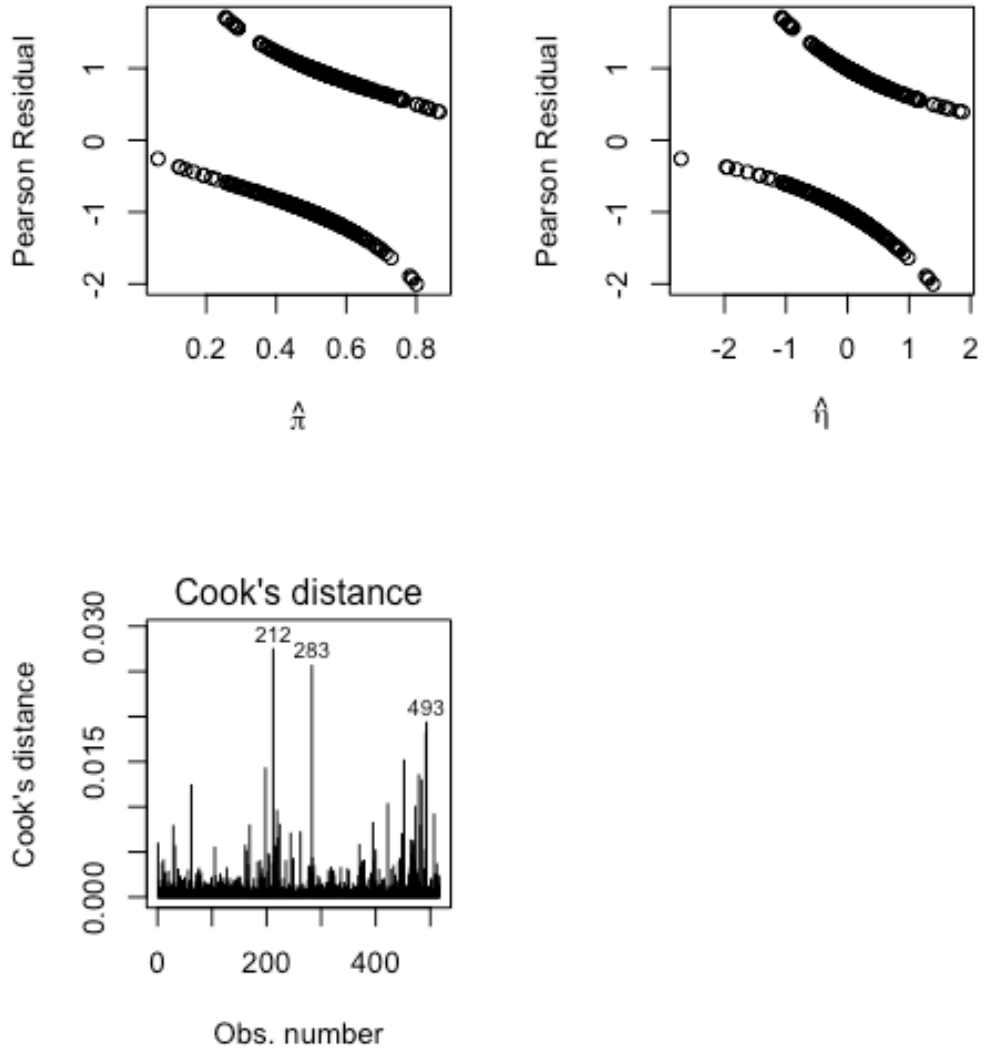
Additionally, we performed another transformation to see if we can get a better fit. In the following figure, we can see that there seems to be a relationship between month and the occurrence of forest fire. However, due to the zeroes in January and November, we couldn't include the factor month in the model. To fit that, we re-classify the 12 months into different seasons and refit model with the season factor. Also, we can see that, there is an obvious pattern of how the number of fires change in four seasons.





However, after stepwise selection, we ended with the exact same model as model (2). So, we conclude that model (2) is the final logistic regression model to analyze the factors influencing the occurrence of forest fire. Next, we need to do model diagnosis.

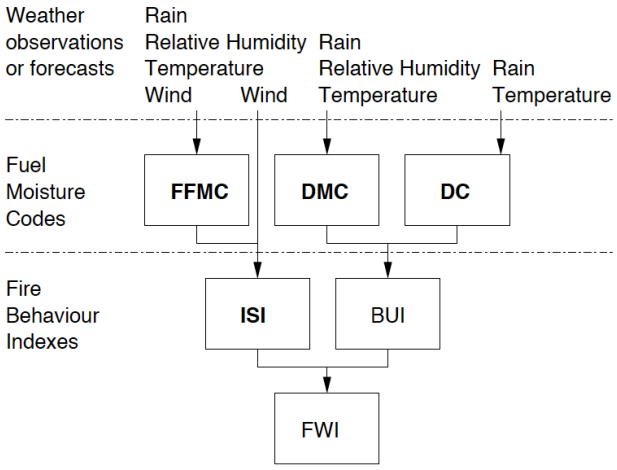
1.3 Model Diagnosis



The plots about show that the model doesn't violate any assumptions. The first plot (top left) shows the distribution of model residuals vs. the probability of two events indicated by response variable (fire occurred and no fire). Since two lines are constant around the origin, the residuals are not showing any obvious patterns. The second plot (top right) is also a residual plot using the logit value as x-axis and the conclusion is the same. The Cook's distance plot shows no obvious outliers in the dataset.

Part 2 Factors influencing fire occurrence in different seasons

2. Multinomial Logit Model



The Fire Weather Index structure shows that wind, temperature and humidity are the most fundamental elements to determine the Fire Weather Index. According to this structure, we decide to use wind temperature and humidity to do the multinomial logit model.

In order to eliminate the effects of zeros, we classify wind, temperature and relative humidity into different levels as factors in the model.

	Wind
<= 3	light
3--6	mid
>= 6	strong

	Temp in Celsius
<= 10	cold
10--20	warm
>= 20	hot

	RH
<= 40	dry
40--60	comfort
>= 60	humid

2.1 Model Selection and Diagnosis

We fit the three weather factors into the multinomial logit model, and use season factors as response variables. After the stepwise selection, there is only one weather factor remains in the model, wind. The result of the refined model is shown below. The LRT shows the full model and reduced model are the same, and the goodness-of-fit test shows that this model is not a good fit.

$$log\frac{\pi_j(x_i)}{\pi_k(x_i)} = log\frac{\pi_j(x_i)}{\pi_1(x_i)} - log\frac{\pi_k(x_i)}{\pi_1(x_i)} = x_i(\beta_j - \beta_k) \quad (3)$$

```
## Call:
## multinom(formula = season ~ cwind, data = datamult)
##
## Coefficients:
##        (Intercept)   cwindlow cwindstrong
## spring  -3.6635600  2.6703111    2.564880
## summer  -0.1670482  0.5347720    0.608908
## winter  -2.2772905 -0.3253825    2.478008
##
## Residual Deviance: 569.9259
## AIC: 587.9259
```

```
## We compare the models before and after selection, and they are the same

pchisq(deviance(multimod1)-deviance(multimod), multimod$edf-multimod1$edf, lo
wer=F)

## [1] 0.480932

## Goodness of fit test

pchisq(deviance(multimod), multimod$edf, lower=F)

## [1] 1.270879e-110
```

The intercepts terms give the probabilities of fire incidence in different seasons when other factors are zero. We can get that $P(\text{fire in autumn}) = 0.506$, $P(\text{fire in spring}) = 0.013$, $P(\text{fire in summer}) = 0.429$, $P(\text{fire in winter}) = 0.052$. The calculations are shown below:

```
cc<- c(0,-3.6635600,-0.1670482,-2.2772905 )
exp(cc)/sum(exp(cc))

## [1] 0.50649283 0.01298702 0.42857338 0.05194678
```

3. Proportional Odds Model

3.1 Check Assumption

Since we think the seasons follow the natural order, and this order may help us to learn how weather factors influencing the fires in forest, we decide to try the proportional odds model to fit the data.

Before fitting the data, we check the proportional odds assumptions for all three of the weather factors. It shows that the temperature factors don't follow proportional odds assumption, but the other two factors seem as constant. Thus, we used humidity and windy factors to fit the proportional odds model.

```
pim<-prop.table(table(databi$ctemp,databi$season),1)
logit(pim[,1])-logit(pim[,1]+pim[,2])
```

```
##       cold        hot       warm
## -2.8622009  0.0000000 -0.8839554

pim<-prop.table(table(datamult$cRH,datamult$season),1)
logit(pim[,1])-logit(pim[,1]+pim[,2])

##    comfort        dry      humid
## -0.4651023 -0.5900750 -0.4390749

pim1<-prop.table(table(datamult$cwind,datamult$season),1)
logit(pim1[,1])-logit(pim1[,1]+pim1[,2])

##       light         low      strong
## -0.2.053422  -0.6143790  -0.9650809
```

3.2 Model Selection and Diagnosis

After fitting the data into proportional odds model, and do the stepwise selection, the model result is shown below.

$$log(\frac{\sum_{j}^{k=1}\pi_k(x)}{\sum_{k=j+1}^{J}\pi_k(x)} = \zeta_j - \eta(x) \tag{4}$$

$$\eta(x) = \beta_1 x_{i1} + \cdots + \beta_{p-1}x_{i,p-1} = x\beta$$

$$\zeta_0 = -\infty < \zeta_1 < \cdots < \zeta_J = \infty$$

The null hypothesis of the LRT is the two models are the same, but based on the Chi-squared test, the p-value is very small, so that we can reject the null hypothesis. The two models are not the same, and the later one (proportional odds model) is better even though it is not a good fit model based on the test p-value.

```
## Call:
## polr(formula = seasons ~ cwind, data = datamult)
##
## Coefficients:
##               Value Std. Error t value
## cwindlow    -0.8071     0.2606  -3.097
## cwindstrong  0.4018     0.4067   0.988
##
## Intercepts:
##               Value   Std. Error t value
## spring|summer -2.8223  0.2911    -9.6957
## summer|autumn -0.2117  0.2114    -1.0013
## autumn|winter  2.2467  0.2882     7.7953
##
## Residual Deviance: 592.225
## AIC: 602.225
```

```r
# Compare this model with the multinomial logit model

pchisq(deviance(propmod1)-deviance(multimod1), multimod1$edf-propmod1$edf, lo
wer=F)

## [1] 0.0001747246

# Goodness-of-fit Test

pchisq(deviance(propmod1), propmod1$edf, lower=F)

## [1] 9.676365e-126
```

Typically, the output from the proportional odds model is easier to interpret. We find out that only low wind and strong wind remain in the model, which is similar to what we got from multinomial logit model, but refine it for two levels of wind.

The coefficients of *cwindlow* shows that the odds of fire happening in spring moving to summer/autumn/winter is increased by a factor of $\exp(-0.8071) = 0.446$ as low wind increase by 1 unit. The coefficents of cwindstrong shows that the odds of fire happening in spring moving to summer/autumn/winter is increased by a factor of $\exp(0.4018) = 1.495$ as strong wind increase by 1 unit.

Based on the result, it shows that strong wind is more dangerous that it will increase the probability of forest fire when season changes from one to the other. So people should pay more attention on the strong wind forecast and reinforce some facilities to prevent it.

Part 3: Factors Influencing Burned Area

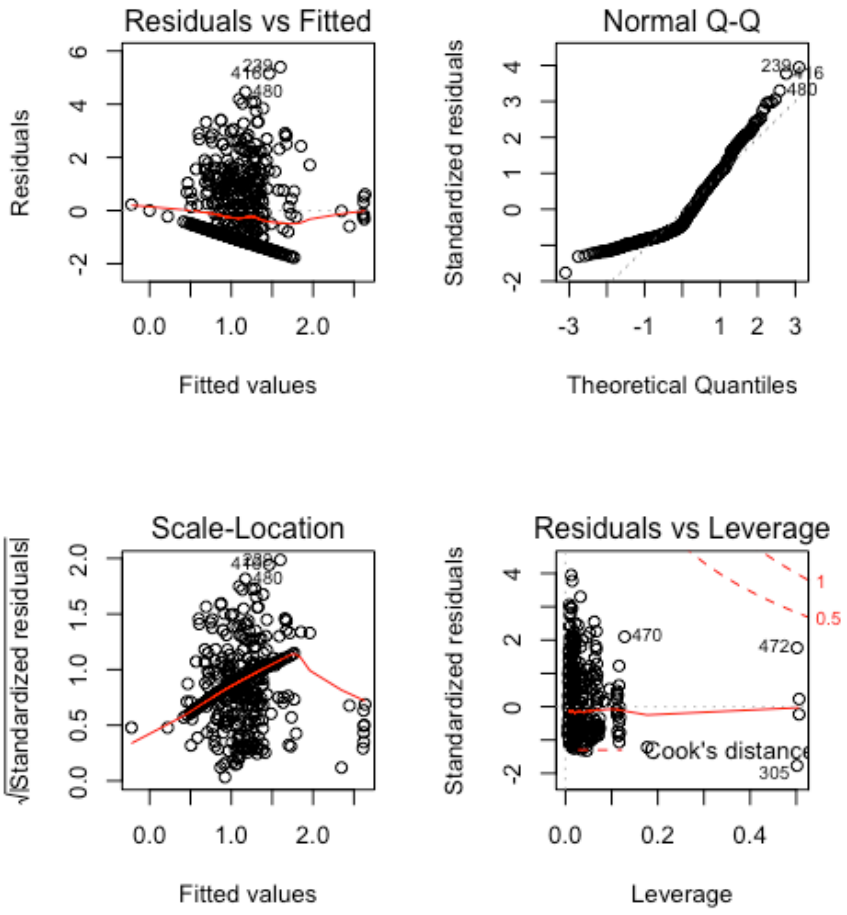
To analysis the factors influencing the amount of area burned in the fire, we took two approached: the first approach is to use all the data in the data set (including the one with zero area), the second approach is to use only the data with positive area value, which is equivalent to analyzing factors influence burned area assuming that the fire has occurred.

Approach 1: With All Data

After including all the factors and stepwise model selection, we only got three significant factors. Most importantly, the model diagnosis shows serious violation to normality and constant variance assumption. In the residual plot, we can clearly see that the data with zero area value are influential to the other positive-area data. In conclusion, we decided to exclude all the data with zero area value and using only positive-area data to analyze the factors influencing the amount of area burned in the fire.

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.461468   0.529446   0.872   0.3838
## monthaug     0.224424   0.778451   0.288   0.7732
## monthdec     2.117783   0.759468   2.789   0.0055 **
```

```
## monthfeb      0.167064   0.554654   0.301   0.7634
## monthjan     -0.581349   1.087122  -0.535   0.5931
## monthjul      0.080785   0.675962   0.120   0.9049
## monthjun     -0.349061   0.628753  -0.555   0.5790
## monthmar     -0.393135   0.496449  -0.792   0.4288
## monthmay      0.713780   1.076777   0.663   0.5077
## monthnov     -0.906345   1.452392  -0.624   0.5329
## monthoct      0.831023   0.948776   0.876   0.3815
## monthsep      0.867144   0.882918   0.982   0.3265
## DMC           0.004186   0.001767   2.369   0.0182 *
## DC           -0.001922   0.001227  -1.566   0.1180
## temp          0.033799   0.014445   2.340   0.0197 *
## wind          0.053007   0.036737   1.443   0.1497
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.375 on 501 degrees of freedom
## Multiple R-squared:  0.06083,    Adjusted R-squared:  0.03271
## F-statistic: 2.163 on 15 and 501 DF,  p-value: 0.006755
```

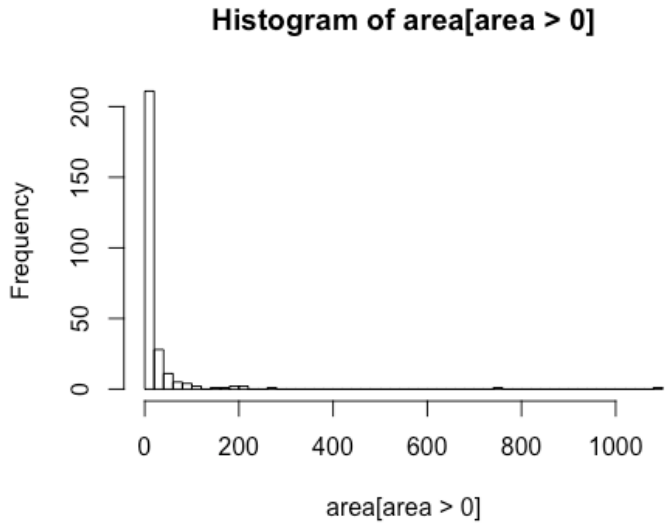


Approach 2: With Only Burned Data

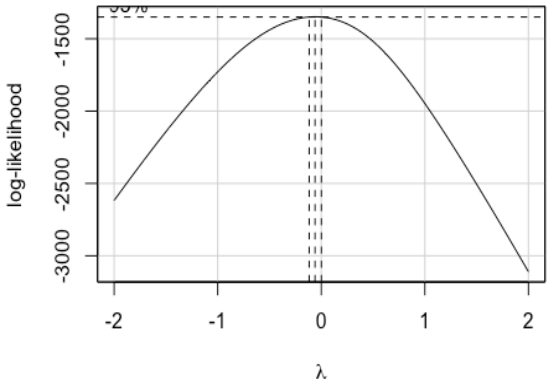
4. Multivariate Linear Regression Model

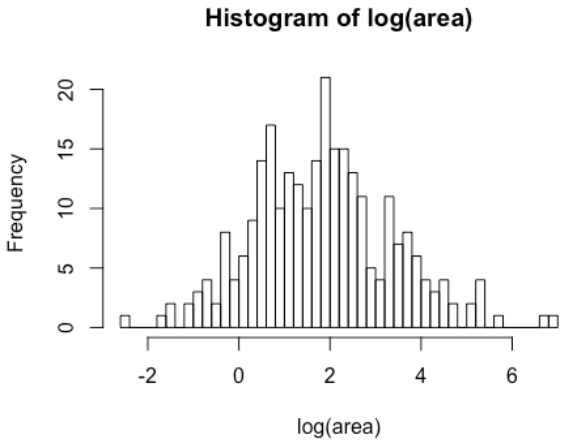
4.1 Box-Cox Transformation

Based on the histogram, we can see that the distribution of burned area is highly skewed. Hence, the Y (area) need to be transformed in order to continue with the analysis.



In order to do so, we used the Box-Cox method to choose the right form of our response variable. The Box-Cox output (see the figure below) suggests that we should take $\lambda = 0$, which means we should apply log transformation on the response variable area. After log transformation, we can see that area is approximately normally distributed.





4.2 Model Selection

First, we include all the variables into model and fit it. Because there are data unavailable in some levels of categorical factors month and day, we exclude these two factors from the model. After stepwise selection, we fitted a logistic regression model with two random variables, *DC* and *wind*. However, the drop1 output shows that the factor *wind* can be removed from the model. Thus, the final logistic regression model was as follow:

$$log(Y_{area}) \sim month + DMC + DC + temp \quad (5)$$

However, based on the output below, we can see that this model's R-square is only 0.01, which indicates that the model is poor fit based on this data. The Goodness of fit test below shows the same conclusion that the model is a bad fit.

```
## lm(formula = log(area) ~ month + DMC + DC + temp, data = datamul)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0282 -0.9929 -0.0772  0.8060  4.5184
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.080453   0.772566   2.693 0.007549 **
## monthaug     0.322943   1.280110   0.252 0.801028
## monthdec     1.667650   1.080894   1.543 0.124099
## monthfeb    -0.269138   0.880843  -0.306 0.760198
## monthjul    -0.192346   1.097080  -0.175 0.860962
## monthjun    -0.666496   1.041825  -0.640 0.522912
## monthmar    -0.492035   0.824906  -0.596 0.551384
## monthmay     1.074614   1.675463   0.641 0.521845
## monthoct     2.814218   1.592373   1.767 0.078363 .
## monthsep     1.593352   1.456284   1.094 0.274926
## DMC          0.009612   0.002658   3.617 0.000359 ***
## DC          -0.004743   0.001999  -2.373 0.018395 *
## temp         0.033362   0.022122   1.508 0.132772
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.488 on 257 degrees of freedom
## Multiple R-squared:  0.09267,    Adjusted R-squared:  0.05031
## F-statistic: 2.187 on 12 and 257 DF,  p-value: 0.01273
```

```
#goodness-of-fit test

pchisq(deviance(lmod12),df.residual(lmod12),lower=F)
```

```
## [1] 1.189773e-25
```

4.3 Model Transformation

Since the last model doesn't perform well regarding explaining the deviance of the response variable, we performed model transformation to improve the model fit. By adding interaction and quadratic terms in model, we were able to increase the R-square from 0.01 to 0.18. After the stepwise model selection, the final logistic regression model with quadratic terms is as follow:

$$log(Y_{area}) \sim month + FFMC + DMC + DC + ISI + temp + RH + wind + DC^2 + temp^2 + wind^2 + FFMC{:}DMC + FFMC{:}temp + FFMC{:}RH + temp{:}RH \quad (6)$$

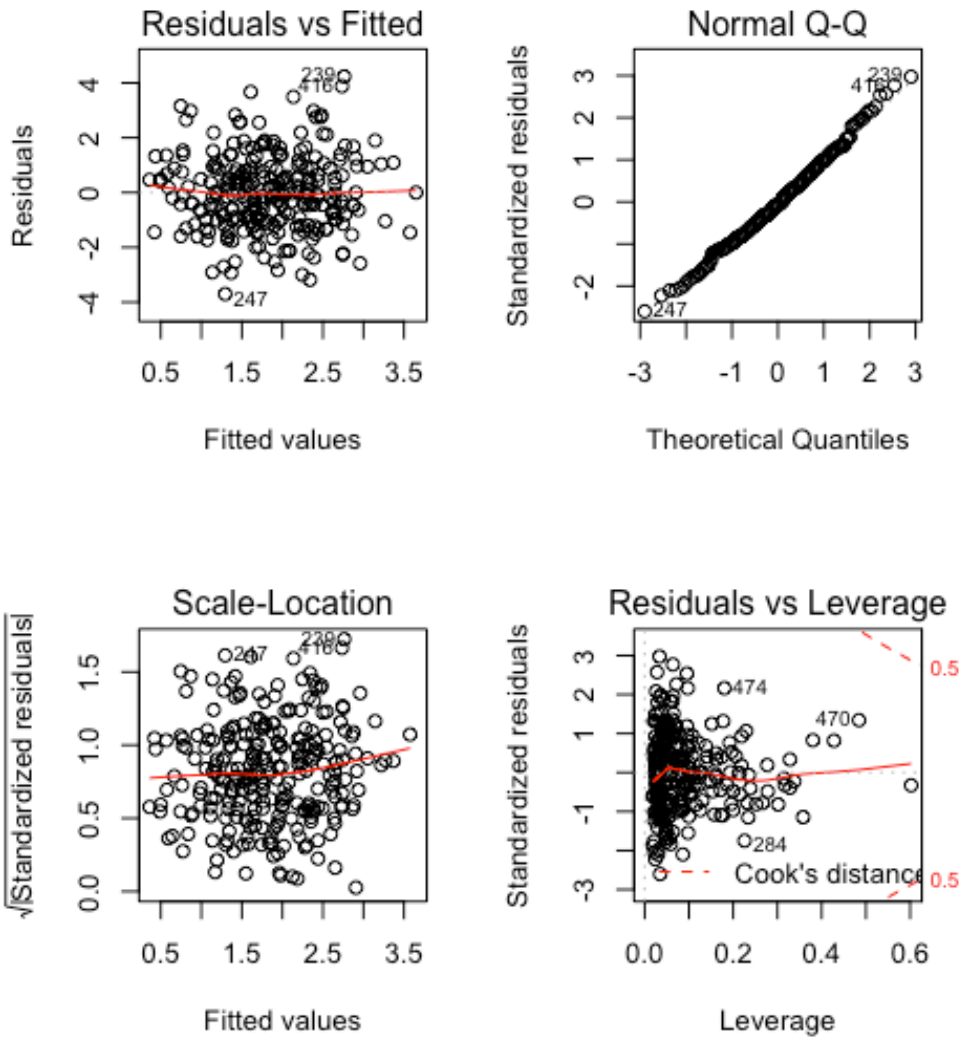
Using the chi-square test to compare model (3) and (4), we reject the null hypothesis and conclude that some of the extra terms in model (4) are significant, so we should choose model (4) instead of model (3). After that, we also tried to replace month with factor season and the chi-square test eventually suggests that we choose model (4).

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.324e+01  2.377e+01  -2.661  0.00831 **
## monthaug    -3.752e+00  2.371e+00  -1.582  0.11490
## monthdec    -2.268e+00  2.116e+00  -1.072  0.28479
## monthfeb     7.482e-02  9.052e-01   0.083  0.93419
## monthjul    -3.763e+00  2.088e+00  -1.802  0.07277 .
## monthjun    -3.663e+00  1.784e+00  -2.054  0.04104 *
## monthmar    -1.613e+00  9.204e-01  -1.753  0.08093 .
## monthmay     8.138e-01  1.711e+00   0.476  0.63468
## monthoct    -7.318e-01  2.503e+00  -0.292  0.77024
## monthsep    -1.898e+00  2.385e+00  -0.796  0.42701
## FFMC         7.755e-01  2.855e-01   2.716  0.00707 **
## DMC         -1.721e-01  7.394e-02  -2.328  0.02073 *
## DC           1.374e-02  8.299e-03   1.655  0.09911 .
## ISI          1.263e-01  1.192e-01   1.059  0.29049
## temp         2.005e+00  7.087e-01   2.830  0.00505 **
## RH           6.064e-01  2.386e-01   2.541  0.01166 *
## wind         4.570e-01  2.075e-01   2.202  0.02857 *
## I(DC^2)     -1.492e-05  6.630e-06  -2.250  0.02531 *
## I(temp^2)    1.236e-02  4.907e-03   2.520  0.01239 *
## I(wind^2)   -4.639e-02  2.217e-02  -2.093  0.03740 *
## FFMC:DMC     2.039e-03  8.123e-04   2.511  0.01270 *
## FFMC:temp   -2.894e-02  9.047e-03  -3.199  0.00156 **
## FFMC:RH     -7.285e-03  2.809e-03  -2.593  0.01008 *
## DC:ISI      -2.747e-04  1.885e-04  -1.457  0.14631
## temp:RH      2.953e-03  1.665e-03   1.774  0.07738 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.448 on 245 degrees of freedom
```

```
## Multiple R-squared:  0.1805, Adjusted R-squared:  0.1003
## F-statistic: 2.249 on 24 and 245 DF,  p-value: 0.001082

#Chi-square test to compare model (3) with model (4)

pchisq(deviance(lmod12)-deviance(lmod3),df.residual(lmod12)-df.residual
(lmod3),lower=F)

## [1] 1.748567e-07
```

4.4 Model Diagnosis

The plots below show that model (4) doesn't violate any assumptions of the model and there are no obvious outliers in the dataset. In conclusion, model (4) is valid.



Implication and Future Study

This case study is a really interesting example because that there're a lot of similar cases in real-life scenario. For instance, when we need to predict a citizen's yearly medical cost, it is common for subjects to have zero cost for an entire year if they did not get sick that entire year. In this case, our study suggests that it is beneficial to separate the data into two groups and analysis the following question respectively:

a. Whether the subject will spend on medical cost or not?

b. If the subject will spend money on medical care, how much will he spend?

This study has several limitations need to be further explored. First, the dataset is not large enough to predict the forest fire. Secondly, we only include quadratic terms in the model and future study can attempt to include cubic terms. Finally, there clearly is information missing in the dataset in order to predict forest fire. Researchers can look into other latent variables in the future.

Reference

Cortez, P., & Morais, A. D. J. R. (2007). A data mining approach to predict forest fires using meteorological data.