STATISTICAL APPROACHES TO ANALYZE FOREST FIRES

Ziyun Ding

PURDUE UNIVERSITY

1. Data Exploration

Forest fires affects forest preservation, create economical and cological lost, and cause human suffering. Therefore, it is neccessary to forecast the forest fire for a successful firefightinng. Weather conditions such as temperature, rain, wind and air humidity are known to affact 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 maintainence 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 leanring repository (http://archive.ics.uci.edu/ml/datasets/Forest+Fires). This data provides the environmental condistions 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*).

By plotting a histogram of the response factor *area* (see figure 1) , we noticed that there is a clear right skewness and hence we might need to try transformation on response variable later. In fact, there are 247 observations (48% of observations) are with area equal to zero.

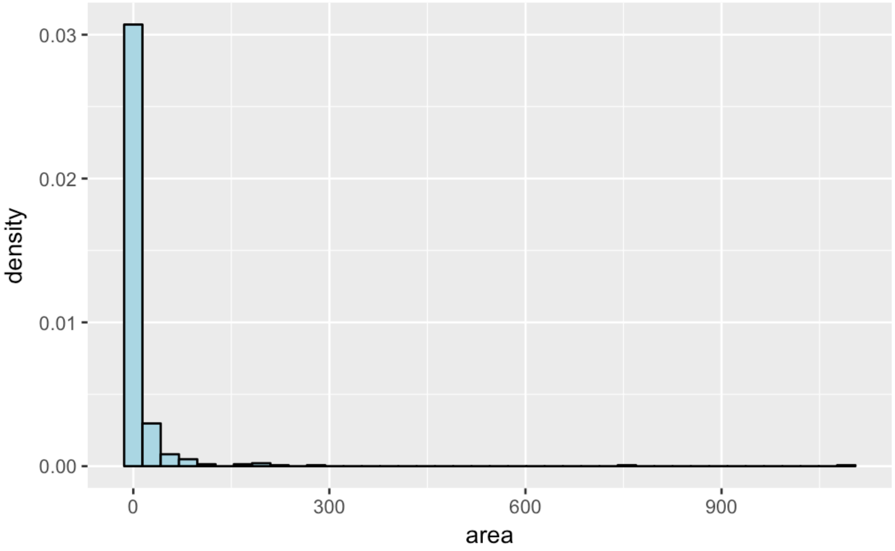

Figure 1. The histogram of the burned area.

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 2), indicating 1-9 on the x-axis and 1-9 on the y-axis.

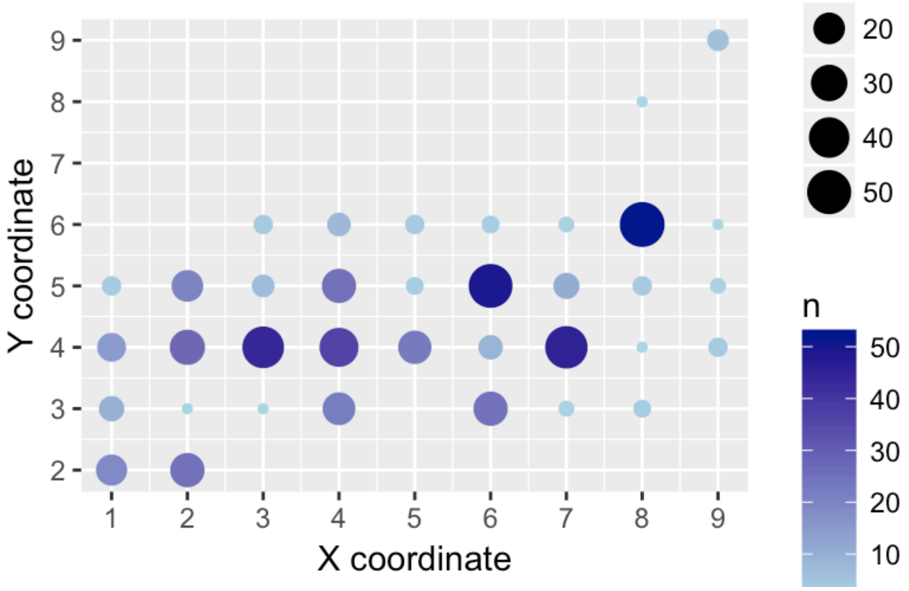

Figure 2: The distribution of the fire (burned area greater than 0)

By using the "tidyverse" library to process the data, the top 10 regions with most fire occurrence, and the top 10 regions with the largest burned area are listed in Table 1. From the tables, we could see that the regions with X and Y coordinates 6_5, 8_6, and 7_4 have the top 3 most fire occurrence and burned area.

Regions with most fire occurrence			**Regions with largest burned area**		
X_Y	**count**	**max_area**	**X_Y**	**count**	**max_area**
6_5	32	1090.84	6_5	32	1090.84
8_6	29	746.28	8_6	29	746.28
7_4	25	278.53	7_4	25	278.53
4_4	20	88.49	1_2	4	212.88
2_4	17	54.29	2_2	11	200.94
3_4	15	35.88	8_8	1	185.76
2_5	14	26.43	4_5	10	154.88
4_3	13	49.59	9_4	4	105.66
5_4	13	24.24	6_4	6	103.39
2_2	11	200.94	1_	4	95.18

Table 1. The Ranking of Regions (Top 10).

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 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).

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.

2. Factors Influencing Fire Occurrence

By creating a dummy variable *factor_area* with two levels (factor_area=1 means the area was bured while factor_area=0 means the area wasn't burned), we are able to analysis the probability of the occurrance of fire with a binomial model. Because there are data unavailable in some levels of categorical factors month, we transform the month factor into the season factor.

2.1 Data Transformation and Model Selection

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 factors in the final logistic regression model (model 1) are all significant:

$$logit(fire) = -0.679 + 0.001X_{DC} - 0.297X_{temp} + 0.099X_{Rh} + 0.009I_{temp^2} - 0.001I_{RH^2} + 0.013I_{wind^2} \tag{1}$$

Since the estimate of the overdispersion parameter is close to 1, we ruled out the possibility of overdispersion.

2.2 Model Diagnosis

The plots in Figure 3 show that the model doesn't violate any assumptions. The first plot 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 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.

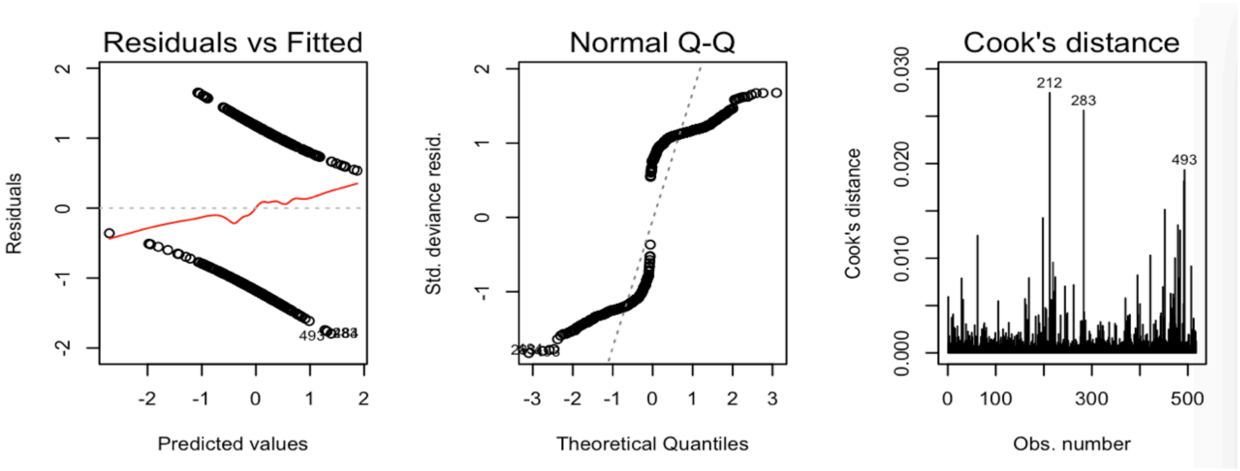

Figure 3: Diagnosis plots for the logistic regression model (1).

2.3 Five-fold Cross Validation

In order to test the model's ability to correctly predict the amount of burned area in real life, we did 5-fold Cross Validation using model (1), the averaged accuracy on the testing data set is 47.38%, indicating that the model is not accurate enough for predicting the occurance of forest fires.

3. Factors Influencing Burned Area

To analysis the factors influencing the amount of area burned in the fire, we used only the data with positive area value, which is equivalent to analyzing factors influence burned area assuming that the fire has occurred.

3.1 Data Transformation

Based on the histogram in Figure 4, 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.

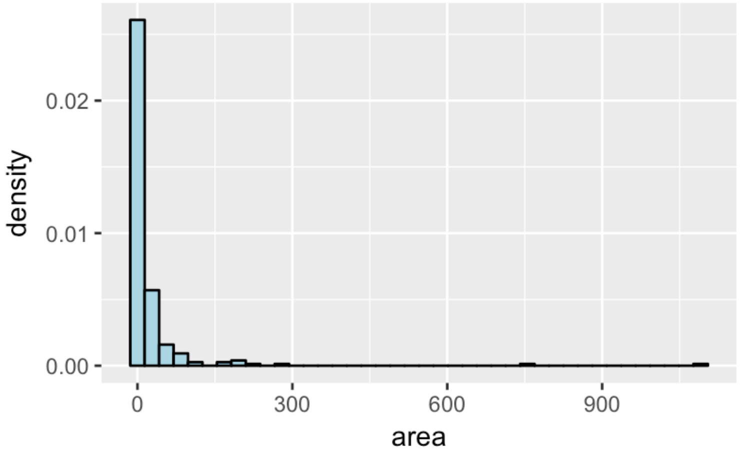

Figure 4: The distribution of all burned area data that larger than zero.

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 5 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.

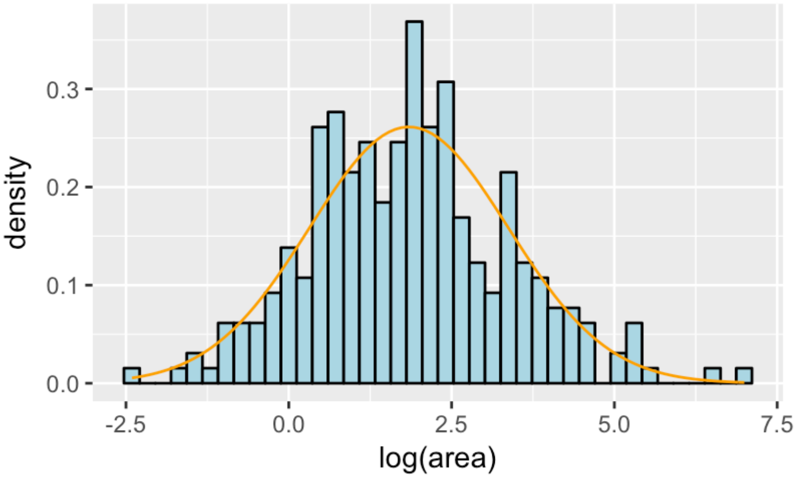

Figure 5: the distribution of burned area data (>0) after log transformation.

3.2 Model Selection

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:

$$log(Y_{area}) = -63.24 - 3.752 monthaug - 2.268 monthdec + 0.075 monthfeb - 3.763 monthjul - 3.663 monthjun - 1.163 monthmar + 81.38 monthmay - 0.732 monthoct - 1.898 monthsep + 0.776 FFMC - 0.172 DMC + 0.014 DC + 0.126 ISI + 2.005 temp + 0.606 RH + 0.457 wind + 0.000\, I_{DC^2} + 2.005 \mathrm{I}_{temp^2} - 0.047 \mathrm{I}_{wind^2} + 0.02 FFMC{:}DMC - 0.029 FFMC{:}temp - 0.007 FFMC{:}RH - 0.000 DC{:}ISI + 0.003 temp{:}RH \qquad (2)$$

3.3 Model Diagnosis

The plots below shows that model (2) doesn’t violate the normality, constant variance and linearity assumptions of the model and there are no obvious outliers in the dataset. In conclusion, model (2) is valid.

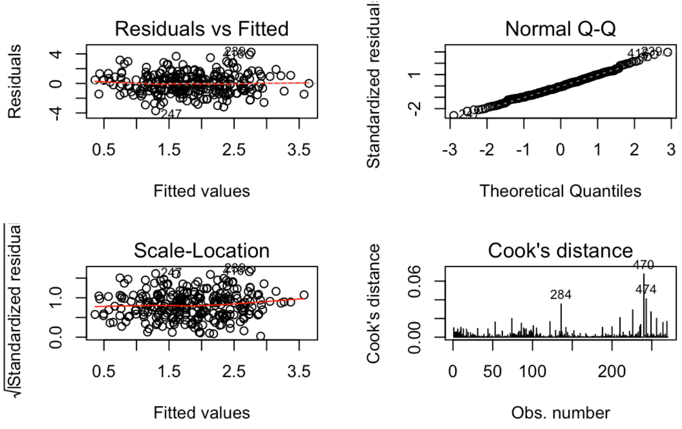

Figure 6: Diagnosis plot for the linear regression model (2).

3.4 Five-fold Cross Validation

In order to test the model's ability to correctly predict the amount of burned area in real life, we did 5-fold Cross Validation to get the proportion of test data that are correctly predicted. Overall, only 43% of the test dataset are correctly predicted, indicating that the model (2) is not accurate enough for predicting the size of the burned area in a forest fire.

4. What we could learn from this study and what we couldn't do

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. The fitting of the models and cross-validation results indicate that the model fitting is not ideal for prediction. 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.

5. Reference

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