STAT151A Final Report

Amanda Wu

"Fire in the Mediterranean is an unavoidable cultural and ecological phenomenon, but an avoidable catastrophe"

STAT151A Linear Modeling

Professor Stoyanov
Amanda Wu
Apr 30th, 2018

STAT151A Final Report

Introduction

Forest fires have become a much more severe problem in many regions of the world as a result of climate change and human activities. They not only affect large areas and people who live nearby, every year, forest fires also release over 3000 million tones of CO2 into atmosphere, contributing to the acceleration of global greenhouse gas emissions.

Thus, it is crucial for us to develop an effective fire prediction model from the data we currently have. Specifically, in this report I will explore the Forest fire data from the Montesinho Natural Park in the Northeast region of Portugal.

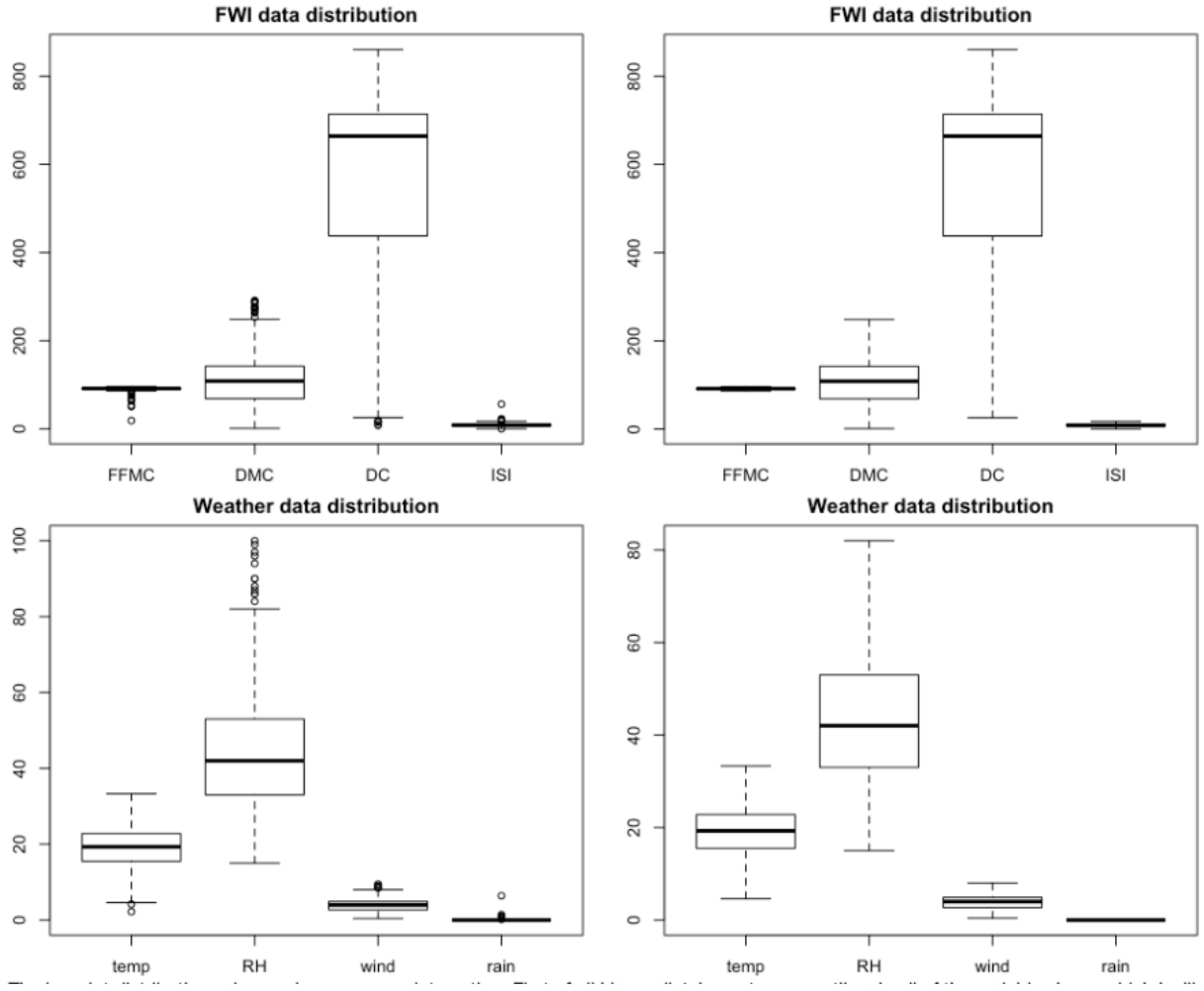
I hope to characterize the current fire regime, understand unobserved patterns in areas burned, and analyze factors that may explain the forest fires to help the government as well as the fire department to answer the following questions:

1. What are the variables that influence the total burned area in forest fires?
2. What is a good predictive equation for predicting the total burned area in terms of given variables?

Explanatory Data Analysis

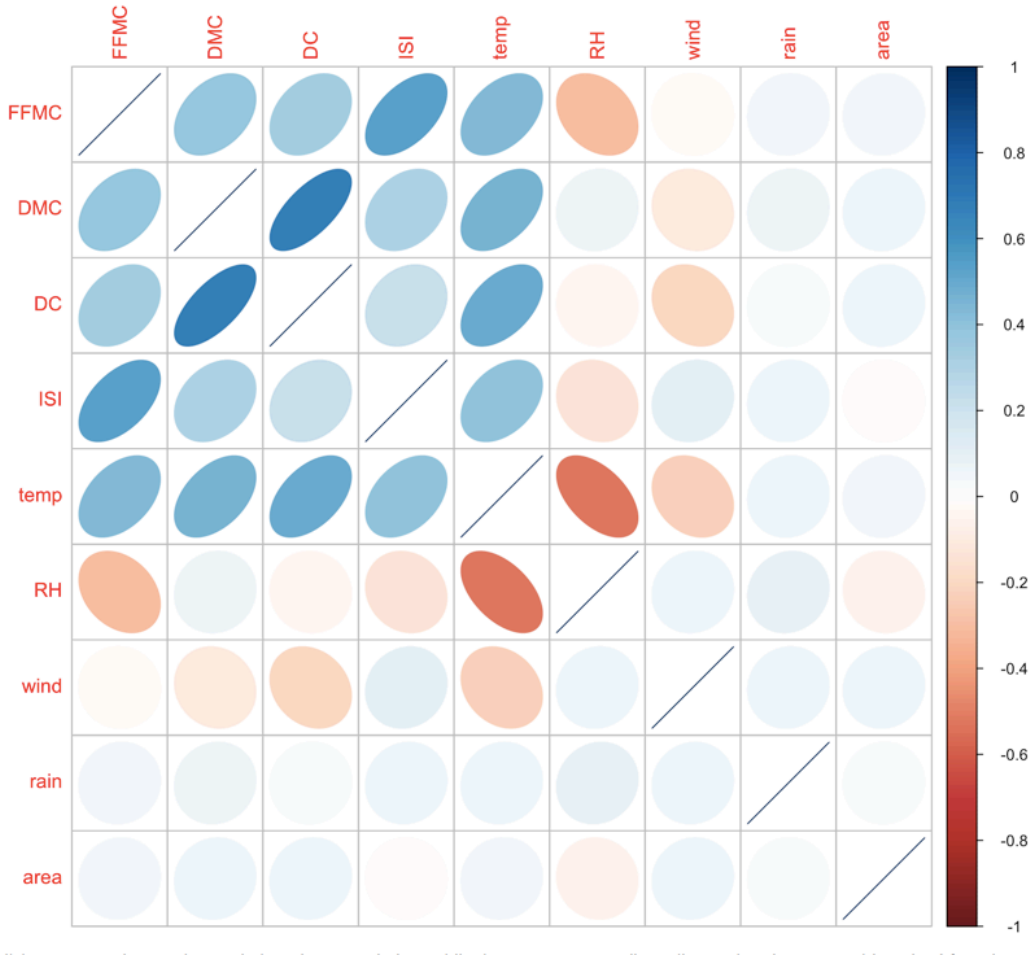
While preparing for the more formal modeling step and the analytics of the data, I want to first examine the predictor and response variables univariately and multivariately because it can help me not only get graphical displays, but also numerical summaries of the data, which is a key initial step in both explanatory and predictive modeling.

The dataset consists of a total of 517 entries with 13 attributed including spatial and temporal attributes including X, Y coordinates, month and day, 4 FWI components that are determined by weather conditions, and 4 meteorological attributes including temperature, humidity, wind and rain. And lastly, the response variable is the burned area.



After carefully examination of the distribution of each variable, while outliers are shown be present in most of the variables, I noticed that skewness is also another big issue. The burned area showed a very positive right skew distribution centered around 0, representing the burned area to be less than 100 square meters. Thus, in order to achieve approximate symmetry and homoscedasticity of the residuals, I transformed

the dependent variably area to log(area + 1), while I kept all the other independent variables the same as they are all taken as fixed, so normality is inapplicable.



Additionally, while doing bivariate Analysis, I spotted some obvious collinearity between my explanatory variables, such as the more shaded area above, which led me to include their interaction terms to represent their interaction and help us understand their effect on each other as well as on our response variable the burned area.

Initial Modeling

Before I start building the model, I went back to my client's question, which to my understanding, asked for two things 1) the variables to be included in an inferential model and 2) a predictive equation.

The goals, the criteria for choice of variables, and choice of methods are very different for these two models. In an inferential model, the independent variables are regarded as causes of dependent variables, so my major goal here to answer the first question is to use current attributes I have to get unbiased estimates of the regression for burned

area, while trying not to omit variables that might have explanatory power in the model. On the other hand, to answer the second question my client has, in a predictive study, my goal is to develop formula for making predictions about the burned area, so the focus is on the association as well as reducing estimation variance.

To start the process, I first used all the variables including the interaction terms to regress on the log transformed burned area, however, the model fit was terrible. So I went back to my univariate analysis to try to catch any missed patterns that might make my model better. With many trials and errors, I found that forest fires occur much more frequently in the weekends and in the summer. Thus, I categorized days into weekend vs weekdays, and months into 4 seasons, and modified my initial model to include these 2 newly added categorical variables, and the fit seemed to improve a lot with a much higher F statistic and r-squared, indicating explanatory power in my independent variables.

Finally, I chose ANOVA for my inferential model and Stepwise Selection for my predictive model. The main reason was that ANOVA measures the relevance of features by their correlation with dependent variable, while Stepwise Selection measures the usefulness of a subset of feature by actually training a model on it. I didn't use ANOVA for my predictive model also because it can make the model more prone to over-fitting, which I absolutely need to avoid for my predictive model, as compared to using subset of features from the Stepwise Selection.

```
Start:  AIC=339.89
area ~ season + DMC + temp + wind

         Df Sum of Sq    RSS    AIC
<none>                971.06 339.89
- DMC     1    7.8151 978.88 342.03
- wind    1    9.4886 980.55 342.92
- temp    1    9.5870 980.65 342.97
- season  3   26.2162 997.28 347.66
```

To conclude, after checking the significance of independent variables one by one, ANOVA method suggests that I include temperature, wind, season, DMC, and the interaction term between DMC and DC as variables that influence the total burned area. And just to satisfy the principle of marginally, I also added DC as it is part of the interacting term and does no harm to the model. On the other hand, Stepwise Selection also give me a similar output only with the exception of the interaction term. It shows that all season, DMC, temp and wind should be included in the multiple linear predictive model, and their coefficient summary are shown as below.

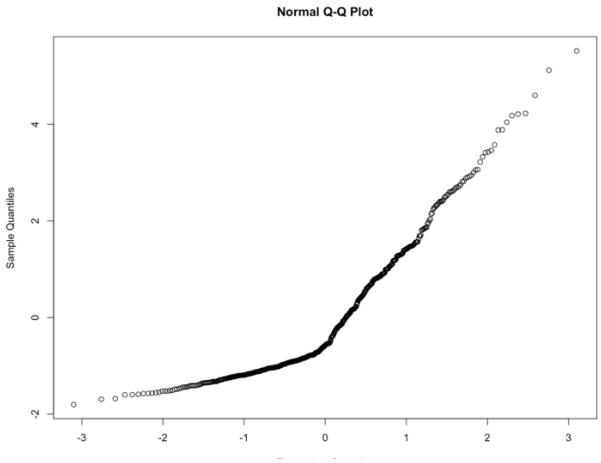
```
lm(formula = area ~ season + DMC + temp + wind, data = data)

Coefficients:
 (Intercept)  seasonspring  seasonsummer  seasonwinter           DMC
    0.041620     -0.080146     -0.390222      0.740523      0.002589
        temp          wind
    0.032042      0.079298
```

Model Diagnostics

For the diagnostics part, I focused mainly on my predictive model since this is the model that was directly asked by my client. To validate my model, I need to check 3 of my error assumptions, including independence, homoscedasticity, and normality, as well as the structure of my model, that is, there is no obvious nonlinearity in the relationship between my explanatory variables and burned area, and lastly, outliers, leverage points and influential points. All the above ideally should be met to conclude that my model is powerful in its predictive strength.

Generally speaking, error assumptions are somewhat met after plotting different graphs. However, when checking the linearity of the model structure, the pattern shown on the graph doesn't indicate linearity, which means that further transformations on the independent variable might be needed. Lastly, as I would already expect from my univariate and multivariate analysis from the beginning, there are many outliers, leverage points and influential points detected through model diagnostics, and many of them overlapped too. This means that I need to consider whether I want to include these observations in the final model, which might depend on various factors and require me to take an even closer look at those specific data points to see if the errors are just an input error.



Conclusion

Based on RMSE for my explanatory model and AIC for my predictive model, both model did relatively descant. Overall, temperature, wind, season and DMC are proven to contribute both to the log(area + 1) as to explain the data as well as to predict. Additionally, more variation that can be explained, I would also recommend adding DMC and DC's interaction term as well as DC.

More generally speaking, throughout the analysis process of this project, it was interesting while also unsurprising to find that forest fires do occur more frequent in the summer, and aside from wind and DMC, the time of the year actually contribute quite a lot to the forest fire as the explanatory variable has the lowest p-value in the F-test. Furthermore, if we un-tranform the fire burned area, we will find out that our explanatory variables contribute to the exponential power of the area of the fire, meaning that, its impact on the fire area is probably a lot more than harmful and damaging than what we have expected. Thus, it is crucial that the fire department are extra prepared in the summer and when they see a rise in temperature, DMC or stronger wind.