PREDICTING FOREST FIRES IN ALGERIAN REGIONS

BY JOYCE CHEN

1. Introduction.

With the increase in the frequency of forest fires worldwide, many towns around forests have been badly damaged, endangering residents' lives and property, and deteriorating air quality. For instance, the occurrence of the Jasper fire in Alta, Canada, highlighted by Asgary[1] (2024), points to the additional problems wildfires bring concerning climate warnings, prompting governments to pay closer attention to this issue. Consequently, governments prefer to predict forest fires to take early precautions. This paper addresses the following questions: What are the key covariates influencing the occurrence of forest fires?
To answer this question, We will use linear models, penalized regression models and additive model to identify the best combination of covariates in each model and employ Generalized Cross-Validation (GCV) to compare and select the best model. By this analysis, we wish to ensure the governments could do some precautions to reduce the occupation of forest fire.

2. Data Analysis.

The paper uses the dataset "Algerian Forest Fires" collected by Abid, Faroudja[2] in 2019. The Algerian Forest Fires Dataset is a collection of 244 data concerning forest fires from June 2012 to September 2012. Every row of the dataset means every day's observation.
The response variable is the probability of forest fire occurrence (Response), calculated using the bootstrap and weight probability method. Initially, we considered 10 covariates: temperature (Temp), relative humidity (RH), wind speed (WS), total daily rain (Rain), Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Buildup Index (BI), and Fire Weather Index (FWI).
When we plot the association of Rain and Response in Figure 1, we find that when the day is not rain (Rain = 0), Response has a huge fluctuation between 0 and 1; And when the day is rain (Rain $\neq$ 0), probability of forest fire always happens (Response = 1). This indicates no reasonable relationship between Rain and Response, which prompting us to remove Rain.

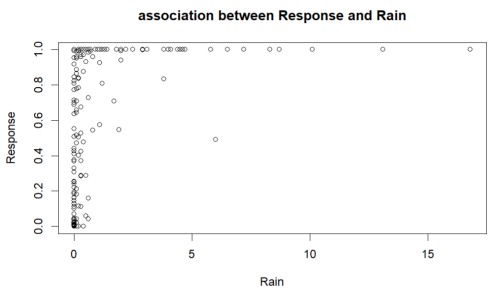

FIG 1. *Plot of Association between Response and Rain*

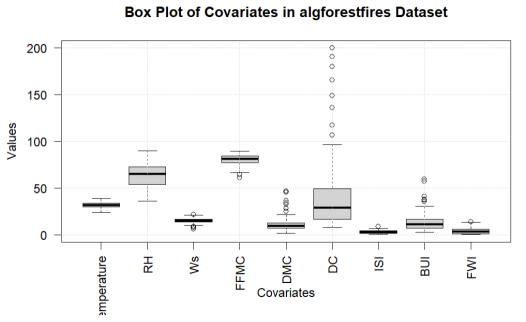

FIG 2. *Boxplot for all covariates*

[1] See Ali 2024

[2] See dataset from UCI 2019

Next, we plot boxplots for the rest of 9 covariates in Figure 2, we find there are several outliers, so we remove them and we get remaining 164 observations. We further reduced the dataset to 80 observations by removing extreme 0s and 1s in the Response variable, which would affect the residual vs. fitted plots for all covariates in Figure 3.

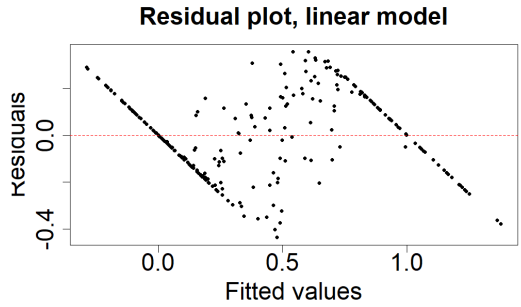

FIG 3. *Risidual vs Fitted plot for All Covariates in Linear Model*

For exploratory analysis, we plot several graphs between the Response and several covariates in Figure 4 the first row, the relationship between the FFMC, ISI, FWI and Response show a rapid increase, indicating linear relations. Therefore, we will test a linear model. Additionally, these trends suggest that these covarites are valuable for predicting the Response. Then, at the last row are two plots between different covriates, revealing positive linear relationships such as between BUI and DMC, and FWI and ISI. This indicates multicollinearity, prompting us to use penalized regression models to fit these covariates. We also need to ensure these variables do not appear together in the same combination for the linear model to maintain their independence. Considering these findings, we will determine feasible combinations of covariates to predict the Response. I expected to find key factors contain one factor in FFMC, ISI and FWI.

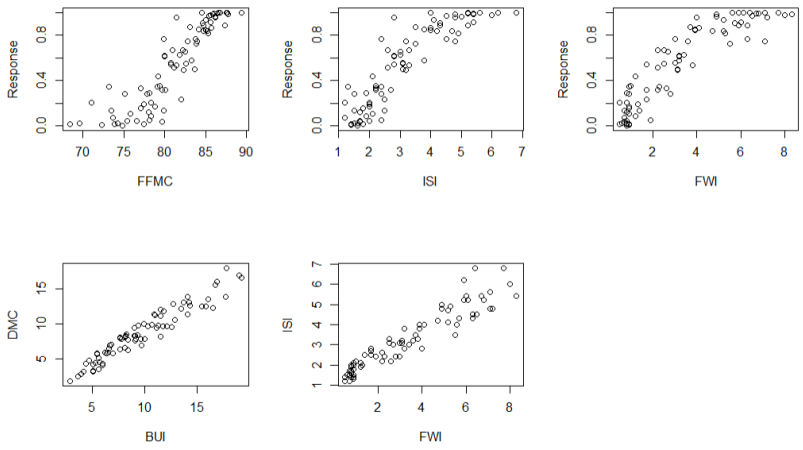

FIG 4. *Plots between different covariates or between Response and covariates*

3. Model.

3.1. *the Best Covariates' Combination in Linear Model by Generalized Cross-Validation.*

At first, we use $X_{ij} = \frac{X_{ij} - \bar{X}_j}{s_j}$, where $\bar{X}_j$ is the average value of X_j and s_j is the standard deviation of X_j, to standardize the dataset. Then, we list all of the possible combination for those covariates, which is around $2^{11} = 2048$ types. After that, we calculate the Generalized Cross-Validation (GCV) for every types by the function $GCV(\hat{y}) = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(1 - \frac{trace(H)}{n})^2}$, where $H = X(X^TX)^{-1}X^T$ and list the best top 10 models in linear model. We then choose the best model and a relatively simpler one. We select the simpler model due to its interpretability, robustness, and computational efficiency.
Besides, we let the selected covariates be $X_1, X_2, ..., X_p$, and the model is $Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$.

3.2. *Penalized Regression Model.*

To reduce the effect of multicollinearity, we add the penalty term, so we use ridge regression, the function is $\beta_{ridge} = argmin_\beta\{\| y - X\beta \|_2^2 + \lambda \| \beta \|_2^2\}$, Lasso regression, the function is $\beta_{lasso} = argmin_\beta\{\| y - X\beta \|_2^2 + \lambda \| \beta \|_1\}$, and the Elastic Net regression, the function is $\beta_{elastic} = argmin_\beta\{\| y - X\beta \|_2^2 + \lambda\alpha \| \beta \|_1 + \lambda(1-\alpha) \| \beta \|_2^2\}$, we will find the best α. Then, we would to compare by GCV score and fit the model with the "best" λ.

3.3. *Additive Model.*

Use generalized additive model to fit the association between Response and all 9 covariates, the function is $Response = \beta_0 + s(Temp) + s(Ws) + s(RH) + s(FFMC) + s(DC) + s(ISI) + s(BUI) + s(DMC) + s(FWI) + \epsilon$, we will change the number of knot of the smoothing function to fit whether the model fit well by the R-square, and we will calculate the $GCV(\hat{y})$.

4. Result.

4.1. *the Best Covariates' Combination in Linear Model.*

From the Generalized Cross-Validation, we get the best top 10 covariates' combination in Figure 5, we choose the first top scoring one, which is "RH, Ws, FFMC, ISI, BUI" and $GCV_{top1\_linear} = 0.0137$, and we choose the simplest one with nearly as high a score in the sixth row, which is "RH, Ws, FFMC, FWI", and $GCV_{simple\_linear} = 0.0141$.

Model <chr>	Score <dbl>
RH, Ws, FFMC, ISI, BUI	0.01371583
RH, Ws, FFMC, DMC, DC, ISI	0.01388943
RH, Ws, FFMC, DC, ISI, BUI	0.01395715
RH, Ws, FFMC, DMC, ISI, BUI	0.01404610
RH, Ws, FFMC, ISI, BUI, FWI	0.01406805
RH, Ws, FFMC, FWI	0.01412733
RH, Ws, FFMC, DC, FWI	0.01418400
RH, Ws, FFMC, DMC, DC, ISI, FWI	0.01421591
RH, Ws, FFMC, DMC, DC, ISI, BUI	0.01426589
RH, FFMC, ISI, BUI	0.01430778

1-10 of 256 rows

FIG 5. *Top 10 best Combinations in Linear Model*

Then, Residual vs fitted plots and QQ-plots for the two models both meet models' assumption, and the p-value for every covariates are smaller than 0.05, which means they are fitted to the linear model.

4.2. *Penalized Regression Model.*

For ridge regression model, for the choice of λ, we find that, large λ did not lead to any explanatory but a higher MSE, so we choose the minimum MSE as the "best" λ, which is around -3, and $\lambda = 0.001$,for the result coefficient estimates, we find that coefficient estimates from ridge regression is similar to that from linear model, there is no 0 in this model, so I do not choose this model.

Then, we make a model for Lasso regression. The choice of λ are all small, the minimum one is -7, which is around 0.0000001 and contain 7 covariates, and the Least Squares Error (LSE) one is -3.5, which is around 0.001 and contain 5 covariates. they are similar, so we choose the relatively less covariates, which is the LSE one. The result coefficient estimates we get in Figure 6, left side is from Lasso regression, right side is linear model, we find that there are only RH, FFMC, DC, ISI, FWI have effective coefficent estimates, and we get $GCV_{Lasso} = 0.0169$.

```
10 x 2 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept) .             0.53646932
Temperature .             0.03707700
RH          3.041806e-02  0.08287165
Ws          .            -0.02543056
FFMC        1.553702e-01  0.14166051
DMC         .             0.09191132
DC          1.316527e-02  0.07419088
ISI         7.866597e-05  0.21851587
BUI         .            -0.02932044
FWI         1.549028e-01 -0.09966644
```

FIG 6. *Coefficient Estimates between Linear Model and Lasso Regression Model*

At last, for Elastic net regression, after we compare those α and we choose $\alpha = 0.98$, which is similar to the Lasso regression ($\alpha = 1$), and we choose the minimum λ since it has relative simpler model, we get around -6.5 and $\lambda = 0.000001$, and GCV score is $GCV_{elastic} = 0.0146$. The choice of covariates is Temp, RH, Ws, FFMC, DMC, DC, ISI.

```
11 x 1 sparse Matrix of class "dgCMatrix"
                     s1
(Intercept)  0.53646932
(Intercept)  .
Temperature  0.03307127
RH           0.07786995
Ws          -0.02446679
FFMC         0.14611661
DMC          0.04147037
DC           0.05022303
ISI          0.14014923
BUI          .
FWI          .
```

FIG 7. *Coefficient Estimates between Linear Model and Elastic Net Regression Model*

4.3. *Additive Model.*

For the additive model, at first we choose the best knots equal to 10, and $GCV_{additive} = 0.0072$, which is negligible. Then, we plot partial effect plot, this plot shows the smooth terms for each predictor variable in a Generalized Additive Model (GAM). The solid lines represent the estimated effect of each predictor, and the dashed lines represent the 95% confidence intervals. From the plot in Figure 8, we can see that most predictor variables have smooth terms that are close to horizontal, and the confidence intervals include zero. This indicates that these variables do not have significant nonlinear effects on the response variable. Based on these plots, we can infer that a GAM model might not be the most suitable choice for this dataset, and a linear model might be more appropriate.

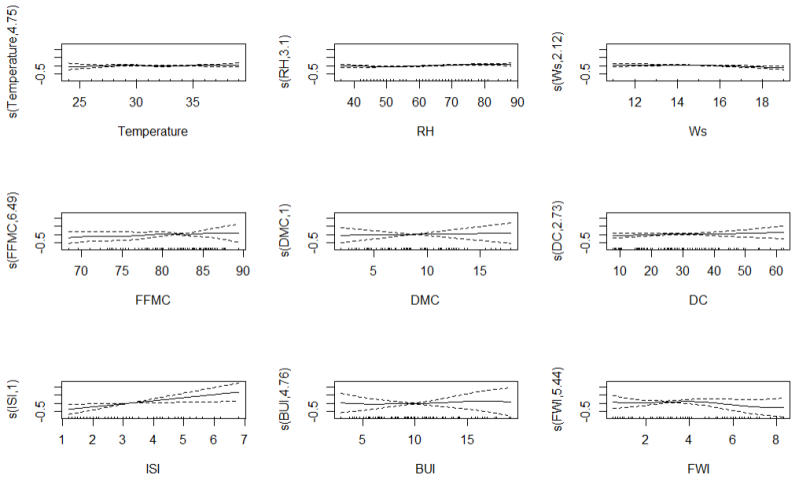

FIG 8. *Partial Effect Plot*

4.4. *Check Model Assumption.*

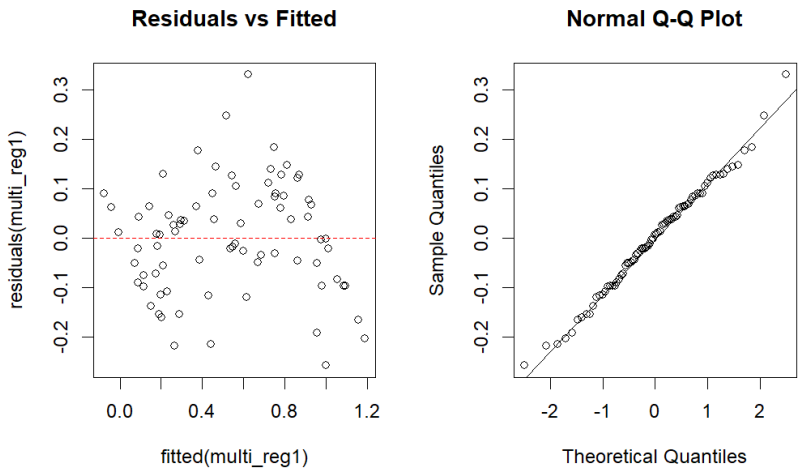

FIG 9. *Risiduals vs Fitted plot and QQ-plot for "RH, Ws, FFMC, FWI" Linear Model*

By analyzing these variables, each variable provides a different perspective on fire risk, and their combined analysis offers a comprehensive understanding of forest fire behavior.
Compare those models from 4.1 to 4.3, I prefer to use the linear model. Then, although the best one has relatively lower GCV score, we prefer to choose the simpler covariates' combination, which is "RH, Ws, FFMC, FWI". Figure 9 shows the risidual vs Fitted plot and QQ-plot, I find:

For Random scatter of residuals: The residuals vs fitted plot shows no clear pattern, indicating that the linear model appropriately fits the data without systematic bias.

For Normal distribution of residuals: The normal QQ-plot indicates that the residuals are approximately normally distributed, which aligns with the assumptions of linear regression.
In summary, RH, Ws, FFMC, and FWI are integral to our research as they collectively enhance the understanding of fire dynamics and contribute to developing effective fire prediction models. The insights gained from analyzing these variables can lead to improved fire prevention, preparedness, and response strategies in the Algerian regions and beyond.

5. Conclusion.

The analysis identified the optimal combination of covariates for predicting forest fire probability in the Algerian regions dataset. The best combination for the linear model included "RH, Ws, FFMC, FWI," achieving a relatively low Generalized Cross-Validation (GCV) score and maintaining simplicity, interpretability, and computational efficiency. This combination proved effective in fitting the data, as evidenced by the residuals vs. fitted plot and the QQ-plot, both indicating no significant deviations from model assumptions.
The study faced several limitations. Firstly, the analysis was limited to data from only two Algerian regions and covered a short time span of just four months in one year, and we remove several observations by different reasons. Consequently, the sample size was relatively small, and the results might be influenced by these constraints, potentially affecting their generalizability. Furthermore, the similarity in region and climate between the two areas analyzed means that the findings may not be applicable to regions with different climatic conditions.
Another limitation was the scope of models used. The analysis employed linear models, penalized regression models, and additive models. While these models provided insights, the study could benefit from exploring other advanced modeling techniques, such as decision trees or more complex machine learning algorithms, to improve predictive accuracy and robustness.
In future, research should aim to address these limitations by collecting more extensive datasets spanning multiple years and various regions. This would enhance the robustness and generalizability of the findings. Additionally, incorporating more advanced modeling techniques could provide a deeper understanding of the factors influencing forest fire occurrence and improve prediction accuracy.

REFERENCES

[1] Ali Asgary (2024, July 31). The jasper fire highlights the risks climate change poses to Canada's World Heritage Sites. The Conversation. https://theconversation.com/the-jasper-fire-highlights-the-risks-climate-change-poses-to-canadas-world-heritage-sites-235622

[2] Abid, Faroudja. (2019). Algerian Forest Fires. UCI Machine Learning Repository. https://doi.org/10.24432/C5KW4N