Linear Regression Analysis of Forest Fire Area

Kerem Zengin

June 5, 2023

Contents

1 Data **2**

2 Methodology **2**

3 Implementation **3**

4 Results **4**

5 Visualizations **5**

6 Conclusion **5**

Abstract

This study presents a linear regression analysis of forest fire areas in relation to three key environmental variables: temperature, relative humidity, and wind speed. The data, compiled into a comprehensive dataset [1], were analyzed using MATLAB's linear regression function. The aim is to identify the relationships between these environmental factors and the size of forest fire areas.

1 Data

We have a comprehensive dataset of forest fires which includes measurements of environmental factors. Dataset comprises measurements of temperature (temp), relative humidity (RH), wind speed (wind), and the size of the area affected by the fire (area).

We cleaned dataset to remove missing data. The total number of observations after cleaning is 517.

2 Methodology

To analyze the data, we used a linear regression model. In this case, the area affected by the forest fire is the dependent variable, while temperature, relative humidity, and wind speed serve as the independent variables. The general form of the linear regression model is as follows

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon \tag{1}$$

[1] http://archive.ics.uci.edu/ml/datasets/Forest+Fires

In this equation:

- y is the dependent variable (the area affected by the forest fire),
- $x_1, x_2, \ldots, x_p$ are the independent variables (temperature, relative humidity, and wind speed),
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the parameters of the model,
- ε is the error term.

Our aim is estimate the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ that minimize the sum of the squared residuals

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

3 Implementation

We implemented the linear regression model using MATLAB.

- **Data Loading and Cleaning:** First, we loaded the data from a CSV file using MATLAB's 'readtable' function. We then removed any rows with missing data using the 'rmmissing' function.
- **Matrix Preparation:** We created a matrix, X, with the predictors. The first column of this matrix was a column of ones for the intercept, and the subsequent columns were for temperature, relative humidity, and wind speed.
- **Regression:** We performed the regression by solving the normal equations, $X^T X\beta = X^T y$, where X is the matrix, y is the variable (area), and β is the vector of coefficients.
- **Fitted Values Calculation:** After obtaining the coefficients, we computed the fitted values by multiplying the matrix, X, with the coefficients, β.
- **Data Visualization:** We visualized the data and the regression fit using MATLAB's plotting functions. Separate plots were created for each predictor versus the response variable.
- **Table:** We also used MATLAB's built-in function 'fitlm' to obtain a detailed summary of the regression results, including the estimated coefficients, standard errors, t-statistics, and p-values for each predictor.

4 Results

The results of the multiple linear regression analysis are summarized in Table 1 [2].The estimated coefficients shows the expected change in the area of forest fire for unit increasing in the corresponding predictor.

	Estimate	SE	tStat	pValue
(Intercept)	-5.3573	19.802	-0.27054	0.78685
Temp	0.98172	0.58126	1.6889	0.091839
RH	-0.11978	0.20193	-0.59319	0.55331
Wind	1.236	1.6044	0.77034	0.44145

Table 1: Estimated coefficients of the regression model.

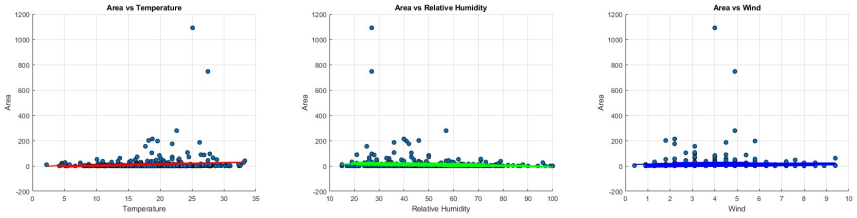
- Temperature has a positive coefficient, implying a positive relationship with the area of forest fire. However, the high p-value of 0.091839 says that this effect is not significant.
- Relative Humidity and Wind also have non-significant p-values, implying that there is not enough statistical evidence to conclude that these predictors have an effect on the area of forest fire.
- The R-squared value of the model is 0.0115, says that the model explains only 1.15% of the variability in the fire area. This value says that the model can not have a strong prediction.

[2]

1. Estimate: This is the estimated coefficient of the corresponding predictor in the regression model.
2. SE (Standard Error): This measures the variability in the estimate for the corresponding predictor's coefficient. A lower standard error indicates a more precise estimate.
3. tStat (t-statistic): This is the value of the test statistic for the hypothesis test that the corresponding predictor's coefficient is equal to zero. It is calculated as the ratio of the estimate to the standard error. The greater the absolute value of the t-statistic, the stronger the evidence against the null hypothesis.
4. pValue (p-value): This is the probability of obtaining a t-statistic as extreme as the observed value under the null hypothesis. A smaller p-value indicates stronger evidence against the null hypothesis.

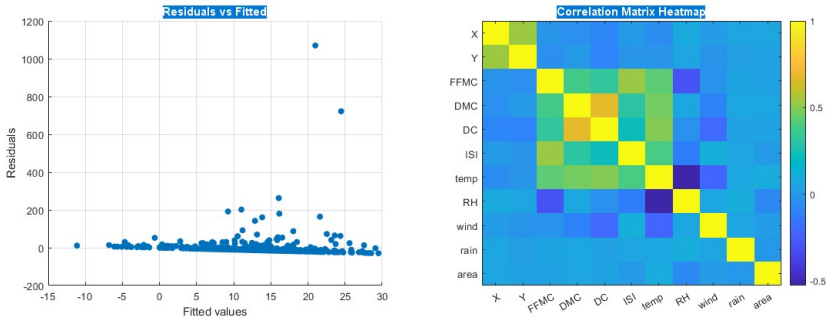
5 Visualizations

In this section, we explore the data and the results of our linear regression model through visualizations.



(a) Area vs Temperature (b) Area vs RH (c) Area vs Wind

Figure 1: Scatterplots of area vs predictors with fitted regression lines.



(a) Residuals vs Fitted (b) Correlation Matrix Heatmap

Figure 2: Residuals vs fitted values and correlation matrix heatmap.

6 Conclusion

In this study, we examined the relationships between forest fire area and three environmental factors: temperature, relative humidity, and wind speed, using a linear regression model. The results say that, while there is some relationship between these predictors and the response, the relationship is not strong, as indicated by the low R-squared value.

Appendix

```
% Load the data
data = readtable('forestfires.csv');

% Remove rows with missing data
data = rmmissing(data);

% Create a  matrix with predictors
X = [ones(height(data),1), data.temp, data.RH, data.wind];

%  Regression Part
beta = X \ data.area;

% Compute the fitted values
fitted = X * beta;

% Compute residuals
residuals = data.area - fitted;

% Create residual plot
figure;
scatter(fitted, residuals);
title('Residuals vs Fitted');

% Compute correlation matrix
dataNumeric = data(:, {'temp', 'RH', 'wind', 'area'});
corrMatrix = corr(dataNumeric{:,:}, 'Rows', 'complete');

% Create correlation matrix heatmap
figure;
heatmap(corrMatrix);
title('Correlation Matrix Heatmap');
```

```
% Load the data
data = readtable('forestfires.csv');

% Remove rows with missing data
data = rmmissing(data);

% Fit the linear model
lm = fitlm(data, 'area ~ temp + RH + wind');

% Display the  results
disp(lm)

% Compute the fitted values
fitted = lm.Fitted;

% Create  figure for Area vs Temperature
figure;
scatter(data.temp, data.area, 'filled', 'MarkerEdgeColor', 'k');
hold on;
plot(data.temp, fitted, 'r-', 'LineWidth', 1.5);
xlabel('Temperature');
ylabel('Area');
title('Area vs Temperature');
grid on;  % Add a grid for better visibility
hold off;

% Create  figure for Area vs Relative Humidity
figure;
scatter(data.RH, data.area, 'filled', 'MarkerEdgeColor', 'k');
hold on;
plot(data.RH, fitted, 'g-', 'LineWidth', 1.5);
xlabel('Relative Humidity');
ylabel('Area');
title('Area vs Relative Humidity');
grid on;
hold off;

% Create figure for Area vs Wind
figure;
scatter(data.wind, data.area, 'filled', 'MarkerEdgeColor', 'k');
```

```
hold on;
plot(data.wind, fitted, 'b-', 'LineWidth', 1.5);
xlabel('Wind');
ylabel('Area');
title('Area vs Wind');
grid on;
hold off;
```

References

1. UCI Machine Learning Repository: Forest Fires Data Set. [Online] Available: http://archive.ics.uci.edu/ml/datasets/Forest+Fires

2. UCI Machine Learning Repository: Forest Fires Data Set, Data Folder. [Online] Available: http://archive.ics.uci.edu/ml/machine-learning-databases/forest-fires