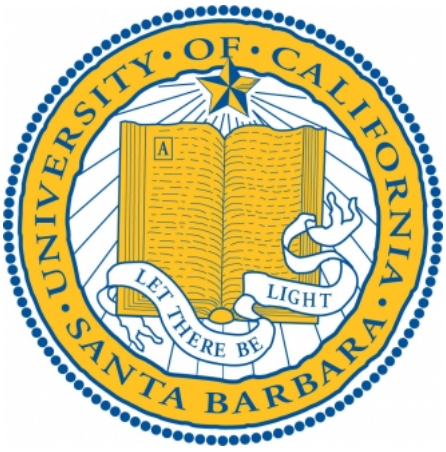
Regression Analysis: Forest Fires

PSTAT 126

Austin Mac



University of California, Santa Barbara
Fall 2019

Introduction

This regression analysis is based around the Forest Fires data set which can be found on the University of California, Irvine's Machine Learning Repository. Attributes of this set include: x & y spatial coordinates, month, date, Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), temperature, relative humidity (RH), wind, rain, and area. Using ISI as the response, this project aims to analyze and interpret relationships between ISI and the other attributes as predictors.

Definitions

The following definitions can be found on the National Wildfire Coordinating Group website.
Fine Fuel Moisture Code: "represents fuel moisture of forest litter fuels under the shade of a forest canopy."
Duff Moisture Code: "represents fuel moisture of decomposed organic material underneath the litter."
Drought Code: "represents drying deep into the soil."

Research Questions

Question 1: Does the effect of wind on ISI depend on the month?

Question 2: What is the mean ISI for areas with the minimum, average, and maximum wind speed, after taking into account average FFMC and DC on a Monday in August?

Regression Methods

To answer each of the research questions, we fit a model with the appropriate predictors. The fitting process involves performing step-wise regression. We conduct multiple partial F-tests and select the most appropriate predictors for the model based off of Akaike's Information Criterion (AIC). Then high influence outliers and leverage points are removed to ensure our model is the one of best fit. Finally, we transform the model to best normalize the distribution of residuals and approximate linearity.

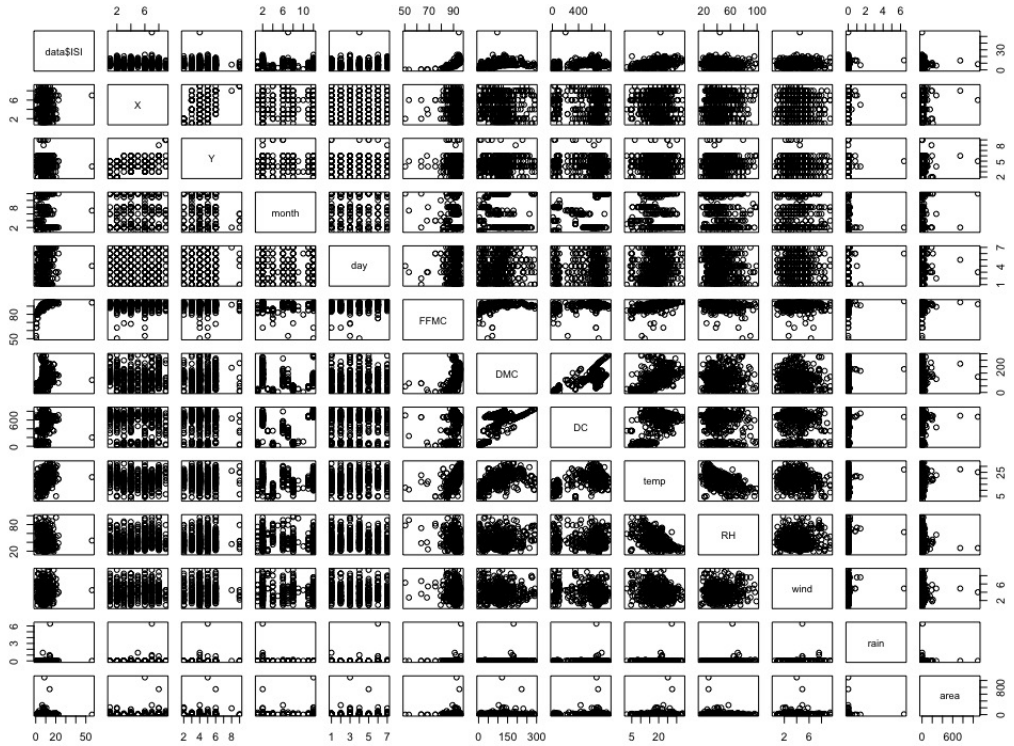
Question 1: To determine whether month is significant in the effect of wind on ISI, we conduct an overall F-test to compare the slopes of interaction terms of month and wind. If at least one of the slope parameters differs significantly from 0, we conclude that the effect of wind on ISI depends on month.

Question 2: To determine the expected ISI, we create a prediction interval which returns a bound which contains the expected ISI with 95% confidence.

Regression Analysis, Results, and Interpretation

Initial Observations

From the initial observation of the scatterplot matrix, it appears that the predictor that has the strongest correlation with the response is FFMC. The correlation matrix confirms this observation, since the correlation is 0.59070367, the highest correlation between any predictor and the response. Additionally, there appears to be some multicollinearity between DMC and DC. The correlation matrix illustrates a correlation of 0.68060366 between DMC and DC, the highest correlation between any two predictors.



Coding Categorical Variables

In order to prepare the data set for regression analysis, we code the categorical variables which correspond to each month.

Month	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11
Jan	1	0	0	0	0	0	0	0	0	0	0
Feb	0	1	0	0	0	0	0	0	0	0	0
Mar	0	0	1	0	0	0	0	0	0	0	0
Apr	0	0	0	1	0	0	0	0	0	0	0
May	0	0	0	0	1	0	0	0	0	0	0
Jun	0	0	0	0	0	1	0	0	0	0	0
Jul	0	0	0	0	0	0	1	0	0	0	0
Aug	0	0	0	0	0	0	0	1	0	0	0
Sep	0	0	0	0	0	0	0	0	1	0	0
Oct	0	0	0	0	0	0	0	0	0	1	0
Nov	0	0	0	0	0	0	0	0	0	0	1
Dec	0	0	0	0	0	0	0	0	0	0	0

We perform the same process for each day of the week.

Day	x1	x2	x3	x4	x5	x6
Mon	1	0	0	0	0	0
Tue	0	1	0	0	0	0
Wed	0	0	1	0	0	0
Thu	0	0	0	1	0	0
Fri	0	0	0	0	1	0
Sat	0	0	0	0	0	1
Sun	0	0	0	0	0	0

Variable Selection

In order to obtain a correctly specified model, and reduce the number of predictors, we perform stepwise regression using partial F-tests and prioritizing the lowest Akaike's Information Criterion (AIC). We arrive at the following model:

$$E(ISI) = \beta_0 + \beta_1 \text{FFMC} + \beta_2 \text{aug} + \beta_3 \text{jun} + \beta_4 \text{wind} + \beta_5 \text{mon} + \beta_6 \text{thu} + \beta_9 \text{sep} + \beta_{10} \text{jul} + \beta_{11} \text{DC} + \beta_{12} \text{oct}$$

Influential Points

In order to clean the data set to better fit our model, we identify any outliers and high leverage points. Then, out of the outliers and high leverage points, we identify any which are also influential.

Outliers

First, we identify the externally studentized residuals of the observations in the data set. If the residual of any of these observations is greater than 3, then the offending observation is flagged as an outlier. In order to determine if a particular outlier is influential, we examine its difference in fits value. If its difference in fits value is greater than $2\sqrt{\frac{10+1}{516-10-1}} = 0.2951757$, then the point is influential. Comparing externally studentized residuals and difference in fits values, we find that the following outliers are influential:

12 23 313

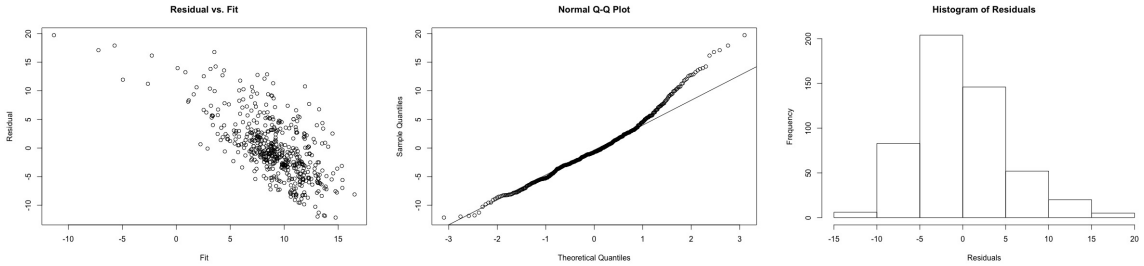
High Leverage Points

Similarly, we identify the leverage of each observation in the data set. If the leverage is greater than 3 times the mean leverage value of the set, then we flag the observation as a high leverage points. To determine if a high leverage point is influential, we compare its difference in fits value to 0.2951757, similarly to above. We find that the influential high leverage points are:

5 13 23 76 98 131 184 200 227 274 300 313 415 443 455 470 475 499

Initial Residual and Normality Analysis

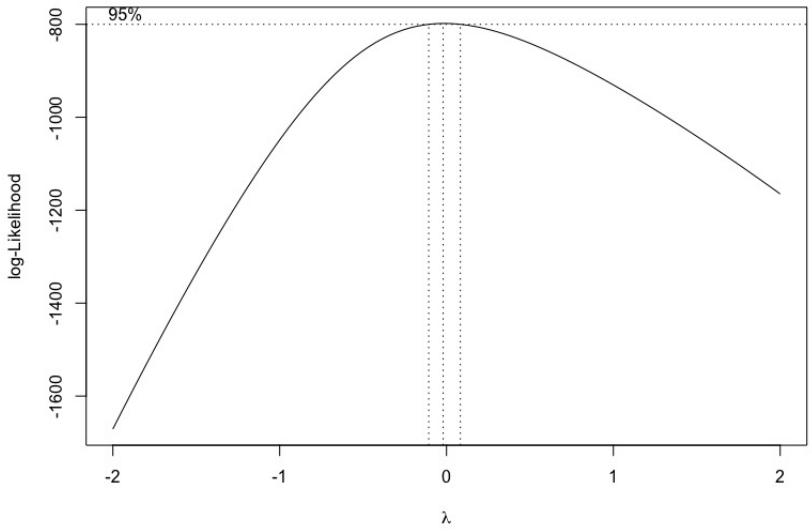
After removing high influence outliers and leverage points, we begin transforming our model by using the residual vs. fit plot, Normal Q-Q plot, and a histogram of residuals as diagnostic tools.



Immediately, we notice issues with all of our plots. The residual vs. fit plot illustrates a strong downward slope, illustrating non-linearity with our model. The Normal Q-Q plot exhibits some non-linearity, and the histogram of residuals is slightly right skewed. Examining the p-value from a Shapiro-Wilk normality test, we observe that $p = 3.027 \times 10^{-9}$. Thus, the distribution is not normal. Further analysis illustrates a multiple R^2 value of 0.4732 and an adjusted R^2 value of 0.4628. Since adjusted R^2 does not increase with the number of predictors added, we can use adjusted R^2 for the interpretation that 46.25% of the variation in the response can be explained by variation in the predictors.

Transformations

An initial Box-Cox diagnosis illustrates that the log-Likelihood is maximized when $\lambda \approx 0$, so we apply a log transformation to our response.

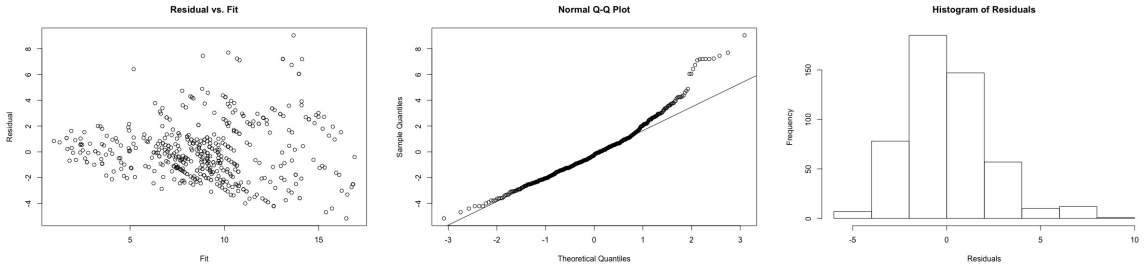


Applying a log transformation to our response significantly worsens our residual vs. fit plot, so instead we proceed with a polynomial model:

$$\begin{aligned} E(ISI) = \beta_0 &+ \beta_1 FFMC^3 + \beta_2 FFMC^2 + \beta_3 FFMC \\ &+ \beta_4 aug + \beta_5 jun + \beta_6 wind + \beta_7 mon + \beta_8 thu + \beta_9 sep \\ &+ \beta_{10} jul + \beta_{11} DC + \beta_{12} oct \end{aligned}$$

Since we include FFMC^3 as a significant predictor in our model, by the Hierarchical Principle, we must include the lower order FFMC terms as well.

By reviewing our diagnostic plots, it is evident that our transformations have made changes.



The points in our residual vs. fit plot have formed a more even horizontal band around the 0 line, indicating better linearity. However, there appears to be a subtle fanning effect, indicating non-constant residual variance. This is supported by a decrease in linearity of our Normal Q-Q plot and an increase in skew in our histogram. Additionally, the p-value of the Shapiro-Wilk Normality test decreased to $p = 4.178 \times 10^{-11}$. Analysis of the summary table for this model indicates a significant increase in the multiple R^2 value, from 0.4732 to 0.6828 and a significant increase in the adjusted R^2 value from 0.4628 to 0.6749. We can now assume that 67.49% of the variation in the response can be attributed to variation in the predictors.

Due to the increase in linearity, and increase in R^2 value, we settle with our final model as:

$$\begin{aligned} E(ISI) = \beta_0 &+ \beta_1 FFMC^3 + \beta_2 FFMC^2 + \beta_3 FFMC \\ &+ \beta_4 aug + \beta_5 jun + \beta_6 wind + \beta_7 mon + \beta_8 thu + \beta_9 sep \\ &+ \beta_{10} jul + \beta_{11} DC + \beta_{12} oct \end{aligned}$$

Research Questions

1. Does the effect of wind on ISI depend on the month?

We begin by formulating the model which includes the interaction terms between our chosen predictor months and wind.

$$\begin{aligned} E(ISI) = \beta_0 &+ \beta_1 FFMC^3 + \beta_2 FFMC^2 + \beta_3 FFMC \\ &+ \beta_4 aug + \beta_5 jun + \beta_6 wind + \beta_7 mon + \beta_8 thu + \beta_9 sep \\ &+ \beta_{10} jul + \beta_{11} DC + \beta_{12} oct + \beta_{13}(wind \times aug) \\ &+ \beta_{14}(wind \times jun) + \beta_{15}(wind \times sep) + \beta_{16}(wind \times jul) \\ &+ \beta_{17}(wind \times oct) \end{aligned}$$

Given this model, we formulate our null and alternative hypothesis from reduced and full models.

$$\begin{aligned} &H_0 : \beta_{13} = \beta_{14} = \beta_{15} = \beta_{16} = \beta_{17} = 0 \\ &H_1 : \text{At least one } \beta_{13}, \beta_{14}, \beta_{15}, \beta_{16}, \beta_{17} \neq 0 \end{aligned}$$

The resulting ANOVA table summarizes the results of our overall F-test.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	496	2547.3			
2	491	2512.5	5	34.887	1.3636	0.2366

The p-value from our F-test, 0.2366, is well above our α value of 0.05. Thus, we fail to reject the our reduced model in favor of the full model. There is no significant impact of month on the effect of wind on ISI.

2. What is the mean ISI for areas with the minimum, average, and maximum wind speed, after taking into account average FFMC and DC on a Monday in August?

We perform three different calculations for confidence intervals, changing only the wind speed while keeping FFMC, DC, day, and month constant. First we calculate a 95% confidence interval for the ISI given minimum wind speed:

```
fit       lwr       upr
7.809869 7.015608 8.604129
```

Thus we are 95% confident that the mean ISI is between 7.015 and 8.604 given minimum wind speed, average FFMC, DC, and an August Monday. Note ISI is a unitless measure. Calculating a 95% confidence interval for ISI given the mean wind speed:

```
fit       lwr       upr
8.99943 8.291687 9.707174
```

We are 95% confident that the mean ISI given mean wind speed and the same conditions as above is between 8.291 and 9.707. Finally, we calculate a 95% confidence interval for ISI given maximum wind speed:

```
fit       lwr       upr
10.79261 9.807299 11.77792
```

We are 95% confident that the mean ISI for maximum wind speed and the same conditions as above is between 9.807 and 11.777.

The confidence intervals given the minimum and maximum wind speeds are the widest since the x_h values used, min(Wind) and max(Wind), are furthest away from the mean wind speed.

Conclusion

The formulated model was used to determine the effect of the interaction between wind and month. Additionally, the model was used to predict intervals for ISI given specific conditions. Results showed that there was no significant impact on ISI from the interaction of wind and month. Furthermore, prediction intervals with higher wind speeds appear to be correlated with greater ISI values. For practical use, the model formulated could be used to predict the rate of spread of a fire, providing valuable insight to first responders.

Appendix

```
data <- read.csv("forestfires.csv")
data <- data[-c(380),] # remove row where ISI = 0
#pairs(cbind(data$ISI,subset(data, select = -c(ISI))))
#cor(cbind(data$ISI,subset(data, select = -c(ISI, month, day))))

data$jan = as.integer(data$month == "jan")
data$feb = as.integer(data$month == "feb")
data$mar = as.integer(data$month == "mar")
data$apr = as.integer(data$month == "apr")
data$may = as.integer(data$month == "may")
data$jun = as.integer(data$month == "jun")
data$jul = as.integer(data$month == "jul")
data$aug = as.integer(data$month == "aug")
data$sep = as.integer(data$month == "sep")
data$oct = as.integer(data$month == "oct")
data$nov = as.integer(data$month == "nov")

data$mon = as.integer(data$day == "mon")
data$tue = as.integer(data$day == "tue")
data$wed = as.integer(data$day == "wed")
data$thu = as.integer(data$day == "thu")
data$fri = as.integer(data$day == "fri")
data$sat = as.integer(data$day == "sat")


fit <- lm(data$ISI ~ data$X+data$Y+data$jan+
            data$feb+
            data$mar+
            data$apr+
            data$may+
            data$jun+
            data$jul+
            data$aug+
            data$sep+
            data$oct+
            data$nov+
            data$mon+
            data$tue+
            data$wed+
            data$thu+
            data$fri+
            data$sat+
            data$FFMC+data$DMC+data$DC+data$temp+
            data$RH+data$wind+data$rain+data$area)

# Stepwise Regression
mod0 <- lm(data$ISI ~ 1)
mod.upper <- lm(data$ISI ~ data$X+data$Y+data$jan+
                  data$feb+
                  data$mar+
                  data$apr+
                  data$may+
                  data$jun+
                  data$jul+
                  data$aug+
                  data$sep+
```

```
                  data$oct+
                  data$nov+
                  data$mon+
                  data$tue+
                  data$wed+
                  data$thu+
                  data$fri+
                  data$sat+
                  data$FFMC+data$DMC+data$DC+data$temp+data$RH
                  +data$wind+data$rain+data$area)

step(mod0, scope = list(lower = mod0, upper = mod.upper))

fit <- lm(data$ISI ~ data$FFMC + data$aug + data$jun + data$wind +
data$mon + data$thu + data$sep + data$jul + data$DC +
data$oct)

# Influential Outliers/Leverage Points
high_leverage <- which(hatvalues(fit) > 3*10/516)
outlier <- which(abs(rstudent(fit)) > 3)

influential <- which(dffits(fit) > 0.2951757)

intersect(high_leverage, influential)
intersect(outlier, influential)

# Removal of Leverage Points
data <- data[-c(5, 13, 23, 76, 98, 131, 184, 200, 227,
274, 300, 313, 415, 443, 455, 470, 475, 499, 12, 23, 313),]

# Residual/ Normality Analysis Initial

yhat <- fitted(fit)
resid <- data$ISI - yhat
plot(resid ~ yhat, main = "Residual vs. Fit", xlab = "Fit",
ylab = "Residual")
qqnorm(resid)
qqline(resid)
hist(resid, main = "Histogram of Residuals", xlab = "Residuals")
summary(fit)
shapiro.test(resid)

library(MASS)
boxcox(data$ISI ~ data$FFMC + data$aug + data$jun + data$wind +
data$mon + data$thu + data$sep + data$jul + data$DC + data$oct)
# Fit2 diagnostics (Post Transformation)

fit2 <- lm(data$ISI ~ I(data$FFMC^3) + I(data$FFMC^2) + data$FFMC +
data$aug + data$jun+ data$wind + data$mon + data$thu + data$sep
+ data$jul + data$DC + data$oct)
yhat <- fitted(fit2)
resid <- data$ISI - yhat
plot(resid ~ yhat, main = "Residual vs. Fit", xlab = "Fit",
ylab = "Residual")
qqnorm(resid)
qqline(resid)
hist(resid, main = "Histogram of Residuals", xlab = "Residuals")
summary(fit2)
```

```
shapiro.test(resid)

#Question 1
x1 <- I(data$FFMC^3)
x2 <- I(data$FFMC^2)
x3 <- data$FFMC
x4 <- data$aug
x5 <- data$jun
x6 <- data$wind
x7 <- data$mon
x8 <- data$thu
x9 <- data$sep
x10 <- data$jul
x11 <- data$DC
x12 <- data$oct

mod_red <- lm(data$ISI ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12)
mod_full <- lm(data$ISI ~
x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+I(data$wind*data$aug) +
I(data$wind*data$jun) + I(data$wind*data$sep) +
I(data$wind*data$jul) + I(data$wind*data$oct))

anova(mod_red, mod_full)

#Question 2
meanFFMC <- mean(data$FFMC)
meanWind <- mean(data$wind)
meanDC <- mean(data$DC)

new0 <- data.frame(x1 = meanFFMC^3,
                   x2 = meanFFMC^2,
                   x3 = meanFFMC,
                   x4 = 1,
                   x5 = 0,
                   x6 = min(data$wind),
                   x7 = 1,
                   x8 = 0,
                   x9 = 0,
                   x10 = 0,
                   x11 = meanDC,
                   x12 = 0)

PImin <- predict(fit, new0, se.fit = TRUE, interval = "prediction",
level = 0.95, type = "response")
PImin$fit

new <- data.frame(x1 = meanFFMC^3,
                   x2 = meanFFMC^2,
                   x3 = meanFFMC,
                   x4 = 1,
                   x5 = 0,
                   x6 = mean(data$wind),
                   x7 = 1,
                   x8 = 0,
                   x9 = 0,
                   x10 = 0,
                   x11 = meanDC,
                   x12 = 0)
```

```
PImean <- predict(fit, new, se.fit = TRUE,
interval = "prediction", level = 0.95, type = "response")
PImean$fit

new2 <- data.frame(x1 = meanFFMC^3,
                   x2 = meanFFMC^2,
                   x3 = meanFFMC,
                   x4 = 1,
                   x5 = 0,
                   x6 = max(data$wind),
                   x7 = 1,
                   x8 = 0,
                   x9 = 0,
                   x10 = 0,
                   x11 = meanDC,
                   x12 = 0)

PImax <- predict(fit, new2, se.fit = TRUE,
interval = "prediction", level = 0.95, type = "response")
PImax$fit
```