Stat 151A Final Project Report

Yue (Julien) Yu

SID: 25234669

Content

(Guide to read the rmd file)

1. Introduction
2. Description of untransformed data
    - a. SLR/MLR of quantitative explanatory variables with respect to “area”
    - b. Linear regression assumptions for the untransformed data
3. Data transformation
    - a. Transformation of the response variable “area”
    - b. R-squared and linear regression assumptions after log transformation of "area"
    - c. Analysis of outlier and high leverage point
    - d. Power transformation of explanatory variables
4. Categorical variables
    - a. Analysis of the categorical variable “month”
    - b. Analysis of the categorical variable “day”
    - c. Analysis of spatial coordinates “X” and “Y”
5. Information criteria, comparison of models and reduction of explanatory variables
    - a. Dummy variables to represent “month” and “day”
    - b. Exclusion of explanatory variables: information criteria and adjusted R-squared
6. Conclusions and final model
7. Reference

I. Introduction

My task is to analyze a forest fire dataset, with 1 response variable “area” denoting the total burned area, 12 explanatory variables, and 517 observations of fires. My objective is to find out whether each of the 12 explanatory variables significantly affects the response variable “area” or not. The 12 explanatory variables are divided into 2 categories: 8 quantitative variables (4 FWI component variables and 4 weather variables) and 4 categorical variables (4 spatial and temporal variables). In my discussion, I will first approach the quantitative explanatory variables, and then the categorical variables.

II. Description of untransformed Data

I did a couple of analyses on the response variable “area” and the 8 quantitative variables. Analyses include fitting simple and multiple linear regression models to the original data, testing the linear regression assumptions using different graphs, and deciding on which transformation I should use for further analysis. For instance, there are 3 different ways to generate the estimated slope of a quantitative explanatory variable: to fit an 8-variable multiple linear regression model, to fit a 4-variable multiple linear regression model by FWI or weather, and to fit a simple linear regression model. I then compare the estimated slopes and their signs to see the direction each variable affects the response variable “area” (Chart 1). From the chart of variance inflation factors (Chart 2), “temp”, “DMC” and “DC” show some collinearity with other variables. I need to fix the collinearity either through transformation or by excluding these variables. The QQ-Plot (Graph 1) shows that the residuals are right-skewed, and the Component residual plots (Graph 2) show that outliers exist and that the simple linear regressions are very often nonlinear. The outlier test (Chart 3) shows that row 239 and 416 are very distinctive outliers. Last but not the least, the plot

of influential points (Graph 3) shows that row 500 is with very high leverage but low residuals and is thereby worth further observation.

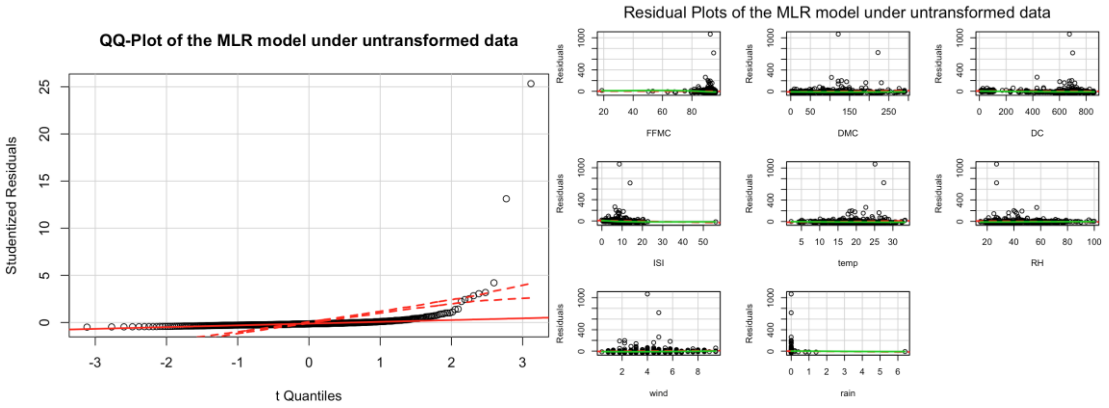
```
##      orig_8 orig_4by4 orig_SLR
## FFMC -0.023    0.3270    0.463
## DMC   0.076    0.0730    0.073
## DC   -0.005   -0.0009    0.013
## ISI  -0.698   -0.3960    0.115
## temp  0.847    1.0100    1.073
## RH   -0.196   -0.1100   -0.295
## wind  1.527    1.2790    0.438
## rain -2.540   -2.8300   -1.584
```

Chart 1: Comparison of estimated slopes.

```
vif(MLR_total)
```

```
##     FFMC      DMC       DC      ISI     temp       RH     wind     rain
## 1.695255 2.330688 2.078205 1.578258 2.661897 1.899989 1.140610 1.044801
```

Chart 2: List of variance inflation factors to measure collinearity.



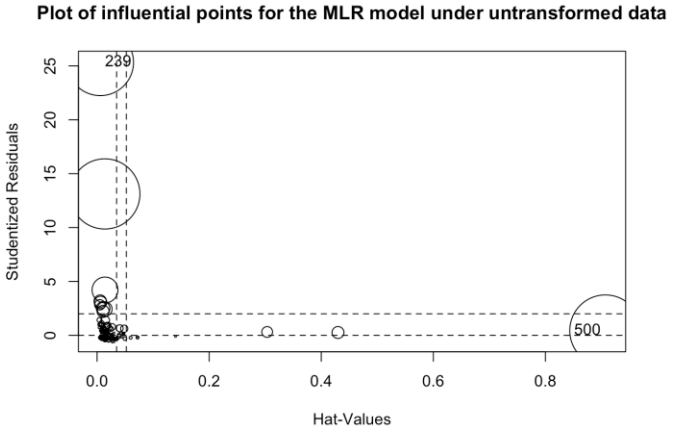
Graph 1: QQ-Plot and skewness to the right.

Graph 2: Component and residual plots of simple linear regressions to test the linearity assumption and to visually represent the outliers.

```
outlierTest(MLR_total) # Row 239, 416 are very distinctive outliers.
```

```
##      rstudent unadjusted p-value Bonferonni p
## 239 25.331541         4.2991e-92   2.2226e-89
## 416 13.121836         4.5256e-34   2.3397e-31
## 480  4.208168         3.0453e-05   1.5744e-02
```

Chart 3: Outlier test shows that row 239 and 416 are very distinctive outliers.



Graph 3: Plot of influential points shows that row 500 is worth further observation.

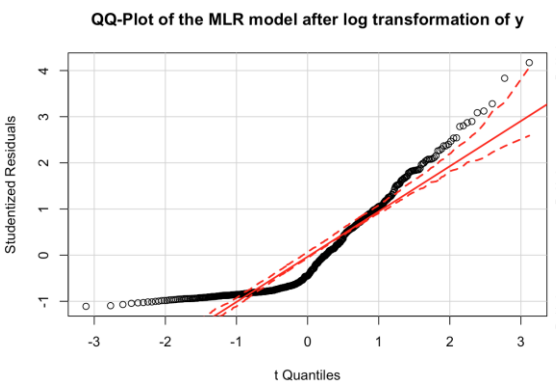
III. Data transformation

The QQ-Plot, residual plots, outlier test and plot of influential points all show that the response variable "area" needs to go down the ladder of transformation. I try the log, square root and cube root transformation and decide to use log transformation for "area". The default model is now MLR_log_total =lm(log_area ~ FFMC + DMC + DC + ISI + temp + RH + wind + rain). I generate similar graphs as in Part II, and it turns out that the log model fits much better to the data than the original model. The multiple R-squared increases as I change from the original multiple linear regression model to the log model (Chart 4).

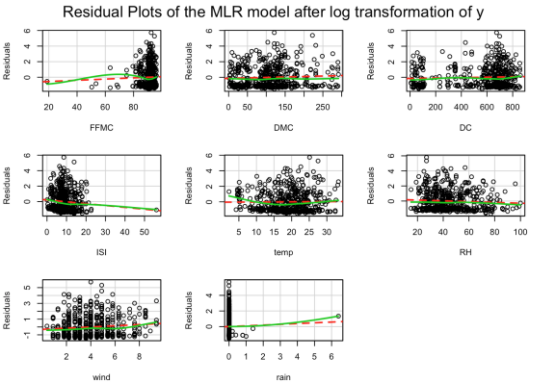
```
##                MLR_total
## log_r_square  0.01988121
## orig_r_square 0.01601285
```

Chart 4: Multiple r-squared before and after log transformation.

Meanwhile, variance inflation factors stay almost the same. The QQ-Plot becomes better, though not perfectly fitted (Graph 4). The component residual plot becomes better in terms of outliers, but shows that certain explanatory variables, such as “FFMC” and “temp”, should either be transformed or be excluded (Graph 5). Also, one single high leverage point (row 500) affects the regression line in the “rain” graph. The outlier at row 416 disappears, and the outlier at row 239 becomes far less significant (Chart 5). Row 500 remains to be a high leverage point because of the heavy rain, and is more obvious visually (Graph 6).



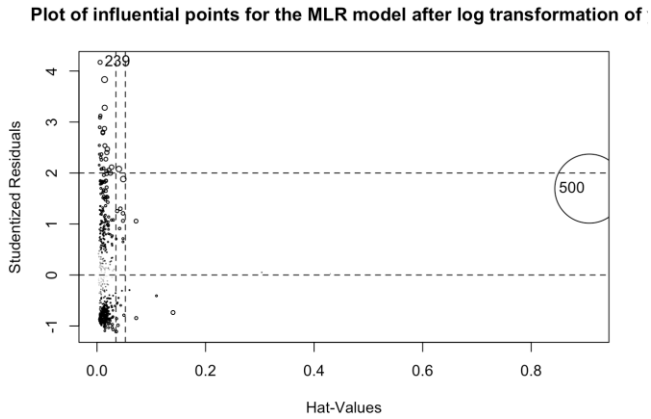
Graph 4: QQ-Plot after log transformation.



Graph 5: Residual plot after log transformation.

```
##     rstudent unadjusted p-value Bonferonni p
## 239 4.171235          3.564e-05      0.018426
```

Chart 5: Outlier test after log transformation.



Graph 6: Plot of influential points after transformation.

I then look at the high leverage point in row 500 (with extreme “rain” variable and regular “area” variable). I decide not to remove that data point because it is the only scenario of a “heavy rain”. Instead, 1-2 dummy variables can be used to replace “rain”:

• rain_1 = 0 for any day without rain, rain_1 = 1 for any day with rain

• (Pending) rain_2 = 0 for any day without heavy rain (>2mm/30min), rain_2 = 1 for any day with heavy rain. Only row 500 has rain_2 = 1 in the dataset.

Last but not the least, power transformations should be made to explanatory variables “FFMC”, “DC” and “temp” to redress the nonlinearity. It turns out that the transformations of these explanatory variables largely increase the multiple R-squared (Chart 6). The default model now becomes MLR_xtransf_total =lm(log_area ~ FFMC^0.4 + DMC + DC^0.1 + ISI + temp^4.5 + RH^0.05 + wind + rain_1 + rain_2), and the adjusted R-squared almost quadruples.

```
summary(MLR_xtransf_total)[9]
```

```
## $adj.r.squared
## [1] 0.01879188
```

```
summary(MLR_log_total)[9]
```

```
## $adj.r.squared
## [1] 0.004446264
```

Chart 6: Adjusted r-squared before and after x transformations.

From then on, adjusted R-squared becomes an important criterion to compare different regression models. I no longer use multiple R-squared because the number of explanatory variables in each model starts to vary from here.

IV. Categorical variables

Having considered the effects of quantitative explanatory variables, I then look at the effects of categorical variables: “month”, “day” and spatial coordinates “X” and “Y”. I classify the months (Chart 7) into 3 levels of risk (high: 8, 9; medium: 3, 7; low: 1, 2, 4, 5, 6, 10, 11, 12) and the days (Chart 8) into 2 levels of risk (high: Fri – Sun; low: Mon – Thu). Therefore, at most 2 dummy variables can be added to represent “month” and 1 dummy variable to represent “day”. Dummy variables are defined as follow:

• month_1 = 0 for month with low risk, month_1 = 1 for month with medium or high risk.

• month_2 = 0 for month with low or medium risk, month_2 = 1 for month with high risk.

• day_0 = 0 for weekdays (Mon - Thu), day_0 = 1 for weekends (Fri - Sun).

	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec
monthly big fires	0	10	19	4	1	8	18	99	97	5	0	9
monthly total fires	2	20	54	9	2	17	32	184	172	15	1	9

Chart 7: Number of fires and number of big fires (>100 square meters) by month

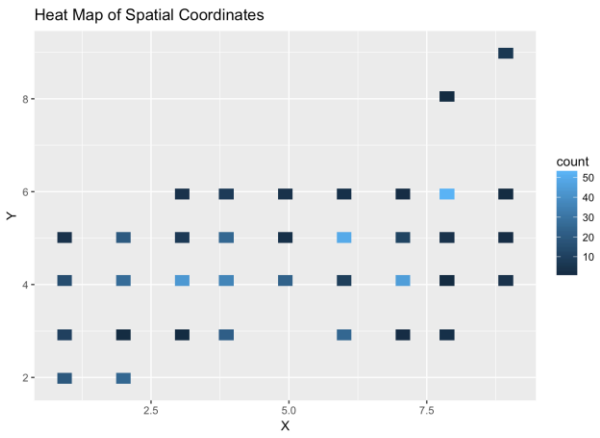
	mon	tue	wed	thu	fri	sat	sun
daily big fires	39	36	32	31	43	42	47
daily total fires	74	64	54	61	85	84	95

Chart 8: Number of fires and number of big fires (>100 square meters) by day

I introduce a heat map to denote whether a certain location has high risk of fire, classifying the spatial coordinates (Graph 7) also into 3 levels of risk (high: light blue; medium: blue; low: dark blue). 2 dummy variables are defined, though not added to the default model.

• coord_1 = 0 for low risk, coord_1 = 1 for medium or high risk.

• coord_2 = 0 for low or medium risk, coord_2 = 1 for high risk.



Graph 7: Heat map for spatial coordinates

The default model then becomes MLR_all_refined =lm(log_area ~ FFMC^0.4 + DMC + DC^0.1 + ISI + temp^4.5 + RH^0.05 + wind + rain_1 + rain_2 + month_1 + month_2 + day_0). There are too many explanatory variables in this model and I am using techniques of information criteria to exclude some of the insignificant variables.

V. Information criteria and comparison of models

In this stage, I am no longer transforming or including variables, but excluding variables. An explanatory variable will be excluded if the model without it simultaneously has higher adjusted R-squared and lower information criteria (both AIC and BIC). 7 proposals are tested, each of which denotes the outcomes of excluding an explanatory variable. Proposals 2, 4, 5, 6, which seek to exclude “DC”, “DMC”, “RH” and “day_0” respectively, all dominate over the default model (row “original”) with smaller information criteria and larger adjusted R-squared. Proposal 7, which seeks to take spatial coordinates “X” and “Y” into consideration, fails both in terms of information criteria and adjusted R-squared. Hence, the final proposal excludes “DC”, “DMC”, “RH”, “day_0” and spatial coordinates. The default model then becomes MLR_all_refined_f =lm(log_area ~ FWI_0 + temp^4.5 + wind + rain_1 + rain_2 + month_1 + month_2), where FWI_0 = FFMC^0.4 + ISI.

```
##                       AIC      BIC Adjusted R-Squared
## Proposal 1      350.7259 401.7024          0.01514358
## Proposal 2      349.2199 400.1964          0.01800837
## Proposal 3      350.1991 401.1756          0.01614664
## Proposal 4      349.1634 400.1399          0.01811560
## Proposal 5      348.8175 399.7940          0.01877238
## Proposal 6      348.7868 399.7633          0.01883063
## Proposal 7      354.0986 417.8193          0.01427992
## original        350.7863 406.0108          0.01688480
## Proposal Final 344.1470 378.1313          0.02019367
```

Chart 9: List of proposals and the exclusion of insignificant explanatory variables

In "MLR_all_refined_f", "rain" was split into two dummy variables to denote the possibilities of no, small and heavy rains, and "month" was split into two dummy variables to denote 3 different levels of risk of fire. P-values of all remaining explanatory variables are also acceptable (<0.25). However, the signs of "rain_2" (positive) and "month_1" (negative) are very counterintuitive because rain_2 = 1 for heavy rains and month_1 = 1 for months with higher risk of fire. One possible reason may be that they are largely collinear with "rain_1" and "month_2" respectively. Therefore, I decide to exclude "rain_2" and "month_1" from the final model, though the adjusted R-squared decreases a little.

```
summary(MLR_all_refined_f)
```

```
##
## Call:
## lm(formula = log_area ~ FWI_0 + newtemp + wind + rain_1 + rain_2 +
##     month_1 + month_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9316 -1.0531 -0.5952  0.8543  5.6633
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.533e-01  2.647e-01   3.602 0.000347 ***
## FWI_0       -1.922e-02  1.460e-02  -1.316 0.188767
## newtemp      1.535e-07  5.918e-08   2.595 0.009739 **
## wind         8.621e-02  3.556e-02   2.425 0.015668 *
## rain_1      -9.891e-01  5.332e-01  -1.855 0.064179 .
## rain_2       2.020e+00  1.485e+00   1.360 0.174381
## month_1     -2.630e-01  2.207e-01  -1.192 0.233843
## month_2      2.746e-01  1.717e-01   1.600 0.110288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.384 on 509 degrees of freedom
## Multiple R-squared:  0.03349,    Adjusted R-squared:  0.02019
## F-statistic: 2.519 on 7 and 509 DF,  p-value: 0.01487
```

Chart 10: "MLR_all_refined_f" model.

```
summary(MLR_all_refined_f_v)
```

```
##
## Call:
## lm(formula = log_area ~ FWI_0 + newtemp + wind + rain_1 + month_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9402 -1.0506 -0.6542  0.8959  5.6571
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.357e-01  2.467e-01   3.388  0.00076 ***
## FWI_0       -2.097e-02  1.456e-02  -1.440  0.15046
## newtemp      1.564e-07  5.904e-08   2.649  0.00833 **
## wind         8.527e-02  3.560e-02   2.395  0.01697 *
## rain_1      -7.603e-01  4.996e-01  -1.522  0.12867
## month_2      1.594e-01  1.415e-01   1.126  0.26059
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.386 on 511 degrees of freedom
## Multiple R-squared:  0.02712,    Adjusted R-squared:  0.0176
## F-statistic: 2.849 on 5 and 511 DF,  p-value: 0.01504
```

Chart 11: "MLR_all_refined_f_v" (final model).

VI. Conclusions and final model

• Final model (Chart 11): MLR_all_refined_f_v =lm(log_area ~ FWI_0 + newtemp + wind + rain_1 + month_2) where FWI_0 = FFMC^0.4 + ISI and newtemp = temp^4.5

• FFMC (fine fuel moisture code), ISI (initial spread index), temp (outside temperature), wind (outside wind speed), rain (outside rain) and month (specific month of the year) are variables that significantly affect the total burned area.

• The adjusted R-squared (0.0176) is still low, but is comparatively higher than:

lm(area ~ FFMC + DMC + DC + ISI + temp + RH + wind + rain): 0.000517

lm(log_area ~ FFMC + DMC + DC + ISI + temp + RH + wind + rain): 0.004446

lm(log_area ~ FFMC^0.4 + DMC + DC^0.1 + ISI + temp^4.5 + RH^0.05 + wind + rain_1 + rain_2 + month_1 + month_2 + day_0): 0.0168848

• The AIC and BIC of the final model are lower than those of any other model (Chart 12).

• The variance inflation factors are closer to 1 than any other model (Chart 13).

• The residual plots are mostly linear, except for the residuals vs rain_1 (Graph 8). That specific graph is nonlinear because the outlier in row 239 cannot be fully represented by 1 dummy variable. That nonlinearity is therefore inevitable unless the outlier is excluded.

• Explanatory variables in the final model have acceptable significance level: the intercept and “newtemp” are highly significant ($p < 0.01$), “wind” is significant ($p < 0.05$), “rain_1” and “FWI_0” have $p < 0.2$, and “month_2” has $p < 0.3$.

• Question: Is the model going to become much more fitted if row 239 (outlier) and row 500 (high leverage point) are taken away?

• Answer (Chart 14): the “rain_1” variable becomes more significant, and the adjusted R-squared only increases a little (from 0.0176 to 0.0179).

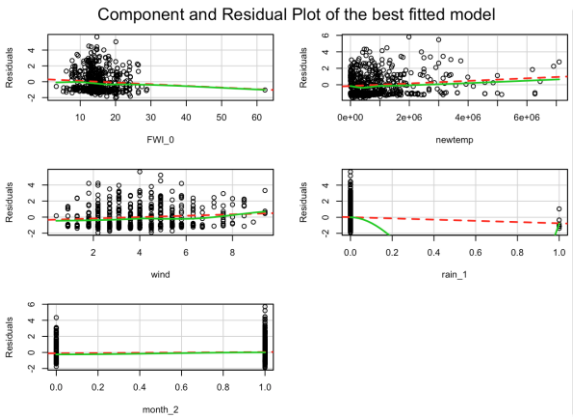
```
##                              AIC      BIC Adjusted R-Squared
## Proposal 1           350.7259 401.7024          0.01514358
## Proposal 2           349.2199 400.1964          0.01800837
## Proposal 3           350.1991 401.1756          0.01614664
## Proposal 4           349.1634 400.1399          0.01811560
## Proposal 5           348.8175 399.7940          0.01877238
## Proposal 6           348.7868 399.7633          0.01883063
## Proposal 7           354.0986 417.8193          0.01427992
## original             350.7863 406.0108          0.01688480
## Proposal Final       344.1470 378.1313          0.02019367
## Final Verified Model 343.5404 369.0286          0.01760149
```

Chart 12: IC of the final verified model.

```
vif(MLR_all_refined_f_v)
```

```
##     FWI_0  newtemp     wind   rain_1  month_2
## 1.229636 1.192271 1.092585 1.023158 1.155243
```

Chart 13: VIF of the final verified model.



Graph 8: Residual plots of the final model.

```
summary(MLR_all_refined_f_v)
```

```
##
## Call:
## lm(formula = log_area ~ FWI_0 + newtemp + wind + rain_1 + month_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8479 -1.0465 -0.6498  0.8906  5.1970
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.327e-01  2.426e-01   3.432 0.000647 ***
## FWI_0       -1.940e-02  1.432e-02  -1.354 0.176227
## newtemp      1.407e-07  5.826e-08   2.415 0.016086 *
## wind         8.387e-02  3.501e-02   2.395 0.016973 *
## rain_1      -1.011e+00  5.245e-01  -1.928 0.054380 .
## month_2      1.433e-01  1.392e-01   1.029 0.303872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.363 on 509 degrees of freedom
## Multiple R-squared:  0.02747,    Adjusted R-squared:  0.01792
## F-statistic: 2.876 on 5 and 509 DF,  p-value: 0.01427
```

Chart 14: Final model (row 239, 500 excluded).

VII. Reference

Fox, John. Applied Regression Analysis and Generalized Linear Models. SAGE, 2016.

Kabacoff, Robert. “Regression Diagnostics.” Quick-R: Regression Diagnostics, www.statmethods.net/stats/rdiagnostics.html.

Kabacoff, Robert. “Graphics with ggplot2.” Quick-R: ggplot2 Graphs, www.statmethods.net/advgraphs/ggplot2.html.