Prediction of Forest Fires

Nakul Vilas Pujari

Introduction

Forest fires are an artificial of naturally occurring phenomenon which result due to several reasons which include campfires, man-made bonfires. However, most forest fires are started by natural causes like lightning, volcanic eruptions, or spontaneous combustion. The aftereffects of theses forest fires are severe, especially considering the current state of our natural habitats. In times like these it becomes more critical to try and prevent events like these or predict these occurrences so that preventive measures can be taken. In the given example, we have tried to estimate the occurrences of these forest fires for a certain location by using several potential predictors. These metrological predictors such as certain moisture, drought codes etc. from the FWI (Fire Weather Index) have been provided with certain data points for analysis. The data points and the predictors are important individually however the analysis done to predict the occurrence of forest fires, uses these predictors and data points together and the result shows that some data points and some predictors may not be useful for prediction. There is however some correlation between the predictors, therefore the prediction improves when these predictors are used as a combination rather than using them individually.

The given problem provides us with the following attribute information:

1. FFMC (x1) – FFMC Index from the FWI System: 18.7 to 96.20
2. DMC (x2) – DMC Index from the FWI System: 1.1 to 291.3
3. DC (x3) – DC Index from the FWI System: 7.9 to 860.6
4. ISI (x4) – ISI Index from the FWI System: 0.0 to 56.10
5. Temp (x5) – Temperature in Celsius degrees: 2.2 to 33.30
6. RH (x6) – Relative humidity in %: 15.0 to 100
7. Wind (x7) – Wind Speed in Km/h: 0.40 to 9.40
8. Rain (x8) – Outside rain in mm/m^2: 0.0 to 6.4
9. Area (x9) – The burned area of the forest (in Ha): 0.00 to 1090.84

The above-mentioned attributes act as the predictor variables for the given analysis. The response variable for the problem is defined as:

$$Y = ln\,(area + 1)$$

(Data plotted in the CSV data file)

Analysis

The initial step in the problem is to analyze the originally provided data using summary table and ANOVA table. (Figure 1)

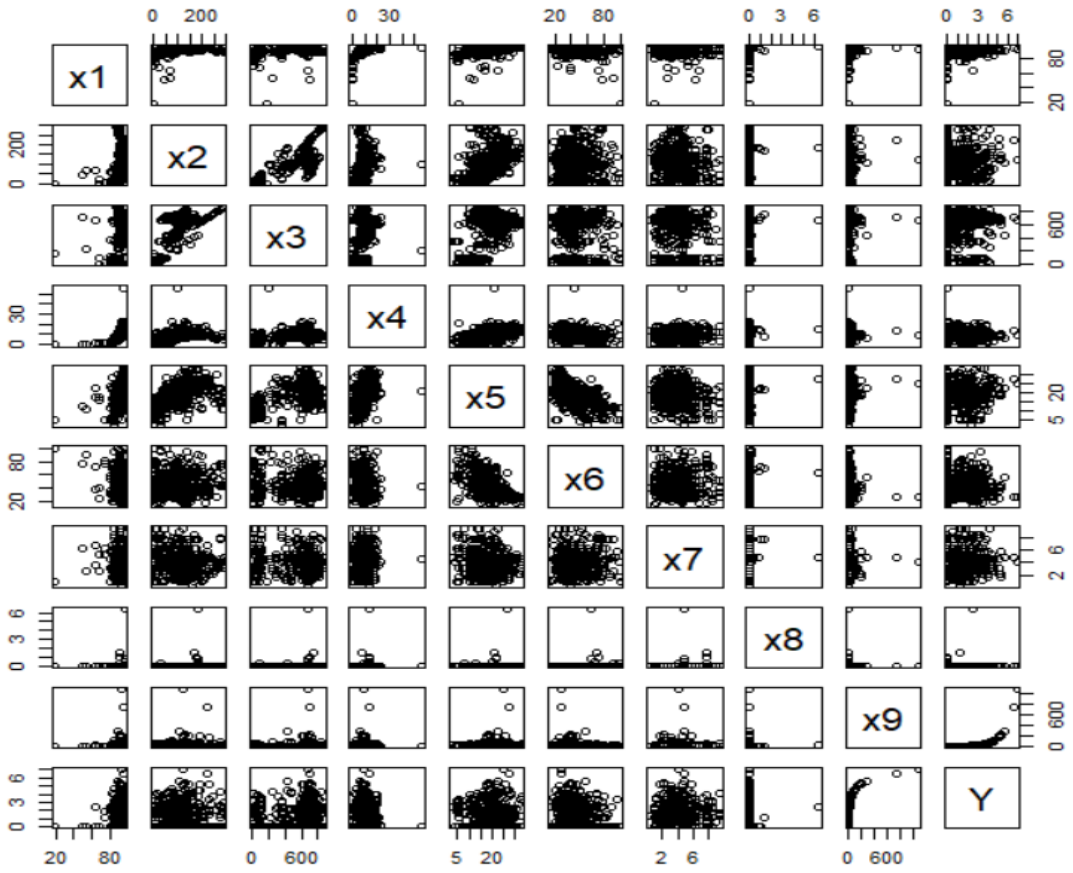
We see from the summary table for the original data, the higher P-Values and the P-Value for the overall model which is greater than a significance level of 0.05 or 0.01. This indicates that the model is not statistically significant. We also observe the R^2 value to be 0.01988 indicating that the percentage variance in Y is explained by X variables to a very small extent. The asterisk next to x7 indicates its importance in explaining the variance which is also indicated by its P-Value which is lower than the significance level of 0.05. We also see the mean square value (MSE) as shown in the ANOVA output table is close to 1.95 which means that the there is a huge amount of variance in the model which is unexplained by the current regression function. Observing the scatter plot and the residual plot, we see presence of outliers as well as non-linearity within the dependent and independent variables. (Figure 2)

```
> summary(allvar.mod)

Call:
lm(formula = Y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8, data = allvar)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5203 -1.1129 -0.6158  0.8787  5.7121

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.2224140  1.3604350   0.163    0.870
x1           0.0077082  0.0144884   0.532    0.595
x2           0.0011915  0.0014642   0.814    0.416
x3           0.0002737  0.0003570   0.767    0.444
x4          -0.0239494  0.0169248  -1.415    0.158
x5           0.0024618  0.0172593   0.143    0.887
x6          -0.0051729  0.0051889  -0.997    0.319
x7           0.0757669  0.0366155   2.069    0.039 *
x8           0.0965122  0.2121461   0.455    0.649
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.395 on 508 degrees of freedom
Multiple R-squared:  0.01988,   Adjusted R-squared:  0.004446
F-statistic: 1.288 on 8 and 508 DF,  p-value: 0.2472
```

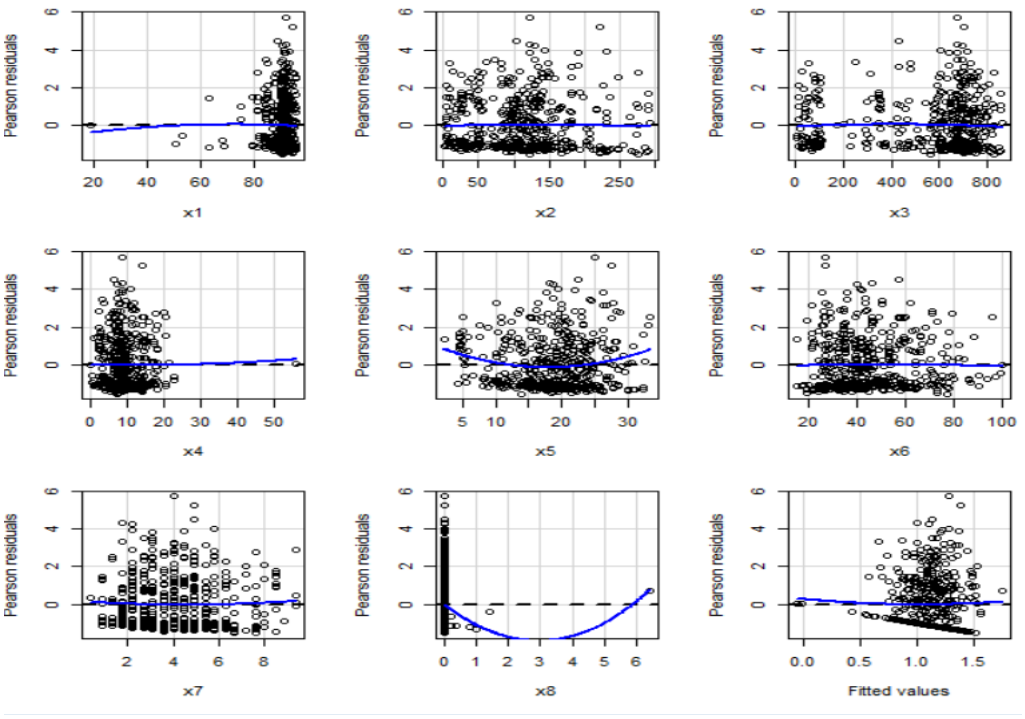
Figure 1

```
> anova(allvar.mod)
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq F value  Pr(>F)
x1          1   2.21  2.2100  1.1351 0.28719
x2          1   2.87  2.8670  1.4726 0.22550
x3          1   0.68  0.6765  0.3474 0.55582
x4          1   2.37  2.3652  1.2149 0.27090
x5          1   0.51  0.5051  0.2594 0.61073
x6          1   2.33  2.3309  1.1972 0.27440
x7          1   8.70  8.7045  4.4709 0.03496 *
x8          1   0.40  0.4029  0.2070 0.64935
Residuals 508 989.04  1.9469
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2



All-var Scatter Plot



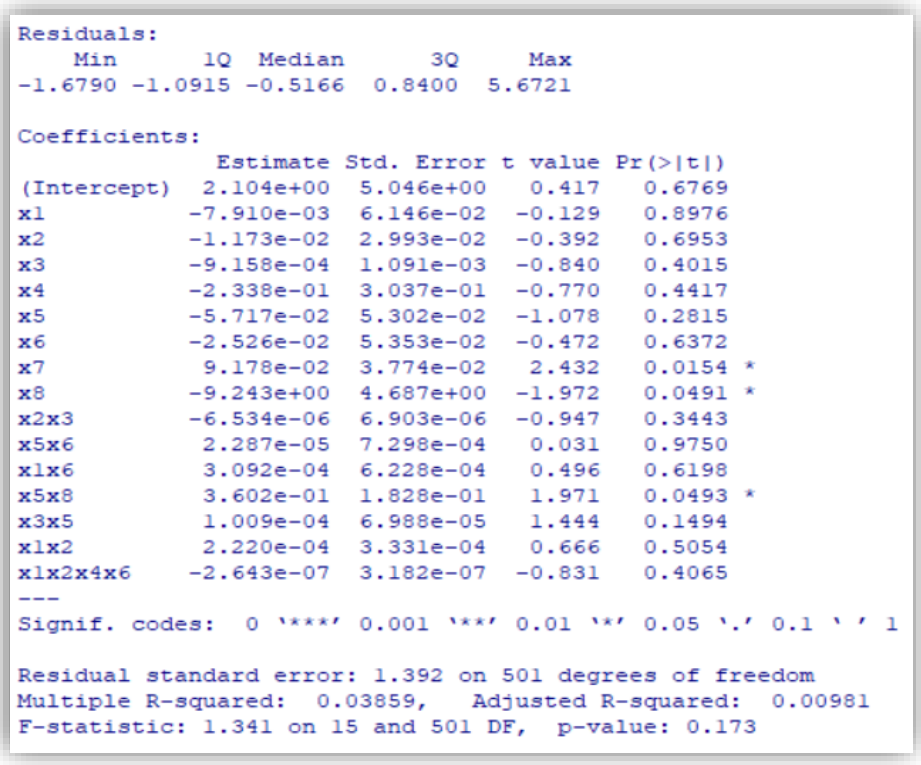
Residual Plot

To improve the model, the linearity, and the amount of variance in the dependent variable, explained by the model must be increased which can be done by including interactive variables for variables having a higher correlation. By including interactive variables, the correlation and linearity between dependent and interactive variables increases making the model a better fit. Another way to reduce the unexplained variance and increase linearity is to observe density plots which indicate the spread of data points for established numeric values. By observing the skewness of the density plots, the variables can be redefined and reintroduced in the model, which will help in improving the model. For the positively skewed variables, redefining the variable as a natural logarithm leads to a more linear model and lesser unexplained variance. This is the case for x3 and x8 where the density plot is more positively skewed. For x8 since several data points have value zero, it makes more sense to redefine the value as natural logarithm of (1+x8). Further, as we observed the scatter plot shown above, we see there is a positive correlation between x2 and x3. We also observe a negative correlation between x5 and x6, however this relation can still be used as an interactive variable due to the correlation present.

Further we observe the P-Values from the original summary output to compare variables. We can observe correlation between them on basis of P-Values, since the P-values indicate the impact of the variable in the model fit, we can include interactive variables having similar P-Values. According to ANOVA output, we take interactive variable combinations as products of one or two variables. When combining variables with similar P-Values, we get interactive variables, like x1x6, x5x8, x3x5, x1x2 and x1x2x4x6. We see that these interactive variable combinations have similar P-Values and hence provide more linearity to the model, reducing the unexplained variance and error. Therefore, we can use

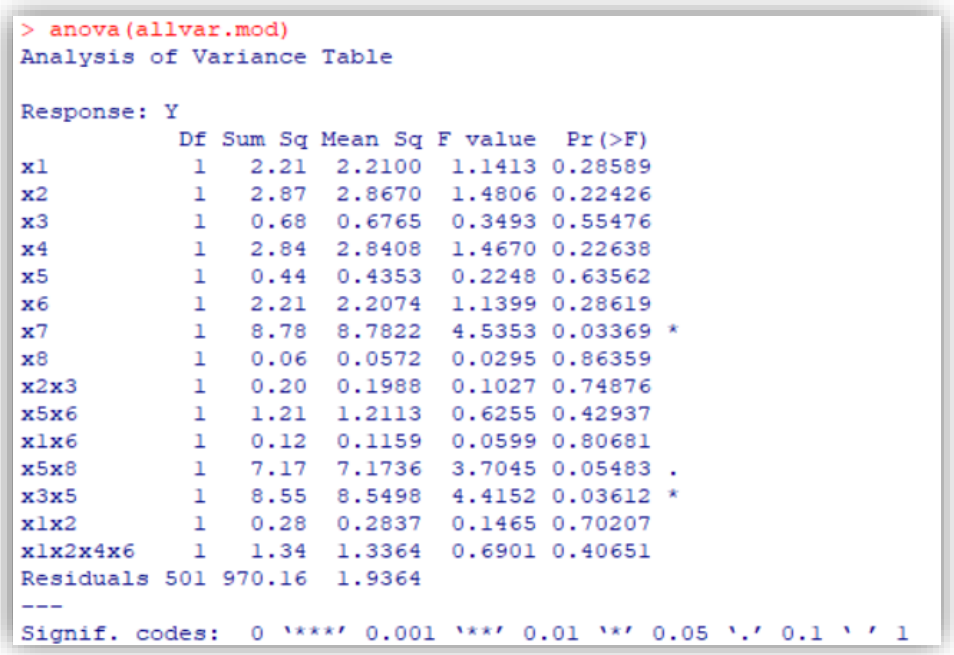
stepwise regression to include these interactive variables and testing for improvement of the model on basis of variance explained and linearity. From the Summary and the ANOVA output of the increased model containing the interactive variables and natural logarithm of positively skewed variables, we see improvement in the model in terms of increased R^2 and reduced Mean Square Error.

However, the model still does not do a good job of predicting the dependent variable with accuracy as is evident by the high P-Value and the outliers as observed in the above residual Plots. So, the next logical step to improve the model is to remove outliers and data points which may negatively affect the model. (Figure 3 and 4)



```
Residuals:
    Min      1Q  Median      3Q     Max
-1.6790 -1.0915 -0.5166  0.8400  5.6721

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.104e+00  5.046e+00   0.417   0.6769
x1          -7.910e-03  6.146e-02  -0.129   0.8976
x2          -1.173e-02  2.993e-02  -0.392   0.6953
x3          -9.158e-04  1.091e-03  -0.840   0.4015
x4          -2.338e-01  3.037e-01  -0.770   0.4417
x5          -5.717e-02  5.302e-02  -1.078   0.2815
x6          -2.526e-02  5.353e-02  -0.472   0.6372
x7           9.178e-02  3.774e-02   2.432   0.0154 *
x8          -9.243e+00  4.687e+00  -1.972   0.0491 *
x2x3        -6.534e-06  6.903e-06  -0.947   0.3443
x5x6         2.287e-05  7.298e-04   0.031   0.9750
x1x6         3.092e-04  6.228e-04   0.496   0.6198
x5x8         3.602e-01  1.828e-01   1.971   0.0493 *
x3x5         1.009e-04  6.988e-05   1.444   0.1494
x1x2         2.220e-04  3.331e-04   0.666   0.5054
x1x2x4x6    -2.643e-07  3.182e-07  -0.831   0.4065
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.392 on 501 degrees of freedom
Multiple R-squared:  0.03859,   Adjusted R-squared:  0.00981
F-statistic: 1.341 on 15 and 501 DF,  p-value: 0.173
```

Figure 3



```
> anova(allvar.mod)
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq F value  Pr(>F)
x1          1   2.21  2.2100  1.1413 0.28589
x2          1   2.87  2.8670  1.4806 0.22426
x3          1   0.68  0.6765  0.3493 0.55476
x4          1   2.84  2.8408  1.4670 0.22638
x5          1   0.44  0.4353  0.2248 0.63562
x6          1   2.21  2.2074  1.1399 0.28619
x7          1   8.78  8.7822  4.5353 0.03369 *
x8          1   0.06  0.0572  0.0295 0.86359
x2x3        1   0.20  0.1988  0.1027 0.74876
x5x6        1   1.21  1.2113  0.6255 0.42937
x1x6        1   0.12  0.1159  0.0599 0.80681
x5x8        1   7.17  7.1736  3.7045 0.05483 .
x3x5        1   8.55  8.5498  4.4152 0.03612 *
x1x2        1   0.28  0.2837  0.1465 0.70207
x1x2x4x6    1   1.34  1.3364  0.6901 0.40651
Residuals 501 970.16  1.9364
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4

We employ Cook's distance as a method to eliminate outliers and other data points which may negatively affect the outcome of the model. We run iterations, where in each iteration, we eliminate data points which have a very high Cook's distance (>0.01-0.0). Then we create a new model again with this modified data to check for accuracy and linearity.

Iteration Sets	R^2 range	Mean Square Error Value range
Iteration 1 – Iteration 3	0.04116 – 0.04169	1.9275 – 1.8751
Iteration 4 – Iteration 6	0.04473 – 0.05359	1.8662 – 1.7723
Iteration 7 – Iteration 9	0.05369 – 0.05455	1.6710 – 1.6417
Iteration 10 – Iteration 12	0.04932 – 0.05656	1.6024 – 1.5318
Iteration 13 – Iteration 15	0.05926 – 0.06234	1.4979 – 1.4724
Iteration 16 – Iteration 18	0.06659 – 0.06597	1.4426 – 1.3981

The above table represents the R^2 and Mean Square Error ranges for the iterations conducted for outlier detection using Cook's Distance method. For each iteration, the data points having Cook's distance values higher than normal have been eliminated and the modified data set has been used for next iteration. We observe in each iteration that removal of the Outliers and data points which negatively impact the model, results in each iteration having a better R^2 value and a lesser Mean Square Error term. This change in values makes sense, since the model improves with each iteration, resulting in removal of outliers and any data point which does not show linearity with the given model. They "pull" the model away from the projected values and thus their removal results in a more linear model and thus reduced unexplained error with each iteration. Hence, we exclude these observations from the

final model. The Influence plot data and the Standardized Residuals plots show the outliers as shown by the bubbles and the higher Cook's Distance values in this case observation number 212 is evidently an outlier. (Figure 5 and 6)

```
> influencePlot(allvar18.mod)
        StudRes         Hat        CookD
4   -0.03558093 0.719249461 0.0002031446
212  2.51566655 0.026116020 0.0104869231
228  3.68161574 0.007932870 0.0065963074
229  3.70304109 0.007509966 0.0063127707
289  0.05809754 0.476469467 0.0001924059
479  1.25672096 0.087336057 0.0094341077
```

Figure 5

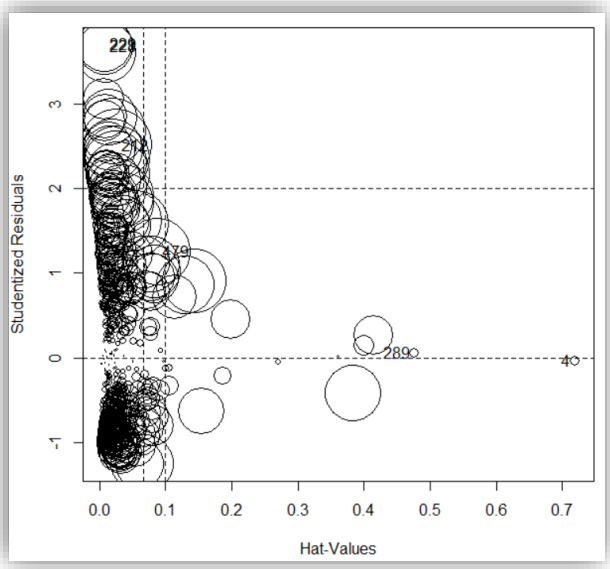


Figure 6

After running iterations, we reach to a point where the R^2 decreases meaning the data points removed are significant for the model and thus we keep these points and complete the outlier removal process.

The data we are left with still contains variables/attributes which may not be related to the data or which may "pull" the data in a negative direction. Thus, the removal of such attributes may improve the model. Although in a real-life scenario all variables play an important role, for prediction using limited data points, it may be best to remove certain un-related variables which may result in improvement of the current model. Thus, we now decide on a model selection criterion.

The presented model is of confirmatory observational type. Since the Planning and data collection along with model exploration is dealt with, we proceed with model selection. For model selection, we choose R^2, Adjusted R^2 (Mean Square Error) and AIC (Akaike's Information Criterion) as basis criterions for model selection. AIC method for model selection is bases on maximum likelihood, therefore the model selected will do a good job of explaining the data. We want the model selected to have a lesser AIC, since the lesser the AIC, the better the model. Hence, we run a stepwise iteration to find a subset model having the lowest AIC value. (Figure 7 and 8)

```
> step(lm(Y~x1+x2+x3+x4+x5+x6+x7+x8+x2x3+x5x6+x1x6+x5x8+x3x5+x1x2+x1x2x4x6, data=allvar16),$
Start:  AIC=193.09
Y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x2x3 + x5x6 + x1x6 +
    x5x8 + x3x5 + x1x2 + x1x2x4x6

           Df Sum of Sq    RSS    AIC
- x1x6      1    0.0065 675.14 191.09
- x6        1    0.0614 675.19 191.13
- x1        1    0.1611 675.29 191.20
- x5x8      1    0.1929 675.32 191.22
- x8        1    0.4269 675.56 191.39
- x4        1    0.5325 675.66 191.47
- x2        1    0.7386 675.87 191.62
- x1x2      1    1.5476 676.68 192.19
- x3        1    2.3765 677.51 192.79
- x2x3      1    2.5358 677.67 192.90
- x1x2x4x6  1    2.5601 677.69 192.92
<none>                  675.13 193.09
- x5x6      1    3.9632 679.09 193.92
- x7        1    4.0927 679.22 194.01
- x3x5      1    8.8360 683.97 197.38
- x5        1   16.9326 692.06 203.07
```

Figure 7

```
           Df Sum of Sq    RSS    AIC
- x5x6      1    2.0148 686.55 183.21
<none>                  684.54 183.78
- x7        1    5.2552 689.79 185.49
- x2        1    7.1265 691.66 186.80
- x1x2      1    8.6995 693.24 187.90
- x3x5      1   12.3937 696.93 190.47
- x1x2x4x6  1   12.5372 697.08 190.57
- x5        1   22.1362 706.67 197.19

Step:  AIC=183.21
Y ~ x2 + x5 + x7 + x3x5 + x1x2 + x1x2x4x6

           Df Sum of Sq    RSS    AIC
<none>                  686.55 183.21
- x7        1    4.3908 690.94 184.29
- x2        1    5.2813 691.83 184.91
- x1x2      1    6.6929 693.25 185.90
- x3x5      1   13.9132 700.47 190.92
- x1x2x4x6  1   14.4531 701.01 191.29
- x5        1   21.6087 708.16 196.20

Call:
lm(formula = Y ~ x2 + x5 + x7 + x3x5 + x1x2 + x1x2x4x6, data = allvar16)

Coefficients:
(Intercept)           x2           x5           x7         x3x5         x1x2     x1x2x4x6
  1.395e+00   -3.503e-02   -7.287e-02    5.580e-02    6.144e-05    4.355e-04   -5.089e-07
```

Figure 8

The above iterations run and provide us with a set of subsets of models. We pick the model having the lowest AIC value. The dataset selected after model exploration (Outlier removal) is now to be modified such that only the variables in the model with the lowest AIC value are to be kept and other variables are to be removed. This is a critical step in the model selection process, since although the R^2 may decrease and Mean Square Error value increases, the remaining variables are the only ones which are statistically significant to the model. The decrease in the indicator criterions is negligible in terms of model improvement on removal of unwanted variables. Thus, the final model is developed after removal of these variables.

We observe from the summary and ANOVA output table, the decrease in R^2 (Final Value – 0.0508) and the increase in Mean Square Error (Final Value – 1.4393). However, we see the P-Value for the model is now 0.0003457 which is significantly lower than 0.05 and 0.01 significance levels and lower than P-Value of the initial starting model (0.2472). To compare we also plot the Y vs Y-Hat plot which further shows, the difference between observed and projected values are better demonstrated by the current model. (Figure 9 and 10)

The scatter plot of the attributes of the final model against the response show a stronger correlation between the variables and the response. All variables show positive correlation, as shown in the scatter plot, furthermore, interactive variables also show strong positive correlation. (Figure 11, 12 and 13)

```
> allvar19<-read.table(file="clipboard", header = T, sep="\t")
> allvar19.mod<-lm(Y~x2+x5+x7+x3x5+x1x2+x1x2x4x6, allvar19)
> summary(allvar19.mod)

Call:
lm(formula = Y ~ x2 + x5 + x7 + x3x5 + x1x2 + x1x2x4x6, data = allvar19

Residuals:
    Min      1Q  Median      3Q     Max
-1.5174 -0.9588 -0.4568  0.7725  4.3413

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.395e+00  2.889e-01   4.830 1.85e-06 ***
x2          -3.503e-02  1.829e-02  -1.916 0.056020 .
x5          -7.287e-02  1.881e-02  -3.875 0.000122 ***
x7           5.580e-02  3.195e-02   1.747 0.081349 .
x3x5         6.144e-05  1.976e-05   3.109 0.001989 **
x1x2         4.355e-04  2.020e-04   2.156 0.031552 *
x1x2x4x6    -5.089e-07  1.606e-07  -3.169 0.001629 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.2 on 477 degrees of freedom
Multiple R-squared:  0.0508,    Adjusted R-squared:  0.03886
F-statistic: 4.255 on 6 and 477 DF,  p-value: 0.0003457
```

Figure 9

```
> anova(allvar19.mod)
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq F value    Pr(>F)
x2          1   0.20  0.1957  0.1360 0.7124586
x5          1   0.27  0.2747  0.1909 0.6623832
x7          1   1.63  1.6267  1.1302 0.2882669
x3x5        1  17.21 17.2086 11.9561 0.0005935 ***
x1x2        1   2.98  2.9827  2.0723 0.1506481
x1x2x4x6    1  14.45 14.4531 10.0416 0.0016287 **
Residuals 477 686.55  1.4393
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 10

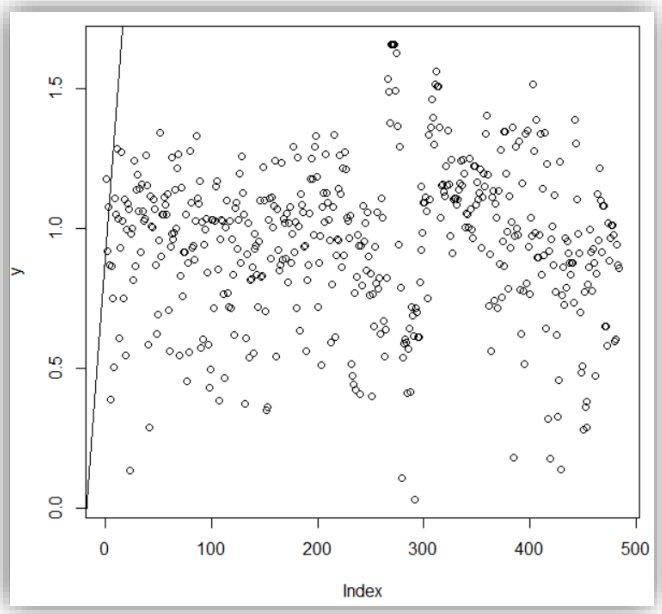


Figure 11

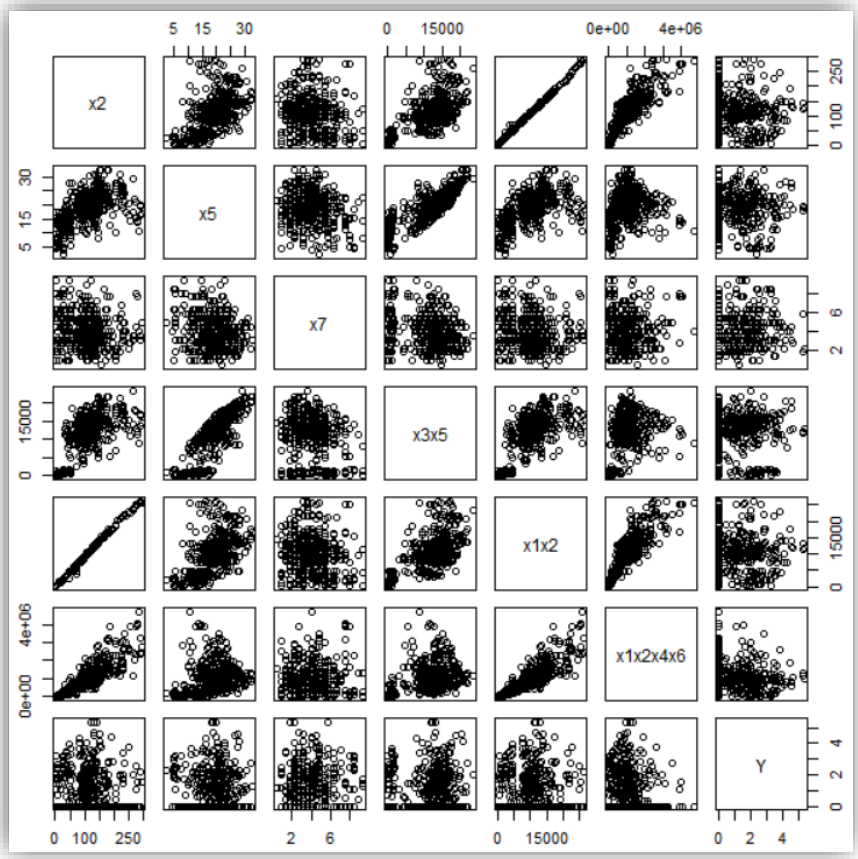


Figure 12

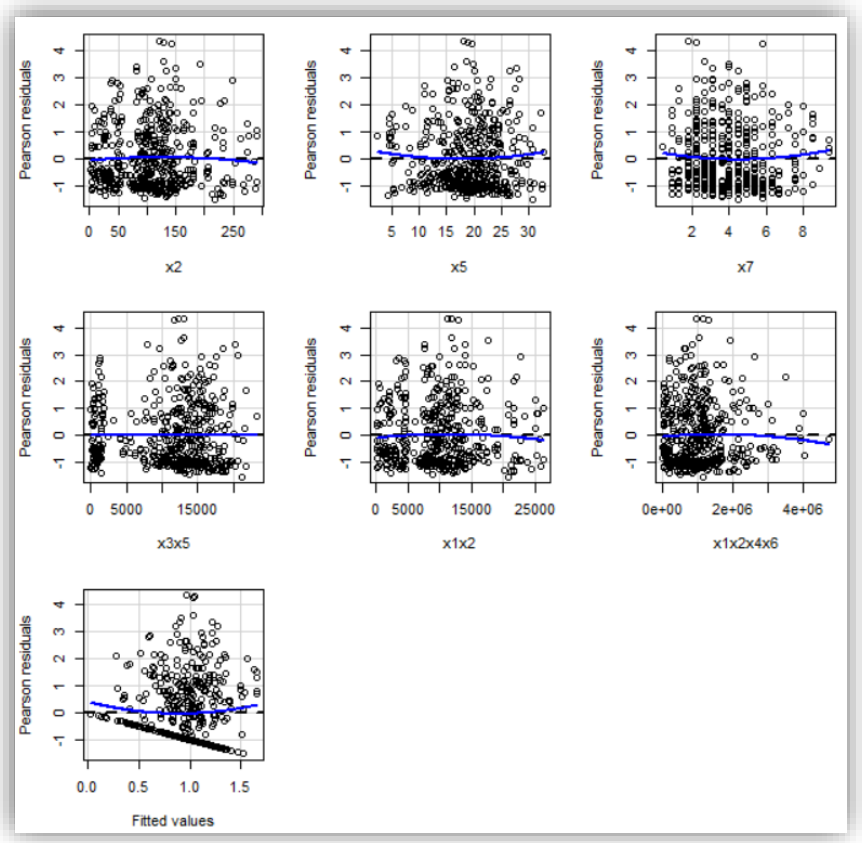


Figure 13

Thus, we see from the above Y vs Y hat graph, scatter plots and the residual plot from above, the correlation between variables has improved, the predicted and observed values are better with the new model and furthermore, the amount of outliers in the model has decreased as is evident from the residuals plot above.

The final regression model is:

$$Y = (1.395e{+}00) - x_2(0.03503) - x_3(0.07287) + x_7(0.05580) + x_3x_5(6.144{*}10^{-5}) + x_1x_2(4.355 * 10^{-4}) - x_1x_2x_4x_6\,(5.089 * 10^{-7})$$

The final R^2 value is:

0.0508

The final MSE value is:

1.4393