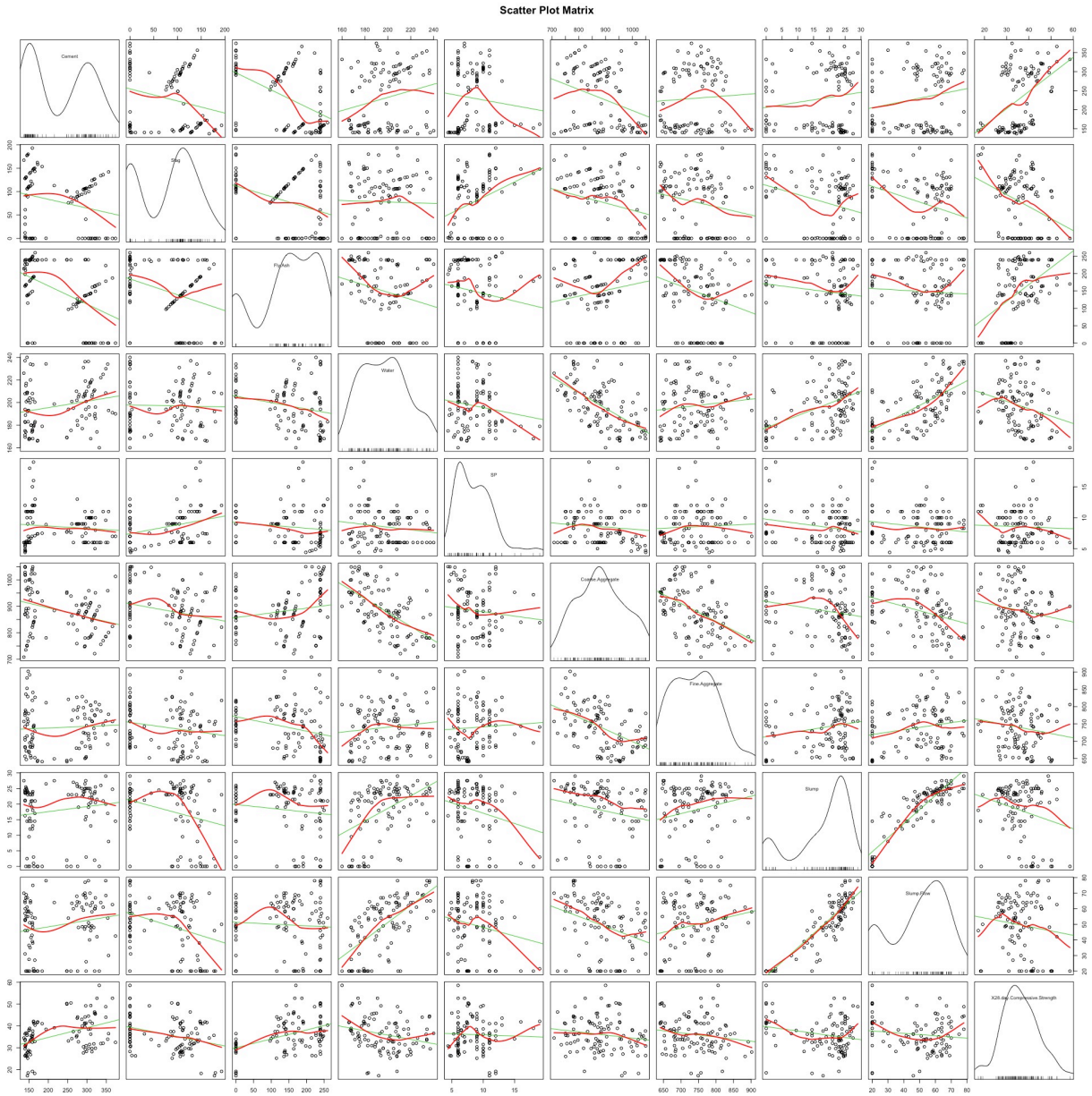
Task 3: Concrete Slump Test Data

1 - Scatterplot Matrix



2 - Build a few models

We selected the following models for our analysis
Output: Slump
fitconc1 <- lm(Slump ~ Water+SP+Slag+Coarse.Aggregate+Fine.Aggregate ,data = dfconc)
fitconc <- lm(Slump ~ Water+SP+Slag,data = dfconc)

Ouput: Slump Flow
fitconc2 <- lm(Slump.Flow ~ Water+Coarse.Aggregate+Slag ,data = dfconc)
fitconc2b <- lm(Slump.Flow ~ Water+Coarse.Aggregate+Slag+Fly.Ash ,data = dfconc)

Ouput: X28.Day Compressive Strength
fitconc3 <- lm(X28.day.Compressive.Strength ~ Water+Fly.Ash+Cement+Slag ,data = dfconc)
fitconc3b <- lm(X28.day.Compressive.Strength ~ Water+Fly.Ash+Cement+Slag+SP ,data = dfconc)

3/4 – Example of how regression diagnostics was performed for each model/ Identifying outliers and unusual behaviour

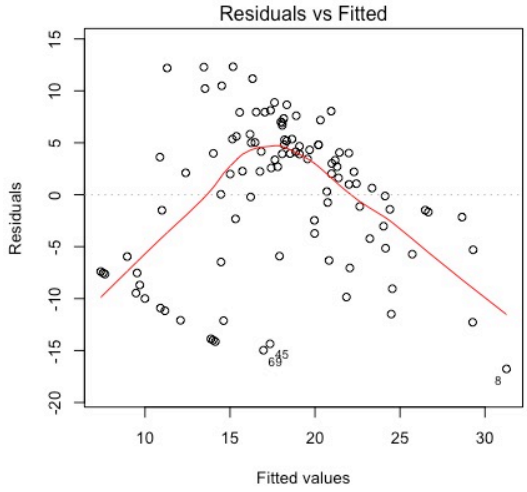
Fitconc1

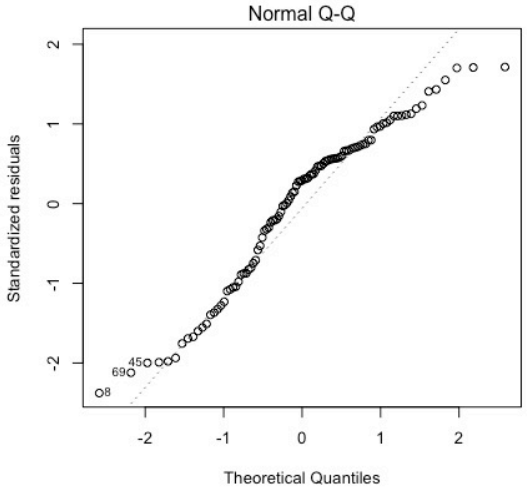
```
Call:
lm(formula = Slump ~ Water + SP + Slag + Coarse.Aggregate + Fine.Aggregate,
    data = dfconc)

Residuals:
    Min      1Q  Median      3Q     Max
-16.766  -5.818   2.105   5.015  12.323

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -64.47383   28.99022  -2.224   0.0285 *
Water              0.23769    0.05068   4.690 8.94e-06 ***
SP                -0.21366    0.28166  -0.759   0.4500
Slag              -0.02377    0.01443  -1.647   0.1028
Coarse.Aggregate   0.02000    0.01397   1.432   0.1555
Fine.Aggregate     0.02928    0.01512   1.937   0.0557 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.388 on 97 degrees of freedom
Multiple R-squared:  0.3222,    Adjusted R-squared:  0.2873
F-statistic: 9.222 on 5 and 97 DF,  p-value: 3.31e-07
```

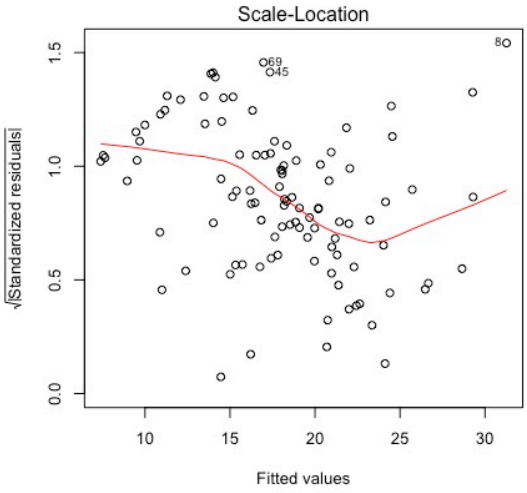
Confidence Intervals

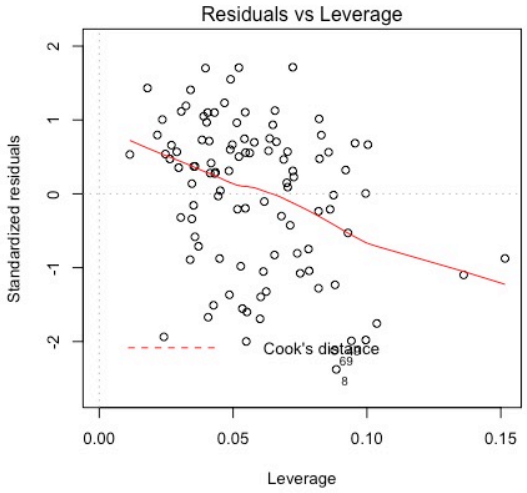
```
                         2.5 %       97.5 %
(Intercept)      -1.220114e+02 -6.936273140
Water             1.371079e-01  0.338272373
SP               -7.726843e-01  0.345365217
Slag             -5.240834e-02  0.004870978
Coarse.Aggregate -7.728598e-03  0.047731280
Fine.Aggregate   -7.232014e-04  0.059280633
```

Residual Plots

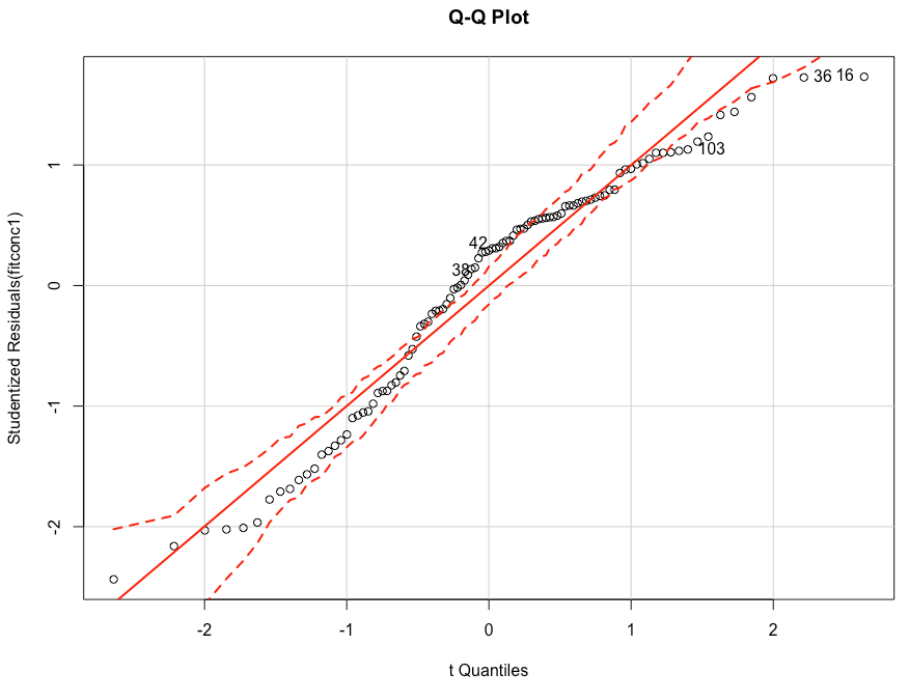


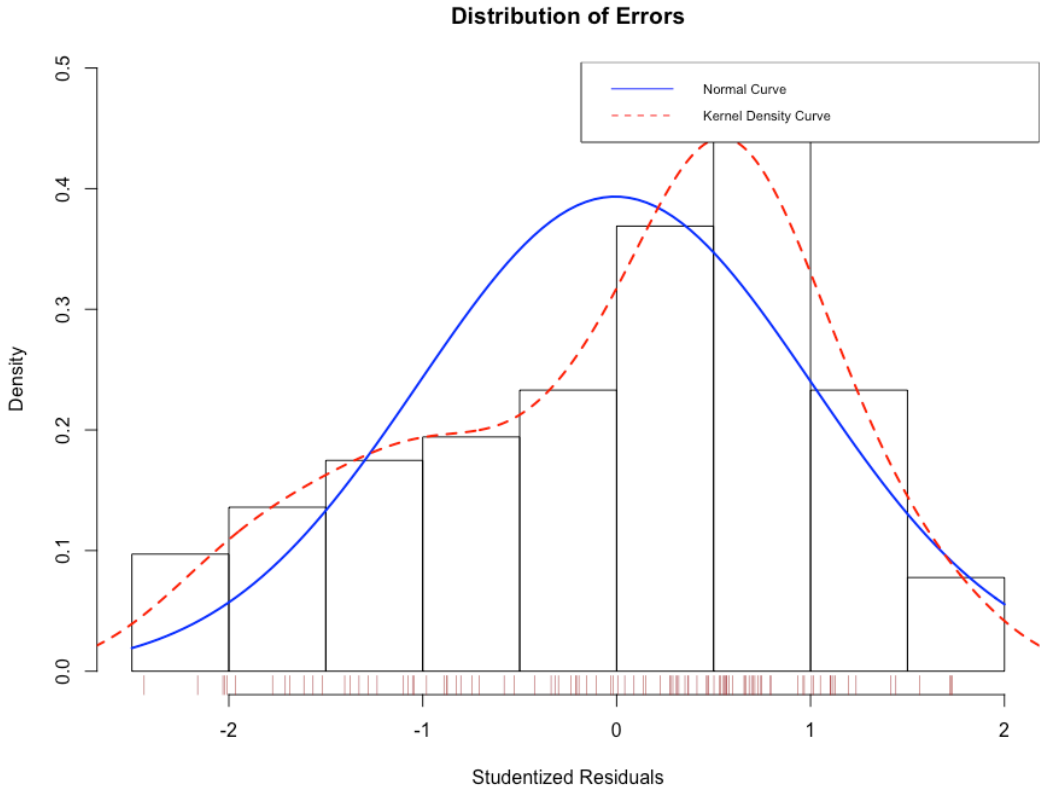






Enhanced Approach





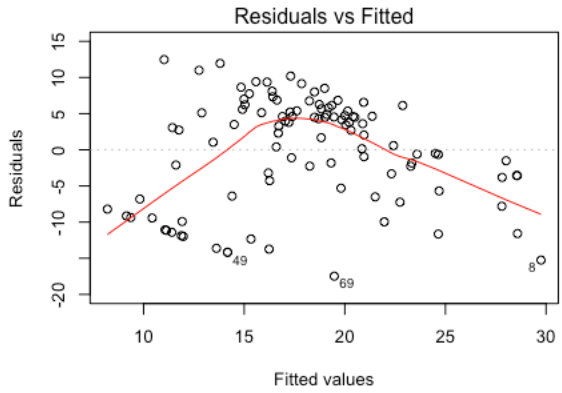
Fitconc

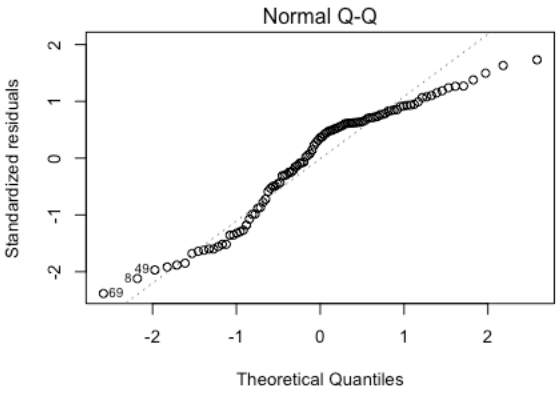
```
Call:
lm(formula = Slump ~ Water + SP + Slag, data = dfconc)

Residuals:
    Min      1Q  Median      3Q     Max
-17.476  -5.500   2.694   5.370  12.479

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.74393    8.00305  -1.967  0.05195 .
Water         0.19469    0.03699   5.263 8.24e-07 ***
SP           -0.20517    0.27964  -0.734  0.46486
Slag         -0.03645    0.01283  -2.840  0.00547 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.456 on 99 degrees of freedom
Multiple R-squared:  0.2953,    Adjusted R-squared:  0.274
F-statistic: 13.83 on 3 and 99 DF,  p-value: 1.331e-07
```

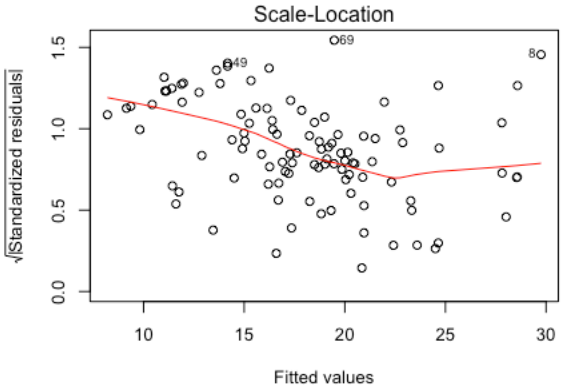
Confidence Intervals

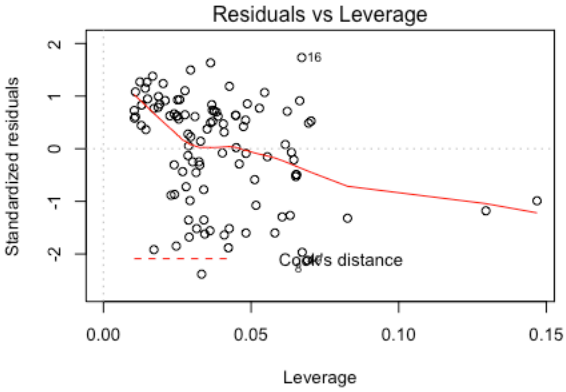
```
                  2.5 %      97.5 %
(Intercept) -31.62370728  0.13585663
Water         0.12128654  0.26809258
SP           -0.76003137  0.34968820
Slag         -0.06190837 -0.01098663
```

Residual Plots

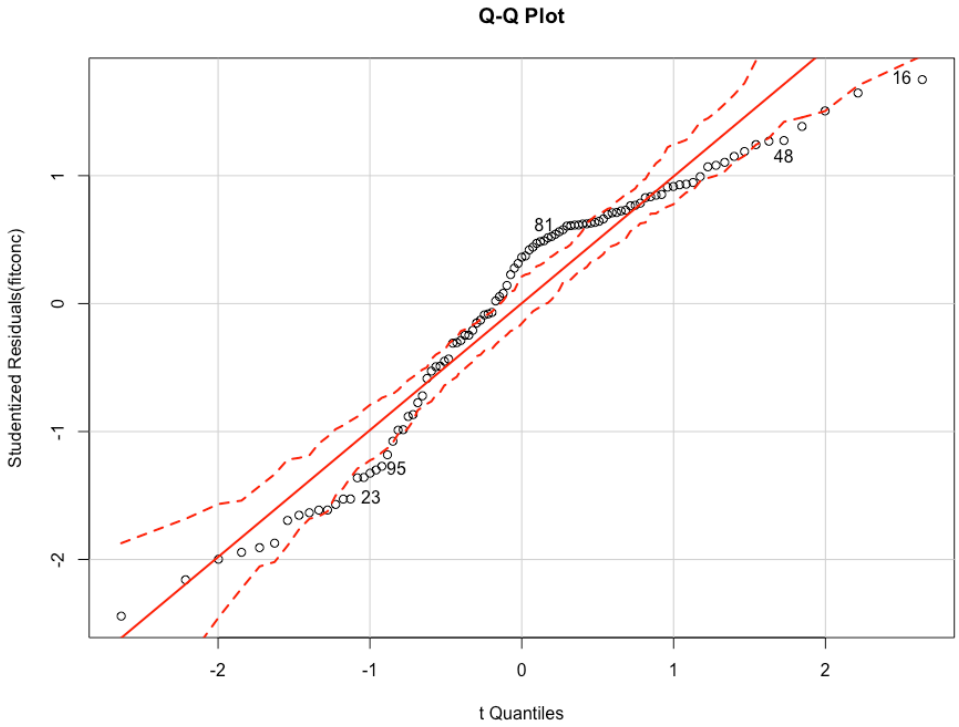


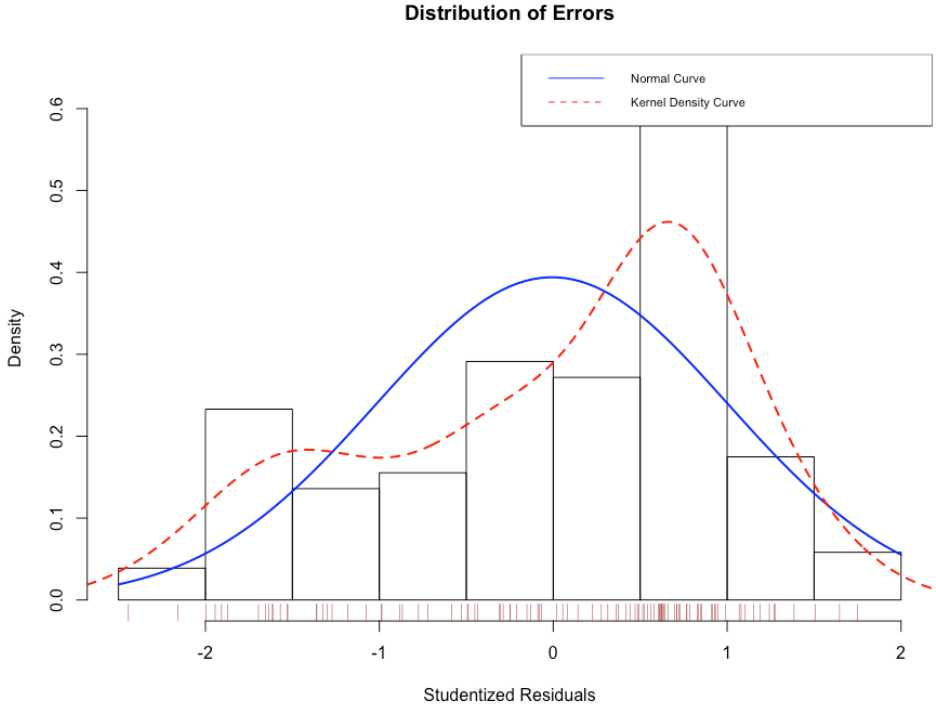






Enhanced Approach





5/6/7 – Select and Interpret the prediction results

ANOVA analysis of both models

```
Analysis of Variance Table

Model 1: Slump ~ Water + SP + Slag
Model 2: Slump ~ Water + SP + Slag + Coarse.Aggregate + Fine.Aggregate
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     99 5504.3
2     97 5294.3  2    209.98 1.9236 0.1516
```

Since it has a significant p value, we cannot reject the null hypothesis that the model 1 is not better than model 2. Hence, we select model 1 = Fitconc as our multiple regression model for the response variable of Slump.

Applying the same methodology for the other model pairs we arrive at:

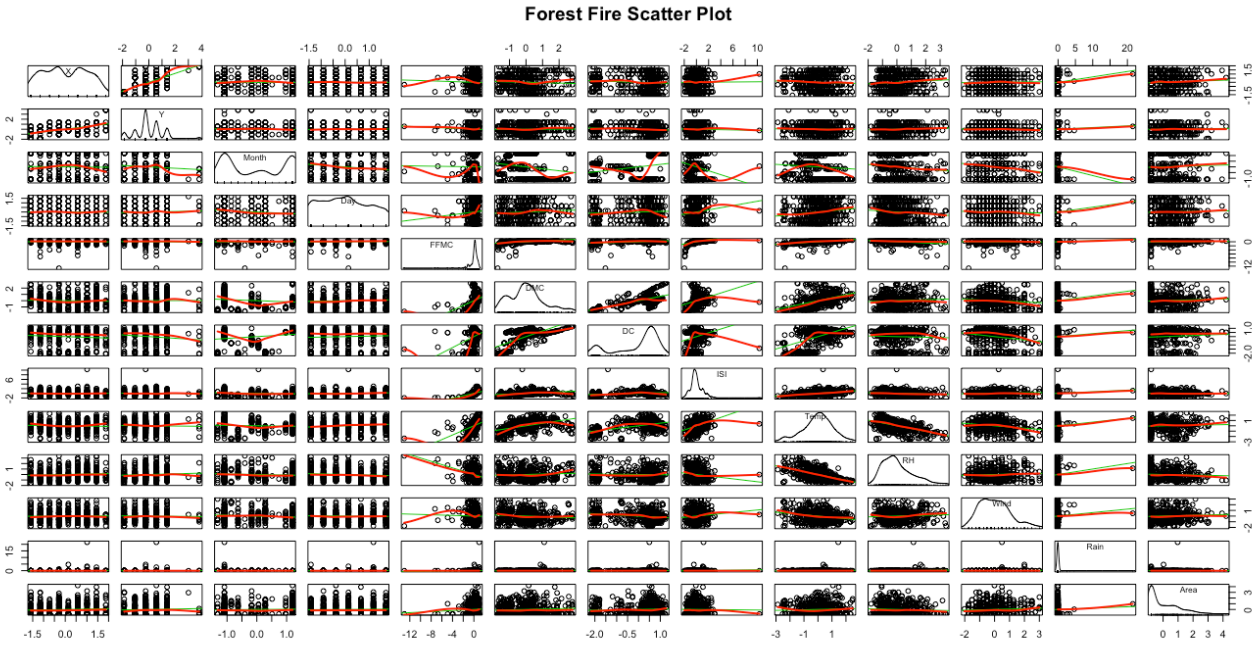
2) Slump Flow - Water+Coarse.Aggregate+Slag+Fly.Ash

3) X28.Day Compressive Strength - Water+Fly.Ash+Cement+Slag

Task 4: Forest Fire Data Set

Before we can plot the scatterplot Matrix, we need to perform data transformation on the data set. First we convert month and days to 1:12 and 1:7 numeric format. Then, we coerce this to numeric type. After which we perform a log(1+x) transformation on the area column as there is a lot of values with 0 and we don't want a heavily skewed regression model.

1 - Scatterplot Matrix



2 - Build a few models

We chose the same models as chosen in the study performed by P. Cortez and A. Morais. "A Data Mining Approach to Predict Forest Fires using Meteorological Data,"

Model 1: STFWTI - Area modelled using all other input variables
Model 2: STM – Area modelled using Spatial, Temporal and Weather Data
Model 3: FWT – Area modelled using FWT
Model 4: M – Area modelled using just weather data

3/4 – Example of how regression diagnostics was performed for each model/ Identifying outliers and unusual behaviour

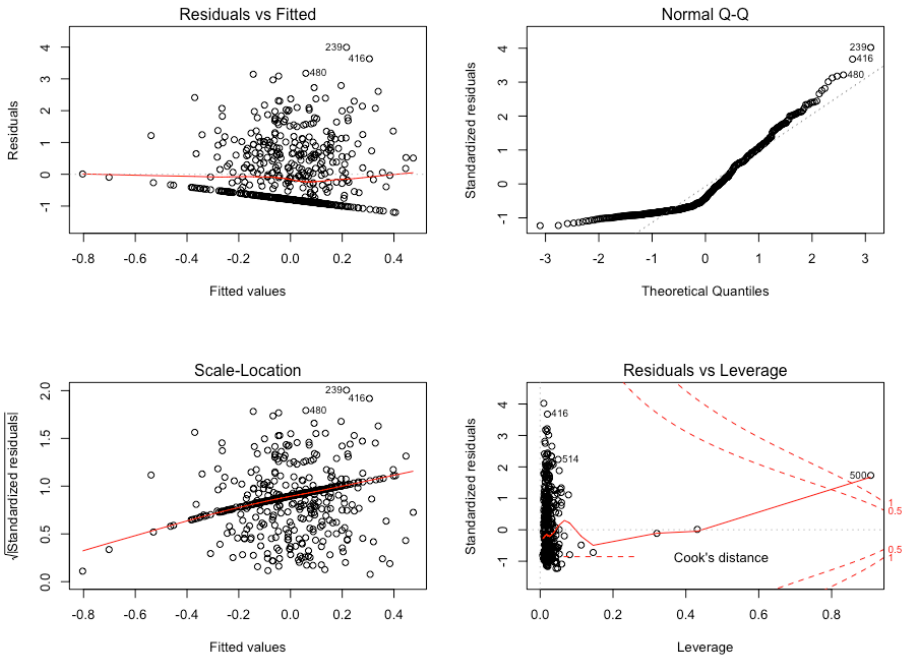
Model 1 - STFWTI

```
Call:
lm(formula = Area ~ X + Y + Month + Day + FFMC + DMC + DC + ISI +
    Temp + RH + Wind + Rain, data = forestds)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2003 -0.7777 -0.4107  0.6321  3.9917

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.638e-16  4.387e-02   0.000   1.0000
X            6.646e-02  5.261e-02   1.263   0.2070
Y            1.202e-02  5.285e-02   0.227   0.8202
Month        5.416e-02  5.322e-02   1.018   0.3094
Day          3.123e-02  4.520e-02   0.691   0.4899
FFMC         2.506e-02  5.746e-02   0.436   0.6629
DMC          8.098e-02  7.238e-02   1.119   0.2637
DC           2.266e-02  7.308e-02   0.310   0.7567
ISI         -7.490e-02  5.571e-02  -1.344   0.1794
Temp         1.766e-02  7.305e-02   0.242   0.8090
RH          -5.827e-02  6.149e-02  -0.948   0.3438
Wind         1.036e-01  4.722e-02   2.194   0.0287 *
Rain         1.610e-02  4.502e-02   0.358   0.7207
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9976 on 504 degrees of freedom
Multiple R-squared:  0.02793,   Adjusted R-squared:  0.004783
F-statistic: 1.207 on 12 and 504 DF,  p-value: 0.2749
```

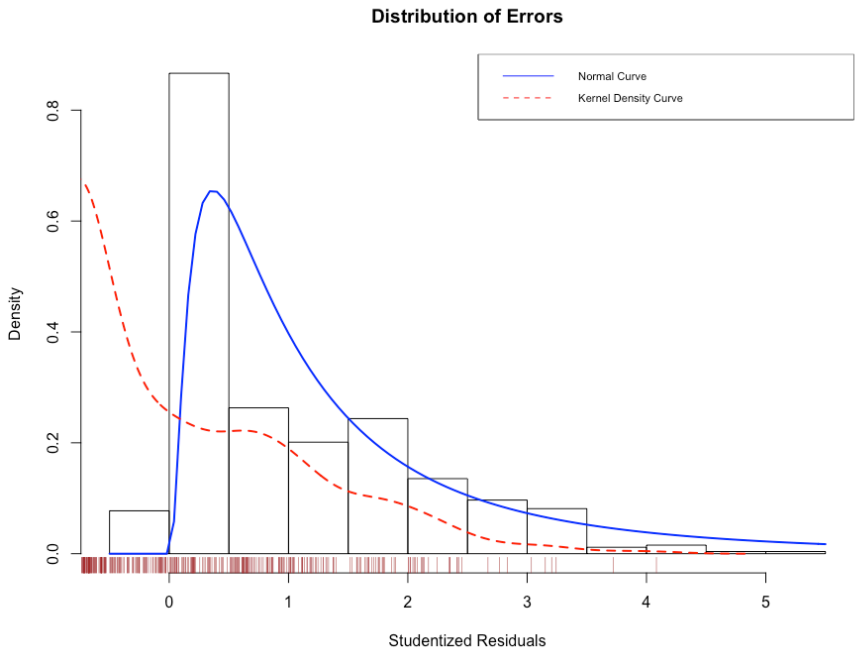
Confidence Intervals

```
                2.5 %     97.5 %
(Intercept) -0.08619971 0.08619971
X           -0.03689272 0.16981377
Y           -0.09181605 0.11585328
Month       -0.05040361 0.15871626
Day         -0.05756805 0.12002852
FFMC        -0.08782598 0.13794685
DMC         -0.06121289 0.22317830
DC          -0.12092423 0.16624139
ISI         -0.18435186 0.03455680
Temp        -0.12585552 0.16118258
RH          -0.17907818 0.06253589
Wind         0.01085088 0.19638374
Rain        -0.07234218 0.10455039
```

Residual Plots



Distribution of Errors



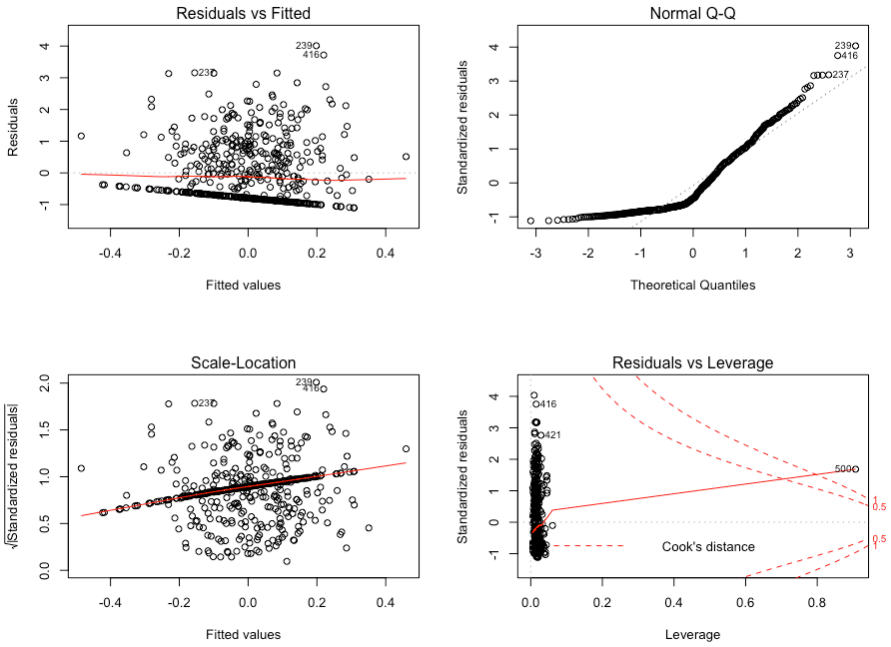
Model 2 - STM

```
Call:
lm(formula = Area ~ X + Y + Month + Day + Temp + RH + Wind +
    Rain, data = forestds)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1029 -0.7854 -0.4504  0.6344  4.0099

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.403e-16  4.391e-02   0.000   1.0000
X            5.789e-02  5.243e-02   1.104   0.2701
Y            1.505e-02  5.239e-02   0.287   0.7741
Month        6.253e-02  4.603e-02   1.358   0.1749
Day          2.867e-02  4.506e-02   0.636   0.5249
Temp         6.416e-02  5.484e-02   1.170   0.2426
RH          -2.416e-02  5.355e-02  -0.451   0.6521
Wind         9.136e-02  4.616e-02   1.979   0.0483 *
Rain         1.468e-02  4.497e-02   0.326   0.7442
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9984 on 508 degrees of freedom
Multiple R-squared:  0.01869,   Adjusted R-squared:  0.003234
F-statistic: 1.209 on 8 and 508 DF,  p-value: 0.2913
```

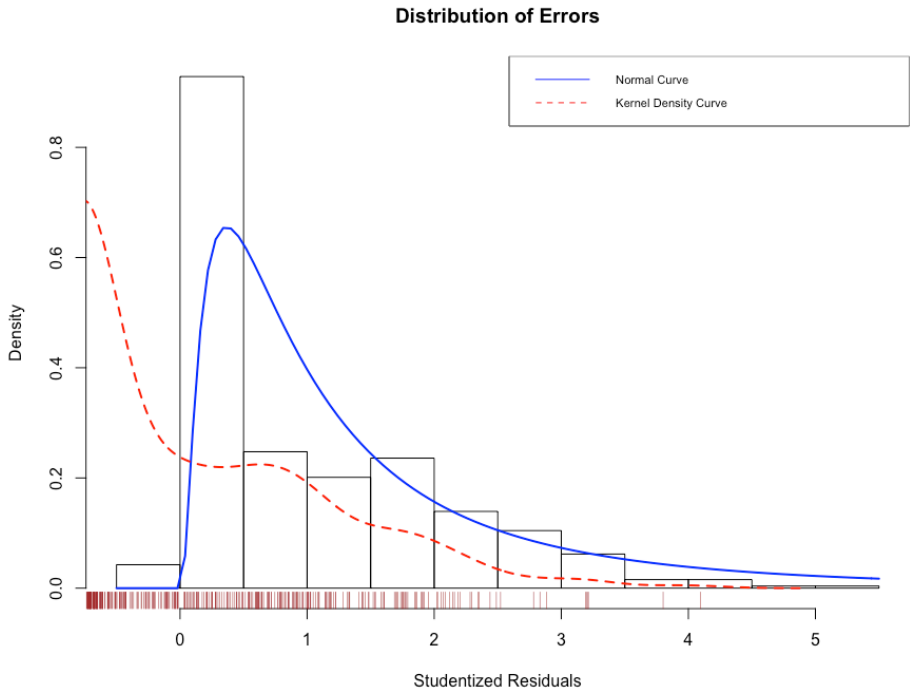
Confidence Intervals

```
                  2.5 %     97.5 %
(Intercept) -0.0862651628 0.08626516
X           -0.0451201642 0.16089948
Y           -0.0878896863 0.11798524
Month       -0.0279052663 0.15296602
Day         -0.0598516514 0.11718875
Temp        -0.0435789659 0.17189594
RH          -0.1293644046 0.08105425
Wind         0.0006700111 0.18204263
Rain        -0.0736644257 0.10302400
```

Residual Plots



Distribution of Errors



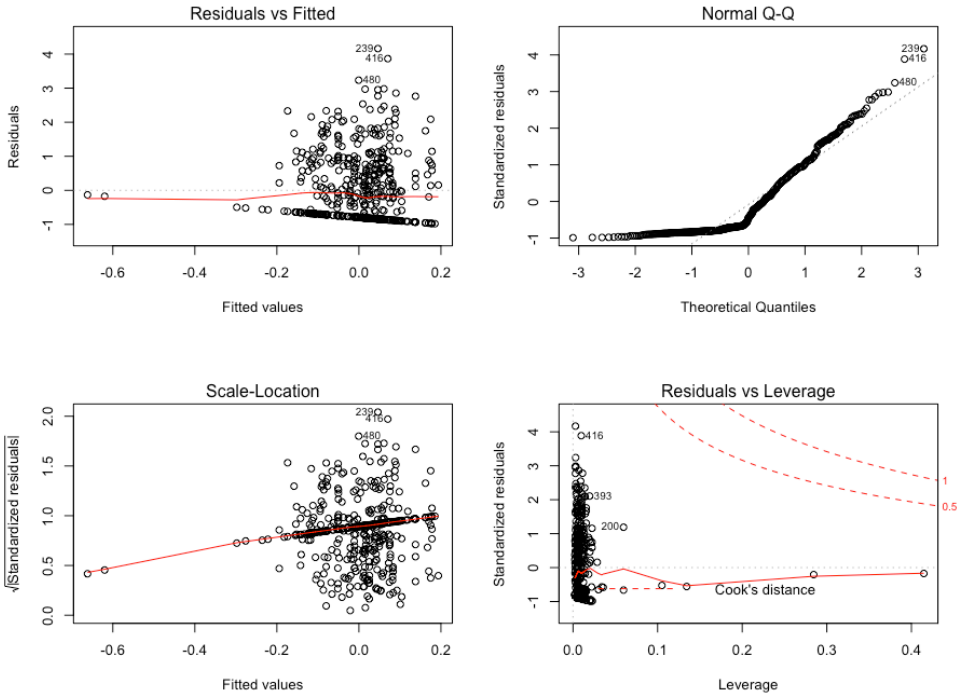
Model 3 - FWI

```
Call:
lm(formula = Area ~ FFMC + DMC + DC + ISI, data = forestds)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9799 -0.8039 -0.4394  0.6351  4.1617

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.588e-16  4.397e-02   0.000    1.000
FFMC         4.998e-02  5.431e-02   0.920    0.358
DMC          4.229e-02  6.223e-02   0.680    0.497
DC           3.420e-02  6.053e-02   0.565    0.572
ISI         -5.767e-02  5.243e-02  -1.100    0.272

Residual standard error: 0.9999 on 512 degrees of freedom
Multiple R-squared:  0.008046,  Adjusted R-squared:  0.0002959
F-statistic: 1.038 on 4 and 512 DF,  p-value: 0.3869
```

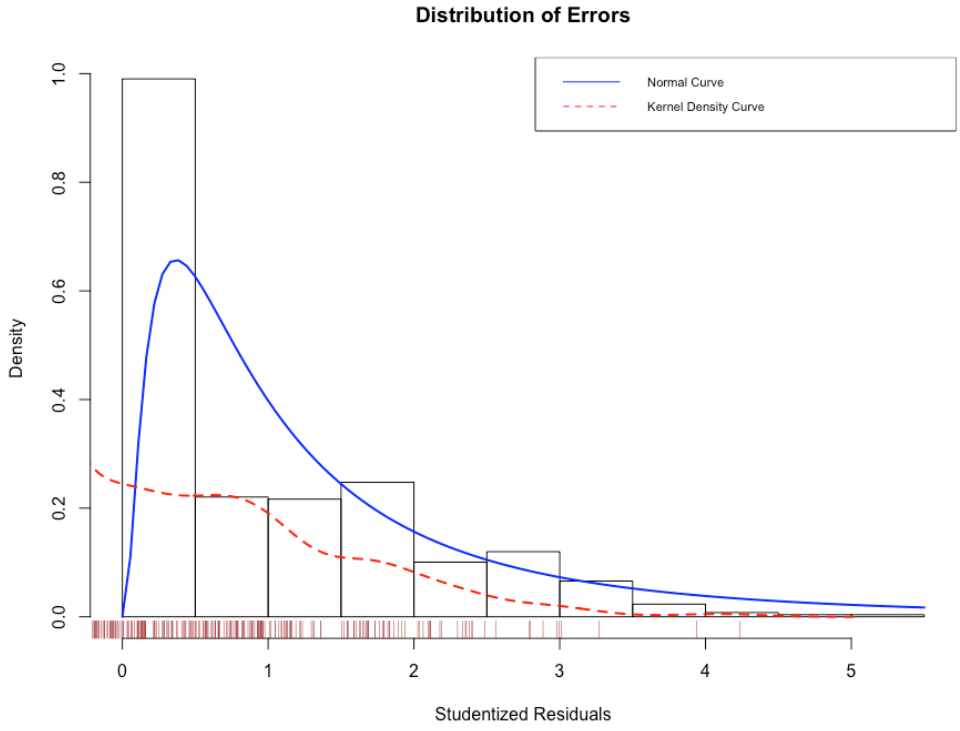
Confidence Intervals

```
                  2.5 %     97.5 %
(Intercept) -0.08639058 0.08639058
FFMC        -0.05670994 0.15667163
DMC         -0.07995893 0.16454177
DC          -0.08471981 0.15312947
ISI         -0.16067558 0.04533677
```

Residual Plots



Distribution of Errors



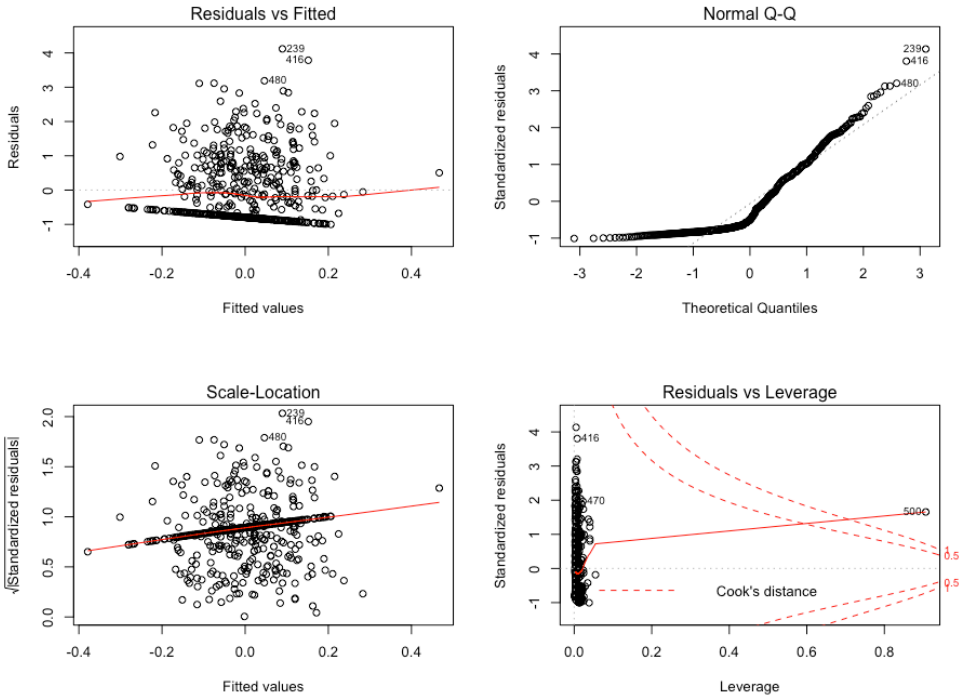
Model 4 - Weather

```
Call:
lm(formula = Area ~ Temp + RH + Wind + Rain, data = forestds)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0006 -0.7850 -0.5063  0.6523  4.1184

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.062e-16  4.392e-02   0.000    1.000
Temp         5.301e-02  5.381e-02   0.985    0.325
RH          -3.307e-02  5.258e-02  -0.629    0.530
Wind         8.021e-02  4.541e-02   1.766    0.078 .
Rain         1.802e-02  4.484e-02   0.402    0.688
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9987 on 512 degrees of freedom
Multiple R-squared:  0.0104,    Adjusted R-squared:  0.002671
F-statistic: 1.345 on 4 and 512 DF,  p-value: 0.2519
```

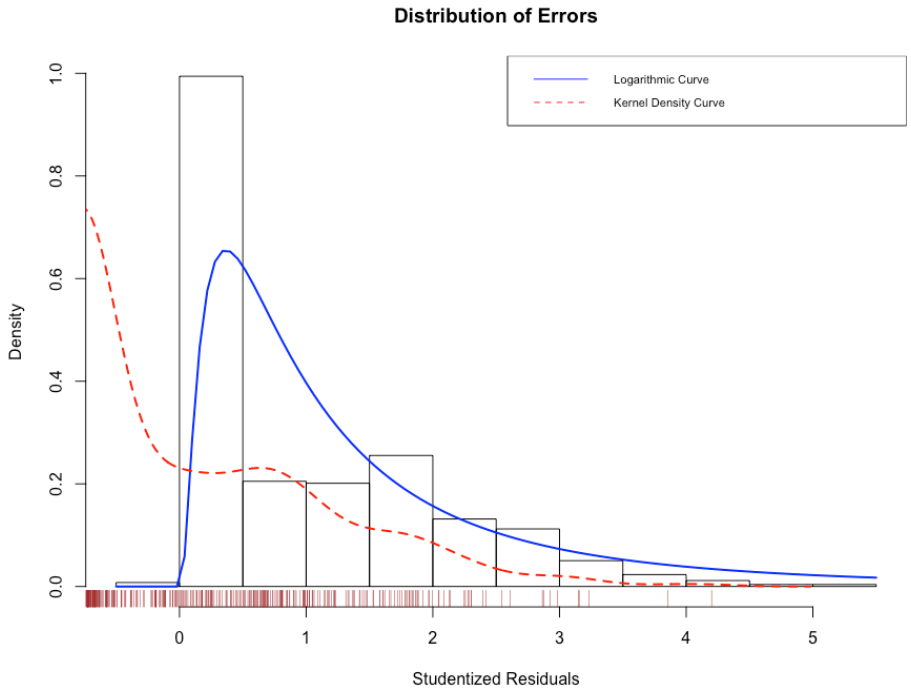
Confidence Intervals

```
                  2.5 %     97.5 %
(Intercept) -0.086287916 0.08628792
Temp        -0.052699195 0.15871717
RH          -0.136361482 0.07021983
Wind        -0.009011874 0.16942489
Rain        -0.070059614 0.10610844
```

Residual Plots



Distribution of Errors



5/6/7 – Select and Interpret the prediction results

The graphs here show the model with following a logarithmic distribution for errors. This is due to the initial log(1+x) transformation we applied to our response variable(Area).
We analyze all our models w.r.t to the first model as it is exhaustive and sufficient.

Model 1 vs Model 2

```
Analysis of Variance Table

Model 1: Area ~ X + Y + Month + Day + Temp + RH + Wind + Rain
Model 2: Area ~ X + Y + Month + Day + FFMC + DMC + DC + ISI + Temp + RH +
    Wind + Rain
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    508 506.36
2    504 501.59  4    4.7681 1.1978 0.3109
```

Model 1 vs Model 3

```
Analysis of Variance Table

Model 1: Area ~ FFMC + DMC + DC + ISI
Model 2: Area ~ X + Y + Month + Day + FFMC + DMC + DC + ISI + Temp + RH +
    Wind + Rain
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    512 511.85
2    504 501.59  8    10.259 1.2886 0.2469
```

Model 1 vs Model 4

```
Analysis of Variance Table

Model 1: Area ~ Temp + RH + Wind + Rain
Model 2: Area ~ X + Y + Month + Day + FFMC + DMC + DC + ISI + Temp + RH +
    Wind + Rain
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    512 510.63
2    504 501.59  8    9.0435 1.1359 0.3373
```

From the ANOVA testing we can see that Model 4 has the most significant p-value. Hence, we select model 4 as our multiple regression model for the response variable of Area. The logic here is that we fail to reject the null hypothesis that (model 4) is not better than the default model 1.