Model Performance on Forest Fire Dataset

Alison Jing Huang

4/15/2018

```
## Warning: package 'caret' was built under R version 3.4.4
```

```
## Warning: package 'randomForest' was built under R version 3.4.4
```

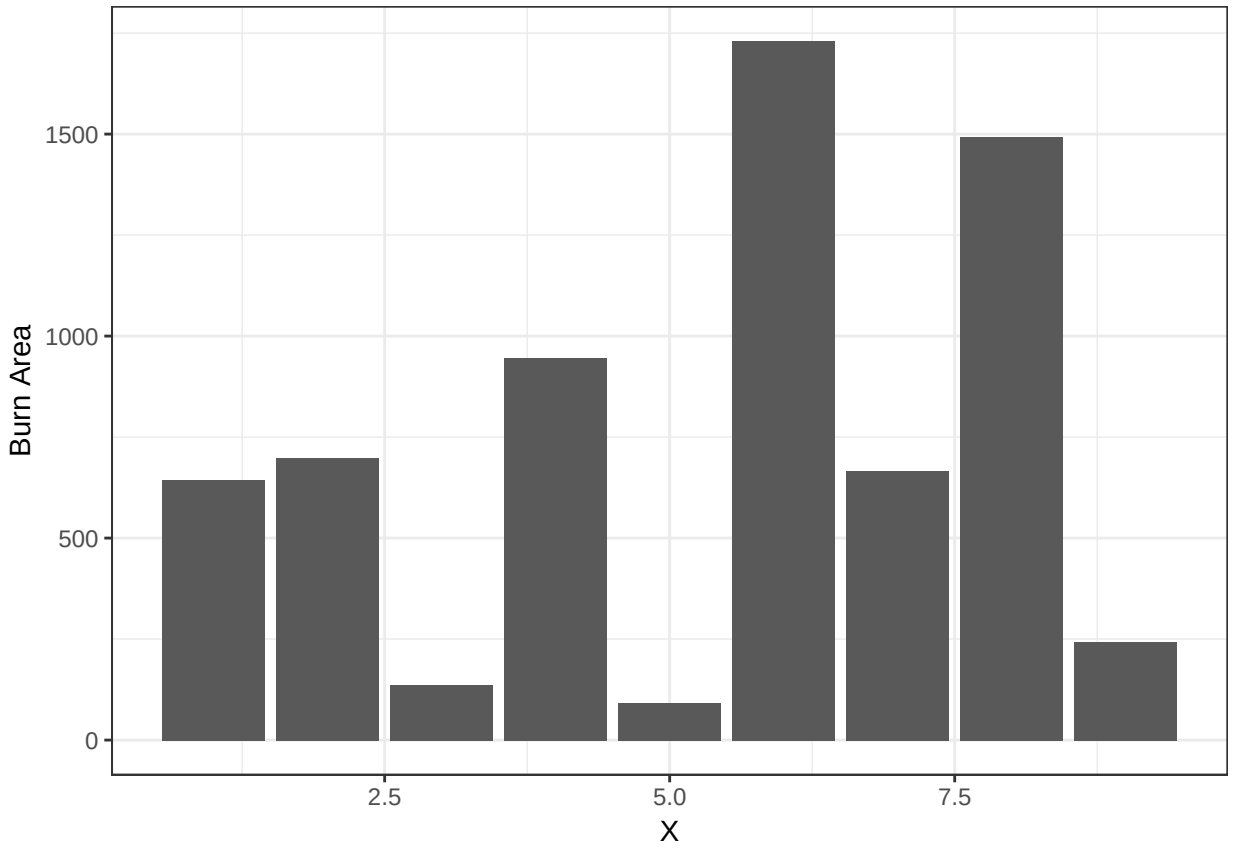
Load the data

```
##   X Y month day FFMC  DMC    DC  ISI temp RH wind rain area
## 1 7 5     8   1 86.2 26.2  94.3  5.1  8.2 51  6.7  0.0    0
## 2 7 4    11   6 90.6 35.4 669.1  6.7 18.0 33  0.9  0.0    0
## 3 7 4    11   3 90.6 43.7 686.9  6.7 14.6 33  1.3  0.0    0
## 4 8 6     8   1 91.7 33.3  77.5  9.0  8.3 97  4.0  0.2    0
## 5 8 6     8   4 89.3 51.3 102.2  9.6 11.4 99  1.8  0.0    0
## 6 8 6     2   4 92.3 85.3 488.0 14.7 22.2 29  5.4  0.0    0
```

After we conducted initial analysis on the dataset, the next step is to create a linear model on the first fires data. Recall earlier we have transform the raw datset to all numerical variables, and renamed it as `fires`, next use ggplot2 to examine different variable with respect to the response variable - **"AREA"**.

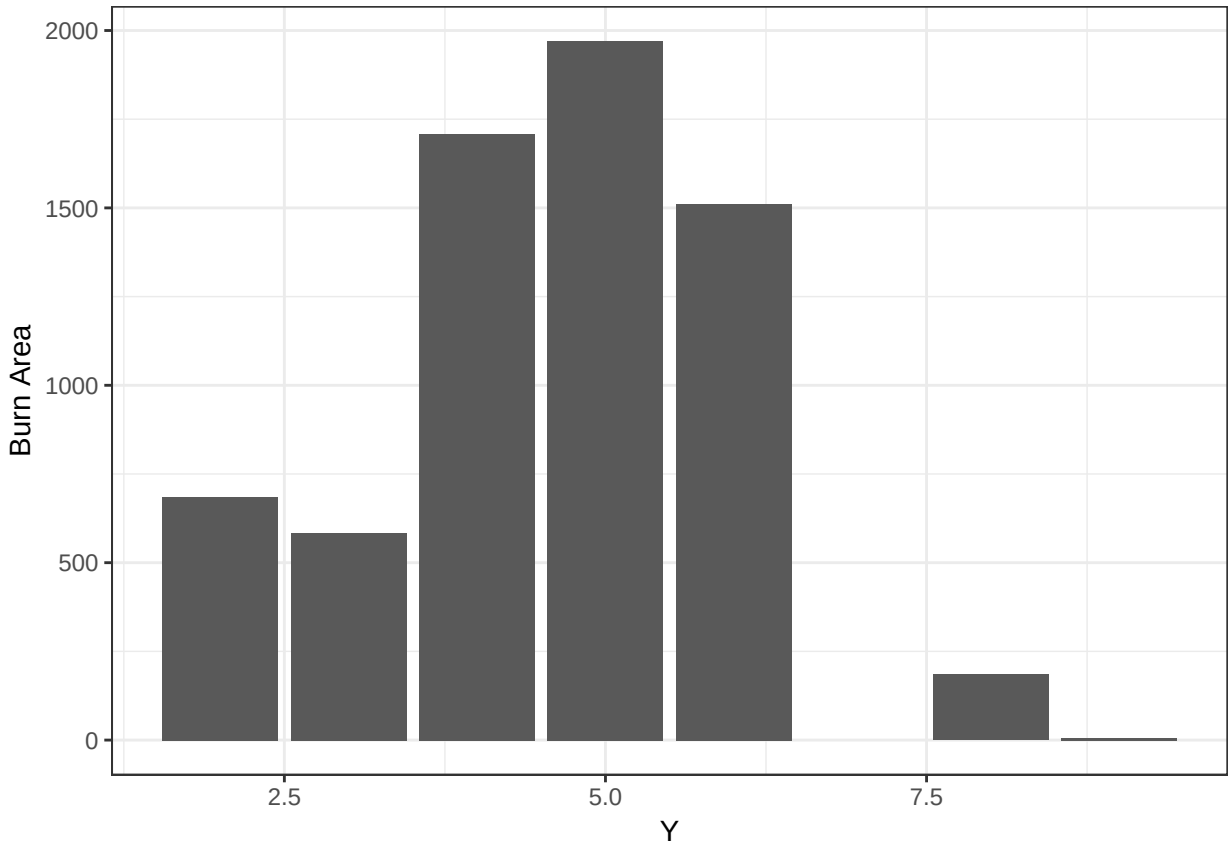
Relationship between X and AREA

```
df %>%
ggplot() +
geom_bar(aes(x = X, y = area),
stat = 'identity') +
labs(x="X", y="Burn Area") +
theme_bw()
```



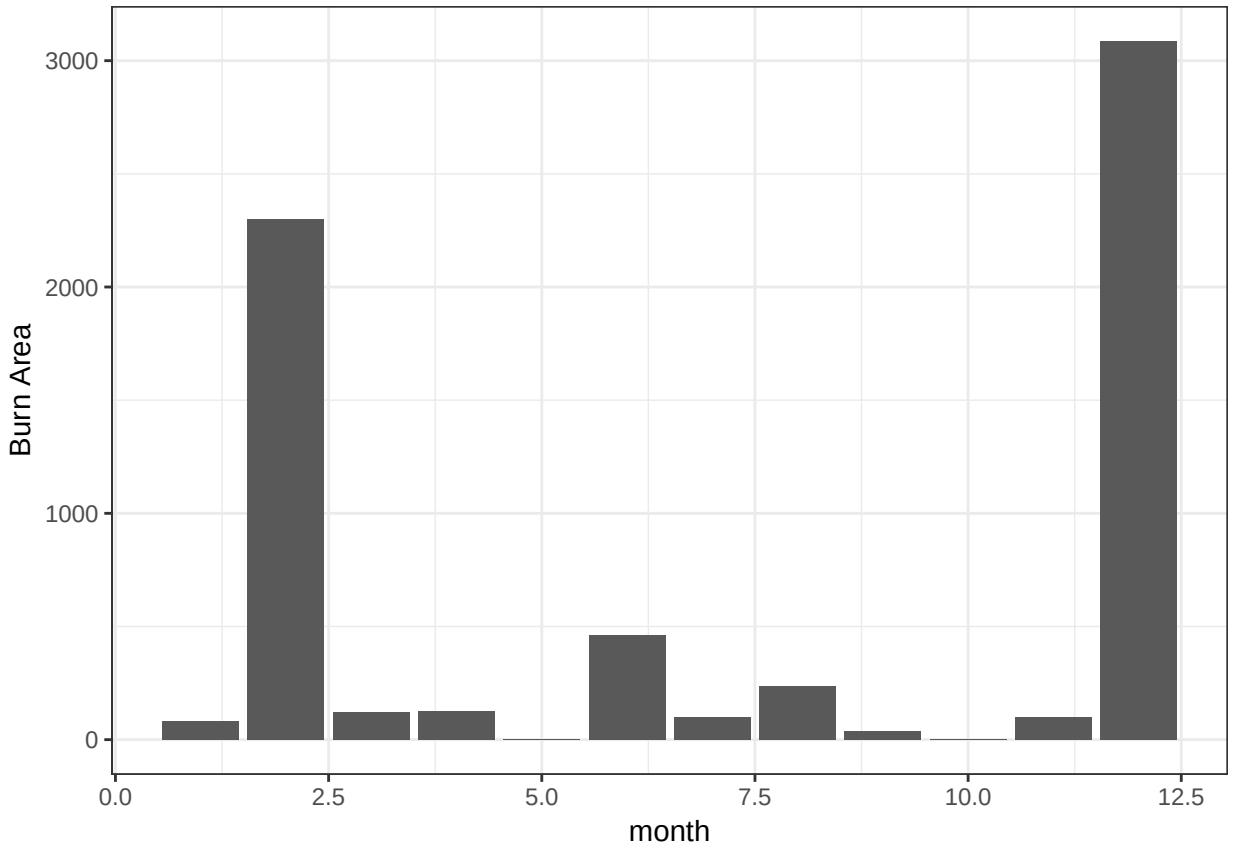
Relationship between Y and AREA

```
df %>%
ggplot() +
geom_bar(aes(x = Y, y = area),
stat = 'identity') +
labs(x="Y", y="Burn Area") +
theme_bw()
```



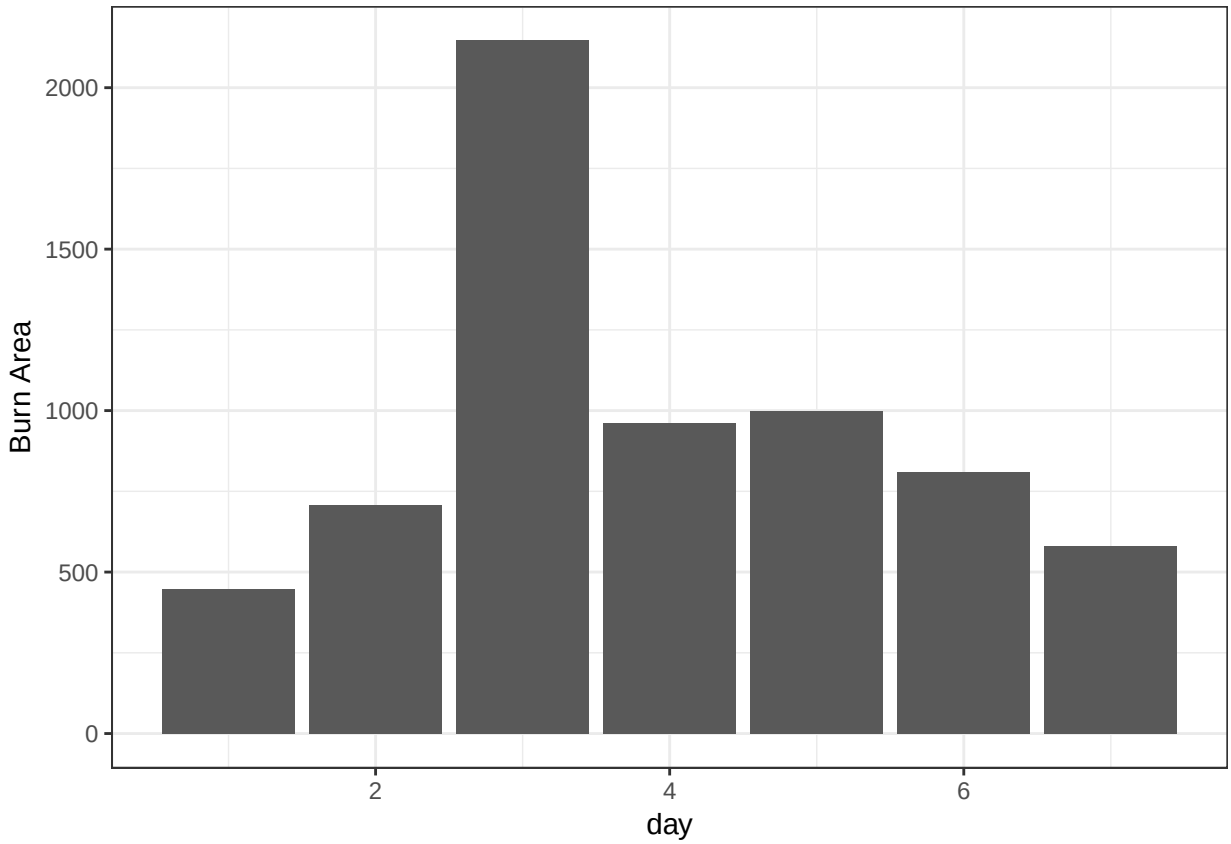
Relationship between MONTH and AREA

```
df %>%
ggplot() +
geom_bar(aes(x = month, y = area),
stat = 'identity') +
labs(x="month", y="Burn Area") +
theme_bw()
```



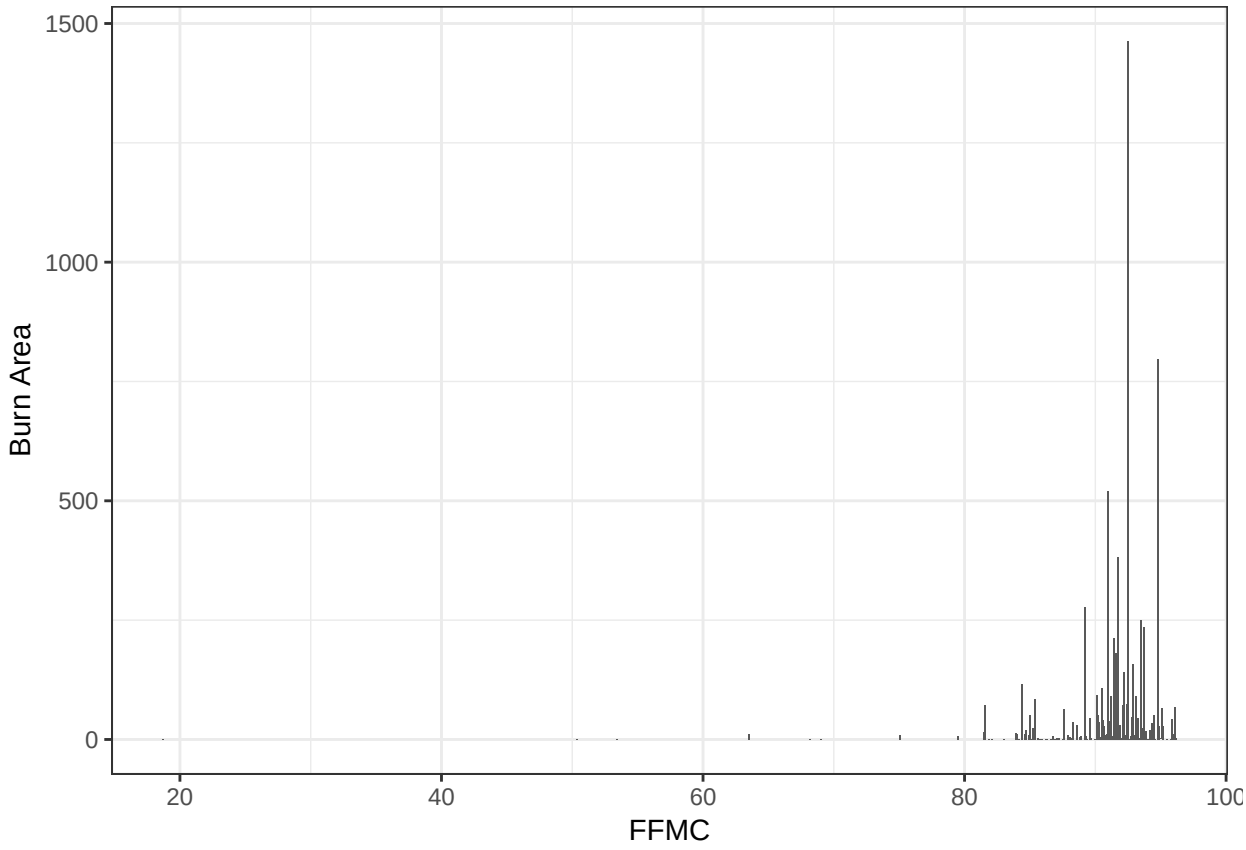
Relationship between DAY and AREA

```
df %>%
ggplot() +
geom_bar(aes(x = day, y = area),
stat = 'identity') +
labs(x="day", y="Burn Area") +
theme_bw()
```



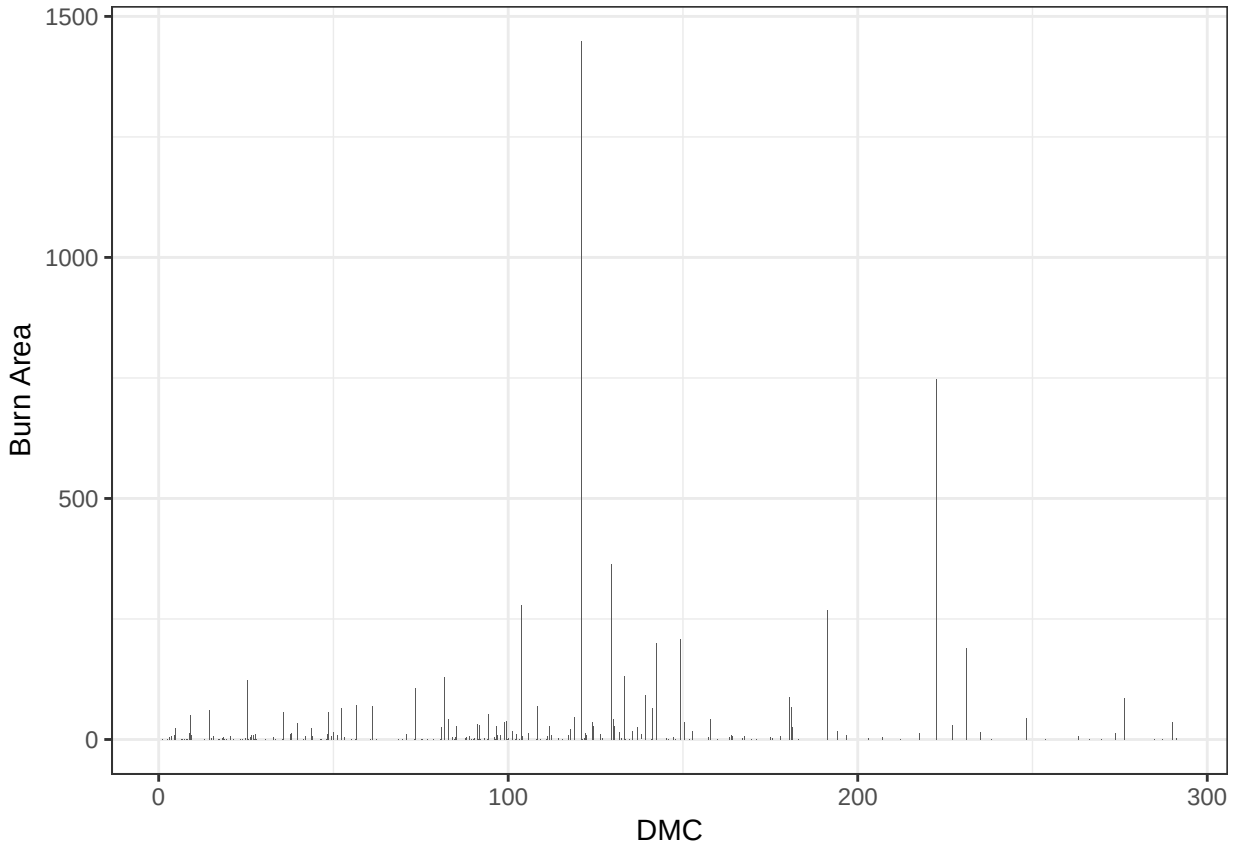
Relationship between FFMC and AREA

```
df %>%
ggplot() +
geom_bar(aes(x = FFMC, y = area),
stat = 'identity') +
labs(x="FFMC", y="Burn Area") +
theme_bw()
```



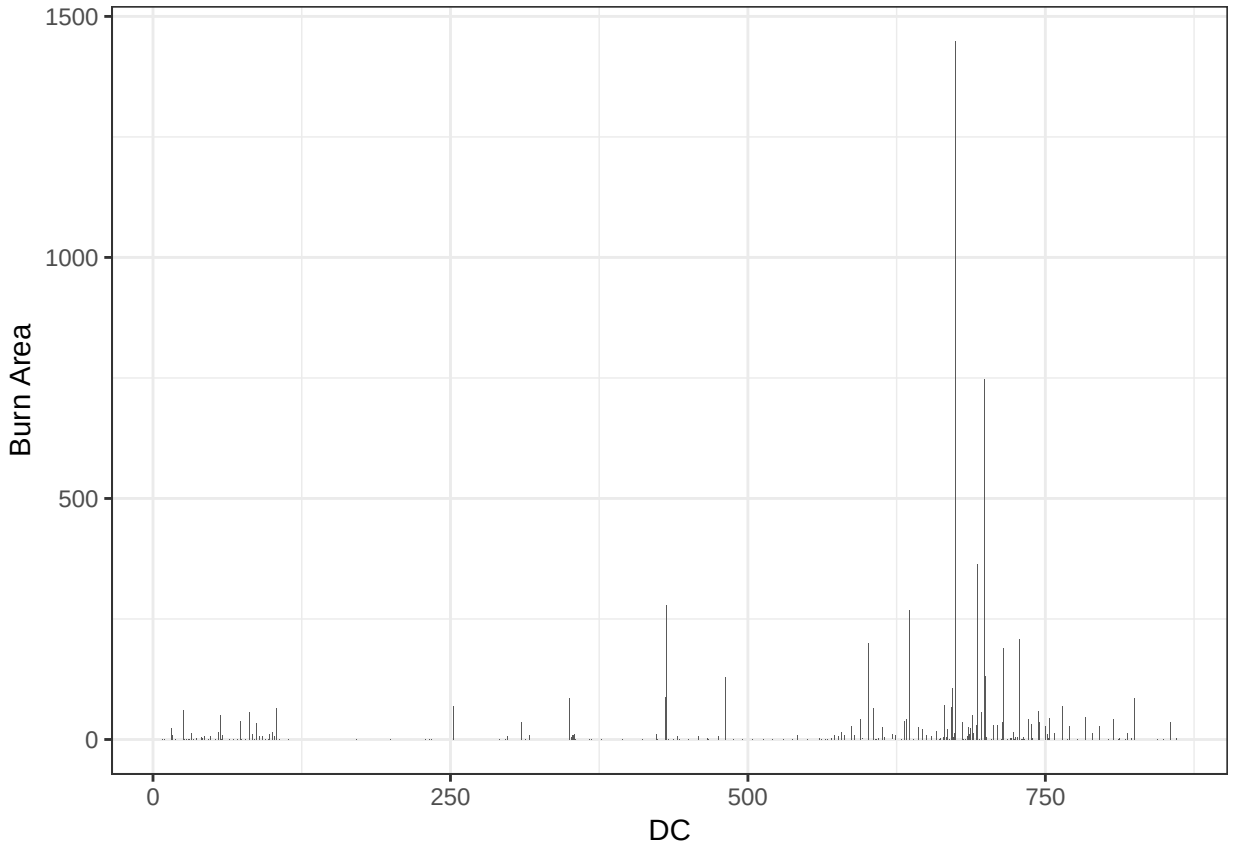
Relationship between DMC and AREA

```
df %>%
ggplot() +
geom_bar(aes(x = DMC, y = area),
stat = 'identity') +
labs(x="DMC", y="Burn Area") +
theme_bw()
```



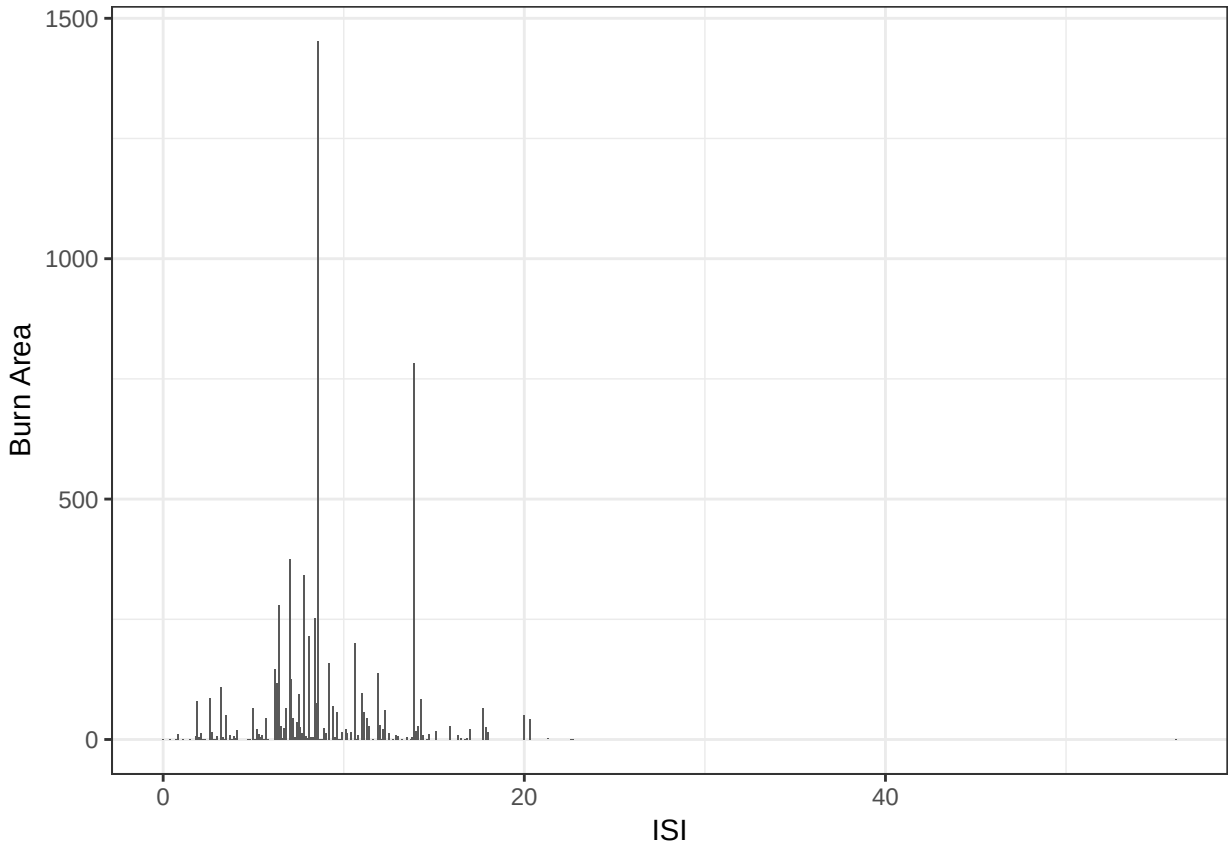
Relationship between DC and AREA

```
df %>%
ggplot() +
geom_bar(aes(x = DC, y = area),
stat = 'identity') +
labs(x="DC", y="Burn Area") +
theme_bw()
```



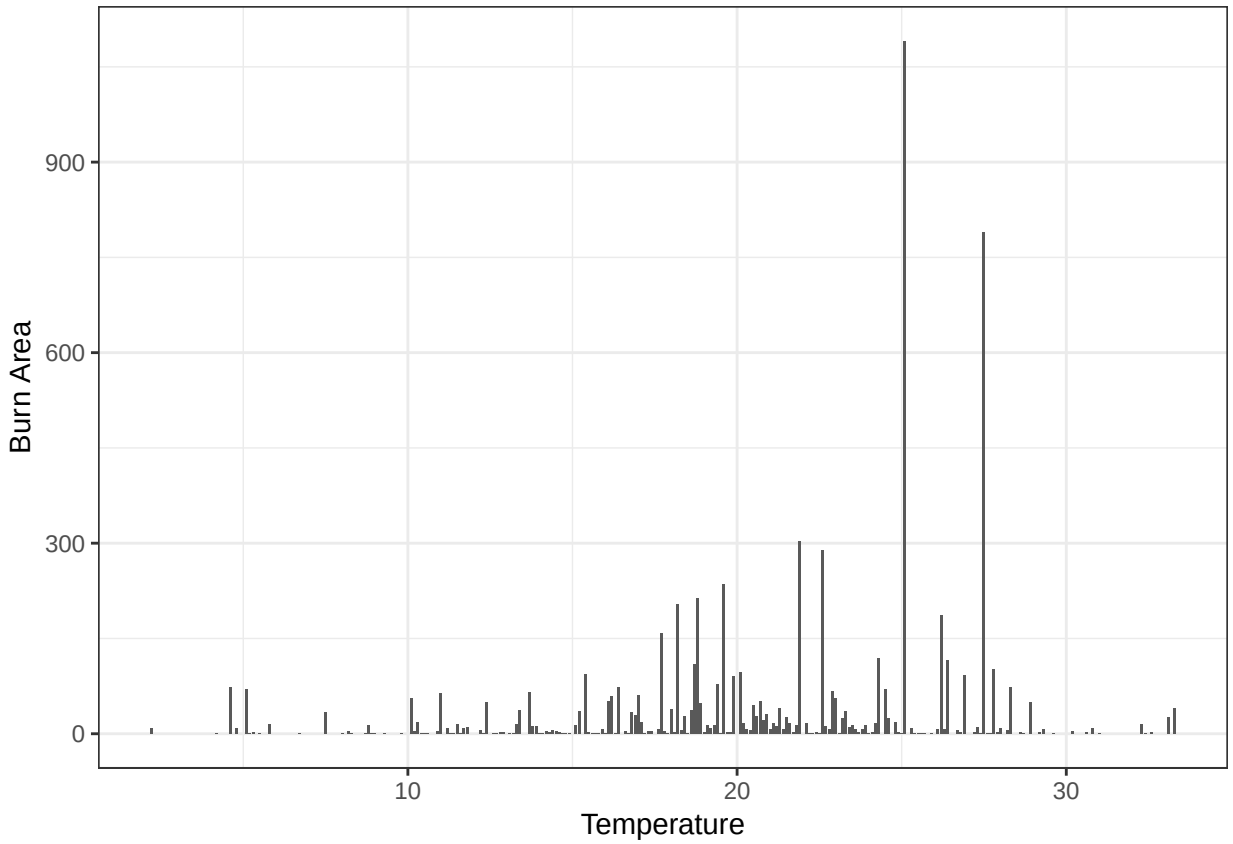
Relationship between ISI and AREA

```
df %>%
ggplot() +
geom_bar(aes(x = ISI, y = area),
stat = 'identity') +
labs(x="ISI", y="Burn Area") +
theme_bw()
```



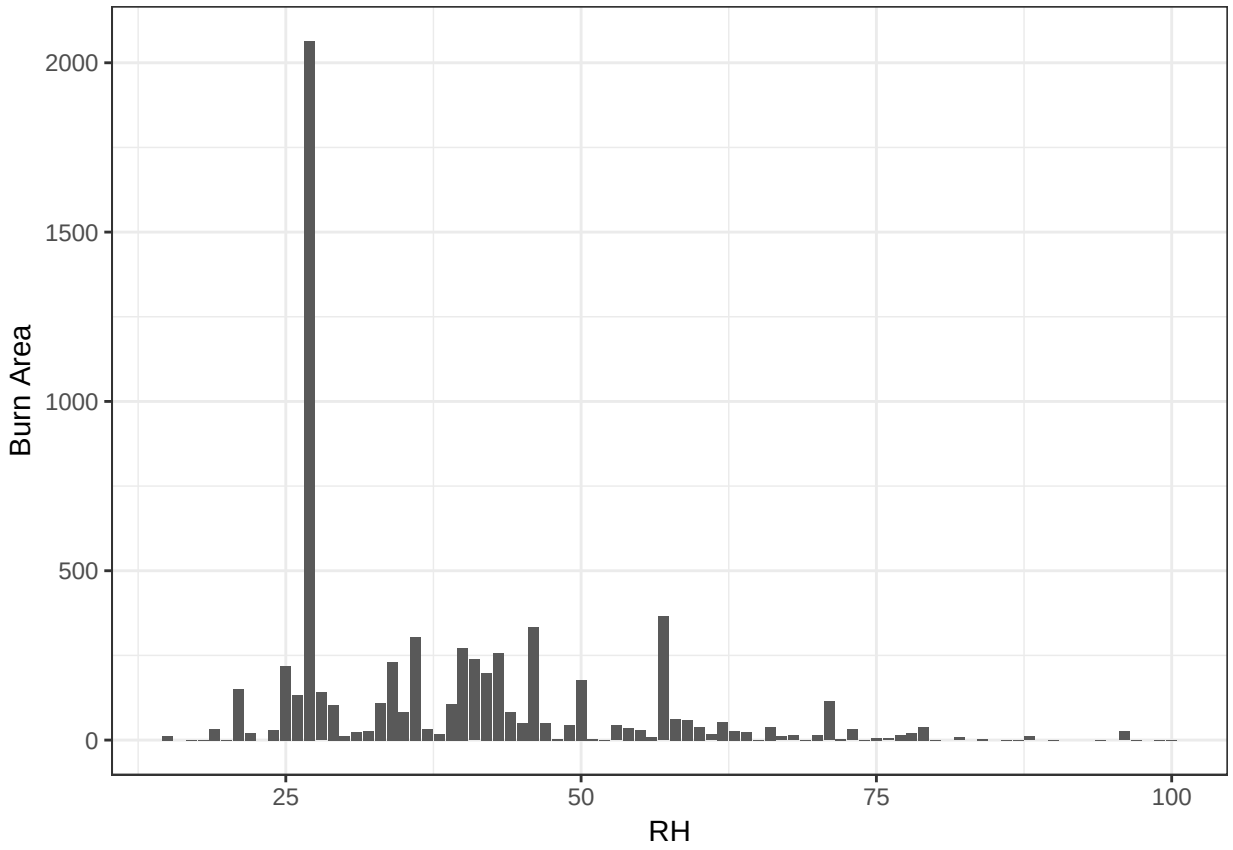
Relationship between TEMPERATURE and AREA

```
df %>%
ggplot() +
geom_bar(aes(x = temp, y = area),
stat = 'identity') +
labs(x="Temperature", y="Burn Area") +
theme_bw()
```



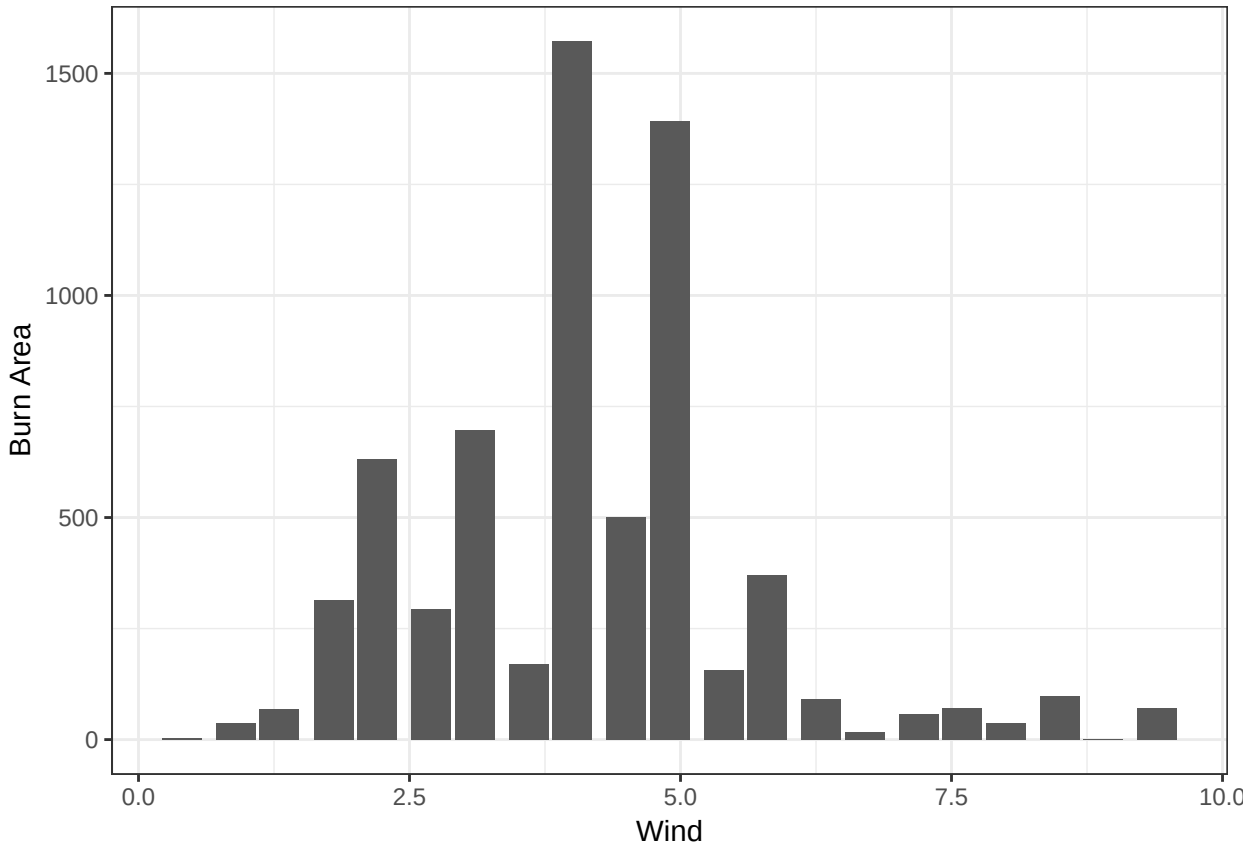
Relationship between RH and AREA

```
df %>%
ggplot() +
geom_bar(aes(x = RH, y = area),
stat = 'identity') +
labs(x="RH", y="Burn Area") +
theme_bw()
```



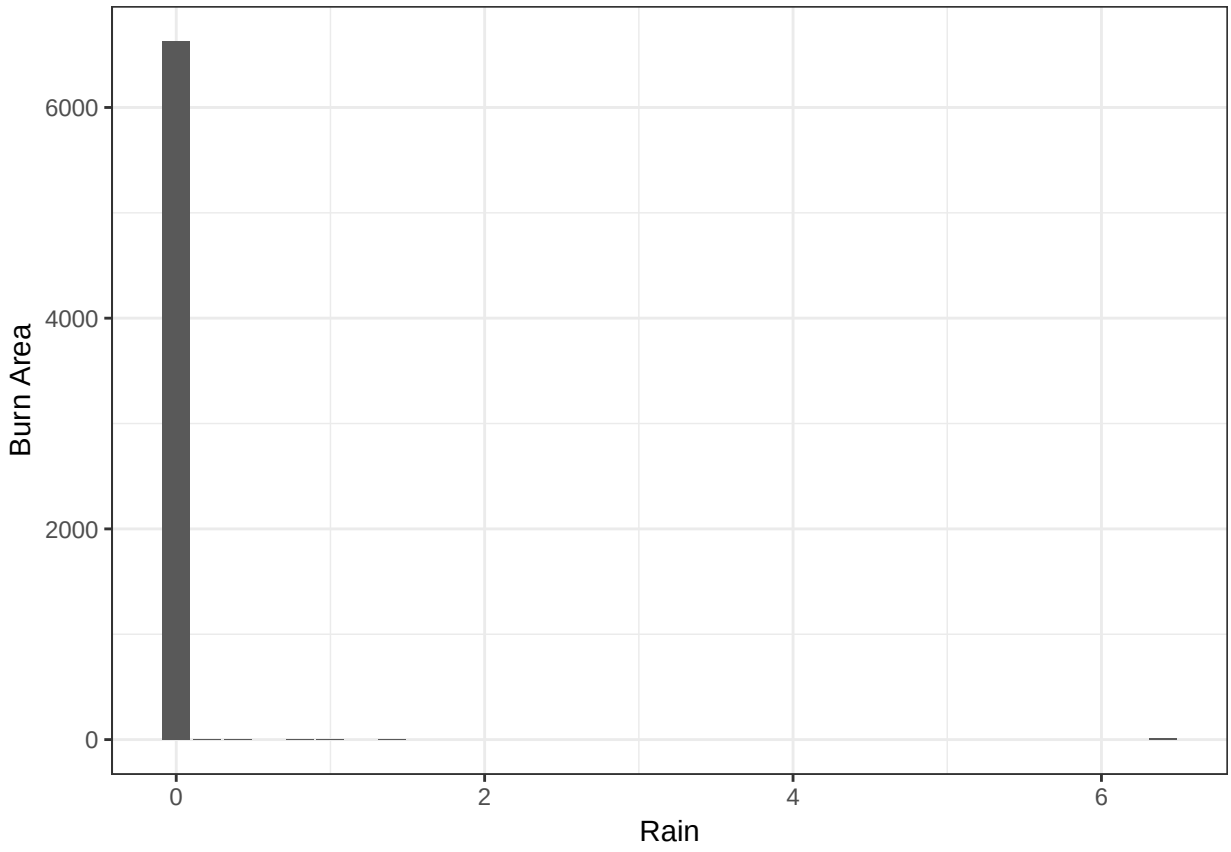
Relationship between wind and AREA

```
df %>%
ggplot() +
geom_bar(aes(x = wind, y = area),
stat = 'identity') +
labs(x="Wind", y="Burn Area") +
theme_bw()
```



Relationship between rain and AREA

```
df %>%
ggplot() +
geom_bar(aes(x = rain, y = area),
stat = 'identity') +
labs(x="Rain", y="Burn Area") +
theme_bw()
```



Model Performance

1. Linear model on Variable "X"

```
lm.x <- lm(df$area ~ df$X, data=df)
lm.x=lm(df$area~df$X)
lm.x
```

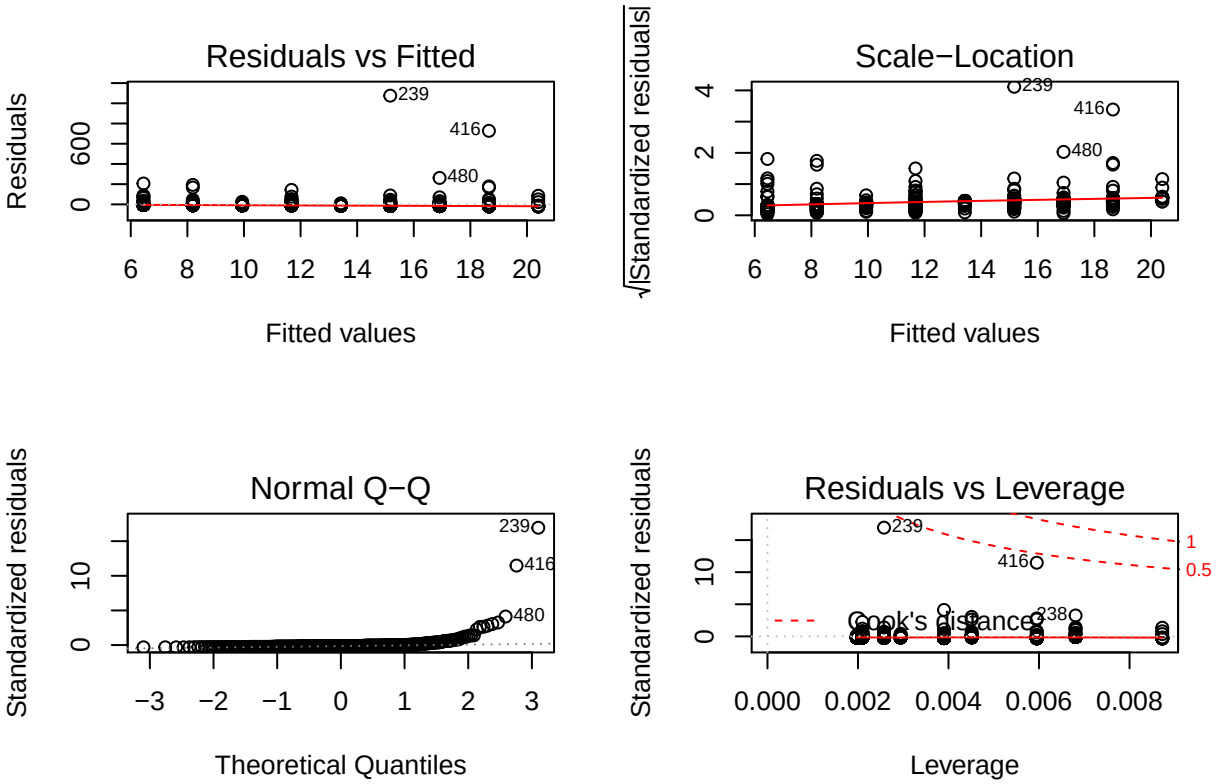
```
##
## Call:
## lm(formula = df$area ~ df$X)
##
## Coefficients:
## (Intercept)         df$X
##       4.705        1.744
```

```
summary(lm.x)
```

```
##
## Call:
## lm(formula = df$area ~ df$X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -20.40  -14.27   -9.94   -5.14 1075.67
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.705      6.304   0.746    0.456
## df$X           1.744      1.210   1.441    0.150
##
## Residual standard error: 63.59 on 515 degrees of freedom
## Multiple R-squared:  0.004018,   Adjusted R-squared:  0.002084
## F-statistic: 2.077 on 1 and 515 DF,  p-value: 0.1501
```

```
layout(matrix(c(1,2,3,4),2,2))
g1 <-plot(lm.x)
```



2. Linear model on Variable "Y"

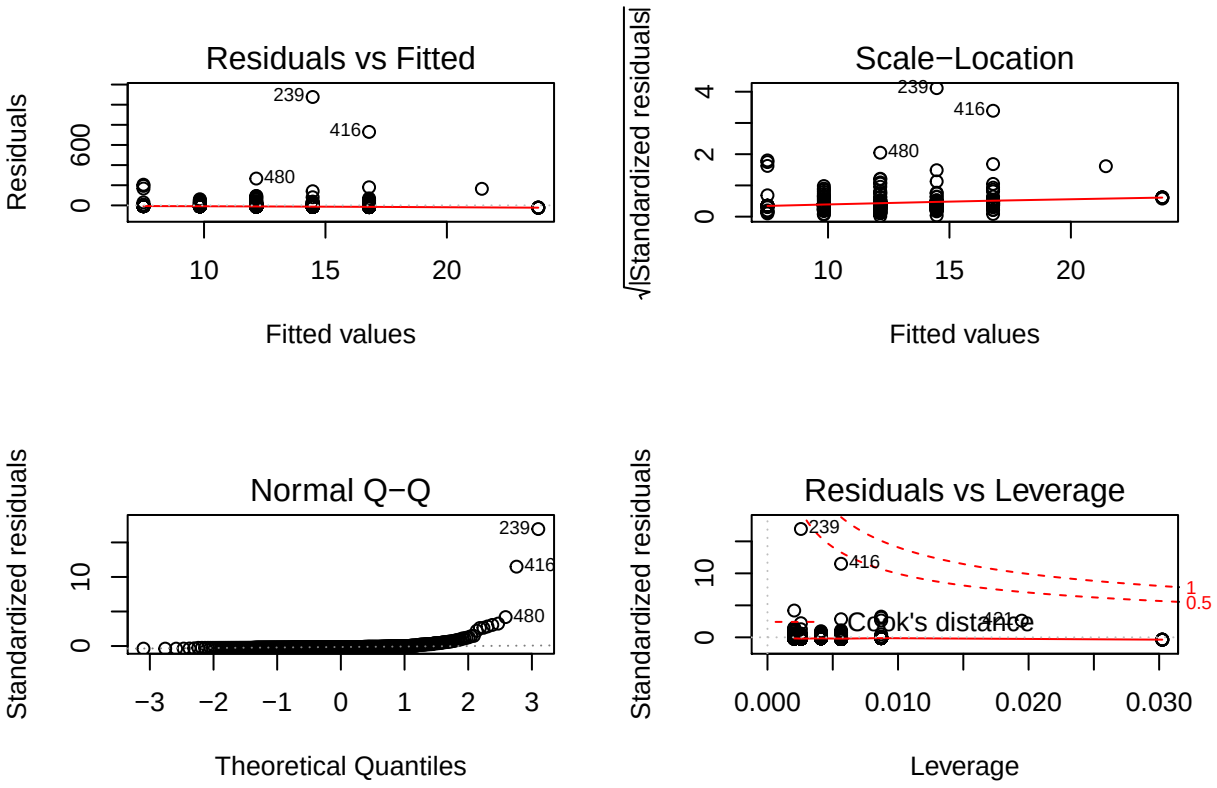
```
lm.y <- lm(df$area ~ df$Y, data=df)
lm.y=lm(df$area~df$Y)
lm.y
```

```
##
## Call:
## lm(formula = df$area ~ df$Y)
##
## Coefficients:
## (Intercept)         df$Y
##       2.861        2.322
```

```r
summary(lm.y)
```

```
##
## Call:
## lm(formula = df$area ~ df$Y)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -23.76  -12.15  -10.50   -6.19 1076.37
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.861     10.189   0.281    0.779
## df$Y           2.322      2.278   1.019    0.309
##
## Residual standard error: 63.65 on 515 degrees of freedom
## Multiple R-squared:  0.002014,   Adjusted R-squared:  7.577e-05
## F-statistic: 1.039 on 1 and 515 DF,  p-value: 0.3085
```

```r
layout(matrix(c(1,2,3,4),2,2))
g2 <-plot(lm.y)
```



3. Linear model on Variable "month"

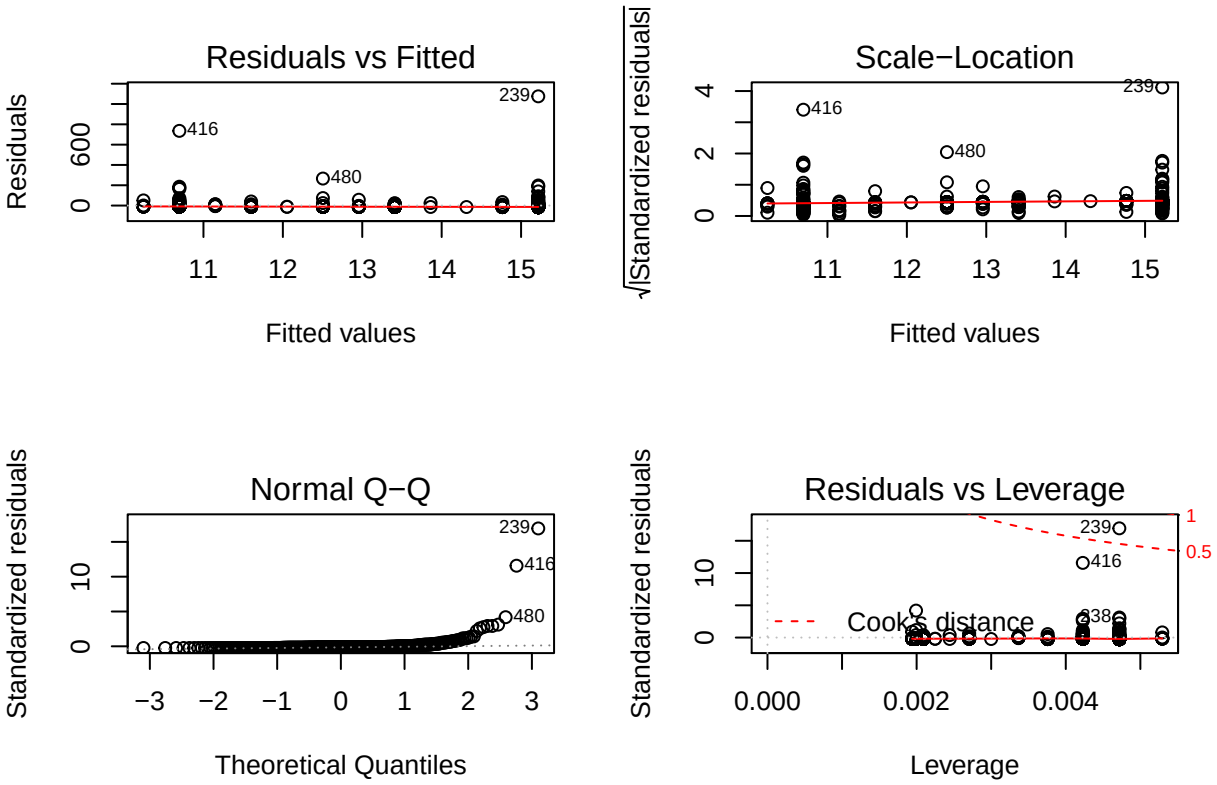
```r
lm.month <- lm(df$area ~ df$month, data=df)
lm.month=lm(df$area~df$month)
lm.month
```

```
##
## Call:
## lm(formula = df$area ~ df$month)
##
## Coefficients:
## (Intercept)     df$month
##       9.793        0.452
```

```
summary(lm.month)
```

```
##
## Call:
## lm(formula = df$area ~ df$month)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -15.22  -13.41  -10.70   -5.93 1075.62
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.7925     5.1592   1.898   0.0582 .
## df$month      0.4520     0.6411   0.705   0.4811
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.69 on 515 degrees of freedom
## Multiple R-squared:  0.0009643,  Adjusted R-squared:  -0.0009755
## F-statistic: 0.4971 on 1 and 515 DF,  p-value: 0.4811
```

```
layout(matrix(c(1,2,3,4),2,2))
g3 <-plot(lm.month)
```



4. Linear model on Variable "day"

```
lm.day <- lm(df$area ~ df$day, data=df)
lm.day=lm(df$area~df$day)
lm.day
```

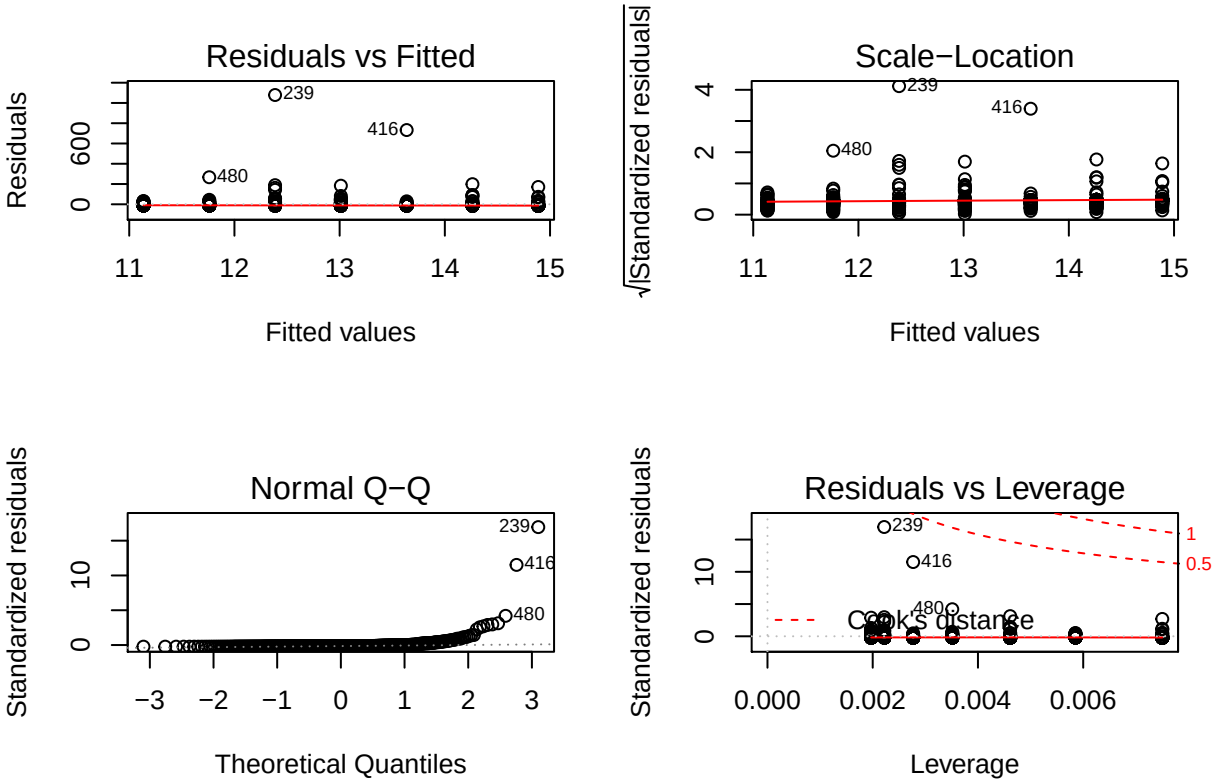
```
##
## Call:
## lm(formula = df$area ~ df$day)
##
## Coefficients:
## (Intercept)       df$day
##     10.5099       0.6255
```

```
summary(lm.day)
```

```
##
## Call:
## lm(formula = df$area ~ df$day)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -14.89  -13.01  -11.14   -6.32 1078.45
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.5099     6.1228   1.717   0.0867 .
## df$day        0.6255     1.4568   0.429   0.6679
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.71 on 515 degrees of freedom
## Multiple R-squared:  0.0003578,  Adjusted R-squared:  -0.001583
## F-statistic: 0.1843 on 1 and 515 DF,  p-value: 0.6679
```

```
layout(matrix(c(1,2,3,4),2,2))
g4 <-plot(lm.day)
```



5. Linear model on Variable "FFMC"

```
lm.FFMC <- lm(df$area ~ df$FFMC, data=df)
lm.FFMC=lm(df$area~df$FFMC)
lm.FFMC
```

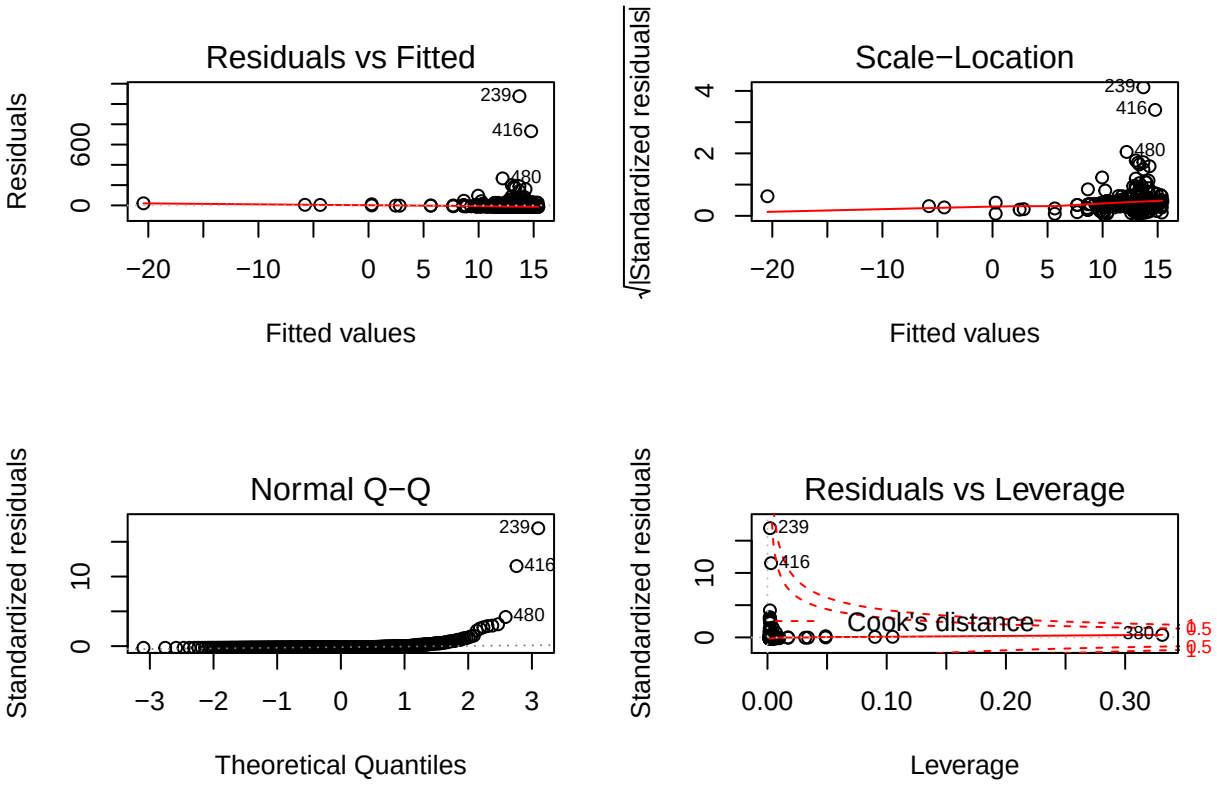
```
##
## Call:
## lm(formula = df$area ~ df$FFMC)
##
## Coefficients:
## (Intercept)      df$FFMC
##    -29.0914       0.4627
```

```
summary(lm.FFMC)
```

```
##
## Call:
## lm(formula = df$area ~ df$FFMC)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -15.42  -13.30  -11.84   -5.81 1077.13
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -29.0914    46.1085  -0.631    0.528
## df$FFMC       0.4627     0.5077   0.911    0.363
##
## Residual standard error: 63.67 on 515 degrees of freedom
## Multiple R-squared:  0.00161,    Adjusted R-squared:  -0.0003288
## F-statistic: 0.8304 on 1 and 515 DF,  p-value: 0.3626
```

```
layout(matrix(c(1,2,3,4),2,2))
g5 <-plot(lm.FFMC)
```



6. Linear model on Variable "DMC"

```
lm.DMC <- lm(df$area ~ df$DMC, data=df)
lm.DMC=lm(df$area~df$DMC)
lm.DMC
```

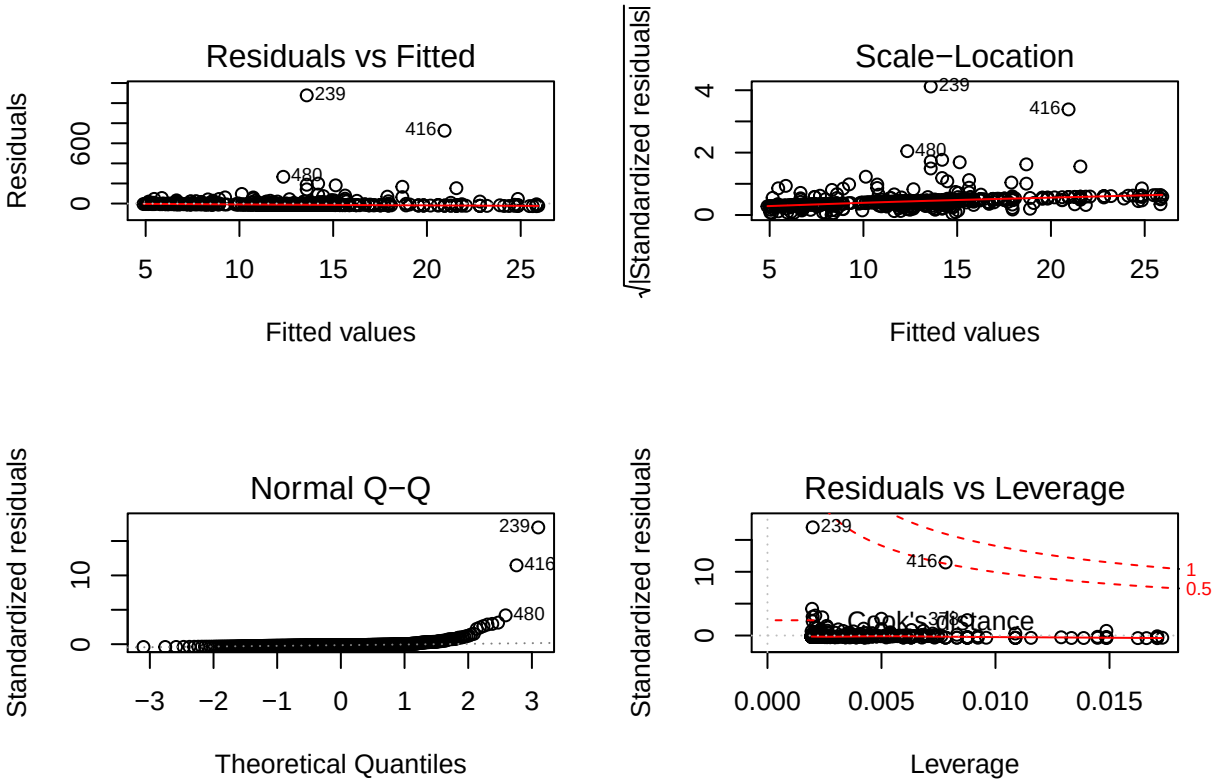
```
##
## Call:
## lm(formula = df$area ~ df$DMC)
##
## Coefficients:
```

```
## (Intercept)       df$DMC
##     4.80361      0.07255
```

```
summary(lm.DMC)
```

```
##
## Call:
## lm(formula = df$area ~ df$DMC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -25.84  -13.48  -10.11   -5.07 1077.25
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.80361    5.59145   0.859   0.3907
## df$DMC       0.07255    0.04368   1.661   0.0973 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.55 on 515 degrees of freedom
## Multiple R-squared:  0.005328,   Adjusted R-squared:  0.003397
## F-statistic: 2.759 on 1 and 515 DF,  p-value: 0.09734
```

```
layout(matrix(c(1,2,3,4),2,2))
g6 <-plot(lm.DMC)
```



7. Linear model on Variable "DC"

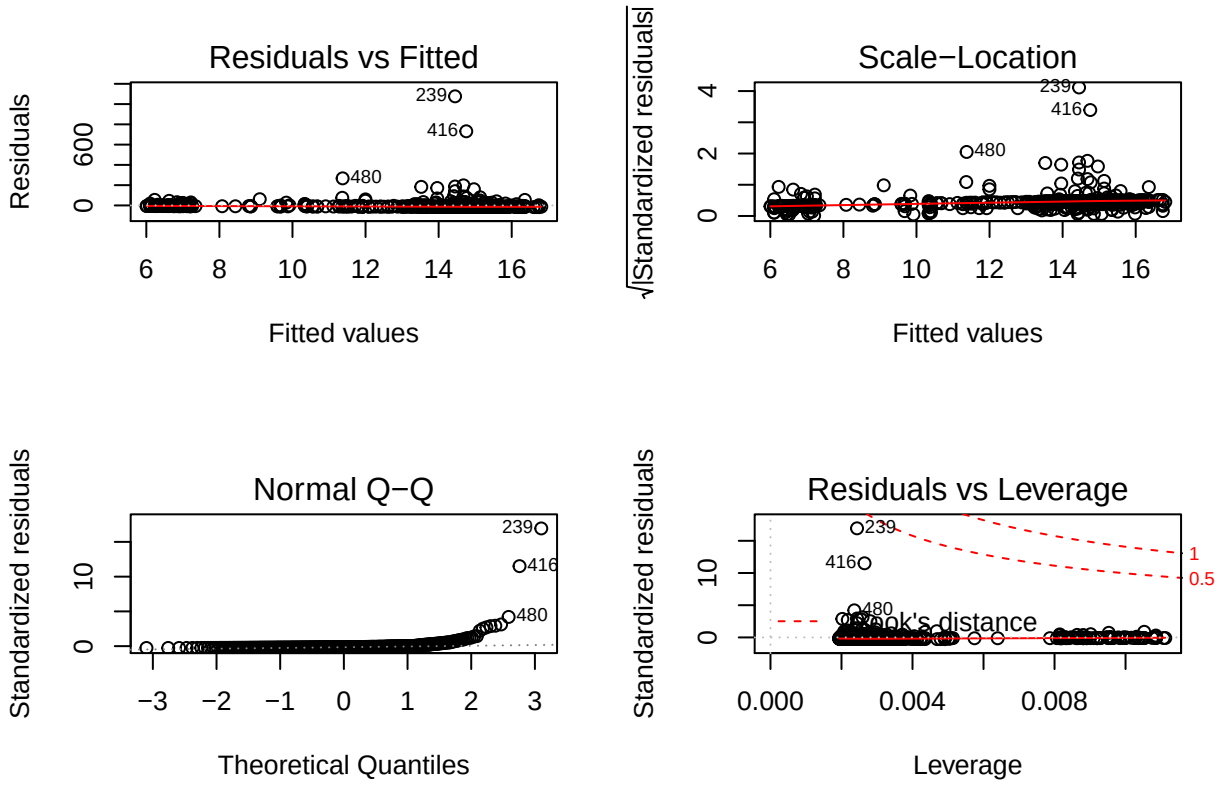
```
lm.DC <- lm(df$area ~ df$DC, data=df)
lm.DC=lm(df$area~df$DC)
lm.DC
```

```
##
## Call:
## lm(formula = df$area ~ df$DC)
##
## Coefficients:
## (Intercept)        df$DC
##     5.90372      0.01267
```

```
summary(lm.DC)
```

```
##
## Call:
## lm(formula = df$area ~ df$DC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -16.74  -14.32  -10.94   -5.36 1076.39
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.90372    6.79180   0.869    0.385
## df$DC        0.01267    0.01129   1.122    0.262
##
## Residual standard error: 63.64 on 515 degrees of freedom
## Multiple R-squared:  0.002439,   Adjusted R-squared:  0.0005017
## F-statistic: 1.259 on 1 and 515 DF,  p-value: 0.2624
```

```
layout(matrix(c(1,2,3,4),2,2))
g7 <-plot(lm.DC)
```



8. Linear model on Variable "ISI"

```
lm.ISI <- lm(df$area ~ df$ISI, data=df)
lm.ISI=lm(df$area~df$ISI)
lm.ISI
```

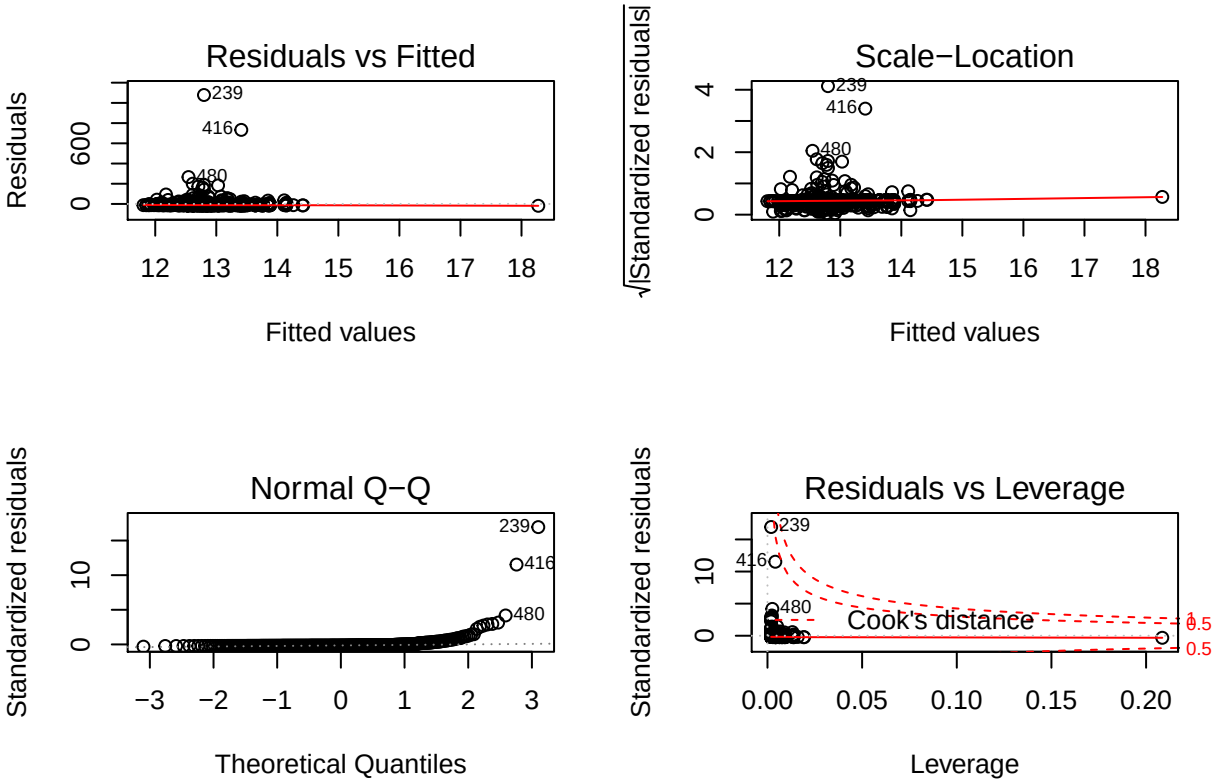
```
##
## Call:
## lm(formula = df$area ~ df$ISI)
##
## Coefficients:
## (Intercept)       df$ISI
##     11.8072       0.1153
```

```
summary(lm.ISI)
```

```
##
## Call:
## lm(formula = df$area ~ df$ISI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -18.27  -12.78  -12.13   -6.19 1078.04
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.8072     6.2173   1.899   0.0581 .
## df$ISI        0.1153     0.6152   0.187   0.8514
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.72 on 515 degrees of freedom
## Multiple R-squared:  6.819e-05,  Adjusted R-squared:  -0.001873
## F-statistic: 0.03512 on 1 and 515 DF,  p-value: 0.8514
```

```
layout(matrix(c(1,2,3,4),2,2))
g8 <-plot(lm.ISI)
```



9. Linear model on Variable "temp"

```
lm.temp <- lm(df$area ~ df$temp, data=df)
lm.temp=lm(df$area~df$temp)
lm.temp
```

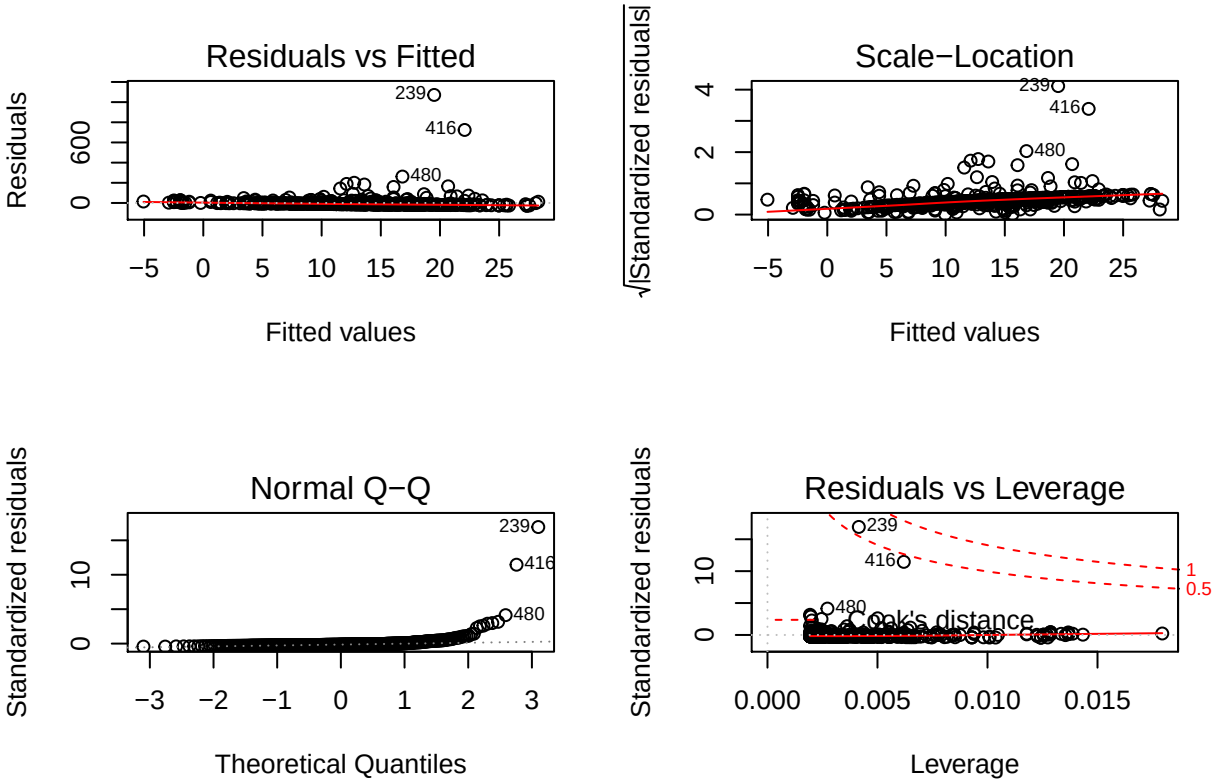
```
##
## Call:
## lm(formula = df$area ~ df$temp)
##
## Coefficients:
## (Intercept)      df$temp
##      -7.414        1.073
```

```
summary(lm.temp)
```

```
##
## Call:
## lm(formula = df$area ~ df$temp)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -27.34  -14.68  -10.39   -3.42 1071.33
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.4138     9.4996  -0.780   0.4355
## df$temp       1.0726     0.4808   2.231   0.0261 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.41 on 515 degrees of freedom
## Multiple R-squared:  0.009573,   Adjusted R-squared:  0.00765
## F-statistic: 4.978 on 1 and 515 DF,  p-value: 0.0261
```

```
layout(matrix(c(1,2,3,4),2,2))
g9 <-plot(lm.temp)
```



10. Linear model on Variable "RH"

```
lm.RH <- lm(df$area ~ df$RH, data=df)
lm.RH=lm(df$area~df$RH)
lm.RH
```

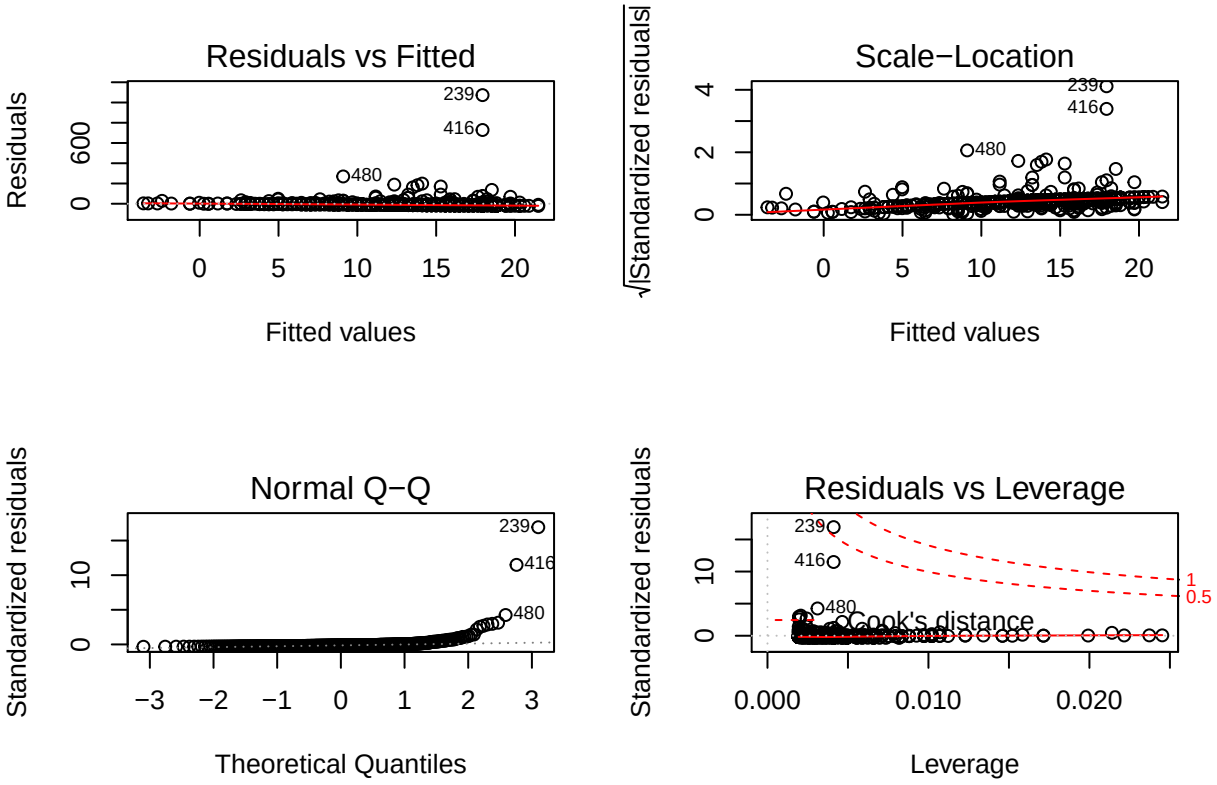
```
##
## Call:
## lm(formula = df$area ~ df$RH)
```

```
##
## Coefficients:
## (Intercept)        df$RH
##     25.8948      -0.2946
```

```
summary(lm.RH)
```

```
##
## Call:
## lm(formula = df$area ~ df$RH)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -21.48  -14.41  -10.58   -3.48 1072.90
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.8948     8.0894   3.201  0.00145 **
## df$RH        -0.2946     0.1714  -1.719  0.08627 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.54 on 515 degrees of freedom
## Multiple R-squared:  0.005703,   Adjusted R-squared:  0.003772
## F-statistic: 2.954 on 1 and 515 DF,  p-value: 0.08627
```

```
layout(matrix(c(1,2,3,4),2,2))
g10 <-plot(lm.RH)
```

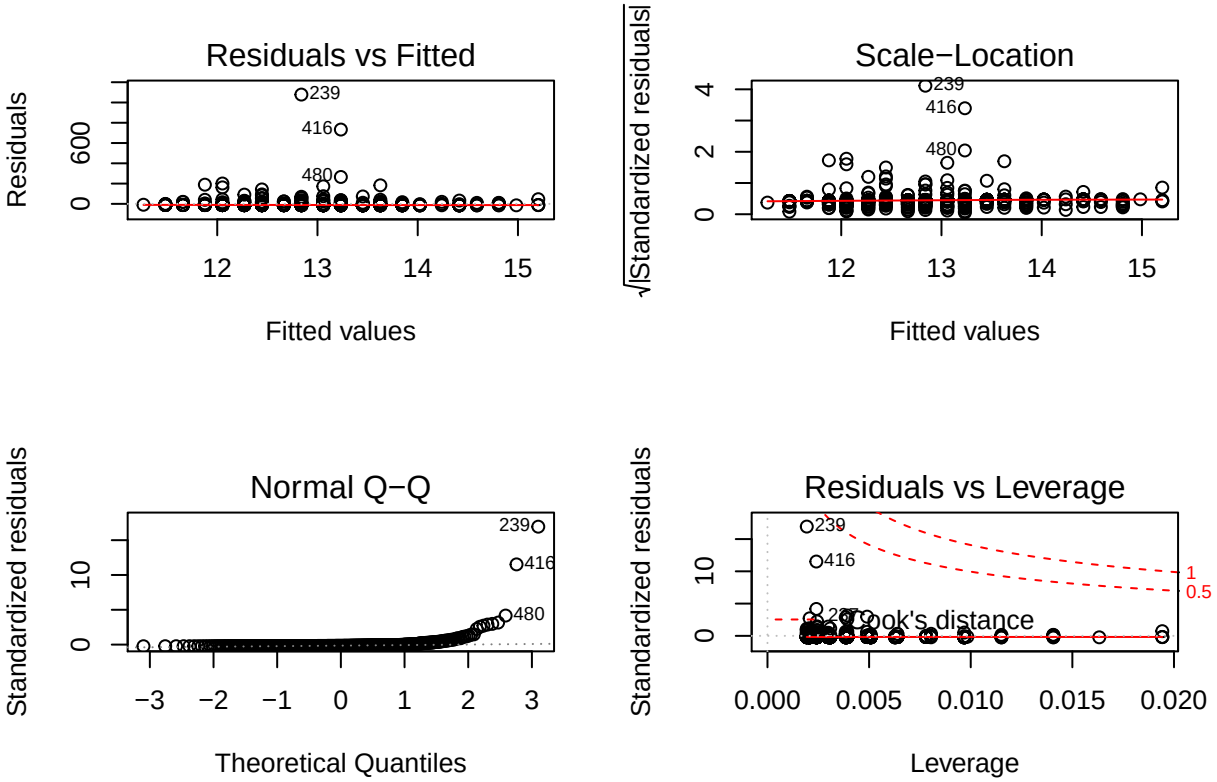


11. Linear model on Variable “wind”

```
lm.wind <- lm(df$area ~ df$wind, data=df)
lm.wind=lm(df$area~df$wind)
summary(lm.wind)
```

```
##
## Call:
## lm(formula = df$area ~ df$wind)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -14.81  -12.84  -11.88   -6.08 1078.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.0891     6.8854   1.611    0.108
## df$wind       0.4376     1.5655   0.280    0.780
##
## Residual standard error: 63.71 on 515 degrees of freedom
## Multiple R-squared:  0.0001517,  Adjusted R-squared:  -0.00179
## F-statistic: 0.07815 on 1 and 515 DF,  p-value: 0.7799
```

```
layout(matrix(c(1,2,3,4),2,2))
g11 <- plot(lm.wind)
```



12. Linear model on Variable "rain"

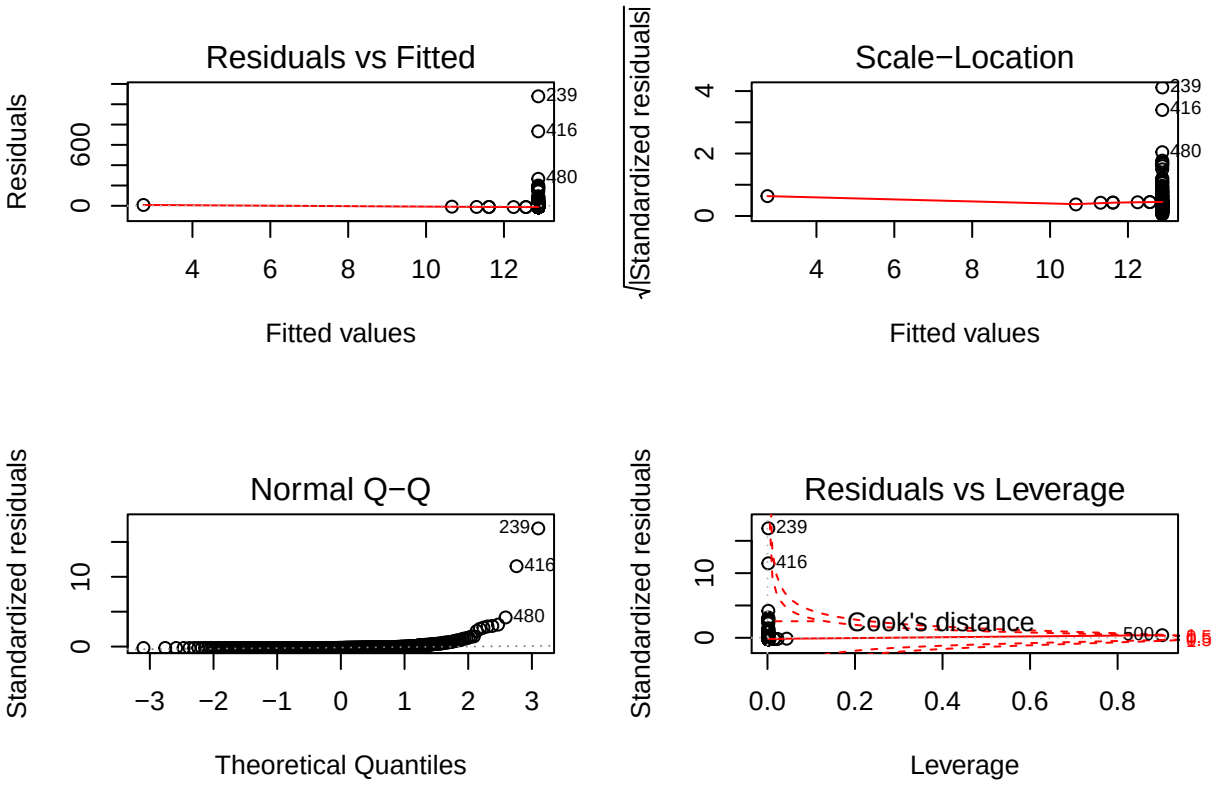
```
lm.rain <- lm(df$area ~ df$rain, data=df)
lm.rain=lm(df$area~df$rain)
lm.rain
```

```
##
## Call:
## lm(formula = df$area ~ df$rain)
##
## Coefficients:
## (Intercept)      df$rain
##      12.882       -1.584
```

```
summary(lm.rain)
```

```
##
## Call:
## lm(formula = df$area ~ df$rain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -12.88  -12.88  -12.25   -6.31 1077.96
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   12.882      2.810   4.585 5.71e-06 ***
## df$rain       -1.584      9.477  -0.167    0.867
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.72 on 515 degrees of freedom
## Multiple R-squared:  5.425e-05,  Adjusted R-squared:  -0.001887
## F-statistic: 0.02794 on 1 and 515 DF,  p-value: 0.8673
```

```
layout(matrix(c(1,2,3,4),2,2))
g12 <- plot(lm.rain)
```



13. Linear model including on the 12 attribute variables

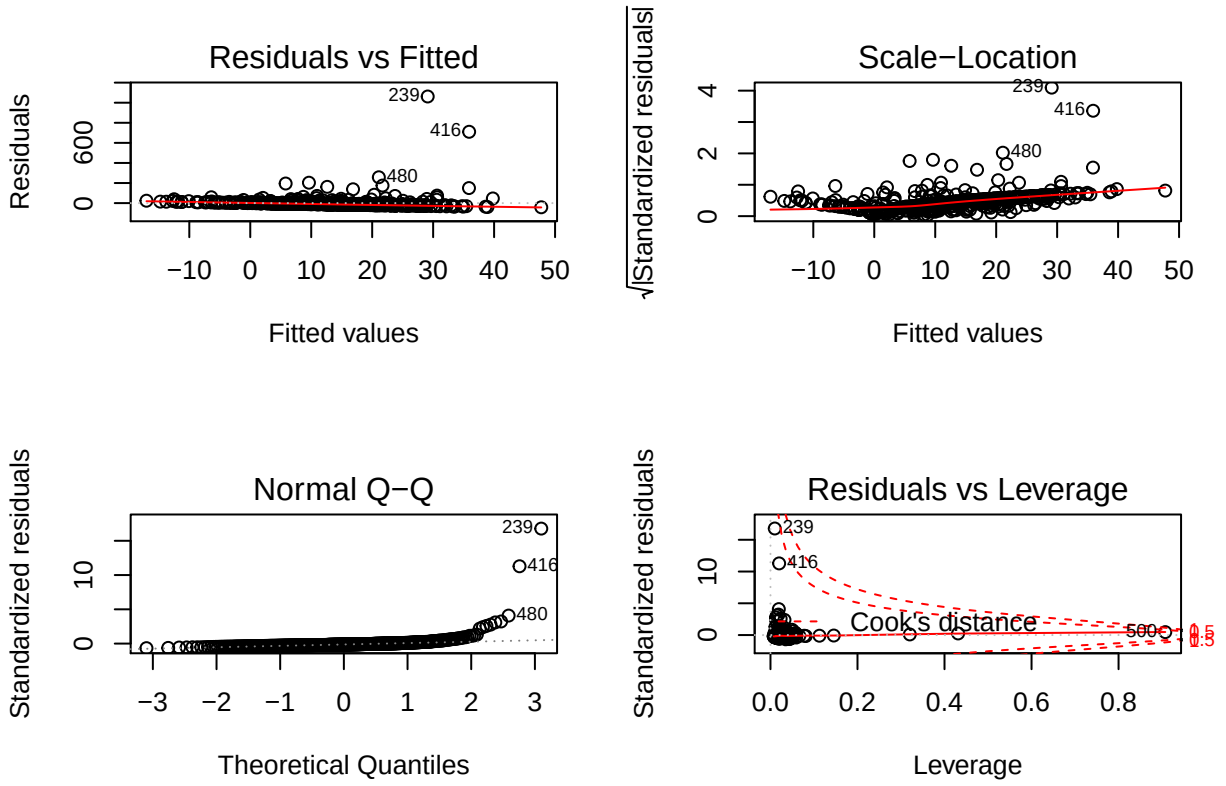
```
lm_all <- lm(df$area ~ ., data=df)
lm_all=lm(df$area~ df$X + df$Y + df$month + df$day + df$FFMC + df$DMC + df$DC + df$ISI + df$temp + df$RH
lm_all
```

```
##
## Call:
## lm(formula = df$area ~ df$X + df$Y + df$month + df$day + df$FFMC +
##     df$DMC + df$DC + df$ISI + df$temp + df$RH + df$wind + df$rain,
##     data = df)
##
## Coefficients:
## (Intercept)          df$X          df$Y      df$month        df$day
##   -12.97338       1.88124       0.52680       0.97328       0.49953
##     df$FFMC        df$DMC         df$DC        df$ISI       df$temp
##    -0.10740       0.10980      -0.01463      -0.61081       0.98013
##       df$RH      df$wind       df$rain
##    -0.18492       1.78229      -3.25171
```

```
summary
```

```
## function (object, ...)
## UseMethod("summary")
## <bytecode: 0x7fd7a02294d8>
## <environment: namespace:base>
```

```
layout(matrix(c(1,2,3,4),2,2))
g13<- plot(lm_all)
```


Residuals vs Fitted
Residuals
Fitted values
239
416
480
Scale−Location
√|Standardized residuals|
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Residuals vs Leverage
Cook's distance
Leverage