Naive Model Performance on Forest Fire Dataset

Alison Jing Huang

4/15/2018

```
#setwd("~/Desktop/CSX415-Data-Science-and-Principles/csx415-project/ForestFire")
#library(ProjectTemplate)
#load.project()
```

```
##   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
```

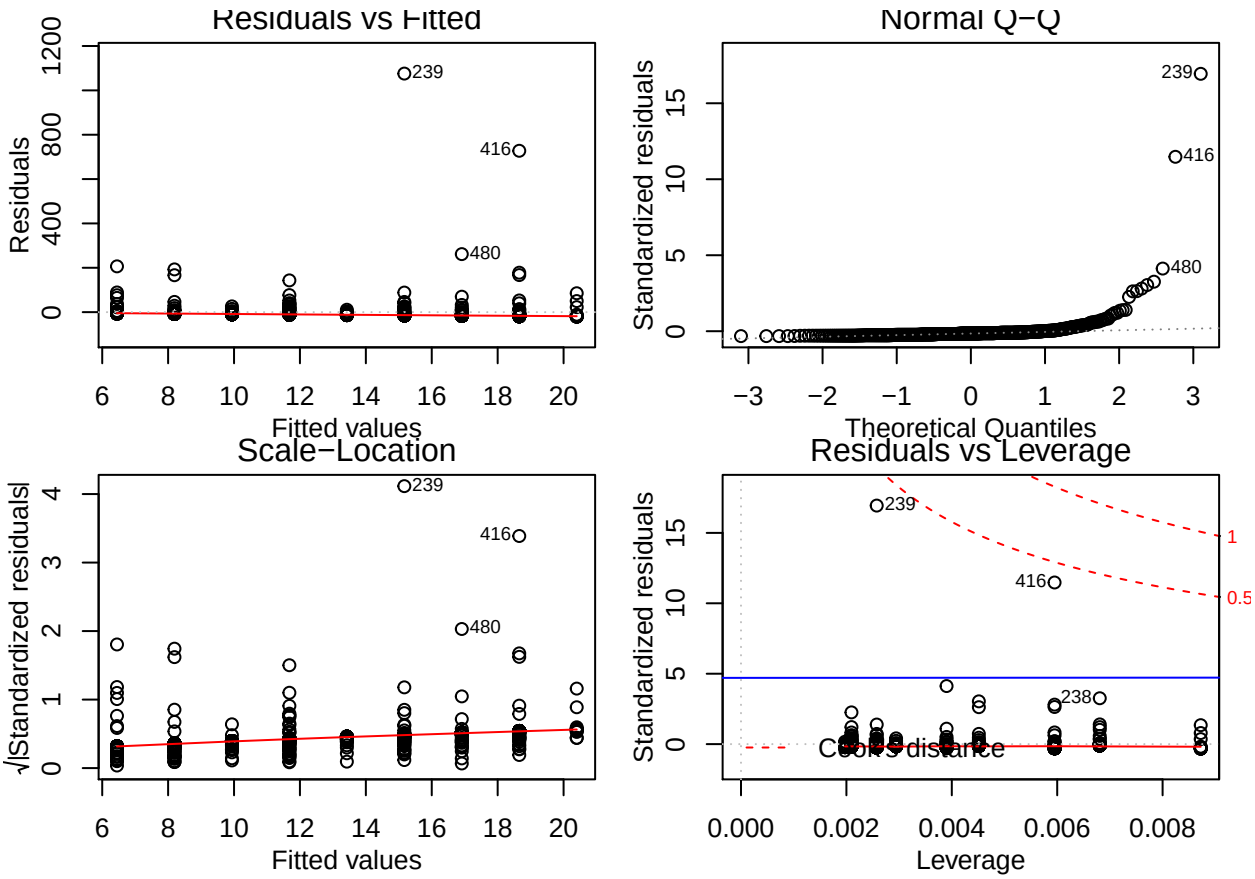
Model Performance on Naive models with Training set

1. Linear model on Variable "X"

```
par(mgp=c(2,1,0), mar=c(3,3,1,1))
require(stats)
lm_x <- lm(area ~ X, data=df)
glance(lm_x)
```

```
##     r.squared adj.r.squared    sigma statistic   p.value df    logLik
## 1 0.004017696    0.00208375 63.58946   2.07746 0.1500965  2 -2879.405
##       AIC      BIC deviance df.residual
## 1 5764.81 5777.554  2082464         515
```

```
par(mfrow=c(2,2))
coeff=coefficients(lm_x)
# equation of the line :
# plot
plot(lm_x)
abline(lm(df$area~df$X), col="blue")
```



Conlusion :The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable.Based on above result, p-value of Variable X gives 0.1501 which is greater than common alpha level of 0.05, this indicates that **variable X is not statistically significant and hence not a very good predictor variable**, and the R^2 gives 0.004018 which is very close to 0.In addition, The F value is the ratio of the mean regression sum of squares divided by the mean error sum of squares that is displaying a value of 2.077.

2. Linear model on Variable "Y"

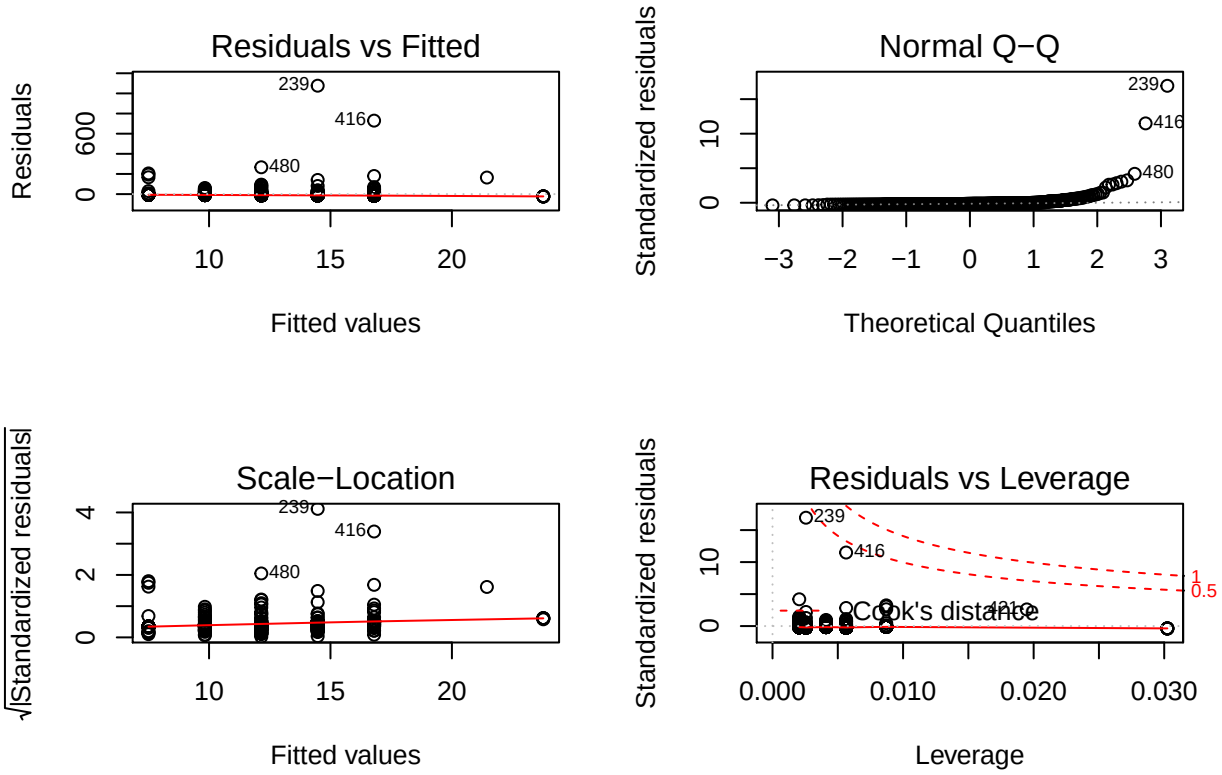
```
lm_y <- lm(area ~ Y, data=df)
lm_y=lm(df$area~df$Y)
lm_y
```

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

```
par(mfrow=c(2,2))
glance(lm_y)
```

```
##      r.squared adj.r.squared    sigma statistic   p.value df    logLik
## 1 0.002013606  7.576868e-05 63.65341    1.0391 0.3085096  2 -2879.924
##        AIC      BIC deviance df.residual
## 1 5765.849 5778.593  2086654         515
```

```
plot(lm_y)
```



Conlusion: variable Y is not statistically significant and hence is a mediocre(not the best) predictor variable since its p-value is 0.3085096, and the R^2 gives 0.002 which is very close to 0.

3. Linear model on Variable "Month"

```
lm_month <- lm(area ~ 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
```

```
par(mfrow=c(2,2))
glance(lm_month)
```

```
##      r.squared adj.r.squared    sigma statistic   p.value df    logLik
## 1 0.0009643454 -0.0009755296 63.68686 0.4971173 0.4810883  2 -2880.196
##        AIC      BIC deviance df.residual
```

```
## 1 5766.392 5779.136  2088848         515
```

Conlusion: **variable Month is not statistically significant and hence is a mediocre(not the best) predictor variable since its p-value is 0.4810**, and the R^2 gives 0.000.

4. Linear model on Variable "Day"

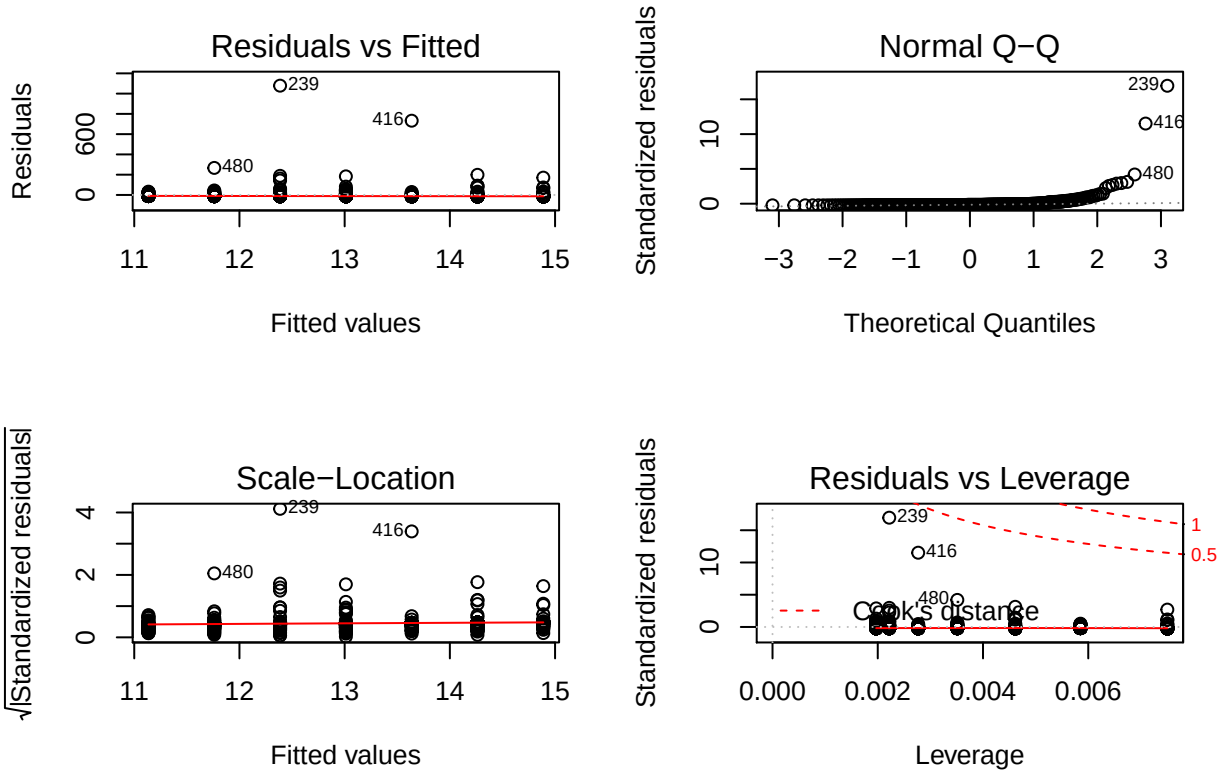
```
lm_day<- lm(area ~ day, data=df)
lm_day=lm(df$area~df$day)
lm_day
```

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

```
par(mfrow=c(2,2))
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
```

```
plot(lm_day)
```



Conlusion: **variable Day is not statistically insignificant because its p-value is 0.6679, therefore making "Day" not a good predictor.**, and the R^2 gives 0.000.

5. Linear model on Variable "FFMC"

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

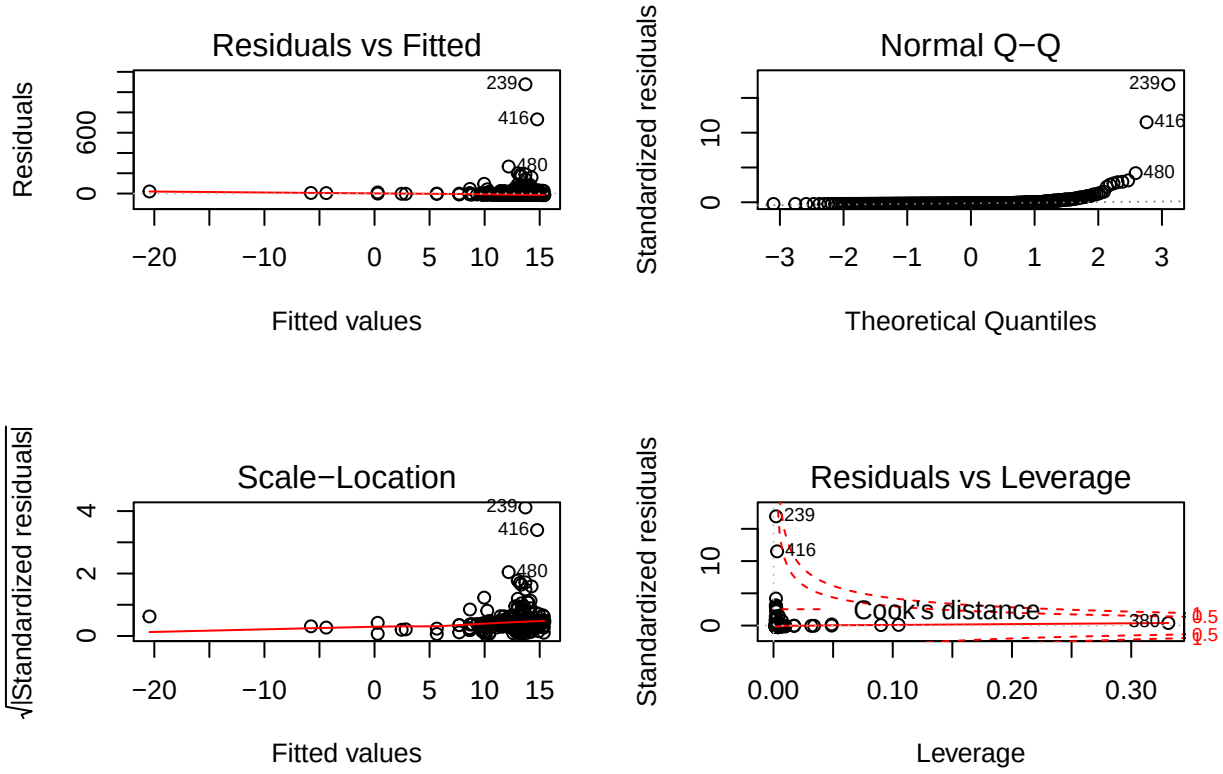
```
##
## Call:
## lm(formula = df$area ~ df$FFMC)
##
## Coefficients:
## (Intercept)      df$FFMC
##    -29.0914       0.4627
```

```
par(mfrow=c(2,2))
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
```

```
plot(lm_ffmc)
```



Conlusion: **variable FFMC is relatively statistically significant because its p-value is 0.3626, therefore making "FFMC"a relatvely good predictor.**, and the R^2 gives 0.000.

6. Linear model on Varialble "DMC"

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

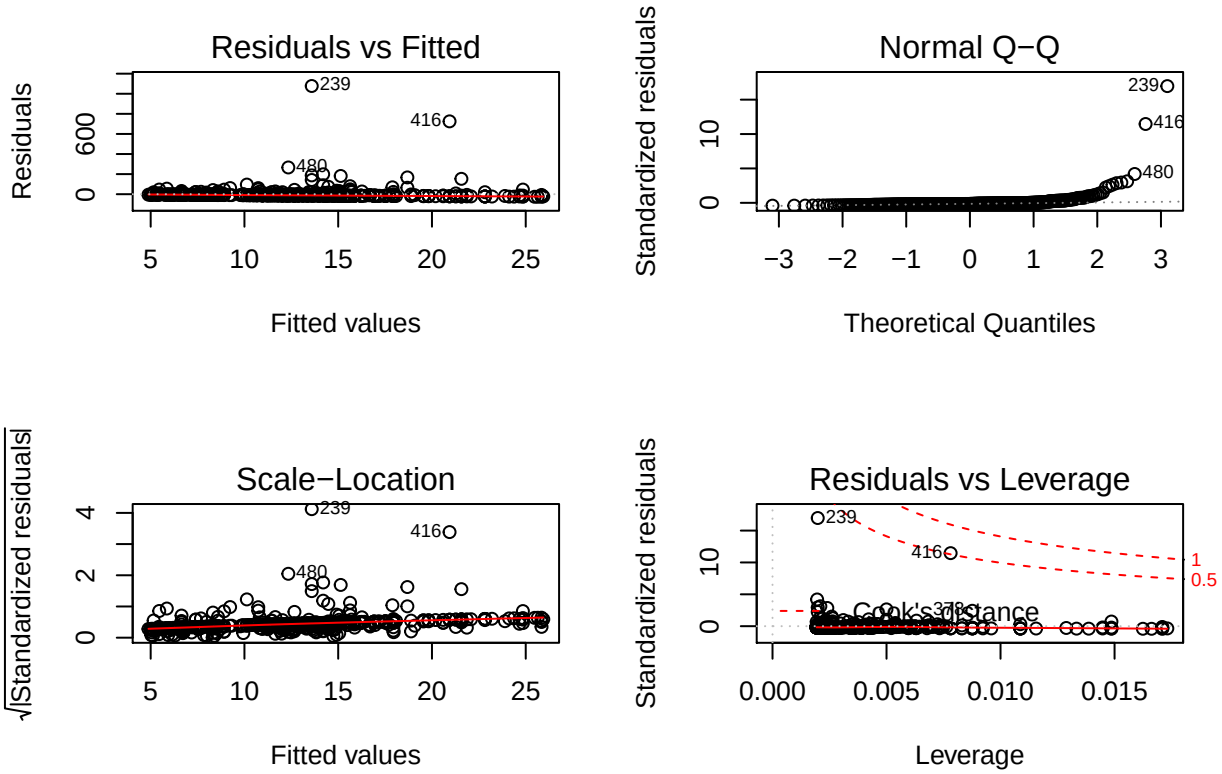
```
##
## Call:
## lm(formula = df$area ~ df$DMC)
##
## Coefficients:
## (Intercept)       df$DMC
##     4.80361      0.07255
```

```
par(mfrow=c(2,2))
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
```

```
plot(lm_dmc)
```



Conlusion: variable DMC is extreme statistically significant because its p-value is 0.09734 which is nearly 0, therefore making "DMC"a very good predictor., and the R^2 gives 0.0005.

7. Linear model on Variable "DC"

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

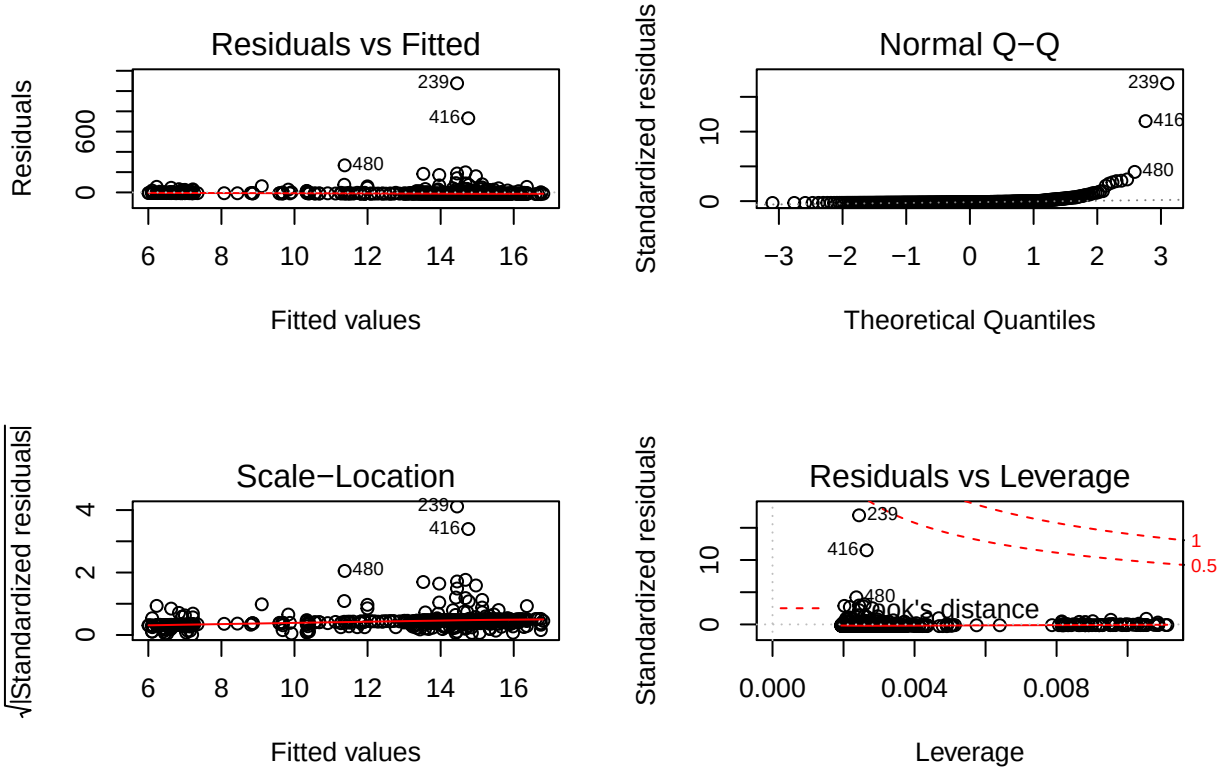
```
##
```

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

```
par(mfrow=c(2,2))
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
```

```
plot(lm_dc)
```



Conlusion: **variable DC is statistically significant because its p-value is 0.2624 which is close to 0, therefore making "DC"a very good predictor.**, and the R^2 gives 0.00024.

8. Linear model on variable "ISI"

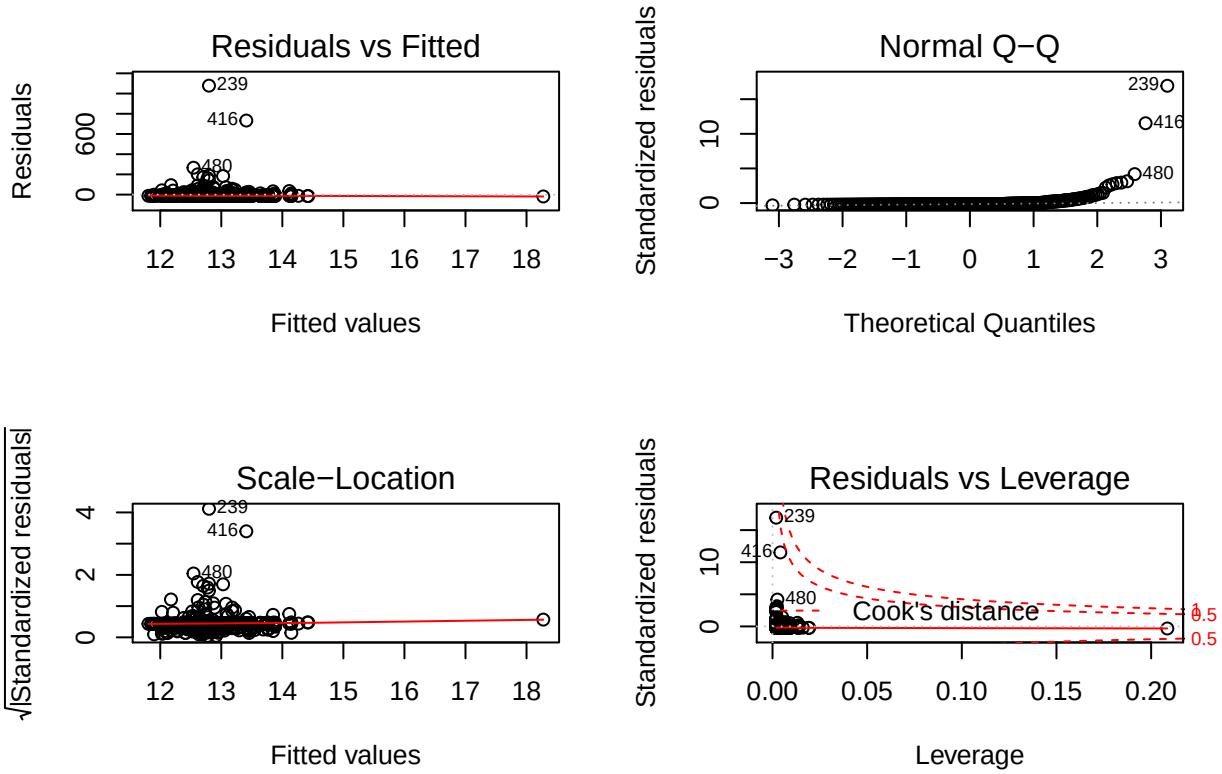
```
lm_isi <- lm(area~ISI, data = df)
lm_isi = lm(df$area~df$ISI)
lm_isi
```

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

```
par(mfrow=c(2,2))
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
```

```
plot(lm_isi)
```



Conlusion: **Variable ISI is statistically insignificant because its p-value is 0.8514 which is close to 1, therefore making "DC"a very good predictor.**, and the R^2 gives 0.00024.

9. Linear model on variable "Temperature"

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

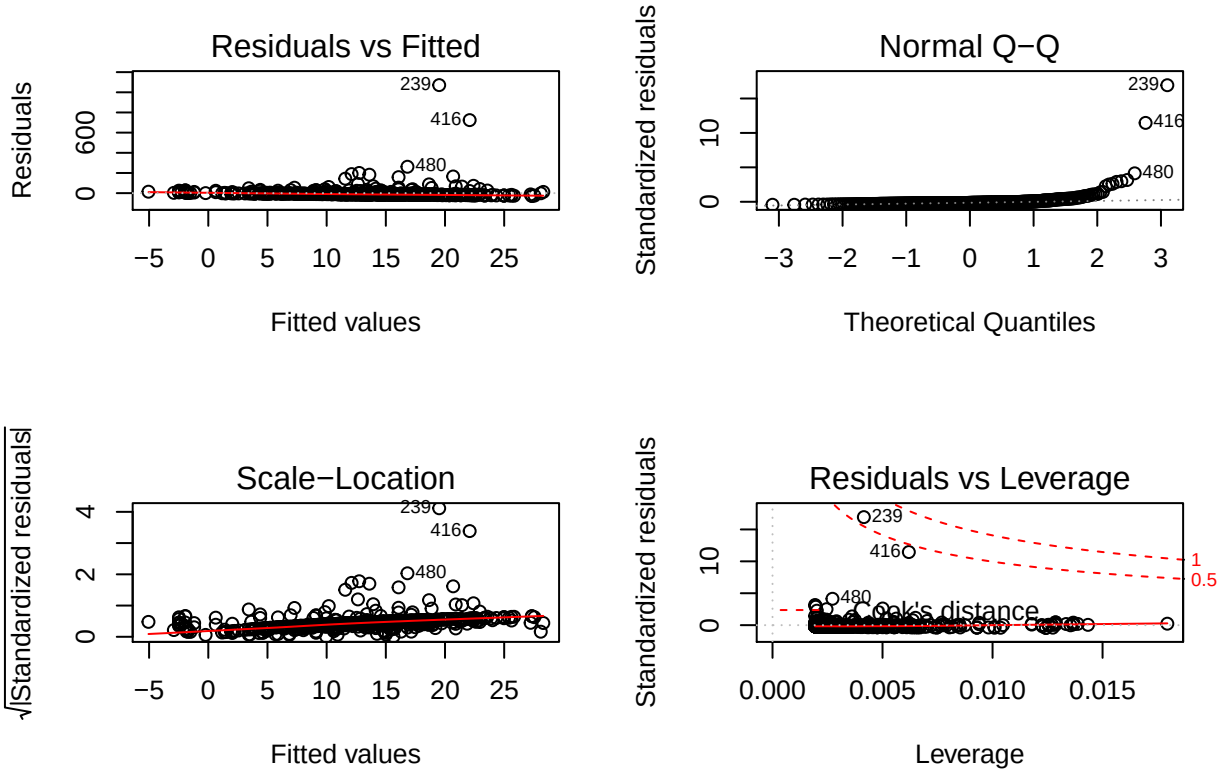
```
##
## Call:
## lm(formula = df$area ~ df$temp)
##
## Coefficients:
## (Intercept)      df$temp
##      -7.414        1.073
```

```
par(mfrow=c(2,2))
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
```

```
plot(lm_temp)
```



Conlusion: **Variable Tmperature is statistically very significant because its p-value is 0.0261(less than alpha level of 0.05) and is close to 0, therefore making "tep"a very good predictor.**, and the R^2 gives 0.00009573.

10. Naive model on variable "RH"

```
lm_RH <- lm(area~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
```

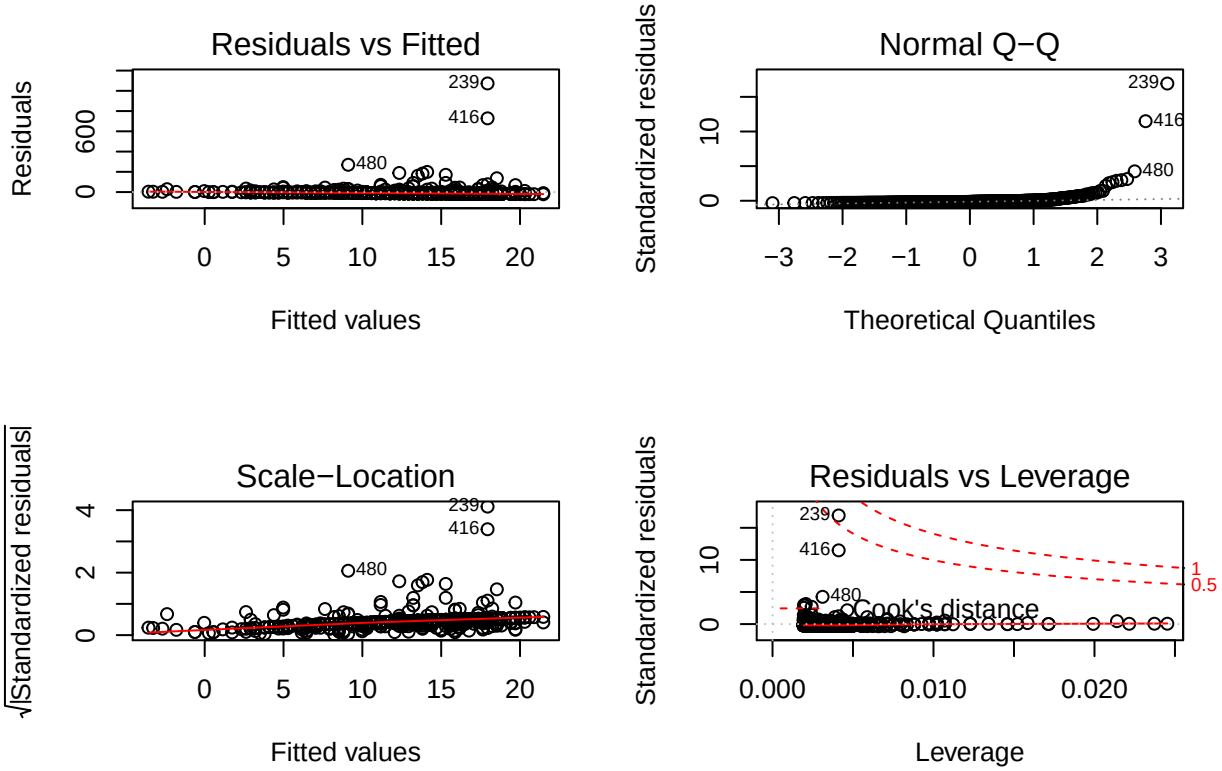
```r
par(mfrow=c(2,2))
tidy(lm_RH)
```

```
##          term   estimate std.error statistic     p.value
## 1 (Intercept) 25.8947848 8.0894377  3.201061 0.001453732
## 2       df$RH -0.2946043 0.1714114 -1.718697 0.086270552
```

```r
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
```

```r
plot(lm_RH)
```



Conlusion: Variable RH(Relative Humidty) is statistically very significant because its p-value

is 0.0145(less than alpha level of 0.05) and is close to 0, therefore making "RH"a very good predictor., and the R^2 gives 0.005703.

11. Linear model on variable "wind"

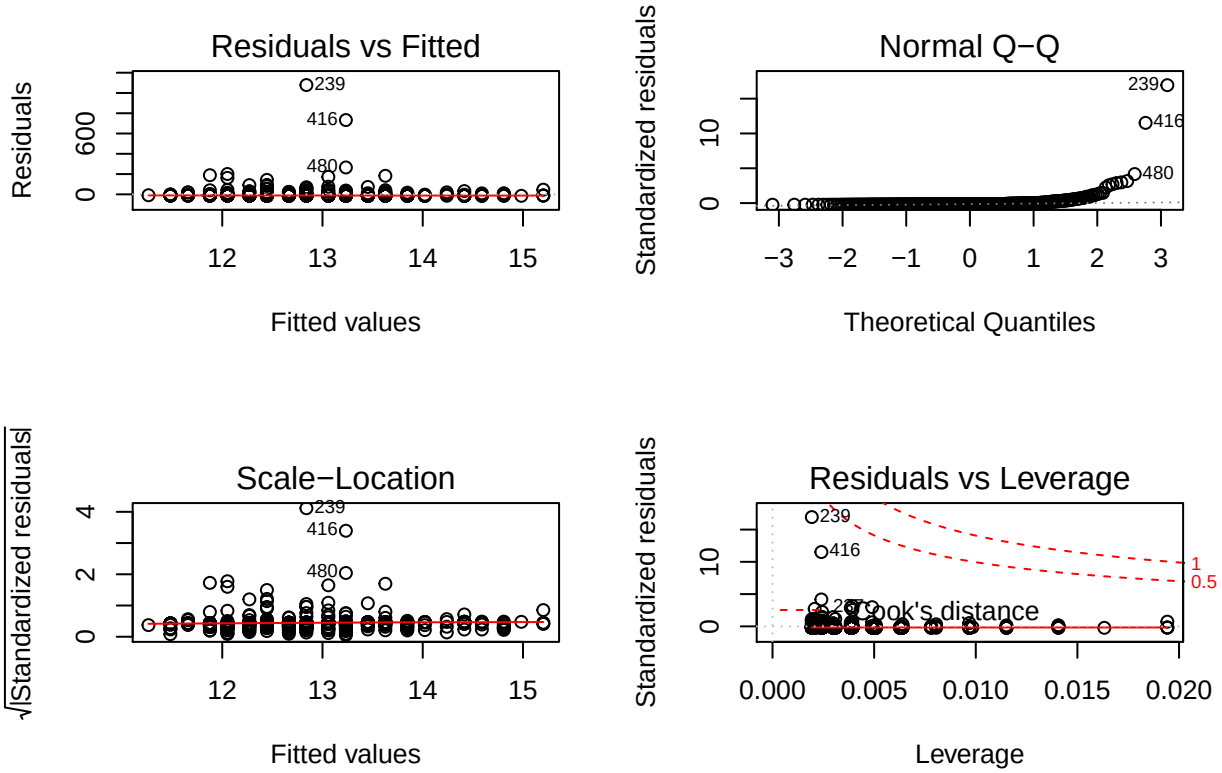
```
lm_wind <- lm(area~wind, data = df)
lm_wind = lm(df$area ~ df$wind)
lm_wind
```

```
##
## Call:
## lm(formula = df$area ~ df$wind)
##
## Coefficients:
## (Intercept)      df$wind
##     11.0891       0.4376
```

```
par(mfrow=c(2,2))
tidy(lm_wind)
```

```
##          term   estimate std.error statistic   p.value
## 1 (Intercept) 11.0891018  6.885434 1.6105159 0.1078980
## 2     df$wind  0.4376219  1.565481 0.2795447 0.7799391
```

```
plot(lm_wind)
```



Conlusion: **Variable Tmperature is statistically very significant because its p-value is 0.0261(less than alpha level of 0.05) and is close to 0, therefore making "tep"a very good predictor.**, and the R^2 gives 0.00009573.

12. Linear model on variable "rain"

```
lm_rain <- lm(area~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
```

```
tidy(lm_rain)
```

```
##          term  estimate std.error  statistic      p.value
## 1 (Intercept) 12.881612  2.809732  4.5846412 5.714847e-06
## 2     df$rain -1.584244  9.477439 -0.1671595 8.673101e-01
```

```
par(mfrow=c(2,2))
plot(lm_rain)
```