

STAT 652 Assignment 1

Dhruv Patel

13/10/2021

Lecture 5 Application

We will now do variable selection with these five variables

1. Use all-subsets regression.

```
rm(list=ls(all=TRUE))
data = na.omit(airquality[,1:4])

data$TWcp = data$Temp*data$Wind
data$TWrat = data$Temp/data$Wind

filter_data = model.matrix(Ozone ~.,data= data)
head(filter_data)
```

```
##      (Intercept) Solar.R Wind Temp  TWcp    TWrat
## 1              1     190  7.4   67 495.8 9.054054
## 2              1     118  8.0   72 576.0 9.000000
## 3              1     149 12.6   74 932.4 5.873016
## 4              1     313 11.5   62 713.0 5.391304
## 7              1     299  8.6   65 559.0 7.558140
## 8              1      99 13.8   59 814.2 4.275362
```

```
library(leaps)
allsub <- regsubsets(x=filter_data,
                    y=data$Ozone, intercept = F)

info.subsets = summary(allsub)
seq.subsets = info.subsets$which
vars.seq.subsets.raw = apply(seq.subsets, 1, function(W){
  vars.list = names(W)[W]
  output = paste0(vars.list, collapse = ", ")})
```

- (a) Report the variables in the best model of each size.

```
print(vars.seq.subsets.raw)
```

```
##              1
##              "TWrat"
##              2
##      "Solar.R, TWrat"
```

```
##                                     3
##                               "(Intercept), Temp, TWrat"
##                                     4
##                               "(Intercept), Solar.R, Temp, TWrat"
##                                     5
##                               "(Intercept), Solar.R, Wind, Temp, TWcp"
##                                     6
## "(Intercept), Solar.R, Wind, Temp, TWcp, TWrat"
```

- (b) Compute BIC on each of these models and report the BIC values for the models.
Answer:

```
print(info.subsets$bic)
```

```
## [1] -185.2244 -189.0768 -204.1878 -207.1195 -204.6274 -202.8590
```

- (c) Identify the best model. What variables are in it?

Answer: The best model is (Intercept),Solar.R,Temp,TWrat since its has minimum BIC value i.e. -207.1195

```
print(min(info.subsets$bic))
```

```
## [1] -207.1195
```

2. Use the hybrid stepwise algorithm that is the default in the step() function. Report the model that it chooses as “best.”

Answer: The best model according to stepwise algorithm is TWrat + Temp + Solar.R.

```
data$TWcp = data$Temp*data$Wind
data$TWrat = data$Temp/data$Wind
head(data)
```

```
##   Ozone Solar.R Wind Temp  TWcp    TWrat
## 1   41      190  7.4   67 495.8 9.054054
## 2   36      118  8.0   72 576.0 9.000000
## 3   12      149 12.6   74 932.4 5.873016
## 4   18      313 11.5   62 713.0 5.391304
## 7   23      299  8.6   65 559.0 7.558140
## 8   19       99 13.8   59 814.2 4.275362
```

```
rows = nrow(data)
```

```
initial <- lm(data=data, formula=Ozone~ 1)
final <- lm(data=data, formula=Ozone~Solar.R+Wind+Temp+TWcp+TWrat)
step <- step(object=initial, scope=list(upper=final), k = log(rows))
```

```
## Start: AIC=781.78
```

```
## Ozone ~ 1
```

```
##
```

```
##           Df Sum of Sq    RSS    AIC
```

```

## + TWrat      1      64323  57479 703.13
## + Temp       1      59434  62367 712.19
## + Wind       1      45694  76108 734.29
## + TWcp       1      24804  96998 761.21
## + Solar.R    1      14780 107022 772.13
## <none>              121802 781.78
##
## Step: AIC=703.13
## Ozone ~ TWrat
##
##           Df Sum of Sq   RSS   AIC
## + Temp     1     12916  44563 679.59
## + Solar.R   1       6542  50938 694.43
## <none>              57479 703.13
## + TWcp     1       1256  56223 705.39
## + Wind     1        332  57147 707.20
## - TWrat    1     64323 121802 781.78
##
## Step: AIC=679.59
## Ozone ~ TWrat + Temp
##
##           Df Sum of Sq   RSS   AIC
## + Solar.R   1     2964.5  41599 676.66
## <none>              44563 679.59
## + TWcp     1       434.8  44128 683.21
## + Wind     1       222.1  44341 683.74
## - Temp     1    12916.3  57479 703.13
## - TWrat    1    17804.4  62367 712.19
##
## Step: AIC=676.66
## Ozone ~ TWrat + Temp + Solar.R
##
##           Df Sum of Sq   RSS   AIC
## <none>              41599 676.66
## - Solar.R   1     2964.5  44563 679.59
## + TWcp     1       508.1  41090 680.00
## + Wind     1       248.0  41351 680.70
## - Temp     1     9339.1  50938 694.43
## - TWrat    1    18045.8  59644 711.94

```

```
summary(step)
```

```

##
## Call:
## lm(formula = Ozone ~ TWrat + Temp + Solar.R, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -56.168 -12.102  -4.424   11.403   77.471
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -93.30421    17.28283  -5.399 4.08e-07 ***
## TWrat        2.86326     0.42026   6.813 5.82e-10 ***

```

```
## Temp          1.25231    0.25551    4.901 3.41e-06 ***
## Solar.R       0.05960    0.02158    2.761 0.00678 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.72 on 107 degrees of freedom
## Multiple R-squared:  0.6585, Adjusted R-squared:  0.6489
## F-statistic: 68.77 on 3 and 107 DF,  p-value: < 2.2e-16
```

3. Use 10-fold CV to estimate the MSPE for the stepwise model selection process. That is,

- (a) Set the seed to 2928893 before running the `sample.int()` function.
 - (b) Create 10 folds
 - (c) Run `step()` on each training set
 - (d) Find the best model, and compute the prediction error on it
 - (e) Report the separate MSPEs from each fold, MSP Ev, $v = 1, \dots, 10$ and the MSPE for the full data.
- Answers:

```
set.seed(2928893)
rows = nrow(data)
V=10
folds = floor((sample.int(rows)-1)*V/rows) + 1
mat_CV_L5 = matrix(NA, nrow=V, ncol=1)

for(v in 1:V){

  initial <- lm(data=data[folds != v,], formula=Ozone~ 1)
  final <- lm(data=data[folds != v,], formula=Ozone~Solar.R+Wind+Temp+TWcp+TWrat)
  rows = nrow(data[folds != v,])
  step <- step(object=initial, scope=list(upper=final), k = log(rows))

  pred = predict(step,newdata=data[folds==v,])
  summary(pred)
  mat_CV_L5[v,1] = mean((data[folds==v,"Ozone"] - pred)^2)
}
```

```
## Start:  AIC=702.68
## Ozone ~ 1
##
##           Df Sum of Sq  RSS   AIC
## + TWrat    1     60815 53471 632.08
## + Temp     1     56213 58073 640.25
## + Wind     1     43966 70320 659.19
## + TWcp     1     25493 88793 682.29
## + Solar.R  1     12398 101888 695.90
## <none>             114286 702.68
##
## Step:  AIC=632.08
```

```

## Ozone ~ TWrat
##
##           Df Sum of Sq   RSS   AIC
## + Temp      1    11732  41738 612.15
## + Solar.R    1     5995  47476 624.90
## <none>                53471 632.08
## + TWcp       1       860  52610 635.06
## + Wind       1       430  53041 635.87
## - TWrat      1    60815 114286 702.68
##
## Step:   AIC=612.15
## Ozone ~ TWrat + Temp
##
##           Df Sum of Sq   RSS   AIC
## + Solar.R    1    2335.5 39403 611.04
## <none>                41738 612.15
## + TWcp       1     329.7 41409 615.96
## + Wind       1     146.4 41592 616.39
## - Temp       1    11732.4 53471 632.08
## - TWrat      1    16334.3 58073 640.25
##
## Step:   AIC=611.04
## Ozone ~ TWrat + Temp + Solar.R
##
##           Df Sum of Sq   RSS   AIC
## <none>                39403 611.04
## - Solar.R    1    2335.5 41738 612.15
## + TWcp       1     461.4 38942 614.47
## + Wind       1     199.8 39203 615.13
## - Temp       1     8073.2 47476 624.90
## - TWrat      1    17102.6 56506 642.14
## Start:   AIC=703.03
## Ozone ~ 1
##
##           Df Sum of Sq   RSS   AIC
## + TWrat      1     59974  47975 626.54
## + Temp       1     51641  56308 642.55
## + Wind       1     38512  69437 663.51
## + TWcp       1     20036  87913 687.10
## + Solar.R    1     12164  95785 695.68
## <none>                107949 703.03
##
## Step:   AIC=626.54
## Ozone ~ TWrat
##
##           Df Sum of Sq   RSS   AIC
## + Temp      1       8751  39224 611.00
## + Solar.R    1       5464  42511 619.05
## + TWcp       1       3069  44906 624.53
## <none>                47975 626.54
## + Wind       1        178  47797 630.77
## - TWrat      1     59974 107949 703.03
##
## Step:   AIC=611

```

```

## Ozone ~ TWrat + Temp
##
##           Df Sum of Sq  RSS    AIC
## + Solar.R  1    2844.4 36380 608.08
## <none>                        39224 611.00
## + Wind     1      40.7 39183 615.50
## + TWcp     1       4.1 39220 615.60
## - Temp     1    8750.6 47975 626.54
## - TWrat    1   17084.4 56308 642.55
##
## Step:  AIC=608.08
## Ozone ~ TWrat + Temp + Solar.R
##
##           Df Sum of Sq  RSS    AIC
## <none>                        36380 608.08
## - Solar.R  1    2844.4 39224 611.00
## + Wind     1     10.6 36369 612.66
## + TWcp     1       4.5 36375 612.67
## - Temp     1    6131.2 42511 619.05
## - TWrat    1   17501.0 53881 642.75
## Start:  AIC=697.92
## Ozone ~ 1
##
##           Df Sum of Sq  RSS    AIC
## + TWrat    1    52346 50221 631.11
## + Temp     1    51128 51439 633.51
## + Wind     1    35862 66705 659.50
## + TWcp     1    18152 84415 683.04
## + Solar.R  1    13264 89303 688.67
## <none>                        102567 697.92
##
## Step:  AIC=631.11
## Ozone ~ TWrat
##
##           Df Sum of Sq  RSS    AIC
## + Temp     1    12215 38006 607.85
## + Solar.R  1     5921 44300 623.17
## <none>                        50221 631.11
## + TWcp     1     1695 48526 632.28
## + Wind     1      155 50066 635.41
## - TWrat    1    52346 102567 697.92
##
## Step:  AIC=607.85
## Ozone ~ TWrat + Temp
##
##           Df Sum of Sq  RSS    AIC
## + Solar.R  1    2256.7 35749 606.33
## <none>                        38006 607.85
## + TWcp     1     111.4 37894 612.16
## + Wind     1      28.3 37978 612.38
## - Temp     1   12215.3 50221 631.11
## - TWrat    1   13432.6 51439 633.51
##
## Step:  AIC=606.33

```

```

## Ozone ~ TWrat + Temp + Solar.R
##
##           Df Sum of Sq  RSS    AIC
## <none>                 35749 606.33
## - Solar.R  1      2256.7 38006 607.85
## + TWcp     1       178.1 35571 610.44
## + Wind     1        50.5 35699 610.80
## - Temp     1      8550.6 44300 623.17
## - TWrat    1     13786.9 49536 634.34
## Start:  AIC=700.65
## Ozone ~ 1
##
##           Df Sum of Sq  RSS    AIC
## + TWrat    1      54886 50521 631.71
## + Temp     1      47981 57426 644.52
## + Wind     1      37187 68220 661.74
## + TWcp     1      19016 86390 685.36
## + Solar.R  1      14196 91210 690.79
## <none>                 105407 700.65
##
## Step:  AIC=631.71
## Ozone ~ TWrat
##
##           Df Sum of Sq  RSS    AIC
## + Temp     1      10331 40190 613.44
## + Solar.R  1        6711 43810 622.06
## <none>                 50521 631.71
## + TWcp     1       1602 48919 633.09
## + Wind     1         81 50440 636.15
## - TWrat    1      54886 105407 700.65
##
## Step:  AIC=613.44
## Ozone ~ TWrat + Temp
##
##           Df Sum of Sq  RSS    AIC
## + Solar.R  1      3642.8 36548 608.54
## <none>                 40190 613.44
## + TWcp     1       180.4 40010 617.59
## + Wind     1        85.2 40105 617.83
## - Temp     1     10330.5 50521 631.71
## - TWrat    1     17235.4 57426 644.52
##
## Step:  AIC=608.54
## Ozone ~ TWrat + Temp + Solar.R
##
##           Df Sum of Sq  RSS    AIC
## <none>                 36548 608.54
## + TWcp     1       242.7 36305 612.48
## + Wind     1       123.9 36424 612.81
## - Solar.R  1      3642.8 40190 613.44
## - Temp     1       7262.2 43810 622.06
## - TWrat    1     17199.5 53747 642.50
## Start:  AIC=690.95
## Ozone ~ 1

```

```

##
##           Df Sum of Sq   RSS   AIC
## + Temp      1      51807 43862 617.57
## + TWrat      1      48459 47210 624.93
## + Wind       1      35136 60533 649.79
## + TWcp       1      17692 77977 675.11
## + Solar.R    1      12912 82757 681.06
## <none>                95669 690.95
##
## Step:  AIC=617.57
## Ozone ~ Temp
##
##           Df Sum of Sq   RSS   AIC
## + TWrat      1      10632 33230 594.42
## + TWcp       1       9593 34269 597.50
## + Wind       1       7685 36176 602.91
## <none>                43862 617.57
## + Solar.R    1       1638 42224 618.37
## - Temp      1      51807 95669 690.95
##
## Step:  AIC=594.42
## Ozone ~ Temp + TWrat
##
##           Df Sum of Sq   RSS   AIC
## + Solar.R    1      2041.6 31188 592.68
## <none>                33230 594.42
## + TWcp       1       787.7 32442 596.63
## + Wind       1       400.5 32829 597.81
## - TWrat      1     10631.5 43862 617.57
## - Temp      1     13980.0 47210 624.93
##
## Step:  AIC=592.68
## Ozone ~ Temp + TWrat + Solar.R
##
##           Df Sum of Sq   RSS   AIC
## <none>                31188 592.68
## - Solar.R    1      2041.6 33230 594.42
## + TWcp       1       828.4 30360 594.60
## + Wind       1       402.1 30786 595.99
## - Temp      1     9954.5 41143 615.78
## - TWrat      1     11035.4 42224 618.37
## Start:  AIC=710.17
## Ozone ~ 1
##
##           Df Sum of Sq   RSS   AIC
## + TWrat      1       60836 55107 640.40
## + Temp      1       57604 58339 646.10
## + Wind      1       43129 72814 668.26
## + TWcp      1       23462 92481 692.17
## + Solar.R    1      13584 102359 702.32
## <none>                115943 710.17
##
## Step:  AIC=640.4
## Ozone ~ TWrat

```



```

##
##           Df Sum of Sq   RSS   AIC
## + Temp      1    13828  41279 616.11
## + Solar.R    1     6051  49056 633.37
## <none>                55107 640.40
## + TWcp       1     1218  53889 642.77
## + Wind        1      378  54729 644.31
## - TWrat       1    60836 115943 710.17
##
## Step:  AIC=616.11
## Ozone ~ TWrat + Temp
##
##           Df Sum of Sq   RSS   AIC
## + Solar.R    1    2736.2 38543 613.86
## <none>                41279 616.11
## + TWcp       1     519.9 40759 619.45
## + Wind        1     296.5 40983 619.99
## - Temp        1   13827.9 55107 640.40
## - TWrat       1   17059.7 58339 646.10
##
## Step:  AIC=613.86
## Ozone ~ TWrat + Temp + Solar.R
##
##           Df Sum of Sq   RSS   AIC
## <none>                38543 613.86
## - Solar.R     1    2736.2 41279 616.11
## + TWcp        1     502.6 38041 617.15
## + Wind         1     257.0 38286 617.79
## - Temp         1   10513.2 49056 633.37
## - TWrat        1   17085.3 55628 645.94
## Start:  AIC=710.48
## Ozone ~ 1
##
##           Df Sum of Sq   RSS   AIC
## + TWrat       1    67132  49163 628.98
## + Temp         1    55714  60581 649.87
## + Wind         1    46340  69955 664.25
## + TWcp         1    23729  92565 692.26
## + Solar.R      1    15806 100488 700.47
## <none>                116294 710.48
##
## Step:  AIC=628.98
## Ozone ~ TWrat
##
##           Df Sum of Sq   RSS   AIC
## + Temp         1    10889  38273 608.55
## + Solar.R       1     5659  43504 621.36
## <none>                49163 628.98
## + TWcp         1     1535  47628 630.42
## + Wind          1      164  48999 633.25
## - TWrat         1    67132 116294 710.48
##
## Step:  AIC=608.55
## Ozone ~ TWrat + Temp

```

```

##
##           Df Sum of Sq   RSS    AIC
## + Solar.R  1    2463.7 35810 606.50
## <none>                        38273 608.55
## + TWcp     1     354.0 37919 612.23
## + Wind     1     183.2 38090 612.67
## - Temp     1   10889.4 49163 628.98
## - TWrat    1   22307.2 60581 649.87
##
## Step:  AIC=606.5
## Ozone ~ TWrat + Temp + Solar.R
##
##           Df Sum of Sq   RSS    AIC
## <none>                        35810 606.50
## - Solar.R  1    2463.7 38273 608.55
## + TWcp     1     414.9 35395 609.94
## + Wind     1     203.8 35606 610.54
## - Temp     1    7694.2 43504 621.36
## - TWrat    1   22096.6 57906 649.96
## Start:  AIC=710.68
## Ozone ~ 1
##
##           Df Sum of Sq   RSS    AIC
## + TWrat    1    63520  53013 636.52
## + Temp     1    56525  60007 648.92
## + Wind     1    48353  68180 661.68
## + TWcp     1    29046  87487 686.62
## + Solar.R  1    13418 103115 703.05
## <none>                        116533 710.68
##
## Step:  AIC=636.52
## Ozone ~ TWrat
##
##           Df Sum of Sq   RSS    AIC
## + Temp     1    10768  42245 618.42
## + Solar.R  1     4493  48519 632.27
## <none>                        53013 636.52
## + Wind     1      936  52076 639.35
## + TWcp     1      348  52665 640.47
## - TWrat    1    63520 116533 710.68
##
## Step:  AIC=618.42
## Ozone ~ TWrat + Temp
##
##           Df Sum of Sq   RSS    AIC
## + Solar.R  1    2356.9 39888 617.29
## <none>                        42245 618.42
## + TWcp     1     837.1 41407 621.03
## + Wind     1     508.4 41736 621.82
## - Temp     1   10768.2 53013 636.52
## - TWrat    1   17762.9 60007 648.92
##
## Step:  AIC=617.29
## Ozone ~ TWrat + Temp + Solar.R

```

```

##
##           Df Sum of Sq   RSS    AIC
## <none>                39888 617.29
## - Solar.R  1      2356.9 42245 618.42
## + TWcp     1        730.7 39157 620.04
## + Wind     1        405.2 39483 620.87
## - Temp     1      8631.7 48519 632.27
## - TWrat    1     17159.3 57047 648.46
## Start:  AIC=711.46
## Ozone ~ 1
##
##           Df Sum of Sq   RSS    AIC
## + TWrat    1      61399  56049 642.09
## + Temp     1      57702  59745 648.48
## + Wind     1      45089  72359 667.63
## + TWcp     1      25377  92070 691.72
## + Solar.R  1      14633 102814 702.76
## <none>                117448 711.46
##
## Step:  AIC=642.09
## Ozone ~ TWrat
##
##           Df Sum of Sq   RSS    AIC
## + Temp     1      12664  43385 621.08
## + Solar.R  1        7059  48989 633.23
## <none>                56049 642.09
## + TWcp     1         914  55135 645.05
## + Wind     1         523  55526 645.76
## - TWrat    1      61399 117448 711.46
##
## Step:  AIC=621.08
## Ozone ~ TWrat + Temp
##
##           Df Sum of Sq   RSS    AIC
## + Solar.R  1      3464.7 39920 617.37
## <none>                43385 621.08
## + TWcp     1        517.6 42867 624.49
## + Wind     1        293.0 43092 625.01
## - Temp     1     12664.1 56049 642.09
## - TWrat    1     16360.6 59745 648.48
##
## Step:  AIC=617.37
## Ozone ~ TWrat + Temp + Solar.R
##
##           Df Sum of Sq   RSS    AIC
## <none>                39920 617.37
## + TWcp     1        635.7 39284 620.37
## - Solar.R  1      3464.7 43385 621.08
## + Wind     1        347.6 39572 621.10
## - Temp     1      9069.4 48989 633.23
## - TWrat    1     16703.5 56624 647.72
## Start:  AIC=698.44
## Ozone ~ 1
##

```

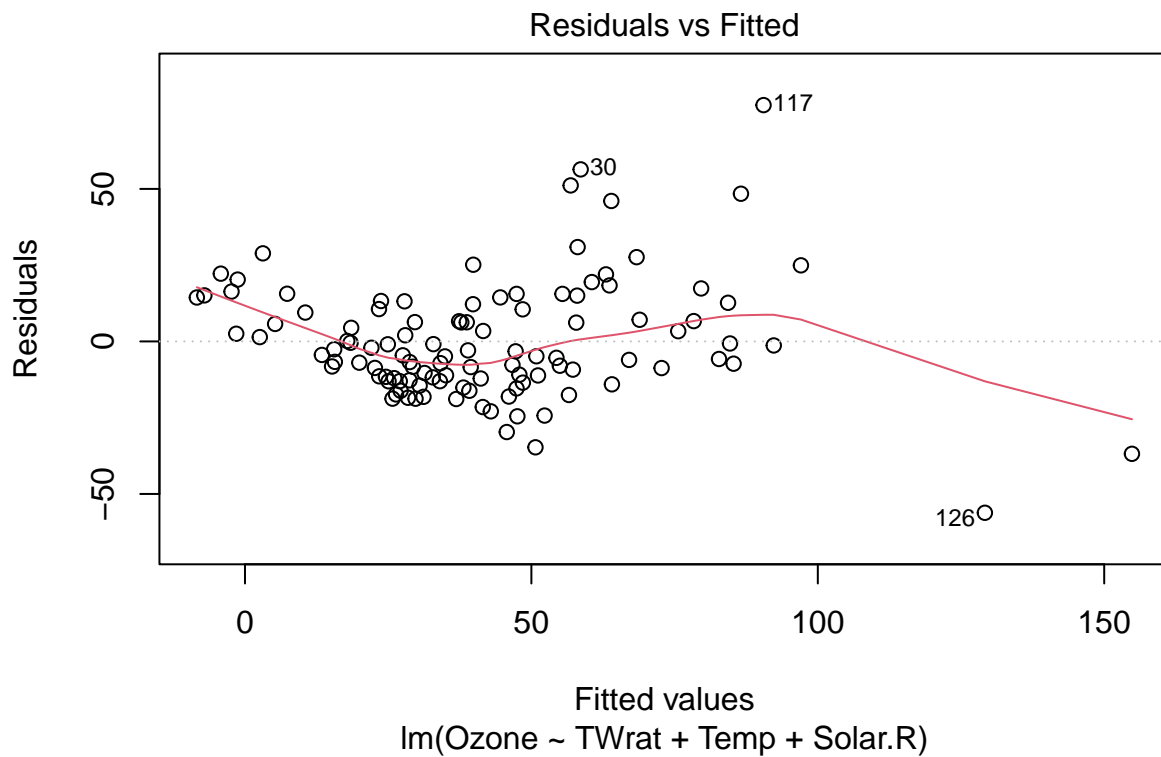
```
##           Df Sum of Sq   RSS   AIC
## + TWrat    1     52732  50375 631.42
## + Temp     1     48457  54650 639.56
## + Wind     1     38890  64217 655.70
## + TWcp     1     22477  80630 678.46
## + Solar.R  1     10598  92509 692.20
## <none>                103107 698.44
##
## Step: AIC=631.42
## Ozone ~ TWrat
##
##           Df Sum of Sq   RSS   AIC
## + Temp     1     10214  40161 613.36
## + Solar.R  1       5153  45222 625.23
## <none>                50375 631.42
## + Wind     1        556  49818 634.91
## + TWcp     1         500  49875 635.03
## - TWrat    1     52732 103107 698.44
##
## Step: AIC=613.36
## Ozone ~ TWrat + Temp
##
##           Df Sum of Sq   RSS   AIC
## + Solar.R  1     2506.3  37655 611.53
## <none>                40161 613.36
## + TWcp     1       692.5  39469 616.23
## + Wind     1       461.0  39700 616.81
## - Temp     1    10213.6  50375 631.42
## - TWrat    1    14489.1  54650 639.56
##
## Step: AIC=611.53
## Ozone ~ TWrat + Temp + Solar.R
##
##           Df Sum of Sq   RSS   AIC
## <none>                37655 611.53
## - Solar.R  1     2506.3  40161 613.36
## + TWcp     1       753.1  36902 614.11
## + Wind     1       487.5  37167 614.83
## - Temp     1     7567.3  45222 625.23
## - TWrat    1    14825.1  52480 640.12
```

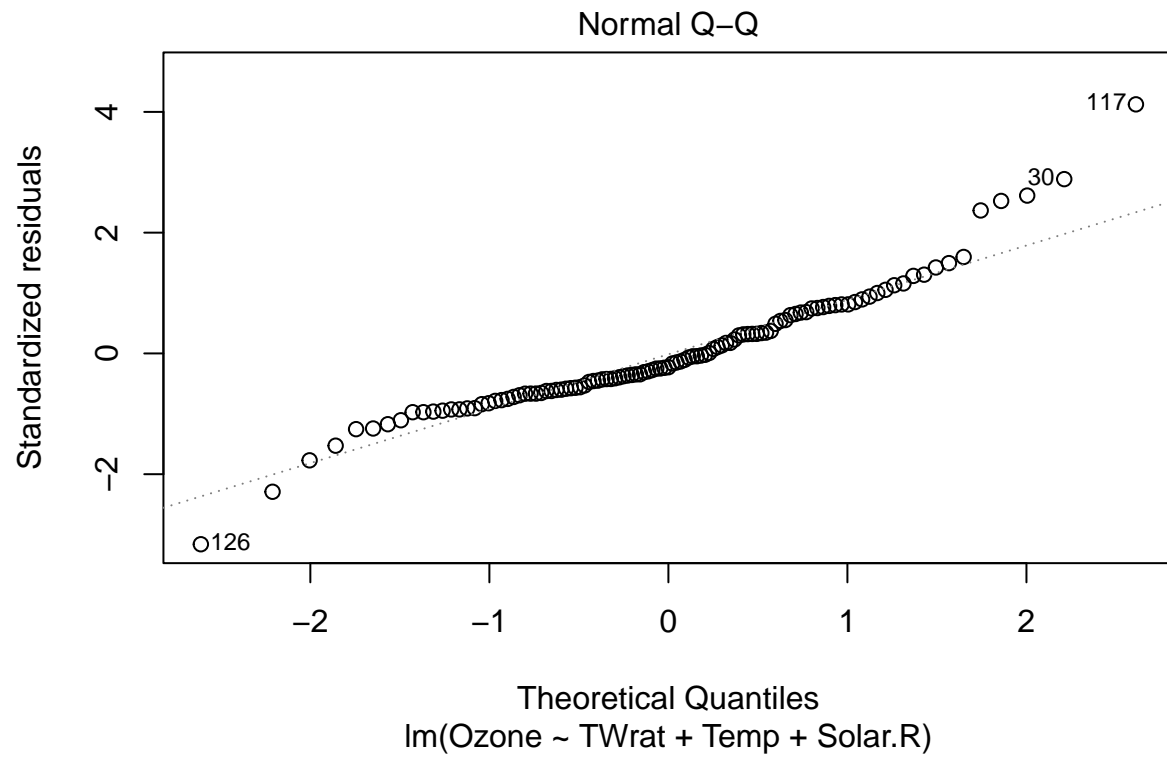
```
lm_best = lm(Ozone ~ TWrat + Temp + Solar.R, data = data)
summary(lm_best)
```

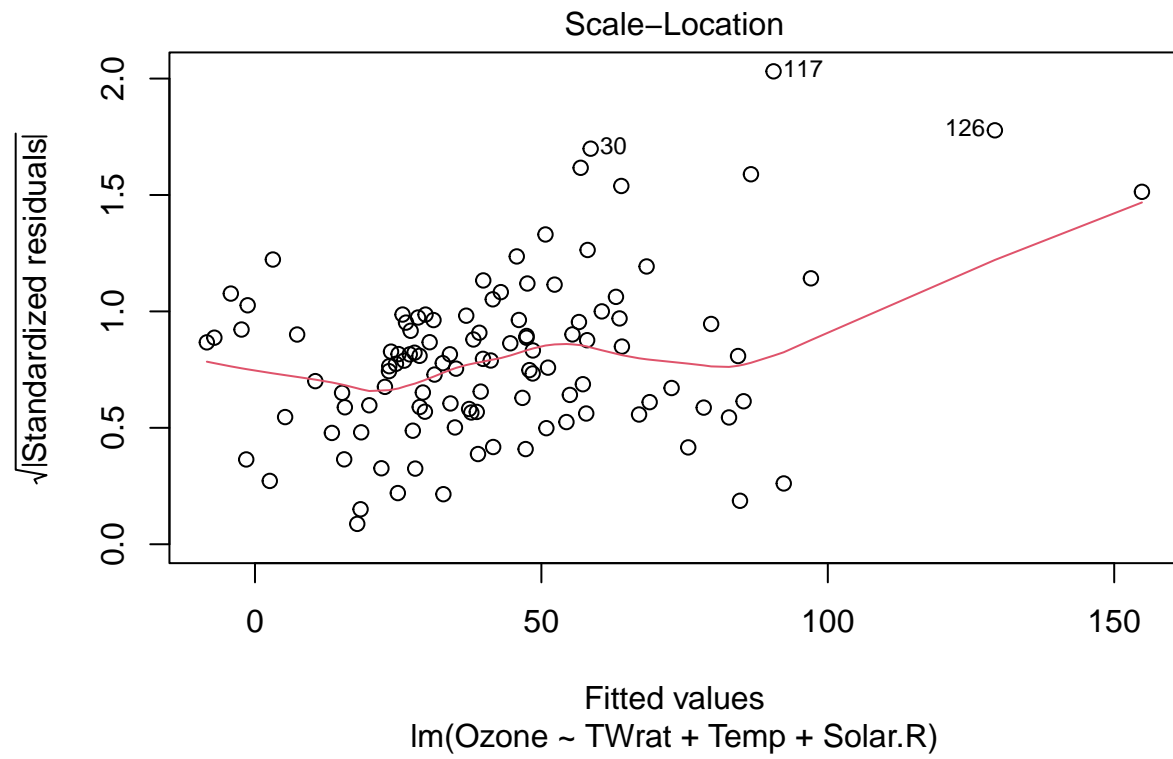
```
##
## Call:
## lm(formula = Ozone ~ TWrat + Temp + Solar.R, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -56.168 -12.102  -4.424   11.403   77.471
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

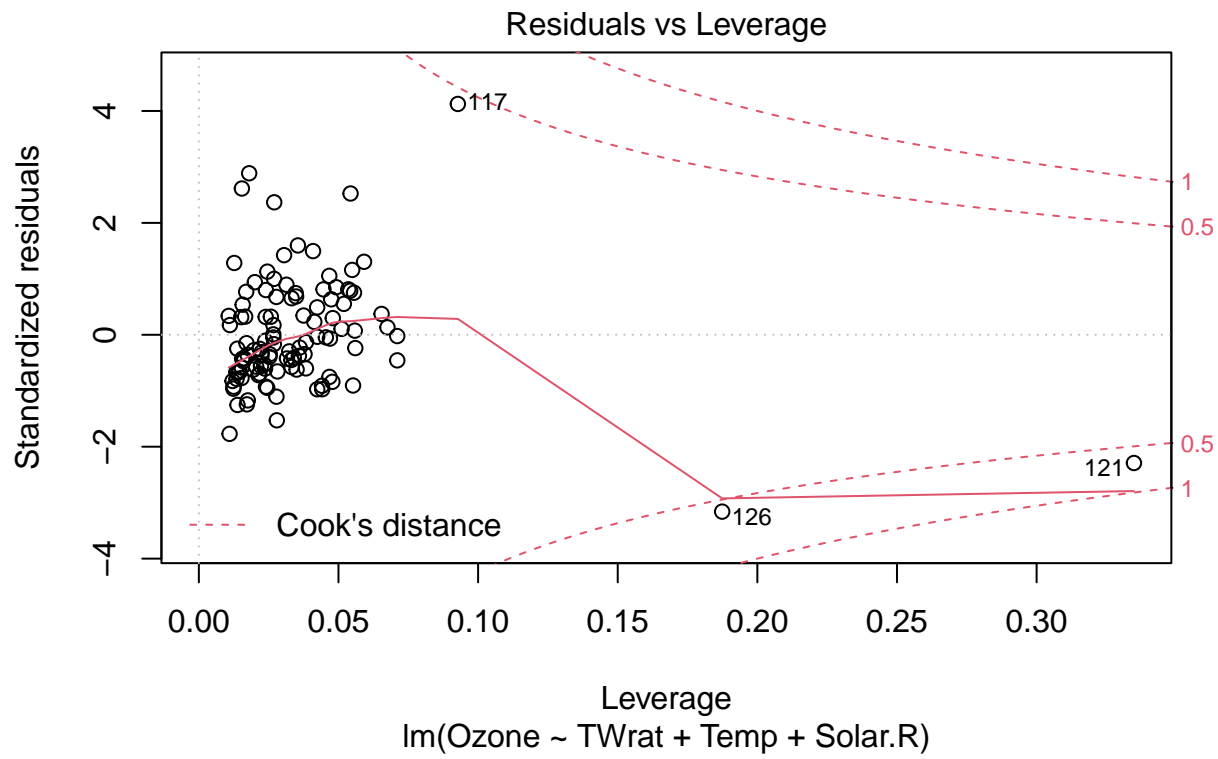
```
## (Intercept) -93.30421 17.28283 -5.399 4.08e-07 ***
## TWrat      2.86326 0.42026 6.813 5.82e-10 ***
## Temp       1.25231 0.25551 4.901 3.41e-06 ***
## Solar.R    0.05960 0.02158 2.761 0.00678 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.72 on 107 degrees of freedom
## Multiple R-squared:  0.6585, Adjusted R-squared:  0.6489
## F-statistic: 68.77 on 3 and 107 DF, p-value: < 2.2e-16
```

```
plot(lm_best)
```









```
colnames(mat_CV_L5) = c('Each fold MPSE')
mat_CV_L5
```

```
##      Each fold MPSE
## [1,]      183.4986
## [2,]      574.0699
## [3,]      558.8930
## [4,]      475.7123
## [5,]     1011.1412
## [6,]      291.4034
## [7,]      665.8734
## [8,]      157.0123
## [9,]      163.6635
## [10,]     370.1384
```

```
mean(mat_CV_L5)
```

```
## [1] 445.1406
```