STAT 652 Assignment 1

Dhruv Patel, 301471961

Lecture 5 Application

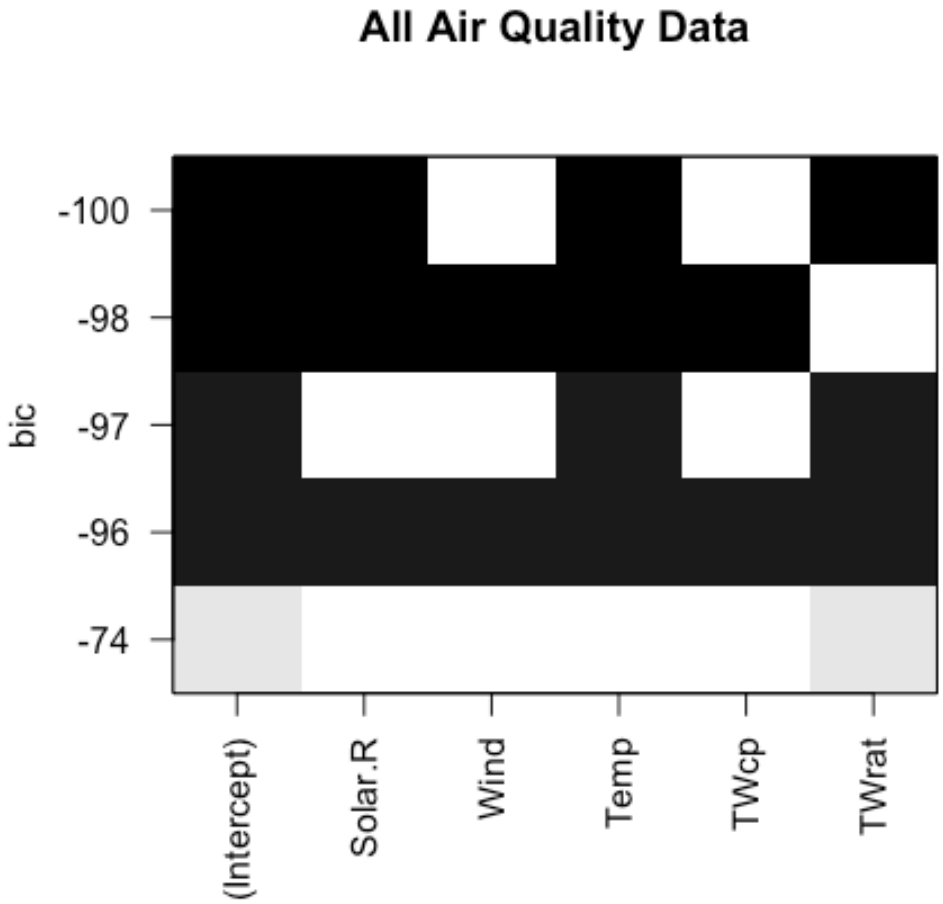
```
## Application
### 1.
#### a).
rm(list=ls(all=TRUE))
data = na.omit(airquality)
filter_data = (data[,1:4])

# Computing new columns TWcp and TWrat from Temp and Wind (Interactions)
filter_data$TWcp = filter_data$Temp*filter_data$Wind
filter_data$TWrat = filter_data$Temp/filter_data$Wind
head(filter_data)

library(leaps)
allsub <- regsubsets(x=filter_data[,2:6],
                     y=filter_data[,1], nbest=1)

summ <- summary(allsub)
sum
#b) Anwers:
#Selection Algorithm: exhaustive
#       Solar.R  Wind Temp TWcp TWrat
#1  ( 1 ) " "    " "  " "  " "  "*"
#2  ( 1 ) " "    " "  "*"  " "  "*"
#3  ( 1 ) "*"    " "  "*"  " "  "*"
#4  ( 1 ) "*"    "*"  "*"  "*"  " "
#5  ( 1 ) "*"    "*"  "*"  "*"  "*"


names(summ)
summ$bic # "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"

bic_table <- data.frame(summ$bic)
bic_table # -73.93871  -97.48091 -100.41253  -97.92049  -96.15211

x11(h=15, w=10, pointsize=12)
par(mfrow=c(1,1))
plot(allsub, main="All Air Quality Data")
#c) According to BIC values Model with only Temp gives good performance.
```

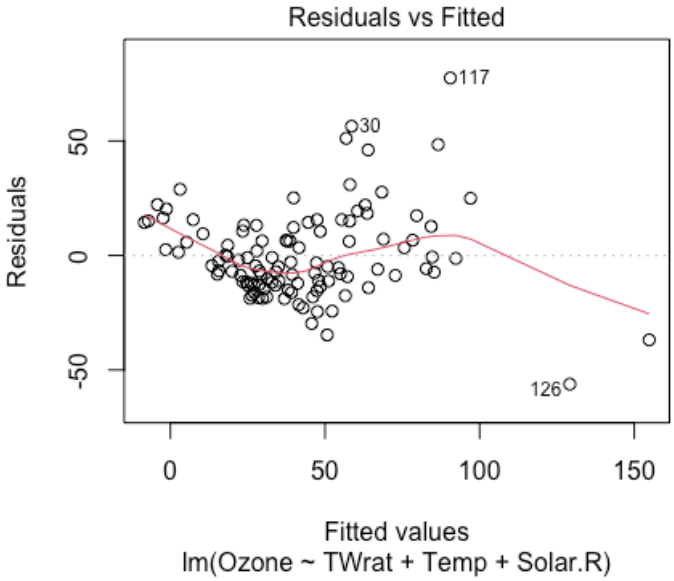
```
#2. Hybrid stepwise algorithm
data$TWcp = data$Temp*data$Wind
data$TWrat = data$Temp/data$Wind
head(data)
rows = nrow(data)

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))
summary(step)

#Answer:
# According to StepWise algorithm below model performs the best:

#  lm(formula = Ozone ~ TWrat + Temp + Solar.R, data = data)

#Coefficients:
#  (Intercept)       TWrat        Temp     Solar.R
#   -93.3042      2.8633      1.2523      0.0596


#3. 10-fold CV to estimate the MSPE for the stepwise model selection process
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)
}


# Best model lm(Ozone ~ TWrat + Temp + Solar.R) and its summary
```

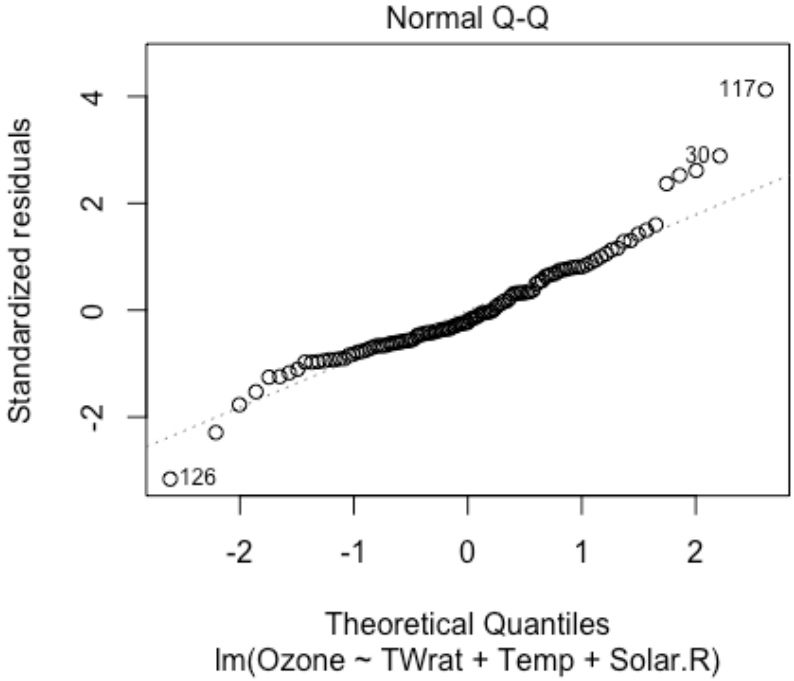
```
# TWrat + Temp + Solar.R
lm_best = lm(Ozone ~ TWrat + Temp + Solar.R, data = filter_data)
summary(lm_best)
plot(lm_best)

# Summary
#Call:
#lm(formula = Ozone ~ TWrat + Temp + Solar.R, data = filter_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




# MPSE for each fold
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
```

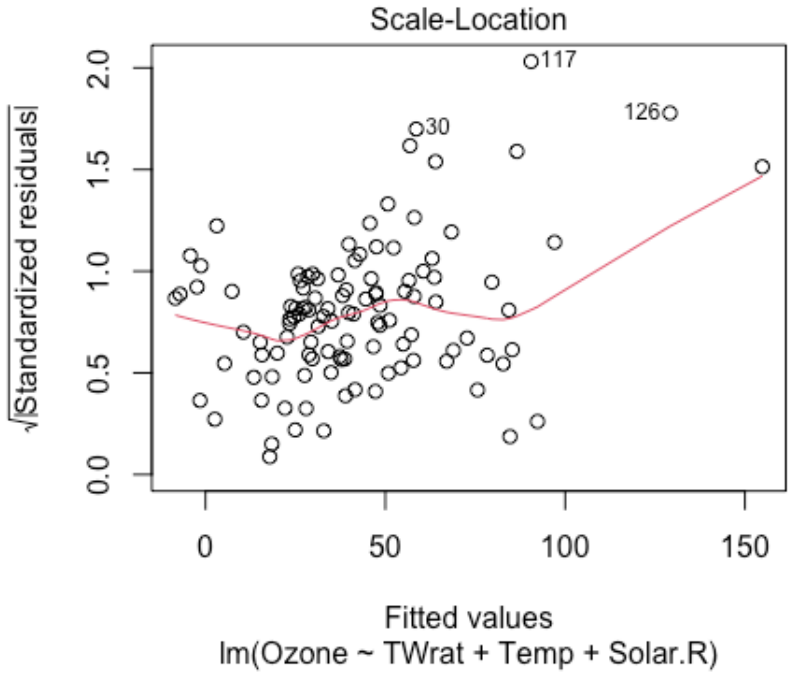
#MPSE for full-data
mean(mat_CV_L5) #445.1406

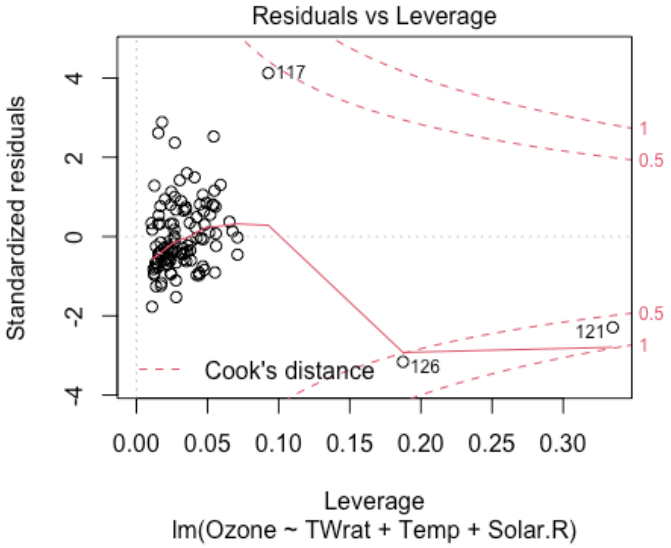
Summary plots




Residuals vs Fitted
117
30
126
Residuals
50
0
-50
0
50
100
150
Fitted values
lm(Ozone ~ TWrat + Temp + Solar.R)


Normal Q-Q
117
30
126
Standardized residuals
4
2
0
-2
-2
-1
0
1
2
Theoretical Quantiles
lm(Ozone ~ TWrat + Temp + Solar.R)


Scale-Location
√|Standardized residuals|
117
126
30
0.0
0.5
1.0
1.5
2.0
0
50
100
150
Fitted values
lm(Ozone ~ TWrat + Temp + Solar.R)


Residuals vs Leverage
Standardized residuals
117
121
126
Cook's distance
1
0.5
0.5
1
-4
-2
0
2
4
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Leverage
lm(Ozone ~ TWrat + Temp + Solar.R)