STAT 46700/CS 5900 Topics in Data Science Spring 2025

Lab 5

[Vaishak Balachandra]

Q) Can you accurately predict insurance costs? A dataset attached with this assignment contains the information on following variables:

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index,
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

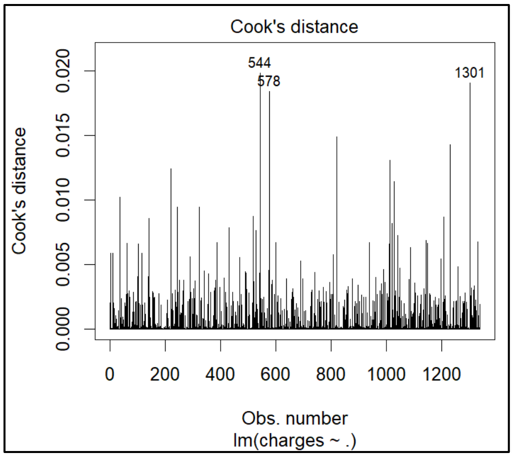
a) Import the data in R and determine its dimension.
b) Fit a multiple linear regression model using charges as a response variable.
c) Calculate the value of R^2, Adj. R^2, AIC, BIC, PRESS statistics
d) Use Stepwise procedure to identify the significant variables.
e) Perform the analysis to determine the influential cases. You may simply draw the Cook's distance from olsrr package.

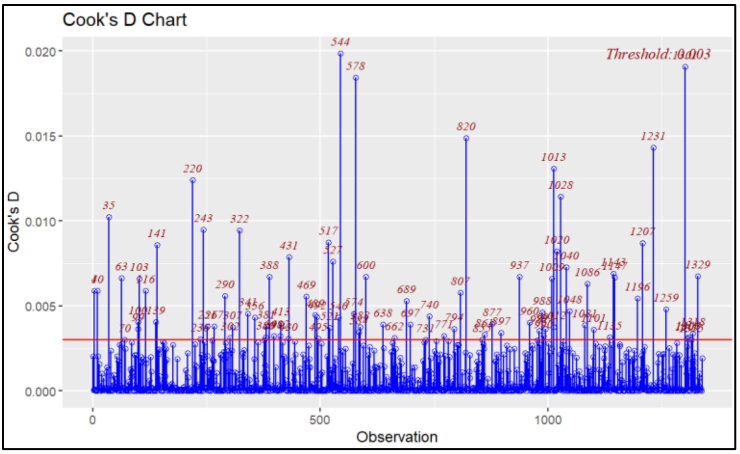
```
> #### Lab
>
>
> # a
> data <- read.csv("C:/Users/PNW_checkout/Downloads/sem2/0. Coursework/Data science/Lab/Lab 5/insur
ance.csv")
> head(data)
  age    sex    bmi children smoker    region   charges
1  19 female 27.900        0    yes southwest 16884.924
2  18   male 33.770        1     no southeast  1725.552
3  28   male 33.000        3     no southeast  4449.462
4  33   male 22.705        0     no northwest 21984.471
5  32   male 28.880        0     no northwest  3866.855
6  31 female 25.740        0     no southeast  3756.622
> dim(data)
[1] 1338    7
> cat("There are 7 columns and 1338 rows in the given dataset.")
There are 7 columns and 1338 rows in the given dataset.
>
>
>
> # b
> names(data)
[1] "age"      "sex"      "bmi"      "children" "smoker"   "region"   "charges"
> attach(data)
> # if we are mentioning the data=data, then attach is optional
> model <- lm(charges~., data=data)
> model

Call:
lm(formula = charges ~ ., data = data)

Coefficients:
   (Intercept)              age          sexmale              bmi         children        smokeryes
      -11938.5            256.9           -131.3            339.2            475.5          23848.5
regionnorthwest  regionsoutheast  regionsouthwest
         -353.0          -1035.0           -960.1
```

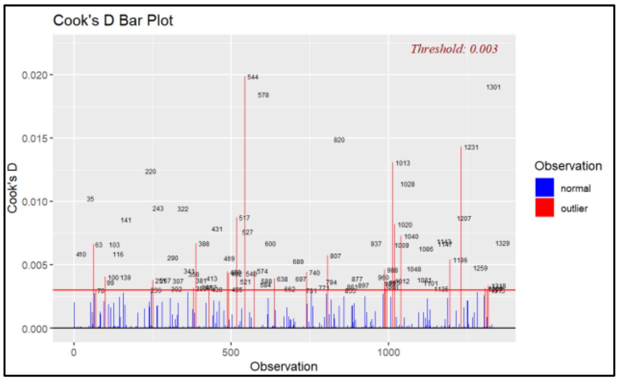
```
> cat("Fitted Model Equation is:
+ charges = -11938.5+(256.9*age)-(131.3*sexmale)+(339.2*bmi)+(475.5* children)+(23848.5*smokeryes)-
(353.0*regionnorthwest)-(1035.0*regionsoutheast)-(960.1*regionsouthwest)")
Fitted Model Equation is:
charges = -11938.5+(256.9*age)-(131.3*sexmale)+(339.2*bmi)+(475.5* children)+(23848.5*smokeryes)-(3
53.0*regionnorthwest)-(1035.0*regionsoutheast)-(960.1*regionsouthwest)
>
>
>
> # c
> summary(model)

Call:
lm(formula = charges ~ ., data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-11304.9  -2848.1   -982.1   1393.9  29992.8 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348    
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769    
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *  
regionsouthwest   -960.0      477.9  -2.009 0.044765 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,	Adjusted R-squared:  0.7494 
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16

> cat("Here, sexmale and regionnorthwest found to be non-significant")
Here, sexmale and regionnorthwest found to be non-significant
> cat("R^2 squared value: 0.7509")
R^2 squared value: 0.7509
> cat("Adjusted R^2 squared value: 0.7494")
Adjusted R^2 squared value: 0.7494
> AIC(model)
[1] 27115.51
> cat("AIC(model): 27115.51")
AIC(model): 27115.51
> BIC(model)
[1] 27167.5
> cat("BIC(model): 27167.5")
BIC(model): 27167.5
>
>
> install.packages("MPV")
> library(MPV)
> PRESS(model)
[1] 49581319689
> cat("PRESS(model): 49581319689")
PRESS(model): 49581319689
>
>
>
> # d
> install.packages("MASS")
> library(MASS)
> stepAIC(model)
Start:  AIC=23316.43
charges ~ age + sex + bmi + children + smoker + region
```

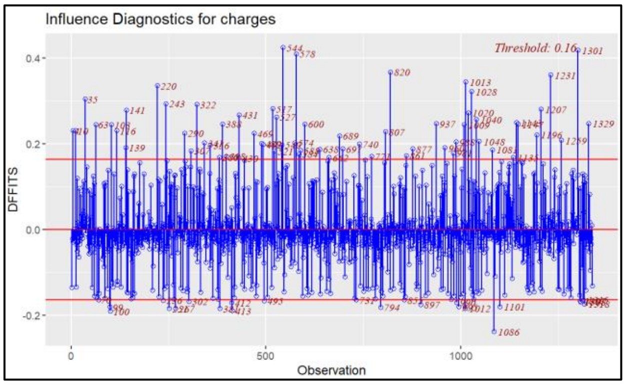
```
           Df  Sum of Sq        RSS   AIC
- sex       1 5.7164e+06 4.8845e+10 23315
<none>                   4.8840e+10 23316
- region    3 2.3343e+08 4.9073e+10 23317
- children  1 4.3755e+08 4.9277e+10 23326
- bmi       1 5.1692e+09 5.4009e+10 23449
- age       1 1.7124e+10 6.5964e+10 23717
- smoker    1 1.2245e+11 1.7129e+11 24993

Step:  AIC=23314.58
charges ~ age + bmi + children + smoker + region

           Df  Sum of Sq        RSS   AIC
<none>                   4.8845e+10 23315
- region    3 2.3320e+08 4.9078e+10 23315
- children  1 4.3596e+08 4.9281e+10 23325
- bmi       1 5.1645e+09 5.4010e+10 23447
- age       1 1.7151e+10 6.5996e+10 23715
- smoker    1 1.2301e+11 1.7186e+11 24996

Call:
lm(formula = charges ~ age + bmi + children + smoker + region, 
    data = data)

Coefficients:
    (Intercept)              age              bmi         children        smokeryes  regionnorthwest  
       -11990.3            257.0            338.7            474.6          23836.3           -352.2  
regionsoutheast  regionsouthwest  
        -1034.4           -959.4  

> cat("Significant variables identified by StepAIC model are: age, bmi, children, smoker, region")
Significant variables identified by StepAIC model are: age, bmi, children, smoker, region
> 
> 
> 
> # e
> plot(model,4)
> 
> install.packages("olsrr")
> library(olsrr)
> ols_plot_cooksd_chart(model)
> # ols_plot_cooksd_chart(model, threshold = 0.002)
> ols_plot_cooksd_bar(model)
> # ols_plot_cooksd_bar(model, threshold = 0.002)
> ols_plot_dffits(model)
> 
> 
> 
> # EXTRA
> install.packages("leaps")
> library(leaps)
> subsets = regsubsets(charges~age+sex+bmi+children+smoker+region, data=data)
> summary(subsets)
Subset selection object
Call: regsubsets.formula(charges ~ age + sex + bmi + children + smoker + 
    region, data = data)
8 Variables  (and intercept)
                Forced in Forced out
age                 FALSE      FALSE
sexmale             FALSE      FALSE
bmi                 FALSE      FALSE
children            FALSE      FALSE
smokeryes           FALSE      FALSE
regionnorthwest     FALSE      FALSE
regionsoutheast     FALSE      FALSE
regionsouthwest     FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         age sexmale bmi children smokeryes regionnorthwest regionsoutheast regionsouthwest
```

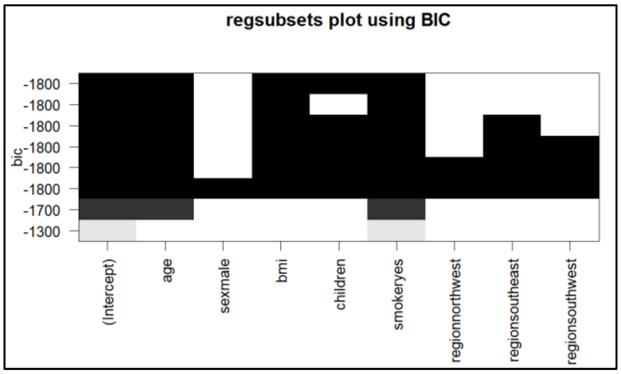
```
1  ( 1 ) " " " "     " " " "      "*"       " "             " "             " "
2  ( 1 ) "*" " "     " " " "      "*"       " "             " "             " "
3  ( 1 ) "*" " "     "*" " "      "*"       " "             " "             " "
4  ( 1 ) "*" " "     "*" "*"      "*"       " "             " "             " "
5  ( 1 ) "*" " "     "*" "*"      "*"       " "             "*"             " "
6  ( 1 ) "*" " "     "*" "*"      "*"       " "             "*"             "*"
7  ( 1 ) "*" " "     "*" "*"      "*"       "*"             "*"             "*"
8  ( 1 ) "*" "*"     "*" "*"      "*"       "*"             "*"             "*"
> plot(subsets, main = "regsubsets plot using BIC")
> plot(subsets, main = "regsubsets plot using Cp", scale = "Cp")
> plot(subsets, main = "regsubsets plot using R^2", scale = "r2")
> plot(subsets, main = "regsubsets plot using R^2 adjusted", scale = "adjr2")
> 
> 
> # outliers
> plot(model,4)
> 
> 
> 
> # OLSRR
> # install.packages("olsrr")
> # library(olsrr)
> ols_plot_dffits(model)
> ols_plot_cooksd_bar(model)
> ols_plot_dfbetas(model)
> ols_regress(charges~., data=data)

                                 Model Summary
-----------------------------------------------------------------------
R                          0.867       RMSE                   6041.680
R-Squared                  0.751       MSE                36501893.007
Adj. R-Squared             0.749       Coef. Var                45.681
Pred R-Squared             0.747       AIC                   27115.506
MAE                     4170.887       SBC                   27167.495
-----------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                       ANOVA
-----------------------------------------------------------------------------------
                        Sum of
                       Squares          DF        Mean Square       F         Sig.
-----------------------------------------------------------------------------------
Regression    147234688724.445           8    18404336090.556    500.811    0.0000
Residual       48839532843.922        1329       36749084.156
Total         196074221568.367        1337
-----------------------------------------------------------------------------------

                                           Parameter Estimates
---------------------------------------------------------------------------------------------------------
          model          Beta    Std. Error    Std. Beta       t        Sig          lower          upper
---------------------------------------------------------------------------------------------------------
    (Intercept)    -11938.539       987.819                 -12.086    0.000    -13876.393    -10000.684
            age       256.856        11.899        0.298     21.587    0.000       233.514       280.199
        sexmale      -131.314       332.945       -0.005     -0.394    0.693      -784.470       521.842
            bmi       339.193        28.599        0.171     11.860    0.000       283.088       395.298
       children       475.501       137.804        0.047      3.451    0.001       205.163       745.838
      smokeryes     23848.535       413.153        0.795     57.723    0.000     23038.031     24659.038
regionnorthwest      -352.964       476.276       -0.013     -0.741    0.459     -1287.298       581.370
regionsoutheast     -1035.022       478.692       -0.038     -2.162    0.031     -1974.097       -95.947
regionsouthwest      -960.051       477.933       -0.034     -2.009    0.045     -1897.636       -22.466
---------------------------------------------------------------------------------------------------------
> 
> # EXAMPLE DATASET:  data(swiss)
> # extractAIC(model) # for AIC
> # extractAIC(model, k = log(n)) # for BIC
```

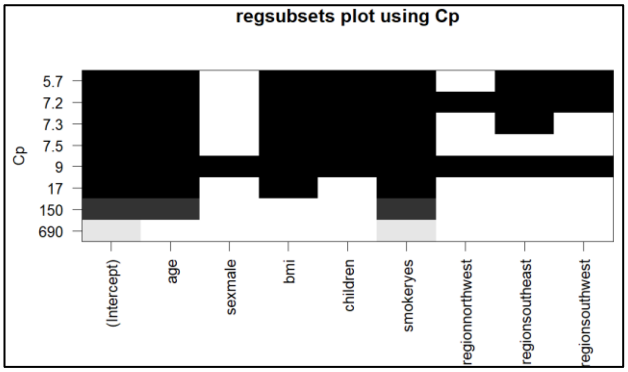

Cook's distance
Cook's distance
0.000
0.005
0.010
0.015
0.020
544
578
1301
0
200
400
600
800
1000
1200
Obs. number
lm(charges ~ .)

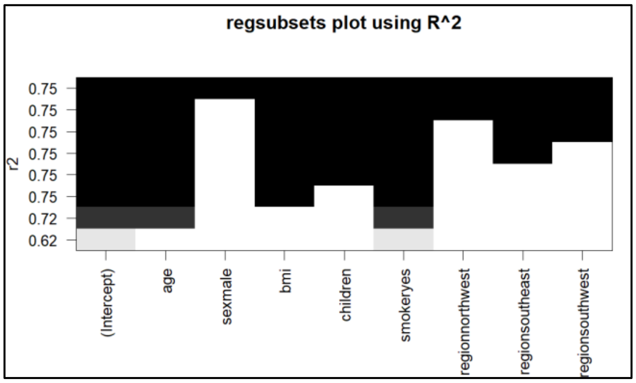

Cook's D Chart
Threshold: 0.003
Cook's D
0.000
0.005
0.010
0.015
0.020
544
578
1301
820
1231
1013
1028
220
35
243
322
141
0
500
1000
Observation

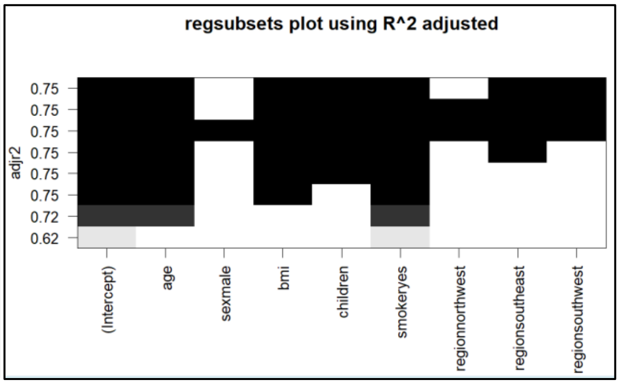

Cook's D Bar Plot
Threshold: 0.003
Cook's D
0.000
0.005
0.010
0.015
0.020
544
578
1301
820
1231
1013
1028
220
0
500
1000
Observation
Observation
normal
outlier

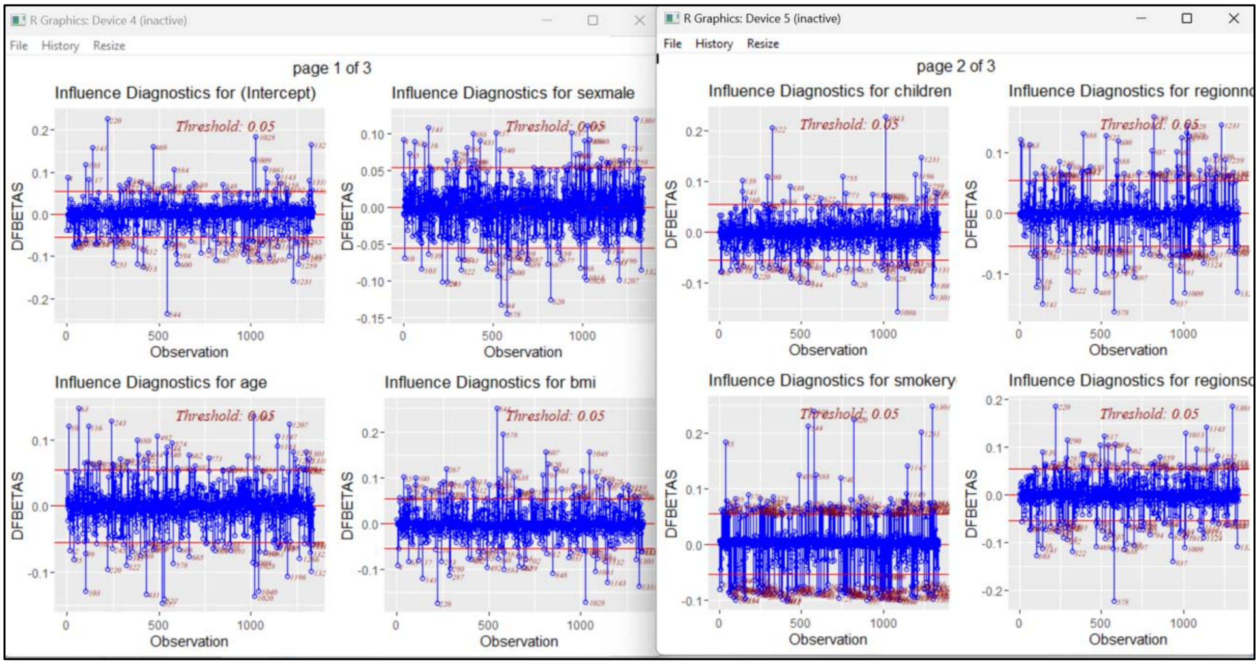

Influence Diagnostics for charges
Threshold: 0.16
DFFITS
-0.2
0.0
0.2
0.4
544
578
1301
820
1231
1013
1028
220
1086
0
500
1000
Observation


regsubsets plot using BIC
bic
-1800
-1800
-1800
-1800
-1800
-1800
-1700
-1300
(Intercept)
age
sexmale
bmi
children
smokeryes
regionnorthwest
regionsoutheast
regionsouthwest

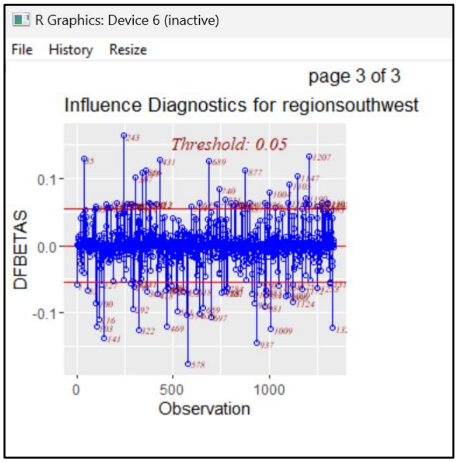

regsubsets plot using Cp
Cp
5.7
7.2
7.3
7.5
9
17
150
690
(Intercept)
age
sexmale
bmi
children
smokeryes
regionnorthwest
regionsoutheast
regionsouthwest


regsubsets plot using R^2
r2
0.75
0.75
0.75
0.75
0.75
0.75
0.72
0.62
(Intercept)
age
sexmale
bmi
children
smokeryes
regionnorthwest
regionsoutheast
regionsouthwest


regsubsets plot using R^2 adjusted
adjr2
0.75
0.75
0.75
0.75
0.75
0.75
0.72
0.62
(Intercept)
age
sexmale
bmi
children
smokeryes
regionnorthwest
regionsoutheast
regionsouthwest



R Graphics: Device 4 (inactive)
File History Resize
page 1 of 3
Influence Diagnostics for (Intercept)
Influence Diagnostics for sexmale
Influence Diagnostics for age
Influence Diagnostics for bmi
Threshold: 0.05
DFBETAS
Observation
0
500
1000
R Graphics: Device 5 (inactive)
File History Resize
page 2 of 3
Influence Diagnostics for children
Influence Diagnostics for regionno
Influence Diagnostics for smokery
Influence Diagnostics for regionsc
Threshold: 0.05
DFBETAS
Observation
0
500
1000



R Graphics: Device 6 (inactive)
File History Resize
page 3 of 3
Influence Diagnostics for regionsouthwest
Threshold: 0.05
DFBETAS
0.1
0.0
-0.1
0
500
1000
Observation