CSCI E-106:Assignment 9

Due Date: November 23, 2020 at 7:20 pm EST

Instructions

Students should submit their reports on Canvas. The report needs to clearly state what question is being solved, step-by-step walk-through solutions, and final answers clearly indicated. Please solve by hand where appropriate.

Please submit two files: (1) a R Markdown file (.Rmd extension) and (2) a PDF document, word, or html generated using knitr for the .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

Solutions:

Problem 1

Refer to Brand preference data, build a model with all independent variables (45 pts, 5 points each)

a-) Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = 0.10$. State the decision rule and conclusion.

No outliers based on the bonferoni test.

```
Brand.Preference <- read.csv("/cloud/project/Brand Preference.csv")
pr1<-lm(Y~X1+X2,data=Brand.Preference)
library(olsrr)
```

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers
```

```
drst<-rstudent(pr1)
tb<-qt(1-0.1/(2*16),16-3-1)
sum(abs(drst)>abs(tb))
```

```
## [1] 0
```

b-) Obtain the diagonal elements of the hat matrix, and provide an explanation for the pattern in these elements.

Max hat value is 0.2375 and the min is 0.1375. The average is 0.19. The compact range, no indication of outliers.

```r
hii <- hatvalues(pr1)
hii
```

```
##      1      2      3      4      5      6      7      8      9     10     11
## 0.2375 0.2375 0.2375 0.2375 0.1375 0.1375 0.1375 0.1375 0.1375 0.1375 0.1375
##     12     13     14     15     16
## 0.1375 0.2375 0.2375 0.2375 0.2375
```

```r
summary(hii)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1375  0.1375  0.1875  0.1875  0.2375  0.2375
```

c-) Are any of the observations outlying with regard to their X values according?

No outliers in direction of X, hat values are less than 2*p/n.

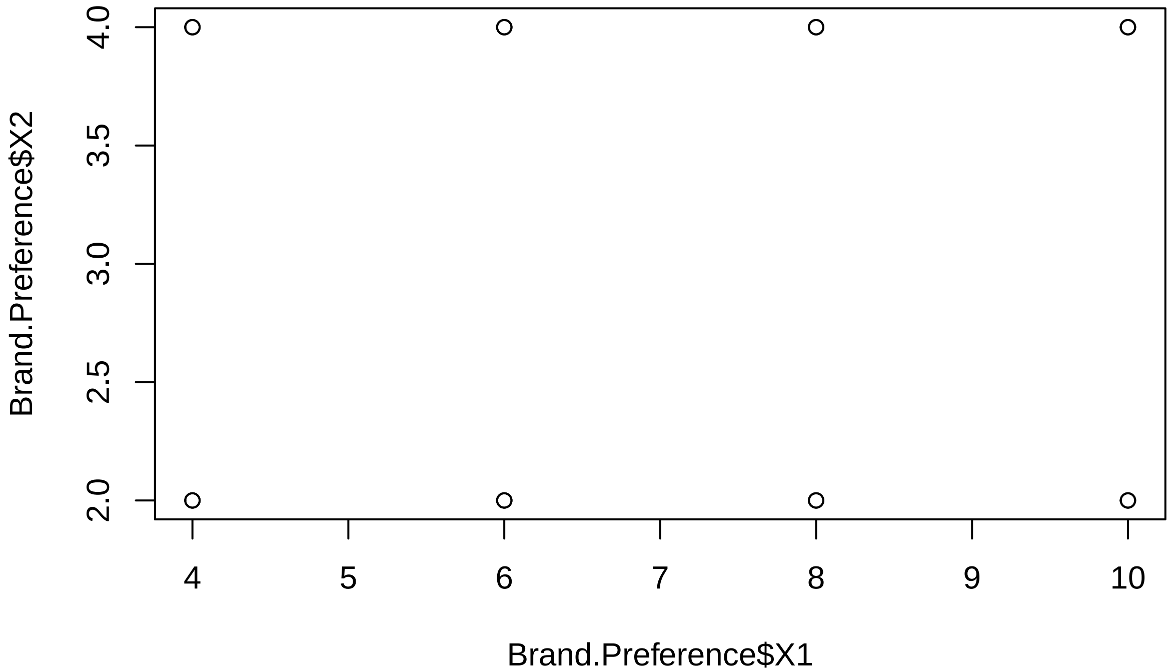
```r
sum(hii>(2*3/16))
```

```
## [1] 0
```

d-) Management wishes to estimate the mean degree of brand liking for moisture content $X_1 = 10$ and sweetness $X_2 = 3$. Construct a scatter plot of X_2 against X_1 and determine visually whether this prediction involves an extrapolation beyond the range of the data. Also, determine whether an extrapolation is involved. Do your conclusions from the two methods agree?

The hat value for the prediction is 0.175 which is within the hat values calculated pat c(max= 0.2375 and min=0.1375). No extrapolation is required.

```r
plot(Brand.Preference$X1,Brand.Preference$X2)
```




```r
X<-model.matrix(pr1)
XXInv<-solve(t(X)%*%X)
Xhnew<-matrix(c(1,10,3),nrow=1,ncol=3)
Hatnew<-Xhnew%*%XXInv%*%t(Xhnew)
Hatnew
```

```
##       [,1]
## [1,] 0.175
```

e-) The largest absolute studentized deleted residual is for case 14. Obtain the DFFITS, DFBETAS, and Cook's distance values for this case to assess the influence of this case. What do you conclude?

Case 14 has the max DFIITS, DFBETAS, and Cooks distance. Cooks distance is 2000 larger than the smallest cooks distance. Indicating influential point.

```r
cd<-influence.measures(pr1)
cd
```

```
## Influence measures of
##   lm(formula = Y ~ X1 + X2, data = Brand.Preference) :
##
##      dfb.1_  dfb.X1  dfb.X2   dffit cov.r   cook.d   hat inf
## 1  -0.02155  0.0157  0.0117 -0.0228 1.667 0.000188 0.238
## 2   0.00868 -0.0235  0.0175  0.0342 1.666 0.000422 0.237
## 3  -0.71785  0.5226  0.3895 -0.7593 1.084 0.180392 0.237
## 4   0.19619 -0.5324  0.3968  0.7735 1.068 0.186258 0.238
## 5  -0.11987  0.0442  0.0988 -0.1465 1.426 0.007666 0.138
## 6   0.02413  0.0800 -0.1790 -0.2655 1.322 0.024547 0.138
## 7  -0.25062  0.0924  0.2065 -0.3063 1.277 0.032297 0.138
## 8  -0.01832 -0.0607  0.1358  0.2015 1.384 0.014354 0.138
## 9   0.07315  0.0560 -0.1252  0.1857 1.397 0.012231 0.138
```

```
## 10  0.12431 -0.0728 -0.1627 -0.2413 1.347 0.020406 0.138
## 11  0.28674  0.2195 -0.4907  0.7279 0.708 0.149828 0.138
## 12 -0.20113  0.1177  0.2632  0.3904 1.171 0.050983 0.138
## 13  0.01467 -0.4378  0.3263 -0.6360 1.225 0.131821 0.237
## 14  0.83881 -0.8077 -0.6020 -1.1735 0.651 0.363412 0.237
## 15 -0.01917  0.5722 -0.4265  0.8314 1.002 0.210661 0.237
## 16 -0.09802  0.0944  0.0704  0.1371 1.643 0.006758 0.237
```

```r
cd$infmat[14,6]/cd$infmat[,6]
```

```
##           1           2           3           4           5           6
## 1936.000000  860.444444    2.014568    1.951121   47.408648   14.804950
##           7           8           9          10          11          12
##   11.252151   25.317340   29.712712   17.809076    2.425527    7.128081
##          13          14          15          16
##    2.756853    1.000000    1.725106   53.777778
```

f-) Calculate the average absolute percent difference in the fitted values with and without case 14. What does this measure indicate about the influence of case 14?

Predicted values are increased by %0.62.

```r
p1<-pr1$fitted.values[-c(14)]
t1<-lm(Y~X1+X2,data=Brand.Preference[-c(14),])
p2<-t1$fitted.values
cbind(Brand.Preference[-c(14),1],p1,p2)
```

```
##           p1        p2
## 1   64 64.10  63.45082
## 2   73 72.85  72.92213
## 3   61 64.10  63.45082
## 4   76 72.85  72.92213
## 5   72 72.95  72.73361
## 6   80 81.70  82.20492
## 7   71 72.95  72.73361
## 8   83 81.70  82.20492
## 9   83 81.80  82.01639
## 10  89 90.55  91.48770
## 11  86 81.80  82.01639
## 12  93 90.55  91.48770
## 13  88 90.65  91.29918
## 15  94 90.65  91.29918
## 16 100 99.40 100.77049
```

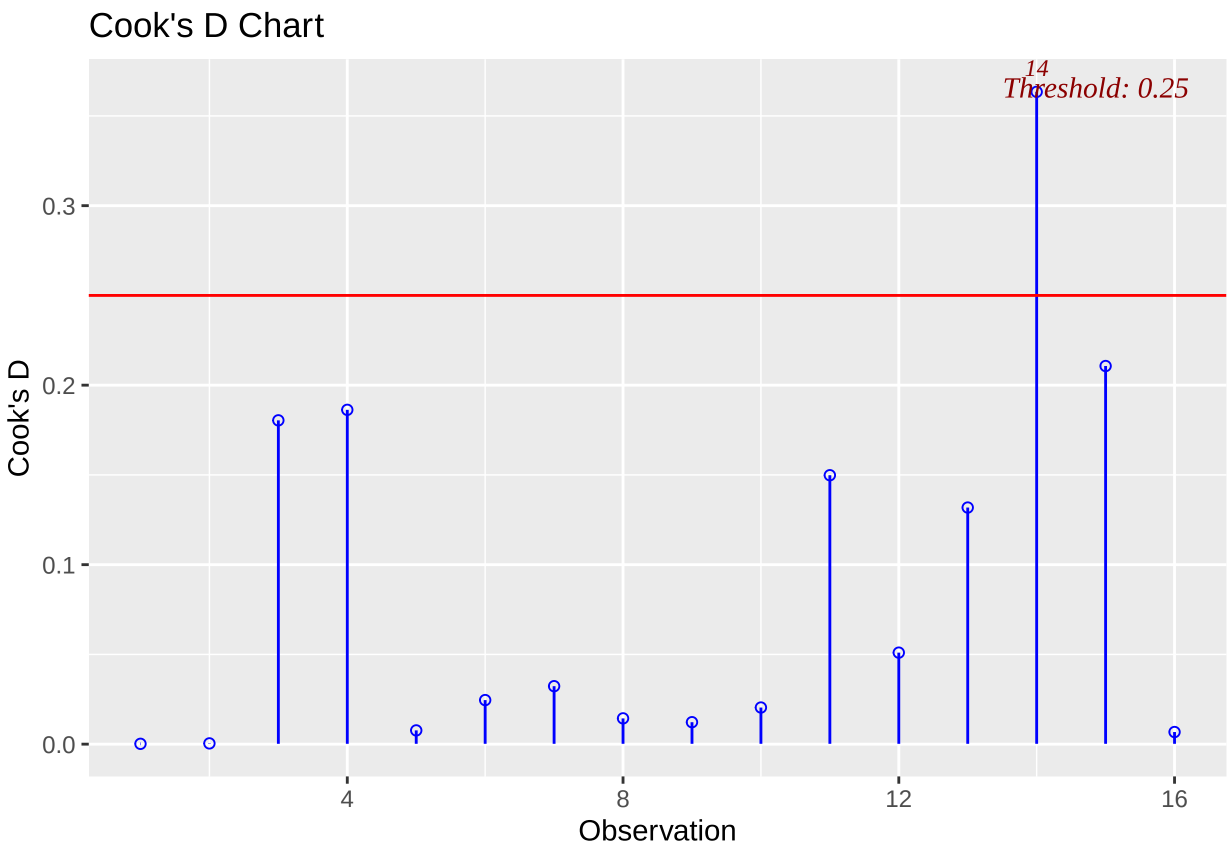
```r
mean((abs(p1-p2)/p2)*100)
```

```
## [1] 0.6284827
```

g-) Calculate Cook's distance D; for each case and prepare an index plot. Are any cases influential according to this measure?

Case 14 is an influential point based on the plot.

```r
ols_plot_cooksd_chart(pr1)
```



h-) Find the two variance inflation factors. Why are they both equal to 1?

X1 and X2 are independent, therefore VIF=1.

```r
library(faraway)
```

```
##
## Attaching package: 'faraway'

## The following object is masked from 'package:olsrr':
##
##     hsb
```

```r
vif(pr1)
```

```
## X1 X2
##  1  1
```

```r
cor(Brand.Preference)
```

```
##            Y        X1        X2
## Y  1.0000000 0.8923929 0.3945807
## X1 0.8923929 1.0000000 0.0000000
## X2 0.3945807 0.0000000 1.0000000
```

Problem 2

Refer to the Lung pressure Data. Increased arterial blood pressure in the lungs frequently leads to the development of heart failure in patients with chronic obstructive pulmonary disease (COPD). The standard method for determining arterial lung pressure is invasive, technically difficult, and involves some risk to the patient. Radionuclide imaging is a noninvasive, less risky method for estimating arterial pressure in the lungs. To investigate the predictive ability of this method, a cardiologist collected data on 19 mild-to-moderate COPD patients. The data includes the invasive measure of systolic pulmonary arterial pressure (Y) and three potential noninvasive predictor variables. Two were obtained by using radionuclide imaging emptying rate of blood into the pumping chamber or the heart (X_1) and ejection rate of blood pumped out of the heart into the lungs (X_2) and the third predictor variable measures blood gas (X_3). (35 points, 5 points each)

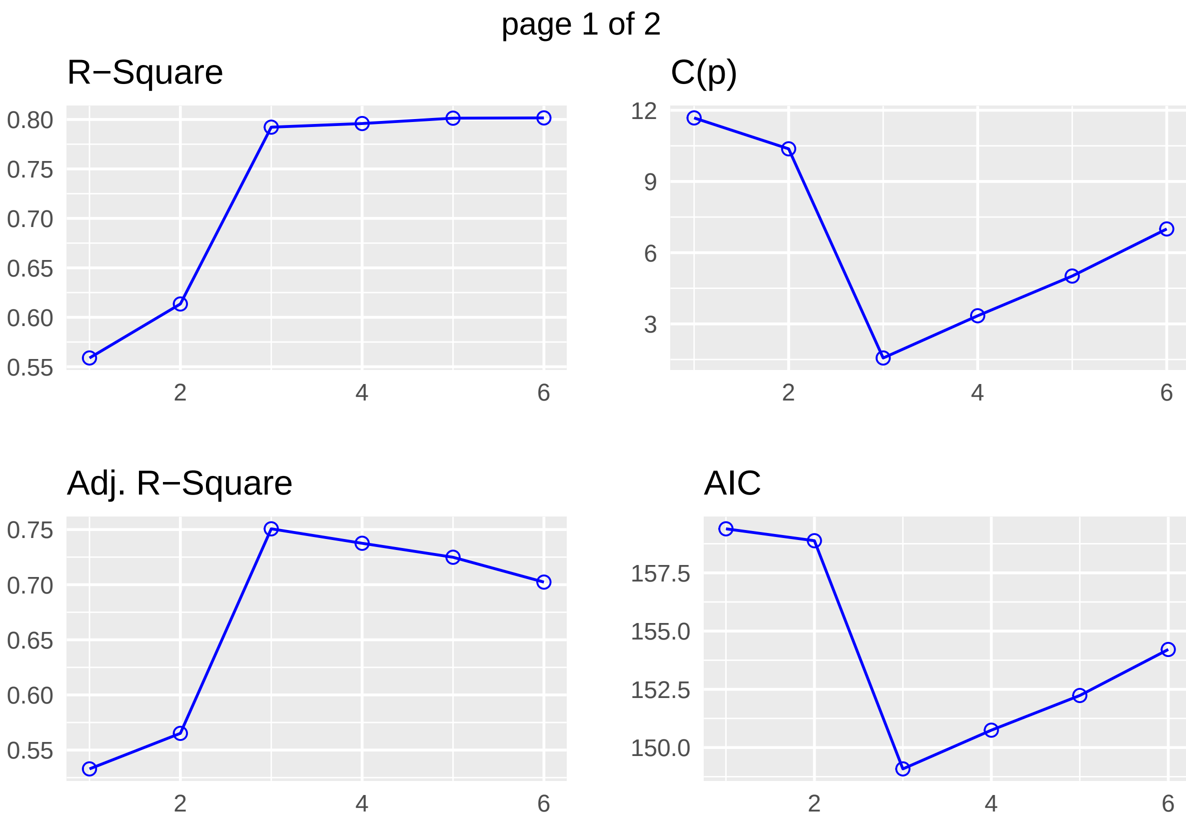
a-) Find the best regression model by using first-order terms and the cross-product term. Ensure that all variables in the model are significant at 5%.

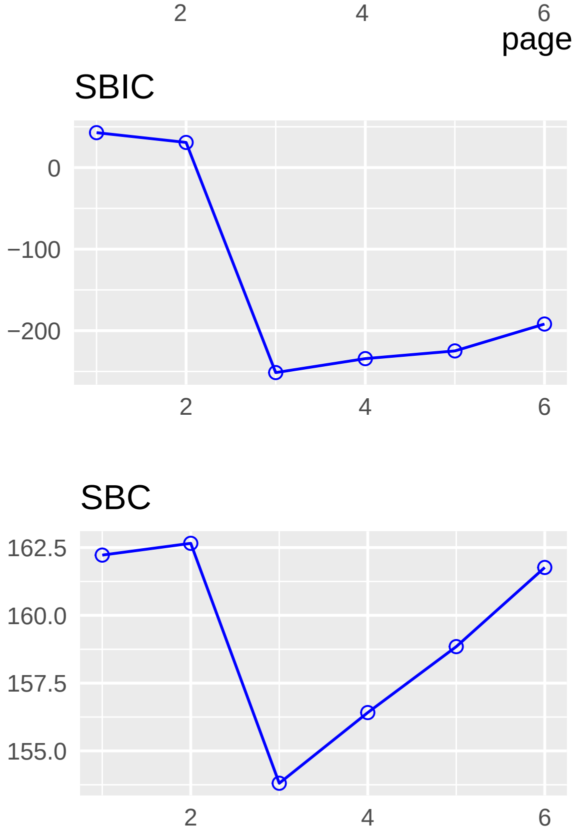
The best subet algorithm is suggesting that the third model is the best model based on Adjusted R square values, CP, SBC and AIC. The model is significant and all variables are significant at 5% level.

```r
Lung.Pressure <- read.csv("/cloud/project/Lung Pressure.csv")
pr2<-lm(Y~.^2,data=Lung.Pressure)
summary(pr2)
```

```
##
## Call:
## lm(formula = Y ~ .^2, data = Lung.Pressure)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.908  -4.817  -2.612   4.623  23.476
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.97317   78.36891   1.850  0.08910 .
## X1           -2.95130    2.76019  -1.069  0.30600
## X2           -1.28415    0.72475  -1.772  0.10179
## X3           -0.23106    1.84130  -0.125  0.90222
## X1:X2         0.03381    0.01017   3.325  0.00605 **
## X1:X3         0.02099    0.06416   0.327  0.74922
## X2:X3        -0.01247    0.01712  -0.729  0.48025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.56 on 12 degrees of freedom
## Multiple R-squared:  0.8016, Adjusted R-squared:  0.7023
## F-statistic: 8.079 on 6 and 12 DF,  p-value: 0.001179
```

```r
library(olsrr)
library(datasets)
k1<-ols_step_best_subset(pr2)
plot(k1)
```





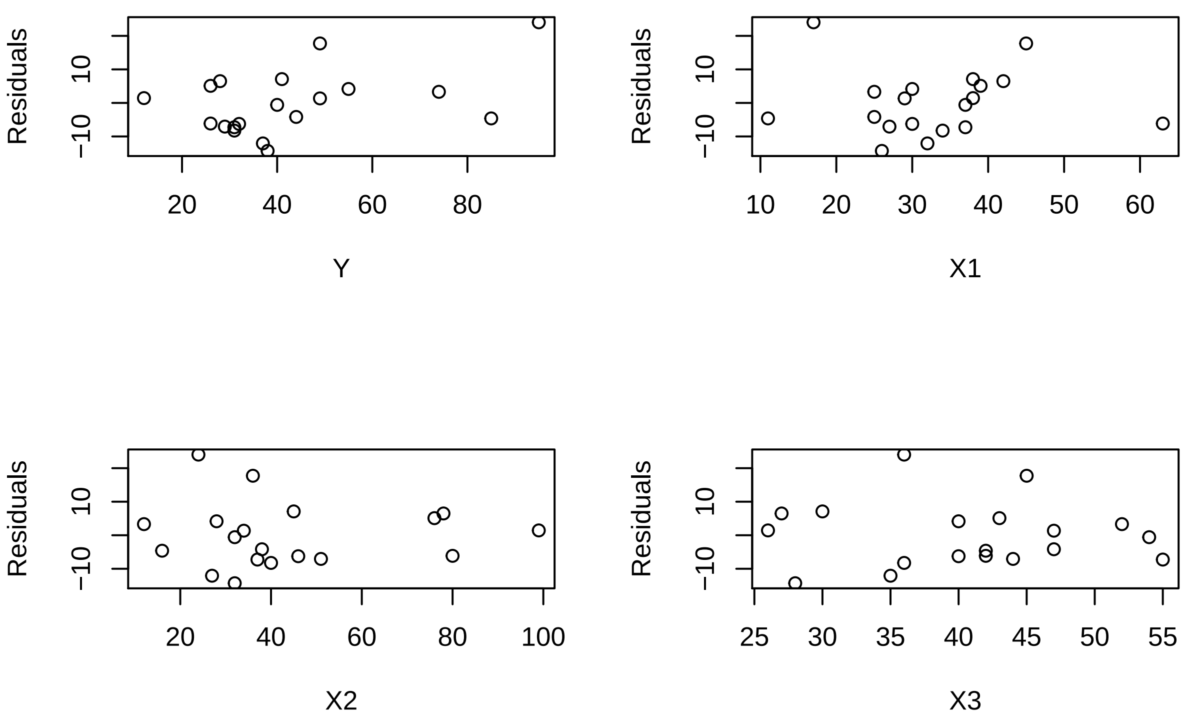

```
pr2.1<-lm(Y~X1+X2+X1:X2,data=Lung.Pressure)
summary(pr2.1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X1:X2, data = Lung.Pressure)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3075  -6.6602  -0.5824   4.6284  24.0398
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 134.399866  15.981599   8.410 4.63e-07 ***
## X1           -2.133022   0.522157  -4.085 0.000975 ***
## X2           -1.699330   0.363669  -4.673 0.000300 ***
## X1:X2         0.033347   0.009283   3.592 0.002667 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.58 on 15 degrees of freedom
## Multiple R-squared:  0.7922, Adjusted R-squared:  0.7507
## F-statistic: 19.06 on 3 and 15 DF,  p-value: 2.233e-05
```

b-) Obtain the residuals and plot them separately against Y and each of the three predictor variables. On the basis of these plots. should any further modification of the regression model be attempted?

No pattern with residuals and X3, indicating that X3 would not increase the power. There are couple of potential outliers in the data.

```r
par(mfrow=c(2,2))
plot(Lung.Pressure$Y,pr2.1$residuals,ylab="Residuals",xlab="Y")
plot(Lung.Pressure$X1,pr2.1$residuals,ylab="Residuals",xlab="X1")
plot(Lung.Pressure$X2,pr2.1$residuals,ylab="Residuals",xlab="X2")
plot(Lung.Pressure$X3,pr2.1$residuals,ylab="Residuals",xlab="X3")
```

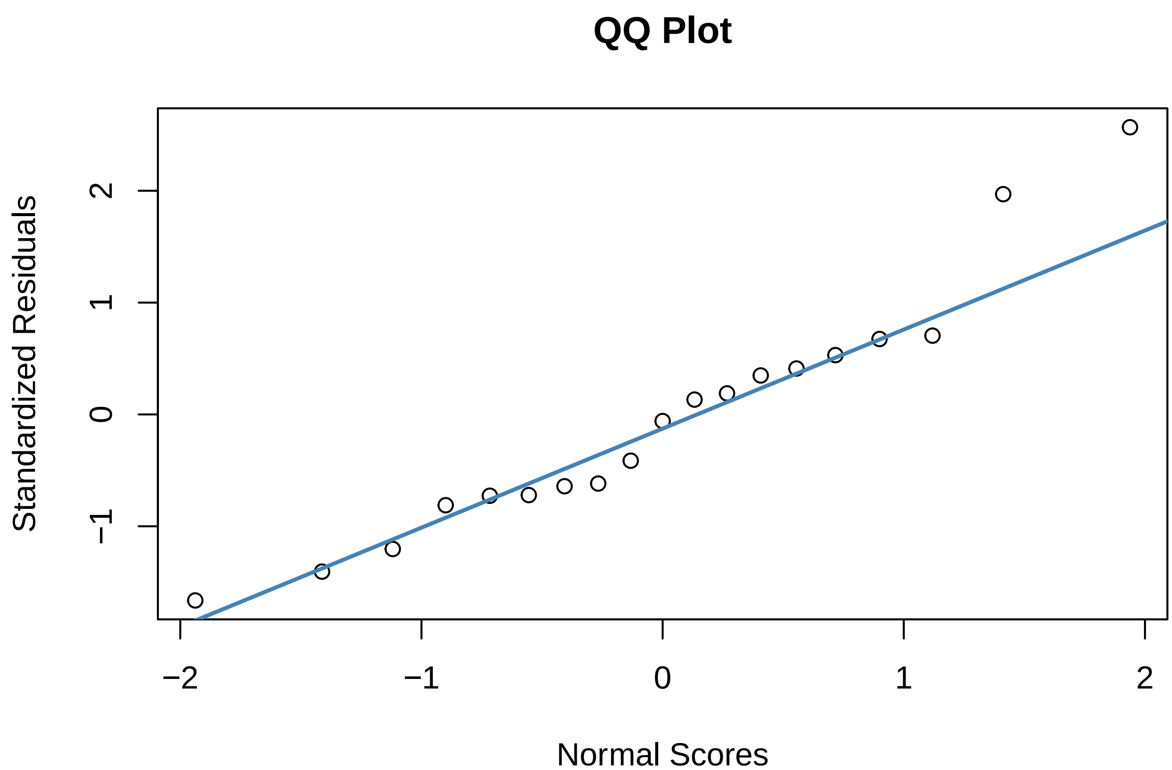


c-)

Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Does the normality assumption appear to be reasonable here?

The correlation is 96%, and graph indicates that the assumption is reasonable.

```
stdei<- rstandard(pr2.1)
qqnorm(stdei,ylab="Standardized Residuals",xlab="Normal Scores", main="QQ Plot")
qqline(stdei,col = "steelblue", lwd = 2)
```




```
a2<-anova(pr2.1)
mse<-a2$`Mean Sq`[4]
ei<-pr2.1$residuals
ei.rank<-rank(ei)
z1<-(ei.rank-0.375)/(19+0.375)
exp.rank<-sqrt(mse)*qnorm(z1)
cor(exp.rank,ei)
```

```
## [1] 0.9606285
```

d-) Obtain the variance inflation factors. Are there any indications that serious multicollinearity problems are present? Explain.

Multicollinearity present VIF>10 for X2 and the interaction term.

```
library(faraway)
vif(pr2.1)
```

```
##        X1        X2     X1:X2
##  5.431477 11.639560 22.474469
```

e-) Obtain the studentized deleted residuals and identify outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = .05$. State the decision rule and conclusion.

No outliers based on the bonforeni test. the largest deleted residual is observation 7, which larger than 3.

```
drst<-rstudent(pr2.1)
tb<-qt(1-0.05/(2*19),19-4-1)
sum(abs(drst)>abs(tb))
```

```
## [1] 0
```

f-) Obtain the diagonal elements of the hat matrix. Are there any outlying X observations? Discuss.

Indicating 3 outliers in X. Observations 3,8 and 15.

```
hii <- hatvalues(pr2.1)
hii
```

```
##          1          2          3          4          5          6          7
## 0.27569243 0.08336965 0.53886673 0.08482945 0.17565769 0.17374756 0.21775095
##          8          9         10         11         12         13         14
## 0.87827870 0.19254581 0.10171037 0.11155424 0.06796196 0.07530137 0.09294148
##         15         16         17         18         19
## 0.47982100 0.08967339 0.14443764 0.13905081 0.07680876
```

```
summary(hii)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06796 0.08725 0.13905 0.21053 0.20515 0.87828
```

```
sum(hii>(2*4/19))
```

```
## [1] 3
```

```
(hii>(2*4/19))
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
##    14    15    16    17    18    19
## FALSE  TRUE FALSE FALSE FALSE FALSE
```

g-) Cases 3, 8, and 15 are moderately far outlying with respect to their X values, and case 7 is relatively far outlying with respect to its Y value. Obtain DFFITS, DFBETAS, and Cook's distance values for these cases to assess their influence. What do you conclude? Case 8 has the largest cooks distance, it is an influential point.Cases 1 and 7 are outliers.

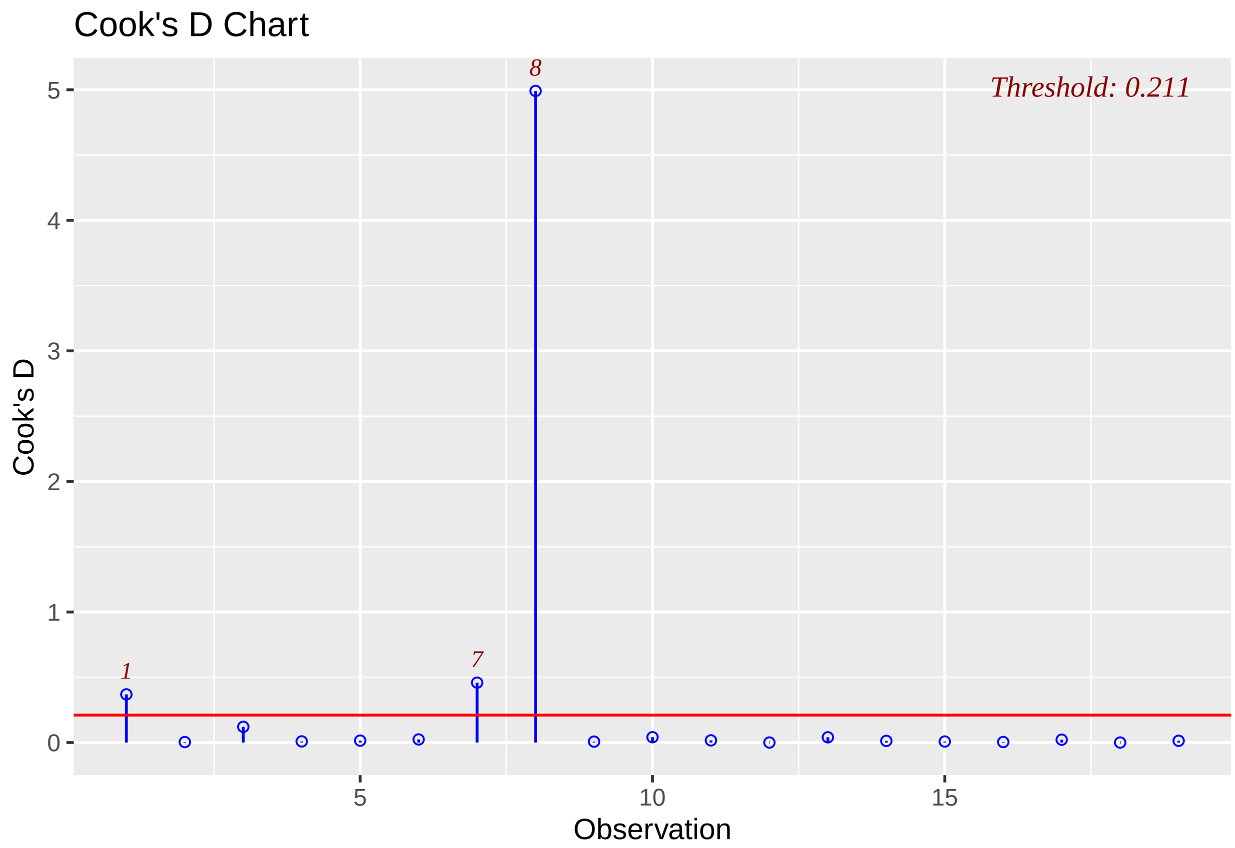
```
cd2<-influence.measures(pr2.1)
cd2
```

```
## Influence measures of
##   lm(formula = Y ~ X1 + X2 + X1:X2, data = Lung.Pressure) :
##
##       dfb.1_    dfb.X1   dfb.X2 dfb.X1.X   dffit cov.r   cook.d    hat inf
## 1  -0.74721  1.086986  0.22392 -0.59551  1.3632 0.550 0.369041 0.2757   *
## 2   0.00722  0.030375 -0.00594 -0.02091  0.1203 1.374 0.003833 0.0834
## 3  -0.65194  0.591913  0.43337 -0.48191 -0.6802 2.556 0.120516 0.5389   *
## 4   0.04063 -0.040524 -0.10592  0.09287 -0.1842 1.299 0.008854 0.0848
## 5  -0.04872 -0.000561  0.10888 -0.04217  0.2387 1.482 0.014982 0.1757
## 6  -0.02764 -0.026282  0.07195  0.01040  0.3035 1.410 0.023920 0.1737
## 7   1.45413 -1.277609 -0.74152  0.84752  1.7486 0.166 0.458917 0.2178   *
## 8  -1.54691  1.186623  3.16227 -3.28579 -4.7798 4.790 4.990815 0.8783   *
## 9   0.10208 -0.046089 -0.10509  0.07011  0.1650 1.580 0.007236 0.1925
## 10  0.03588 -0.171698  0.00924  0.10165 -0.4116 0.977 0.040999 0.1017
## 11  0.10895 -0.171977 -0.06084  0.11835 -0.2534 1.285 0.016600 0.1116
## 12  0.00127  0.006001  0.00576 -0.00950  0.0346 1.407 0.000320 0.0680
## 13 -0.12604  0.051336 -0.01204  0.03230 -0.4159 0.810 0.040228 0.0753
## 14 -0.10749  0.144648  0.07972 -0.10974  0.2215 1.270 0.012712 0.0929
## 15 -0.01551 -0.035251  0.07715 -0.01570  0.1749 2.510 0.008170 0.4798   *
## 16 -0.02274  0.017421 -0.03719  0.02835 -0.1262 1.383 0.004218 0.0897
## 17  0.05185 -0.018569 -0.19203  0.14258 -0.2914 1.337 0.021956 0.1444
## 18  0.00912 -0.015709 -0.00358  0.00987 -0.0230 1.529 0.000142 0.1391
## 19  0.08055 -0.124148 -0.08285  0.11431 -0.2314 1.193 0.013708 0.0768
```

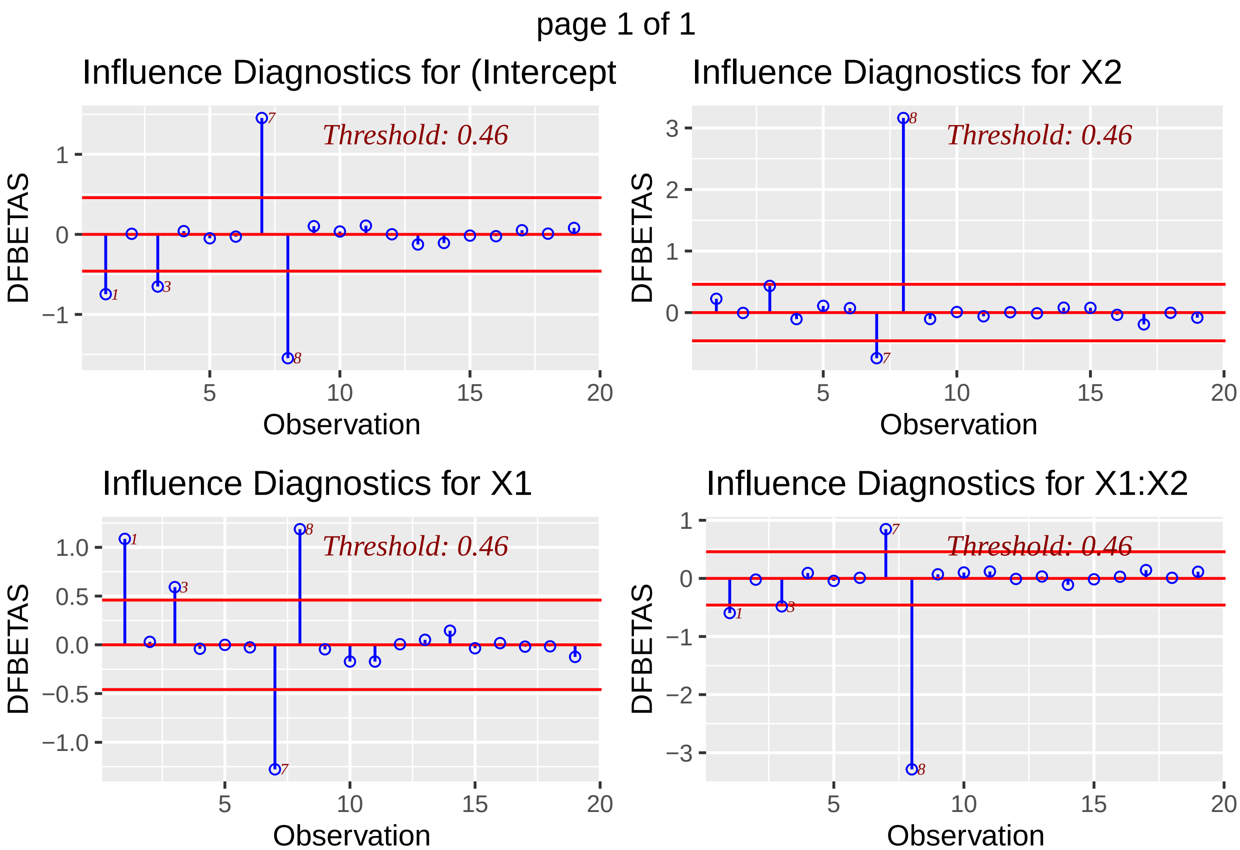
```r
cd3<-cd2$infmat
cd3[c(3,7,8,15),]
```

```
##        dfb.1_      dfb.X1      dfb.X2    dfb.X1:X      dffit     cov.r
## 3  -0.6519371  0.59191342  0.43337176 -0.48191103 -0.6801824 2.5561254
## 7   1.4541305 -1.27760852 -0.74151968  0.84752328  1.7485509 0.1661137
## 8  -1.5469080  1.18662253  3.16226530 -3.28579003 -4.7797848 4.7895257
## 15 -0.0155059 -0.03525106  0.07714703 -0.01569977  0.1748573 2.5095274
##         cook.d       hat
## 3  0.120515509 0.5388667
## 7  0.458917058 0.2177509
## 8  4.990814979 0.8782787
## 15 0.008170411 0.4798210
```

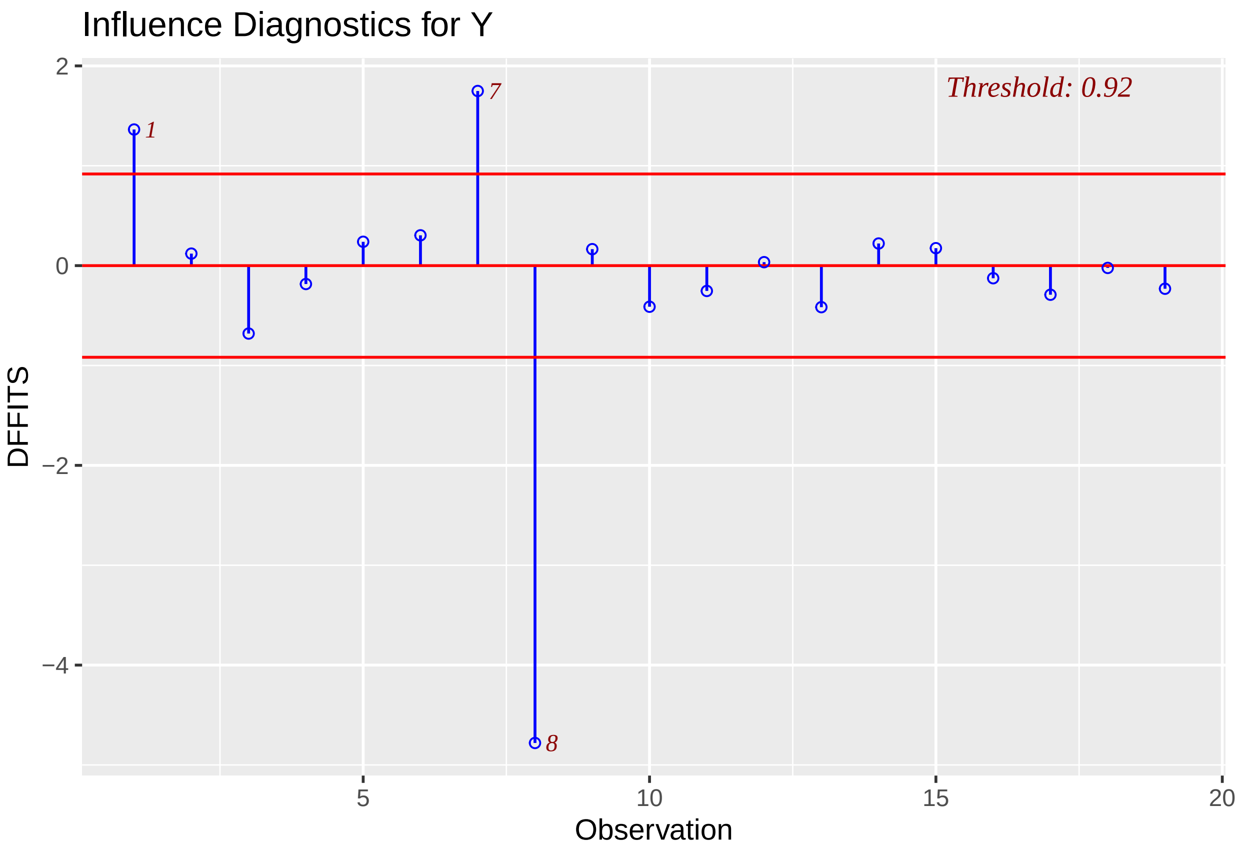
```r
ols_plot_cooksd_chart(pr2.1)
```



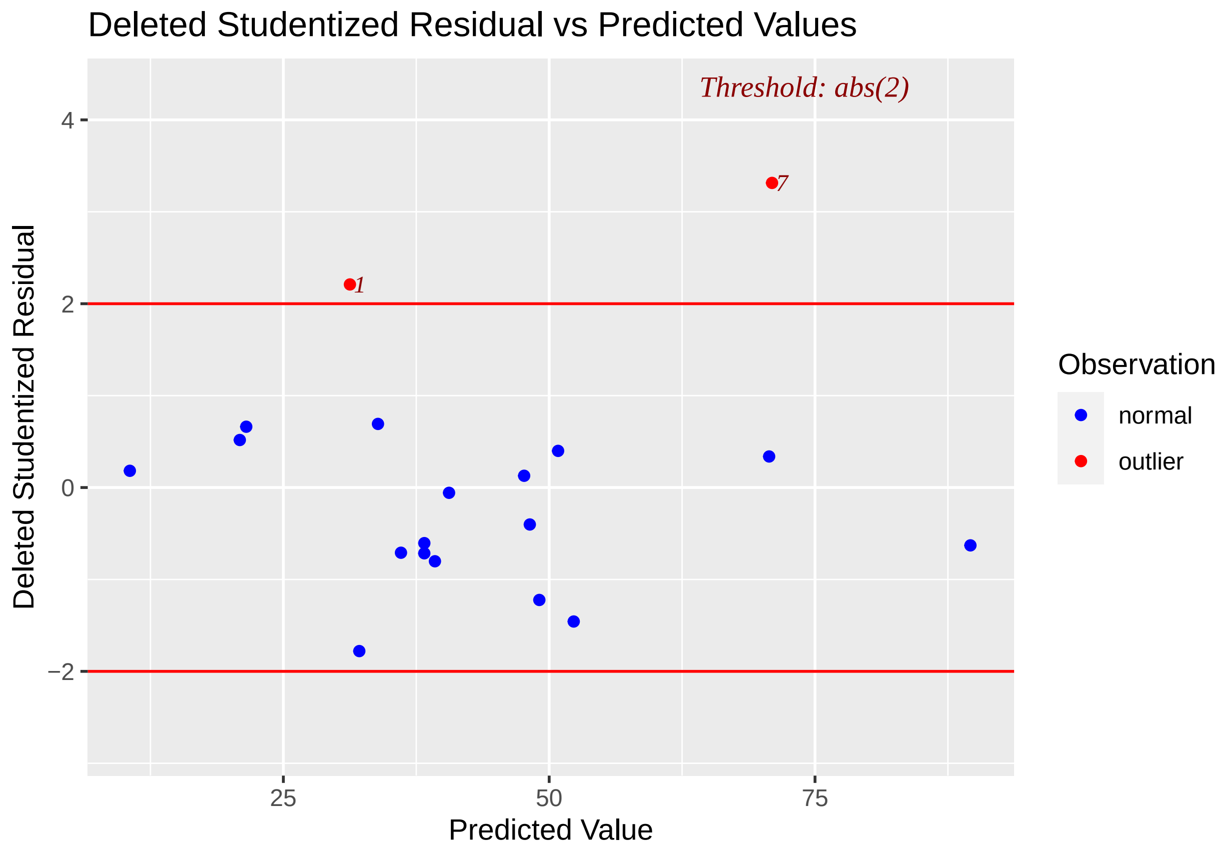

```
ols_plot_dfbetas(pr2.1)
```




```
ols_plot_dffits(pr2.1)
```




```r
ols_plot_resid_stud_fit(pr2.1)
```



Problem 3

Refer to the Prostate cancer data set. Serum prostate-specific antigen (PSA) was determined in 97 men with advanced prostate cancer. PSA (Y) is a well-established screening test for prostate cancer and the oncologists wanted to examine the correlation between level of PSA and a number of clinical measures for men who were about to undergo radical prostatectomy. The measures are cancer volume (X_1), prostate weight (X_2), patient age (X_3), the amount of benign prostatic hyperplasia (X_4), seminal vesicle invasion (X_5), capsular penetration (X_6), and Gleason score (X_7). (20 points, 5 points each)

a-) Select a random sample of 65 observations to use as the model-building data set. Develop a best subset model for predicting PSA. Justify your choice of model. Assess your model's ability to predict and discuss its usefulness to the oncologists.

We tried two variable selection methdoldogies, best subset and stepwise. Both methods are suggesting Y=X1+X6 to be the best model.

```
library(olsrr)
PROSTATE.CANCER <- read.csv("/cloud/project/PROSTATE.CANCER.csv")
set.seed(567)
sample.ind <- sample(1:nrow(PROSTATE.CANCER), size = 65)
devq5 <- PROSTATE.CANCER[sample.ind,]
holdoutq5 <- PROSTATE.CANCER[-sample.ind,]
pr3<-lm(Y~X1+X2+X3+X4+X5+X6+X7,data=devq5)
ols_step_both_p(pr3,prem=0.05,details=TRUE)
```

```
## Stepwise Selection Method
```

```
## ---------------------------
##
## Candidate Terms:
##
## 1. X1
## 2. X2
## 3. X3
## 4. X4
## 5. X5
## 6. X6
## 7. X7
##
## We are selecting variables based on p value...
##
##
## Stepwise Selection: Step 1
##
## - X6 added
##
##                          Model Summary
## ---------------------------------------------------------------
## R                       0.696       RMSE               29.360
## R-Squared               0.485       Coef. Var         121.436
## Adj. R-Squared          0.477       MSE               861.994
## Pred R-Squared          0.263       MAE                17.101
## ---------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## ----------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square      F         Sig.
## ----------------------------------------------------------------------
## Regression     51158.364         1      51158.364    59.349    0.0000
## Residual       54305.592        63        861.994
## Total         105463.956        64
## ----------------------------------------------------------------------
##
##                                  Parameter Estimates
## ---------------------------------------------------------------------------------------
##       model     Beta    Std. Error    Std. Beta      t       Sig      lower     upper
## ---------------------------------------------------------------------------------------
## (Intercept)    8.258         4.187                 1.972    0.053    -0.110    16.625
##          X6    7.469         0.970        0.696    7.704    0.000     5.532     9.407
## ---------------------------------------------------------------------------------------
##
##
##
## Stepwise Selection: Step 2
##
## - X1 added
##
```

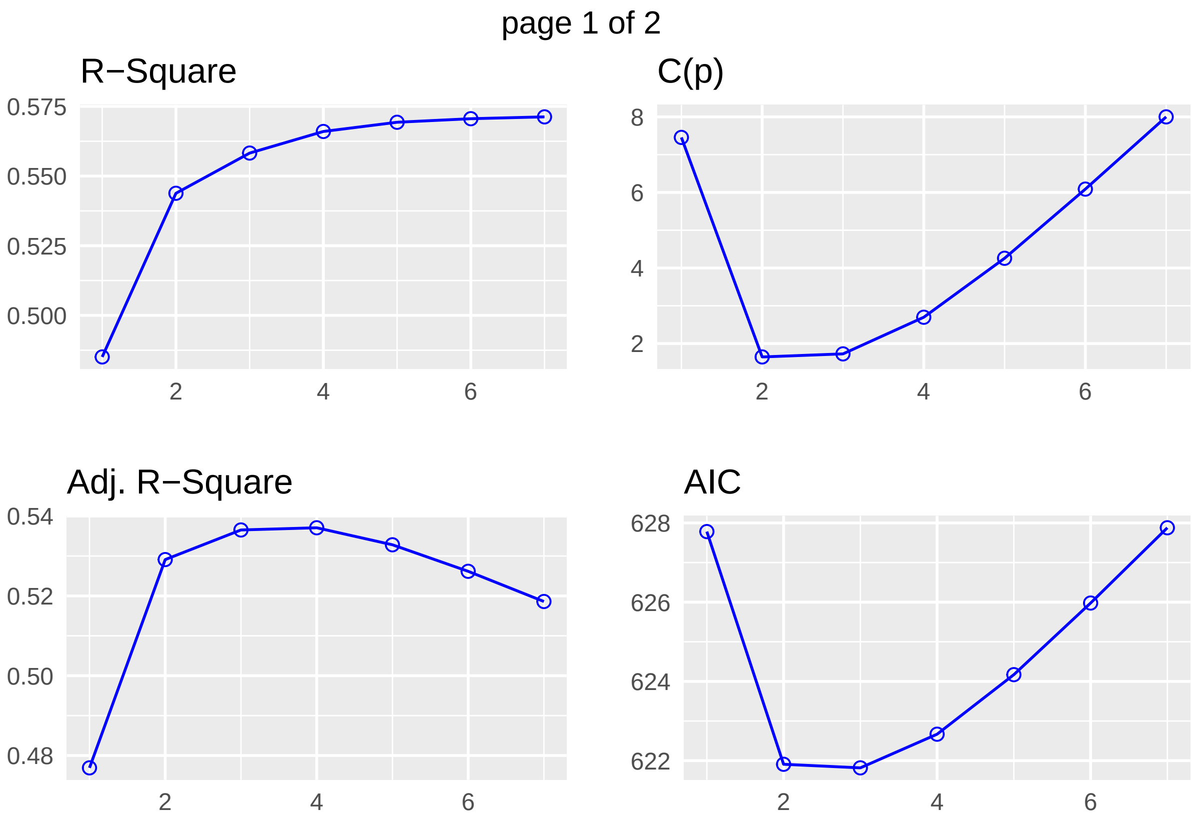
```
##                           Model Summary
## ---------------------------------------------------------------
## R                        0.737       RMSE               27.856
## R-Squared                0.544       Coef. Var         115.218
## Adj. R-Squared           0.529       MSE               775.976
## Pred R-Squared           0.305       MAE                15.294
## ---------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                  ANOVA
## ----------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square      F         Sig.
## ----------------------------------------------------------------------
## Regression    57353.472        2      28676.736    36.956    0.0000
## Residual      48110.483       62        775.976
## Total        105463.956       64
## ----------------------------------------------------------------------
##
##                                   Parameter Estimates
## ----------------------------------------------------------------------------------------
##       model     Beta    Std. Error    Std. Beta      t        Sig      lower     upper
## ----------------------------------------------------------------------------------------
## (Intercept)    2.012         4.546                 0.442    0.660    -7.076    11.100
##          X6    4.906         1.292        0.457    3.797    0.000     2.324     7.489
##          X1    1.567         0.555        0.340    2.826    0.006     0.458     2.676
## ----------------------------------------------------------------------------------------
##
##
##
##                           Model Summary
## ---------------------------------------------------------------
## R                        0.737       RMSE               27.856
## R-Squared                0.544       Coef. Var         115.218
## Adj. R-Squared           0.529       MSE               775.976
## Pred R-Squared           0.305       MAE                15.294
## ---------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                  ANOVA
## ----------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square      F         Sig.
## ----------------------------------------------------------------------
## Regression    57353.472        2      28676.736    36.956    0.0000
## Residual      48110.483       62        775.976
## Total        105463.956       64
## ----------------------------------------------------------------------
##
##                                   Parameter Estimates
```

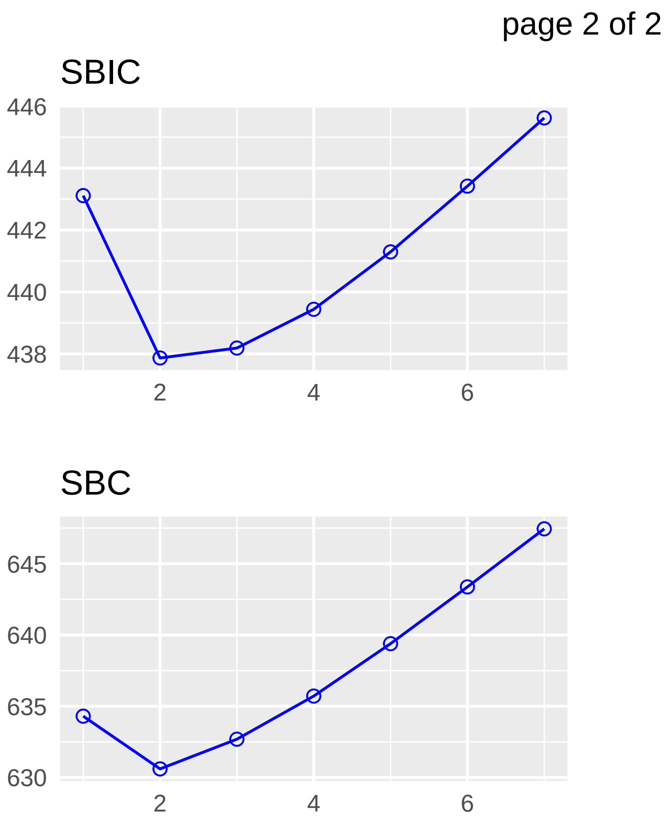
```
## ----------------------------------------------------------------------------------------
##       model     Beta    Std. Error    Std. Beta      t       Sig      lower     upper
## ----------------------------------------------------------------------------------------
## (Intercept)    2.012         4.546                 0.442    0.660    -7.076    11.100
##          X6    4.906         1.292        0.457    3.797    0.000     2.324     7.489
##          X1    1.567         0.555        0.340    2.826    0.006     0.458     2.676
## ----------------------------------------------------------------------------------------
##
##
##
## No more variables to be added/removed.
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## --------------------------------------------------------------
## R                       0.737       RMSE              27.856
## R-Squared               0.544       Coef. Var        115.218
## Adj. R-Squared          0.529       MSE              775.976
## Pred R-Squared          0.305       MAE               15.294
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## ---------------------------------------------------------------------
##                  Sum of
##                 Squares        DF    Mean Square      F         Sig.
## ---------------------------------------------------------------------
## Regression    57353.472         2      28676.736    36.956    0.0000
## Residual      48110.483        62        775.976
## Total        105463.956        64
## ---------------------------------------------------------------------
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------
##       model     Beta    Std. Error    Std. Beta      t       Sig      lower     upper
## ----------------------------------------------------------------------------------------
## (Intercept)    2.012         4.546                 0.442    0.660    -7.076    11.100
##          X6    4.906         1.292        0.457    3.797    0.000     2.324     7.489
##          X1    1.567         0.555        0.340    2.826    0.006     0.458     2.676
## ----------------------------------------------------------------------------------------

##
##                               Stepwise Selection Summary
## ---------------------------------------------------------------------------------------
##                    Added/                    Adj.
## Step    Variable    Removed     R-Square    R-Square     C(p)        AIC        RMSE
## ---------------------------------------------------------------------------------------
##    1       X6       addition       0.485       0.477    7.4560    627.7817    29.3597
##    2       X1       addition       0.544       0.529    1.6460    621.9084    27.8563
```

```
## ---------------------------------------------------------------------------------
```

```
k1<-ols_step_best_subset(pr3)
plot(k1)
```






```r
pr31<-lm(Y~X1+X6,data=devq5)
summary(pr31)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X6, data = devq5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.117  -5.059   1.852   6.041 123.528
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0117     4.5463   0.442 0.659674
## X1            1.5673     0.5547   2.826 0.006343 **
## X6            4.9062     1.2920   3.797 0.000335 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.86 on 62 degrees of freedom
## Multiple R-squared:  0.5438, Adjusted R-squared:  0.5291
## F-statistic: 36.96 on 2 and 62 DF,  p-value: 2.711e-11
```

b-) Perform appropriate diagnostic checks to evaluate outliers and assess their influence.

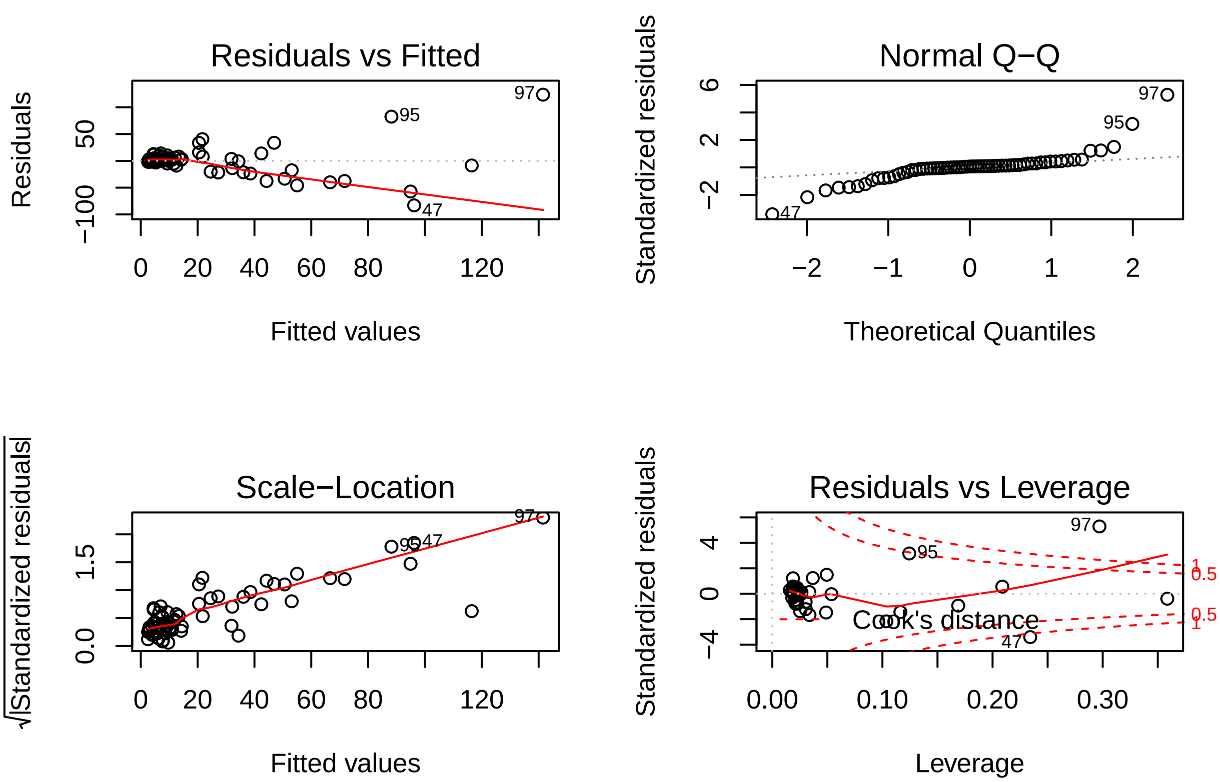
The model is significant with 54% R-Square. QQ plot show problem with the normal distribution. Residual vs Fitted graph shows a megaphone shape indicating un equal variances. Cook's distance graph show that observation 33 and 8 are influential points. Observations 33,61,8, and 44 are outliers.

```r
influence.measures(pr31)
```

```
## Influence measures of
##   lm(formula = Y ~ X1 + X6, data = devq5) :
##
##       dfb.1_    dfb.X1    dfb.X6     dffit cov.r   cook.d    hat inf
## 73  0.038312 -1.53e-02  4.77e-03  0.039858 1.066 5.38e-04 0.0191
## 85  0.069107 -3.31e-02  2.48e-02  0.078374 1.053 2.07e-03 0.0187
## 61  0.068109 -2.20e-02 -7.62e-03  0.069283 1.063 1.62e-03 0.0226
## 46  0.010425  2.36e-03 -6.31e-03  0.013449 1.072 6.13e-05 0.0210
## 25  0.007133 -2.39e-03 -7.21e-04  0.007238 1.074 1.77e-05 0.0228
## 23  0.014668 -6.31e-03 -2.37e-04  0.014698 1.076 7.32e-05 0.0249
## 7  -0.016318  4.30e-03  2.71e-03 -0.016841 1.073 9.61e-05 0.0218
## 47 -0.160876  1.03e+00 -1.95e+00 -2.074925 0.735 1.19e+00 0.2342   *
## 53  0.026063 -1.41e-02  5.45e-03  0.026644 1.073 2.40e-04 0.0231
## 48  0.013864 -3.85e-03 -9.17e-04  0.014385 1.070 7.01e-05 0.0191
## 30 -0.088578  3.10e-02 -1.18e-01 -0.221547 0.982 1.61e-02 0.0252
## 24 -0.088232  5.35e-02 -5.85e-02 -0.115993 1.041 4.51e-03 0.0213
## 31  0.002809 -1.25e-03  1.44e-04  0.002822 1.075 2.70e-06 0.0230
## 50  0.018118 -3.67e-03 -2.89e-03  0.019133 1.070 1.24e-04 0.0192
## 2  -0.010160  4.58e-03 -2.60e-05 -0.010168 1.077 3.50e-05 0.0255
## 6  -0.002336  1.06e-03 -1.04e-05 -0.002338 1.078 1.85e-06 0.0256
## 43  0.012176 -5.38e-03  8.95e-04  0.012290 1.073 5.12e-05 0.0220
## 60  0.049607 -8.02e-03 -1.27e-02  0.052916 1.065 9.47e-04 0.0208
## 3  -0.012075  5.17e-03  2.14e-04 -0.012101 1.076 4.96e-05 0.0249
## 92  0.064361  2.64e-01 -2.62e-01  0.343506 0.990 3.85e-02 0.0494
## 10 -0.006098  2.20e-03  4.77e-04 -0.006160 1.075 1.29e-05 0.0233
## 40  0.000504 -2.21e-04  4.93e-05  0.000512 1.073 8.87e-08 0.0211
## 17 -0.007979  4.26e-03 -1.11e-03 -0.008030 1.077 2.18e-05 0.0252
## 29 -0.026201  1.09e-02 -2.82e-03 -0.026907 1.069 2.45e-04 0.0199
## 83  0.004558  1.78e-02 -1.37e-02  0.024340 1.086 2.01e-04 0.0336
## 80 -0.000330 -6.80e-03  5.34e-03 -0.008102 1.110 2.22e-05 0.0537
## 65  0.010699  9.71e-03 -1.30e-02  0.020137 1.078 1.37e-04 0.0265
## 59  0.063681 -1.96e-02 -7.99e-03  0.064988 1.064 1.43e-03 0.0224
## 28  0.017795 -7.49e-03 -4.28e-04  0.017844 1.076 1.08e-04 0.0247
## 66  0.026856 -3.24e-03 -5.35e-03  0.029397 1.067 2.93e-04 0.0180
## 36  0.001225 -5.35e-05 -4.45e-04  0.001373 1.072 6.39e-07 0.0204
## 90  0.148772 -7.12e-02  5.34e-02  0.168721 0.995 9.41e-03 0.0187
## 97 -1.005084 -2.64e-01  3.37e+00  4.603614 0.247 3.94e+00 0.2971   *
## 72  0.066976 -8.02e-03 -1.97e-02  0.072630 1.059 1.78e-03 0.0206
## 94  0.127646 -2.62e-01  1.09e-01 -0.289155 1.625 2.83e-02 0.3587   *
## 58  0.063590 -2.06e-02 -7.11e-03  0.064685 1.065 1.41e-03 0.0226
## 57  0.010453 -6.00e-03  3.37e-03  0.011083 1.073 4.16e-05 0.0221
## 41  0.026703 -7.79e-03 -3.75e-03  0.027358 1.072 2.53e-04 0.0221
## 18 -0.051454 -5.16e-02  4.81e-02 -0.106189 1.045 3.79e-03 0.0208
## 1  -0.012898  5.58e-03  1.82e-04 -0.012922 1.077 5.66e-05 0.0250
## 19  0.010665 -4.60e-03 -1.61e-04  0.010686 1.077 3.87e-05 0.0250
## 84 -0.020188 -1.53e-02 -4.44e-02 -0.112711 1.061 4.28e-03 0.0303
## 70  0.048609 -1.05e-02 -7.18e-03  0.051102 1.064 8.83e-04 0.0193
## 86  0.196062 -2.97e-01 -2.44e-01 -0.760963 0.924 1.81e-01 0.1036   *
## 39 -0.060084 -3.98e-03 -1.64e-01 -0.317919 0.946 3.27e-02 0.0337
## 13 -0.037709 -7.79e-04  1.27e-02 -0.043896 1.064 6.52e-04 0.0181
## 16 -0.021717 -2.49e-03  1.10e-02 -0.026355 1.070 2.35e-04 0.0205
```

```
## 91 -0.053970  2.71e-01 -2.22e-01  0.285168 1.307 2.74e-02 0.2089   *
## 38  0.023927 -7.74e-03 -2.68e-03  0.024339 1.073 2.01e-04 0.0226
## 56  0.035525 -2.33e-03 -1.22e-02  0.039426 1.068 5.26e-04 0.0204
## 32  0.019766 -7.24e-03 -1.45e-03  0.019951 1.074 1.35e-04 0.0234
## 75  0.031293 -1.54e-01 -5.80e-02 -0.337188 0.992 3.72e-02 0.0488
## 63 -0.018241 -1.07e-01 -1.18e-03 -0.216101 1.007 1.54e-02 0.0304
## 49  0.008995  2.95e-03 -6.26e-03  0.012293 1.073 5.12e-05 0.0216
## 67 -0.047842  2.54e-02 -3.67e-02 -0.071800 1.060 1.74e-03 0.0208
## 55  0.064157 -3.94e-01  3.28e-01 -0.420206 1.211 5.90e-02 0.1688   *
## 35  0.011264 -5.42e-03  9.20e-04  0.011303 1.076 4.33e-05 0.0240
## 8  -0.014651  4.04e-03  2.27e-03 -0.015070 1.073 7.69e-05 0.0219
## 77  0.026657  4.43e-03 -6.35e-03  0.035901 1.063 4.36e-04 0.0159
## 54 -0.069321  3.96e-02 -6.71e-02 -0.118515 1.044 4.71e-03 0.0229
## 95  0.015090 -3.48e-01  1.07e+00  1.292260 0.707 4.74e-01 0.1244   *
## 82 -0.093129  2.60e-01 -4.78e-01 -0.525584 1.074 9.05e-02 0.1163
## 33 -0.000866  5.22e-05  3.01e-04 -0.000963 1.072 3.14e-07 0.0204
## 93  0.006438  1.72e-01 -7.28e-02  0.241858 1.012 1.93e-02 0.0368
## 42  0.006436  2.47e-04 -2.81e-03  0.007507 1.072 1.91e-05 0.0204
```

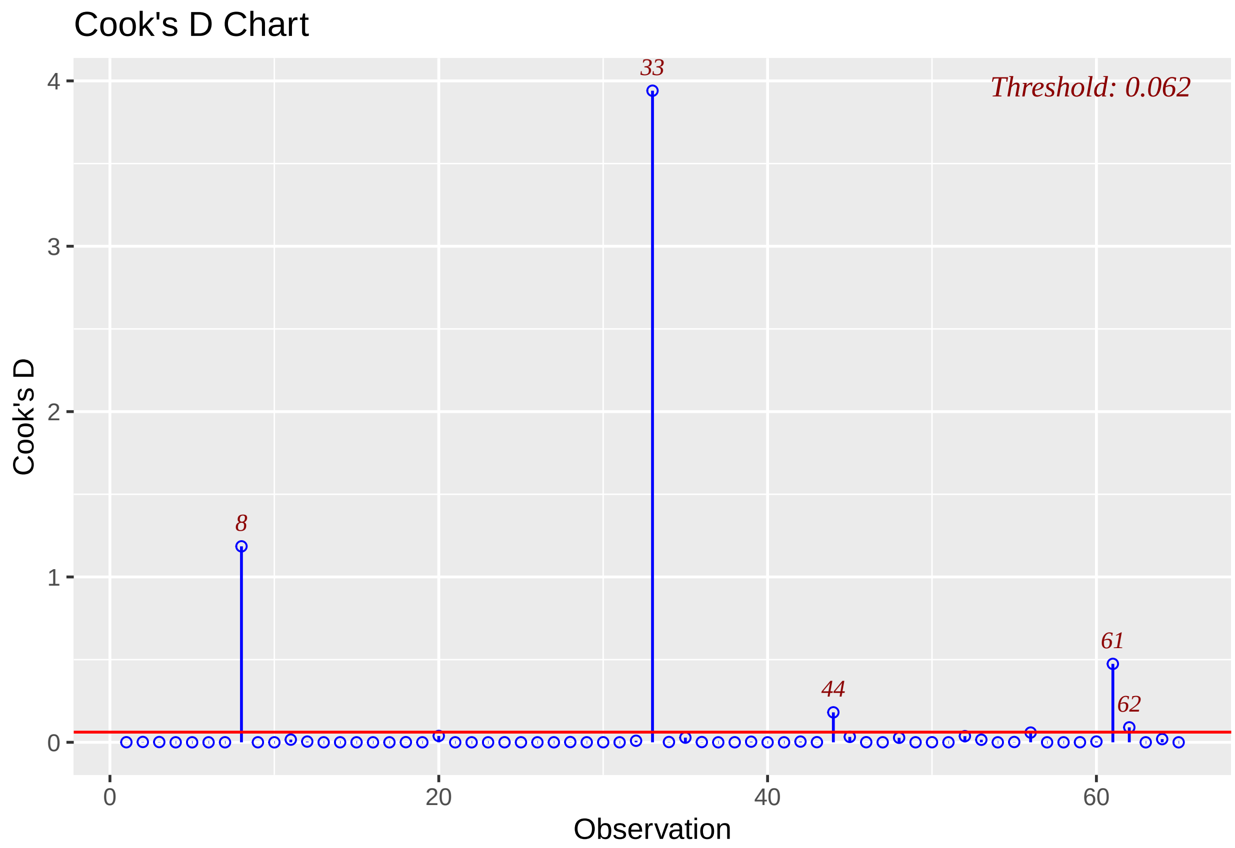
```
par(mfrow=c(2,2))
plot(pr31)
```



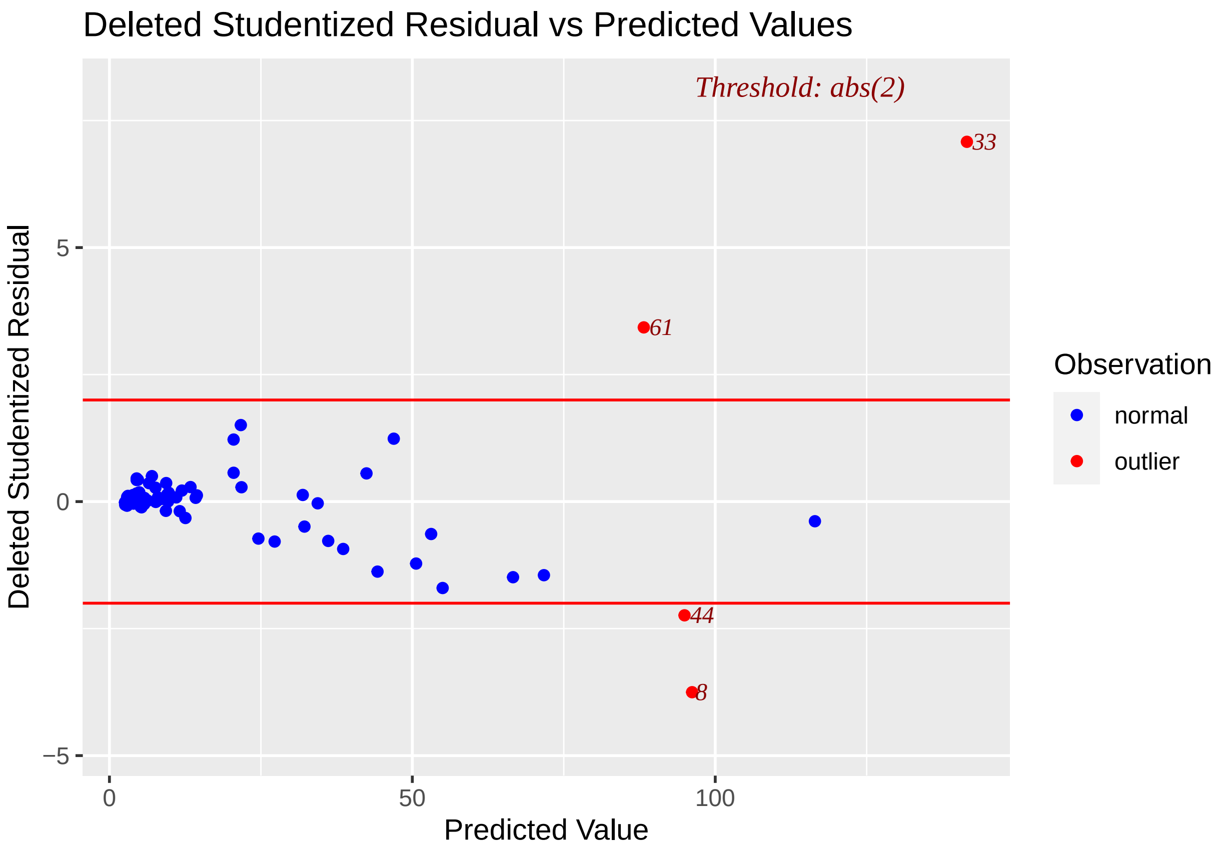

```
vif(pr31)
```

```
##       X1       X6
## 1.972479 1.972479
```

```r
ols_plot_cooksd_chart(pr31)
```




```r
ols_plot_resid_stud_fit(pr31)
```



c-) Fit the regression model identified in part a to the validation data set. Compare the estimated regression coefficients and their estimated standard errors with those obtained in part a. Also compare the error mean square and coefficients of multiple determination. Does the model fitted to the validation data set yield similar estimates as the model fitted to the model-building data set?

Capsular.penetration (X_6) becomes insignificant and Rsquare decreases. MSE increased from 776 to 1149.8.Indicating problem with the model stability.

```r
f31<-lm(Y~X1+X6,data=holdoutq5)
summary(f31)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X6, data = holdoutq5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.762  -9.941   0.629   6.181 150.292
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.010      9.118  -0.659  0.51504
## X1             6.467      1.682   3.844  0.00061 ***
## X6            -4.129      2.424  -1.703  0.09918 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 33.91 on 29 degrees of freedom
## Multiple R-squared:  0.3844, Adjusted R-squared:  0.342
## F-statistic: 9.055 on 2 and 29 DF,  p-value: 0.0008805
```

```
anova(f31)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## X1         1  17486 17486.2 15.2074 0.000525 ***
## X6         1   3337  3336.6  2.9017 0.099180 .
## Residuals 29  33346  1149.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(pr31)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1  46164   46164  59.491 1.246e-10 ***
## X6         1  11190   11190  14.420 0.0003352 ***
## Residuals 62  48110     776
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
MSPR<-sum(f31$residuals^2)/length(f31$residuals)
MSPR
```

```
## [1] 1042.05
```

d-) Calculate the mean squared prediction error and compare it to MSE obtained from the model-building data set. Is there evidence of a substantial bias problem in MSE here?

MSPR=1042.05 and MSE=776.Indicating problem with the model stability.

```
MSPR<-sum(f31$residuals^2)/length(f31$residuals)
MSPR
```

```
## [1] 1042.05
```