CS-E-106: Data Modeling

Assignment 8

Instructor: Hakan Gogtas
Submitted by: *Saurabh Kulkarni*

Due Date: *11/25/2019*

Question 1 Refer to Brand preference data, build a model with all independent variables (45 pts)

(a) Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with = .10. State the decision rule and conclusion. (5pts)

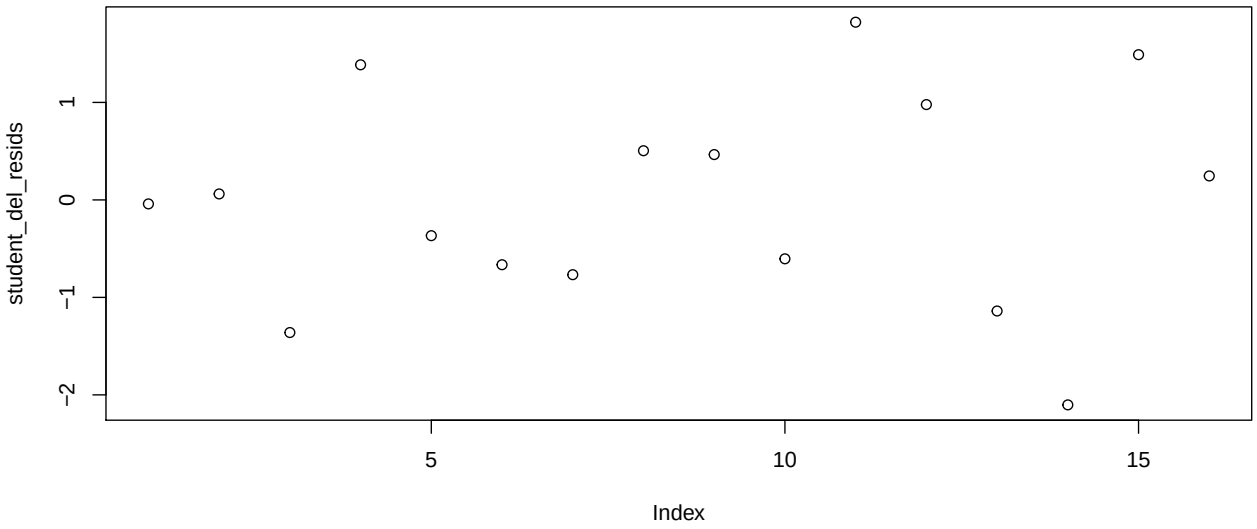
Solution:

```
brand_data = read.csv("Brand Preference.csv")
lm_brand = lm(Y~., data=brand_data)
summary(lm_brand)
```

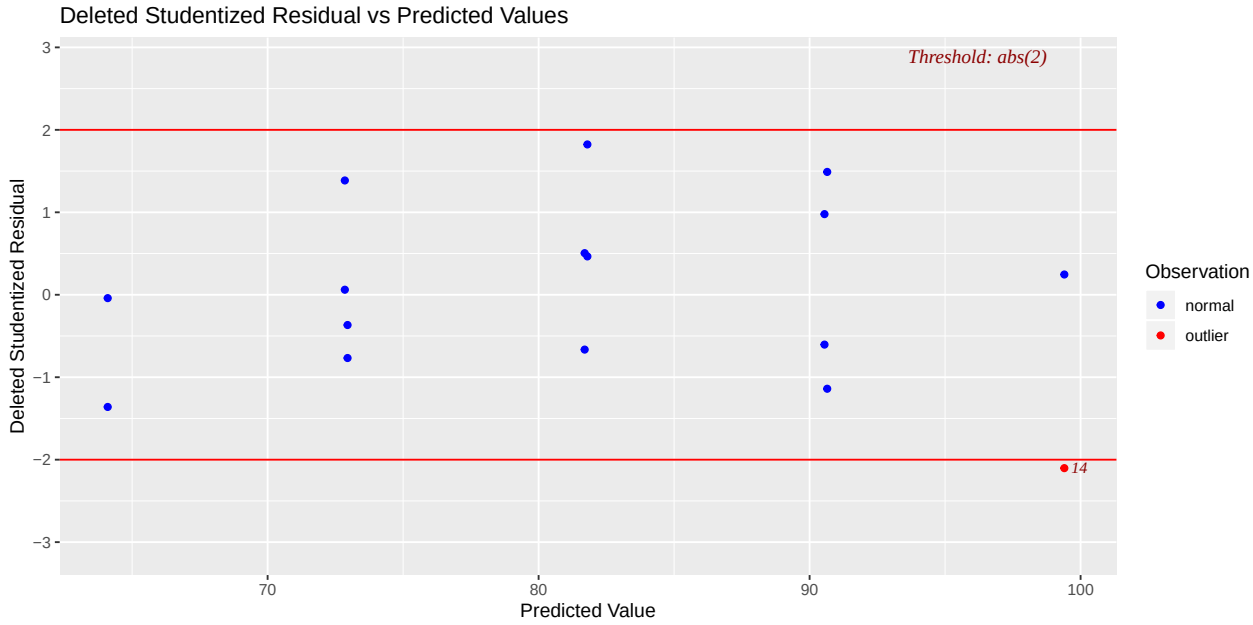
```
##
## Call:
## lm(formula = Y ~ ., data = brand_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.400 -1.762  0.025  1.587  4.200
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.6500     2.9961  12.566 1.20e-08 ***
## X1            4.4250     0.3011  14.695 1.78e-09 ***
## X2            4.3750     0.6733   6.498 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.693 on 13 degrees of freedom
## Multiple R-squared:  0.9521, Adjusted R-squared:  0.9447
## F-statistic: 129.1 on 2 and 13 DF,  p-value: 2.658e-09
```

Studentized deleted residuals:

```
student_del_resids = rstudent(lm_brand)
plot(student_del_resids)
```




```
ols_plot_resid_stud_fit(lm_brand)
```



Interpretation

We can see that case 14 is an outlier with regard to the Y observations, based on the studentized deleted residuals.

Bonferroni Outlier Test:

H_0 : No outlier H_1 : Atleast one outlier

```
n = nrow(brand_data)
p = length(lm_brand$coefficients)
alpha = 0.1
tTest = qt(1-alpha/(2*n),n-p-1)
tTest
```

```
## [1] 3.307783
```

```
any(abs(student_del_resids)>=abs(tTest))
```

```
## [1] FALSE
```

```
which(abs(student_del_resids)>=abs(tTest))
```

```
## named integer(0)
```

Decision Rule:

- If $|t_i| \leq t(1-\alpha/2n; n-p-1)$, $n=i$ is not an outlier.
- If $|t_i| > t(1-\alpha/2n; n-p-1)$, $n=i$ is an outlier.

Result:

Since none $|t_i| > t(1-\alpha/2n; n-p-1)$, we conclude H_0. There are no outliers with $\alpha = 0.1$.

(b) Obtain the diagonal elements of the hat matrix, and provide an explanation for the pattern in these elements. (5pts)

Solution:

```
hii = hatvalues(lm_brand)
index = hii>2*p/n
index
```

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

Interpretation:

The hat matrix measures the distance between X_i and $\bar{X}$. The fact that none of the $h_{ii} > \frac{2*p}{n}$ means that all the X_is are more or less close to their means.

(c) Are any of the observations outlying with regard to their X values according? (5pts)

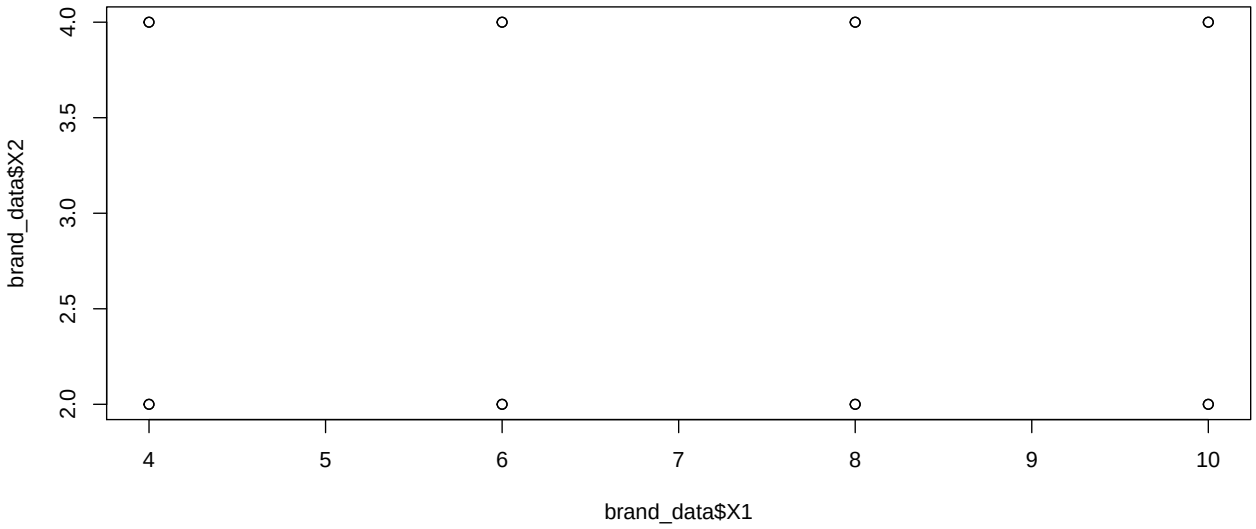
Solution:

We do not see any observations that are outliers with regard to their X values.

(d) Management wishes to estimate the mean degree of brand liking for moisture content X1 = 10 and sweetness X2 = 3. Construct a scatter plot of X2 against X1 and determine visually whether this prediction involves an extrapolation beyond the range of the data. Also, use (10.29) to determine whether an extrapolation is involved. Do your conclusions from the two methods agree? (5pts)

Solution:

```
plot(brand_data$X1, brand_data$X2)
```



Interpretation:

We can see that X_{1new} and X_{2new} are both well within the range of the given dataset, so we don't need any extrapolation beyond the range of the data already given to us.

```
X = rep(c(1))
X = cbind(X, data.matrix(brand_data[,-(names(brand_data)%in%c("Y"))]))
X
```

```
##       X X1 X2
##  [1,] 1  4  2
##  [2,] 1  4  4
##  [3,] 1  4  2
##  [4,] 1  4  4
##  [5,] 1  6  2
##  [6,] 1  6  4
##  [7,] 1  6  2
##  [8,] 1  6  4
##  [9,] 1  8  2
## [10,] 1  8  4
## [11,] 1  8  2
## [12,] 1  8  4
## [13,] 1 10  2
## [14,] 1 10  4
## [15,] 1 10  2
## [16,] 1 10  4
```

```
XTX = crossprod(X)
XTX_inv = solve(XTX)
XTX_inv
```

```
##           X            X1      X2
## X   1.2375 -8.750000e-02 -0.1875
## X1 -0.0875  1.250000e-02  0.0000
## X2 -0.1875  2.602085e-17  0.0625
```

```
Xh = c(1,10,3)
H_extrap = t(Xh)%*%XTX_inv%*%Xh
H_extrap
```

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

Thus, we see that $h_{new.new} = 0.175$

```
range(hii)
```

```
## [1] 0.1375 0.2375
```

Interpretation:

From the above result, we see the value of $h_{new.new}$ is well within the range of the leverage values h_{ii} for the cases in the data set, so no hidden extrapolation is involved for this estimate.

(e) The largest absolute studentized deleted residual is for case 14. Obtain the DFFlTS, DFBETAS, and Cook's distance values for this case to assess the influence of this case. What do you conclude? (5pts)

Solution:

```
influence_results = influence.measures(lm_brand)
influence_results$infmat[14,]
```

```
##     dfb.1_     dfb.X1     dfb.X2      dffit      cov.r     cook.d
##  0.8388068 -0.8076796 -0.6020088 -1.1735312  0.6506614  0.3634123
##        hat
##  0.2375000
```

Interpretation:

|DFBETAS| are all < 1, so case 14 does not have a big influence on betas. |DFFITS| > 1, so case 14 has considerable influence on Y_{14}. According to Cook's Distance, case 14 has little influence on all the fitted values, since $0.1 < P(F(p, n-p) \leq Cook'sDistance) < 0.5$.

(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? (10pts)

Solution:

```
new_df = brand_data[-c(14),]
new_lm_brand = lm(Y~., data=new_df)
newer_result = mean(abs((new_lm_brand$fitted.values-lm_brand$fitted.values)/lm_brand$fitted.values))
```

```
## Warning in new_lm_brand$fitted.values - lm_brand$fitted.values: longer
## object length is not a multiple of shorter object length
```

```
newer_new_result = (100*newer_result)/nrow(brand_data)
print(newer_new_result)
```

```
## [1] 0.2455057
```

Interpretation:

We see that the measure indicated that case 14 has a big influence on the fitted regression function in the range of X observations directly.

(g) Calculate Cook's distance D; for each case and prepare an index plot. Are any cases influential according to this measure? (5pts)

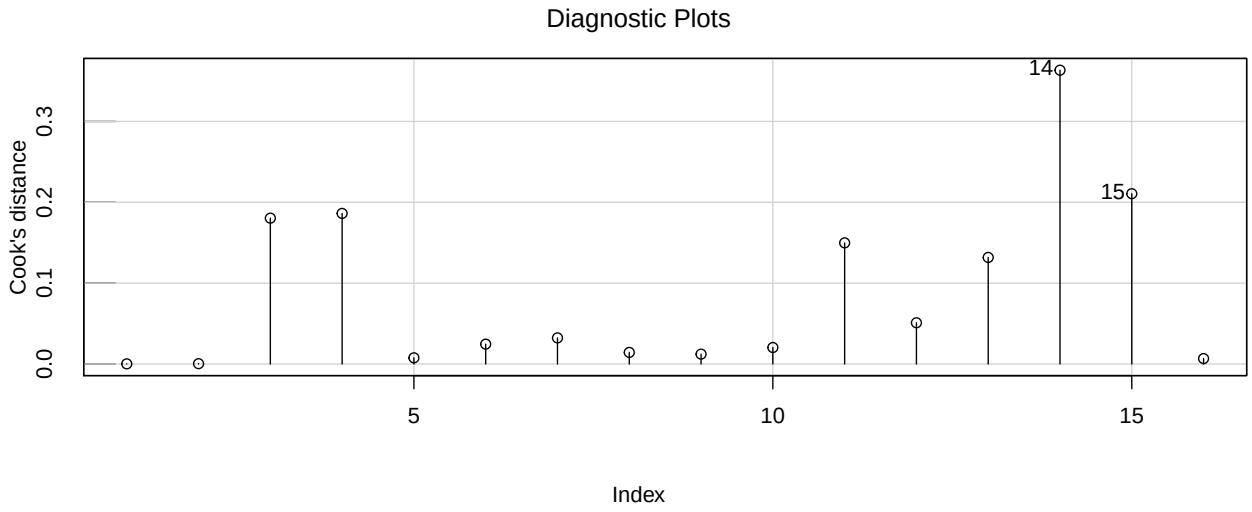
Solution:

```
influence_results$infmat[,"cook.d"]
```

```
##            1            2            3            4            5
## 0.0001877130 0.0004223542 0.1803921815 0.1862582123 0.0076655286
```

```
##            6            7            8            9           10
## 0.0245466787 0.0322971439 0.0143542862 0.0122308711 0.0204060192
##           11           12           13           14           15
## 0.1498281704 0.0509831969 0.1318214458 0.3634123447 0.2106609008
##           16
## 0.0067576676
```

```
influenceIndexPlot(lm_brand, vars=c("Cook"))
```



Interpretation:

Cases 14 (Cook's Distance Value: 0.3634123447) and 15 (Cook's Distance Value: 0.2106609008) seem to be more influential compared to the others.

(h) Find the two variance inflation factors. Why are they both equal to 1? (5pts)

Solution:

```
vif(lm_brand)
```

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

Interpretation:

$(VIF)_k = (1 - R_k^2)^{-1}$

Thus, all VIFs = 1 implies that there is no linear association between either Xs.

Question 2 Refer to the Lung pressure Data and Homework 7. The subset regression model containing first-order terms for X1 and X2 and the cross-product term X1X2 is to be evaluated in detail. (35 pts)

(a) 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? (5pts)

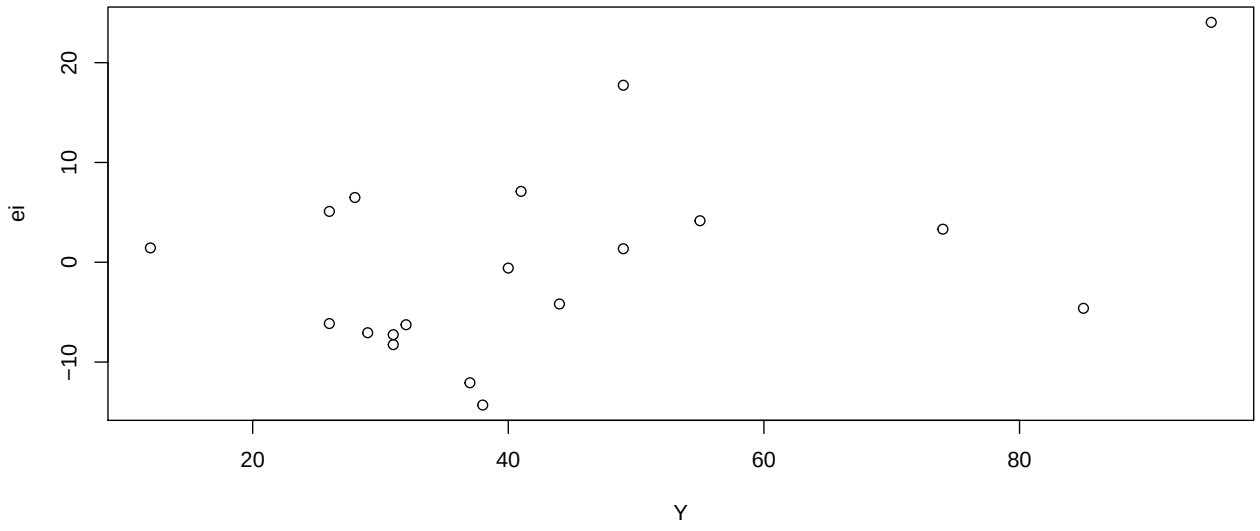
Solution:

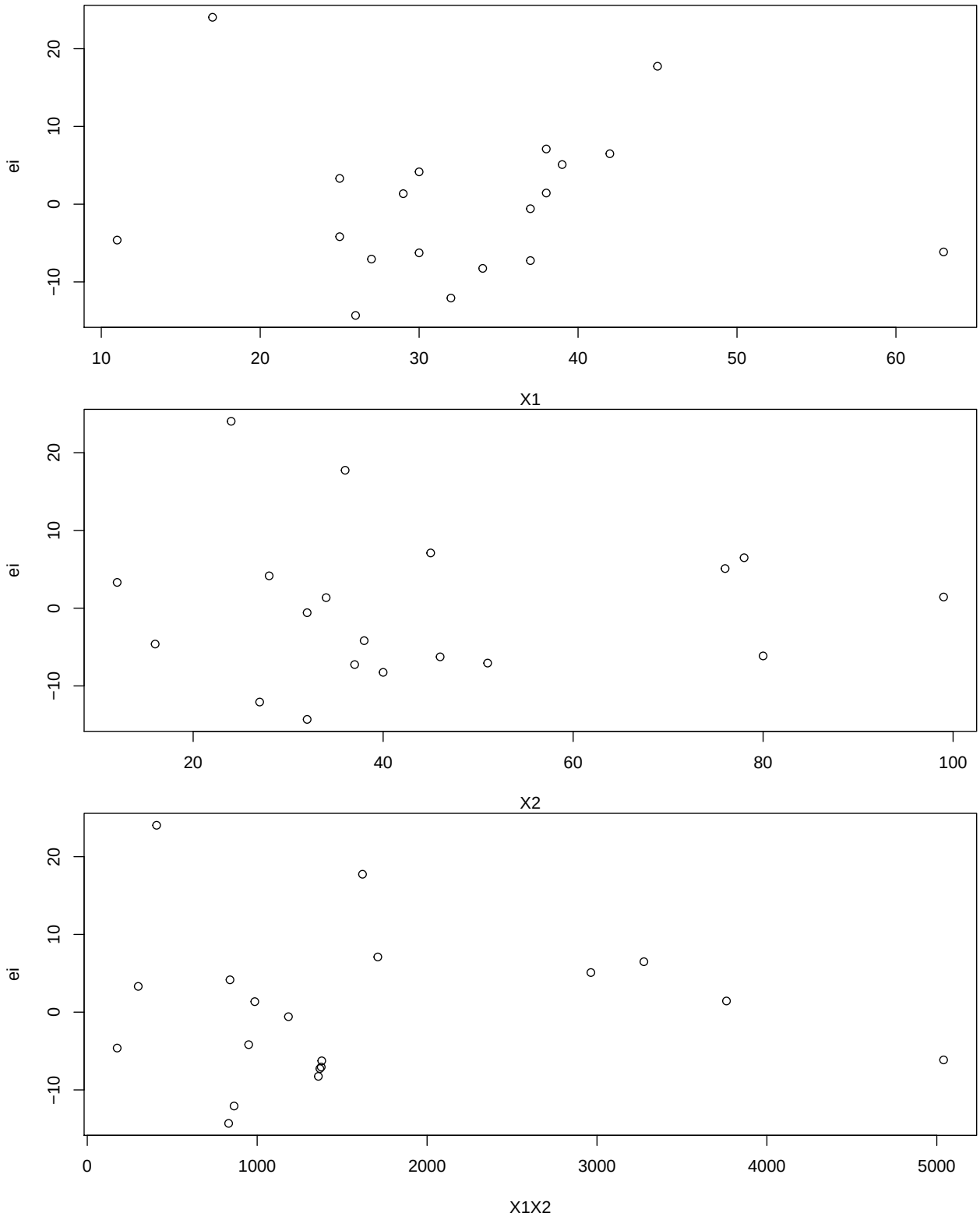
```
lung_data = read.csv("Lung Pressure.csv")
lung_data["X1X2"] = lung_data$X1*lung_data$X2
lm_lung = lm(Y~X1+X2+X1X2, data=lung_data)
summary(lm_lung)
```

```
##
## Call:
```

```
## lm(formula = Y ~ X1 + X2 + X1X2, data = lung_data)
##
## 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 ***
## X1X2          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
```

```
vars = c("Y", "X1", "X2", "X1X2")
ei = lm_lung$residuals
for(v in vars) {
  plot(lung_data[[v]], ei, xlab=v)
}
```





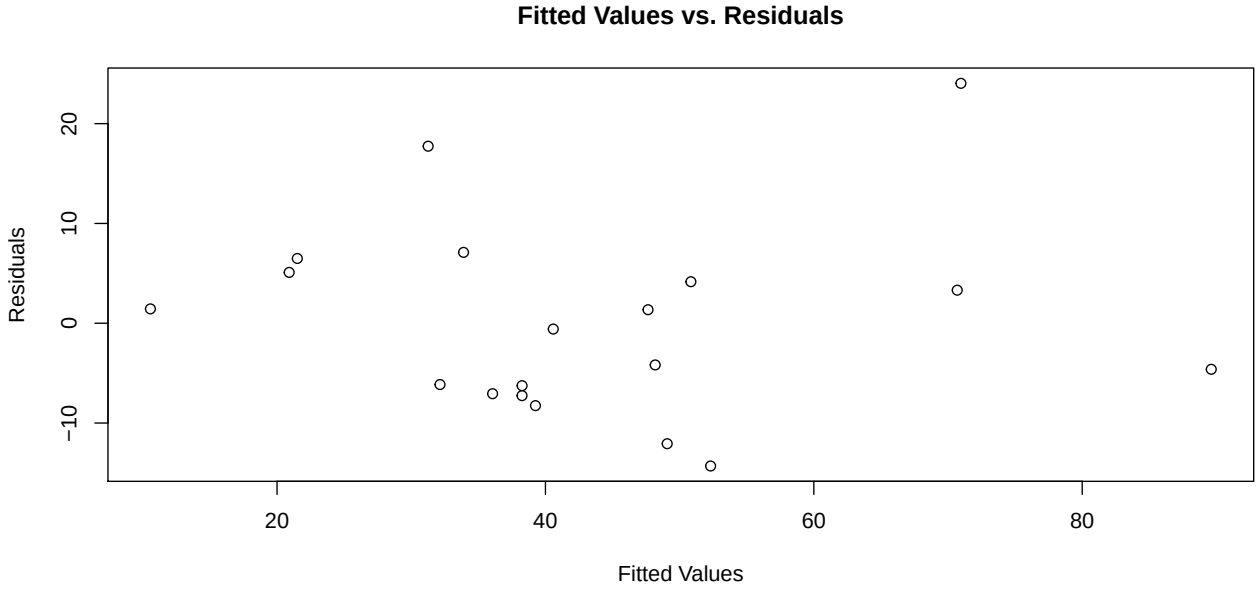
Interpretation:

We see that the plots dont have much pattern. However, we do see some outliers for X1 and the plot for the interaction term X1X2 seems to have non-constant error.

(b) 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? (5pts)

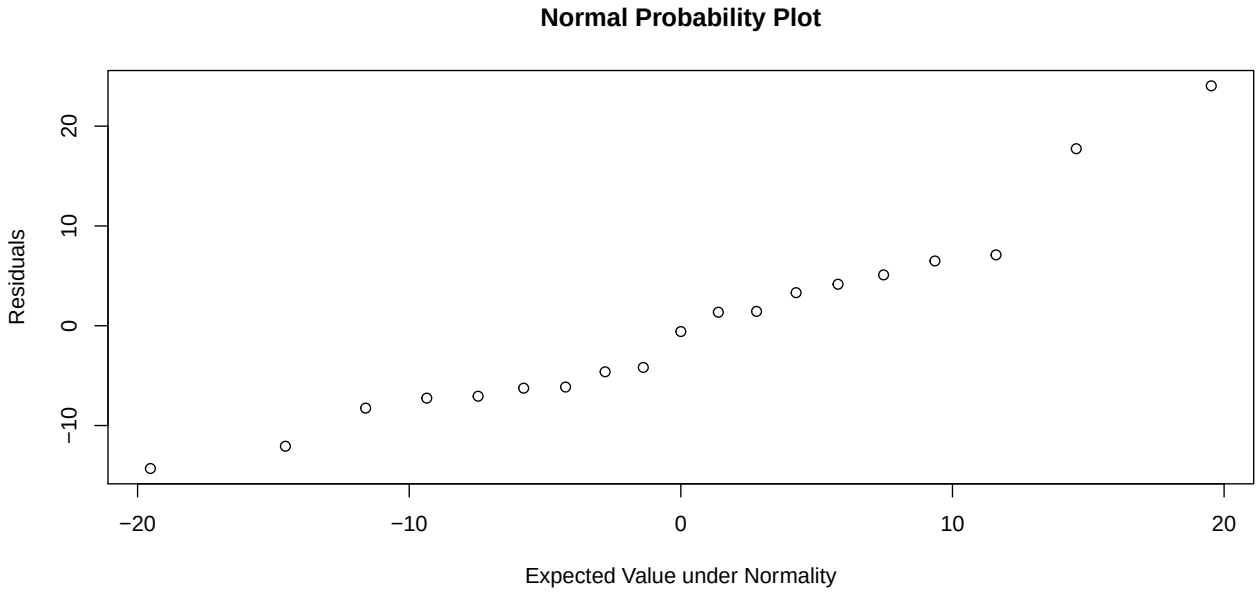
Solution:

```
build_residual_qq <- function(lm, df, rse){
  ei = lm$residuals
  fitted_values = lm$fitted.values

  par(mfrow=c(1,1))
  plot(fitted_values, ei, xlab="Fitted Values", ylab="Residuals")
  title(main="Fitted Values vs. Residuals")


  ri = rank(ei)
  n = nrow(df)
  zr = (ri-0.375)/(n+0.25)

  #residual standard error from summary(lm) above
  zr1 = rse*qnorm(zr)

  print(cor.test(zr1, ei))

  plot(zr1, ei, xlab="Expected Value under Normality",ylab="Residuals")
  title(main="Normal Probability Plot")

}

build_residual_qq(lm=lm_lung, df=lung_data, rse=10.58)
```

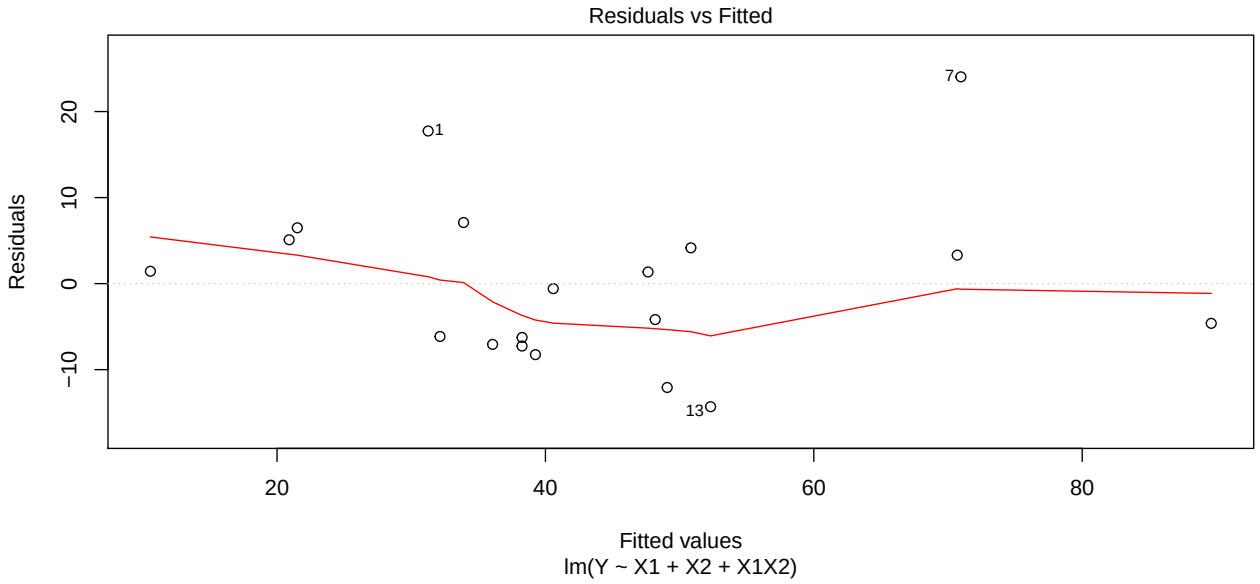


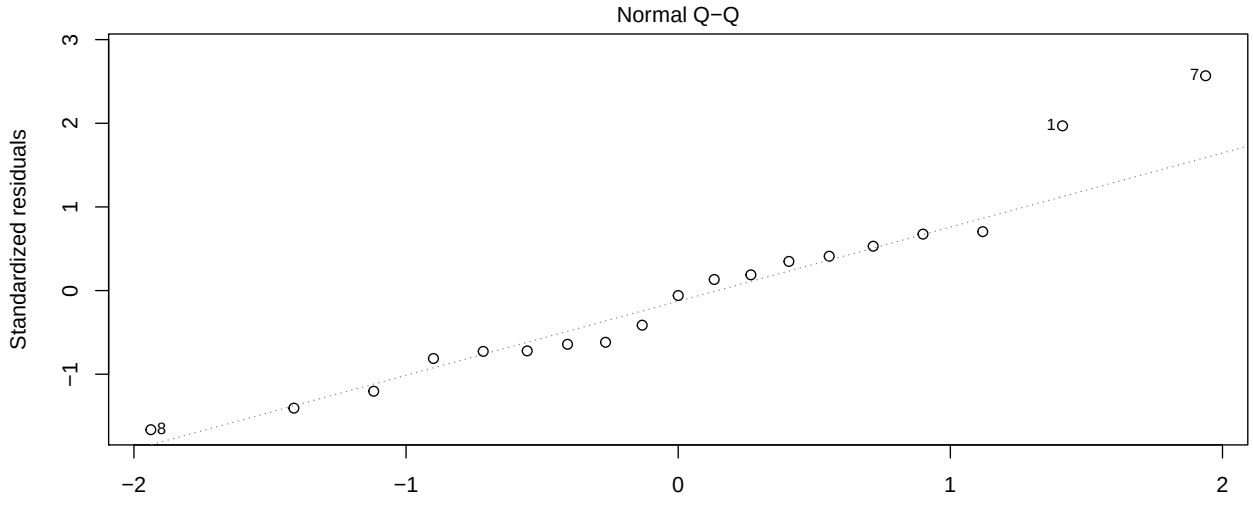

```
##
##  Pearson's product-moment correlation
##
## data:  zr1 and ei
## t = 14.813, df = 17, p-value = 3.779e-11
```

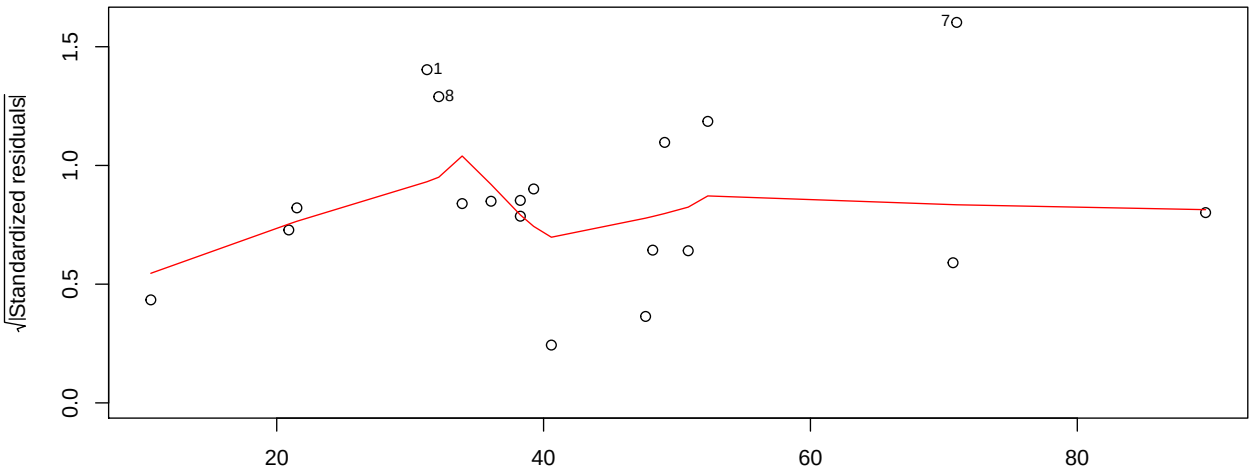
```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9053026 0.9860949
## sample estimates:
##       cor
## 0.9633751
```

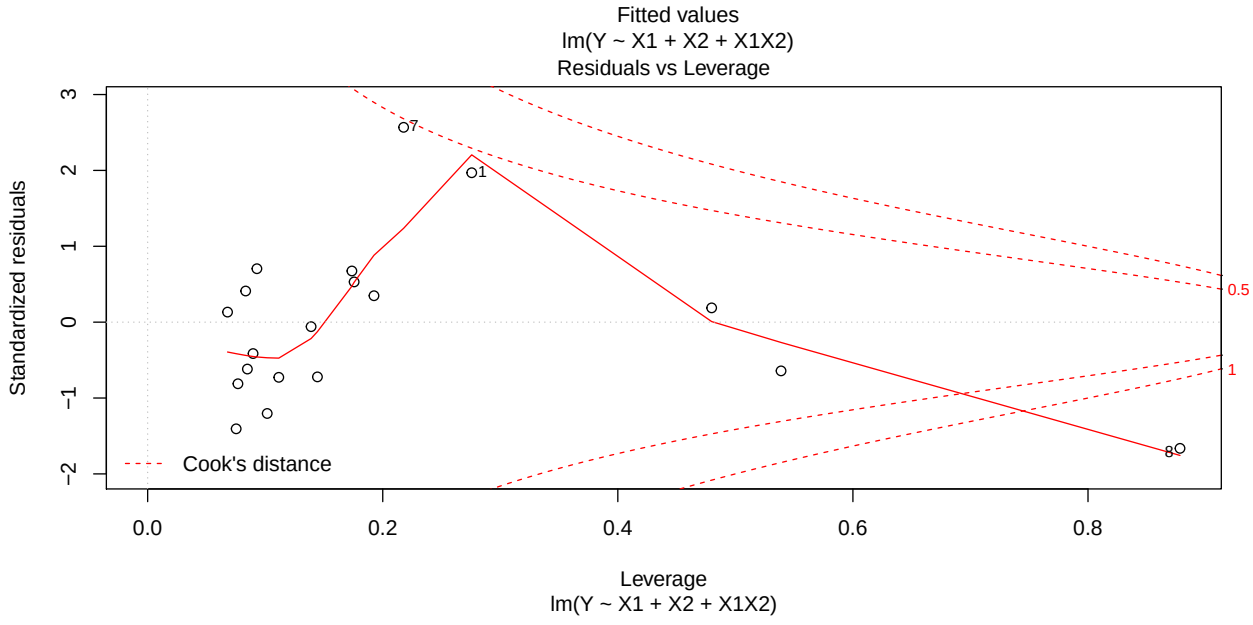



```
plot(lm_lung)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Y ~ X1 + X2 + X1X2)


Scale−Location
√|Standardized residuals|
lm(Y ~ X1 + X2 + X1X2)


Residuals vs Leverage
Standardized residuals
Cook's distance
0.5
1
Leverage
lm(Y ~ X1 + X2 + X1X2)

Interpretation:

We can see that the plot is not linear and the residuals do not conform with the assumptions of normality.

(c) Obtain the variance inflation factors. Are there any indications that serious multicollinearity problems are present? Explain. (5pts)

Solution:

```
vif(lm_lung)
```

```
##        X1        X2      X1X2
##  5.431477 11.639560 22.474469
```

Interpretation:

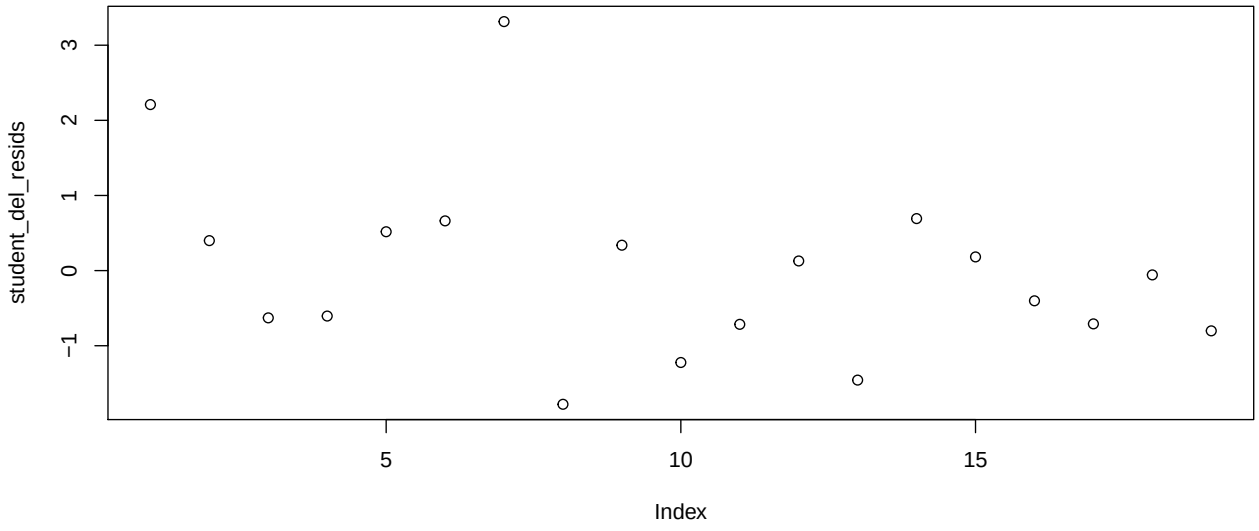
Since the variance inflation factors for all the coefficients in the model are > 1, we can say that there is serious multi-collinearity present.

(d) Obtain the studentized deleted residuals and identify outlying Y observations. Use the Bonferroni outlier test procedure with = .05. State the decision rule and conclusion. (5pts)

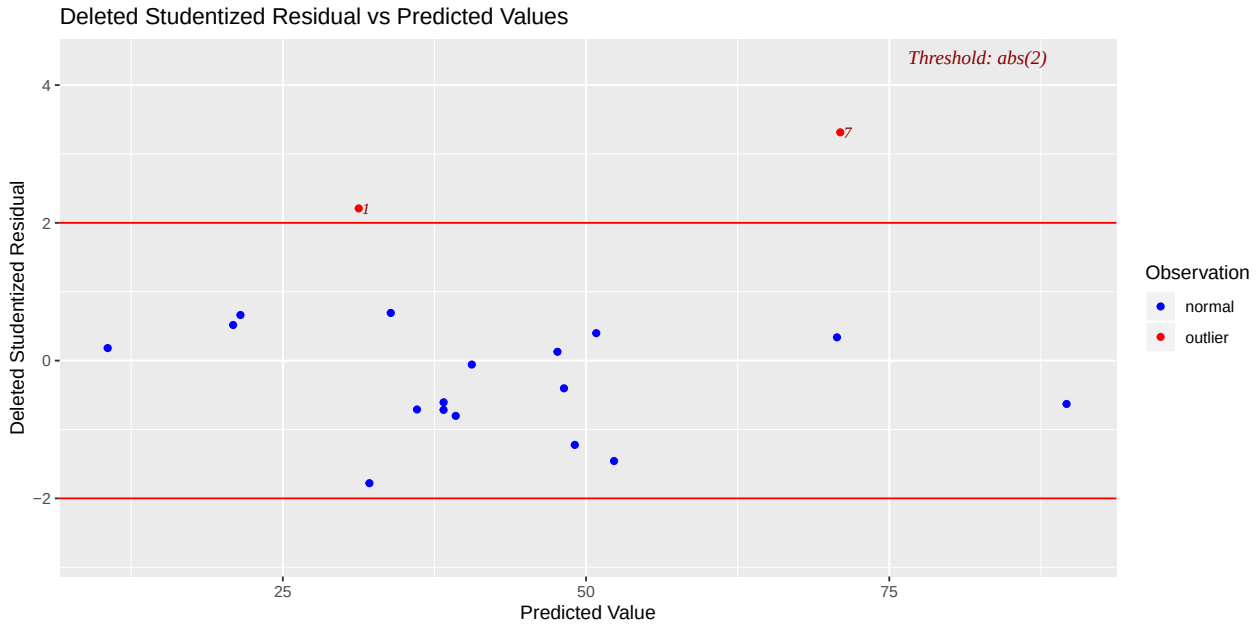
Solution:

Studentized deleted residuals:

```
student_del_resids = rstudent(lm_lung)
plot(student_del_resids)
```




```
ols_plot_resid_stud_fit(lm_lung)
```



Bonferroni Outlier Test:

Test value $= t(1-\alpha/2n; n-p-1)$

H_0 : No outlier H_1 : Atleast one outlier

```
n = nrow(lung_data)
p = length(lm_lung$coefficients)
alpha = 0.1
tTest = qt(1-alpha/(2*n),n-p-1)
tTest
```

```
## [1] 3.299917
```

```
any(abs(student_del_resids)>=abs(tTest))
```

```
## [1] TRUE
```

```
which(abs(student_del_resids)>=abs(tTest))
```

```
## 7
## 7
```

Decision Rule:

- If $|t_i| \leq t(1-\alpha/2n; n-p-1)$, $n=i$ is not an outlier.
- If $|t_i| > t(1-\alpha/2n; n-p-1)$, $n=i$ is an outlier.

Result:

We can see that $|t_i| > t(1-\alpha/2n; n-p-1)$ for case #7, we conclude H_1 for that case. Case #7 is an outlier.

(e) Obtain the diagonal elell1ents of the hat matrix. Are there any outlying X observations? Discuss. (5pts)

Solution:

```
hii = hatvalues(lm_lung)
index = hii>2*p/n
index
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12
```

```
## FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
##    13    14    15    16    17    18    19
## FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
```

Interpretation:

The hat matrix measures the distance between X_i and $\bar{X}$. We see that there are quite a few outlying X observations.

(f) 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 inf1uence. What do you conclude? (10pts)

Solution:

```
influence_results = influence.measures(lm_lung)
influence_results$infmat
```

```
##          dfb.1_        dfb.X1       dfb.X2     dfb.X1X2       dffit
## 1  -0.747205465  1.0869860015  0.223924583 -0.595508140  1.36315255
## 2   0.007215245  0.0303754712 -0.005937180 -0.020907997  0.12029768
## 3  -0.651937120  0.5919134167  0.433371759 -0.481911033 -0.68018240
## 4   0.040625335 -0.0405241241 -0.105922648  0.092872907 -0.18417401
## 5  -0.048719889 -0.0005608422  0.108881432 -0.042166912  0.23874956
## 6  -0.027643506 -0.0262819114  0.071947319  0.010397011  0.30346848
## 7   1.454130524 -1.2776085214 -0.741519685  0.847523283  1.74855091
## 8  -1.546908014  1.1866225267  3.162265299 -3.285790033 -4.77978481
## 9   0.102084510 -0.0460886172 -0.105089555  0.070106489  0.16502598
## 10  0.035880293 -0.1716977361  0.009243400  0.101652209 -0.41160962
## 11  0.108949526 -0.1719767566 -0.060836664  0.118347760 -0.25344986
## 12  0.001272909  0.0060013168  0.005761819 -0.009501834  0.03460742
## 13 -0.126037755  0.0513357150 -0.012036907  0.032298115 -0.41589916
## 14 -0.107493533  0.1446477355  0.079720416 -0.109737168  0.22154556
## 15 -0.015505900 -0.0352510604  0.077147027 -0.015699767  0.17485734
## 16 -0.022744857  0.0174212358 -0.037190574  0.028348436 -0.12621246
## 17  0.051850664 -0.0185689176 -0.192026035  0.142575494 -0.29140135
## 18  0.009124219 -0.0157090310 -0.003580971  0.009868227 -0.02302799
## 19  0.080551739 -0.1241477527 -0.082846252  0.114313382 -0.23135890
##        cov.r       cook.d        hat
## 1  0.5498713 0.3690412874 0.27569243
## 2  1.3741269 0.0038327437 0.08336965
## 3  2.5561254 0.1205155088 0.53886673
## 4  1.2987531 0.0088542919 0.08482945
## 5  1.4820600 0.0149819558 0.17565769
## 6  1.4100371 0.0239195505 0.17374756
## 7  0.1661137 0.4589170583 0.21775095
## 8  4.7895257 4.9908149785 0.87827870
## 9  1.5798700 0.0072356817 0.19254581
## 10 0.9773165 0.0409990640 0.10171037
## 11 1.2849113 0.0165996962 0.11155424
## 12 1.4072860 0.0003204295 0.06796196
## 13 0.8099602 0.0402283355 0.07530137
## 14 1.2699110 0.0127121277 0.09294148
## 15 2.5095274 0.0081704112 0.47982100
## 16 1.3826267 0.0042181308 0.08967339
## 17 1.3374694 0.0219561921 0.14443764
## 18 1.5292131 0.0001420082 0.13905081
```

```
## 19 1.1926432 0.0137076489 0.07680876
```

```
cases = c(3,7,8,15)

for(c in cases) {
  influence_results = influence.measures(lm_lung)
  print(paste("Case #", c))
  print(influence_results$infmat[c,])
}
```

```
## [1] "Case # 3"
##     dfb.1_     dfb.X1     dfb.X2   dfb.X1X2      dffit      cov.r
## -0.6519371  0.5919134  0.4333718 -0.4819110 -0.6801824  2.5561254
##     cook.d        hat
##  0.1205155  0.5388667
## [1] "Case # 7"
##     dfb.1_     dfb.X1     dfb.X2   dfb.X1X2      dffit      cov.r
##  1.4541305 -1.2776085 -0.7415197  0.8475233  1.7485509  0.1661137
##     cook.d        hat
##  0.4589171  0.2177509
## [1] "Case # 8"
##     dfb.1_     dfb.X1     dfb.X2   dfb.X1X2      dffit      cov.r
## -1.5469080  1.1866225  3.1622653 -3.2857900 -4.7797848  4.7895257
##     cook.d        hat
##  4.9908150  0.8782787
## [1] "Case # 15"
##       dfb.1_       dfb.X1       dfb.X2     dfb.X1X2        dffit
## -0.015505900 -0.035251060  0.077147027 -0.015699767  0.174857335
##        cov.r       cook.d          hat
##  2.509527439  0.008170411  0.479820998
```

Interpretation:

Case #3:

|DFFITS|<1 implies that this case does not have significant influence on Y_3.

|DFBETAS|<1 which implies this case does not have significant influence on the coefficients.

Cook Distance: According to Cook's Distance, this case has little influence on all the fitted values, since $0.1 < P(F(p, n-p) \leq Cook'sDistance) < 0.5$.

Case #7:

|DFFITS|>1 implies that this case does have significant influence on Y_7.

|DFBETAS|>1 for β_1 only and remaining |DFBETAS|<1. Which implies this case only has significant influence on β_1 the coefficients.

Cook Distance: According to Cook's Distance, this case has little influence on all the fitted values, since $0.1 < P(F(p, n-p) \leq Cook'sDistance) < 0.5$.

Case #8:

|DFFITS|>1 implies that this case does have significant influence on Y_8.

|DFBETAS|>1 for all variables. Which implies this case has significant influence on all the coefficients.

Cook Distance: According to Cook's Distance, this case has major influence on all the fitted values, since $P(F(p, n-p) \leq Cook'sDistance) > 0.5$.

Case #15:

|DFFITS|<1 implies that this case does not have significant influence on Y_{15}.

|DFBETAS|<1 for all variables. Which implies this case does not have significant influence on any of the coefficients.

Cook Distance: According to Cook's Distance, this case has no influence on all the fitted values, since $P(F(p, n-p) \leq Cook's Distance) < 0.1$.

Summary: - Case #8 seems to be the biggest outlier compared to other cases seen above. - Case #7 has an impact only on β_1, which means that only X_7 might the outlier here. - Case #3 and #15 don't seem to have much influence on the model and should not be consedered as outliers.

Question 3 Refer to the Prostate Cancer data set in Appendix C.6 and Homework 7. For the best subset model developed in Homework 7, perform appropriate diagnostic checks to evaluate outliers and assess their influence. Do any serious multicollinearity problems exist here? (20pts)

Solution:

```
prostate_data = read.csv("Prostate Cancer.csv")
prostate_data$Seminal.vesicle.invasion = as.factor(prostate_data$Seminal.vesicle.invasion)
prostate_data$Gleason.score = as.factor(prostate_data$Gleason.score)
summary(prostate_data)
```

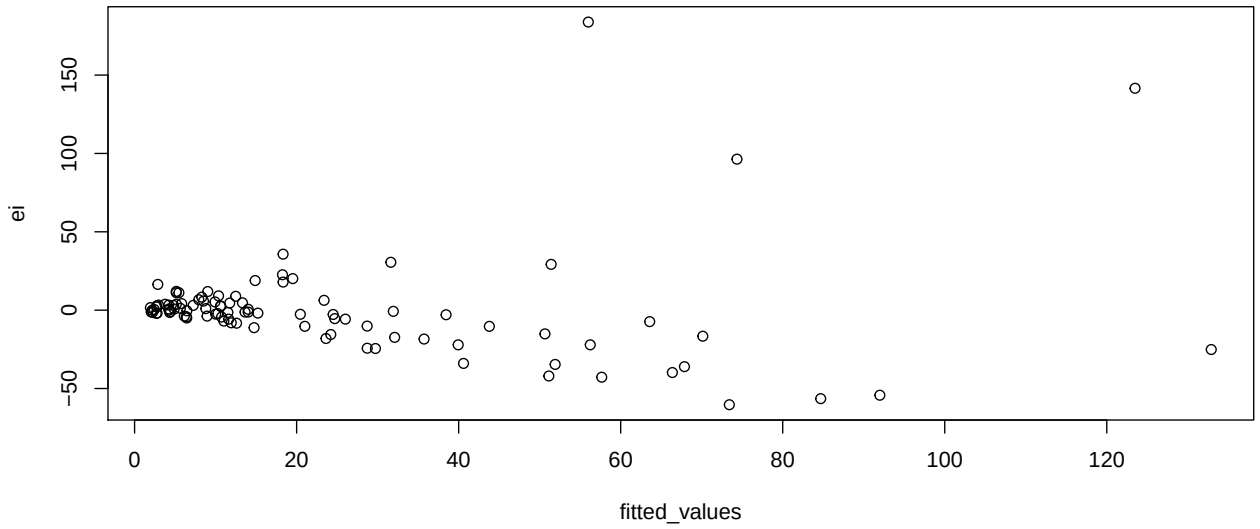
```
##    PSA.level       Cancer.volume         Weight            Age
##  Min.   :  0.651   Min.   : 0.2592   Min.   : 10.70   Min.   :41.00
##  1st Qu.:  5.641   1st Qu.: 1.6653   1st Qu.: 29.37   1st Qu.:60.00
##  Median : 13.330   Median : 4.2631   Median : 37.34   Median :65.00
##  Mean   : 23.730   Mean   : 6.9987   Mean   : 45.49   Mean   :63.87
##  3rd Qu.: 21.328   3rd Qu.: 8.4149   3rd Qu.: 48.42   3rd Qu.:68.00
##  Max.   :265.072   Max.   :45.6042   Max.   :450.34   Max.   :79.00
##  Benign.prostatic.hyperplasia Seminal.vesicle.invasion
##  Min.   : 0.000               0:76
##  1st Qu.: 0.000               1:21
##  Median : 1.350
##  Mean   : 2.535
##  3rd Qu.: 4.759
##  Max.   :10.278
##  Capsular.penetration Gleason.score
##  Min.   : 0.0000      6:33
##  1st Qu.: 0.0000      7:43
##  Median : 0.4493      8:21
##  Mean   : 2.2454
##  3rd Qu.: 3.2544
##  Max.   :18.1741
```

```
lm_prostate_best = lm(PSA.level~Cancer.volume+Capsular.penetration, data=prostate_data)
summary(lm_prostate_best)
```

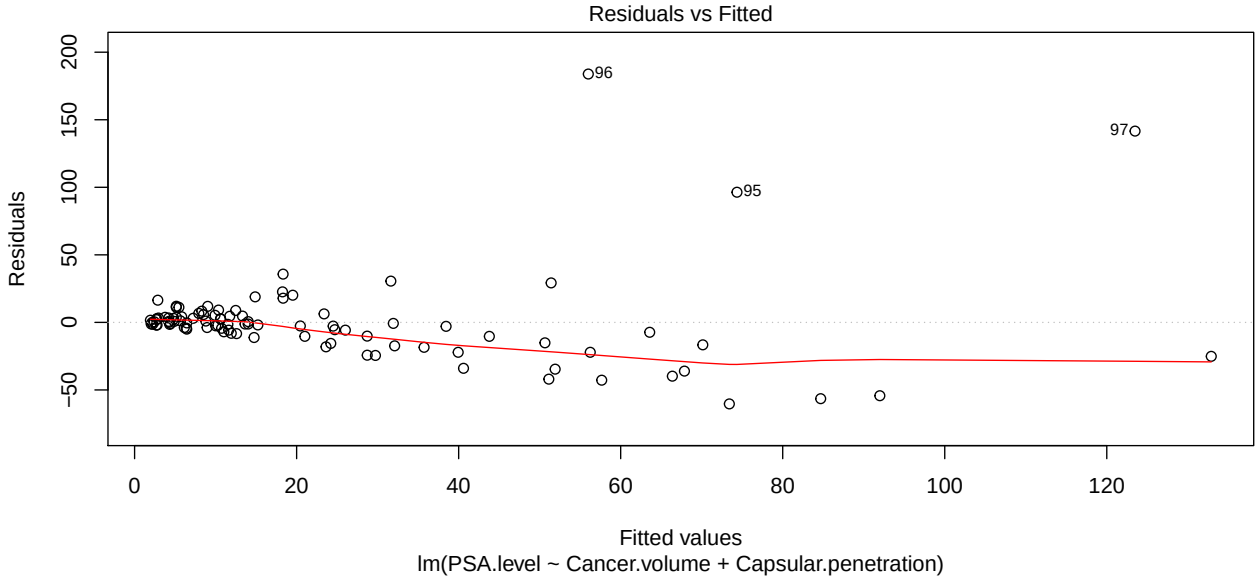
```
##
## Call:
## lm(formula = PSA.level ~ Cancer.volume + Capsular.penetration,
##     data = prostate_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.346  -8.324  -1.205   4.159 183.843
##
## Coefficients:
```

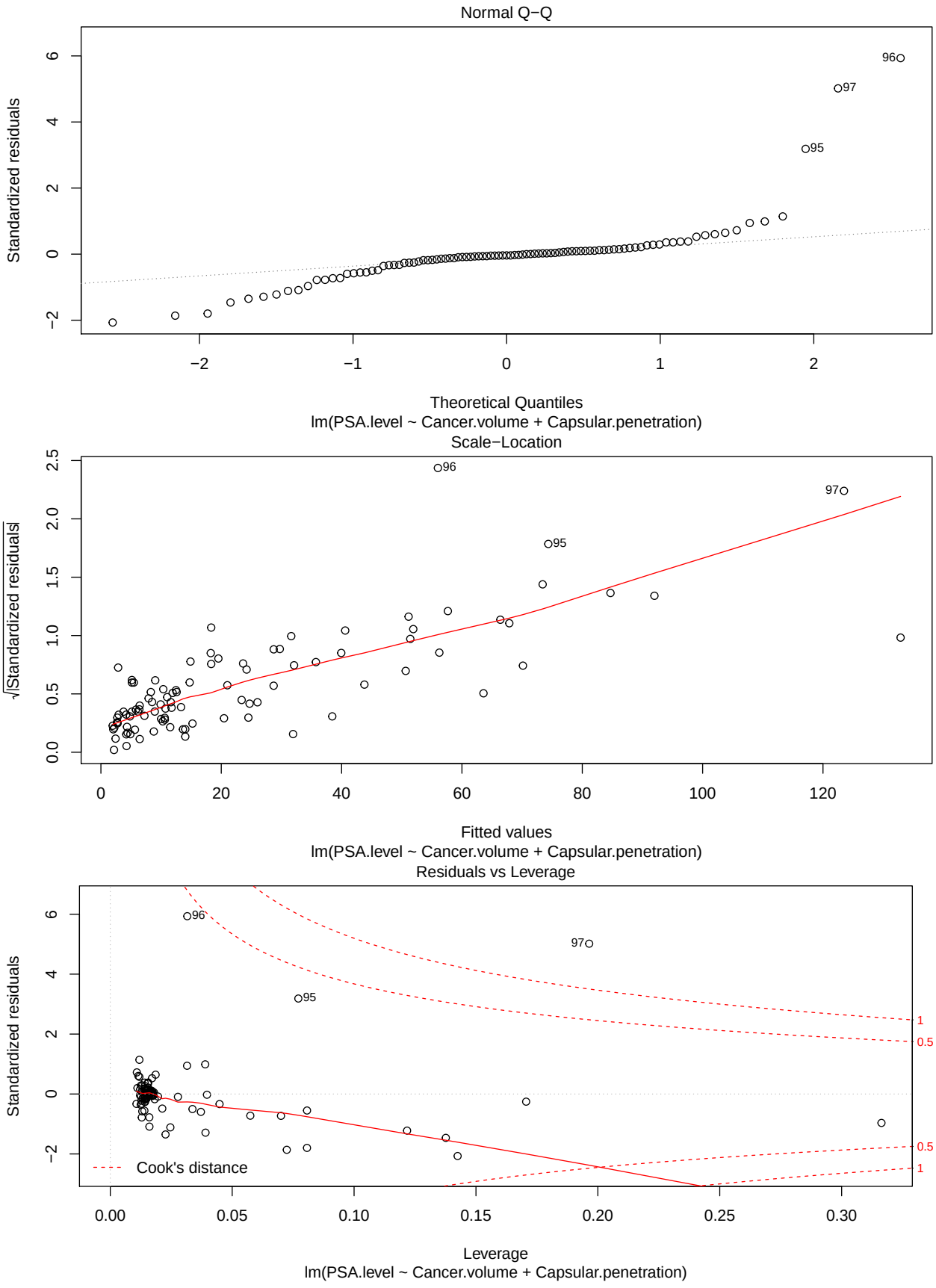
```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.3276     4.2861   0.310    0.757
## Cancer.volume           2.4139     0.5655   4.269 4.69e-05 ***
## Capsular.penetration    2.4533     1.1779   2.083    0.040 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.48 on 94 degrees of freedom
## Multiple R-squared:  0.4165, Adjusted R-squared:  0.4041
## F-statistic: 33.55 on 2 and 94 DF,  p-value: 1.01e-11
```

```
ei = lm_prostate_best$residuals
fitted_values = lm_prostate_best$fitted.values
plot(fitted_values, ei)
```




```
plot(lm_prostate_best)
```



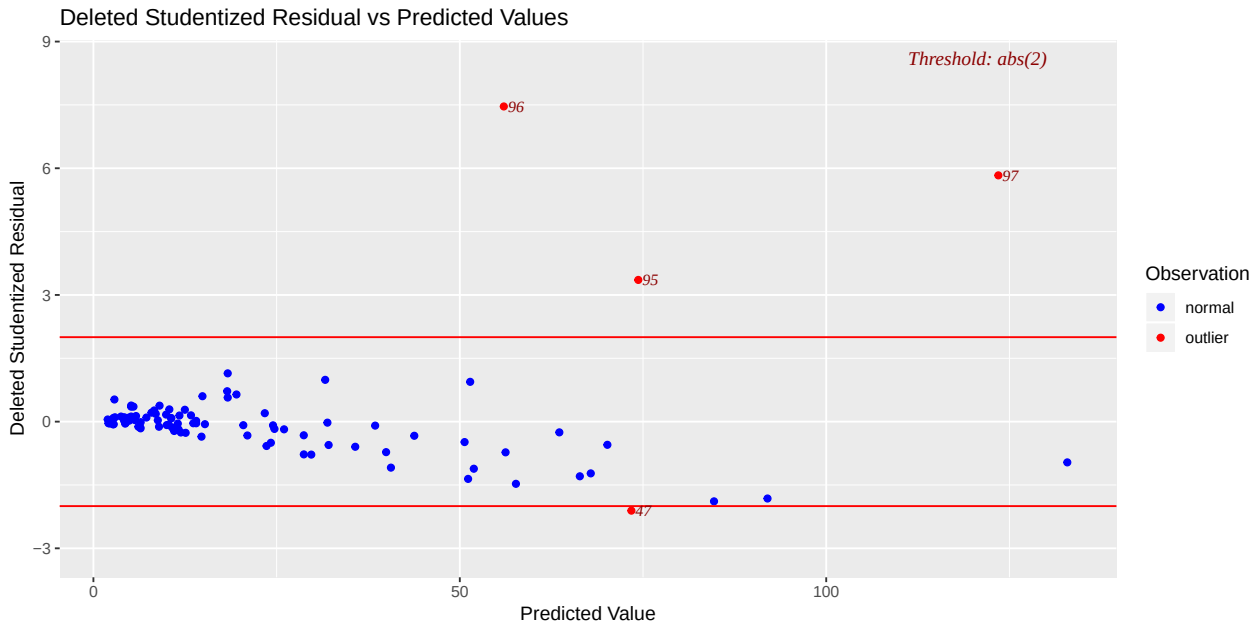

Normal Q−Q
Standardized residuals
96
97
95
−2
−1
0
1
2
Theoretical Quantiles
lm(PSA.level ~ Cancer.volume + Capsular.penetration)
Scale−Location
√|Standardized residuals|
96
97
95
0.0
0.5
1.0
1.5
2.0
2.5
0
20
40
60
80
100
120
Fitted values
lm(PSA.level ~ Cancer.volume + Capsular.penetration)
Residuals vs Leverage
Standardized residuals
96
97
95
1
0.5
0.5
1
Cook's distance
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Leverage
lm(PSA.level ~ Cancer.volume + Capsular.penetration)

Interpretation:

Residual Plot: We can see that the residuals are clustered near the lower values of the fitted Y's. We can also see a few outliers (like case #95, 96, 97).

Normal QQ plot: The plot seems to be linear at the center, however, it deviates from linearity at the tails. Thus, it does not conform with the assumptions of normality.

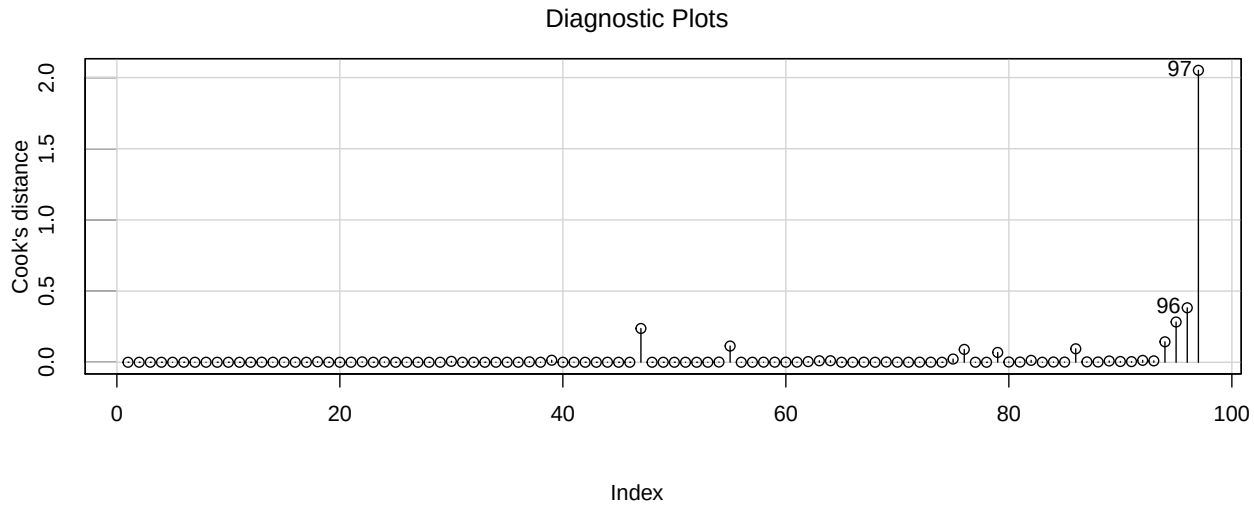
```
ols_plot_resid_stud_fit(lm_prostate_best)
```



Interpretation:

Case #95, 96 and 97 again show up as outliers according to the Studentized Deleted Residuals, along with case #47.

```
influenceIndexPlot(lm_prostate_best, vars=c("Cook"))
```



Cook's Distance also spots the same cases for the outliers as above (95,96,97 and possibly 47), with case #97 being the biggest influencer compared to the other cases.

Bonferroni Outlier Test:

Test value $= t(1 - \alpha/2n; n - p - 1)$

```
n = nrow(prostate_data)
p = length(lm_prostate_best$coefficients)
alpha = 0.05
tTest = qt(1-alpha/(2*n),n-p-1)
tTest
```

```
## [1] 3.598447
```

```
student_del_resids = rstudent(lm_prostate_best)
any(abs(student_del_resids)>=abs(tTest))
```

```
## [1] TRUE
```

```
which(abs(student_del_resids)>=abs(tTest))
```

```
## 96 97
## 96 97
```

Interpretation:

According to the Bonferroni test, we get case # 96 and 97 as the outliers at $\alpha = 0.05$.

```
hii = hatvalues(lm_prostate_best)
index = hii>2*p/n
which(index==TRUE)
```

```
## 47 55 76 79 82 86 89 91 94 95 97
## 47 55 76 79 82 86 89 91 94 95 97
```

Interpretation:

According to just the X values we get the above indices as the outliers. Interestingly, it does not contain case #96.

```
vif(lm_prostate_best)
```

```
##        Cancer.volume Capsular.penetration
##             1.923468             1.923468
```

```
cor(prostate_data[,c("Cancer.volume", "Capsular.penetration")])
```

```
##                      Cancer.volume Capsular.penetration
## Cancer.volume            1.0000000            0.6928967
## Capsular.penetration     0.6928967            1.0000000
```

Interpretation:

Based on the VIF of the model in cosideration and the correlation matrix above, we see that there exists multi-collinearity in the dataset.

```
cases = c(47,94,95,96,97)
influence_results = influence.measures(lm_prostate_best)
influence_results$infmat[cases,]
```

```
##         dfb.1_   dfb.Cnc.   dfb.Cps.      dffit     cov.r    cook.d
## 47 -0.02674524  0.3703131 -0.7900548 -0.8590224 1.0468984 0.2372859
## 94  0.30635415 -0.6121617  0.2763227 -0.6563720 1.4659340 0.1437134
## 95 -0.06069678 -0.1439826  0.7424481  0.9700406 0.7941423 0.2827951
## 96 -0.13991634  0.9791187 -0.3076477  1.3476189 0.2608362 0.3826700
## 97 -0.87553836  0.2509163  1.8417311  2.8833207 0.5045327 2.0510035
```

```
##           hat
## 47 0.14247293
## 94 0.31633848
## 95 0.07712797
## 96 0.03157409
## 97 0.19644430
```

Interpretation

- Looking at the influence measures (DFFITS, Cook's distances, DFBETAS) above, case #96 and 97 have higher influence that the other cases.
- Looking at all the diagnostics so far, we can conclude the case #96 and 97 have greater influence on the overall model compared to the other cases that have shown up as outliers.