An example from the Internet: Credit to Math.Dartmouth.Edu

Math 50 Fall 2017, Homework #7

An example from Regression Diagnostics:

Identifying Influential Data and Sources of Collinearity (Belsley, Kuh and Welsch)

[,1] sr numeric aggregate personal savings [,2] pop15 numeric % of population under 15 [,3] pop75 numeric % of population over 75 [,4] dpi numeric real per-capita disposable income [,5] ddpi numeric % growth rate of dpi

```
data(LifeCycleSavings)  #data set available in R
lm.SR <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data = LifeCycleSavings)
summary(inflm.SR <- influence.measures(lm.SR))
```

```
## Potentially influential observations of
##   lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = LifeCycleSavings) :
##
##                dfb.1_ dfb.pp15 dfb.pp75 dfb.dpi dfb.ddpi dffit   cov.r
## Chile          -0.20   0.13     0.22    -0.02    0.12    -0.46    0.65_*
## United States  0.07  -0.07     0.04    -0.23   -0.03    -0.25    1.66_*
## Zambia         0.16  -0.08    -0.34     0.09    0.23     0.75    0.51_*
## Libya          0.55  -0.48    -0.38    -0.02   -1.02_*  -1.16_*  2.09_*
##                cook.d hat
## Chile          0.04   0.04
## United States  0.01   0.33_*
## Zambia         0.10   0.06
## Libya          0.27   0.53_*
```

```
inflm.SR
```

```
## Influence measures of
##   lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = LifeCycleSavings) :
##
##                  dfb.1_ dfb.pp15 dfb.pp75  dfb.dpi  dfb.ddpi   dffit cov.r
## Australia       0.01232 -0.01044 -0.02653  0.04534 -0.000159  0.0627 1.193
## Austria        -0.01005  0.00594  0.04084 -0.03672 -0.008182  0.0632 1.268
## Belgium        -0.06416  0.05150  0.12070 -0.03472 -0.007265  0.1878 1.176
## Bolivia         0.00578 -0.01270 -0.02253  0.03185  0.040642 -0.0597 1.224
## Brazil          0.08973 -0.06163 -0.17907  0.11997  0.068457  0.2646 1.082
## Canada          0.00541 -0.00675  0.01021 -0.03531 -0.002649 -0.0390 1.328
## Chile          -0.19941  0.13265  0.21979 -0.01998  0.120007 -0.4554 0.655
## China           0.02112 -0.00573 -0.08311  0.05180  0.110627  0.2008 1.150
## Colombia        0.03910 -0.05226 -0.02464  0.00168  0.009084 -0.0960 1.167
## Costa Rica     -0.23367  0.28428  0.14243  0.05638 -0.032824  0.4049 0.968
## Denmark        -0.04051  0.02093  0.04653  0.15220  0.048854  0.3845 0.934
## Ecuador         0.07176 -0.09524 -0.06067  0.01950  0.047786 -0.1695 1.139
## Finland        -0.11350  0.11133  0.11695 -0.04364 -0.017132 -0.1464 1.203
## France         -0.16600  0.14705  0.21900 -0.02942  0.023952  0.2765 1.226
## Germany        -0.00802  0.00822  0.00835 -0.00697 -0.000293 -0.0152 1.226
## Greece         -0.14820  0.16394  0.02861  0.15713 -0.059599 -0.2811 1.140
## Guatamala       0.01552 -0.05485  0.00614  0.00585  0.097217 -0.2305 1.085
## Honduras       -0.00226  0.00984 -0.01020  0.00812 -0.001887  0.0482 1.186
## Iceland         0.24789 -0.27355 -0.23265 -0.12555  0.184698 -0.4768 0.866
## India           0.02105 -0.01577 -0.01439 -0.01374 -0.018958  0.0381 1.202
## Ireland        -0.31001  0.29624  0.48156 -0.25733 -0.093317  0.5216 1.268
## Italy           0.06619 -0.07097  0.00307 -0.06999 -0.028648  0.1388 1.162
## Japan           0.63987 -0.65614 -0.67390  0.14610  0.388603  0.8597 1.085
## Korea          -0.16897  0.13509  0.21895  0.00511 -0.169492 -0.4303 0.870
## Luxembourg     -0.06827  0.06888  0.04380 -0.02797  0.049134 -0.1401 1.196
## Malta           0.03652 -0.04876  0.00791 -0.08659  0.153014  0.2386 1.128
## Norway          0.00222 -0.00035 -0.00611 -0.01594 -0.001462 -0.0522 1.168
## Netherlands     0.01395 -0.01674 -0.01186  0.00433  0.022591  0.0366 1.229
## New Zealand    -0.06002  0.06510  0.09412 -0.02638 -0.064740  0.1469 1.134
## Nicaragua      -0.01209  0.01790  0.00972 -0.00474 -0.010467  0.0397 1.174
## Panama          0.02828 -0.05334  0.01446 -0.03467 -0.007889 -0.1775 1.067
## Paraguay       -0.23227  0.16416  0.15826  0.14361  0.270478 -0.4655 0.873
## Peru           -0.07182  0.14669  0.09148 -0.08585 -0.287184  0.4811 0.831
## Philippines    -0.15707  0.22681  0.15743 -0.11140 -0.170674  0.4884 0.818
## Portugal       -0.02140  0.02551 -0.00380  0.03991 -0.028011 -0.0690 1.233
## South Africa    0.02218 -0.02030 -0.00672 -0.02049 -0.016326  0.0343 1.195
```

```
## South Rhodesia  0.14390 -0.13472 -0.09245 -0.06956 -0.057920  0.1607 1.313
## Spain          -0.03035  0.03131  0.00394  0.03512  0.005340 -0.0526 1.208
## Sweden          0.10098 -0.08162 -0.06166 -0.25528 -0.013316 -0.4526 1.086
## Switzerland     0.04323 -0.04649 -0.04364  0.09093 -0.018828  0.1903 1.147
## Turkey         -0.01092 -0.01198  0.02645  0.00161  0.025138 -0.1445 1.100
## Tunisia         0.07377 -0.10500 -0.07727  0.04439  0.103058 -0.2177 1.131
## United Kingdom  0.04671 -0.03584 -0.17129  0.12554  0.100314 -0.2722 1.189
## United States   0.06910 -0.07289  0.03745 -0.23312 -0.032729 -0.2510 1.655
## Venezuela      -0.05083  0.10080 -0.03366  0.11366 -0.124486  0.3071 1.095
## Zambia          0.16361 -0.07917 -0.33899  0.09406  0.228232  0.7482 0.512
## Jamaica         0.10958 -0.10022 -0.05722 -0.00703 -0.295461 -0.3456 1.200
## Uruguay        -0.13403  0.12880  0.02953  0.13132  0.099591 -0.2051 1.187
## Libya           0.55074 -0.48324 -0.37974 -0.01937 -1.024477 -1.1601 2.091
## Malaysia        0.03684 -0.06113  0.03235 -0.04956 -0.072294 -0.2126 1.113
##                  cook.d    hat inf
## Australia      8.04e-04 0.0677
## Austria        8.18e-04 0.1204
## Belgium        7.15e-03 0.0875
## Bolivia        7.28e-04 0.0895
## Brazil         1.40e-02 0.0696
## Canada         3.11e-04 0.1584
## Chile          3.78e-02 0.0373   *
## China          8.16e-03 0.0780
## Colombia       1.88e-03 0.0573
## Costa Rica     3.21e-02 0.0755
## Denmark        2.88e-02 0.0627
## Ecuador        5.82e-03 0.0637
## Finland        4.36e-03 0.0920
## France         1.55e-02 0.1362
## Germany        4.74e-05 0.0874
## Greece         1.59e-02 0.0966
## Guatamala      1.07e-02 0.0605
## Honduras       4.74e-04 0.0601
## Iceland        4.35e-02 0.0705
## India          2.97e-04 0.0715
## Ireland        5.44e-02 0.2122
## Italy          3.92e-03 0.0665
## Japan          1.43e-01 0.2233
## Korea          3.56e-02 0.0608
## Luxembourg     3.99e-03 0.0863
## Malta          1.15e-02 0.0794
```

```
## Norway         5.56e-04 0.0479
## Netherlands    2.74e-04 0.0906
## New Zealand    4.38e-03 0.0542
## Nicaragua      3.23e-04 0.0504
## Panama         6.33e-03 0.0390
## Paraguay       4.16e-02 0.0694
## Peru           4.40e-02 0.0650
## Philippines    4.52e-02 0.0643
## Portugal       9.73e-04 0.0971
## South Africa   2.41e-04 0.0651
## South Rhodesia 5.27e-03 0.1608
## Spain          5.66e-04 0.0773
## Sweden         4.06e-02 0.1240
## Switzerland    7.33e-03 0.0736
## Turkey         4.22e-03 0.0396
## Tunisia        9.56e-03 0.0746
## United Kingdom 1.50e-02 0.1165
## United States  1.28e-02 0.3337   *
## Venezuela      1.89e-02 0.0863
## Zambia         9.66e-02 0.0643   *
## Jamaica        2.40e-02 0.1408
## Uruguay        8.53e-03 0.0979
## Libya          2.68e-01 0.5315   *
## Malaysia       9.11e-03 0.0652
```

```
which(apply(inflm.SR$is.inf, 1, any))
```

```
##         Chile United States        Zambia         Libya
##             7            44            46            49
```

```
rstandard(lm.SR)
```

```
##      Australia        Austria        Belgium        Bolivia         Brazil 
##     0.23520105     0.17282943     0.61085760    -0.19245030     0.96858807 
##         Canada          Chile          China       Colombia     Costa Rica 
##    -0.09083873    -2.20907436     0.69453131    -0.39319153     1.40168682 
##        Denmark        Ecuador        Finland         France        Germany 
##     1.46686216    -0.65379142    -0.46394723     0.70042898    -0.04974135 
##         Greece      Guatamala       Honduras        Iceland          India 
##    -0.86217889    -0.91031261     0.19259259    -1.69401854     0.13881900 
##        Ireland          Italy          Japan          Korea     Luxembourg 
##     1.00475012     0.52442520     1.57595468    -1.65713877    -0.45967116 
##          Malta         Norway    Netherlands    New Zealand      Nicaragua 
##     0.81536209    -0.23495632     0.11735008     0.61802723     0.17443311 
##         Panama       Paraguay           Peru    Philippines       Portugal 
##    -0.88366877    -1.66987256     1.77851567     1.81461452    -0.21267488 
##   South Africa South Rhodesia          Spain         Sweden    Switzerland 
##     0.13140922     0.37072635    -0.18374340    -1.19700295     0.67944806 
##         Turkey        Tunisia United Kingdom  United States      Venezuela 
##    -0.71532499    -0.77031393    -0.75327449    -0.35811077     0.99934066 
##         Zambia        Jamaica        Uruguay          Libya       Malaysia 
##     2.65091534    -0.85634746    -0.62681420    -1.08705199    -0.80805950
```

```
rstudent(lm.SR)
```

```
##      Australia        Austria        Belgium        Bolivia         Brazil 
##     0.23271611     0.17095506     0.60655220    -0.19037831     0.96790816 
##         Canada          Chile          China       Colombia     Costa Rica 
##    -0.08983197    -2.31342946     0.69048169    -0.38946778     1.41731062 
##        Denmark        Ecuador        Finland         France        Germany 
##     1.48644473    -0.64957871    -0.45986445     0.69640933    -0.04918692 
##         Greece      Guatamala       Honduras        Iceland          India 
##    -0.85967533    -0.90854545     0.19051919    -1.73119989     0.13729730 
##        Ireland          Italy          Japan          Korea     Luxembourg 
##     1.00485886     0.52015744     1.60321582    -1.69103214    -0.45560591 
##          Malta         Norway    Netherlands    New Zealand      Nicaragua 
##     0.81227407    -0.23247367     0.11605663     0.61373189     0.17254242 
##         Panama       Paraguay           Peru    Philippines       Portugal 
##    -0.88147653    -1.70488128     1.82391409     1.86382587    -0.21040432 
##   South Africa South Rhodesia          Spain         Sweden    Switzerland 
##     0.12996586     0.36714512    -0.18175853    -1.20293404     0.67532922 
##         Turkey        Tunisia United Kingdom  United States      Venezuela 
##    -0.71138840    -0.76677907    -0.74959873    -0.35461507     0.99932569 
##         Zambia        Jamaica        Uruguay          Libya       Malaysia 
##     2.85355834    -0.85376418    -0.62253411    -1.08930326    -0.80489153
```

```
# dfbetas(lm.SR)
dffits(lm.SR)
```

```
##      Australia        Austria        Belgium        Bolivia         Brazil 
##     0.06271756     0.06324405     0.18780542    -0.05967770     0.26464755 
##         Canada          Chile          China       Colombia     Costa Rica 
##    -0.03897262    -0.45535788     0.20077524    -0.09602160     0.40493458 
##        Denmark        Ecuador        Finland         France        Germany 
##     0.38451126    -0.16946909    -0.14641688     0.27653834    -0.01521770 
##         Greece      Guatamala       Honduras        Iceland          India 
##    -0.28114772    -0.23053977     0.04816829    -0.47676403     0.03808618 
##        Ireland          Italy          Japan          Korea     Luxembourg 
##     0.52157524     0.13884474     0.85965081    -0.43025048    -0.14006342 
##          Malta         Norway    Netherlands    New Zealand      Nicaragua 
##     0.23855360    -0.05216187     0.03663477     0.14694487     0.03972980 
##         Panama       Paraguay           Peru    Philippines       Portugal 
##    -0.17751461    -0.46547654     0.48109398     0.48840149    -0.06901872 
##   South Africa South Rhodesia          Spain         Sweden    Switzerland 
##     0.03429664     0.16071740    -0.05261883    -0.45256252     0.19034296 
##         Turkey        Tunisia United Kingdom  United States      Venezuela 
##    -0.14453378    -0.21765669    -0.27221843    -0.25095085     0.30708996 
##         Zambia        Jamaica        Uruguay          Libya       Malaysia 
##     0.74823509    -0.34555773    -0.20513659    -1.16013341    -0.21262745
```

```
covratio(lm.SR)
```

```
##      Australia        Austria        Belgium        Bolivia         Brazil
##      1.1928303      1.2678392      1.1761879      1.2238199      1.0823332
##         Canada          Chile          China       Colombia     Costa Rica
##      1.3283009      0.6547098      1.1498637      1.1666845      0.9681384
##        Denmark        Ecuador        Finland         France        Germany
##      0.9344047      1.1393880      1.2031561      1.2262654      1.2256855
##         Greece      Guatamala       Honduras        Iceland          India
##      1.1396174      1.0852720      1.1855450      0.8658808      1.2024438
##        Ireland          Italy          Japan          Korea     Luxembourg
##      1.2680432      1.1624611      1.0845999      0.8695843      1.1961844
##          Malta         Norway    Netherlands    New Zealand      Nicaragua
##      1.1282611      1.1680616      1.2285315      1.1336998      1.1742677
##         Panama       Paraguay           Peru    Philippines       Portugal
##      1.0667255      0.8732040      0.8312741      0.8177726      1.2331038
##   South Africa South Rhodesia          Spain         Sweden    Switzerland
##      1.1945449      1.3130954      1.2081541      1.0864869      1.1471125
##         Turkey        Tunisia United Kingdom  United States      Venezuela
##      1.1003557      1.1314365      1.1886236      1.6554816      1.0945955
##         Zambia        Jamaica        Uruguay          Libya       Malaysia
##      0.5116454      1.1995171      1.1872025      2.0905736      1.1126445
```

Question-1

Chapter 6, Problem 15.

First check the following page from R project documentation (for various plots to visualize the influence measures):

https://cran.r-project.org/web/packages/olsrr/vignettes/influence_measures.html (https://cran.r-project.org/web/packages/olsrr/vignettes/influence_measures.html)

Note: You might need libraries such as olsrr for some of the plots below.

a. Plot : Cook's D chart, DFBETAs Panel, DFFITS Plot and Standardized Residual Chart that are shown in the above link.
b. Find the points with high leverage and Cook's distance.

c. Plot “Studentized Residuals vs Leverage Plot” that you see in the above link. Which regions in this plot corresponds to leverage points, pure leverage and influential regions. Detect the points in each region.

d. What do you think are the most influential points? (You can use the stats shown above or plots in previous parts.)

e. Comment about the normality assumption using probability plot. Remove the most influential points (that you suggested in part-d) and discuss the change/improvements on normality assumption (comparing probability plots).

Answer:

```
library(olsrr)
```

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

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

```
library(MPV)
```

```
##
## Attaching package: 'MPV'
```

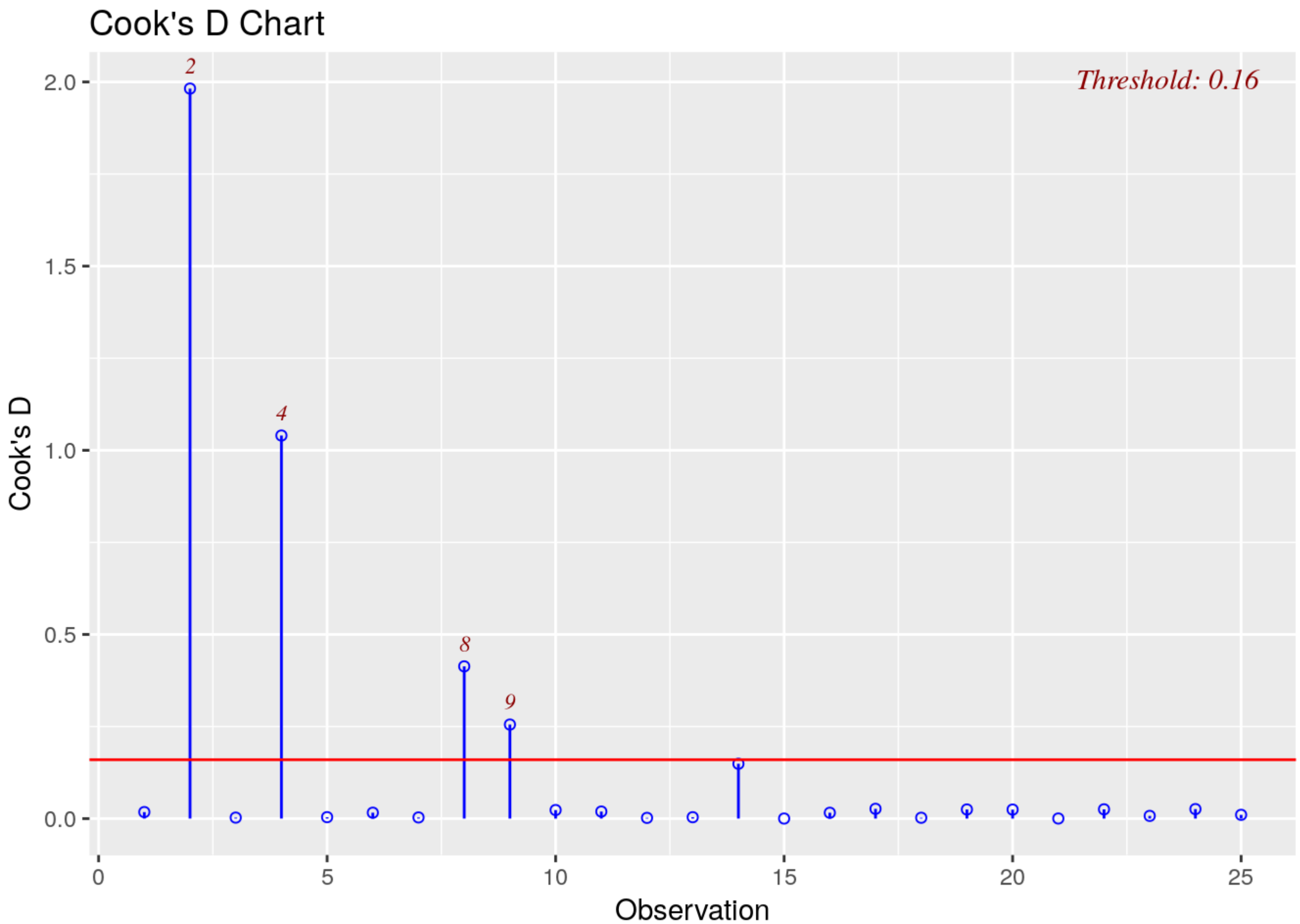
```
## The following object is masked from 'package:olsrr':
##
##     cement
```

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

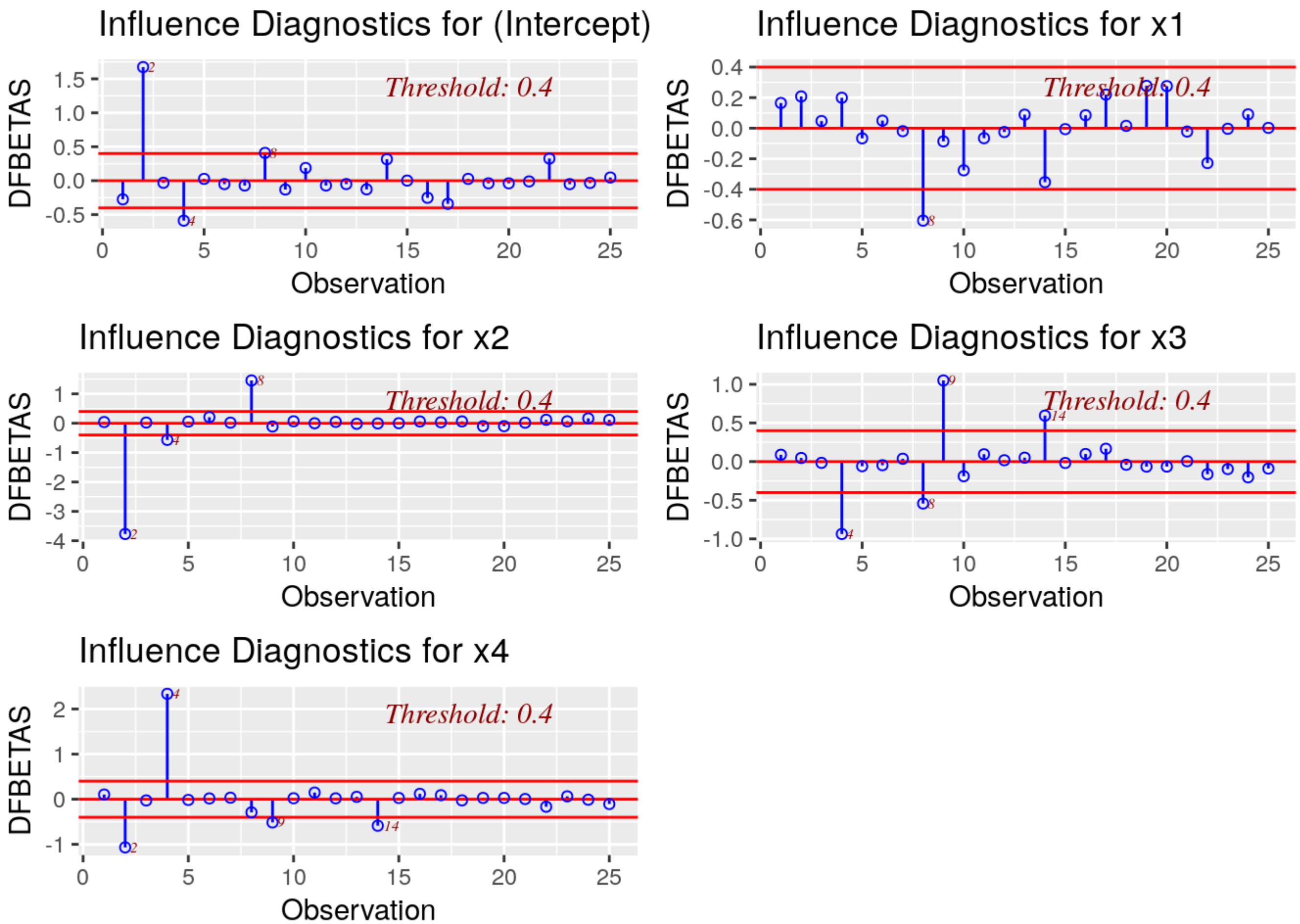
```
data(table.b14)
fitted=lm(y~x1+x2+x3+x4, data=table.b14)


cat ("Part a.")
```

```
## Part a.
```

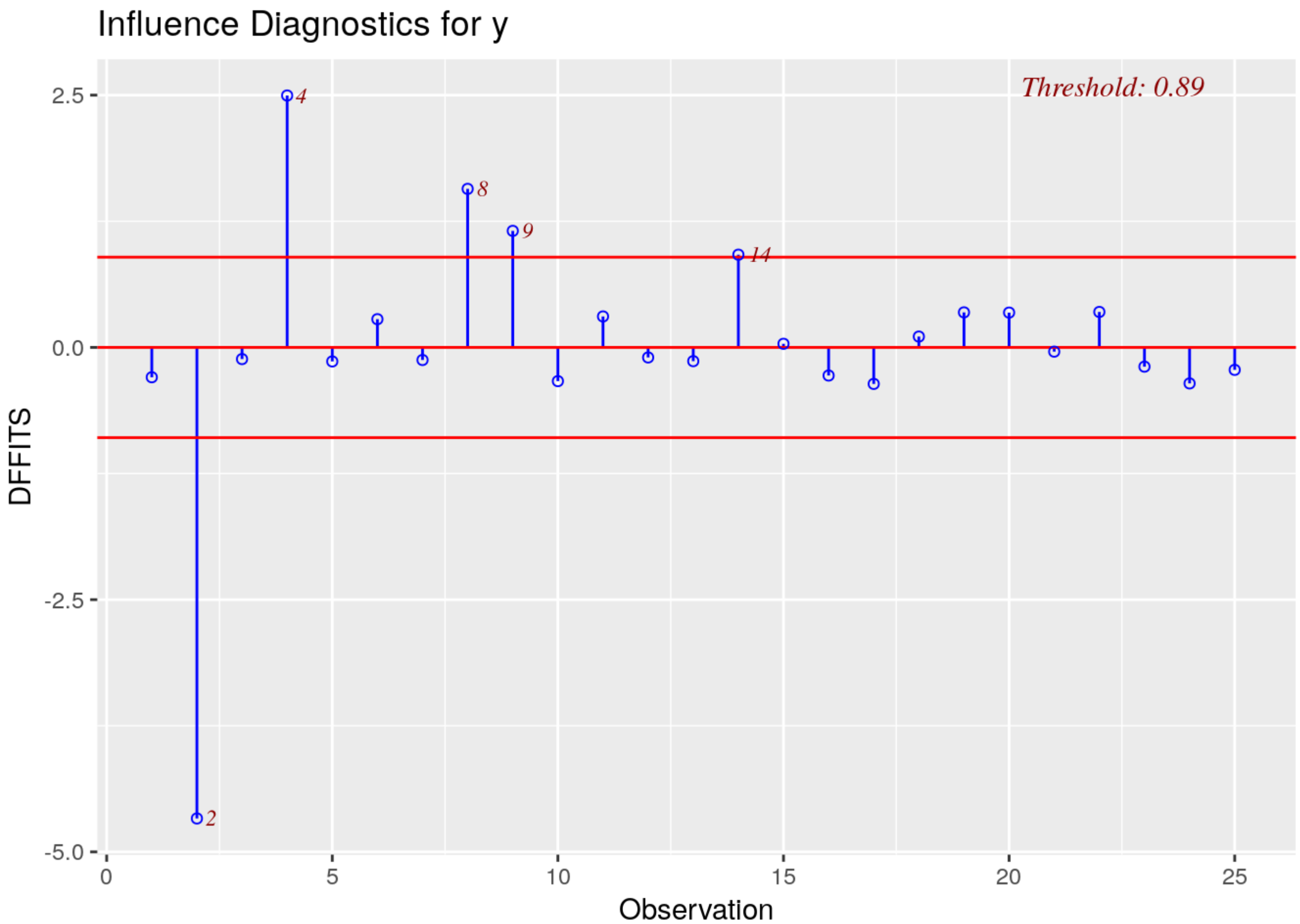
```
# Cooks D
ols_cooksd_chart(fitted)
```



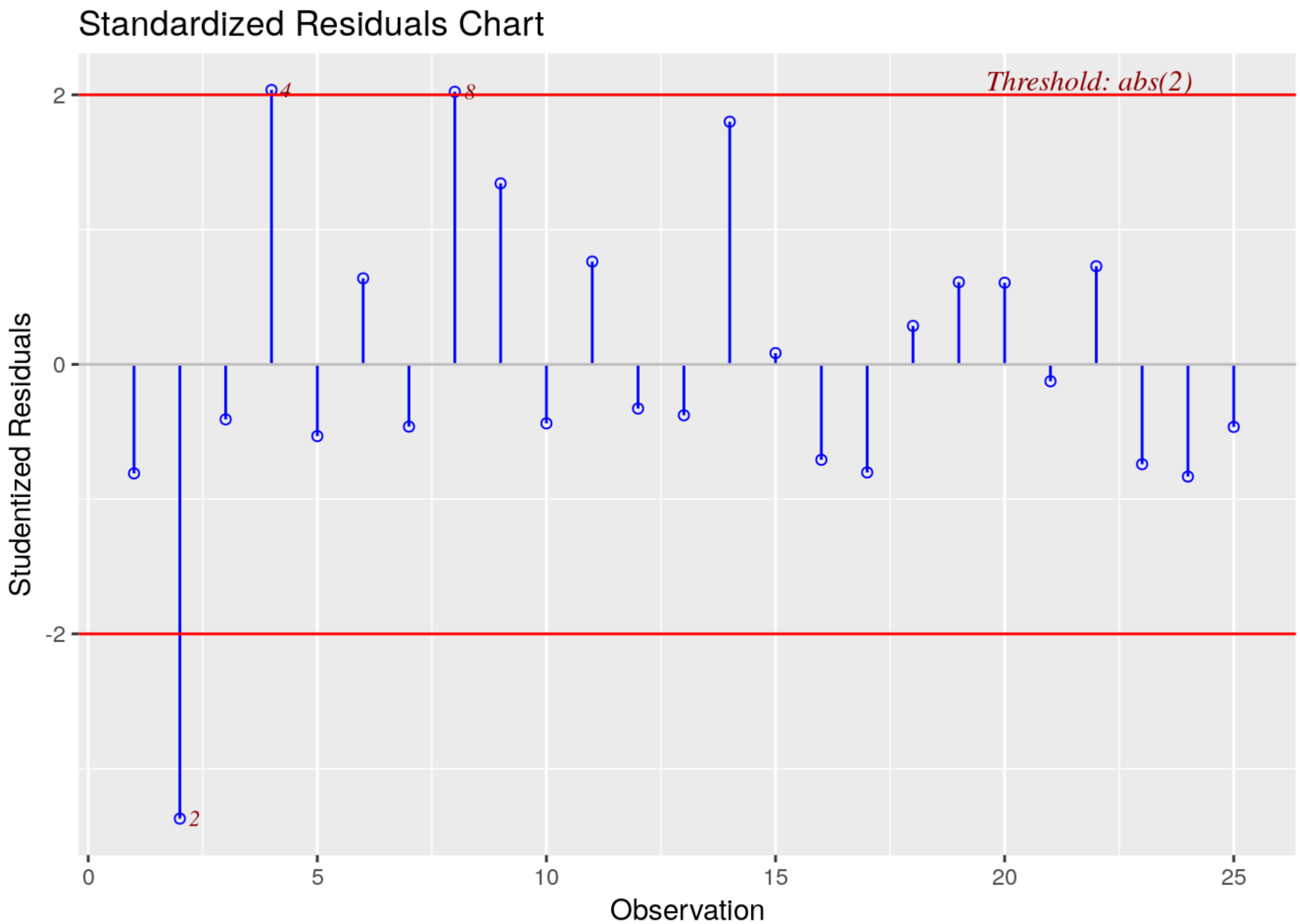
```
#   DFBETAs Panel
ols_dfbetas_panel(fitted)
```




```
# DFFITS Plot
ols_dffits_plot(fitted)
```




```
# Standardized Residuals Chart
ols_srsd_chart(fitted)
```



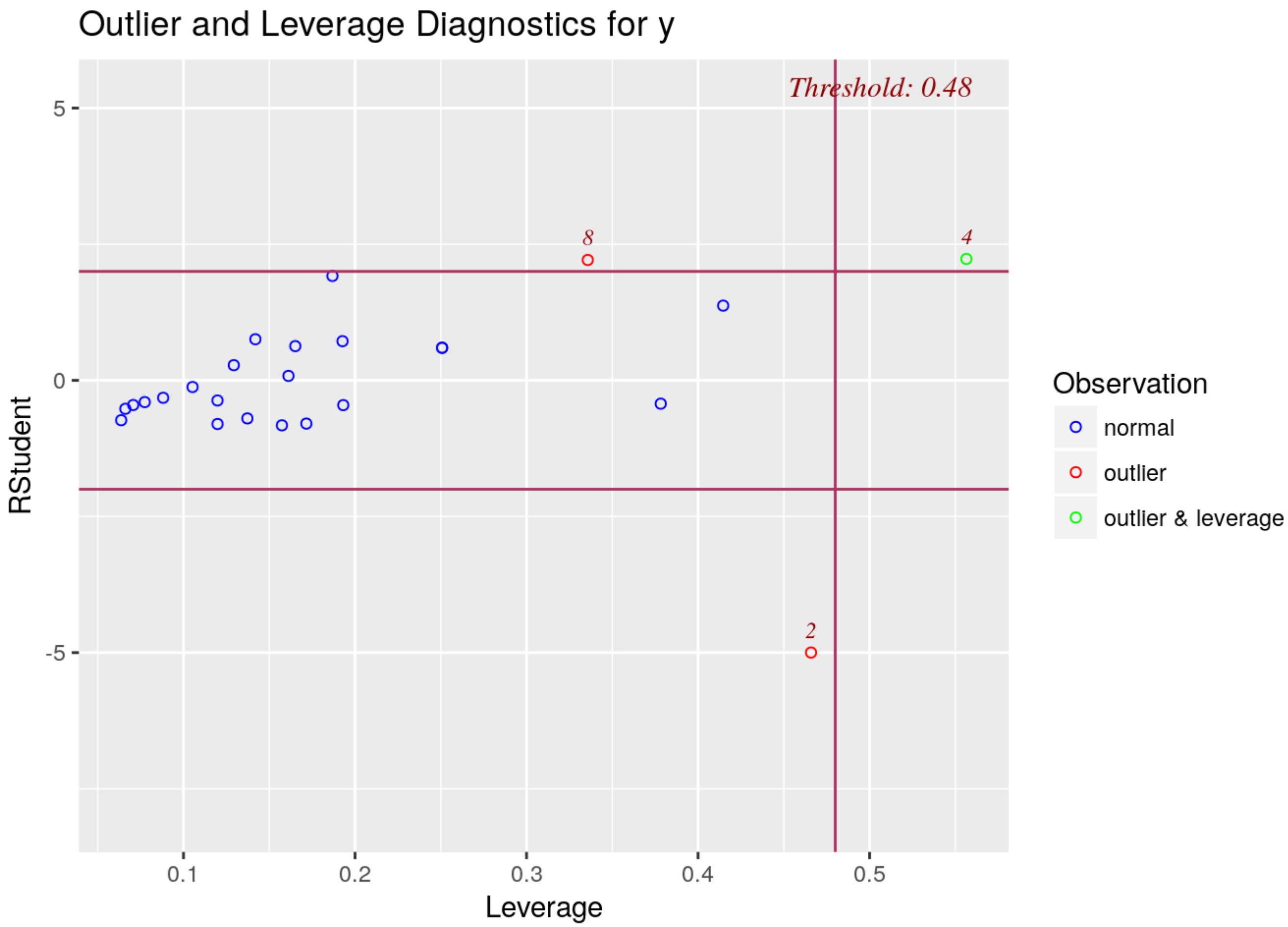

```
cat ("Part b.
     2,4,8 and 9 seems to have high Cook's D value. From the outlier and leverage diagnostics plot we see that point 4 has t
he highest leverage, and point 2 is the next one that can also be included as a leverage point since it seems just a little
 below the threshold. 2 and 4 are the two points with highest Cook's D.")
```

```
## Part b.
##      2,4,8 and 9 seems to have high Cook's D value. From the outlier and leverage diagnostics plot we see that point 4 ha
s the highest leverage, and point 2 is the next one that can also be included as a leverage point since it seems just a litt
le below the threshold. 2 and 4 are the two points with highest Cook's D.
```

```
cat("Part c.
    The points on the right side of the vertical line are  “leverage points” which denote the points that are at remote loca
tions in x-space. 4 is in this region and 2 is very close to it.  The points with high leverage but consistent with the fitt
ed model (thus low residual) are pure leverage points. This corresponds the middle section of the right side of the vertical
line. The remaining upper and lower sections (on the right of the vertical line) are the regions for influential points. Obs
ervation 4 is in this region.
There are two horizontal line let us denote as the upper and lower one. The regions above the upper horizontal line and belo
w the lower horizontal line are outliers. 2,4,8 are in this region.
")
```

```
## Part c.
##     The points on the right side of the vertical line are  “leverage points” which denote the points that are at remote l
ocations in x-space. 4 is in this region and 2 is very close to it.  The points with high leverage but consistent with the f
itted model (thus low residual) are pure leverage points. This corresponds the middle section of the right side of the verti
cal line. The remaining upper and lower sections (on the right of the vertical line) are the regions for influential points.
Observation 4 is in this region.
## There are two horizontal line let us denote as the upper and lower one. The regions above the upper horizontal line and b
elow the lower horizontal line are outliers. 2,4,8 are in this region.
```

```
ols_rsdlev_plot(fitted)
```



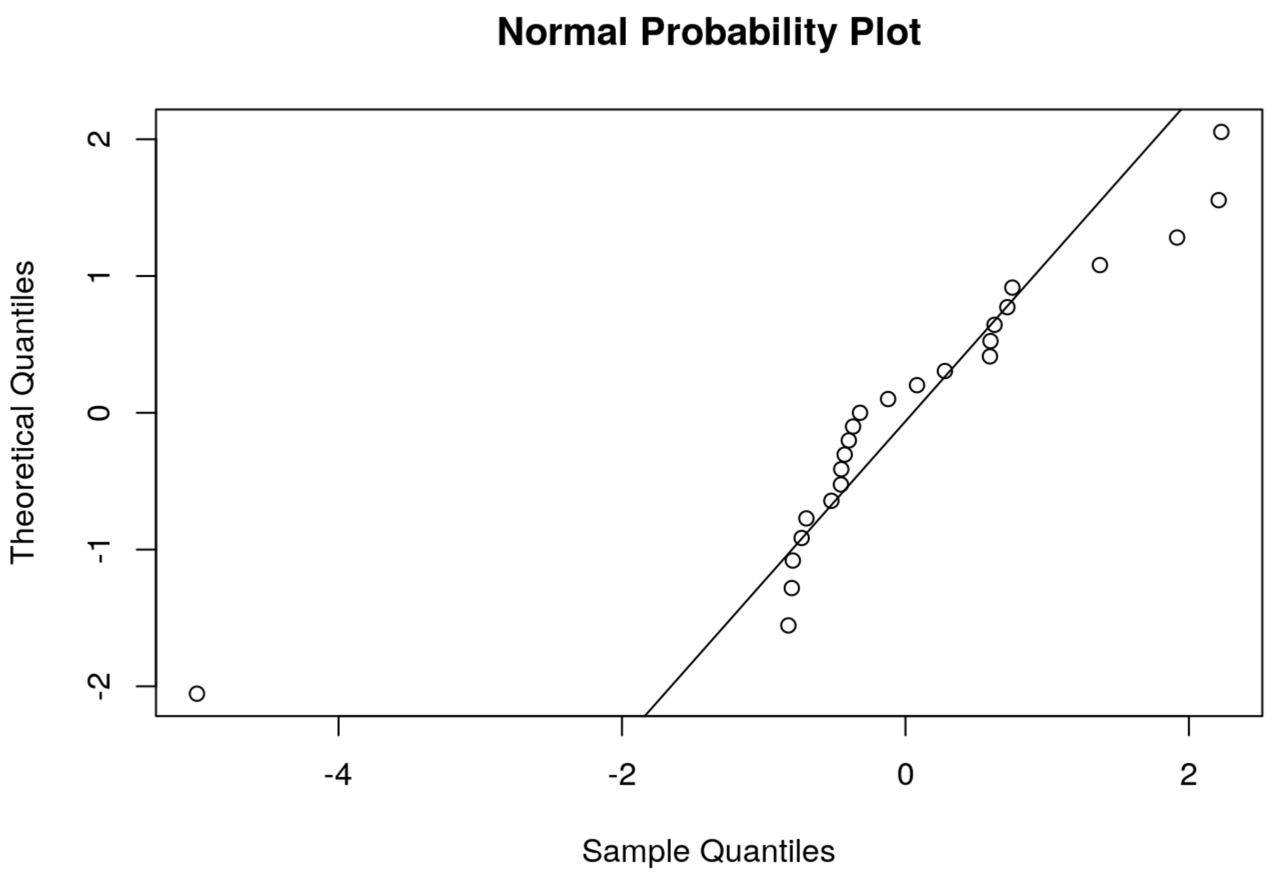

```
cat("Part d.
    The points 2 and 4 are considered as leverage points and they are also outliers therefore we can conclude they are the i
nfluential points. It is possible to add more by considering different cut-off values. ")
```

```
## Part d.
##     The points 2 and 4 are considered as leverage points and they are also outliers therefore we can conclude they are th
e influential points. It is possible to add more by considering different cut-off values.
```

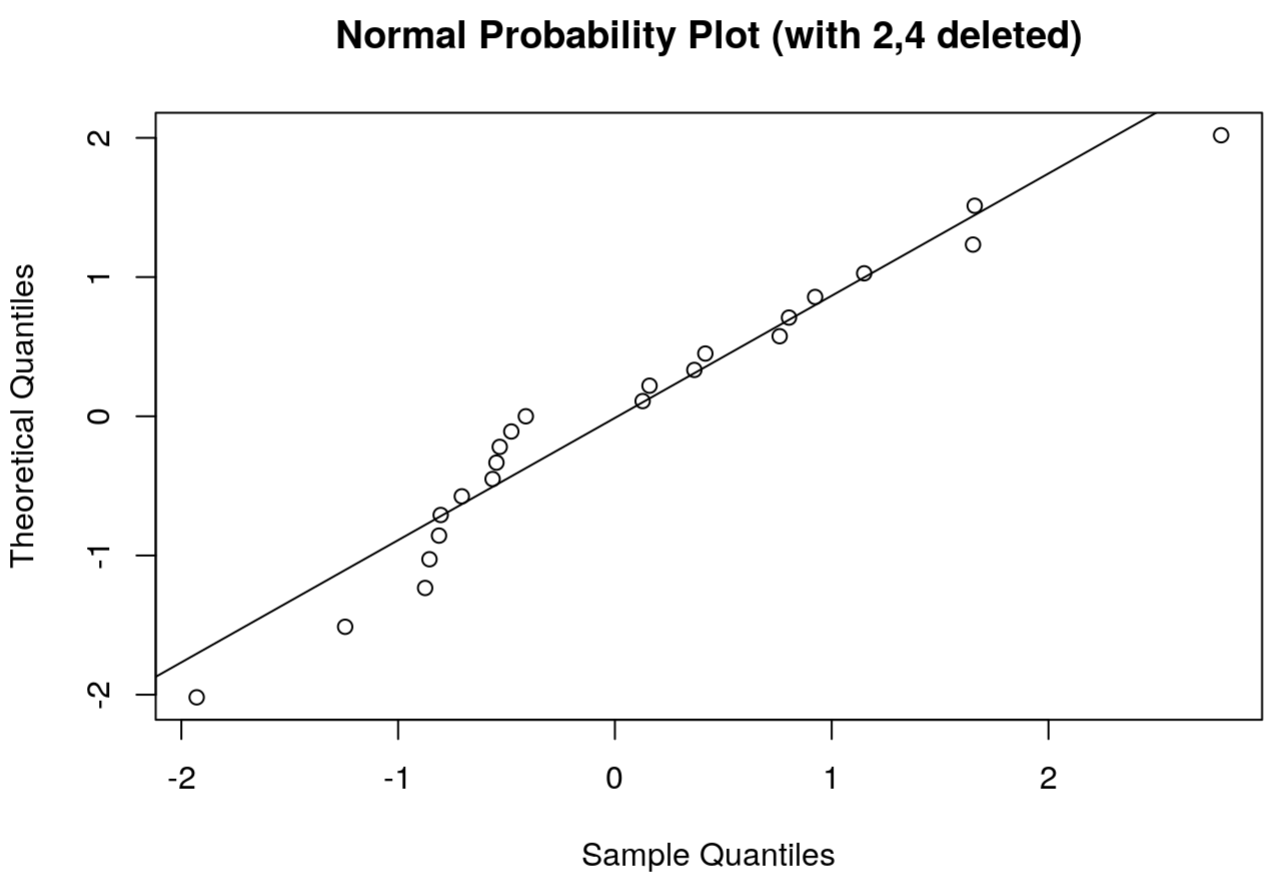
```
cat("Part e.
    As we can see from the probability plot there is a violation of normality assumption. The plot improves slightly with th
e removal of outliers that we are studying.")
```

```
## Part e.
##     As we can see from the probability plot there is a violation of normality assumption. The plot improves slightly with the removal of outliers that we are studying.
```

```
rStuRes = rstudent(fitted)
qqnorm(rStuRes, datax = TRUE, main="Normal Probability Plot")
qqline(rStuRes, datax = TRUE)
```



```
delTable=table.b14[-c(2,4),]
fitted=lm(y~x1+x2+x3+x4, data=delTable)
rStuRes = rstudent(fitted)
qqnorm(rStuRes, datax = TRUE, main="Normal Probability Plot (with 2,4 deleted)")
qqline(rStuRes, datax = TRUE)
```




```
```