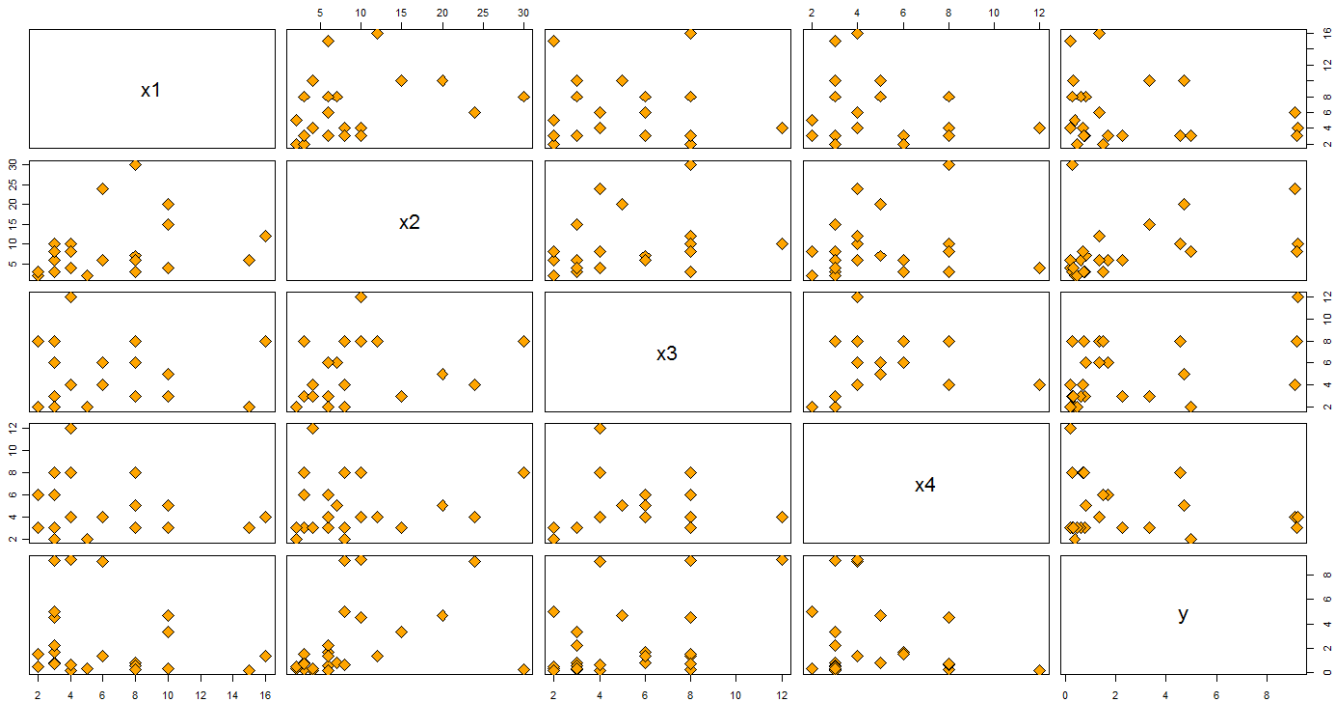
REGRESSION MODELS – HW 2
CHAPTER 4-5 EXCERCISES

Thanh Doan – Student ID 0159701

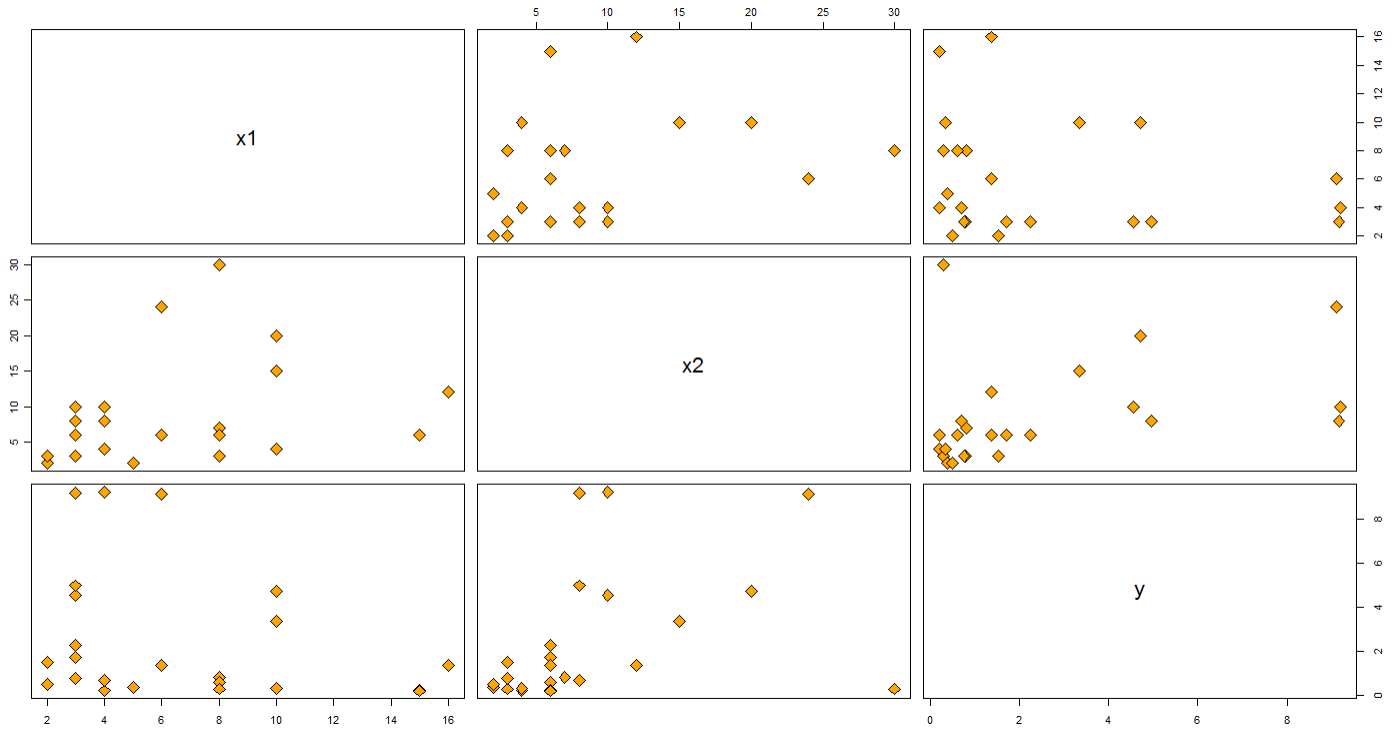
EXERCISE 4.22

a - Fit a multiple regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$
and investigate the adequacy of the model

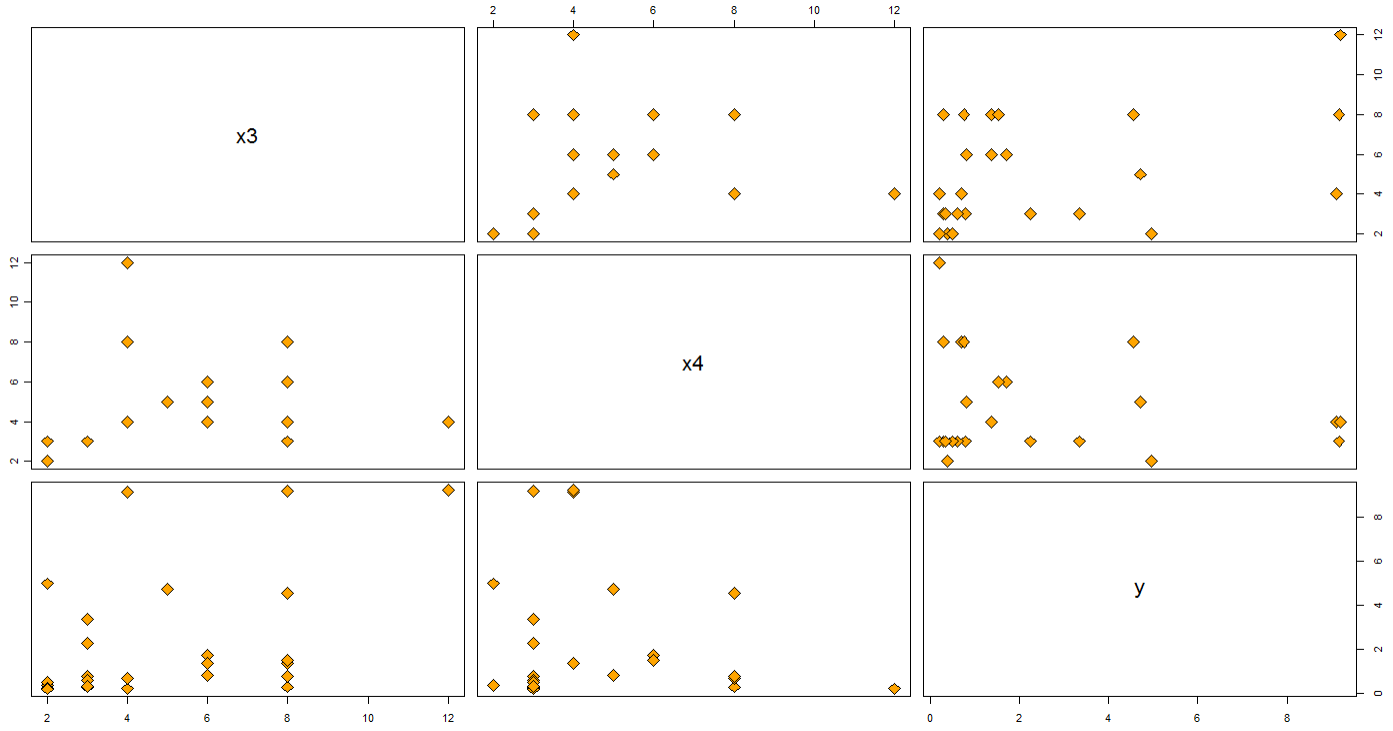
- Plot the scatter plot to have a bird-eye view on the relationship between response and explanatory variables



- Let us zoom in by creati8ng a scatter plot between response and x1, x2 variables



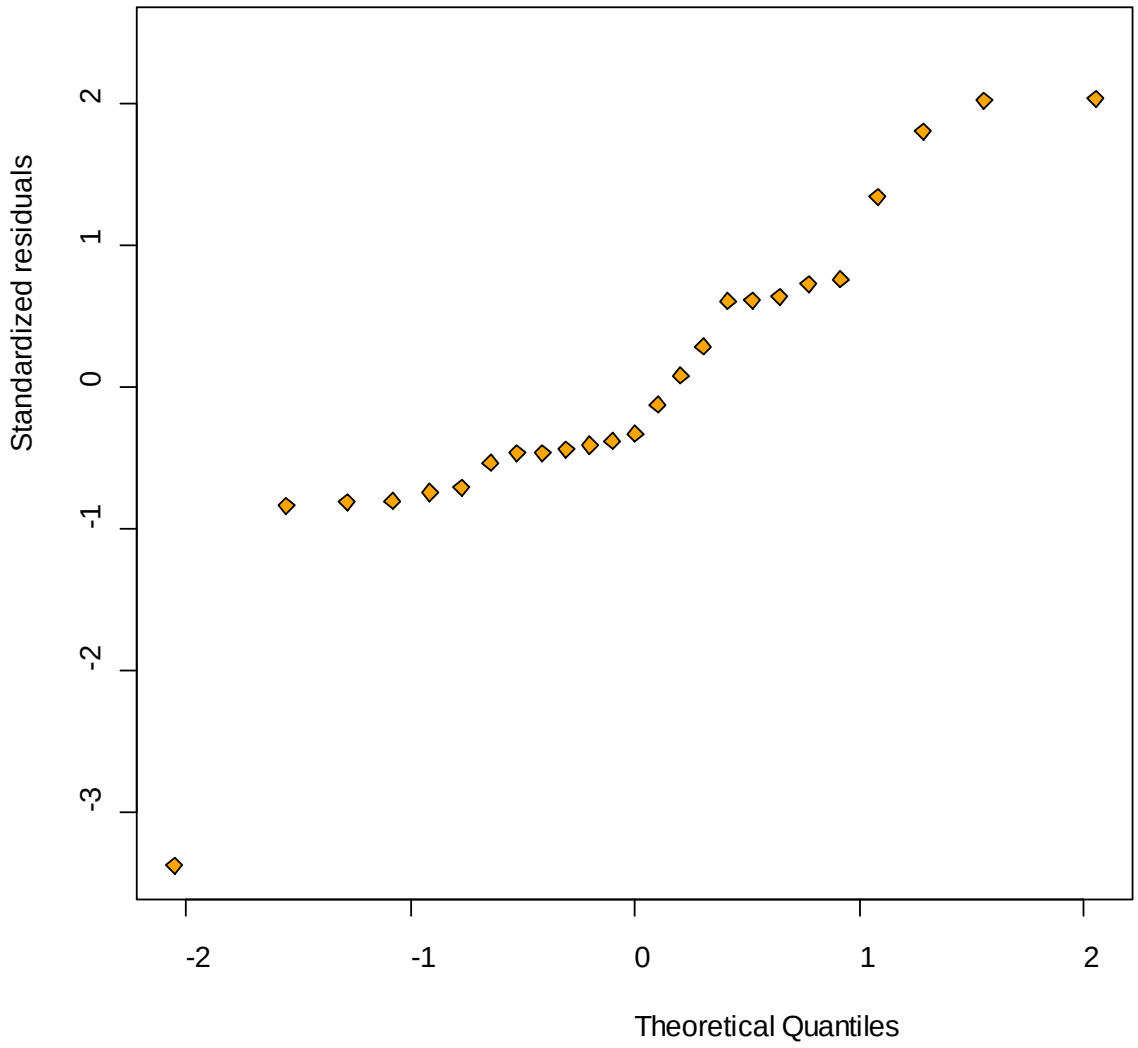
- Scatter plot between response and x3, x4 variables



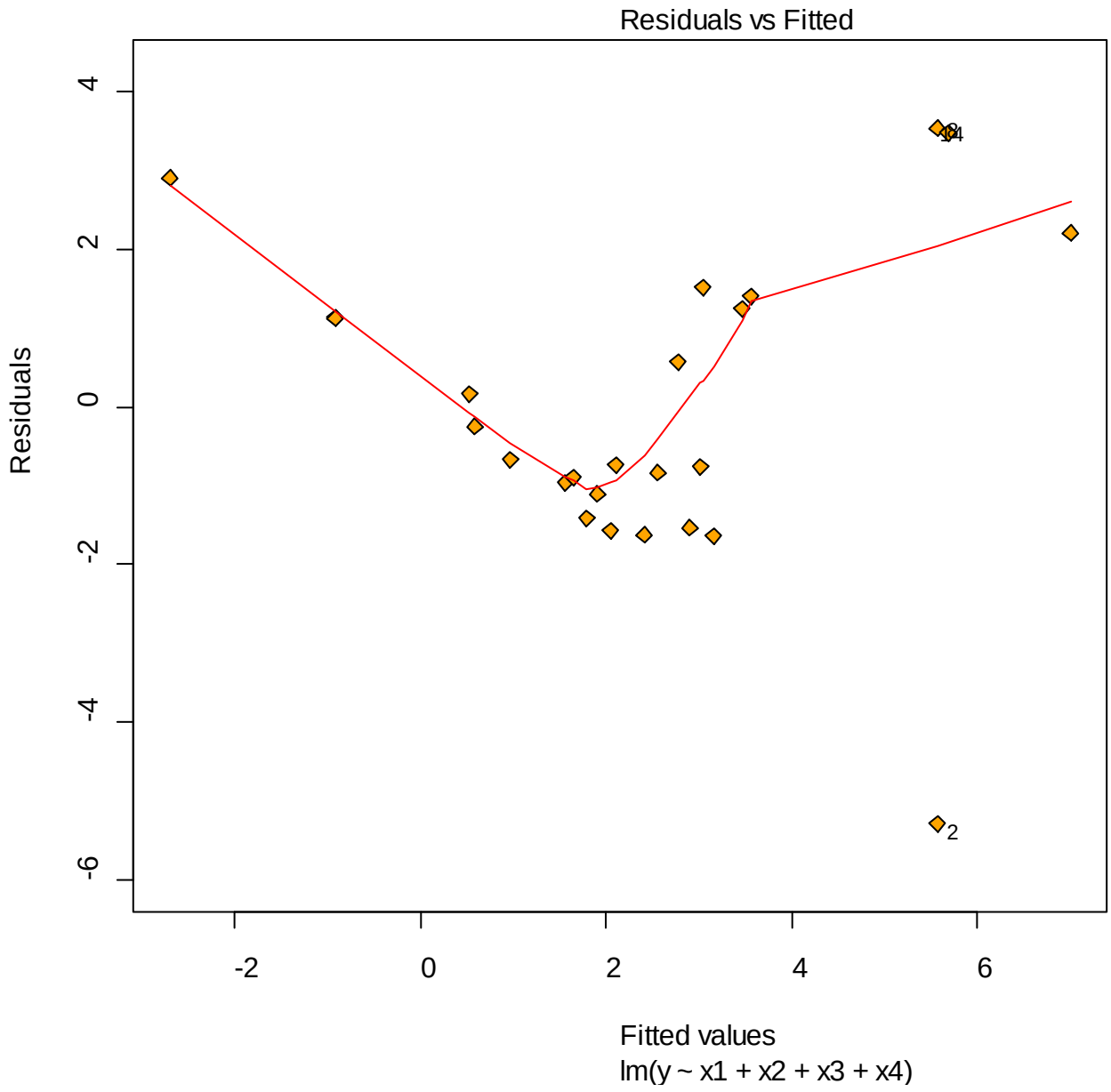
- Fit a multiple regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$

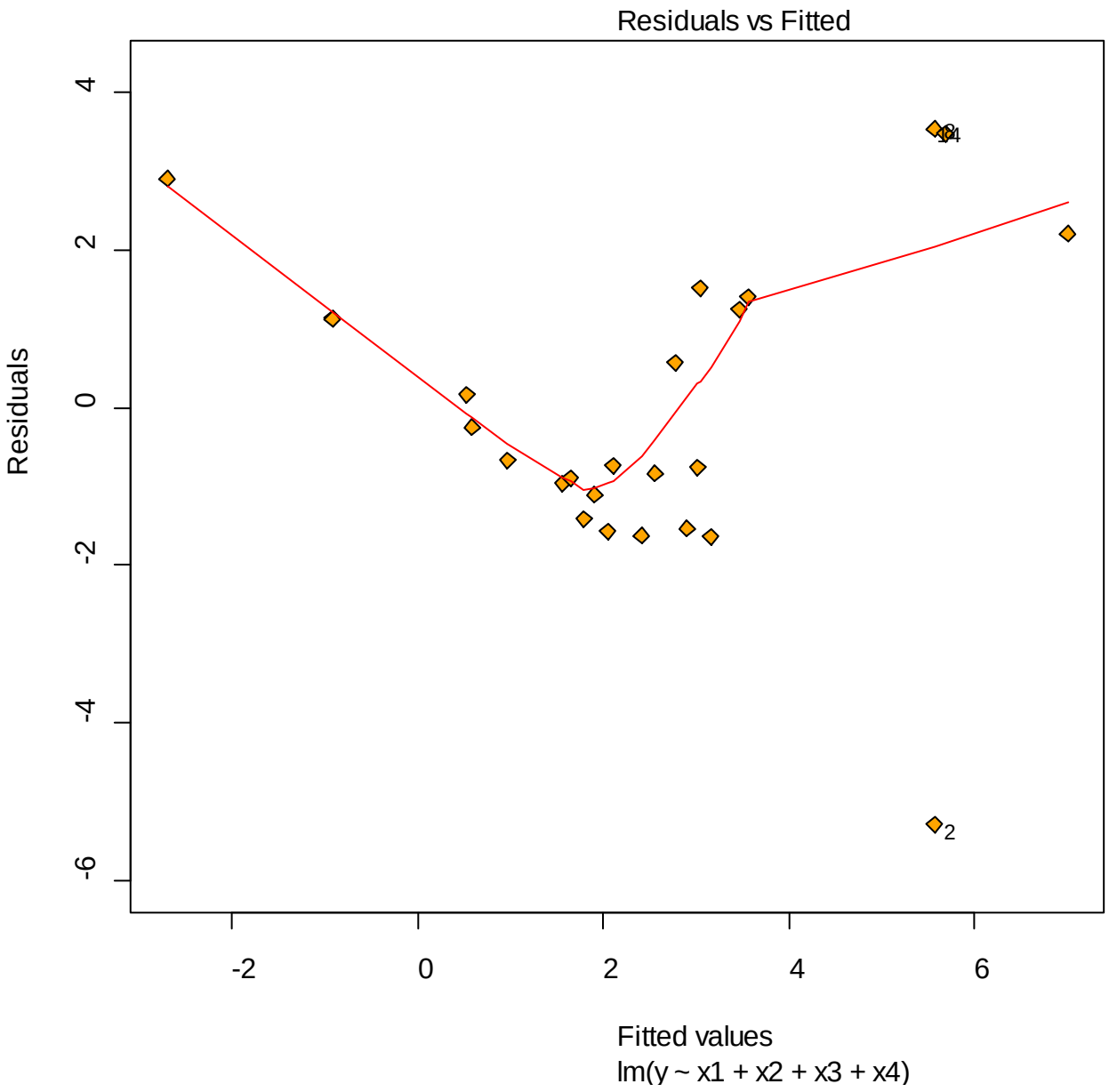
```
> lm(y ~ x1+x2+x3+x4, data=b14);

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = b14)

Coefficients:
(Intercept)           x1           x2           x3           x4
     3.1482      -0.2900       0.1992       0.4554      -0.6092
```

- Normal Probability Plot of the residuals

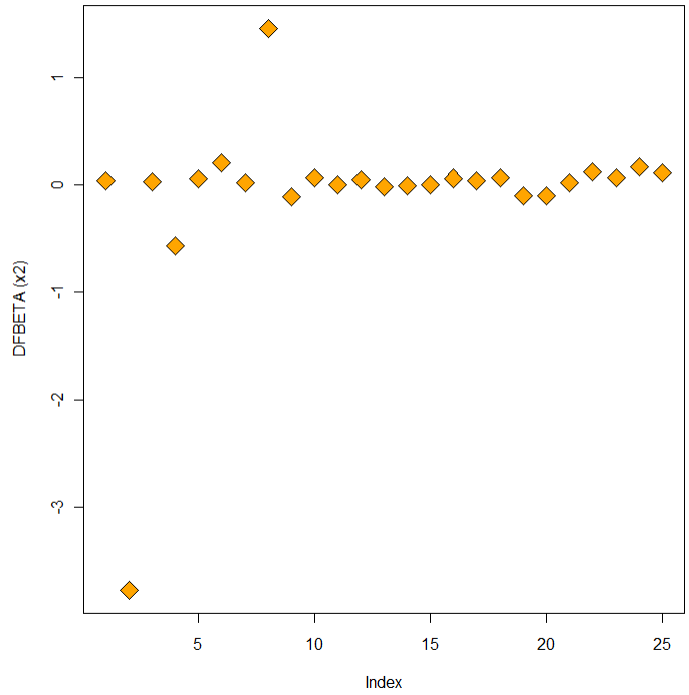


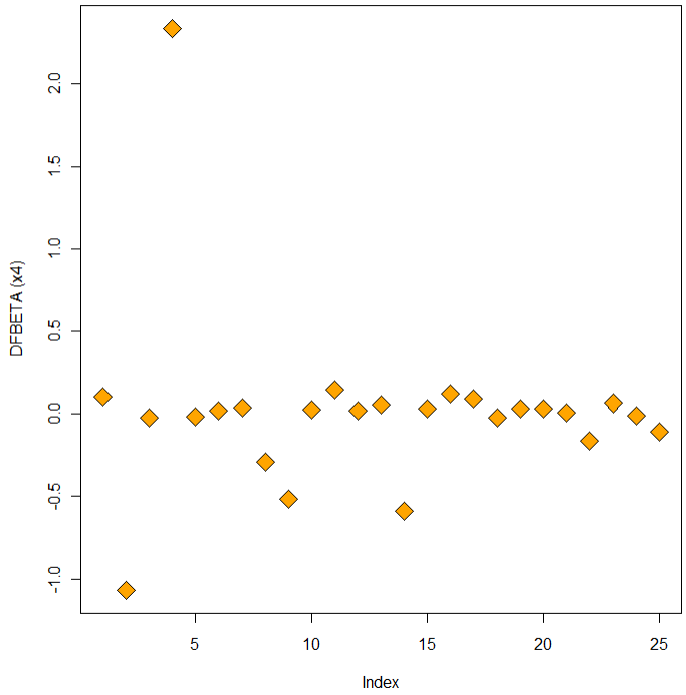
- Plot of the residuals against fitted values



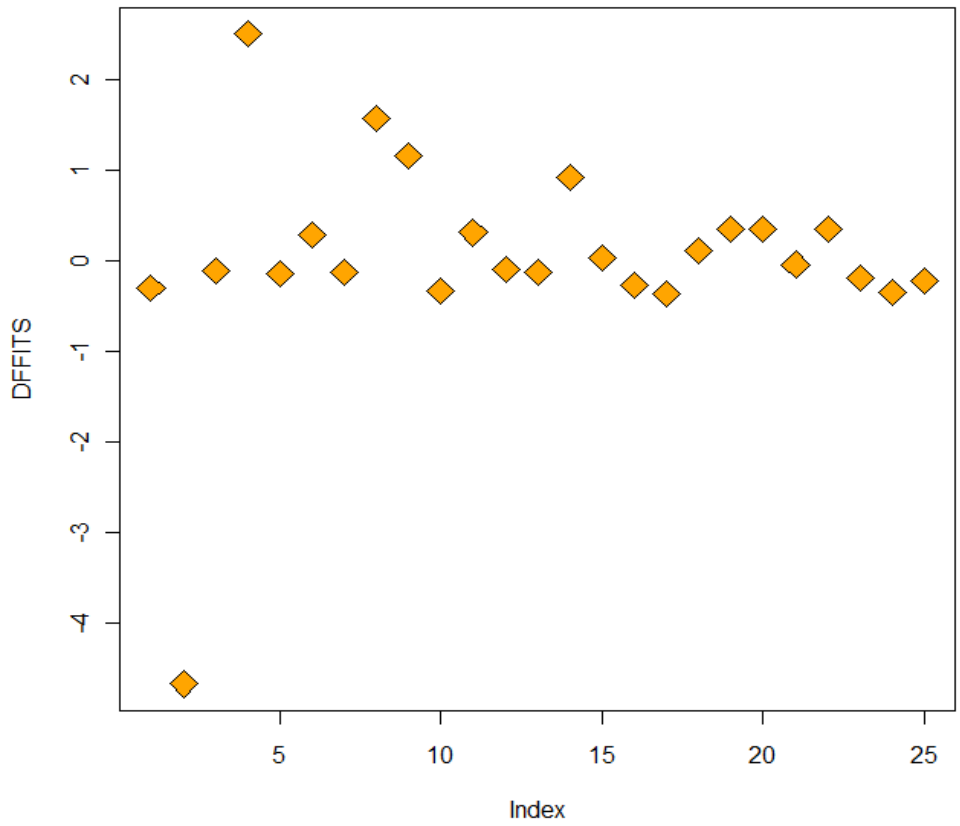

Residuals vs Fitted
Residuals
Fitted values
lm(y ~ x1 + x2 + x3 + x4)
4
2
0
-2
-4
-6
-2
0
2
4
6
2

- DFBETA Plots

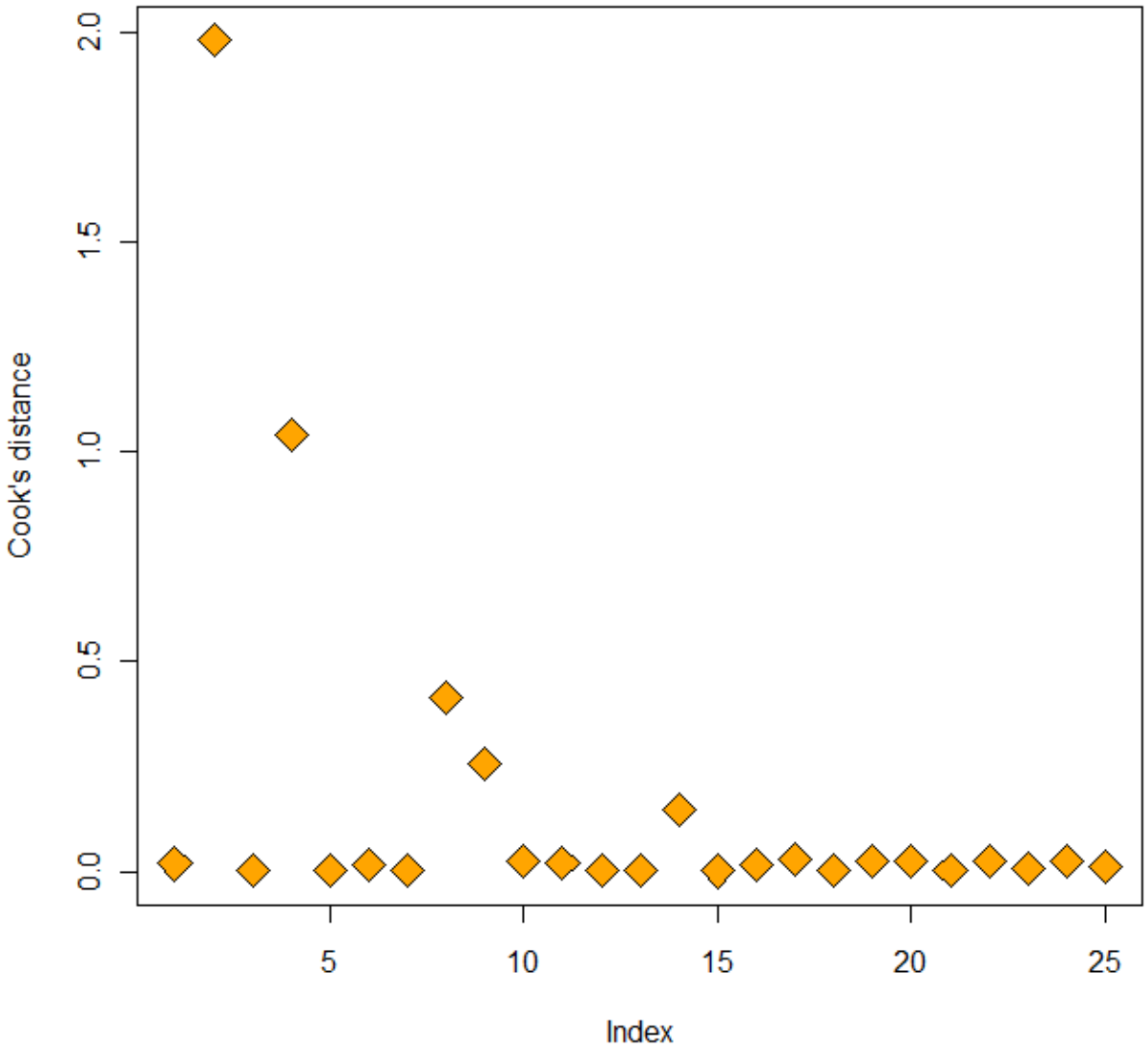




- DFFITS Plot



- Cook's distance Plot



- Bonferroni outlier test

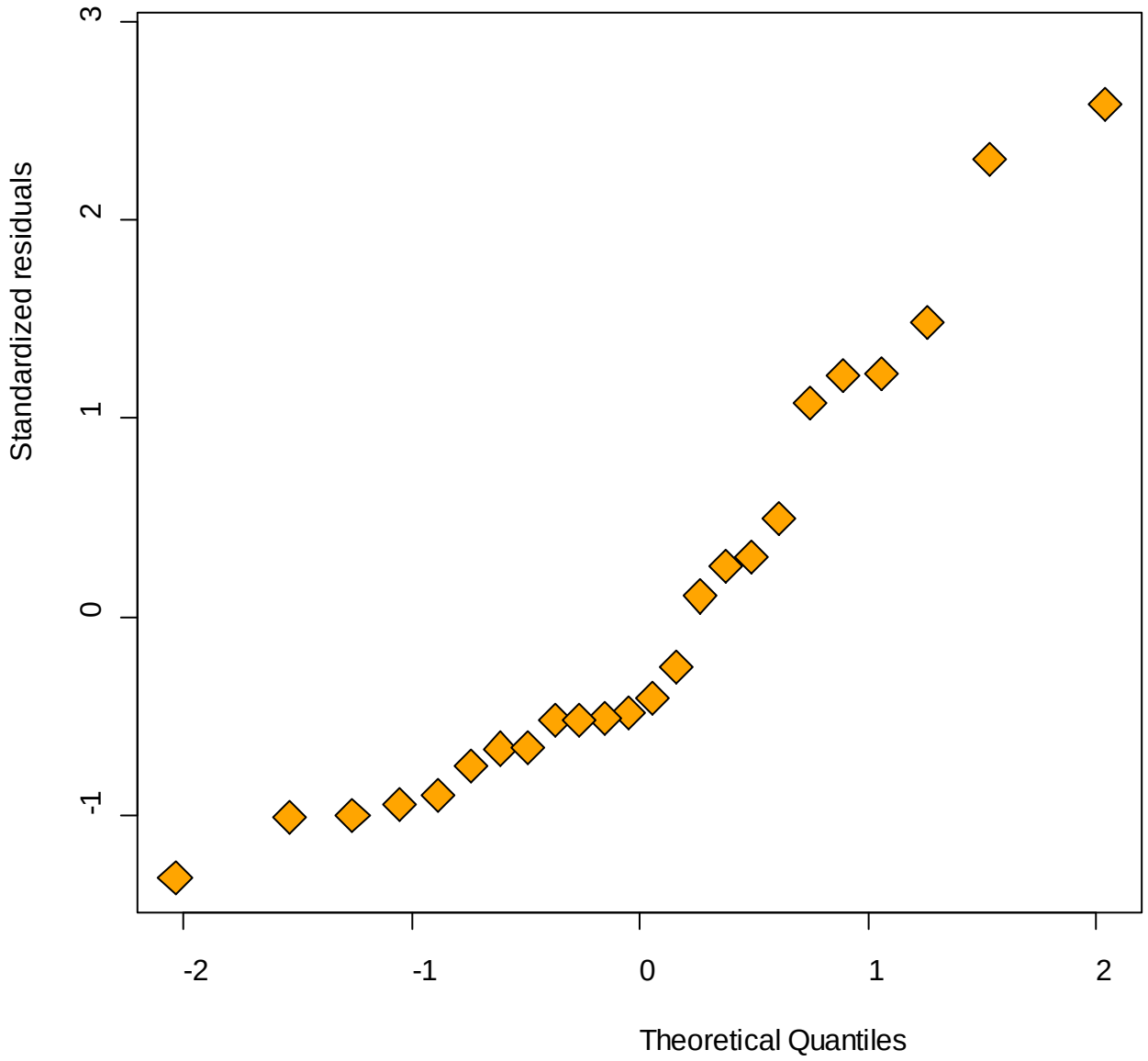
```
> # Bonferroni outlier test
> outlierTest(m422);
   rstudent unadjusted p-value Bonferonni p
2 -4.998983          7.968e-05     0.001992
>
```

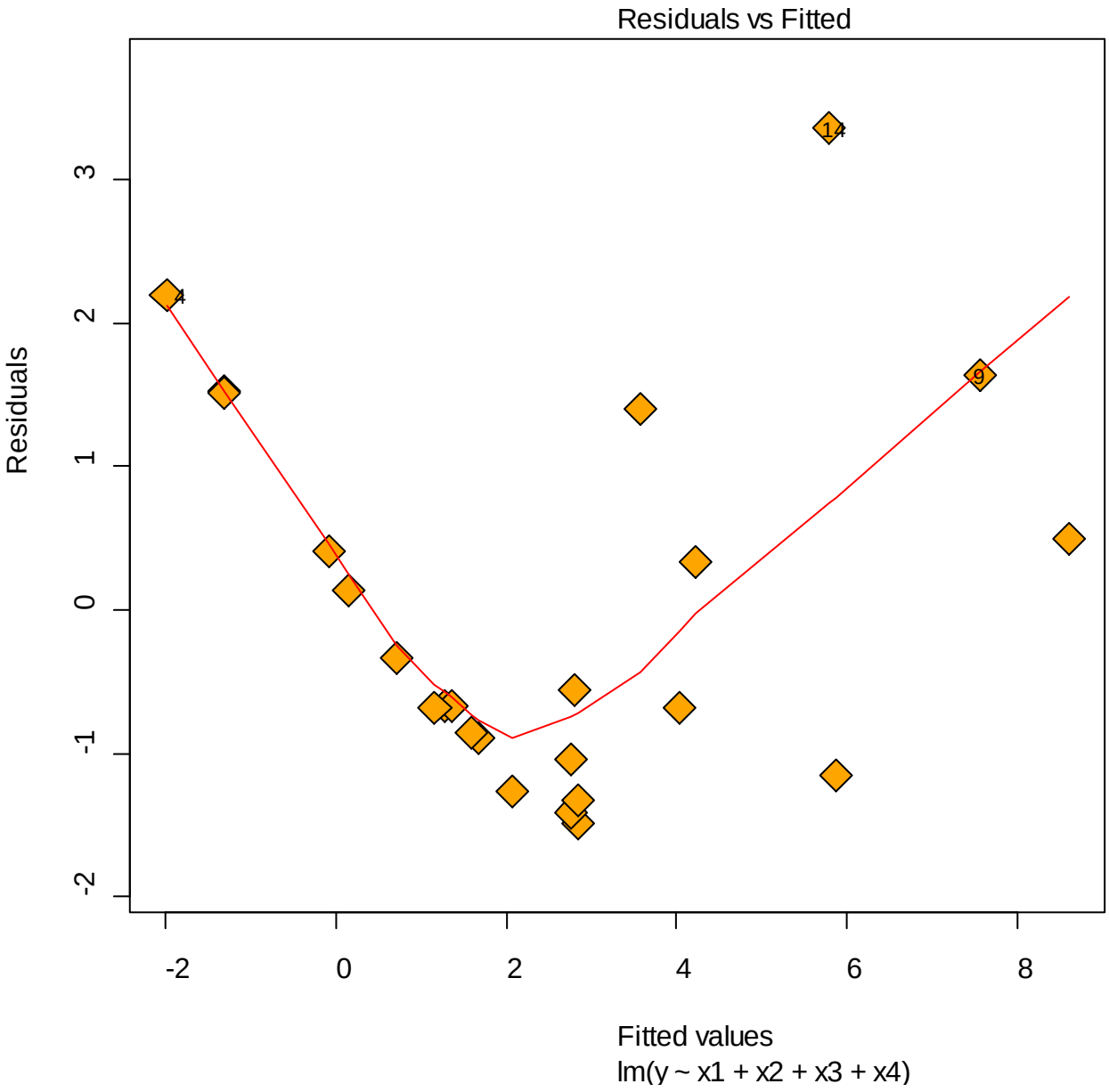
Comment:

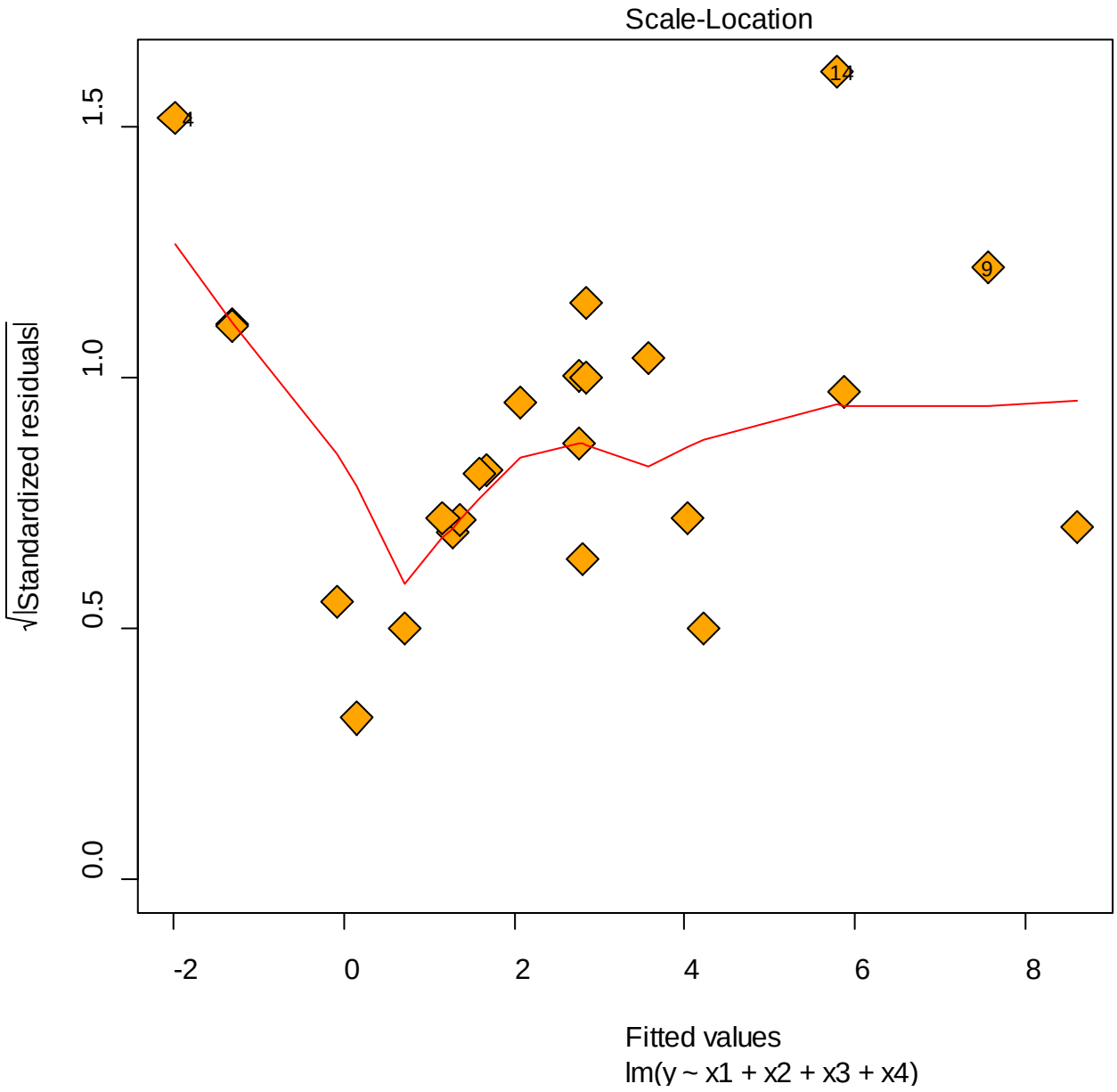
Diagnostic plots & Bonferroni outlier test suggest *Observation 2* is a potential outlier and the model does not fit very well

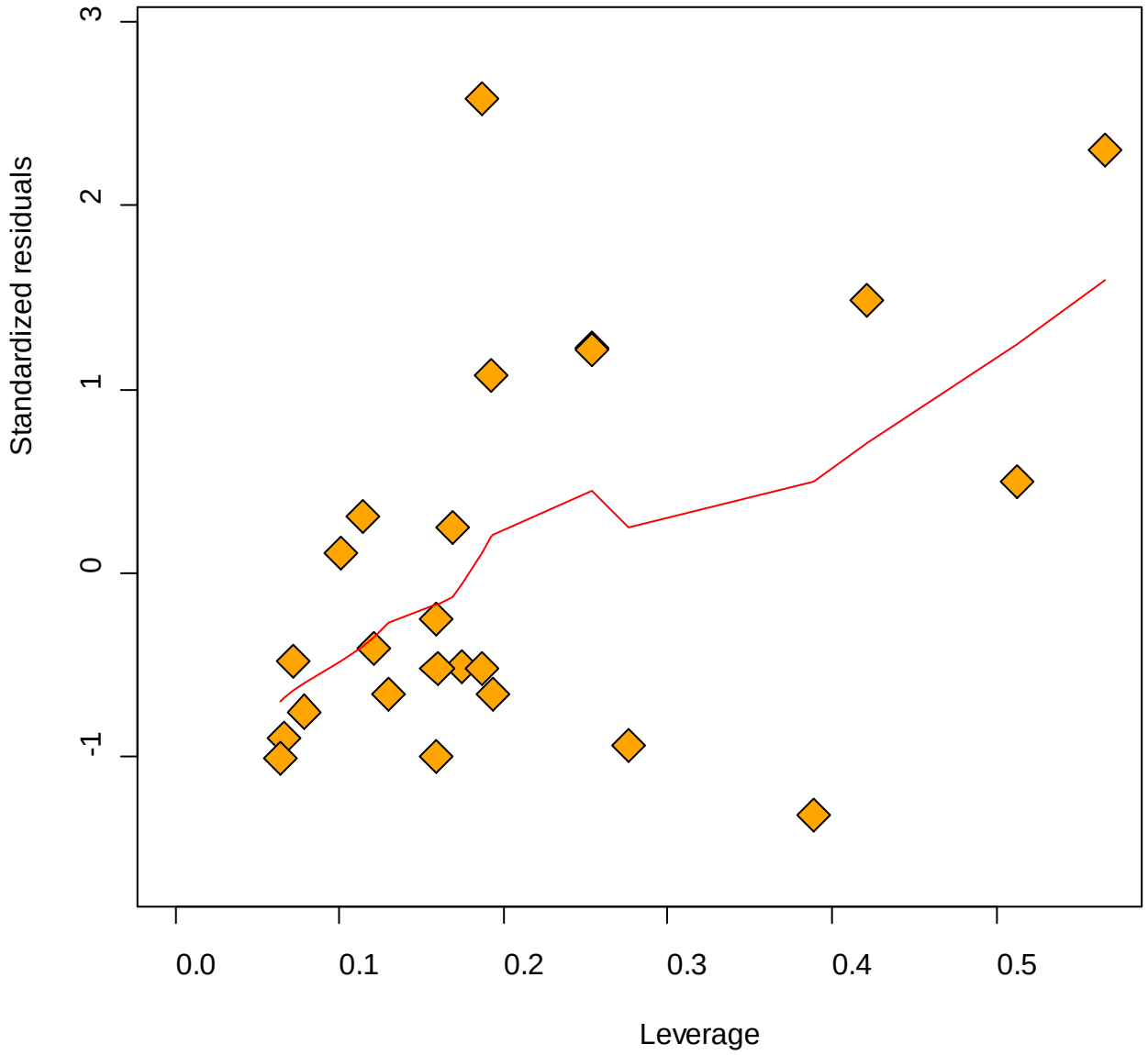
b – Delete observation 2 and refit the model, perform residual analysis.

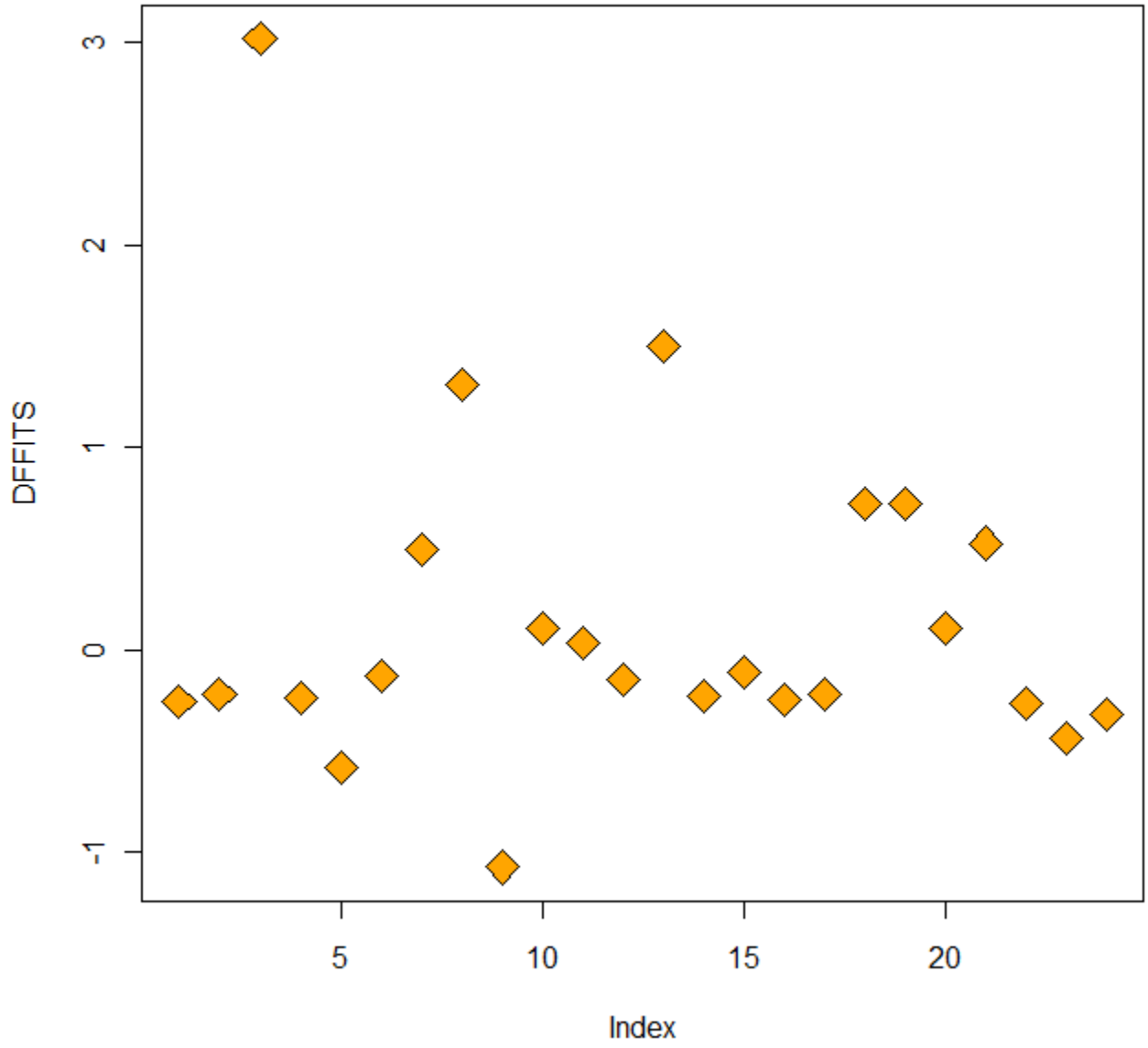
```
> subs = c(1,seq(3,25));
> b14b = b14[subs,];
> m422b=lm(y ~ x1+x2+x3+x4, data=b14b);
> # old model
> m422;

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = b14)

Coefficients:
(Intercept)           x1           x2           x3           x4
     3.1482      -0.2900       0.1992       0.4554      -0.6092

> # new model - without observation 2
> m422b;

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = b14b)

Coefficients:
(Intercept)           x1           x2           x3           x4
     1.5243      -0.3061       0.3744       0.4496      -0.4656
```

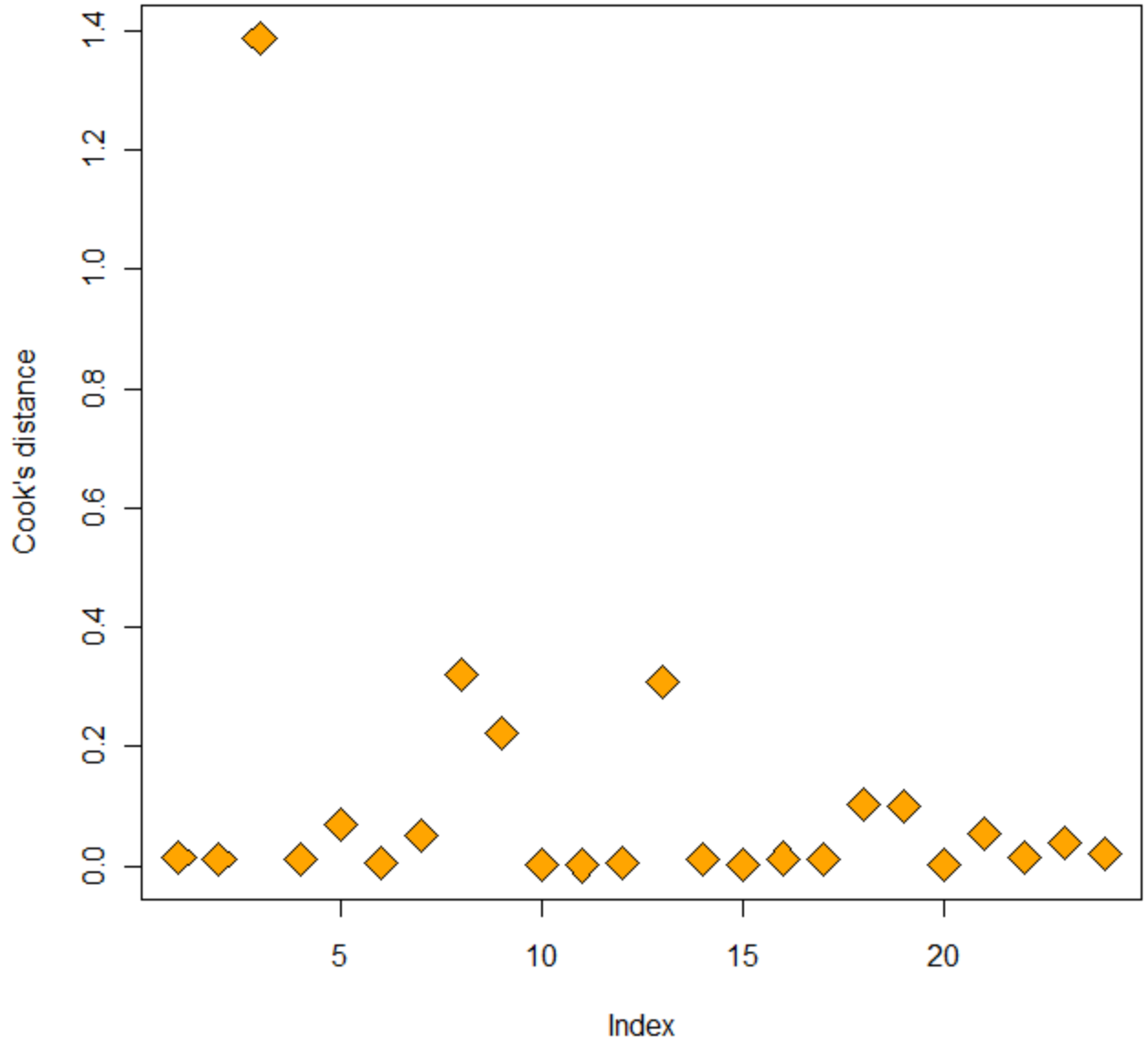

Standardized residuals
3
2
1
0
-1
-2
-1
0
1
2
Theoretical Quantiles


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ x1 + x2 + x3 + x4)
14
4
9
-2
-1
0
1
2
3
-2
0
2
4
6
8


Scale-Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
-2
0
2
4
6
8
Fitted values
lm(y ~ x1 + x2 + x3 + x4)
4
14
9


Standardized residuals
3
2
1
0
-1
0.0
0.1
0.2
0.3
0.4
0.5
Leverage


DFFITS
3
2
1
0
-1
5
10
15
20
Index


Cook's distance
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
5
10
15
20
Index

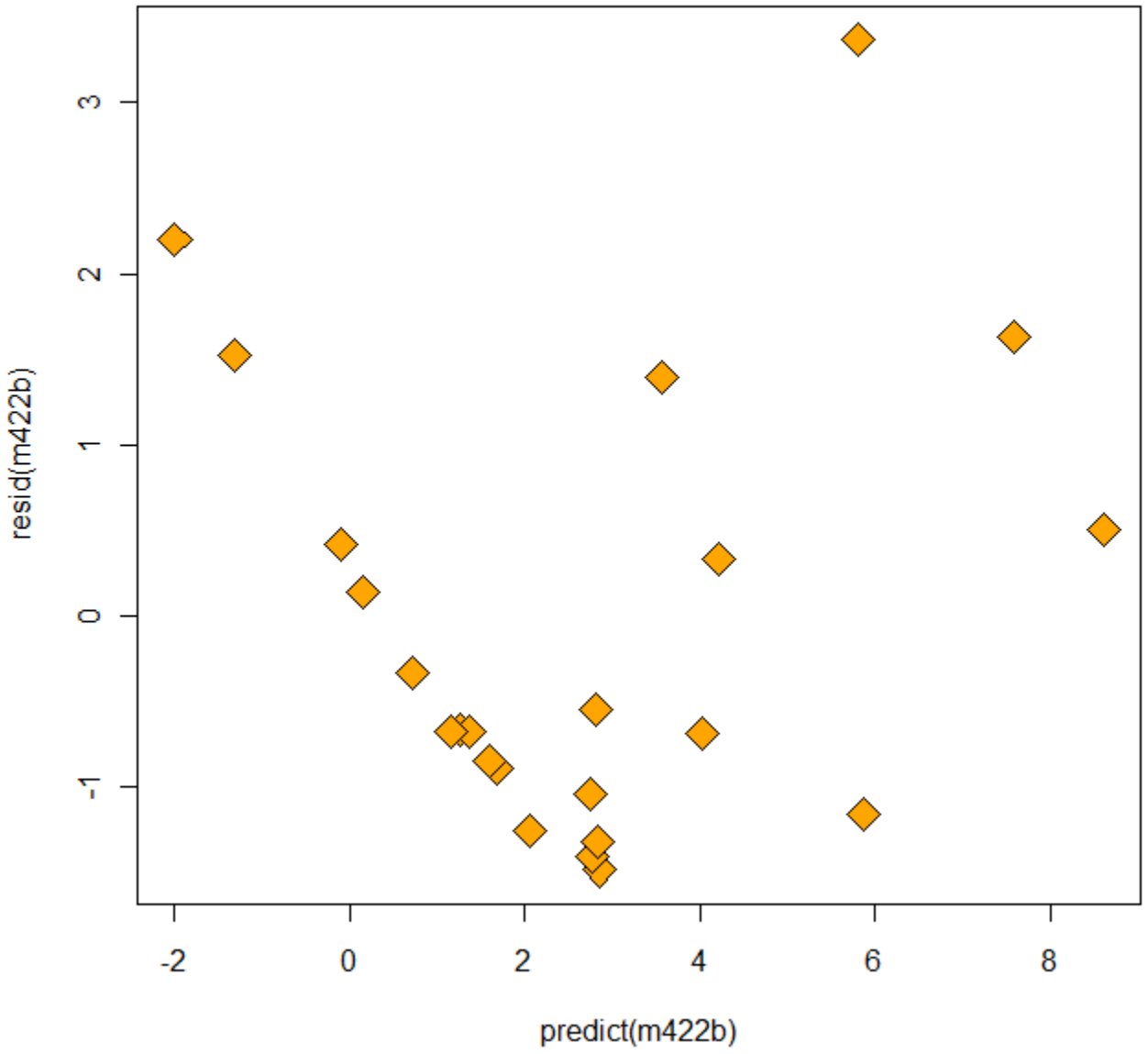
```
> # Bonferroni outlier test
> outlierTest(m422b);

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferonni p
14 3.120792          0.0059044      0.14171
>
> #All the influence measures
> influence.measures(m422b);
Influence measures of
         lm(formula = y ~ x1 + x2 + x3 + x4, data = b14b) :

     dfb.1_    dfb.x1    dfb.x2    dfb.x3    dfb.x4    dffit cov.r   cook.d    hat inf
1  -0.23604   0.13211   0.06986   0.07204   0.09740 -0.2527 1.340 0.013170 0.1301
3  -0.04835   0.08805   0.02306 -0.02992 -0.05407 -0.2171 1.221 0.009652 0.0782
4  -0.79463   0.22223 -0.30450 -1.12663   2.82125   3.0188 0.586 1.386161 0.5662   *
5   0.04987 -0.11303   0.06539 -0.10212 -0.03118 -0.2398 1.128 0.011616 0.0666
6   0.19342 -0.06419 -0.48637   0.07879 -0.10426 -0.5803 1.425 0.067780 0.2763
7  -0.07584 -0.02115   0.02817   0.03838   0.03815 -0.1310 1.329 0.003578 0.0723
8   0.00684 -0.16621   0.47133 -0.14097 -0.01201   0.4949 2.519 0.051046 0.5122   *
9  -0.18957 -0.10354 -0.00457   1.18100 -0.53343   1.3104 1.221 0.320459 0.4208   *
10  0.62010 -0.86503   0.06063 -0.59754   0.02938 -1.0741 1.328 0.221285 0.3890
11 -0.03630 -0.02345   0.02626   0.03096   0.05670   0.1106 1.550 0.002575 0.1687
12  0.01877   0.00840 -0.01899 -0.00553 -0.00860   0.0338 1.453 0.000241 0.1008
13 -0.13432   0.09527 -0.00896   0.05509   0.05799 -0.1476 1.426 0.004559 0.1206
14  0.47171 -0.57701   0.02352   0.96866 -0.92368   1.4971 0.185 0.307053 0.1871   *
15  0.01122   0.04148 -0.04272   0.11031 -0.18086 -0.2318 1.480 0.011184 0.1747
16 -0.09702   0.02839   0.04117   0.03452   0.04936 -0.1064 1.534 0.002380 0.1595
17 -0.23149   0.13945   0.05955   0.10724   0.07078 -0.2444 1.500 0.012434 0.1867
18 -0.01597 -0.02418 -0.15128   0.07692   0.02620 -0.2213 1.452 0.010197 0.1596
19 -0.05128   0.57636 -0.21413 -0.13342   0.04239   0.7226 1.166 0.101556 0.2538
20 -0.05102   0.57350 -0.21307 -0.13276   0.04218   0.7190 1.171 0.100627 0.2538
21  0.03338   0.05517 -0.05380 -0.01284 -0.01797   0.1069 1.443 0.002400 0.1139
22  0.46496 -0.34248   0.14251 -0.24474 -0.24460   0.5278 1.186 0.055224 0.1927
23 -0.07168 -0.00602   0.08441 -0.13309   0.08919 -0.2633 1.065 0.013852 0.0640
24 -0.05327   0.10766   0.18804 -0.24626 -0.00283 -0.4346 1.190 0.037789 0.1593
25  0.06175   0.00286   0.13235 -0.12740 -0.14680 -0.3155 1.449 0.020541 0.1932
>
```

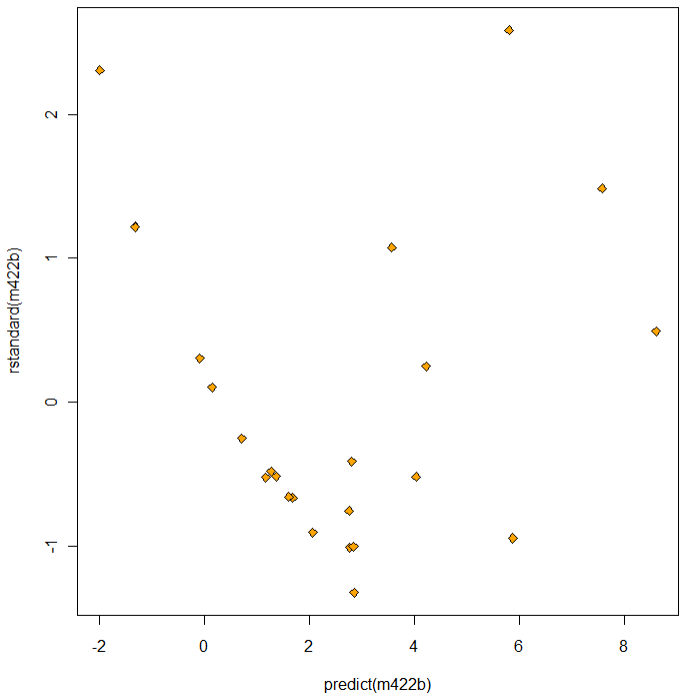
Comment:

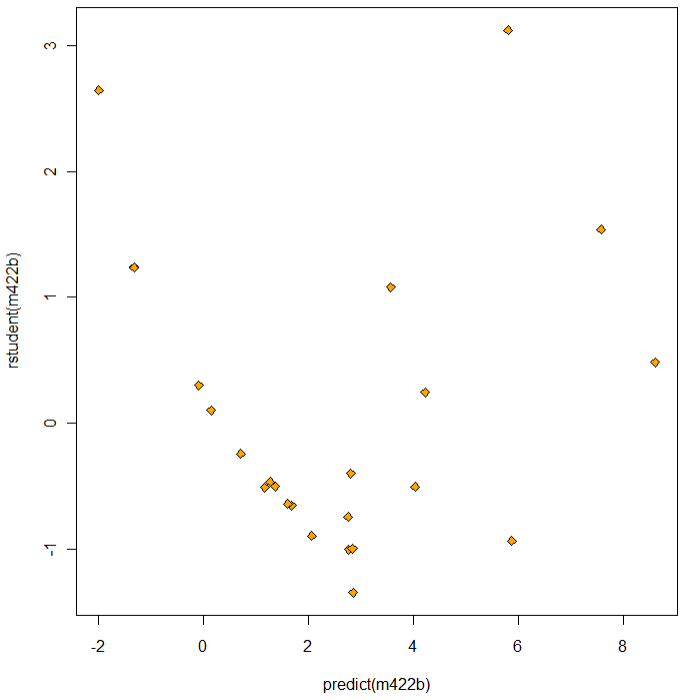
Residual analysis plots and other diagnostic plot suggest the model still does not fit very well. Residuals still do not follow normality and there seems to be non linear patterns. Observations 4, 8, 9, 14 seem to have high influences.

EXERCISE 5.16

a – Plot the original residuals, the studentized residuals and R-student vs the predicted response



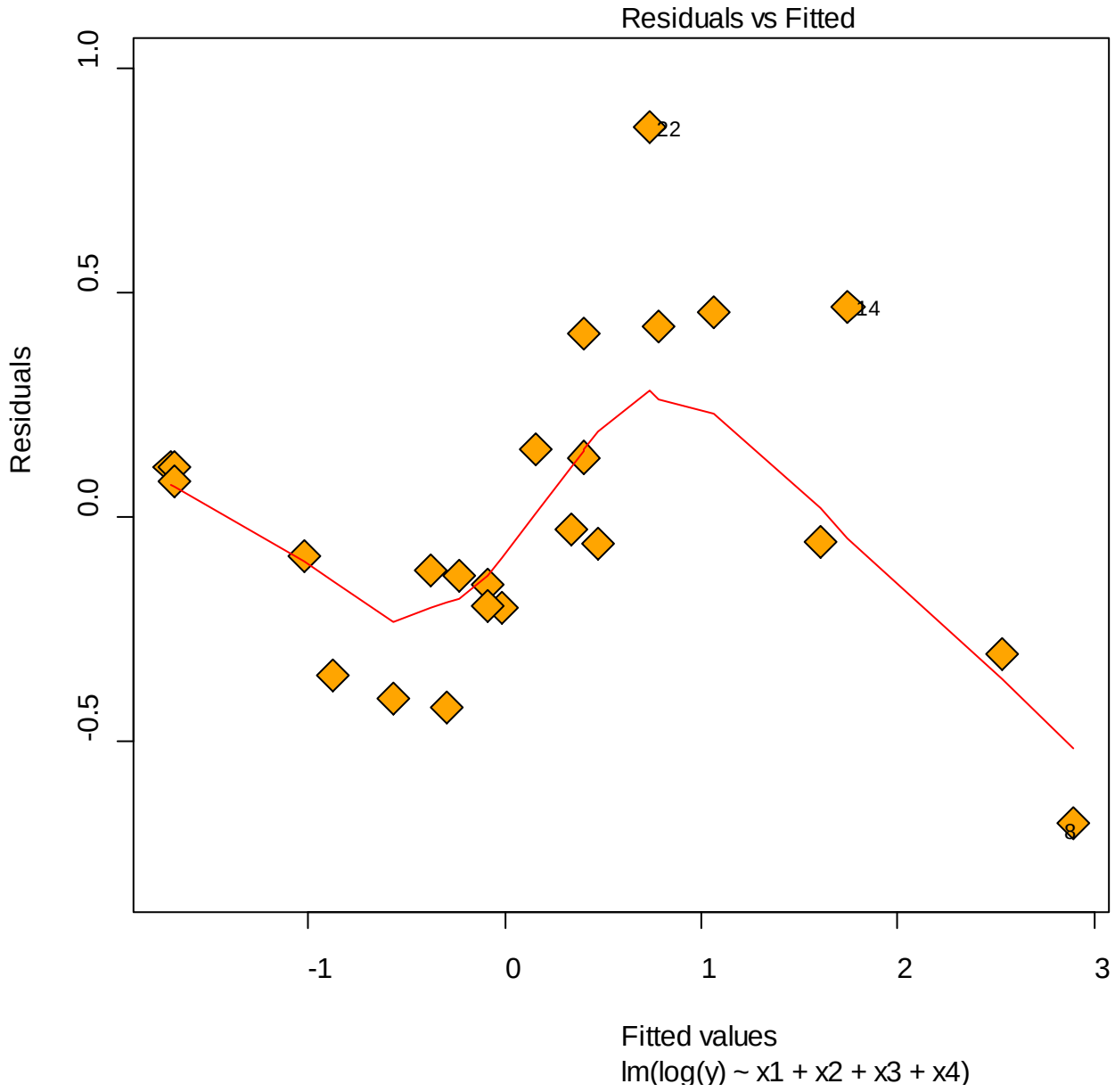


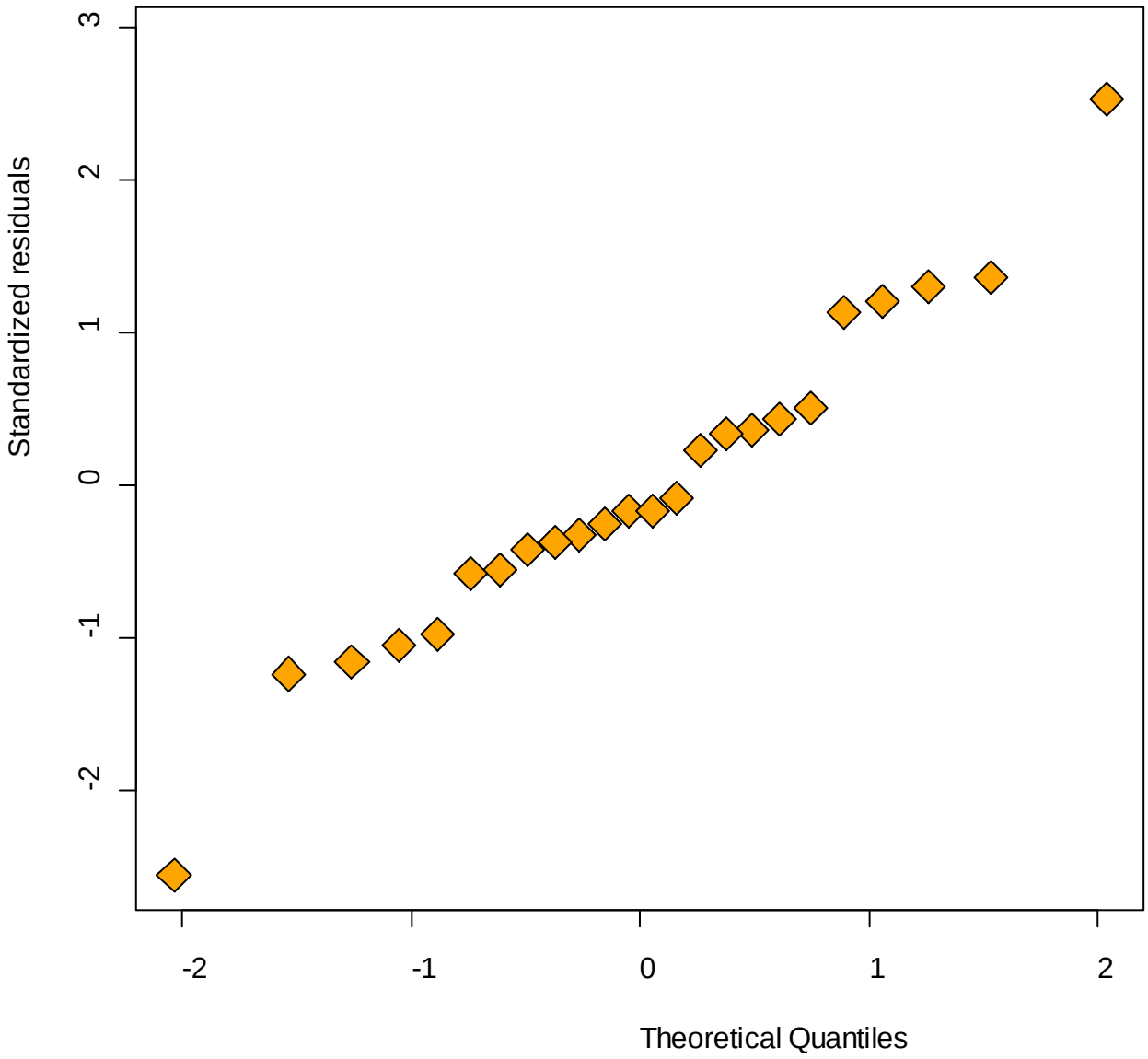


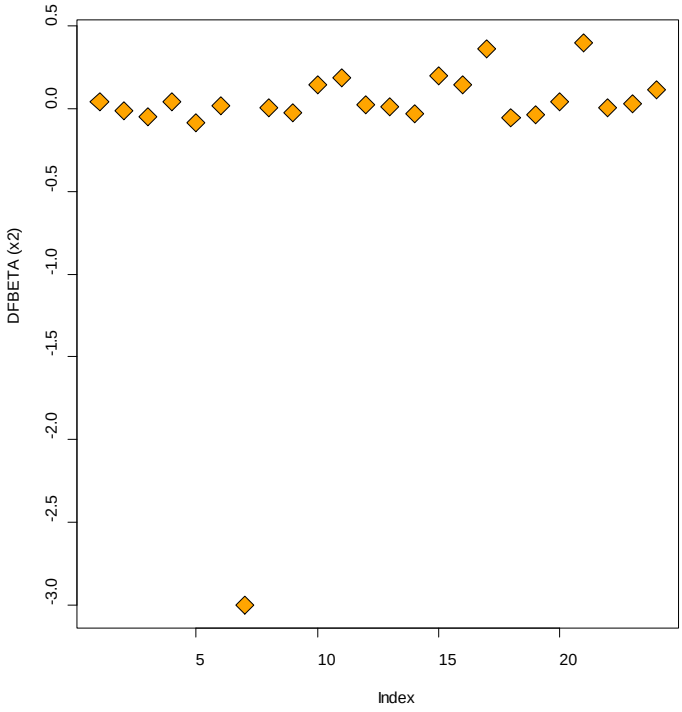
Comment:

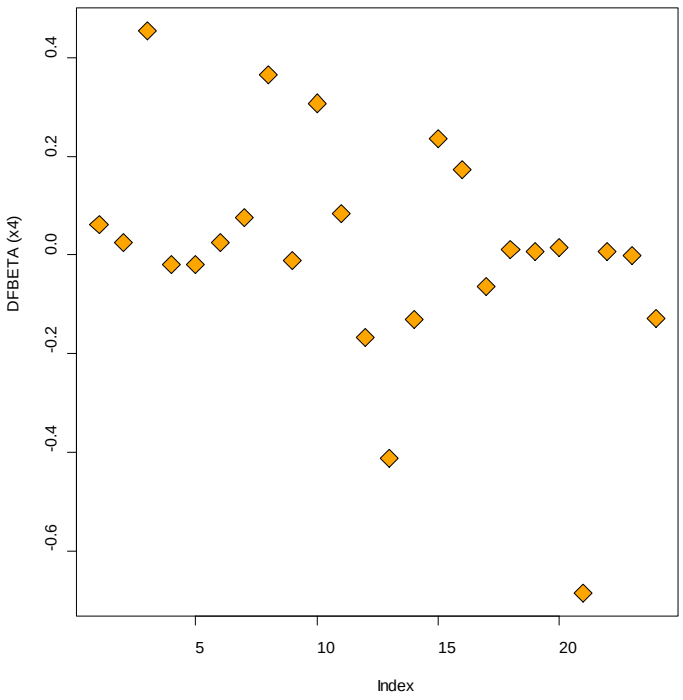
Patterns for residual plots are not satisfactory. There is a non-linear pattern to the residuals.

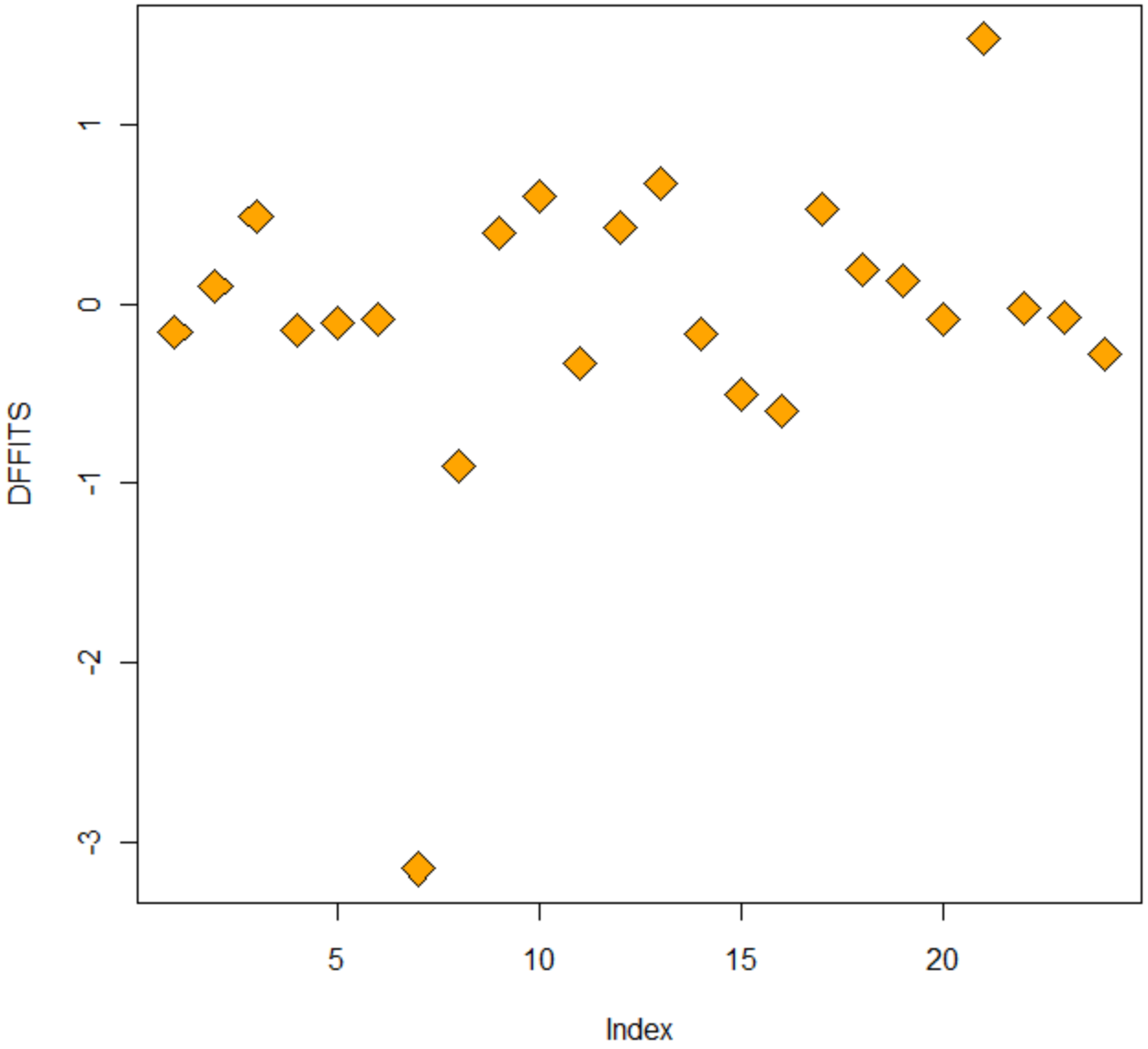
b – From the (somewhat) parabola shape in the residual plot, let us try to transform the response y using log transformation



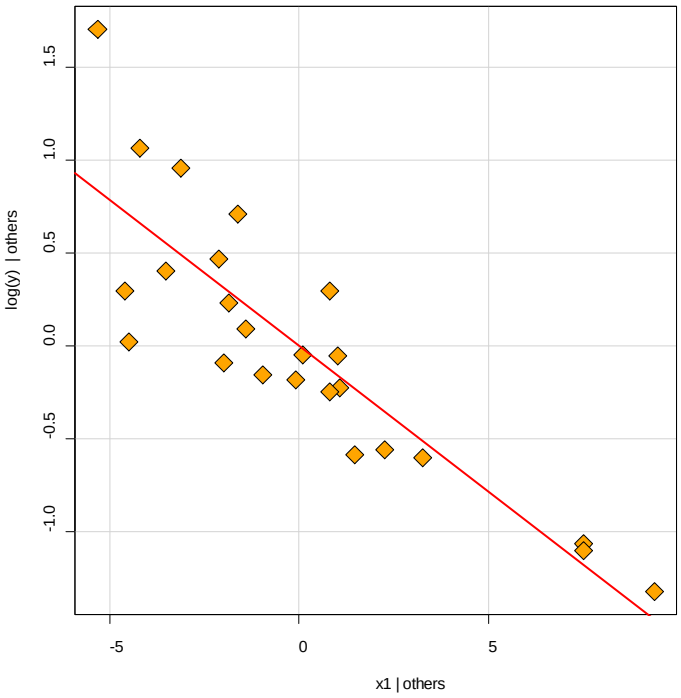

Standardized residuals
3
2
1
0
-1
-2
-2
-1
0
1
2
Theoretical Quantiles

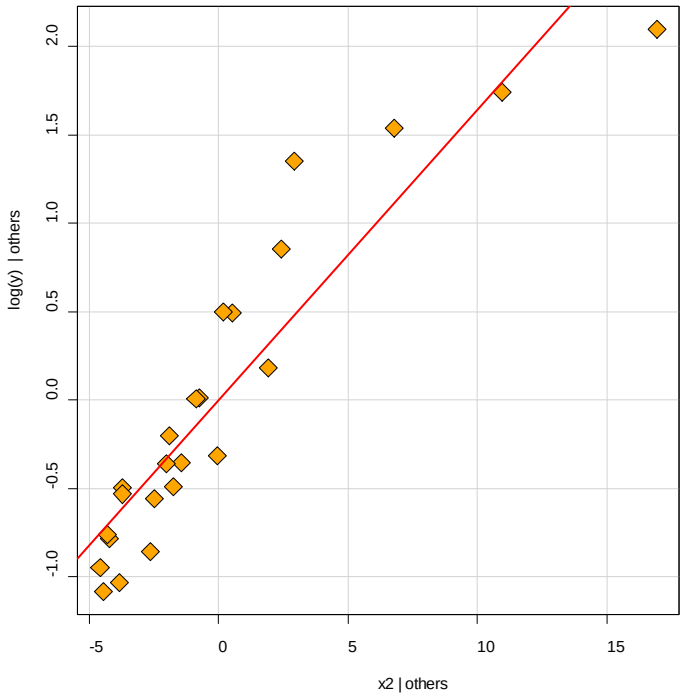

DFBETA (x2)
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
5
10
15
20
Index

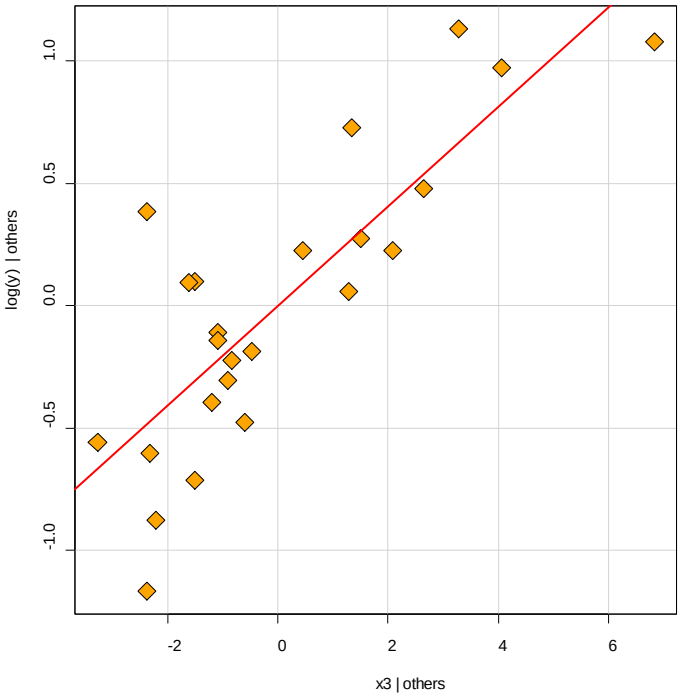

DFBETA (x4)
0.4
0.2
0.0
-0.2
-0.4
-0.6
5
10
15
20
Index

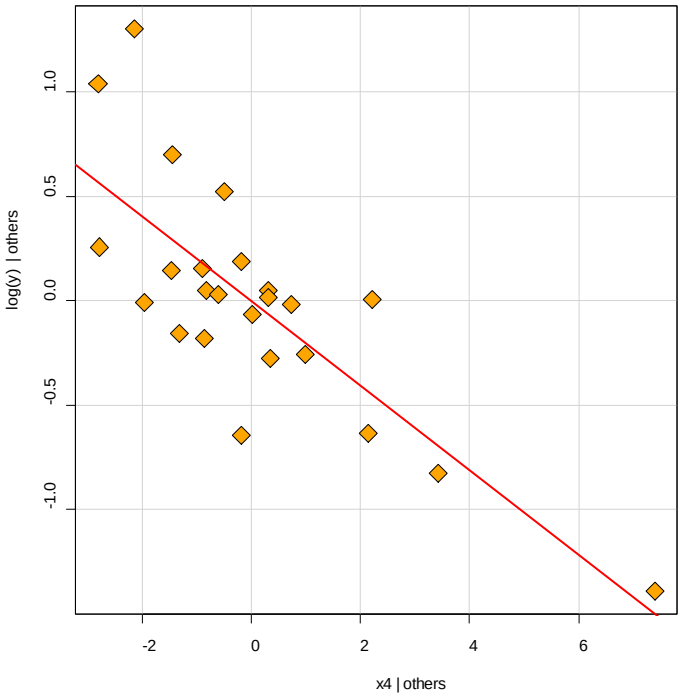

DFFITS
1
0
-1
-2
-3
5
10
15
20
Index

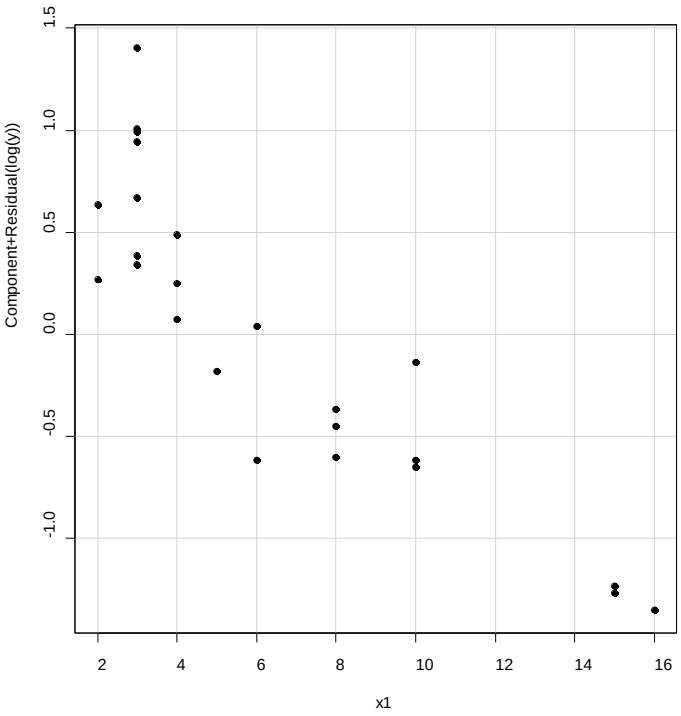
c –using partial regression or partial residuals plots to aid the study of suitable transformation on both response and repressors variables.

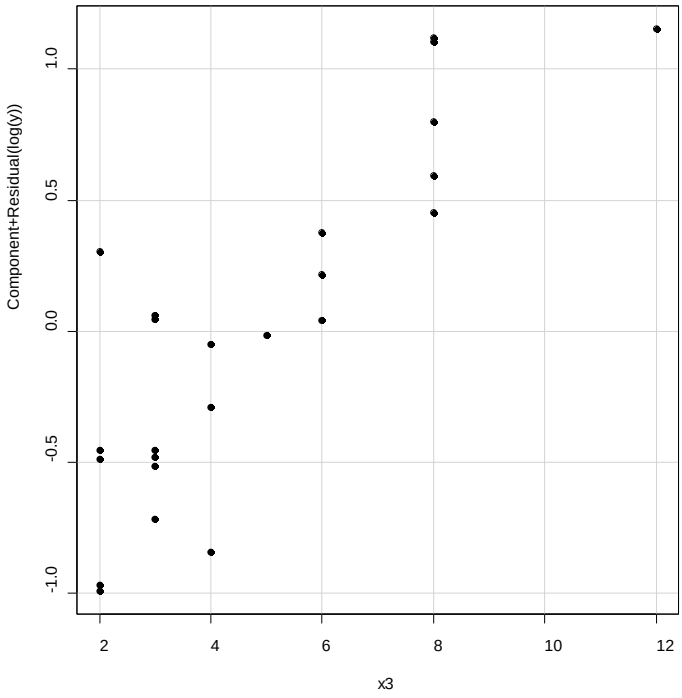




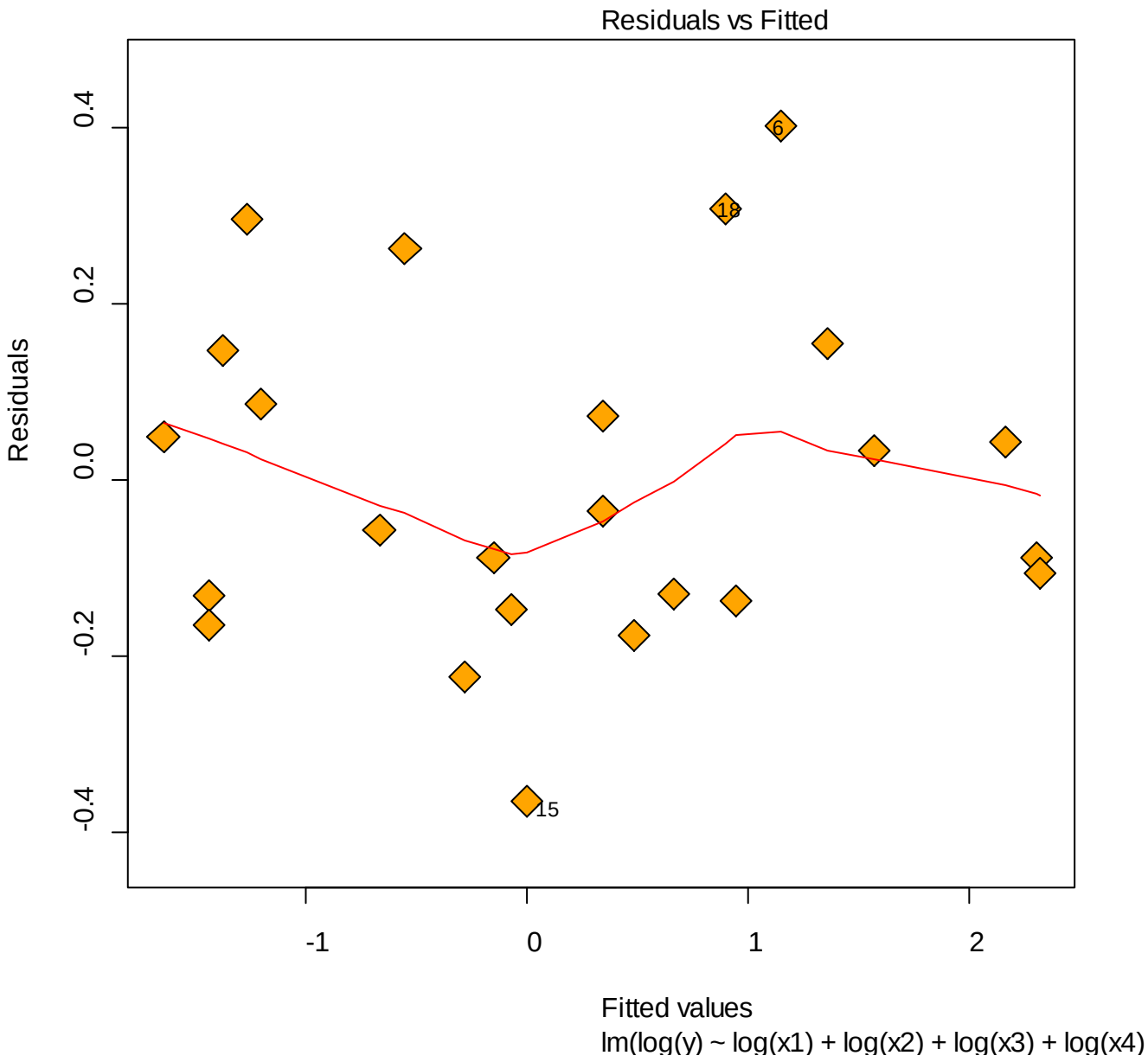

log(y) | others
x3 | others
1.0
0.5
0.0
-0.5
-1.0
-2
0
2
4
6

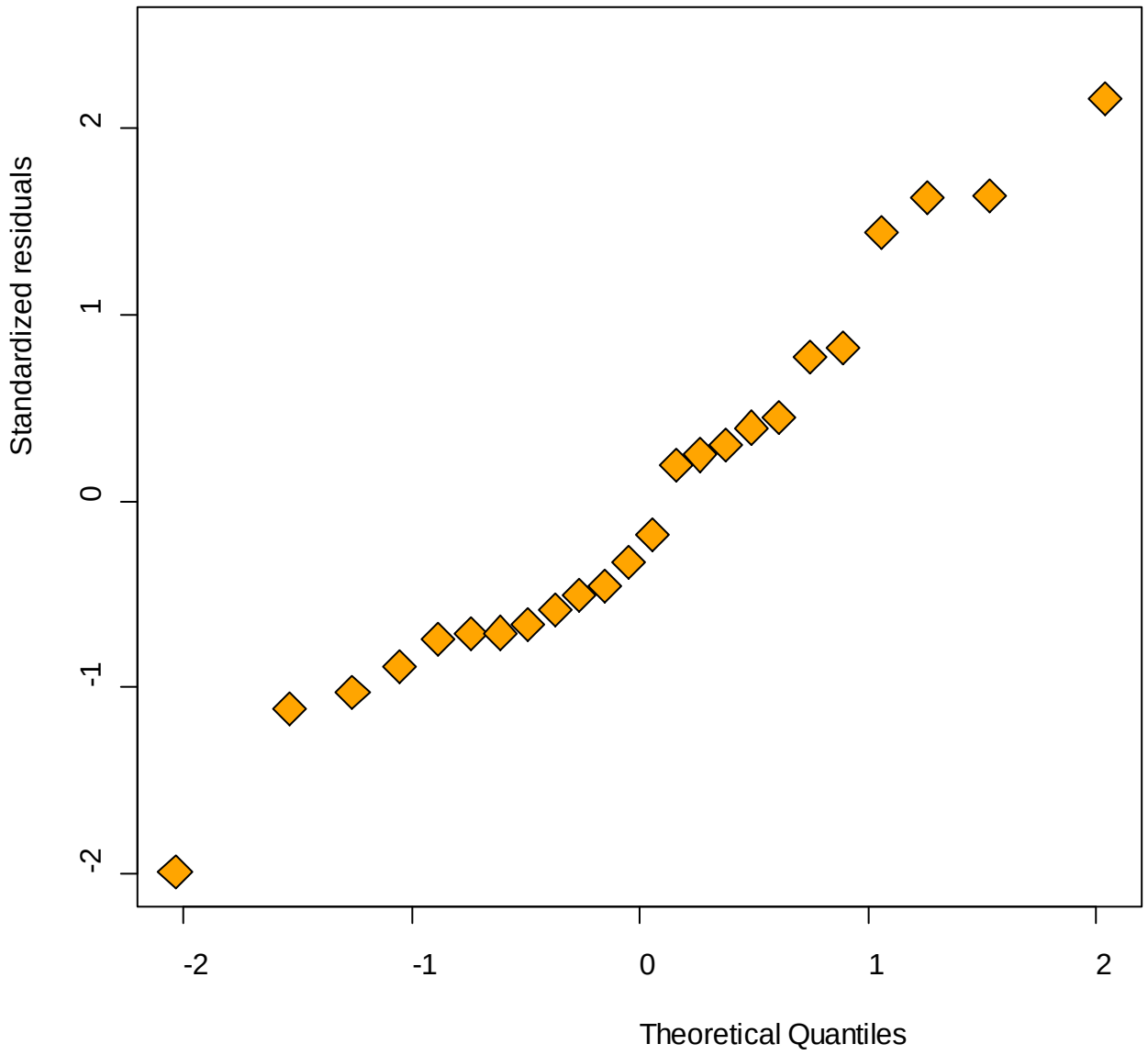

log(y) | others
x4 | others
1.0
0.5
0.0
-0.5
-1.0
-2
0
2
4
6

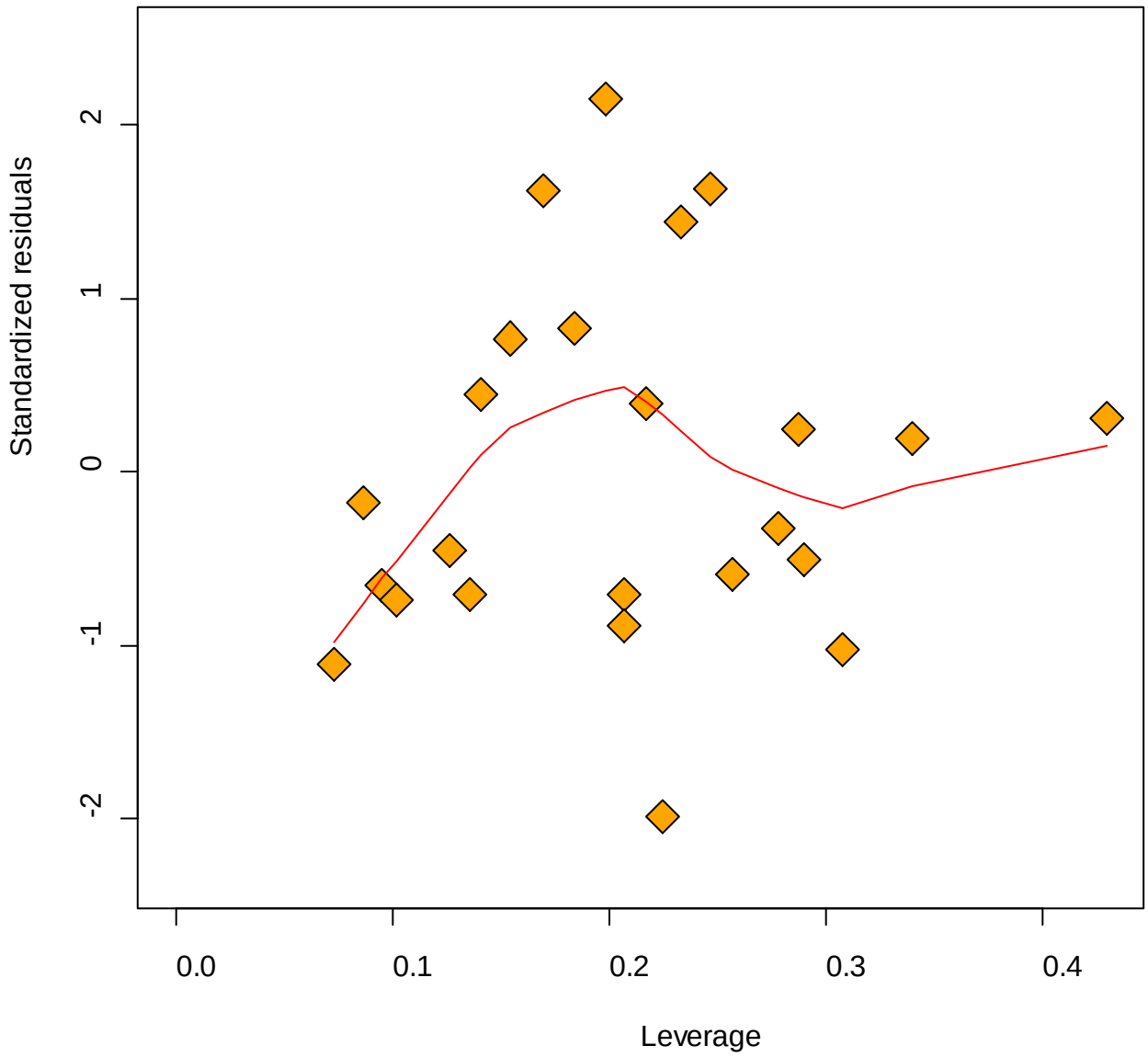

Component+Residual(log(y))
x1
1.5
1.0
0.5
0.0
-0.5
-1.0
2
4
6
8
10
12
14
16



Now using log transformation on each of repressors and also use log transformation on response variable.


Residuals vs Fitted
Residuals
0.4
0.2
0.0
-0.2
-0.4
6
18
15
-1
0
1
2
Fitted values
lm(log(y) ~ log(x1) + log(x2) + log(x3) + log(x4)


Standardized residuals
2
1
0
-1
-2
-2
-1
0
1
2
Theoretical Quantiles



Comment:

After using log transformation both repressors and response variables, the final model looks satisfactory.