Regression Analysis lab 5

1 Transformations and weighting

1.1 Import data

```
windmill<-read.csv(file="D:/chilo/Regression 5/windmill.csv", header=T)
windmill
```

```
   observation velocity DCoutput
1            1     5.00    1.582
2            2     6.00    1.822
3            3     3.40    1.057
4            4     2.70    0.500
5            5    10.00    2.236
6            6     9.70    2.386
7            7     9.55    2.294
8            8     3.05    0.558
9            9     8.15    2.166
10          10     6.20    1.866
11          11     2.90    0.653
12          12     6.35    1.930
13          13     4.60    1.562
14          14     5.80    1.737
15          15     7.40    2.088
16          16     3.60    1.137
17          17     7.85    2.179
18          18     8.80    2.112
19          19     7.00    1.800
20          20     5.45    1.501
21          21     9.10    2.303
22          22    10.20    2.310
23          23     4.10    1.194
24          24     3.95    1.144
25          25     2.45    0.123
```

```
utility<-read.csv(file="D:/chilo/Regression 5/utility.csv", header=T)
utility
```

```
   customer usage demand
1         1   679   0.79
2         2   292   0.44
3         3  1012   0.56
4         4   493   0.79
5         5   582   2.70
```

```
6          6  1156   3.64
7          7   997   4.73
8          8  2189   9.50
9          9  1097   5.34
10        10  2078   6.85
11        11  1818   5.84
12        12  1700   5.21
13        13   747   3.25
14        14  2030   4.43
15        15  1643   3.16
16        16   414   0.50
17        17   354   0.17
18        18  1276   1.88
19        19   745   0.77
20        20   435   1.39
21        21   540   0.56
22        22   874   1.56
23        23  1543   5.28
24        24  1029   0.64
25        25   710   4.00
26        26  1434   0.31
27        27   837   4.20
28        28  1748   4.88
29        29  1381   3.48
30        30  1428   7.58
31        31  1255   2.63
32        32  1777   4.99
33        33   370   0.59
34        34  2316   8.19
35        35  1130   4.79
36        36   463   0.51
37        37   770   1.74
38        38   724   4.10
39        39   808   3.94
40        40   790   0.96
41        41   783   3.29
42        42   406   0.44
43        43  1242   3.24
44        44   658   2.14
45        45  1746   5.71
46        46   468   0.64
47        47  1114   1.90
48        48   413   0.51
49        49  1787   8.33
50        50  3560  14.94
```

```
51        51  1495   5.11
52        52  2221   3.85
53        53  1526   3.93
```

```
food<-read.csv(file="D:/chilo/Regression 5/food.csv", header=T)
food
```

```
   income expense
1   81464    3000
2   72661    3150
3   72344    3085
4   90743    5225
5   98588    5350
6   96507    6090
7  126574    8925
8  114133    9015
9  115814    8885
10 123181    8950
11 131434    9000
12 140564   11345
13 151352   12275
14 146926   12400
15 130963   12525
16 144630   12310
17 147041   13700
18 179021   15000
19 166200   15175
20 180732   14995
21 178187   15050
22 185304   15200
23 155931   15150
24 172579   16800
25 188851   16500
26 192424   17830
27 203112   19500
28 192482   19200
29 218715   19000
30 214317   19350
```

2 windmill data

2.1 linear model

```r
attach(windmill)
wfit1 <- lm(DCoutput ~ velocity, data=windmill)
summary(wfit1)
```

```
Call:
lm(formula = DCoutput ~ velocity, data = windmill)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5987 -0.1410  0.0606  0.1726  0.3218

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.131      0.126    1.04     0.31
velocity       0.241      0.019   12.66  7.5e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.236 on 23 degrees of freedom
Multiple R-squared:  0.874,Adjusted R-squared:  0.869
F-statistic:  160 on 1 and 23 DF,  p-value: 7.55e-12
```

```r
wfit1$fit  # fitted values
```

```
     1      2      3      4      5      6      7      8      9     10
1.3366 1.5778 0.9508 0.7820 2.5424 2.4700 2.4338 0.8664 2.0962 1.6260
    11     12     13     14     15     16     17     18     19     20
0.8302 1.6622 1.2402 1.5295 1.9154 0.9990 2.0239 2.2530 1.8189 1.4451
    21     22     23     24     25
2.3253 2.5906 1.1196 1.0834 0.7217
```

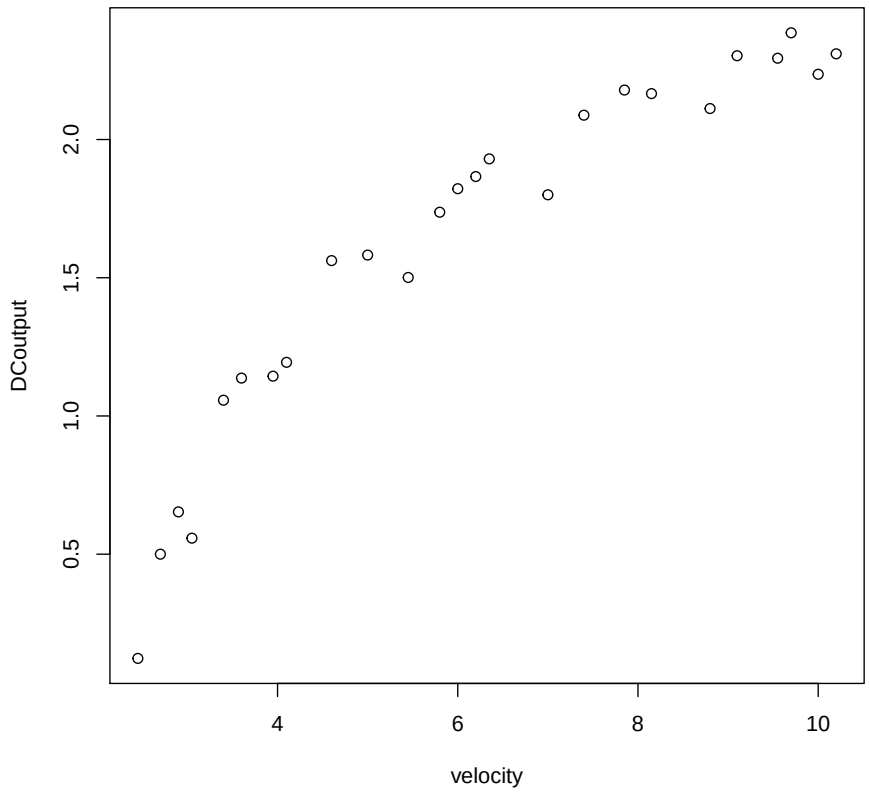
```r
wfit1$res  # residuals
```

```
       1        2        3        4        5        6        7        8
 0.24538  0.24423  0.10622 -0.28198 -0.30636 -0.08402 -0.13985 -0.30838
       9       10       11       12       13       14       15       16
 0.06976  0.24000 -0.17721  0.26783  0.32184  0.20746  0.17262  0.13799
      17       18       19       20       21       22       23       24
 0.15511 -0.14099 -0.01892  0.05586 -0.02233 -0.28059  0.07441  0.06059
      25
-0.59869
```

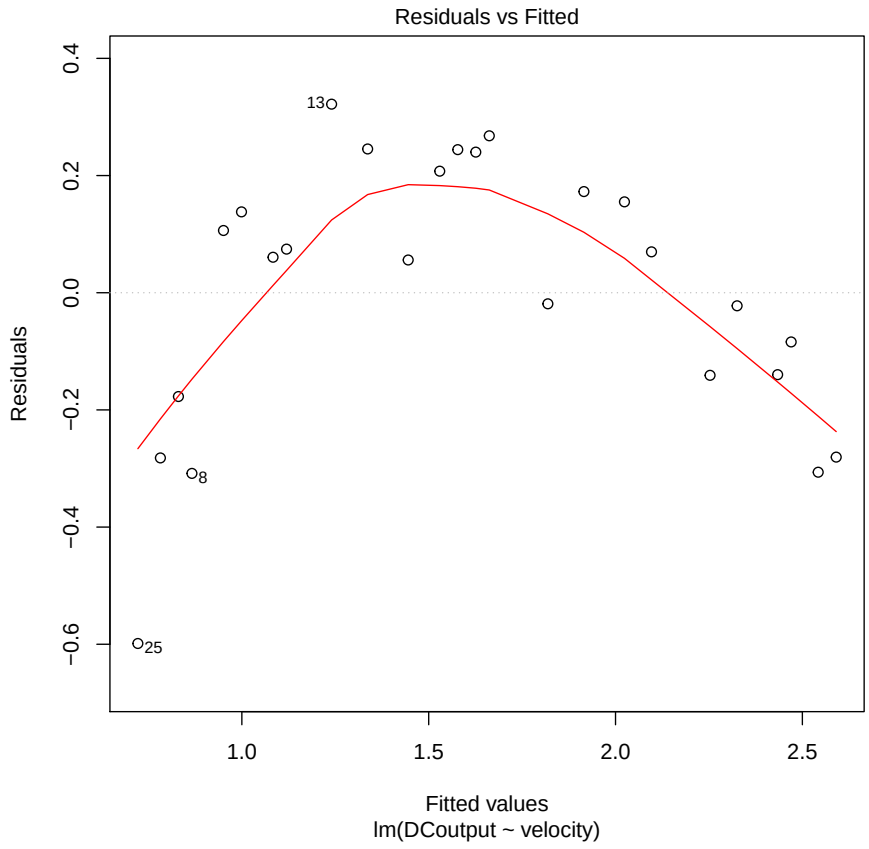
```r
R2_1<-summary(wfit1)$r.squared
R2_1
```

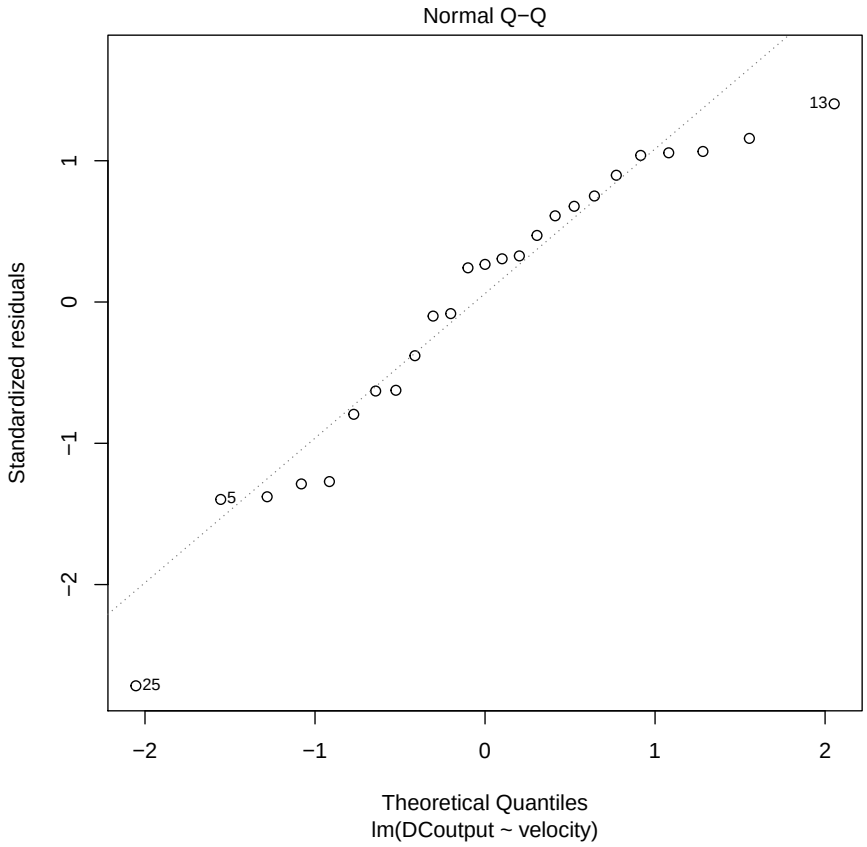
```
[1] 0.8745
```

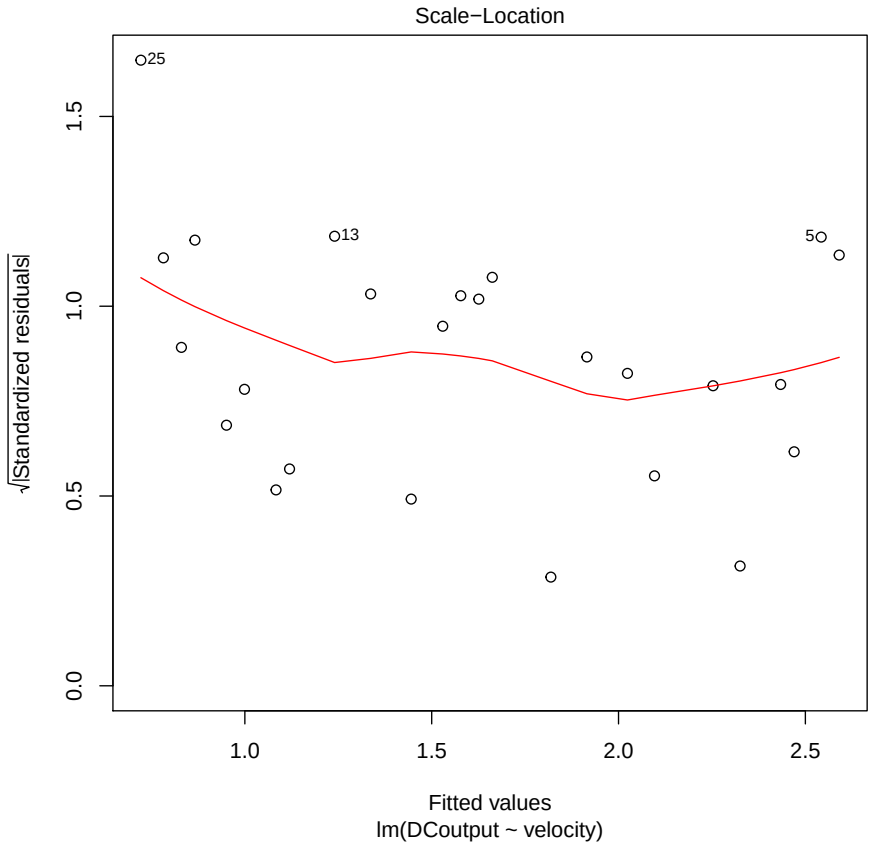
```r
R2.adj_1<-summary(wfit1)$adj.r.squared
R2.adj_1

[1] 0.869

sigmahat_1<-summary(wfit1)$sigma
sigmahat_1

[1] 0.2361

sigmahat2_1<-sigmahat_1^2
MSE_1<-sigmahat2_1
MSE_1

[1] 0.05572

plot(DCoutput ~ velocity)
```

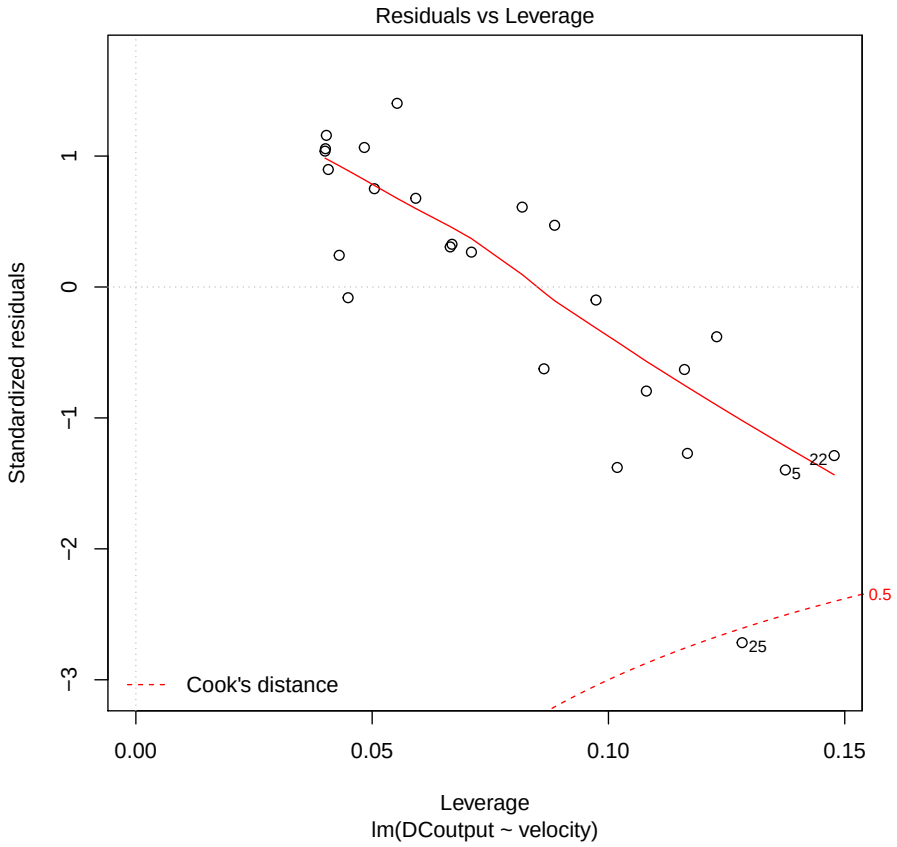


```
plot(wfit1)
```


Residuals vs Fitted
Residuals
0.4
0.2
0.0
−0.2
−0.4
−0.6
13
8
25
1.0
1.5
2.0
2.5
Fitted values
lm(DCoutput ~ velocity)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(DCoutput ~ velocity)
13
5
25
−2
−1
0
1
2


Scale−Location
25
13
5
√|Standardized residuals|
0.0
0.5
1.0
1.5
1.0
1.5
2.0
2.5
Fitted values
lm(DCoutput ~ velocity)



2.2 quadratic model

```
x2<-velocity^2
wfit2 <- lm(DCoutput ~ velocity + x2, data=windmill)
summary(wfit2)


Call:
lm(formula = DCoutput ~ velocity + x2, data = windmill)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2635 -0.0254  0.0126  0.0391  0.1990

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.1559     0.1746   -6.62  1.2e-06 ***
velocity      0.7229     0.0614   11.77  5.8e-11 ***
x2           -0.0381     0.0048   -7.95  6.6e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.123 on 22 degrees of freedom
Multiple R-squared:  0.968,Adjusted R-squared:  0.965
F-statistic:  328 on 2 and 22 DF,  p-value: <2e-16
```

```
wfit2$fit  # fitted values
```

```
     1      2      3      4      5      6      7      8      9     10
1.5058 1.8094 0.8614 0.5181 2.2614 2.2698 2.2714 0.6944 2.2039 1.8609
    11     12     13     14     15     16     17     18     19     20
0.6200 1.8976 1.3630 1.7547 2.1063 0.9526 2.1700 2.2539 2.0367 1.6518
    21     22     23     24     25
2.2660 2.2520 1.1673 1.1049 0.3865
```

```
wfit2$res  # residuals
```

```
        1         2         3         4         5         6         7
 0.076241  0.012635  0.195594 -0.018127 -0.025372  0.116214  0.022580
        8         9        10        11        12        13        14
-0.136437 -0.037945  0.005062  0.032981  0.032385  0.199031 -0.017743
       15        16        17        18        19        20        21
-0.018328  0.184376  0.008956 -0.141856 -0.236730 -0.150817  0.036972
       22        23        24        25
 0.058049  0.026673  0.039083 -0.263474
```

```
R2_2<-summary(wfit2)$r.squared
R2_2
```

```
[1] 0.9676
```

```
R2.adj_2<-summary(wfit2)$adj.r.squared
R2.adj_2
```

```
[1] 0.9646
```

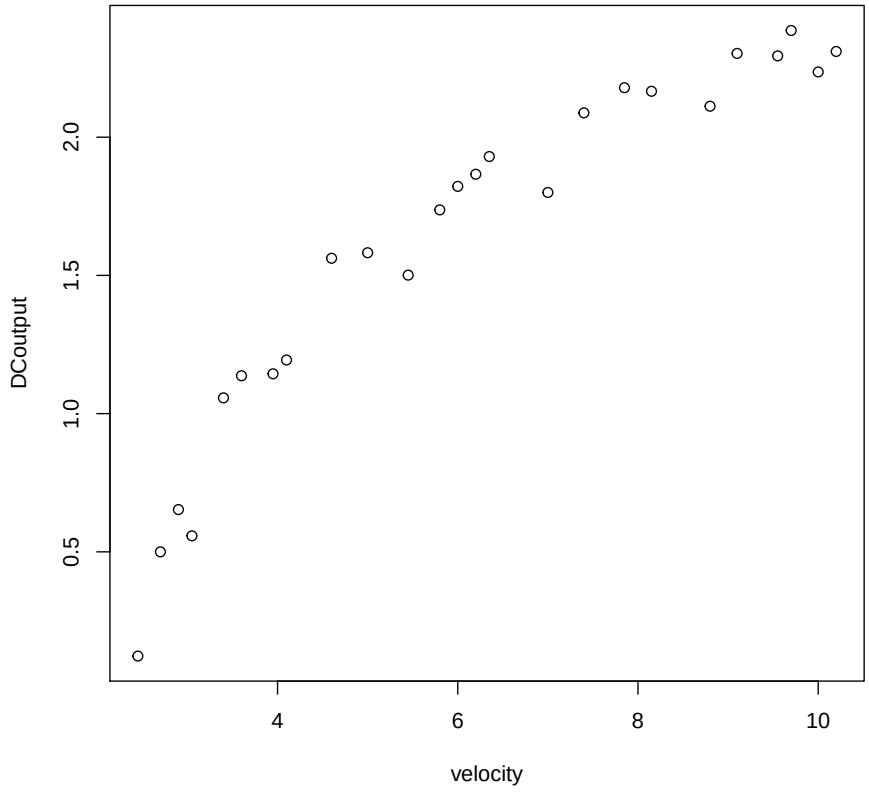
```
sigmahat_2<-summary(wfit2)$sigma
sigmahat_2
```

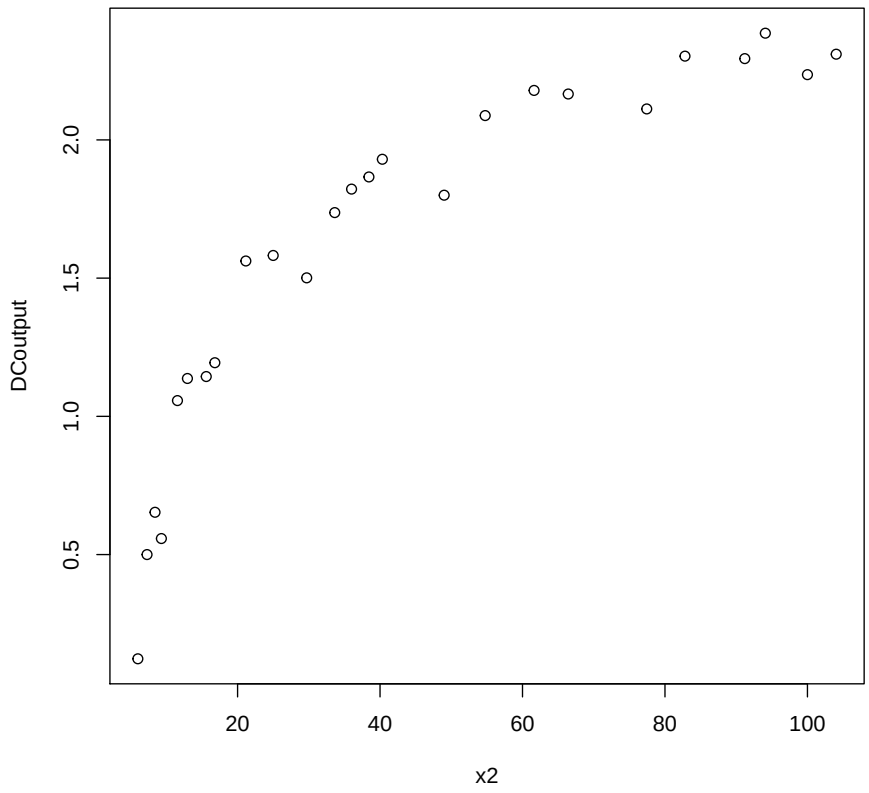
```
[1] 0.1227
```

```
sigmahat2_2<-sigmahat_2^2
MSE_2<-sigmahat2_2
MSE_2
```

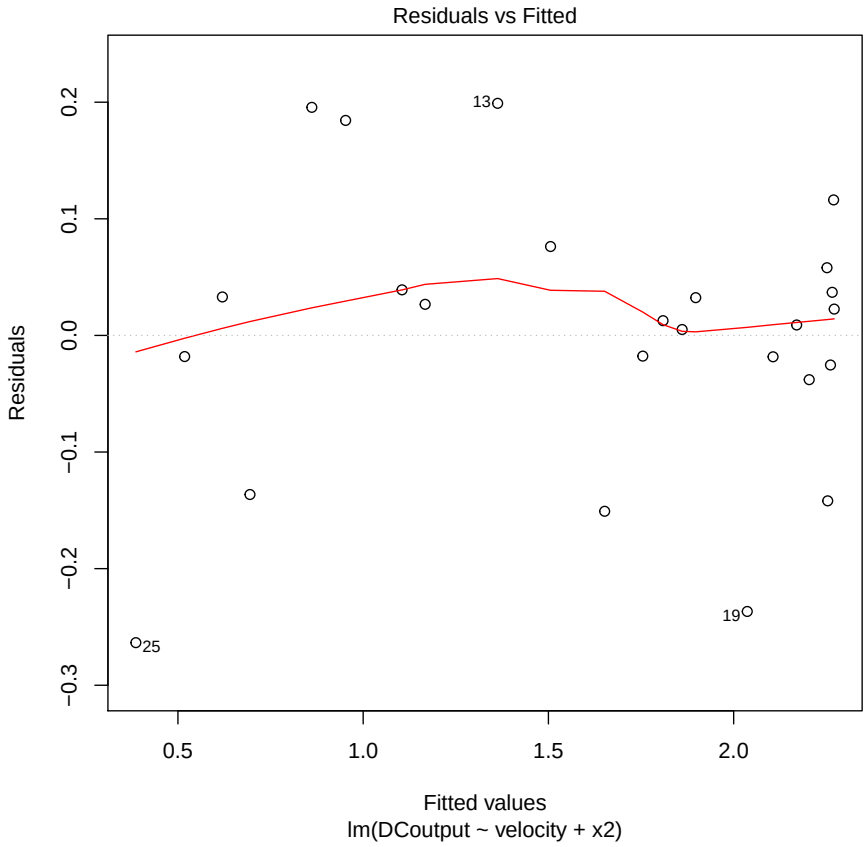
```
[1] 0.01505
```

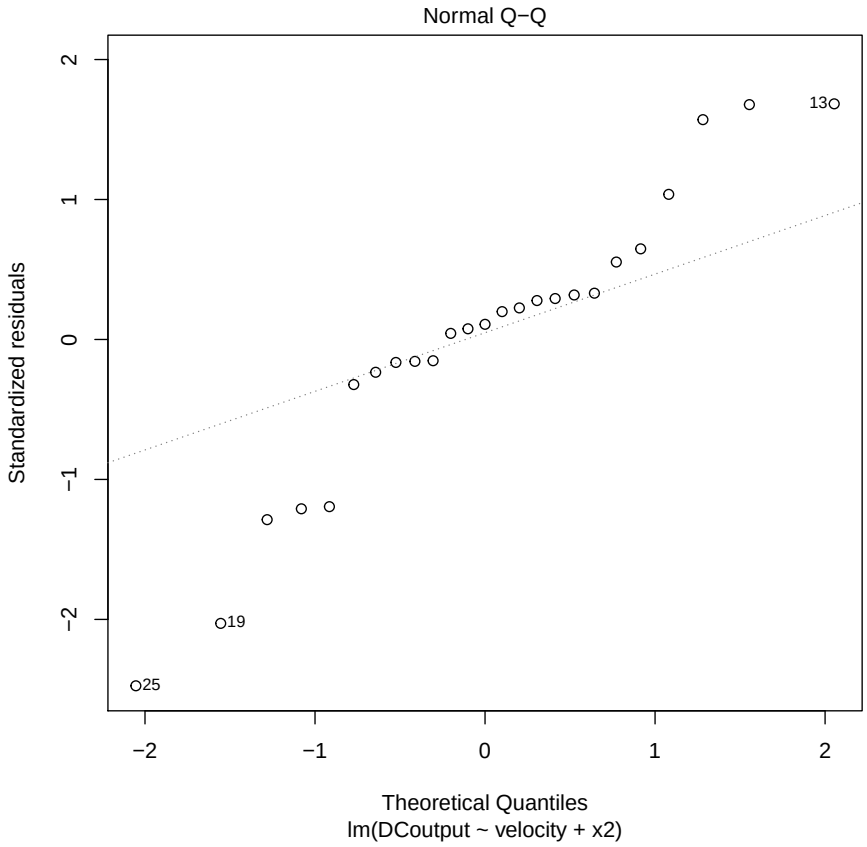
```
plot(DCoutput ~ velocity + x2)
```

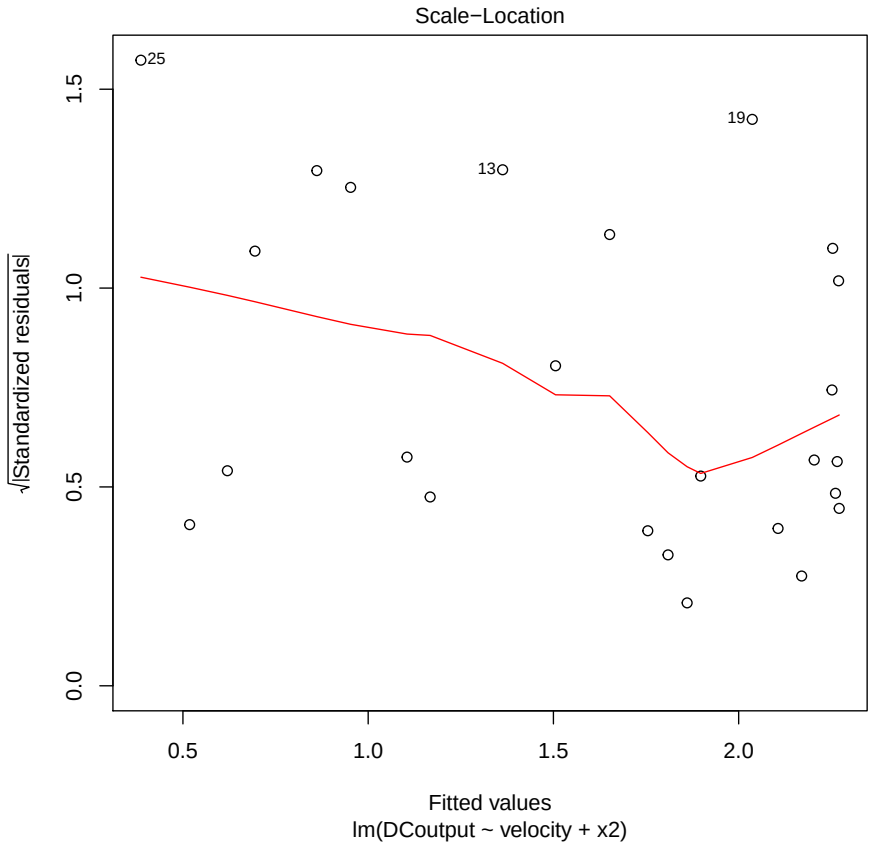


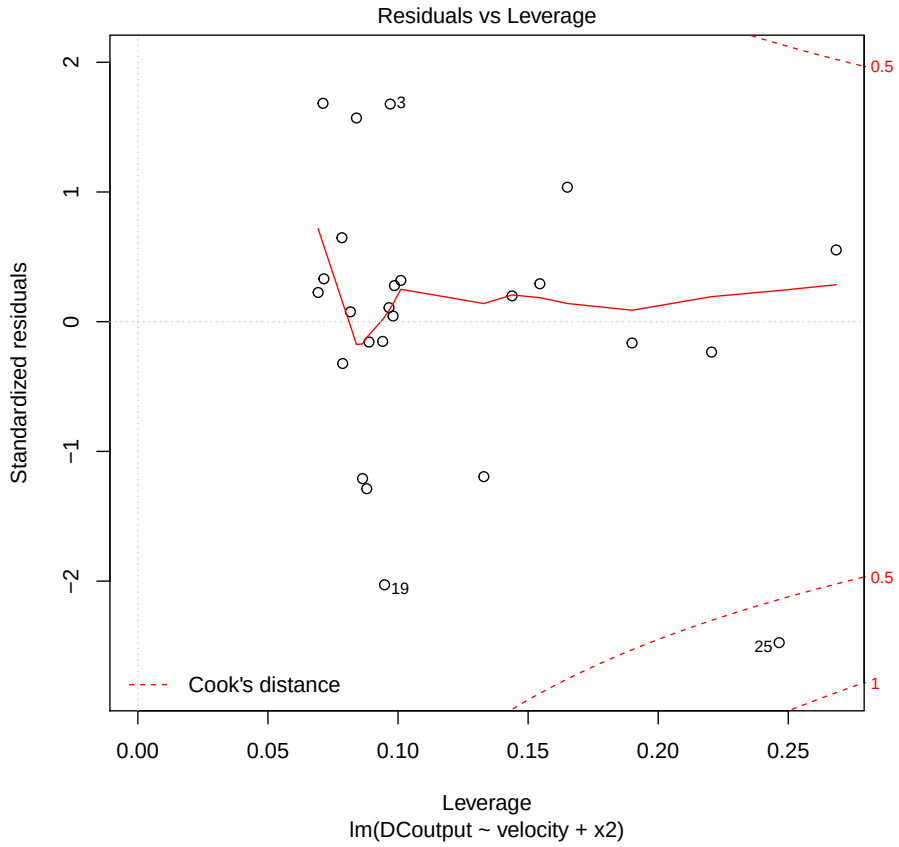



```
plot(wfit2)
```


Residuals vs Fitted
Residuals
0.2
0.1
0.0
−0.1
−0.2
−0.3
13
25
19
0.5
1.0
1.5
2.0
Fitted values
lm(DCoutput ~ velocity + x2)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(DCoutput ~ velocity + x2)
13
19
25
−2
−1
0
1
2


Scale−Location
25
19
13
1.5
1.0
0.5
0.0
√|Standardized residuals|
0.5
1.0
1.5
2.0
Fitted values
lm(DCoutput ~ velocity + x2)



2.3 linear model with x'=1/x

```
x<-1/velocity
wfit3 <- lm(DCoutput ~ x, data=windmill)
summary(wfit3)


Call:
lm(formula = DCoutput ~ x, data = windmill)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2055 -0.0494  0.0110  0.0835  0.1220

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9789     0.0449    66.3   <2e-16 ***
x            -6.9345     0.2064   -33.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0942 on 23 degrees of freedom
Multiple R-squared:  0.98,Adjusted R-squared:  0.979
F-statistic: 1.13e+03 on 1 and 23 DF,  p-value: <2e-16
```

```
wfit3$fit  # fitted values
```

```
     1      2      3      4      5      6      7      8      9     10
1.5920 1.8231 0.9393 0.4105 2.2854 2.2640 2.2527 0.7052 2.1280 1.8604
    11     12     13     14     15     16     17     18     19     20
0.5876 1.8868 1.4713 1.7832 2.0418 1.0526 2.0955 2.1908 1.9882 1.7065
    21     22     23     24     25
2.2168 2.2990 1.2875 1.2233 0.1484
```

```
wfit3$res  # residuals
```

```
        1         2         3         4         5         6         7
-0.009951 -0.001102  0.117713  0.089491 -0.049405  0.122042  0.041270
        8         9        10        11        12        13        14
-0.147238  0.038005  0.005615  0.065363  0.043195  0.090650 -0.046249
       15        16        17        18        19        20        21
 0.046241  0.084403  0.083522 -0.078843 -0.188211 -0.205466  0.086178
       22        23        24        25
 0.010997 -0.093507 -0.079279 -0.025433
```

```
R2_3<-summary(wfit3)$r.squared
R2_3
```

```
[1] 0.98
```

```
R2.adj_3<-summary(wfit3)$adj.r.squared
R2.adj_3
```

```
[1] 0.9792
```

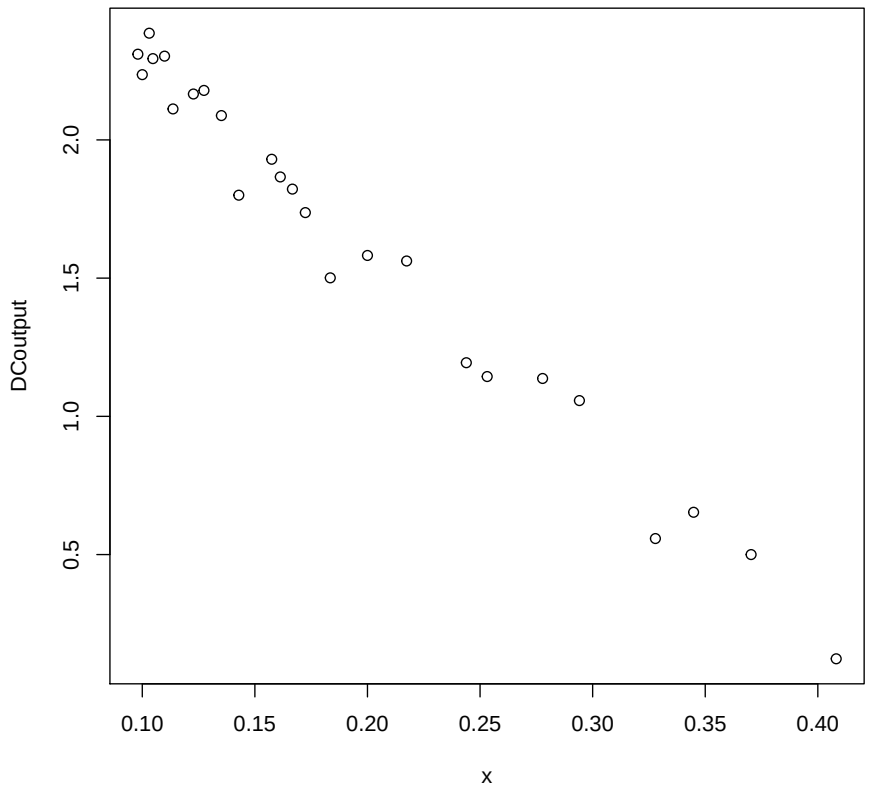
```
sigmahat_3<-summary(wfit3)$sigma
sigmahat_3
```

```
[1] 0.09417
```

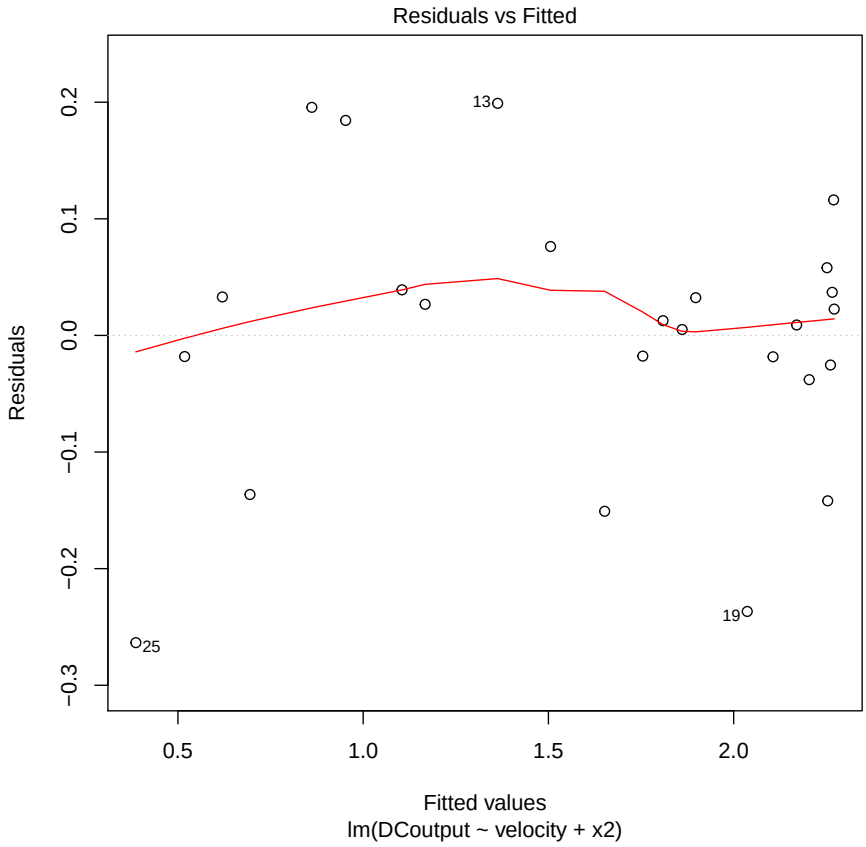
```
sigmahat2_3<-sigmahat_3^2
MSE_3<-sigmahat2_3
MSE_3
```

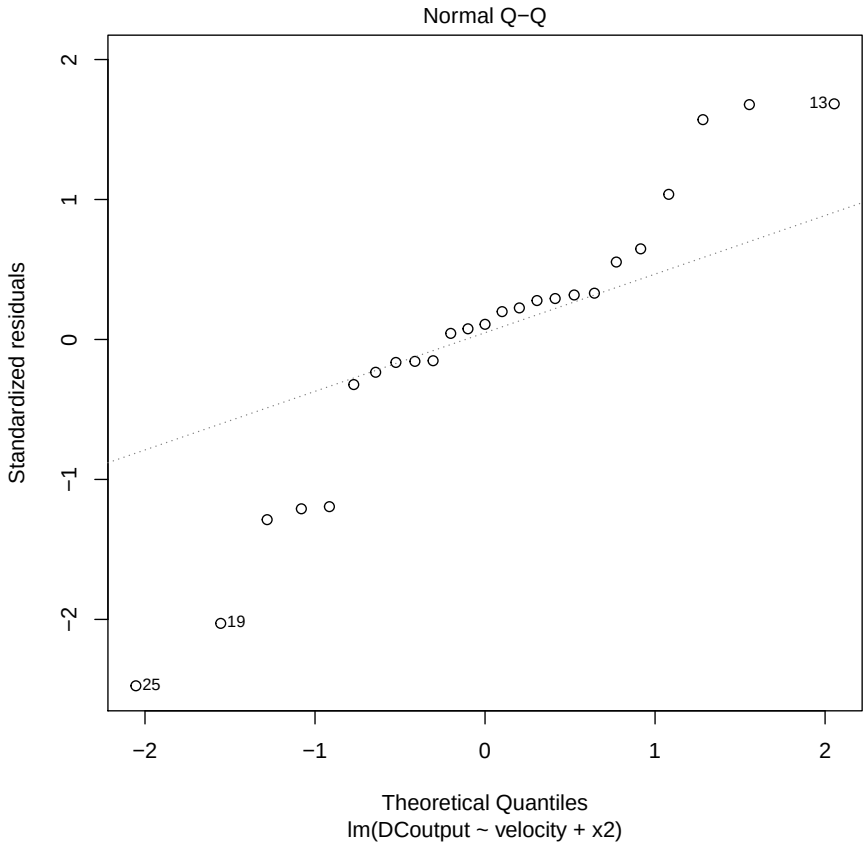
```
[1] 0.008868
```

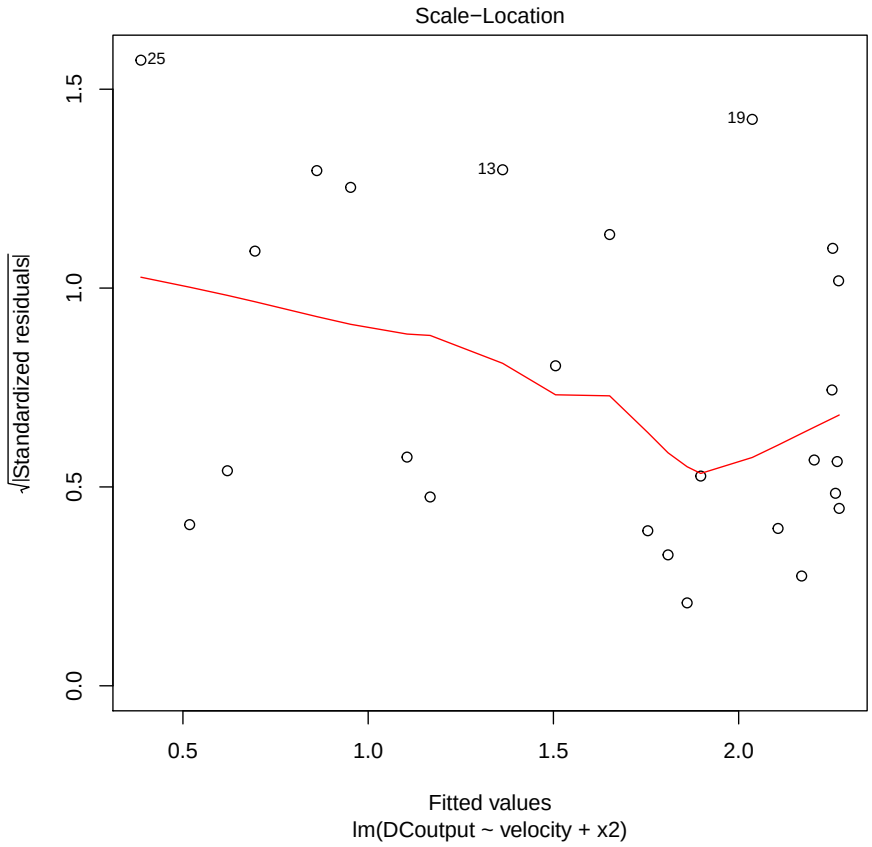
```
plot(DCoutput ~ x)
```

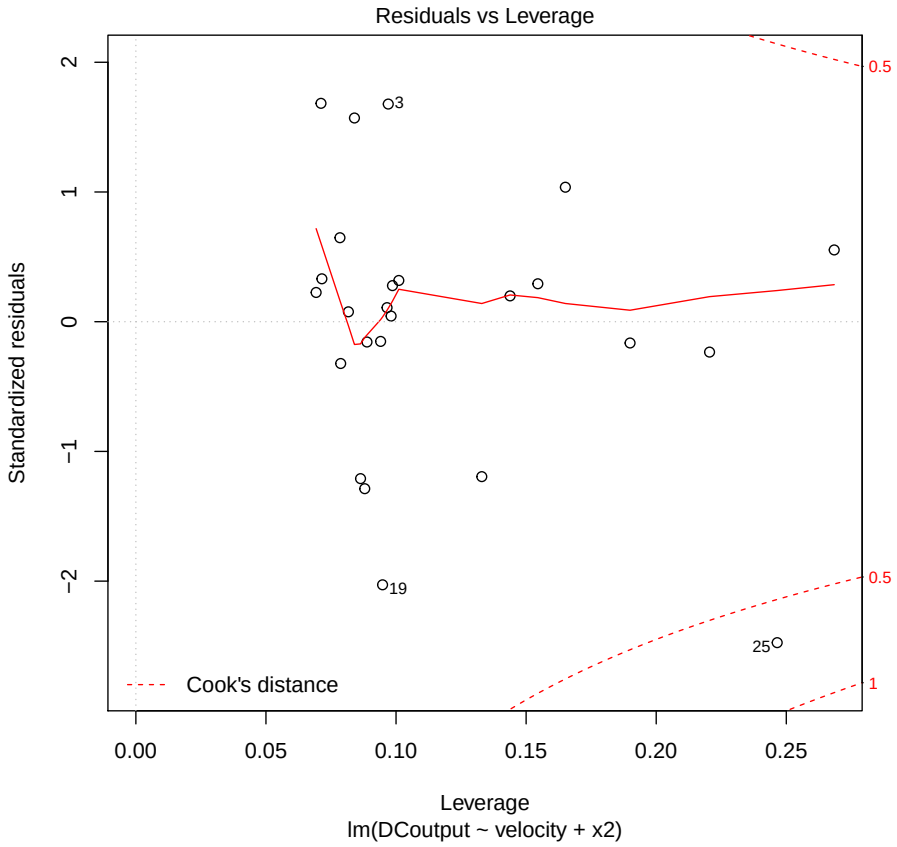



```
plot(wfit2)
```


Residuals vs Fitted
Residuals
0.2
0.1
0.0
−0.1
−0.2
−0.3
13
19
25
0.5
1.0
1.5
2.0
Fitted values
lm(DCoutput ~ velocity + x2)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(DCoutput ~ velocity + x2)
13
19
25
−2
−1
0
1
2


Scale−Location
25
19
13
1.5
1.0
0.5
0.0
√|Standardized residuals|
0.5
1.0
1.5
2.0
Fitted values
lm(DCoutput ~ velocity + x2)



3 electric utility data

3.1 linear model

```
attach(utility)
ufit1 <- lm(demand ~ usage, data=utility)
summary(ufit1)


Call:
lm(formula = demand ~ usage, data = utility)

Residuals:
   Min     1Q Median     3Q    Max
-4.140 -0.828 -0.193  1.238  3.152
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.831304   0.441612   -1.88    0.065 .
usage        0.003683   0.000334   11.03  4.1e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.58 on 51 degrees of freedom
Multiple R-squared:  0.705,Adjusted R-squared:  0.699
F-statistic:  122 on 1 and 51 DF,  p-value: 4.11e-15
```

```
ufit1$fit  # fitted values
```

```
      1       2       3       4       5       6       7       8       9
 1.6693  0.2441  2.8957  0.9843  1.3121  3.4261  2.8405  7.2304  3.2088
     10      11      12      13      14      15      16      17      18
 6.8216  5.8641  5.4295  1.9198  6.6449  5.2196  0.6934  0.4724  3.8680
     19      20      21      22      23      24      25      26      27
 1.9124  0.7707  1.1574  2.3875  4.8513  2.9583  1.7835  4.4499  2.2512
     28      29      30      31      32      33      34      35      36
 5.6063  4.2547  4.4278  3.7907  5.7131  0.5313  7.6982  3.3303  0.8739
     37      38      39      40      41      42      43      44      45
 2.0045  1.8351  2.1444  2.0781  2.0524  0.6639  3.7428  1.5920  5.5989
     46      47      48      49      50      51      52      53
 0.8923  3.2714  0.6897  5.7499 12.2796  4.6745  7.3483  4.7887
```

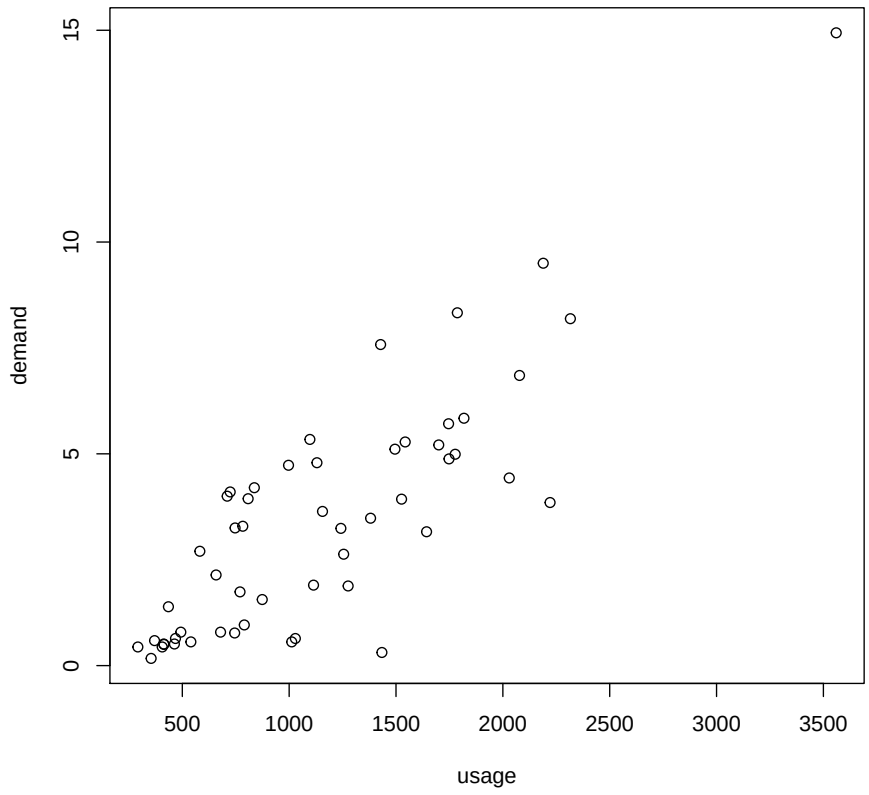
```
ufit1$res  # residuals
```

```
       1        2        3        4        5        6        7        8
-0.87935  0.19591 -2.33573 -0.19434  1.38789  0.21394  1.88951  2.26956
       9       10       11       12       13       14       15       16
 2.13122  0.02836 -0.02410 -0.21953  1.33022 -2.21487 -2.05961 -0.19339
      17       18       19       20       21       22       23       24
-0.30242 -1.98800 -1.14241  0.61927 -0.59743 -0.82750  0.42868 -2.31834
      25       26       27       28       29       30       31       32
 2.21649 -4.13989  1.94876 -0.72631 -0.77470  3.15220 -1.16066 -0.72311
      33       34       35       36       37       38       39       40
 0.05865  0.49184  1.45969 -0.36385 -0.26449  2.26493  1.79557 -1.11814
      41       42       43       44       45       46       47       48
 1.23764 -0.22393 -0.50279  0.54799  0.11106 -0.25227 -1.37138 -0.17971
      49       50       51       52       53
 2.58006  2.66038  0.43545 -3.49829 -0.85871
```

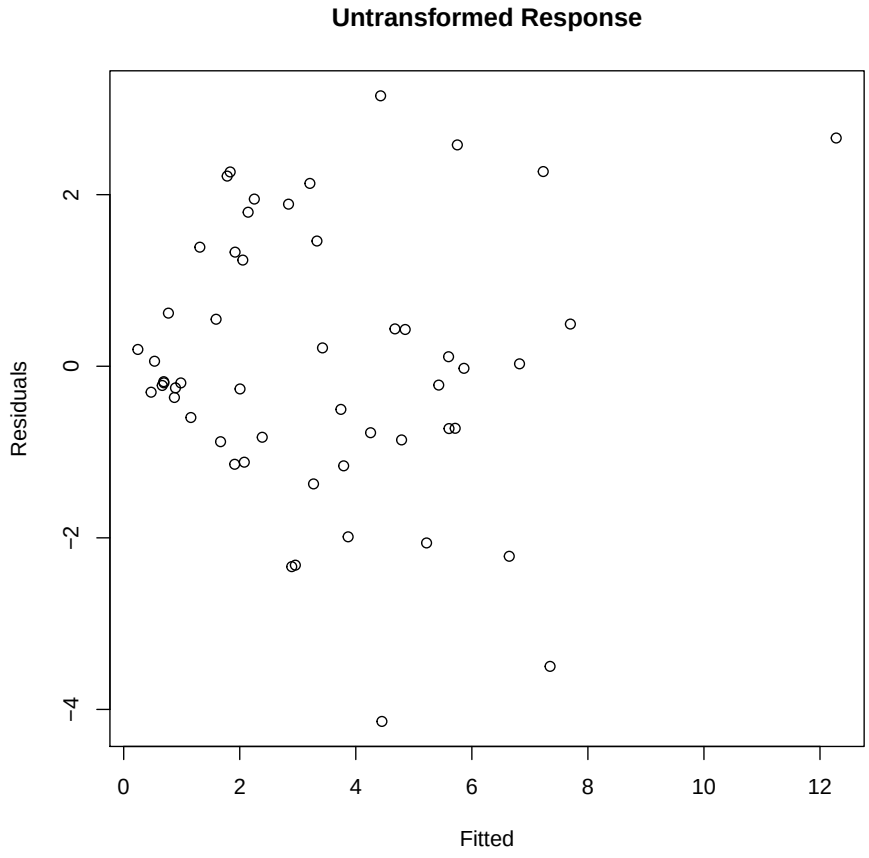
```
R2_1<-summary(ufit1)$r.squared
R2_1
```

```
[1] 0.7046
```

```
R2.adj_1<-summary(ufit1)$adj.r.squared
R2.adj_1

[1] 0.6988

sigmahat_1<-summary(ufit1)$sigma
sigmahat_1

[1] 1.577

sigmahat2_1<-sigmahat_1^2
MSE_1<-sigmahat2_1
MSE_1

[1] 2.488

plot(demand ~ usage)
```



```
plot(ufit1$fit,ufit1$res,xlab="Fitted",ylab="Residuals",
main="Untransformed Response")
```



3.2 linear model with yy=sqrt(y)

```
ufit2 <- lm(sqrt(demand) ~ usage, data=utility)
summary(ufit2)


Call:
lm(formula = sqrt(demand) ~ usage, data = utility)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3918 -0.3058 -0.0388  0.2538  0.8103
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.82e-01   1.30e-01    4.48  4.2e-05 ***
usage       9.53e-04   9.82e-05    9.70  3.6e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.464 on 51 degrees of freedom
Multiple R-squared:  0.648,Adjusted R-squared:  0.642
F-statistic: 94.1 on 1 and 51 DF,  p-value: 3.61e-13
```

```
ufit2$fit  # fitted values
```

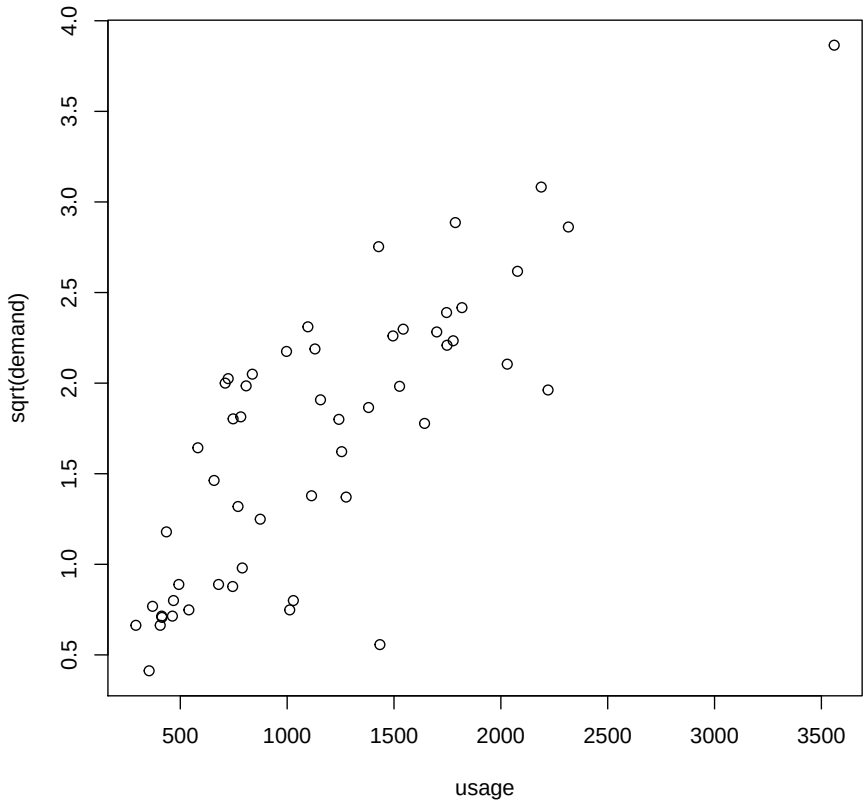
```
     1      2      3      4      5      6      7      8      9     10
1.2292 0.8605 1.5465 1.0520 1.1368 1.6837 1.5322 2.6680 1.6275 2.5623
    11     12     13     14     15     16     17     18     19     20
2.3145 2.2021 1.2940 2.5165 2.1478 0.9767 0.9195 1.7981 1.2921 0.9967
    21     22     23     24     25     26     27     28     29     30
1.0968 1.4150 2.0525 1.5627 1.2588 1.9486 1.3798 2.2478 1.8981 1.9429
    31     32     33     34     35     36     37     38     39     40
1.7781 2.2755 0.9348 2.7890 1.6590 1.0234 1.3159 1.2721 1.3521 1.3350
    41     42     43     44     45     46     47     48     49     50
1.3283 0.9691 1.7657 1.2092 2.2459 1.0282 1.6437 0.9758 2.2850 3.9744
    51     52     53
2.0068 2.6985 2.0363
```

```
ufit2$res  # residuals
```

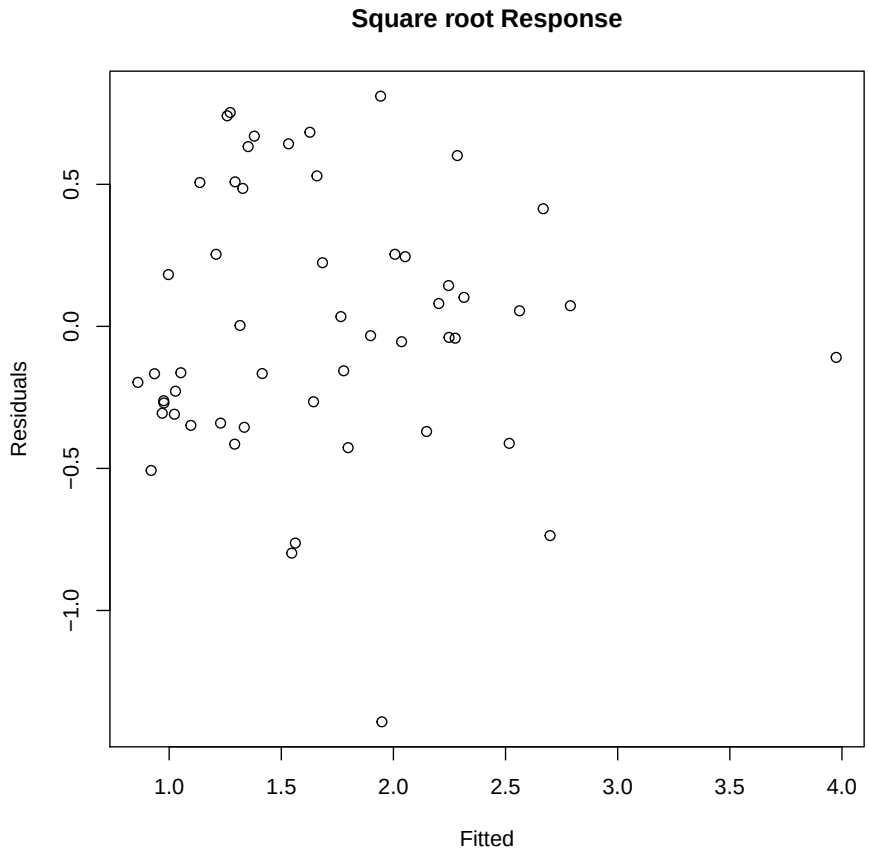
```
        1         2         3         4         5         6         7
-0.340398 -0.197136 -0.798188 -0.163166  0.506378  0.224148  0.642630
        8         9        10        11        12        13        14
 0.414173  0.683332  0.054984  0.102086  0.080456  0.508764 -0.411773
       15        16        17        18        19        20        21
-0.370134 -0.269603 -0.507227 -0.426943 -0.414609  0.182263 -0.348438
       22        23        24        25        26        27        28
-0.166025  0.245338 -0.762718  0.741244 -1.391849  0.669621 -0.038751
       29        30        31        32        33        34        35
-0.032648  0.810271 -0.156336 -0.041625 -0.166669  0.072770  0.529650
       36        37        38        39        40        41        42
-0.309257  0.003163  0.752750  0.632807 -0.355189  0.485521 -0.305762
       43        44        45        46        47        48        49
 0.034323  0.253667  0.143643 -0.228164 -0.265306 -0.261614  0.601189
       50        51        52        53
-0.109174  0.253781 -0.736384 -0.053866
```

```
R2_2<-summary(ufit2)$r.squared
R2_2
```

```
[1] 0.6485

R2.adj_2<-summary(ufit2)$adj.r.squared
R2.adj_2

[1] 0.6416

sigmahat_2<-summary(ufit2)$sigma
sigmahat_2

[1] 0.464

sigmahat2_2<-sigmahat_2^2
MSE_2<-sigmahat2_2
MSE_2

[1] 0.2153

plot(sqrt(demand) ~ usage)
```

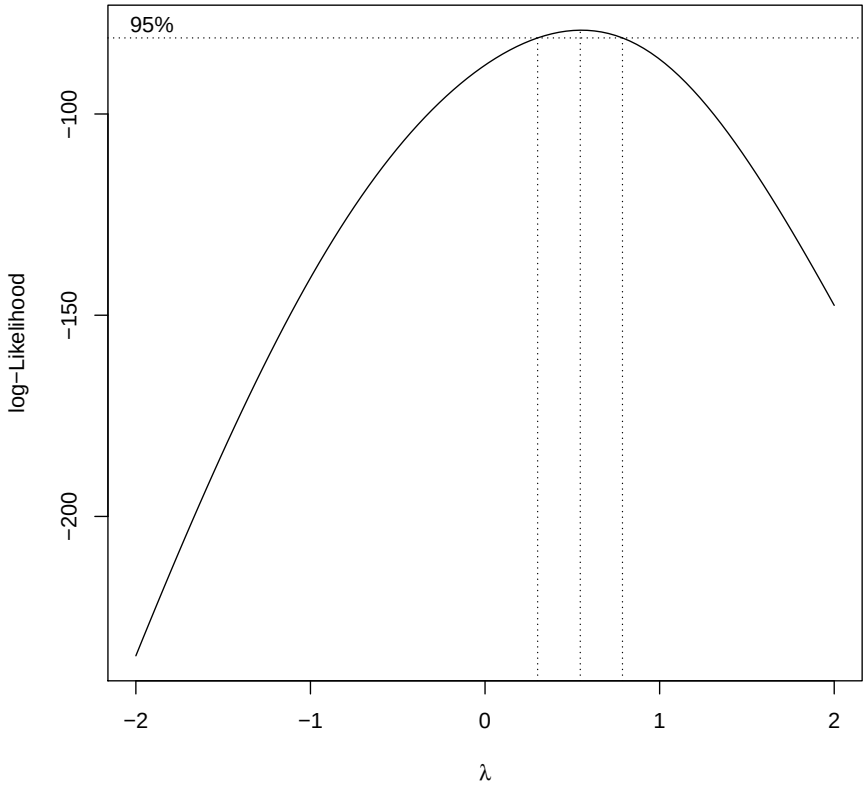



```
plot(ufit2$fit,ufit2$res,xlab="Fitted",ylab="Residuals",
main="Square root Response")
```

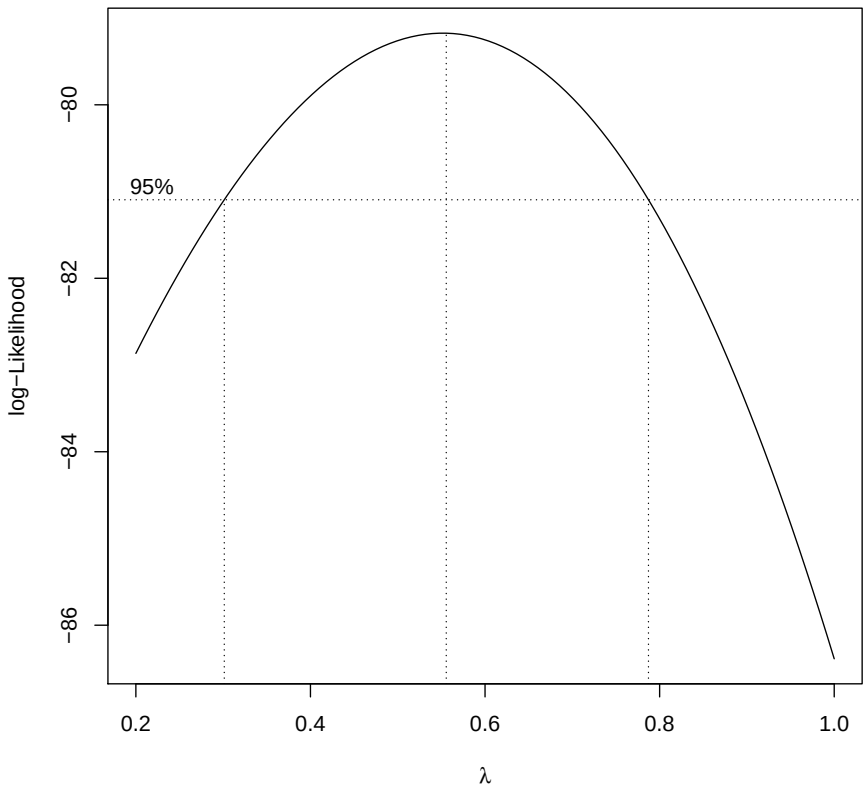


4 Box-Cox transformation

```
attach(utility)

## The following objects are masked from utility (position 3):
##
##     customer, demand, usage

library(MASS)

## Warning:  package 'MASS' was built under R version 3.0.3

ufit1 <- lm(demand ~ usage, data=utility)
boxcox(ufit1,plotit=T)
```




```
# the plot is too broad
boxcox(ufit1,plotit=T,lambda=seq(0.2,1.0,by=0.1))
```

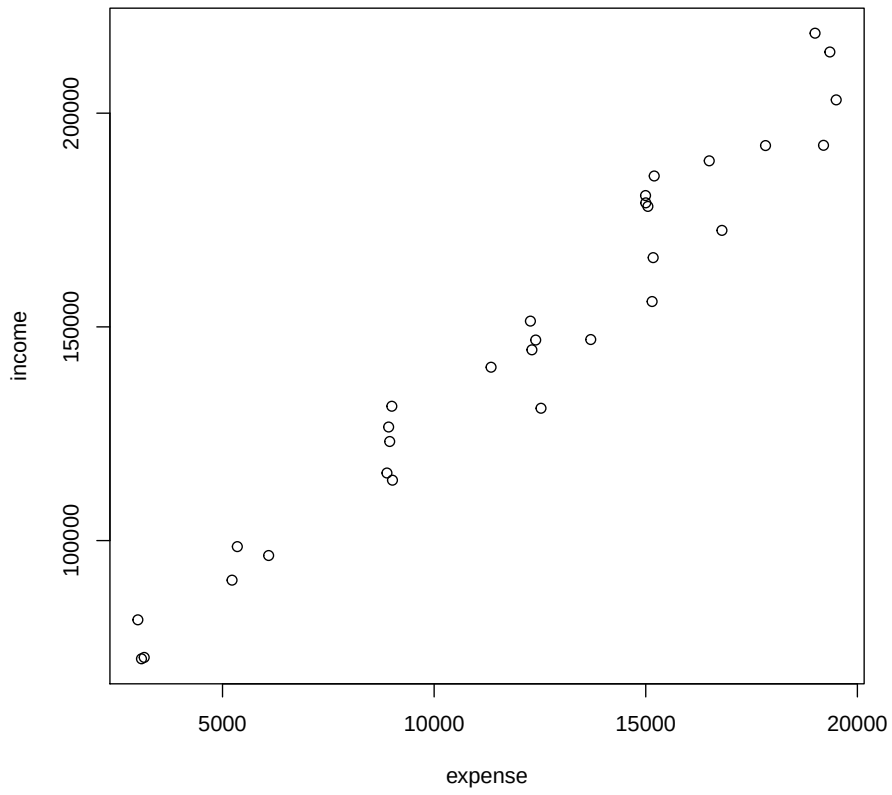



```
# the 95% C.I. for lambda is [0.3, 0.78]
```

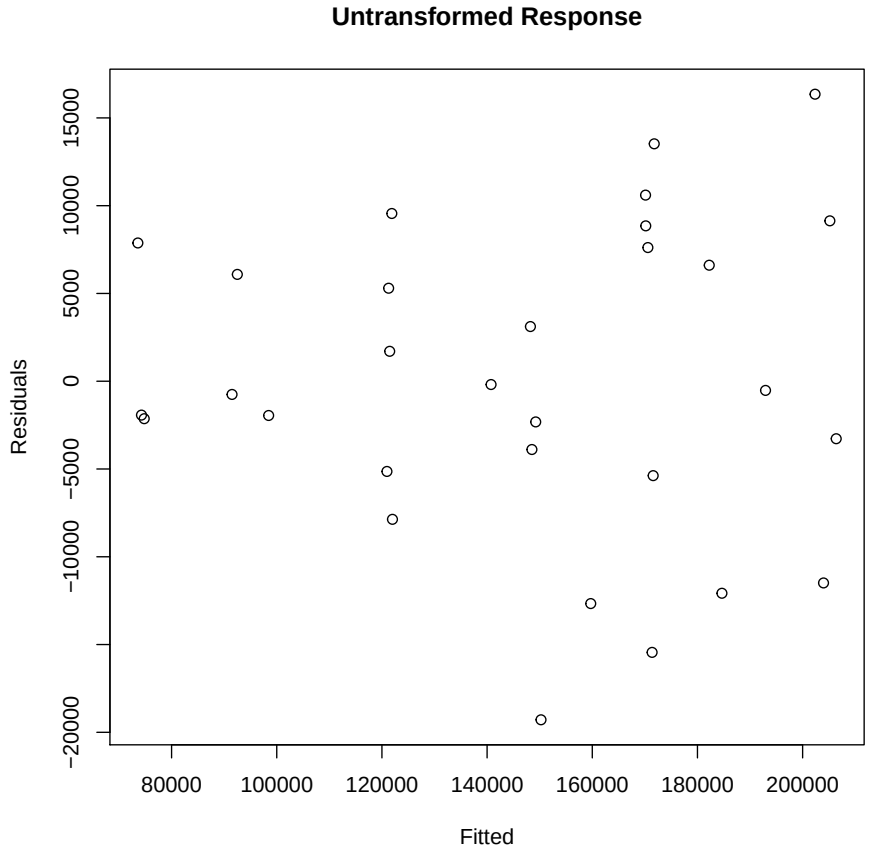
5 weighted least squares

```
attach(food)
ffit1 <- lm(income ~ expense, data=food)
summary(ffit1)

##
## Call:
## lm(formula = income ~ expense, data = food)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -19287  -4827   -638   7363  16351
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.94e+04   4.29e+03    11.5  3.8e-12 ***
## expense     8.05e+00   3.27e-01    24.6  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9000 on 28 degrees of freedom
## Multiple R-squared:  0.956,Adjusted R-squared:  0.954
## F-statistic:  608 on 1 and 28 DF,  p-value: <2e-16

ffit1$fit  # fitted values

##      1      2      3      4      5      6      7      8      9     10
##  73589  74796  74273  91497  92503  98458 121276 122000 120954 121477
##     11     12     13     14     15     16     17     18     19     20
## 121879 140753 148238 149244 150250 148520 159707 170170 171579 170130
##     21     22     23     24     25     26     27     28     29     30
## 170572 171780 171377 184657 182243 192947 206388 203974 202364 205181

ffit1$res  # residuals

##        1        2        3        4        5        6        7        8
##   7875.3  -2135.0  -1928.8   -753.5   6085.4  -1951.4   5298.3  -7867.1
##        9       10       11       12       13       14       15       16
##  -5139.8   1704.0   9554.6   -189.0   3114.0  -2318.1 -19287.1  -3889.7
##       17       18       19       20       21       22       23       24
## -12666.1   8851.0  -5378.5  10602.2   7614.5  13524.3 -15446.3 -12078.2
##       25       26       27       28       29       30
##   6608.3   -523.1  -3276.0 -11491.5  16351.2   9136.2

plot(income ~ expense)
```




```
plot(ffit1$fit,ffit1$res,xlab="Fitted",ylab="Residuals",
main="Untransformed Response")
```



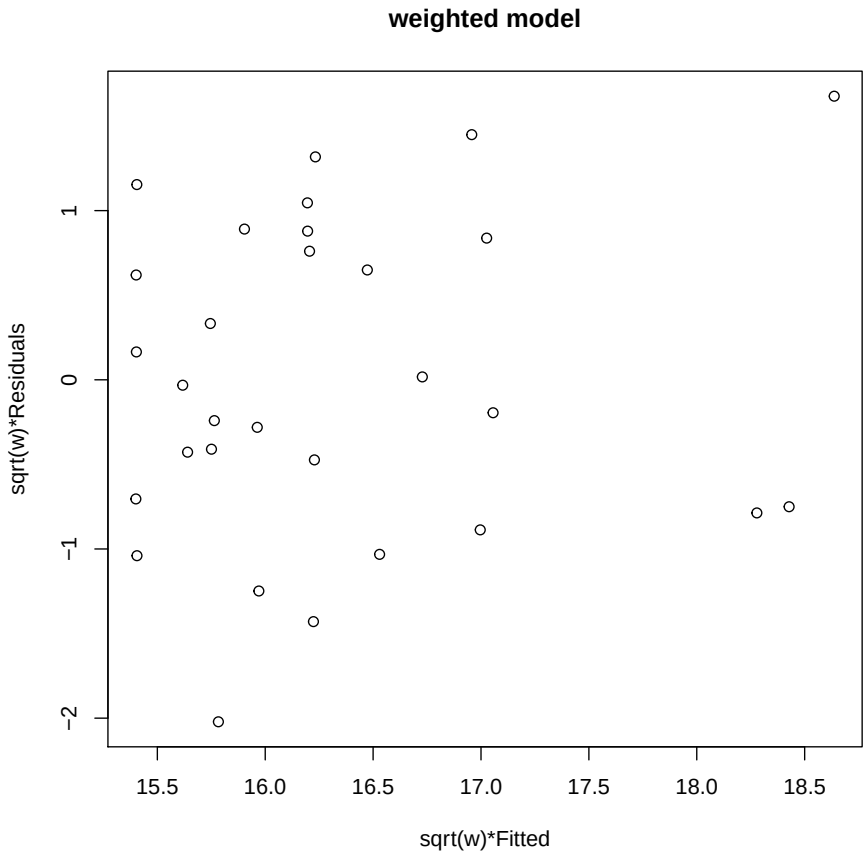

```
sy2<--7376216.04+7819.77*expense
w<-sy2^-1
w
```

```
##  [1] 6.218e-08 5.795e-08 5.971e-08 2.987e-08 2.902e-08 2.485e-08 1.602e-08
##  [8] 1.584e-08 1.610e-08 1.597e-08 1.587e-08 1.229e-08 1.129e-08 1.116e-08
## [15] 1.104e-08 1.125e-08 1.002e-08 9.097e-09 8.986e-09 9.101e-09 9.065e-09
## [22] 8.970e-09 9.001e-09 8.065e-09 8.220e-09 7.573e-09 6.891e-09 7.005e-09
## [29] 7.082e-09 6.948e-09
```

```
ffit2 <- lm(income ~ expense, data=food, weights=sy2^-1)
summary(ffit2)
```

```
##
## Call:
## lm(formula = income ~ expense, data = food, weights = sy2^-1)
```

```
##
## Weighted Residuals:
##    Min     1Q Median     3Q    Max
## -2.022 -0.739 -0.113  0.818  1.676
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.10e+04   2.51e+03    20.3   <2e-16 ***
## expense     7.92e+00   2.53e-01    31.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.962 on 28 degrees of freedom
## Multiple R-squared:  0.972,Adjusted R-squared:  0.971
## F-statistic:  979 on 1 and 28 DF,  p-value: <2e-16

fit<-sqrt(w)*ffit2$fit  # fitted values
res<-sqrt(w)*ffit2$res  # residuals
plot(fit,res,xlab="sqrt(w)*Fitted",ylab="sqrt(w)*Residuals",
main="weighted model")
```


weighted model
sqrt(w)*Residuals
sqrt(w)*Fitted
−2
−1
0
1
15.5
16.0
16.5
17.0
17.5
18.0
18.5