Homework 9

2022-12-13

(3)Use a simulation to empirically estimate the relative efficiencies of the following three estimators of the population mean: (i) arithmetic mean, (ii) median, (iii) Huber's M-estimator with = 1.5 and based on scale estimated by the MAD separately for the following three distributions:

N(0,1)

```
library(robustbase)
```

```
## Warning: il pacchetto 'robustbase' è stato creato con R versione 4.2.2
```

```
set.seed(1234)
dataN<-list()
rstN.mean=rep(NA,1000)
rstN.median=rep(NA,1000)
rstN.Hub=rep(NA,1000)
rstN.mean<-rstN.median<-rstN.Hub<-numeric(0)

for (i in 1:1000) {
  dataN[[i]]<-rnorm(20,0,1)
  rstN.mean[i] <-mean(dataN[[i]])
  rstN.median[i]<-  median(dataN[[i]])
  rstN.Hub[i]<- (huberM(dataN[[i]], k=1.5,se = TRUE))$mu
}

var.test(rstN.mean,rstN.median)
```

```
##
##  F test to compare two variances
##
## data:  rstN.mean and rstN.median
## F = 0.64953, num df = 999, denom df = 999, p-value = 1.111e-11
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5737325 0.7353502
## sample estimates:
## ratio of variances
##          0.6495339
```

The ratio is computed as var(rstN.mean)/var(rstN.median), so since the result is 0.651 we can say that the estimator median has a bigger variance and so it is preferable.

```r
var.test(rstN.mean,rstN.Hub)
```

```
##
##  F test to compare two variances
##
## data:  rstN.mean and rstN.Hub
## F = 0.94867, num df = 999, denom df = 999, p-value = 0.4051
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8379597 1.0740090
## sample estimates:
## ratio of variances
##          0.9486708
```

The ratio is 0.949 so in this case the huber's estimator is preferable. But is a value close to 1 so the work in a quite similar way.

```r
var.test(rstN.median,rstN.Hub)
```

```
##
##  F test to compare two variances
##
## data:  rstN.median and rstN.Hub
## F = 1.4605, num df = 999, denom df = 999, p-value = 2.413e-09
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.290094 1.653507
## sample estimates:
## ratio of variances
##           1.460541
```

We can say finally that the estimator median is the best, and the mean is the worst. The higher the variance the better the estimator.

t2

```r
set.seed(1234)
dataT2<-list()
rstT2.mean=rep(NA,1000)
rstT2.median=rep(NA,1000)
rstT2.Hub=rep(NA,1000)
rstT2.mean<-rstT2.median<-rstT2.Hub<-numeric(0)


for (i in 1:1000) {

  dataT2[[i]]<-rt(20,2,1)
  rstT2.mean[i] <-mean(dataT2[[i]])
  rstT2.median[i]<-  median(dataT2[[i]])
  rstT2.Hub[i]<- (huberM(dataT2[[i]], k=1.5,se = TRUE))$mu
}

var.test(rstT2.mean,rstT2.median)
```

```
##
##  F test to compare two variances
##
## data:  rstT2.mean and rstT2.median
## F = 7.1585, num df = 999, denom df = 999, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  6.323120 8.104314
## sample estimates:
## ratio of variances
##            7.15853
```

Here the estimator median has a variance very small with respect to the variance of the other for the mean. In fact the first one is 7 times the other.

```
var.test(rstT2.mean,rstT2.Hub)
```

```
##
##  F test to compare two variances
##
## data:  rstT2.mean and rstT2.Hub
## F = 6.4807, num df = 999, denom df = 999, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  5.724425 7.336969
## sample estimates:
## ratio of variances
##           6.480735
```

Also the Huber's estimator is very small with respect to the other one.

```
var.test(rstT2.median,rstT2.Hub)
```

```
##
##  F test to compare two variances
##
## data:  rstT2.median and rstT2.Hub
## F = 0.90532, num df = 999, denom df = 999, p-value = 0.1161
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7996648 1.0249267
## sample estimates:
## ratio of variances
##          0.9053164
```

Here we can see that the median is little bit worst that the Huber's. So we can say that the best is absolutely the estimator mean, and the worst is the Median in this situation.

t4

```r
set.seed(1234)
dataT4<-list()
rstT4.mean=rep(NA,1000)
rstT4.median=rep(NA,1000)
rstT4.Hub=rep(NA,1000)
rstT4.mean<-rstT4.median<-rstT4.Hub<-numeric(0)


for (i in 1:1000) {

  dataT4[[i]]<-rt(20,2,1)
  rstT4.mean[i] <-mean(dataT4[[i]])
  rstT4.median[i]<-  median(dataT4[[i]])
  rstT4.Hub[i]<- (huberM(dataT4[[i]], k=1.5,se = TRUE))$mu
}

var.test(rstT4.mean,rstT4.median)
```

```
##
##  F test to compare two variances
##
## data:  rstT4.mean and rstT4.median
## F = 7.1585, num df = 999, denom df = 999, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  6.323120 8.104314
## sample estimates:
## ratio of variances
##            7.15853
```

The variance of the mean is very big with respect to the variance of the median. The mean is preferable.

```r
var.test(rstT4.mean,rstT4.Hub)
```

```
##
##  F test to compare two variances
##
## data:  rstT4.mean and rstT4.Hub
## F = 6.4807, num df = 999, denom df = 999, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  5.724425 7.336969
## sample estimates:
## ratio of variances
##           6.480735
```

Alse here the mean is preferable to the Huber's.

```r
var.test(rstT4.median,rstT4.Hub)
```

```
##
```

```
##  F test to compare two variances
##
## data:  rstT4.median and rstT4.Hub
## F = 0.90532, num df = 999, denom df = 999, p-value = 0.1161
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7996648 1.0249267
## sample estimates:
## ratio of variances
##          0.9053164
```

The results are very similar with the previous case with t2-distribution. The mean is the best and the median is the worst.

(4)Investigate the sensitivity of the two estimators to outliers. (a) There is an obvious outlier in the data set. Create a new data set identical to the original one, but with this outlier removed.

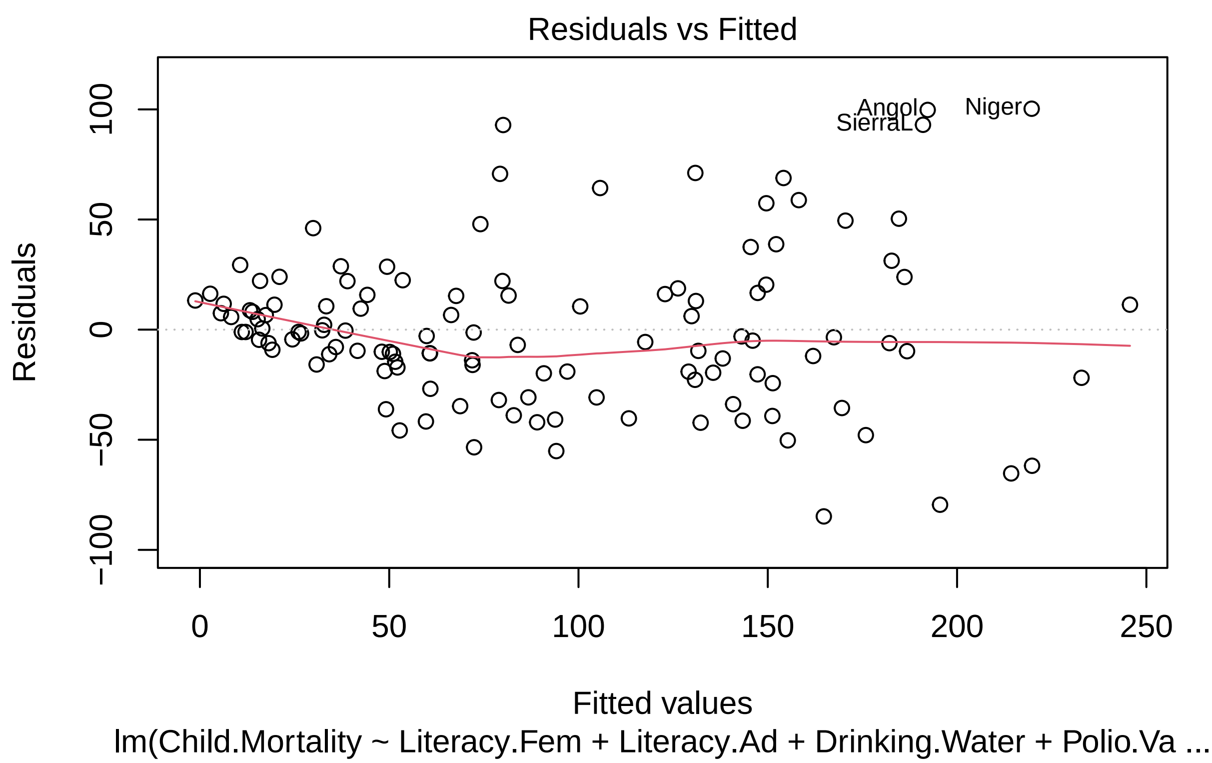
```
data <- read.csv("C:/Users/Utente/OneDrive/Desktop/bigData/datasets/unicef97.dat", sep="")
```

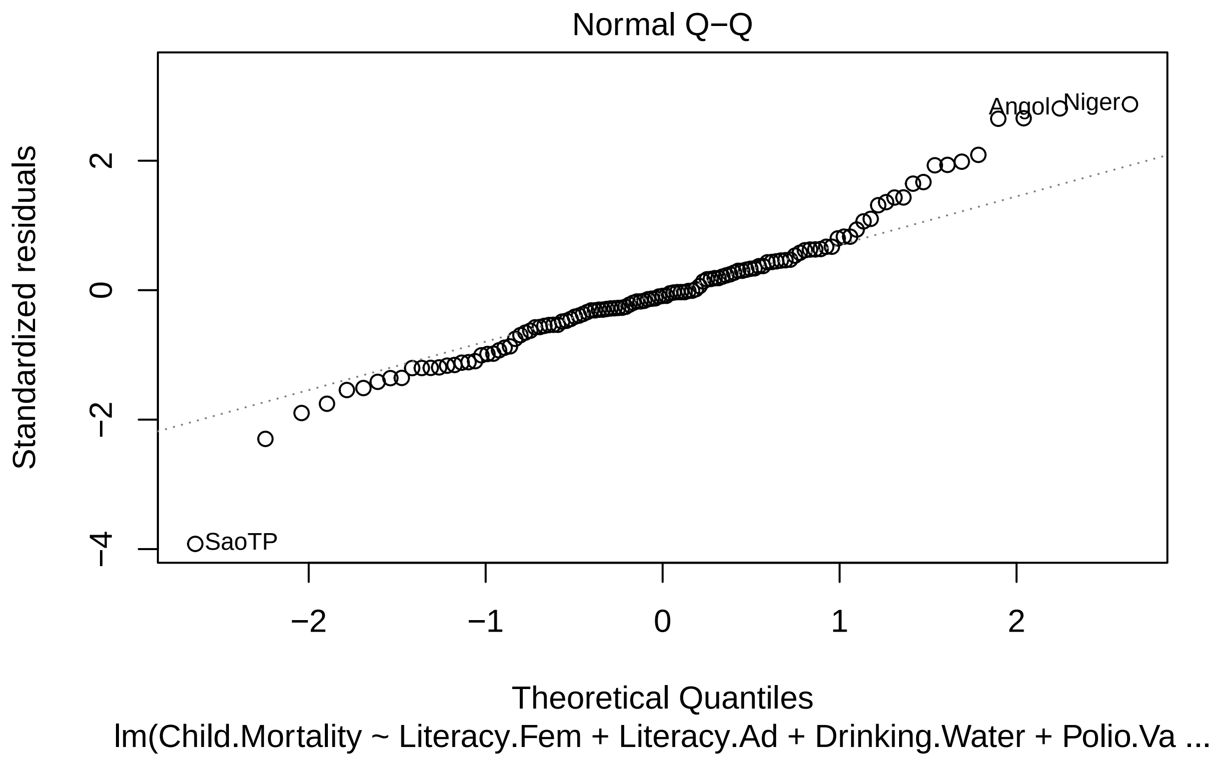
LM

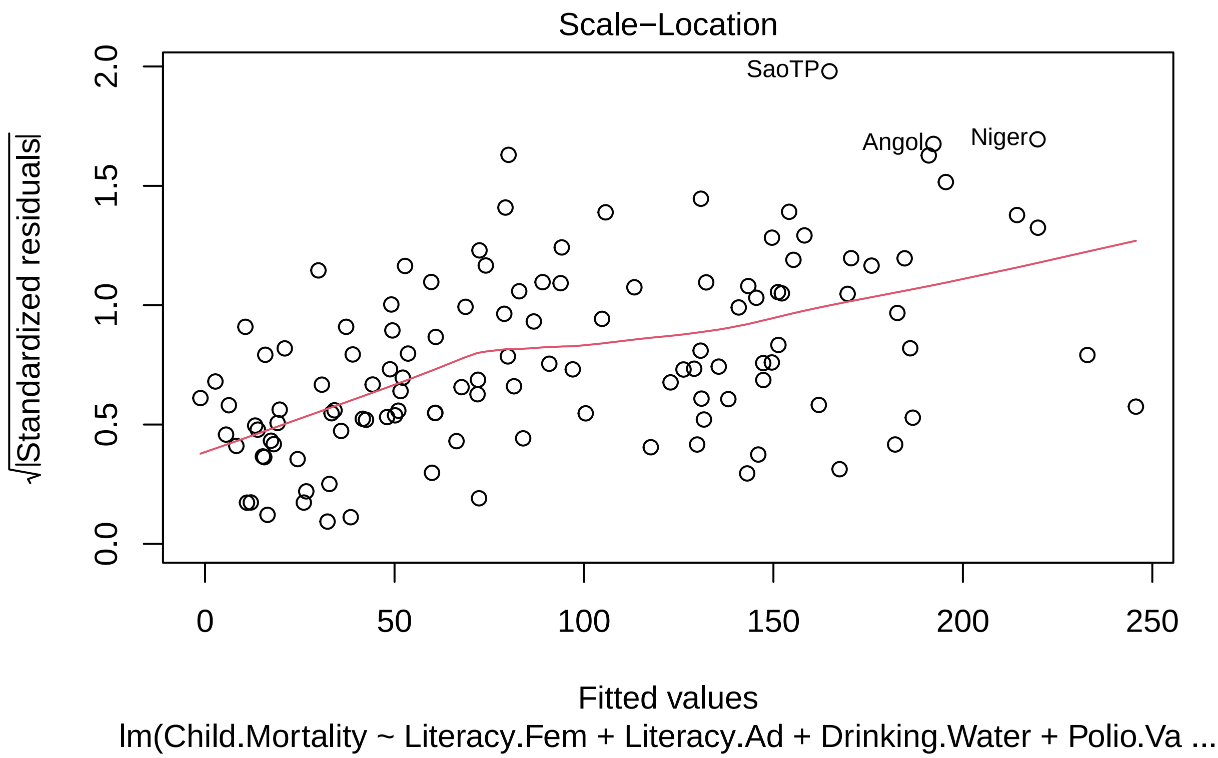
```
set.seed(1234)
data.lm <- lm(Child.Mortality~Literacy.Fem+Literacy.Ad+Drinking.Water+ Polio.Vacc+Tetanus.Vacc.Preg+Urba
summary.lm(data.lm)
```

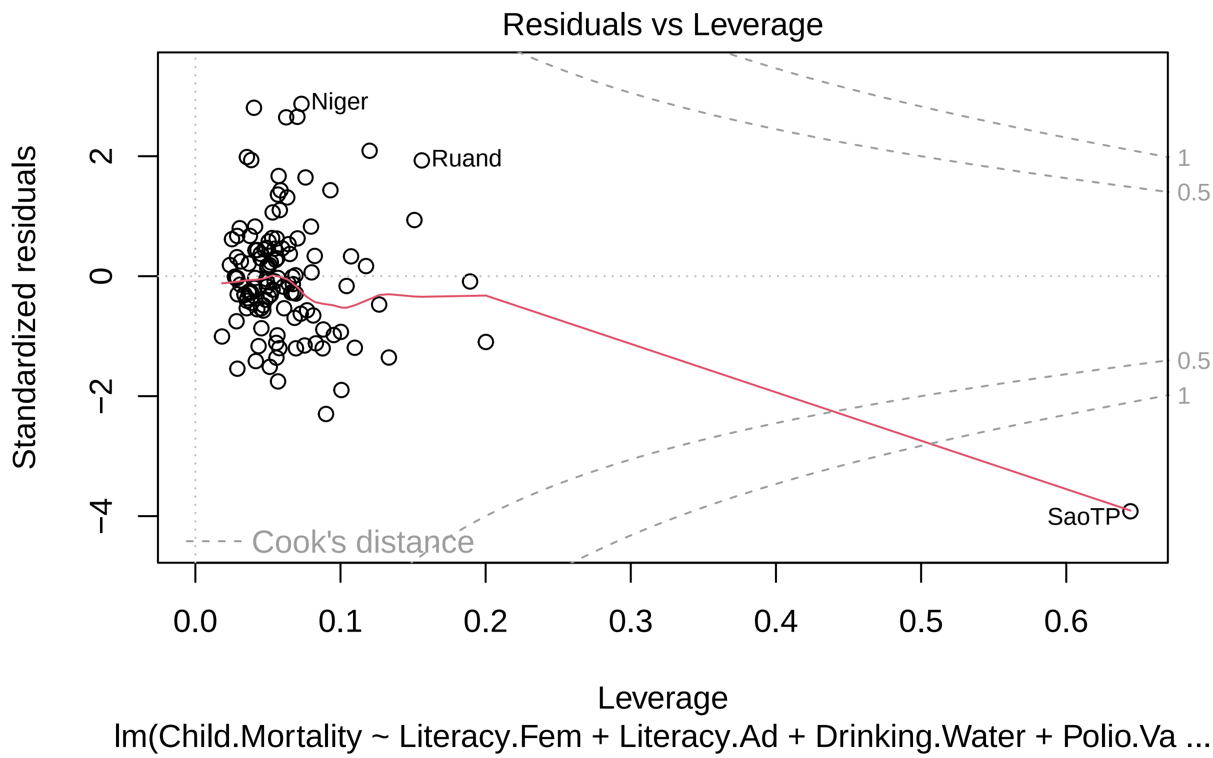
```
##
## Call:
## lm(formula = Child.Mortality ~ Literacy.Fem + Literacy.Ad + Drinking.Water +
##     Polio.Vacc + Tetanus.Vacc.Preg + Urban.Pop + Foreign.Aid,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.802 -19.570  -3.072  16.142 100.297
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       333.4750    16.7638  19.893  < 2e-16 ***
## Literacy.Fem       -1.1577     0.4432  -2.612  0.01021 *
## Literacy.Ad        -0.2405     0.4167  -0.577  0.56497
## Drinking.Water     -0.8695     0.2004  -4.339 3.13e-05 ***
## Polio.Vacc         -0.7159     0.2362  -3.031  0.00302 **
## Tetanus.Vacc.Preg  -0.0985     0.1593  -0.618  0.53750
## Urban.Pop          -0.4112     0.1952  -2.107  0.03736 *
## Foreign.Aid         0.2878     0.1759   1.636  0.10459
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.27 on 113 degrees of freedom
## Multiple R-squared:  0.7587, Adjusted R-squared:  0.7437
## F-statistic: 50.75 on 7 and 113 DF,  p-value: < 2.2e-16
```

```
plot(data.lm)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Child.Mortality ~ Literacy.Fem + Literacy.Ad + Drinking.Water + Polio.Va ...
Angola
SierraL
Niger
100
50
0
−50
−100
0
50
100
150
200
250


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Child.Mortality ~ Literacy.Fem + Literacy.Ad + Drinking.Water + Polio.Va ...
SaoTP
Angol
Niger
−4
−2
0
2
−2
−1
0
1
2


Scale−Location
SaoTP
Angol
Niger
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
50
100
150
200
250
Fitted values
lm(Child.Mortality ~ Literacy.Fem + Literacy.Ad + Drinking.Water + Polio.Va ...



SaoTP seems to be a bad leverage point.

MM

```
set.seed(1234)
data.mm <- lmrob(Child.Mortality~Literacy.Fem+Literacy.Ad+Drinking.Water+ Polio.Vacc+Tetanus.Vacc.Preg+U
summary(data.mm)
```

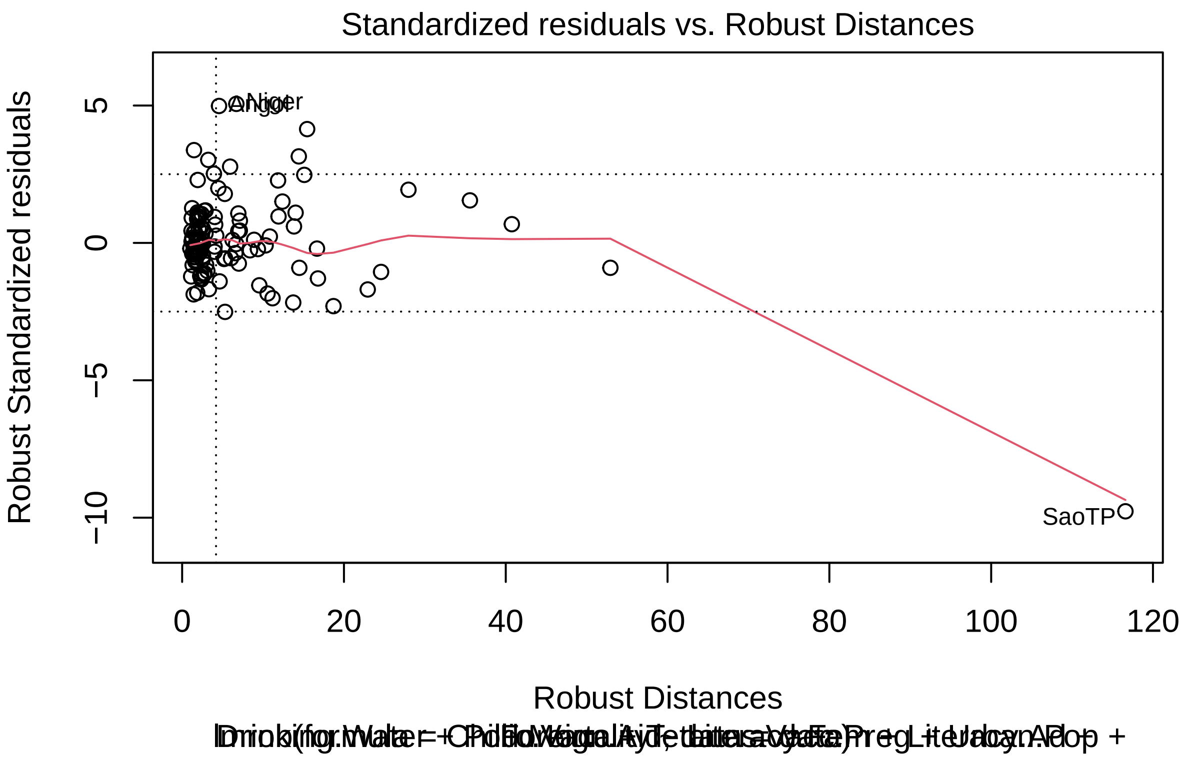
```
##
## Call:
## lmrob(formula = Child.Mortality ~ Literacy.Fem + Literacy.Ad + Drinking.Water +
##     Polio.Vacc + Tetanus.Vacc.Preg + Urban.Pop + Foreign.Aid, data = data)
##  \--> method = "MM"
## Residuals:
##       Min        1Q    Median        3Q       Max
## -238.8820  -14.2924   -0.4143   21.3896  123.7362
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        277.88469   34.15661   8.136 5.91e-13 ***
## Literacy.Fem        -1.14738    0.55415  -2.071 0.040683 *
## Literacy.Ad          0.01122    0.43620   0.026 0.979529
## Drinking.Water      -0.61264    0.19972  -3.067 0.002702 **
## Polio.Vacc          -0.63284    0.36036  -1.756 0.081775 .
## Tetanus.Vacc.Preg   -0.15987    0.13705  -1.166 0.245872
## Urban.Pop           -0.32653    0.16752  -1.949 0.053752 .
```

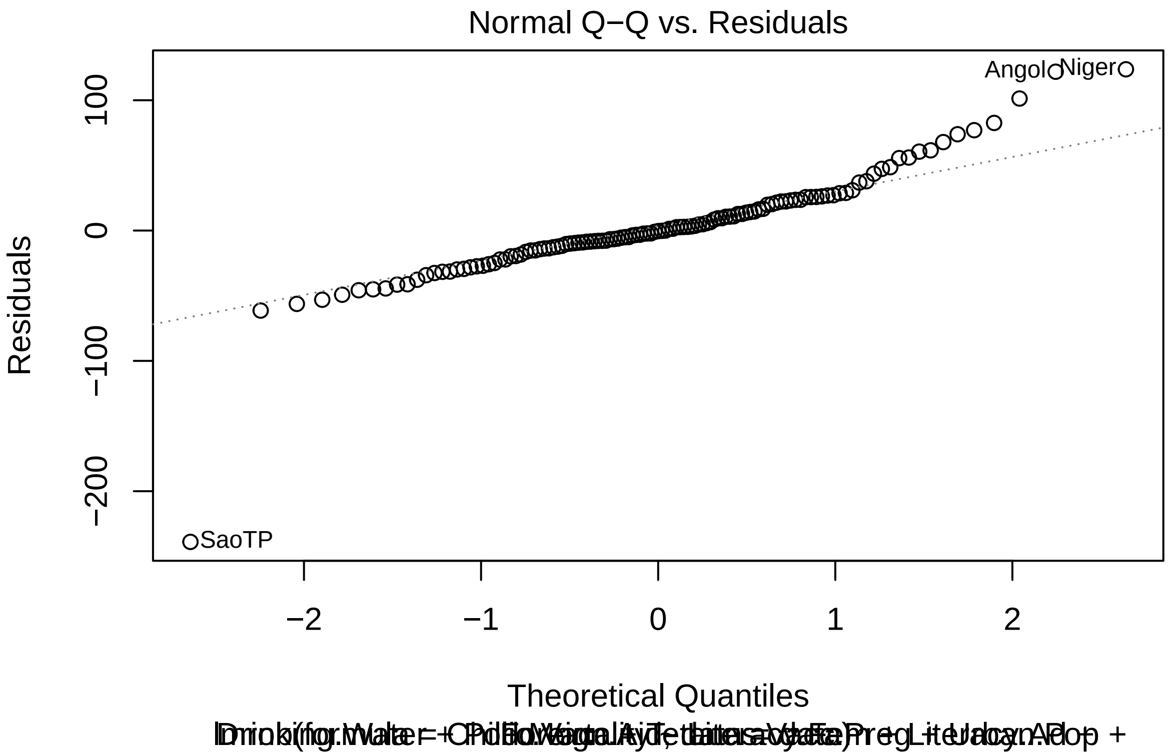
```
## Foreign.Aid         1.25256    0.31866   3.931 0.000146 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Robust residual standard error: 24.46
## Multiple R-squared:  0.8142, Adjusted R-squared:  0.8027
## Convergence in 24 IRWLS iterations
##
## Robustness weights:
##  3 observations c(4,80,91) are outliers with |weight| = 0 ( < 0.00083);
##  9 weights are ~= 1. The remaining 109 ones are summarized as
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04766 0.85490 0.94130 0.87120 0.98690 0.99900
## Algorithmic parameters:
##        tuning.chi                bb        tuning.psi        refine.tol
##         1.548e+00         5.000e-01         4.685e+00         1.000e-07
##           rel.tol         scale.tol         solve.tol       eps.outlier
##         1.000e-07         1.000e-10         1.000e-07         8.264e-04
##             eps.x warn.limit.reject warn.limit.meanrw
##         3.165e-10         5.000e-01         5.000e-01
##      nResample         max.it       best.r.s       k.fast.s          k.max
##            500             50              2              1            200
##    maxit.scale      trace.lev            mts     compute.rd fast.s.large.n
##            200              0           1000              0           2000
##                   psi           subsampling                   cov
##            "bisquare"         "nonsingular"         ".vcov.avar1"
## compute.outlier.stats
##                  "SM"
## seed : int(0)
```

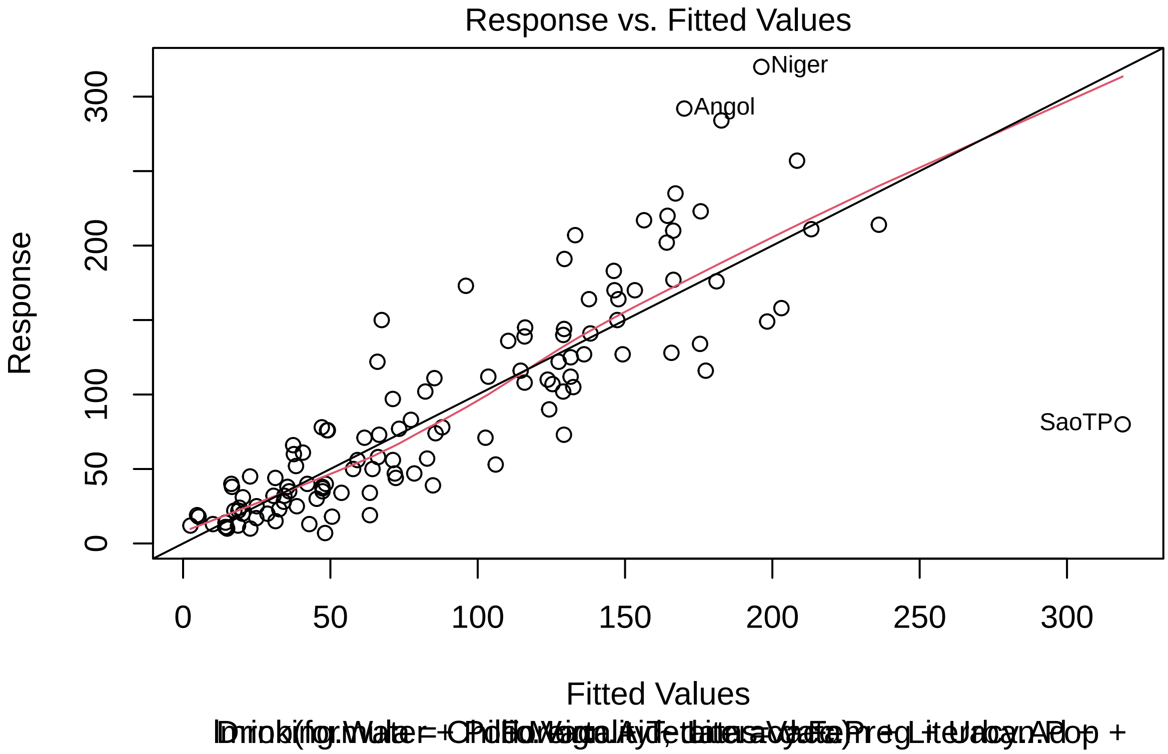
```
plot(data.mm)
```

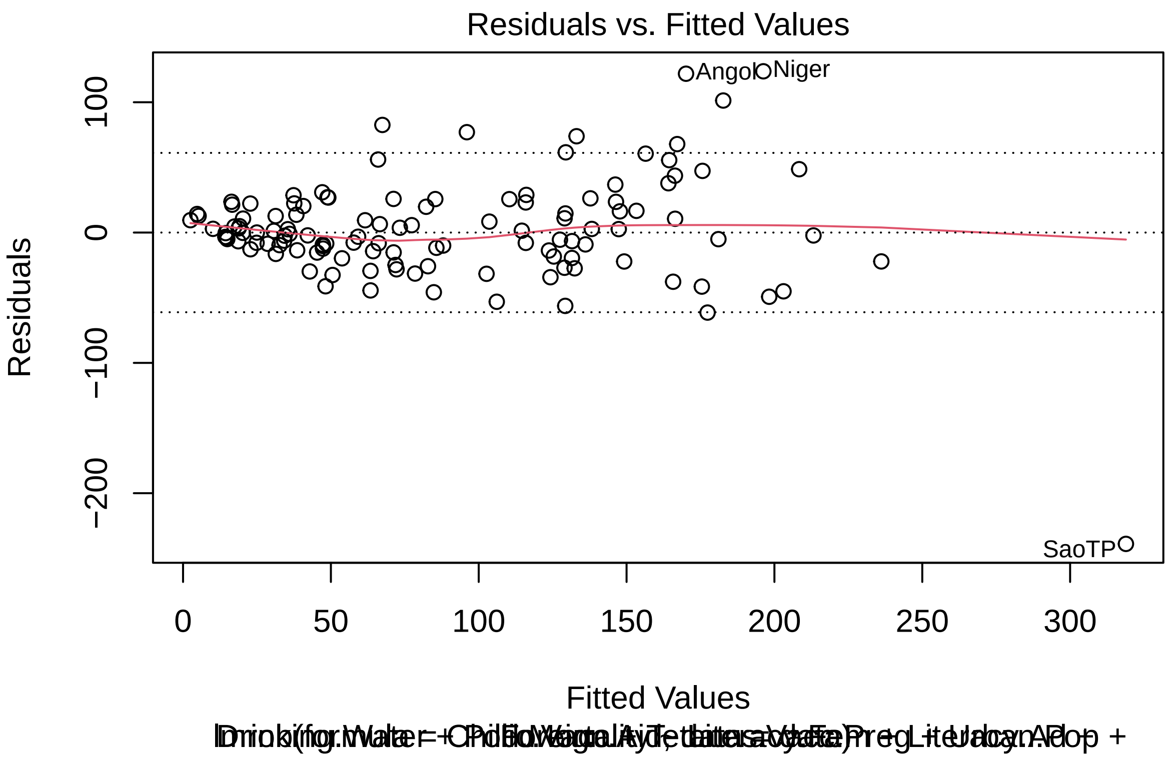
```
## recomputing robust Mahalanobis distances
```

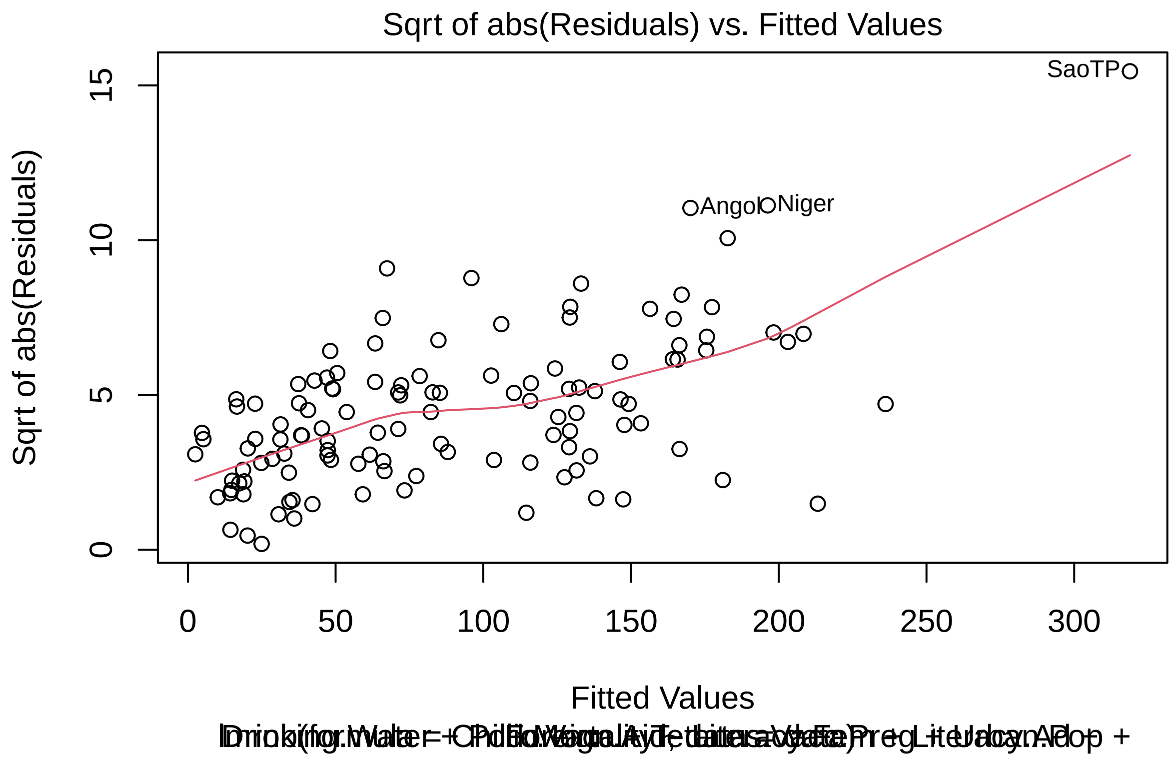
```
## saving the robust distances 'MD' as part of 'data.mm'
```


Standardized residuals vs. Robust Distances
Robust Standardized residuals
5
0
−5
−10
Niger
SaoTP
0
20
40
60
80
100
120
Robust Distances


Normal Q−Q vs. Residuals
Residuals
Theoretical Quantiles
100
0
−100
−200
−2
−1
0
1
2
Angola
Niger
SaoTP


Response vs. Fitted Values
Response
Fitted Values
Niger
Angol
SaoTP
0
50
100
200
300
0
50
100
150
200
250
300


Residuals vs. Fitted Values
Residuals
100
0
−100
−200
Angol
Niger
SaoTP
0
50
100
150
200
250
300
Fitted Values
lm(formula = ChildMortality ~ literacyFem + literacyAd +
Drinking.Water + Polio.Vacc + Tetanus.Vacc.Preg + Urban.Pop +



Robustness weights: 3 observations c(4,80,91) are outliers with |weight| = 0 In particular the sao tp seems to be a bad leverage, in both cases.

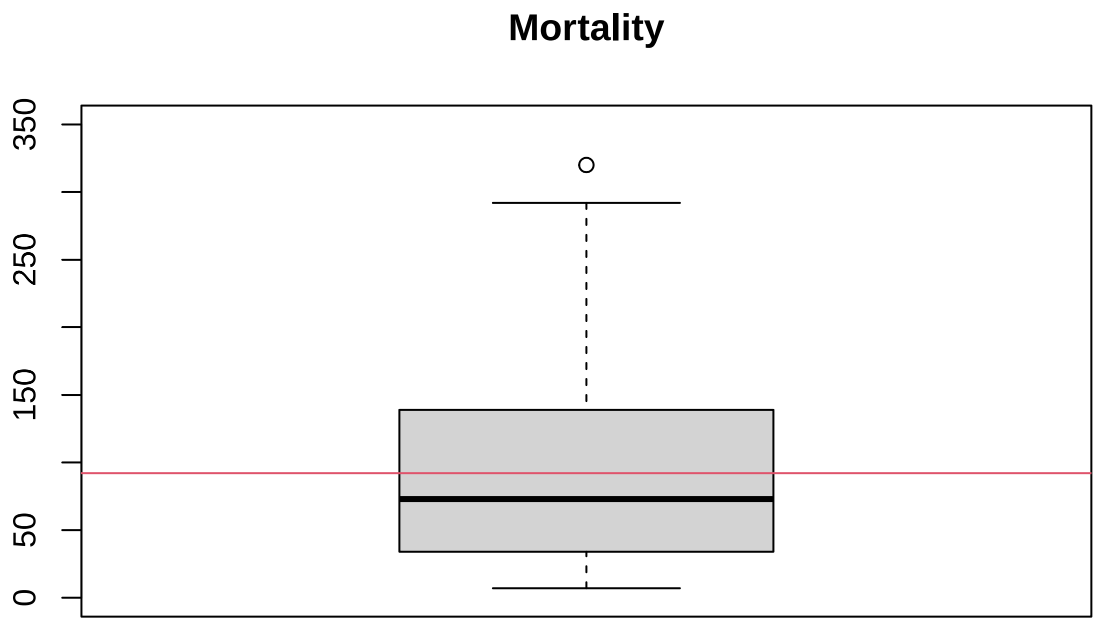
```
set.seed(1234)
mean(data$Child.Mortality)
```

```
## [1] 92.08264
```

```
sum(data$Child.Mortality>92.08264)
```

```
## [1] 51
```

```
boxplot(data$Child.Mortality,main="Mortality",ylim=c(0,350))
abline(mean(data$Child.Mortality),0,col=2)
```



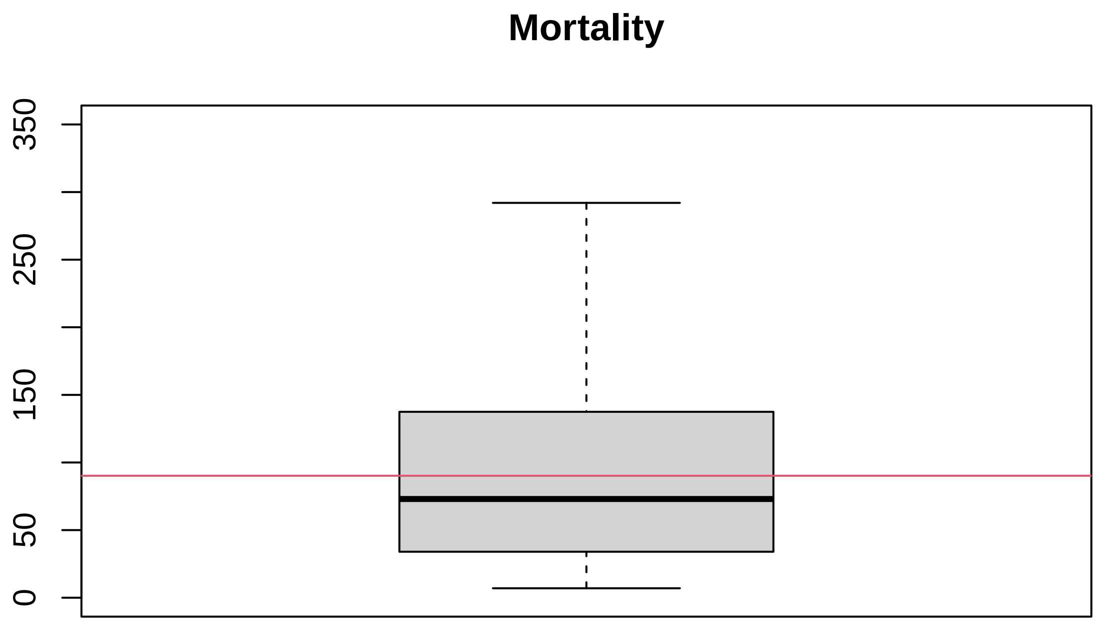

```
which.max(data$Child.Mortality)
```

```
## [1] 80
```

```
mean(data$Child.Mortality[-80])
```

```
## [1] 90.18333
```

```
boxplot(data$Child.Mortality[-80],main="Mortality",ylim=c(0,350))
abline(mean(data$Child.Mortality[-80]),0,col=2)
```



The outlier that is shown in the first plot has been removed.

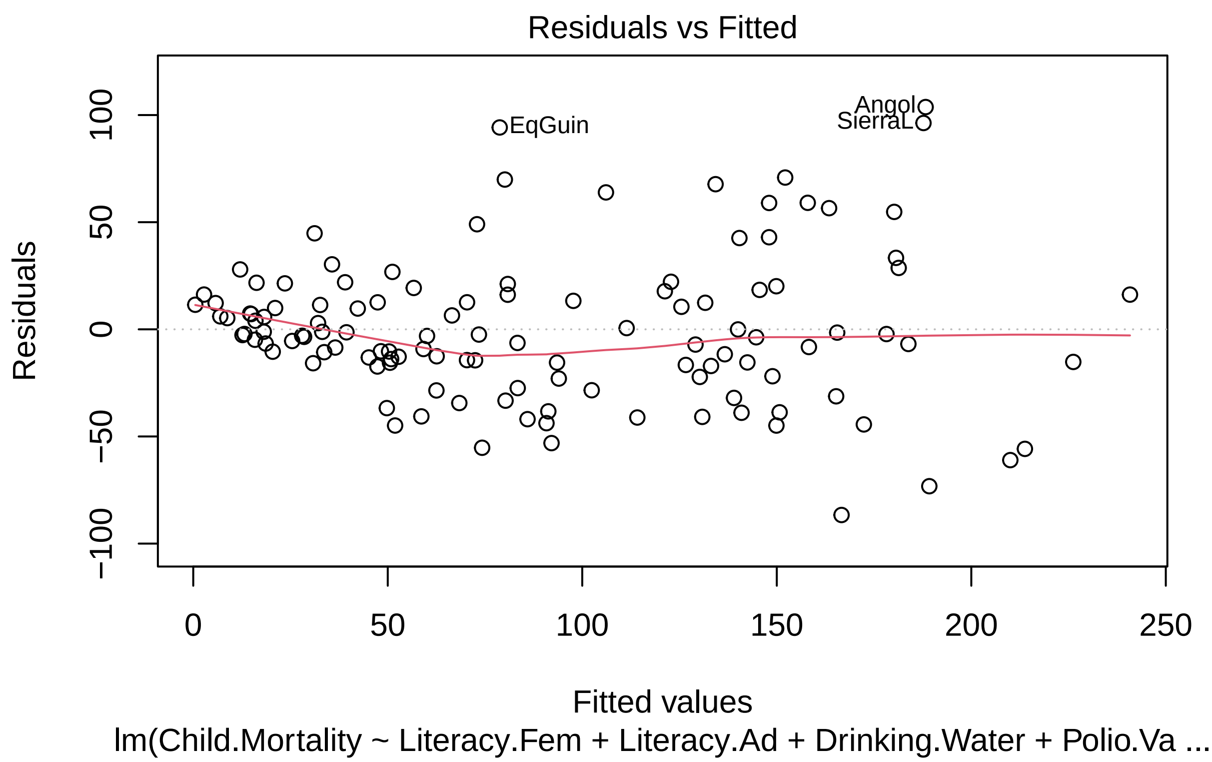
Run both lm and lmrob on the new data set, and compare the results to the results of the same regression method on the original data by (i) commenting on how t-test results of the variables have changed qualitatively. (ii) comparing the vectors of estimators of the regression coefficients. Comment on what the results mean regarding the sensitivity of the two regressions.

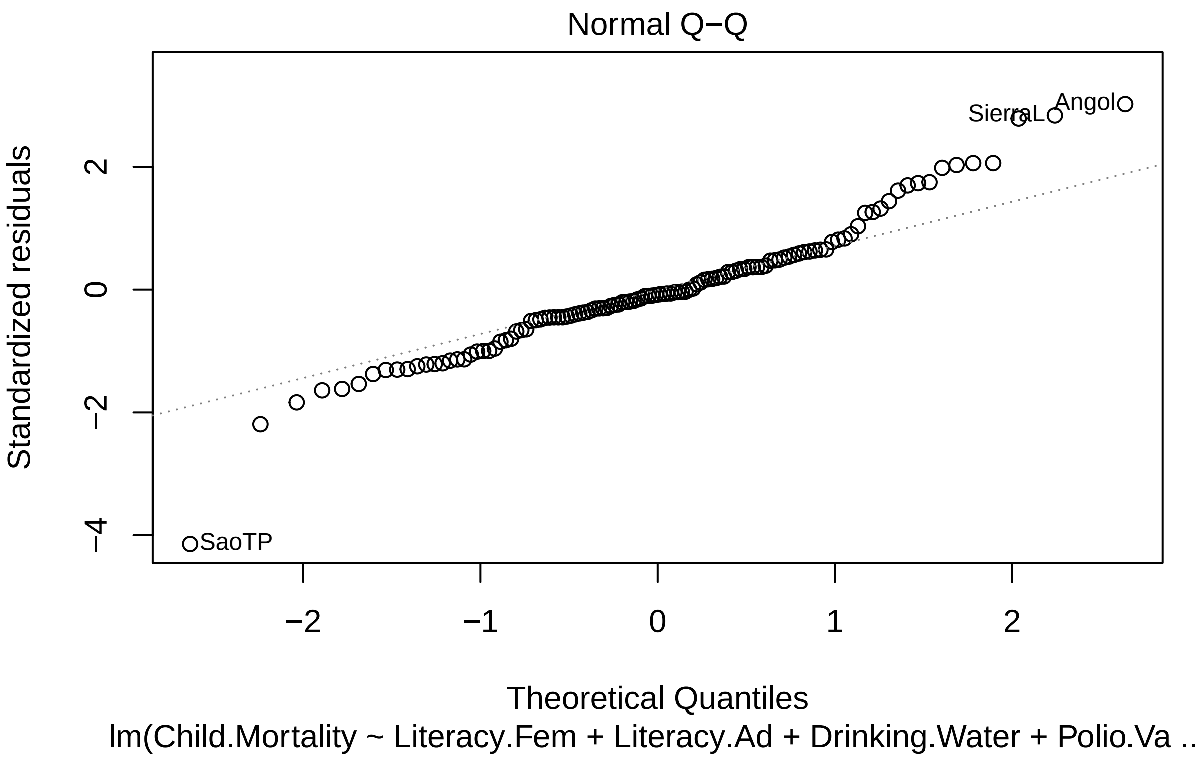
```
set.seed(1234)
data.2 <- data[-c(80),]
data.2lm <- lm(Child.Mortality~Literacy.Fem+Literacy.Ad+Drinking.Water+Polio.Vacc+Tetanus.Vacc.Preg+Urba
summary.lm(data.2lm)
```

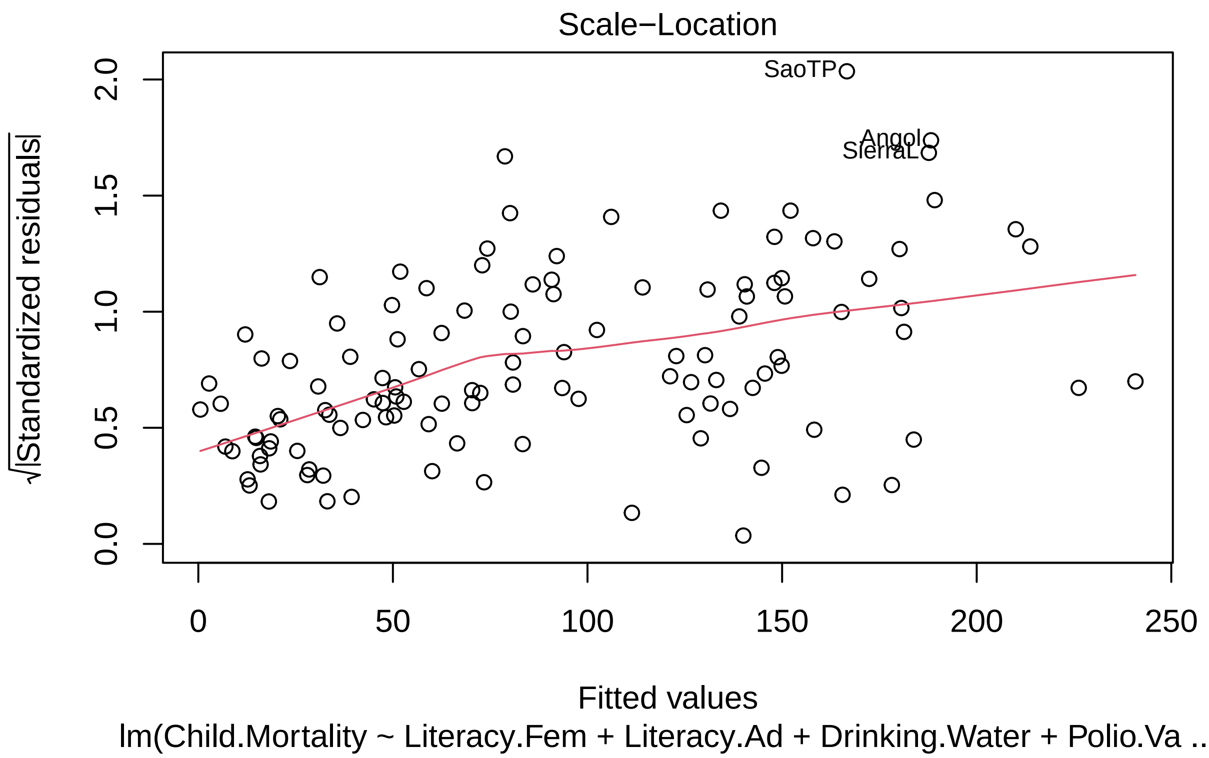
```
##
## Call:
## lm(formula = Child.Mortality ~ Literacy.Fem + Literacy.Ad + Drinking.Water +
##     Polio.Vacc + Tetanus.Vacc.Preg + Urban.Pop + Foreign.Aid,
##     data = data.2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -86.644 -16.739  -2.835  16.212 103.738
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        324.0990    16.5169  19.622  < 2e-16 ***
## Literacy.Fem        -1.1738     0.4286  -2.739  0.00718 **
## Literacy.Ad         -0.1328     0.4046  -0.328  0.74341
```

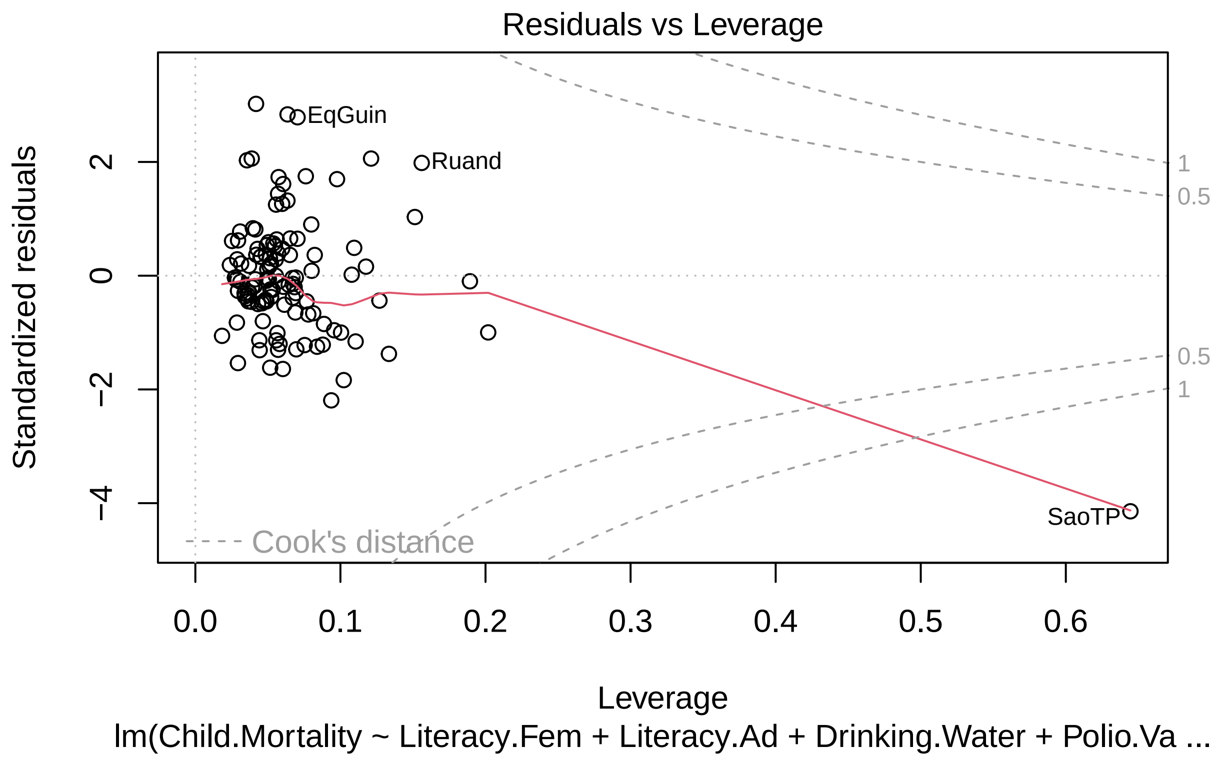
```
## Drinking.Water     -0.9109     0.1943  -4.688 7.84e-06 ***
## Polio.Vacc         -0.6572     0.2293  -2.867  0.00496 **
## Tetanus.Vacc.Preg  -0.1025     0.1540  -0.665  0.50726
## Urban.Pop          -0.3999     0.1888  -2.118  0.03640 *
## Foreign.Aid         0.3195     0.1704   1.875  0.06345 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.08 on 112 degrees of freedom
## Multiple R-squared:  0.7555, Adjusted R-squared:  0.7402
## F-statistic: 49.44 on 7 and 112 DF,  p-value: < 2.2e-16
```

```
plot(data.2lm)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Child.Mortality ~ Literacy.Fem + Literacy.Ad + Drinking.Water + Polio.Va ...
SierraL
Angol
SaoTP
2
0
−2
−4
−2
−1
0
1
2


Scale−Location
SaoTP
Angol
SierraL
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
50
100
150
200
250
Fitted values
lm(Child.Mortality ~ Literacy.Fem + Literacy.Ad + Drinking.Water + Polio.Va ...




```
set.seed(1234)
data.2mm <- lmrob(Child.Mortality~Literacy.Fem+Literacy.Ad+Drinking.Water+Polio.Vacc+Tetanus.Vacc.Preg+U
summary(data.2mm)
```

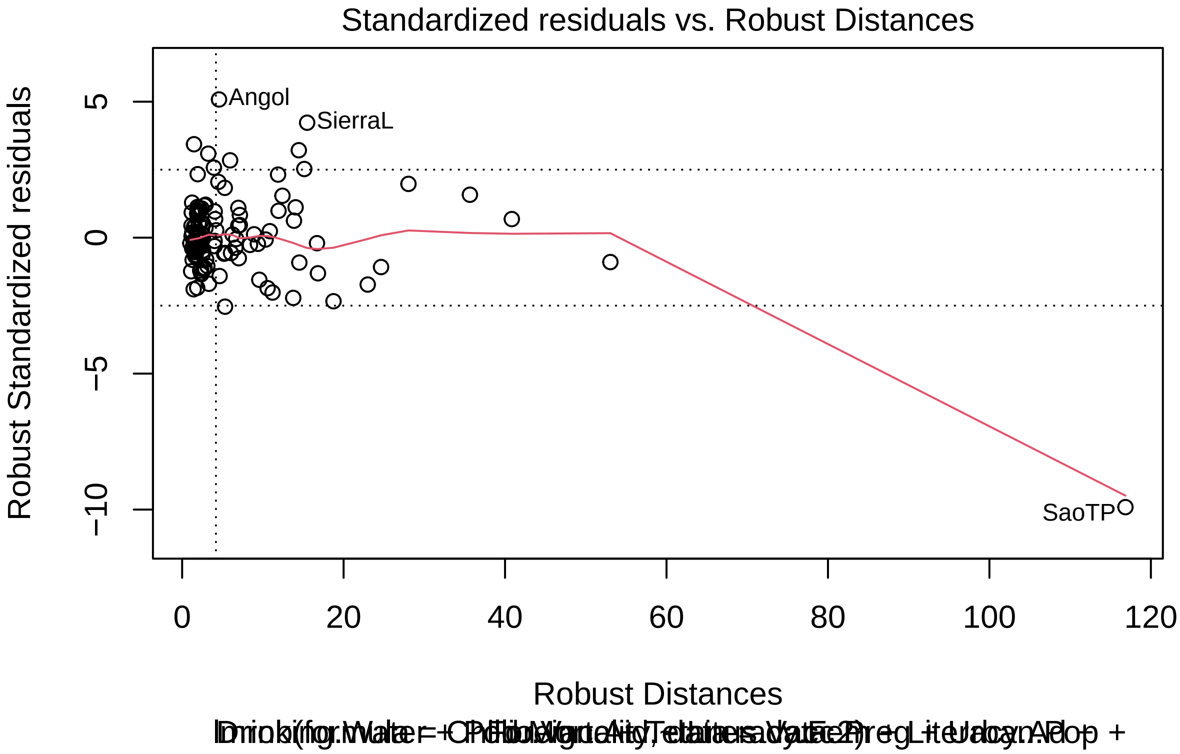
```
##
## Call:
## lmrob(formula = Child.Mortality ~ Literacy.Fem + Literacy.Ad + Drinking.Water +
##     Polio.Vacc + Tetanus.Vacc.Preg + Urban.Pop + Foreign.Aid, data = data.2)
##  \--> method = "MM"
## Residuals:
##       Min        1Q    Median        3Q       Max
## -238.4432  -14.4463   -0.9357   20.6700  122.2843
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.770e+02  3.522e+01   7.863 2.54e-12 ***
## Literacy.Fem      -1.131e+00  5.629e-01  -2.010 0.046850 *
## Literacy.Ad       -9.009e-04  4.415e-01  -0.002 0.998375
## Drinking.Water    -6.113e-01  2.009e-01  -3.043 0.002922 **
## Polio.Vacc        -6.245e-01  3.695e-01  -1.690 0.093790 .
## Tetanus.Vacc.Preg -1.638e-01  1.368e-01  -1.197 0.233744
## Urban.Pop         -3.287e-01  1.674e-01  -1.964 0.052015 .
## Foreign.Aid        1.252e+00  3.199e-01   3.914 0.000156 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

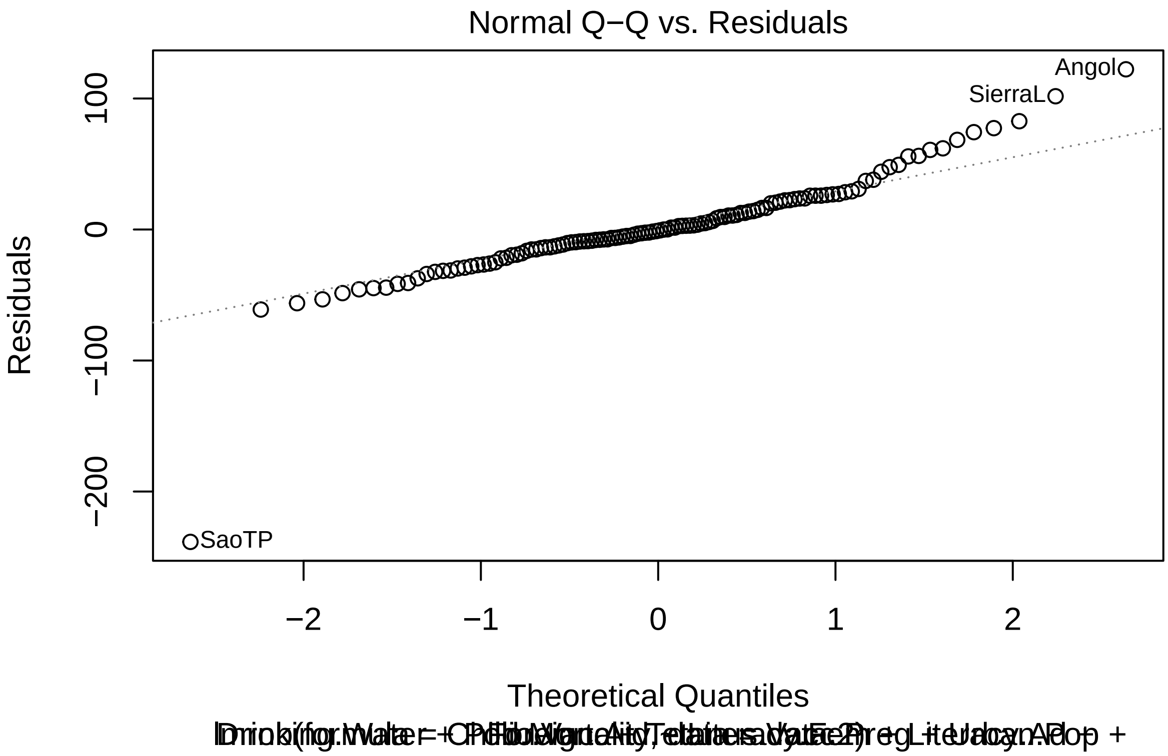
```
##
## Robust residual standard error: 24.06
## Multiple R-squared:  0.8158, Adjusted R-squared:  0.8043
## Convergence in 24 IRWLS iterations
##
## Robustness weights:
##  2 observations c(4,90) are outliers with |weight| = 0 ( < 0.00083);
##  9 weights are ~= 1. The remaining 109 ones are summarized as
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03456 0.85200 0.94010 0.86760 0.98690 0.99900
## Algorithmic parameters:
##        tuning.chi                bb        tuning.psi        refine.tol
##         1.548e+00         5.000e-01         4.685e+00         1.000e-07
##           rel.tol         scale.tol         solve.tol       eps.outlier
##         1.000e-07         1.000e-10         1.000e-07         8.333e-04
##             eps.x warn.limit.reject warn.limit.meanrw
##         3.165e-10         5.000e-01         5.000e-01
##      nResample         max.it       best.r.s       k.fast.s          k.max
##            500             50              2              1            200
##    maxit.scale      trace.lev            mts     compute.rd fast.s.large.n
##            200              0           1000              0           2000
##                   psi           subsampling                   cov
##            "bisquare"         "nonsingular"         ".vcov.avar1"
## compute.outlier.stats
##                  "SM"
## seed : int(0)
```

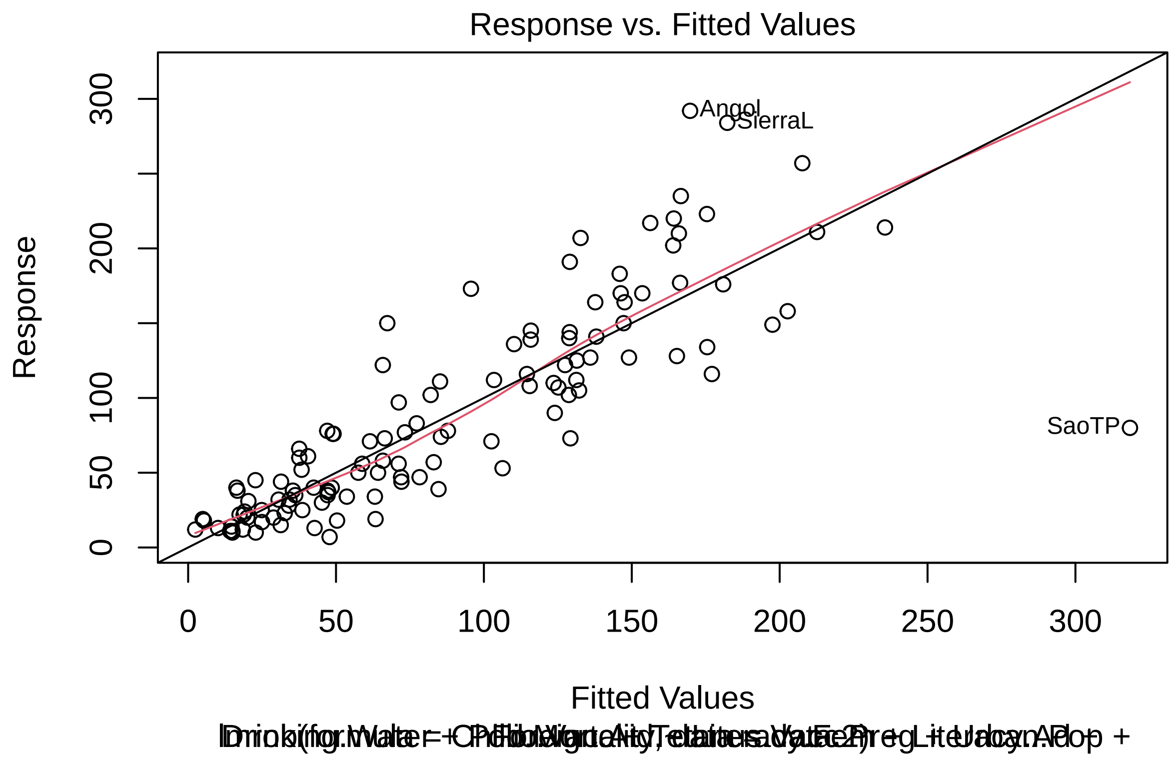
```
plot(data.2mm)
```

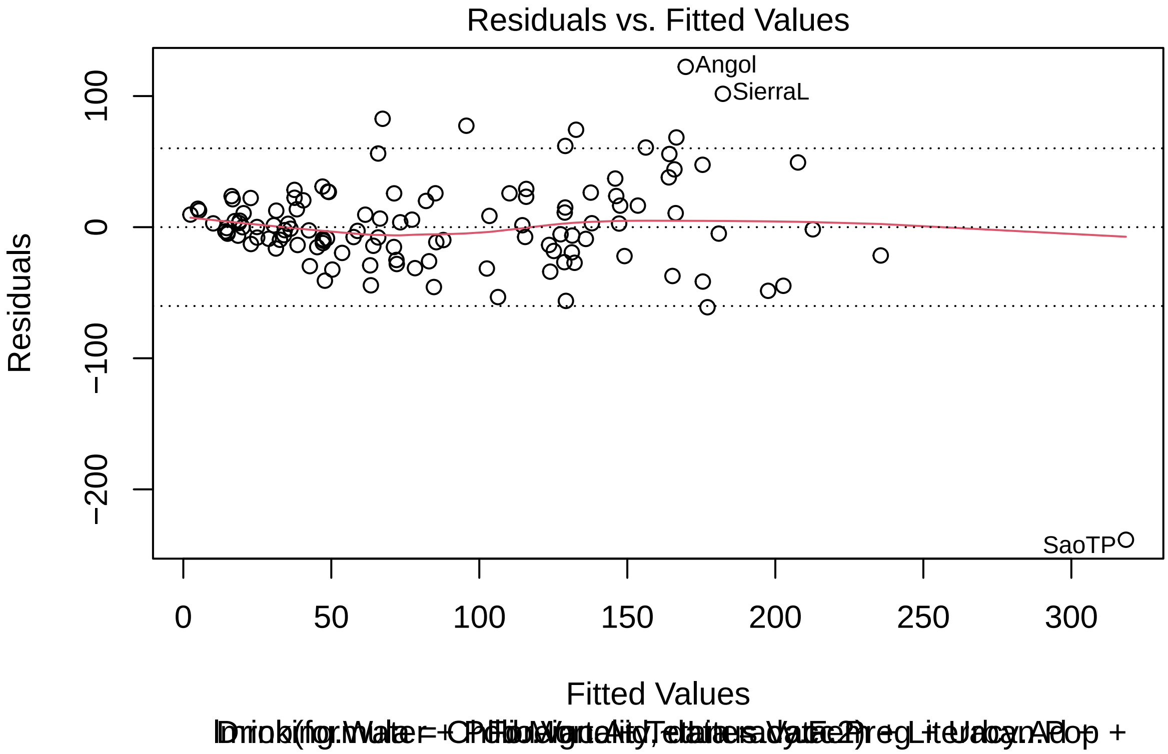
```
## recomputing robust Mahalanobis distances
```

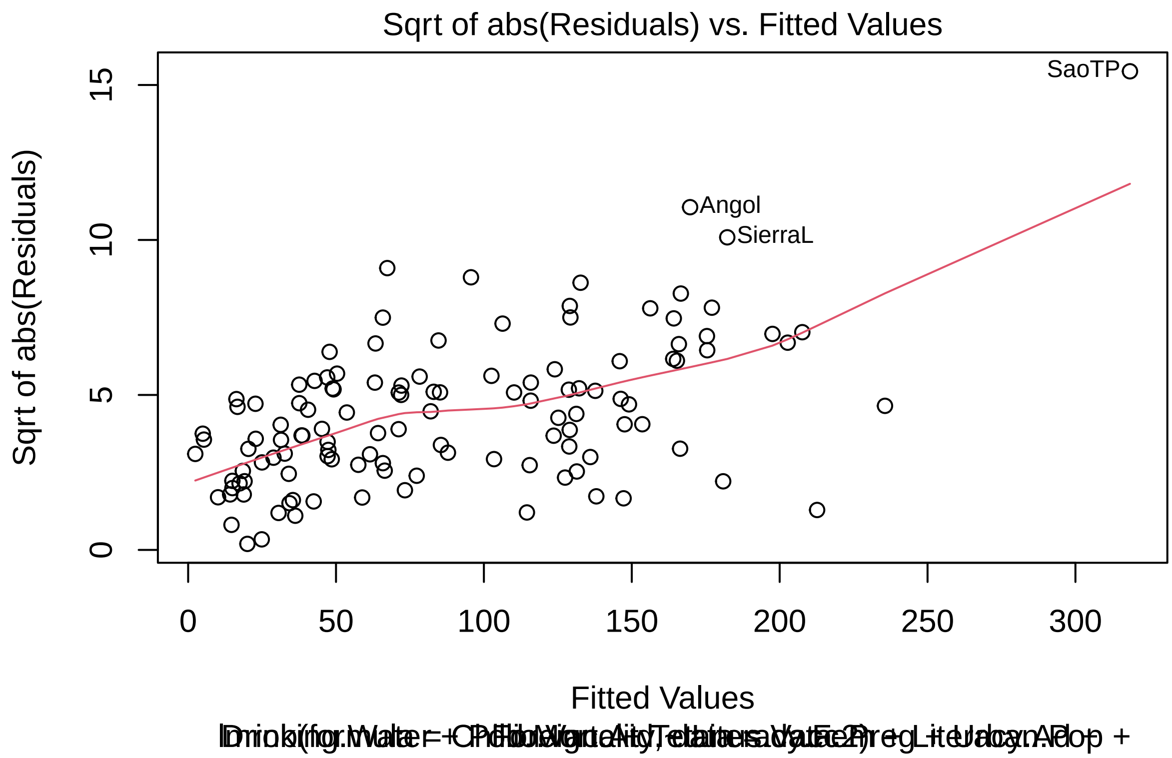
```
## saving the robust distances 'MD' as part of 'data.2mm'
```


Standardized residuals vs. Robust Distances
Robust Standardized residuals
Robust Distances
Angol
SierraL
SaoTP
5
0
−5
−10
0
20
40
60
80
100
120
Drinking.Water + Pollution.Air.Total + Water.Freq + Literacy.Adult +
Imm(formula = CHDMortality/Populations.Pop ~ Urban.Pop +


Normal Q−Q vs. Residuals
Residuals
100
0
−100
−200
Angol
SierraL
SaoTP
−2
−1
0
1
2
Theoretical Quantiles


Response vs. Fitted Values
Response
Fitted Values
Angol
SierraL
SaoTP
0
50
100
150
200
250
300


Residuals vs. Fitted Values
Angol
SierraL
SaoTP
100
0
−100
−200
Residuals
0
50
100
150
200
250
300
Fitted Values



The point 4 is an outlier also with the modified dataset, while is show the point 90 instead of the 91.

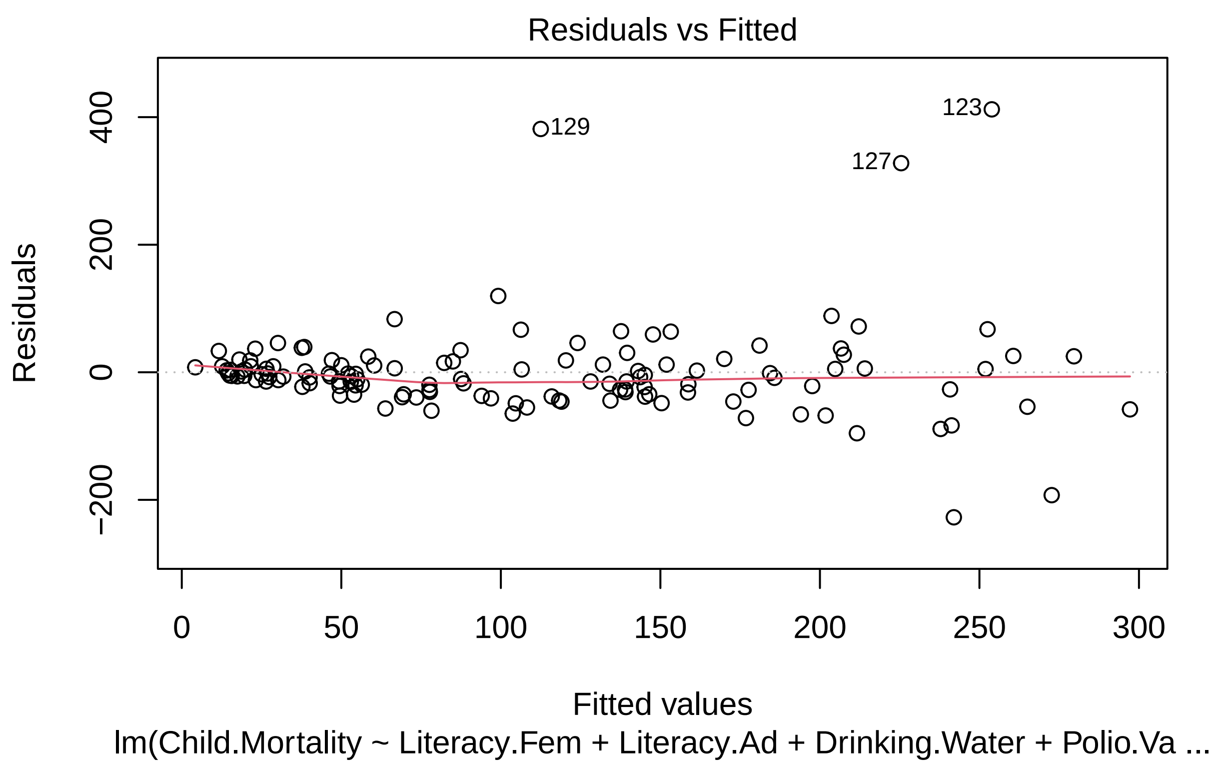
(b)Create a new data set that adds 10 outliers to the original data set, which are randomly generated. Run both lm and lmrob on the new data set, and compare the results to the results of the same regression method on the original data by (i) commenting on how t-test results of the variables have changed qualitatively (ii) comparing the estimators of the regression coefficients. Comment on what the results mean regarding the sensitivity of the two regressions.

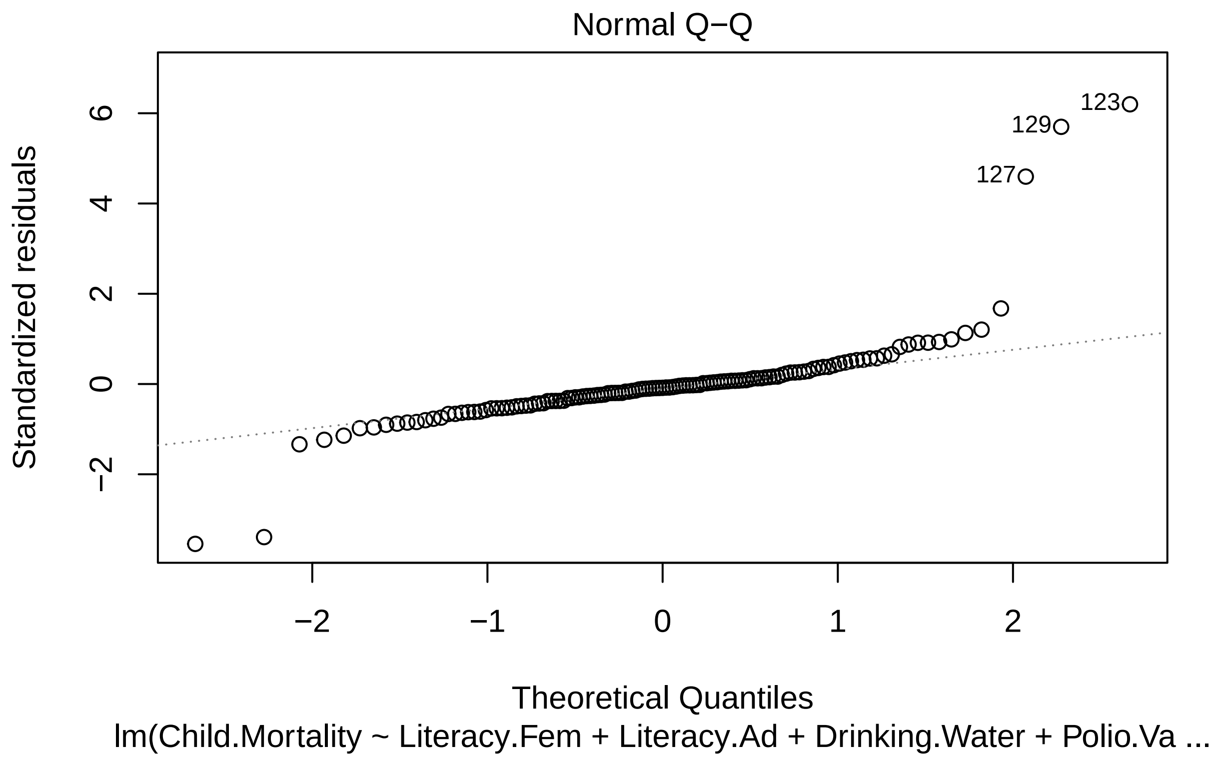
```
set.seed(1234)
data3 <- data

for(i in 1:10){
  x <- runif(1,0,1000)
  for(i in 2:8)
    x[i] <- runif(1,0,100)
  data3 <- rbind(data3,x)
}

set.seed(1234)
data.3lm<- lm(Child.Mortality~Literacy.Fem+Literacy.Ad+Drinking.Water+Polio.Vacc+Tetanus.Vacc.Preg+Urban
summary(data.3lm)
```

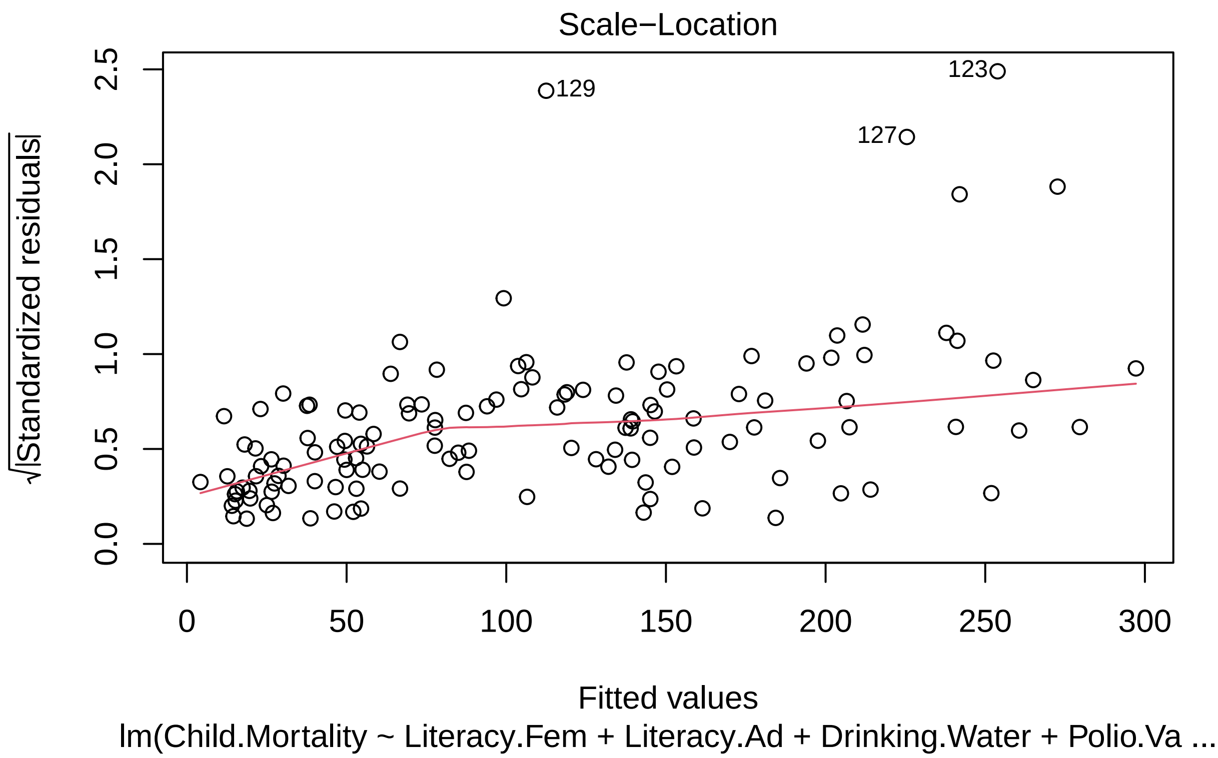
```
##
## Call:
## lm(formula = Child.Mortality ~ Literacy.Fem + Literacy.Ad + Drinking.Water +
##     Polio.Vacc + Tetanus.Vacc.Preg + Urban.Pop + Foreign.Aid,
##     data = data3)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -227.35  -29.20   -6.07   13.40  412.21
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       346.69512   32.02327  10.826  < 2e-16 ***
## Literacy.Fem       -1.00542    0.62517  -1.608  0.11035
## Literacy.Ad        -0.77612    0.55155  -1.407  0.16190
## Drinking.Water     -0.46689    0.39910  -1.170  0.24433
## Polio.Vacc         -1.20114    0.41342  -2.905  0.00435 **
## Tetanus.Vacc.Preg   0.07995    0.29150   0.274  0.78434
## Urban.Pop          -0.09667    0.36144  -0.267  0.78956
## Foreign.Aid         0.80687    0.31618   2.552  0.01194 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.7 on 123 degrees of freedom
## Multiple R-squared:  0.5209, Adjusted R-squared:  0.4936
## F-statistic: 19.11 on 7 and 123 DF,  p-value: < 2.2e-16
```

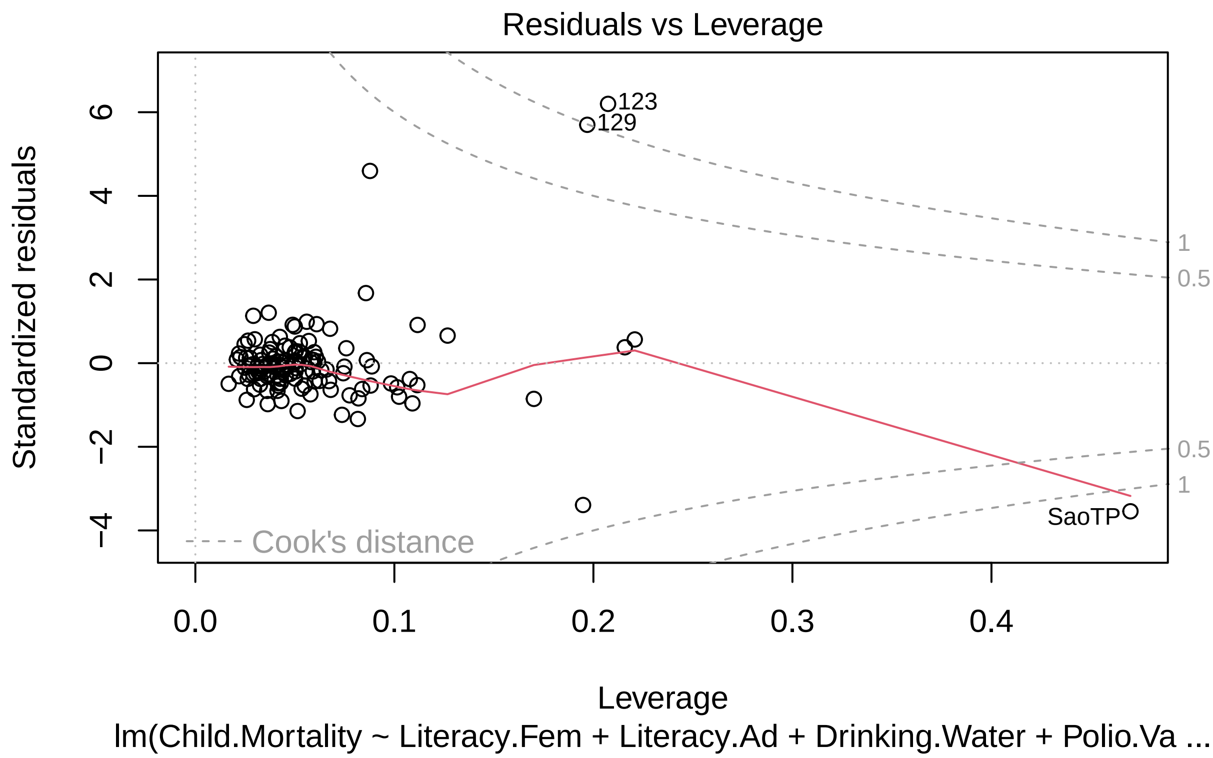
```
plot(data.3lm)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
123
129
127
6
4
2
0
−2
−2
−1
0
1
2
lm(Child.Mortality ~ Literacy.Fem + Literacy.Ad + Drinking.Water + Polio.Va ...



Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
0
50
100
150
200
250
300
Fitted values
lm(Child.Mortality ~ Literacy.Fem + Literacy.Ad + Drinking.Water + Polio.Va ...
129
123
127




```
set.seed(1234)
data.3mm <- lmrob(Child.Mortality~Literacy.Fem+Literacy.Ad+Drinking.Water+Polio.Vacc+Tetanus.Vacc.Preg+U
summary(data.3mm)
```

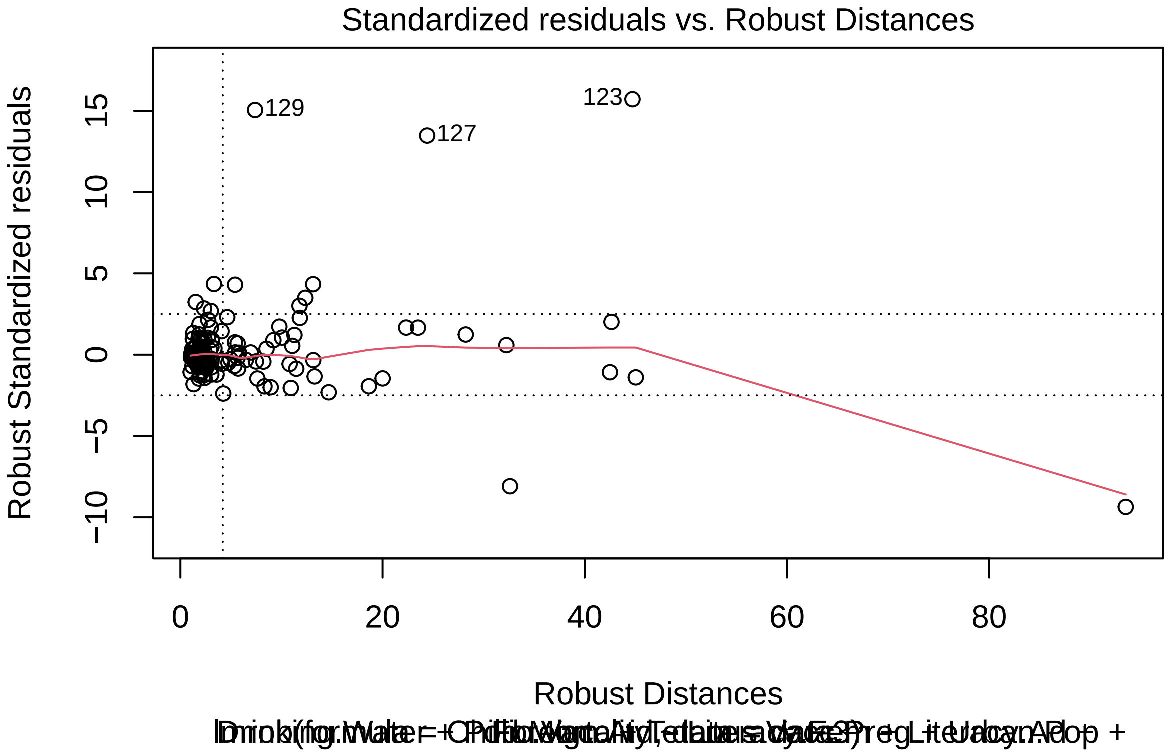
```
##
## Call:
## lmrob(formula = Child.Mortality ~ Literacy.Fem + Literacy.Ad + Drinking.Water +
##     Polio.Vacc + Tetanus.Vacc.Preg + Urban.Pop + Foreign.Aid, data = data3)
##  \--> method = "MM"
## Residuals:
##       Min        1Q    Median        3Q       Max
## -249.6158  -16.3566   -0.4436   24.3018  418.9458
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        283.7133    42.2362   6.717 6.11e-10 ***
## Literacy.Fem        -0.9203     0.6519  -1.412  0.16053
## Literacy.Ad         -0.3440     0.5106  -0.674  0.50181
## Drinking.Water      -0.6691     0.2023  -3.307  0.00124 **
## Polio.Vacc          -0.6301     0.3605  -1.748  0.08300 .
## Tetanus.Vacc.Preg  -0.1710     0.1531  -1.116  0.26641
## Urban.Pop           -0.1812     0.2119  -0.855  0.39408
## Foreign.Aid          1.3134     0.3969   3.309  0.00123 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

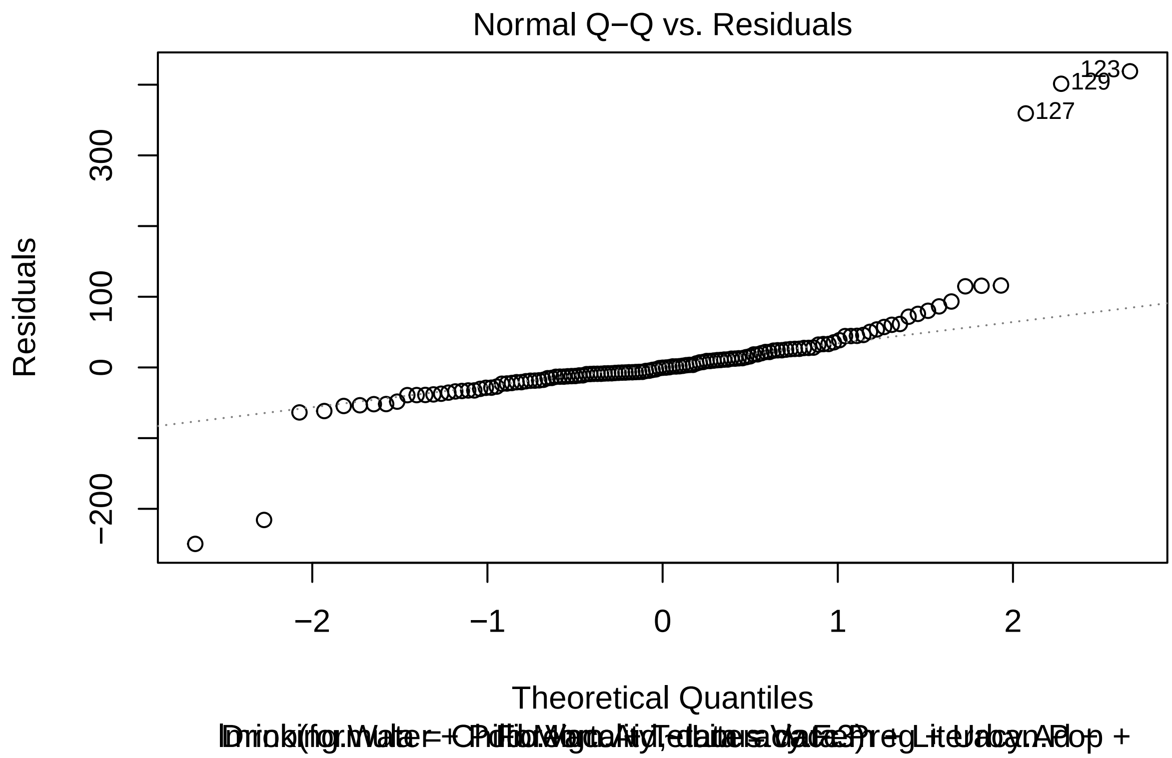
```
##
## Robust residual standard error: 26.67
## Multiple R-squared:  0.8226, Adjusted R-squared:  0.8125
## Convergence in 26 IRWLS iterations
##
## Robustness weights:
##  5 observations c(91,123,127,129,131) are outliers with |weight| = 0 ( < 0.00076);
##  9 weights are ~= 1. The remaining 117 ones are summarized as
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01942 0.83060 0.94480 0.85870 0.98650 0.99870
## Algorithmic parameters:
##        tuning.chi                bb        tuning.psi        refine.tol
##         1.548e+00         5.000e-01         4.685e+00         1.000e-07
##           rel.tol         scale.tol         solve.tol       eps.outlier
##         1.000e-07         1.000e-10         1.000e-07         7.634e-04
##             eps.x warn.limit.reject warn.limit.meanrw
##         3.165e-10         5.000e-01         5.000e-01
##      nResample         max.it       best.r.s       k.fast.s          k.max
##            500             50              2              1            200
##    maxit.scale      trace.lev            mts     compute.rd fast.s.large.n
##            200              0           1000              0           2000
##                   psi           subsampling                   cov
##            "bisquare"         "nonsingular"         ".vcov.avar1"
## compute.outlier.stats
##                  "SM"
## seed : int(0)
```

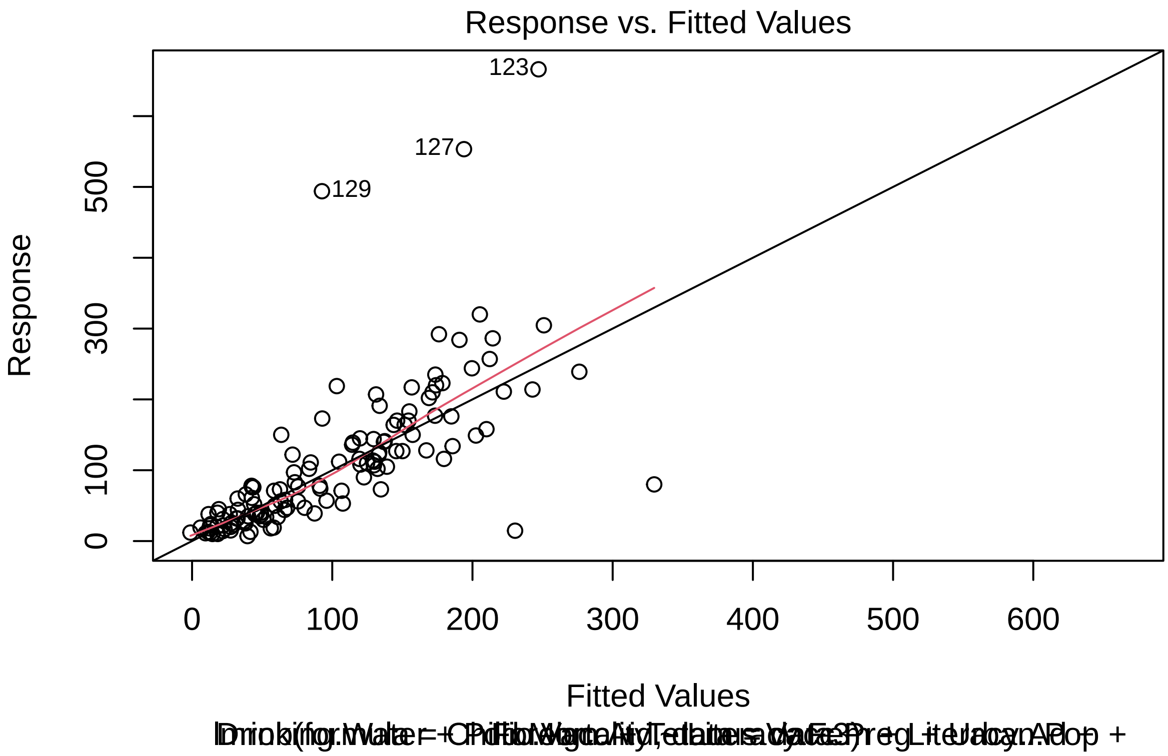
```
plot(data.3mm)
```

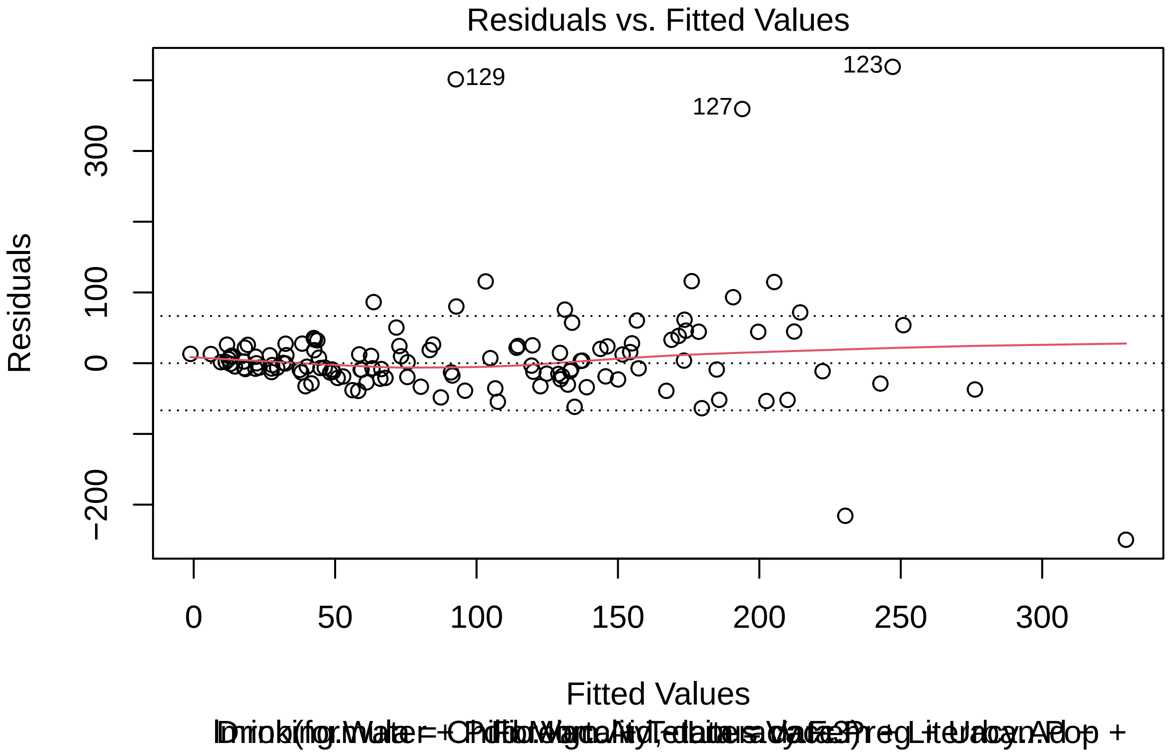
```
## recomputing robust Mahalanobis distances
```

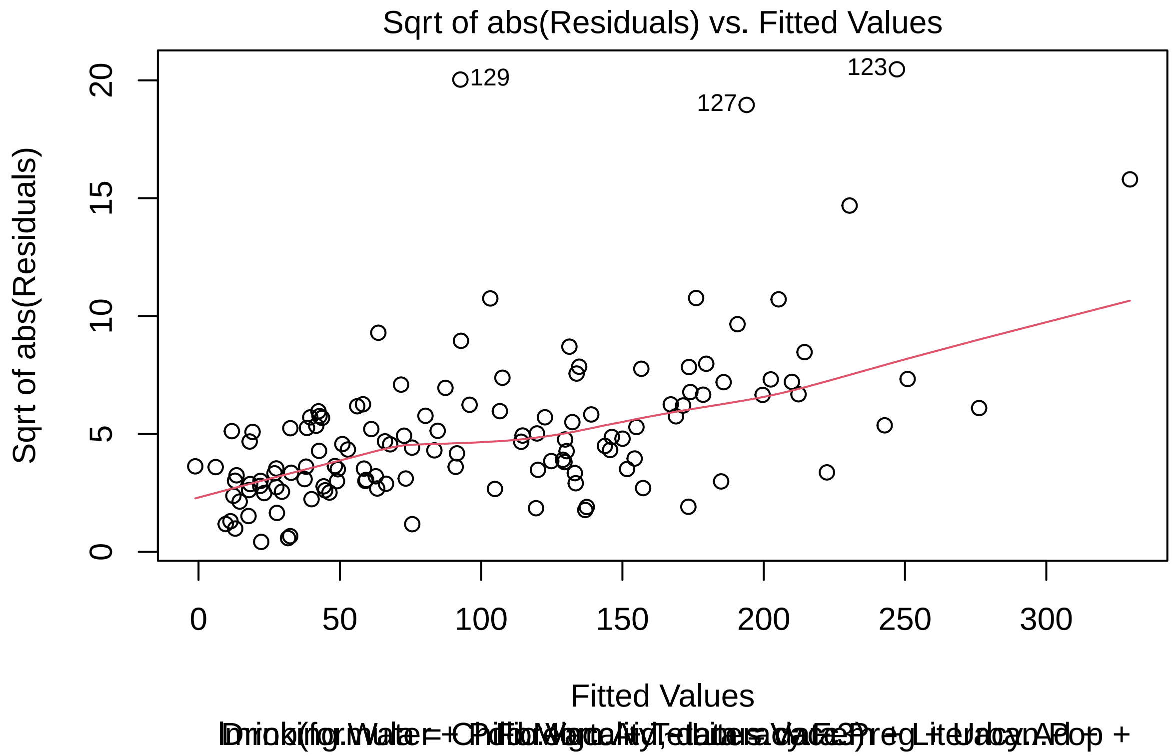
```
## saving the robust distances 'MD' as part of 'data.3mm'
```


Standardized residuals vs. Robust Distances
Robust Standardized residuals
Robust Distances
129
127
123
−10
−5
0
5
10
15
0
20
40
60
80


Normal Q−Q vs. Residuals
Residuals
−200
0
100
300
123
129
127
−2
−1
0
1
2
Theoretical Quantiles


Response vs. Fitted Values
123
127
129
Response
0
100
300
500
0
100
200
300
400
500
600
Fitted Values


Residuals vs. Fitted Values
Residuals
Fitted Values
−200
0
100
300
0
50
100
150
200
250
300
129
127
123



How many of the added outliers does lmrob identify as outliers? Is the original "obvious outlier" still identified as outlier?

There are very influent leverage points (127,123,129) that can be distinguished in both plots. The function lmrob has found only 5 outliers. 5 observations c(91,123,127,129,131) are outliers. In the dataset[-80] there were 4 and 91, but here the point 4 is no longer considered as an outlier, only 91 remains.