EDA- Final Project

Nini Curcione

2022-11-01

Loading Libraries

```
library(mosaic)
```

```
## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'tibble'

## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'pillar'

## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'hms'

## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2

##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.

##
## Attaching package: 'mosaic'

## The following objects are masked from 'package:dplyr':
##
##     count, do, tally

## The following object is masked from 'package:Matrix':
##
##     mean

## The following object is masked from 'package:ggplot2':
##
##     stat

## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
```

```
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
library(readr)
library(ggplot2);
library(readxl)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.4     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x mosaic::count()            masks dplyr::count()
## x purrr::cross()             masks mosaic::cross()
## x mosaic::do()               masks dplyr::do()
## x tidyr::expand()            masks Matrix::expand()
## x dplyr::filter()            masks stats::filter()
## x ggstance::geom_errorbarh() masks ggplot2::geom_errorbarh()
## x dplyr::lag()               masks stats::lag()
## x tidyr::pack()              masks Matrix::pack()
## x mosaic::stat()             masks ggplot2::stat()
## x mosaic::tally()            masks dplyr::tally()
## x tidyr::unpack()            masks Matrix::unpack()
```

```
library(stats)
library(mosaic)
library(dplyr)
```

Importing Dataset

```
Dataset1 <- read.csv("~/Desktop/DATA231/Project /Data/Dataset1.csv")
```

5 Number Summary and Histogram for Life Expectancy

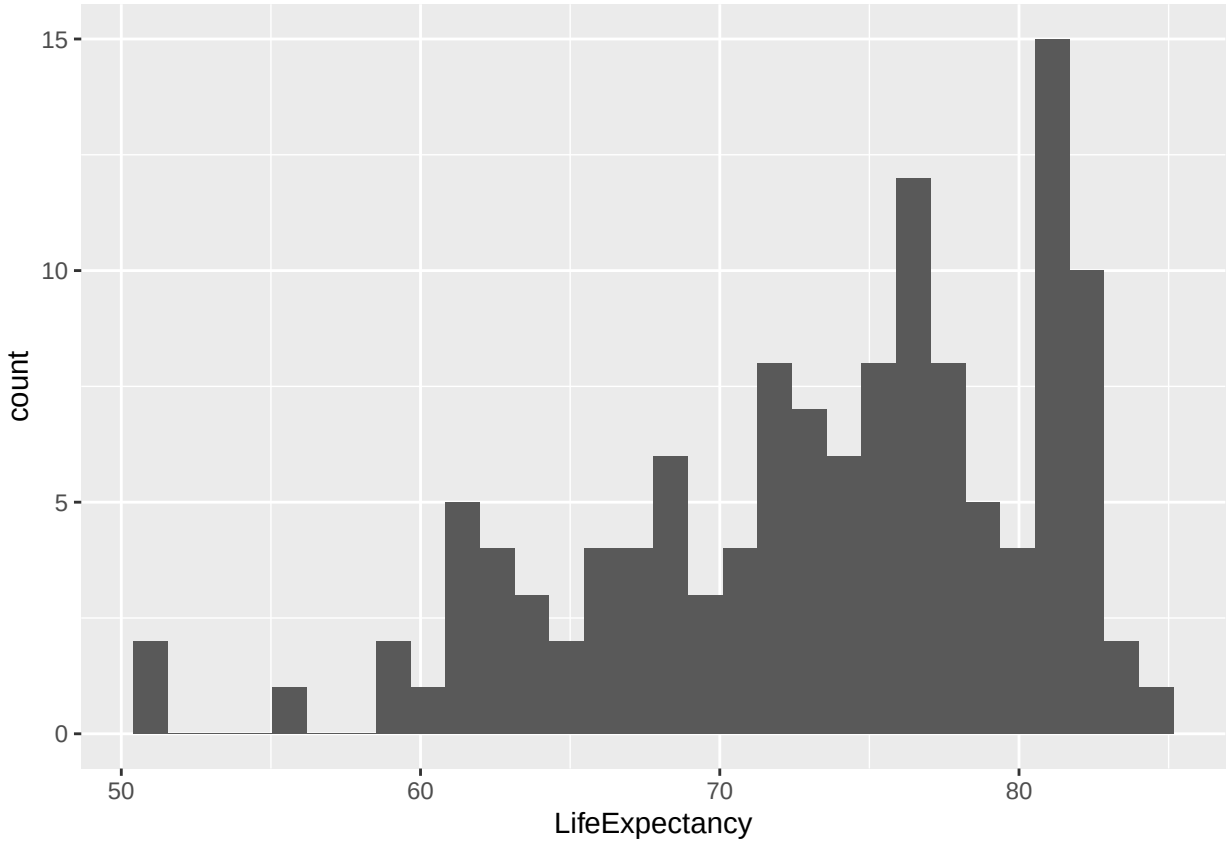
```
favstats(Dataset1$LifeExpectancy)
```

```
##   min   Q1 median    Q3  max     mean       sd   n missing
##  50.5 68.6   74.9 79.35 84.1 73.44567 7.352451 127       4
```

```
ggplot(data=Dataset1) +
  geom_histogram(mapping = aes(x=LifeExpectancy))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



5 Number Summary and Histogram for GDP Per Capita

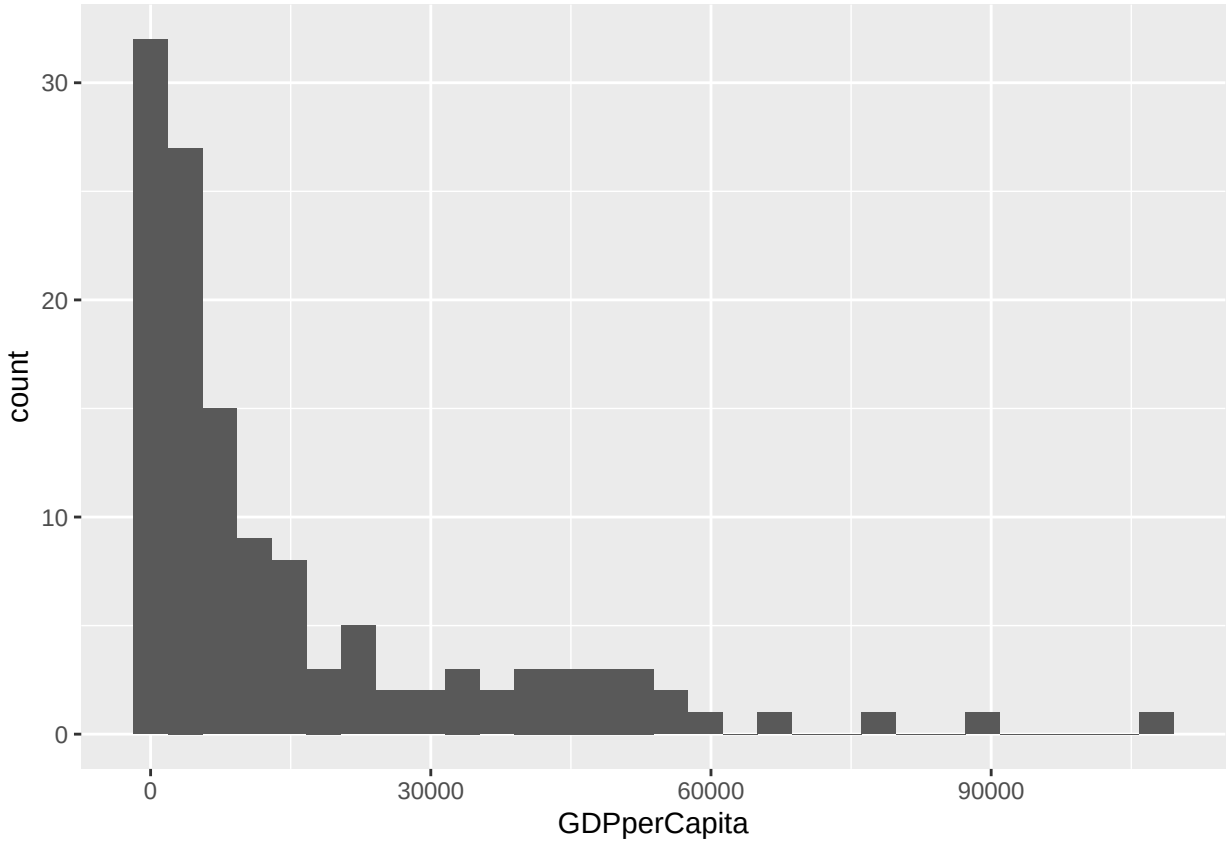
```
favstats(Dataset1$GDPperCapita)
```

```
##  min   Q1 median    Q3    max     mean       sd   n missing
##  347 1955   6380 21400 108000 15547.69 20224.24 127       4
```

```
ggplot(data=Dataset1) +
  geom_histogram(mapping = aes(x=GDPperCapita))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



5 Number Summary and Histogram for Food Supply

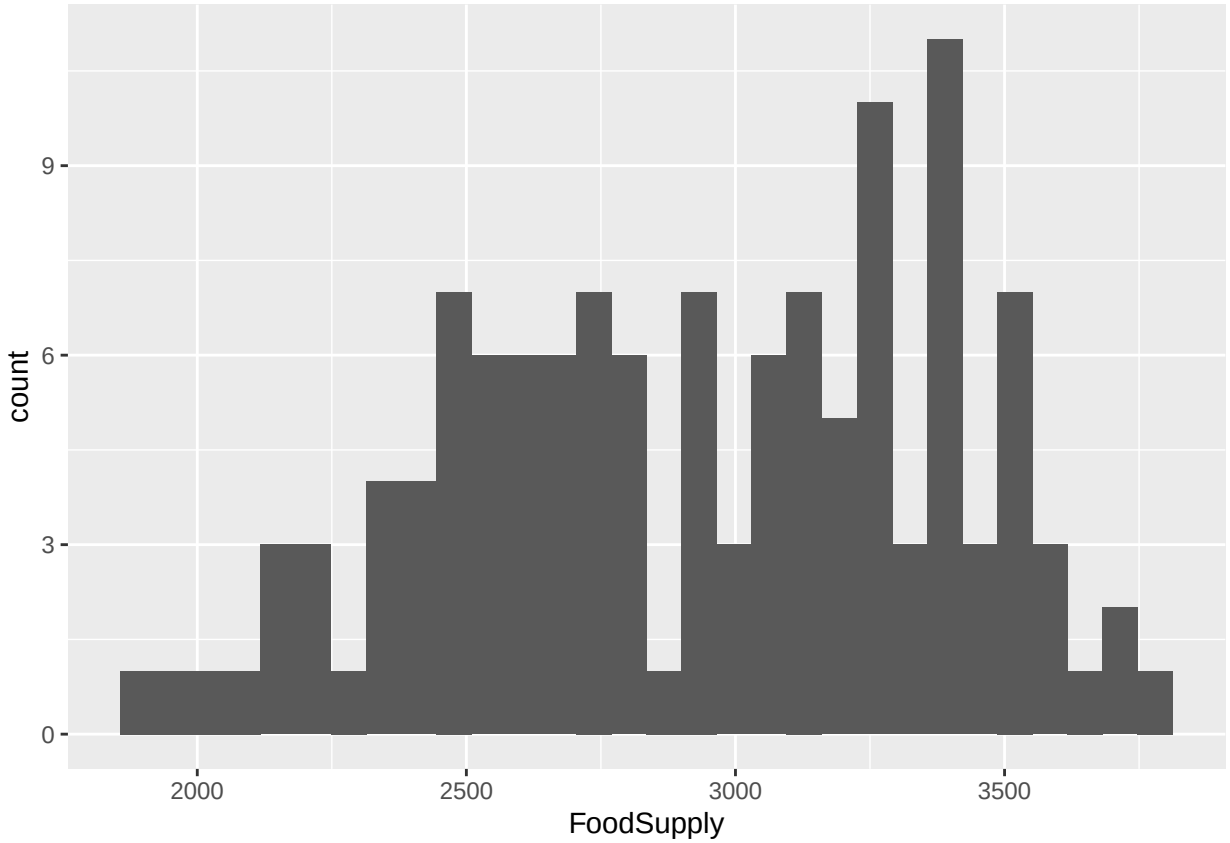
```
favstats(Dataset1$FoodSupply)
```

```
##   min   Q1 median   Q3  max     mean      sd   n missing
##  1880 2575   2950 3280 3770 2923.701 447.795 127       4
```

```
ggplot(data=Dataset1) +
  geom_histogram(mapping = aes(x=FoodSupply))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



5 Number Summary and Histogram for High Tech Export

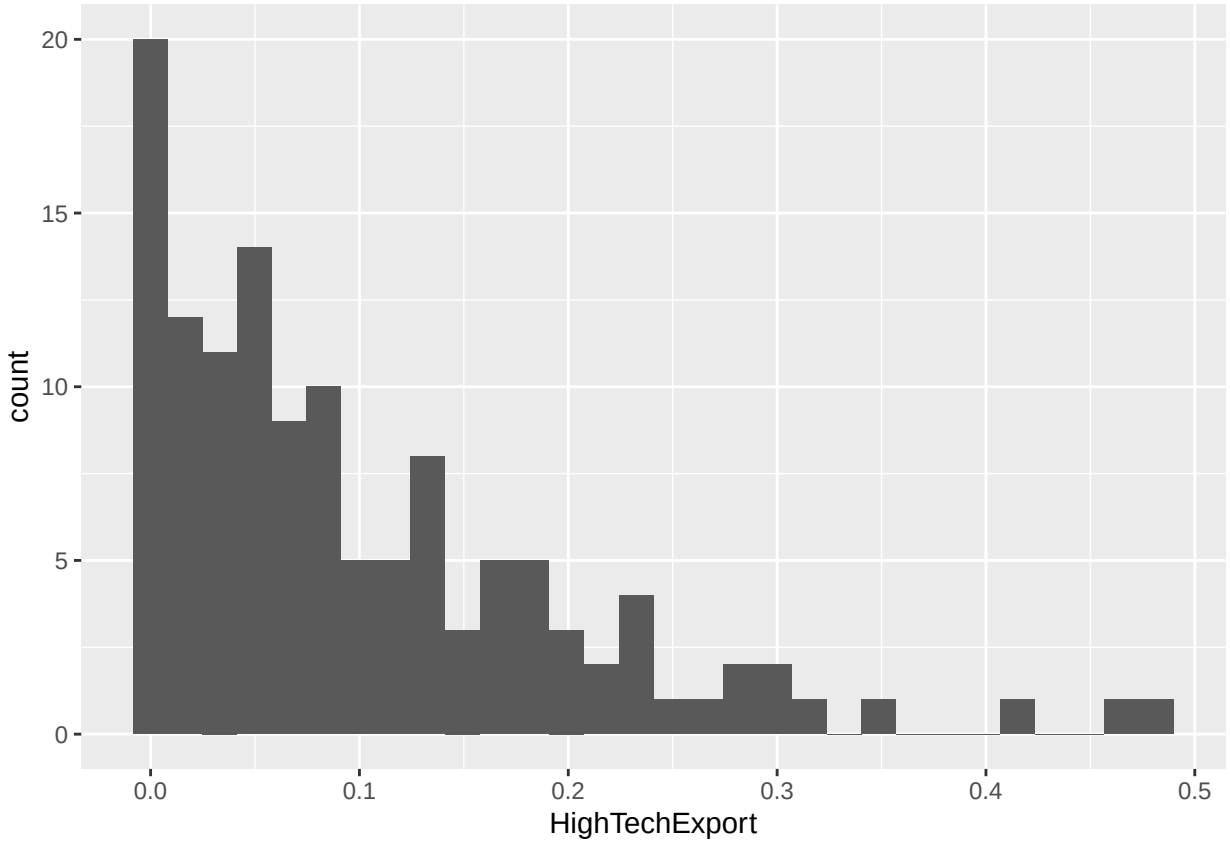
```
favstats(Dataset1$HighTechExport)
```

```
##       min    Q1 median    Q3   max       mean        sd   n missing
##  0.000554 0.025 0.0697 0.147 0.482 0.09970608 0.1001542 127       4
```

```
ggplot(data=Dataset1) +
  geom_histogram(mapping = aes(x=HighTechExport))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



5 Number Summary and Histogram for School Years

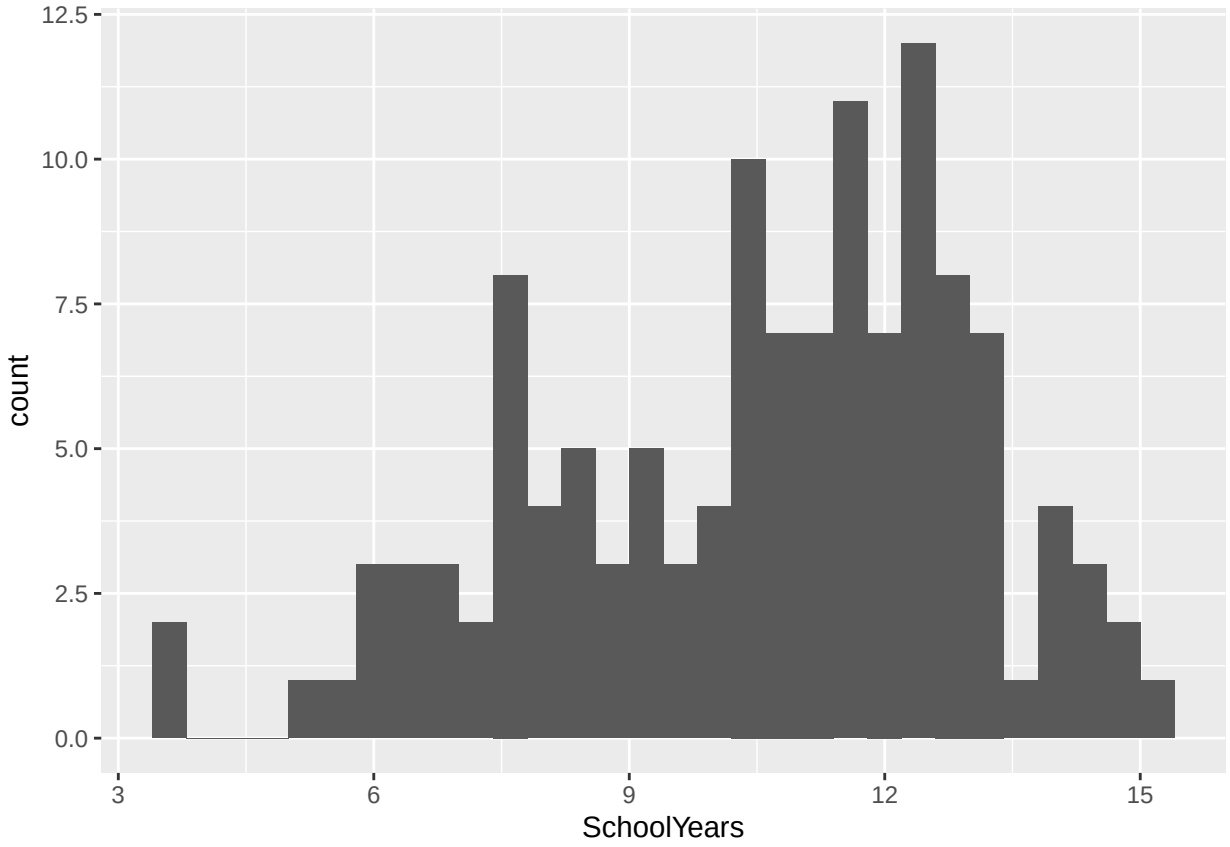
```
favstats(Dataset1$SchoolYears)
```

```
##  min    Q1 median   Q3  max    mean       sd   n missing
##  3.6 8.535     11 12.4 15.2 10.5552 2.499803 127       4
```

```
ggplot(data=Dataset1) +
  geom_histogram(mapping = aes(x=SchoolYears))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



5 Number Summary and Histogram for Basic Sanitation

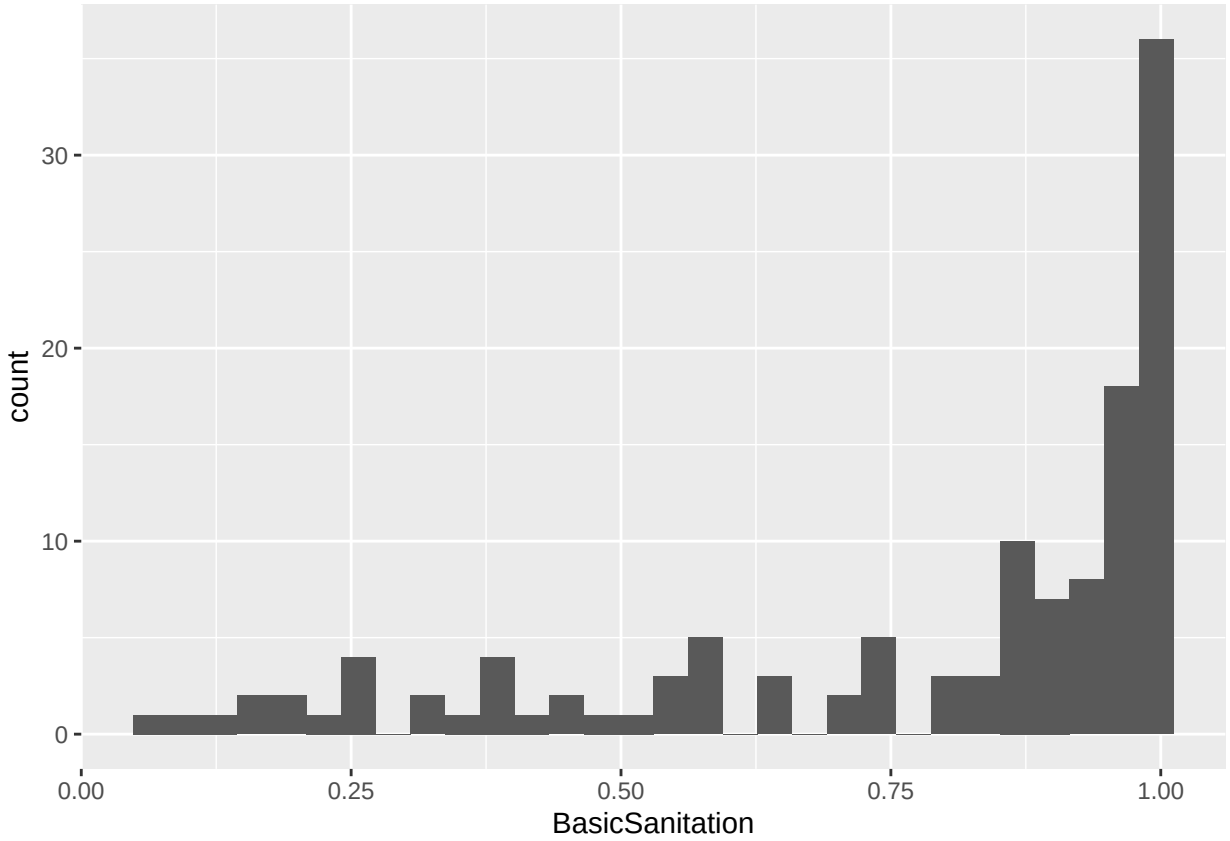
```
favstats(Dataset1$BasicSanitation)
```

```
##     min     Q1 median     Q3 max      mean        sd   n missing
##  0.0686 0.6135   0.91 0.9835   1 0.7811339 0.2700192 127       4
```

```
ggplot(data=Dataset1) +
  geom_histogram(mapping = aes(x=BasicSanitation))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



5 Number Summary and Histogram for Alcohol Consumption

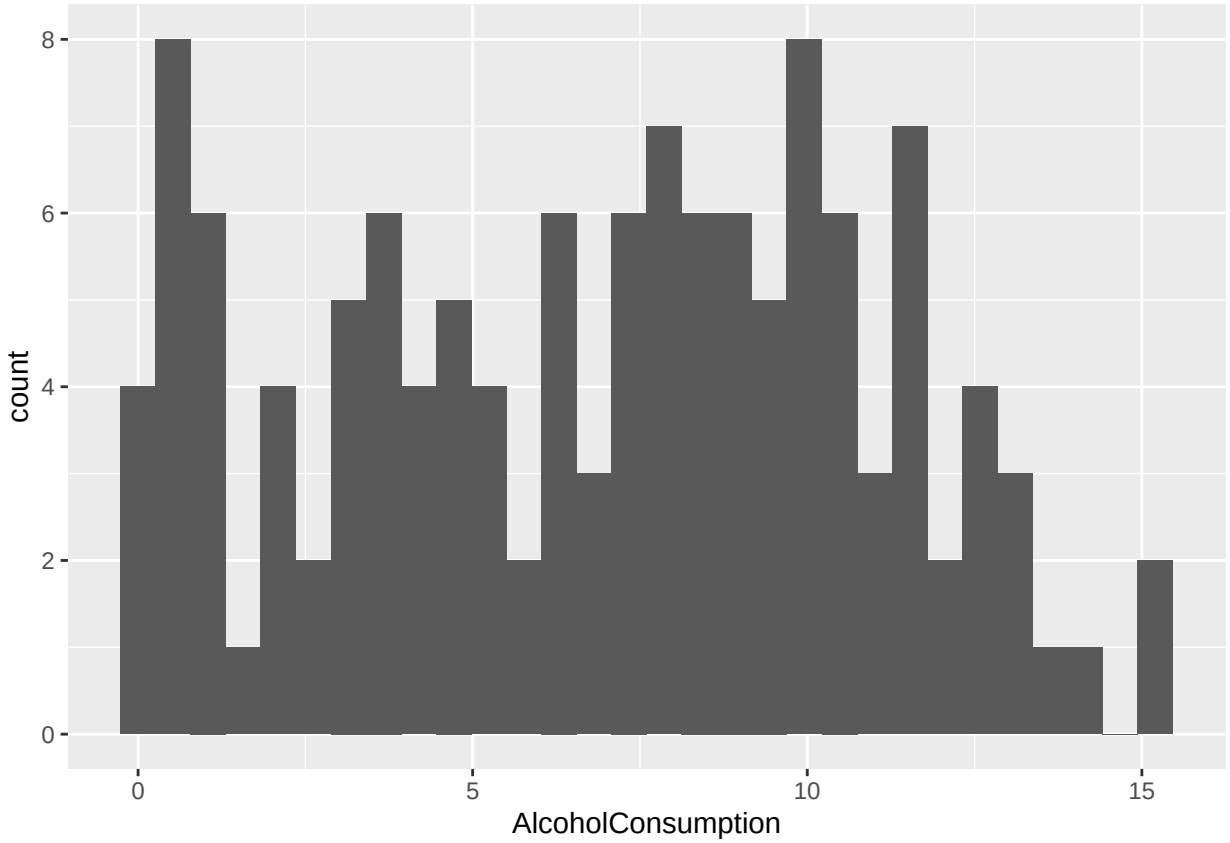
```
favstats(Dataset1$AlcoholConsumption)
```

```
##  min   Q1 median  Q3  max     mean      sd   n missing
##    0 3.75    7.5 9.9 15.2 6.866142 4.04179 127       4
```

```
ggplot(data=Dataset1) +
  geom_histogram(mapping = aes(x=AlcoholConsumption))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



5 Number Summary and Histogram for BroadBand Subscribers

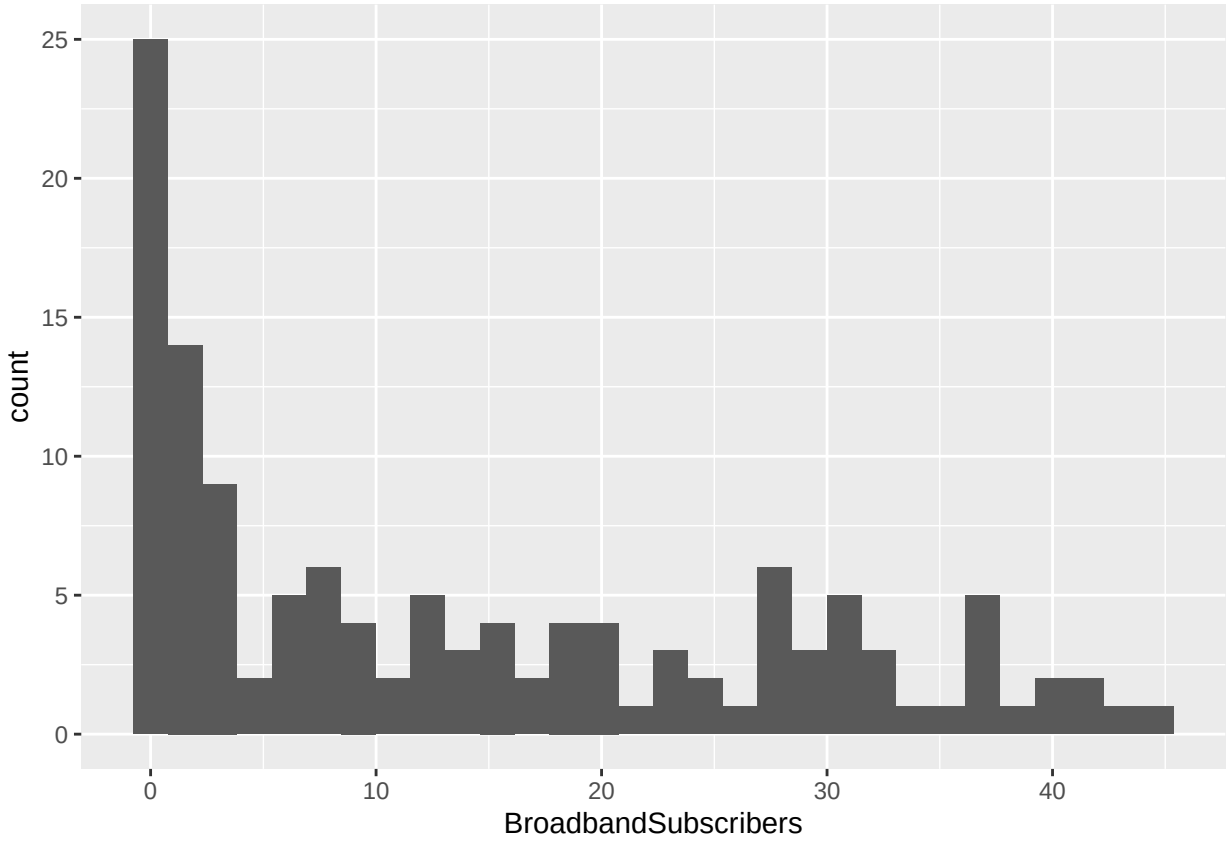
```
favstats(Dataset1$BroadbandSubscribers)
```

```
##      min   Q1 median   Q3  max     mean       sd   n missing
##  0.00875 1.45   9.42 25.8 44.6 13.86721 13.41206 127       4
```

```
ggplot(data=Dataset1) +
  geom_histogram(mapping = aes(x=BroadbandSubscribers))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



5 Number Summary and Histogram for MCV

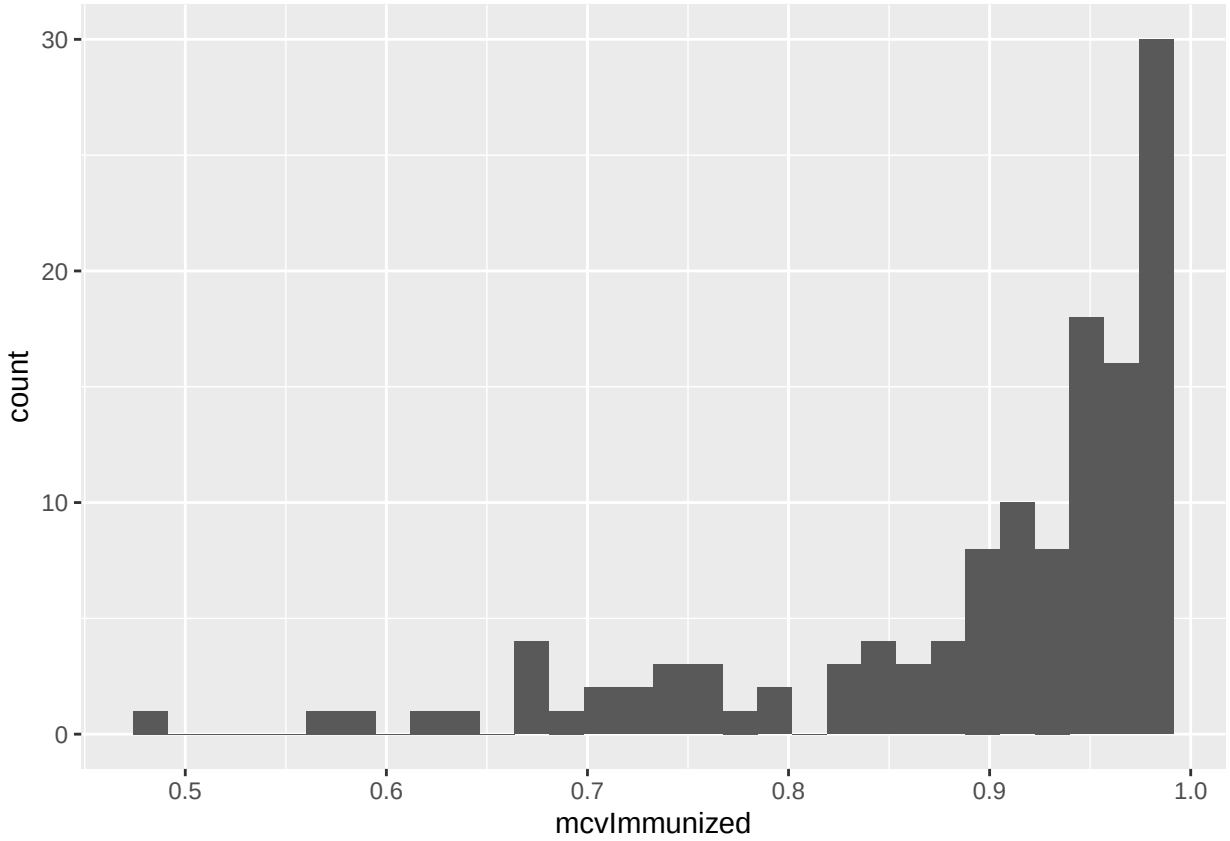
```
favstats(Dataset1$mcvImmunized)
```

```
##   min   Q1 median   Q3  max      mean        sd   n missing
##  0.49 0.87   0.94 0.97 0.99 0.8954331 0.1073989 127       4
```

```
ggplot(data=Dataset1) +
  geom_histogram(mapping = aes(x=mcvImmunized))
```

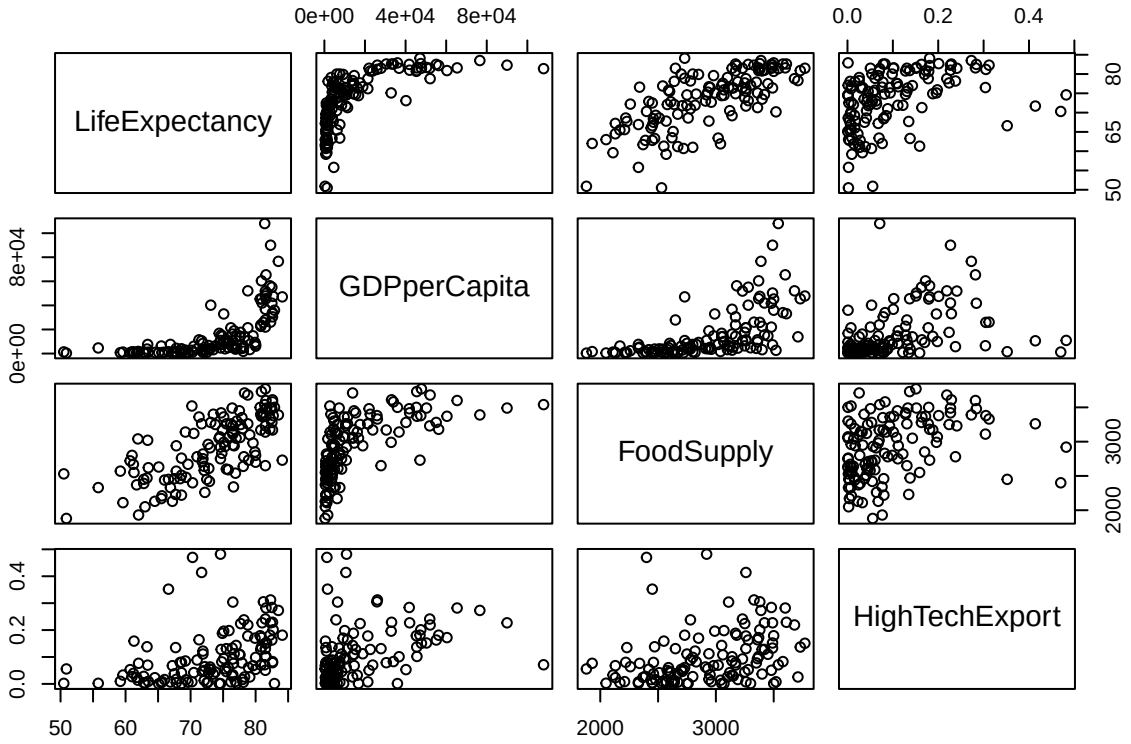
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```

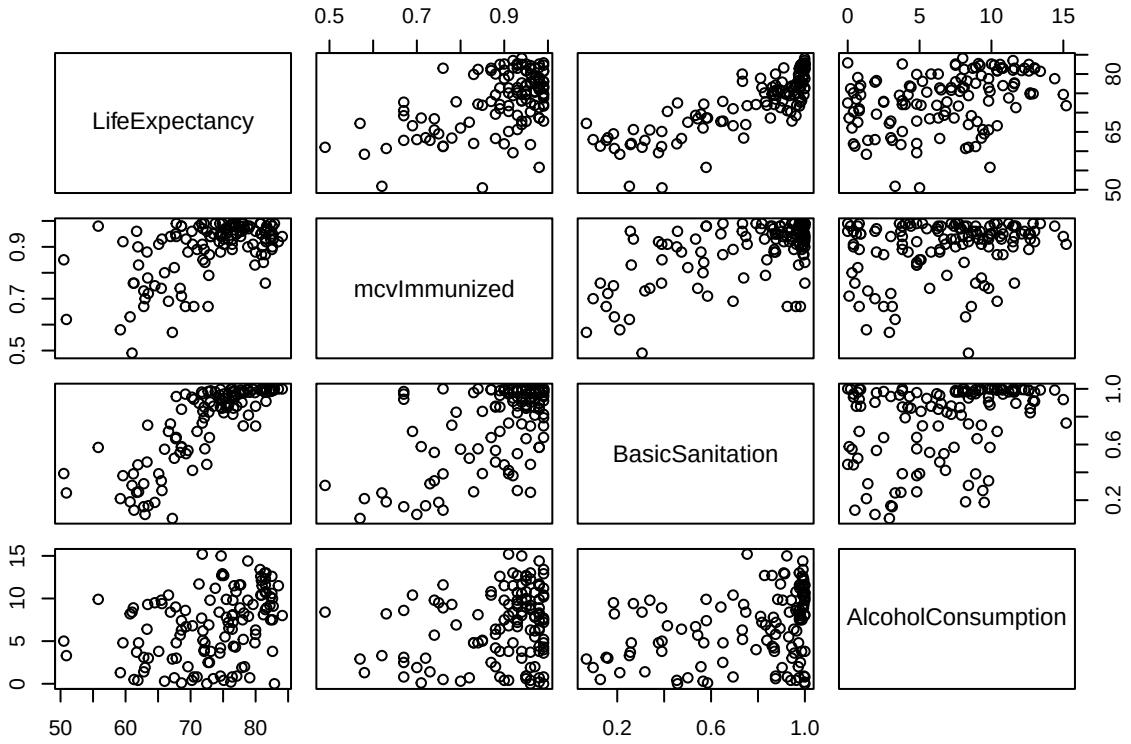


We will plot LifeExpectancy against each of the potential explanatory variables.

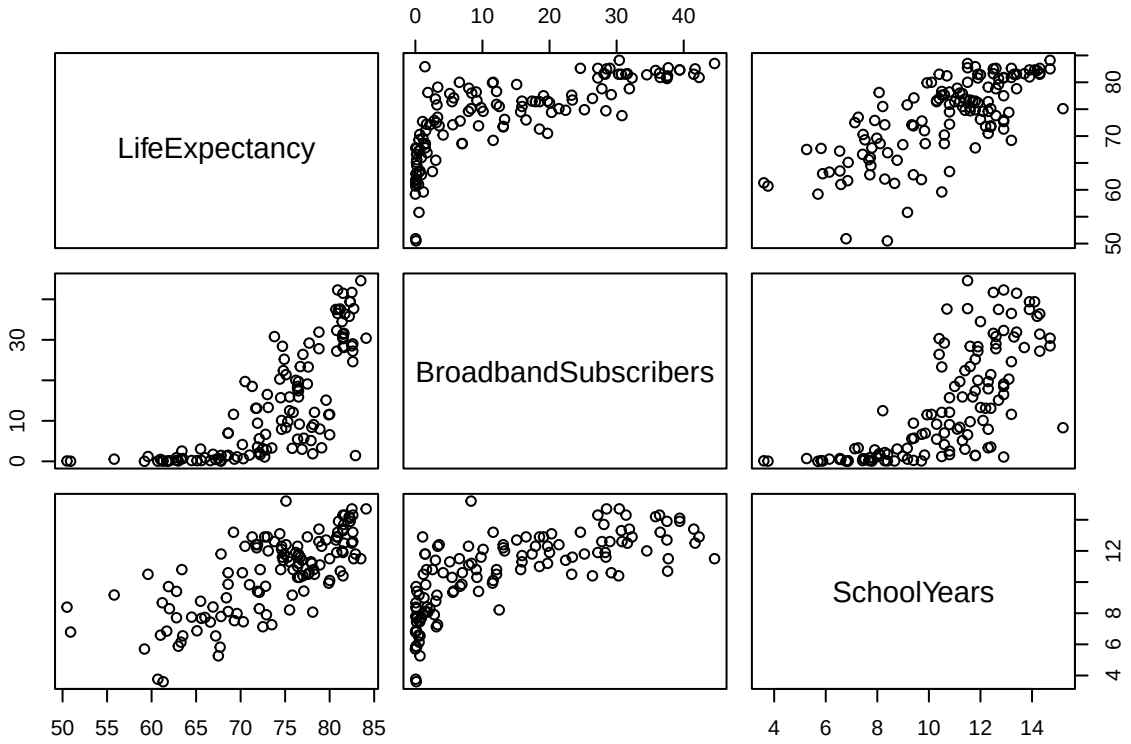
```
pairs(~LifeExpectancy + GDPperCapita + FoodSupply + HighTechExport, data = Dataset1)
```




```
pairs(~LifeExpectancy + mcvImmunized + BasicSanitation + AlcoholConsumption, data = Dataset1)
```




```
pairs(~LifeExpectancy + BroadbandSubscribers + SchoolYears, data = Dataset1)
```



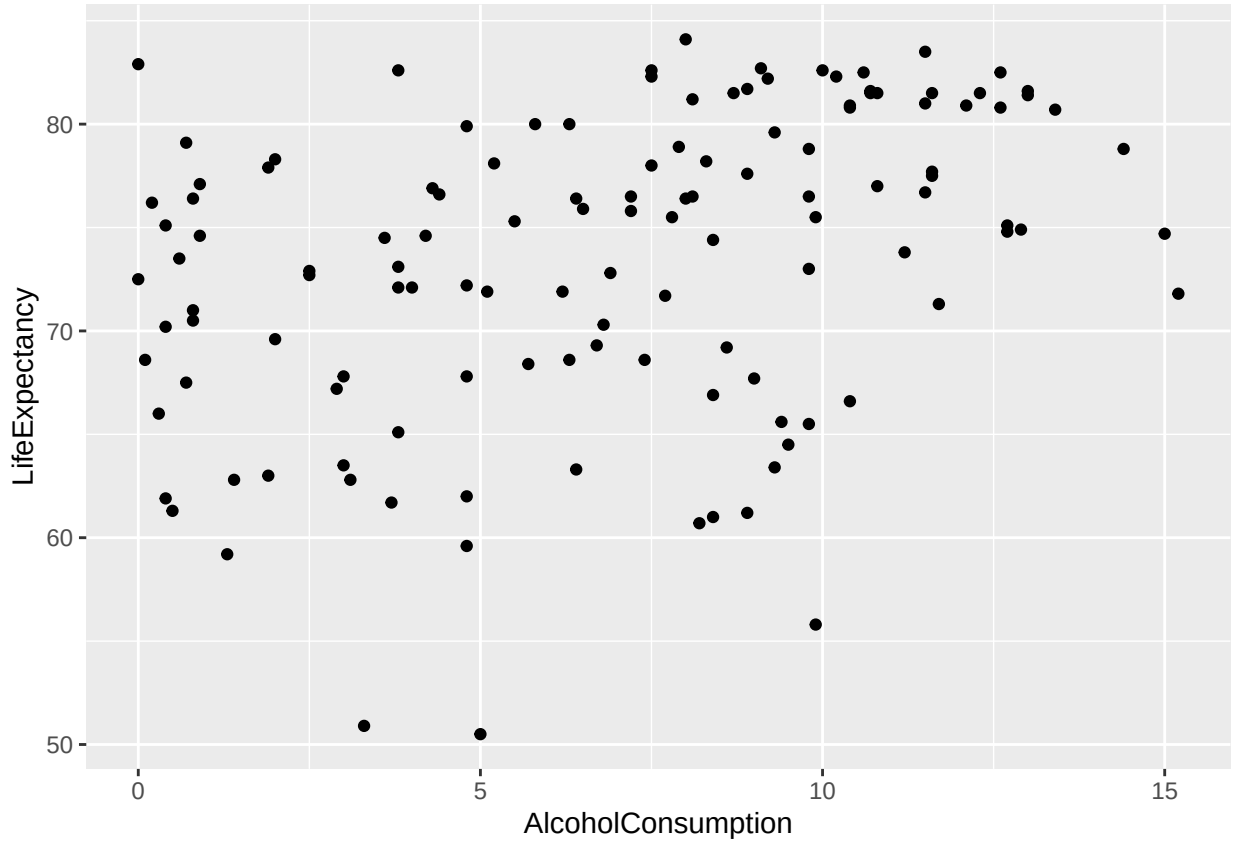
Most of the variables seem to have some kind of positive linear assosiation with LifeExpectancy, and GDPperCapita, HighTechExport and broadband subscribers have curverd relationships.

Scatterplots of each predictor against LifeExpectancy

AlcoholConsumption

```
ggplot(data=Dataset1) +
  geom_point(mapping = aes(x=AlcoholConsumption, y=LifeExpectancy)) +
    geom_abline()
```

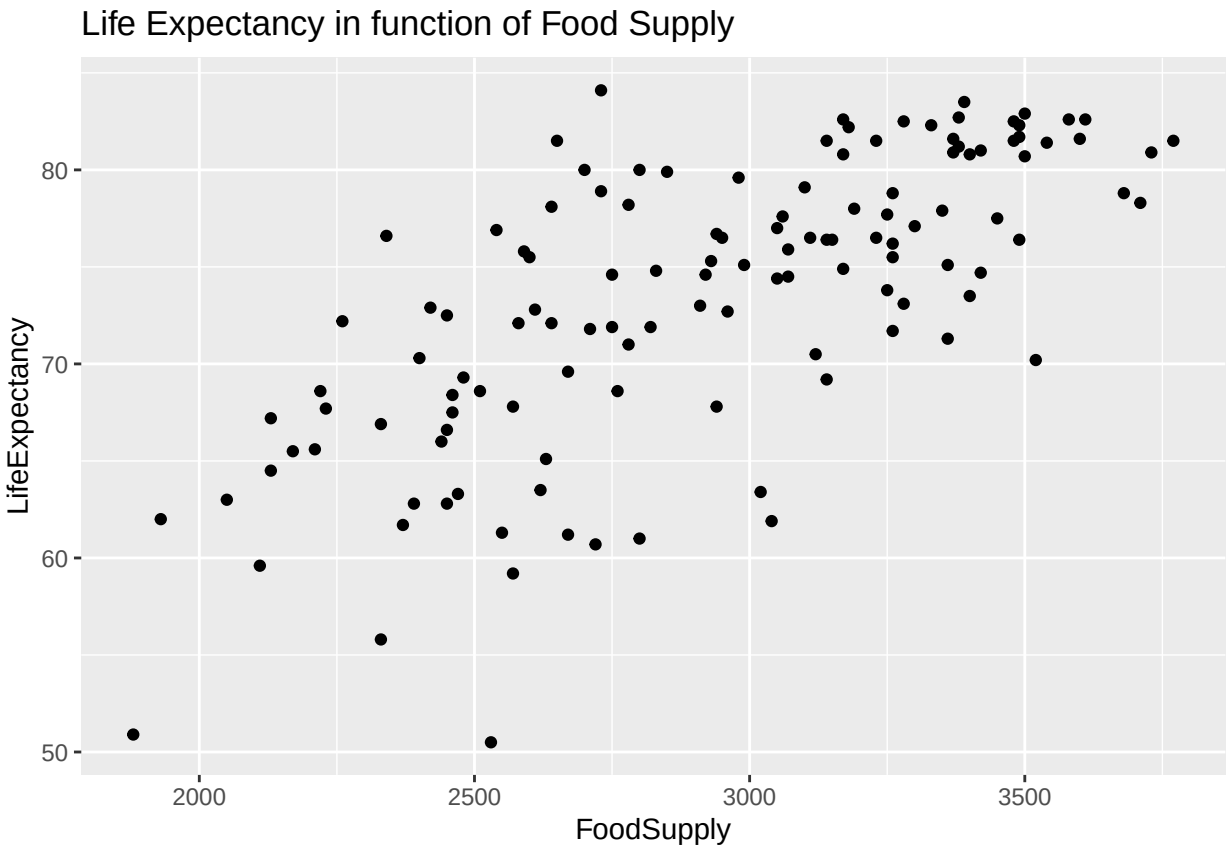
```
## Warning: Removed 4 rows containing missing values (geom_point).
```



FoodSupply

```
ggplot(data=Dataset1) +
  geom_point(mapping = aes(x=FoodSupply, y=LifeExpectancy)) + ggtitle("Life Expectancy in function of F
```

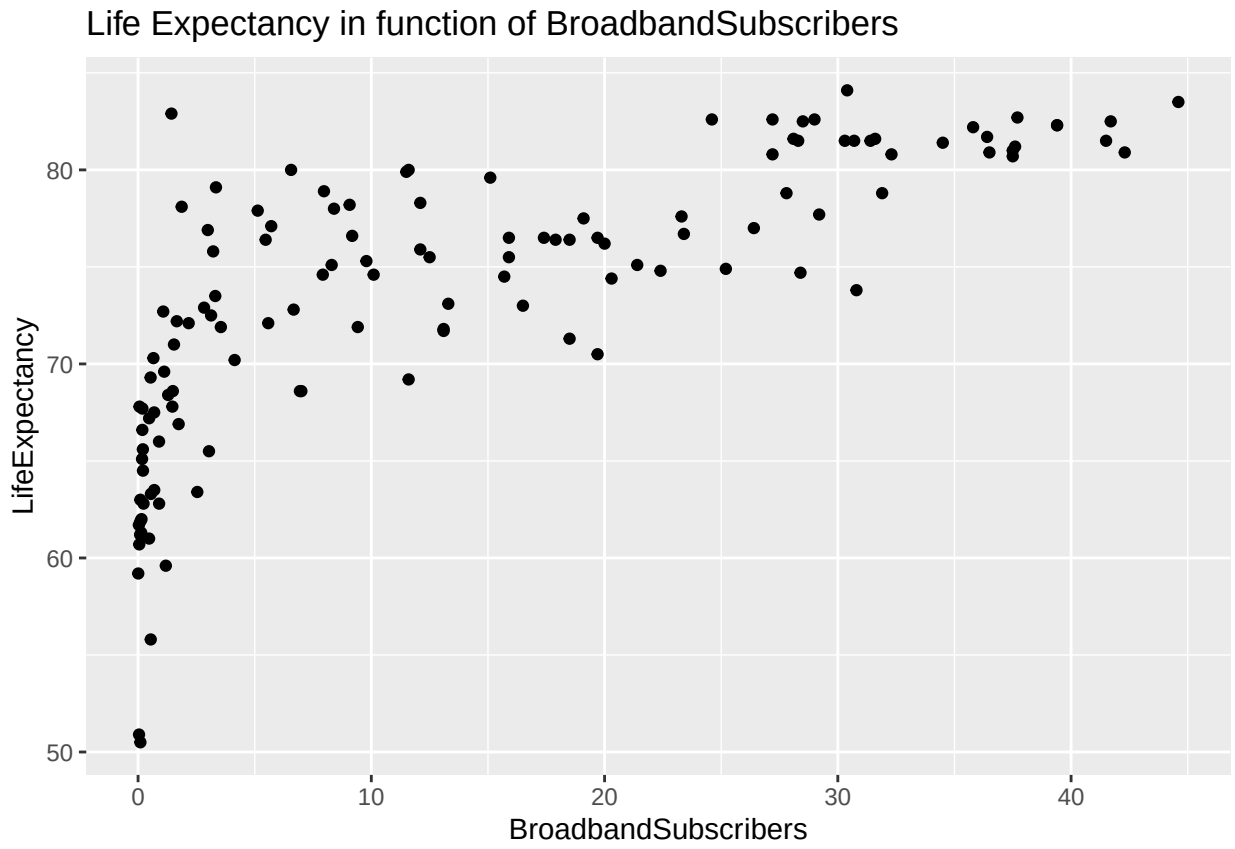
```
## Warning: Removed 4 rows containing missing values (geom_point).
```



BroadbandSubscribers

```
ggplot(data=Dataset1) +
  geom_point(mapping = aes(x=BroadbandSubscribers, y=LifeExpectancy)) + ggtitle("Life Expectancy in fun
```

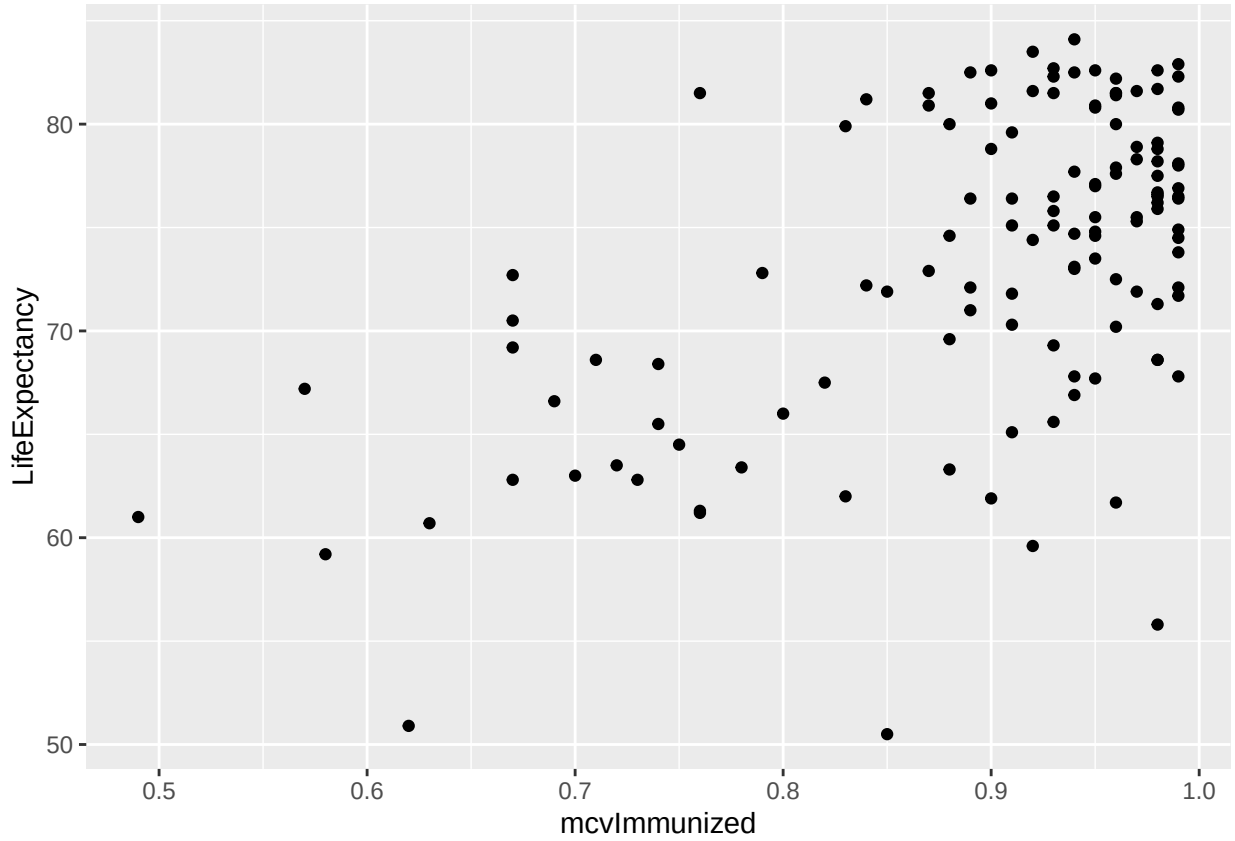
```
## Warning: Removed 4 rows containing missing values (geom_point).
```



mcvImmunized

```
ggplot(data=Dataset1) +
  geom_point(mapping = aes(x=mcvImmunized, y=LifeExpectancy)) +
    geom_abline()
```

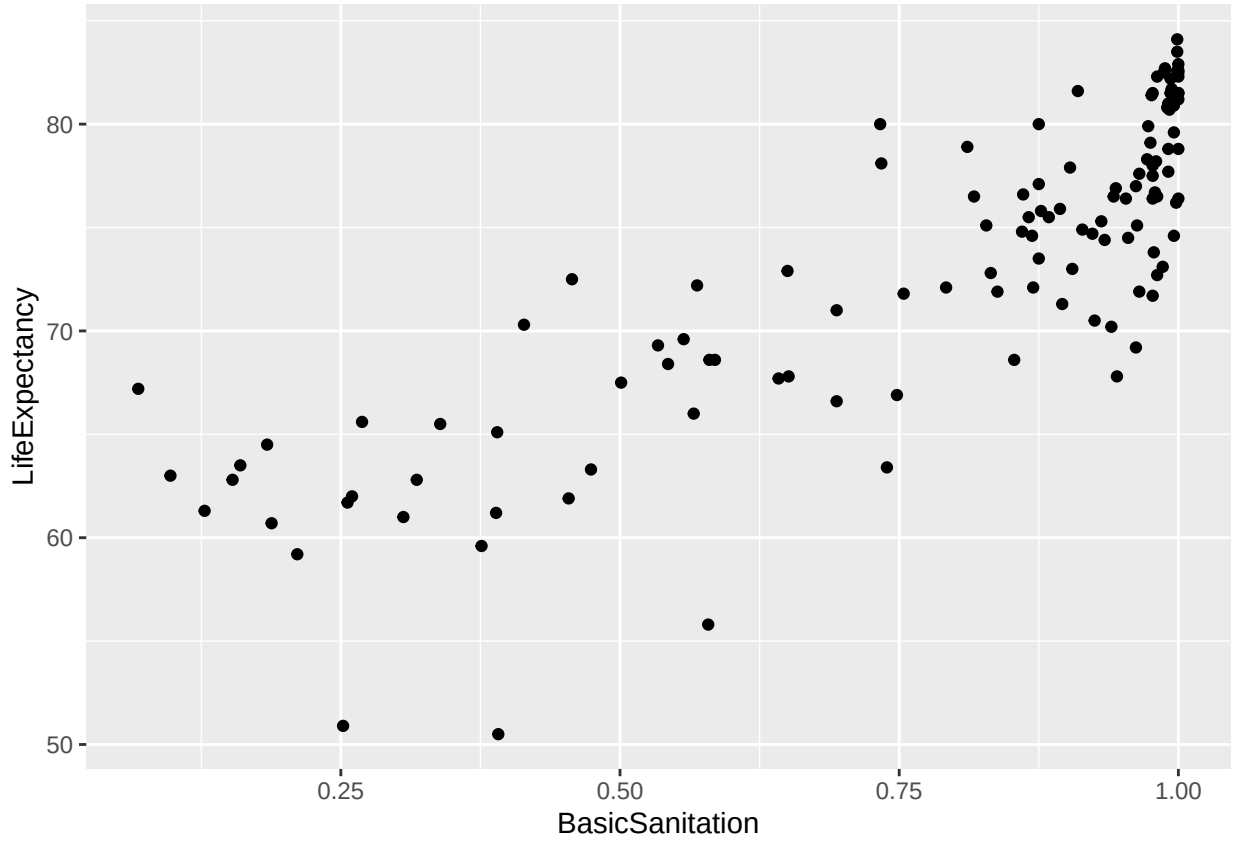
```
## Warning: Removed 4 rows containing missing values (geom_point).
```



BasicSanitation

```
ggplot(data=Dataset1) +
  geom_point(mapping = aes(x=BasicSanitation, y=LifeExpectancy)) +
    geom_abline()
```

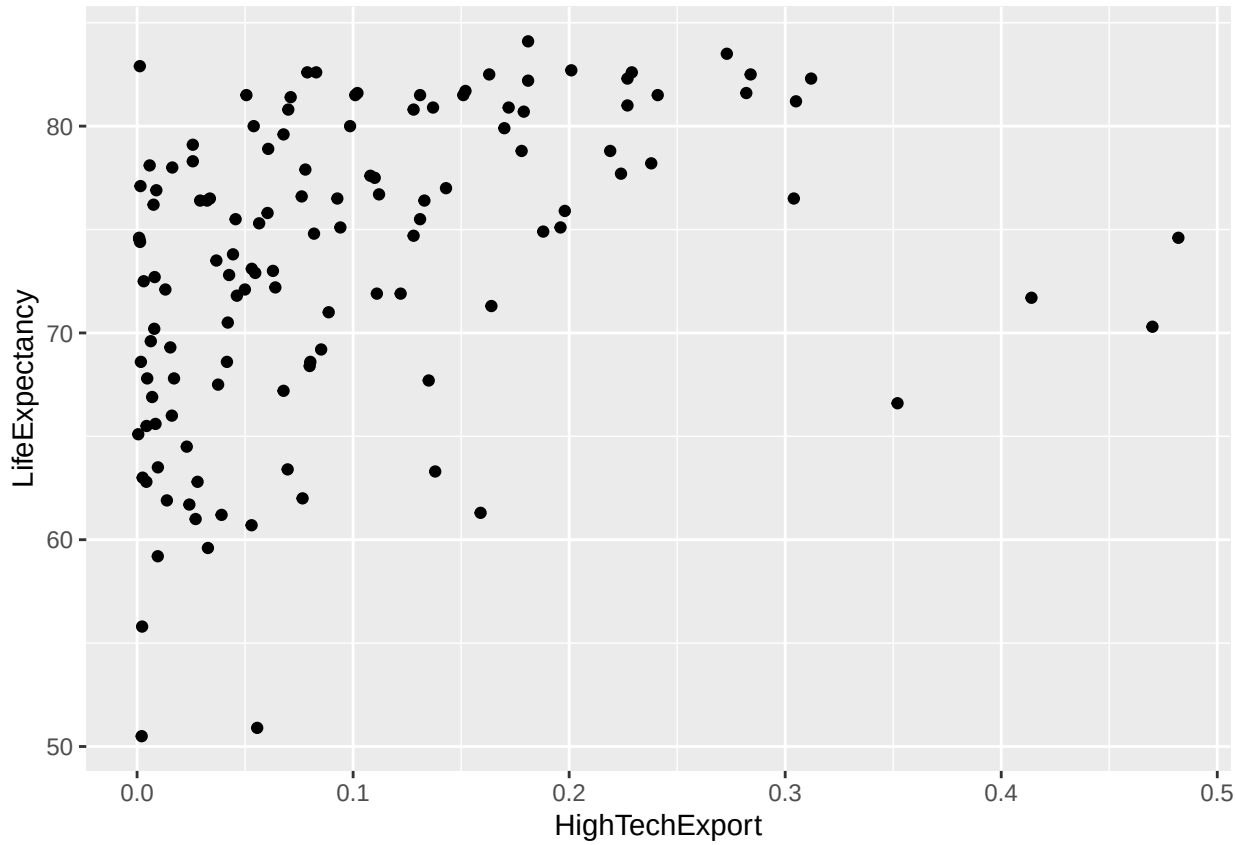
```
## Warning: Removed 4 rows containing missing values (geom_point).
```



HighTechExport

```
ggplot(data=Dataset1) +
  geom_point(mapping = aes(x=HighTechExport, y=LifeExpectancy)) +
    geom_abline()
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```

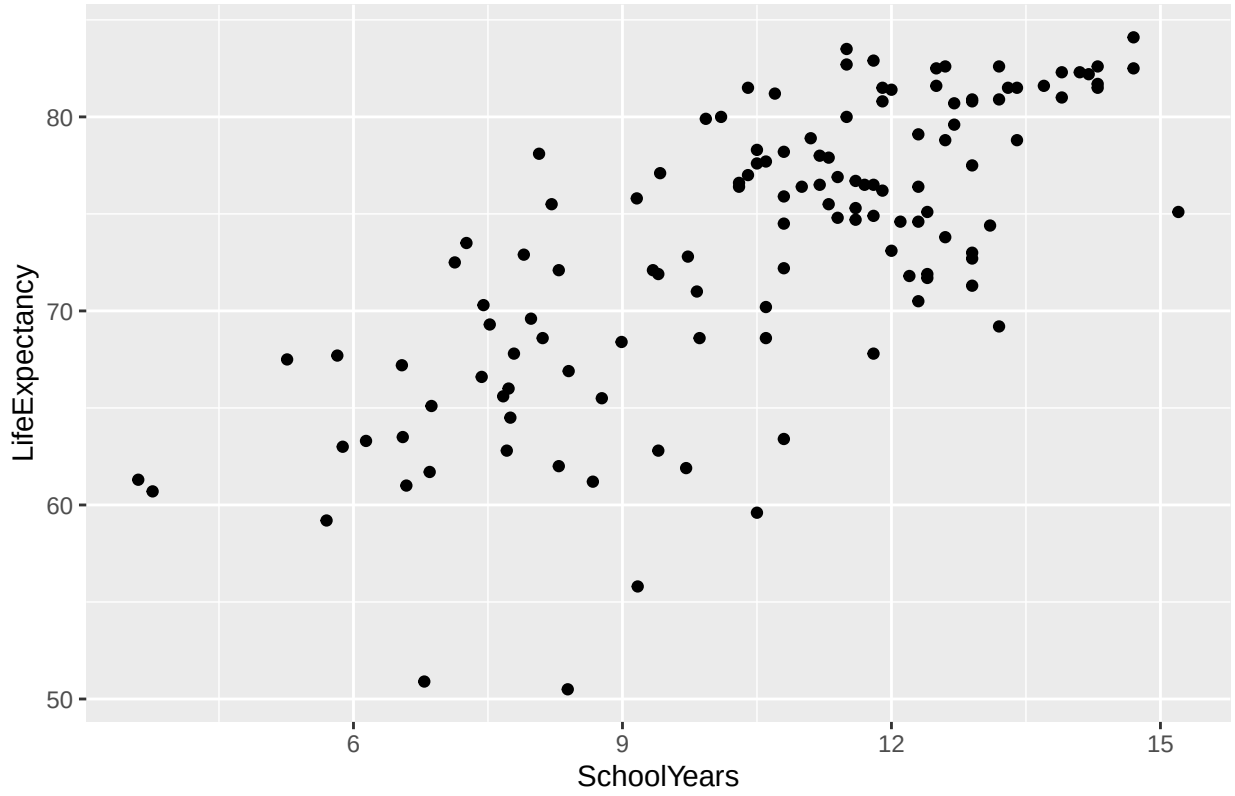


SchoolYears

```
ggplot(data=Dataset1) +
  geom_point(mapping = aes(x=SchoolYears, y=LifeExpectancy)) + ggtitle("Figure 3: Life Expectancy in fu
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```

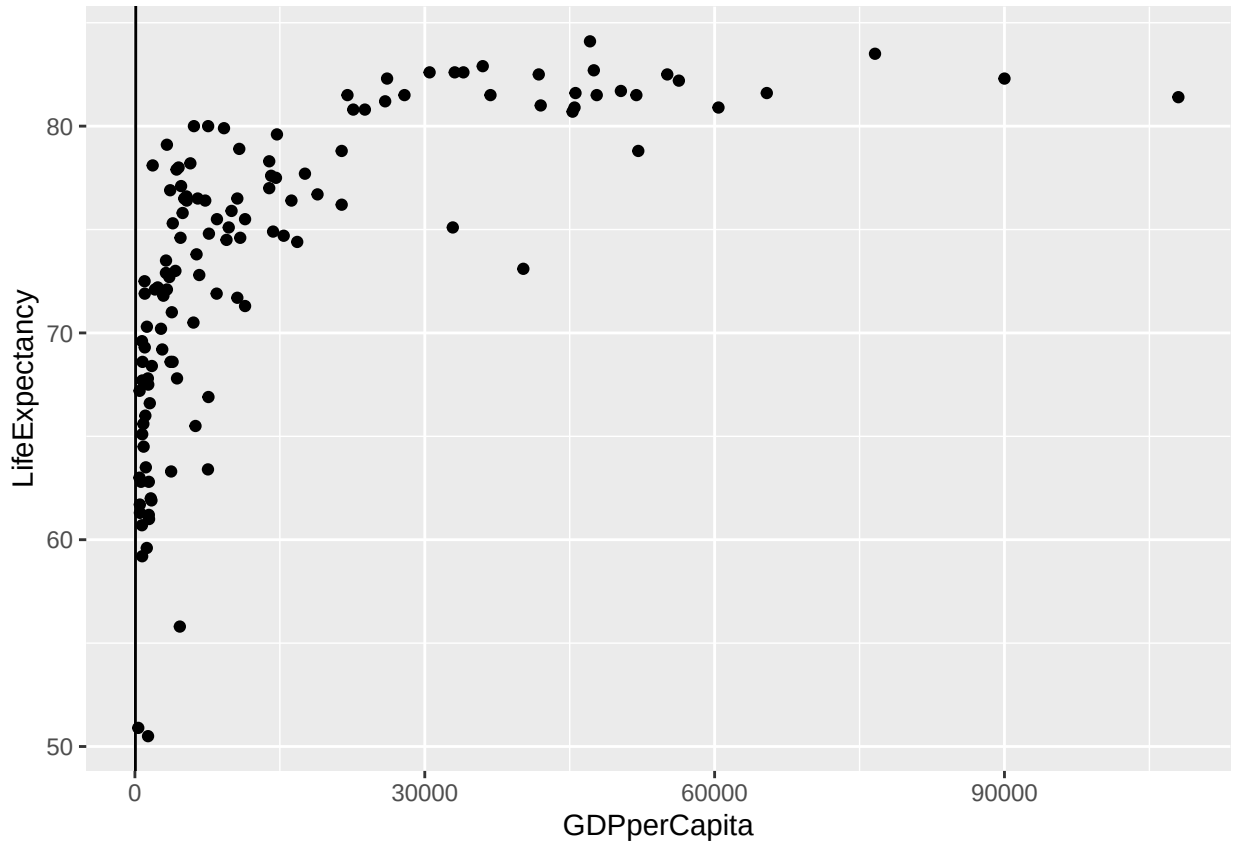
Figure 3: Life Expectancy in function of Average School Years



GDPperCapita

```
ggplot(data=Dataset1) +
  geom_point(mapping = aes(x=GDPperCapita, y=LifeExpectancy)) +
    geom_abline()
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```

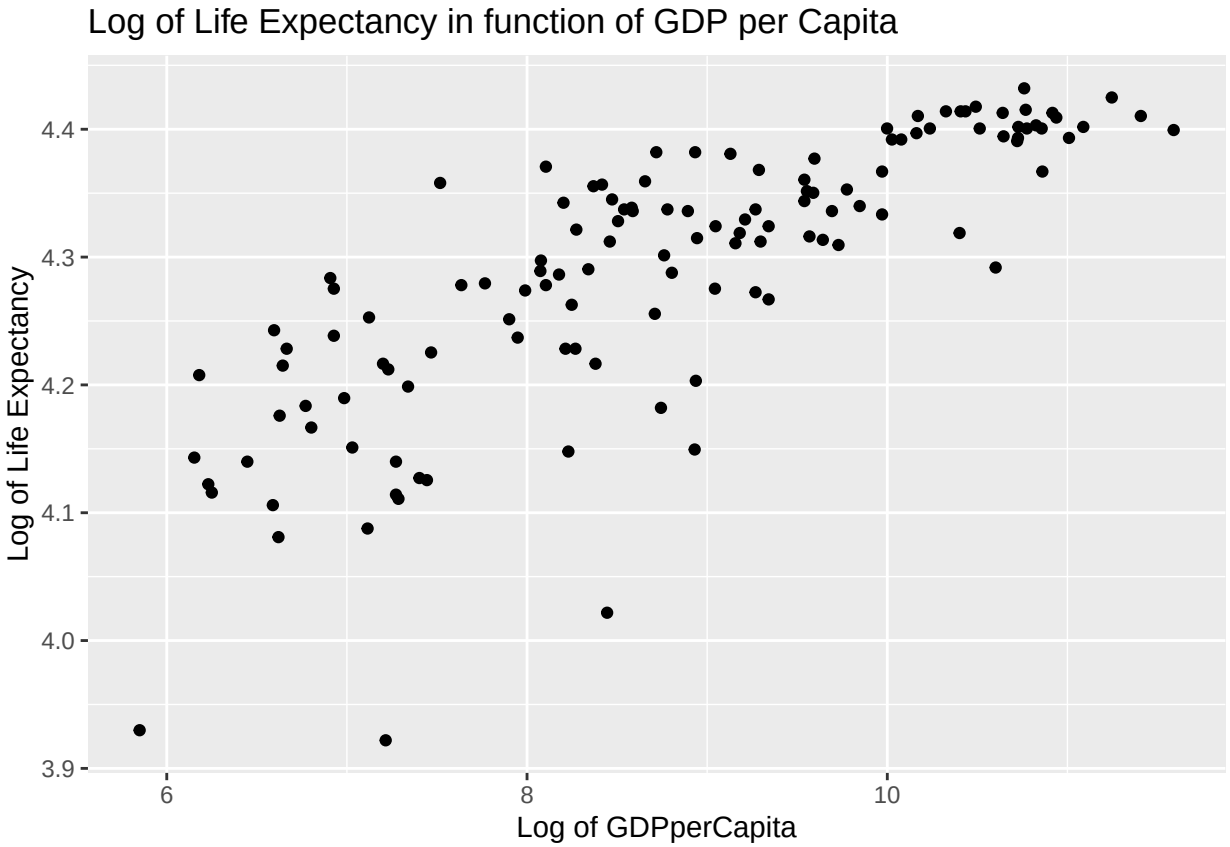


Transformations

GDPperCapita and HighTechExport bear some sort of curved relationship with Life Expectancy. We can visualize plots of the logged values of the predictor and explanatory variable.

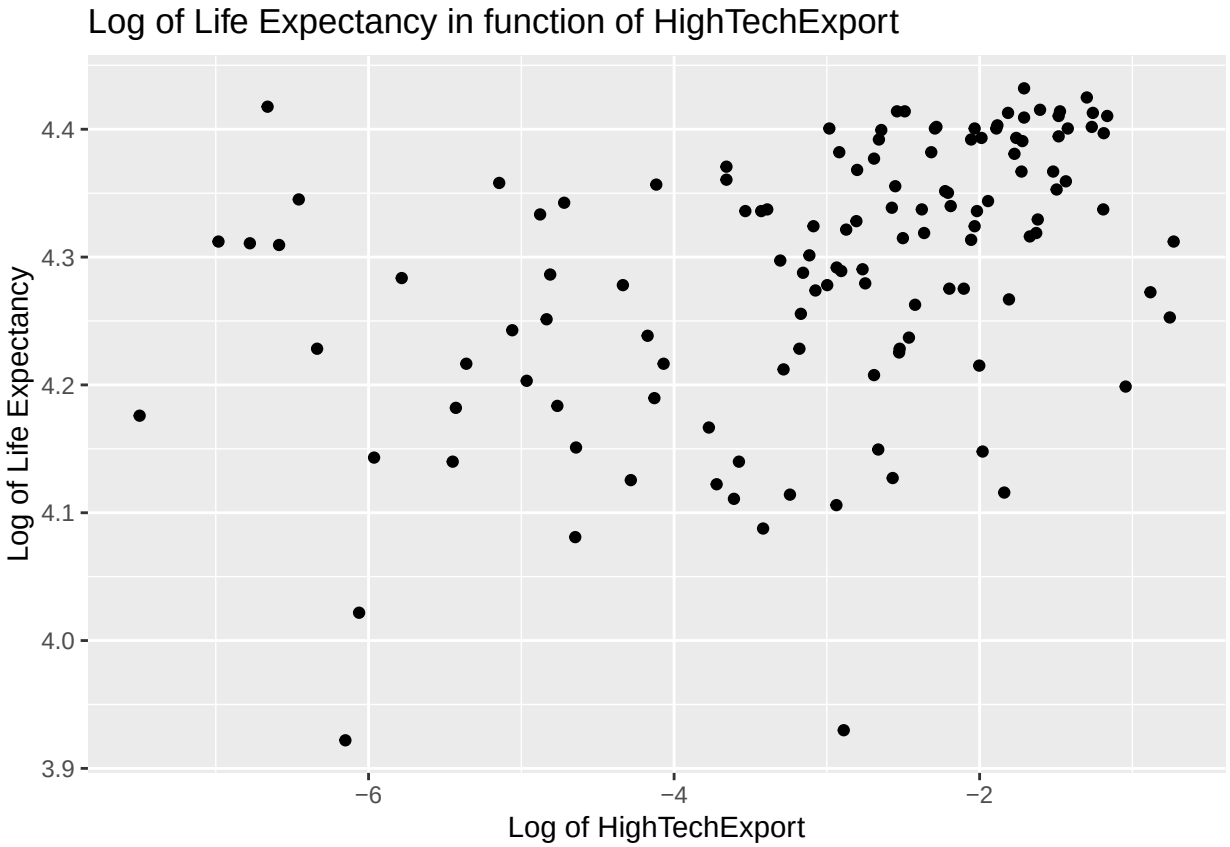
```
ggplot(data = Dataset1) + geom_point(mapping = aes(x = log(GDPperCapita), y = log(LifeExpectancy))) + g
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```




```
ggplot(data = Dataset1) + geom_point(mapping = aes(x = log(HighTechExport), y = log(LifeExpectancy))) +
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```

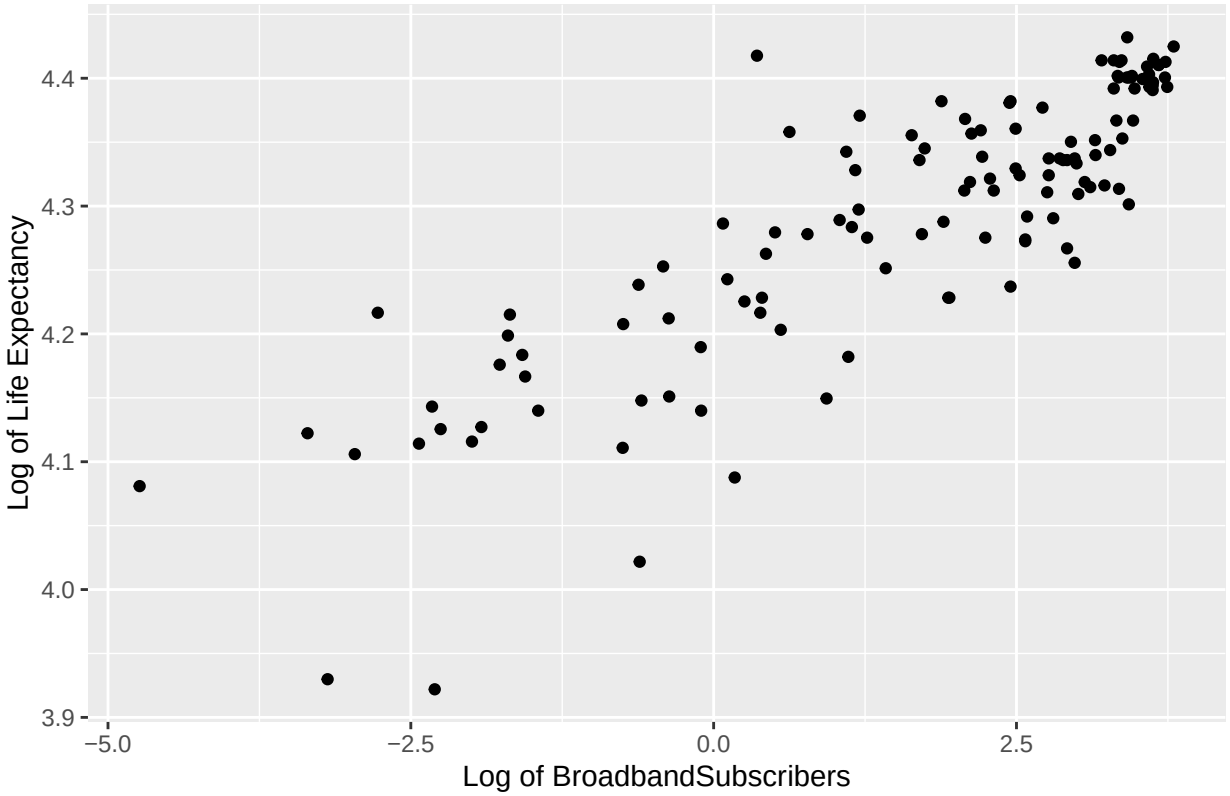


BroadbandSubscribers also seems to have some curved relationship with Life Expectancy. Lets also visualize plots of the logged values.

```
ggplot(data = Dataset1) + geom_point(mapping = aes(x = log(BroadbandSubscribers), y = log(LifeExpectanc
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```

Figure 4: Log of Life Expectancy in function of BroadbandSubscribers



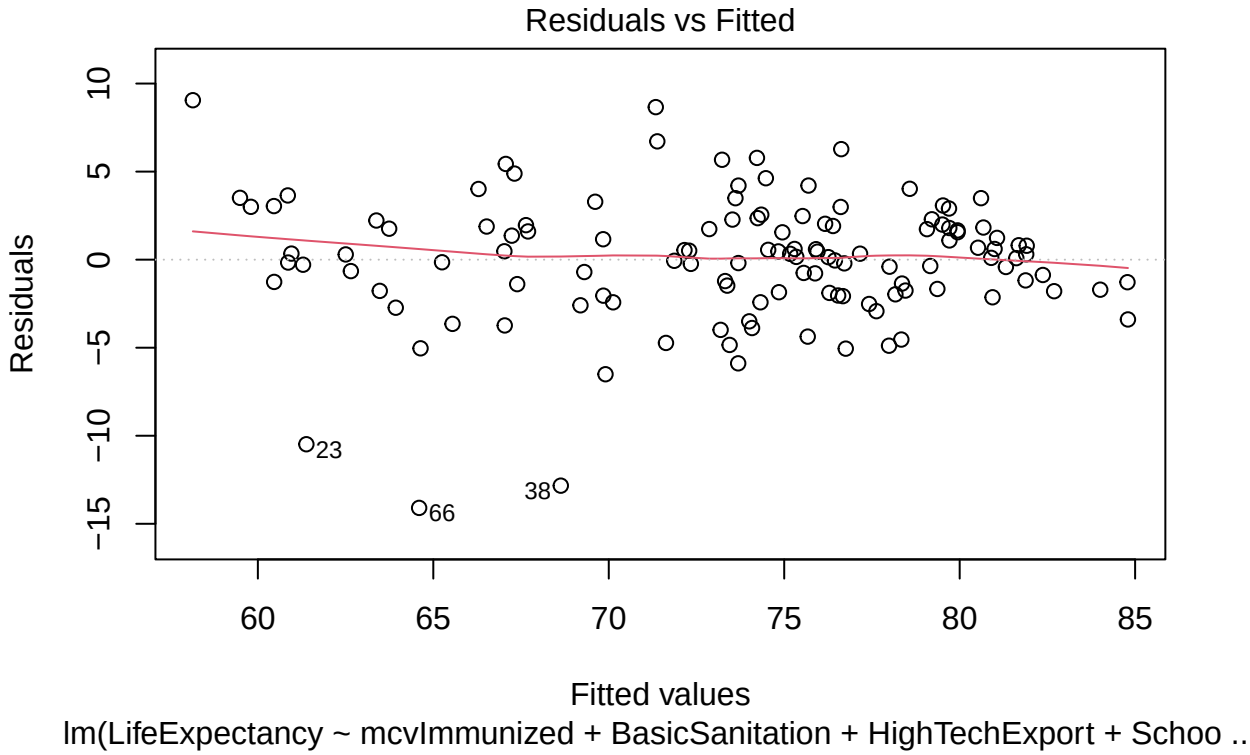
Model of logged of the other variables

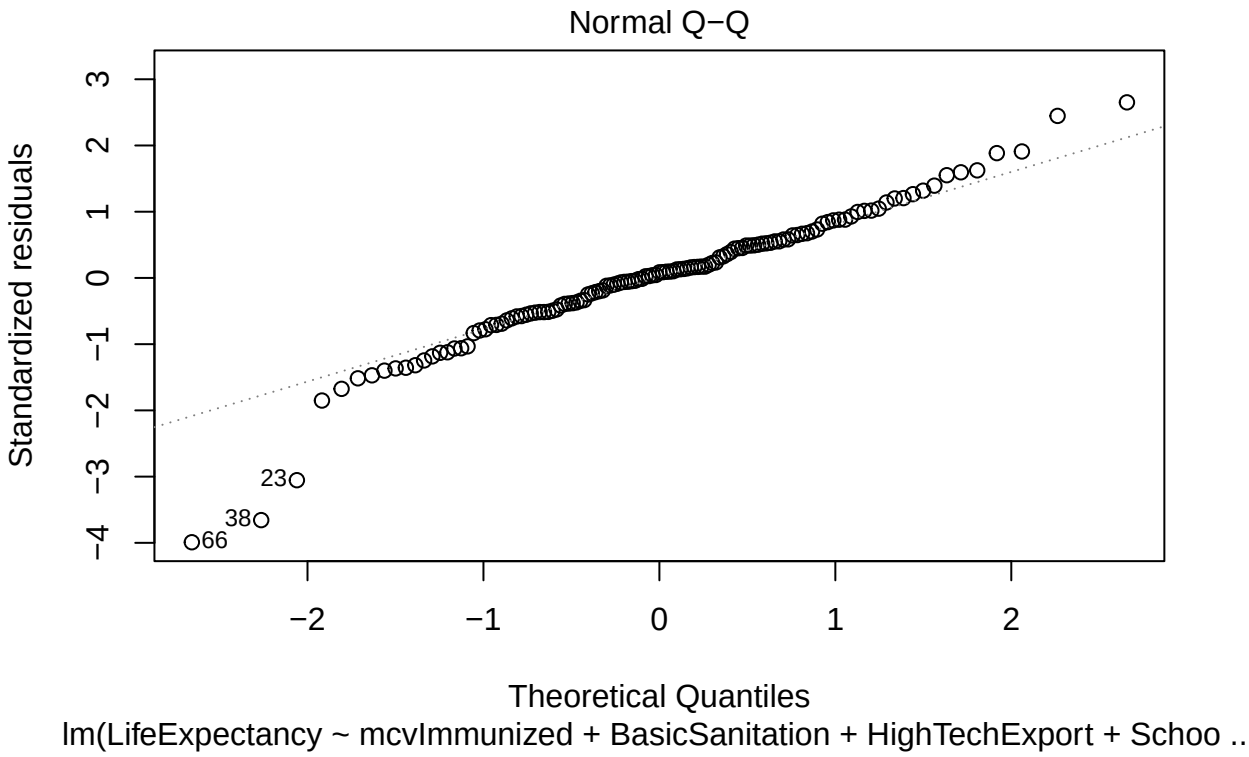
```
Model=lm(LifeExpectancy~ mcvImmunized + BasicSanitation + HighTechExport + SchoolYears + BroadbandSubsc
         + GDPperCapita + FoodSupply, data=Dataset1)
summary(Model)
```

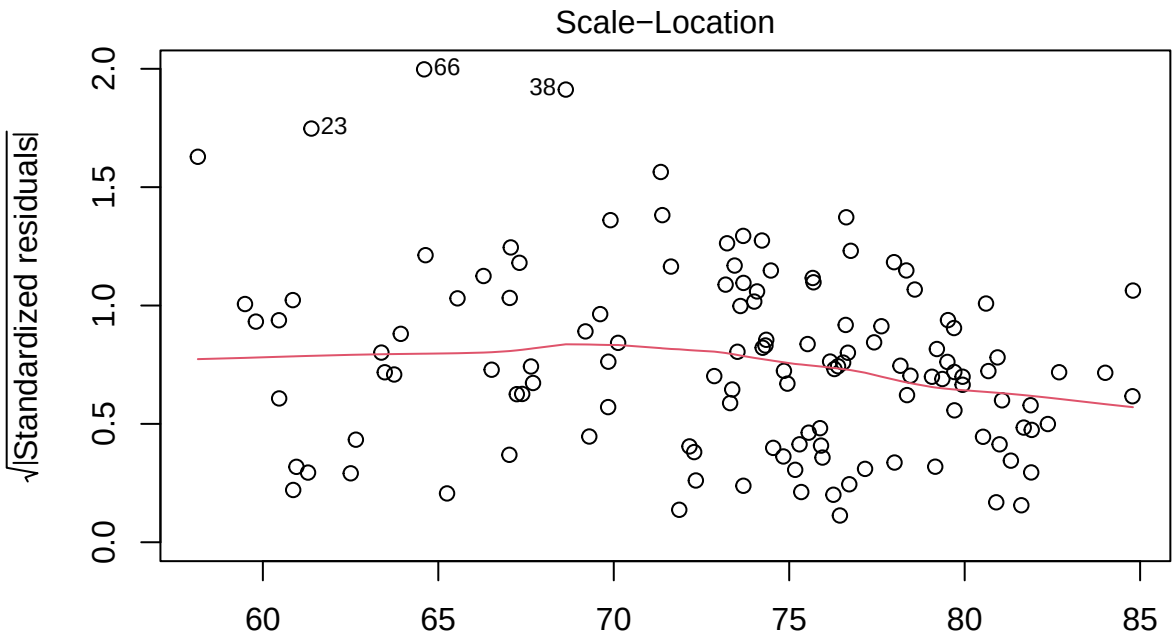
```
##
## Call:
## lm(formula = LifeExpectancy ~ mcvImmunized + BasicSanitation +
##     HighTechExport + SchoolYears + BroadbandSubscribers + GDPperCapita +
##     FoodSupply, data = Dataset1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.0964  -1.7829   0.2958   1.9345   9.0527
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.519e+01  4.307e+00  12.815  < 2e-16 ***
## mcvImmunized          6.499e+00  3.809e+00   1.706  0.09060 .
## BasicSanitation       1.623e+01  2.653e+00   6.118 1.25e-08 ***
## HighTechExport        1.320e+00  3.688e+00   0.358  0.72113
## SchoolYears          -2.211e-01  2.435e-01  -0.908  0.36571
## BroadbandSubscribers  1.328e-01  4.983e-02   2.664  0.00878 **
```

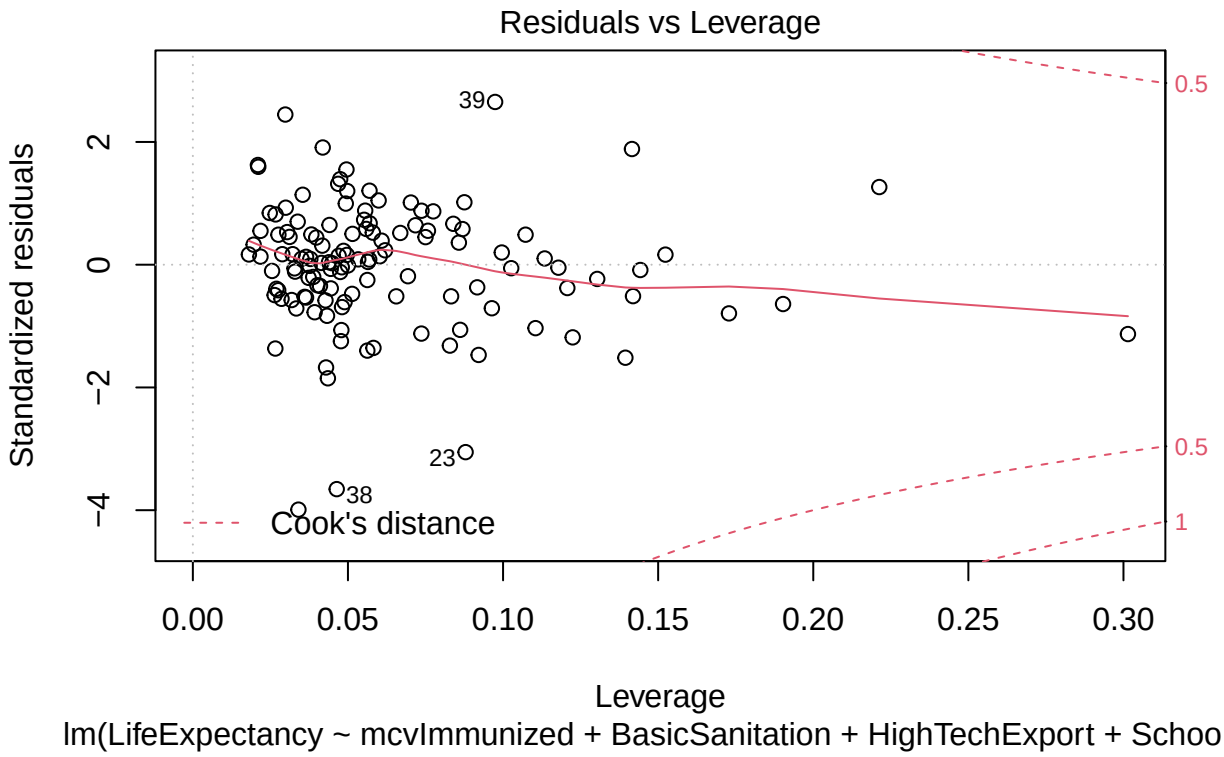
```
## GDPperCapita           6.010e-05  2.679e-05   2.243  0.02673 *
## FoodSupply            -2.792e-04  1.253e-03  -0.223  0.82404
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.594 on 119 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7744, Adjusted R-squared:  0.7611
## F-statistic: 58.34 on 7 and 119 DF,  p-value: < 2.2e-16
```

```
plot(Model)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(LifeExpectancy ~ mcvImmunized + BasicSanitation + HighTechExport + Schoo ..
23
38
66


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
66
38
23
60
65
70
75
80
85
Fitted values
lm(LifeExpectancy ~ mcvImmunized + BasicSanitation + HighTechExport + Schoo ..



Matrix Correlation of all Variables

```
a <- Dataset1 %>% select(LifeExpectancy, BasicSanitation, BroadbandSubscribers, AlcoholConsumption, mcv
cor <- round(cor(a), 2)
cor
```

```
##                       LifeExpectancy BasicSanitation BroadbandSubscribers
## LifeExpectancy                     1              NA                    NA
## BasicSanitation                   NA               1                    NA
## BroadbandSubscribers              NA              NA                     1
## AlcoholConsumption                NA              NA                    NA
## mcvImmunized                      NA              NA                    NA
## FoodSupply                        NA              NA                    NA
## SchoolYears                       NA              NA                    NA
## GDPperCapita                      NA              NA                    NA
## HighTechExport                    NA              NA                    NA
##                       AlcoholConsumption mcvImmunized FoodSupply SchoolYears
## LifeExpectancy                        NA           NA         NA          NA
## BasicSanitation                       NA           NA         NA          NA
## BroadbandSubscribers                  NA           NA         NA          NA
## AlcoholConsumption                     1           NA         NA          NA
## mcvImmunized                          NA            1         NA          NA
## FoodSupply                            NA           NA          1          NA
## SchoolYears                           NA           NA         NA           1
```

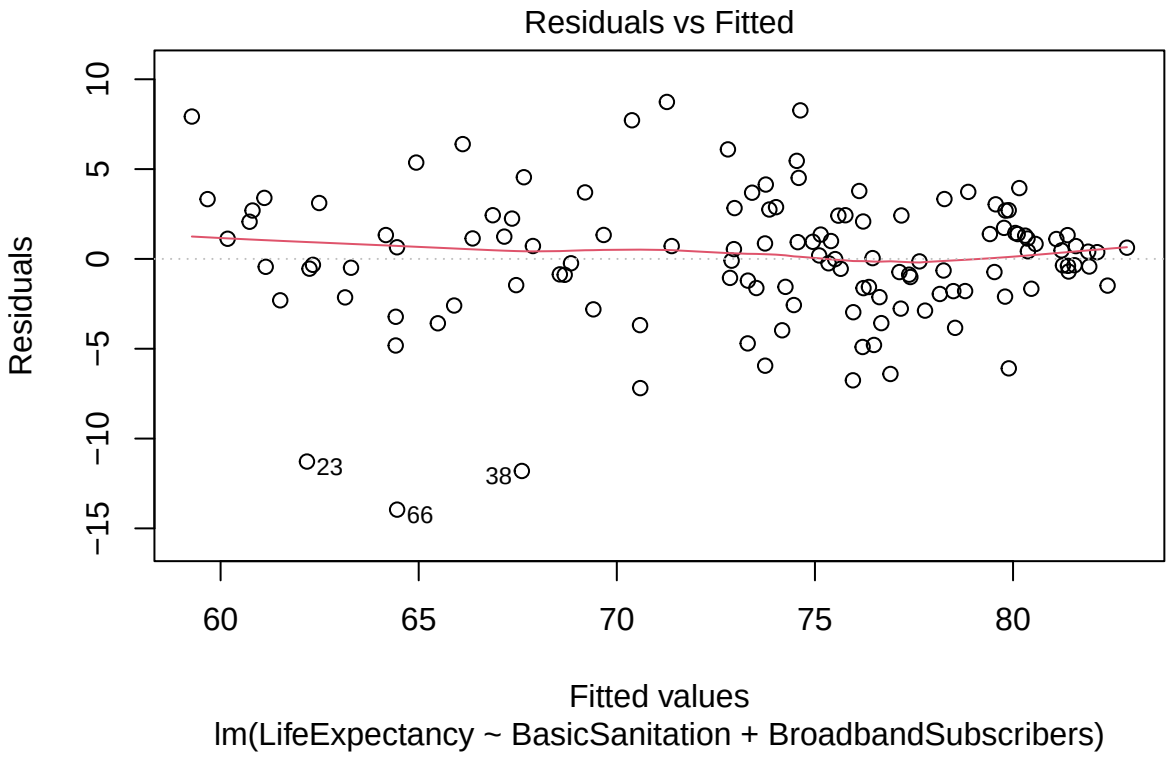
```
## GDPperCapita                          NA           NA         NA           NA
## HighTechExport                        NA           NA         NA           NA
##                       GDPperCapita HighTechExport
## LifeExpectancy                  NA             NA
## BasicSanitation                 NA             NA
## BroadbandSubscribers            NA             NA
## AlcoholConsumption              NA             NA
## mcvImmunized                    NA             NA
## FoodSupply                      NA             NA
## SchoolYears                     NA             NA
## GDPperCapita                     1             NA
## HighTechExport                  NA              1
```

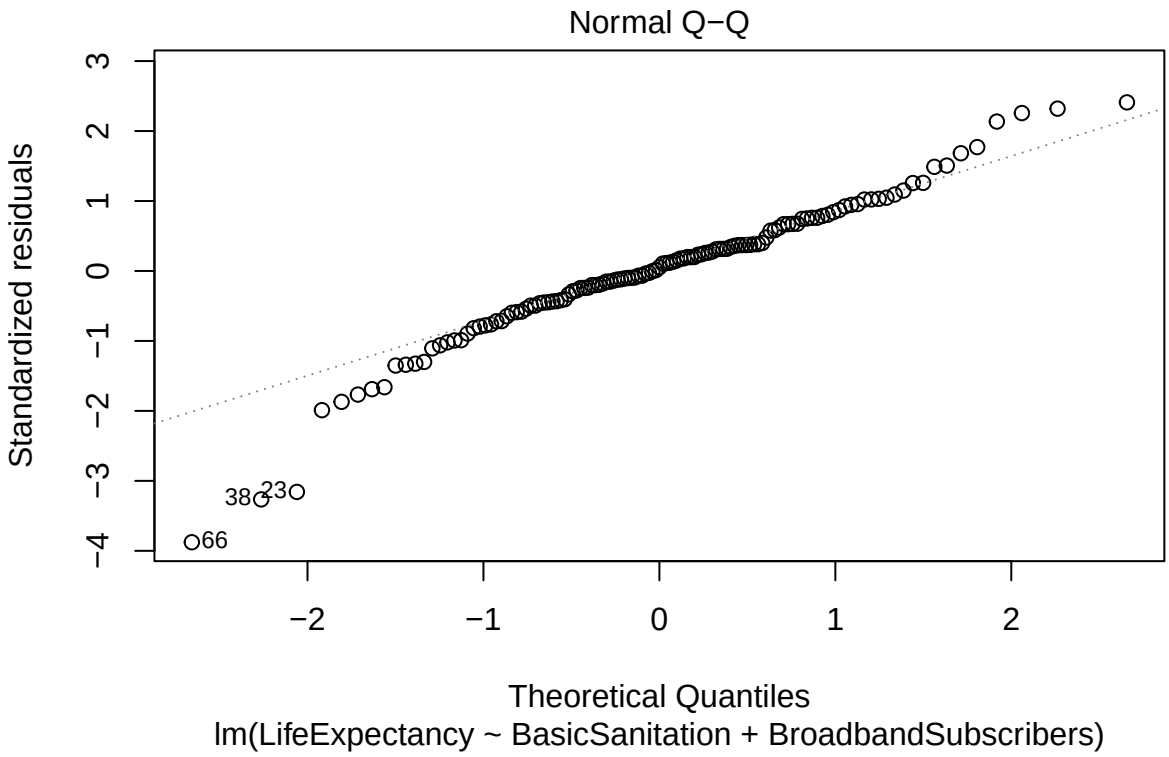
It can be seen that despite the curved relationship between BroadbandSubscribers and LifeExpectancy, the conditions are met and the R-squared is high

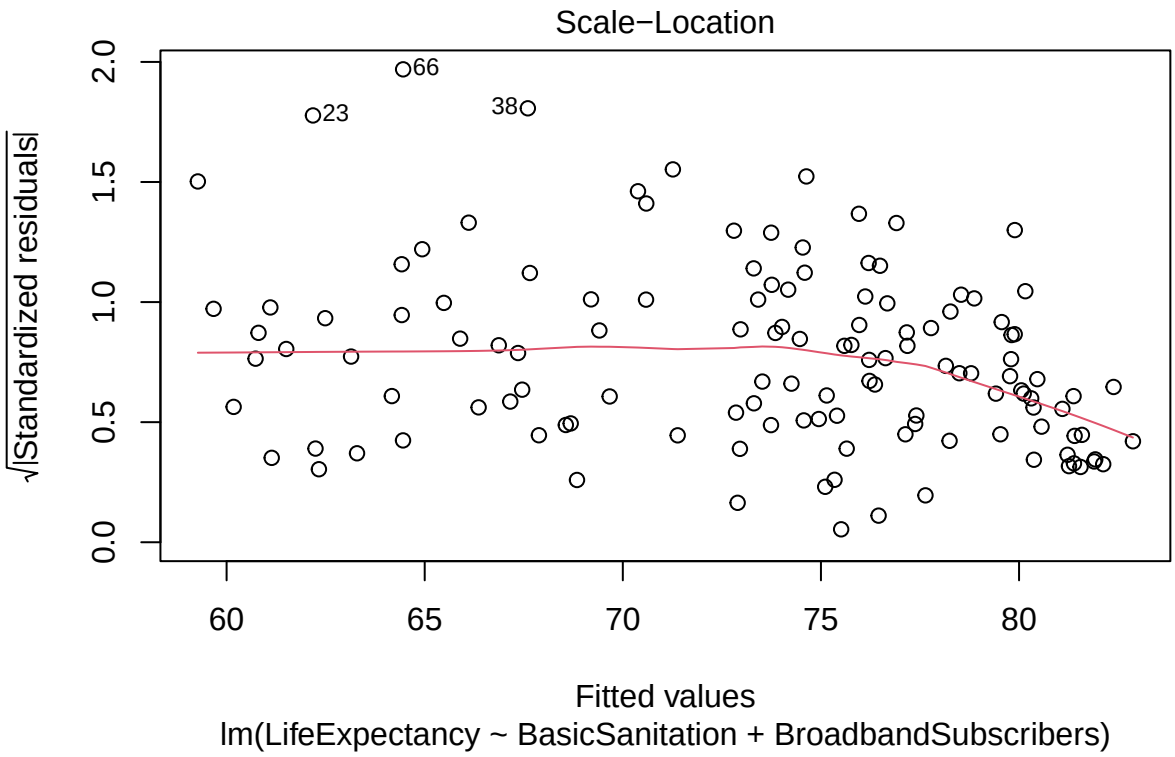
```
model1 <- lm(LifeExpectancy~BasicSanitation + BroadbandSubscribers, data=Dataset1)
summary(model1)
```

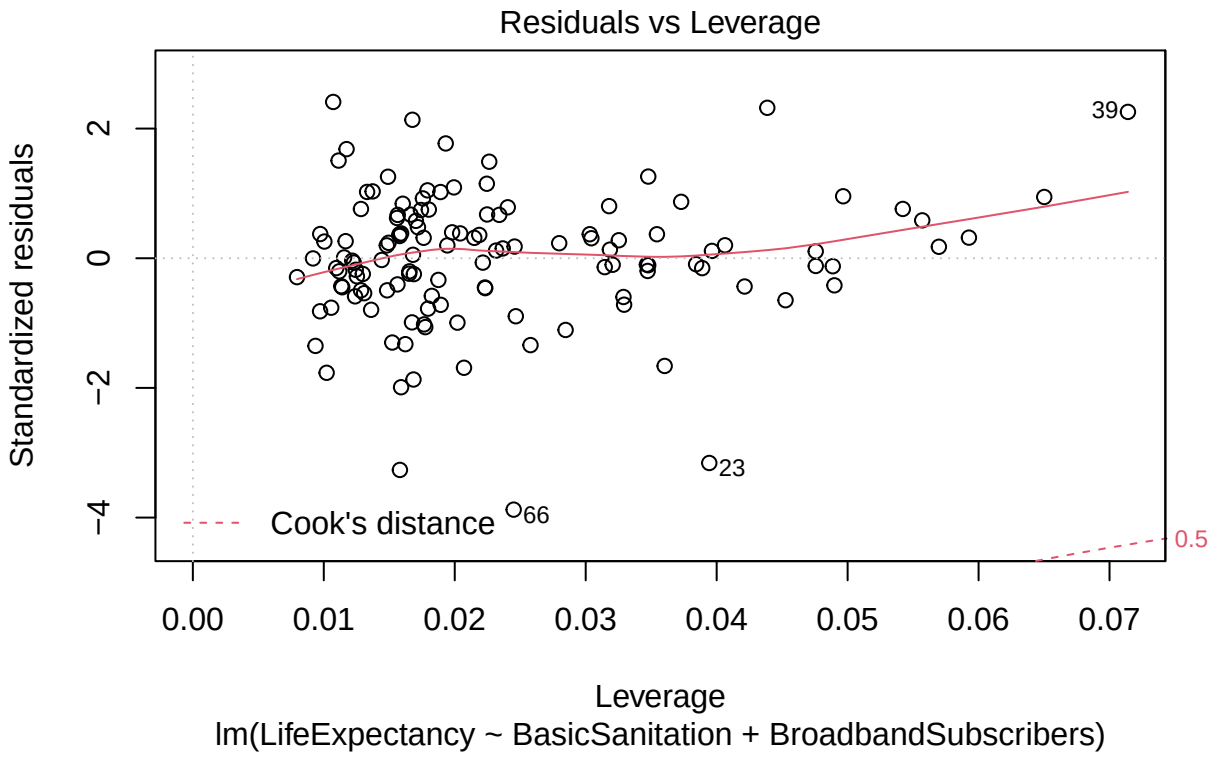
```
##
## Call:
## lm(formula = LifeExpectancy ~ BasicSanitation + BroadbandSubscribers,
##     data = Dataset1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.9556  -1.6436   0.1926   2.1637   8.7365
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          58.06597    1.06702   54.42  < 2e-16 ***
## BasicSanitation      16.29275    1.61317   10.10  < 2e-16 ***
## BroadbandSubscribers  0.19131    0.03248    5.89 3.39e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.644 on 124 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7582, Adjusted R-squared:  0.7543
## F-statistic: 194.4 on 2 and 124 DF,  p-value: < 2.2e-16
```

```
plot(model1)
```


Residuals vs Fitted
Residuals
10
5
0
−5
−10
−15
23
38
66
60
65
70
75
80
Fitted values
lm(LifeExpectancy ~ BasicSanitation + BroadbandSubscribers)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(LifeExpectancy ~ BasicSanitation + BroadbandSubscribers)
38
23
66


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
66
23
38
60
65
70
75
80
Fitted values
lm(LifeExpectancy ~ BasicSanitation + BroadbandSubscribers)



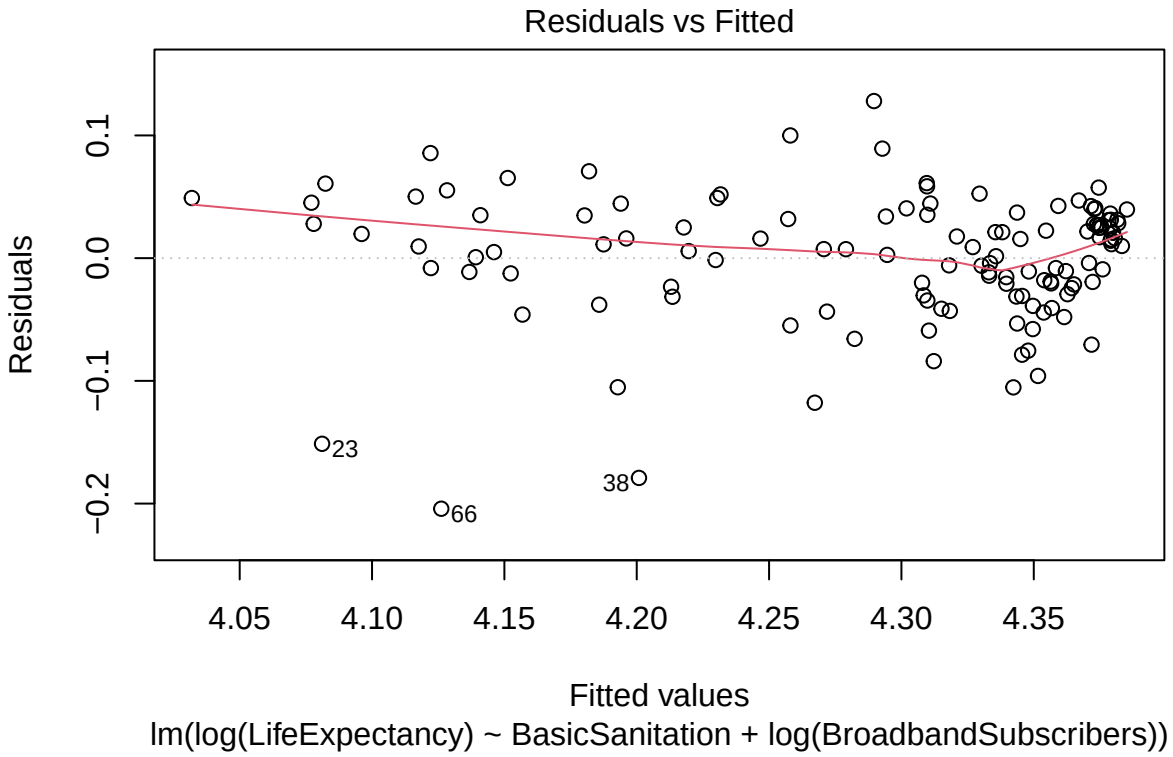
After logging the variables the result was worse than the original

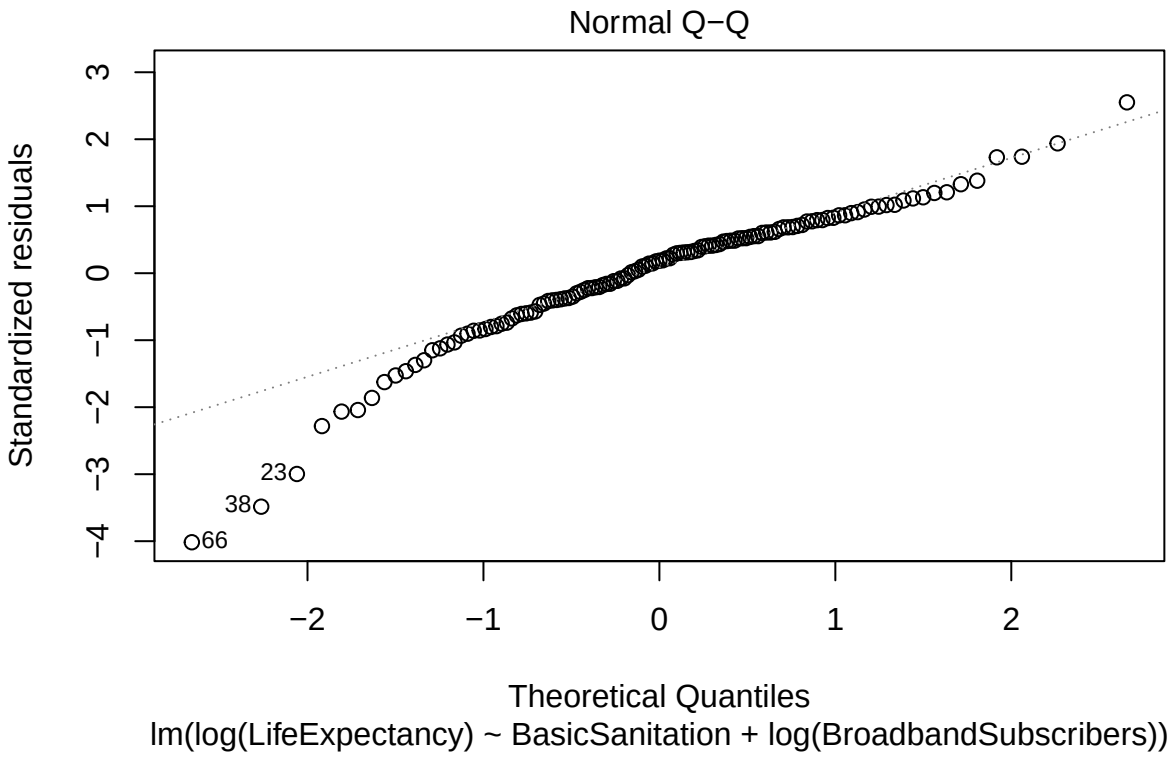
```
logmodel1 <- lm(log(LifeExpectancy)~BasicSanitation + log(BroadbandSubscribers), data=Dataset1)
summary(logmodel1)
```

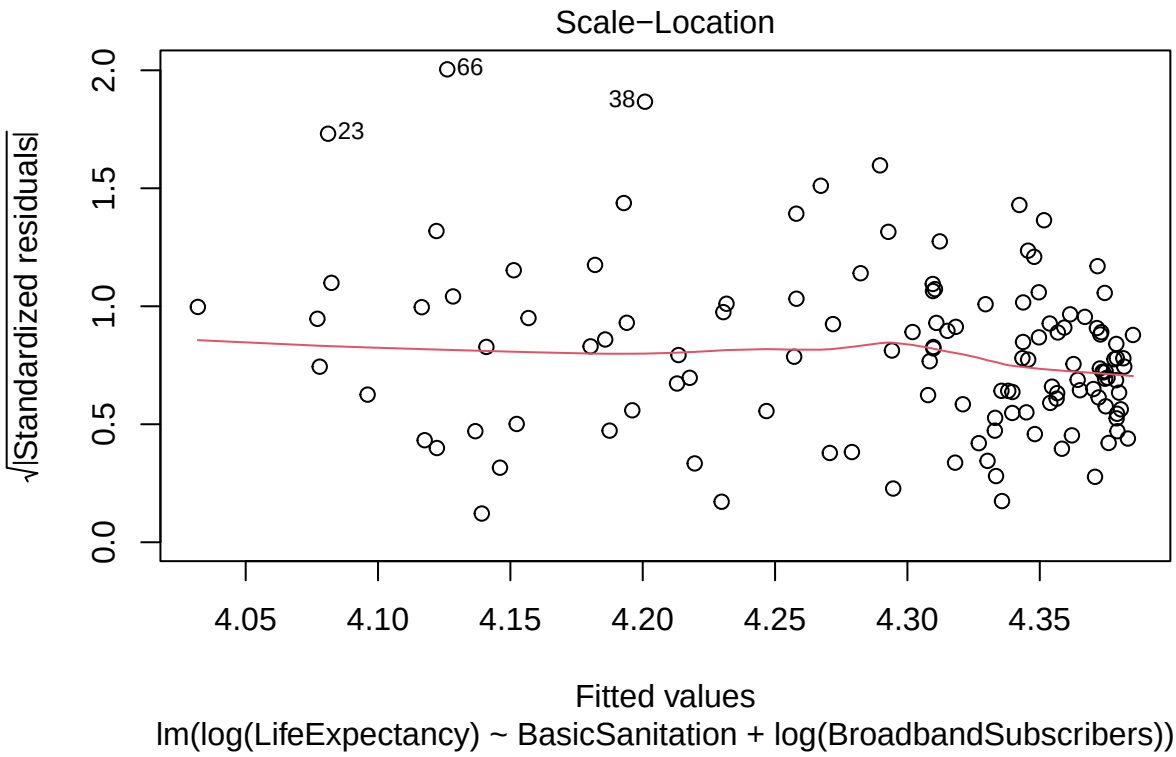
```
##
## Call:
## lm(formula = log(LifeExpectancy) ~ BasicSanitation + log(BroadbandSubscribers),
##     data = Dataset1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.204178 -0.023860  0.009509  0.032916  0.128001
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                4.132774   0.020765 199.029  < 2e-16 ***
## BasicSanitation            0.146909   0.033226   4.421 2.12e-05 ***
## log(BroadbandSubscribers) 0.027823   0.004481   6.209 7.36e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05186 on 124 degrees of freedom
##   (4 observations deleted due to missingness)
```

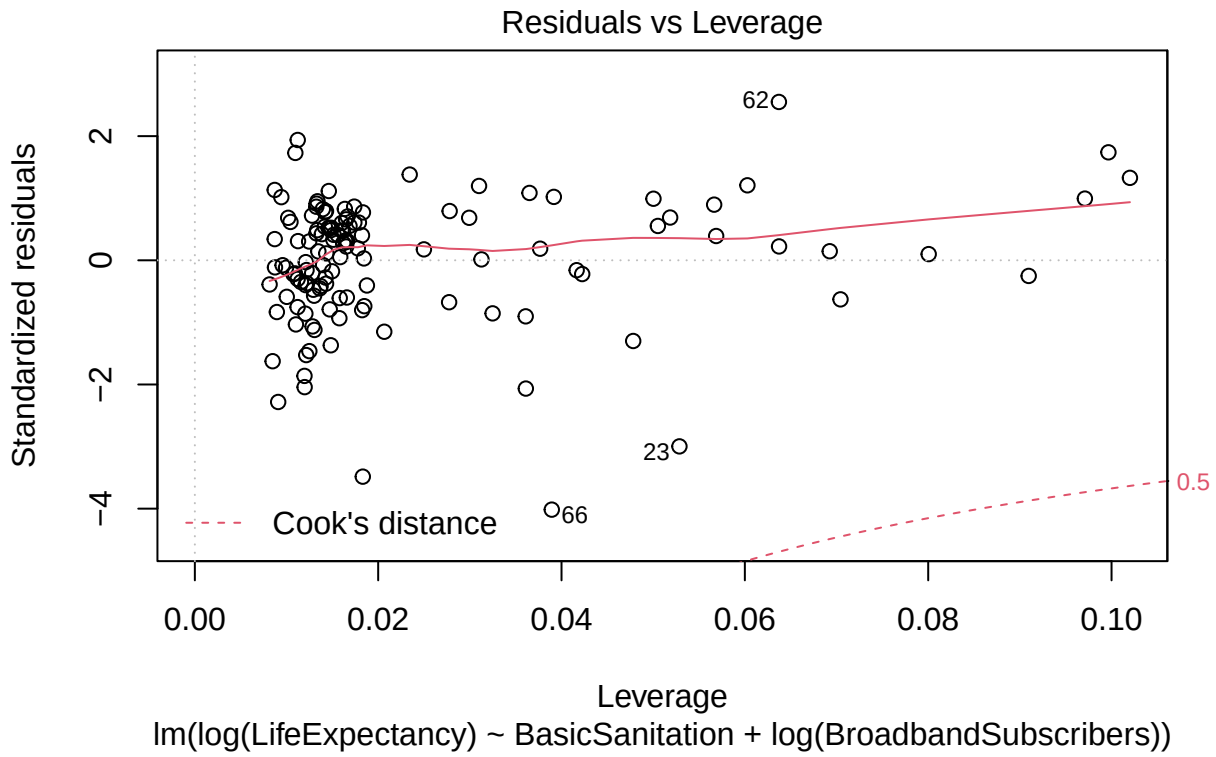
```
## Multiple R-squared:  0.7618, Adjusted R-squared:  0.758
## F-statistic: 198.3 on 2 and 124 DF,  p-value: < 2.2e-16
```

```
plot(logmodel1)
```




Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
−4
23
38
66
−2
−1
0
1
2
Theoretical Quantiles
lm(log(LifeExpectancy) ~ BasicSanitation + log(BroadbandSubscribers))


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
66
38
23
4.05
4.10
4.15
4.20
4.25
4.30
4.35
Fitted values
lm(log(LifeExpectancy) ~ BasicSanitation + log(BroadbandSubscribers))



From stepwise and the backward and forward methods, it is possible to see that the best model contains BasicSanitation, BroadbandSubscribers, GDPperCapita, mcvImmunized and AlcoholConsumption as predictors. Despite the lower mallow cp, it still has the same R-squared as the two variable model containing only BasicSanitation and BroadbandSubscribers

```
all <- lm(LifeExpectancy~mcvImmunized + BasicSanitation + HighTechExport + SchoolYears + BroadbandSubsc
        + GDPperCapita + FoodSupply + AlcoholConsumption, data=Dataset1)
MSE <- (summary(all)$sigma)^2

step(all, scale=MSE, direction="backward")
```

```
## Start:  AIC=9
## LifeExpectancy ~ mcvImmunized + BasicSanitation + HighTechExport +
##     SchoolYears + BroadbandSubscribers + GDPperCapita + FoodSupply +
##     AlcoholConsumption
##
##                        Df Sum of Sq    RSS      Cp
## - HighTechExport        1      2.49 1497.4  7.1964
## - FoodSupply            1      2.95 1497.8  7.2332
## - SchoolYears           1      9.01 1503.9  7.7112
## <none>                              1494.9  9.0000
## - AlcoholConsumption    1     42.09 1537.0 10.3221
## - mcvImmunized          1     43.45 1538.3 10.4296
## - GDPperCapita          1     47.71 1542.6 10.7664
```

```
## - BroadbandSubscribers  1    133.52 1628.4 17.5392
## - BasicSanitation       1    456.07 1951.0 43.0008
##
## Step:  AIC=7.2
## LifeExpectancy ~ mcvImmunized + BasicSanitation + SchoolYears +
##     BroadbandSubscribers + GDPperCapita + FoodSupply + AlcoholConsumption
##
##                         Df Sum of Sq    RSS      Cp
## - FoodSupply             1      3.23 1500.6  5.4514
## - SchoolYears            1      9.56 1506.9  5.9513
## <none>                               1497.4  7.1964
## - AlcoholConsumption     1     41.25 1538.6  8.4526
## - mcvImmunized           1     42.72 1540.1  8.5689
## - GDPperCapita           1     48.96 1546.3  9.0613
## - BroadbandSubscribers   1    147.76 1645.1 16.8598
## - BasicSanitation        1    462.72 1960.1 41.7216
##
## Step:  AIC=5.45
## LifeExpectancy ~ mcvImmunized + BasicSanitation + SchoolYears +
##     BroadbandSubscribers + GDPperCapita + AlcoholConsumption
##
##                         Df Sum of Sq    RSS      Cp
## - SchoolYears            1      9.03 1509.6  4.1641
## <none>                               1500.6  5.4514
## - AlcoholConsumption     1     38.78 1539.4  6.5128
## - mcvImmunized           1     44.41 1545.0  6.9570
## - GDPperCapita           1     46.29 1546.9  7.1057
## - BroadbandSubscribers   1    146.83 1647.4 15.0412
## - BasicSanitation        1    531.26 2031.9 45.3872
##
## Step:  AIC=4.16
## LifeExpectancy ~ mcvImmunized + BasicSanitation + BroadbandSubscribers +
##     GDPperCapita + AlcoholConsumption
##
##                         Df Sum of Sq    RSS      Cp
## <none>                               1509.6  4.1641
## - AlcoholConsumption     1     40.59 1550.2  5.3680
## - GDPperCapita           1     41.63 1551.3  5.4504
## - mcvImmunized           1     45.71 1555.3  5.7723
## - BroadbandSubscribers   1    139.24 1648.9 13.1550
## - BasicSanitation        1    700.53 2210.2 57.4612

##
## Call:
## lm(formula = LifeExpectancy ~ mcvImmunized + BasicSanitation +
##     BroadbandSubscribers + GDPperCapita + AlcoholConsumption,
##     data = Dataset1)
##
## Coefficients:
##          (Intercept)          mcvImmunized       BasicSanitation
##            5.406e+01             7.162e+00             1.413e+01
## BroadbandSubscribers          GDPperCapita    AlcoholConsumption
##            1.789e-01             4.767e-05            -1.882e-01
```

```r
none <- lm(LifeExpectancy~1,data=Dataset1)
step(none, scope=list(upper=all), scale=MSE, direction="forward")
```

```
## Start:  AIC=412.66
## LifeExpectancy ~ 1
##
##                         Df Sum of Sq    RSS      Cp
## + BasicSanitation        1    4703.8 2107.5  43.361
## + BroadbandSubscribers   1    3809.9 3001.4 113.921
## + SchoolYears            1    3541.3 3270.1 135.127
## + FoodSupply             1    3409.5 3401.9 145.530
## + GDPperCapita           1    2738.3 4073.1 198.515
## + mcvImmunized           1    2064.7 4746.6 251.681
## + HighTechExport         1    1009.3 5802.1 334.994
## + AlcoholConsumption     1     906.5 5904.9 343.110
## <none>                               6811.4 412.664
##
## Step:  AIC=43.36
## LifeExpectancy ~ BasicSanitation
##
##                         Df Sum of Sq    RSS      Cp
## + BroadbandSubscribers   1    460.79 1646.8  8.9883
## + GDPperCapita           1    424.37 1683.2 11.8636
## + HighTechExport         1     95.53 2012.0 37.8203
## + FoodSupply             1     94.80 2012.7 37.8781
## + AlcoholConsumption     1     45.59 2061.9 41.7624
## + SchoolYears            1     39.67 2067.9 42.2303
## <none>                               2107.5 43.3614
## + mcvImmunized           1     17.41 2090.1 43.9869
##
## Step:  AIC=8.99
## LifeExpectancy ~ BasicSanitation + BroadbandSubscribers
##
##                       Df Sum of Sq    RSS      Cp
## + GDPperCapita         1    57.271 1589.5  6.4676
## + AlcoholConsumption   1    50.822 1595.9  6.9766
## + mcvImmunized         1    36.696 1610.0  8.0916
## <none>                             1646.8  8.9883
## + SchoolYears          1     5.917 1640.8 10.5212
## + HighTechExport       1     2.352 1644.4 10.8027
## + FoodSupply           1     0.036 1646.7 10.9854
##
## Step:  AIC=6.47
## LifeExpectancy ~ BasicSanitation + BroadbandSubscribers + GDPperCapita
##
##                       Df Sum of Sq    RSS     Cp
## + mcvImmunized         1    39.266 1550.2 5.3680
## + AlcoholConsumption   1    34.145 1555.3 5.7723
## <none>                             1589.5 6.4676
## + SchoolYears          1    11.990 1577.5 7.5211
## + HighTechExport       1     1.718 1587.8 8.3320
## + FoodSupply           1     1.305 1588.2 8.3646
##
```

```
## Step:  AIC=5.37
## LifeExpectancy ~ BasicSanitation + BroadbandSubscribers + GDPperCapita +
##     mcvImmunized
##
##                       Df Sum of Sq    RSS     Cp
## + AlcoholConsumption  1    40.589 1509.6 4.1641
## <none>                             1550.2 5.3680
## + SchoolYears         1    10.834 1539.4 6.5128
## + HighTechExport      1     2.234 1548.0 7.1917
## + FoodSupply          1     0.468 1549.8 7.3311
##
## Step:  AIC=4.16
## LifeExpectancy ~ BasicSanitation + BroadbandSubscribers + GDPperCapita +
##     mcvImmunized + AlcoholConsumption
##
##                   Df Sum of Sq    RSS     Cp
## <none>                         1509.6 4.1641
## + SchoolYears     1    9.0289 1500.6 5.4514
## + HighTechExport  1    3.3017 1506.3 5.9034
## + FoodSupply      1    2.6956 1506.9 5.9513


##
## Call:
## lm(formula = LifeExpectancy ~ BasicSanitation + BroadbandSubscribers +
##     GDPperCapita + mcvImmunized + AlcoholConsumption, data = Dataset1)
##
## Coefficients:
##          (Intercept)       BasicSanitation  BroadbandSubscribers
##            5.406e+01             1.413e+01             1.789e-01
##         GDPperCapita          mcvImmunized    AlcoholConsumption
##            4.767e-05             7.162e+00            -1.882e-01
```

```
step(none, scope=list(upper=all), scale=MSE)
```

```
## Start:  AIC=412.66
## LifeExpectancy ~ 1
##
##                         Df Sum of Sq    RSS      Cp
## + BasicSanitation       1    4703.8 2107.5  43.361
## + BroadbandSubscribers  1    3809.9 3001.4 113.921
## + SchoolYears           1    3541.3 3270.1 135.127
## + FoodSupply            1    3409.5 3401.9 145.530
## + GDPperCapita          1    2738.3 4073.1 198.515
## + mcvImmunized          1    2064.7 4746.6 251.681
## + HighTechExport        1    1009.3 5802.1 334.994
## + AlcoholConsumption    1     906.5 5904.9 343.110
## <none>                              6811.4 412.664
##
## Step:  AIC=43.36
## LifeExpectancy ~ BasicSanitation
##
##                         Df Sum of Sq    RSS       Cp
## + BroadbandSubscribers  1     460.8 1646.8   8.9883
```

```
## + GDPperCapita          1     424.4 1683.2  11.8636
## + HighTechExport        1      95.5 2012.0  37.8203
## + FoodSupply            1      94.8 2012.7  37.8781
## + AlcoholConsumption    1      45.6 2061.9  41.7624
## + SchoolYears           1      39.7 2067.9  42.2303
## <none>                              2107.5  43.3614
## + mcvImmunized          1      17.4 2090.1  43.9869
## - BasicSanitation       1    4703.8 6811.4 412.6643
##
## Step:  AIC=8.99
## LifeExpectancy ~ BasicSanitation + BroadbandSubscribers
##
##                         Df Sum of Sq    RSS       Cp
## + GDPperCapita          1     57.27 1589.5   6.4676
## + AlcoholConsumption    1     50.82 1595.9   6.9766
## + mcvImmunized          1     36.70 1610.0   8.0916
## <none>                              1646.8   8.9883
## + SchoolYears           1      5.92 1640.8  10.5212
## + HighTechExport        1      2.35 1644.4  10.8027
## + FoodSupply            1      0.04 1646.7  10.9854
## - BroadbandSubscribers  1    460.79 2107.5  43.3614
## - BasicSanitation       1   1354.68 3001.4 113.9213
##
## Step:  AIC=6.47
## LifeExpectancy ~ BasicSanitation + BroadbandSubscribers + GDPperCapita
##
##                         Df Sum of Sq    RSS       Cp
## + mcvImmunized          1     39.27 1550.2   5.3680
## + AlcoholConsumption    1     34.15 1555.3   5.7723
## <none>                              1589.5   6.4676
## + SchoolYears           1     11.99 1577.5   7.5211
## + HighTechExport        1      1.72 1587.8   8.3320
## + FoodSupply            1      1.30 1588.2   8.3646
## - GDPperCapita          1     57.27 1646.8   8.9883
## - BroadbandSubscribers  1     93.70 1683.2  11.8636
## - BasicSanitation       1   1382.41 2971.9 113.5898
##
## Step:  AIC=5.37
## LifeExpectancy ~ BasicSanitation + BroadbandSubscribers + GDPperCapita +
##     mcvImmunized
##
##                         Df Sum of Sq    RSS      Cp
## + AlcoholConsumption    1     40.59 1509.6  4.1641
## <none>                              1550.2  5.3680
## - mcvImmunized          1     39.27 1589.5  6.4676
## + SchoolYears           1     10.83 1539.4  6.5128
## + HighTechExport        1      2.23 1548.0  7.1917
## + FoodSupply            1      0.47 1549.8  7.3311
## - GDPperCapita          1     59.84 1610.0  8.0916
## - BroadbandSubscribers  1     98.66 1648.9 11.1555
## - BasicSanitation       1    782.12 2332.3 65.1055
##
## Step:  AIC=4.16
## LifeExpectancy ~ BasicSanitation + BroadbandSubscribers + GDPperCapita +
```

```
##     mcvImmunized + AlcoholConsumption
##
##                         Df Sum of Sq    RSS      Cp
## <none>                               1509.6  4.1641
## - AlcoholConsumption     1     40.59 1550.2  5.3680
## - GDPperCapita           1     41.63 1551.3  5.4504
## + SchoolYears            1      9.03 1500.6  5.4514
## - mcvImmunized           1     45.71 1555.3  5.7723
## + HighTechExport         1      3.30 1506.3  5.9034
## + FoodSupply             1      2.70 1506.9  5.9513
## - BroadbandSubscribers   1    139.24 1648.9 13.1550
## - BasicSanitation        1    700.53 2210.2 57.4612

##
## Call:
## lm(formula = LifeExpectancy ~ BasicSanitation + BroadbandSubscribers +
##     GDPperCapita + mcvImmunized + AlcoholConsumption, data = Dataset1)
##
## Coefficients:
##          (Intercept)       BasicSanitation  BroadbandSubscribers
##            5.406e+01             1.413e+01             1.789e-01
##         GDPperCapita          mcvImmunized    AlcoholConsumption
##            4.767e-05             7.162e+00            -1.882e-01
```

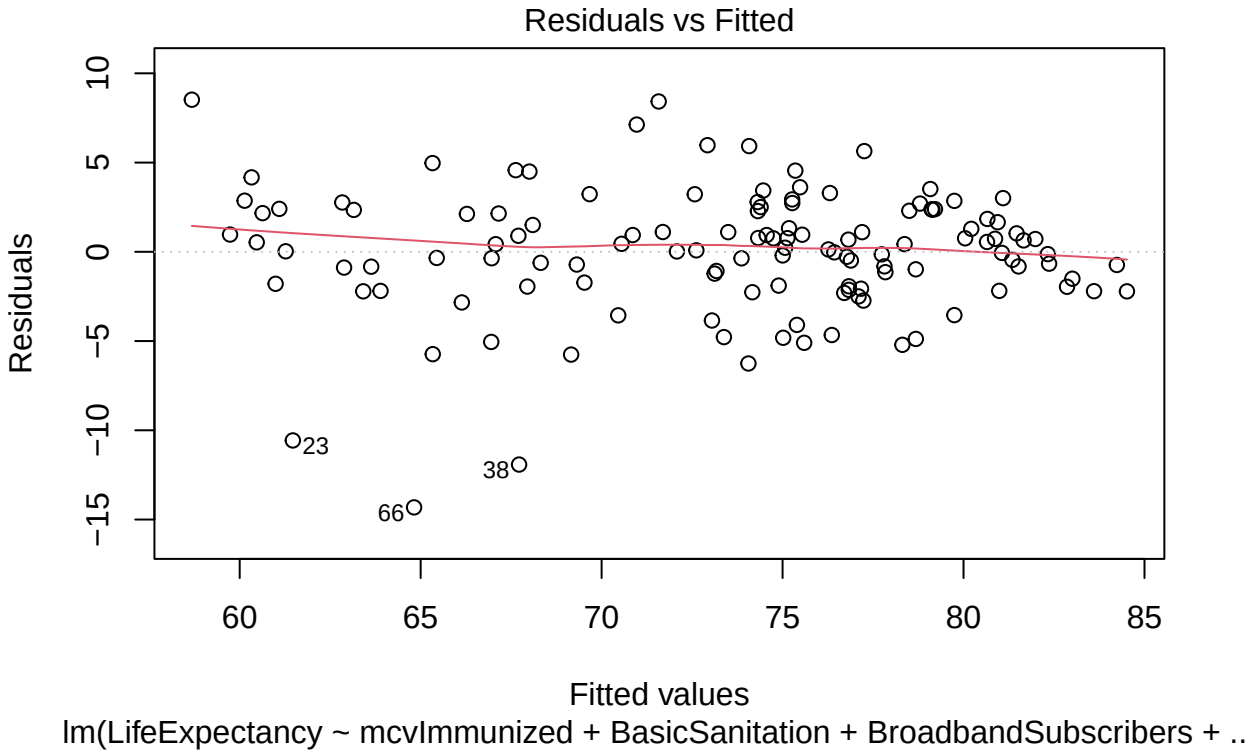
To balance R-squared, complexity and Mallow Cp, I am going to try models combining GDPperCapita, mcvImmunized and AlcoholConsumption added to the two variable model to see which combination is best.

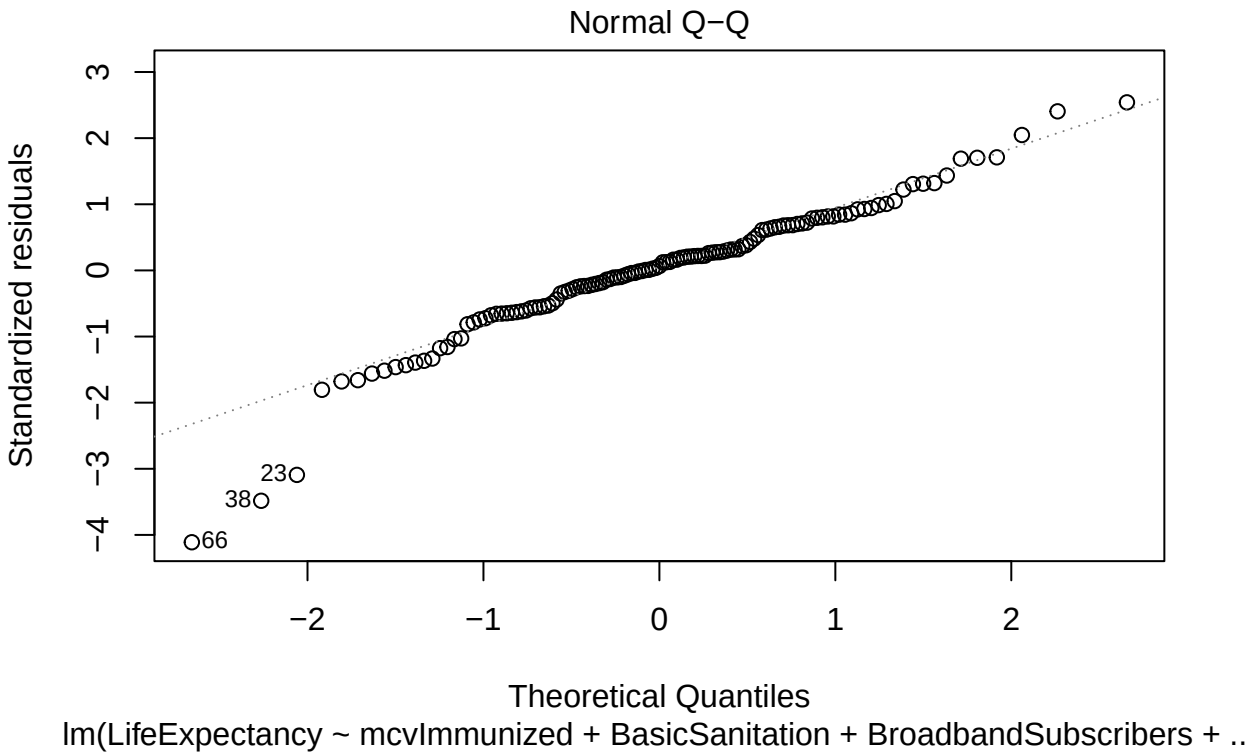
```
stepmodel <- lm(LifeExpectancy~mcvImmunized + BasicSanitation + BroadbandSubscribers
         + GDPperCapita + AlcoholConsumption, data=Dataset1)
summary(stepmodel)
```

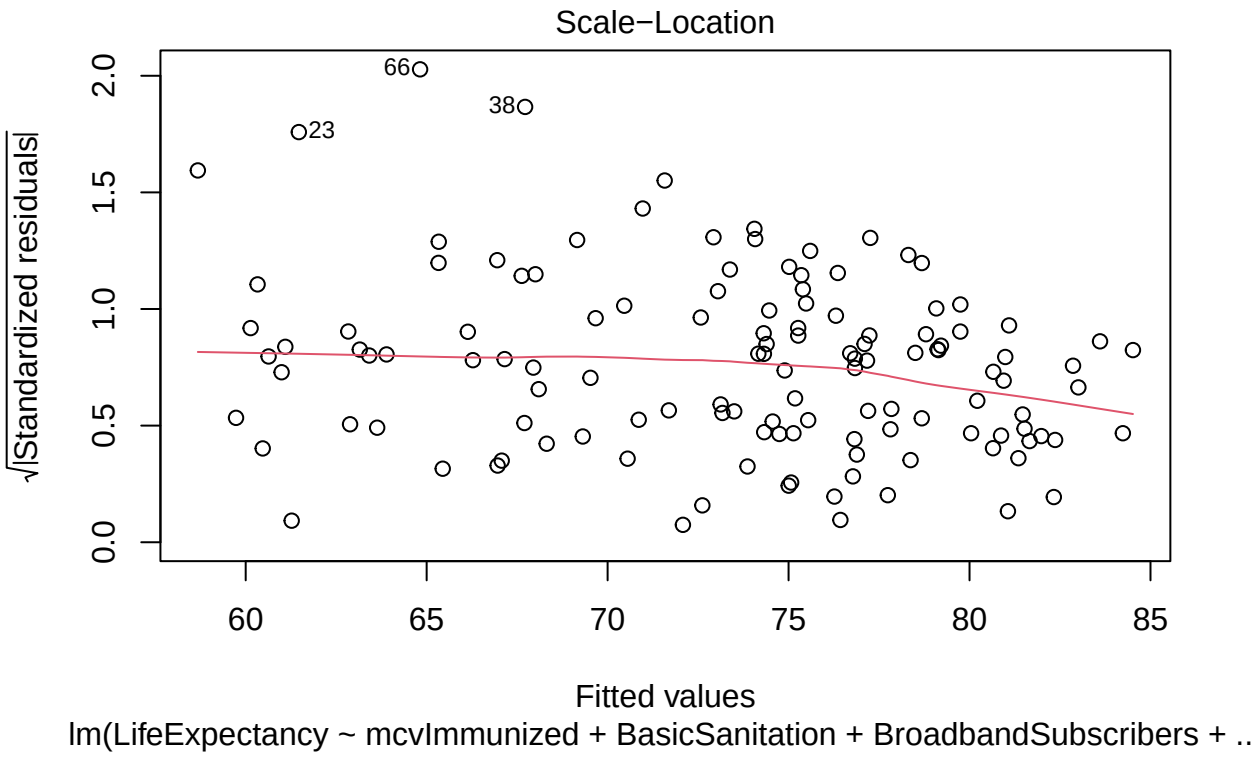
```
##
## Call:
## lm(formula = LifeExpectancy ~ mcvImmunized + BasicSanitation +
##     BroadbandSubscribers + GDPperCapita + AlcoholConsumption,
##     data = Dataset1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3188  -1.9129   0.2291   2.2899   8.5220
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.406e+01  2.818e+00  19.188  < 2e-16 ***
## mcvImmunized          7.162e+00  3.742e+00   1.914  0.05797 .
## BasicSanitation       1.413e+01  1.886e+00   7.493 1.22e-11 ***
## BroadbandSubscribers  1.789e-01  5.356e-02   3.341  0.00111 **
## GDPperCapita          4.767e-05  2.609e-05   1.827  0.07021 .
## AlcoholConsumption   -1.882e-01  1.044e-01  -1.804  0.07377 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

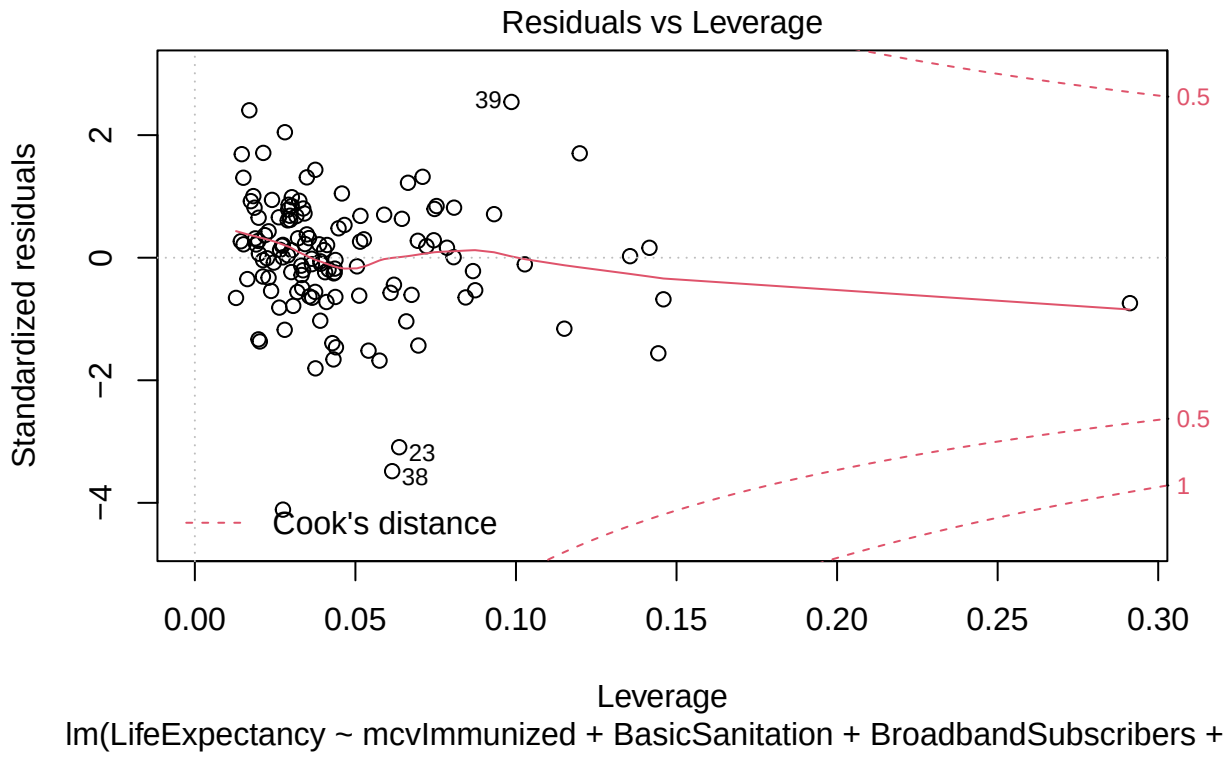
```
##
## Residual standard error: 3.532 on 121 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7692
## F-statistic: 84.99 on 5 and 121 DF,  p-value: < 2.2e-16
```

```
plot(stepmodel)
```




Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
−4
23
38
66
−2
−1
0
1
2
Theoretical Quantiles
lm(LifeExpectancy ~ mcvImmunized + BasicSanitation + BroadbandSubscribers + ..


Scale−Location
66
38
23
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
60
65
70
75
80
85
Fitted values
lm(LifeExpectancy ~ mcvImmunized + BasicSanitation + BroadbandSubscribers + ..




```
extractAIC(stepmodel, scale=MSE)
```

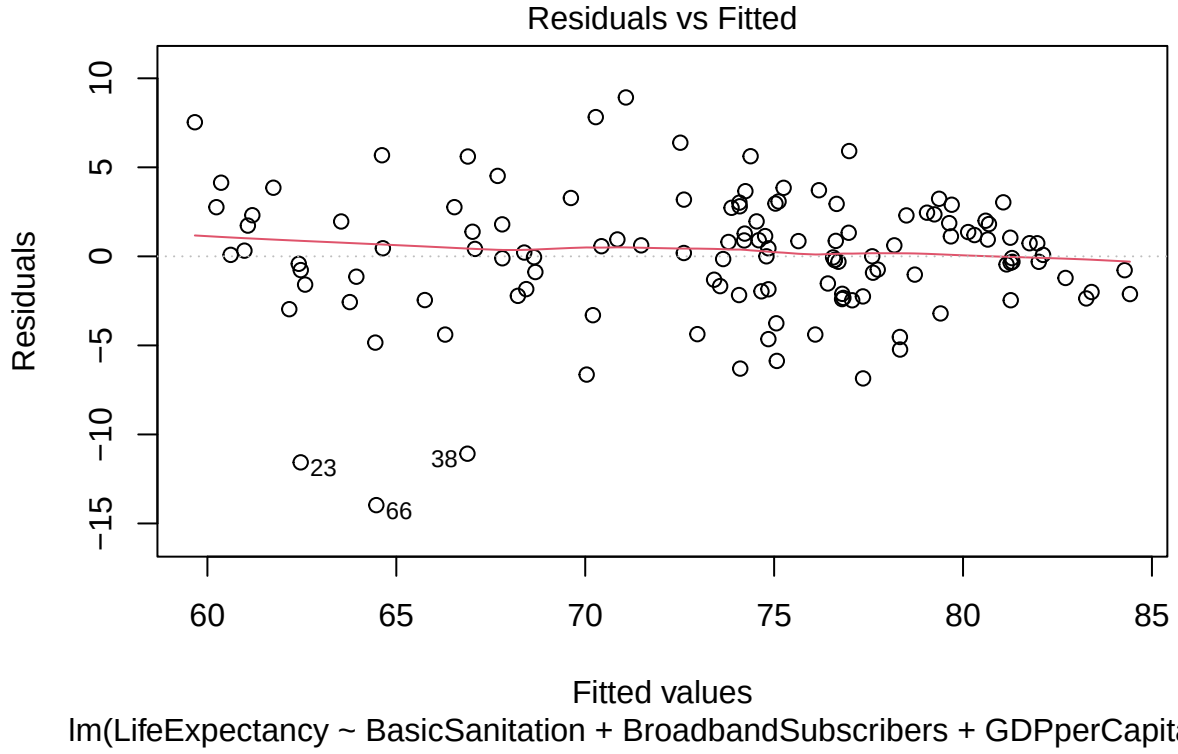
```
## [1] 6.000000 4.164073
```

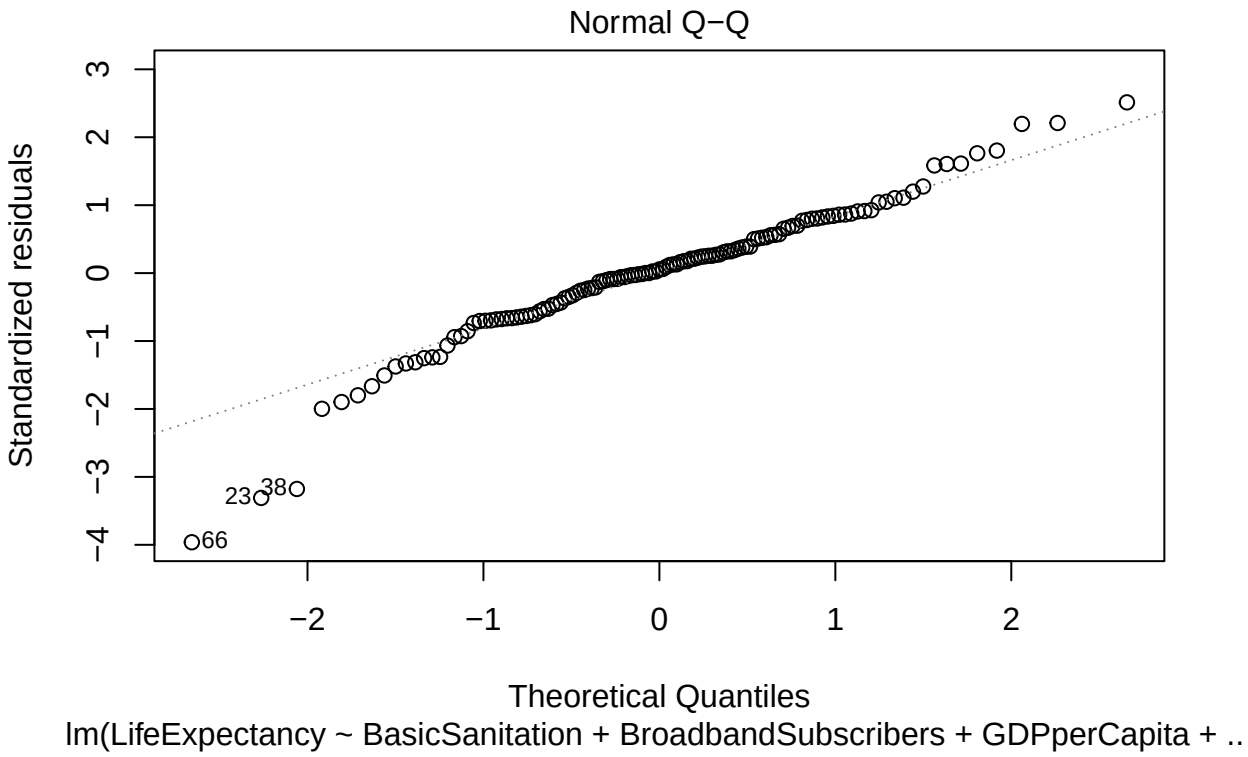
```
stepmodel2 <- lm(LifeExpectancy~BasicSanitation + BroadbandSubscribers
         + GDPperCapita + AlcoholConsumption, data=Dataset1)
summary(stepmodel2)
```

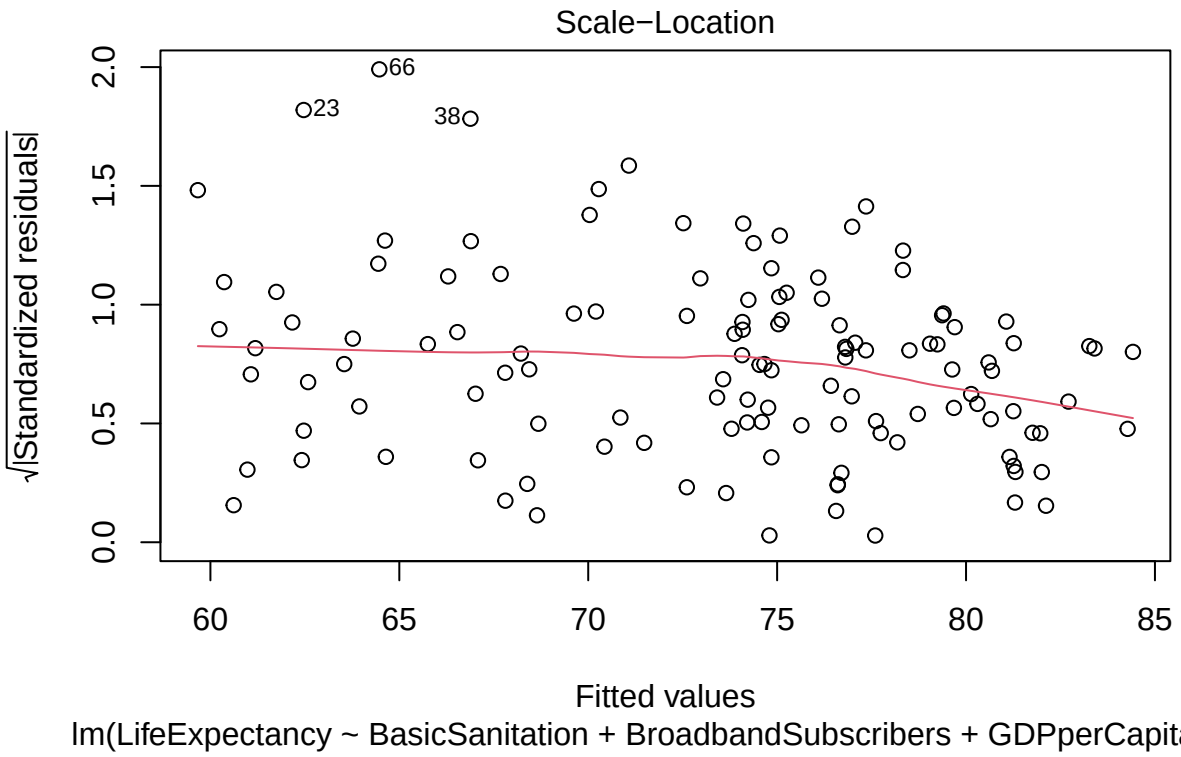
```
##
## Call:
## lm(formula = LifeExpectancy ~ BasicSanitation + BroadbandSubscribers +
##     GDPperCapita + AlcoholConsumption, data = Dataset1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.9704  -1.9064   0.1896   1.9829   8.9238
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.896e+01  1.196e+00  49.297  < 2e-16 ***
## BasicSanitation        1.609e+01  1.602e+00  10.045  < 2e-16 ***
## BroadbandSubscribers   1.708e-01  5.397e-02   3.165  0.00196 **
## GDPperCapita           4.707e-05  2.638e-05   1.784  0.07684 .
## AlcoholConsumption    -1.721e-01  1.051e-01  -1.637  0.10430
```

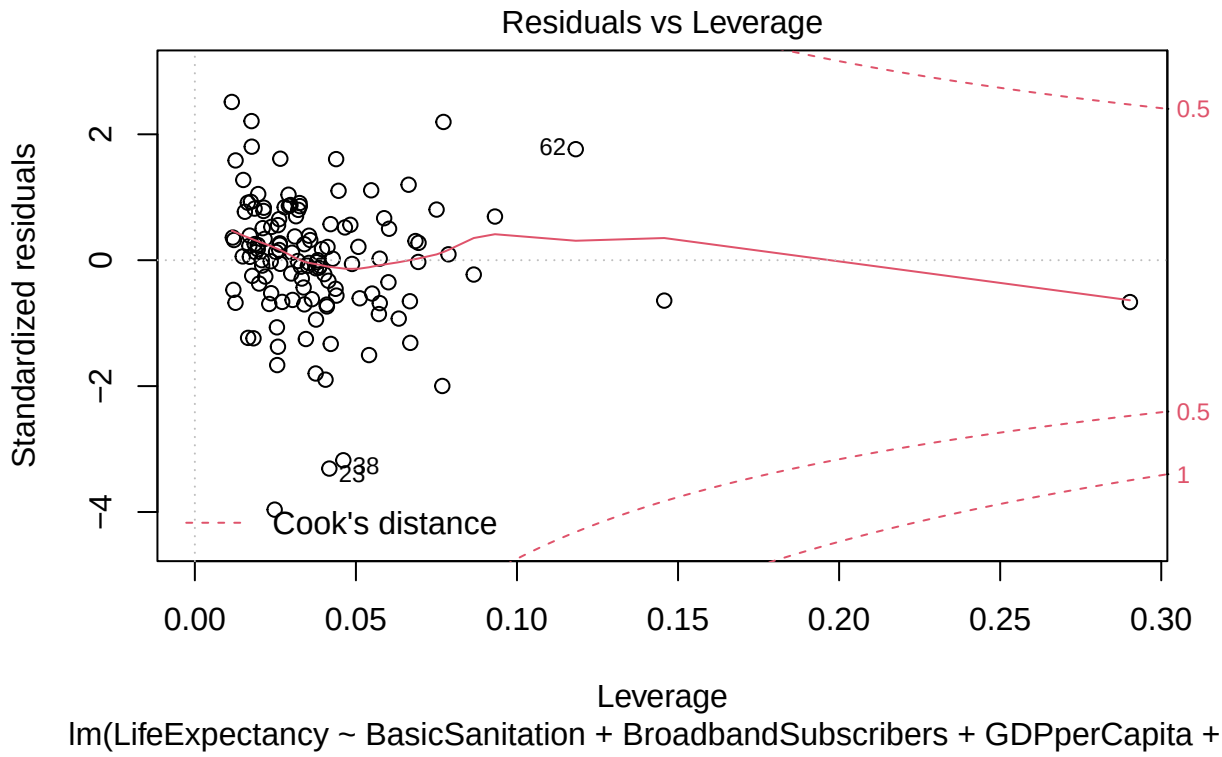
```
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.571 on 122 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7717, Adjusted R-squared:  0.7642
## F-statistic: 103.1 on 4 and 122 DF,  p-value: < 2.2e-16
```

```
plot(stepmodel2)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(LifeExpectancy ~ BasicSanitation + BroadbandSubscribers + GDPperCapita + ..
23
38
66


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
66
23
38
60
65
70
75
80
85
Fitted values
lm(LifeExpectancy ~ BasicSanitation + BroadbandSubscribers + GDPperCapita + ..




```
extractAIC(stepmodel2, scale=MSE)
```

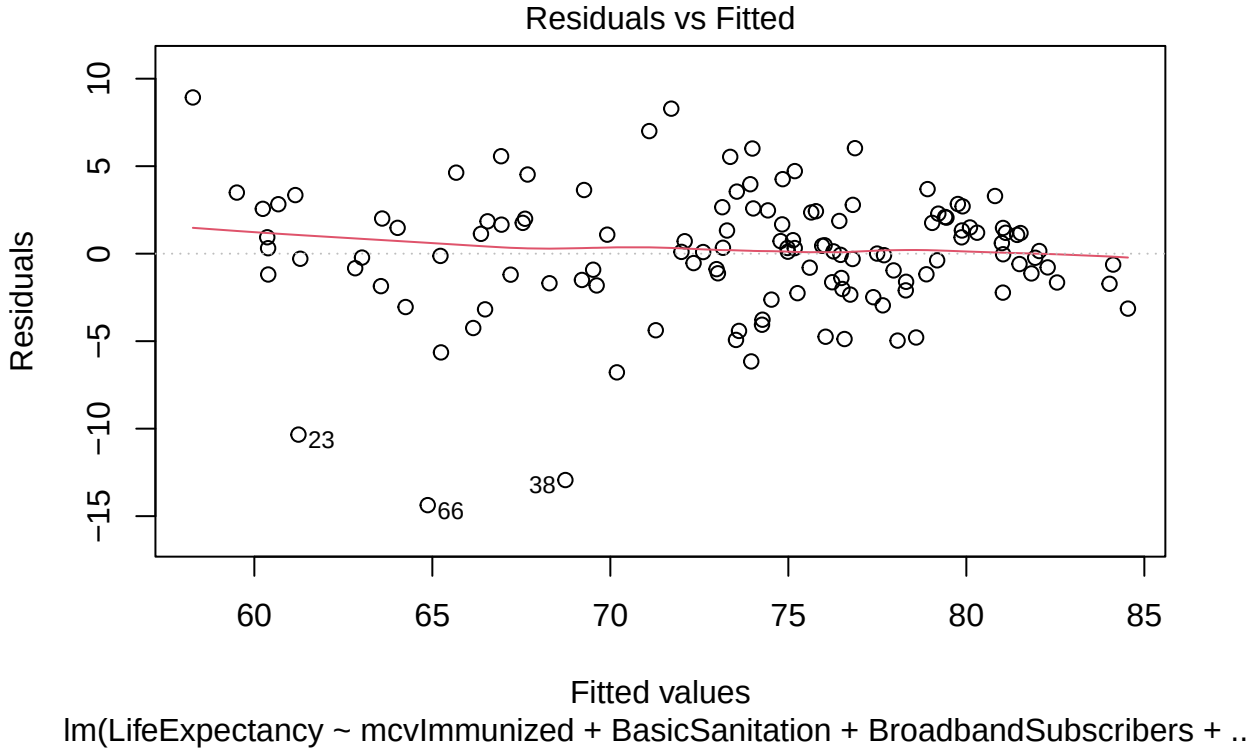
```
## [1] 5.000000 5.772259
```

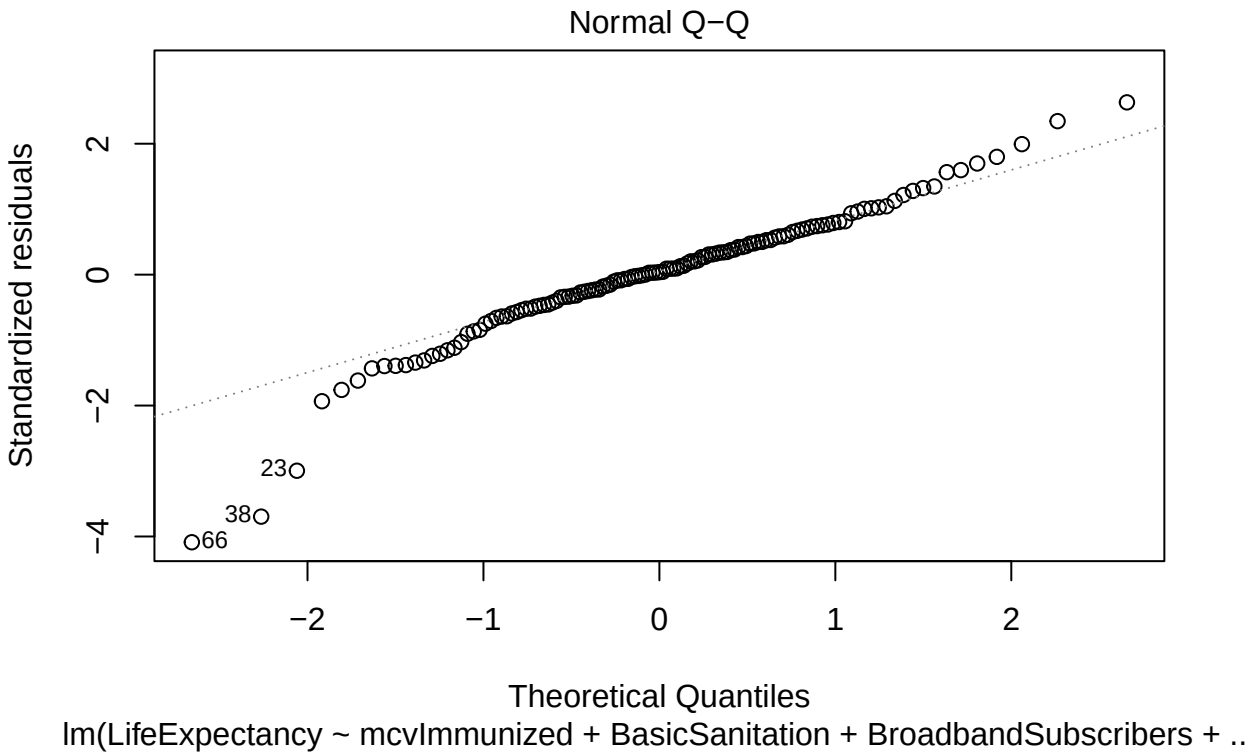
```
stepmodel3 <- lm(LifeExpectancy~mcvImmunized + BasicSanitation + BroadbandSubscribers
         + GDPperCapita, data=Dataset1)
summary(stepmodel3)
```

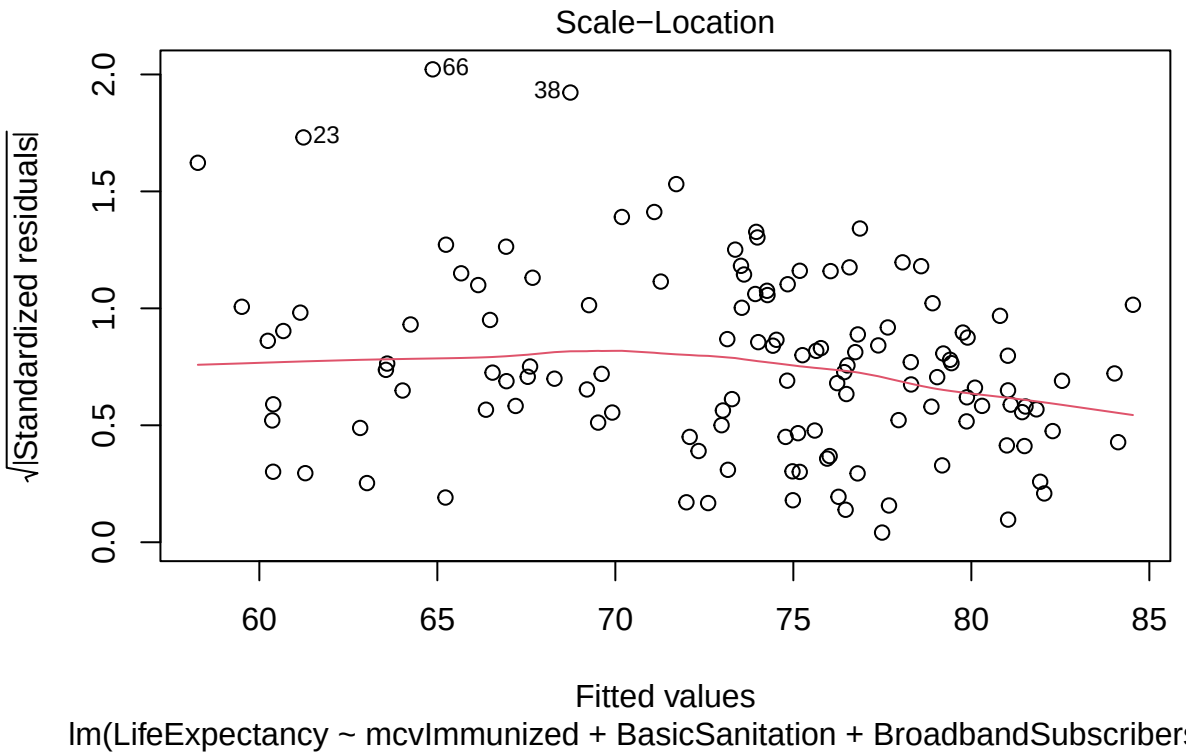
```
##
## Call:
## lm(formula = LifeExpectancy ~ mcvImmunized + BasicSanitation +
##     BroadbandSubscribers + GDPperCapita, data = Dataset1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3705  -1.6394   0.1318   2.0022   8.9273
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.341e+01  2.819e+00  18.941  < 2e-16 ***
## mcvImmunized         6.616e+00  3.764e+00   1.758  0.08127 .
## BasicSanitation      1.471e+01  1.875e+00   7.845 1.85e-12 ***
## BroadbandSubscribers 1.262e-01  4.529e-02   2.786  0.00618 **
## GDPperCapita         5.620e-05  2.590e-05   2.170  0.03194 *
```

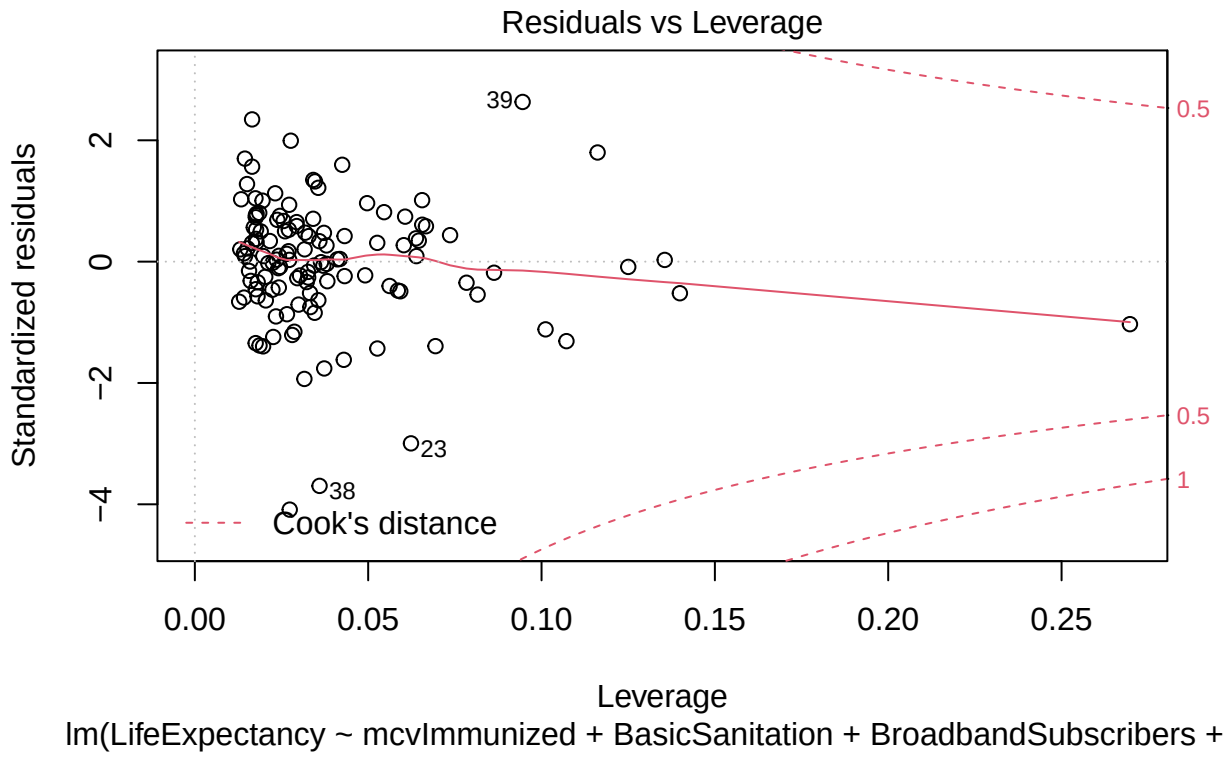
```
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.565 on 122 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7724, Adjusted R-squared:  0.7649
## F-statistic: 103.5 on 4 and 122 DF,  p-value: < 2.2e-16
```

```
plot(stepmodel3)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(LifeExpectancy ~ mcvImmunized + BasicSanitation + BroadbandSubscribers + ..
23
38
66
−4
−2
0
2
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
66
38
23
60
65
70
75
80
85
Fitted values
lm(LifeExpectancy ~ mcvImmunized + BasicSanitation + BroadbandSubscribers + ..




```
extractAIC(stepmodel3, scale=MSE)
```

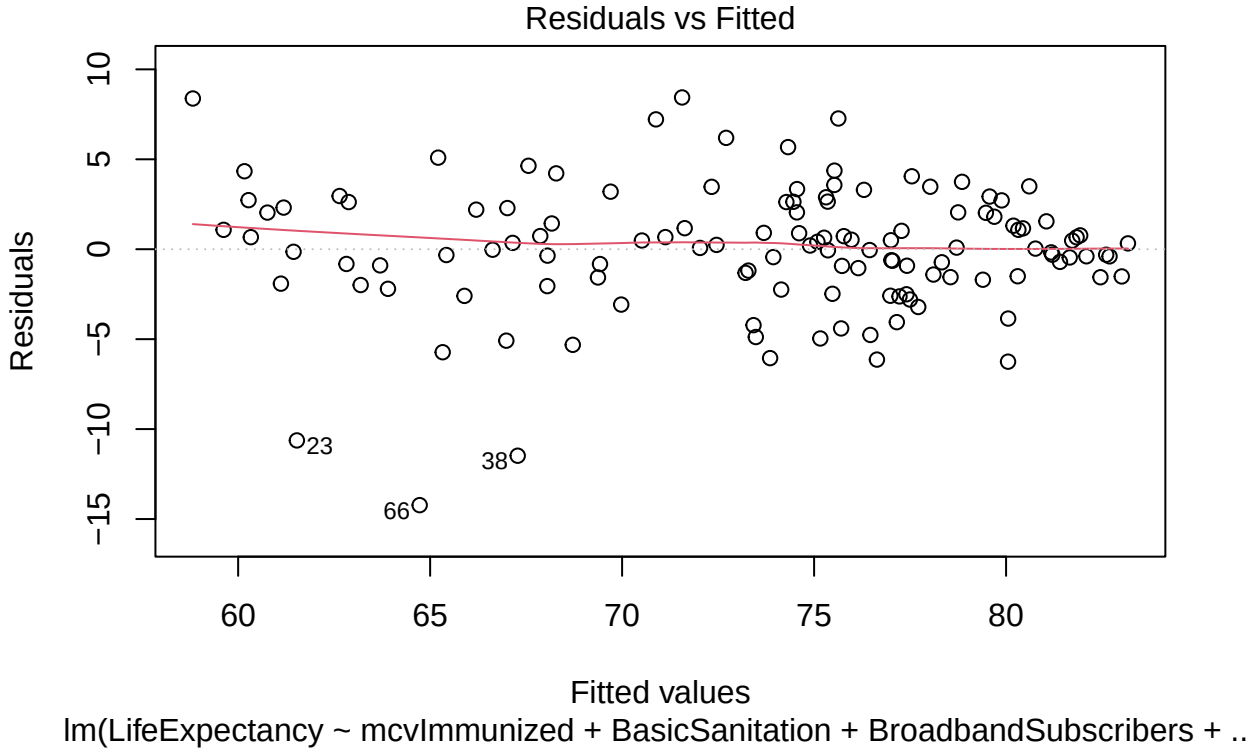
```
## [1] 5.000000 5.368035
```

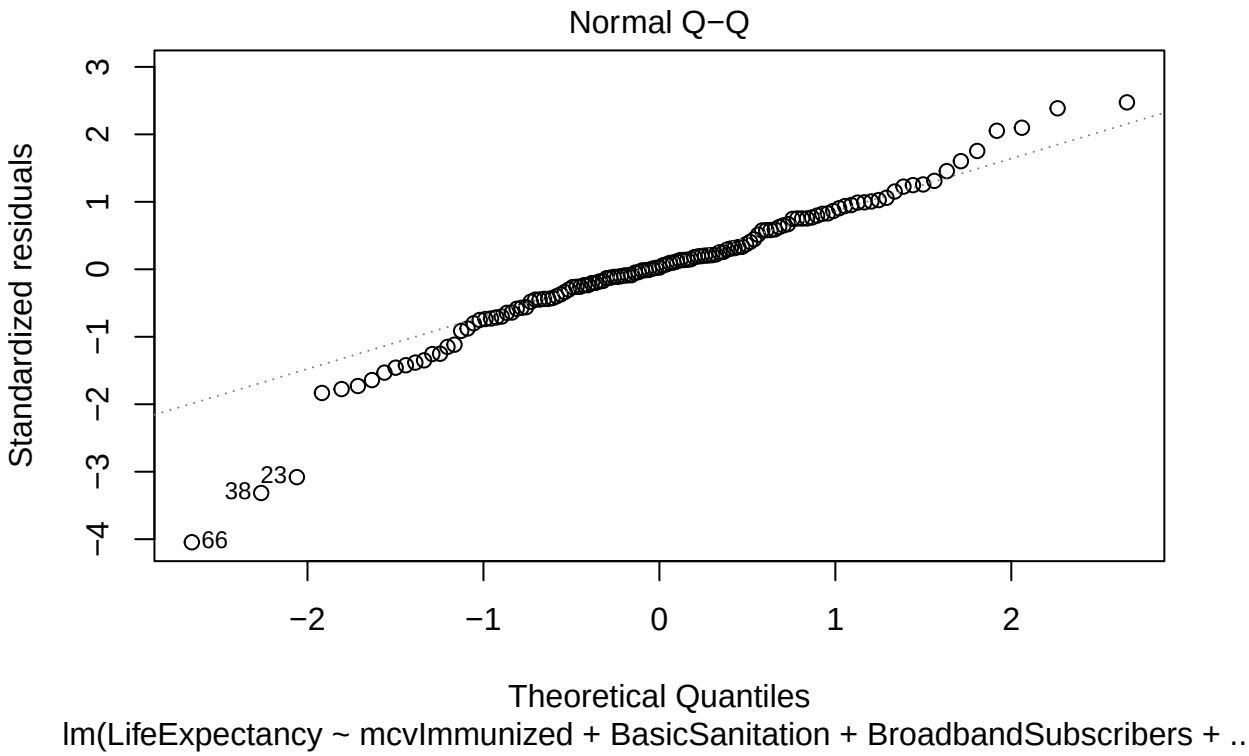
```
stepmodel4 <- lm(LifeExpectancy~mcvImmunized + BasicSanitation + BroadbandSubscribers
          + AlcoholConsumption, data=Dataset1)
summary(stepmodel4)
```

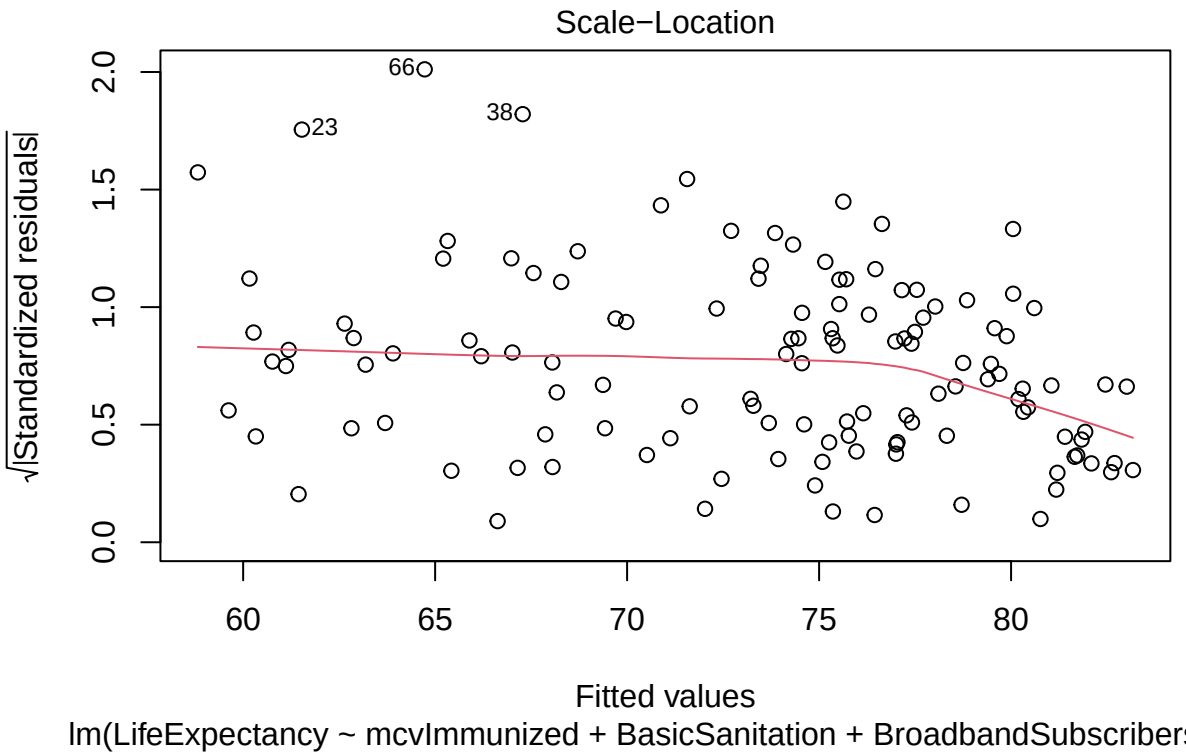
```
##
## Call:
## lm(formula = LifeExpectancy ~ mcvImmunized + BasicSanitation +
##     BroadbandSubscribers + AlcoholConsumption, data = Dataset1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2275  -1.5556   0.0889   2.1239   8.4373
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          54.35986    2.83976  19.142  < 2e-16 ***
## mcvImmunized          7.07960    3.77706   1.874   0.0633 .
## BasicSanitation      13.91149    1.89996   7.322 2.87e-11 ***
## BroadbandSubscribers  0.24587    0.03944   6.235 6.74e-09 ***
## AlcoholConsumption   -0.22279    0.10360  -2.150   0.0335 *
```

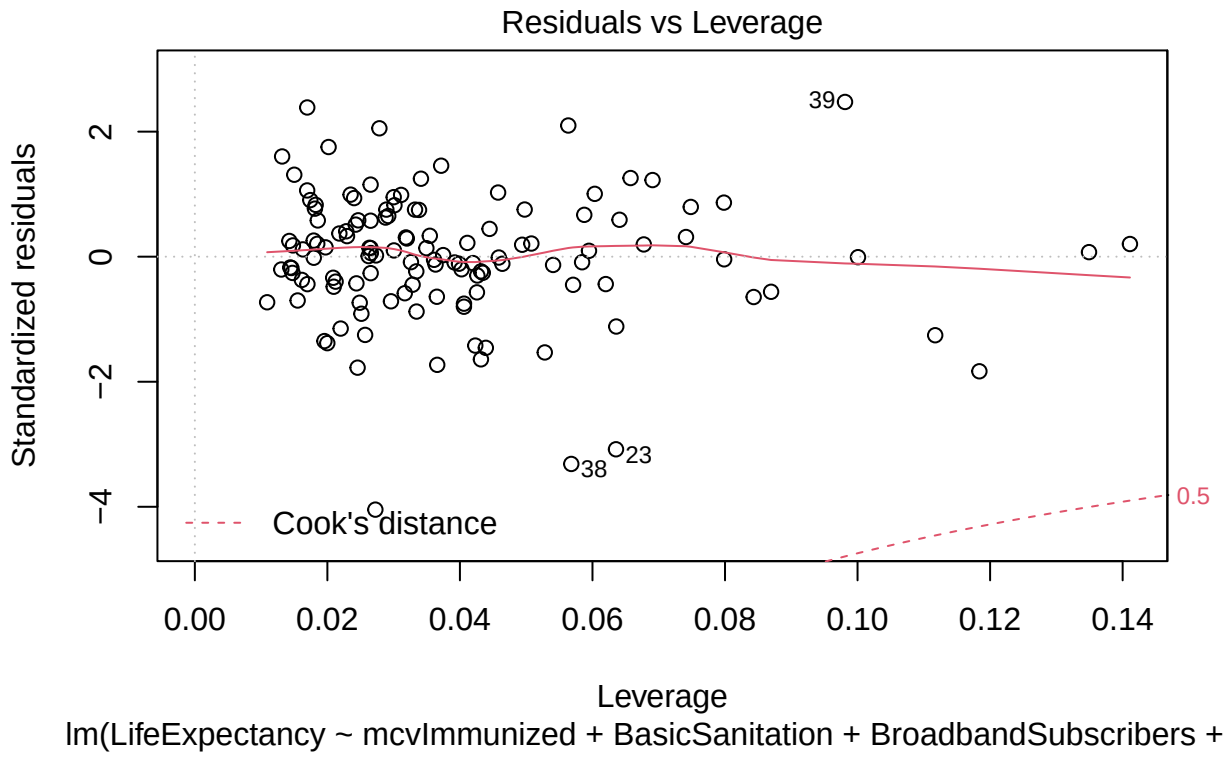
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.566 on 122 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7723, Adjusted R-squared:  0.7648
## F-statistic: 103.4 on 4 and 122 DF,  p-value: < 2.2e-16
```

```
plot(stepmodel4)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(LifeExpectancy ~ mcvImmunized + BasicSanitation + BroadbandSubscribers + ..
38
23
66


Scale−Location
66
38
23
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
60
65
70
75
80
Fitted values
lm(LifeExpectancy ~ mcvImmunized + BasicSanitation + BroadbandSubscribers + ..




```
extractAIC(stepmodel4, scale=MSE)
```

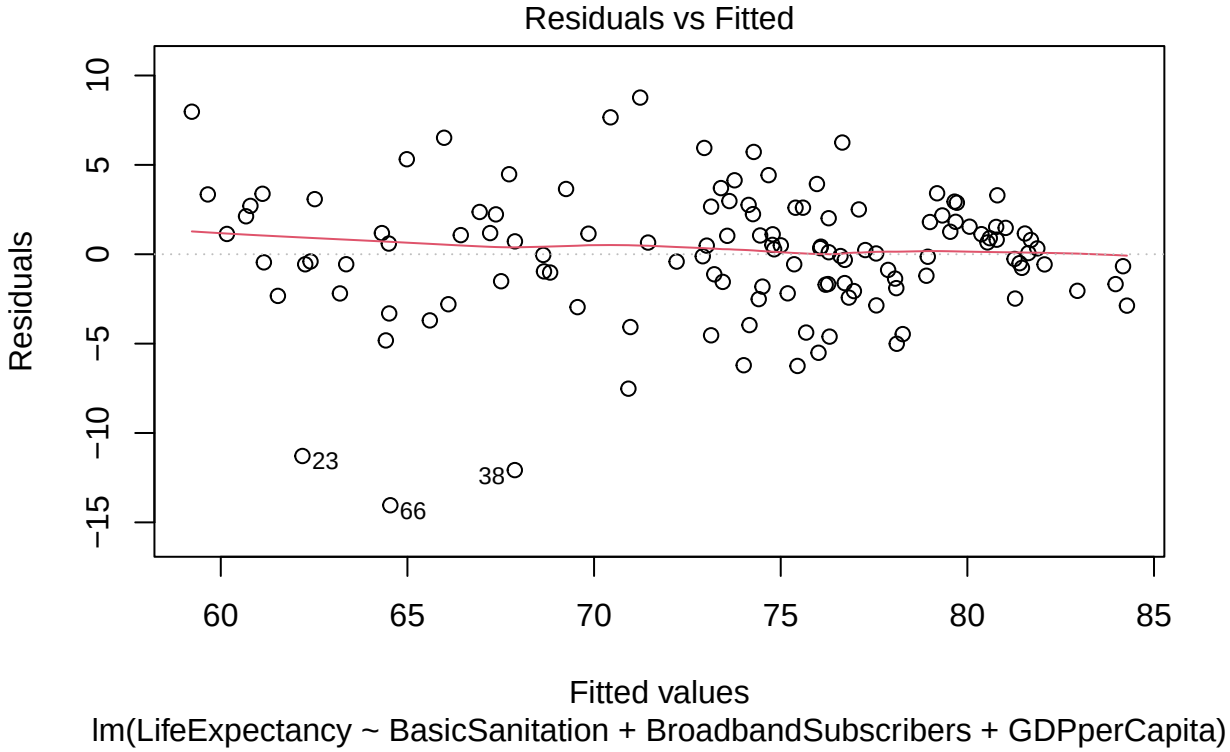
```
## [1] 5.000000 5.450383
```

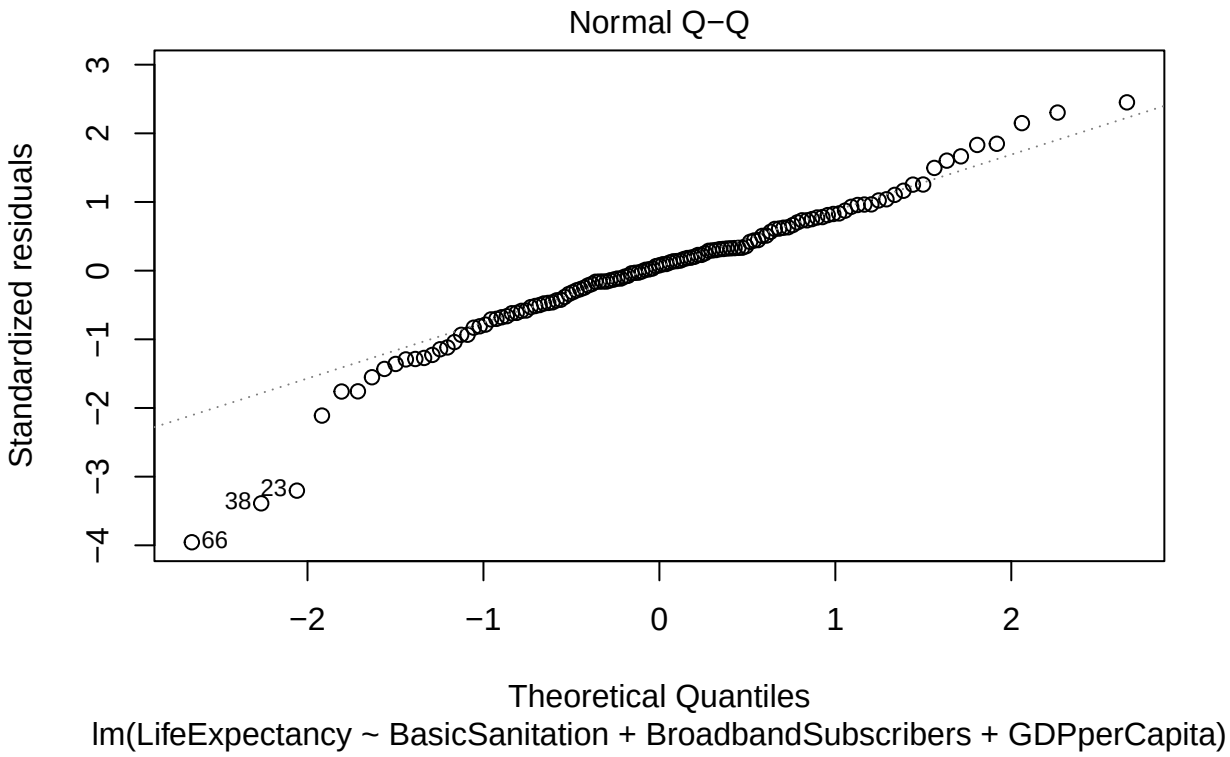
```
stepmodel5 <- lm(LifeExpectancy~ BasicSanitation + BroadbandSubscribers
        + GDPperCapita, data=Dataset1)
summary(stepmodel5)
```

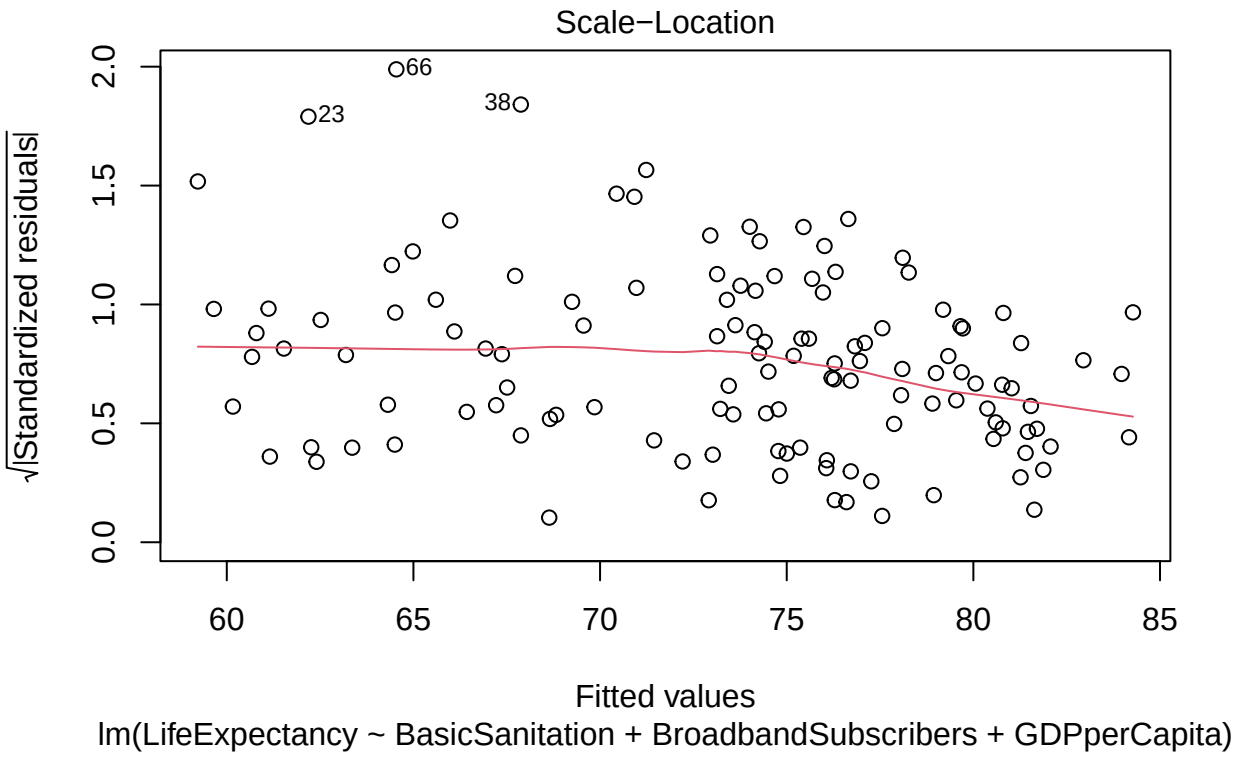
```
##
## Call:
## lm(formula = LifeExpectancy ~ BasicSanitation + BroadbandSubscribers +
##     GDPperCapita, data = Dataset1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.0424  -1.6870   0.2769   2.1469   8.7645
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.801e+01  1.053e+00  55.095  < 2e-16 ***
## BasicSanitation      1.649e+01  1.594e+00  10.343  < 2e-16 ***
## BroadbandSubscribers 1.229e-01  4.564e-02   2.693  0.00808 **
## GDPperCapita         5.496e-05  2.611e-05   2.105  0.03731 *
## ---
```

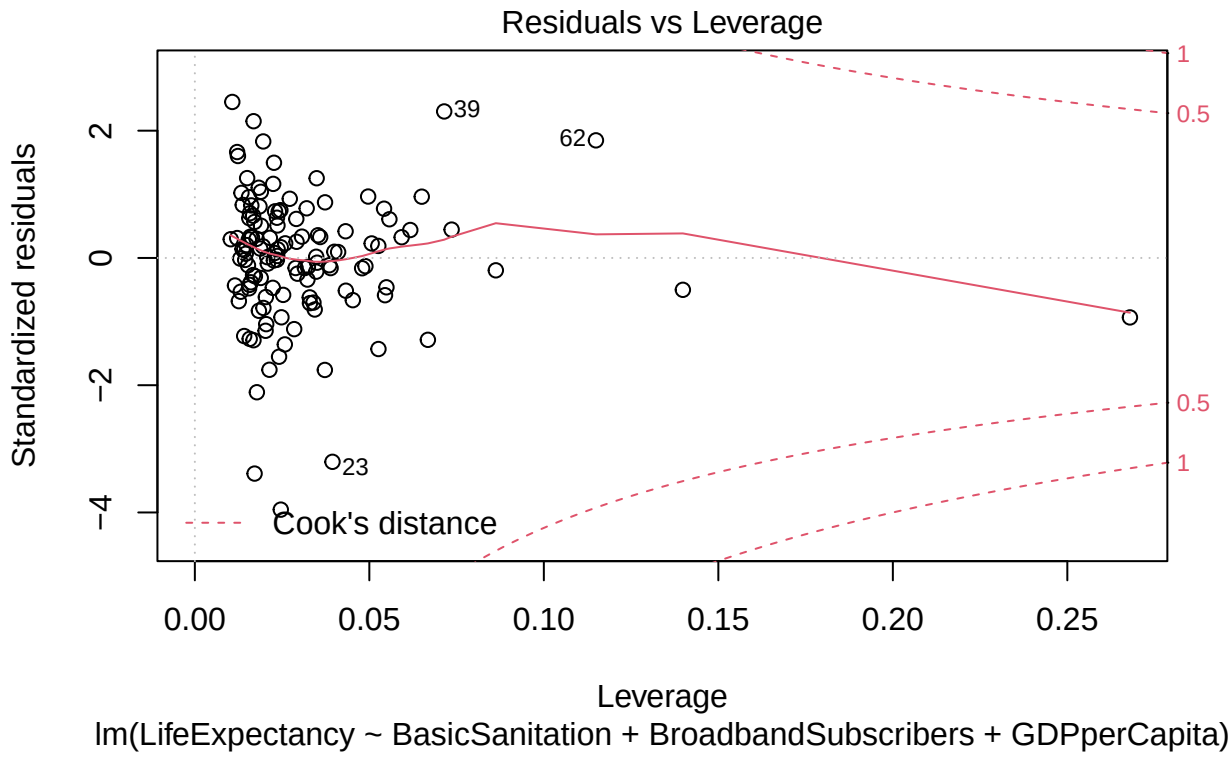
```
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.595 on 123 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7666, Adjusted R-squared:  0.761
## F-statistic: 134.7 on 3 and 123 DF,  p-value: < 2.2e-16
```

```
plot(stepmodel5)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(LifeExpectancy ~ BasicSanitation + BroadbandSubscribers + GDPperCapita)
38
23
66


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
66
23
38
60
65
70
75
80
85
Fitted values
lm(LifeExpectancy ~ BasicSanitation + BroadbandSubscribers + GDPperCapita)




```
extractAIC(stepmodel5, scale=MSE)
```

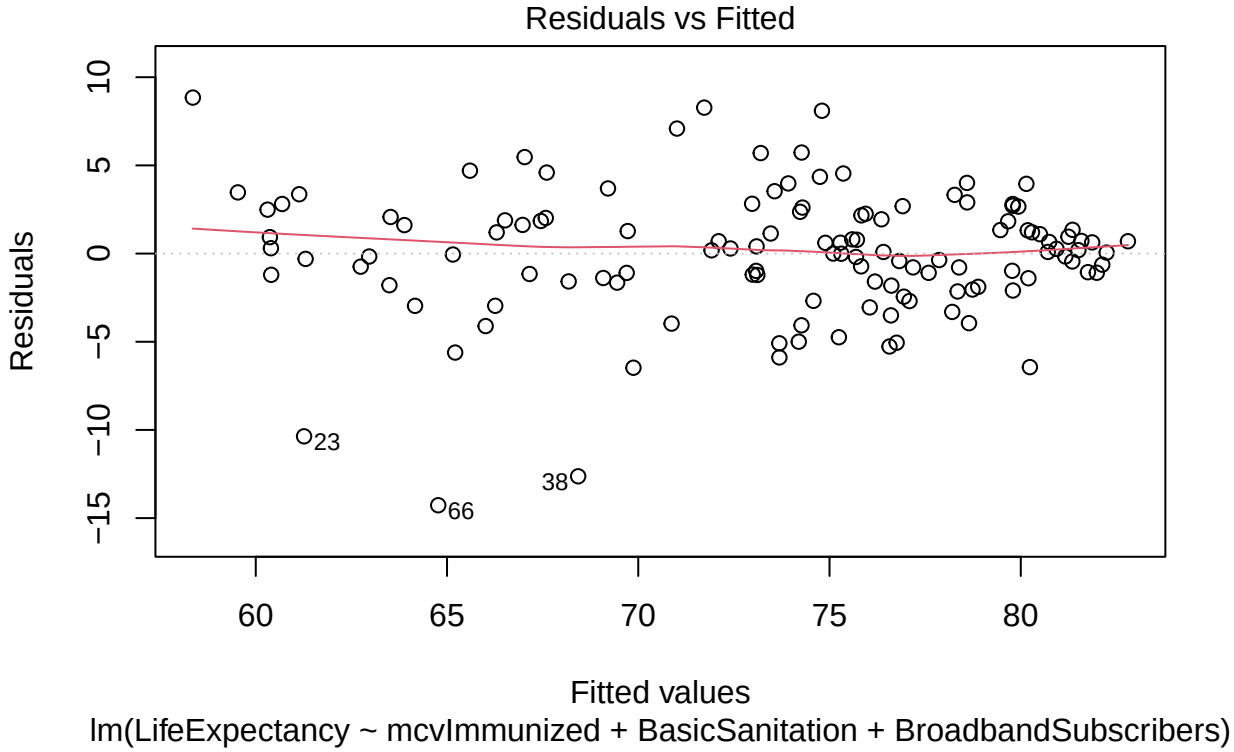
```
## [1] 4.000000 6.467558
```

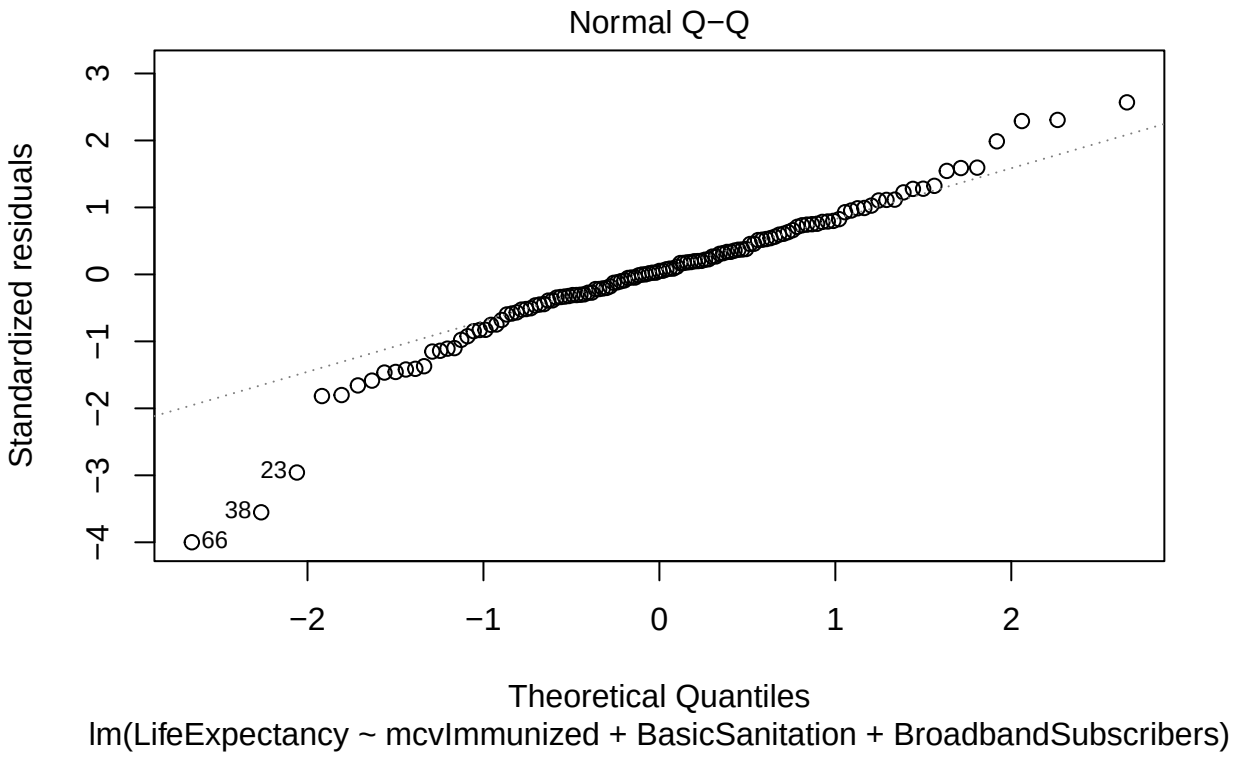
```
stepmodel6 <- lm(LifeExpectancy~mcvImmunized + BasicSanitation
                 + BroadbandSubscribers, data=Dataset1)
summary(stepmodel6)
```

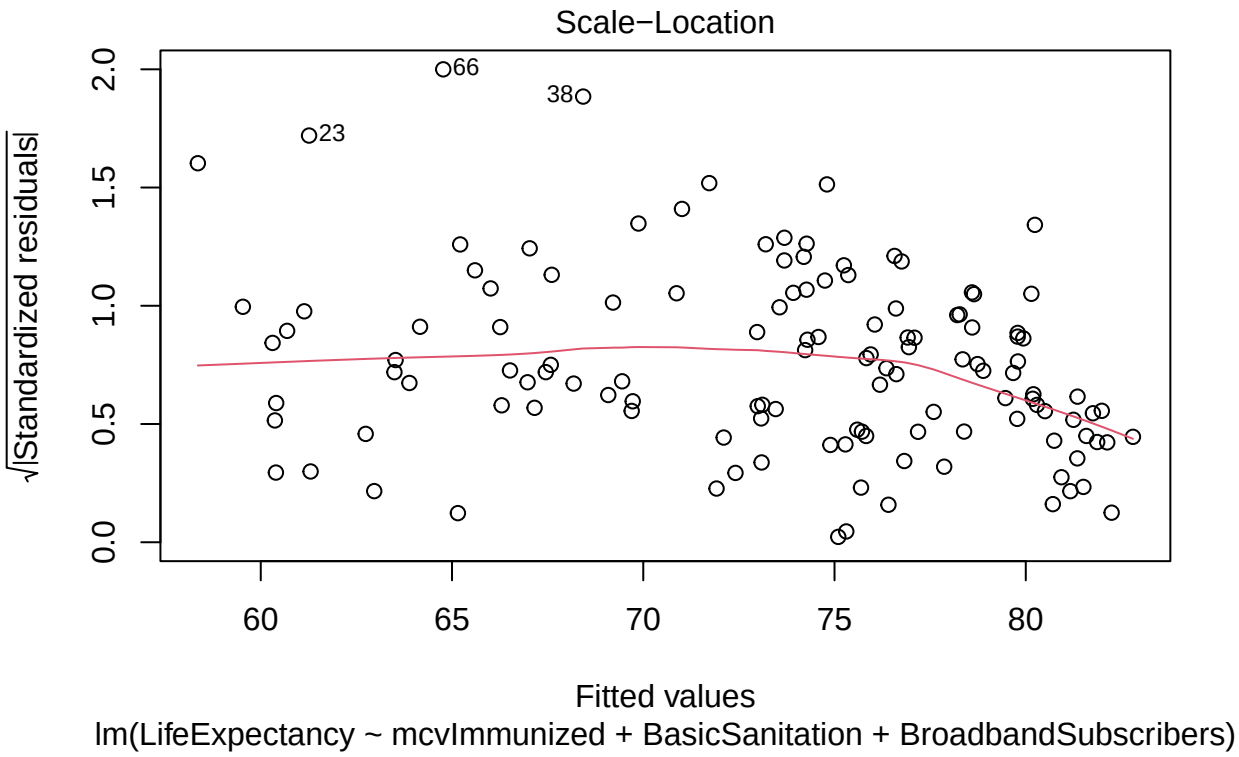
```
##
## Call:
## lm(formula = LifeExpectancy ~ mcvImmunized + BasicSanitation +
##     BroadbandSubscribers, data = Dataset1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2708  -1.5839   0.1841   2.0466   8.8449
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          53.61804    2.85996  18.748  < 2e-16 ***
## mcvImmunized          6.39365    3.81862   1.674   0.0966 .
## BasicSanitation      14.57429    1.90221   7.662 4.71e-12 ***
## BroadbandSubscribers  0.19601    0.03237   6.056 1.57e-08 ***
## ---
```

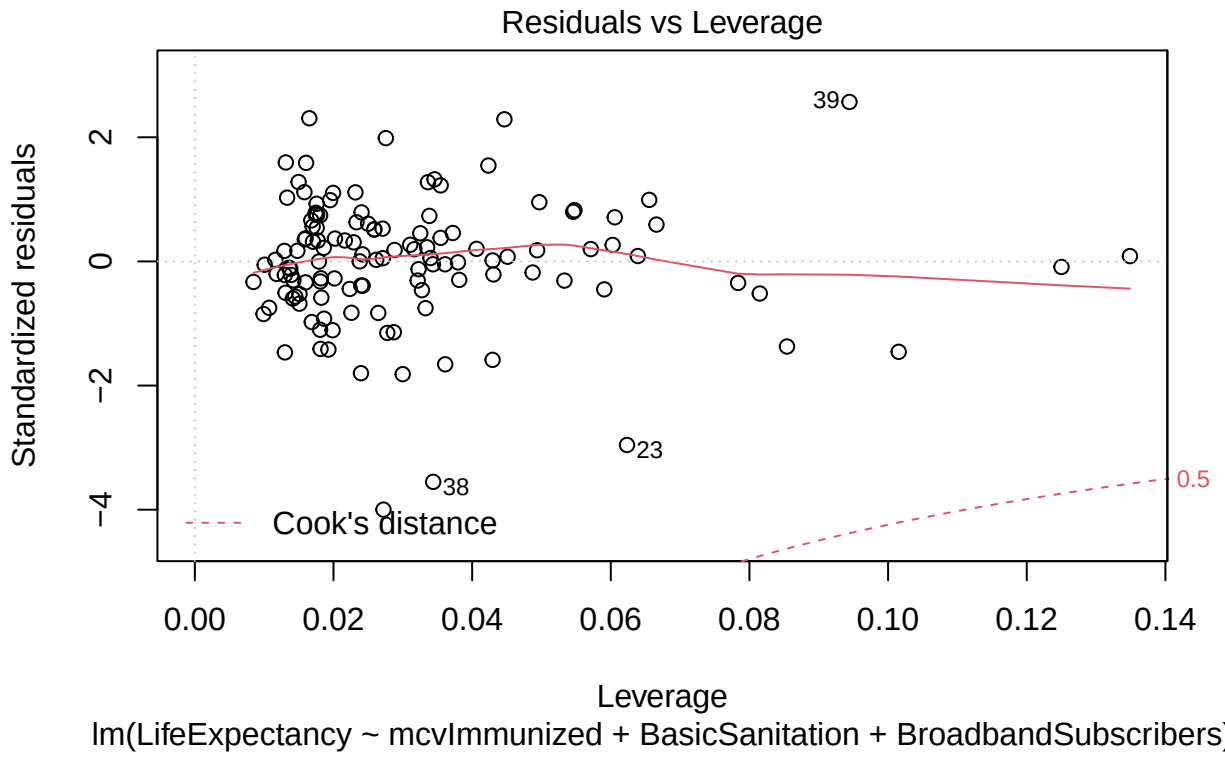
```
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.618 on 123 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7636, Adjusted R-squared:  0.7579
## F-statistic: 132.5 on 3 and 123 DF,  p-value: < 2.2e-16
```

```
plot(stepmodel6)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(LifeExpectancy ~ mcvImmunized + BasicSanitation + BroadbandSubscribers)
23
38
66


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
66
38
23
60
65
70
75
80
Fitted values
lm(LifeExpectancy ~ mcvImmunized + BasicSanitation + BroadbandSubscribers)




```
extractAIC(stepmodel6, scale=MSE)
```

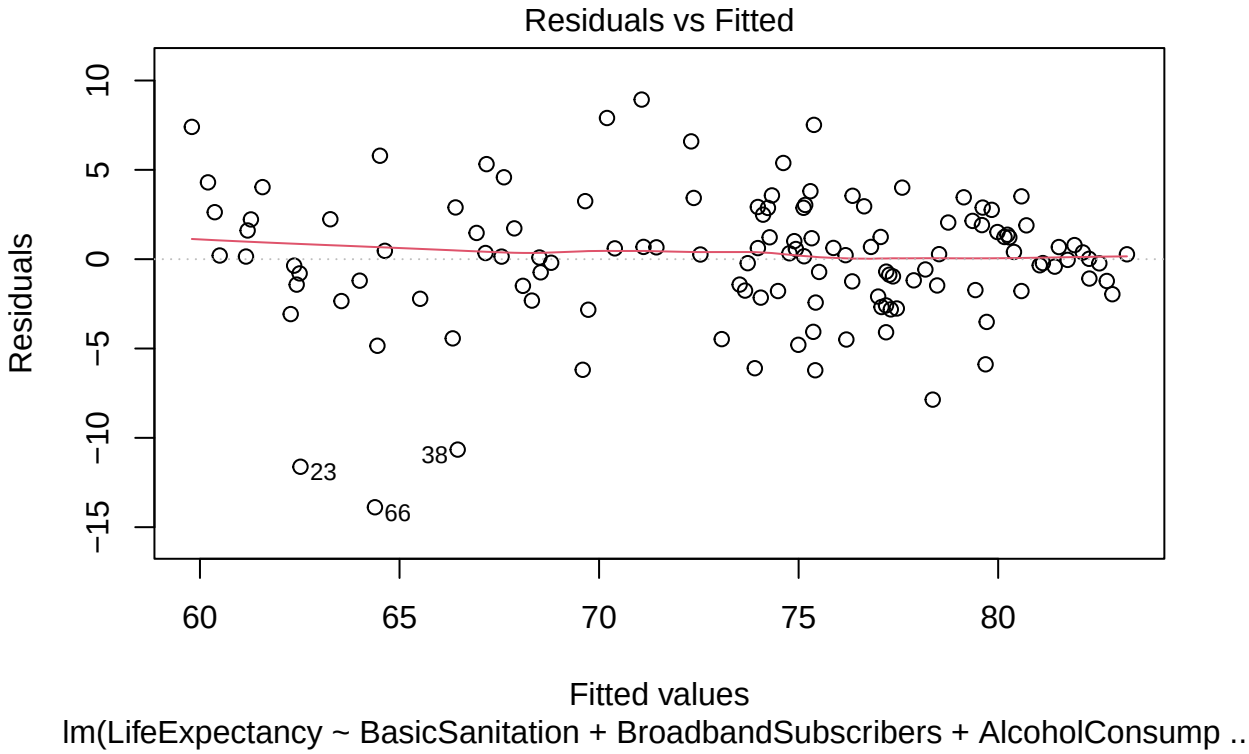
```
## [1] 4.000000 8.091641
```

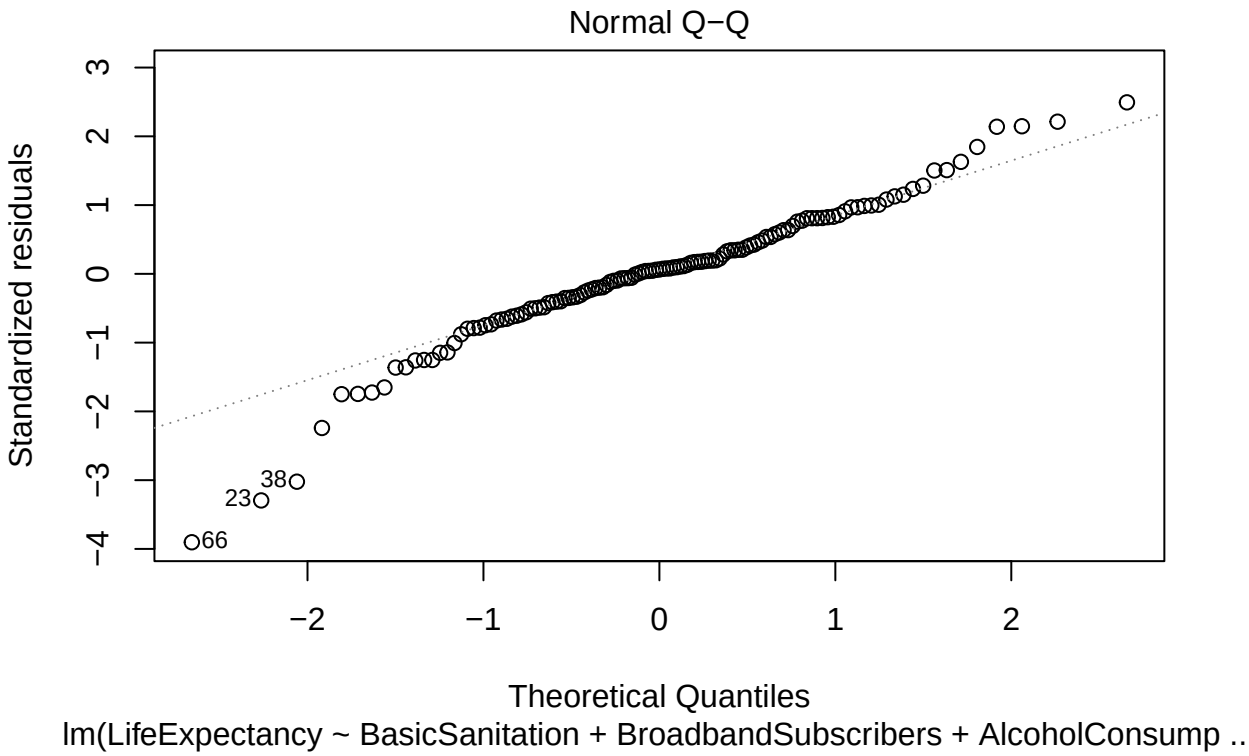
```
stepmodel7 <- lm(LifeExpectancy~BasicSanitation + BroadbandSubscribers
            + AlcoholConsumption, data=Dataset1)
summary(stepmodel7)
```

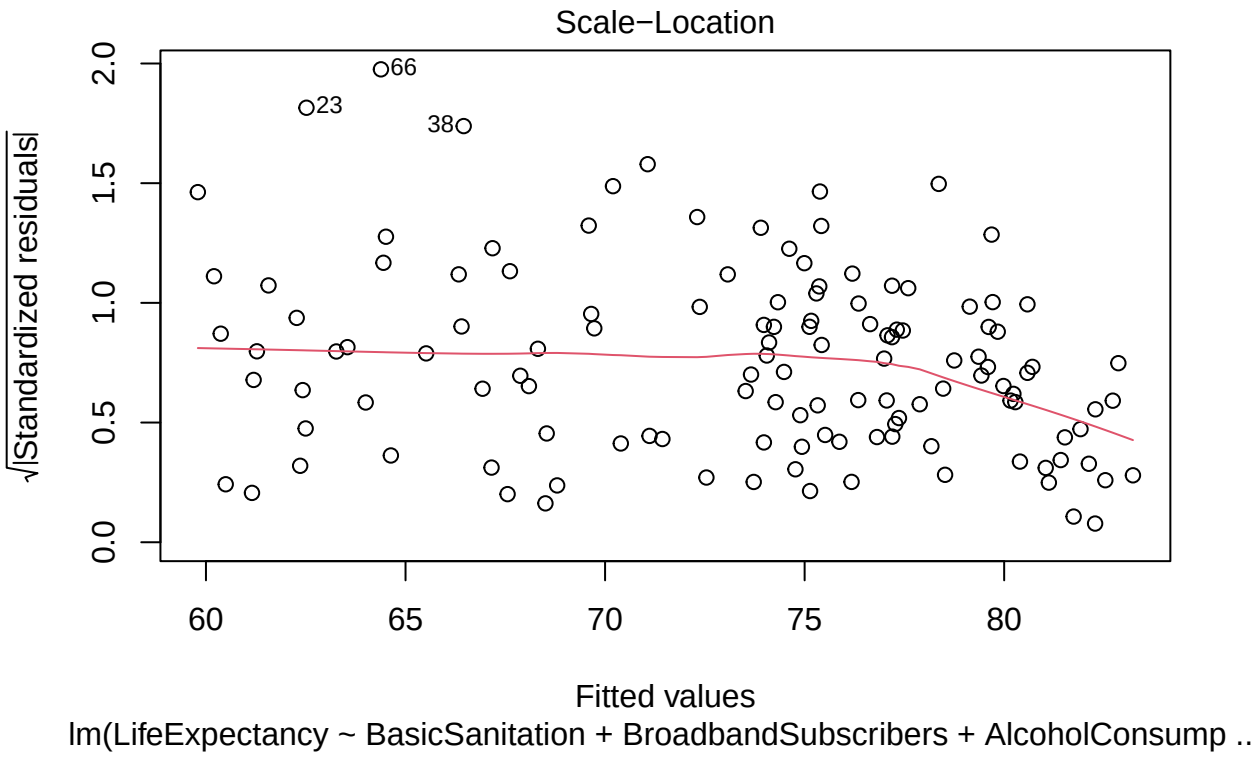
```
##
## Call:
## lm(formula = LifeExpectancy ~ BasicSanitation + BroadbandSubscribers +
##     AlcoholConsumption, data = Dataset1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.8842  -1.7433   0.2238   2.0963   8.9324
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          59.19525    1.19914  49.365  < 2e-16 ***
## BasicSanitation      15.84953    1.61017   9.843  < 2e-16 ***
## BroadbandSubscribers  0.23703    0.03955   5.993 2.11e-08 ***
## AlcoholConsumption   -0.20639    0.10428  -1.979     0.05 .
## ---
```

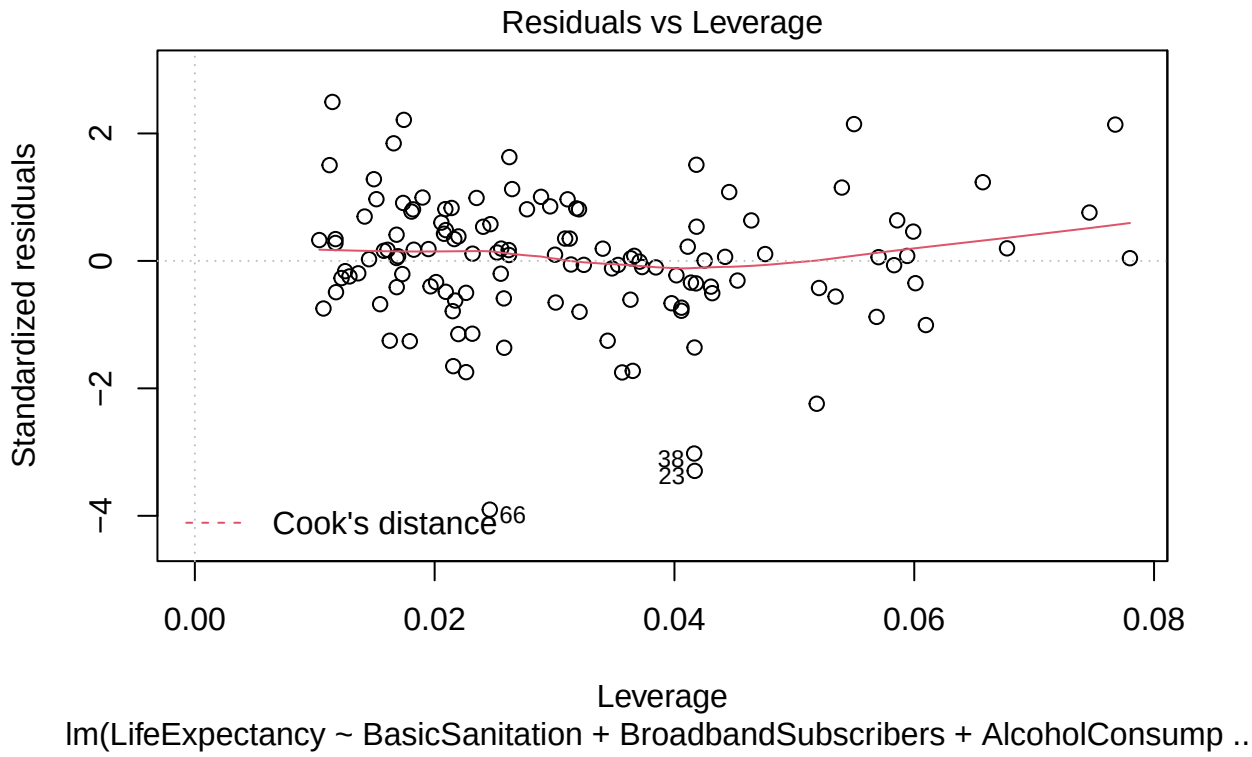
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.602 on 123 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7657, Adjusted R-squared:   0.76
## F-statistic:   134 on 3 and 123 DF,  p-value: < 2.2e-16
```

```
plot(stepmodel7)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(LifeExpectancy ~ BasicSanitation + BroadbandSubscribers + AlcoholConsump ..
38
23
66


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
66
23
38
60
65
70
75
80
Fitted values
lm(LifeExpectancy ~ BasicSanitation + BroadbandSubscribers + AlcoholConsump ..




```
extractAIC(stepmodel7, scale=MSE)
```

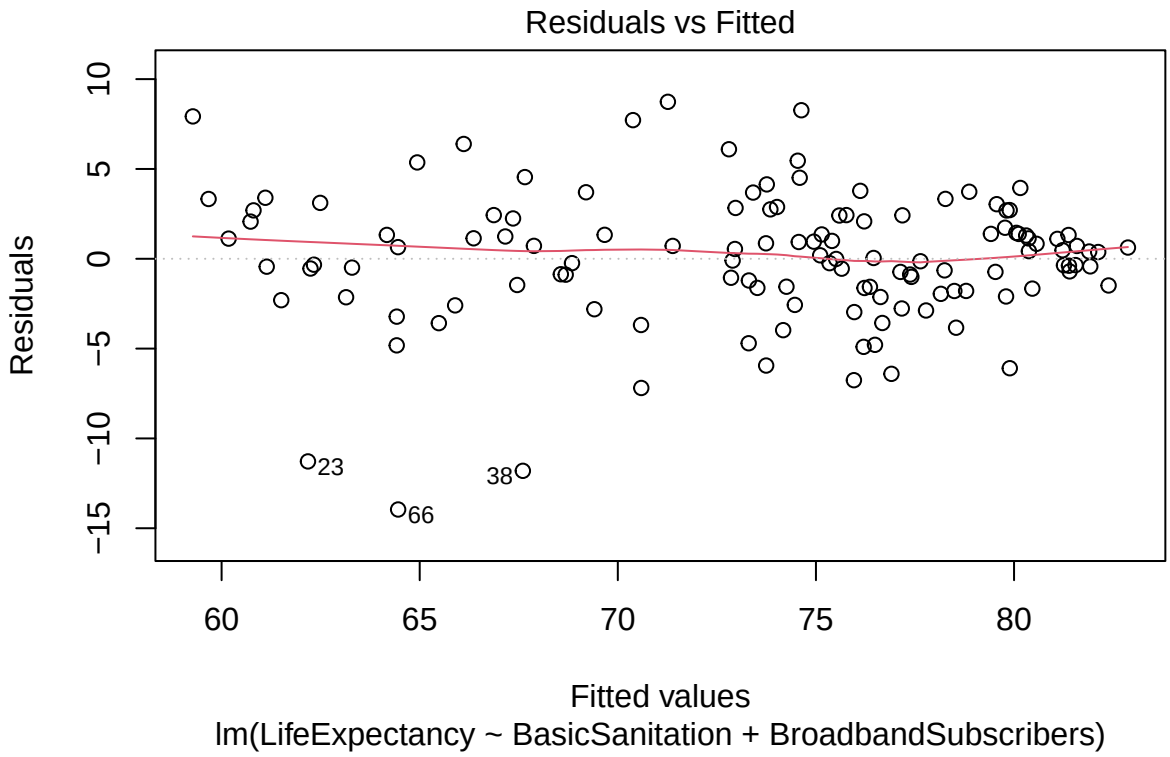
```
## [1] 4.000000 6.976595
```

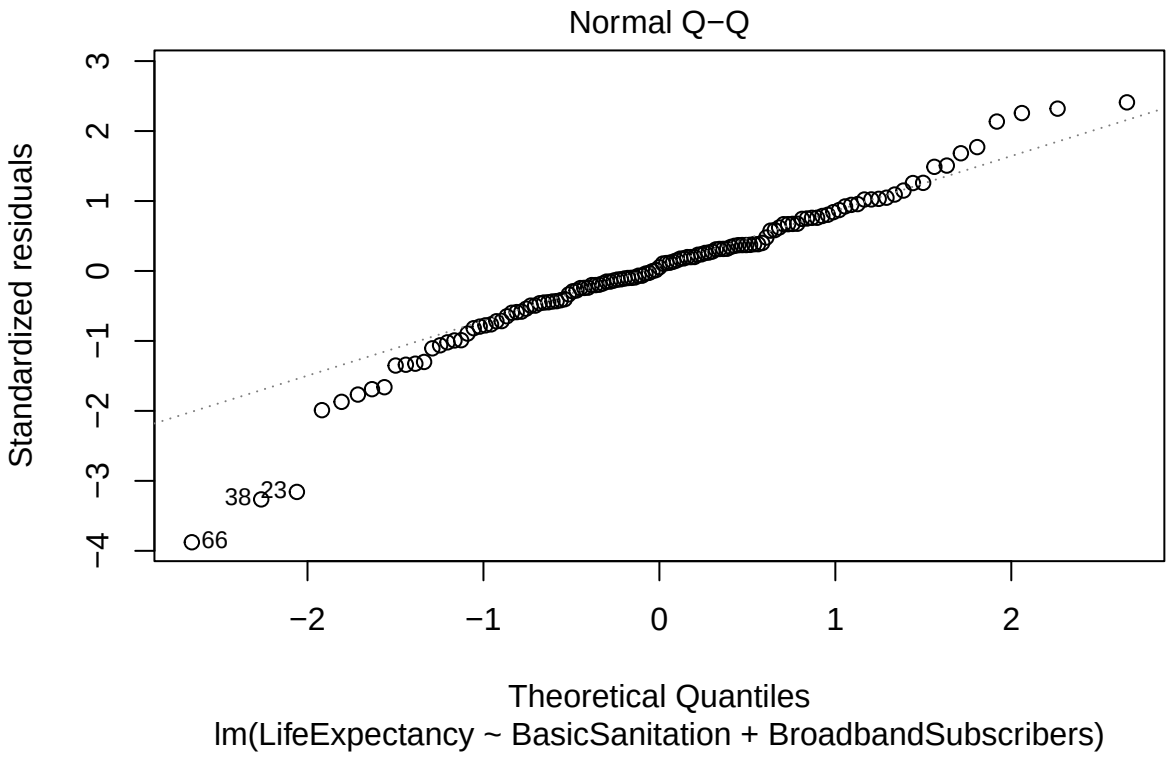
```
stepmodel8 <- lm(LifeExpectancy~BasicSanitation + BroadbandSubscribers, data=Dataset1)
summary(stepmodel8)
```

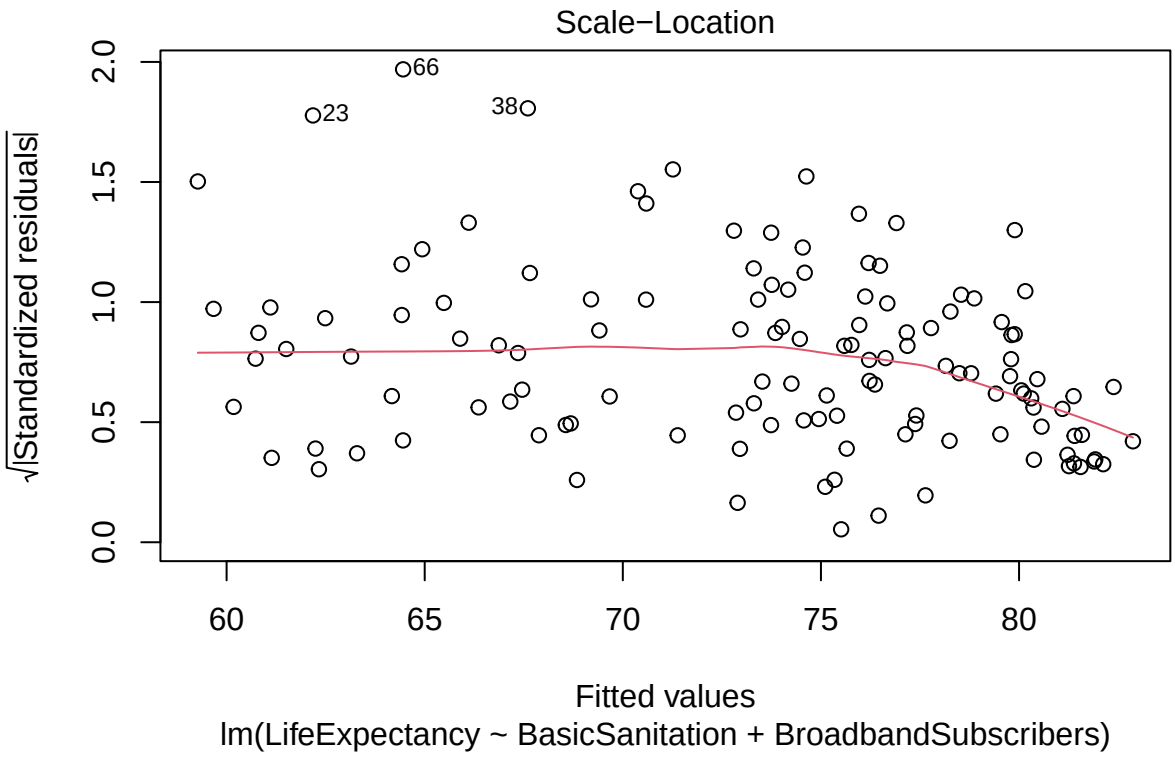
```
##
## Call:
## lm(formula = LifeExpectancy ~ BasicSanitation + BroadbandSubscribers,
##     data = Dataset1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.9556  -1.6436   0.1926   2.1637   8.7365
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          58.06597    1.06702   54.42  < 2e-16 ***
## BasicSanitation      16.29275    1.61317   10.10  < 2e-16 ***
## BroadbandSubscribers  0.19131    0.03248    5.89 3.39e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

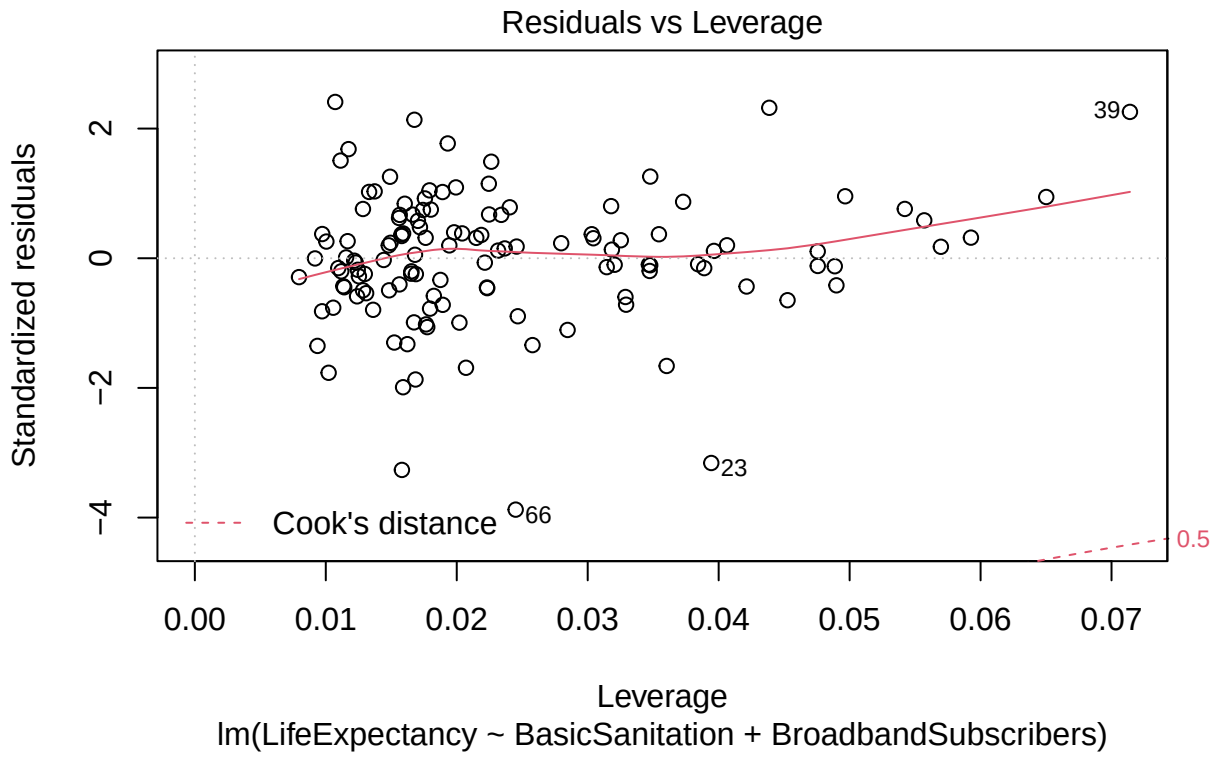
```
## Residual standard error: 3.644 on 124 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7582, Adjusted R-squared:  0.7543
## F-statistic: 194.4 on 2 and 124 DF,  p-value: < 2.2e-16
```

```
plot(stepmodel8)
```




Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
−4
38
23
66
−2
−1
0
1
2
Theoretical Quantiles
lm(LifeExpectancy ~ BasicSanitation + BroadbandSubscribers)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
66
23
38
60
65
70
75
80
Fitted values
lm(LifeExpectancy ~ BasicSanitation + BroadbandSubscribers)




```
extractAIC(stepmodel8, scale=MSE)
```

```
## [1] 3.000000 8.988307
```

Final Model

The final model selected consists of BasicSanitation, BroadbandSubscribers and mcvImmunized as predictors.

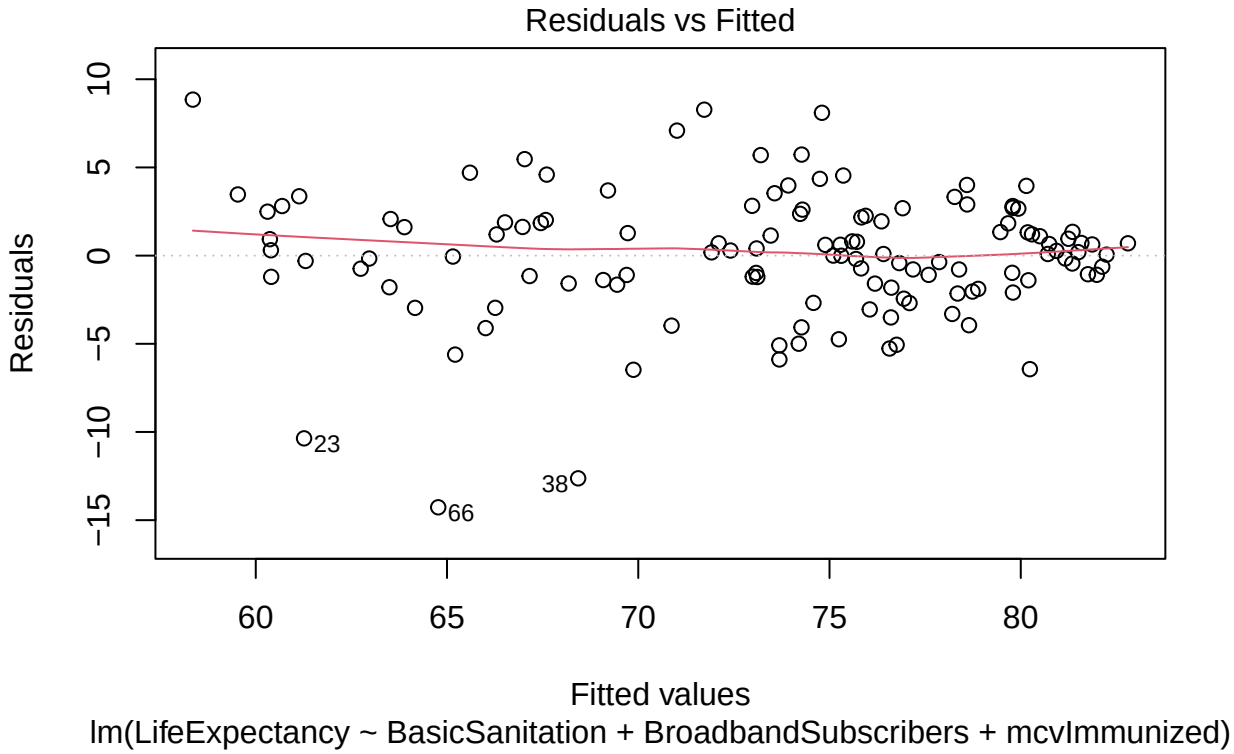
The obtained model has high R-squared, a relatively mallow Cp, and the linearity, constant variance, and normality conditions seem to be met.

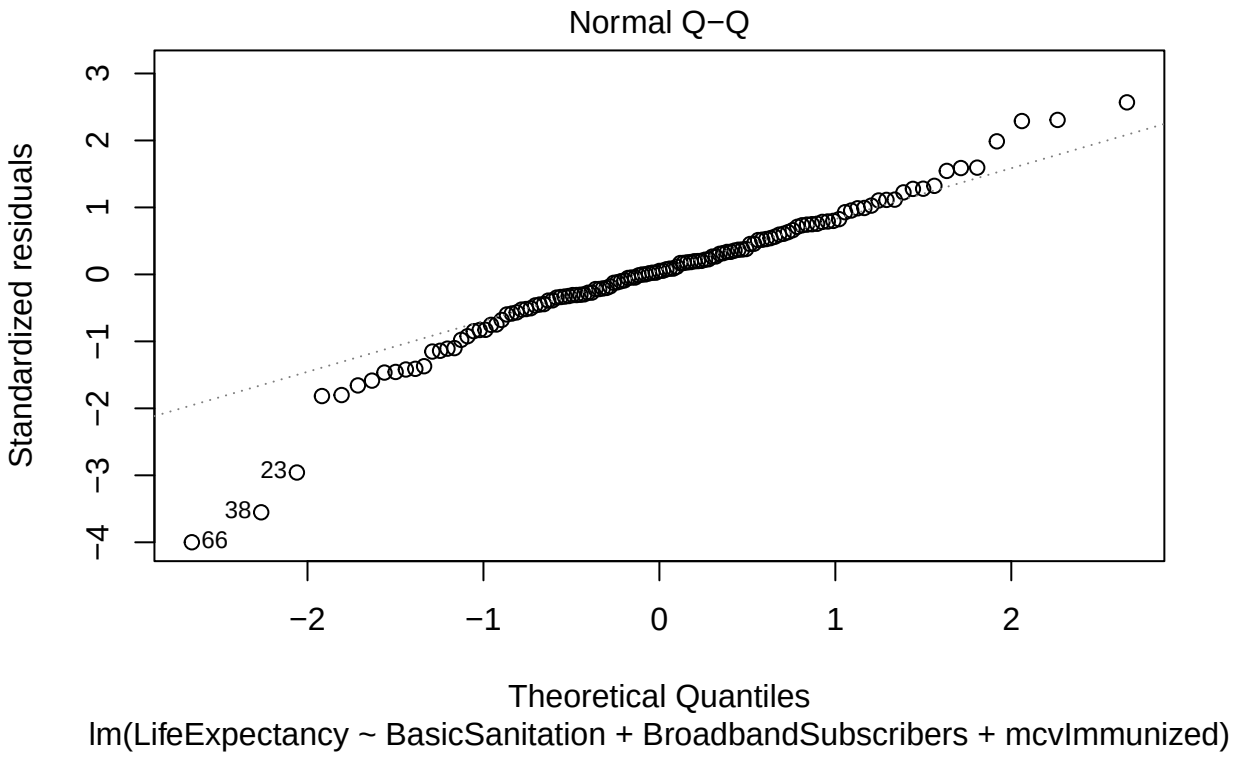
```
final <- lm(LifeExpectancy~ BasicSanitation +
            BroadbandSubscribers + mcvImmunized, data=Dataset1)
summary(final)
```

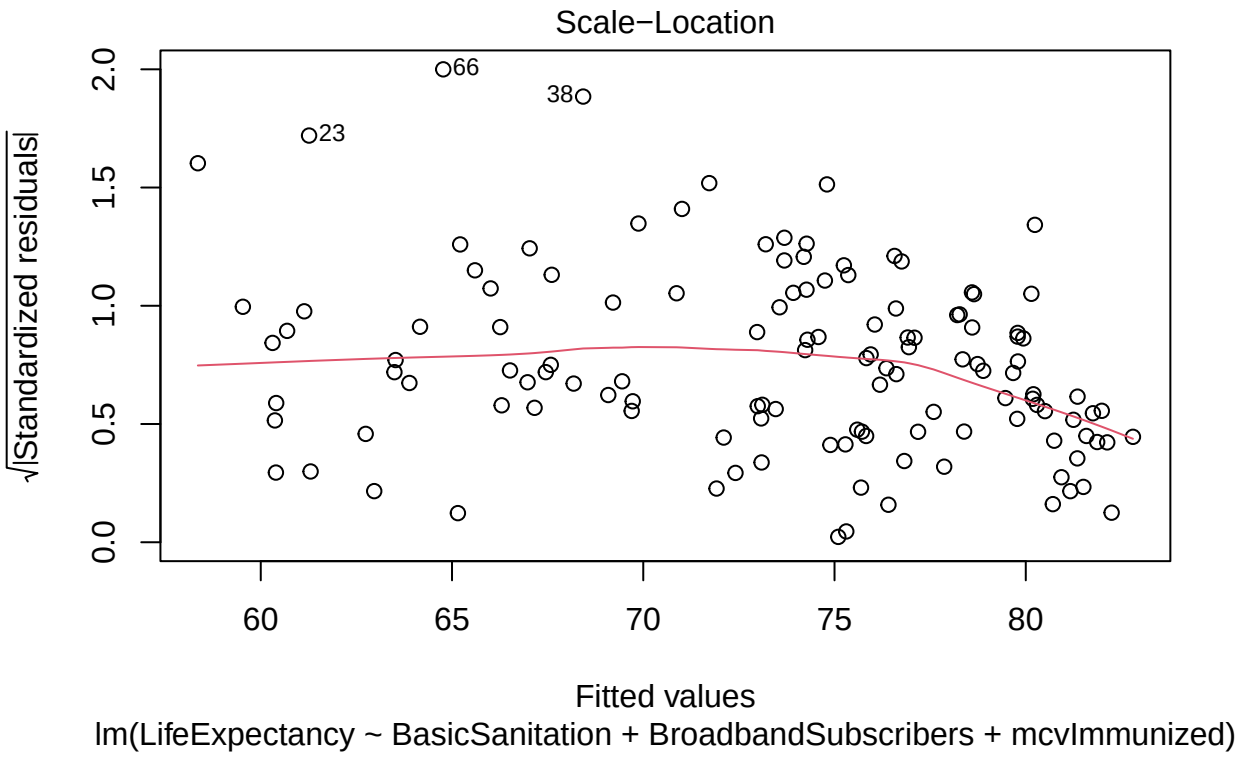
```
##
## Call:
## lm(formula = LifeExpectancy ~ BasicSanitation + BroadbandSubscribers +
##     mcvImmunized, data = Dataset1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2708  -1.5839   0.1841   2.0466   8.8449
```

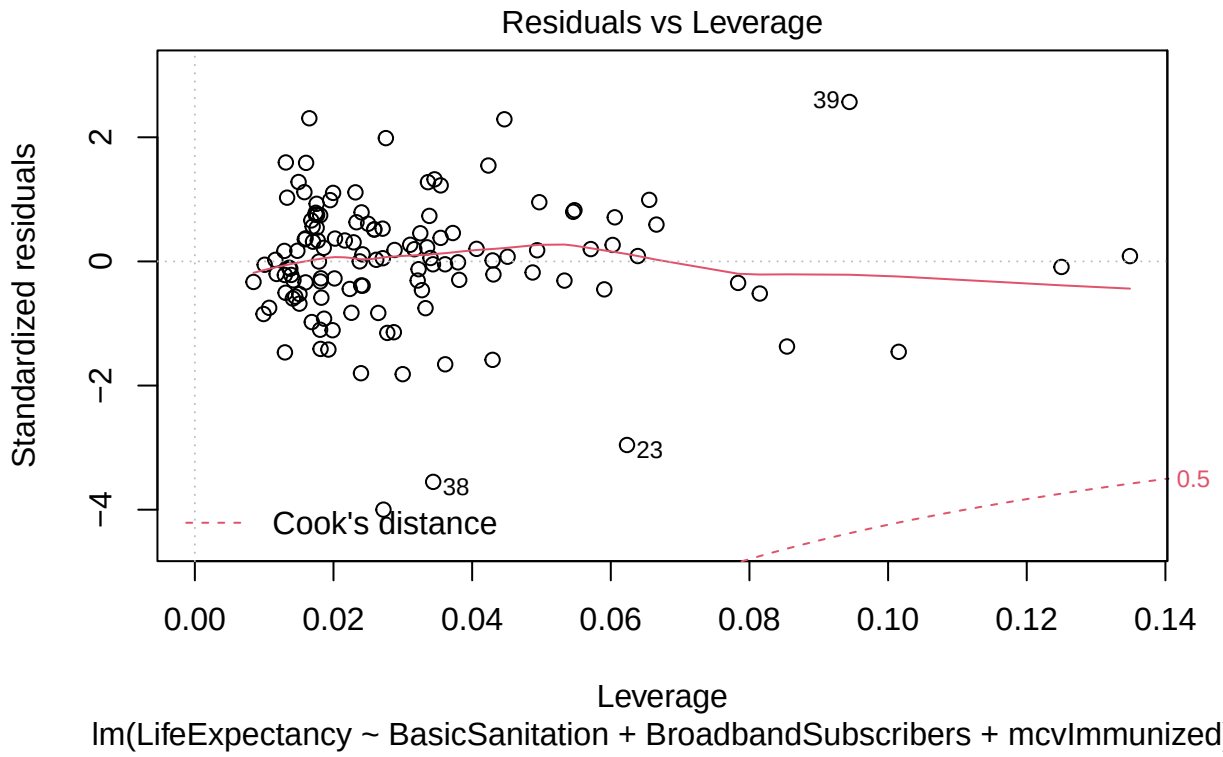
```
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           53.61804    2.85996  18.748  < 2e-16 ***
## BasicSanitation       14.57429    1.90221   7.662 4.71e-12 ***
## BroadbandSubscribers   0.19601    0.03237   6.056 1.57e-08 ***
## mcvImmunized           6.39365    3.81862   1.674   0.0966 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.618 on 123 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7636, Adjusted R-squared:  0.7579
## F-statistic: 132.5 on 3 and 123 DF,  p-value: < 2.2e-16
```

```
plot(final)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(LifeExpectancy ~ BasicSanitation + BroadbandSubscribers + mcvImmunized)
23
38
66


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
66
38
23
60
65
70
75
80
Fitted values
lm(LifeExpectancy ~ BasicSanitation + BroadbandSubscribers + mcvImmunized)




```
extractAIC(final, scale=MSE)
```

```
## [1] 4.000000 8.091641
```

```
anova(final)
```

```
## Analysis of Variance Table
##
## Response: LifeExpectancy
##                        Df Sum Sq Mean Sq  F value    Pr(>F)
## BasicSanitation         1 4703.8  4703.8 359.3490 < 2.2e-16 ***
## BroadbandSubscribers    1  460.8   460.8  35.2021  2.81e-08 ***
## mcvImmunized            1   36.7    36.7   2.8034   0.09661 .
## Residuals             123 1610.1    13.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Confidence Interval for final model

```
confint(final)
```

```
##                              2.5 %     97.5 %
```

```
## (Intercept)          47.9569320 59.2791546
## BasicSanitation      10.8089801 18.3396011
## BroadbandSubscribers  0.1319408  0.2600723
## mcvImmunized         -1.1650601 13.9523696
```