Project 1: regression 1 (fishermen - do not use the variable MeHg)

Urvan Christen, Amandine Goffeney, Joseph Vermeil, Lucile Vigué

March 17, 2019

Data description

Source: (Al-Majed and Preston 2000)

Description: Factors related to mercury levels among fishermen and a control group of non-fishermen.

Variables/names

- Fisherman indicator (*fisherman*)
- Age in years (*age*)
- Residence Time in years (*restime*)
- Height in cm (*height*)
- Weight in kg (*weight*)
- Fish meals per week (*fishmlwk*)
- Parts of fish consumed: 0=none, 1=muscle tissue only, 2=mt and sometimes whole fish, 3=whole fish (*fishpart*)
- Methyl Mercury in mg/g (*MeHg*)
- Total Mercury in mg/g (*TotHg*)

Imports and loading data

We have the following continuous variables: - Age in years (*age*) - Residence Time in years (*restime*) - Height in cm (*height*) - Weight in kg (*weight*) - Fish meals per week (*fishmlwk*) - Total Mercury in mg/g (*TotHg*)

We have the following categorical or boolean variables: - Fisherman indicator (*fisherman*) - Parts of fish consumed: 0=none, 1=muscle tissue only, 2=mt and sometimes whole fish, 3=whole fish (*fishpart*)

NB: pour *restime* et *fishmlwk* je ne sais pas trop si on les compte comme catégoriques ou pas.

Let's separate the table into the table of interest (the fisherman) vs the control table (the non fisherman).

We can now do some regression on both datasets and compare the results with some statistical tests?

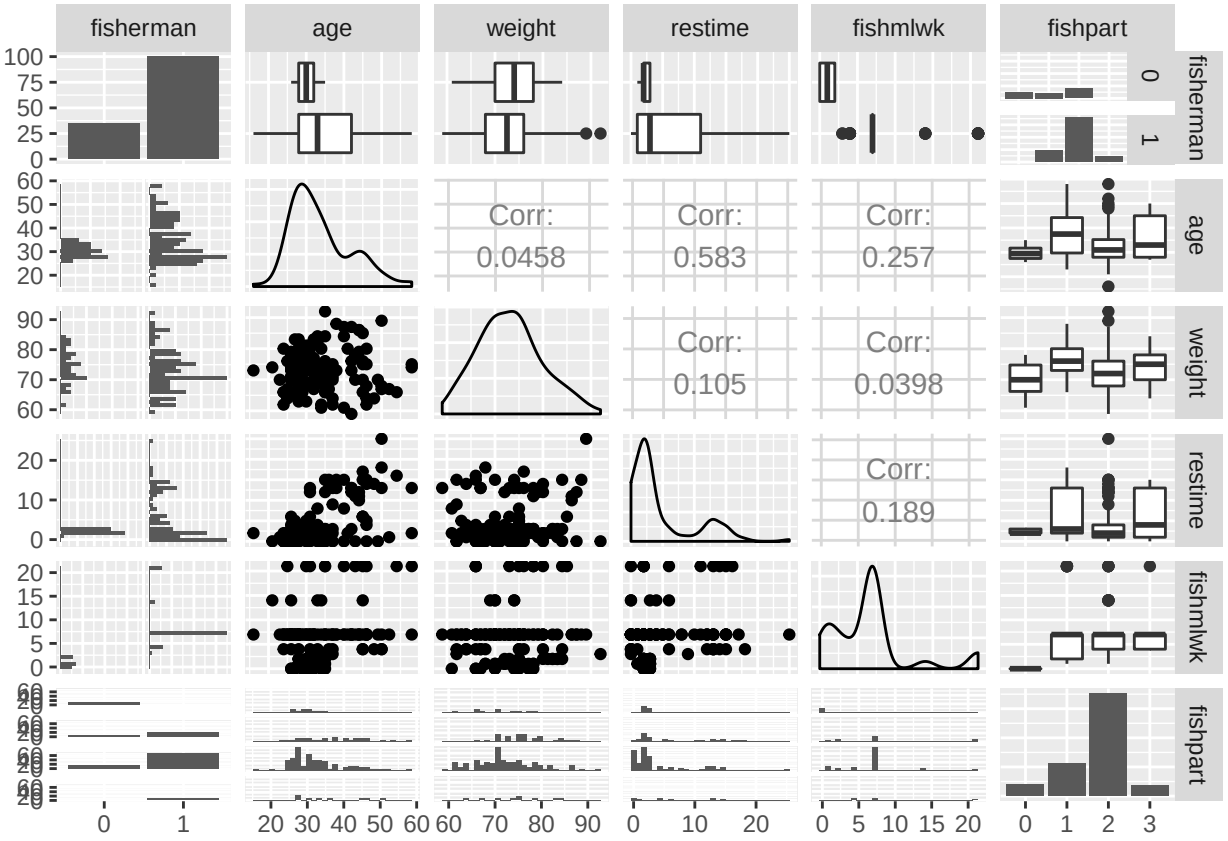
Exploratory analysis

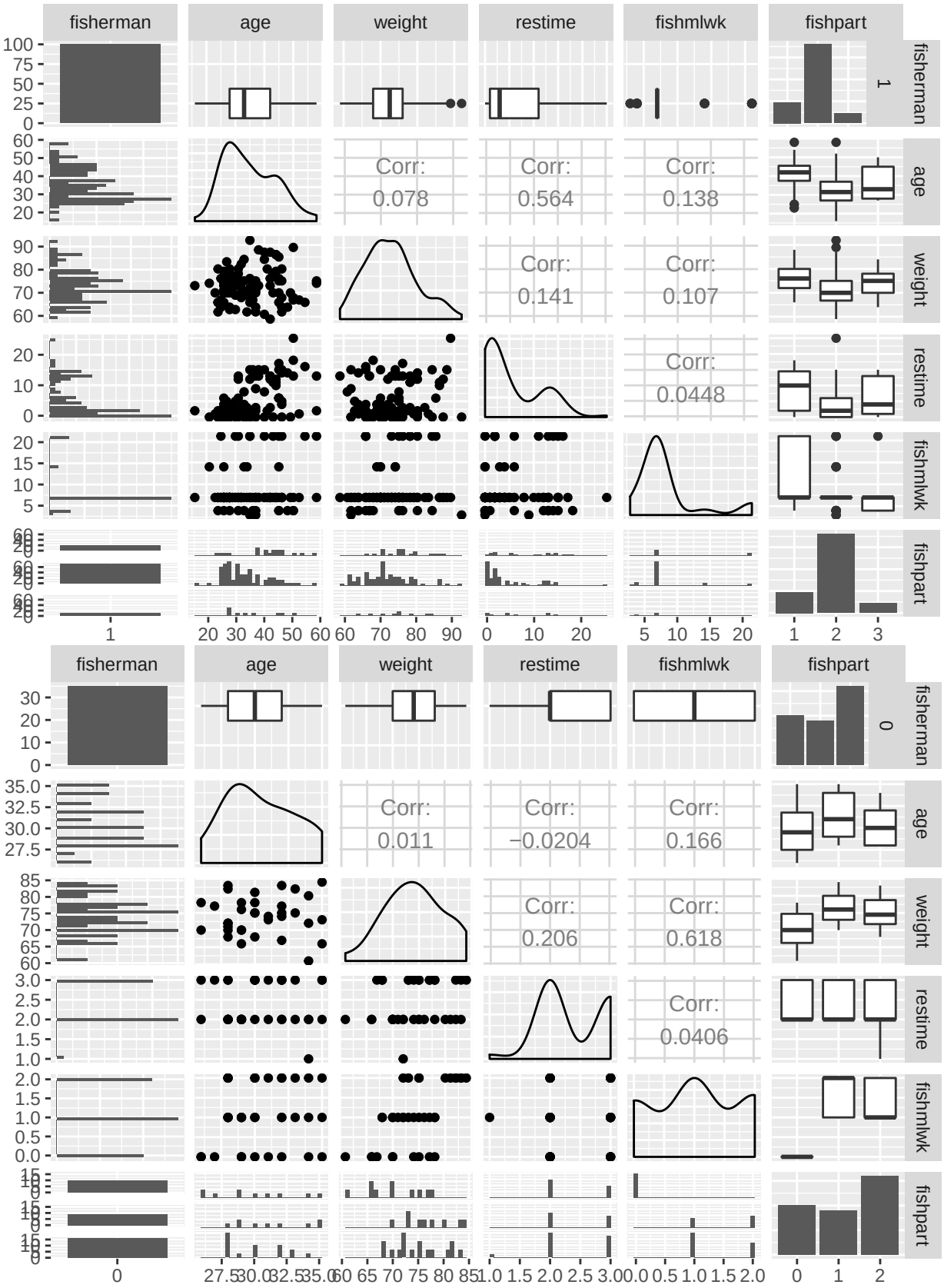
Summarize the dataset

```
##  fisherman      age            restime           height
##  0: 35     Min.   :16.00   Min.   : 0.000   Min.   :154.0
##  1:100     1st Qu.:28.00   1st Qu.: 2.000   1st Qu.:170.0
##            Median :32.00   Median : 2.000   Median :175.0
##            Mean   :33.76   Mean   : 4.593   Mean   :174.4
##            3rd Qu.:37.50   3rd Qu.: 6.000   3rd Qu.:180.0
##            Max.   :58.00   Max.   :25.000   Max.   :195.0
##      weight         fishmlwk      fishpart     TotHg
##  Min.   :59.00   Min.   : 0.000   0:10     Min.   : 0.025
##  1st Qu.:68.50   1st Qu.: 2.000   1:28     1st Qu.: 1.904
##  Median :73.00   Median : 7.000   2:88     Median : 3.006
##  Mean   :73.16   Mean   : 6.526   3: 9     Mean   : 3.775
##  3rd Qu.:77.00   3rd Qu.: 7.000            3rd Qu.: 4.688
##  Max.   :92.00   Max.   :21.000            Max.   :17.788
##     LogTotHg        logFishmlwk
##  Min.   :-3.6889   Min.   :0.000
##  1st Qu.: 0.6433   1st Qu.:1.099
##  Median : 1.1006   Median :2.079
##  Mean   : 1.0346   Mean   :1.755
##  3rd Qu.: 1.5450   3rd Qu.:2.079
##  Max.   : 2.8785   Max.   :3.091
```

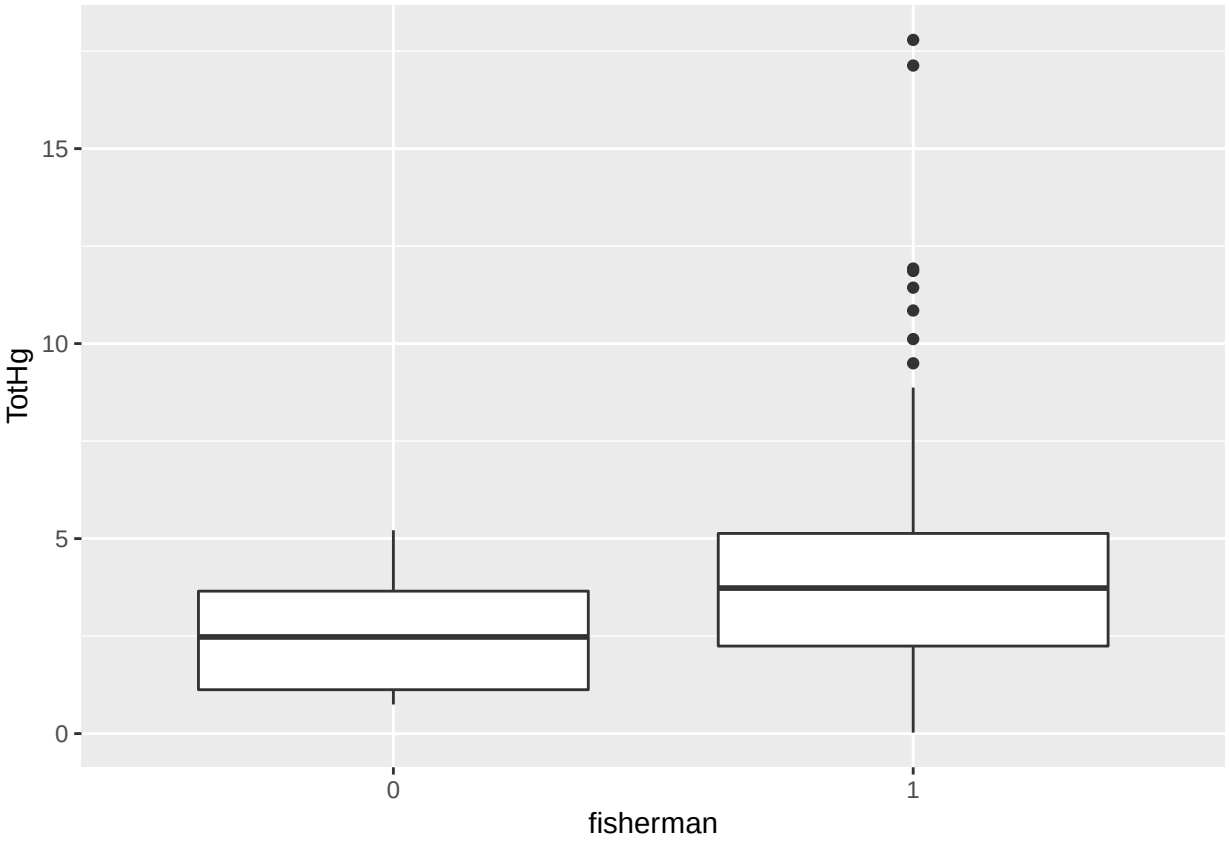
Plots

Pairwise behaviour of explnatory variables

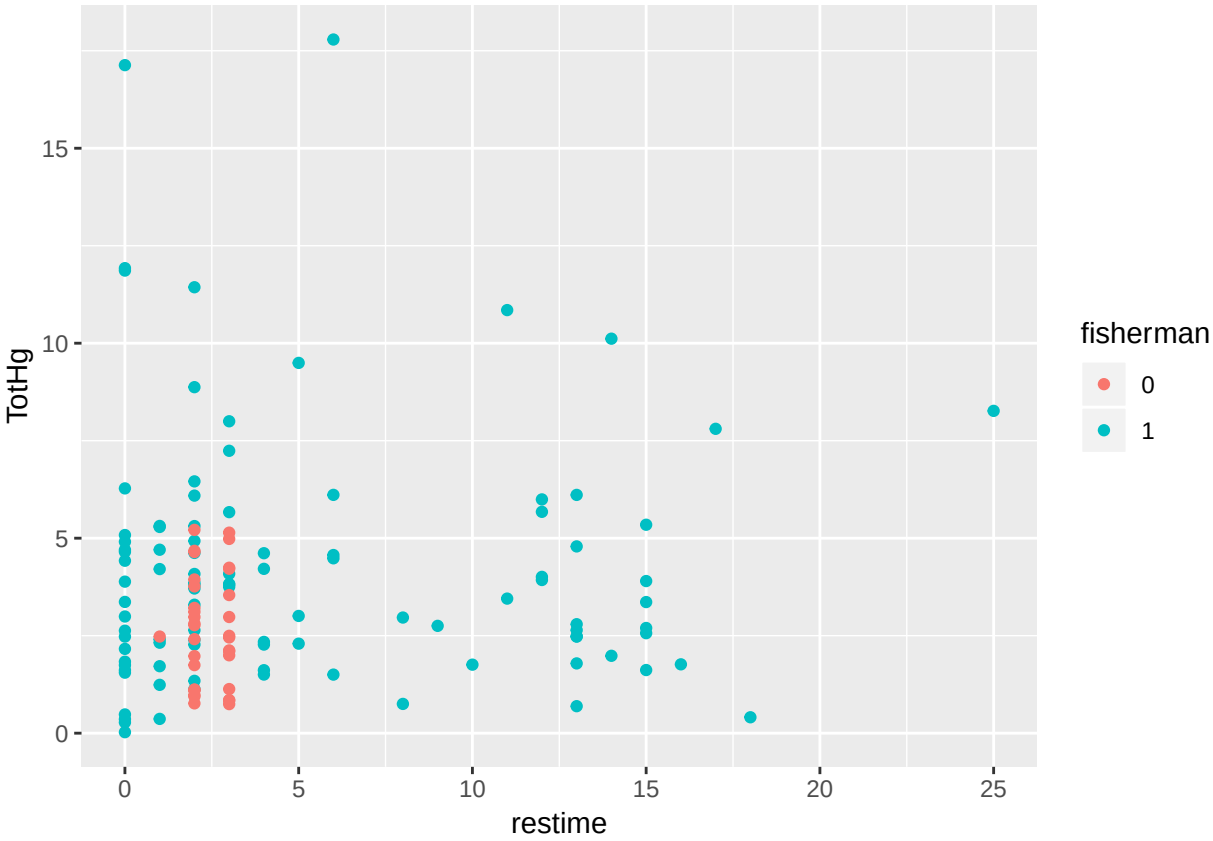


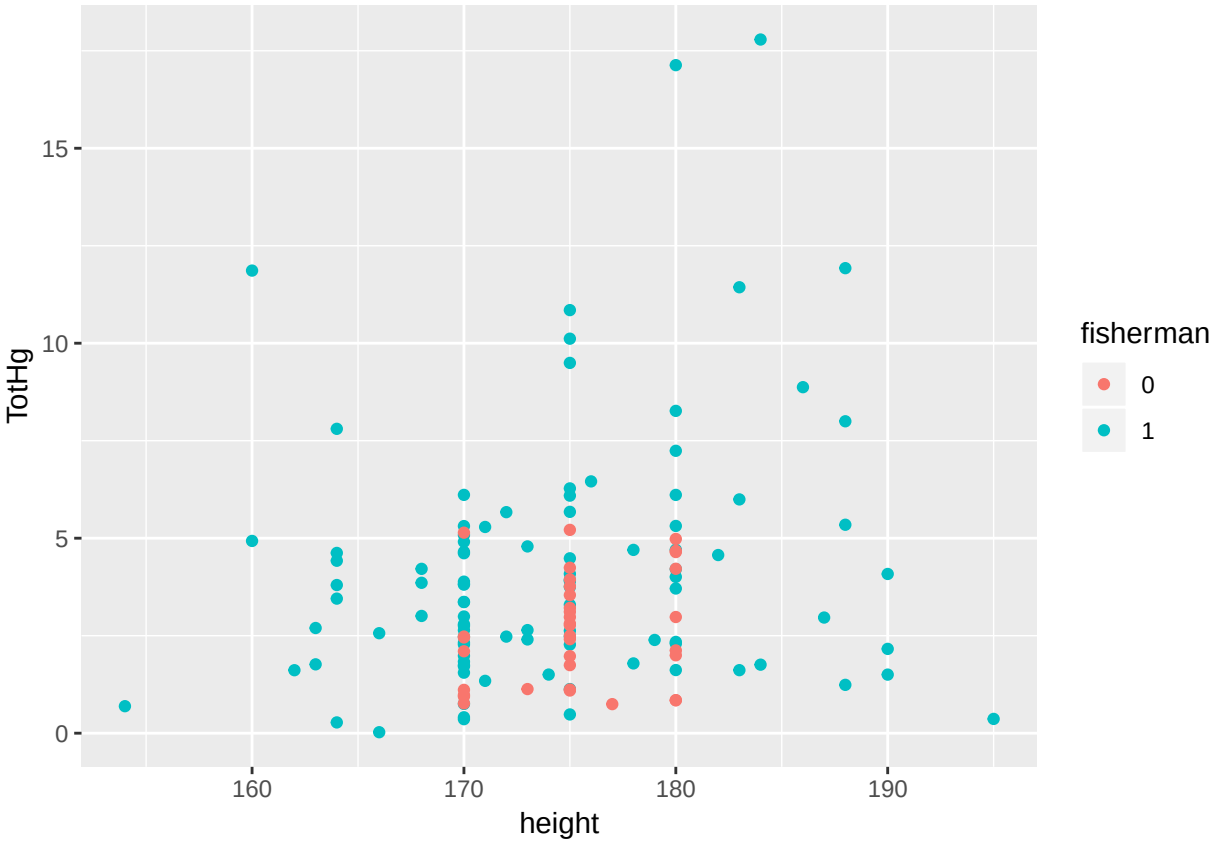

fisherman
age
weight
restime
fishmlwk
fishpart
1
0
Corr:
0.078
Corr:
0.564
Corr:
0.138
Corr:
0.141
Corr:
0.107
Corr:
0.0448
Corr:
0.011
Corr:
−0.0204
Corr:
0.166
Corr:
0.206
Corr:
0.618
Corr:
0.0406

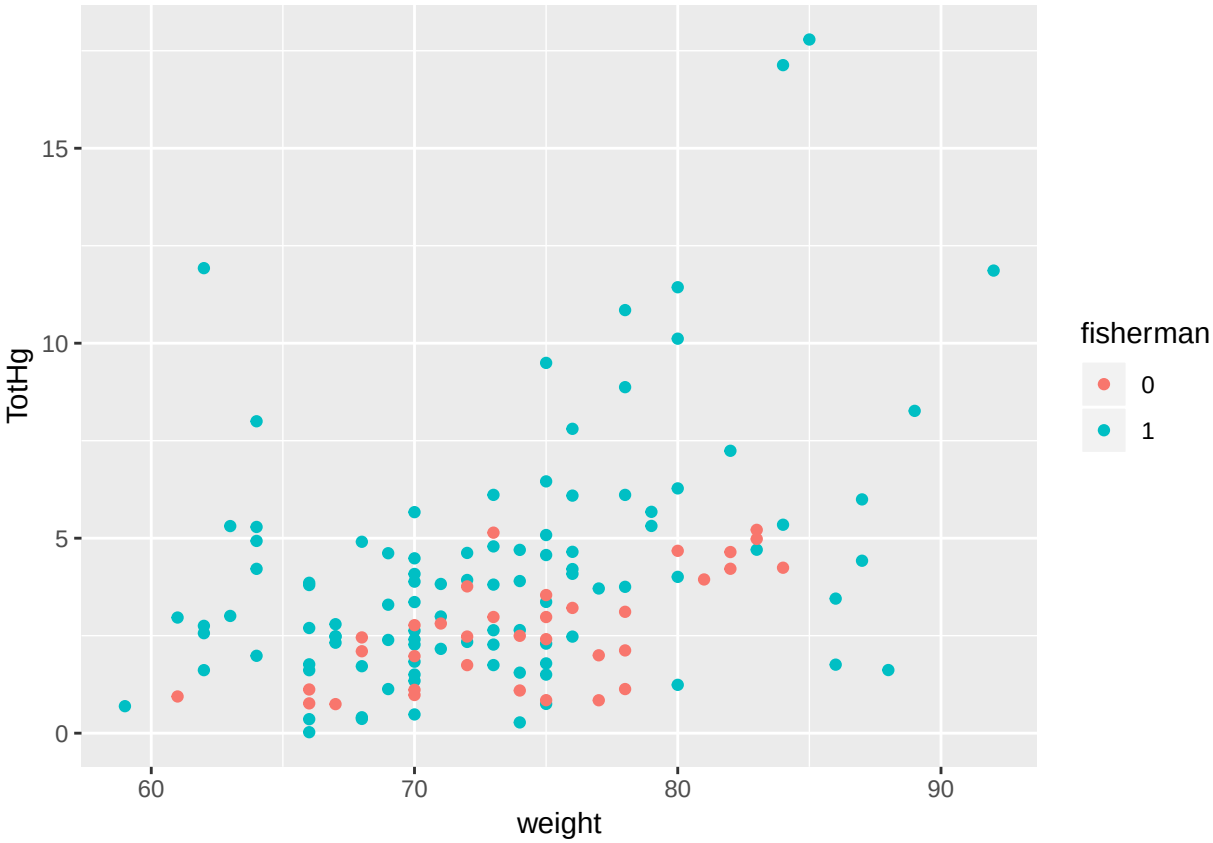
Plots of factored data

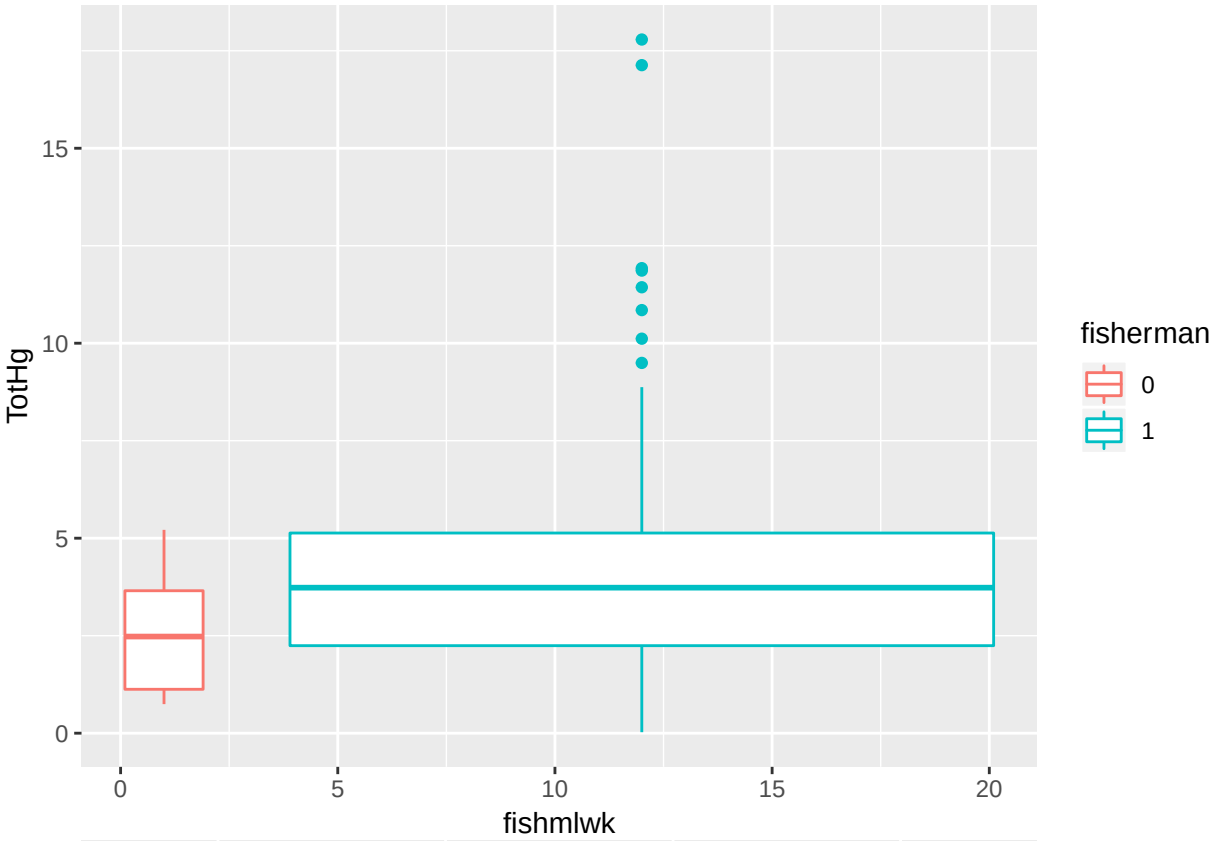


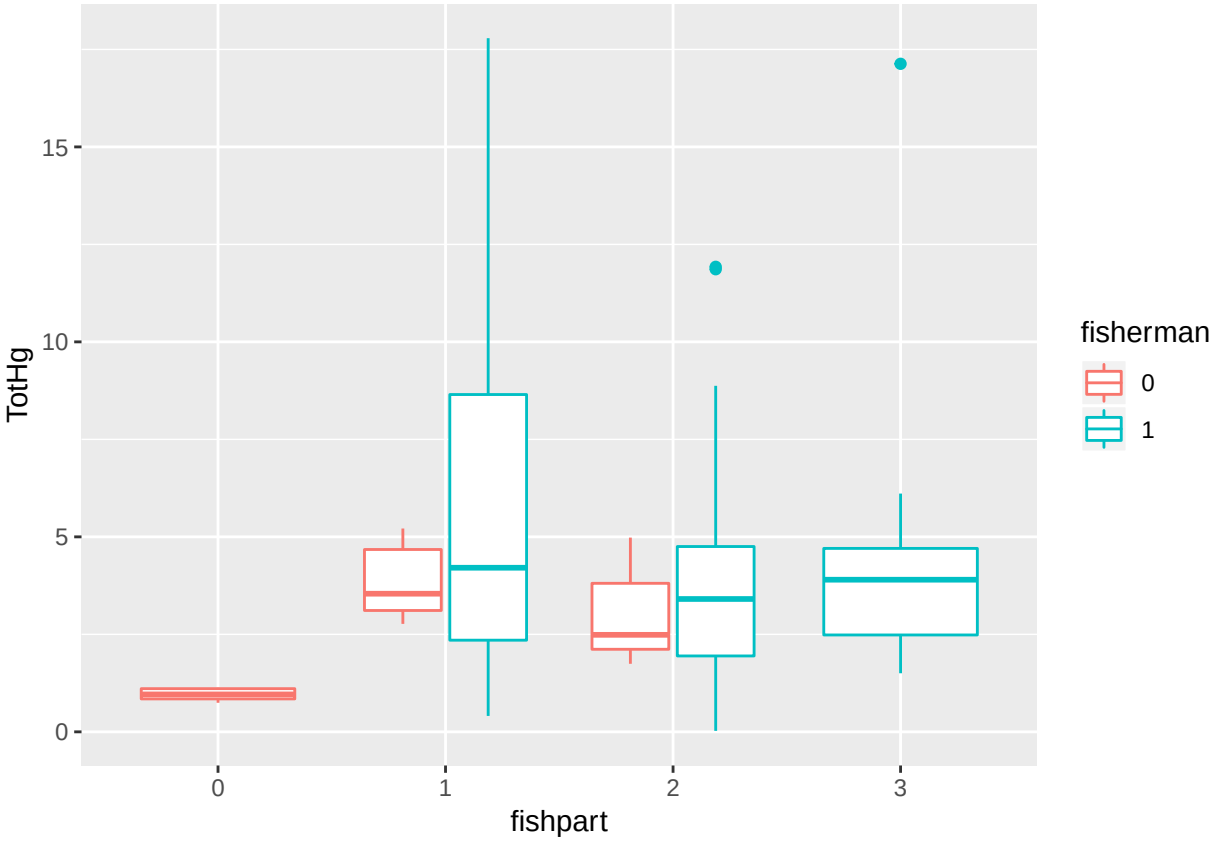

TotHg
15
10
5
0
20
30
40
50
60
age
fisherman
0
1


TotHg
15
10
5
0
0
5
10
15
20
25
restime
fisherman
0
1


TotHg
15
10
5
0
160
170
180
190
height
fisherman
0
1


TotHg
15
10
5
0
60
70
80
90
weight
fisherman
0
1


TotHg
15
10
5
0
0
5
10
15
20
fishmlwk
fisherman
0
1

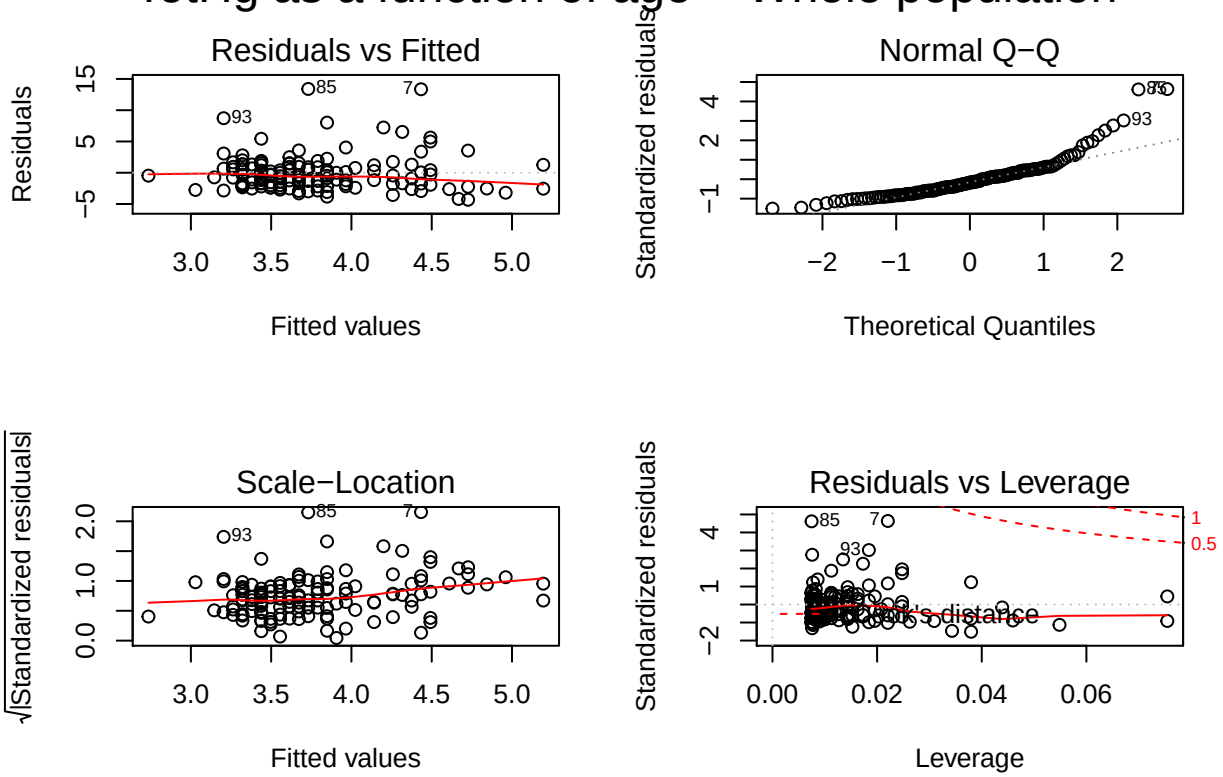

TotHg
15
10
5
0
0
1
2
3
fishpart
fisherman
0
1

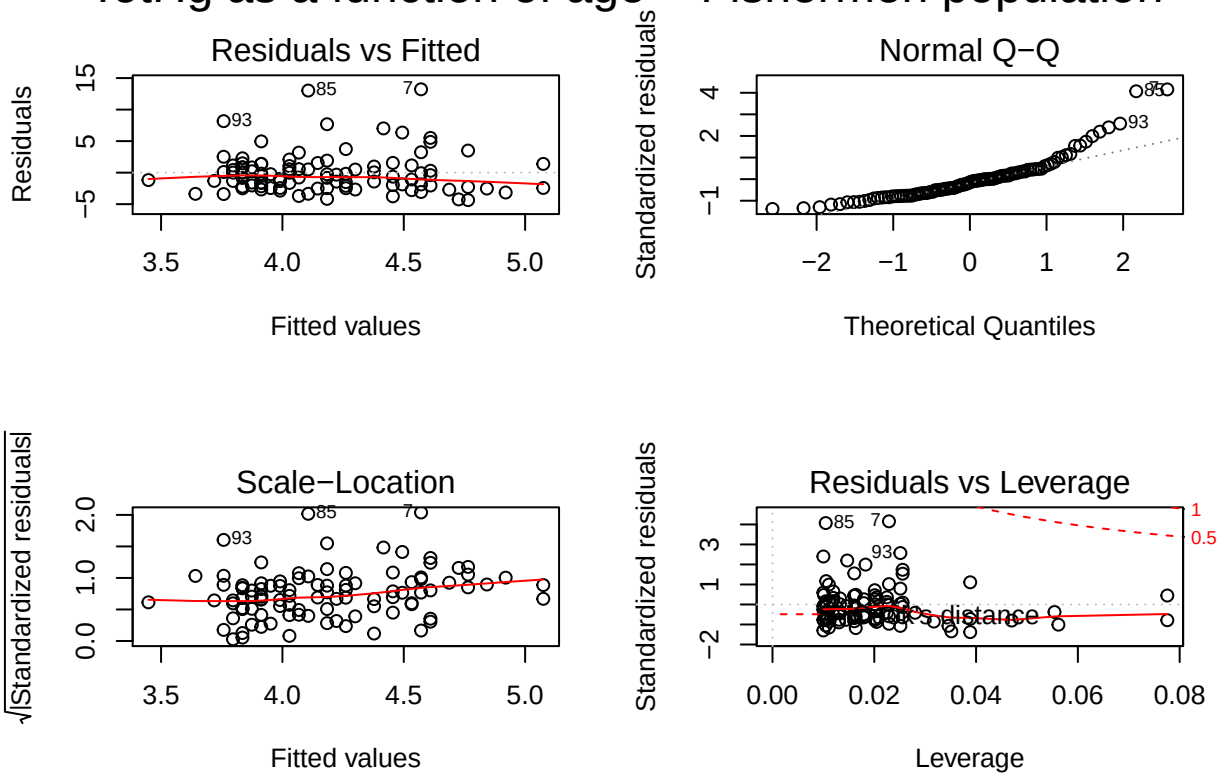
Analyze

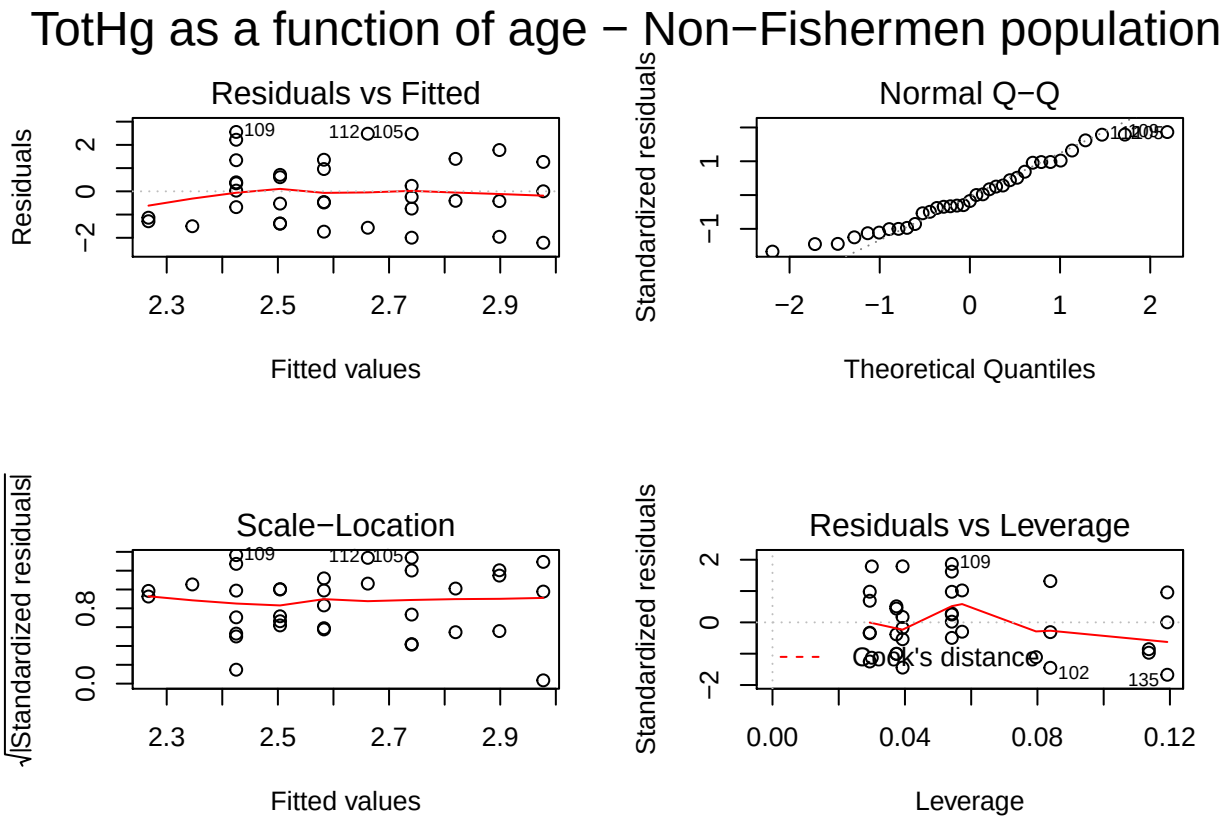
1-D linear models

Age

Code


TotHg as a function of age − Whole population
Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


TotHg as a function of age − Fishermen population
Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage



Comments:

- Residuals seem all in all well distributed in all three populations suggesting that the distribution of TotHg as a function of age is homoscedastic.
- Normal Q-Q fits a line for Non-Fishermen population reinforcing the assumption that the distribution of TotHg given an age follows a normal distribution.
- Normal Q-Q suggest however a heavy right tail for Fishermen population on this same distribution

Test of homoskedasticity with a Breusch-Pagan test (Joseph)

- This is a test for homoskedasticity of the data. The null hypothesis is homoskedasticity, and ncvTest calculates a p-value.

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 10.93327, Df = 1, p = 0.00094453
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 4.561776, Df = 1, p = 0.032693
```

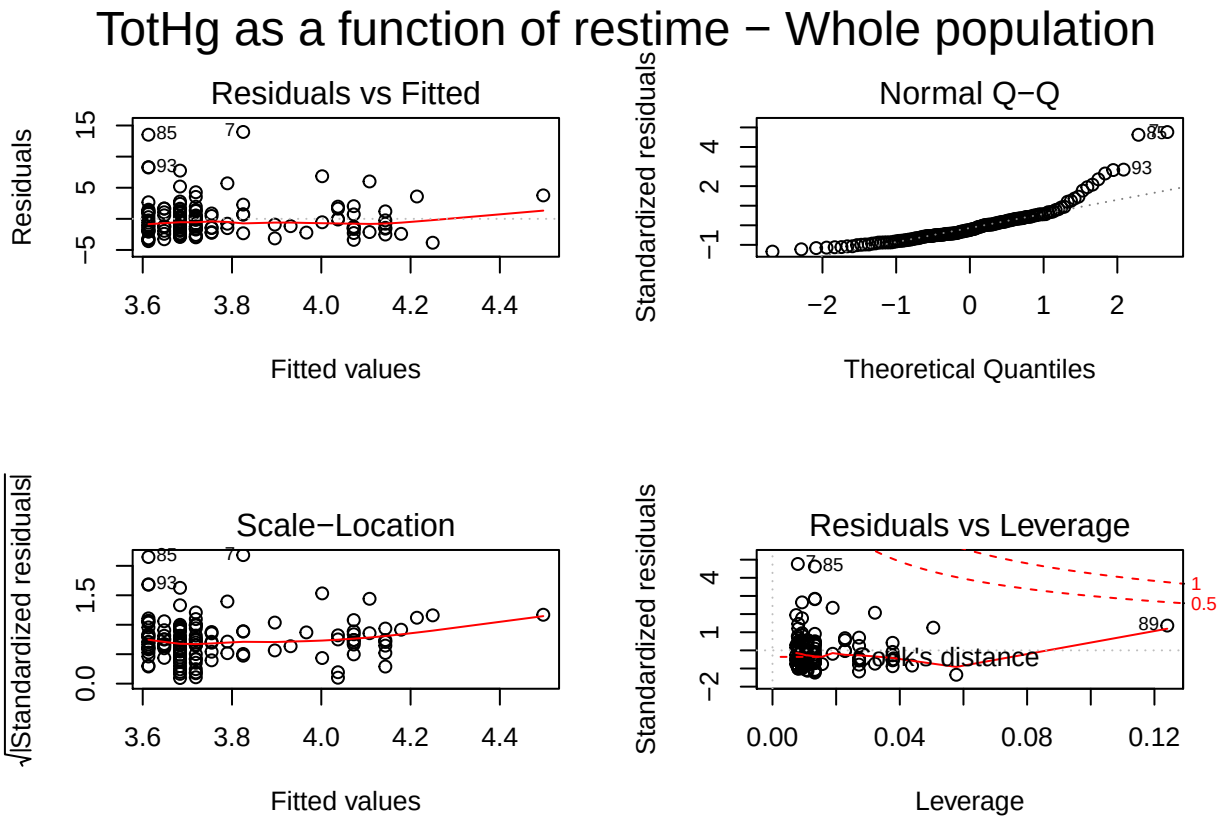
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
```

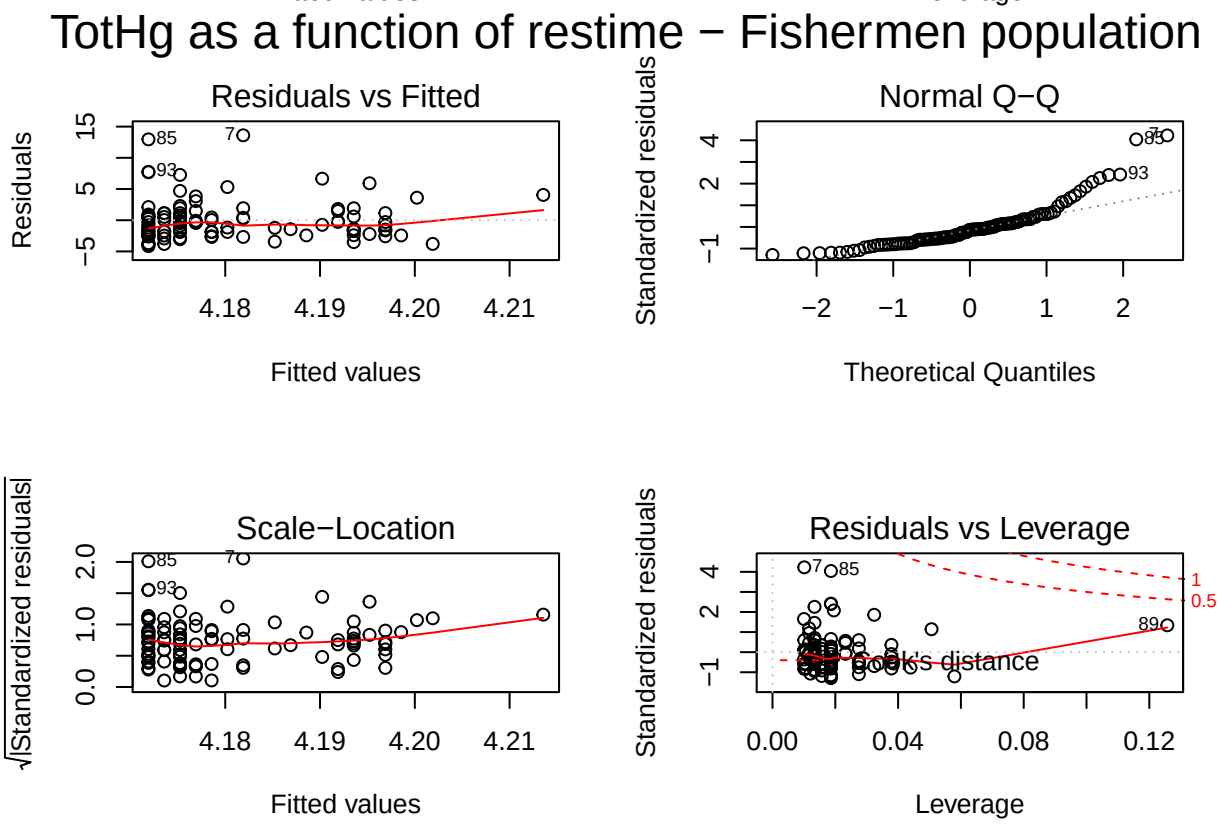
```
## Chisquare = 0.2004155, Df = 1, p = 0.65439
```

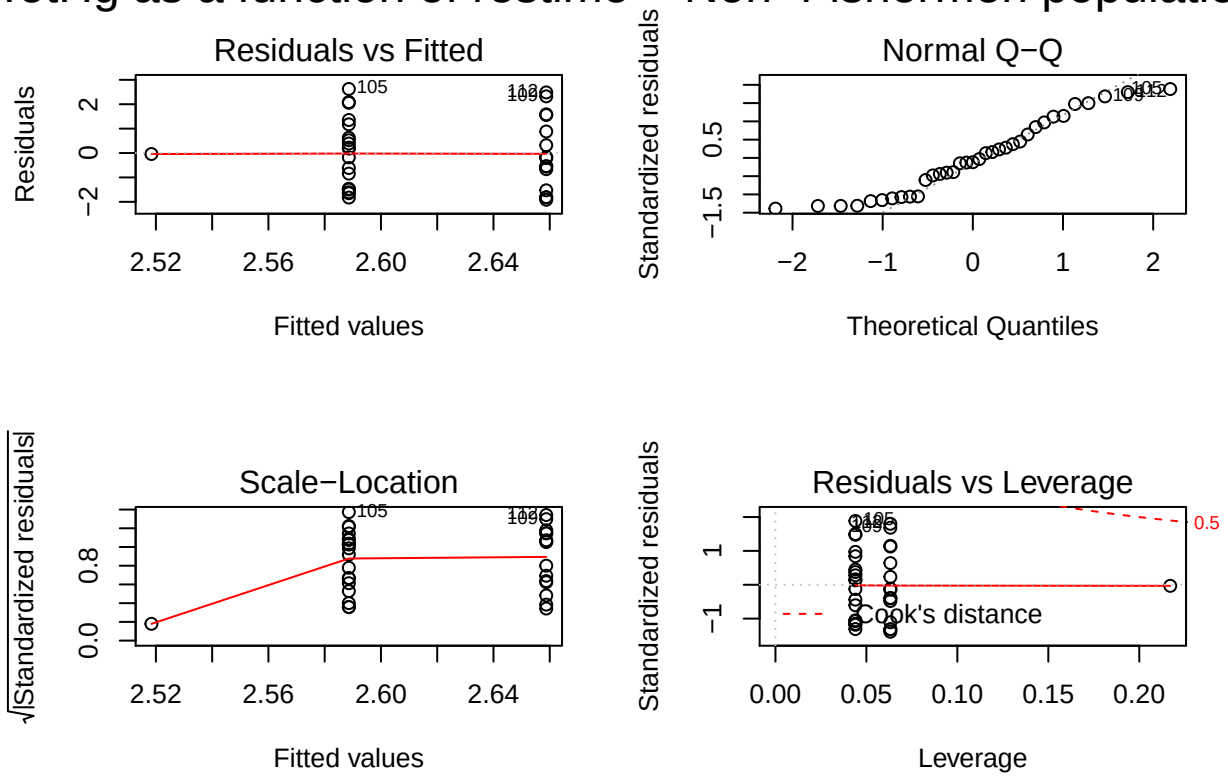
These results suggest that only the non-Fisherman pop has homoskedastic HgTot vs. Age values.

Restime

Code


TotHg as a function of restime − Whole population
Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


TotHg as a function of restime − Fishermen population
Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage



Comments

- There is a great difference of the distribution of *restime* between Fishermen and control population. In the control population it takes a complete range of values whereas in the control population, the distribution is discrete and takes only 3 (even 2) different values. Thus it might be difficult to draw conclusions on whether *restime* is correlated or not with the *TotHg* value.
- There also is this problem in Fishermen population of the right long tail and the left short tail for *TotHg* distribution. **It might be useful to use a log scale.**
- Residuals in Fishermen population for fitted *TotHg* as a function of *restime* are all in all well distributed around 0 for all values of *restime* suggesting an homoscedasticity of the distribution of *TotHg* according to *restime*.
- However, there are some residuals with very high positive values when there are none with very "high" negative values, suggesting some possible bias in the distribution.

Test of homoskedasticity with a Breusch-Pagan test (Joseph)

- This is a test for homoskedasticity of the data. The null hypothesis is homoskedasticity, and ncvTest calculates a p-value.

```
## Non-constant Variance Score Test
```

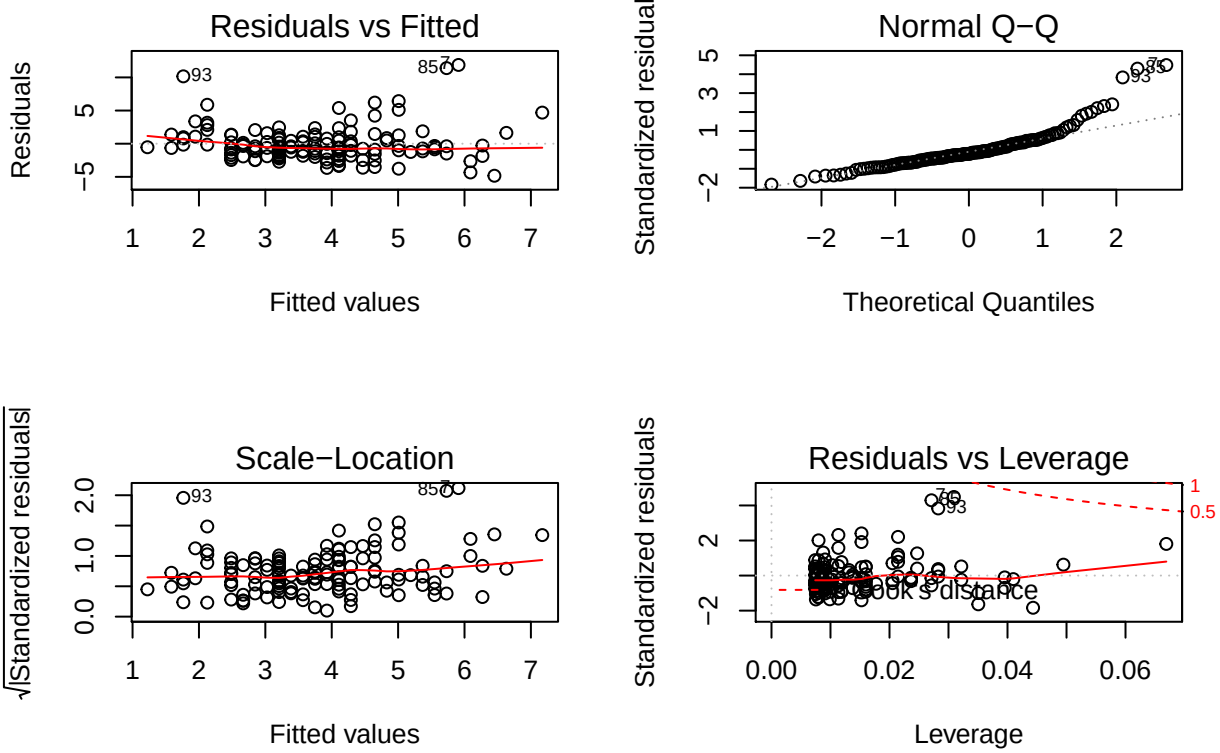
```
## Variance formula: ~ fitted.values
## Chisquare = 0.2252327, Df = 1, p = 0.63508

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.9487475, Df = 1, p = 0.33004

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.2388775, Df = 1, p = 0.62502
```

This time those results suggest that all 3 pop have homoskedastic HgTot vs. restime values.

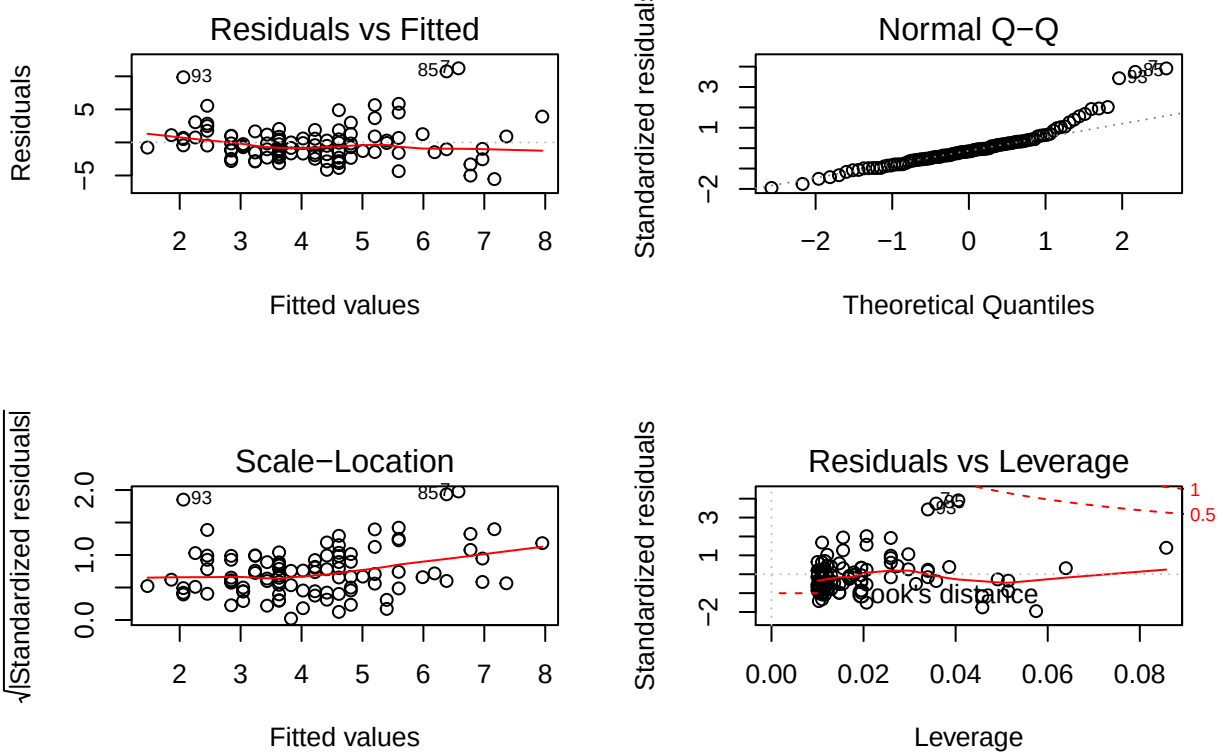
Weight

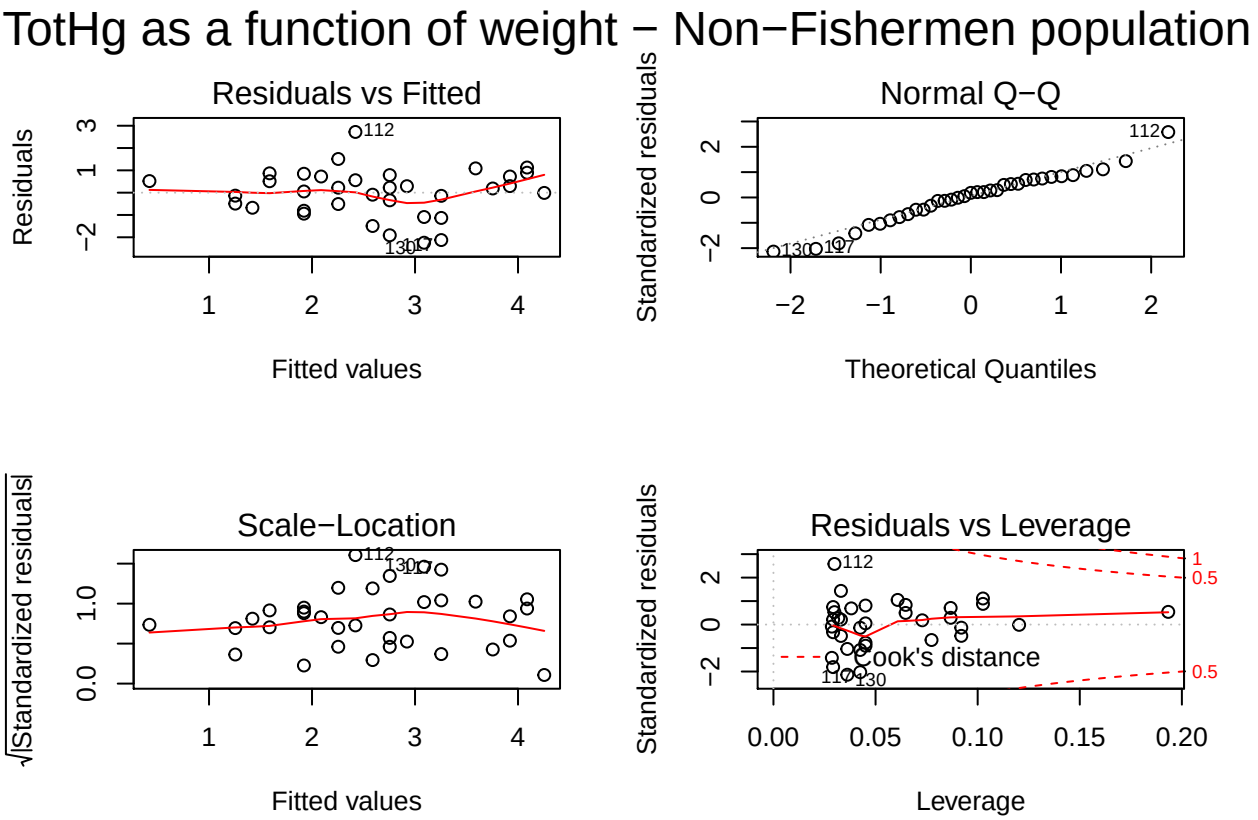
Code

TotHg as a function of weight − Whole population



TotHg as a function of weight − Fishermen population



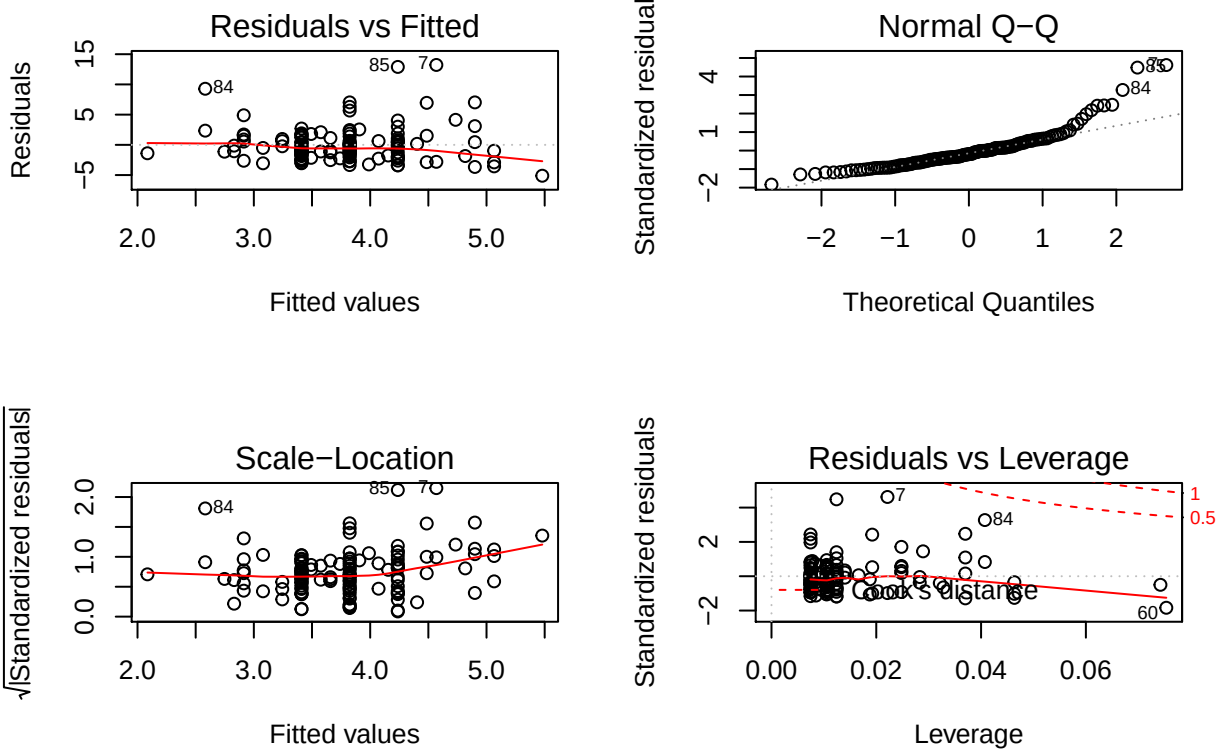


Comments

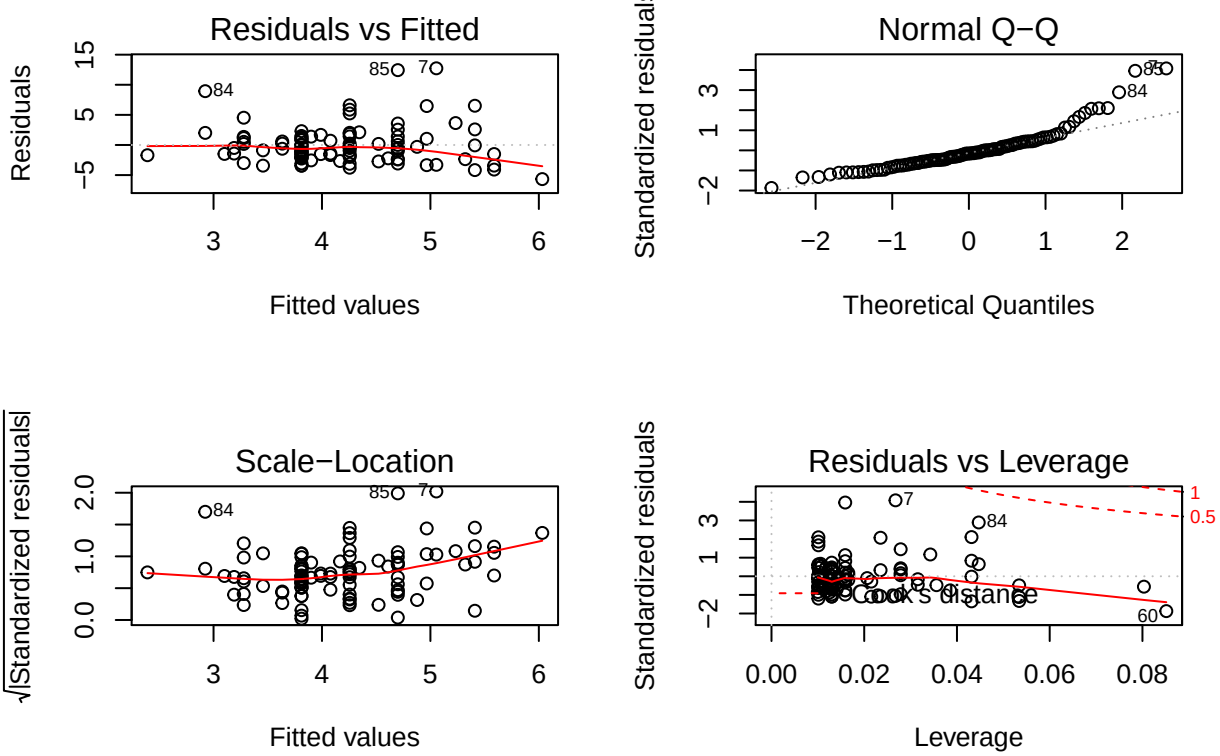
Height

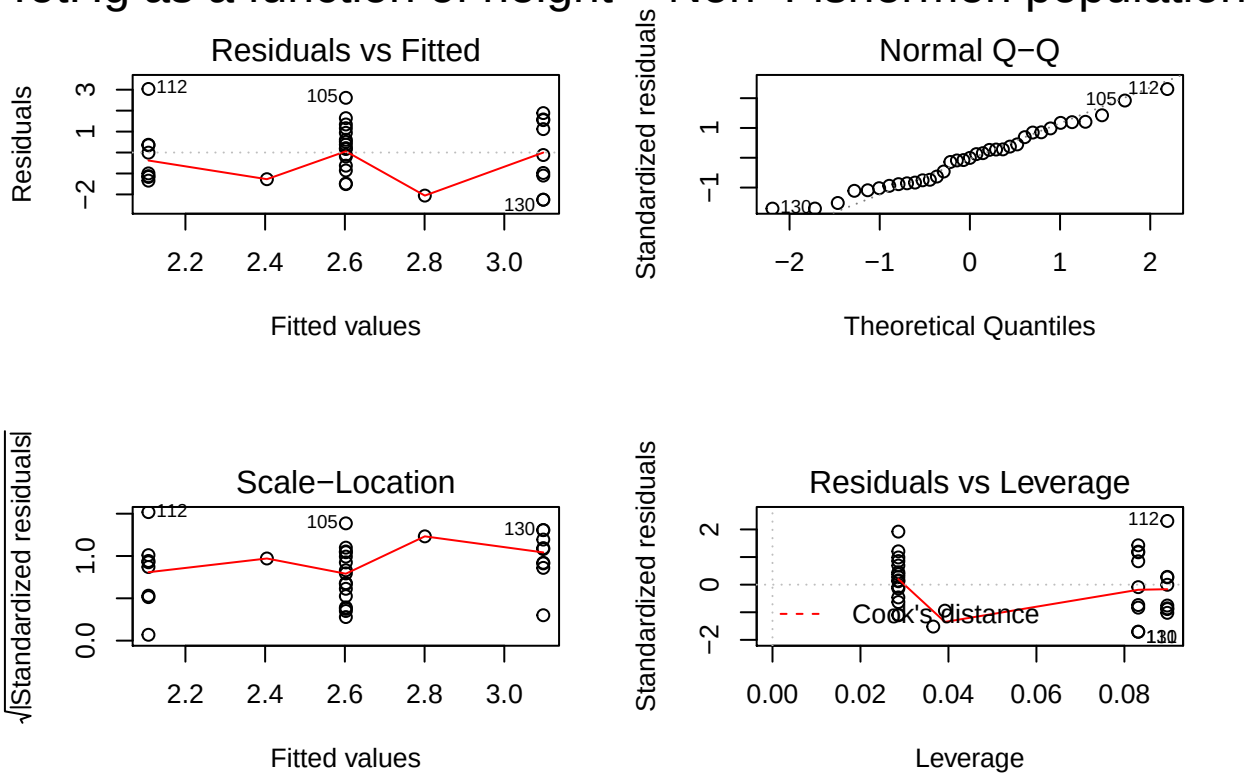
Code

TotHg as a function of height − Whole population



TotHg as a function of height − Fishermen population



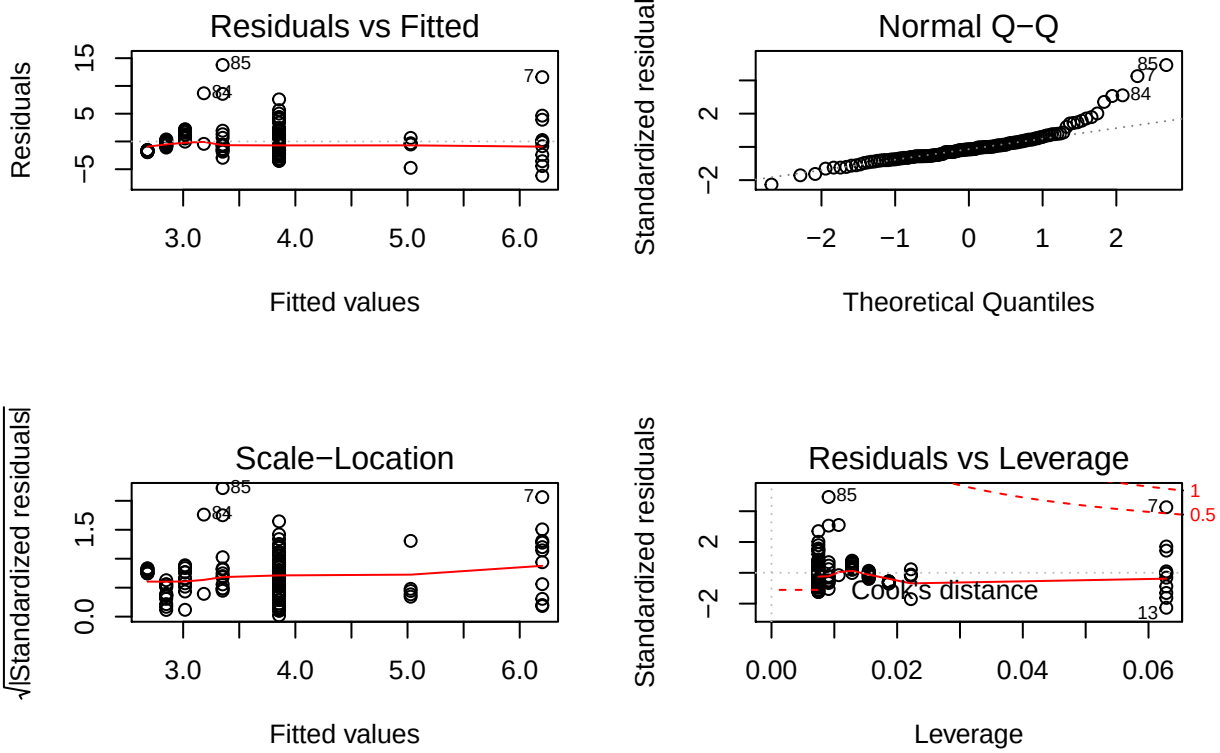


Comments

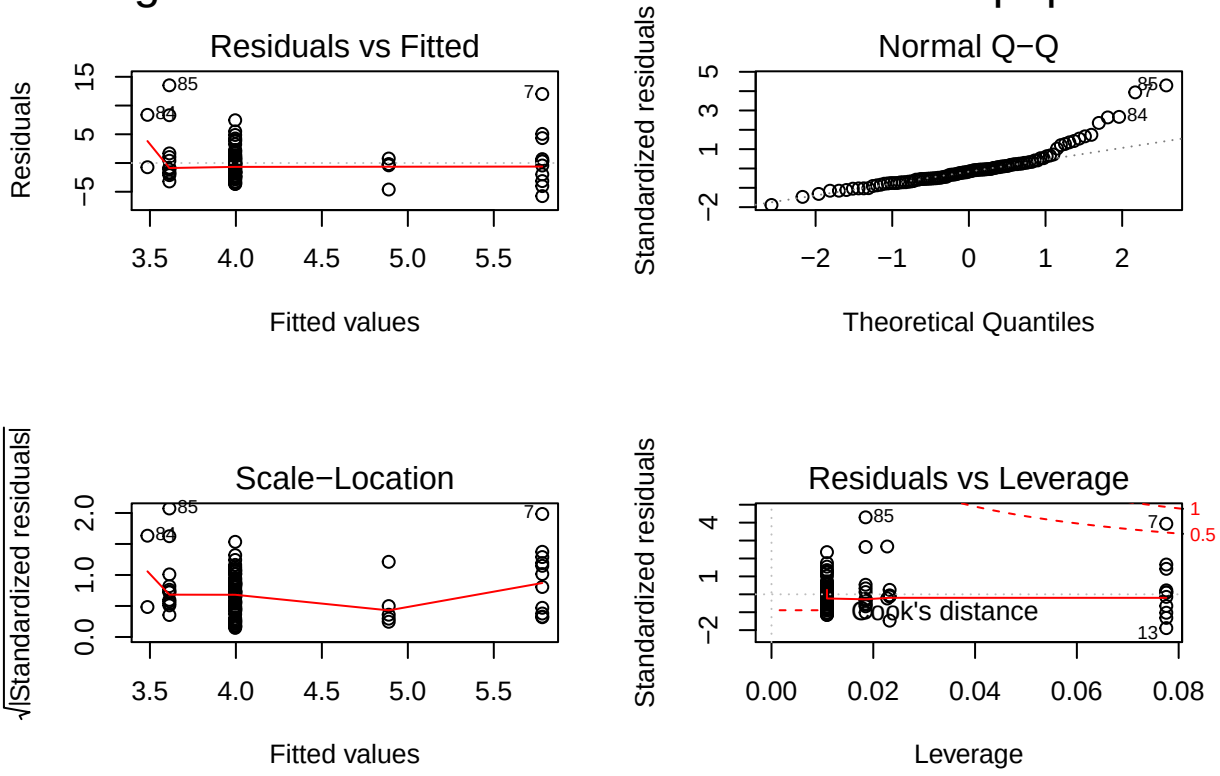
Fish meal per week

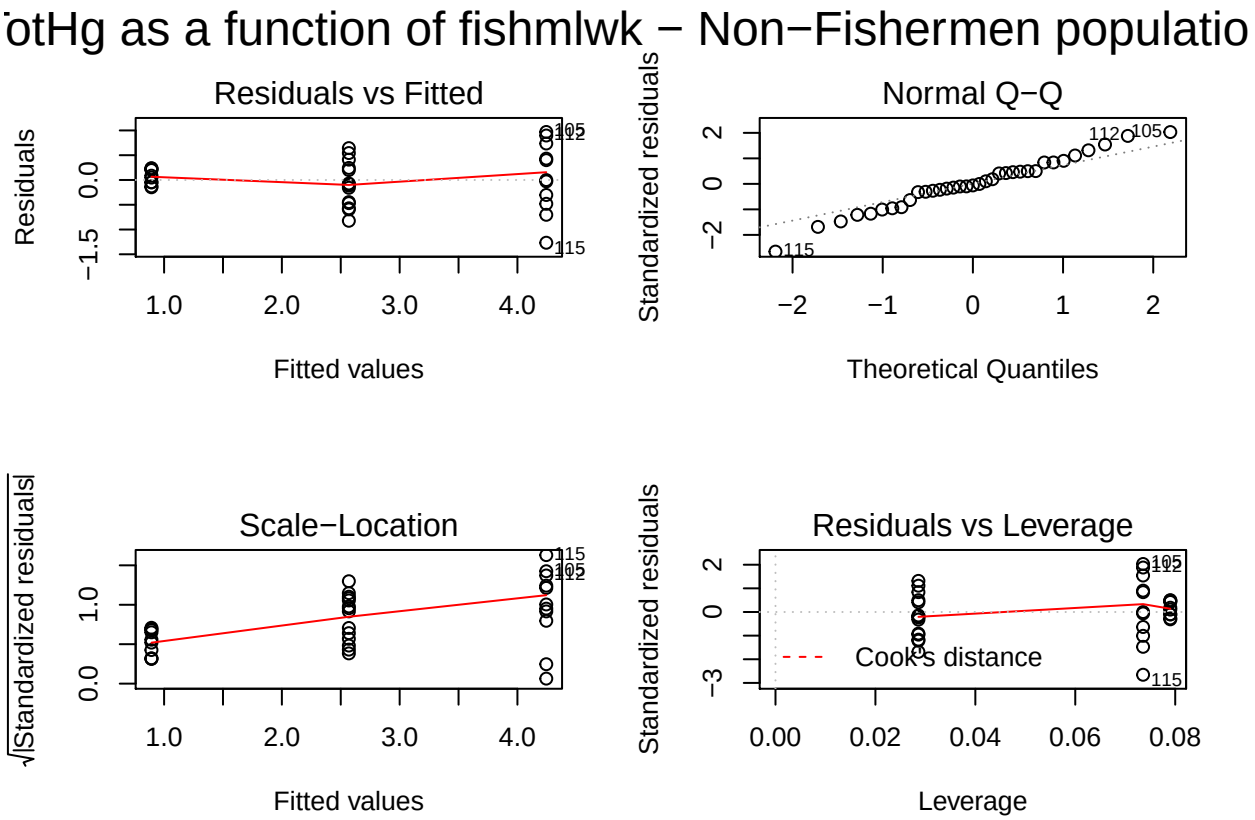
Code

TotHg as a function of fishmlwk − Whole population



TotHg as a function of fishmlwk − Fishermen population





Comments

Two groups significantly different

The fishermen have higher levels of mercury in their hair.

Test of the difference between fishermen and non-fishermen

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## fisherman     1   63.4   63.43   7.714 0.00627 **
## Residuals   133 1093.7    8.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is indeed a significant difference between these two groups. What are the differences between the two populations that can explain such observations?

Comments:

- The population of non-fishermen is between 25 and 35 years while the population of fishermen is between 15 and 60 years. As there seem to be a little correlation between age and mercury levels this could affect our other results.

- The variable restime seems difficult to interpret (poor correlation with Hg levels, narrow range of values for non-fishermen).
- The height indicator for non-fishermen is not very precise (it takes only 3 different values: 170cm, 175cm, 180cm).
- There seems to be a correlation between weight and mercury levels. Should we study mercury levels per kg instead? (that may be not very relevant because the mercury levels are the mercury levels in the hair so there have no reason to be linearly correlated with the total mass of the body)
- There is a clear difference in way of life between fishermen and non-fishermen: the first ones eat fish more often than the second. We have to be very careful in interpreting our results because any correlation found between high levels of fish consumption and mercury can reflect the correlation between being a fisherman and having high levels of mercury without meaning that it is fish consumption that causes high levels of Hg. However, among the non-fishermen population there seems to be a clear trend between fish consumption and Hg mercury.
- No clear trend between fishpart and Hg levels, maybe we need to put in relation fishpart and fish consumption.
- Unbalanced design (more fishermen than controls)

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## fishpart      3  149.1   49.71   6.461 0.000411 ***
## Residuals   131 1007.9    7.69
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Possible analysis

- Add a correlation coefficient to the scatterplots
- Check homoscedasticity
- Fit a linear model to the data
- Model selection:
- Compare models using F-tests, AIC, BIC
- If the number of variables is small enough, could compare all possible models. Usually this is not practical, use automatic procedures: forward selection, backward elimination, stepwise selection
- Adjusted R^2, ANOVA
- Look for influential points (studentized residuals, Cook's distance)

- Other diagnostic plots: residuals against predicted values, normal QQ-plot, scale location, residual vs leverage

Plan

- VIF to check for multicolinearity between variables + choose which we want to keep
- stepwise selection on the model with interactions
- fit the model, with the whole population, the fishermen, and the non fishermen
- diagnostic plots
- eventually robust regression
- conclude for the values of the parameters + some nice plots

Selection of the model

```
##                          GVIF Df GVIF^(1/(2*Df))
## age              2.351239e+03  1       48.489574
## restime          5.596218e+03  1       74.807877
## height           2.482155e+02  1       15.754855
## weight           8.393475e+02  1       28.971495
## fishpart         5.277202e+09  3       41.725411
## age:restime      4.950791e+01  1        7.036186
## age:height       2.340909e+03  1       48.382946
## age:weight       5.747451e+02  1       23.973841
## age:fishpart     8.447199e+05  3        9.722668
## restime:height   1.898520e+03  1       43.572007
## restime:weight   3.103624e+02  1       17.617106
## restime:fishpart 5.298681e+04  3        6.128600
## height:weight    8.693067e+02  1       29.484008
## height:fishpart  1.214946e+10  3       47.946785
## weight:fishpart  7.634263e+07  3       20.596552
```

- The Variance Inflation Factors tends to show there is no case of too high colinearity here. (BUT problem with the fact we have categorical var vs. continuous var ?)

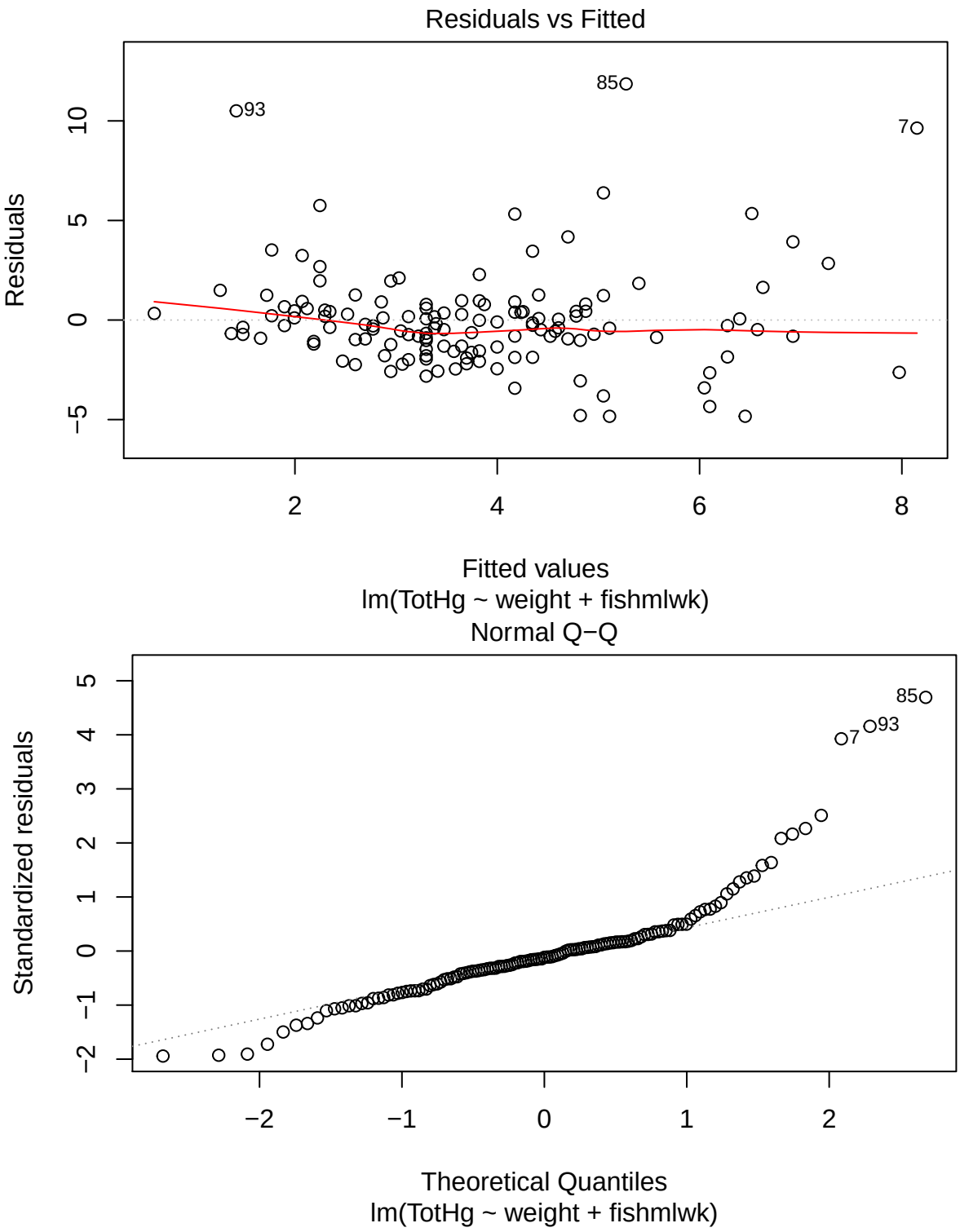
Whole population, non-squared model

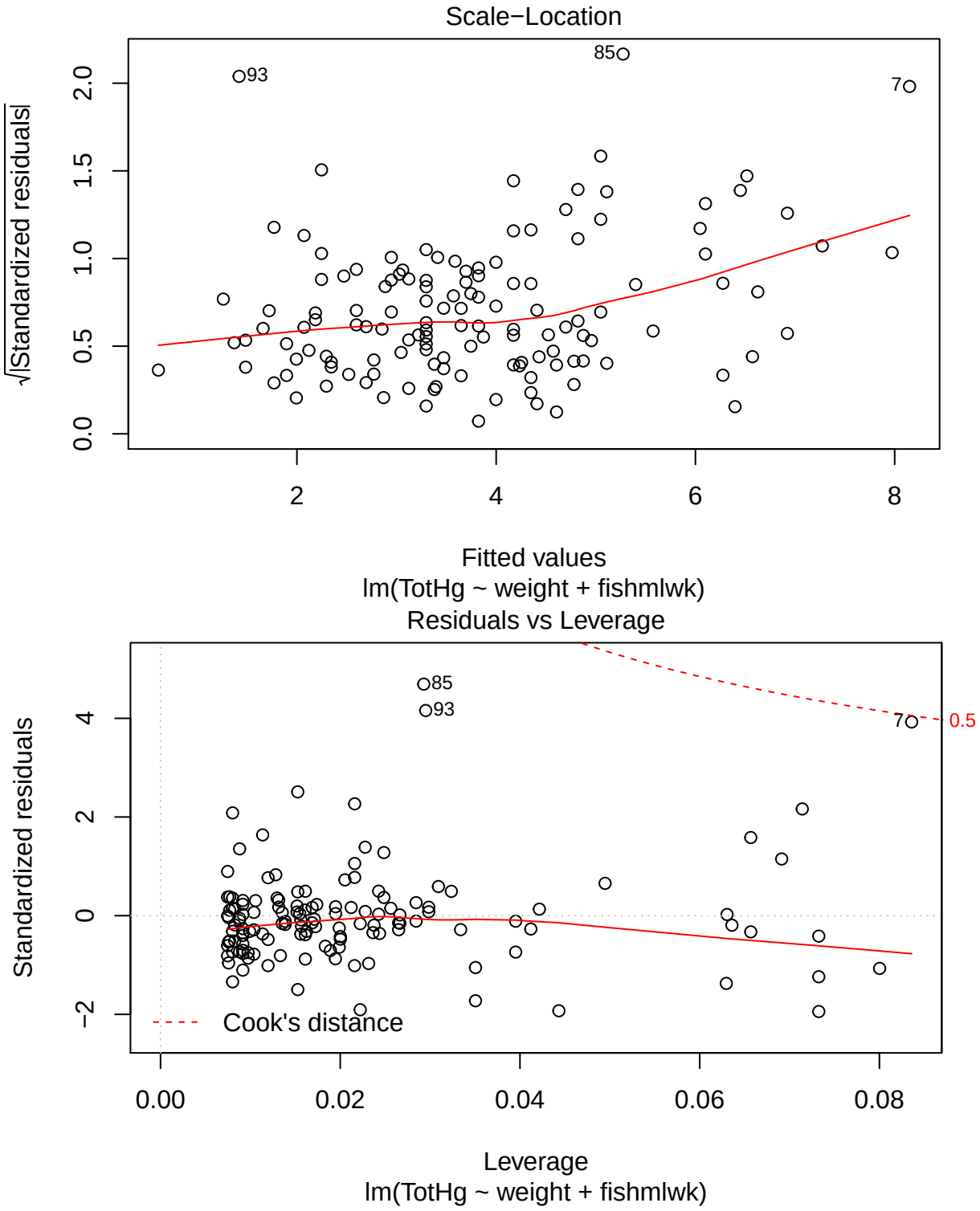
```
## Start:  AIC=260.18
## TotHg ~ age + restime + height + weight + fishmlwk + fishpart
##
##            Df Sum of Sq    RSS    AIC
## - height    1     4.793 816.53 258.97
```

```
## - age       1     8.814 820.55 259.63
## - fishpart  3    33.660 845.39 259.66
## - restime   1    10.460 822.19 259.90
## <none>                  811.73 260.18
## - fishmlwk  1    46.822 858.55 265.75
## - weight    1   115.001 926.73 276.06
##
## Step:  AIC=258.97
## TotHg ~ age + restime + weight + fishmlwk + fishpart
##
##            Df Sum of Sq    RSS    AIC
## - age       1     9.289 825.81 258.50
## - fishpart  3    36.427 852.95 258.86
## <none>                  816.53 258.97
## - restime   1    12.200 828.73 258.97
## + height    1     4.793 811.73 260.18
## - fishmlwk  1    45.206 861.73 264.25
## - weight    1   139.057 955.58 278.20
##
## Step:  AIC=258.5
## TotHg ~ restime + weight + fishmlwk + fishpart
##
##            Df Sum of Sq    RSS    AIC
## - restime   1     4.799 830.61 257.28
## <none>                  825.81 258.50
## + age       1     9.289 816.53 258.97
## - fishpart  3    40.857 866.67 259.02
## + height    1     5.267 820.55 259.63
## - fishmlwk  1    52.912 878.73 264.88
## - weight    1   135.101 960.92 276.95
##
## Step:  AIC=257.28
## TotHg ~ weight + fishmlwk + fishpart
##
##            Df Sum of Sq    RSS    AIC
## - fishpart  3    37.460 868.07 257.24
## <none>                  830.61 257.28
## + height    1     6.330 824.28 258.25
## + restime   1     4.799 825.81 258.50
## + age       1     1.888 828.73 258.97
## - fishmlwk  1    49.571 880.19 263.11
## - weight    1   133.220 963.83 275.36
##
## Step:  AIC=257.24
## TotHg ~ weight + fishmlwk
```

```
##
##             Df Sum of Sq     RSS    AIC
## <none>                    868.07 257.24
## + fishpart  3    37.460  830.61 257.28
## + height    1     9.170  858.90 257.80
## + age       1     5.744  862.33 258.34
## + restime   1     1.402  866.67 259.02
## - fishmlwk  1    95.208  963.28 269.29
## - weight    1   182.853 1050.93 281.04
```

```
##
## Call:
## lm(formula = TotHg ~ weight + fishmlwk, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8344 -1.3096 -0.2953  0.6279 11.8572
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.07682    2.44481  -4.122 6.60e-05 ***
## weight        0.17518    0.03322   5.273 5.34e-07 ***
## fishmlwk      0.15884    0.04175   3.805 0.000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.564 on 132 degrees of freedom
## Multiple R-squared:  0.2498, Adjusted R-squared:  0.2384
## F-statistic: 21.97 on 2 and 132 DF,  p-value: 5.787e-09
```


Residuals vs Fitted
Residuals
Fitted values
lm(TotHg ~ weight + fishmlwk)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(TotHg ~ weight + fishmlwk)
85
93
7



With this method of stewise selection, it seems that the best model would be : TotHg ~ weight + fishmlwk The Multiple R-squared is only 0.2498: it seems we are unable to explain most of the variability of TotHg between individuals. However as the p-values for those two parameters are excellents, their influence on TotHg seems well established.

Attempt of backward selection

```
##
## Call:
## lm(formula = hg.form.custom, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8344 -1.3096 -0.2953  0.6279 11.8572
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.07682    2.44481  -4.122 6.60e-05 ***
## weight        0.17518    0.03322   5.273 5.34e-07 ***
## fishmlwk      0.15884    0.04175   3.805 0.000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.564 on 132 degrees of freedom
## Multiple R-squared:  0.2498, Adjusted R-squared:  0.2384
## F-statistic: 21.97 on 2 and 132 DF,  p-value: 5.787e-09
```

Whole population, squared model

```
## Start:  AIC=258.55
## TotHg ~ (age + restime + height + weight + fishpart)^2
##
##                     Df Sum of Sq    RSS    AIC
## - restime:fishpart  3    15.272 638.75 255.82
## - age:restime       1     0.119 623.59 256.58
## - restime:height    1     2.433 625.91 257.08
## - height:weight     1     3.361 626.83 257.28
## <none>                          623.47 258.55
## - age:height        1    11.375 634.85 259.00
## - restime:weight    1    18.614 642.09 260.53
## - height:fishpart   3    39.646 663.12 260.88
## - age:fishpart      3    58.751 682.22 264.71
## - weight:fishpart   3    76.886 700.36 268.25
## - age:weight        1    61.658 685.13 269.29
##
## Step:  AIC=255.82
## TotHg ~ age + restime + height + weight + fishpart + age:restime +
##     age:height + age:weight + age:fishpart + restime:height +
##     restime:weight + height:weight + height:fishpart + weight:fishpart
##
```

```
##                     Df Sum of Sq    RSS    AIC
## - age:restime        1     0.295 639.04 253.88
## - restime:height     1     0.535 639.28 253.93
## - height:weight      1     1.013 639.76 254.03
## - age:height         1     5.954 644.70 255.07
## <none>                           638.75 255.82
## + restime:fishpart   3    15.272 623.47 258.55
## - age:fishpart       3    44.040 682.79 258.82
## - restime:weight     1    28.860 667.61 259.79
## - height:fishpart    3    51.426 690.17 260.27
## - weight:fishpart    3    77.865 716.61 265.35
## - age:weight         1    84.002 722.75 270.50
##
## Step:  AIC=253.88
## TotHg ~ age + restime + height + weight + fishpart + age:height +
##     age:weight + age:fishpart + restime:height + restime:weight +
##     height:weight + height:fishpart + weight:fishpart
##
##                     Df Sum of Sq    RSS    AIC
## - restime:height     1     0.627 639.67 252.02
## - height:weight      1     1.148 640.19 252.13
## - age:height         1     5.974 645.02 253.14
## <none>                           639.04 253.88
## + age:restime        1     0.295 638.75 255.82
## + restime:fishpart   3    15.449 623.59 256.58
## - age:fishpart       3    43.860 682.90 256.85
## - restime:weight     1    29.585 668.63 257.99
## - height:fishpart    3    51.140 690.18 258.28
## - weight:fishpart    3    77.617 716.66 263.36
## - age:weight         1    83.824 722.86 268.52
##
## Step:  AIC=252.02
## TotHg ~ age + restime + height + weight + fishpart + age:height +
##     age:weight + age:fishpart + restime:weight + height:weight +
##     height:fishpart + weight:fishpart
##
##                     Df Sum of Sq    RSS    AIC
## - height:weight      1     1.533 641.20 250.34
## - age:height         1     6.040 645.71 251.28
## <none>                           639.67 252.02
## + restime:height     1     0.627 639.04 253.88
## + age:restime        1     0.386 639.28 253.93
## - age:fishpart       3    43.394 683.06 254.88
## + restime:fishpart   3    13.517 626.15 255.13
## - height:fishpart    3    51.922 691.59 256.55
```

```
## - restime:weight    1    32.422 672.09 256.69
## - weight:fishpart   3    77.093 716.76 261.38
## - age:weight        1    85.904 725.57 267.03
##
## Step:  AIC=250.34
## TotHg ~ age + restime + height + weight + fishpart + age:height +
##     age:weight + age:fishpart + restime:weight + height:fishpart +
##     weight:fishpart
##
##                     Df Sum of Sq    RSS    AIC
## - age:height         1     6.116 647.32 249.62
## <none>                           641.20 250.34
## + height:weight      1     1.533 639.67 252.02
## + restime:height     1     1.012 640.19 252.13
## + age:restime        1     0.601 640.60 252.21
## - age:fishpart       3    42.060 683.26 252.92
## + restime:fishpart   3    10.706 630.49 254.07
## - height:fishpart    3    52.538 693.74 254.97
## - restime:weight     1    37.724 678.92 256.06
## - weight:fishpart    3    75.564 716.76 259.38
## - age:weight         1   100.902 742.10 268.07
##
## Step:  AIC=249.62
## TotHg ~ age + restime + height + weight + fishpart + age:weight +
##     age:fishpart + restime:weight + height:fishpart + weight:fishpart
##
##                     Df Sum of Sq    RSS    AIC
## <none>                           647.32 249.62
## + age:height         1     6.116 641.20 250.34
## + height:weight      1     1.610 645.71 251.28
## - age:fishpart       3    38.213 685.53 251.36
## + restime:height     1     0.415 646.90 251.53
## + age:restime        1     0.377 646.94 251.54
## + restime:fishpart   3     7.928 639.39 253.96
## - height:fishpart    3    64.139 711.45 256.38
## - restime:weight     1    43.861 691.18 256.47
## - weight:fishpart    3    77.043 724.36 258.80
## - age:weight         1    94.791 742.11 266.07
##
## Call:
## lm(formula = TotHg ~ age + restime + height + weight + fishpart +
##     age:weight + age:fishpart + restime:weight + height:fishpart +
##     weight:fishpart, data = dataset)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4743 -1.4451 -0.2005  1.3316  7.2385
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       60.495422  39.416759   1.535  0.12756
## age               -1.863893   0.569443  -3.273  0.00140 **
## restime            1.501286   0.564413   2.660  0.00892 **
## height            -0.049950   0.236701  -0.211  0.83324
## weight            -0.728835   0.292295  -2.493  0.01406 *
## fishpart1        -15.866695  39.630479  -0.400  0.68962
## fishpart2        -18.762969  37.665865  -0.498  0.61933
## fishpart3         20.476896  43.797249   0.468  0.64099
## age:weight         0.026830   0.006510   4.121  7.1e-05 ***
## age:fishpart1      0.037588   0.359720   0.104  0.91696
## age:fishpart2     -0.046435   0.356420  -0.130  0.89657
## age:fishpart3      0.317690   0.384146   0.827  0.40993
## restime:weight    -0.021109   0.007529  -2.804  0.00593 **
## height:fishpart1   0.068815   0.252066   0.273  0.78534
## height:fishpart2   0.127417   0.240461   0.530  0.59720
## height:fishpart3  -0.521829   0.301334  -1.732  0.08598 .
## weight:fishpart1   0.067272   0.254018   0.265  0.79161
## weight:fishpart2   0.002856   0.237454   0.012  0.99042
## weight:fishpart3   0.865055   0.327122   2.644  0.00932 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.362 on 116 degrees of freedom
## Multiple R-squared:  0.4406, Adjusted R-squared:  0.3538
## F-statistic: 5.075 on 18 and 116 DF,  p-value: 2.293e-08
```

Fit of the selected model

Code

```
##
## Call:
## lm(formula = selected.model, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7826 -1.0144 -0.2124  0.8760  6.1297
##
## Coefficients:
```

```
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       24.981784  53.208315   0.470 0.639644
## age               -0.412173   0.994228  -0.415 0.679274
## restime            1.795995   2.394176   0.750 0.454779
## height             0.140765   0.319834   0.440 0.660723
## weight            -0.704960   0.292921  -2.407 0.017781 *
## fishmlwk          -3.241359   1.202236  -2.696 0.008129 **
## fishpart1         12.627404  49.726018   0.254 0.800021
## fishpart2         -7.524995  46.664268  -0.161 0.872189
## fishpart3         10.527728  51.509107   0.204 0.838433
## age:height        -0.007766   0.005278  -1.471 0.144063
## age:weight         0.025053   0.006602   3.795 0.000243 ***
## age:fishpart1      0.067387   0.340625   0.198 0.843545
## age:fishpart2     -0.068695   0.333678  -0.206 0.837275
## age:fishpart3      0.468567   0.361365   1.297 0.197487
## restime:weight    -0.017359   0.007773  -2.233 0.027571 *
## restime:fishmlwk   0.012259   0.008325   1.472 0.143784
## restime:fishpart1 -0.891040   2.296907  -0.388 0.698824
## restime:fishpart2 -0.594226   2.297002  -0.259 0.796357
## restime:fishpart3 -0.903354   2.301023  -0.393 0.695390
## height:fishmlwk    0.019042   0.006978   2.729 0.007407 **
## height:fishpart1  -0.094828   0.306716  -0.309 0.757780
## height:fishpart2   0.065931   0.289712   0.228 0.820405
## height:fishpart3  -0.463851   0.339728  -1.365 0.174950
## weight:fishpart1   0.085785   0.246705   0.348 0.728721
## weight:fishpart2   0.018290   0.230795   0.079 0.936980
## weight:fishpart3   0.819267   0.315924   2.593 0.010812 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.211 on 109 degrees of freedom
## Multiple R-squared:  0.5396, Adjusted R-squared:  0.434
## F-statistic: 5.109 on 25 and 109 DF,  p-value: 1.025e-09
```

```
##
## Call:
## lm(formula = selected.model, data = dataset.fisherman)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3642 -1.5448 -0.0712  1.4663  5.6957
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        41.714299  46.590508   0.895  0.37333
```

```
## age                -0.456802   1.130905  -0.404  0.68736
## restime             0.843434   0.721336   1.169  0.24581
## height             -0.003519   0.270079  -0.013  0.98964
## weight             -0.545177   0.309911  -1.759  0.08243 .
## fishmlwk           -2.916617   1.523614  -1.914  0.05921 .
## fishpart2         -21.229382  20.833440  -1.019  0.31131
## fishpart3          -1.770745  32.746970  -0.054  0.95701
## age:height         -0.007250   0.006113  -1.186  0.23917
## age:weight          0.025474   0.007897   3.226  0.00183 **
## age:fishpart2      -0.140504   0.095471  -1.472  0.14508
## age:fishpart3       0.379636   0.186099   2.040  0.04470 *
## restime:weight     -0.017988   0.009052  -1.987  0.05037 .
## restime:fishmlwk    0.017535   0.010367   1.691  0.09469 .
## restime:fishpart2   0.355986   0.156873   2.269  0.02598 *
## restime:fishpart3   0.078375   0.219395   0.357  0.72187
## height:fishmlwk     0.016742   0.008912   1.879  0.06399 .
## height:fishpart2    0.204607   0.129904   1.575  0.11924
## height:fishpart3   -0.333149   0.243637  -1.367  0.17538
## weight:fishpart2   -0.161967   0.144947  -1.117  0.26720
## weight:fishpart3    0.643322   0.293728   2.190  0.03146 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.521 on 79 degrees of freedom
## Multiple R-squared:  0.5108, Adjusted R-squared:  0.3869
## F-statistic: 4.124 on 20 and 79 DF,  p-value: 3.014e-06
```

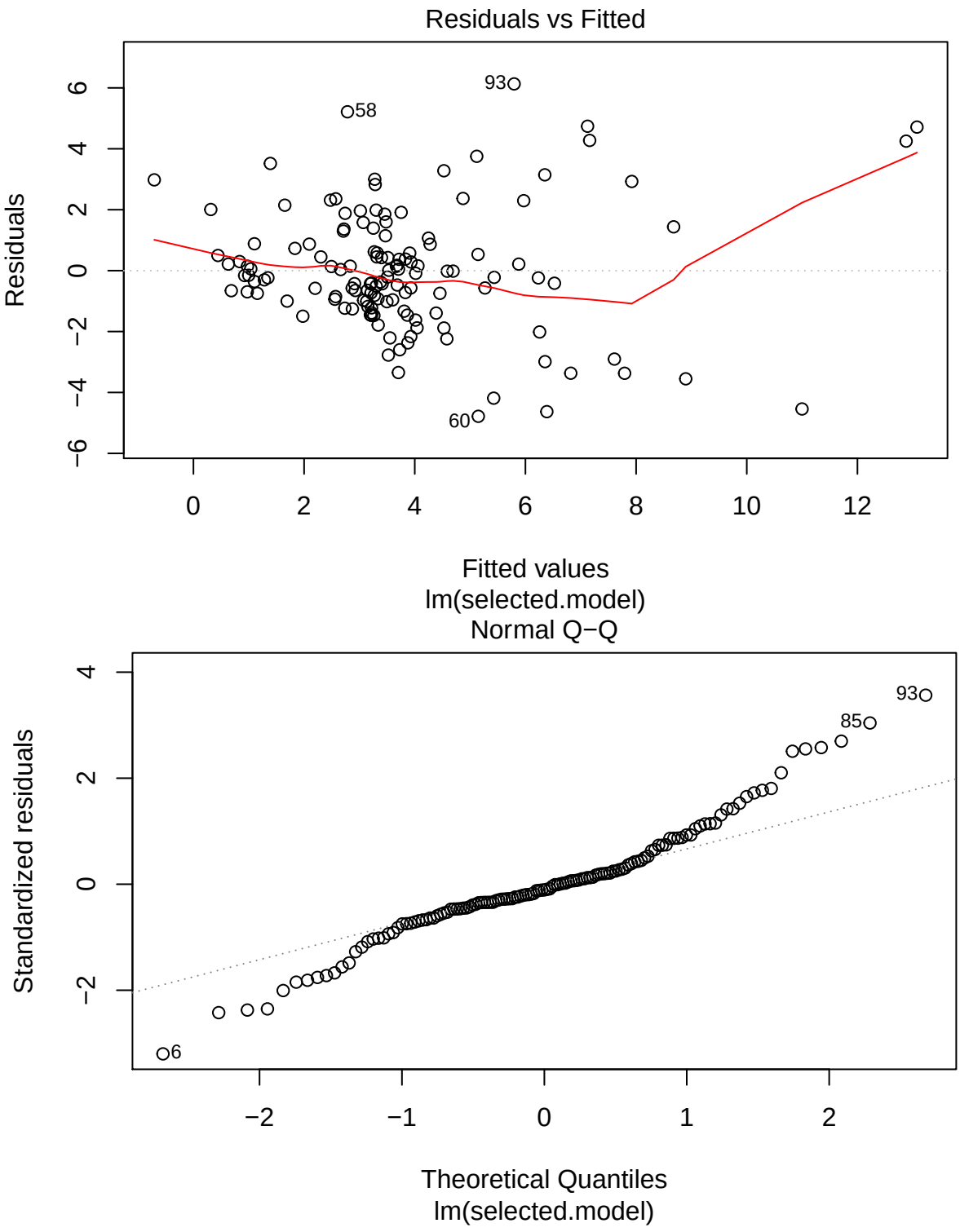
```
##
## Call:
## lm(formula = selected.model, data = dataset.non_fisherman)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62234 -0.09610 -0.00379  0.08914  0.55459
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        36.596286  80.012135   0.457  0.65441
## age                -1.555857   2.604720  -0.597  0.55983
## restime            -0.319986   4.142193  -0.077  0.93952
## height             -0.424264   0.498117  -0.852  0.40869
## weight              0.571077   0.250525   2.280  0.03883 *
## fishmlwk          -24.011624  17.203224  -1.396  0.18453
## fishpart1          95.745727  45.785381   2.091  0.05523 .
## fishpart2          44.283630  29.031066   1.525  0.14943
```

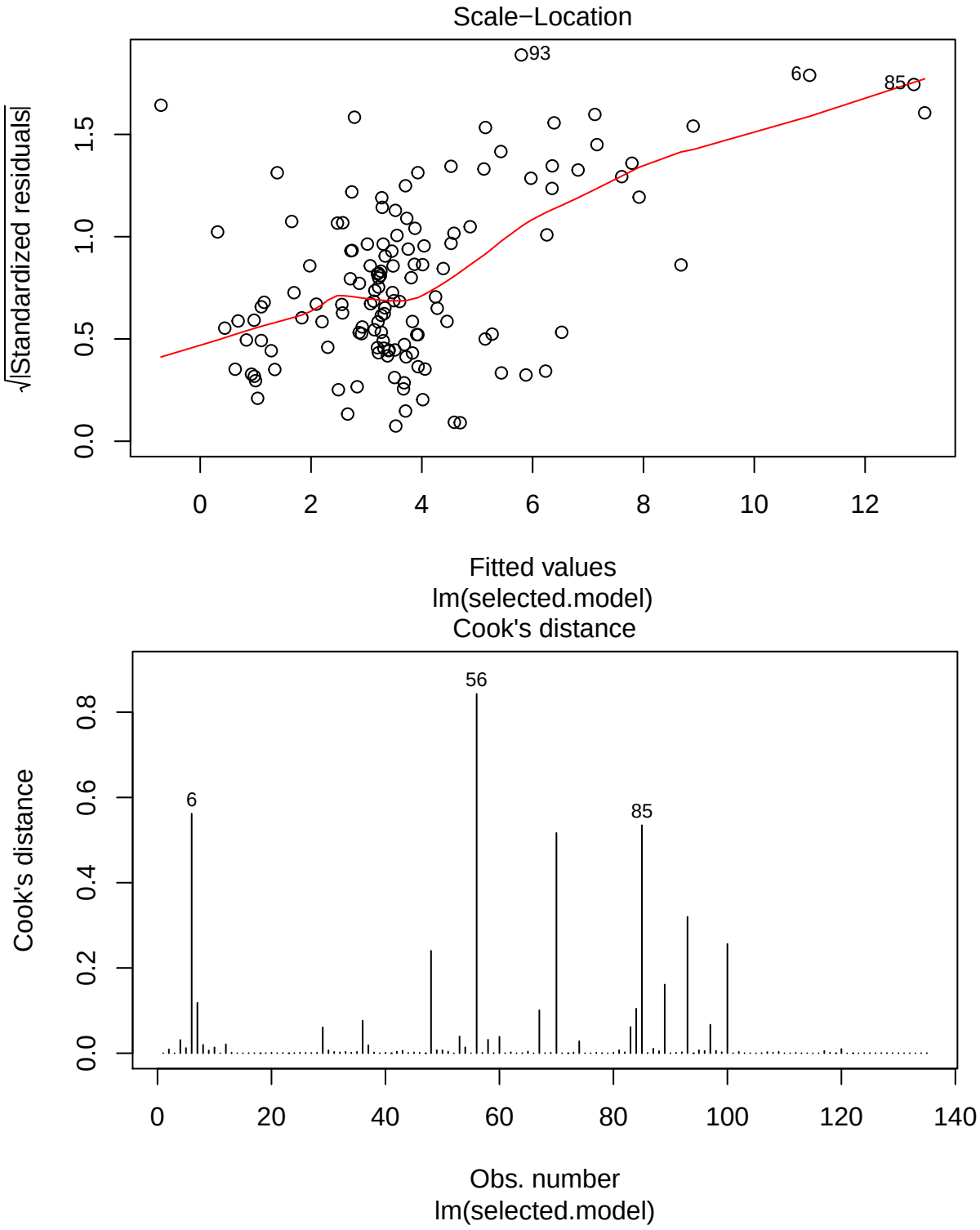
```
## age:height          0.016179   0.017723   0.913  0.37674
## age:weight         -0.018116   0.009150  -1.980  0.06773 .
## age:fishpart1       0.485522   0.135223   3.591  0.00295 **
## age:fishpart2      -0.001470   0.087674  -0.017  0.98686
## restime:weight     -0.005438   0.055128  -0.099  0.92282
## restime:fishmlwk    0.332154   0.732277   0.454  0.65708
## restime:fishpart1  -1.104359   1.388306  -0.795  0.43962
## restime:fishpart2   0.267880   0.916348   0.292  0.77432
## height:fishmlwk     0.139328   0.104784   1.330  0.20489
## height:fishpart1   -0.589596   0.273646  -2.155  0.04910 *
## height:fishpart2   -0.278770   0.172470  -1.616  0.12832
## weight:fishpart1   -0.047403   0.055493  -0.854  0.40736
## weight:fishpart2    0.056973   0.053737   1.060  0.30700
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3245 on 14 degrees of freedom
## Multiple R-squared:  0.978,  Adjusted R-squared:  0.9466
## F-statistic: 31.12 on 20 and 14 DF,  p-value: 2.396e-08
```

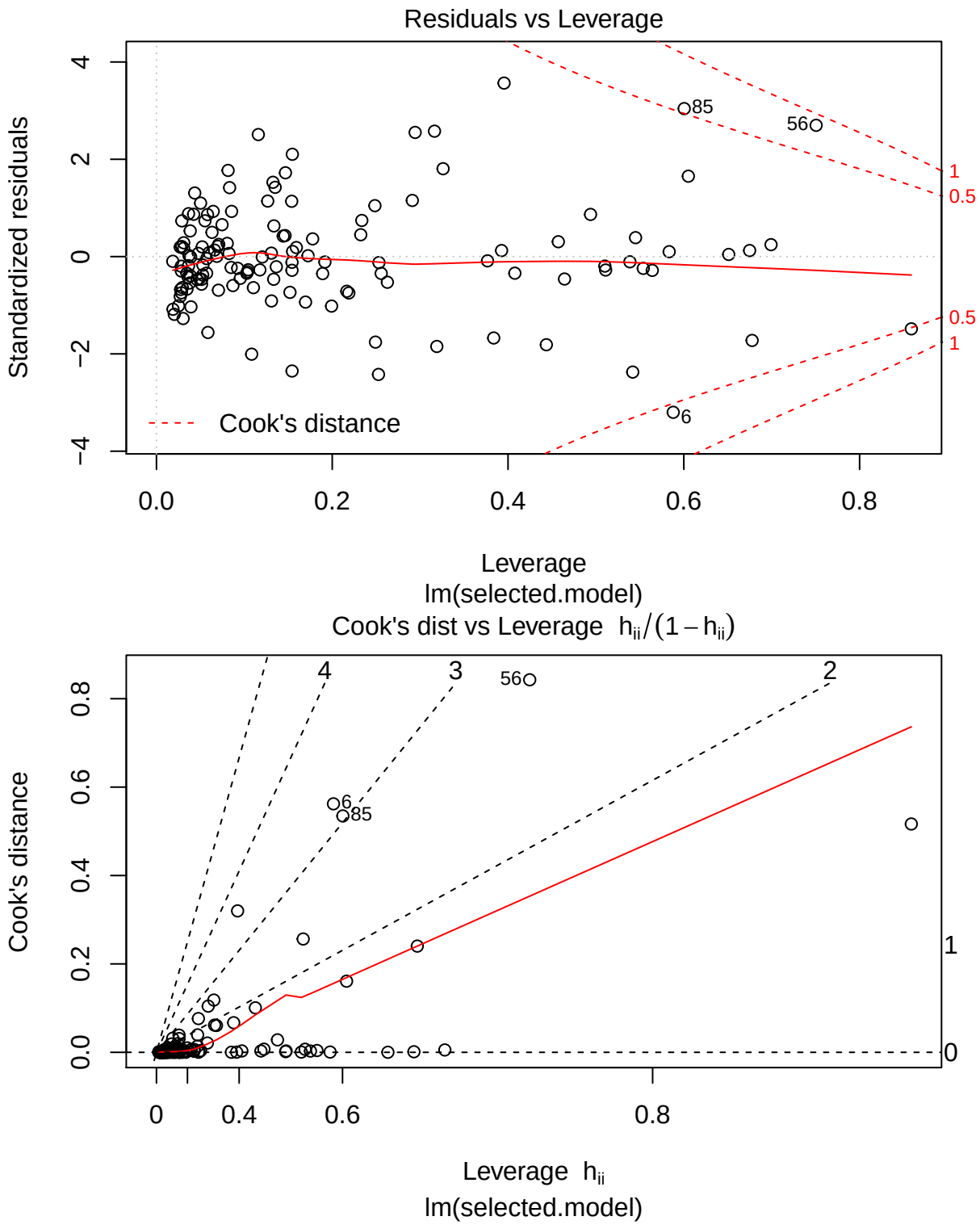
Comments

Diagnostic plots

Code


Residuals vs Fitted
Residuals
Fitted values
lm(selected.model)
93
58
60
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(selected.model)
93
85
6


Scale−Location
√|Standardized residuals|
93
6
85
0.0
0.5
1.0
1.5
0
2
4
6
8
10
12
Fitted values
lm(selected.model)
Cook's distance
Cook's distance
56
6
85
0.0
0.2
0.4
0.6
0.8
0
20
40
60
80
100
120
140
Obs. number
lm(selected.model)



Comments

It seems that there are some influencial points and we sould maybe remove them to analyse the results without them.

Removing the influencial points

New fit

```
##
## Call:
## lm(formula = selected.model, data = dataset.without.inf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6566 -1.0609 -0.1059  0.7876  6.9176
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.661664  49.596348   0.054 0.957303
## age                 0.256819   0.954104   0.269 0.788325
## restime             0.861760   2.208321   0.390 0.697155
## height              0.195453   0.295391   0.662 0.509630
## weight             -0.509945   0.273526  -1.864 0.065067 .
## fishmlwk           -4.746572   1.188419  -3.994 0.000121 ***
## fishpart1          33.540025  45.927883   0.730 0.466848
## fishpart2           7.675486  42.974327   0.179 0.858591
## fishpart3          44.135629  78.246045   0.564 0.573915
## age:height         -0.008736   0.004948  -1.766 0.080358 .
## age:weight          0.017727   0.006355   2.790 0.006269 **
## age:fishpart1       0.323090   0.319042   1.013 0.313537
## age:fishpart2      -0.054656   0.305984  -0.179 0.858576
## age:fishpart3       0.577663   0.701373   0.824 0.412024
## restime:weight     -0.007359   0.007576  -0.971 0.333617
## restime:fishmlwk    0.003968   0.008192   0.484 0.629140
## restime:fishpart1 -0.932007   2.106658  -0.442 0.659102
## restime:fishpart2 -0.335415   2.107046  -0.159 0.873827
## restime:fishpart3 -0.429848   2.121973  -0.203 0.839863
## height:fishmlwk     0.028101   0.006921   4.060 9.46e-05 ***
## height:fishpart1  -0.255528   0.284106  -0.899 0.370491
## height:fishpart2  -0.033702   0.266874  -0.126 0.899749
## height:fishpart3  -0.759303   1.004839  -0.756 0.451553
## weight:fishpart1    0.080102   0.226317   0.354 0.724097
## weight:fishpart2    0.030988   0.211663   0.146 0.883883
## weight:fishpart3    0.959723   1.296247   0.740 0.460718
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.027 on 105 degrees of freedom
## Multiple R-squared:  0.5519, Adjusted R-squared:  0.4452
```

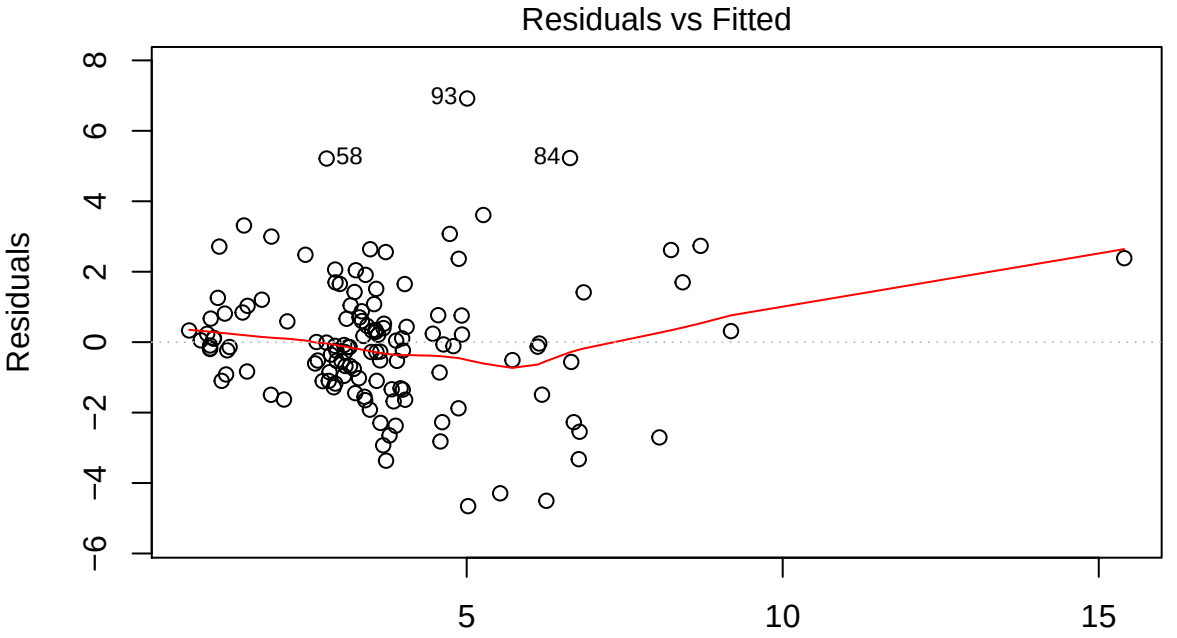
```
## F-statistic: 5.173 on 25 and 105 DF,  p-value: 1.021e-09
```

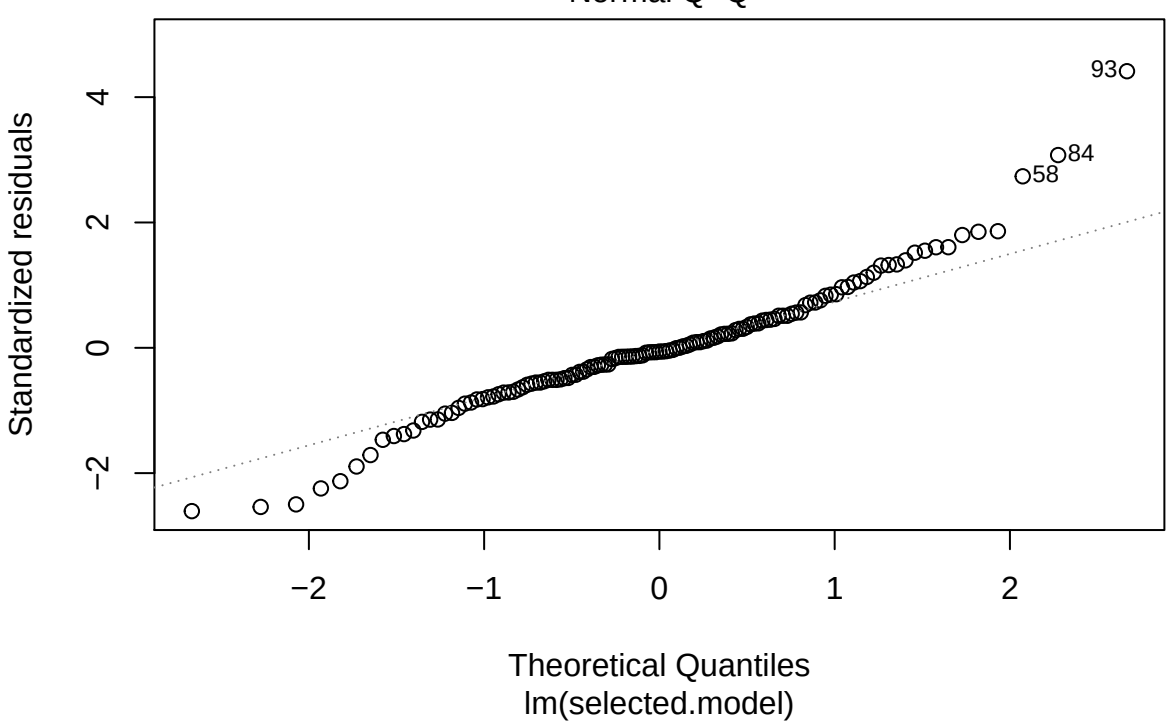
```
##
## Call:
## lm(formula = selected.model, data = dataset.fisherman.without.inf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2003 -1.3465  0.0311  1.1577  6.6744
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        43.872355  43.336649   1.012 0.314620
## age                 0.524010   1.112808   0.471 0.639086
## restime            -0.171606   0.730707  -0.235 0.814966
## height             -0.137731   0.251259  -0.548 0.585207
## weight             -0.339437   0.291889  -1.163 0.248558
## fishmlwk           -4.676130   1.529576  -3.057 0.003096 **
## fishpart2         -29.417772  19.330365  -1.522 0.132254
## fishpart3           1.048212  74.634176   0.014 0.988832
## age:height         -0.008473   0.005757  -1.472 0.145256
## age:weight          0.018041   0.007607   2.372 0.020281 *
## age:fishpart2      -0.386600   0.114424  -3.379 0.001158 **
## age:fishpart3       0.196775   0.729221   0.270 0.788022
## restime:weight     -0.007513   0.008904  -0.844 0.401480
## restime:fishmlwk    0.006829   0.010209   0.669 0.505615
## restime:fishpart2   0.682339   0.174803   3.903 0.000205 ***
## restime:fishpart3   0.587924   0.324639   1.811 0.074144 .
## height:fishmlwk     0.027426   0.008983   3.053 0.003131 **
## height:fishpart2    0.287966   0.121858   2.363 0.020714 *
## height:fishpart3   -0.361388   1.099962  -0.329 0.743415
## weight:fishpart2   -0.165095   0.133261  -1.239 0.219252
## weight:fishpart3    0.688826   1.488082   0.463 0.644780
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.316 on 75 degrees of freedom
## Multiple R-squared:  0.522,  Adjusted R-squared:  0.3945
## F-statistic: 4.095 on 20 and 75 DF,  p-value: 4.24e-06
```

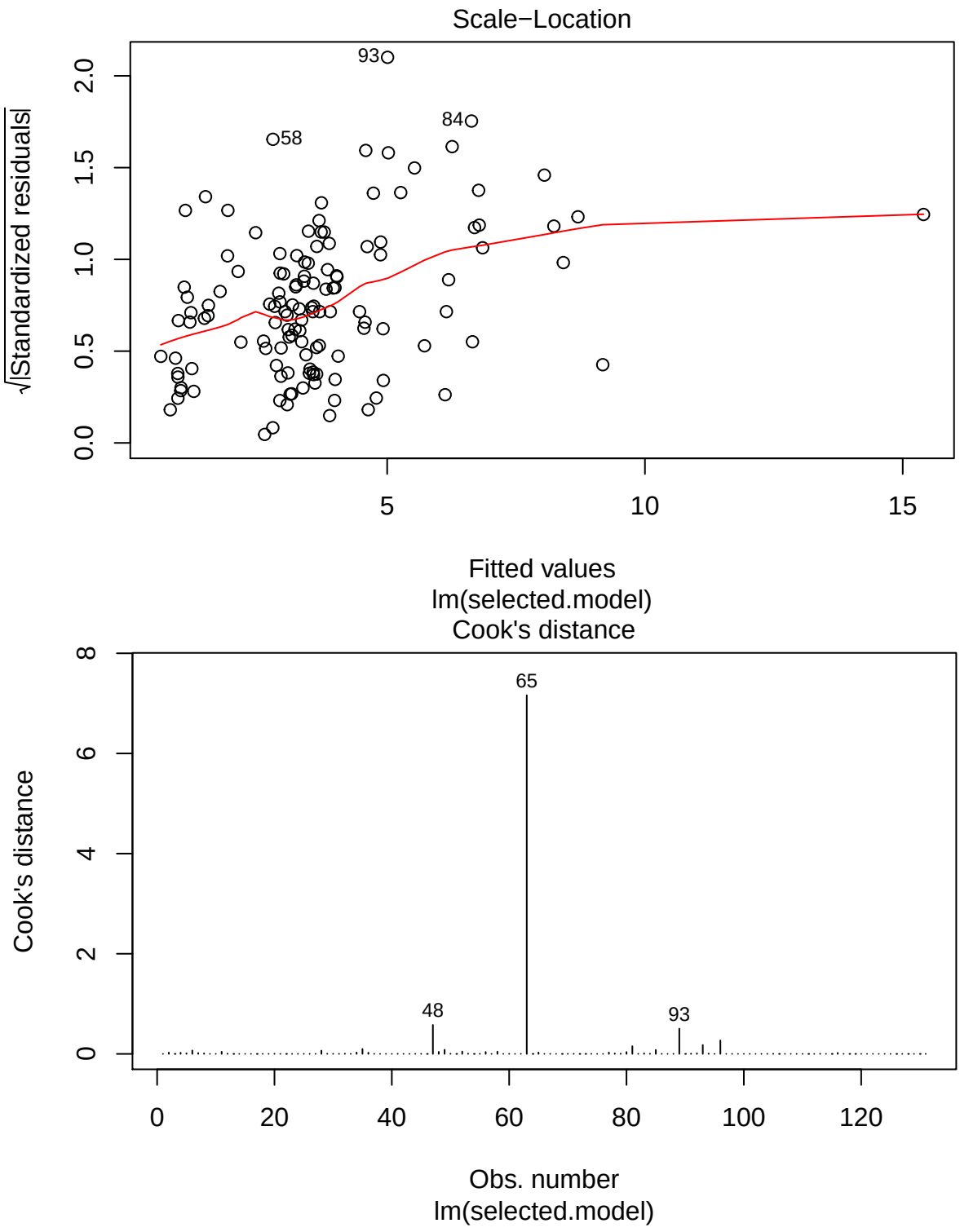
```
##
## Call:
## lm(formula = selected.model, data = dataset.non_fisherman.without.inf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.62736 -0.07481 -0.02919  0.06583  0.55047
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         33.734918  82.126444   0.411  0.68793
## age                 -1.412222   2.680184  -0.527  0.60713
## restime              0.298330   4.375769   0.068  0.94668
## height              -0.421240   0.510383  -0.825  0.42407
## weight               0.591213   0.259017   2.283  0.03993 *
## fishmlwk           -23.097006  17.696314  -1.305  0.21446
## fishpart1           91.163570  47.570751   1.916  0.07757 .
## fishpart2           40.558609  30.429616   1.333  0.20547
## age:height           0.015324   0.018218   0.841  0.41546
## age:weight          -0.017776   0.009394  -1.892  0.08093 .
## age:fishpart1        0.468472   0.141629   3.308  0.00566 **
## age:fishpart2       -0.020205   0.095457  -0.212  0.83565
## restime:weight      -0.012227   0.057682  -0.212  0.83542
## restime:fishmlwk     0.403277   0.760222   0.530  0.60473
## restime:fishpart1   -1.331245   1.475208  -0.902  0.38325
## restime:fishpart2    0.069161   0.999403   0.069  0.94588
## height:fishmlwk      0.133199   0.107878   1.235  0.23879
## height:fishpart1    -0.551152   0.288096  -1.913  0.07802 .
## height:fishpart2    -0.245249   0.185915  -1.319  0.20989
## weight:fishpart1    -0.062699   0.062673  -1.000  0.33537
## weight:fishpart2     0.041797   0.060956   0.686  0.50495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3325 on 13 degrees of freedom
## Multiple R-squared:  0.9778, Adjusted R-squared:  0.9436
## F-statistic: 28.62 on 20 and 13 DF,  p-value: 1.148e-07
```

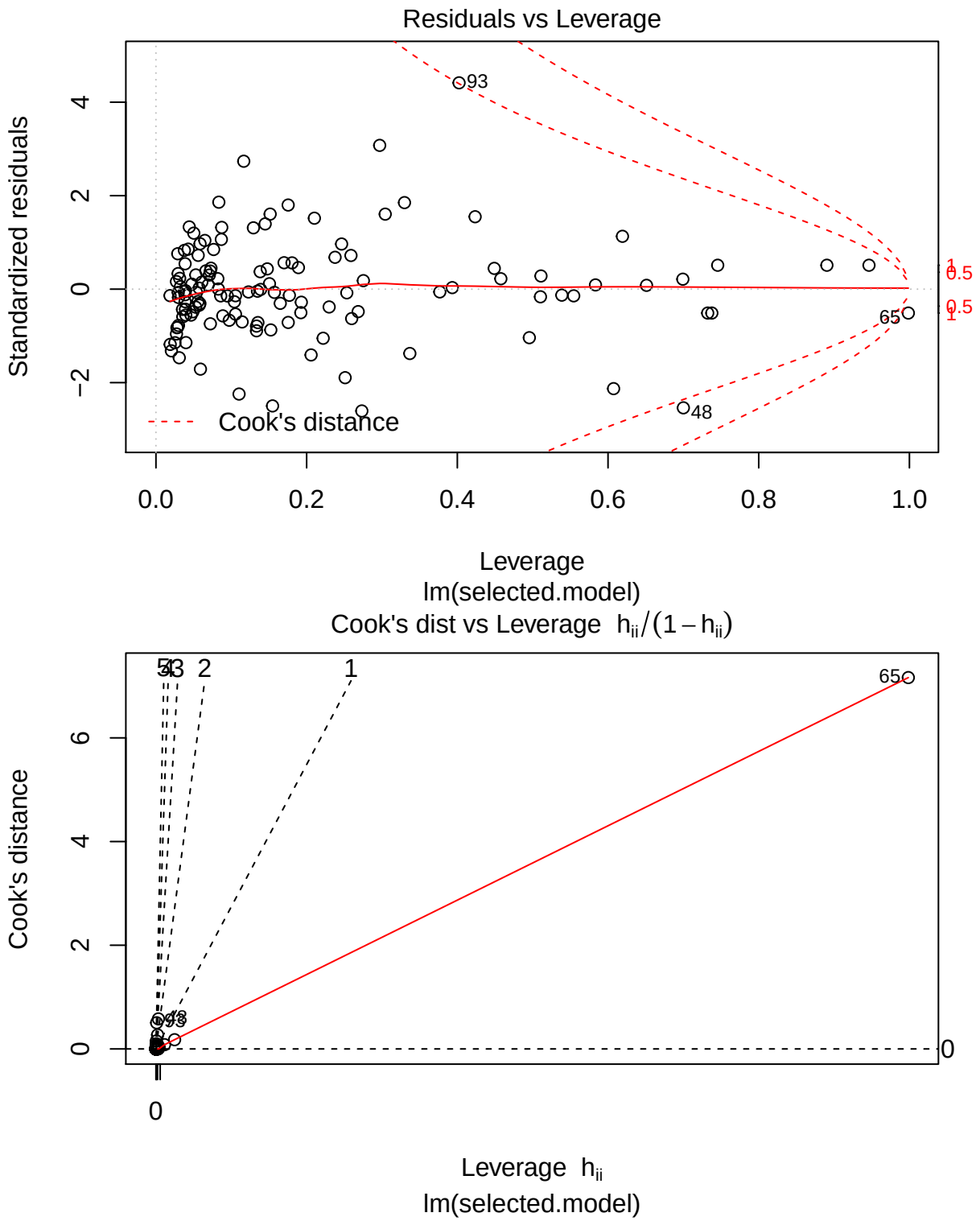
New diagnostic plots








```
## Warning in sqrt(crit * p * (1 - hh)/hh): production de NaN

## Warning in sqrt(crit * p * (1 - hh)/hh): production de NaN
```



We should maybe use the robust regression because we have now other influencial points.

Robust regression

```
##
## Call: rlm(formula = selected.model, data = dataset)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.07936 -0.75778 -0.07816  0.84256  9.23039
##
## Coefficients:
##                   Value   Std. Error t value
## (Intercept)       20.4116 45.1313     0.4523
## age               -0.4640  0.8433    -0.5502
## restime            0.6663  2.0307     0.3281
## height             0.0164  0.2713     0.0603
## weight            -0.3164  0.2485    -1.2734
## fishmlwk          -3.9053  1.0197    -3.8297
## fishpart1         21.0513 42.1776     0.4991
## fishpart2          9.3541 39.5807     0.2363
## fishpart3         29.3766 43.6901     0.6724
## age:height        -0.0020  0.0045    -0.4432
## age:weight         0.0114  0.0056     2.0290
## age:fishpart1      0.1756  0.2889     0.6077
## age:fishpart2     -0.0361  0.2830    -0.1276
## age:fishpart3      0.2872  0.3065     0.9371
## restime:weight    -0.0072  0.0066    -1.0976
## restime:fishmlwk   0.0113  0.0071     1.6070
## restime:fishpart1 -0.5709  1.9482    -0.2930
## restime:fishpart2 -0.1594  1.9483    -0.0818
## restime:fishpart3 -0.4054  1.9517    -0.2077
## height:fishmlwk    0.0229  0.0059     3.8759
## height:fishpart1  -0.1753  0.2602    -0.6737
## height:fishpart2  -0.0701  0.2457    -0.2851
## height:fishpart3  -0.3925  0.2882    -1.3620
## weight:fishpart1   0.1062  0.2093     0.5076
## weight:fishpart2   0.0818  0.1958     0.4179
## weight:fishpart3   0.4477  0.2680     1.6709
##
## Residual standard error: 1.214 on 109 degrees of freedom
```

Now we have no significant result anymore. Do we remove by hand some more influencial points so that we don't remove everything?

=> *What is the scientific question we want to answer?* => In our model, how do we deal with correlated explanatory variables??

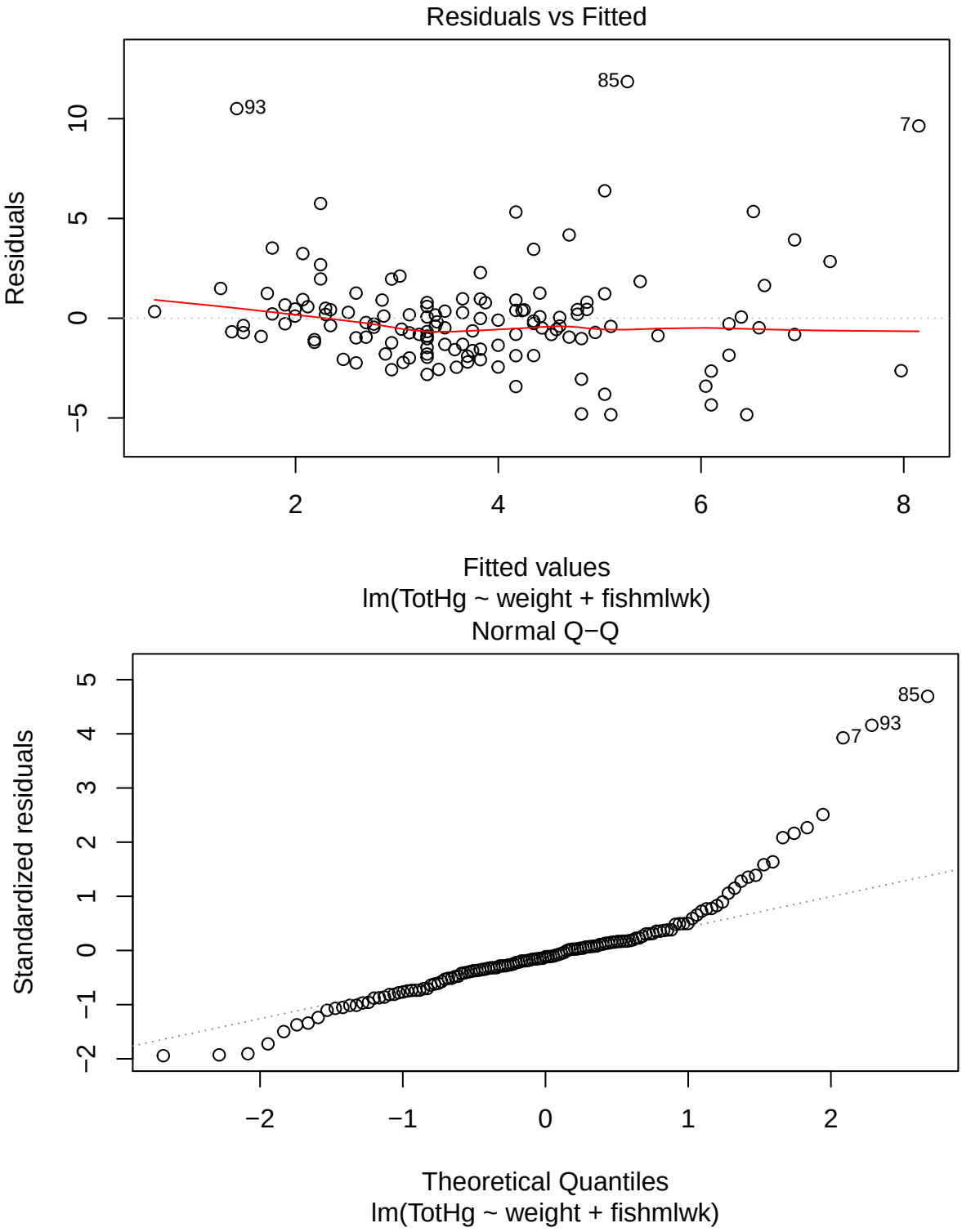
What I think we should do to answer the "scientific question" (Joseph)

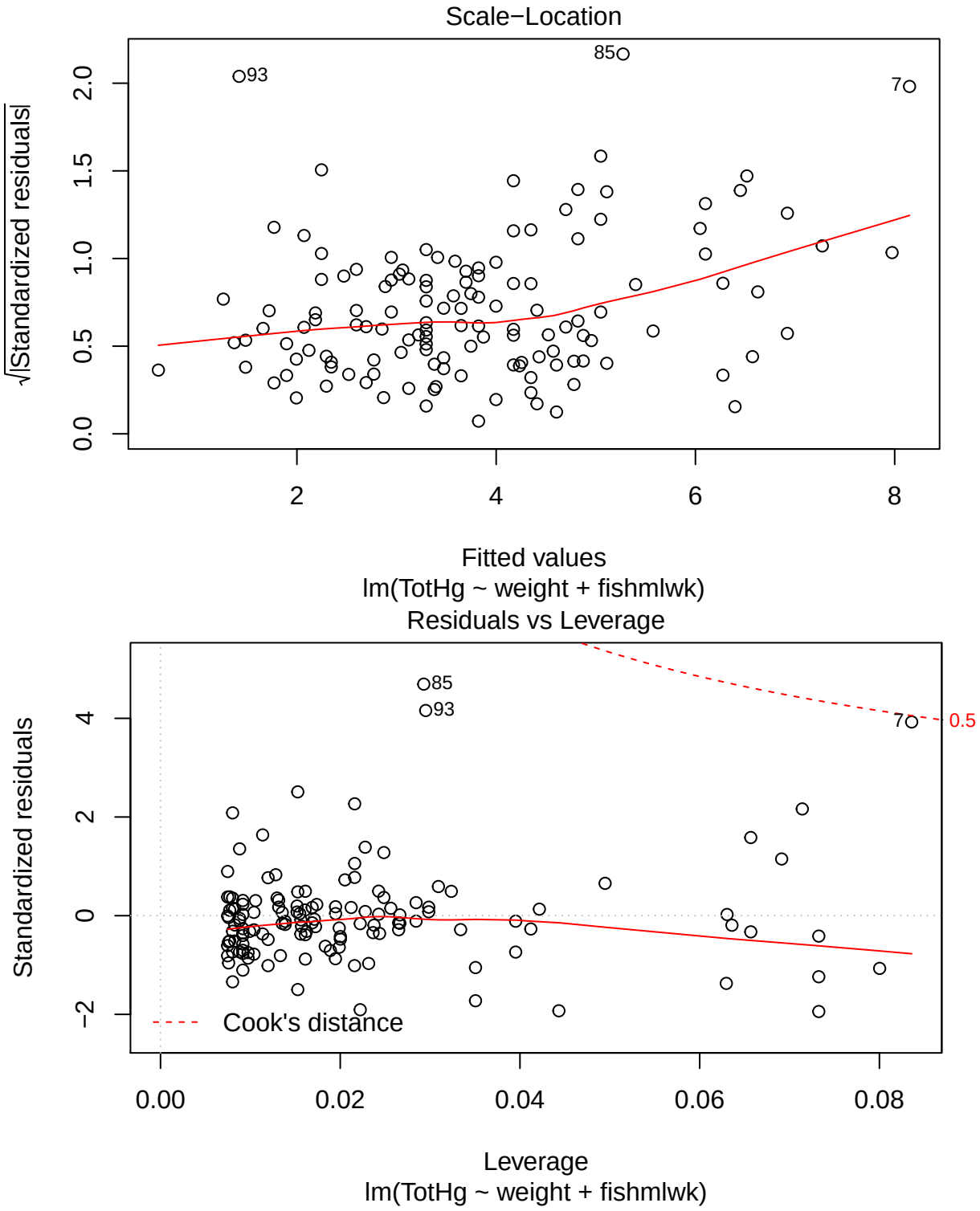
Model selection

Whole population

```
## Start:  AIC=260.18
## TotHg ~ age + restime + height + weight + fishmlwk + fishpart
##
##            Df Sum of Sq    RSS    AIC
## - height    1     4.793 816.53 258.97
## - age       1     8.814 820.55 259.63
## - fishpart  3    33.660 845.39 259.66
## - restime   1    10.460 822.19 259.90
## <none>                  811.73 260.18
## - fishmlwk  1    46.822 858.55 265.75
## - weight    1   115.001 926.73 276.06
##
## Step:  AIC=258.97
## TotHg ~ age + restime + weight + fishmlwk + fishpart
##
##            Df Sum of Sq    RSS    AIC
## - age       1     9.289 825.81 258.50
## - fishpart  3    36.427 852.95 258.86
## <none>                  816.53 258.97
## - restime   1    12.200 828.73 258.97
## + height    1     4.793 811.73 260.18
## - fishmlwk  1    45.206 861.73 264.25
## - weight    1   139.057 955.58 278.20
##
## Step:  AIC=258.5
## TotHg ~ restime + weight + fishmlwk + fishpart
##
##            Df Sum of Sq    RSS    AIC
## - restime   1     4.799 830.61 257.28
## <none>                  825.81 258.50
## + age       1     9.289 816.53 258.97
## - fishpart  3    40.857 866.67 259.02
## + height    1     5.267 820.55 259.63
## - fishmlwk  1    52.912 878.73 264.88
## - weight    1   135.101 960.92 276.95
##
## Step:  AIC=257.28
## TotHg ~ weight + fishmlwk + fishpart
```

```
##
##             Df Sum of Sq    RSS    AIC
## - fishpart  3    37.460 868.07 257.24
## <none>                  830.61 257.28
## + height    1     6.330 824.28 258.25
## + restime   1     4.799 825.81 258.50
## + age       1     1.888 828.73 258.97
## - fishmlwk  1    49.571 880.19 263.11
## - weight    1   133.220 963.83 275.36
##
## Step:  AIC=257.24
## TotHg ~ weight + fishmlwk
##
##             Df Sum of Sq     RSS    AIC
## <none>                   868.07 257.24
## + fishpart  3    37.460  830.61 257.28
## + height    1     9.170  858.90 257.80
## + age       1     5.744  862.33 258.34
## + restime   1     1.402  866.67 259.02
## - fishmlwk  1    95.208  963.28 269.29
## - weight    1   182.853 1050.93 281.04
##
## Call:
## lm(formula = TotHg ~ weight + fishmlwk, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8344 -1.3096 -0.2953  0.6279 11.8572
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.07682    2.44481  -4.122 6.60e-05 ***
## weight        0.17518    0.03322   5.273 5.34e-07 ***
## fishmlwk      0.15884    0.04175   3.805 0.000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.564 on 132 degrees of freedom
## Multiple R-squared:  0.2498, Adjusted R-squared:  0.2384
## F-statistic: 21.97 on 2 and 132 DF,  p-value: 5.787e-09
```


Residuals vs Fitted
Residuals
Fitted values
lm(TotHg ~ weight + fishmlwk)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(TotHg ~ weight + fishmlwk)
85
93
7

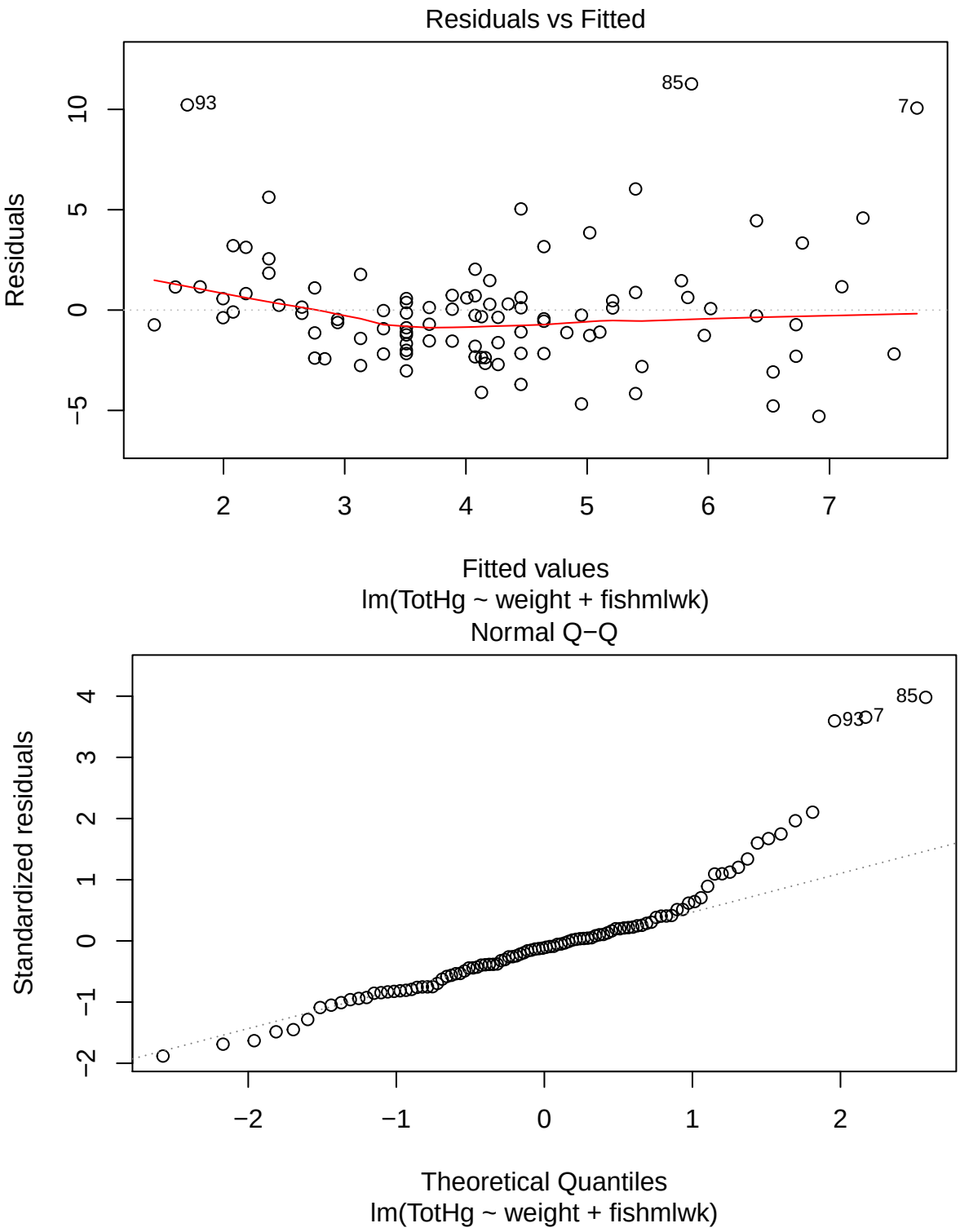


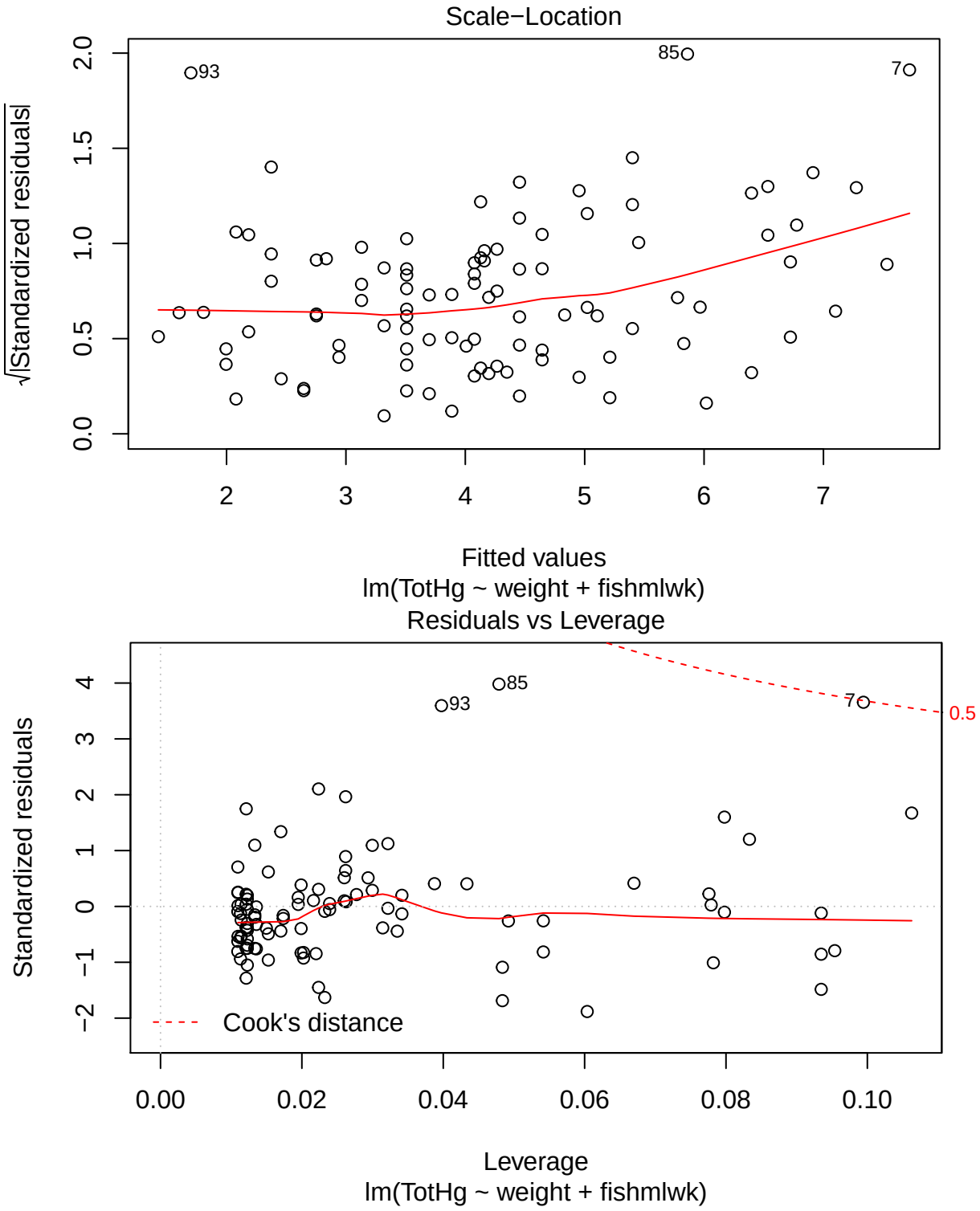
With this method of stewise selection, it seems that the best model would be : TotHg ~ weight + fishmlwk The Multiple R-squared is only 0.2498: it seems we are unable to explain most of the variability of TotHg between individuals. However as the p-values for those two parameters are excellents, their influence on TotHg seems well established.

Fishermen only

```
## Start:  AIC=221.2
## TotHg ~ age + restime + height + weight + fishmlwk + fishpart
##
##            Df Sum of Sq    RSS    AIC
## - fishpart  2    12.485 790.84 218.79
## - height    1     5.743 784.09 219.94
## - age       1     7.787 786.14 220.20
## - fishmlwk  1    13.061 791.41 220.87
## - restime   1    14.307 792.66 221.02
## <none>                  778.35 221.20
## - weight    1   123.311 901.66 233.91
##
## Step:  AIC=218.79
## TotHg ~ age + restime + height + weight + fishmlwk
##
##            Df Sum of Sq    RSS    AIC
## - height    1     9.578 800.41 218.00
## - restime   1    10.815 801.65 218.15
## - age       1    11.071 801.91 218.18
## <none>                  790.84 218.79
## - fishmlwk  1    19.323 810.16 219.21
## + fishpart  2    12.485 778.35 221.20
## - weight    1   142.341 933.18 233.34
##
## Step:  AIC=218
## TotHg ~ age + restime + weight + fishmlwk
##
##            Df Sum of Sq    RSS    AIC
## - age       1    12.296 812.71 217.52
## - restime   1    13.146 813.56 217.62
## <none>                  800.41 218.00
## - fishmlwk  1    18.054 818.47 218.23
## + height    1     9.578 790.84 218.79
## + fishpart  2    16.320 784.09 219.94
## - weight    1   177.424 977.84 236.02
##
## Step:  AIC=217.52
## TotHg ~ restime + weight + fishmlwk
##
##            Df Sum of Sq    RSS    AIC
## - restime   1     4.006 816.72 216.01
## <none>                  812.71 217.52
## + age       1    12.296 800.41 218.00
```

```
## + height    1    10.803 801.91 218.18
## - fishmlwk  1    22.809 835.52 218.29
## + fishpart  2    20.425 792.29 218.97
## - weight    1   175.956 988.67 235.12
##
## Step:  AIC=216.01
## TotHg ~ weight + fishmlwk
##
##            Df Sum of Sq    RSS    AIC
## <none>                  816.72 216.01
## + height    1    11.883 804.83 216.55
## - fishmlwk  1    22.257 838.97 216.70
## + restime   1     4.006 812.71 217.52
## + age       1     3.156 813.56 217.62
## + fishpart  2    14.974 801.74 218.16
## - weight    1   171.982 988.70 233.12
##
## Call:
## lm(formula = TotHg ~ weight + fishmlwk, data = dataset.fisherman)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2935 -1.7079 -0.3086  0.7564 11.2698
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.41533    3.04772  -3.417 0.000925 ***
## weight        0.18909    0.04184   4.520 1.75e-05 ***
## fishmlwk      0.09829    0.06045   1.626 0.107221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.902 on 97 degrees of freedom
## Multiple R-squared:  0.2045, Adjusted R-squared:  0.1881
## F-statistic: 12.47 on 2 and 97 DF,  p-value: 1.52e-05
```


Residuals vs Fitted
Residuals
Fitted values
lm(TotHg ~ weight + fishmlwk)
93
85
7
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(TotHg ~ weight + fishmlwk)



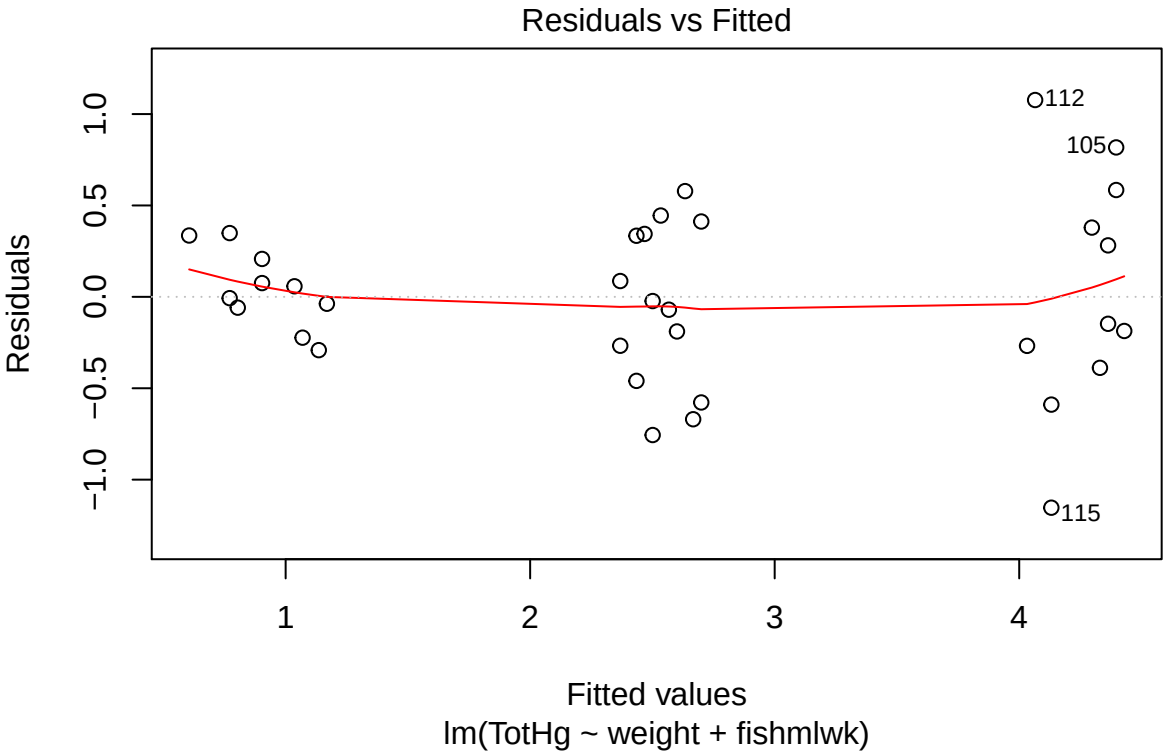
For the fishermen population only, with the same method of stewise selection, the same best model is obtained : TotHg ~ weight + fishmlwk The Multiple R-squared is still very low, and the p-values is good only for the weight parameter. Moreover the weight coefficient is higher than it was in the whole population model. Therefore it seems weight is the dominant explanatory variable among fishermen. Also we have among fishermen a non-negligible amount of very high values for TotHg > The model do not explain that obviously (maybe we

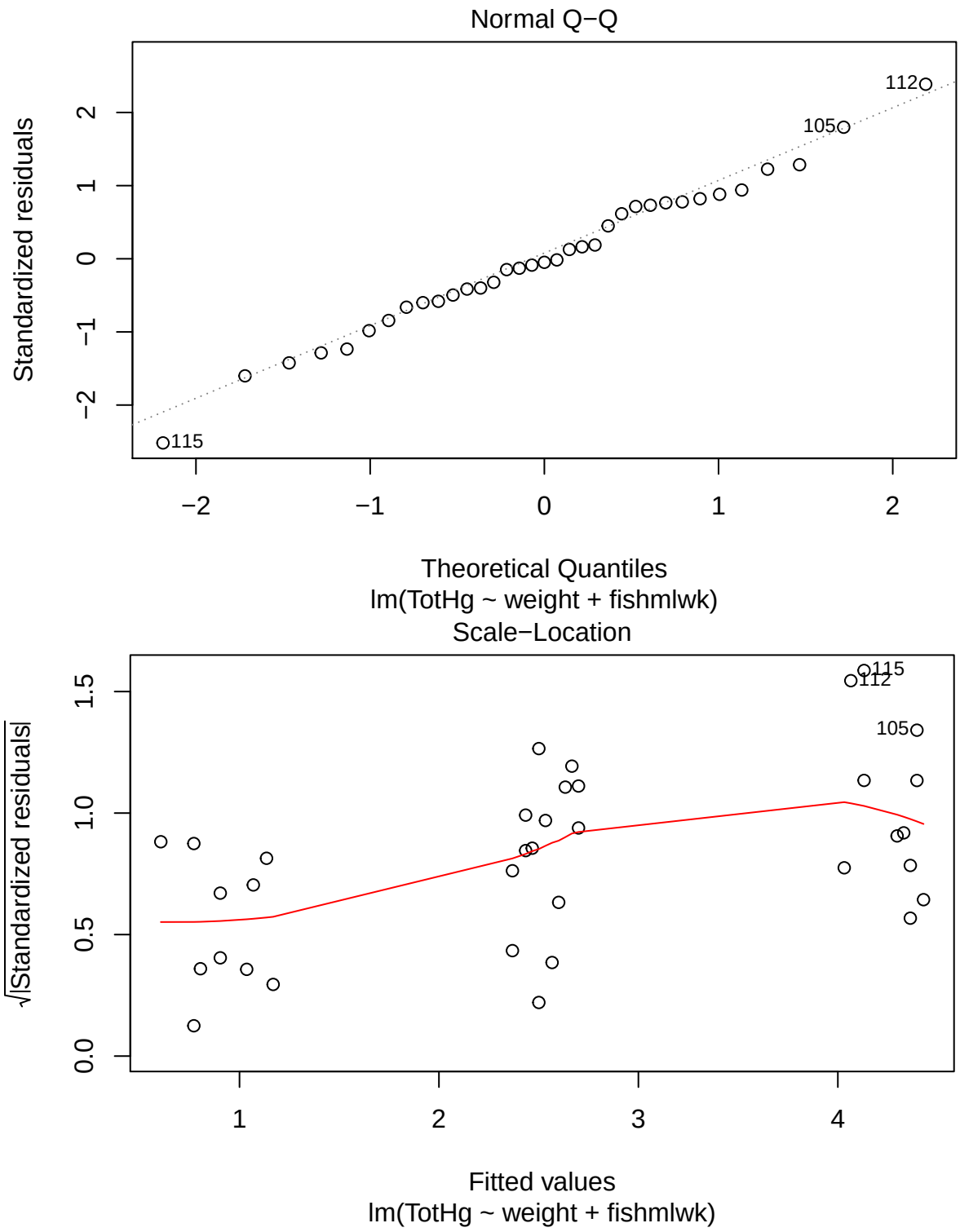
can imagine that a small proportion of the pop are fixing Hg much more / eliminating Hg much slower. . .)

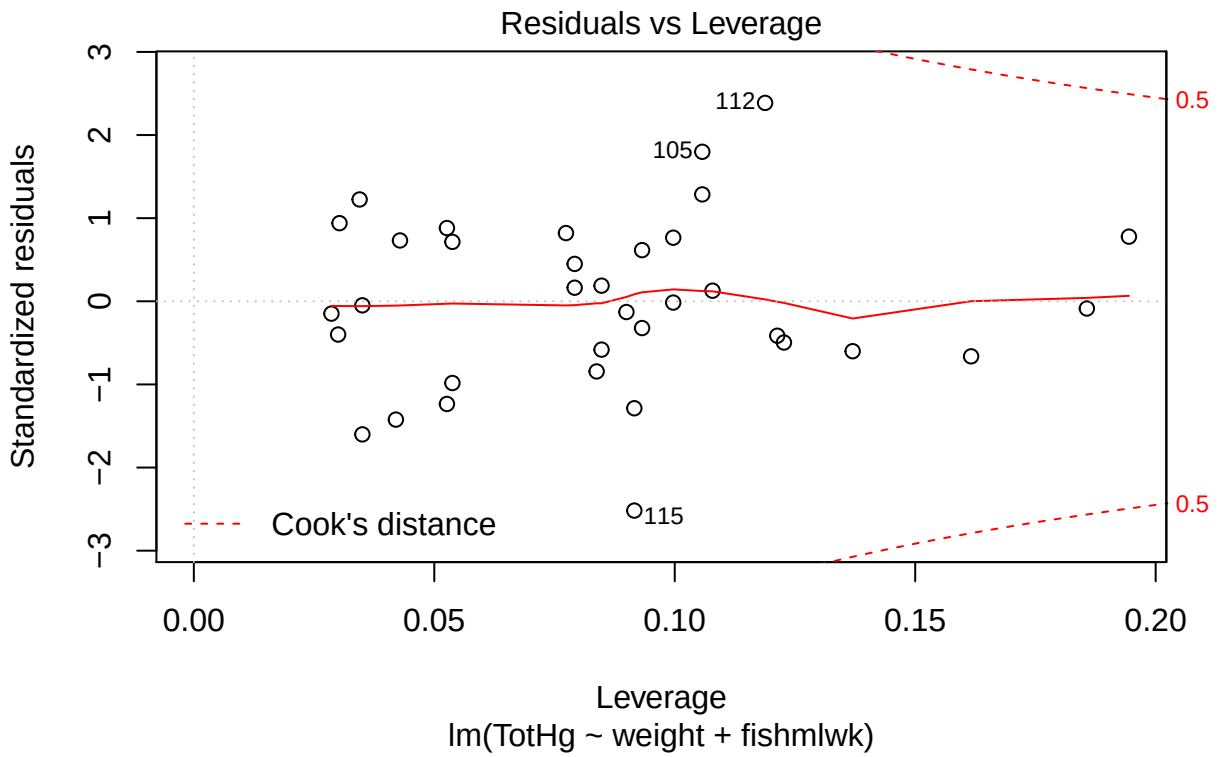
Non-Fishermen only

```
## Start:  AIC=-43.85
## TotHg ~ age + restime + height + weight + fishmlwk
##
##            Df Sum of Sq    RSS     AIC
## - age       1    0.0027  7.099 -45.841
## - restime   1    0.0239  7.120 -45.737
## - height    1    0.2142  7.310 -44.813
## <none>                   7.096 -43.854
## - weight    1    0.9896  8.085 -41.285
## - fishmlwk  1   26.9473 34.043   9.030
##
## Step:  AIC=-45.84
## TotHg ~ restime + height + weight + fishmlwk
##
##            Df Sum of Sq    RSS     AIC
## - restime   1    0.0242  7.123 -47.722
## - height    1    0.2118  7.310 -46.812
## <none>                   7.099 -45.841
## + age       1    0.0027  7.096 -43.854
## - weight    1    0.9936  8.092 -43.255
## - fishmlwk  1   28.3067 35.405   8.403
##
## Step:  AIC=-47.72
## TotHg ~ height + weight + fishmlwk
##
##            Df Sum of Sq    RSS     AIC
## - height    1    0.2709  7.394 -48.415
## <none>                   7.123 -47.722
## + restime   1    0.0242  7.099 -45.841
## + age       1    0.0031  7.120 -45.737
## - weight    1    0.9821  8.105 -45.201
## - fishmlwk  1   28.5577 35.680   6.674
##
## Step:  AIC=-48.42
## TotHg ~ weight + fishmlwk
##
##            Df Sum of Sq    RSS     AIC
## <none>                   7.394 -48.415
## + height    1    0.2709  7.123 -47.722
## - weight    1    0.7145  8.108 -47.187
```

```
## + restime   1    0.0832  7.310 -46.812
## + age       1    0.0004  7.393 -46.417
## - fishmlwk  1   30.3881 37.782   6.677

##
## Call:
## lm(formula = TotHg ~ weight + fishmlwk, data = dataset.non_fisherman)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15344 -0.26774 -0.02297  0.33987  1.07683
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.41565    1.31955  -1.073   0.2914
## weight       0.03313    0.01884   1.759   0.0882 .
## fishmlwk     1.53107    0.13351  11.468 7.17e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4807 on 32 degrees of freedom
## Multiple R-squared:  0.8897, Adjusted R-squared:  0.8828
## F-statistic:   129 on 2 and 32 DF,  p-value: 4.806e-16
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(TotHg ~ weight + fishmlwk)
112
105
115
Scale−Location
√|Standardized residuals|
Fitted values
lm(TotHg ~ weight + fishmlwk)



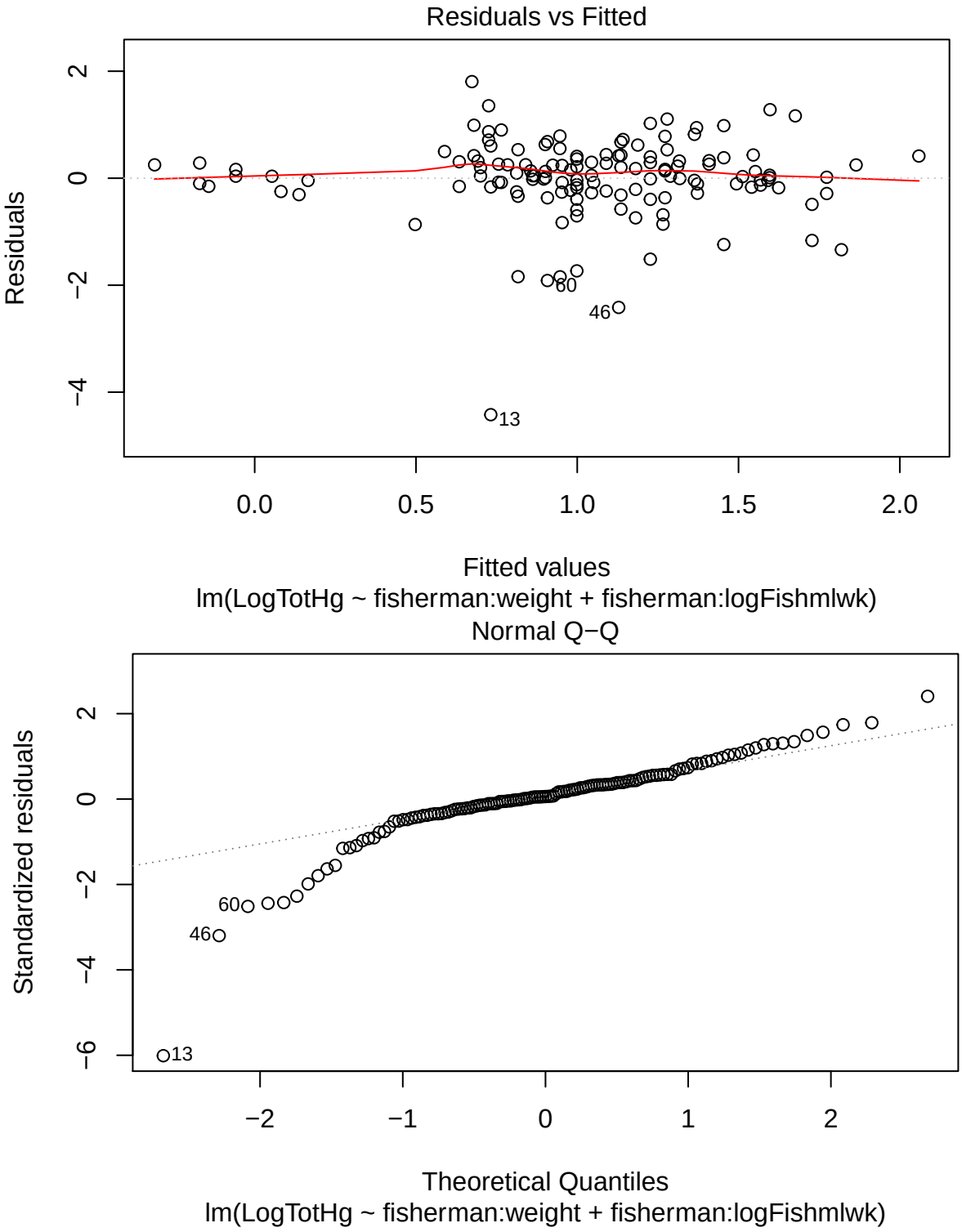
For the non-fishermen population only, with the same method of stewise selection, a model with weight, fishmlwk and fishpart is obtained. However the values of the parameters for fishpart are absurd, or too strange to be interpreted. If fishpart is removed from the formula, the model obtained is the same as before : TotHg ~ weight + fishmlwk The Multiple R-squared is now very good, and the p-values is good only for the fishmlwk parameter. Moreover the fishmlwk coefficient is higher than it was in the whole population model. Therefore it seems fishmlwk is the dominant explanatory variable by a very large margin among non-fishermen.

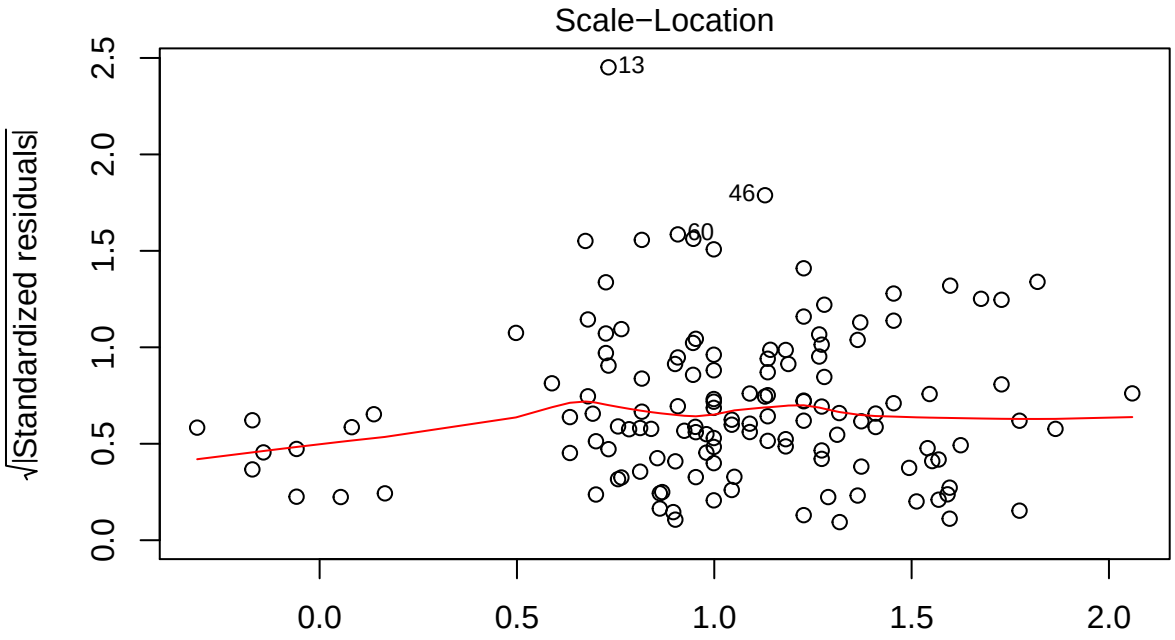
Whole population, degree 1

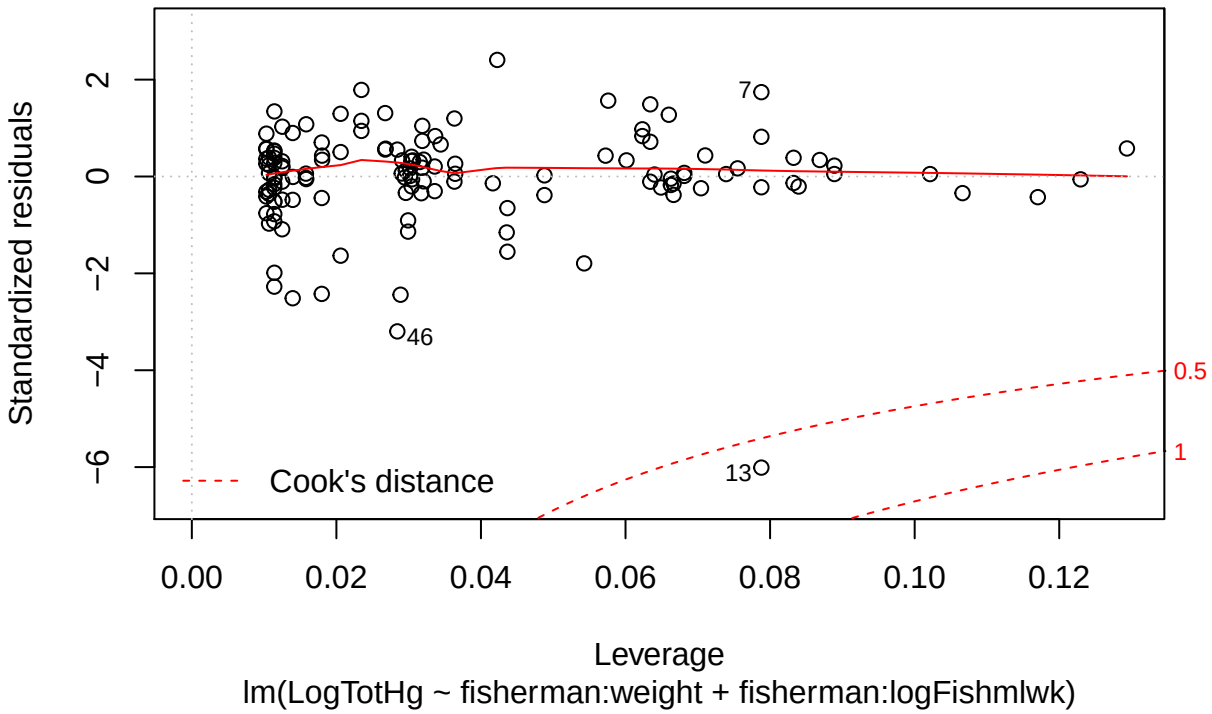
```
##                             GVIF Df GVIF^(1/(2*Df))
## fisherman:age          102.85719  2        3.184628
## fisherman:restime       24.57676  2        2.226543
## fisherman:weight       102.72273  2        3.183586
## fisherman:logFishmlwk  12.49073  2        1.879953
```

```
## Start:  AIC=-59.23
## LogTotHg ~ fisherman:(age + restime + weight + logFishmlwk)
##
##                          Df Sum of Sq    RSS     AIC
## - fisherman:restime       2    0.0296 76.219 -63.175
## - fisherman:age           2    0.1215 76.311 -63.012
## <none>                                 76.189 -59.227
## - fisherman:logFishmlwk   2    7.1744 83.364 -51.078
## - fisherman:weight        2   10.7780 86.967 -45.365
```

```
##
## Step:  AIC=-63.17
## LogTotHg ~ fisherman:age + fisherman:weight + fisherman:logFishmlwk
##
##                          Df Sum of Sq    RSS     AIC
## - fisherman:age           2    0.1354 76.354 -66.935
## <none>                                 76.219 -63.175
## + fisherman:restime       2    0.0296 76.189 -59.227
## - fisherman:logFishmlwk   2    7.4716 83.691 -54.550
## - fisherman:weight        2   10.9210 87.140 -49.098
##
## Step:  AIC=-66.94
## LogTotHg ~ fisherman:weight + fisherman:logFishmlwk
##
##                          Df Sum of Sq    RSS     AIC
## <none>                                 76.354 -66.935
## + fisherman:age           2    0.1354 76.219 -63.175
## + fisherman:restime       2    0.0435 76.311 -63.012
## - fisherman:logFishmlwk   2    7.4895 83.844 -58.303
## - fisherman:weight        2   14.3208 90.675 -47.729

##
## Call:
## lm(formula = LogTotHg ~ fisherman:weight + fisherman:logFishmlwk,
##     data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4208 -0.2184  0.0450  0.3705  1.8052
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -2.01729    0.79877  -2.525 0.012755 *
## fisherman0:weight        0.02798    0.01180   2.370 0.019234 *
## fisherman1:weight        0.04557    0.01037   4.396 2.27e-05 ***
## fisherman0:logFishmlwk   1.17520    0.32944   3.567 0.000506 ***
## fisherman1:logFishmlwk -0.08354    0.18855  -0.443 0.658453
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7664 on 130 degrees of freedom
## Multiple R-squared:  0.2519, Adjusted R-squared:  0.2289
## F-statistic: 10.94 on 4 and 130 DF,  p-value: 1.114e-07
```


Residuals vs Fitted
Residuals
Fitted values
lm(LogTotHg ~ fisherman:weight + fisherman:logFishmlwk)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(LogTotHg ~ fisherman:weight + fisherman:logFishmlwk)
13
46
60






```
##
## Call:
## lmRob(formula = step$call$formula, data = dataset)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -5.01436 -0.32630  0.02197  0.23154  1.73561
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -2.096131   0.597814  -3.506 0.000624 ***
## fisherman0:weight       0.029103   0.008775   3.317 0.001181 **
## fisherman1:weight       0.038270   0.007970   4.802 4.26e-06 ***
## fisherman0:logFishmlwk  1.167915   0.224733   5.197 7.64e-07 ***
## fisherman1:logFishmlwk  0.289806   0.140491   2.063 0.041121 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.396 on 130 degrees of freedom
## Multiple R-Squared: 0.3308
##
## Test for Bias:
##             statistic  p-value
## M-estimate      3.191 0.670604
## LS-estimate    18.828 0.002069
```

I tried three things here:

- I set a log scale on *TotHg* since it is a value ranging from zero to infinity possibly taking arbitrarily low values, but bottom-bounded by zero: thus it couldn't possibly have a normal distribution for any valuation of the observables.
- I set a (approximate) log scale on *fishmlwk* since it is a count value ranging from 0 to 21. I followed the idea from the course 5, even if *fishmlwk* is not a response (y) but an observation (x).
- I used the robust regression from the *robust* package in order to obtain the *p-values* for the coefficients of the robust regression, and set a reduced weight to observations classified as outliers by the Cook distance.

I obtained several interesting results:

- The VIFs suggest that no colinearity is to be observed between the observables.
- The right tail of residuals is very reduced and almost fits the normal distribution.
- The left tail of residuals is somewhat expanded, but less than the right tail with non-log scale.
- The residuals **do not appear to be homoscedactically distributed**... Namely, there is an island with small fitted values with a very small variance of residuals...

The final model got by the regression is *LogTotHg ~ fisherman:weight + fisherman:logFishmlwk*.

References

Al-Majed, NB, and MR Preston. 2000. "Factors Influencing the Total Mercury and Methyl Mercury in the Hair of the Fishermen of Kuwait." *Environmental Pollution* 109 (2). Elsevier: 239–50.