Project 1: regression 1

Urvan Christen, Amandine Goffeney, Joseph Vermeil, Lucile Vigué

March 20, 2019

Introduction

Mercury is a metal present in the environment whose harmful effects on human health are well assessed (Park and Zheng 2012). In a study led in 2000 (Al-Majed and Preston 2000), scientists collected data on total mercury and methyl mercury levels in the hair of 100 fishermen of Kuwait, aged 16 to 58 years, comparing them to those of a control population of 35 non-fishermen, aged 26 to 35 years. The aim of this report is to analyse the factors influencing the levels of mercury in both populations. For the sake of simplicity, we will only focus on total Hg, leaving out methyl mercury, as both variables are strongly correlated.

The dataset gathers information about six numerical variables (age, height, weight, number of fish meals per week and residence time in Kuwait) and two categorical ones (being a fisherman or not, fish consumption habits). There is no additional information about gender as all the participants in the study are males.

A first insight at the data shows that the fishermen population exhibits higher levels of mercury in their hair. The significance of the difference between the means of both distributions is assessed by a *Welch Two Sample t-test* with the alternative hypothesis that the fishermen population has a average greater level of mercury than the control population (p-value = 7.473e-05).

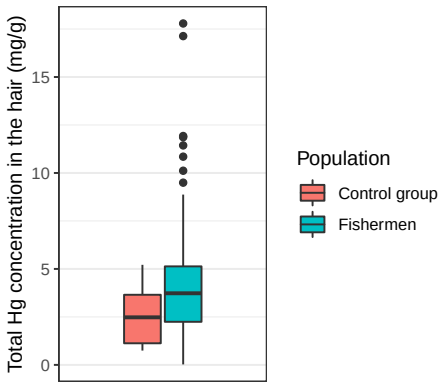


Figure 1: Boxplots of Hg levels

Exploratory analysis (Amandine)

Overview of the data

Before fitting a model, let's have a look at the data. The following table shows the 8 possible values of fishmlwk according to the 2 possible values of fisherman.

```
##
##       0  1  2  3  4  7 14 21
##   0 10 14 11  0  0  0  0  0
##   1  0  0  0  2 12 70  5 11
```

First, let's notice that for some values, we have a very few people, for instance only 2 people eat fish 3 times a week. We see also that the data is really unbalanced, because the fishermen and non fishermen are separated on the variable "fishmlwk": the non fishermen don't eat fish more than 2 times a week, whereas the fishermen don't eat fish less than 3 times a week. Besides, we face the same unbalanced pattern for the variables *age* and *restime*.

Multicolinearity

To check whether we have multicolinear variables, we use the VIF function. If all the variables have a result below 5, we keep them.

```
##                GVIF Df GVIF^(1/(2*Df))
## fisherman 2.176756  1        1.475383
## age       1.609314  1        1.268587
## restime   1.609630  1        1.268712
## height    1.138614  1        1.067059
## weight    1.220641  1        1.104826
## fishmlwk  1.754183  1        1.324456
## fishpart  1.745741  3        1.097312
```

Indeed, we observe that all the variables have a VIF below 2, so we don't eliminate any variable, yet.

Model selection (Joseph)

Now we are going to use and compare the different methods of model selection to select the more relevant parameters to explain the TotHg variations within the population.

Stepwise selection

The first attempt is a stepwise selection based on a formula with all the parameters except for *fisherman*. Indeed, the preliminary exploration has shown that the *fisherman* variable have considerable impact on many of the other variables. Therefore to understand what is the reason behind the TotHg difference between the 2 groups, it seems natural to start with a model without the *fisherman* variable.

```
##
## 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
```

With stepwise selection, the selected model is : TotHg ~ weight + fishmlwk. The intercept and the two coefficients are very significant. Moreover the signs of the coefficients are not absurd: while it is not really intuitive that the weight coefficient should be positive or negative, the coefficient of *fishmlwk* had to be positive, and it's the case here.

Now to determine possible differences between the fishermen and non-fishermen, a model based on the interactions between the *fisherman* variable and all the others is proposed. This could show if a certain variable is very relevant concerning fishermen but less when it comes to non-fishermen, for instance.

```
##
## Call:
## lm(formula = TotHg ~ fisherman:weight + fisherman:fishmlwk, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0705 -1.1391 -0.2026  0.6525 11.4266
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -9.26610    2.48392  -3.730 0.000284 ***
## fisherman0:weight    0.14464    0.03671   3.939 0.000133 ***
## fisherman1:weight    0.17361    0.03419   5.078 1.29e-06 ***
## fisherman0:fishmlwk  1.09393    0.60324   1.813 0.072076 .
## fisherman1:fishmlwk  0.09691    0.05274   1.837 0.068420 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.532 on 130 degrees of freedom
## Multiple R-squared:  0.2797, Adjusted R-squared:  0.2575
## F-statistic: 12.62 on 4 and 130 DF,  p-value: 1.053e-08
```

The best model is now : TotHg ~ fisherman:weight + fisherman:fishmlwk, ie the interactions between fisherman and the variables of the previous model. While the coefficients for fisherman0:weight and fisherman1:weight stays very significant, fisherman0:fishmlwk and

fisherman1:fishmlwk are barely significant (but their p-value is very close to 0.05) ; this comes from the fact that within each of the subpopulation {fishermen} and {non-fishermen} the range in which fishmlwk varies is very reduced compared to the whole population (see the Exploratory analysis). Moreover, when looking at the values of the coefficients, it appears that for the non-fishermen, the number of fish meal per week has a very preponderant effect, while among fishermen, the contribution of weight and fish meal per week are almost even.

In order to check the dependency of the selected model to the selection technique, we have tried backward and forward selection and obtained similar models.

Results and discussion (Urvan)

Fit model (* fishermen) (I picked the model given by stepwise selection)

```
##
## Call:
## lm(formula = hg.form.selected, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2935 -1.1455 -0.1474  0.5763 11.2698
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## fisherman0           -1.41565    6.93854  -0.204 0.838654
## fisherman1          -10.41533    2.65475  -3.923 0.000141 ***
## fisherman0:weight     0.03313    0.09907   0.334 0.738597
## fisherman1:weight     0.18909    0.03644   5.189    8e-07 ***
## fisherman0:fishmlwk   1.53107    0.70201   2.181 0.030997 *
## fisherman1:fishmlwk   0.09829    0.05266   1.867 0.064235 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.528 on 129 degrees of freedom
## Multiple R-squared:  0.7325, Adjusted R-squared:  0.7201
## F-statistic: 58.89 on 6 and 129 DF,  p-value: < 2.2e-16
```

This regression shows several intersting results:

First, we get two coefficients very significantly different from 0 ($p < 0.001$), both concerning fisherman population. The fact that there are less significant coefficients for non-fisherman population could be explained by several causes:

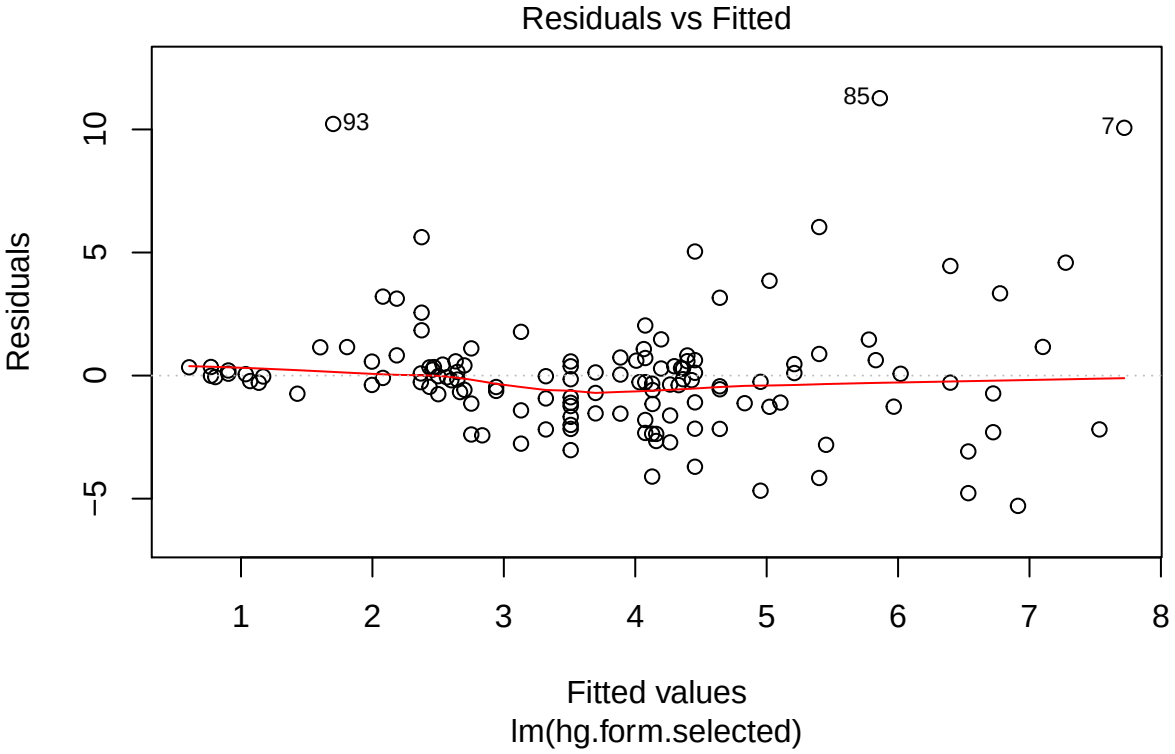
- The non-fisherman population could be too small to properly show the size effect of those observables.

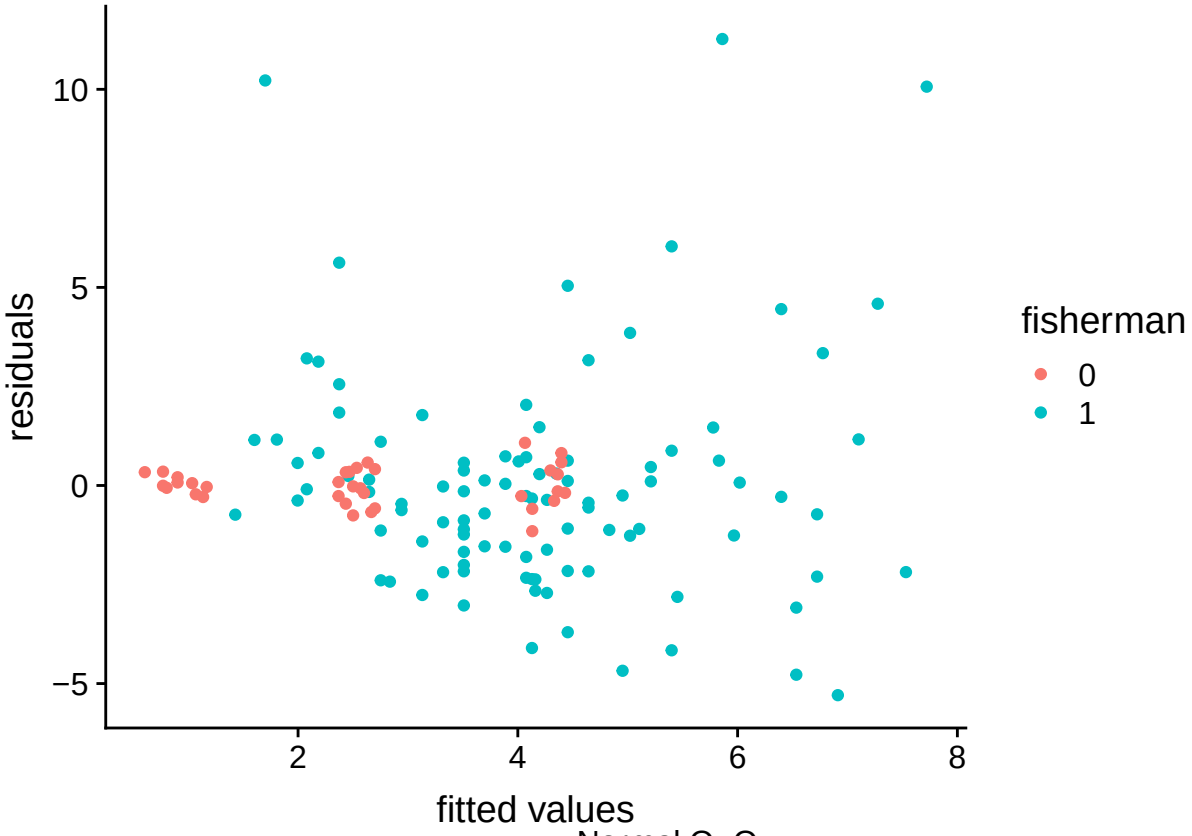
- The non-fisherman population may not have settled long enough in the place to have repercussion on the mercury levels.

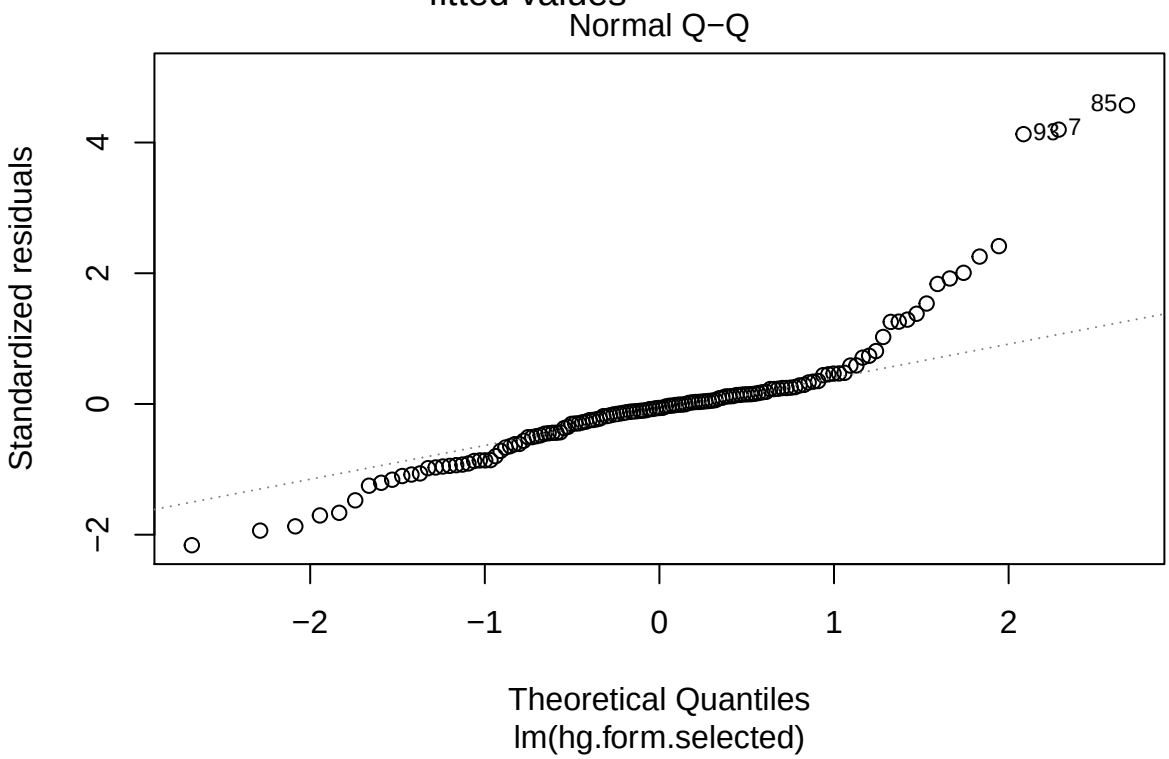
However, those are only suppositions in order to explain the distribution of the p-values, but none of them have been proven.

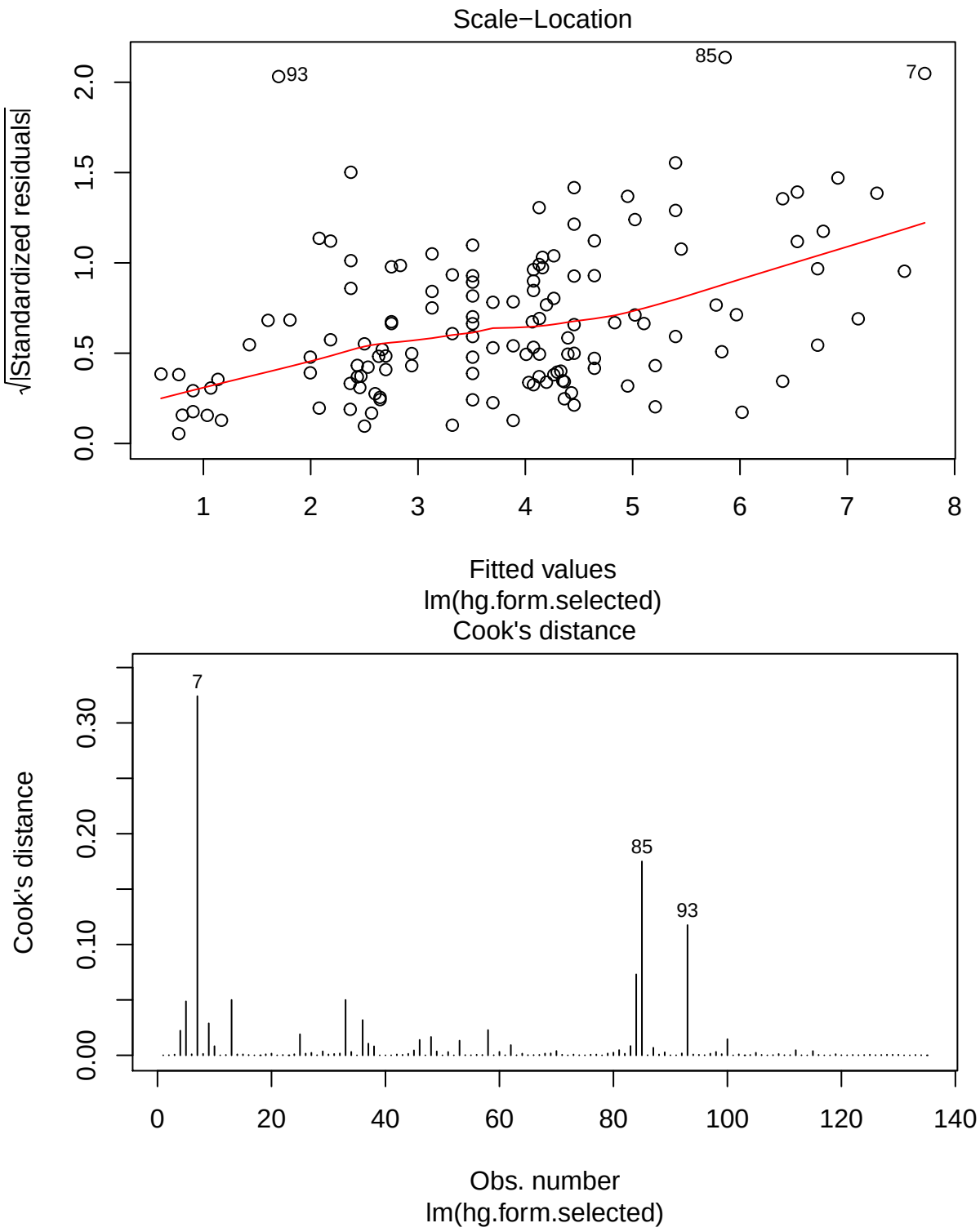
On the other side, the number of meals composed of fish (*fishmlwk*) seems to be contributing to the mercury levels of both populations (p-values of 0.03 for non-fisherman and 0.06 for fisherman). However, we can see that there is a huge difference between the two contributions of a factor 16.

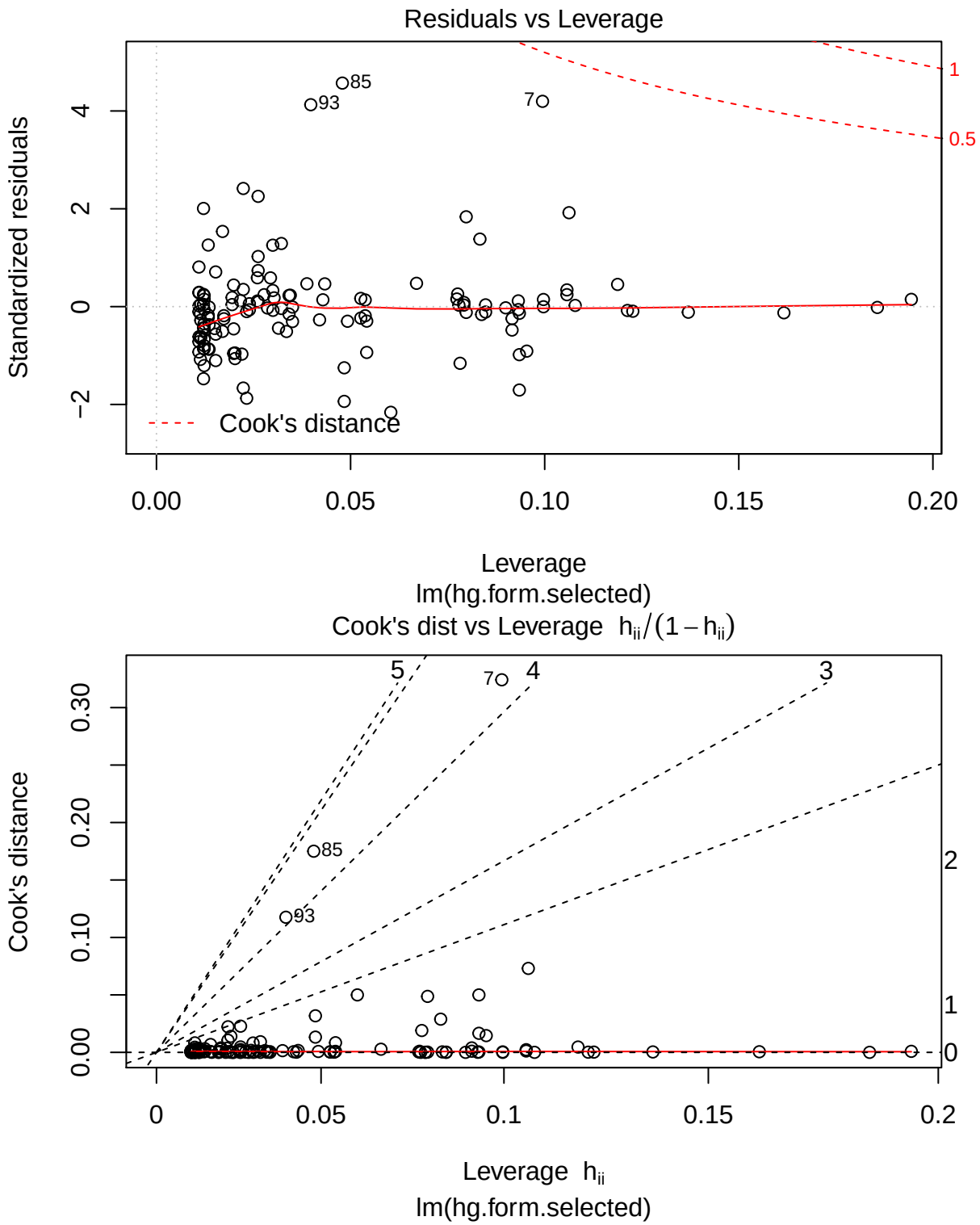
However, the most significant coefficient is for *weight* for fisherman population, with a p-value of 8e-07.




residuals
10
5
0
−5
2
4
6
8
fitted values
fisherman
0
1


Normal Q−Q
Standardized residuals
4
2
0
−2
−2
−1
0
1
2
Theoretical Quantiles
lm(hg.form.selected)
85
93
7


Scale−Location
√|Standardized residuals|
Fitted values
lm(hg.form.selected)
93
85
7
Cook's distance
Cook's distance
Obs. number
lm(hg.form.selected)



Conclusion (Lucile)

3 lignes

Les principales variables explicatives sont:

- La consommation de poisson => logique
- Le poids => plus étonnant mais on sait que les cellules graisseuses stockent les toxiques

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.

Park, Jung-Duck, and Wei Zheng. 2012. "Human Exposure and Health Effects of Inorganic and Elemental Mercury." *Journal of Preventive Medicine and Public Health* 45 (6). Korean Society for Preventive Medicine: 344.