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 greater average level of mercury than the control population (p-value = 7.473e-05).

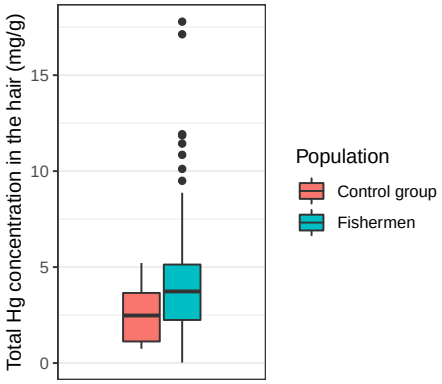


Figure 1: Boxplots of Hg levels

Exploratory analysis

Overview of the data

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

Number of fish meal per week	**0**	**1**	**2**	**3**	**4**	**7**	**14**	**21**
Fishermen	0	0	0	2	12	70	5	11
Non-fishermen	10	14	11	0	0	0	0	0

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 *fisherman* and the *fishmlwk* variable are very highly correlated: 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 variance inflation factor (VIF). We set the following criteria: we keep only the variables that have a result below 5. We observe that all the variables have a variance inflation factor below 2, so we don't eliminate any variable, yet.

Model selection

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.

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 has considerable impact on many 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.

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -10.0768196 2.44480942 -4.121720 6.603741e-05
## weight        0.1751823 0.03322237  5.273023 5.336914e-07
## fishmlwk      0.1588378 0.04174543  3.804915 2.161017e-04
```

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* has 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. We also include the *fisherman* variable itself, that will allow to include an adjustment term between the values of the intercepts of the two groups if it is relevant.

```
##                          Estimate Std. Error    t value     Pr(>|t|)
## fisherman0           -1.41565206 6.93853651 -0.2040275 8.386536e-01
## fisherman1          -10.41533001 2.65474931 -3.9232819 1.412643e-04
## fisherman0:weight     0.03313256 0.09907139  0.3344312 7.385973e-01
## fisherman1:weight     0.18908832 0.03644361  5.1885177 8.003708e-07
```

```
## fisherman0:fishmlwk   1.53107486 0.70200877  2.1809911 3.099742e-02
## fisherman1:fishmlwk   0.09828757 0.05265742  1.8665473 6.423521e-02
```

The best model is now : TotHg ~ fisherman + fisherman:weight + fisherman:fishmlwk - 1. While the coefficients for fisherman1:weight and fisherman0:fishmlwk stays very significant, fisherman1:fishmlwk is barely significant and fisherman0:weight, fisherman1 and the intercept are far from being significant. 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 (and reliable) effect compared to the weight influence, while among fishermen, the contribution of weight is twice the one of *fishmlwk*. Eventually the value of the fisherman1 parameter is surprising: it implies that the fact of being a fisherman gives you -8.99 mg/g Hg compared to non-fishermen. Yet this comes from the increased value of fisherman1:weight, which will make the overall Hg concentration more important among fishermen, as expected.

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

We are now discussing the results of the model: TotHg ~ fisherman + fisherman:weight + fisherman:fishmlwk - 1.

Difference between fisherman and control populations

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. Further experiments are needed in order to show whether or not those observables have a really different effect on both populations.

Number of fish meals per week

On the other side, the number of meal 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.

Here again, we can come up with some possible explanations, needing further inquiries:

- The distribution of *fishmlwk* is very different between both populations and thus may lead to different coefficients if the effect of this observables is not truly linear (*e.g.* a logarithmic effect that could take some ceiling effects into account, *i.e.* the fact that past a certain dose, the hair cannot absorb more mercury).
- The observable does not reflect entirely the quantity of fish eaten, since one can eat more or less fish per meal. The weight of fish eaten per week, might be a more accurate observable to study.

Here again, more experiments are required to confirm or reject those hypotheses.

Weight

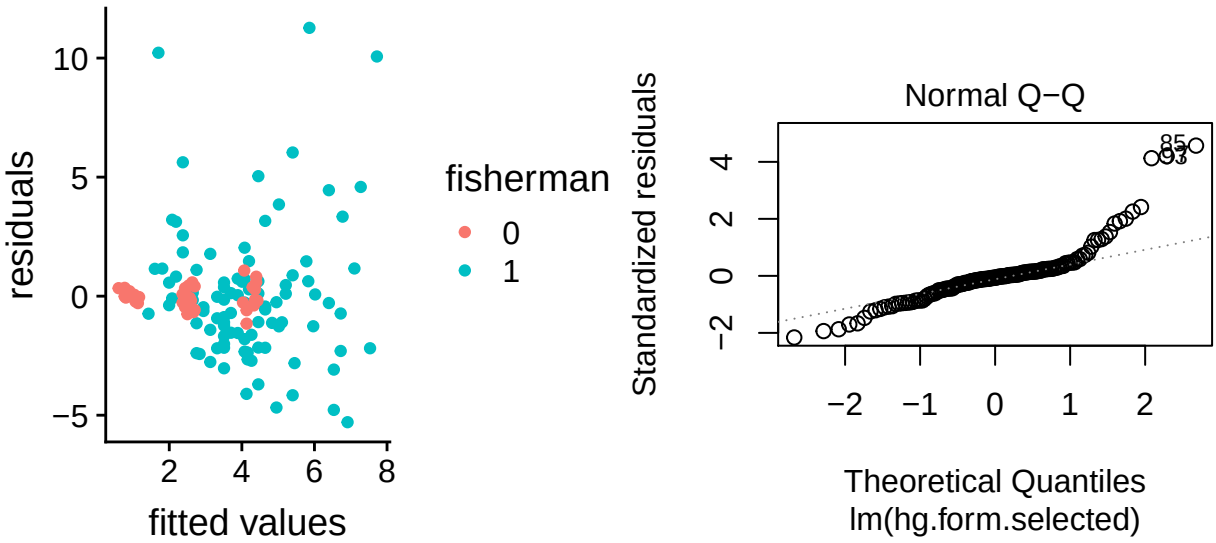
However, the most significant coefficient is for *weight* for fisherman population, with a p-value of 8e-07. This coefficient suggests a high positive correlation between the weight of the fisherman and the concentration of mercury in its hair.

The fact that the weight has a positive influence on this concentration was unexpected, since a concentration and not an absolute quantity was measured.

However, even though it was unexpected it has many possible explanations such as the fact that weight is much likely correlated with adiposity more susceptible to catch toxins than other tissues. Another explanation could be that the fatter, the more one eats and possibly ingests mercury that could fix in the hair; since hair weight isn't likely to be correlated with body weight, it could explain the high mercury concentration in hair.

Here again, further experiments are needed in order to support or reject those hypotheses.

Diagnostic plots



The model does not seem to be really homoscedastic. Variance is much higher for fishermen than for non-fishermen. However, within each class, the variance is overall well distributed, even if it tends to be a little more spread for high fitted values. But, since our model is split into two submodels corresponding to fisherman population and non-fisherman population, the difference of variances between both classes doesn't matter.

We have a heavy tails distribution of residuals, with a very heavy right tail. The fact that we have two submodels merged in one might explain the heavy tails. However it cannot explain the difference between right and left tails. This could be caused by an insufficient sample size. Or it could be explained by a non-linear relation between the observables and the concentration of mercury, thus possibly unbalancing the residuals distribution.

Conclusion

We have achieved to build a simple model that can help explaining the levels of mercury observed in a fishermen population compared to a control group. It appears that the variables that have the most significant influence over the measured levels of mercury are the weight of the individual and the frequency at which they eat fish. The former can seem surprising even if some hypotheses can be formed to account for the influence of weight on mercury levels. The latter may be the main explanation for the differences observed between our two groups: fishermen use to eat fish much more often than non-fishermen, as fish is a well-known source of mercury it seems logical to see a positive correlation between fish meal frequency and mercury levels and thus to observe higher mercury levels in fishermen populations compared to non-fishermen.

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.