Project 1: regression 1

Urvan Christen, Amandine Goffeney, Joseph Vermeil, Lucile ViguÃ©

March 27, 2019

Introduction

Mercury is a metal present in the environment whose harmful effects on human health are well known (Park and Zheng 2012). In (Al-Majed and Preston 2000), total mercury and methyl mercury levels in the hair of 100 fishermen of Kuwait, aged 16 to 58 years, were compared to those of a control population of 35 non-fishermen, aged 26 to 35 years. The aim of our study 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, since both variables are strongly correlated (shown in the paper).

The dataset contains six numerical variables (age, height, weight, number of fish meals per week and residence time in Kuwait) and two categorical variables (being a fisherman or not, fish consumption habits). All study participants are male.

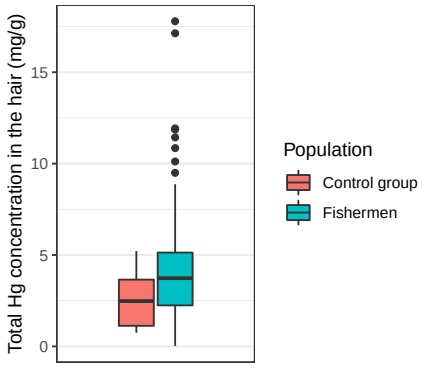


Figure 1: Boxplots of Hg levels

Exploratory analysis

Overview of the data

Before fitting a model, we first have a look at the data. Table 1 shows the distribution of individuals according to the number of fish meals per week and the two groups.

We note that for some values, there are very few people. For instance only two people eat fish three times per week. We also see that the number of fish meals per week is completely separable by population group. *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 do not eliminate any variable, for the moment.

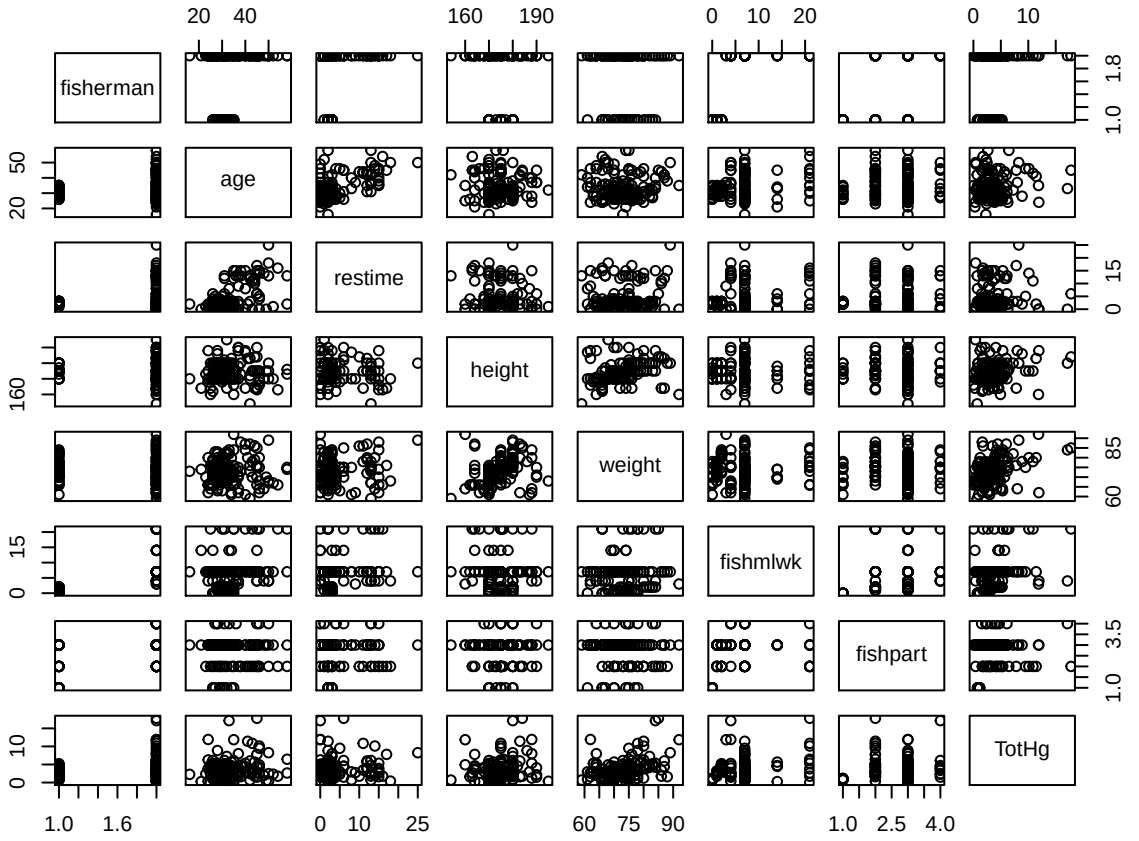


Figure 2: Pair plots

Table 1: Distribution of the number of fish meals accross fishermen and non-fishermen populations

	0	1	2	3	4	7	14	21
non-fisherman	10	14	11	0	0	0	0	0
fisherman	0	0	0	2	12	70	5	11

Table 2: Correlation matrix

	fisherman	age	restime	height	weight	fishmlwk	fishpart	TotHg
fisherman	1.00	0.25	0.25	-0.06	-0.09	0.61	0.46	0.23
age	0.25	1.00	0.58	0.00	0.05	0.26	-0.01	0.16
restime	0.25	0.58	1.00	-0.05	0.11	0.19	0.00	0.06
height	-0.06	0.00	-0.05	1.00	0.30	-0.04	-0.03	0.19
weight	-0.09	0.05	0.11	0.30	1.00	0.04	-0.05	0.41
fishmlwk	0.61	0.26	0.19	-0.04	0.04	1.00	0.19	0.30
fishpart	0.46	-0.01	0.00	-0.03	-0.05	0.19	1.00	0.11
TotHg	0.23	0.16	0.06	0.19	0.41	0.30	0.11	1.00

Model selection

We now use the different methods of model selection to select the more relevant variables to explain the *TotHg* variations within the population.

We first apply a stepwise selection based on a formula with all the variables. 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.

With stepwise selection, the selected model is : $TotHg = \beta_0 + \beta_1 \cdot fisherman_1 + \beta_2 \cdot age + \beta_3 \cdot restime + \beta_4 \cdot weight + \beta_5 \cdot fishmlwk$. The intercept and the two coefficients are highly 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 we 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.

The best model is now $TotHg = \beta_0 + \beta_1 \cdot fisherman_1 + \beta_2 \cdot fisherman_0 \cdot weight + \beta_3 \cdot fisherman_1 \cdot weight + \beta_4 \cdot fisherman_0 \cdot fishmlwk + \beta_5 \cdot fisherman_1 \cdot fishmlwk$. While the coefficients for *fisherman1:weight* and *fisherman0:fishmlwk* are 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

Table 3: Full model regression results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-12.68	2.71	-4.68	7.0e-06
fisherman1	1.11	0.65	1.70	9.1e-02
age	0.05	0.03	1.43	1.6e-01
restime	-0.08	0.05	-1.45	1.5e-01
weight	0.19	0.03	5.58	1.4e-07
fishmlwk	0.10	0.05	1.82	7.2e-02

Table 4: Model regression results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.42	6.94	-0.20	8.4e-01
fisherman1	-9.00	7.43	-1.21	2.3e-01
fisherman0:weight	0.03	0.10	0.33	7.4e-01
fisherman1:weight	0.19	0.04	5.19	8.0e-07
fisherman0:fishmlwk	1.53	0.70	2.18	3.1e-02
fisherman1:fishmlwk	0.10	0.05	1.87	6.4e-02

that for the non-fishermen, the number of fish meal per week has a very preponderant effect compared to the weight influence, while among fishermen, the contribution of *weight* is twice the one of *fishmlwk*. Eventually the value of β_1 (corresponding to *fisherman1* variable) 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 applied backward and forward selection and obtained similar models.

Results and discussion

We now turn to discussing the results of the model: $TotHg = \beta_0 + \beta_1 \cdot fisherman_1 + \beta_2 \cdot fisherman_0 \cdot weight + \beta_3 \cdot fisherman_1 \cdot weight + \beta_4 \cdot fisherman_0 \cdot fishmlwk + \beta_5 \cdot fisherman_1 \cdot fishmlwk$.

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

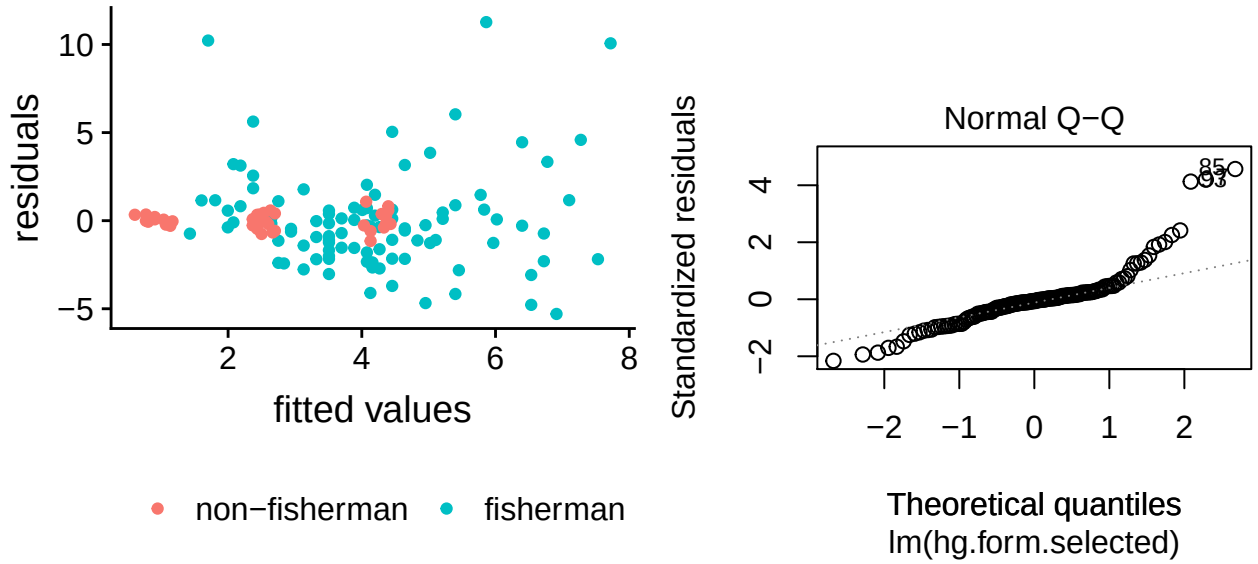
However, the most significant coefficient is for *weight* for fisherman population, with a p-value of 8e-07. This coefficient suggests a high positive correlation between the weight of the fisherman and the concentration of mercury in its hair.

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.

We have a heavy tailed distribution of residuals, with a very heavy right tail. It could be explained by a non-linear relation between the variables and the concentration of mercury.

Conclusion

We have built a simple model that can help to explain 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 though 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 eat fish much more often than non-fishermen and since 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.