Logit and Regression Analysis of Energy Consumption

Groupe 9

Midterm Assignment

42008 Introductory Econometrics F20

Prepared for: Niels Framroze Møller

Team Members and their contribution to the report :

Thomas Yarrington (s196618@student.dtu.dk) : contributed to R code implementation and report writing of Exercise 1 (especially questions 1 to 4)

Reda Lahlou (s192431@student.dtu.dk) : contributed to R code implementation and report writing of Exercise 1 (especially questions 4 to 7)

Silas Brack (s174433@student.dtu.dk) : contributed to R code implementation and report writing of Exercise 2

Pierre Puppnick (s191965@student.dtu.dk) : contributed to R code implementation and report writing of Exercise 2

Contents

Exercise 1: Logit analysis of energy consumption awareness

For the first part of this report, we will analyze a random selection of 1200 observations from the EEHA data set [2], which is analysed thoroughly in Baldini et al.[1] The data set contains a binary variable "$know_{el}$" which is a variable that returns a value of 1 if the consumers currently knows their annual electricity consumption, and a 0 if they do not. The goal of this section is to analyze what determines whether or not consumers are aware of their electricity consumption. Our Y_i discrete regressand is this variable $know_{el}$. The details of the models and the variables used are shown respectively in Table 1.19 and Table 1.20 in the appendix.

Question 1

First, we consider the "age" variable in the dataset representing the age of the consumer. To provide a representation of the data, we keep Y=$know_{el}$ and set X=age, and plot the conditional sample frequency of knowing the electricity consumption "knowing" $\hat{f}(y = 1|x)$ against age. To do this we must first cross tabulate our data. This gives us the following table (Table 1.1):

	18-29	30-39	40-49	50-59	60 or older
0	60	40	58	88	58
1	31	81	203	307	274

Table 1.1: Cross-tabulation showing Y=$know_{el}$ conditionally to X=age

We can easily see that for our data, the number of consumers between the age 18-29 that do not know their annual electricity consumption is 60. From here, we plot the joint sample frequency distribution on a 3 dimensional histogram Figure 1.1:

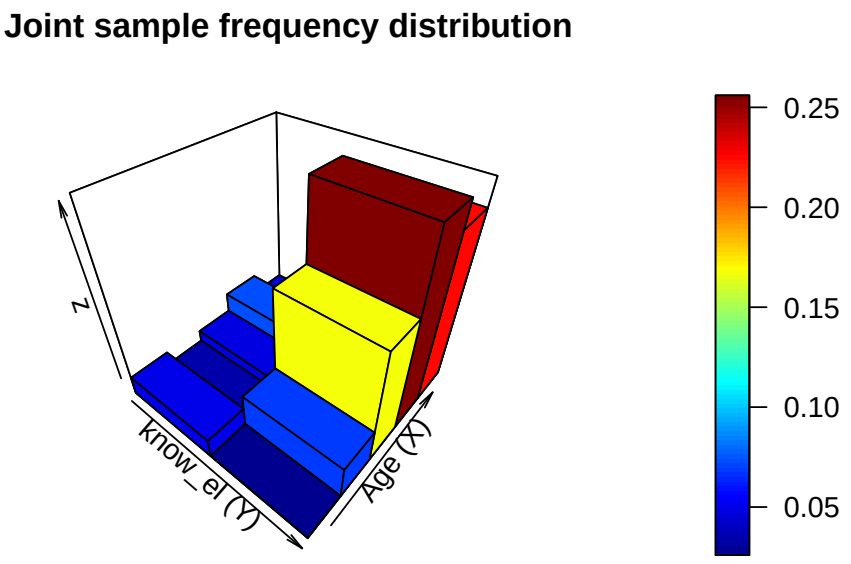


Figure 1.1: Joint sample frequency distribution of $Y = know_{el}$ conditionally to X=age

Table 1.2 shows which percentage of each age group knows their annual electricity bill with the sample conditional distribution of Y given X.

	18-29	30-39	40-49	50-59	60 or older
0	0.659	0.331	0.222	0.223	0.175
1	0.341	0.669	0.778	0.777	0.825

Table 1.2: Sample conditional distribution of $Y = know_{el}$ given X=age

It can be seen here that of the 60 or older age group, almost 83% of them know their annual electricity bill. To make this data more comprehensible, the sample frequency of knowing the annual electricity bill $\hat{f}(y = 1|x)$ against age is plotted below (left figure), as well as the odds of knowning (right figure):

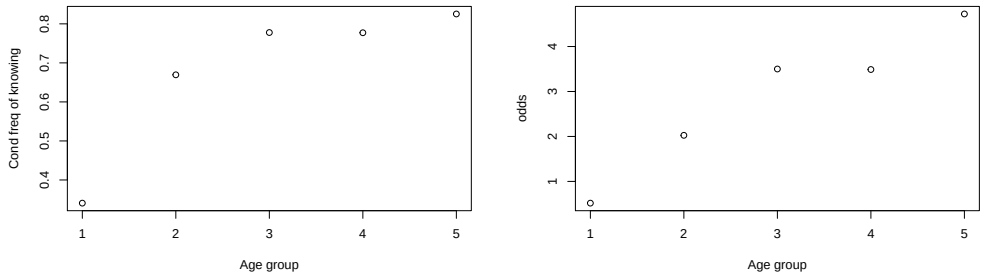


Figure 1.2: Sample frequency on "knowing" $\hat{f}(y = 1|x)$ against age (**left**) and odds of knowing against age (**right**)

We would now like to develop a figure which shows a prediction line for the odds that a certain age group will know their annual electricity consumption. To do this we simply take the log of the odds and plot a trendline as can be seen in Figure 1.3 below.

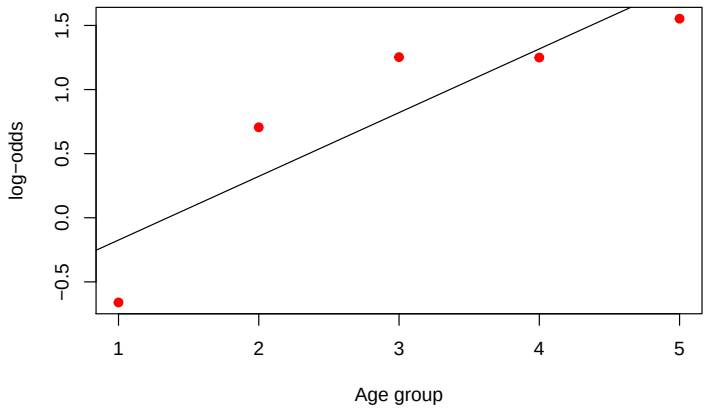


Figure 1.3: Log-odds of $Y = know_{el}$ depending on X=age

According to the logit model, the logarithm of the odds should be a linear function of the age group X. Figure 1.3 shows that the model does not provide a good fit (the value of the linear correlation coefficient is r = 0.89, which is low).

Question 2

Next we will consider the variable age, and construct age.n considering the simple logit model $p(X_i) = \beta_1 + \beta_2 X_i$ where X_i=age.n, where age.n represents the age group indexed from 0 to 4 instead of 1 to 5 (between 18 and 25 is now group 0 instead of group 1). We will refer to this model as M1.a. A summary of this model is tabularized below in Table 3:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.007	0.147	-0.045	0.964
age.n	0.438	0.055	7.920	0.000

Table 1.3: Summary of M1.a

It should be first noted that the p-values of the 2 estimates are very different. β_2 is significantly different from 0 (i.e. *"we are able to reject the null hypothesis, which is that β_2 is equal to 0"*), while β_1 is not (i.e. *"we are unable to reject the null hypothesis, which is that β_1 is equal to 0"*). Interpreting these estimates based on the logit model, this means that:

- The logarithm of the odds of knowing your electricity consumption given that you are between 18 and 29 (first group of age) is not significantly different from 0
- The change in the logarithm of the odds of knowing your electricity consumption from moving from one age group to the next one is significantly different from 0 (with an estimated change of $\beta_2 = 0.43$)

Going back to the log-odds $p(X_i)$, it can be shown that its derivative is expressed as follows:

$$\frac{\delta p(x)}{\delta x} = \beta_2 \frac{exp(\beta_1 + \beta_2 x)}{(1 + exp(\beta_1 + \beta_2 x))^2} \tag{1}$$

The sign of the derivative is thus given by β_2 (here negative).This confirms what has been hinted in question 1: the older one is, the higher the odds of them knowing their electricity consumption are.

Question 3

We will now test to see whether or not X_i should enter the logit function linearly, in other words, we will be testing for misspecification. To do this we have created dummy variables for the different age groups of age.n. We will refer to this model as M1. To test M1.a against M1, we have used a likelihood ratio test. This is possible because M1.a is nested in M1. Taking the notations from Table 1.4 (see below), M1.a can be obtained from M1 through the following restrictions: $a_2 = 2a_1$, $a_3 = 3a_1$ and $a_4 = 4a_1$

X.n	M1.a	M1 (general model)	M1 with restrictions
0	β_1	b	b
1	$\beta_1 + \beta_2$	$b + a_1$	$b + a_1$
2	$\beta_1 + 2\beta_2$	$b + a_2$	$b + 2a_1$
3	$\beta_1 + 3\beta_2$	$b + a_3$	$b + 3a_1$
4	$\beta_1 + 4\beta_2$	$b + a_4$	$b + 4a_1$

Table 1.4: Illustration of the restrictions needed to get from M1 to M1.a

The results of the likelihood ratio test between M1 and M1.a are summarized in Table 1.5 below:

#Df	LogLik	Df	Chisq	Pr(>Chisq)
2	-646.881	NA	NA	NA
5	-636.726	3	20.31	0

Table 1.5: Likelihood ratio test between M1 and M1.a

The p-value here is much lower than 5% (0 in the table indicate a p-value lower than 10^{-3}), therefore it can be concluded that the M1.a is significantly different from M1 (i.e. *"we are able to reject the null hypothesis, which is that the difference between the 2 likelihood ratio is equal to 0"*). In other terms, the general model M1 is a more accurate model than its restricted model M1.a.

Question 4

The log-odds plot in Figure 1.3 shows that the model M1.a does not seem to follow the main assumption of the logit model (i.e. that the log-odds of the regressand is a linear function of the explanatory variable). Furthermore, the lr-test in Table 1.5 shows that the general model M1 is significantly more accurate than the M1.a. Thus, it makes sense to augment M1.a. A model M1.b was created by adding a single dummy variable for the first age groupe called D.young. The model is then: know_el ~ age.n + D.young.

As M1.a, M1.b can be attained by restricting the general model M1. While M1.a had 2 degrees of freedom (related to 3 restricting equations for M), M1.b has 3 degrees of freedom (related to 2 restricting equations for M). The restrictions needed to get M1.b from M1 are: $a_1 = 2a_2 - a_3$ and $a_1 = 3a_3 - 2a_4$.

X.n	M1.b	M1 (general model)
0	$\beta_1 + \beta_0$	b
1	$\beta_1 + \beta_2$	$b + a_1$
2	$\beta_1 + 2\beta_2$	$b + a_2$
3	$\beta_1 + 3\beta_2$	$b + a_3$
4	$\beta_1 + 4\beta_2$	$b + a_4$

Table 1.6: Illustration of the restrictions needed to get from M1 to M1.b

We can test M1.b against M1 in the same way we did with M1.a and the likelihood ratio test. The results of this test are summarized in Table 1.7

#Df	LogLik	Df	Chisq	Pr(>Chisq)
3	-637.821	NA	NA	NA
5	-636.726	2	2.189	0.335

Table 1.7: Likelihood ratio test between M1 and M1.b

It can be seen that the p-value of this test is much higher than 5%, indicating that these models are not significantly different (i.e. *"we are unable to reject the null hypothesis, which is that the difference between the 2 likelihood ratio is equal to 0"*)

Question 5

Occam Razor's law states that, "If you have two equally likely solutions to a problem, choose the simplest." [3] Our model M1.b is simpler than M1 and we will therefore stick with it for now. We will now augment M1.b with more regressors from the data. We would like to see if the profile of a person based on their lifestyle plays a role in that person's knowledge of their annual electricity bill. Therefore we included *light_score* and *EE_index* - which are both indicators of energy performance, so possibly impacting our regressand -, number of inhabitants in the household $qty_{inhabitants}$, type of household *house_type*.1 - both sociological indicators, so possibly impacting our regressand, - and income - an economic indicator, so possibly impacting our regressand. The summary of this initial model $M1.c_0$ is shown in Table 1.8 below:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.033	0.390	-2.651	0.008
age.n	0.262	0.080	3.273	0.001
D.young	-1.005	0.326	-3.083	0.002
light_score	0.764	0.232	3.300	0.001
EE_index	1.124	0.389	2.888	0.004
qty_inhabitants	0.042	0.082	0.514	0.607
house_type.1farmhouse	0.524	0.299	1.751	0.080
house_type.1single family home	0.213	0.189	1.128	0.259
house_type.1town_SD_row	-0.290	0.221	-1.315	0.188
income	0.054	0.040	1.363	0.173

Table 1.8: Summary of $M1.c_0$

Some variables show a p-value above 5%, meaning that the estimate of their coefficient is not significantly different from 0. In particular, the levels of the variable *house_type*.1 appear as insignificant. We choose to first drop this variable, along with $qty_{inhabitants}$, which leads us to the model $M1.c_1$. Its summary is shown in Table 1.9 below:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.063	0.374	-2.842	0.004
age.n	0.263	0.074	3.536	0.000
D.young	-1.006	0.323	-3.114	0.002
light_score	0.778	0.229	3.390	0.001
EE_index	1.131	0.388	2.914	0.004
income	0.093	0.033	2.832	0.005

Table 1.9: Summary of $M1.c_1$

All the variables in this model are significant (p-value below 5%). However, before concluding that this is our final model, we need to perform a likelihood ratio test to compare it with $M1.c_0$. Indeed, it can happen that regressors are highly correlated. This means that dropping 1 variable from our model can affect the significance of the other variables - which is exactly what happened going from models $M1.c_0$ to $M1.c_1$ for the variable income, which became significant. To justify our choice of dropped variables, we need to make sure that the model M1.c we end up with is not significantly worse that the model $M1.c_0$ it is nested in - hence the likelihood ratio test shown in Table 1.10 below:

#Df	LogLik	Df	Chisq	Pr(>Chisq)
6	-622.262	NA	NA	NA
10	-617.555	4	9.414	0.052

Table 1.10: Likelihood ratio test between $M1.c_0$ and $M1.c_1$

The p-value of the test is above 5%, meaning that these models are not significantly different (i.e. *"we are unable to reject the null hypothesis, which is that the difference between the 2 likelihood ratio is equal to 0"*). Based again on Occam Razor's law, we should keep the simplest model, which is $M1.c_1$ - which we will call from now on M1.c. Based on the estimated coefficients, we can thus conclude that variables *light_score*, *EE_index* and income all positively affect the probablity of knowing the electricity bill.

Question 6

The aim of this question is to answer the following question: "does the number of kids you have have an impact on your knowledge of your electricity bill ?". This idea seems to be far-fetched, so our hypothesis would be that there is no significant impact. First, an indicator nk was created, which calculates the number of individuals below the age of 18 in each household. A model Mk.a was created by augmenting model M1.c by this variable. Its summary is shown in Table 1.11 below:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.102	0.380	-2.898	0.004
age.n	0.281	0.081	3.464	0.001
D.young	-0.953	0.336	-2.836	0.005
light_score	0.769	0.230	3.346	0.001
EE_index	1.121	0.389	2.884	0.004
income	0.088	0.034	2.603	0.009
nk	0.056	0.099	0.562	0.574

Table 1.11: Summary of Mk.a

As shown by its p-value above 5%, the new variable nk is not significant. In order to test for functional misspecification, we construct 6 binary variables nk_i, which are equal to 1 if the household has i number of kids - the maximum number of kids being 5. Augmenting M1.c by these variables, we end up with model Mk.b, which summary is shown below in Table 1.12:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	10.789	324.744	0.033	0.973
age.n	0.286	0.082	3.493	0.000
D.young	-0.947	0.339	-2.792	0.005
light_score	0.776	0.232	3.352	0.001
EE_index	1.187	0.391	3.031	0.002
income	0.090	0.034	2.637	0.008
nk0	-11.954	324.744	-0.037	0.971
nk1	-11.985	324.744	-0.037	0.971
nk2	-11.426	324.744	-0.035	0.972
nk3	-12.643	324.744	-0.039	0.969
nk4	-11.760	324.746	-0.036	0.971

Table 1.12: Summary of Mk.b

The new variables are all insignificant. If we dropped them all, we would end up with model M1.c. As a result, we should perform a likelihood ratio test between M1.c and Mk.b. This is shown in Table 1.13 below:

#Df	LogLik	Df	Chisq	Pr(>Chisq)
6	-622.262	NA	NA	NA
11	-617.986	5	8.552	0.128

Table 1.13: Likelihood ratio test between M1.c and Mk.b

The p-value is above 5%, meaning that we should stick with M1.c (the simplest model). This shows that the number of kids in a household does not have any significant impact on the knowledge of the electricity bill. Our hypothesis is validated.

Question 7

The aim of this question is to answer the following question: "does the zone you live in have an impact on your knowledge of your electricity bill ?". Rumors say that Jutlanders are more stingy than Sealanders. Taking this statement as our hypothesis, this means that we would expect either a significantly positive estimate for the coefficients for zones 6 to 9 (Jutlanders would know more about their electricity bill), and/or a significantly negative estimate for zone 1 to 4 (Sealanders would know less about their electricity bill) (see Appendix "Figure 1.4" for the Denmark's geographical zones) . An indicator zone.n was calculated, which is equal to the zone the household is in minus 1 (by doing so, the first level of zone.n is 0). A model Mz.a was created by augmenting model M1.c by these variables. Its summary is shown in Table 1.14 below:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.034	0.398	-2.598	0.009
age.n	0.263	0.074	3.542	0.000
D.young	-1.007	0.323	-3.117	0.002
light_score	0.776	0.230	3.383	0.001
EE_index	1.128	0.388	2.903	0.004
income	0.093	0.033	2.832	0.005
zone.n	-0.006	0.029	-0.217	0.828

Table 1.14: Summary of Mz.a

As shown by its p-value above 5%, the new variable z.n is not significant. In order to test for functional misspecification, binary indicators z.i were constructed, which are equal to 1 of the household is in zone i. Augmenting M1.c by these variables, we end up with model Mz.b, which summary is shown below in Table 1.15:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.910	0.409	-2.225	0.026
age.n	0.252	0.076	3.341	0.001
D.young	-1.028	0.327	-3.141	0.002
light_score	0.793	0.233	3.395	0.001
EE_index	1.086	0.394	2.752	0.006
income	0.089	0.033	2.677	0.007
z.1	-2.129	0.625	-3.406	0.001
z.2	-0.008	0.258	-0.030	0.976
z.3	0.493	0.409	1.205	0.228
z.4	0.172	0.261	0.659	0.510
z.5	-0.254	0.283	-0.897	0.370
z.6	-0.291	0.272	-1.070	0.285

	Estimate	Std. Error	z value	Pr(>|z|)
z.7	-0.090	0.290	-0.308	0.758
z.8	-0.246	0.253	-0.975	0.330

Table 1.15: Summary of Mz.b

Among the new variables, only z.1 is significant. We drop all the other new variables to build a model Mz.c. Its summary is hown in Table 1.16 below:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.986	0.377	-2.612	0.009
age.n	0.251	0.075	3.352	0.001
D.young	-1.037	0.326	-3.183	0.001
light_score	0.787	0.231	3.413	0.001
EE_index	1.109	0.391	2.840	0.005
income	0.091	0.033	2.759	0.006
z.1	-2.067	0.602	-3.432	0.001

Table 1.16: Summary of Mz.c

All the variables are significant. Before concluding, we still need to perform a likelihood ratio test, because we dropped several variables at once. In fact, 2 tests were conducted: one against Mz.b (see Table 1.17), and the other against M1.c (see Table 1.18):

#Df	LogLik	Df	Chisq	Pr(>Chisq)
7	-615.434	NA	NA	NA
14	-611.653	7	7.562	0.373

Table 1.17: Likelihood ratio test between Mz.c and Mz.b

#Df	LogLik	Df	Chisq	Pr(>Chisq)
6	-622.262	NA	NA	NA
7	-615.434	1	13.656	0

Table 1.18: Likelihood ratio test between Mz.c and M1.c

Table 1.17 shows a value above 5%, meaning that we should stick with model Mz.c rather than Mz.b (simplest model). Table 1.18 shows a value below 5%, meaning that our new model Mz.c has a likelihood ratio that is significantly different from M1.c. This is logical, as Mz.c is M1.c augmented by a significant variable (z.1)

This allows us to say that living in zone 1 has a significant impact on the knowledge of the electricity bill, since the estimate for the coefficient is negative (see Table 1.16). This means that people living in the region of Copenhaguen know significantly less than the rest their electricity bill. This is not exacly our initial hypothesis, but this is still an interesting result. At first sight, one could interpret this by saying that this is logical, since people living in Copenhaguen have a higher income - so they care less about their bills. Yet, our final includes both the variables *income* and z.1, and it explains the variance in the data better than the model M1.c, which did not have z.1 but still had *income*. Adding z1 improved the fit of the

model, which would not have been the case if the 2 variables were totally correlated. The indicator "living in the Copenhaguen area" brings extra information which helps explain the variance in the data in a way the variable *income* could not.

Exercise 2 : Regression analysis of energy consumption

For the second part of this report, we will look into the Residential Energy Consumption Survey (RECS) data [6]. This data collected by the U.S Energy Information Administration (EIA) contains energy-related information about households as well as demographic and economic ones. In this exercise we will analyse 300 random samples (out of 5686) from the RECS data set. The goal in this exercise is to build a predictive model of the electricity usage of an household based on other variables concerning this household in the RECS data set.

Question 1

First, we divided the regressand variable (Total electricity consumption in KWh : KWH) by the number of household members (NHSLDMEM), and then we logged the regressand variable as its distribution was left skewed. We then analyzed the resulting variable (LKWH.pers) by looking at its distribution :

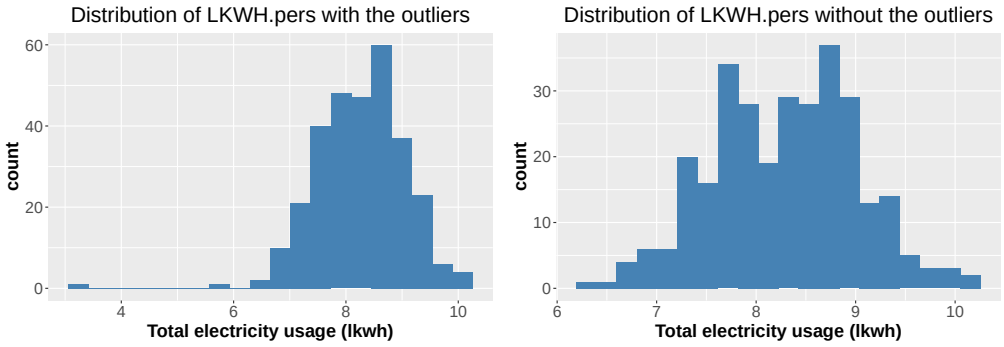
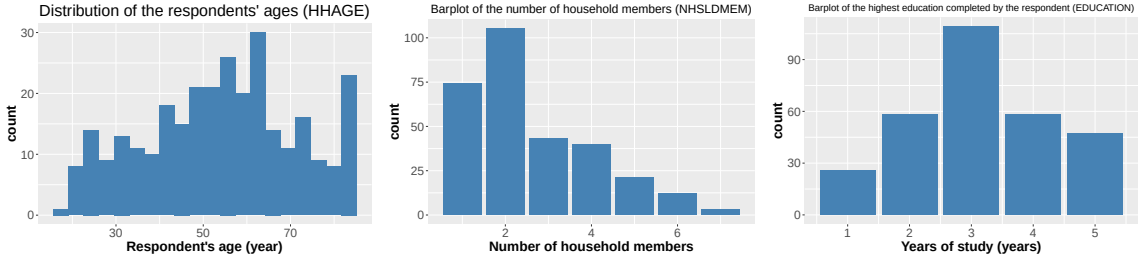


Figure 2.1 : Distribution of the total electricity usage of the household (LKWH.pers) with (left) and without (right) outliers

We can see that the distribution is approximately normal, with 2 outliers on the right of the main distribution. The 1^{st} outlier is the only sample to have "9" as the number of household members (NHSLDMEM), which could explain why it is an outlier, and we could not find any significant difference for the 2^{nd} outlier (even in the bigger set of variables). We decided to suppress the two outliers (values of LKWH.pers < 6), as the model we will build afterwards could be affected by those values. Furthermore, suppressing those two samples won't greatly affect the model, as we still have 298 samples remaining.

Then, we described (see Figure 2.2) the regressors (X_i) which we will implement later in the model to explain the total electricity usage of an household



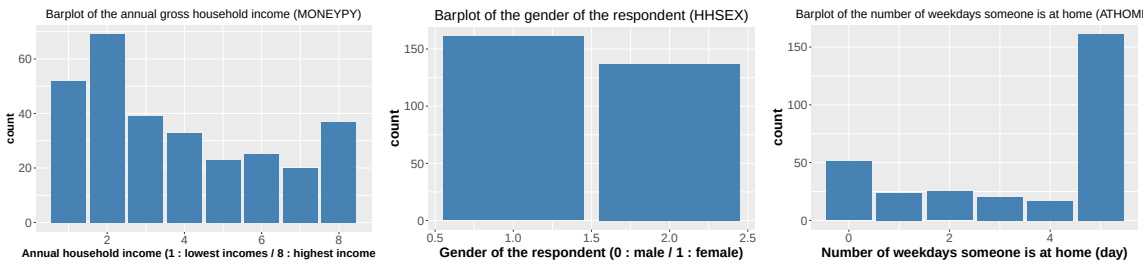


Figure 2.2 : Distribution of the regressors (HHAGE, NHSLDMEM, EDUCATION, MONEYPY, HHSEX and ATHOME)

For each of the variables, each of the sub-group are represented, therefore there is no issue with including these variables in the future model.

Question 2

After verifying and cleaning the variables, we first built a linear model with one regressor variable, the number of household members (NHSLDMEM) and the regressand being the log of the total electricity usage in the household (LKWH.pers). We called this model M2.a

Question 3

Here is the slope estimate $\hat{\beta}_2$ from Q2.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.862	0.077	115.735	0
NHSLDMEM	-0.225	0.026	-8.709	0

Table 2.1 : Summary of the regression : LKWH.pers NHSLDMEM

The zeros obtains in the p.value column are actually values $< 10^3$. Table 2.1 shows that the number of household members (NHSLDMEM) has a significant influence on the electricity usage of this household (LKWH.pers). Also, the higher the number of household members, the lower the electricity consumption per person ($\hat{\beta}_2 < 0$). This is understandable, as if there are more people in an household, then the electricity usage is shared by more people, which reduces the overall consumption per person (even though more electricity might be used overall, the electricity usage per head is lower).

Question 4

Here are the asumptions of the M2.a linear model :

i. Independence of the (X_i , Y_i) pairs
ii. Conditional normality
iii. Exogeneity
iv. Parameter space

First, given n data points, the pairs of data points (X1, Y1), (X2, Y2), . . . , (Xn, Yn) are mutually independent. The second assumption consists of the assumption of conditional normality of the regressand on the

regressors. The third assumption states that the conditioning variable is exogenous, meaning that this variable can be treated as fixed points without errors. Finally, the fourth assumption expresses the assumption that the statistical parameters used in the model are real numbers, existing in a real parameter space, with the constant conditional variance having the additional constraint of being positive.

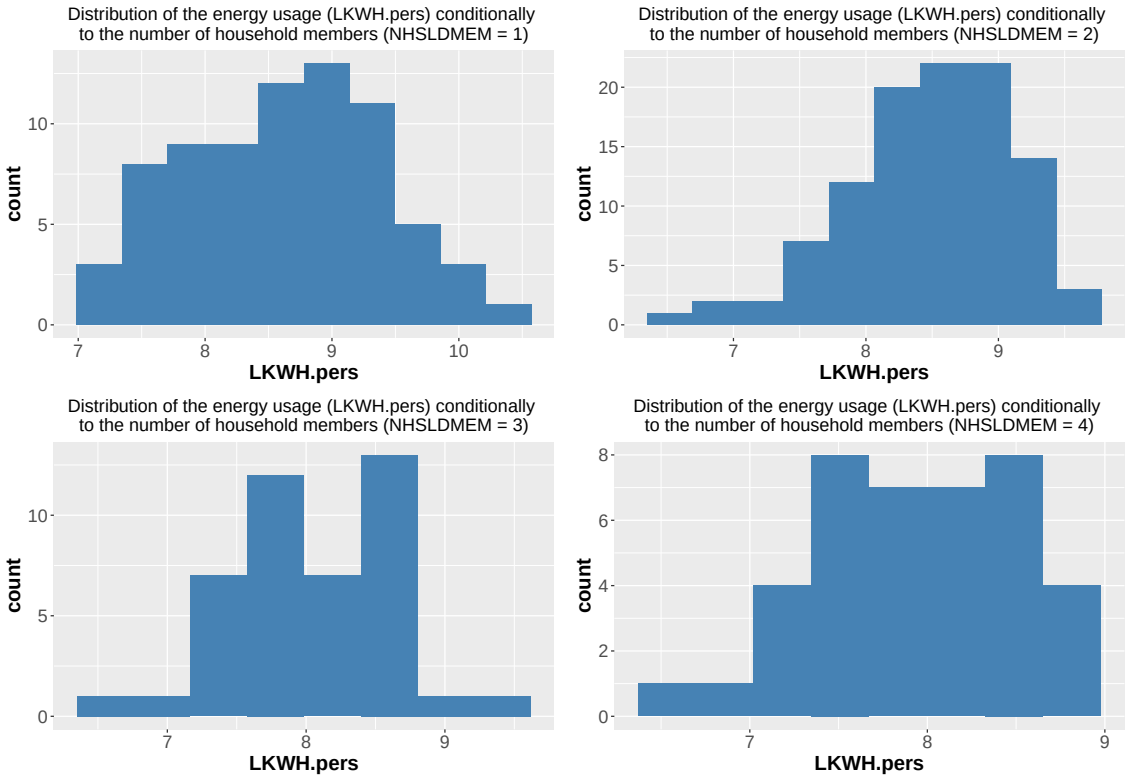
Additionally, (i) and (ii) together implicitly express the additional assumption that the residuals are also conditionally normally distributed and independent.

These assumptions are important to the process of inference. Notably, the assumption of normality of the residuals is important for the assumption of a normal distribution for the model parameters and is therefore important to maintain precision in the calculation of these parameters and of confidence intervals.

Assumption (iii) allows for conditional statements regarding fixed values for the regressors. These assumptions are therefore important for determining the posterior distribution for the model and the conditional expectation of the parameters.

Question 5

Supposing that the assumption of independance (i) and the assumption of exogeneity of all regressors (iii) hold, we tested the normality asumption (ii), first by having a look at the distribution of the number of the total electricity usage (LKWH.pers) conditionally to the number of household members.



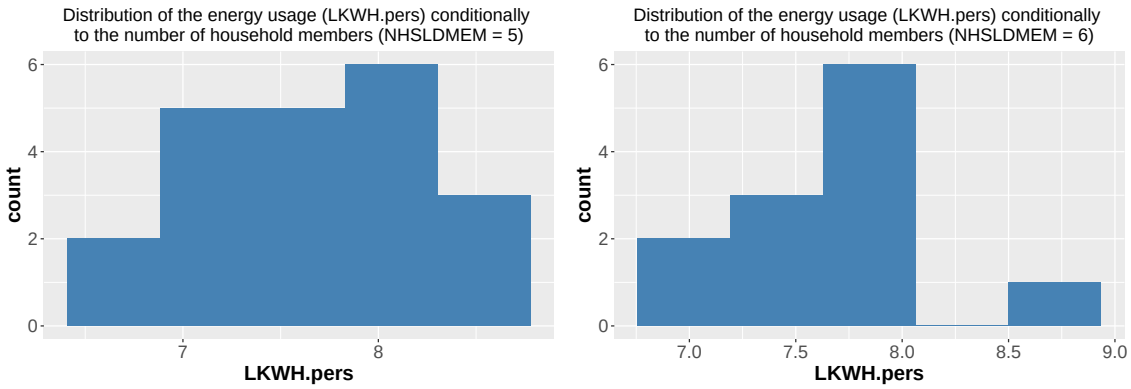


Figure 2.3 : Distributions of the energy usage (LKWH.pers) conditionally to the number of household members (NHSLDMEM = 1:6)

As there are really few samples (for X = 7 : 3 samples), the conditional distribution for the X = 7 subcategoy is not normal and does not look like a real distribution, but just some peaks. The plotted conditional distributions could be described as normal (except for X = 6), which partly confirm the conditionnal normality asumption (ii).

After, we assessed how the means of the samples vary for each category of NHSLDMEM.

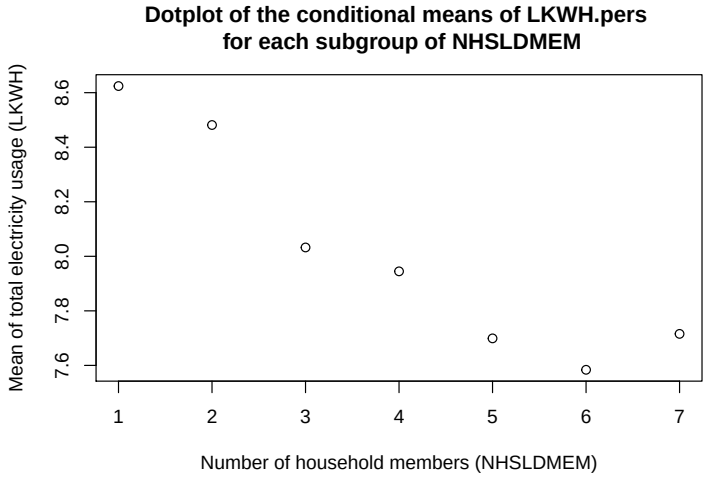


Figure 2.4 : Dotplot of the conditional means of LKWH.pers for each subgroup of NHSLDMEM

We can see from this graph that the sample mean decreases with the number of household members. This is coherent with the negative estimate we obtained at Q3, Table 2.1 ($\hat{\beta}_2$ = -0.225).

Then, we used the Jarque-Bera (JB) test to assess numerically if the residuals were distributed normally :

X-squared	df	p.value
4.255	2	0.119

Table 2.2 : Summary of the Jarque-Bera test on the residuals of the regression of LKWH.pers (Y) on NHSLDMEM

(X) (Table 2.1)

The p-value obtained is roughly 0.12. Therefore we cannot reject the null hypothesis (H_0 = "The residuals are normally distributed"), as the commonly choosen critical p-value to reject the null hypothesis is 0.05. Therefore, we can conclude that the residuals are normally distributed.

Question 6

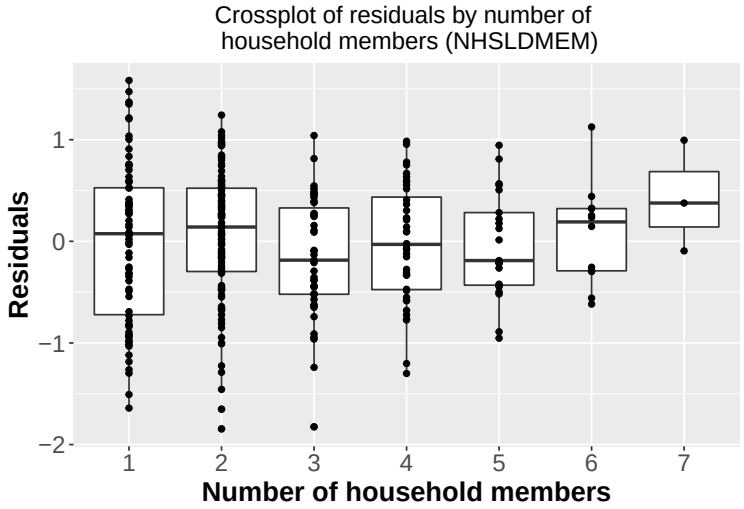


Figure 2.5 : Crossplot of the residuals by number of household members (NHSLDMEM)

The White's test gives us a F-statistic of 5.745, which correspond to a p-value of roughly 0.004. Therefore we can reject the null hypothesis (H_0 = "Homoskedasticity is respected"), and conclude that there is evidence of heteroskedasticity, which means that the conditional variance of the residuals given the number of household members varies with the number of household members. We could see this by looking at the "Crossplot of residuals by number of household members" just above (Figure 2.5), for example the residuals are much more spread for the subgroup "1" than for the other ones. The consequence of heteroskadasticity is that inferences made from the model could be misleading or false [3].

Question 7

Looking at the "Crossplot of residuals by number of household members" in Q6 (Figure 2.5), we can see that some residuals are extreme, such as the ones below the value of -1.5. When removing them, we obtain a higher p-value for the Jarque-Bera test (which means an even more normal distribution of the residuals), but the p-value for the White Test remain below 0.05, therefore the residual conditional variance still varies conditionally the number of household members, and we do not observe major changes by removing those suspected outliers.

	Jarque-Bera (p-val)	White (p-val)
All the samples (n = 300)	$< 10^{-16}$	0.3598
Q1's outliers removal (n = 298)	0.1191	0.0036
Residual outliers (< -1.5) removal (n = 293)	0.1318	0.0023

Table 2.3 : Results of the Jarque-Bera and White tests on different sample sets (outliers' removal)

However (as shown in Table 2.3), when computing the p-value of the Jarque-Bera test with all the 300 samples (with the two outliers removed in Q1), we obtain a p-value below 0.05, which confirms that we had to remove them from the initial set.

Question 8

In order to better specify the model, more variables were added to the model, variables which are relevant to explain the total electricity usage (LKWH.pers). Namely, we added : the highest education completed by the respondent (EDUCATION), the annual gross household income for the last year (MONEYPY), the respondent's gender (HHSEX), the respondent's age (HHAGE) and the number of weekdays someone is at home (ATHOME), each of these variables are further described in Table 2.9. This new model is called M2.b.

From the new Jarque-Bera test, we can conclude that with the addition of the new variables to the model, the residuals are even more normally distributed (the p-value with only one explanatory variable (NHSLDMEM) was 0.12 and now the p-value is roughly 0.30). Furthermore, from the new White test, we can conclude that with the addition of the variables, we cannot reject the null hypothesis anymore (H_0 = "Homoskedasticity is respected"), and therefore, the conditional variance doesn't vary significantly conditionally to each subgroup of each variable. From those two mis-specification test, we can conclude that the augmented model is well-specified.

Question 9

Interpretation of the estimates

As seen in the first model (M2.a), the number of household members significantly negatively impact the electricty usage per person (LKWH.pers), which is coherent as more people will benefit from the same energy comsumption (which will split the consumption between the recipients). The education (EDUCATION) also significantly impact negatively LKWH.pers, which means that higher educated people tend to consume less electricty than lower educated people. A simple explanation of this might be that higher educated people are more aware of the consequences of the use of electricity (global warming for example). The annual gross household income (MONEYPY) significantly impact positively LKWH.pers, which is cohenrent as the richer an household, the bigger the house might be and the more electrical appliances the household might own. As one could assume, the gender of the respondent is not significantly correlated to LKWH.pers, as the gender of the respondent does not condition the gender distribution within the household (which could be linked to LKWH.pers but it would remain to be proven). The respondent's age (HHAGE) significantly affect positively LKWH.pers, which is hardly explainable we would not think that it influence LKWH.pers (as with the respondent gender). After, further digging in the data, we realise that HHAGE is significantly linked to the number of children / adults (NUMCHILD / NUMADULT, so globally the age distribution in the household) in the household, which are variables making more sense in explaining LKWH.pers (however we will keep HHAGE in the model as it is a simpler variable). And finally, one might have expected the number of weekdays someone is at home (ATHOME) to be significantly correlated to LKWH.pers, however it is not. One reason could be that we do not know how many people are present at once. For instance, let's say five people might be staying at home on Mondays in one household while one person, in an other one, might be staying on its own on Mondays. So in this example, this variable (ATHOME) is not telling us enough, perhaps we would need to know the number of people staying at home each day for the variable to be more insighful.

Overall the regressors which impact LKWH.pers are (from the most impactful to the less impactful) : the number of household members (NHSLDMEM), the education of the respondent (EDUCATION), the annual gross household income (MONEYPY) and the age of the respondent (HHAGE).

We removed the non-significant varibles (HHSEX and ATHOME) from the model (M2.b_modif), and we can observe in Table 2.4 that their removal did not affect the other estimates and p-values.

	Estimate (with HHSEX/ATHOME)	Estimate (without HHSEX/ATHOME)	p.value (with HHSEX/ATHOME)	p.value (without HHSEX/ATHOME)
NHSLDMEM	-0.201	-0.200	$< 10^{-3}$	$< 10^{-3}$
EDUCATION	-0.101	-0.101	0.006	0.006
MONEYPY	0.058	0.056	0.002	0.002
HHAGE	0.011	0.011	$< 10^{-3}$	$< 10^{-3}$
HHSEX	-0.071	0	0.331	NA
ATHOME	-0.008	0	0.679	NA

Table 2.4 : Results of the modification of the model M2.b (estimates and p-values) before and after removal of HHSEX and ATHOME

We chose to add two other relevant variables to our model (M2.c). DIVISION which describes the location of the respondent's dwelling and TYPEHUQ which describe the type of housing. Two factors which influence the electricity consumption according to Jones et al. 2015 [4].

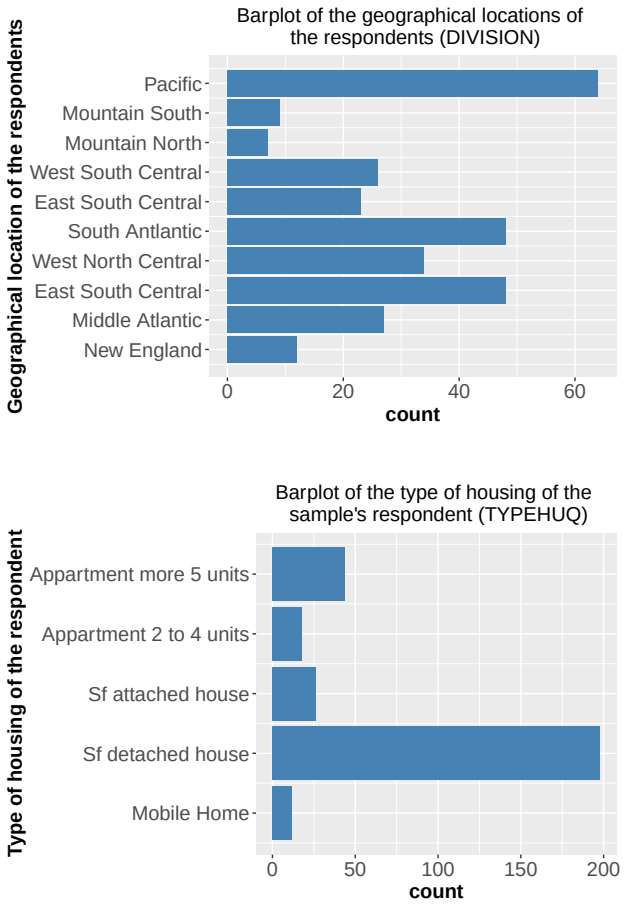


Table 2.6 : Barplot of the two additionnal variables (DIVISION and TYPEHUQ)

As shown in the barplot in Table 2.6, DIVISION and TYPEHUQ have no major problem and they are both significantly linked to LKWH.pers (p-value of the F-statistic equals to $7.66 * 10^{-10}$ and $9.23 * 10-6$ respectively), as mentionned in Jones et al. 2015 [4].

Question 10

Table 2.5 shows that, concerning the respondents' geographical location, respondents living in East South Central, West North Central, South Antlantic and West South Central use significantly more electricty than respondents in other locations (this is coherent with Figure 2.7 in appendix, where we can see that in these precise census divisions, electricity consumption is higher than anywhere else in the US). Concerning the type of housing, we can see that respondents leaving in mobile home, single-family attached houses and single-family detached houses use significantly more electricity than the other respondents. We therefore created an other model (M2.c_modif) in which we kept only the variables of the M2.c model significantly influencing the electricity usage of a household (LKWH.pers).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.014	0.190	42.188	0.000
NHSLDMEM	-0.211	0.024	-8.735	0.000
EDUCATION	-0.062	0.032	-1.897	0.059
MONEYPY	0.030	0.017	1.781	0.076
HHAGE	0.008	0.002	3.539	0.000
New.England	-0.208	0.175	-1.192	0.234
Middle.Atlantic	-0.123	0.132	-0.936	0.350
East.South.Central	0.439	0.138	3.193	0.002
West.North.Central	0.264	0.121	2.187	0.030
South.Antlantic	0.430	0.111	3.864	0.000
West.South.Central	0.482	0.132	3.657	0.000
Mountain.North	-0.201	0.219	-0.915	0.361
Mountain.South	-0.222	0.197	-1.125	0.262
Pacific	-0.088	0.104	-0.843	0.400
Mobile.Home	0.485	0.175	2.780	0.006
Sf.detached.house	0.471	0.091	5.170	0.000
Sf.attached.house	0.312	0.129	2.415	0.016

Table 2.5 : Summary of the M2.c model

To assess if a regressor variable is significantly linked to a regressand, one can also perform a t-test. We performed the t-test on the number of household members (NHSHLDMEM), with the null hypothesis being $H_0 : \beta_2 = 0$ and the alternative hypothesis being $H_a : \beta_2 < 0$ (as we expect that the more household members there are, the less electricity consumption per person there will be, see Table 2.1 in Q1). The t-test gives us a t-value of -8.71 which correspond to a p-value of $2.21 * 10^{-16}$. Therefore, we can reject the null hypothesis and opt for the H_a ($\beta_2 < 0$), confirming the result we obtained before (Table 2.1)

Question 11 / 12

In order to know if the addition / removal of one or multiple regressors impact significantly our model, one can perform a Log-Likelihood Ration Test in order to compare two different models. We compared the 5 models (cf. Table 2.9 for the details of the models) built during this exercise (i.e : M2.a, M2.b, M2.b_modif, M2.c and M2.c_modif). Here are the p-values associated to the log-likelihood ration test statistic, for the different comparisons :

	M2.a vs. M2.b	M2.b vs. M2.b_modif	M2.b_modif vs. M2.c	M2.c_modif vs. M2.c
p-value	$9.51 * 10^{-7}$	0.55	$6.73 * 10^{-6}$	0.77

Table 2.6 : Results of the LR test (p.value of the log likelihood ratio test statistic) comparing 4 models (M2.a, M2.b, M2.b modif and M2.c)

- M2.a vs. M2.b : we obtain a p-value of $9.51 * 10^{-7}$, we can therefore reject the null hypothesis (H_0 : "The estimates of the added varibales equal 0"), and conclude that the added variables have a significant impact on total electricity usage of a household.
- M2.b vs. M2.b_modif : we obtain a p-value of 0.55, therefore we cannot reject the null hypothesis, which means that the two variables (ATHOME and HHSEX) do not add any significance to the model (as we have concluded in Q9. Table 2.4).
- M2.b_modif vs. M2.c : we obtain a p-value of $6.73 * 10^{-6}$, which tells us that the two variables (REGIONC and TYPEHUQ) matter in explaining the electricity usage of a household.
- M2.c_modif vs. M2.c : for this last LR-test, we obtain a p-value of 0.77, which means that there is no significant different between the two model, we will thereofre keep simple simplest one (Occam Razor's law)

Question 13

Our final model is M2.c_modif (Table 2.7). All the selected variables are significantly linked to the total electricity usage of an household.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.940	0.181	43.834	0.000
NHSLDMEM	-0.211	0.024	-8.826	0.000
EDUCATION	-0.064	0.032	-1.989	0.048
MONEYPY	0.031	0.017	1.854	0.065
HHAGE	0.007	0.002	3.477	0.001
East.South.Central	0.529	0.121	4.387	0.000
West.North.Central	0.353	0.101	3.482	0.001
South.Antlantic	0.520	0.089	5.832	0.000
West.South.Central	0.571	0.114	5.020	0.000
Mobile.Home	0.475	0.173	2.753	0.006
Sf.detached.house	0.477	0.089	5.355	0.000
Sf.attached.house	0.323	0.128	2.518	0.012

Table 2.7 : Summary of M2.c modif

One can then wonder if the parameters, we selected are structural parameters. In econometrics, in order for a parameter to be structural, (1) it needs to be invariant upon intervention on the economy (new regulation for example), (2) it should not vary when the sample is extended and (3) it should not be derived from more basic parameters. In this report, we have not tested asumptions (1) and (2). However, we have discussed the asumpiton (3) in Q9, where we have concluded that HHAGE (the age of the respondent) was significantly correlated with the number of children or adults in a household, these variables being therefore variables from which HHAGE is derived. So in that sense, HHAGE is not structural. Furthermore, to determine if a variable is structural, we used the LR-test which showed us the essential variables in the models, which led us to M2.c_modif. Also removing some variables of the model and assessing the impact on the estimates of the other variables (see Table 2.4 in Q9), was one way of assessing if our parameters are structural and therefore that they are not linked to one another.

References

[1] M. Baldini, A. Trivella, and J. W. Wente. "The impact of socioeconomic and behavioural factors for purchasing energy efficient household appliances: A case study for Denmark". In: *Energy Policy* (2018). ISSN: 03014215. DOI: 10.1016/j.enpol.2018.05.048.

[2] EEHA. *EEHA Data Set*. <URL: https://cn.inside.dtu.dk/cnnet/filesharing/download/ca146e8e-3895-4ae5-9fa9-9216d496701a. (n.d.).>.

[3] D. F. Hendry and B. Nielsen. *Econometric modeling: A likelihood approach*. 2012. ISBN: 9781400845651.

[4] R. V. Jones, A. Fuertes, and K. J. Lomas. *The socio-economic, dwelling and appliance related factors affecting electricity consumption in domestic buildings*. 2015. DOI: 10.1016/j.rser.2014.11.084.

[5] W. M. Thorburn. *Occam's Razor*. 1915, pp. 287-288.

[6] U.S Energy Information Administration (EIA). *Residential Energy Consumption Survey (RECS) data*. 2015. <URL: https://www.eia.gov/consumption/residential/data/2015/index.php?view=microdata.>.

Appendix

Model name	Model variables
M1.a	age.n
M1	age.n.0 / age.n.1 / age.n.2 / age.n.3 / age.n.4
M1.b	age.n / D.young
$M1.c_0$	age.n / D.young / *light_score* / *EE_index* / *qty_inhabitants* / *house_type* / income
$M1.c$	age.n / D.young / *light_score* / *EE_index* / income
$Mk.a$	age.n / D.young / *light_score* / *EE_index* / income / nk
$Mk.b$	age.n / D.young / *light_score* / *EE_index* / income / nk0 / nk1 / nk2 / nk3 / nk4
$Mz.a$	age.n / D.young / *light_score* / *EE_index* / income / zone.n
$Mz.b$	age.n / D.young / *light_score* / *EE_index* / income / z.1 / z.2 / z.3 / z.4 / z.5 / z.6 / z.7 / z.8
$Mz.c$	age.n / D.young / *light_score* / *EE_index* / income / z.1

Table 1.19 : Summary of the models used in Exercise 1

Variable name	Variable meaning
$know_{el}$	Knowledge of own electricity consumption: 1 if yes, 0 if no
age	Age of the consumer (5 levels: 18–29, 30–39, 40–49, 50–59, 60 or older)
light_score	Energy efficiency lighting ownership, in [0,1]
EE_index	Behavioural energy efficiency index, in [0,1]
qty_inhabitants	Number of household inhabitants
house_type	4 levels: apartment, farmhouse, single house, townhouse
income	Gross household income

Table 1.20 : Names and meanings of the variables used in Exercise 1

Model name	Model variables
M2.a	NHSLDMEM
M2.b	NHSLDMEM / EDUCATION / MONEYPY / HHAGE / HHSEX / ATHOME
M2.b_modif	HSLDMEM / EDUCATION / MONEYPY / HHAGE
M2.c	NHSLDMEM / EDUCATION / MONEYPY / HHAGE / DIVISION (each category) / TYPEHUQ (each category)
M2.c_modif	NHSLDMEM / EDUCATION / MONEYPY / HHAGE / East.South.Central / West.North.Central / South.Antlantic / West.South.Central / Mobile.Home / Sf.detached.house / Sf.attached.house

Table 2.8 : Summary of the models used in Exercise 2

Variable name	Variable meaning
KWH	Total site electricity usage, in 2015, in kilowatthours
LKWH.pers	Logarithm of KWH/NHSLDMEM where KWH is total electricity usage in kwhs.
NHSLDMEM	Number of household members
EDUCATION	Highest education completed by respondent (1-5)
MONEYPY	Annual gross household income for the last year (1-8)
HHSEX	Respondent gender : 0 if male and 1 if female.
HHAGE	Respondent age, 18-110
ATHOME	Number of weekdays someone is at home (1-5)
DIVISION	Census region of the respondent (house) : 1 : New England / 2 : Middle Atlantic / 3 : East North Central / 4 : West North Central / 5 : South Atlantic / 6 : East South Central / 7 : West South Central / 8 : Moutain North / 9 : Moutain South / 10 : Pacific
TYPEHUQ	Type of housing unit of the respondent (1 : Mobile Home / 2 : Single-family detached house / 3 : Single-family attached house / 4 : Appartment in a building with 2 to 4 units / 5 : Apartment in a building with 5 or more units)

Table 2.9 : Names and meanings of the variables used in Exercise 2

Danmarkskort med postnumre og grænser

Figure 1.4 : Denmark's geographical zones

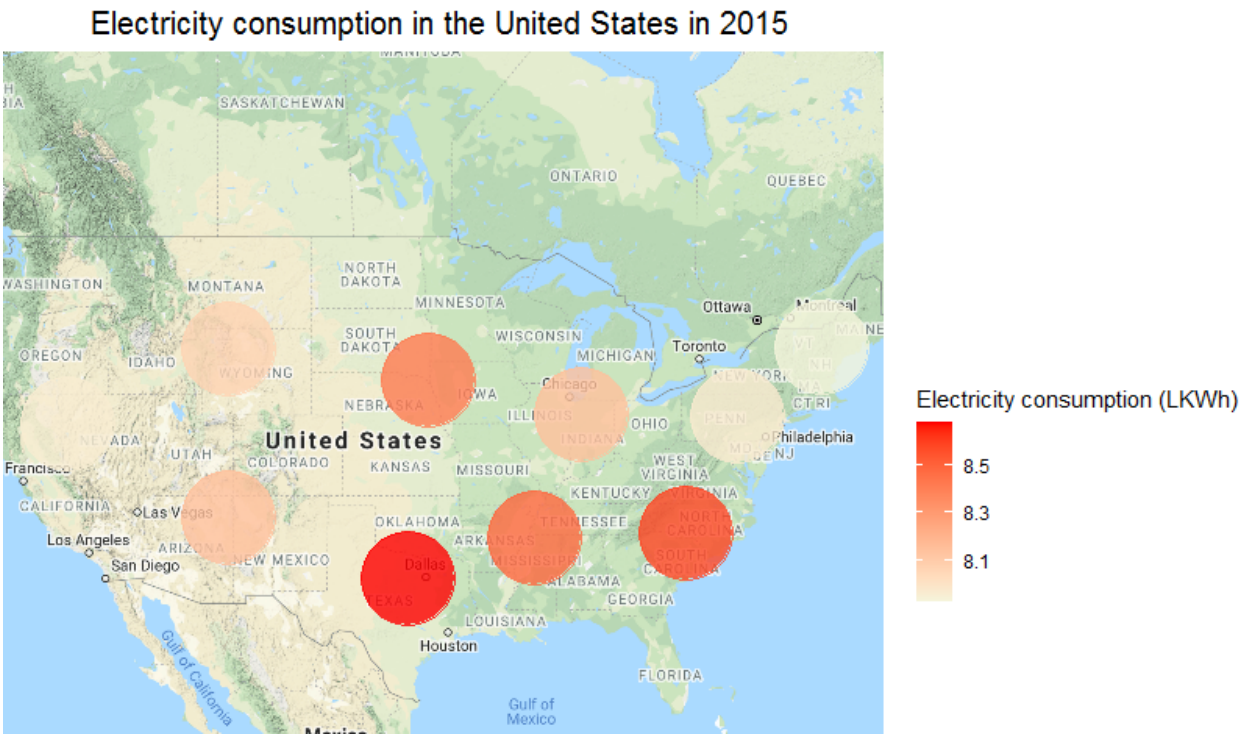


Figure 2.7 : Electricity consumption in the United States in 2015 (KWh)