Advanced Statistics for Finance Project

Silvia Adaggio* Beatrice Saccucci†

Paolo Manenti‡ Alessandro Lanzo§

May 24, 2022

Identification numbers:
*2011167
†1786623
‡1999957
§1995464

Contents

1 Introduction

From the data survey "Household finance and consumption (HFCS)" conducted in 2016 by the Bank of Italy, we have aimed to proceede with our statistical investigation trying to analyse the wealth through the explaination of the "Current price of the main residence" considering seven different regressors. The phenomena has been studied through a sample of 4176 observations. We start explaining it in relation to the behaviours of the other phenomena through the implementation of a multiple linear regression model. This type of model better fits the collective phenomena, due to their complex and numerous relationships. The dependent variable Y can hence be expressed as a function of the other explanatory variables, also with the presence of an error term ϵ that shows the not captured information in the regression and henceforth, all the other factors other than the Xs that can affect Y. The multiple regression model equation can be defined as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon \tag{1}$$

where:

- Y: phenomenon of interest
- B_i (i=0, 1, ..., n): parameter estimation
- X_j (j=1, ..., n): explanatory variables
- ϵ: error term

2 Variables definition

The data we are working with can be identified as cross-sectional: a collection of data obtained observing many subjects through the same period of time (2016). We wanted to explain the "Current price of the main residence" through:

- *Square meters of the house* that is referred to each main residence considered in the survey. What is asked is to give the precise measure in square meters of the accomodation the interviewee spends the most of the time during a year.
- *Years spent in the household* may catch the aptitude of the interviewee of changing its residence or not, by specifying the precise number of years spent in it.
- *Total value of cars* represents the value that may be obtained by selling the interviewee own cars.

- *Value of sight accounts* that is given by the precise amount of money owned in these accounts. It may be considered one of the main wealth indicators.
- *Total gross amount from financial investments* it considers the amount obtained from the financial investments of the last 12 months.
- *Food expenses* is referred to the mean of the total amount of money spent in food and beverages at home each month, calculated for 12 months in total.
- *HMR/imputed rent* it considers the possible amount of money the interviewee is willing to pay monthly in order to rent its own house.

For a easier comprehension during the analysis, we are naming the covariates used with their X correspondent number.

3 Model fitting analysis

In the light of the above, the multiple linear regression model is more amenable to ceteris paribus analysis because it allows to check how and how many other factors simultaneously affect the dependent variable, that in our specific case is the current price of the main residence. As happens in the simple regression, it is possible to define:

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{2}$$

$$SSE = \sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2 \tag{3}$$

$$SSR = \sum_{i=1}^{n}\hat{u_i}^2 \tag{4}$$

where they respectively represent the total sum of squares, the explained sum of squares and the residual sum of squares. Computing the ratio between the SSR and the SST, it gives us the R^2. It represents a measure of the linear relationship between the regressors and the response variable. This statistics allows us to check how well the model has fitted the analysed data.

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

For its interpretation, it is necessary to consider that its value is always between 0 and 1 where, if the result is near 0 it does not represent a good fit of the regressors onto the

explained variable, and henceforth following this reasoning, if it is near 1, otherwise. Moreover, it generally increases when including more variables in the model given that the SSR decreases, but sometimes is better to rely on another version of the R^2 defined as the $adjustedR^2$ because it properly adjusts its previous version considered. It is obtained as:

$$\bar{R}^2 = 1 - \frac{SSE/(n-k)}{SST/(n-1)} \tag{6}$$

considering n as the size of the observed sample and k as the number of the explanatory variables.

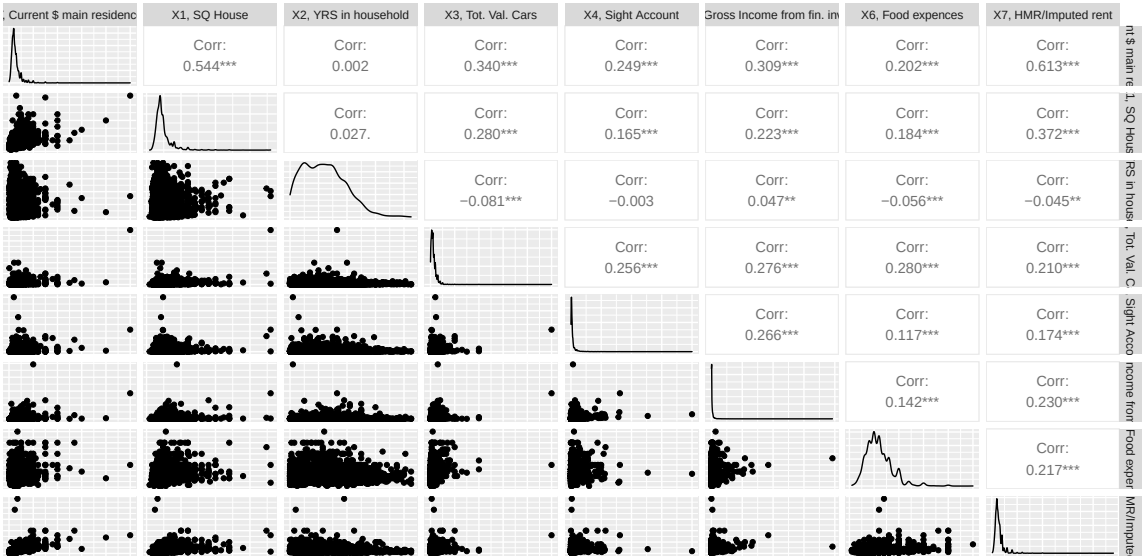

Figure 1: Correlation matrix

Through the above shown matrix, it is possible to highlight the different levels of correlation between the dependent variable and the covariates, but also between the covariates themselves. Looking at the results obtained it is easy to say that all the regressors choosen are highly statistically significant but for "Years spent in the household". Anyways, we wanted to include it in the very first model to see how it behaves compared to the others.

	Int.	*log(X1)*	*X2*	*log(X3)*	*X4*	*log(X5)*	*log(X6)*	*log(X7)*	*Adj R2*
Mod. 1	5.60e+0	4.50e-1	-2.06e-4	8.91e-3	7.77e-7	2.38e-2	-4.87e-3	6.89e-1	0.577
	<2e-16***	<2e-16***	0.620	0.168	0.000***	0.000***	0.741	2e-16***	
Mod. 2	5.60e+0	4.49e-1	-	8.96e-3	7.89e-7	2.34e-2	-	6.90e-1	0.578
	<2e-16***	<2e-16***	-	0.145	0.000***	0.000***	-	<2e-16***	

Signif. codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '·', 0.1 ' ', 1

Table 1: Basic models

Subsequently to the consideration of the correlation matrix and the following elimination of the covariate $X2$ (Figure 1) in the following model, we computed a table to compare Model 1 and Model 2. Firstly, given the high level of discrepances of the values of the observations shown, we applied the log transformation to the most of the covariates considered, but for $X2$ and $X4$. Given the results obtained that highlight the various significance levels, we also deleted the $log(X6)$; with this change, an slight improvement of the model has been obtained, and it can be easily said by looking at the values of the Adjusted R^2.

4 Residual analysis

The analysis of residuals plays an important role in validating the regression model. The i_{th} residual is the difference between the observed value of the dependent variable, y_i, and the value predicted by the estimated regression equation, $\hat{y}_i$. These residuals, computed from the available data, are treated as estimates of the model error, ϵ. As such, they are used by statisticians to validate the assumptions concerning the error term. Henceforth, it is important to focus on the residuals to check, afterwards, the plausible presence of heteroskedasticity and autocorrelation that may influence the OLS estimation.

Going on with the analysis, we start with the analysis of raw residuals, that are exactly the residuals obtained from the linear model with regard the observed variables. In the second graph, we proceed by dividing these raw residuals by the standard deviation in order to get the so called standardized residuals which play an important role in the Normality assumptions and into the removing heteroskedasticity step. Lastly, we consider the studentized residuals that are obtained from ratio between the removed residuals and the standard deviation. In our case, the results of these three type of residuals are shown in the following graphs.

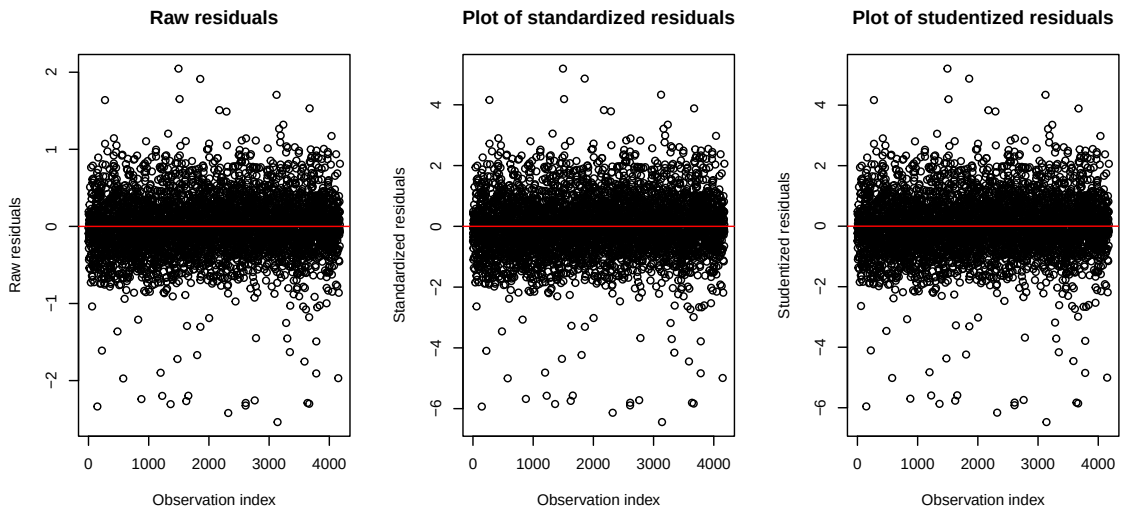

Figure 2: Residuals

There is no evidence of a systematic pattern, the point clouds seems randomly distributed to be near zero, thus it can be stated that the linearity assumption holds. Moreover, a plausible presence of heteroskedasticity can be pointed out, and for this reason, we have deepened the analysis computing other graphs that show the residuals confronted with each regressor, fitted values confronted with both raw and standardized residuals, and lastly the squared residuals confronted with the fitted values and each regressor.

In the first case, the same assumption obtained through the very first analysis of the residuals is considered. For the the fitted values, the variability of the residuals increases seems to be slightly higher. As further confirmation of the plausible presence of heteroskedascitity, in the squared residuals examination, there are some residuals away from the points clouds.

4.1 Leverage points and Cook's distance

The points highlighted in the previous analysis can be categorized in three different ways:

- Leverage points: They represent a type of observations that with a slight variation can have a massive impact onto the residuals analysis.
- Influence points: The significant impact of their elimination, can be due to their high value of p_{ii}.

· Outliers: These points, generally, do not have a strong impact on the parameter estimation, even their poor fit characterization of the model.

Subsequently to the residuals analysis, it is worth of noting that the amount of influence applied by the i_{th} observation is equal to leverage p_{ii}, which with the increase of this latter, increases its value in the regression too. The average of p_{ii} is k/n, thus the i_{th} observation is a high leverage point if $p_{ii} > \frac{2k}{n}$. The Cook's distance is used to measure the influence of a given point and can be defined as follows:

$$D_i = \frac{|\hat{y}_i - \hat{y}_{(i)}|^2}{ks^2} = \frac{1}{k} {e_i}^2 \frac{p_{ii}}{1 - p_{ii}} \tag{7}$$

where high values of D_i are obtained only with high values of e and p_{ii}. The value of the Cook's distance can be defined as high if $D_i > 4/n$, considering n as the number of the observations, this value implies the possible elimination of these specific observations.

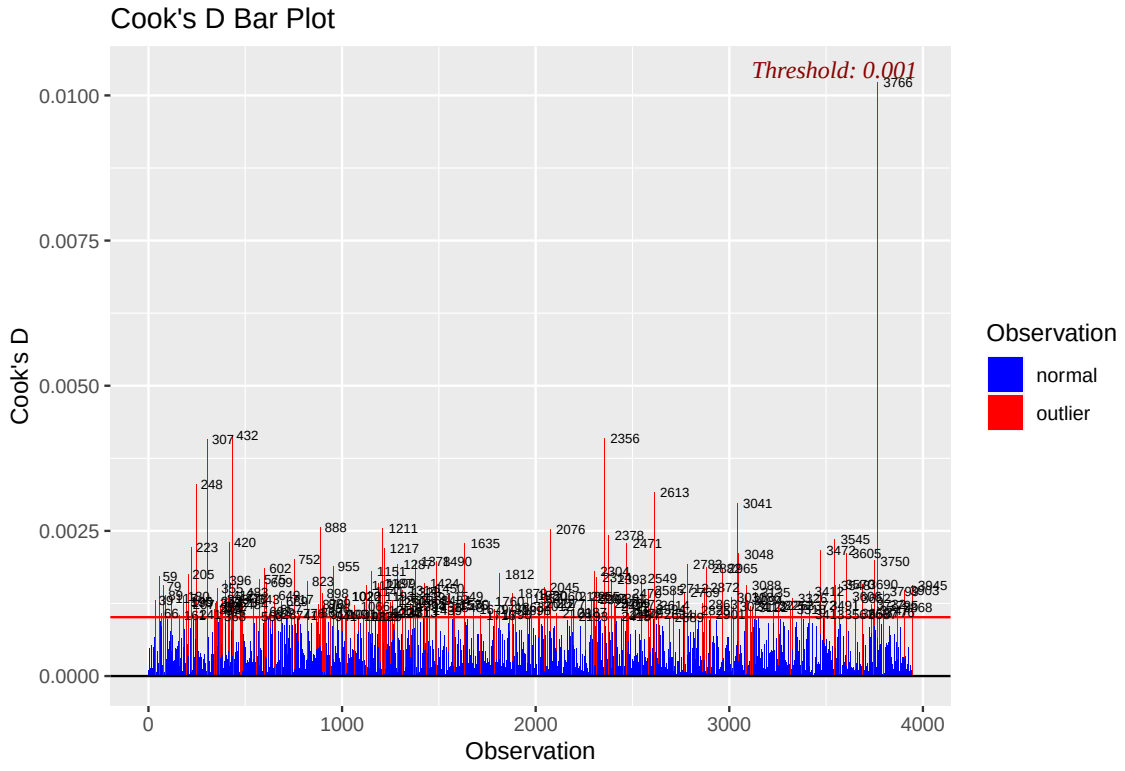

Figure 3: Cook's Distance Model 2

4.2 Normality of residuals

The possible non-normality of residuals is coherent with the non-normality of Y and ϵ. With the aim of our analysis, we are considering some normality test.

- Jarque-Bera: It is an asympthotic test and its H_0 imposes the normal distribution for the residuals.
- Kolmogorov-Smirnov: It compares the observed sample distribution to the theorical one where the H_0 represents the normality of the residuals.

Before computing these tests, we have claimed that the normality assumption does not hold given the heavy tails in the representation of the distribution; this also identify the presence of outliers. Subsequently to the QQ plot and the test analysis, the lack of Normality assumption holds. In order to obtain a normal distribution for the residuals, there exist three different paths: checking the presence of outliers or influential observations; transforming the dependent variable; or finding an alternative specification of the model. As mentioned before, even though the dependent variable is already transformed into its log version, we compute the Bonferroni test which thanks to the the p-value for the most extreme observations, enables us to identify the outliers. Before excluding them, we have tried to re-examine these values in the model through a creation of a dummy, but it was not enough, hence we deleted these value in order to overcome the non normality of the residuals. The model that contains the outliers is better than that which does not include them.

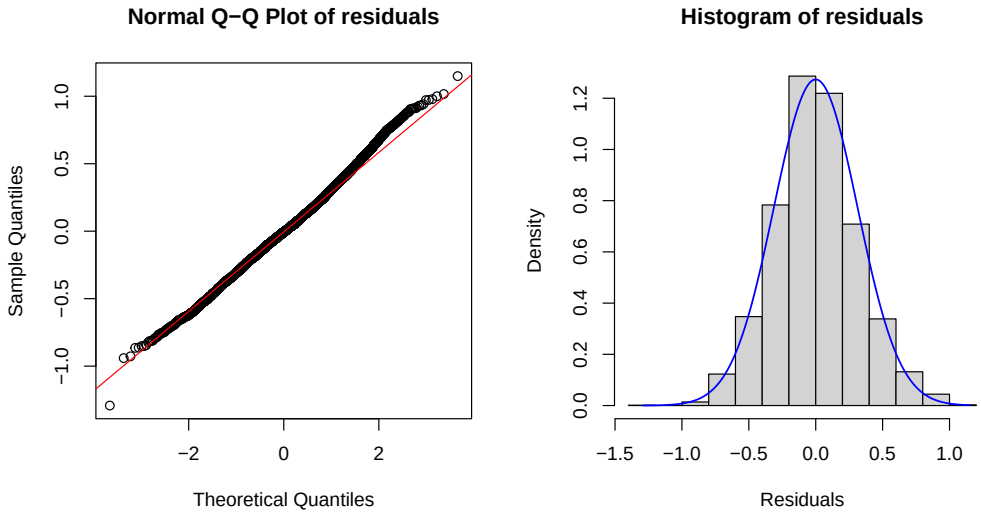

Figure 4: Normal Residuals

5 Mis-specification tests

These type of tests are applied when there may exist specification errors in the analysis such as: model underfitting (omission of relevant variable), model overfitting (inclusion of irrelevant variable), mis-specification of the functional form. To overcome these latter issues listed, we are proceeding using the Durbin-Watson test, that detects the autocorrelation between an excluded variable and the model, and the VIF, that quantifies the level of multicollinearity by providing an index that measures how much the variance of an estimated regression coefficient is increased due to the presence of collinearity.

Starting from the underfitting analysis, it is possible to state that the regressor excluded, $X2$, seems effectively not to be relevant given that the DW test shows that the residuals are not correlated. On the other hand, the same thing can not be stated for $X6$ which, indeed, has to be readmitted in the model. Furthermore, considering the overfitting examination with the computation of the VIF test, all the values computed are acceptable, hence we can confidently say that there is no multicollinearity.

To detect the presence of functional mis-specification due to non-linear regressor effects, we can use the RESET test through a particular application of the F-test with the exclusion constraints. If the model specification is correct and the hypothesis concerning the zero conditional expected value of the disturbance holds, no non-linear function of the explanatory variable should be significant when added to the model. Here, the computed RESET test shows that no relation in the model has to be changed (p-value > 0.05), because some regressors act in linear way on y and we already computed some transformation at the beginning. According to the last model obtained, we did again all the procedures to check again the behaviour of the residuals: to do so, we proceed with the heteroskedasticity check.

5.1 Homoskedasticity and heteroskedasticity

The definition of homoskedasticity states that the variance of the unobserved error, conditional on the explanatory variables, is constant. It is needed to justify the F-tests and the confidence intervals for the OLS estimation of the linear regression model. This condition no longer holds when the variance of ϵ changes within different segments of the population, determined by different values of the explanatory variables. This latter concept is referred to the heteroskedasticity.

The homoskedasticity condition is fundamental to correct the variance of the parameters and it is implemented to verify the accuracy of the linear regression model. This requirement ensures that the residual variances are computed correctly and that the OLS estimators are efficient. To obtain its empirical proof, we can proceed through the implementation of the Breusch-Pagan test, that checks the plausible presence of heteroskedasticity in the model by assuming as null hypothesis the model as homoskedastic. This is not realized, hence we proceed with a perform test procedure with alternative estimates of variance and covariance matrix. We go for the White test.

5.1.1 White test

Considering the presence of heteroskedasticity, conditional on the predictors and a new set of variable named W, the variance ϵ_i can be expressed as a function of W and a vector of parameters δ.

$$Var(\epsilon_i|X, W) = {\sigma_i}^2 = f(W, \delta) \tag{8}$$

where $f(.) > 0$ and $\forall i$ and $\forall \delta$

Given $Var(\epsilon_i) = E({\epsilon_i}^2)$, we have ${\epsilon_i}^2 = E({\epsilon_i}^2|X, W) + u_i$ with $E(u_i|X, W) = 0$ $\forall i$. It is possible to test the homoskedasticity hypothesis assessing the significance of the estimated δ, through the adoption of a specific function $f(W, \delta)$ and choosing W variables. The easiest type considers the estimated y_i and their squares as regressors which, being a f of all the predictors, the squared contain both the squares of the x_j and the cross products. Auxiliary regression:

$$e^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + u \tag{9}$$

where the $H_0 : \delta_1 = \delta_2 = 0$ is tested. If the test does not reject the null, it indicates the significance of the auxiliary regression but it does not imply the heteroskedasticity as the only cause (eg: misspecification of the functional form).

When the homoskedasticity hypothesis is rejected, and we do not have enough information to consistently estimate the variances and, to overcome the heteroskedasticity issue, we have proceeded correcting the standard errors. Furthermore, we have re-examined a model including a different elaboration of the standard errors: the robust version. The main reason of this applied procedure is to check the presence of any significant differences between the two type of errors. According to our results, no significant difference in the standard errors among Model 6 and Model6.robust, this means that we could be confident in the results based on homoskedasticity.

		Estimate	*Std. Error*	*t value*	*P-value*
Mod. 6	*Intercept*	5.572e+00	8.843e-02	63.013	<2e-16***
	log(X1)	4.478e-01	1.485e-02	30.166	<2e-16***
	log(X3)	1.305e-02	5.313e-03	2.456	0.01410*
	X4	6.577e-07	2.424e-07	2.714	0.00668**
	log(X5)	2.045e-02	3.411e-03	5.994	2.24e-09***
	log(X7)	6.883e-01	1.214e-02	56.716	<2e-16***
	X6	3.933e-05	2.076e-05	1.895	0.05822·
Mod. 6 robust	*Intercept*	5.5725e+00	8.9382e-02	62.3441	<2.2e16***
	log(X1)	4.4784e-01	1.4478e-02	30.9316	<2.2e-16***
	log(X3)	1.3049e-02	4.8406e-03	2.6957	0.0070541**
	X4	6.5767e-07	1.8255e-07	3.6027	0.0003188***
	log(X5)	2.0447e-02	3.2579e-03	6.2760	3.849e-10***
	log(X7)	6.8828e-01	1.2537e-02	54.9001	<2.2e-16***
	X6	3.9326e-05	2.1607e-05	1.8200	0.0688330·

Signif. codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '·', 0.1 ' ', 1

Table 2: Final model comparison

6 Conclusions

In the light of the above, the last and the best model considered, Model 6 robust, is characterized by six out of seven of the covariates taken in consideration previously, and it can be stated that it can adequately explain the current price of the main residence. Through the performed tests, the Jarque-Bera, availing the Central Limit Theorem, confirms the normality of the residuals, also confirmed in the QQ-plots and histograms. Carrying out the analysis, the Durbin-Watson test has showed an underfitting result, including again the $X6$ regressor that results statistically significant in the last model. When it comes to the Variance Inflation Factor test, it proves the lack of multicollinearity in the analysis, having all the regressors' values < 2. The homoskedasticity has been obtained through the application of the White test, after a non satisfactory result given by the Breusch-Pagan, that hence, testified the heteroskedastic characteristics.

7 R Code

```r
rm(list = ls())
graphics.off()

setwd("/Users/alessandro/Desktop/")

library(car)
library(dplyr)
library(sf)
library(magrittr)
library(stargazer)

data <- read.table("datBI.txt", sep = ";")
dataset <- data.frame(data$HB0900, data$HB0100, data$HB0200, data$HB4400,
                      data$HD1110, data$HG0410, data$HI0100, data$HNB0920)
summary(dataset)

dataset1 <- na.omit(dataset)
rm(data, dataset)
Dataset <- dataset1
rm(dataset1)

rownames(Dataset) <- 1:nrow(Dataset)
dataset <- Dataset
dataset$Observations <- 1:nrow(dataset)

rm(Dataset)
colnames(dataset) <- c("Y, Current $ main residence", "X1, SQ House", "X2, YRS in household",
                       "X3, Tot. Val. Cars", "X4, Sight Account",
                       "X5, Gross Income from fin. invest.", "X6, Food expences",
                       "X7, HMR/Imputed rent", "Observations")

# VARIABLES OF INTEREST
Y <- dataset$`Y, Current $ main residence`
X1 <- dataset$`X1, SQ House`
X2 <- dataset$`X2, YRS in household`
X3 <- dataset$`X3, Tot. Val. Cars`
X4 <- dataset$`X4, Sight Account`
X5 <- dataset$`X5, Gross Income from fin. invest.`
X6 <- dataset$`X6, Food expences`
X7 <- dataset$`X7, HMR/Imputed rent`

summary(dataset)

###### MODEL SPECIFICATIONS #######

pairs(dataset, upper.panel = panel.smooth)
```

```
GGally::ggpairs(dataset[, c(1:8)], axisLabels = "none")

model1 <- lm(log(Y) ~ log(X1) + X2 + log(X3) + X4 + log(X5)
             + log(X6) + log(X7), data = dataset, x = T, y = T)
summary(model1)

# We eliminate X2 and log(X6) because of their high p-values
transformation with the form 1/X7
model2 <- update(model1, .~. -X2 - log(X6))
summary(model2)

MASS::stepAIC(model2) # As a confirmation, the model2 seems to be the best and we eliminate
the other from the environment
rm(model1)

res <- model2$residuals

# Plots of residuals
par(mfrow = c(1,3))
plot(res, ylab = "Raw residuals", xlab = "Observation index", main = "Raw residuals")
abline(h = mean(res), col = "red")
plot(rstandard(model2), ylab = "Standardized residuals", xlab = "Observation index",
        main = "Plot of standardized residuals")
abline(h = mean(res), col = "red")
plot(rstudent(model2), ylab = "Studentized residuals", xlab = "Observation index",
        main = "Plot of studentized residuals")
abline(h = mean(res), col = "red")
# The distribution of the point clouds seems to be near to 0, thus the linearity assumption is valid
# There is a plausible presence of heteroskedasticity

par(mfrow = c(1,1))

# Residuals vs each regressors
head(model2$x)
par(mfrow=c(3,2))
for (i in c(2:6)){
 plot(model2$x[,i], rstandard(model2), ylab='Standardized Residuals',
xlab=colnames(model2$x)[i],main=paste('Standardized residuals vs',colnames(model2$x)[i], sep=' ') )
}
# As a further confirmation, also the plots of residuals vs each regressors
# show the confirmation of the linearity hypothesis,
# and the plausible presence of heteroskedasticity and outliers

# Fitted values and scatterplots
par(mfrow = c(1,2))
plot(model2$fitted.values, model2$residuals, ylab = "Residuals",
xlab = "Fitted values", main = "Plot of fitted values and residuals")
abline(h = mean(res), col = 2)
plot(model2$fitted.values, rstandard(model2), ylab = "Standardized residuals",
xlab = "Fitted values", main = "Plot of standardized fitted values and residuals")
```

```r
abline(h = mean(res), col = 2)

par(mfrow = c(1,1))
# Squared residuals
sq_res <- res^2
# y hat
plot(model2$fitted.values, sq_res, xlab = expression(paste("Estimated",
    sep = " ", hat("y"))), ylab = "Residuals")
# As further confirmation of the plausible presence of heteroskedascitity
# we compute the squared residuals

# Squared residuals vs each regressor
par(mfrow = c(3,2))
for (i in c(2:6)){
  plot(model2$x[, i], sq_res, ylab = expression(paste("Residuals"^2)), xlab = colnames(model2$x)[i])
}
par(mfrow = c(1,1))

plot(model2) # The scale location and the Cook's distance plots help us to
# identify the presence of outliers

# We can obtain Leverages by the function hat (it stands for hat values) applied to the model.matrix:
lev <- hat(model.matrix(model2))
plot(lev, ylab = "Leverages", main = "Index plot of Leverages")
lev.t <- 2*ncol(model.matrix(model2))/nrow(model.matrix(model2))
abline(h = lev.t, col = "red")
# Observations with high leverage:
h.l <- cbind(which(lev > lev.t), lev[c(which(lev > lev.t))]) # there are 269 cases

#Cook's distance
cooksD <- cooks.distance(model2)
plot(cooksD)
abline(h = 4/length(cooksD), col = "green")
d.inf <- cooksD <= 4/length(cooksD)
table(d.inf) # there are 231 influential observations

# Check for NORMALITY ASSUMPTION
qqnorm(model2$residuals, main = "Normal Q-Q Plot of residuals")
qqline(model2$residuals, col = "red")
# The normality assumption is not satisfied given the heavy tails
# in the representation of the distribution
# We need to compare the normal qqplot with the student-t one
# to point out the presence of outliers
car::qqPlot(model2, ylab = "", main = "Studentized residuals Q-Q plot")

# Tests for the normality assumption
tseries::jarque.bera.test(resid(model2)) # Ho: Normality
# since the p-value is less than alpha, we have to reject Ho
# non-parametric test
ks.test(model2$residuals, "pnorm") # Ho: Normality (REJECTED)
```

```r
# The Shapiro - Wilk test is not performed due to the dimension of the dataset
# As can be seen form the plots and also from the tests, the residuals are NOT normally distributed,
# thus we proceed with the elimination of the influential observations

# Bonferroni outlier Test
# let's try to see how many of these are outliers:
print(outl <- car::outlierTest(model2, labels = dataset$Observations))
# For the interpretation of the results, the observations characterized by a p-value < 0.05
# are highlighted by the test.

d.outl <-
dataset$Observations == names(outl$rstudent)[1] | dataset$Observations == names(outl$rstudent)[2]|
  dataset$Observations == names(outl$rstudent)[3] | dataset$Observations == names(outl$rstudent)[4]|
  dataset$Observations == names(outl$rstudent)[5] | dataset$Observations == names(outl$rstudent)[6]|
  dataset$Observations == names(outl$rstudent)[7] | dataset$Observations == names(outl$rstudent)[8]|
  dataset$Observations == names(outl$rstudent)[9] | dataset$Observations == names(outl$rstudent)[10]

# Before excluding the outliers found, we elaborate a different version of the model
# that includes the dummy variable d.outl that might be a solution for this issue

model3 <- update(model2, .~. + d.outl)
tseries::jarque.bera.test(resid(model3))
shapiro.test(resid(model3))
ks.test(model3$residuals, "pnorm")
# The model with the outliers still presents non normality of the residuals,
# thus we proceed with the elimination of them.

model4 <- update(model2, .~., subset = d.outl == 0)
summary(model4)
summary(model3) # best model which contains the outliers
summary(model2)

# there is not the normality yet
tseries::jarque.bera.test(resid(model4))
shapiro.test(resid(model4))
ks.test(model4$residuals, "pnorm")

# we can try to exclude all the influential observations
model5 <- update(model4, subset = d.inf)
summary(model5)

qqnorm(model5$residuals)
qqline(model5$residuals, col = "red")
hist(model5$residuals)

tseries::jarque.bera.test(resid(model5))


# UNDERFITTING
```

```r
# After obtaining the normality of the residuals we can proceed with the Mis-specifications analysis

lmtest:: dwtest(model5, order.by = dataset$`X2, YRS in household`[cooksD<(4/length(cooksD))])
lmtest:: dwtest(model5, order.by = dataset$`X6, Food expences`[cooksD<(4/length(cooksD))])

#  One out of two regressors that has been excluded from the analysis seem not
# to be relevant in the analysis
# the Durbin - Watson test, in fact, shows that the residuals are not correlated.
# Only X6 needs to be readmitted
# HO: autocorrelation of the disturbances is 0 (ACCEPTED)

model6 <- update(model5, .~. + I(X6))
summary(model6)
tseries::jarque.bera.test(model6$residuals)

# OVERFITTING

vif(model6) #Considering that values under 2 are still considered acceptable,
# we can be confident that the problem of multicollinearity is not present

View(cor(model.matrix(model6)[, c(2:7)]))
# Considering the correlation matrix, the absence of multicollinearity is confirmed
# because of the low values of the correlation among the regressors.

# LINEARITY

lmtest::resettest(model6, power = 2:3, type = "fitted")
# The RESET test shows that none of the relations has to be changed (p-value > 0.05)


res2 <- model6$residuals

par(mfrow = c(1,3))
plot(res2, ylab = "Raw residuals", xlab = "Observation index", main = "Raw residuals")
abline(h = mean(res2), col = "red")
plot(rstandard(model6), ylab = "Standardized residuals",
# xlab = "Observation index", main = "Plot of standardized residuals")
abline(h = mean(res2), col = "red")
plot(rstudent(model6), ylab = "Studentized residuals",
# xlab = "Observation index", main = "Plot of studentized residuals")
abline(h = mean(res2), col = "red")
# The mean of the residuals is concentrated in 0 and there may be presence of heteroskedasticity

# Residuals vs each regressors
head(model6$x)
par(mfrow=c(3,2))
for (i in c(2:7)){
  plot(model6$x[,i], rstandard(model6), ylab='Standardized Residuals',
  xlab=colnames(model6$x)[i],main=paste('Standardized residuals vs',colnames(model6$x)[i], sep=' '))
```

```
}

# Fitted values and scatterplots
par(mfrow = c(1,2))
plot(model6$fitted.values, model6$residuals, ylab = "Residuals",
xlab = "Fitted values", main = "Plot of fitted values and residuals")
abline(h = mean(res2), col = 2)
plot(model6$fitted.values, rstandard(model6), ylab = "Standardized residuals",
xlab = "Fitted values", main = "Plot of standardized fitted values and residuals")
abline(h = mean(res2), col = 2)

par(mfrow = c(1,1))

# Squared residuals
sq_res2 <- res2^2
# y hat
plot(model6$fitted.values, sq_res2, xlab = expression(paste("Estimated", sep = " ",
hat("y"))), ylab = "Residuals")

# Squared residuals vs each regressor
par(mfrow = c(3,2))
for (i in c(2:7)){
  plot(model6$x[, i], sq_res2, ylab = expression(paste("Residuals"^2)), xlab = colnames(model6$x)[i])
}

par(mfrow = c(1,1))

plot(model6)
plot(model6, which = 4, col = "red") #We can see a smaller number of influential observations

#Check for normality
par(mfrow=c(1,2))

qqnorm(model6$residuals, main = "Normal Q-Q Plot of residuals")
qqline(model6$residuals, col = "red")

hist(model6$residuals, main = "Histogram of residuals", xlab = "Residuals", freq = F, breaks = 10)
xfit <- seq(min(model6$residuals), max(model6$residuals), length = 1000)
yfit <- dnorm(xfit, mean = mean(model6$residuals), sd = sd(model6$residuals)) #normal distribution
lines(xfit, yfit, type = "l", col = "blue", lwd = 1.5)

car::qqPlot(model6, ylab = "", main = "Studentized residuals Q-Q plot")

# The normality is now obtained just by looking at the qqplot, that shows
# correspondence among the theoretical distribution and the realized one

# The normality is also confirmed by the tests
tseries::jarque.bera.test(resid(model6))
ks.test(model6$residuals, "pnorm")
```

```r
olsrr::ols_plot_cooksd_bar(model6)

# HOMOSKEDASTICITY TEST

# Towards predicted y values
plot(sq_res2 ~ fitted(model6), xlab = expression(paste("fitted values", sep = " ",
hat("y"))), ylab = expression(paste("squared residuals", sep = " ", hat("e"))))

# Towards each regressor
x.m6 <- as.data.frame(model.matrix(model6))
par(mfrow = c(3, 2))
for (i in 2:7)  {
  plot(sq_res2 ~ x.m6[, i], ylab = expression(paste("residuals", sep = " ",
  hat("e")^2)), xlab = colnames(x.m6)[i])
}

# Breusch-Pagan test

lmtest::bptest(model6, studentize = T)
lmtest::bptest(model6, studentize = F)
# The residuals are then heteroskedastic (p-value < 0.05)

m.VH = sandwich::vcovHC(model6)
# Then, to perform test procedure with alternative estimates of variance and covariance matrix:

model6.robust <- lmtest::coeftest(model6, vcov=vcovHC(model6, type="HC0"), df=3938) # White correction

summary(model6) #No significant difference in the Std. Error among model 6 and the model6.robust.
# Thus, model6.robust has also improved in significance, with an higher p-value in log(X3) and X4
model6.robust


# Model6.robust is thus the best.
```