Advanced Statistics for Finance

Project Work

Denise Citti 1839271, Gianlorenzo Gattinara 1979008

Abstract

The goal of our research is to analyze the harmonized sample survey 'Household Finance and Consumption Survey' (HFCS), conducted in 2016 by National Central Banks (Bank of Italy, in our case), which collects information on wealth, income and consumption of households in the euro area. The analysis was conducted using a multiple linear regression approach and only a handful of variables have been included in the final regression. We have undertaken a broad analysis, focusing on residuals and the link between the dependent variable and the regressors, starting with an initial model; and finally, verifying the goodness of fit of the model and its validity.

1. Introduction

We aim at modeling the effect of a given set of explanatory variables $x_1,.....,x_k$ on a variable y of primary interest. The variable of primary interest y is called response or dependent variable and the explanatory variables are also called covariates, independent variables, or regressors. The types of response variables (continuous, binary, categorical, or counts) and the varied types of covariates (also continuous, binary, or categorical) distinguish the various models. The linear regression model is especially applicable when the response variable y is continuous and shows an approximately normal distribution (conditional on the covariates). When the response variable is binary, the effect of covariates is nonlinear, or when geographical or cluster-specific heterogeneity must be considered, more general regression models are necessary. A main characteristic of regression models is that the relationship between the response variable y and the covariates is not a deterministic function $f(x_1,....,x:k)$, but rather shows random errors. One main goal of regression is to analyze the influence of the covariates on the mean value of the response variable. In other words, we model the (conditional) expected value $E(y|x_1,....,x_k)$ of y depending on the covariates. Hence, the expected value is a function of the covariates:

$$E(y|x_1,....,x_k) = f(x1,.....,x_k)$$

It is then possible to decompose the response into:

$$y = E(y|x_1,....,x_k) + \varepsilon = f(x1,.....,x_k) + \varepsilon$$

where ε is the random deviation from the expected value and it is also called random or stochastic component, disturbance, or error term, while the expected value $E(y|x_1,....,x_k) = f(x1,.....,x_k)$ is often denoted as the systematic component of the model. In the classical linear model, we assume that the error term does not depend on covariates. The most common class is the linear regression model represented by the following formulation:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

where:
- Y is the dependent variable,
- X_1, X_2 and X_3 are the independent variables;
- β_0 is the intercept;
- β_1, β_2,and β_3 are the slopes, which meaure the expected change in Y for a unit change in X;
- ε is the error term, which considers all the other factors not explicitly introduced in the model.

In our analysis, we implemented a Classical Linear Model with the following characteristics:
• $E(\varepsilon_t) = 0$ The errors have zero mean
• $var(\varepsilon_t) = \sigma^2$, The variance of the errors is constant and finite over all values of xt
• $cov(\varepsilon_i, \varepsilon_j) = 0$ The errors are linearly indipendent of one another
• $cov(\varepsilon_t, x_t) = 0$ There is no relationship between the error and corresponding x variate
• ε_t are normally distributed

According to what we stated above, our analysis will follow this path: in the upcoming section we will present an overview of the dependent variable that we want to explain and all the regressors. We will then perform the linear regression, having a particular attention on the variables and explaining the model; then we will perform the analysis on the residuals of the focusing on the CLRM assumption, such as normality, homoscedasticity, etc. We will illustrate all the performed tests and figures. In the conclusion of our analysis, we will see the WLS regression. In the Appendix, we will insert the R code used for our analysis.

1.1 Variables

Variables			
	Name	code	type
Y	Net wealth	DN3001	continous
X_1	level of education	PA0200	categoric
X_2	total gross income	DI2000	continous
X_3	residence size	HB0100	discrete
X_4	employee income	PG0100	dummy
X_5	is income normal in reference period	HG0700	categoric
X_6	investment in mutual funds	HD1300	dummy
X_7	Time in main job	PE0700	discrete

Dependent variable
We decided to choose as dependent variable Y “Net Wealth” which can be defined as total household assets excluding public and occupational pension wealth minus total outstanding household’s liabilities

Independent variables
The independent variables can be identified as regressors or explanatory variables. They are basically factors able to influence the phenomenon. In this section, you will find a description of all the predictors inserted in the linear regression.
Highest level of education completed (X_1) is associated to the “Highest level of education completed”. Categories based on ISCED-97 classification: 1 – Primary or below; 2 – Lower secondary; 3 – Upper secondary; 4- Post-secondary; 5 – Tertiary; 6-Second stage tertiary. The second regressor is total gross income (X_2), an individual’s gross income (sometimes known as gross pay on a paycheck) is their total earnings before taxes and other deductions. This encompasses all sources of income, not only employment, and is not restricted to cash income; it also includes property or services received. Size of main residence (X_3) is referred to the primary residence of a household is the dwelling in which the members of the household regularly reside, which is commonly a house or an apartment. At any given moment, a household can only have one main residence, albeit it may be shared with people who are not members of the household. It can happen sometimes that the households is not properly clear because of travelers or people who live in more than one house (multiple housing). In these instances, rather than rigid regulations, the criterion for determining the primary residence of the household would consist primarily of guidelines.
Received employee income (X_4) is associated to “Received employee income”. The variable is associate to the receive of any sort of employee income during the last 12 months/ last calendar year.
Income normal in reference period (X_5) is associated to “Is income normal in reference period?”. It means to state if your (household’s) income during the last 12 months was high, on average or low relative to what you’d anticipate in a ”typical” year. Households own investment in mutual funds (X_6) is associated investments in mutual funds. According to Regulation ECB/2008/32 and Regulation ECB/2007/8 mutual fund and money market funds are the same thing. Money market funds are collective investment undertakings the shares/units of which are, in liquidity term, close substitutes for deposits. This encompasses investing in money market instruments and in MMF shares/units or in transferable debt instruments. The residual

maturity up to one year (included) and in bank deposits which aim for a return that is comparable to the interest rate on money market instruments. When we talking about investment funds, we usually refer to a type of investment which is collective. It is able to collect money from a large number of investors and invests it in stocks, short-term money market instruments, securities and bonds. Finally, Time in main job (X_7) is related to the number of years the respondent has worked for the company where he or she is employed at the time of the interview. If you've worked for this company for less than a year, you'll get a zero. 1) a change in position within the company, 2) off-duty leaves during which the employment relationship has not been paused and has not lasted longer than one year, 3) parental leaves, or 4) changes in the company's name due to ownership changes or mergers and acquisitions have no bearing on the duration of current employment.

2. Statistical analysis of our variables

We'll look at some statistics on our variables to better understand how they behave, as this will affect the linear regression model we'll create.

Table 1. summary statistics

	Mean	SD	Skewness	Kurtosis
Net wealth	311472	684101.9	19.67628	650.2126
Gross income	44775.4	34357.98	2.993476	17.90356
Residence size	105.4	55.76498	4.168061	39.2076
Time in main job	18.47	12.12669	0.3388178	2.27401
Employee income	1.224	0.4171701	1.322063	2.747852
Income normal in reference	2.1444	0.4452749	0.6259129	4.12782
Level of education	2.987	1.122358	0.6073842	2.661745
Investments in mutual funds	1.906	0.291987	-2.780807	8.732885

Skewness can be of two types: negative and positive skewness. The latter shows asymmetry in the right tail of the distribution and the negative one suggests imbalance in the left tail. The skewness values are all different from zero, indicating that our variable distribution is not symmetric. The kurtosis metrics are associated and compared to the Standard normal distribution in which we can identify three types:
• *Leptokurtosis*: the distribution has heavier tails than the Gaussian distribution $K > 3$
• *Mesokurtosis*: the kurtosis statistic is similar to that of the normal distribution $K = 3$
• *Platykurtosis*: the tails are shorter than the normal distribution $K < 3$

2.1 Correlation

Proceeding with our analysis, we checked whether there were any correlations between the variables examined. Since our variables are not Normally distributed, Pearson's method can not be used, so we proceeded to check for correlations using Spearman's method, which does not require the variables to be continuous and Normally distributed. The table below is the correlation matrix and displays the correlation coefficients for each pair of variables.

Table 2. Correlation matrix

	Net.wealth	Gross.income	Residence.size	Time.job
Net.wealth	1.00	0.59	0.56	0.28
Gross.income	0.59	1.00	0.42	0.21
Residence.size	0.56	0.42	1.00	0.21
Time.job	0.28	0.21	0.21	1.00

There is a moderate relationship between the regressors and the response variable, which are positively correlated; while the correlation between the regressors is not high, which means that we shouldn't expect any problem with multicollinearity. In addition to the correlation matrix, the scatterplot show some of the relationship between the variables, we only compared a selection of numerical variables. As we can see, none of the plots display a "clear" linear relationship and some plots have the observation clustered in one portion of the graph. So, we evince that there might be some problem with the functional form.

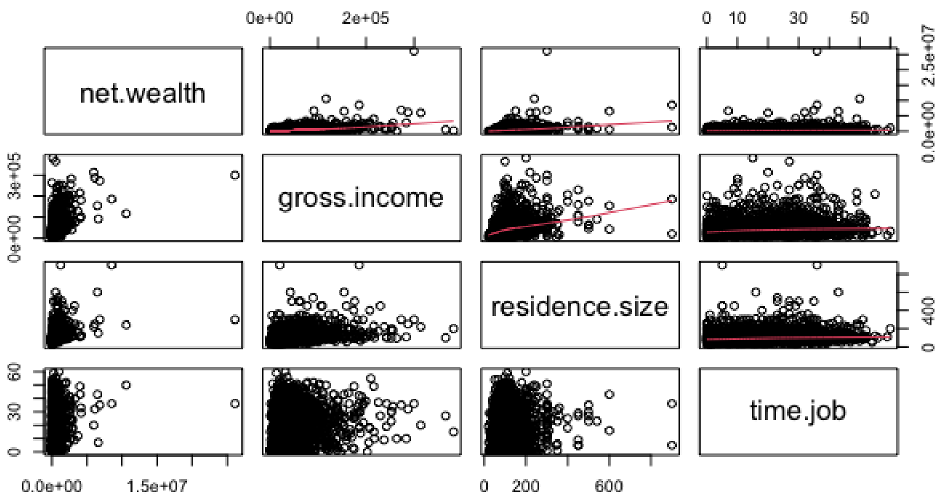

Figure 1. scatterplot

3. Linear Regression Model

We built our model, first, by identifying and discarding those variables which are not statistically significant and, then, by testing several forms, so by trying many models. During this phase we compared nested models with "anova" test and we performed the "Ramsey Reset test" to check the functional form (we will discuss about this test later on in the analysis). The best model that we found is "m2 sqr", which considers a square root transformation of the dependent variable. As matter of fact, since the standard deviation of the response variable Y is proportional to the mean (the ratio is equal to 2), we can use the square root on the y variable.

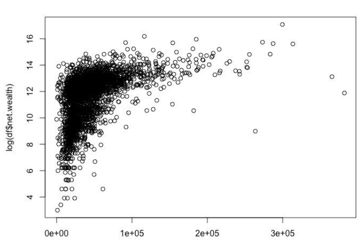

Figure 2. log Y vs gross income

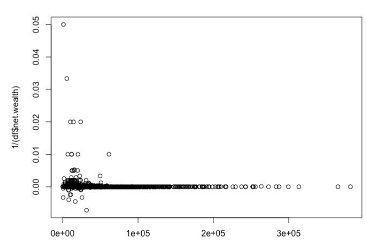

Figure 3. 1/Y vs gross income

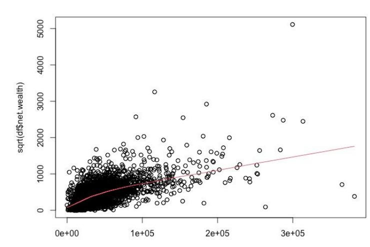

Figure 4. $\sqrt{Y}$ vs gross income

Tranformation used: Square root of Y (net wealth). $\mu_y = 315999.3$; $\sigma_y = 687861.8$; $\frac{\sigma_y}{\mu_y} = 2.176783$. It seems that there is a proportionality between the sd and mean.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	9.7572	37.0386	0.26	0.7922
gross.income	0.0035	0.0001	25.49	0.0000
as.character(education)2	31.3451	19.1682	1.64	0.1021
as.character(education)3	87.2788	18.7360	4.66	0.0000
as.character(education)5	122.2842	20.6226	5.93	0.0000
residence.size	1.8745	0.0783	23.94	0.0000
employee.income	84.4411	10.0095	8.44	0.0000
mutual.funds	-84.2543	14.2519	-5.91	0.0000
time.job	3.9771	0.3448	11.53	0.0000
Residual standard error: 229.5 on 3240 degrees of freedom				
Multiple R-squared: 0.5058, Adjusted R-squared: 0.5046				
F-statistic: 414.6 on 8 and 3240 DF, $p-value < 2.2e-16$				

The adjusted $R^2 = 0.5046$, meaning that the 50.46% of the variability of net wealth is explained by the variability of the regressors. As we can see from the F-statistic and the $p-value < 2.2e-16$, the β estimators are jointly significant.

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			170683420.10	35332.09		
gross.income	1	34239209.20	204922629.30	35924.08	649.95	0.0000
as.character(education)	3	3814880.12	174498300.22	35397.91	24.14	0.0000
residence.size	1	30199119.57	200882539.66	35859.39	573.26	0.0000
employee.income	1	3749100.18	174432520.27	35400.68	71.17	0.0000
mutual.funds	1	1841140.53	172524560.62	35364.95	34.95	0.0000
time.job	1	7009142.97	177692563.07	35460.84	133.05	0.0000

3.1 RESET Test

Diagnostic tests serve to check whether the CLRM assumptions are verified for the chosen model. The satisfaction or otherwise of the assumptions leads to different conclusions about the estimation of coefficients, their standard deviations and thus the reliability of the significance tests. The RESET test serves to verify the first of the four CLRM assumptions: that the model is linear in its parameters. We implemented the Ramsey Regression Equation Specification Error Test also known as ”RESET” to look for model mis-specification. The test is used to examine if non linear combinations of fitted values help explain the response variable. This happens also in the case in which independent variables provide in a good manner all the explanatory description of the dependent variable. Furthermore, according to the test, the model is mis-specified if the combinations of the explanatory variables are not linear and can explain the response variable. As for the alternative hypothesis of the test we can say that the model is suffering from an omitted variable problem. As we can see from the table we push for the rejection of the null hypothesis meaning that all the regressors doesn’t act in a linear way. On the other hand, we can still affirm that the situation gets better when we perform the transformation of the Y.

Table 3. RESET test results

Model	RESET	df1	df2	p-value
Mod1	180.75	2	3282	$< 2.2e-16$
Mod2	167.49	2	3285	$< 2.2e-16$
Mod3	172.87	2	3285	$< 2.2e-16$
Mod4	164.46	2	3286	$< 2.2e-16$
Mod5	164.46	2	3286	$< 2.2e-16$
Mod6	116.88	2	3281	$< 2.2e-16$
Mod7	237.98	2	3283	$< 2.2e-16$
Mod8	175.3	2	3284	$< 2.2e-16$
Mod9	180.92	2	3283	$< 2.2e-16$
Mod10	172.49	2	3280	$< 2.2e-16$
Mod11	268.65	2	3279	$< 2.2e-16$
Mod12	264.96	2	3281	$< 2.2e-16$

Table 4. RESET test results

Model	RESET	df1	df2	p-value
Mod13	285.56	2	3278	$< 2.2e-16$
Mod14	268.56	2	3278	$< 2.2e-16$
Mod15	282.63	2	3280	$< 2.2e-16$
Mod16	90.45	2	3285	$< 2.2e-16$
Mod17	257.92	2	3285	$< 2.2e-16$
Mod19	169.96	2	3284	$< 2.2e-16$
Mod20	158.92	2	3287	$< 2.2e-16$
Mod21	162.91	2	3285	$< 2.2e-16$
Mod22	165.55	2	3286	$< 2.2e-16$
Mod23	160.22	2	3286	$< 2.2e-16$
m1 sqr	24.206	2	3236	$< 3.677e-11$
m2 sqr	24.344	2	3238	$< 3.209e-11$

4. Residuals analysis

After the model has been described, the residuals must be examined to see if the model provides a correct representation of the dependent variable and if the model's essential assumptions are met: linearity, normality, homoscedasticity, and independency. We can do that by looking at graphical representations. Homoscedasticity means that all variables have the same variance and if the variance doesn't remain constant across the variables then, the distribution is called heteroschedastic.

4.1 Graphical analysis

We can plainly discern 3 main types of residuals:
• Studentized: achieved if the unknown quantity σ^2 is replaced by a suitable approximation during standardization and the Student-t distribution will apply to the standardized residuals.
• Raw: simple residuals obtained by the formula: $e_i = y_i - \hat{y}_i = (1 - p_{ii})\varepsilon_i$, where p_{ii} is the leverage.
• Standardized: we used three alternative plots for the distribution to check for homoskedasticity from a graphic standpoint. The plain distribution is represented by one plot, while the others were calculated using two different error terms, resulting in a standardised residuals via normalisation to unit variance using the overall error variance of the residuals and a Studentized residuals.

Following, the scatter plot. The residuals should show a constant variation and should be randomly distributed around zero, if the conditions are met.

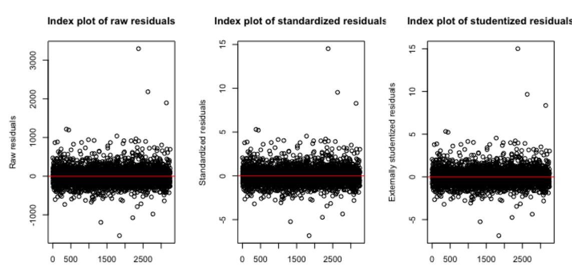

Figure 5. Residuals

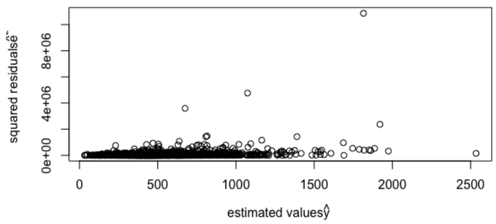

Figure 6. Squared Residuals

From figures 5 and 6 it is not clear if there is heteroschedasticity, because the residuals seem to lie around zero and to be constant; however there are some data points which are more dispersed. Following the plots against fitted values.

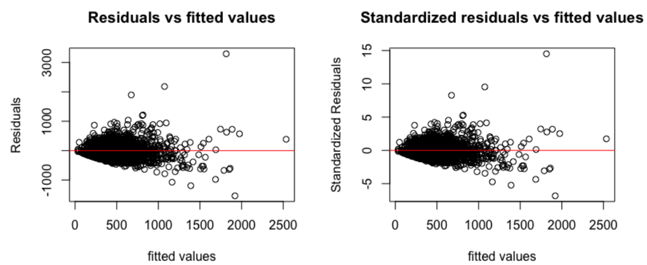

Figure 7. Fitted values

The presence of heteroschedasticity is more clear here. In fact, the greater the $\hat{y}$, the greater is the dispersion from 0 of the squared residuals. Furthermore in the Q-Q Plots, the straight line represents the theoretical quantiles of a Gaussian distribution, while the dots represent the quantiles of the residuals.

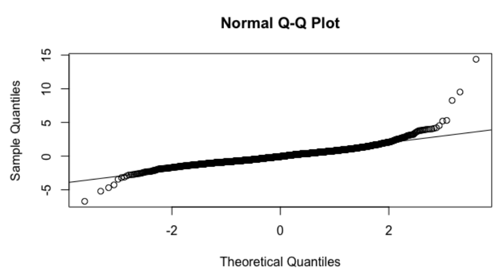

Figure 8. Quantile-Quantile plots

4.2 Outliers and leverage points

We can classify into three categories all the points that can be seen in both QQplots and residuals plots:
• Influence points: they have a great impact in the linear regression model and have high values of p_{ii} and residuals.
• Outliers: the model doesn't fit well but sometimes it can't affect parameters estimation.
• Leverage points: with high p_{ii} values but they may not have a significant impact on parameters estimation.

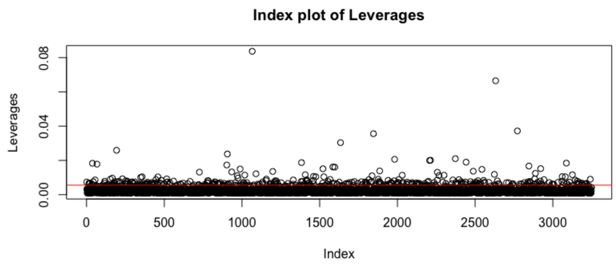

Figure 9. Leverages

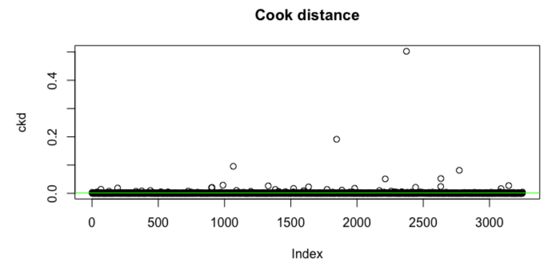

Figure 10. Cook's distance

To properly quantify the degree of impact, the Cook's distance proves to be a useful tool for identifying influential data points that should be double-checked for model validity.
D_i is the Cook's distance and it can be explained by the sum of all the changes in the regression model.We can see several points over the straight line on the graph created by plotting our data defining them as influential observations. These observation have been removed in our analysis.

4.3 Normality check

After removing the influential observations, we move on with the analysis of normality of the residuals, which is one of the basic hypotheses of linear regression. While this assumption didn't hold in prior QQplots, we can see that now the quantiles are closer to the theoretical one, even though the left tail is far from normality. In the figures, we presented the graphical representation of the estimated residuals with an histogram, followed by a QQ-plot, which represents the ordered values of the residuals vs. the theoretical quantiles.

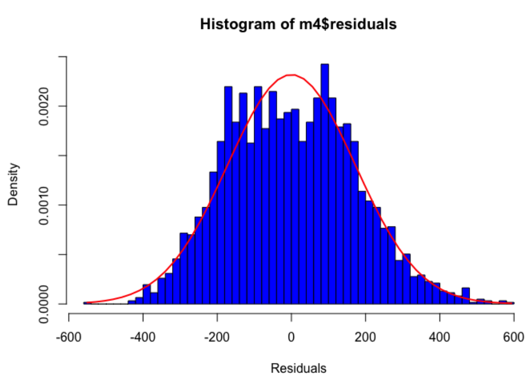

Figure 11. Histograms

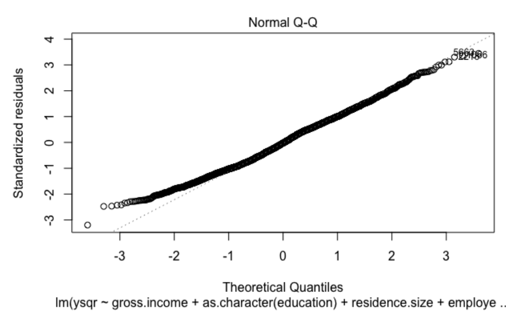

Figure 12. QQplots

Normality assumption can also be checked by performing the Shapiro-Wilk and Jarque Bera test which have the following hypothesis:
• $H0$: Normality of residuals
• $H1$: Non normality of residuals
We reject the null hypothesis for both tests implying that Net Wealth is not constant.

	Shapiro Wilk Test	Jarque Bera Test
Statistic	0.99435	29.376
p-value	1.643e-09	4.18e-07

Table 5. Shapiro Wilk Test and Jarque Bera Test

5. Multicollinearity

The Variance Inflation Factor (VIF) quantifies the severity of multicollinearity in an OLS regression analysis. It provides an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity. The threshold indicating multicollinearity between the regressors is equal to 10. Thus, if for a variable in the estimated model the test produces a $VIF > 10$, that variable should be eliminated from the model because it does not provide any additional information.

Variables	VIF
Gross.income	1.368257
As.character(education)	1.153177
Residence.size	1.181157
Employee.income	1.048905
Mutual.funds	1.071104
Time.job	1.069939

There is not a significant collinearity and we do not have an overfitting problem.

6. Mis-specification test

Model mis-specification refers to all of the ways in which a multiple regression linear model may fail to accurately describe a given circumstance. We can misspecify a regression model following various paths:
• Measurement errors;
• Model underrfitting (we omit all the variables which are relevant);
• Model overfitting (we include variables which are not relevant);
• Mis-specificatin of disturbance term;
• Functional form's mis-specification.
During this phase we would like to check how efficient is our linear approximation, simply by evaluating problems related to underfitting and over fitting. In order to do so we used the Durbin Watson test having as a statistic DW and taking values between 0 and 4.
• If DW is equal to 2 there is no autocorrelation;
• If $2 < DW < 4$ there is negative autocorrelation;
• If $0 < DW < 2$ there is positive autocorrelation.

	Durbin Watson
statistic	1.855
p-value	2.811e-05

Table 6. Durbin Watson test

Alternative hypothesis: true autocorrelation is greater than 0

7. Homoschedasticity

If the residuals' variance is constant we are in the condition in which we have homoschedasticity.In order to assess it we can take a look at the plot of squared residuals; if a trend is shown we are able to affirm the existence of heteroschedasticiy. We can detect the presence of a variance which is constant using many many tests: one example is the Breusch-Pagan test. The latter has been elaborated by Adrian Pagan and Trevor Breusch in 19794. The null hypothesis of this test basically implies the presence of homoschedasticity.

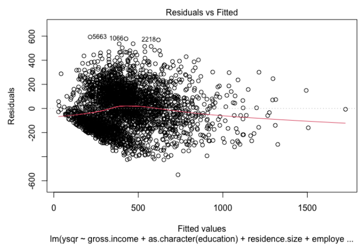

Figure 13. Fitted values vs residuals

	Breusch Pagan
statistic	48.305
df	8
p-value	8.64e-08

Table 7. Results

8. WLS

WLS regression compensates for violations of the homoscedasticity assumption by weighting instances differently: examples with large variances on the independent variable(s) count less, while those with small variances count more in estimating the regression coefficients. In other words, cases with higher weights contribute more to the regression line's fit. As a result, the predicted coefficients are frequently quite near to what they would be in OLS regression, but the standard errors are reduced in WLS regression. WLS regression is sometimes used to adjust fit to give less weight to remote points and outliers, or to provide less weight to observations believed to be less reliable, in addition to its basic function of correcting for heteroscedasticity. Since heteroscedasticity is present, we will perform weighted least squares by defining the weights in such a way that the observations with lower variance are given more weight.

	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	-46.4914	33.5412	-1.39	0.1658
gross.income	0.0039	0.0002	25.05	0.0000
as.character(education)2	27.0734	13.1647	2.06	0.0398
as.character(education)3	79.6716	13.1215	6.07	0.0000
as.character(education)5	94.0249	15.1275	6.22	0.0000
residence.size	2.2938	0.0884	25.95	0.0000
employee.income	70.6080	8.6077	8.20	0.0000
mutual.funds	-75.1645	14.1138	-5.33	0.0000
time.job	3.8527	0.2670	14.43	0.0000

Residual standard error: 1.107 on 3092 degrees of freedom
Multiple R-squared: 0.5453, Adjusted R-squared: 0.5441
F-statistic: 463.4 on 8 and 3092 DF, $p-value :< 2.2e-16$

	RESET	Jarque-Bera	Shapiro-Wilk	Breusch Pagan
statistic	30.958	44.386	0.99597	0.029906
p-value	4.874e-14	2.299e-10	2.023e-07	1

Table 8. Results

Finally, we made an attempt also by transforming the predictors with logarithms. We can see that we achieve better results with some tests (e.g. Reset test).

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	-2102.9794	67.0908	-31.35	0.0000
log(gross.income)	129.3266	5.7048	22.67	0.0000
as.character(education)2	-4.1737	14.0236	-0.30	0.7660
as.character(education)3	40.7911	14.0121	2.91	0.0036
as.character(education)5	84.7878	15.8908	5.34	0.0000
log(residence.size)	243.4087	8.8380	27.54	0.0000
employee.income	93.0555	8.6522	10.76	0.0000
mutual.funds	-93.8579	14.0907	-6.66	0.0000
log.time.job	39.1999	3.2927	11.91	0.0000

Residual standard error: 1.091 on 3024 degrees of freedom
Multiple R-squared: 0.5533, Adjusted R-squared: 0.5521
F-statistic: 468.2 on 8 and 3024 DF, $p - value :< 2.2e - 16$

	RESET	Jarque-Bera	Shapiro-Wilk	Breusch Pagan
statistic	6.0373	60.376	0.99472	0.030436
p-value	0.002417	7.749e-14	5.651e-09	1

Table 9. Results

9. Conclusion

During our analysis we focused on residuals and the connection between the dependent variable and all the various regressors. Our starting point was the initial model and we tried to verify its effectiveness, validity and goodness of fit.We estimated a model trying to respect all the assumptions that the OLS theory imposed on us. Not all of them have been met and our initial assumptions were only partially fulfilled. Furthermore, we checked whether there were any correlations between the variables examined. Ours were not normally distributed and we implemented and used Spearman's method to look for correlations. We found out a moderate relationship between the regressors and the response variable, which are positively correlated. After trying many models, we were able to say that the best one is m2 sqr, which consider a square root transformation of the dependent variable. In addition we used diagnostic tests in order to meet the CLRM assumptions; implementing Ramsey Equation Specification Error Test. With the latter, the regressors didn't act in a linear way, although we were able to improve the result. After the description of the model we examined the residuals. Linearity, normality, homoscedasticity, and independency must be met. We identifyied and eliminated the influential data points. We also notice that we didn't find significant collinearity and we did not have problems related to overfitting. One of the key assumptions of linear regression is that the residuals are distributed with equal variance at each level of the predictor variable. This assumption is known as homoscedasticity. When this assumption is violated, we say that heteroscedasticity is present in the residuals. When this occurs, the results of the regression become unreliable. One way to handle this issue is to instead use weighted least squares regression, which places weights on the observations such that those with small error variance are given more weight since they contain more information compared to observations with larger error variance. Since, we found evidence of heteroschedasticity we decided to use the WLS regression. Although we were not able to solve all the problems, our analysis show an improvement in the results considering the starting point and the arrival.

10. Code

Here it is presented the R code used for the project.

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

setwd("/Users/gianlorenzo/Desktop/FINASS/Advanced Statistics for finance/Project/Data")
# install.packages("dplyr")    # alternative installation of the %>%
library(magrittr) # needs to be run every time you start R and want to use %>%
library(dplyr)    # alternatively, this also loads %>%
library(purrr)
library(tidyverse)
library(data.table)
library(car)
library(ggstatsplot)
library(lmtest)
library(zoo)
library(moments)
library(tseries)
# library(rstatix)
library(MASS)
library(gap)
library(xtable)
library(stargazer)

data <- read.delim("datBI.txt", header = TRUE, sep = ";")
# Dataset
df <-
  data.frame(data$DN3001, data$DI2000, data$PA0200, data$HB0100,
             data$PG0100, data$HG0700, data$HD1300, data$PE0700) #data$HG0400
df <- na.omit(df)
# rename columns
newnames <- c("net.wealth", "gross.income","education", "residence.size",
              "employee.income", "inc.norm", "mutual.funds", "time.job")  #"fin.inv"
setnames(df, colnames(df), new = newnames)

###### EXPLORATORY ANALYSIS ######

#summary dataset
summary(df)
str(df)

### Net wealth
summary(df$net.wealth)
sd(df$net.wealth)
skewness(df$net.wealth)
kurtosis(df$net.wealth)
hist(df$net.wealth)

# Gross Price density distribution
ggplot(df)+
  geom_density(aes(x= df$net.wealth), fill = 'darkorange', color = 'black', alpha =.6)+
  theme_minimal()+
  stat_function(fun = dnorm, colour='blue', size=1,
```

```
                   args = list(mean=mean(df$net.wealth),
                               sd = sd(df$net.wealth)))+
  labs(y = "Density", x = "Net Wealth")

# Q-Q plot
par(mfrow=c(1,1))
qqnorm(df$net.wealth) # Empirical
qqline(df$net.wealth, col="red") # Theoretical

# Shapiro-Wilk test: H0 - normality of data
shapiro.test(df$net.wealth) # Rejection of H0 -> net.wealth is not Normal

# Jarque-Bera tes: H0 - normality of data
jarque.bera.test(df$net.wealth) # Rejection of H0 -> net.wealth is not Normal

boxplot(df$net.wealth)


### gross.income
summary(df$gross.income)
sd(df$gross.income)
skewness(df$gross.income)
kurtosis(df$gross.income)
hist(df$gross.income)

# Gross Price density distribution
ggplot(df)+
  geom_density(aes(x= df$gross.income), fill = 'darkorange', color = 'black', alpha =.6)+
  theme_minimal()+
  stat_function(fun = dnorm, colour='blue', size=1,
                   args = list(mean=mean(df$gross.income),
                               sd = sd(df$gross.income)))+
  labs(y = "Density", x = "Net Wealth")

# Q-Q plot
par(mfrow=c(1,1))
qqnorm(df$gross.income) # Empirical
qqline(df$gross.income, col="red") # Theoretical

# Shapiro-Wilk test: H0 - normality of data
shapiro.test(df$gross.income) # Rejection of H0 -> gross.income is not Normal

# Jarque-Bera tes: H0 - normality of data
jarque.bera.test(df$gross.income) # Rejection of H0 -> gross.income  is not Normal

boxplot(df$gross.income)
boxplot(df$gross.income)$out


### residence.size
summary(df$residence.size)
sd(df$residence.size)
skewness(df$residence.size)
kurtosis(df$residence.size)
hist(df$residence.size)
```

```r

# Gross Price density distribution
ggplot(df)+
  geom_density(aes(x= df$residence.size), fill = 'darkorange', color = 'black', alpha =.6)+
  theme_minimal()+
  stat_function(fun = dnorm, colour='blue', size=1,
                args = list(mean=mean(df$residence.size),
                            sd = sd(df$residence.size)))+
  labs(y = "Density", x = "Net Wealth")

# Q-Q plot
par(mfrow=c(1,1))
qqnorm(df$residence.size) # Empirical
qqline(df$residence.size, col="red") # Theoretical

# Shapiro-Wilk test: H0 - normality of data
shapiro.test(df$residence.size) # Rejection of H0 -> residence.size is not Normal

# Jarque-Bera tes: H0 - normality of data
jarque.bera.test(df$residence.size) # Rejection of H0 -> residence.size  is not Normal

boxplot(df$residence.size)


### time.job
summary(df$time.job)
sd(df$time.job)
skewness(df$time.job)
kurtosis(df$time.job)
hist(df$time.job)

# Gross Price density distribution
ggplot(df)+
  geom_density(aes(x= df$time.job), fill = 'darkorange', color = 'black', alpha =.6)+
  theme_minimal()+
  stat_function(fun = dnorm, colour='blue', size=1,
                args = list(mean=mean(df$time.job),
                            sd = sd(df$time.job)))+
  labs(y = "Density", x = "Net Wealth")

# Q-Q plot
par(mfrow=c(1,1))
qqnorm(df$time.job) # Empirical
qqline(df$time.job, col="red") # Theoretical

# Shapiro-Wilk test: H0 - normality of data
shapiro.test(df$time.job) # Rejection of H0 -> time.job is not Normal

# Jarque-Bera tes: H0 - normality of data
jarque.bera.test(df$time.job) # Rejection of H0 -> time.job  is not Normal

boxplot(df$time.job)
############################################
```

```
plot(df$gross.income,df$net.wealth)
lines(panel.smooth(df$gross.income,df$net.wealth))

plot(df$residence.size,sqrt(df$net.wealth))
lines(panel.smooth(df$residence.size,df$net.wealth))

plot(df$time.job,sqrt(df$net.wealth))
lines(panel.smooth(df$time.job,df$net.wealth))

summary(lm(net.wealth ~ gross.income, data = df))
summary(lm(net.wealth ~ residence.size, data = df))


#-------------------------------------------------------------------------------
# Correlation Matrix
cor_var <- df[,c(1,2,4,8)]
#View(cor_var)
cor_mat <- cor(cor_var, method = "spearman")
round(cor_mat, 2)

# scatterplot for relationship between variables
pairs(df, upper.panel = panel.smooth)

# selection of numerical variables
sel = df[,c(1,2,4,8)]
pairs(sel, upper.panel = panel.smooth)

#-------------------------------------------------------------------------------
# MODELS ESTIMATION

# M1: complete model:
mod1<-lm(net.wealth ~ gross.income + as.character(education) + residence.size
         + employee.income + as.character(inc.norm) + mutual.funds + time.job, data = df, x=T, y=T)
summary(mod1)

# M2: NO education:
mod2 <- update(mod1, .~. -as.character(education))
summary(mod2)
# compare 2 nested model full vs reduced
anova(mod2,mod1) # restricted model
lrtest(mod1,mod2)

# M3: No education No income:
mod3<- update(mod2, .~. -as.character(inc.norm))
summary(mod3)
anova(mod3, mod2)

# M4: No mutual funds:
mod4<- update(mod2, .~. -mutual.funds)
summary(mod4)
anova(mod4, mod2) # restricted model

# M5: No income No mutual funds:
mod5<- update(mod2, .~. -mutual.funds)
summary(mod5)
```

```r
anova(mod5, mod2) # restricted

# M6: gross.income power two
mod6<- update(mod1, .~. + I(gross.income^2))
summary(mod6)
anova(mod2, mod6) # larger model

# M7: gross.income power three
mod7<- update(mod2, .~.  + I(gross.income^2) + I(gross.income^3))
summary(mod7)

# M8: power two on residence size
mod8<- update(mod2, .~. + I(residence.size^2))
summary(mod8)

# M9: residence size power three
mod9<- update(mod2, .~. + I(residence.size^2) + I(residence.size^3))
summary(mod9)

#M10: power two on residence size and power two on gross.income
mod10<- update(mod6, .~. + I(residence.size^2))
summary(mod10)
anova(mod6, mod10) # larger model

#M11: power 2 on residence size and power 3 on gross.income
mod11<- update(mod10, .~. + I(gross.income^3))
summary(mod11)
anova(mod10, mod11) # larger model

#M12
mod12<- update(mod11, .~. - as.character(inc.norm))
summary(mod12)
anova(mod12, mod11) # restricted model

#M13: power two on time.job
mod13<- update(mod11, .~. + I(time.job^2))
summary(mod13)
anova(mod11, mod13) # larger

#M14: exp  on time.job
mod14<- update(mod11, .~. + I(exp(time.job)))
summary(mod14)
anova(mod11, mod14) # restricted

#M15
mod15<- update(mod13, .~. -as.character(inc.norm))
summary(mod15)
anova(mod15, mod13) # restricted

#M16 -18
mod16<- update(mod2, .~. -gross.income + log(gross.income))
summary(mod16)

mod17<- update(mod16, .~. -residence.size + log(residence.size))
summary(mod17)
```

```
mod18<- update(mod17, .~. -time.job + log(time.job)) # error
summary(mod18)


# RESET TEST:
resettest(mod1, power = 2:3, type = c("fitted"))
resettest(mod2, power = 2:3, type = c("fitted"))
resettest(mod3, power = 2:3, type = c("fitted"))
resettest(mod4, power = 2:3, type = c("fitted"))
resettest(mod5, power = 2:3, type = c("fitted"))
resettest(mod6, power = 2:3, type = c("fitted"))
resettest(mod7, power = 2:3, type = c("fitted"))
resettest(mod8, power = 2:3, type = c("fitted"))
resettest(mod9, power = 2:3, type = c("fitted"))
resettest(mod10, power = 2:3, type = c("fitted"))
resettest(mod11, power = 2:3, type = c("fitted"))
resettest(mod12, power = 2:3, type = c("fitted"))
resettest(mod13, power = 2:3, type = c("fitted"))
resettest(mod14, power = 2:3, type = c("fitted"))
resettest(mod15, power = 2:3, type = c("fitted"))
resettest(mod16, power = 2:3, type = c("fitted"))
resettest(mod17, power = 2:3, type = c("fitted"))
# p-value < 2.2e-16


# let's consider other type of interactions between the independent variables
# M19: complete model:
mod19<-lm(net.wealth ~ gross.income* employee.income + as.character(education)
         + mutual.funds + time.job, data = df, x=T)
summary(mod19)

# M20: NO education:
mod20 <- update(mod19, .~. -as.character(education))
summary(mod20)

# M21: complete model:
mod21<-lm(update(mod20, .~. + as.character(inc.norm)))
summary(mod21)

# M22: No mutual funds:
mod22<- update(mod20, .~. -mutual.funds + gross.income*mutual.funds)
summary(mod4)
anova(mod22, mod20) # restricted model

# M23: complete model:
mod23<-lm(update(mod20, .~. -mutual.funds+ as.character(inc.norm)))
summary(mod23)
#############
resettest(mod19, power = 2:3, type = c("fitted"))
resettest(mod20, power = 2:3, type = c("fitted"))
resettest(mod21, power = 2:3, type = c("fitted"))
resettest(mod22, power = 2:3, type = c("fitted"))
resettest(mod23, power = 2:3, type = c("fitted"))
# p-value < 2.2e-16
```

```r
327
328 #---------------------------------------------------------------------------
329 # Transformations of the dependent variable (Y)
330
331 # Let's now consider the following relationships:
332 plot(df$gross.income,log(df$net.wealth))
333 plot(df$gross.income,1/(df$net.wealth))
334
335 plot(df$gross.income,sqrt(df$net.wealth))
336 lines(panel.smooth(df$gross.income,sqrt(df$net.wealth)))
337
338 plot(df$residence.size,sqrt(df$net.wealth))
339 lines(panel.smooth(df$residence.size,sqrt(df$net.wealth)))
340
341 plot(df$time.job,sqrt(df$net.wealth))
342 lines(panel.smooth(df$time.job,sqrt(df$net.wealth)))
343
344 #--------
345 ### transform y into log(y)
346 any(df$net.wealth==0)
347 mod<-lm(log(net.wealth+1) ~ gross.income + as.character(education) + residence.size
348          + employee.income + as.character(inc.norm) + mutual.funds + time.job, data = df, x=T)
349 summary(mod)
350
351 # alternative
352 log_y <- log(df$net.wealth)
353 df$log_y = log_y
354 df <- na.omit(df)
355 mod<-lm(log_y ~ gross.income + as.character(education) + residence.size
356         + employee.income + as.character(inc.norm) + mutual.funds + time.job, data = df, x=T)
357 summary(mod)
358
359 ########
360 # Logs do not resolve our problems, so we move on with other type of transformations
361 ########
362
363 #--------
364 # Note: in case you run lines 353 to 358, the re-run lines 26 to 33
365 ### Square root of Y
366 mean(df$net.wealth)
367 # mean Y = 315999.3
368 sd(df$net.wealth)
369 # sd Y = 687861.8
370 # sigma/mu = 2.176783
371 # it seems that there is a proportionality between Y sd and Y mean
372 ysqr <- sqrt(df$net.wealth)
373 df$ysqr = ysqr
374 df <- na.omit(df)
375
376 m1_sqr<-lm(ysqr ~ gross.income + as.character(education) + residence.size
377         + employee.income + as.character(inc.norm) + mutual.funds + time.job, data = df, x=T)
378 summary(m1_sqr)
379
380 m2_sqr<-lm(ysqr ~ gross.income + as.character(education) + residence.size
381         + employee.income + mutual.funds + time.job, data = df, x=T)
```

```
summary(m2_sqr)
anova(m2_sqr, m1_sqr) #restricted

resettest(m1_sqr, power = 2:3, type = c("fitted"))
resettest(m2_sqr, power = 2:3, type = c("fitted")) # slightly improve compared to all the previous models
drop1(m2_sqr, test="F")

#-------------------------------------------------------------------------------
# set mod = choosen model
mod=m2_sqr
xtable(mod)
##

#-------------------------------------------------------------------------------
### RESIDUALS ANALYSIS ###
par(mfrow=c(1,1))
resid<-residuals(mod)
t.test(resid)
shapiro.test(resid)
qqnorm(scale(resid))
abline(0,1)

model<-formula(mod)
bptest(model,data=df)
dwtest(model,data=df)

confint(mod)
coeftest(mod)

yfit<-fitted(mod)

## Homoschedasticity
# if there is homoschedasticity the inclination of the line should be zero
summary(lm(abs(resid) ~ yfit))
## serial correlation
# independence if there is no pattern
n<-length(resid)
plot(resid[-n], resid[-1])
dwtest(mod)
#
## Let's check more in detail:
# Index plots:
par(mfrow=c(1,3))
plot(mod$residuals, ylab='Raw residuals', xlab='Observation index',
     main='Index plot of raw residuals')
abline(h=0, col='red')
plot(rstandard(mod), ylab='Standardized residuals', xlab='Observation index',
     main='Index plot of standardized residuals')
abline(h=0, col='red')
plot(rstudent(mod), ylab='Externally studentized residuals', xlab='Observation index',
     main='Index plot of studentized residuals')
abline(h=0, col='red')
#
# scatterplots vs fitted values:
par(mfrow=c(1,2))
```

```r
plot(mod$fitted.values, mod$residuals, ylab='Residuals', xlab='fitted values',
     main='Residuals vs fitted values')
abline(h=0, col='red')
plot(mod$fitted.values, rstandard(mod), ylab='Standardized Residuals', xlab='fitted values',
     main='Standardized residuals vs fitted values')
abline(h=0, col='red')
#
# scatter plots vs each covariate (are there some non-linearities?):
head(mod$x)
par(mfrow=c(3,2))
for (i in c(2:6)){
  plot(mod$x[,i], rstandard(mod), ylab='Standardized Residuals', xlab=colnames(mod$x)[i],
       main=paste('Standardized residuals vs',colnames(mod$x)[i], sep=' ') )
}

#squared residuals
Sq.res<-resid(mod)^2
par(mfrow=c(1,1))
plot(Sq.res~fitted(mod), xlab=expression(paste('estimated values',hat('y'), sep=' ')),
     ylab=expression(paste('squared residuals',hat('e')^2, sep=' ')))


# We can obtain leverages by the function hat (it stands for hatvalues) applied to the model.matrix:
lev<-hat(model.matrix(mod))
plot(lev,ylab="Leverages",main="Index plot of Leverages")
lev.t<-2*ncol(model.matrix(mod))/nrow(model.matrix(mod)) # threshold leverage, twice the average level
abline(h=lev.t, col='red')
# units with high leverage:
h.l<-cbind(which(lev > lev.t),lev[c(which(lev > lev.t))]) #there are 25 cases with high leverage
h.l
#
# Cook's distances are given by the following function:
ckd<-cooks.distance(mod)
ckd
plot(ckd, main="Cook distance")
abline(h=4/length(ckd), col='green')
d.inf<-ckd<= 4/length(ckd)
table(d.inf) # influential observations
#

qqPlot(mod, main = "Q-Q Plot for Standardized residuals, model", col = "darkgrey")


# let's try to see how many of these are outliers:
print(outl<-outlierTest(mod))

# let's then try to exclude all influential observations:
m2 = update(mod, subset = ckd<=4/length(ckd))
summary(m2)
plot(m2)

#Normality of residuals
summary(m2$residuals)
skewness(m2$residuals)
kurtosis(m2$residuals)
```

```
# skewness(m4$residuals) = 0.1891498
# kurtosis(m4$residuals) = 2.706209
plot(m2, which = 2) # almost normal
shapiro.test(resid(m2))
ks.test(m2$res, "pnorm")
jarque.bera.test(m2$residuals)

# Histogram + normal curve
h <- hist(m2$residuals, col = "blue", xlab = "Residuals", freq=F, nclass = 50)
xfit <- seq(min(m2$residuals), max(m2$residuals), length=50)
yfit <- dnorm(xfit, mean=mean(m2$residuals), sd=sd(m2$residuals))
lines(xfit, yfit, col="red", lwd=2)

# VIF
vif(m2)
cor(m2$model[,c(1,2,5,7)])

#HOMOSCEDASTICITY
# Residuals vs fitted - we should have randomly located points
plot(m2, which=1)

# Breusch-Pagan test (it yeilds the rejection of H0, thus there is Heteroscedasticity)
#-> H0: homoscedastic residuals
bptest(m2, studentize = TRUE)
bptest(m2, studentize = FALSE)

resid<-residuals(m2)
t.test(resid)
yfit<-fitted(m2)
summary(lm(abs(resid) ~ yfit))

#serial correlation
n<-length(resid)
plot(resid[-n], resid[-1])
dwtest(m2)

# Underfitting, check the autocorrelation of the excluded variable
dwtest(m2, order.by=df$inc.norm[ckd<(4/length(ckd))])


#-----------------------------------------------------------------------------
# Weighted Least Squares Regression (in order to solve heteroscedasticity)
# define weights to use
wt <- 1/lm(abs(mod$residuals) ~ mod$fitted.values)$fitted.values^2

# weighted least squares regression
wls_model <-lm(ysqr ~ gross.income + as.character(education) + residence.size
         + employee.income + mutual.funds + time.job, data = df, weights=wt)
summary(wls_model)
xtable(wls_model)
plot(wls_model)

bptest(wls_model, studentize = T)

# We can obtain leverages by the function hat (it stands for hatvalues) applied to the model.matrix:
```

```
lev<-hat(model.matrix(wls_model))
plot(lev,ylab="Leverages",main="Index plot of Leverages")
lev.t<-2*ncol(model.matrix(wls_model))/nrow(model.matrix(wls_model)) # threshold leverage, twice the average level
abline(h=lev.t, col='red')
# units with high leverage:
h.l<-cbind(which(lev > lev.t),lev[c(which(lev > lev.t))]) #there are 25 cases with high leverage
h.l
#
# Cook's distances are given by the following function:
ckd<-cooks.distance(wls_model)
ckd
plot(ckd, main="Cook distance")
abline(h=4/length(ckd), col='green')
d.inf<-ckd<= 4/length(ckd)
table(d.inf) # 148 influential observations
#
qqPlot(wls_model, main = "Q-Q Plot for Standardized residuals, model", col = "darkgrey")

# let's try to see how many of these are outliers:
print(outl<-outlierTest(wls_model))

# let's then try to exclude all influential observations:
wls2 = update(wls_model, subset = ckd<=4/length(ckd))
summary(wls2)
xtable(wls2)
plot(wls2)

#RESIDUALS ANALYSIS
summary(wls2$residuals)
skewness(wls2$residuals)
kurtosis(wls2$residuals)
shapiro.test(wls2$residuals)
jarque.bera.test(wls2$residuals)

resettest(wls2, power = 2:3, type = c("fitted"))

# Overfitting, we use the Variance Inflation Factor (VIF) to check whether or not there is multicollinearity
vif(wls2)

### HETEROSKEDASTICITY ###
# Breusch-Pagan test
bptest(wls2, studentize = T)  # the residuals are homoskedastic
plot(wls2)


# DW test
# Error in dwtest(wls2) : weighted regressions are not supported
### RESIDUALS ANALYSIS ###
par(mfrow=c(1,1))
resid<-residuals(wls2)
t.test(resid)
shapiro.test(resid)
qqnorm(scale(resid))
abline(0,1)
```

```
model<-formula(wls2)
bptest(model,data=df)
dwtest(model,data=df)

yfit<-fitted(wls2)

# Homoschedasticity
summary(lm(abs(resid) ~ yfit))
#serial correlation
n<-length(resid)
plot(resid[-n], resid[-1])


#------------------------------------------------------------
# WLS using transformations of the independent variables
df$log.time.job = log(df$time.job)
df <- df[!is.infinite(rowSums(df)),]

m3_sqr<-lm(ysqr ~ log(gross.income) + as.character(education) + log(residence.size)
           + employee.income + mutual.funds + log.time.job, data = df, x=T)
summary(m3_sqr)

mod=m3_sqr
wt <- 1/lm(abs(mod$residuals) ~ mod$fitted.values)$fitted.values^2

# weighted least squares regression
wls_model <-lm(ysqr ~ log(gross.income) + as.character(education) + log(residence.size)
               + employee.income + mutual.funds + log.time.job, data = df, weights=wt)
summary(wls_model)
xtable(wls_model)
plot(wls_model)

# We can obtain leverages by the function hat (it stands for hatvalues) applied to the model.matrix:
lev<-hat(model.matrix(wls_model))
plot(lev,ylab="Leverages",main="Index plot of Leverages")
lev.t<-2*ncol(model.matrix(wls_model))/nrow(model.matrix(wls_model)) # threshold leverage, twice the average level
abline(h=lev.t, col='red')
# units with high leverage:
h.l<-cbind(which(lev > lev.t),lev[c(which(lev > lev.t))]) #there are 25 cases with high leverage
h.l
#
# Cook's distances are given by the following function:
ckd<-cooks.distance(wls_model)
ckd
plot(ckd, main="Cook distance")
abline(h=4/length(ckd), col='green')
d.inf<-ckd<= 4/length(ckd)
table(d.inf) # 148 influential observations
#
qqPlot(wls_model, main = "Q-Q Plot for Standardized residuals, model", col = "darkgrey")

# let's try to see how many of these are outliers:
print(outl<-outlierTest(wls_model))

# let's then try to exclude all influential observations:
```

```r
wls2 = update(wls_model, subset = ckd<=4/length(ckd))
summary(wls2)
xtable(wls2)
plot(wls2)

#RESIDUALS ANALYSIS
summary(wls2$residuals)
skewness(wls2$residuals)
kurtosis(wls2$residuals)
shapiro.test(wls2$residuals)
jarque.bera.test(wls2$residuals)

resettest(wls2, power = 2:3, type = c("fitted"))

# Overfitting, we use the Variance Inflation Factor (VIF) to check whether or not there is multicollinearity
vif(wls2)

### HETEROSKEDASTICITY ###
# Breusch-Pagan test
bptest(wls2, studentize = T)  # the residuals are homoskedastic
plot(wls2)
```