MULTIVARIATE LINEAR REGRESSION MODEL ESTIMATION ON THE EUROPEAN CENTRAL BANK CORE VARIABLES CATALOGE IN 2012

Namazli Ilkin, Singhateh Bubacarr, Foojan Mehraban, Emin Aker, Seyedeghbal Ghodsmofidi

**PROJECT WORK MAY 2022**

**ABSTRACT:**

The aim of this paper is to make a multivariate linear regression analysis of the survey done in Germany in the year 2012 for household consumption and income. We have selected some variables of our interest to compute our analysis by checking the effects of the dependent variable on the independent or explanatory variables. The source of this data set variables are from the European Central Bank Eurosystem, and HFCS core variables catalogue.

This analysis will be capture by looking first for the stationarity of the data and to perform some statistical inferential regression analysis.

**INTRODUCTION**:

This multivariate statistical analysis that we conducted in this work that have enable us to make some empirical interpretation of our results. It has helps us to know the relationship between the dependent variables and the explanatory variables. There are many statistical programming applications or software’s that can be used to perform this analysis but we have perform our analysis on the R studio.

The form of the regression model is as follows:

Y = ƒ(X1,X2,……,Xm) + ϵ = ƒ(X)+ϵ

Y is the phenomenon that will explain the component ƒ(X), represent the relationship that connect X and Y variables of the model while ϵ is the random term of the disturbance and represents all those factors that induce effects on the Y variable.

If the linear relationship is made explicit, then we obtain this equation:

Y = β1X1+β2X2+……+βKXK + ϵ

In doing this, the linear regression model was applied to some of the variables provided by the dataset. This consisted in selecting suitable set of explanatory variables to construct the model of linear regression that would be compatible with the economic theory of our analysis.

**Description of Variables**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| **HI0100** | Food consumption at home, numerical value in EUR |
| **HB0900** | Current price of household main residence, numerical value in EUR |
| **HB2300** | Monthly amount paid as rent, numerical value in EUR |
| **DN3001** | Net wealth, numerical value in EUR |
| **PA0100\_1** | Maritul status: single/never married, dummy variable |
| **PA0100\_2** | Maritul status: maried, dummy variable |
| **PA0100\_4** | Widowed, dummy variable |

**Numerical Summary of Variables**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Min** | **1st Quantile** | **Median** | **Mean** | **3rd Quantile** | **Max** |
| **HI0100** | 50 | 400 | 500 | 556,7 | 700 | 2.000 |
| **HB0900** | 5.000 | 137.000 | 200.000 | 238.042 | 300.000 | 3.000.000 |
| **HB2300** | 9 | 200 | 350 | 349.7 | 466.5 | 20.000 |
| **DN3001** | -44.600 | 91.000 | 210.584 | 312.524 | 375.750 | 6.480.000 |
| **PA0100** | 1 | 2 | 2 | 2,503 | 4 | 5 |

**Summary of the regression model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Coefficients** | **Estimate** | **Std. Error** | **t-value** | **p-value** |
| **Intercept** | 4.112e+02 | 2.722e+01 | 15.107 | < 2e-16 \*\*\* |
| **DN3001** | 1.155e-04 | 1.246e-05 | 9.273 | < 2e-16 \*\*\* |
| **HB0900** | 2.370e-05 | 3.012e-05 | 0.787 | 0.4314 |
| **HB2300** | 4.618e-02 | 2.744e-02 | 1.683 | 0.0925 . |
| **PA0100\_1** | -5.083e+01 | 3.282e+01 | -1.549 | 0.1216 |
| **PA0100\_2** | 1.583e+02 | 2.571e+01 | 6.157 | 9.36e-10 \*\*\* |
| **PA0100\_4** | -7.084e+01 | 2.779e+01 | -2.549 | 0.0109 \* |
| \*\*\* = significant at 1%  \*\* = significant at 5%  Adjusted = 0.2145 | | | | |

The above table presents the summary of the regression model for each coefficients. As can be seen from the table, two variables; DN3001 and PA0100\_2 are statistically significant at 1% level, as well as the intercept. Additionally, PA0100\_4 is significant at 5% level and HB2300 at 10% level. The other two variables are not statistically significant. Also, the adjusted value is around 21%. This means that only 21% of the variability is explained by the regression model.

Before going into the interpretation of the regression output, we need to check several assumptions to assess the validity of the results.

**Assumptions for Multiple Linear Regression**

In this section, linearity, multicollinearty, homoscedasticity, and normality assumptions are analyzed.

**Linearity:** There must be a linear relationship between the dependent variable and the independent variables. So the parameters of the regression equation is as the following:

HI0100 = 𝛽₀ + 𝛽₁(DN3001) + 𝛽₂(HB0900) + 𝛽₃(HB2300) + α1(PA0100\_1) + α2(PA0100\_2) + α3(PA0100\_4) + u

A picture containing text, sky, map

Description automatically generated

As can be seen from the scatter plots, each independent variables has a linear relationship with the dependent variable. Hence, the linearity assumption is met.

**Multicollinearity**: If there are high correlations among the variables, it can be said that there is a multicollinearity problem in the analysis. To check this issue, Variance Inflation Factor(VIF) is used. The formula for VIF is given below. It theoretically suggests that if the VIF value is higher than 10, then the analysis contains a serious multicollinearity. For our variables the VIF values are lower than 10.

VIF =

**VIF Values**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **DN3001** | **HB0900** | **HB2300** | **PA0100\_1** | **PA0100\_2** | **PA0100\_4** |
| 1.024264 | 1.000458 | 1. 003048 | 2.127924 | 4.313189 | 3.683605 |

Moreover, to see this relationship visually, the correlation matrix is also provided. As can be seen from the chart below there is no serious correlation among the variables. Therefore, this assumption is also satisfied.

Chart, bubble chart

Description automatically generated

**Homoscedasticity** refers to the situation where the variance of error term is constant so that they don’t vary too much depending on the change in the predictor variable. Therefore, ideally, the model should be homoscedastic, rather than being heteroscedastic. In order to check the presence of heteroscedasticity, Goldfeld-Quandt test is used. The null and alternative hypothesis are as the following:

H0: variances of the errors are constant (homoscedastic)

Ha: variances of the errors are not constant (heteroscedastic)

and the output can be seen in the table.

|  |  |  |  |
| --- | --- | --- | --- |
| GQ = 0.79423 | df1 = 794 | df2 = 794 | p-value = 0.9994 |

Since the p-value is not less than 0.05, we fail to reject the null hypothesis. We do not have sufficient evidence to say that homoscedasticty is present in the regression model.

In order to be confident with this result, we may also see Scale-Location Plot. It helps to check “homoscedasticity” assumption for the residuals of the regression model. To meet the assumpion of homoscedasticity, the red line must be horizontal. For our model, the plot has not a fully horizontal line and deviates too much. So one can conclude that the homoscedasticity assumption is not met in this case.

Chart, scatter chart

Description automatically generated

**Normality** refers to normal distribution of error terms. There are several ways to test this assumption. First, let’s have a look at the distribution of error terms.

Chart, histogram

Description automatically generated

As can be seen from the histogtram, distribution of the residuals seems positvely skewed rather than following a normal distribution. To be confident with this finding, Jarque Bera Test can be conducted.

The formula for this test as following,

Where S is for skewness and K is for Kurtosis.

Table X shows the output of JB-test.

|  |  |  |
| --- | --- | --- |
| X-squared = 1143.3 | df = 2 | p-value < 2.2e-16 |

The p-value of this test also suggests that the residuals are not normally distributed.

Lastly, Normal Q-Q Plot is useful to see the distribution of residuals in the regression model. If the points are on the dashed line or not deviate too much from the line, it can be said that the distribution of residuals are normal. Therefore, from the plot above, the residuals of this model are not normally distributed since quite high number of observations lie outside the normality line(dashed line).

Chart, line chart

Description automatically generated

**Linearity of the residuals**: Residuals vs. Fitted Plot is used to see if there is a non-linear pattern among the residuals. If the red line is roughly horizontal and in line with the dashed line, then it can be said that the residuals are linear. In our case, most of the points are accumulated on the lower bounds and their behaivor seems quite linear. Only the outliers lead to a deviation in the red line. However, from a general look, we can say that the residuals follow a linear pattern and thus linear regression model can be used in this case.

Chart, scatter chart

Description automatically generated

Before going into conclusion, we can also examine the influencial observations. Residuals vs. Leverage Plot helps to detect influential observations. The dashed red lines corresponds to the Cook’s distance. Hence, if some points fall outside of the lines, then it can be said that those points are influential observations. As can be seen from the above plot, there are quite high number of observations that fall outside of the Cook’s distance. Therefore, there are influential points in the dataset.

Chart, line chart, scatter chart

Description automatically generated

**New Model**

Since the original model has some problems such as heteroscedasticity and non-normal distribution of the error terms, we conducted a BoxCox Transformation. It is a statistical technique known to have a highly corrective effect on skewed data. The function determines the type of transformation from the non-normal distribution to the normal distribution according to the lambda (λ) parameter. For Box-Cox, the lambda (λ) parameter has the range -5 <λ <5.

After this transformation, the model is run again but excluding one insignificant variable which is HB0900. After this change, the new model becomes as the following:

**New\_model:** HI0100 = 𝛽₀ + 𝛽₁(DN3001) + 𝛽₂( HB2300) + α1(PA0100\_1) + α2(PA0100\_2) + α3(PA0100\_4) + ε

After we implemented this Box-Cox, distributions of the residuals became normal. This can be seen from the new histogram below.Chart, histogram

Description automatically generatedAlso QQ-Plot shows a similar recovery. As can be seen, more observations are now close to the dashed line.

Chart, line chart

Description automatically generated

In the case of heteroscedasticity, we rerun the Goldfeld-Quandt test for the new model. The resuls can be seen below.

|  |  |  |  |
| --- | --- | --- | --- |
| GQ = 1.0663 | df1 = 795 | df2 = 795 | p-value = 0.1827 |

Box-Cox transformation did not eliminate the heteroscedasticity problem but it led to a better output. Because the p-value is now lower than the original one(0.9994). However, it is still below the 5% benchmark and thus we still do not have sufficient evidence to say that homoscedasticity is present in the regression model.

The below table summarizes the new regression output. As can be seen from the table, the new model is better than the original one. Because all the variables are statistically significant now. Also, the adjusted R2 value increased by 4 points. Therefore, the model now explains 25% of the variance of the dependent variable around its mean.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Coefficients** | **Estimate** | **Std. Error** | **t-value** | **p-value** |
| **Intercept** | 1.076e+01 | 1.431e-01 | 75.224 | < 2e-16 \*\*\* |
| **DN3001** | 5.540e-07 | 6.798e-08 | 8.150 | 7.30e-16 \*\*\* |
| **HB2300** | 2.617e-04 | 1.497e-04 | 1.748 | 0.08062 . |
| **PA0100\_1** | -3.448e-01 | 1.791e-01 | -1.926 | 0.05434 . |
| **PA0100\_2** | 1.017e+00 | 1.403e-01 | 7.247 | 6.59e-13 \*\*\* |
| **PA0100\_4** | -4.899e-01 | 1.517e-01 | -3.230 | 0.00126 \*\* |
| \*\*\* = significant at 1%  \*\* = significant at 5%  Adjusted = 0.2569 | | | | |

**Conclusion**

To conclude, net wealth(DN3001), marital status of married(PA0100\_2), marital status of widowed(PA0100\_4), and as well as the intercept term is statistically significant. Additionally, although the other variables are not statistically significant at 1-5% levels, they are significant at 10% levels. Therefore, it is not rational to avoid their importance. Explaining the contributions of the independent variables, firstly, net wealth(DN3001) has a positive impact on the amount spent on food at home(HI0100). However, the degree of this relationship is quite low(5.540e-07 plus the intercept coefficient). Secondly, the monthly amount paid as rent(HB2300) is also positively affecting the dependent variable. But, yet, the impact of it is quite low as well(2.617e-04). The last three variables represent the dummies for marital status. The first one is for single people(PA0100\_1) and it can be seen that its negatively but slightly affects the response variable(-3.448e-01). The second one is for married people(PA0100\_2) and its impact on the response variable is positive but low(1.017e+00). The last one is for widowed people(PA0100\_4) and its impact in the model is similar to the one for single people, negative and low(-4.899e-01). These results are quite significant, however, there is still the problem of heteroscedasticity. Therefore a further test must be conducted to eliminate this problem. Some new variables can be included in the equation to achieve a better model.

**R-Script**

library(tidyverse)

library(forecast)

library(urca)

library(TSstudio)

library(quantmod)

library(aTSA)

library(AER)

library(stats)

library(moments)

library(tswge)

library(dplyr)

library(strucchange)

library(plotrix)

library(readxl)

library(xts)

library(highfrequency)

library(tseries)

library(ggplot2)

library(fBasics)

library(rugarch)

library(rumidas)

library(FinTS)

library(MCS)

library(psych)

library(car)

library(caTools)

library(MASS)

library(corrplot)

library(lattice)

library(strucchange)

library(lmtest)

library(nortest)

attach(datBI.txt)

#Dependent Variable

HI0100 <- na.omit(datBI.txt$HI0100) #amount spent on food at home

#Independent Variables

HB0900 <- na.omit(datBI.txt$HB0900) # current price of household main residence

DN3001 <- na.omit(datBI.txt$DN3001) # net wealth

PA0100 <- na.omit(datBI.txt$PA0100) # maritul status

HB2300 <- na.omit(datBI.txt$HB2300) # monthly amount paid as rent

#Arranging the data

DN3001 <- DN3001[-(1603:7951)]

HB0900 <- HB0900[-(1603:5954)]

HI0100 <- HI0100[-(1603:33333)]

PA0100 <- PA0100[-(1603:32423)]

HB2300 <- as.numeric(HB2300)

#Some summary statistics

summary(DN3001)

summary(HB0900)

summary(HB2300)

summary(HI0100)

summary(PA0100)

#Checking for normality

qqnorm(DN3001, main='Normal')

qqline(DN3001)

qqnorm(HB0900, main='Normal')

qqline(HB0900)

qqnorm(HB2300, main='Normal')

qqline(HB2300)

qqnorm(HI0100, main='Normal')

qqline(HI0100)

#Creating the dummies

install.packages('fastDummies')

library('fastDummies')

data <- data.frame(HI0100,DN3001, HB0900, HB2300, PA0100)

data\_dummy <- dummy\_cols(data, select\_columns = 'PA0100')

head(data\_dummy)

data\_dummy$PA0100 <- NULL

plot(model)

#Building the model

model <- lm(HI0100 ~ DN3001 + HB0900 + HB2300 + PA0100\_1 + PA0100\_2+ PA0100\_4, data = data\_dummy)

summary(model)

avPlots(model, main = "Linearity in the Parameters")

confint(new\_model, level=0.99)

vif\_values <- vif(model)

barplot(vif\_values, main = "VIF Values", horiz = TRUE, col = "steelblue") #create horizontal bar chart to display each VIF value

cor(data\_dummy[,-8], method = "pearson")

#Visualizing the cor matrix

install.packages("corrplot")

library(corrplot)

corrplot(cor(data\_dummy[,-8], method = "pearson"))

#Chow Test

sctest(HI0100 ~ DN3001 + HB0900 + HB2300 + PA0100\_1+ PA0100\_2+ PA0100\_4, data = data\_dummy, type = "Chow", point = 10)

#Goldfeld Quandt test to check heteroscedasticity

gqtest(model, order.by = ~DN3001 + HB0900 + HB2300 + PA0100\_1+ PA0100\_2+ PA0100\_4, data = data\_dummy)

#Breusch-Pagan Test for the same purpose

bptest(model)

residualPlots(model)

hist(residuals(model),col="peachpuff", border="black", prob = TRUE,xlab = "temp")

lines(density(residuals(model)),

lwd = 2, col = "chocolate3")

qqnorm(resid(model), main = "Normal Q-Q Plot")

qqline(resid(model), col = "blue")

#Residuals

jarque.bera.test(model$residuals)

layout(matrix(c(1,2,3,4),2,2))

plot(model)

ols\_plot\_resid\_hist(model)

chisq.test(model$fitted.values)

#finding optimal lambda for Box-Cox transformation

bc <- boxcox(HI0100 ~ DN3001 + HB2300 + PA0100\_1 + PA0100\_2+ PA0100\_4, data = data\_dummy)

lambda <- bc$x[which.max(bc$y)]

#fitting new linear regression model using the Box-Cox transformation

new\_model <- lm(((HI0100^lambda-1)/lambda) ~ DN3001 + HB2300 + PA0100\_1 + PA0100\_2+ PA0100\_4, data = data\_dummy)

summary(new\_model)

plot(new\_model)

hist(residuals(new\_model),col="peachpuff", border="black", prob = TRUE,xlab = "temp")

lines(density(residuals(new\_model)),

lwd = 2, col = "chocolate3")

#Goldfeld Quandt test to check heteroscedasticity

gqtest(new\_model, order.by = ~DN3001 + HB2300 + PA0100\_1+ PA0100\_2+ PA0100\_4, data = data\_dummy)