Influence Factors of the Economic Development Level Across European Countries

Diana Ioana POPA (diana.popa@insse.ro)
National Institute of Statistics Romania/
MA in Statistics student – Bucharest University of Economic Studies

ABSTRACT

The economic development level of a country refers to the measure of the progress in an economy that could be measured, especially through GDP or GDP per capita. The level of these indicators can be influenced by many factors as a large scale, from social and economical to environmental and government policies factors.

The paper aims to investigate some of these influence factors of the economic development level, represented in this case by GDP per capita, across European countries in the context of the most recently crisis, named the Great Recession (2008) and after, when the economies are starting to recover (2013).

Using linear regression in R (lm function), the goal is to explain the relationship between the interest variable (GDP per capita) and certain independent variables. It is expected that even tough the estimators are to be different – as level – in both cases studied, the relationship type between them to be the same. The goodness of fit for the models used will be made based on ANOVA.

Key words: *Economic Development Level, Multiple Linear Regression, R*
JEL Classification: *C1, C6, O1*

INTRODUCTION

GDP per capita is a measure of a country's economic development level that can be influenced by a multitude of factors. The main purpose of this paper is to discover through statistical tools available in R potential factors that influence it by analysing data from 31 European countries, in two different moments in time (2008 – the debut year of the most recent financial crisis and 2013 – the year considered the one where economies are starting to give signs of recovery). Several potential factors were considered from the range of socio-economic indicators, with the best fitted model being chosen for analysis.

The Multiple Linear Regression is used for modelling the relationship between a dependent variable and two or more independent variables. It uses the Least Squares Method for estimating the parameters for each independent variable.

METHODS

Linear Regression is a method used for modelling the relationship between a dependent variable and one or more independent (explanatory) variables.

In this research study Multiple Linear Regression models were applied to observe the influence of certain socio-economic factors on the economic development level. The goal is to explain the relationship between the interest variable (in this case GDP per capita was chosen as a measure to represent the economic development level of a country) and certain independent variables. Several independent variables were chosen (a total of 11) for testing. The best fitted model contains three of these variables.

In theory, the model takes the following form, given n observations:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, i = 1, 2, \dots, n,$$

with the estimated equation having the following form:

$$\hat{y} = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}, i = 1, 2, \dots, n$$

The b_0 parameter, also called intercept, shows the value of the dependent variable considering the independent variables as null. It is more useful in calculation since it does not have a definite significance.

The rest of b_i parameters show how the dependent variable (y) modifies when an independent variable (x_i) changes with one unit, when all other variables remain constant. The sign of the parameters shows the type of relationship between the dependent and independent variables; when the parameter is below 0 they are related in a negative linear sense, when the sign is positive there is a positive linear sense.

The confidence interval gives an estimated range of values for the estimated parameters, with a selected probability.

DATA SOURCE AND SOFTWARE USED

Data from the Eurostat database (http://ec.europa.eu/eurostat/data/database) was used in this research. As such, the comparability between countries is insured by the application of common definitions. The data refers to 31 European countries (Belgium, Bulgaria, Czech Republic, Denmark, Germany, Estonia, Ireland, Greece, Spain, France, Croatia, Italy, Cyprus, Latvia, Lithuania, Luxembourg, Hungary, Malta, Netherlands, Austria, Poland, Portugal, Romania, Slovenia, Slovakia, Finland, Sweden, United Kingdom, Iceland, Norway, Switzerland) in the reference years 2008, respectively 2013.

To compute the multiple linear regression model the `lm` function in R was used. The goodness of fit of the model was established using the `anova` function in R, and the confidence interval for the estimators was established using the `confint` function in R.

VARIABLES OF THE MODEL

The dependent variable representing the level of economic development is GDP per capita. The proposed independent variables for the model were (a total of eleven): the unemployment rate, the NEET rates (for the age groups 15 – 29 and 15 – 24 years of age), share of the working age population with upper secondary and post-secondary non-tertiary education (levels 3 and 4 ISCED), share of the working age population with tertiary education (levels 5-8 ISCED), share of GDP allocated to research and development (R&D), R&D expenditure per capita, life expectancy at birth, public health expenditure, the fertility rate and the rate of people at risk of poverty.

After testing various models with different combinations of independent variables, the best fitted model found has the following explanatory variables: the unemployment rate, the share of the working age population with tertiary education (levels 5-8 ISCED) and life expectancy at birth.

The final variables used are explained using the definitions from the metadata available on Eurostat in the rows below.

Dependent variable - GDP per capita:

- **GDP - Gross domestic product (*gpd_cap*):** GDP at market prices is the final result of the production activity of resident producer units. Data is transmitted in accordance with the European System of Accounts - ESA 2010. Population - a key auxiliary indicator - is used to derive GDP per capita, which is used as an indicator of economic welfare or material well-being.

Independent variables:

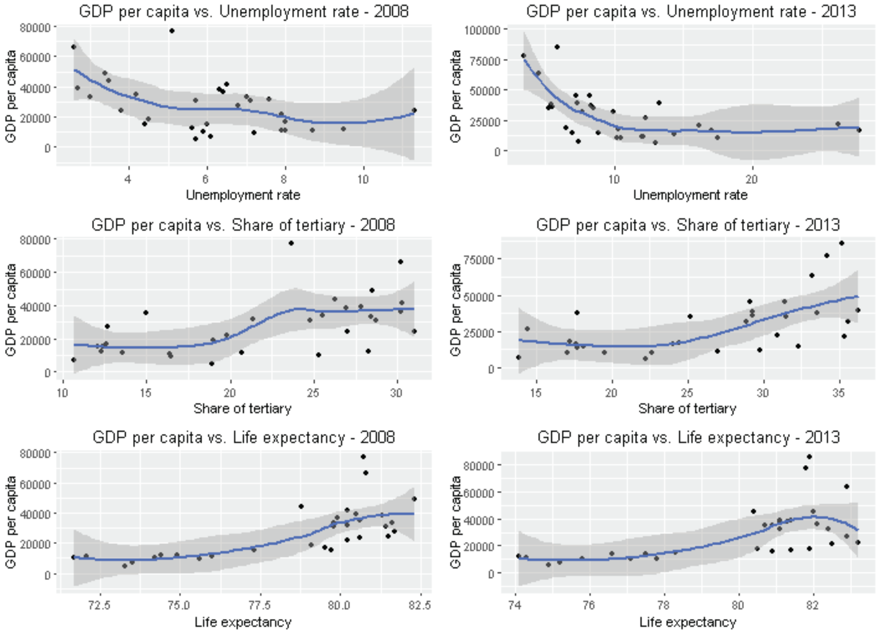
1. **Unemployment rate (*unempl_rate*):** Unemployment rates represent unemployed persons (persons aged 15-74 who were without work during the reference week, were currently available for work and were either actively seeking work in the past four weeks or had already found a job to start within the next three months) as a percentage of the active population (economically active population (labor force) comprises employed and unemployed persons). The estimates for the main labor market characteristics are provided from The European Union Labor Force Survey (EU-LFS). For this study, the unemployment rate of the working age population (15-64 years of age) was chosen.

2. **Share of the working age population with tertiary education (levels 5-8 ISCED) (sh_tert):** The educational attainment level of an individual is the highest ISCED (International Standard Classification of Education) level successfully completed .The tertiary education this covers ISCED 2011 levels 5, 6, 7 and 8 (short-cycle tertiary education, bachelor's or equivalent level, master's or equivalent level, doctoral or equivalent level, online code ED5-8 'tertiary education'). For this study, the share of the working age population (15-64 years of age) with tertiary education was chosen.
3. **Life expectancy at birth (life_ex):** Life expectancy at certain ages represents the mean number of years still to be lived by a person who has reached a certain exact age, if subjected throughout the rest of his or her life to the current mortality conditions (age-specific probabilities of dying). For this study, the life expectancy at birth was chosen.

The relationship between GPD per capita and the independent variables can be observed in **Figure 1** below. GDP per capita and the unemployment rate seem to be related in a negative linear sense, the higher the unemployment rate a decrease in GDP per capita is observed. Between GDP per capita and the share of the working age population with tertiary education appears to be a relationship with a positive linear sense, the same which applies to the GDP per capita and life expectancy at birth. When the share of the working age population with tertiary education or, respectively, life expectancy, increases, the GDP per capita takes higher values.

The relationship between GDP per capita vs. the independent variables, in 2008 and 2013

Figure 1



Source data: Eurostat

ANALISYS

The multiple regression equation of the best fitted model takes the following form:

$$gdp\_cap_i = \beta_0 + \beta_1 * unempl\_rate_i + \beta_2 * sh\_tert_i + \beta_3 * life\_ex_i + \varepsilon_i$$

The estimated equation can be written as follows:

$$\widehat{gdp\_cap_i} = b_0 + b_1 * unempl\_rate_i + b_2 * sh\_tert_i + b_3 * life\_ex_i$$

Two equations were computed initially, one for each year of interest. The full output obtained in R can be found in the Annex – Model 1 (2008), Model 2 (2013). In **Table 1** the coefficient estimates can be found, along with their standard errors and the confidence intervals. The estimated equations are:

Model 1 (2008):

$$\widehat{gdp\_cap} = -197405.1 - 1917.9 * unemp\_rate_i + 722.1 * sh\_tert_i + 2807.8 * \ life\_ex_i \quad (1)$$

Model 2 (2013):

$$\widehat{gdp\_cap} = -244511.7 - 1447.2 * unemp\_rate_i + 1001.4 * sh\_tert_i + 3283.2 * life\_ex_i \quad (2)$$

The multiple R- squared takes the value of 0.6259 for the 2008 model and 0.6935 in 2013, which suggests that over 60% of the variation is explained by the independent variables. The p-value in both cases suggests that the model is valid from a statistical point of view.

Coefficient estimates for Model 1 (2008) and Model 2 (2013)

Table 1

Variables / Year		Coeff. / Std. error	Confidence interval	
			Lower	Upper
Intercept	2008	-197405.1 (53977.5)	-308157.83119	-86652.4392
	2013	-244511.7 (61956.4)	-371635.7936	-117387.5940
Unemployment rate	2008	-1917.9 (1045.1)	-4062.40158	226.5135
	2013	-1447.2 (374.4)	-2215.4376	-679.0198
Share of tertiary	2008	722.1 (331.1)	42.75333	1401.3967
	2013	1001.4 (321.6)	341.6494	1661.1880
Life expectancy	2008	2807.8 (689.1)	1393.87693	4221.6541
	2013	3283.2 (805.5)	1630.4356	4935.9367

Source data: Eurostat

Coefficients interpretation – Model 1 (2008):

For the year 2008, the coefficients estimates are b_0 = -197405.1 (intercept), b_1 = -1917.9 (unemployment rate), b_2 = 722.1 (share of tertiary) and b_3 = 2807.8 (life expectancy). The p-value for the estimates suggests that all the estimates can be distinguished from 0, for the intercept, share of tertiary and life expectancy with a 95% of higher level of confidence, and for the unemployment rate with a 90% level of confidence.

- **b_1 = -1917.9**

As the unemployment rate increases with a percentage point, the GDP per capita decreases with 1917.9 Euro, when all other variables remain constant. The 95% confidence interval for the estimate ranges from -4062.4 to 226.51 Euro.

- **b_2 = 722.1**

As the share of the working age population with tertiary education increases with a percentage point, the GDP per capita increases with 722.1 Euro, when all other variables remain constant. The 95% confidence interval for the estimate ranges from 42.75 to 1401.39 Euro.

- **b_3 = 2807.8**

As the life expectancy at birth increases with one year, the GDP per capita increases with 2807.8 Euro, when all other variables remain constant. The 95% confidence interval for the estimate ranges from 1393.87 to 4221.65 Euro.

Coefficients interpretation – Model 2 (2013):

For the year 2013, the coefficients estimates are b_0 = -244511.7 (intercept), b_1 = -1447.2 (unemployment rate), b_2 = 1001.4 (share of tertiary) and b_3 = 2383.2 (life expectancy). The p-value for the estimates suggests that all the estimates can be distinguished from 0 with a 99% of higher level of confidence.

- **b_1 = -1447.2**

As the unemployment rate increases with a percentage point, the GDP per capita decreases with 1447.2 Euro, when all other variables remain constant. The 95% confidence interval for the estimate ranges from -2215.43 to -679.01 Euro.

- **b_2 = 1001.4**

As the share of the working age population with tertiary education increases with a percentage point, the GDP per capita increases with 1001.4 Euro, when all other variables remain constant. The 95% confidence interval for the estimate ranges from 341.64 to 1661.18 Euro.

- **b_3 = 2383.2**

As the life expectancy at birth increases with one year, the GDP per capita increases with 2383.2 Euro, when all other variables remain constant. The 95% confidence interval for the estimate ranges from 1630.13 to 4935.935 Euro.

Taking account of outliers

Because Luxembourg's GDP per capita is almost three times higher than the mean of the GDP per capita of the European countries selected, it is considered an outlier. Given the fact that outliers can negatively impact the results, another two models were computed after the elimination of the outlier. The new estimated equations are:

Model 3 (2008):

$$\widehat{gdp\_cap} = -178170.6 - 1690.8 * unemp\_rate_i + 736.7 * sh\_tert_i + 2523.5 * \ life\_ex_i \quad (3)$$

Model 4 (2013):

$$\widehat{gdp\_cap} = -231050.8 - 1294.6 * unemp\_rate_i + 812.8 * sh\_tert_i + 3141.0 * \ life\_ex_i \quad (4)$$

The full output obtained in R can be found in the Annex – Model 3 (2008), Model 4 (2013). In **Table 2** the coefficient estimates can be found, along with their confidence intervals.

Coefficient estimates for Model 3 (2008) and Model 4 (2013)

Table 2

Variables / Year		Coeff. / Std. error	Confidence interval	
			Lower	Upper
Intercept	2008	-178170.6 (38001.3)	-256283.3340	-100057.7691
	2013	-231050.8 (50556.4)	-334970.9131	-127130.6638
Unemployment rate	2008	-1690.8 (733.8)	-3199.0499	-182.5346
	2013	-1294.6 (307.4)	-1926.3737	-662.8098
Share of tertiary	2008	736.7 (232.1)	259.6531	1213.7082
	2013	812.8 (266.3)	265.3323	1360.2085
Life expectancy	2008	2523.5 (485.9)	1524.8130	3522.2138
	2013	3141.0 (656.8)	1791.0197	4490.9592

Source data: Eurostat

The multiple R- squared takes the value of 0.7495 for the 2008 model and 0.7276 in 2013, which suggests that over 70% of the variation is explained by the independent variables, an increase from the original models. The p-value in both cases suggests that the model is valid from a statistical point of view.

For the 2008 model (Model 3), the coefficients estimates are b_0 = -178170.6 (intercept), b_1 = -1690.8 (unemployment rate), b_2 = 736.7 (share of tertiary) and b_3 = 2523.5 (life expectancy). The p-value for the estimates suggests that all the estimates can be distinguished from 0 with a 95% or higher level of confidence. An increase in the level of significance for the variables can be observed, most notably in the case of the unemployment rate and the share of tertiary.

For the 2013 model (Model 4), the coefficients estimates are b_0 = -231050.8 (intercept), b_1 = -1294.6 (unemployment rate), b_2 = 812.8 (share of tertiary) and b_3 = 3141.0 (life expectancy). The p-value for the estimates suggests that all the estimates can be distinguished from 0 with a 99% or higher level of confidence, but in this case there are no notable differences from the previous model tested.

As level, the coefficients estimates are different in the years studied, but as expected, the relationship type between the dependent and independent variables remains the same. It is expected for the values of the estimates to be different; the data analyzed in the two cases is five years apart and in different socio-economic conditions – 2008 is the debut year of the most recent financial crisis, named the Great Recession; 2013 is considered to be the year when most of the countries economies are starting to recover, which was the reason behind it being chosen. The values of the estimates have changed in the two years studied as expected. The unemployment rate appears to have a slightly higher negative impact in 2008 on GDP per capita, which is to be expected; the life expectancy at birth increased in 2013 from 2008, and that has a somewhat higher positive impact, whereas the influence of the share of tertiary doesn't appear to influence the GDP level very much.

CONCLUSIONS AND RECOMMENDATION

From the eleven proposed independent variables only three were contained in the best fitted model. Thus was discovered that the unemployment rate, the share of the working age population and life expectancy at birth all influence the economic development level of a country.

As expected, in both years studied the relationship type between the dependent and independent variables remains the same; the increase of the unemployment rate has a negative impact on the economic development level, while the increase in the share of tertiary education and life expectancy has a positive effect.

As level, the coefficients estimates are different in the years studied, the data analyzed in the two cases is five years apart and in different socio-economic conditions; the unemployment rate appears to have a higher negative impact in 2008.

The present study takes into account just a few of the possible factors of influence of the economic development level, but many more factors are involved, from social and economical to environmental and government policies. The papers could serve as a starting point for future research that includes new variables from more than just the socio-economic sphere.

REFERENCES

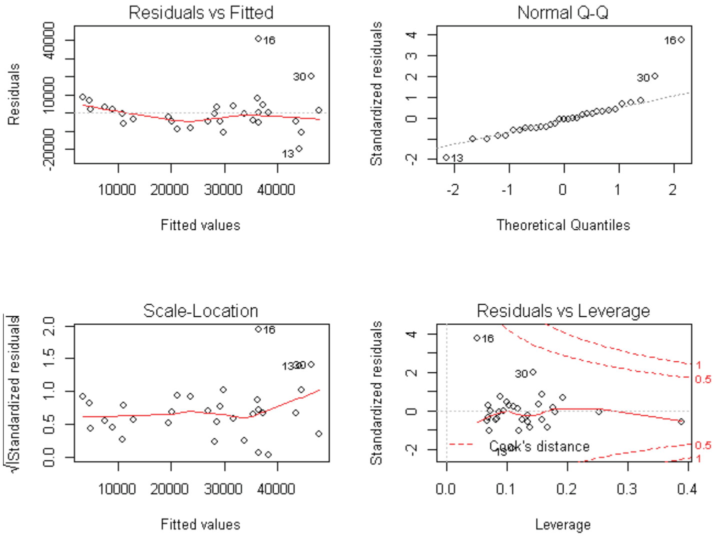
1. **Maindonald, J., Braun, W.J.**, 2010, *Data Analysis and Graphics Using R – an Example-Based Approach*, Third Edition, Cambridge University Press, New York, ISBN-13 978-0-511-71286-9
2. **Caragea, Nicoleta & Alexandru, Ciprian Antoniade & Dobre, Ana Maria**, 2012, “*Bringing New Opportunities to Develop Statistical Software and Data Analysis Tools in Romania*,” MPRA Paper 48772, University Library of Munich, Germany.
3. **Dobre Ana Maria & Caragea Nicoleta & Alexandru Ciprian Antoniade**, 2013, *“R versus Other Statistical Software,”* Ovidius University Annals, Economic Sciences Series, Ovidius University of Constantza, Faculty of Economic Sciences, vol. 0(1), pages 484-488, May.
3. **Nicoleta Caragea & Antoniade-Ciprian Alexandru & Ana Maria Dobre, 2014**, *“R – a Global Sensation in Data Science,”* Romanian Statistical Review, Romanian Statistical Review, vol. 62(2), pages 7-16, June.
4. **Caragea, N.**, 2015, *Statistica - Concepte şi metode de analiză a datelor*, Editura Mustang, Bucureşti, ISBN 978-606-652-063-8

ANNEX

Model 1 (2008):

```
> liniar_model1 <- lm(y ~ x1+x5+x8, data=date_08)
> summary(liniar_model1)

Call:
lm(formula = y ~ x1 + x5 + x8, data = date_08)

Residuals:
   Min     1Q Median     3Q    Max 
-20097  -5037   -636   3062  40487 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -197405.1    53977.5  -3.657 0.001088 ** 
x1            -1917.9     1045.1  -1.835 0.077531 .  
x5              722.1      331.1   2.181 0.038071 *  
x8             2807.8      689.1   4.075 0.000363 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11020 on 27 degrees of freedom
Multiple R-squared:  0.6259, Adjusted R-squared:  0.5844 
F-statistic: 15.06 on 3 and 27 DF,  p-value: 5.905e-06

> liniar_model2 <- lm(y ~ x1+x5+x8, data=date_13)
> summary(liniar_model2)

> anova(liniar_model1)
Analysis of Variance Table

Response: y
          Df     Sum Sq    Mean Sq F value    Pr(>F)    
x1         1 1890903989 1890903989  15.566 0.0005112 ***
x5         1 1580393166 1580393166  13.010 0.0012395 ** 
x8         1 2016789481 2016789481  16.602 0.0003632 ***
Residuals 27 3279820460  121474832                      
---
```

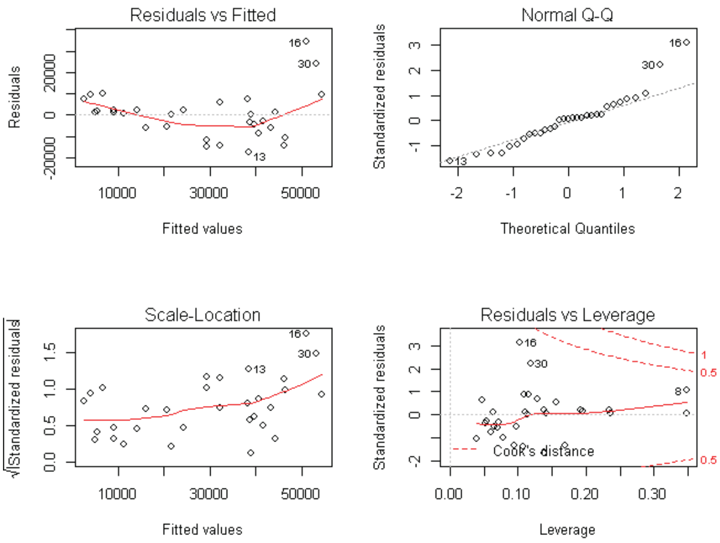
```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> confint(liniar_model1)
                    2.5 %      97.5 %
(Intercept) -308157.83119 -86652.4392
x1            -4062.40158    226.5135
x5               42.75333   1401.3967

x8             1393.87693   4221.6541
```



Model 2 (2013):

```
Call:
lm(formula = y ~ x1 + x5 + x8, data = date_13)

Residuals:
   Min     1Q Median     3Q    Max
-17501  -5974    552   4061  34207

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -244511.7    61956.4  -3.947 0.000510 ***
x1            -1447.2      374.4  -3.865 0.000631 ***
x5             1001.4      321.6   3.114 0.004333 **
x8             3283.2      805.5   4.076 0.000362 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11550 on 27 degrees of freedom
Multiple R-squared:  0.6935, Adjusted R-squared:  0.6594
F-statistic: 20.36 on 3 and 27 DF,  p-value: 4.199e-07
> anova(liniar_model2)
Analysis of Variance Table

Response: y
          Df     Sum Sq    Mean Sq F value    Pr(>F)
x1         1 2699351580 2699351580  20.220 0.0001177 ***
x5         1 3238490205 3238490205  24.259 3.723e-05 ***
x8         1 2217855769 2217855769  16.613 0.0003619 ***
```

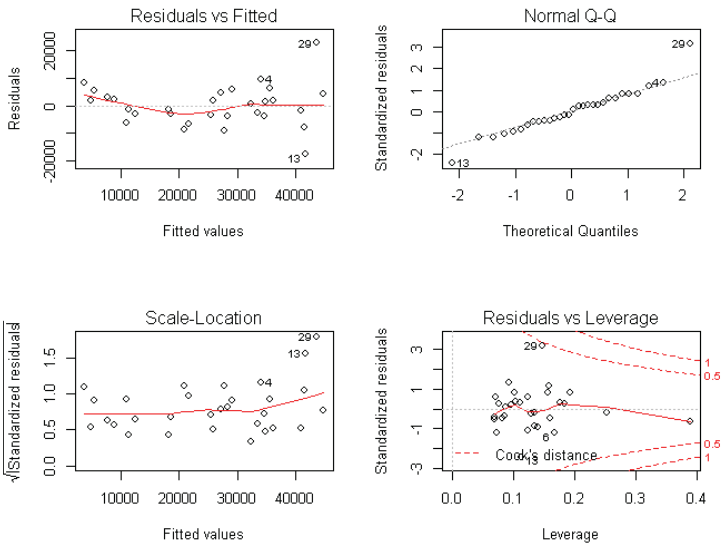
```
Residuals 27 3604439220  133497749
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> confint(liniar_model2)
                  2.5 %       97.5 %
(Intercept) -371635.7936 -117387.5940
x1            -2215.4376    -679.0198
x5              341.6494    1661.1880

x8             1630.4356    4935.9367
```



Model 3 (2008 – after outlier elimination):

```
> liniar_model3 <- lm(y ~ x1+x5+x8, data=date_08_fo)
> summary(liniar_model3)

Call:
lm(formula = y ~ x1 + x5 + x8, data = date_08_fo)

Residuals:
     Min       1Q   Median       3Q      Max
-17736.7  -3769.3   -209.9   3904.4  22819.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -178170.6    38001.3  -4.689 7.65e-05 ***
x1            -1690.8      733.8  -2.304  0.02945 *
x5              736.7      232.1   3.174  0.00384 **
x8             2523.5      485.9   5.194 2.02e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7725 on 26 degrees of freedom
Multiple R-squared:  0.7495, Adjusted R-squared:  0.7206
F-statistic: 25.93 on 3 and 26 DF,  p-value: 5.606e-08
```

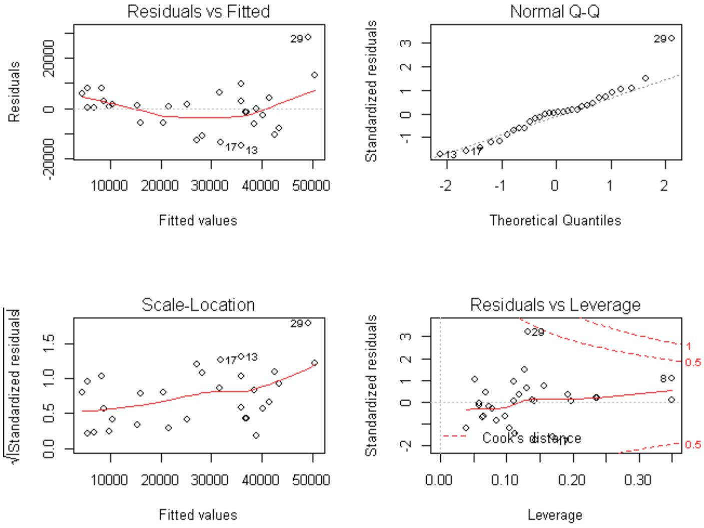
```
> anova(liniar_model3)
Analysis of Variance Table

Response: y
          Df     Sum Sq    Mean Sq F value    Pr(>F)
x1         1 1535780048 1535780048  25.735 2.772e-05 ***
x5         1 1496692275 1496692275  25.080 3.290e-05 ***
x8         1 1609854689 1609854689  26.977 2.016e-05 ***
Residuals 26 1551571656   59675833
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> confint(liniar_model3)
                   2.5 %       97.5 %
(Intercept) -256283.3340 -100057.7691
x1            -3199.0499    -182.5346
x5              259.6531    1213.7082

x8             1524.8130    3522.2138
```



Model 4 (2013 – after outlier elimination):

```
> liniar_model4 <- lm(y ~ x1+x5+x8, data=date_13_fo)
> summary(liniar_model4)

Call:
lm(formula = y ~ x1 + x5 + x8, data = date_13_fo)

Residuals:
     Min       1Q   Median       3Q      Max
-15010.0  -5680.6    465.9   3746.0  28252.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -231050.8    50556.4  -4.570 0.000105 ***
x1            -1294.6      307.4  -4.212 0.000268 ***
x5              812.8      266.3   3.052 0.005188 **
x8             3141.0      656.8   4.783 5.97e-05 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9405 on 26 degrees of freedom
Multiple R-squared:  0.7276, Adjusted R-squared:  0.6962
F-statistic: 23.15 on 3 and 26 DF,  p-value: 1.645e-07


> anova(liniar_model4)
Analysis of Variance Table

Response: y
          Df     Sum Sq    Mean Sq F value    Pr(>F)
x1         1 1899672971 1899672971  21.474 8.832e-05 ***
x5         1 2220222154 2220222154  25.098 3.275e-05 ***
x8         1 2023449200 2023449200  22.873 5.965e-05 ***
Residuals 26 2300029341   88462667
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> confint(liniar_model4)
                   2.5 %       97.5 %
(Intercept) -334970.9131 -127130.6638
x1            -1926.3737    -662.8098
x5              265.3323    1360.2085
x8             1791.0197    4490.9592
```


Residuals vs Fitted
Residuals
20000
0
-20000
10000
20000
30000
40000
50000
Fitted values
29
17
13
Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-1
0
1
2
Theoretical Quantiles
29
13
17
Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
10000
20000
30000
40000
50000
Fitted values
29
17
13
Residuals vs Leverage
Standardized residuals
3
2
1
0
-2
0.00
0.10
0.20
0.30
Leverage
29
8
17
1
0.5
0.5
Cook's distance