Social factors and GDP: a Multiple Regression approach

Agnese Adorante, Martina Del Gaudio

Introduction and research question

Measuring the prosperity and wellbeing of a country is not an easy task. There are multiple indicators of national richness nowadays, but one of the most used by economists is GDP, which measures the monetary value of final goods and services produced by a country in a year. It is a measure of the output of a country and even if it is not all that matters, it is an important indicator of welfare. Understanding in which way governments could improve the economy of their country would solve lots of issues, but factors impacting the growth are difficult to identify due also to the huge differences between countries themselves. Of course, both economic and social factors contribute to the GDP growth, but in this paper we will be mainly concerned about the social ones. More precisely, we want to see if it is indeed possible to use multiple linear regression to prove evidence about the influence of these factors to the economy, for the final aim of constructing a production function.
Firstly, we will do a multiple regression between GDP and labor force with capital, which are nowadays the recognized most influential factors on GDP and see if the multiple regression will agree and return evidence for the correlation between these variables. For the project we will indeed transform a non-linear function in a linear one to perform a regression and the consistency of the results is not immediate.
After that we will do the actual regression using social factors as independent variables. We decided to cover different aspects of society: variations in the population (defined by mortality rates of male and female and net migration), unemployment, fertility rate (a gender indicator) and education (primary school enrollments and number of people who have completed at least lower secondary). At the end of the paper these variables are better explained.

Data selection and data cleaning

For the project we collected data from almost 120 countries for a period of 60 years (from 1960 to 2021). Almost every data was collected from The World Bank Data, except for the labor force which comes from OECD. We tried to select explanatory variables that would represent without any discrimination more or less every country, both developed and undeveloped ones. Our variables model general features of a country's society as a whole, and indicate just some of the many variables influencing GDP. Also, we looked for data which were available from 1960, and this restricted our research even more.
We downloaded all the csv files from the internet and then we opened them on R. Here we did a lot of coding to glue all the columns with matching year, country and country code and we ended up with a 13.149 x 10 data frame. After that we canceled all the rows with missing values, and we reduced it to a 390x10 data frame. At a certain point we had to take the logarithm of the data for the purpose of the project, and this produced new NA values due to negative values, but these were few and were removed.
The absence of data limited our research, but still we remained with sufficient data to perform a statistically significant regression. This having said, the final sample was representative of more or less every country and covered a period of almost 50 years, thus we can safely say, statistically speaking, that our analysis will be sufficiently reliable.

Cobb-Douglas production function using multiple linear regression

A Cobb-Douglas production function models the relationship between production output and production inputs (factors). The general form of a Cobb-Douglas production function for a set of n inputs is:

$$Y = f(x_1, x_2, \ldots, x_n) = \gamma \prod_{i=1}^{n} x_i^{\alpha_i}$$

where Y is the output, x_i are the inputs, α_i the elasticity parameters for good i and γ the efficiency parameter.

The original and most used Cobb-Douglas production function is the one that relates GDP with labor force (L) and capital (K), and is of the form:

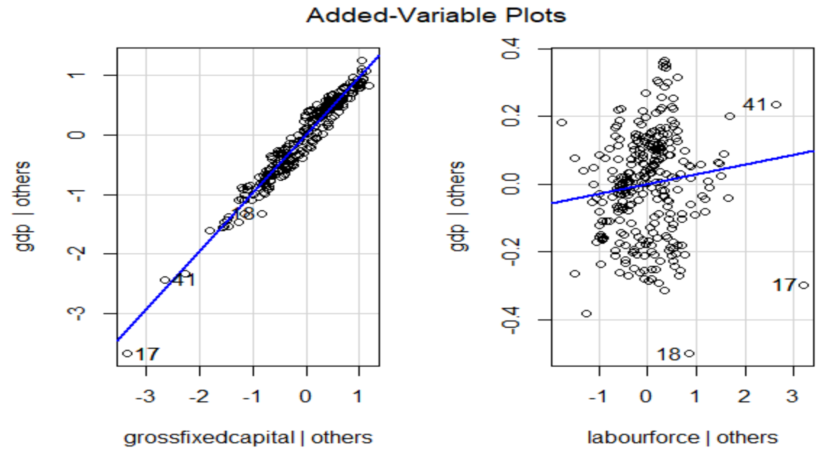
$$Y = AL^{\beta}K^{\alpha}$$

For the purpose of our study, we converted the production function to a linear model by taking the logarithm of both sides of the equation and obtaining the following formula:

$$logY = logA + \beta logL + \alpha logK$$

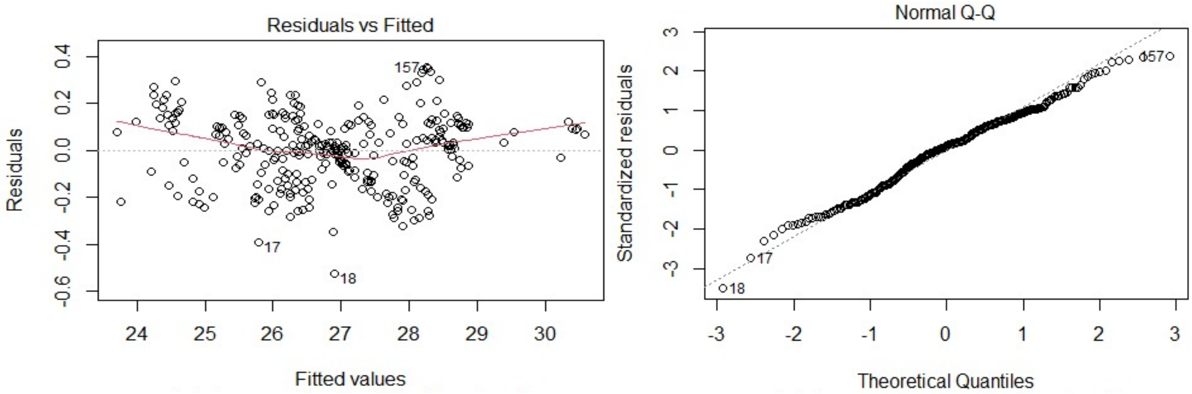
Note that this is still a linear model since the parameters are linear (functions of variables are allowed in linear models).

After taking the logarithm of all the data we proceed with some plotting to better visualize them, and we compute the correlation between the variables. From now on we will refer to log(data) as data to simplify the exposure. We get a correlation of 0.7774794 with labor and one of 0.9886621, which are both high indexes. Note that this doesn't mean necessarily that the variables will be significant in the model, since correlation and predictive power are different things. Doing the plots however it seems reasonable to suspect a relation between the variables, especially with capital where linear relation is almost obvious.



We then perform the linear regression and analyze the results [1]: all variables are statistically significant since their p-values are less than the 0.05 significance level and, as suspected, capital has a stricter relation with GDP than labor, which has a much larger p-value. What is quite surprising is the adjusted R-squared error of 0.9878, which indicates that our model explains almost perfectly well the change of variance in GDP. Really high R-squared usually is a symptom of overfitting, but our model only contains 2 variables, and they all are significant, thus making the hypothesis of overfitting unlikable.

To drive conclusions about the statistical significance of our test however we need to firstly check assumptions on normality and homoscedasticity of the residuals. The first thing we notice is that the residuals have a median approximately of 0 (mean of residuals is 0 and symmetric distribution have median =mean) and the 1Q and 3Q along with Min and Max have similar magnitudes. To investigate normality, we will look graphically at the Normal QQ-plot and for the spread of variance we will look at the "fitted vs residuals" graph. The results are quite comforting: they suggest no significant anomalies in homoscedasticity and univariate normality of the residuals. To be more precise we also performed some tests to check normality, among which only the Kolmogorov-Smirnov one did not reject the null hypothesis of the residuals being normally distributed with a p-value of 0.2660 (above the 0.05 significance level) [2].



We could have gotten better results about the residuals, but we must keep in mind that usually the p-values for normality tests of this kind and with such a lot of data are almost 0, so our analysis can be regarded as statistically significant.

We can now conclude our study going back to our original Cobb-Douglas function and putting the estimates we found in the regression to build an approximative production function. The result is

$$Y \;=\; 1.7486\,K^{0.97940}\;L^{0.02902}$$

Actually what we found is even more surprising: in Cobb-Douglas functions, assuming perfect competition and $\alpha + \beta = 1$, α and β can be shown to be capital's and labor's shares of output! In our model $\alpha + \beta \sim 1$, strengthening even more our hypothesis that regressions can be used for estimations of this kind.

"Social" production function using multiple linear regression

Now that we proved linear regression is indeed a reliable method for constructing production functions, we are ready to perform the multivariate linear regression on our social variables. The data we are using are still the log(data), and we will refer to them as data. As for before, we visualize our data by plotting each explanatory variable with the dependent variable and adding the corresponding regression line [3]. Actually, it doesn't seem to be that much of correlation except for net migration, while for other factors a more marked spread of variance might lead to a smaller adjusted R-squared. Also, we identify some outliers, but before removing them we will look if they are indeed influential in the analysis.

We will study both high-leverage points and outliers [4]. For the latter, since y-outliers are more common than x-outliers, we analyze the boxplot specifically for the "gdp" variable from which delightfully we can deduce the regularity of the data. For the first we begin with the regression and then analyze the "residuals vs leverage" plot but all points lay well inside the Cook's line, which barely appears in the graph, indicating low-leverage levels. We conclude there are no influential points in the model worth removing because this will not significantly change the results and eliminating data is always something to be taken care of.

```
Residuals:
     Min       1Q   Median       3Q      Max 
-2.14725 -0.48536 -0.08516  0.43860  2.37662 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          18.06642    5.79946   3.115 0.002028 ** 
fertilityrate         0.14449    0.29213   0.495 0.621264    
netmigration          0.64540    0.03019  21.376  < 2e-16 ***
unemployment          0.30307    0.11561   2.621 0.009234 ** 
primaryeducation     -0.07597    1.07661  -0.071 0.943796    
mortalityratefemale -0.01479    0.43993  -0.034 0.973212    
mortalityratemale    -0.30533    0.38117  -0.801 0.423796    
secondaryeducation    0.71837    0.20447   3.513 0.000515 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8044 on 282 degrees of freedom
Multiple R-squared:  0.6617,	Adjusted R-squared:  0.6533 
F-statistic:  78.8 on 7 and 282 DF,  p-value: < 2.2e-16
```

At a first glance the results seem more comforting than expected, but before driving conclusions we analyze the residuals to see if the assumptions are satisfied. As before the information about residuals in the summary of the regression are positive, since they have a median almost equal to 0 and the extremes are similar in magnitudes, thus indicating symmetry. We thus proceed with some plotting [5]: the histogram of residuals pretty much resembles a normal distribution, and the normal Q-Q plot also indicates some sort of normality except for the right tail which seems heavier than that of a normal distribution since points deviate from the 45-degree reference line. Moreover, the “residuals vs fitted” graph displays no significant change in the spread of the residuals, meeting the assumptions of the OLS regression. Before concluding we perform some normality hypothesis testing, but this time all the null hypotheses of the residuals being normally distributed are rejected: this is something we expected since the data we are taking into consideration take a large range of values. Despite of the good graphic result we should thus take in mind that the results we are going to obtain are not statistically significant.

From the summary of the regression, we see that there are 3 relevant independent variables: net migration (as expected from the plots), unemployment and secondary education, for which the t-test all gave p-values smaller than 0.05. A curious fact that we will investigate in the conclusions is that they all have a positive correlation with GDP, also unemployment, which we expected to have a negative correlation instead. As suspected the adjusted R-squared is smaller than in the first case, but still of some relevance, so to avoid the possibility of an overfitting which caused a high result we will perform a model selection. Finally, we look at the p-value obtained from the F-test: this is below the significance level, thus indicating that our model fits the data better than a model with only the intercept.

For model selection we try two different approaches: the first one uses hypothesis testing for every covariate, while the second one uses subset selection, so it chooses the best subset of independent variables matching some statistical criteria, which are highest adjusted R-squared and lowest penalty terms.
For hypothesis testing we use 3 methods: step-forward, step-backward and step-both.
The general idea of stepwise regression is to build a regression model by adding/removing predictors step-by-step, until the pre-set significance level is met for all predictors, by testing the null hypothesis of the estimate being 0 for every explanatory variable with a t-test.
The step forward method [6] starts with the null model and iteratively adds variables whose inclusion gives the most statistically significant improvement, stopping when the new variable added doesn’t improve the statistical significance of the model.
The step backward method [7] starts with the full model and at each step drops the variables with lowest significance level until all the variables remaining are statistically significant.
The step-both method is a combination of the preceding two models, and thus prevents two kinds of drawbacks: ignoring the fact that adding new variables to the model may render some of the existing ones not significant and not allowing a dropped variable to be included again in the model.

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	netmigration	addition	0.614	0.612	36.1280	733.1862	0.8507
2	secondaryeducation	addition	0.648	0.646	9.3410	708.0377	0.8132
3	unemployment	addition	0.656	0.652	5.0270	703.7280	0.8058
4	mortalityratemale	addition	0.661	0.657	2.2890	700.9008	0.8005

For all three stepwise testing, we get that the best model is the one with the four variables “net migration”, “secondary education”, “unemployment” and “mortality rate of male”. The last

variable became significant only after model selection: this can be explained by the fact that sometimes adding some variables may render the existing ones insignificant, so restricting the model can sometimes bring better predictive results.
The best subset method aims to find the subset of independent variables which best predict the dependent one. For this testing we use the ols_step_best_subset() function from the "olsrr" library [8]. We see that as for stepwise regression most of the tests performed indicate that the best fitting model is the one with the 4 variables aforementioned, with an adjusted R-squared of 0.6566 and lowest penalties for all methods, except for the SBC one which reaches the lowest value (722.0774) at the subset containing 3 variables ("mortality rate of male" is excluded).
To conclude we would choose a model with the four explanatory variables since most of the tests gave evidence for it.

Going back to our primary goal of constructing a production function depending on social variables, we run again a linear regression with the four most significant variables ("net migration", "secondary education", "unemployment" and "mortality rate of male") [9]. Again, we take the estimates found in the regression to build an approximative production function:

$$Y = 17.84096\, X^{0.64534}\, Y^{0.71898}\, Z^{0.29617} W^{-0.32651}$$

where 17.84096 is the intercept, X stands for net migration, Y for secondary education, Z for unemployment and W for male mortality rate.

Conclusions and Limitations of the study

We conclude with some final remarks. Multiple linear regression is indeed a good way for estimating production functions, given that the assumptions for a linear regression are satisfied. In the first case results on residuals indicated reliability of the study, while in the second one the situation about residuals was less precise. As mentioned at the beginning, we had to cut a lot of data from the original data frame and this surely impacted our study. This difference could also be due to the different nature of the explanatory variables: while capital and labor have a direct incidence on the measurement of GDP, the social factors we chose have to go by another step before actually impacting GDP: they influence society and society influences capital and labor, which then determine GDP. This however does not mean that our model is incorrect: we found a relationship between the dependent and independent variables, and this indicates that our study could be improved to give it statistical significance. One possible way to improve the study could be to focus on one country at a time, so that with less data the R-squared should be higher because of a more homogeneous variance. Also, we have to take into account that while for capital and labor there is precise and complete data for almost every country, it is not the same for social variables: probably with these factors we can make a reliable production function mostly for developed countries.
Nevertheless, the results of our study deserve to be discussed.
The negative correlation with male mortality rate is probably because in poorer countries (low GDP), mortality rates are higher because of lack in the health systems and in some countries labor force is still made mainly of men rather than women, thus explaining maybe why the correlation was less marked with female mortality rates. This study moreover shows that higher levels of education can predict a higher GDP, thus indicating that governments may invest in schools and incentivize a pursuit in the studies after mandatory education.
Net migration also has a positive correlation with GDP. This nowadays is an important topic of discussion in worldwide politics: some governments even in the most developed countries are adopting anti-immigration policies based on the belief of immigrants hindering the economic growth of the country. What evidence says, however, is different. We are not saying that there are no migration drawbacks, but we should also take into consideration the benefits of it: they account for the increase in the workforce and contribute to labor-market flexibility,

and they also potentially bring new perspectives and innovation to a country, thus contributing to economic growth.

Finally, we tried to give two explanations of the positive correlation between GDP and unemployment. The first one is purely statistical and concerns the lack of data for poorer countries: it is reasonable to think that the screening of employed people is more precise in richer countries. The second one is more an economic interpretation we gained from our previous studies: technological progress and automation of production processes decrease the need of human capital and thus decrease employment rates.

This however is a hypothesis that should be further investigated, maybe setting a model with unemployment as dependent variable and some technological indicators as independent ones.

Sitography (variables)

gdp - https://data.worldbank.org/indicator/NY.GDP.MKTP.CD
netmigration - https://data.worldbank.org/indicator/SM.POP.NETM
labourforce - https://data.oecd.org/emp/labour-force.htm
fertilityrate - https://data.worldbank.org/indicator/SP.DYN.TFRT.IN
unemployment - https://data.worldbank.org/indicator/SL.UEM.TOTL.NE.ZS
primaryeducation - https://data.worldbank.org/indicator/SE.PRM.ENRR
mortalityratefemale - https://data.worldbank.org/indicator/SP.DYN.AMRT.FE
mortalityratemale - https://data.worldbank.org/indicator/SP.DYN.AMRT.MA
grossfixedcapital - https://data.worldbank.org/indicator/NE.GDI.FTOT.CD
secondaryeducation - https://data.worldbank.org/indicator/SE.SEC.CUAT.LO.ZS

Description of the variables

- **gdp** GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. Data are in current U.S. dollars.
- **netmigration** Net migration is the net total of migrants during the period, that is, the number of immigrants minus the number of emigrants, including both citizens and noncitizens.
- **labourforce** The labor force, or currently active population, comprises all persons who fulfil the requirements for inclusion among the employed (civilian employment plus the armed forces) or the unemployed. The unemployed are defined as people without work but actively seeking employment and currently available to start work. It is measured in thousand of people.
- **fertilityrate** Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years. This indicator can indicate the status of women within households and a woman's decision about the number and spacing of children.
- **unemployment** Unemployment refers to the share of the labor force that is without work but available for and seeking employment.
- **primaryeducation** Primary school enrollment ratio is the ratio of total enrollment in primary school, regardless of age, to the population of age 6 to 11 years old.
- **mortalityratefemale** Female adult mortality rate is the probability of dying between the ages of 15 and 60, subject to age-specific mortality rates.
- **mortalityratemale** Male adult mortality rate is the probability of dying between the ages of 15 and 60, subject to age-specific mortality rates.
- **grossfixedcapital** Gross fixed capital formation includes land improvements, plant, machinery, and equipment purchases; and the construction of roads, railways, and the like, including schools, offices, hospitals, private residential dwellings, and commercial and industrial buildings. Data are in current U.S. dollars.
- **secondaryeducation** Lower secondary educational attainment is the percentage of population ages 25 and over that attained or completed lower secondary education.

Appendix

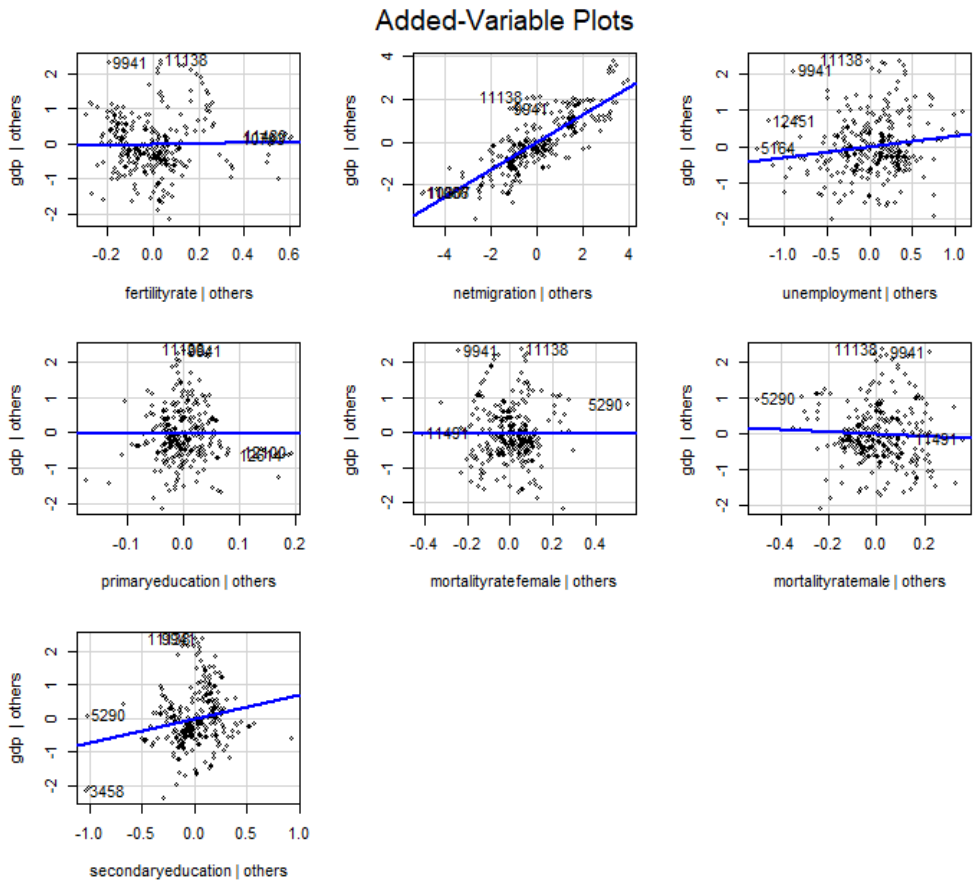
[1] Regression (1)

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.52570 -0.11209  0.01623  0.10846  0.35702 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        1.74866    0.22563   7.750  1.6e-13 ***
grossfixedcapital  0.97940    0.01280  76.526  < 2e-16 ***
labourforce        0.02902    0.01415   2.051   0.0411 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1508 on 287 degrees of freedom
Multiple R-squared:  0.9879,	Adjusted R-squared:  0.9878 
F-statistic: 1.172e+04 on 2 and 287 DF,  p-value: < 2.2e-16
```

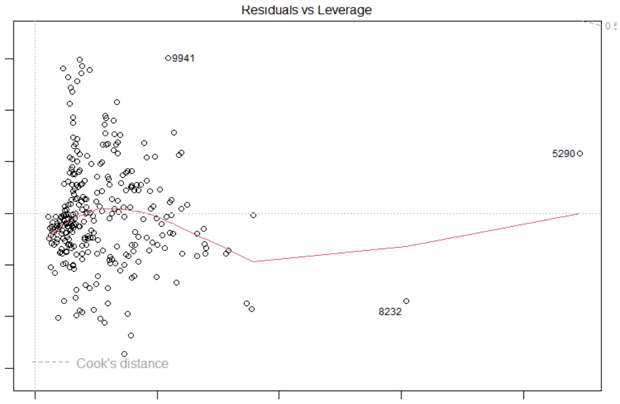
[2] Normality test for residuals (1)

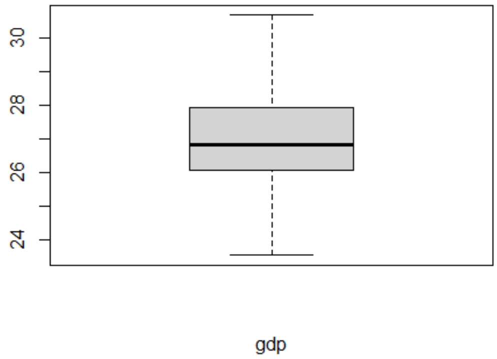
Test	Statistic	pvalue
Shapiro-Wilk	0.9898	0.0401
Kolmogorov-Smirnov	0.0589	0.2660
Cramer-von Mises	70.4516	0.0000
Anderson-Darling	1.0966	0.0070

[3] Added variable plot for regression with social factors

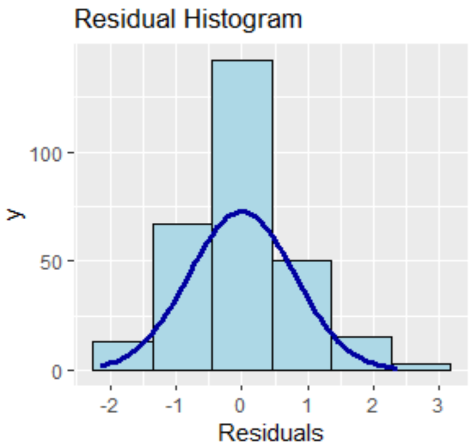


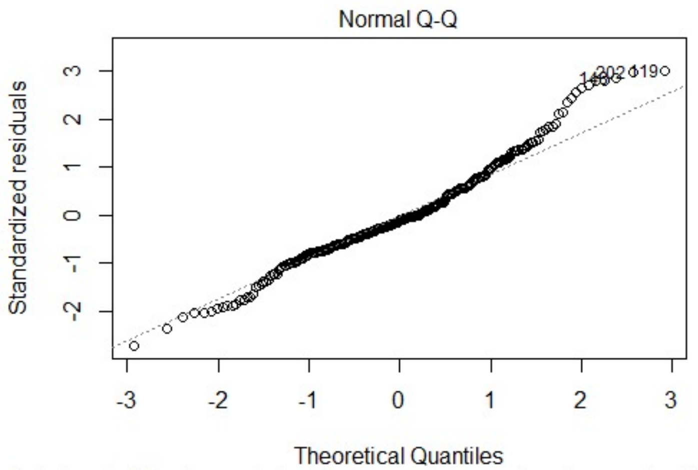
[4] Residuals vs Leverage Box plot of GDP

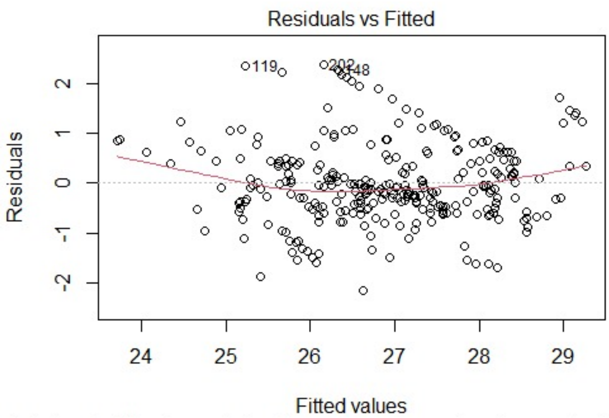




[5] Normality tests on residuals







Test	Statistic	pvalue
Shapiro-Wilk	0.9733	0.0000
Kolmogorov-Smirnov	0.0837	0.0344
Cramer-von Mises	31.1441	0.0000
Anderson-Darling	2.5365	0.0000

[6] Model selection, step forward method

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	netmigration	0.6136	0.6122	36.1283	733.1862	0.8507
2	secondaryeducation	0.6481	0.6457	9.3405	708.0377	0.8132
3	unemployment	0.6557	0.6521	5.0271	703.7280	0.8058
4	mortalityratemale	0.6614	0.6566	2.2890	700.9008	0.8005

[7] Model selection, step backward method

Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	mortalityratefemale	0.6617	0.6545	6.0011	704.6050	0.8029
2	primaryeducation	0.6617	0.6557	4.0066	702.6106	0.8015
3	fertilityrate	0.6614	0.6566	2.2890	700.9008	0.8005

[8] Model selection, best subsets method

Best Subsets Regression

Model Index	Predictors
1	netmigration
2	netmigration secondaryeducation
3	netmigration unemployment secondaryeducation
4	netmigration unemployment mortalityratemale secondaryeducation
5	fertilityrate netmigration unemployment mortalityratemale secondaryeducation
6	fertilityrate netmigration unemployment primaryeducation mortalityratemale secondaryeducation
7	fertilityrate netmigration unemployment primaryeducation mortalityratefemale mortalityratemale secondaryeducation

Subsets Regression Summary

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
1	0.6136	0.6122	0.6077	36.1283	733.1862	-90.2170	744.1959	209.8647	0.7287	0.0025	0.3918
2	0.6481	0.6457	0.6407	9.3405	708.0377	-115.0153	722.7172	191.7779	0.6681	0.0023	0.3593
3	0.6557	0.6521	0.6461	5.0271	703.7280	-119.1738	722.0774	188.3087	0.6583	0.0023	0.3540
4	0.6614	0.6566	0.6489	2.2890	700.9008	-121.8118	722.9201	185.8524	0.6519	0.0023	0.3505
5	0.6617	0.6557	0.6468	4.0066	702.6106	-120.0352	728.2998	186.3225	0.6558	0.0023	0.3526
6	0.6617	0.6545	0.644	6.0011	704.6050	-117.9839	733.9640	186.9796	0.6603	0.0023	0.3550
7	0.6617	0.6533	0.6406	8.0000	706.6038	-115.9282	739.6327	187.6443	0.6648	0.0023	0.3575

AIC: Akaike Information Criteria
SBIC: Sawa's Bayesian Information Criteria
SBC: Schwarz Bayesian Criteria
MSEP: Estimated error of prediction, assuming multivariate normality
FPE: Final Prediction Error
HSP: Hocking's Sp
APC: Amemiya Prediction Criteria

[9] Regression (3)

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.10997 -0.47787 -0.09619  0.41776  2.38223

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         17.84096    1.22277  14.591  < 2e-16 ***
netmigration         0.64534    0.02935  21.984  < 2e-16 ***
secondaryeducation   0.71898    0.16166   4.447 1.25e-05 ***
unemployment         0.29617    0.10949   2.705  0.00724 **
mortalityratemale   -0.32651    0.14929  -2.187  0.02954 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8005 on 285 degrees of freedom
Multiple R-squared:  0.6614,    Adjusted R-squared:  0.6566
F-statistic: 139.2 on 4 and 285 DF,  p-value: < 2.2e-16
```