HUMBOLDT-UNIVERSITÄT ZU BERLIN

School of Business and Economics
Insitute of Statistics

A Cross Country Analysis on the Effect of Income on Child Mortality

A report of

Grzegorzthehero, K. M. , J. S.

in the seminar "Statistical Programming Languages"
Summer 2018

Supervisor
XYZ

---

21. September 2018

1 Introduction

Up until the year 2000 new born deaths around the globe were rarely mentioned in global policy and programmes. Shortly after the 21st century had started, global organizations and country governments began to increase their attention for child mortality and especially new born deaths.

In this work we will analyze the relation between the child mortality rates for children under 5 years of age and the average Gross Domestic Product (GDP) in different countries around the globe.

Our hypotheses is, that the higher the GDP in a country/ region the lower the under – 5 child mortality rate. Therefore, we will analyze different datasets from CHERG and the Worldbank throughout our paper to see if we can confirm our hypotheses.

2 Literature Review

In Liu et. al. (2012) the total number of deaths of children up to 59 months around the globe are analyzed. Therefore, two groups are build. The first group is made up of the child mortality of children from 0 to 27 days, the neonates, and the second group from 1 to 59 months. The number of deaths are categorized for the different causes of deaths and diseases for different countries.

In the year 2000 9,6 million children died before reaching their fifth birthday. This number decreased significantly to only 7,6 million deaths of children under 5 years in 2010 due to the collective reduction in infectious causes. About 5 million, that is around 64,0%, out of those 7,6 million deaths were caused by infections, of which pneumonia with 14,1%, diarrhoea with 9,9% and malaria with 7,4%. were the main infectious disorders. 40,3% of the deaths occurred in neonates. Leading causes for neonatal death were preterm

birth complications with 14,1%, intrapartum-related complications with 9,4% and sepsis or meningitis with 5,2%.

The most children under 5 died in Africa (3.6 million) and southeast Asia (2,1 million) in 2010.

Medically certified vital registration data about the deaths in children under 5 years are available for only 61 countries, which corresponds only 2,7% of the 7,6 million deaths in 2010.

To generate the total numbers of deaths the country – specific under – 5 mortality rates (U5MR) and neonatal mortality rates are needed. The U5MRs used in Lui et al. (2012) are from the publication of the UN Integragency Group for Child Mortality Estimation (IGME), published in 2010. The neonatal mortality rates are developed by WHO. The estimated number of livebirths are used from the UN Population Devision 2010 revision.

In Lui et al. (2010), the total number of deaths in neonates and children under 5 years is estimated by putting the data of U5MPs by IGME and WHO about mortality rates in relation with the population of children younger than 5 years and the number of livebirths.

For different countries with different conditions about the data availability various methods are used to estimate and generate useful datasets. For example, for countries without adequate death registration and with an U5MR of 35 or below per 1000 livebirths in 2000 to 2010, multinomial logistic regression model is used to estimate the distribution of causes of deaths for neonates and older children separately. For older children univariate meta – regression methods and multi – variate stepwise forward ordinary – least - squares regression models are used.

On the other hand, causes of neonatal deaths are estimated by using a revised verbal – autopsy – data – based multi – cause model (VAMCM) for high – mortality countries with a U5MR above 35 per 1000 livebirths without adequate death registration from 2000

to 2010. Also within the different diseases in the eight categories, various methods are used to generate the numbers as precise as possible.

While using cross validation, 90% of the data is used to build the model and 10% is used as a test set.

The Child Health Epidemiology Reference Group (CHERG) is working together with WHO and UNICEF in order to be able to develop methodologies and processes that help estimating and publishing the total number of child mortality every year and their reasons.

Also in Liu et al. (2015), the distributions of causes of child mortality is estimated separately for neonates and children from 1 – 59 months for the time period of 2000 to 2013. New vital registration data and verbal autopsy data is included in order to generate cause – specific mortality fractions. The findings are similar to the findings of Lui et. al. (2012).

Moreover, estimated predictions of the numbers of child mortality for the years 2030 and 2035 due to updated available data are made.

In Lawn et al. (2012) it is stated that between 2000 and 2010 the regional variation of neonatal mortality rates change ranged from 3,0% per year in developed countries to 1,5% per year in sub – Saharan Africa. In some countries changes in income or fertility are the reason for the changing numbers of child mortality. Also, the amount of new born survival data and the evidence based increased, as well as recognition in donor funding.

The data visualizes that a progress in reducing most causes of death since 2000 has been made, mainly for tetanus, pneumonia and diarrhoea. However, deaths due to preterm birth complications are decreasing more slowly. They are now the second leading cause of child deaths.

3 Theory and Design

1. Tests

In our work we decided to conduct several statistical tests in order to check whether some data is normal distributed, whether the heteroskedasticity exists while regressing the data and whether the correlation of some variables is significant. For the first one we decided to test whether the "death_under_5_per_birth" is well-modeled by a normal distribution. In order to achieve this goal the Jarque-Bera Test on normal distribution, testing whether the sample data match the skewness and kurtosis of normal distribution. The test is approximated with Chi-squared test in R. The H_0 hypothesis of this test is: The variables are normal distributed, whereas the H_1: The variables are not normal distributed. Furthermore, to test the significance of correlation between two variables the correlation test has been conducted using one of Pearson's product moment correlation coefficient, Spearman's rho or Kendall's tau. As the last one the test on the homoskedasticity of the data the Goldfeld – Quandt test was conducted on three regressions. This Test analyzes the variances of 2 submodels split apart by a breakpoint and rejects if the variances are not even. Test follows F statistics and under H_0 the variance does not increase from segment 1 to 2.

2. Diagrams

In order to visualize data properly we decided to use such a diagram types as boxplots, dotplots with linear regression line, smoothed density function plot, and a world map plot showing the data for every country.

Boxplots "display only the most important information about the frequency distribution. Specifically, the boxplot contains the smallest and the largest observed values x1 and x2 and three quartiles $x_{0.25}$, $x_{0.5}$ and $x_{0.75}$. The second quartile $x_{0.5}$ is of course the median.

The quartiles are denoted by a line and the first and third quartile are connected so that we obtain a box." (Härdle 2015, 64)

On the other hand, a dotplot is a „two-dimensional graphical display of one-dimensional data. On horizontal axis, you find the observed value. The value on the vertical axis is arbitrary (usually randomly chosen)." (Härdle 2015, 510). With regression line within it can depict a trend in a data.

The smoothed density function plot "is (an approximate) density function of the (continuous) variable X". (Härdle 2015, 119)

In order to depict an observation for every country a map plot had to be used. It can visualize how strong the attribute is present in every country.

3. Regression

We use an ordinary least square regression to estimate the average effect of income on child mortality. Several conditions must be fulfilled for an ordinary least square regression to produce consistent results: the regressors must be exogenous and the errors should be homoscedastic and serially uncorrelated. We will not tackle the problem of exogeneity in our analysis. This means that in theory we cannot say if the effects we find are causal. Still it seems plausible to assume that correlation of child mortality and GDP is mainly driven by the GDP. A high child mortality rate in one year, should not have too much of a causal effect on GDP in the same year. To check if heteroscedasticity might bias our error, we will perform a statistical test to check for the homoscedasticity of the errors.

4 Implementation/Simulation

2. Quantlet: SPL_Correlation

In order to create diagrams with ggplot the package „ggplot2" is needed. After installing the package in line 8 it loads alongside with the package "RColorBrewer" which contains various color palletes which are used in the later code. The code in line 21 creating the graphic "d1" consists of "ggplot". In its brackets the data source with "aes()" function can be found. The latter describes how variables in the data are mapped to visualize the properties. By adding the funtion "geom_point()" in line 24 we define that every observation is placed on a diagram as a point. Inside of the brackets of "Aes()" the "color=final.df$region_who" gives every WHO region a different color so that it is possible to distinguish different regions on the graph. "Geom_smooth()" in line 25 adds the moving average to the graph and the "labs()" in line 26 formats the graph in such a way that the labels of x and y-axis can changed from the names of columns to the names in inverted commas. In line 28 "Title" gives the title to the graph and in line 29 "col" changes the title of the legend.

Both of the graphs "d2" and "d3" are dot plots which show the correlation between two variables in one region. Inside of "aes()" in lines 39 and 57 the variables, which should be displayed in the plot form the porper dataframes are defined. "geom_point()" creates a dot plot and inside of its brackets the color of the dots is set. In line 43, 44 and 61, 62 "ylim()" and "xlim()" set up the length of the coordinates and the "geom_smooth(method=lm)" produces a regression line. Again with the help of "labs()" the description of the coordinates and the title can be edited in lines 46 and 63.

In „d2ver2" in line 49, the second version of graph "d2", different lengths of the coordinates are set through function "ylim()" and "xlim()".

In order to make the comparison of both of the graphs easier, they can be arranged side by side on the same page. To accomplish this, a package "gridExtra" needs to be installed and loaded. In line 76 with help of "grid.arrange()" function a gtable layout can be set up to place multiple grobs on a page. The number of rows and columns of the diagrams can be also defined by the function.

3. Quantlet: SPL_Distribution

The next 2 graphs "density1" and "density2" in lines 3 and 14 are the kernel density diagrams for the distribution of a variable. With "fill" which demonstrates the usage of the subgroup of aesthetics the function is used to encode a vector as a factor. In lines 6 and 17 with help of "geom_density()" a new diagram type is defined and inside the brackets, with "fill" its color and with "Alpha" color transparency is modified. Again the corresponding labels are modified within "labs()" in lines 7 and 18.

The next graph "boxp1" in line 29 is a boxplot displaying the distribution of data showing the shape, variability and median of data. Inside of aesthetics in line 29 the x-axis and y-axis are defined. The statement "color = final. df$region_who" in line 31 gives every WHO region a different color. By adding "geom_boxplot()" in line 32 a box plot is produced in output. Again inside of "labs()" in line 33 the labels are described alongside with the title of the graph. If unnecessary, the legend of every diagram can be removed with "theme(legend.position = none)" as in line 36. In the end the "coord_flip()" in line 37 flips the coordinates so that the data are better visualized.

4. Quantlet: SPL_Probability

Thanks to proper calculations the probability of dying before being 5 when being born in 2010 can be calculated assumed that the data are not going to change in the next five years.

In order to do so a new column containing the probability for every country in any dataframe has to be added in line 8. A distribution of this probability can be depicted, line 11, in a boxplot. Different colors can be given to every category inside of "aes()" in line 11. By specifying the type of the plot, in line 14, the boxplot is produced in the output. Thanks to "coord_flip()" in line 18 the coordinates can be flipped in order to make the interpretation easier.

The best method to visualize data for every country is to plot it on the world map. In order to do it a package "rworldmap" needs to be installed, line 24. If it is not already installed the code in the "if" statement does so and with command "library()" the package loads. In line 29 the function "joinCountryData2Map()" joins user data referenced by country codes. The data frame has to be defined on the first place, then "joinCode = "IS03" tells how countries are referenced and in the next line the "nameJoinColumn" is the name of the column containing the country referencing, country codes. In line 32 the "par()" has to be performed only once for a session and it makes sure that the whole available space is going to be used while displaying the map.

"mapCountryData()", line 33, draws a map showing country-level data. Inside of its brackets "mapped_data" refers to previously coded statement from line 29, whereas "nameColumnToPlot" is the name of the column containing the data which should appear on the map. Moreover a title of the map is added with "mapTitle()" and color of countries for which the data is not available and the ocean color are added with "missingCountryCol" and "oceanCol" respectively.

5. Quantlet: SPL_Tests

Normal distriburion

In line 14 function conducting a test on normal distribution of data is defined. It uses the Jarque-Bera test on normal distribution. In line 16 inside of "jarque.bera.test()"

brackets “complete.cases()” makes sure that all rows with missing values are deleted and not considered in the test. The test can be applied by putting a variables inside of function “nor()” brackets.

Correlation

In line 26 “cor(x,y)” function tests the correlation between two variables. It is composed of “cor.test(x,y)”, test for association between paired samples, using one of Pearson’s product moment correlation coefficient, Kendall’s tau or Spearman’s rho. To use the function the arguments have to be put inside “cor()”, one as x and the second one as y.

Heteroskedasticity

To conduct another test on heteroskedasticity the package “lmtest” is needed. If it hasn’t been already installed the “if()” statement in line installs it in line 49 and “library()” loads it. To test whether the heteroskedasticity problem exists the function “het(x,y)” in line 51 is implemented. First in line 52 the function models a linear regression with “modell = lm(y∼x)” whereas x is the independent variable and y is the dependent variable. And in the next line it conducts a Goldfeld-Quandt Test against heteroskedasticity.

6. Quantlet: SPL_Regression

In order to test our main hypothesis that the GDP has a statistically negative effect on child mortality, we implement an ordinary least square model. In the simplest specification, implemented in line 33 we directly regress the under 5 child mortality in 2010 on the GDP. The estimated model is:

$$CHM_i = \beta_0 + \beta_1 ln(GDP)_i + \epsilon_i \tag{1}$$

where i is the country index, CHM is the number of children dying under the age of five, as a share of the number of birth and ln(GDP) is the natural logarithm of the Gross

Domestic Product. β_1 is the regression coefficient, β_0 is a constant and ϵ_i is the error term.

As a second step, we add two control variables to the regression, the income inequality and the birth rate. The model is coded on lines 38 - 41 The estimated model is:

$$CHM_i = \beta_0 + \beta_1 ln(GDP)_i + \beta_2 gini_i + \beta_3 BirthR_i + \epsilon_i \tag{2}$$

where gini is the gini coefficient and BirthR is the birthrate. β_2 and β_3 are regression coefficients.

In a next step we add region dummies which are 1 if the country is in the respective region and 0 otherwise. The model is coded on lines 45 - 52. The estimated model is:

$$CHM_i = \beta_0 + \beta_1 ln(GDP)_i + I_i + \epsilon_i \tag{3}$$

where I is the set of regional dummies.

In the ultimate specification we estimate the control variables and the region dummies together in one model. The model is coded on lines 57 - 66. The estimated model is:

$$CHM_i = \beta_0 + \beta_1 ln(GDP)_i + +\beta_2 Gini_i + \beta_3 BirthR_i + \beta_k I_i + \epsilon_i \tag{4}$$

.

where β_k is the set of regression coefficients for the region dummies.

5 Empirical Study/ Testing

1. Quantlet: SPL_Load_data

The datasets we load are from three different sources. The first source is Liu et al. (2012) for the total number of death children in 2010. This source has already been described extensively in the literature review. The second source we use is the Standardized World Income Inequality Database (SWIID) created by Solt (2016). From this source we derive the gini coefficient for 2010. The third source is the Worldbank data repository. We use it for the income data and for population and birthrate data. As loading the data is quite straightforward, we will not discuss it in detail here. To merge the observations from all three data sources, it is necessary to find a unique common identifier. The countrycode package provides help for this task (Arel-Bundock 2018). The package provides a function, which takes as input various different sets of country names or country codes and transforms the input set into another specified set. The package is first applied on line 54. The unique identifier is the English country name.

2. Quantlet: SPL_Correlation

In lines 15 and 16 data frames "final.df_Africa" and "final.df_Europe" are set as subsets of "final.df" with countries from African and European region respectively. New data frames are created containing only the information from European and African region. In "d1", line 22, "GDP_PPP_2010" is placed on x-axis and the "death_under_5_per_birth" on the y-axis. A dot diagram is expected to be produced in the output what actually happens. The graph shows the relation between the total child mortality and the income in the whole world. The diagram shows that the lower the GDP in every WHO region the higher the percentage of death under 5 per birth. It is visible that the percentage of death under 5 in the various regions can be differently high at the same GDP. For example, at an GDP of $\approx$ 2860 US - $ the under – 5 death in the WPRO region lies between 1% to

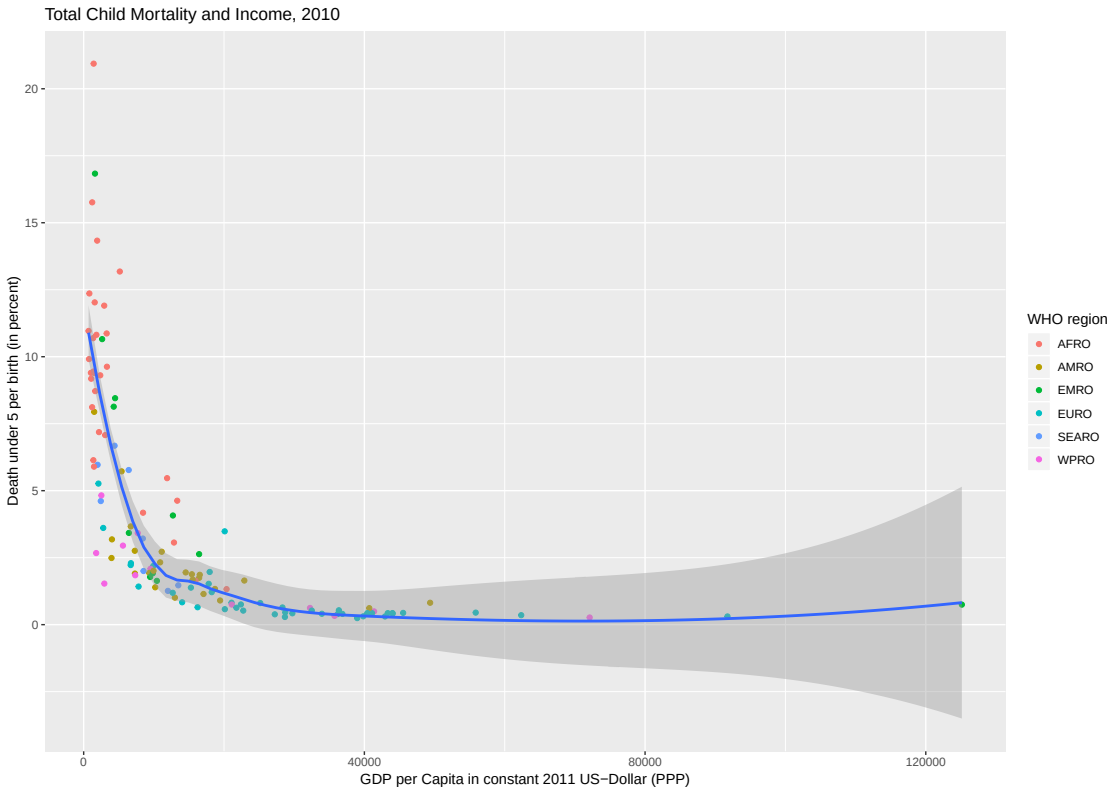

Figure 1: Correlation between GDP and child mortality across all countries

SPL_Load_data SPL_correlation

5%, whereat the under – 5 death in Africa lies between 6% up to over 20%. It can be seen that the title of the graph was changed from “final.df$region_who” to “WHO region”.

The next graph “d2” should depict the correlation of data in Europe. That’s why only the subset dataframe “final.df_Europe” is taken. Therefor inside of ggplot’s brackets in line 39 this data frame is put and after a comma the aes() is defined.

Again the same graphic type is defined by “geom_point()” but this time a specific color was chosen for the points on a diagram in line 41. As expected the dot plot is produced. The graph shows the realtion between the total child mortality and the income in Europe for the year 2010. A linear regression is drawn showing negative correlation between GDP per Capita in US - $ (PPP) and the percentage of death under 5 per birth. The straight linear line reaches the zero – point on the y – axis at x $\approx$ 46951 US - $ GDP per Capita in US - $ (PPP). It can be seen that for the poorest country in Europe the mortality

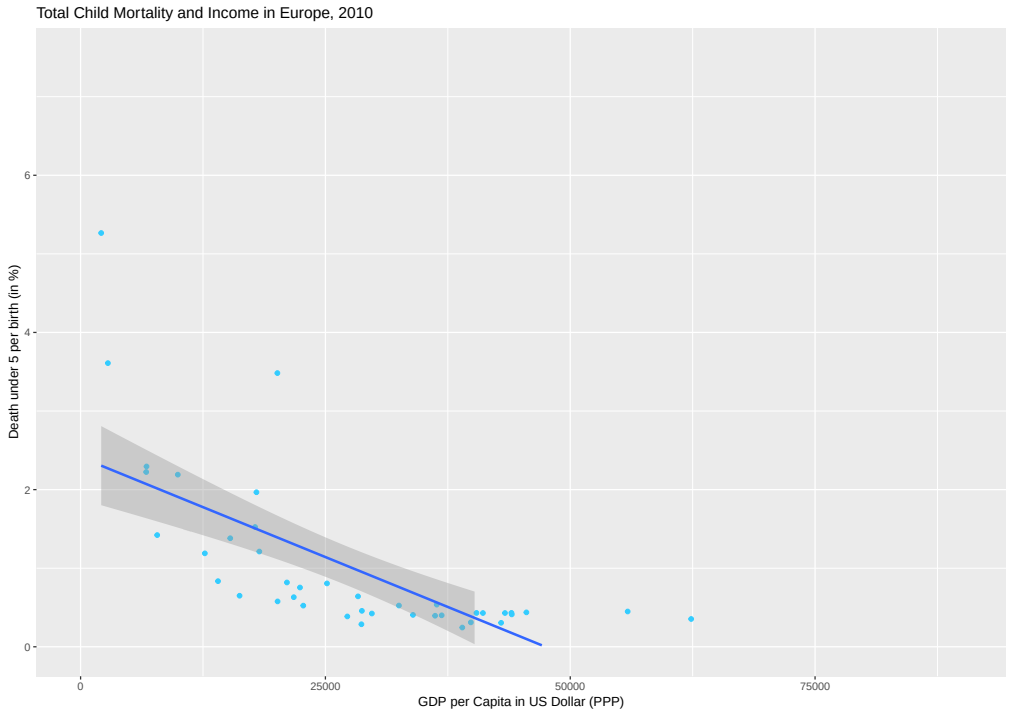

Figure 2: Correlation between GDP and child mortality for European countries

SPL_Load_data SPL_Correlation

rate before 5 can be higher than 5% falling to less than 1% as the GDP per Capita grows above 25000 US $. The correlation between GDP per Capita and child mortality is clearly visible in Europe.

The next diagram “d3” should be of the same type as the previous one showing the correlation between “GDP_PPP_2010” and “death_under_5_per_birth” in Africa as defined in “aes()” in line 57. In line 59, in “geom_point()” a red color is set with the code "#FF3300” to visualize observations for African countries. Moreover, different lengths of coordinates are set in lines 61 and 62 since the countries in Africa tend to be less well of (shorter x-Axis) and tend to have more deaths under 5 per birth (longer y-axis). The description of the labels of coordinates - "GDP per Capita in US Dollar (PPP)" and "Death under 5 per birth (in %)", and the title - "Total Child Mortality and Income in Africa, 2010" are added in “labs()” respectively, line 63. As expected very similar linear regression for the same x – and y – variables for Africa is drawn with the data from year

2010. In the poorest country for which the GDP per Capita does not exceed 2000 US $ the mortality rate can reach even more than 20% and reaching the value of approximately 10% for the countries with GDP per Capita of around 3000 US $. The mortality gradually falls with higher GDP being on its lowest level for the richest country whose GDP per capita exceeds 20000 US $.

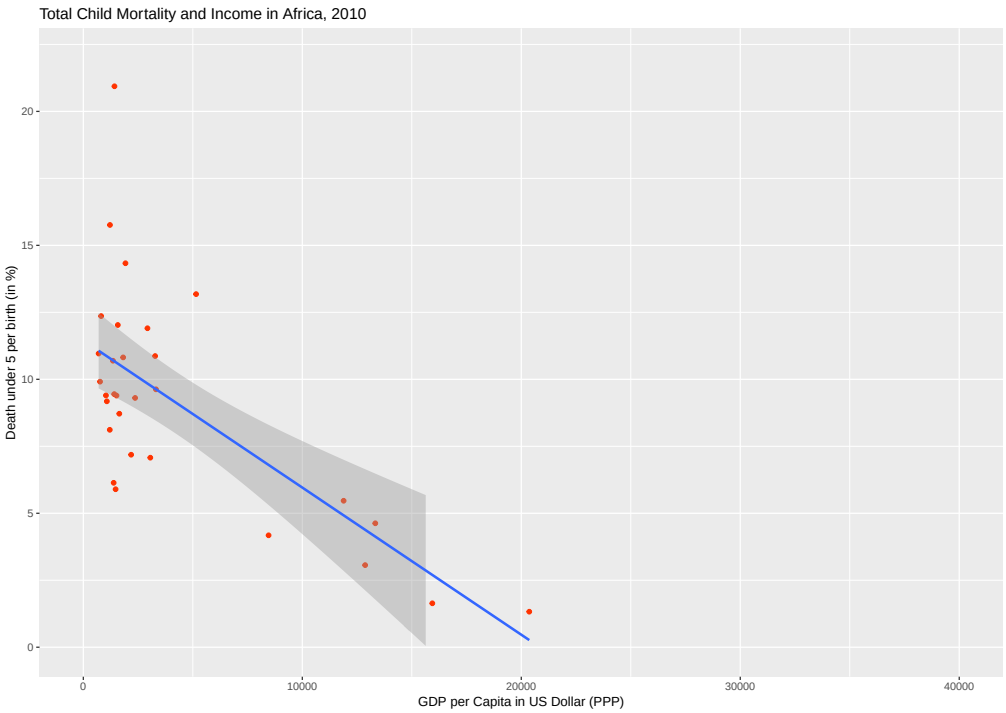

Figure 3: Correlation between GDP and child mortality for African countries

SPL__Load__data SPL__Correlation

For the purpose of juxtaposition of those both correlations, in line 76, inside of brackets of “grid.arrange()” function graphs “d2” and “d3” are defined alongside with the number of columns with the graphs. In this case there is only one row and two columns with one diagram in every of them. The starting point is with 10,9% death under 5 per birth significantly higher than in Europe with 2,1% at an GDP of 2105 US - $. However, the percentage of 2,1% death under 5 per birth is reached in Africa according to the linear regression at an GDP of 16333 US - $ per Capita (PPP). The difference in GDP per Capita (PPP) and in mortality rates between two regions is clearly visible.

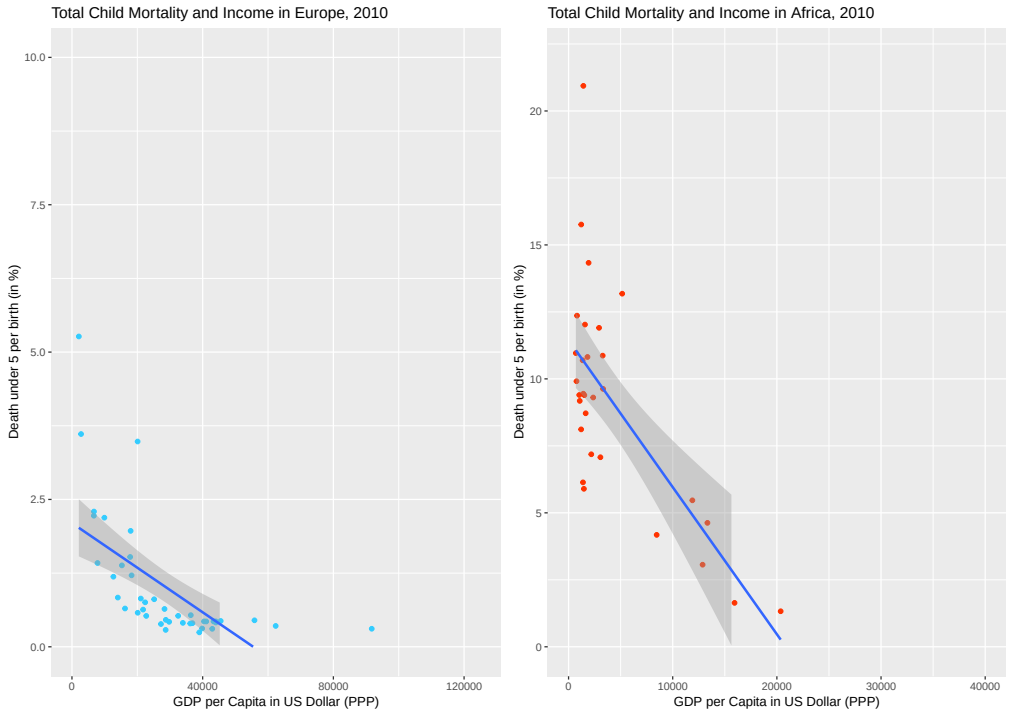

Figure 4: Correlation between GDP and child mortality for Africa and Europe next to each other

SPL_Load_data SPL_Correlation

3. Quantlet: SPL_Distribution

In order to show the distribution of GDP per Capita in Europe and in Africa the kernel density diagrams are plotted. The first one "density1" should show the distribution of data for Europe. In line 3 in "ggplot()" the data frame "final.df_Europe" is set to depict the data for European countries. In "aes()" the distribution of "GDP_PPP_2010" is chosen. In line 5 inside of "as.factor()" function the vector "region_who" is set as a factor. If alpha is 0.3 it means that the transparency is 0.3 of its maximum 1. Again the labels and the title are modified within "labs()" . In "density2" similarly to "density1" the data frame "final.df_Africa" is chosen and within "aes()" in line 15 the distribution of "GDP_PPP_2010". As a factor the WHO region is chosen, in this case African countries, line 16. The graph is of yellow color and with transparency 0.3 set inside of "geom_density()", line 17. On both of the plots in line 23 the "grid.arrange()" function

is used to juxtapose them next to each other. In output it can be clearly seen that the biggest amount of countries in Europe has the GDP per Capita (PPP) of almost 25000 US $ whereas for African countries it is 2000 US $. Moreover, there are some outliers in Africa with GDP per capita (PPP) higher than 10000 US $.

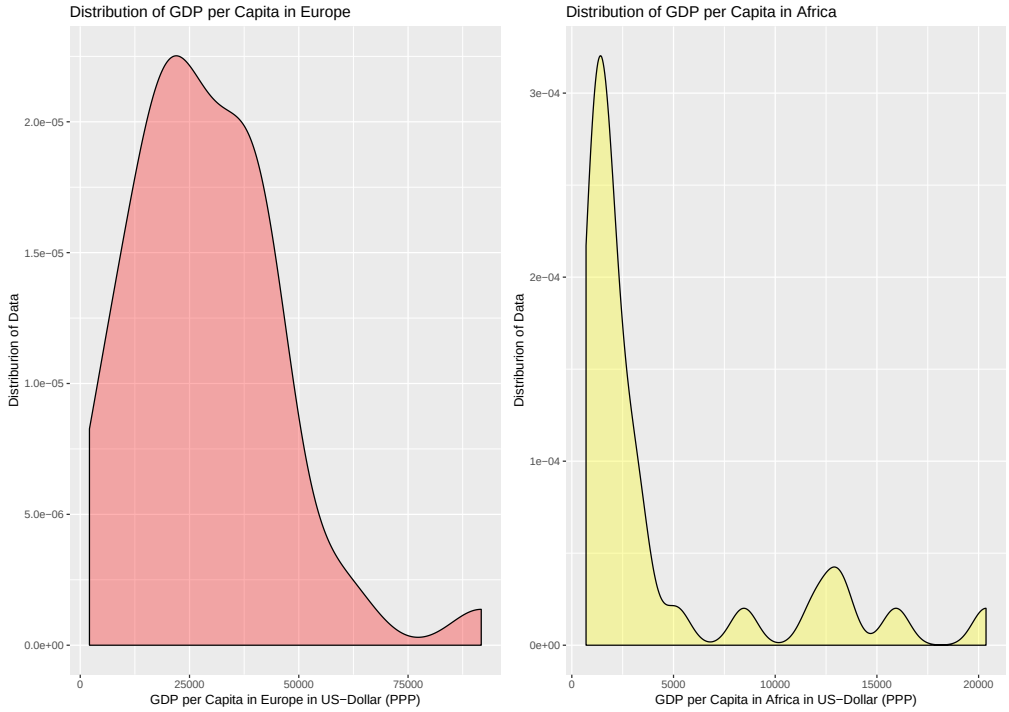

Figure 5: Density distribution of GDP per ca

SPL_Load_data SPL_Distribution

In order to create the boxplot in line 29 showing the income distribution for every WHO region the data frame for all countries “final.df” is set on the first place inside of “ggplot()”.

If the legend is unnecessary, it should be removed. In this case it is indeed unnecessary because the names of the regions are on the y-axis. As expected all of the box plots are shown on one graph.

The boxplot for the WPRO region (Western Pacific region) is the boxplot with the widest interquartile range. The median is closer to the 25% - quartile than to the 75% - quartile and is with x(quer) $\approx$ 7830 US - $ GDP per Capita (PPP) similar to the medians of the SEARO region (South Asia region) and the EMRO region (Eastern Mediterranean).

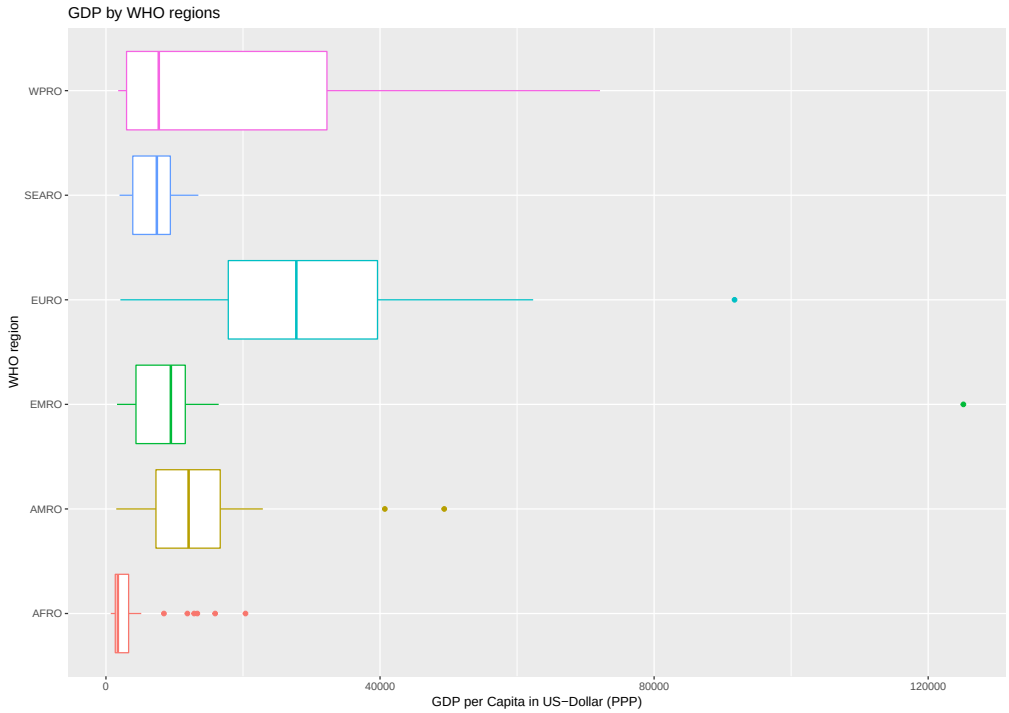

Figure 6: Distribution of GDP per capita in every WHO region

SPL_Load_data SPL_Probability

The whisker towards the maximum on the x – axis is very long and even longer than the boxplot itself. The minimum is approximately 1740 US - $ and the maximum is the biggest one of all WHO – regions in the graph excluding the maximums counting as outliner with approximately 71340 US - $ GDP per Capita (PPP).

The boxplot for the SEARO region is after the boxplot for the AFRO region (Africa) the second smallest in the diagram. Its median is closer to the 75% - quartile and its interquartile range is one of the smallest ranges shown. The median is about

x(quer) ≈ 6960 US - $, the minumum x(min) ≈ 1740 US - $, the same as for the WPRO region, and x (max) ≈ 13050 US - $ GDP per Capita (PPP).

The boxplot of the EURO region (Europe) is after the boxplot of the WPRO region the second biggest one of all of the graphs. Its median with approximately 27840 US - $ is only minimally closer to the 25% - quartile than to the 75% - quartile and is the highest

one of all medians of the different regions shown in the plot. Its whiskers towards the minimum and maximum are quite long and x(min) ≈ 17450 US - $ and

x(max) ≈ 61770 US - $. There is one outlier at x ≈ 91350 US - $ GDP per Capita (PPP).

The boxplot of the EMRO region is of medium size. The minimum is approximately 1740 US - $, the median, that is closer to the 75% - quartile is x(med) ≈ 9570 US - $ and the maximum is at x(max) ≈ 16530 US - $ GDP per Capita (PPP).

The AMRO plot (two Americas) has its median almost in the middle of the boxplot

with x ≈ 1749 US - $. X(min) is approximately 1740 US - $ and x(max) ≈ 22679 US - $. There are two outliers at 4046 US - $ GDP per Capita (PPP).

The boxplot of the AFRO region (Africa) is the smallest one out of the six boxplots shown. The minimum is approximately 870 US - $ and close to the median of x(med) ≈ 1740 US - $. The maximum is x(max) ≈ 5220 US - $ and there are various outliners between 8700 and 20880 US - $ GDP per Capita (PPP).

4. Quantlet: SPL_Probability

In line 8 a new column “death_probability” is added to the dataframe “final.df”. In line 11 the diagram showing the distribution of probability of dying before 5 in every WHO region is plotted and then the world map is plotted to show the outcome for every country. The graph “d5” is the boxplot. Also in line 11, inside of “ggplot()”brackets the data frame “final.df” is specified and inside “aes()” x is specified as WHO region and y as death probability. By adding the “color = final.df$region_who” in line 13 a different color is ascribed to every WHO region. With “geom_boxplot()” the type of the diagram is specified and inside of “labs()” the labels for coordinates and the title are described. In line 19 legend is set to “none” with “theme(legend.position = “none”)”.

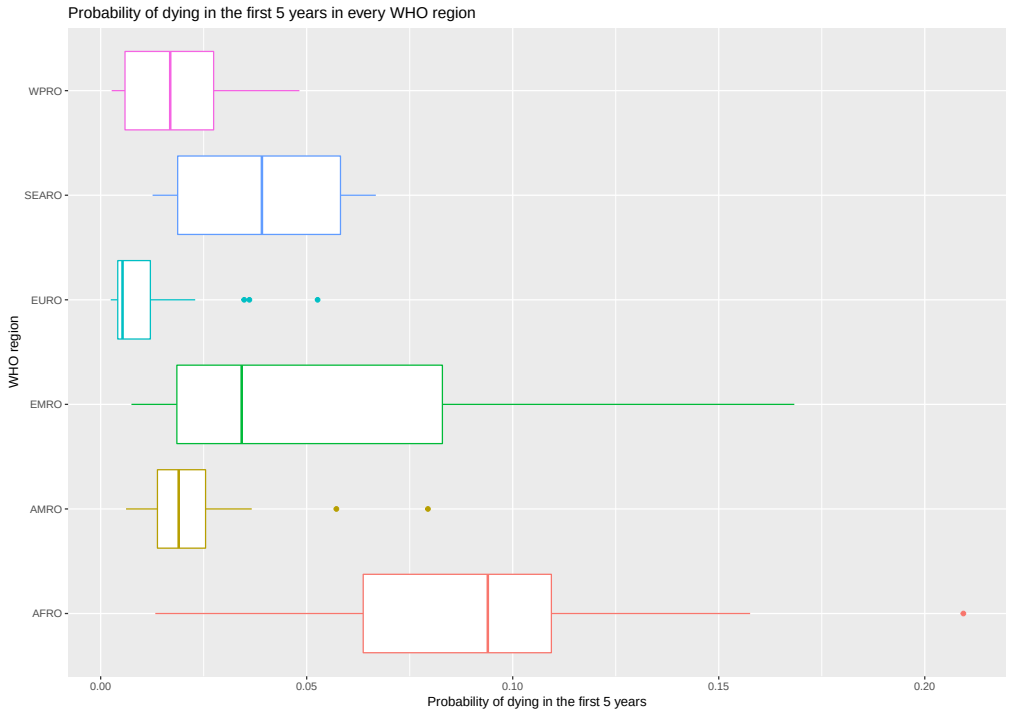

Figure 7: Distribution of probability of dying before 5 in every WHO region

SPL_Load_data SPL_Probability

Having a look at the different boxplots of the six different WHO regions about the probability of a child dying in the first five years per region a few observations become obvious. The highest probability for the burden of death within the first 60 months after birth have children in Africa with a median probability of approximately 9,6%. The second highest probability has the SEARO region with a median probability of $\approx$ 4%, the third highest has the EMRO region with $\approx$ 3,4%, followed by the AMRO region with a median probability of 2%. The second lowest median probability for death in the first 5 years has the WPRO region with $\approx$ 1,8% and the highest chances to survive within the first 5 years have children of the EURO region with a median probability of $\approx$ 0,44% of dying.

Willing to depict the probability of death for every country in the world, the map plot has to be used. In line 29 inside of the "joinCountryData2Map()" brackets the data frame "final.df" is specified. As the name of the column containing the country referencing the "Country.Code"is chosen. In order to depict the probability for every country inside of

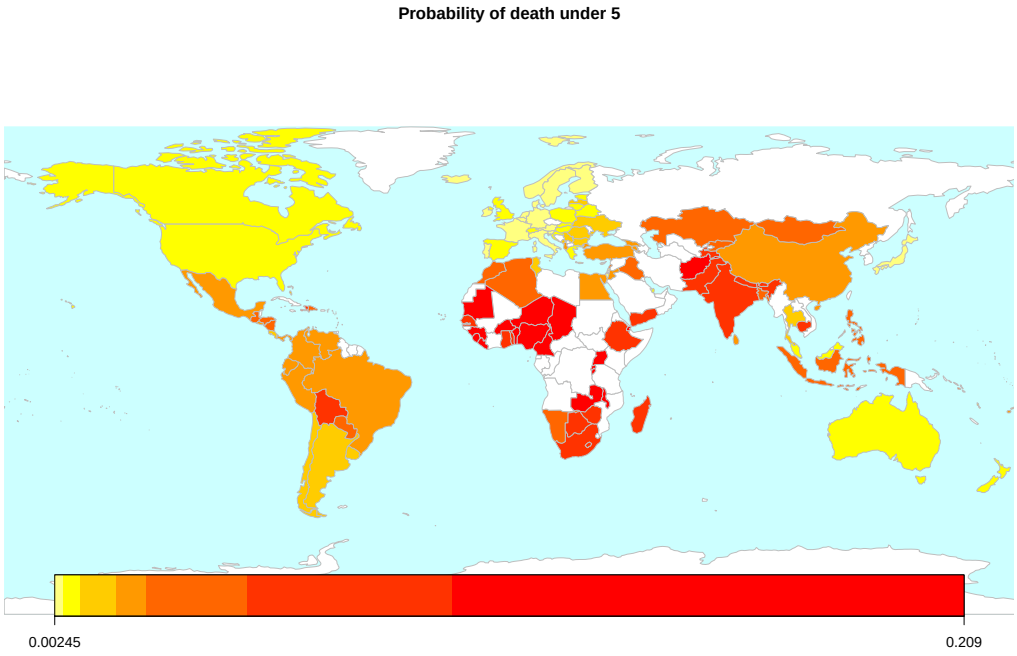

Figure 8: Map whith probability of dying before the age of 5

SPL_Load_data SPL_Probability

"mapCountryData()" function, the code in line 34 has to be performed "nameColumnToPlot = death_probability".

In the next line the title is set and in the next two lines "#FFFFFF" stands for white color and "#CCFFFF" stands for light blue color which are added to the map as the "missingCountryCol" and "oceanCol" respectively.

It is clearly visible that the highest probability of dying before 5 can be found in African countries and come Asian. In AFRO region Tunisia and Egypt stand out as the countries with light orange color meaning that the probability is relatively small there.

5. Quantlet: SPL_Tests

In lines 31, 36 and 41 the test is carried out three times. The first time for correlation between "GDP_PPP_2010" and "death_under_5_per_birth" in the whole world, where the data is taken from the data frame. For the second time between the same variables

but only in Europe and for the third time also for the same variables but only in Africa. In every test the alternative hypothesis comes out to be true what means that the correlation is not equal to 0. Running the test one can also read the value to the T-test, number of degrees of freedom, the p-value, 95percent confidence interval and the correlation. In every case the correlation is negative what implies that when one variable increases the other decreases, and vice versa. It confirms the statement that when the GDP per capita increases, the child mortality rate decreases on average.

In order to check whether in our correlation models the heteroskedasticity occurs the tests on heteroskedasticity are conducted in lines 59, 63 and 67.

Out of three conducted tests, for two of them the H_0 is right and there is no heteroskedasticity. For the whole world and for Africa

while regressing "final.df$death_under_5_per_birth" on "final.df$GDP_PPP_2010" and

"final.df_Africa$death_under_5_per_birth" on "final.df_Africa$GDP_PPP_2010" there is no heteroskedasticity problem because the p-values of the test are greater than 0,05 and we reject the alternative hypothesis. Only for Europe while regressing "final.df_Europe$death_under_5_per_birth" on "final.df_Europe$GDP_PPP_2010" the p-value is smaller than 0,05 and the H0 has to be rejected and the alternative hypothesis "variance increases from segment 1 to 2" accepted.

6. Quantlet: SPL_Regression

The model described in the implementation section is now applied to the dataset of variables measured in 2010. Following our hypothesis, we would expect a higher GDP to decrease the child mortality rate. Moreover, we would expect a higher child mortality, where income is less evenly distributed. We believe an unevenly distributed income to cause more poverty in society. In countries with an unevenly distributed income, we would therefore expect the child mortality rate to be higher, as child death is mostly a

phenomenon of poverty. For the birthrate, we would expect it to increase the child mortality rate. Where there are more children born per women, parents have less resources to care for every individual newborn, thereby increasing the chance, that the child suffers from malnutrition or some related disease. For the region dummies, we would expect dummies for low income regions to be associated with a higher child mortality rate, than regions with a high income. Table 1 represents the regression results. For all four specifications, GDP is negatively affecting the child mortality rate as the effect is significant at the 1% level. This confirms our main hypothesis, that a higher income decreases the child mortality rate. The size of the effect differs, when control variables are added. In column 2, when region dummies are not added, the coefficient has a negative sign. A higher gini coefficient is reporting a less unequally distributed income. Therefore, the effect of a more unequal society is a decreased child mortality rate. This leaves us with a puzzle. When regional dummies are added in column 4, the effect vanishes. The effect of the birthrate is as expected. A higher birthrate does result in a higher child mortality rate. The regional dummies show ambiguous effects and only an example will be explained here. For countries in the WHO region AFRO which is made of sub-Saharan African countries, the coefficient is positive, signaling that the child mortality is higher for African countries, than for the rest of the sample, even if the GDP is already controlled for.

To transform the regression results into a nice LaTeX table, we use the stargazer package (Halavac 2018).

6 Conclusion

The hypotheses and research question of our paper is that the higher the GDP in a country/ region of the world the lower the under – 5 child mortality rate.

In the eLibrary of the World Bank Group a paper called ‘Child Mortality and Public

Table 1: Child Mortality and Income

	Dependent variable:			
	Children dead under 5(in percent)			
	(1)	(2)	(3)	(4)
log(GDP in 2010)	−2.802***	−0.944***	−1.994***	−0.997***
	(0.180)	(0.284)	(0.213)	(0.286)
Gini Coefficient in 2010		−0.048**		−0.045
		(0.024)		(0.030)
Birthrate in 2010		0.258***		0.193***
		(0.032)		(0.039)
WHO Region AFRO			4.354***	2.759***
			(0.784)	(0.788)
WHO Region AMRO			0.527	0.828
			(0.744)	(0.702)
WHO Region EMRO			3.123***	2.085**
			(0.880)	(0.826)
WHO Region EURO			0.681	0.848
			(0.707)	(0.691)
WHO Region SEARO			0.847	1.619*
			(0.969)	(0.898)
WHO Region WPRO				
Constant	29.423***	8.700**	20.359***	9.082**
	(1.662)	(3.429)	(2.072)	(3.601)
Observations	127	127	127	127
R^2	0.659	0.783	0.762	0.806
Adjusted R^2	0.657	0.777	0.750	0.793

NOTE: *p<0.1; **p<0.05; ***p<0.01

tandard errors in paranthesis.

Data source as described in "1. Quantlet: SPL_Load_data.

Spending on Health: How Much Does Money Matter?' by Deon Filmer and Lant Pritchett, 1997, can be found. In this paper is stated: „Roughly 95 percent of cross-national variation in child or infant mortality can be explained by a country's per capita income [...]'. The latter paper is not the only source we read that let us suspect that there probably will be a correlation between GDP and the under – 5 child mortality rate in the different regions/countries in the world.

Due to our work we are able to say that the lower the GDP in every WHO region the higher the percentage of death under 5 per birth even though the percentage of death under 5 in the various regions can be differently at the same GDP.

In Europe there is a clearly visible negative correlation between GDP per Capita and the percentage of death under 5 per birth.

Even though the dimensions and averages of the GDP per Capita and the number of children born and dying under 5 are different to Europe, the same conclusion and correlation can be drawn for Africa. The mortality gradually falls with higher GDP being on its lowest level for the richest country whose GDP exceeds 20000 US - $.

In output it can be clearly seen that the biggest amount of countries in Europe has the GDP per Capita (PPP) of almost 25000 US $ whereas for African countries it is 2000 US $, but there are also some outliners.

In every case the correlation is negative what implies that when one variable increases the other decreases, and vice versa. It confirms the statement that when the GDP per capita increases, the child mortality rate decreases on average.

The highest probability for the burden of death within the first 60 months after birth have children in Africa. In the AFRO regions Tunisia and Egypt the chances of surviving within the first 60 months are clearly higher than in the other AFRO states.

The second highest probability has the SEARO region, the third highest has the EMRO region, followed by the AMRO region. The second lowest median probability for death in the first 5 years has the WPRO and the highest chances to survive within the first 5 years have children of the EURO region.

The test for heteroskedasticity shows that there is none for the whole world nor for Africa, but there is heteroskedasticity for Europe.

Facing potential problems that might occur in our work we can say on the one hand side that while using the fifth function that checks for normal distribution we cut out 5 tests due to the result 'infinity'. Moreover we could have used the Gini coefficient to check further for correlation. Furthermore we could have installed the Shiny package to optimize our results by tuning our performance.

7 Appendix: Complete R-Code

1. Quantlet: SPL_Load_data

```
# Set the working directory

setwd("C:/Users/Privat/SPL/Code/R_Working")

#The stargazer package is needed to produce nice LaTex tables

if (any(grepl("stargazer", installed.packages())) == FALSE){
  install.packages("stargazer")
  }

library(stargazer)

#List off all WHO regions
```

```
14
15 who.regions = final.df$region_who
16 who.regions = unique(who.regions)
17
18 #Create regional dummies for all six WHO regions
19
20 final.df$region_dummy_EMRO = ifelse(final.df$region_who == "EMRO",1,0)
21 final.df$region_dummy_EURO = ifelse(final.df$region_who == "EURO",1,0)
22 final.df$region_dummy_AFRO = ifelse(final.df$region_who == "AFRO",1,0)
23 final.df$region_dummy_AMRO = ifelse(final.df$region_who == "AMRO",1,0)
24 final.df$region_dummy_WPRO = ifelse(final.df$region_who == "WPRO",1,0)
25 final.df$region_dummy_SEARO = ifelse(final.df$region_who == "SEARO"
     ,1,0)
26
27 #Generate natural log of GDP with sapply
28
29 final.df$ln_GDP_PPP_2010 = sapply(final.df$GDP_PPP_2010, log)
30
31 #Estimate linear model with logaritmic child mortality as the
     dependend variable
32
33 lreg1 = lm(final.df$death_under_5_per_birth~final.df$ln_GDP_PPP_2010)
34
35 #Add the gini coefficient and the birthrate as control
36 #variables
37
38 lreg2 = lm(final.df$death_under_5_per_birth
39            ~final.df$ln_GDP_PPP_2010
40            +final.df$gini_disp
41            +final.df$birthrate_2010)
42
43 #Control for region dummies
44
45 lreg3 = lm(final.df$death_under_5_per_birth
```

```
            ~final.df$ln_GDP_PPP_2010
            +final.df$region_dummy_AFRO
            +final.df$region_dummy_AMRO
            +final.df$region_dummy_EMRO
            +final.df$region_dummy_EURO
            +final.df$region_dummy_SEARO
            +final.df$region_dummy_WPRO)

#Controlled for the gini coefficient and the birthrate,
#As well as for the region dummies

lreg4 = lm(final.df$death_under_5_per_birth
             ~final.df$ln_GDP_PPP_2010
             +final.df$gini_disp
             +final.df$birthrate_2010
             +final.df$region_dummy_AFRO
             +final.df$region_dummy_AMRO
             +final.df$region_dummy_EMRO
             +final.df$region_dummy_EURO
             +final.df$region_dummy_SEARO
             +final.df$region_dummy_WPRO)

#Create Latex table

stargazer(lreg1, lreg2, lreg3, lreg4,
           title="Child Mortality and Income",
           out = "C:/Users/Privat/SPL/Code/R_Working/reg_table.tex",
           out.header = TRUE,
           align=TRUE, dep.var.labels=c("Children dead under 5(in
             percent)"),
           covariate.labels=c("log(GDP in 2010)",
                              "Gini Coefficient in 2010",
                              "Birthrate in 2010",
                              "WHO Region AFRO",
```

```
                                "WHO Region AMRO",
                                "WHO Region EMRO",
                                "WHO Region EURO",
                                "WHO Region SEARO",
                                "WHO Region WPRO"),
          omit.stat=c("LL","ser","f"), no.space=TRUE)
```

2. Quantlet: SPL_Correlation

```
setwd("/Users/grzegorzantkiewicz/Documents/R_Working")



#Install the package ggplot2


if (any(grepl("ggplot2", installed.packages())) == FALSE){
  install.packages("ggplot2")
}

library("ggplot2")

library("RColorBrewer")

final.df_Africa = subset(final.df, region_who == "AFRO")
final.df_Europe = subset(final.df, region_who == "EURO")


#Correlation between GDP per Capita and Death
#under 5 per birth worldwide and moving average
d1 = ggplot(final.df,
            aes(x=final.df$GDP_PPP_2010,
                y=final.df$death_under_5_per_birth)) +
  geom_point(aes(color=final.df$region_who))  +
  geom_smooth() +
```

```
  labs(x="GDP per Capita in constant 2011 US-Dollar (PPP)",
       y ="Death under 5 per birth (in percent)",
       title = "Total Child Mortality and Income, 2010",
       col = "WHO region")
?ggplot



d1

#Correlation between GDP per Capita and Death
#under 5 per birth in Europe and Africa


d2 = ggplot(final.df_Europe, aes(x=final.df_Europe$GDP_PPP_2010,
                                 y=final.
                                   df_Europe$death_under_5_per_birth))
                                   +
  geom_point(color="#33CCFF") +
  theme(plot.background = element_blank()) +
  ylim(0, 7.5) +
  xlim(0, 125000)+
  geom_smooth(method=lm) +
  labs( x="GDP per Capita in US Dollar (PPP)",
        y ="Death under 5 per birth (in %)",
        title = "Total Child Mortality and Income in Europe, 2010")
d2
?xlim

d2ver2 = d2 +
  ylim(0, 7.5) +
  xlim(0, 90000)
d2ver2

d3 = ggplot(final.df_Africa, aes(x=final.df_Africa$GDP_PPP_2010,
```

```
                                    y=final.
                                      df_Africa$death_under_5_per_birth))
                                      +
  geom_point(color="#FF3300") +
  geom_smooth(method=lm) +
  ylim(0, 22) +
  xlim(0, 40000) +
  labs(x="GDP per Capita in US Dollar (PPP)",
       y ="Death under 5 per birth (in %)",
       title = "Total Child Mortality and Income in Africa, 2010")

d3


#Arranging 2 graphs on the same page
if (any(grepl("gridExtra", installed.packages())) == FALSE){
  install.packages("gridExtra")
}

library("gridExtra")
grid.arrange(d2, d3, ncol = 2)
```

3. Quantlet: SPL_Distribution

```
setwd("/Users/grzegorzantkiewicz/Documents/R_Working")

density1 = ggplot(final.df_Europe,
                  aes(GDP_PPP_2010,
                  fill = as.factor(final.df_Europe$region_who)))+
  geom_density(fill = "red", alpha = .3) +
  labs(x = "GDP per Capita in Europe in US-Dollar (PPP)",
       y = "Distriburion of Data",
       title = "Distribution of GDP per Capita in Europe")
density1
```

```
11
12 region = as.factor(final.df_Africa$region_who)
13
14 density2 = ggplot(final.df_Africa,
15                   aes(GDP_PPP_2010,
16                       fill = as.factor(final.df_Africa$region_who))) +
17   geom_density(fill = "yellow", alpha = .3) +
18   labs(x = "GDP per Capita in Africa in US-Dollar (PPP)",
19        y = "Distriburion of Data",
20        title = "Distribution of GDP per Capita in Africa")
21
22 density2
23 grid.arrange(density1, density2 , ncol = 2)
24
25 #Box-plot
26
27
28
29 boxp1 = ggplot(final.df, aes(x = final.df$region_who,
30                              y = final.df$GDP_PPP_2010,
31                              color = final.df$region_who)) +
32   geom_boxplot() +
33   labs(y="GDP per Capita in US-Dollar (PPP)",
34        x = "WHO region" ,
35        title = "GDP by WHO regions") +
36   theme(legend.position = "none")+
37   coord_flip()
38
39 boxp1
```

4. Quantlet: SPL_Probability

```
1 #Set working directory
2 setwd("/Users/grzegorzantkiewicz/Documents/R_Working")
```

```
3
4
5  #Probability of dying before being 5
6
7  final.df$death_probability =
8    (final.df$death_under_5_total/final.df$birth_total)
9
10 d5 = ggplot(final.df, aes(x=final.df$region_who,
11                           y=final.df$death_probability,
12                           color = final.df$region_who)) +
13   geom_boxplot() +
14   labs(x = "WHO region",
15        y = "Probability of dying in the first 5 years",
16        title="Probability of dying in the first 5 years in every WHO
          region") +
17   coord_flip() +
18   theme(legend.position = "none")
19
20 d5
21
22 if (any(grepl("rworldmap", installed.packages())) == FALSE){
23   install.packages("rworldmap")
24 }
25 library("rworldmap")
26
27
28 mapped_data <- joinCountryData2Map(final.df, joinCode = "ISO3",
29                                    nameJoinColumn = "Country.Code")
30
31 par(mai=c(0,0,0.2,0),xaxs="i",yaxs="i")
32 mapCountryData(mapped_data,
33                nameColumnToPlot = "death_probability",
34                mapTitle = "Probability of death under 5",
35                missingCountryCol = "#FFFFFF",
```

```
oceanCol = "#CCFFFF")
```

5. Quantlet: SPL_Tests

```
#Set working directory
setwd("/Users/grzegorzantkiewicz/Documents/R_Working")


#Function testing the normal distribution of the data

if (any(grepl("tseries", installed.packages())) == FALSE){
  install.packages("tseries")
  YES
}
library("tseries")


nor = function(x) {

  jarque.bera.test(complete.cases(x))

}

#Testing the normality of data for the world, Europe and Africa

nor(final.df$death_under_5_per_birth)


#Function testing the significance of correlation
cor = function(x,y){
  cor.test(x, y)
}
#Test on significance of correlation in all countries

```

```r
cor(final.df$GDP_PPP_2010,
    final.df$death_under_5_per_birth)

#Test on significance of correlation in Europe

cor(final.df_Europe$GDP_PPP_2010,
    final.df_Europe$death_under_5_per_birth)

#Test on significance of correlation in Africa

cor(final.df_Africa$GDP_PPP_2010,
    final.df_Africa$death_under_5_per_birth)


#Function testing the heteroskedasticity
if (any(grepl("lmtest", installed.packages())) == FALSE){
  install.packages("lmtest")
}
library("lmtest")

het = function(x, y) {
  modell = lm(y~x)
  gqtest(modell)

}

#Test on heteroskedasticity in the world

het(final.df$GDP_PPP_2010, final.df$death_under_5_per_birth)

#Test on heteroskedasticity in Europe

het(final.df_Europe$GDP_PPP_2010, final.
  df_Europe$death_under_5_per_birth)
```

6. Quantlet: SPL_Regression

```
# Set the working directory

setwd("C:/Users/Privat/SPL/Code/R_Working")

#The stargazer package is needed to produce nice LaTex tables

if (any(grepl("stargazer", installed.packages())) == FALSE){
  install.packages("stargazer")
  }

library(stargazer)

#List off all WHO regions

who.regions = final.df$region_who
who.regions = unique(who.regions)

#Create regional dummies for all six WHO regions

final.df$region_dummy_EMRO = ifelse(final.df$region_who == "EMRO",1,0)
final.df$region_dummy_EURO = ifelse(final.df$region_who == "EURO",1,0)
final.df$region_dummy_AFRO = ifelse(final.df$region_who == "AFRO",1,0)
final.df$region_dummy_AMRO = ifelse(final.df$region_who == "AMRO",1,0)
final.df$region_dummy_WPRO = ifelse(final.df$region_who == "WPRO",1,0)
final.df$region_dummy_SEARO = ifelse(final.df$region_who == "SEARO"
  ,1,0)

#Generate natural log of GDP with sapply

final.df$ln_GDP_PPP_2010 = sapply(final.df$GDP_PPP_2010, log)

```

```
#Estimate linear model with logaritmic child mortality as the
  dependend variable

lreg1 = lm(final.df$death_under_5_per_birth~final.df$ln_GDP_PPP_2010)

#Add the gini coefficient and the birthrate as control
#variables

lreg2 = lm(final.df$death_under_5_per_birth
           ~final.df$ln_GDP_PPP_2010
           +final.df$gini_disp
           +final.df$birthrate_2010)

#Control for region dummies

lreg3 = lm(final.df$death_under_5_per_birth
           ~final.df$ln_GDP_PPP_2010
           +final.df$region_dummy_AFRO
           +final.df$region_dummy_AMRO
           +final.df$region_dummy_EMRO
           +final.df$region_dummy_EURO
           +final.df$region_dummy_SEARO
           +final.df$region_dummy_WPRO)

#Controlled for the gini coefficient and the birthrate,
#As well as for the region dummies

lreg4 = lm(final.df$death_under_5_per_birth
            ~final.df$ln_GDP_PPP_2010
            +final.df$gini_disp
            +final.df$birthrate_2010
            +final.df$region_dummy_AFRO
            +final.df$region_dummy_AMRO
            +final.df$region_dummy_EMRO
```

```
              +final.df$region_dummy_EURO
              +final.df$region_dummy_SEARO
              +final.df$region_dummy_WPRO)

#Create Latex table

stargazer(lreg1, lreg2, lreg3, lreg4,
          title="Child Mortality and Income",
          out = "C:/Users/Privat/SPL/Code/R_Working/reg_table.tex",
          out.header = TRUE,
          align=TRUE, dep.var.labels=c("Children dead under 5(in
            percent)"),
          covariate.labels=c("log(GDP in 2010)",
                             "Gini Coefficient in 2010",
                             "Birthrate in 2010",
                             "WHO Region AFRO",
                             "WHO Region AMRO",
                             "WHO Region EMRO",
                             "WHO Region EURO",
                             "WHO Region SEARO",
                             "WHO Region WPRO"),
omit.stat=c("LL","ser","f"), no.space=TRUE)
```

8 References

Arel-Bundock, V. (2018). countrycode: Convert Country Names and Country Codes. R-Package 1.00.0. https://github.com/vincentarelbundock/countrycode

Härdle, W.K., Klinke, S., Rönz, B. (2015). Introduction to Statistics. Springer International Publishing AG Switzerland.

Hlavac, M. (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

Liu, L., Johnson, H.L., Cousens, S., Perin, J., Scott, S., Lawn, J.E., Rudan, I., Campbell, H., Cibulskis, R., Li, M., Mathers, C., Black, R.E. (2012). Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. The Lancet, 379(9832), 2151 – 2161.

Liu, L., Oza, S., Hogan, D., Perin, J., Rudan, I., Lawn, J.E., Cousens, S., Mathers, C., Black, R.E. (2015). Global, regional, and national causes of child mortality in 2000-13, with projections to inform post-2015 priorities: an updated systematic analysis. The Lancet, 385(9966), 420.

Lawn, J.E., Kinney, M.V., Black, R.E., Pitt, C., Cousens, K., Kerber, K., Corbett, E., Moran, A.C., Morrissey, C.S., Oestergaard, M.Z. (2012). Newborn survival: a multi-country analysis of a decade of change. Health Policy and Planning, 27, iii6 – iii28.

Filmer, D., Pritchett, L. (1997). Child Mortality and Public Spending on Health: How Much Does Money Matter? World Bank., no. WPS 1864.

Solt, F. (2016). The Standardized World Income Inequality Database. Social Science Quarterly 97(5):1267-1281.

9 Declaration of Authorship

We hereby confirm that we have authored this Seminar paper independently and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

Berlin, 12.08.2018

Grzegorzthehero, K. M. , J. S.