STATISTICAL ANALYSIS TO UNDERSTAND LIFE EXPECTANCY

Abdallah Zaher, Cristina Mousinho, Gabriel Ravi, Nicolae-Radu

NOVA Information Management School and Universidade Nova de Lisboa

Introduction

Life expectancy is a statistical measure of the average time an organism is expected to live, based on the year of its birth, its current age and other demographic factors including gender. Life expectancy has been growing for the last 10 decades, but not at the same rate in all countries.

Objectives

This study is going to analyse 17 sustainable variables for 195 countries to understand which actions and cares can prolong the Life Expectancy in the world. The techniques that will be applied are **Stepwise Regression** and **Principal Components analysis**. This study contains:

- Descriptive analysis of all variables
- Analysis of the most important effects on Life Expectancy

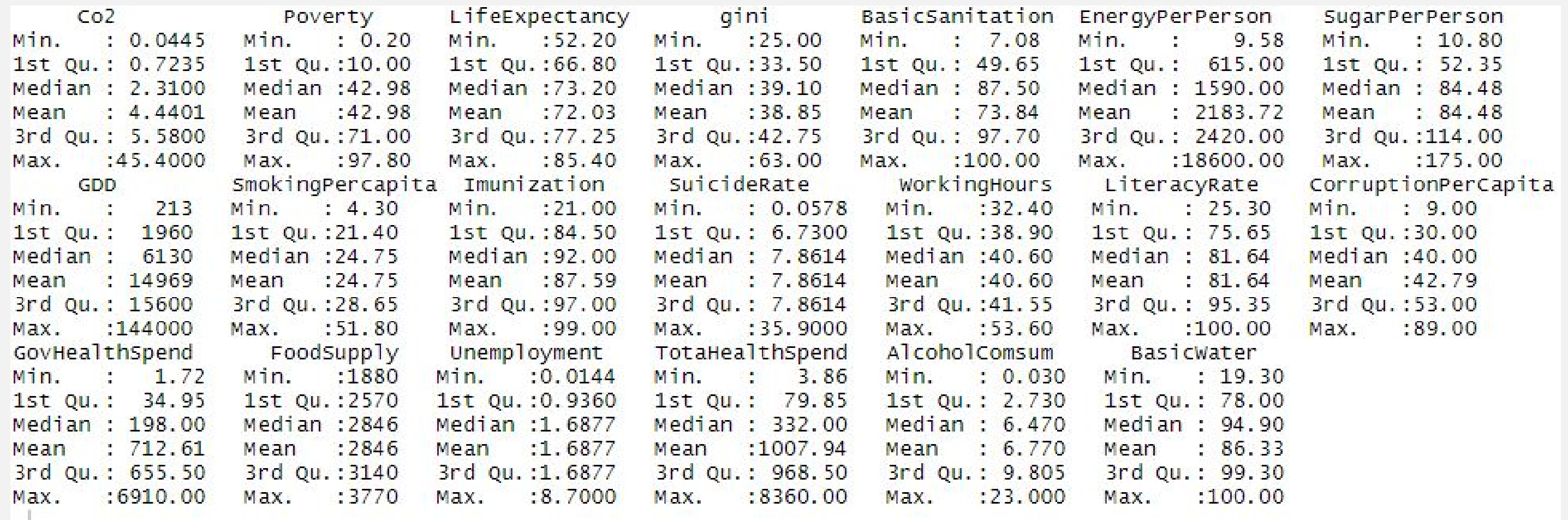
Data Exploration

The data was obtained from Open Access DataWorldBank: `https://data.worldbank.org/` and hosted on: `https://github.com/GRaviSantos79/StatsForDS`.

The MicroWorldBankdata Library is a collection of data sets from the World Bank and other international, regional and national organizations. This data set contains information about the most important indicators of sustainability of the world.

This analysis consists of a longitudinal collection of 20 indicators of sustainability about 195 countries and response variables, Life Expectancy.

The image below contains the descriptive analysis of all varaibles that may hav impact on Life Expectancy.



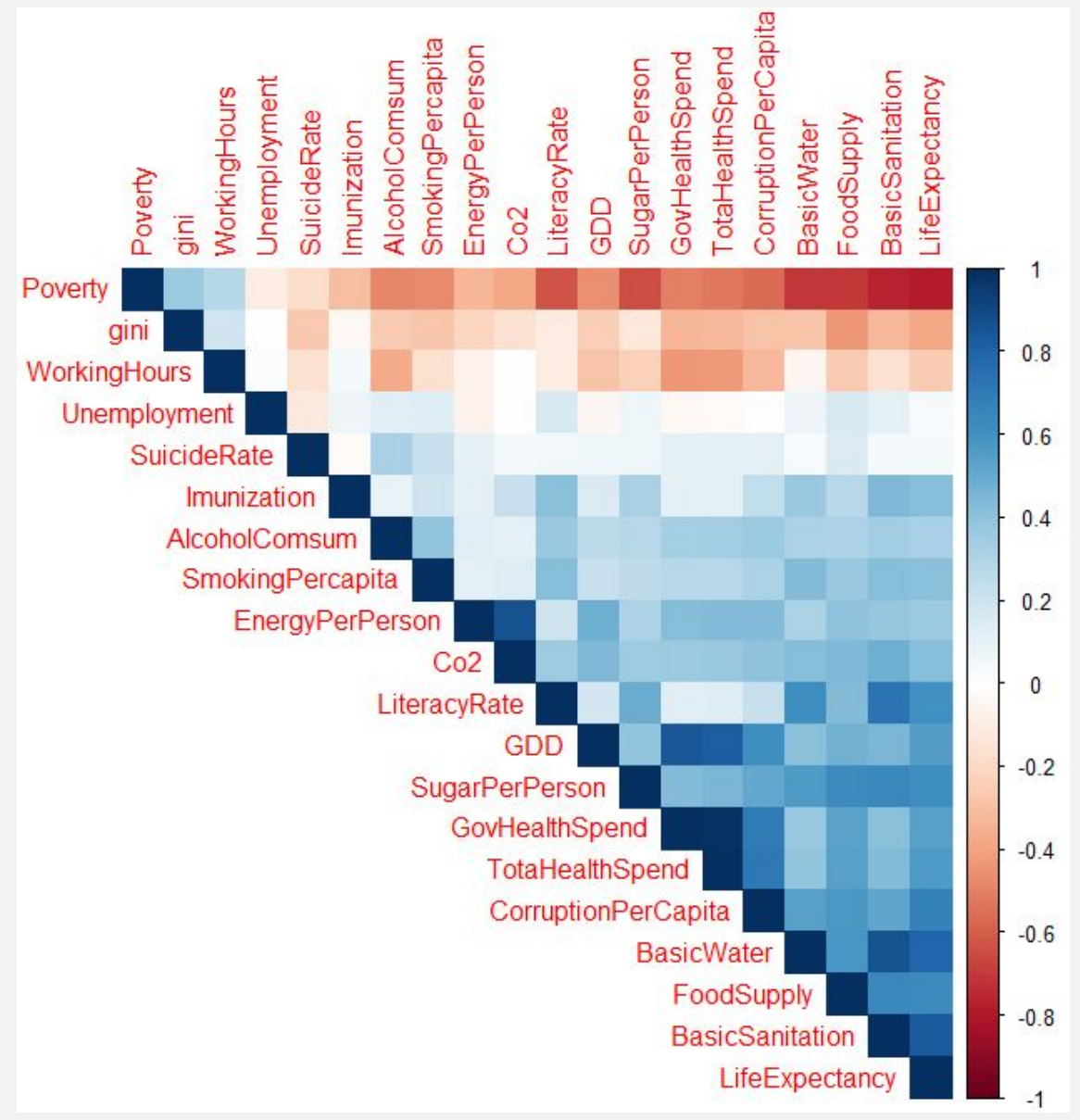
```
     Co2               Poverty         LifeExpectancy        gini         BasicSanitation  EnergyPerPerson    SugarPerPerson
Min.   : 0.0445   Min.    : 0.20   Min.    :52.20   Min.    :25.00   Min.    :  7.08   Min.    :    9.58   Min.    : 10.80
1st Qu.: 0.7235   1st Qu.:10.00   1st Qu.:66.80   1st Qu.:33.50   1st Qu.: 49.65   1st Qu.:  615.00   1st Qu.: 52.35
Median : 2.3100   Median :42.98   Median :73.20   Median :39.10   Median : 87.50   Median : 1590.00   Median : 84.48
Mean   : 4.4401   Mean    :42.98   Mean    :72.03   Mean    :38.85   Mean    : 73.84   Mean    : 2183.72   Mean    : 84.48
3rd Qu.: 5.5800   3rd Qu.:71.00   3rd Qu.:77.25   3rd Qu.:42.75   3rd Qu.: 97.70   3rd Qu.: 2420.00   3rd Qu.:114.00
Max.   :45.4000   Max.    :97.80   Max.    :85.40   Max.    :63.00   Max.    :100.00   Max.    :18600.00   Max.    :175.00
     GDD          SmokingPercapita  Imunization      SuicideRate        WorkingHours     LiteracyRate     CorruptionPerCapita
Min.   :   213   Min.    : 4.30   Min.    :21.00   Min.    : 0.0578   Min.    :32.40   Min.    : 25.30   Min.    : 9.00
1st Qu.:  1960   1st Qu.:21.40   1st Qu.:84.50   1st Qu.: 6.7300   1st Qu.:38.90   1st Qu.: 75.65   1st Qu.:30.00
Median :  6130   Median :24.75   Median :92.00   Median : 7.8614   Median :40.60   Median : 81.64   Median :40.00
Mean   : 14969   Mean    :24.75   Mean    :87.59   Mean    : 7.8614   Mean    :40.60   Mean    : 81.64   Mean    :42.79
3rd Qu.: 15600   3rd Qu.:28.65   3rd Qu.:97.00   3rd Qu.: 7.8614   3rd Qu.:41.55   3rd Qu.: 95.35   3rd Qu.:53.00
Max.   :144000   Max.    :51.80   Max.    :99.00   Max.    :35.9000   Max.    :53.60   Max.    :100.00   Max.    :89.00
GovHealthSpend      FoodSupply      Unemployment     TotaHealthSpend    AlcoholComsum      BasicWater
Min.   :   1.72   Min.    :1880   Min.    :0.0144   Min.    :   3.86   Min.    : 0.030   Min.    : 19.30
1st Qu.:  34.95   1st Qu.:2570   1st Qu.:0.9360   1st Qu.:  79.85   1st Qu.: 2.730   1st Qu.: 78.00
Median : 198.00   Median :2846   Median :1.6877   Median : 332.00   Median : 6.470   Median : 94.90
Mean   : 712.61   Mean    :2846   Mean    :1.6877   Mean    :1007.94   Mean    : 6.770   Mean    : 86.33
3rd Qu.: 655.50   3rd Qu.:3140   3rd Qu.:1.6877   3rd Qu.: 968.50   3rd Qu.: 9.805   3rd Qu.: 99.30
Max.   :6910.00   Max.    :3770   Max.    :8.7000   Max.    :8360.00   Max.    :23.000   Max.    :100.00
```

Data Exploration

Check the summary of the variables bellow:

BasicWater	% of Sugar/person
CO2	CO2 emissions
GINI	Dist. of whealth
BasicSanitation	% Sanitation Available
Energy	% of Energy/person
Sugar	% of Water/person
GDD	Growing Degree Day
Smoking	% pop that smokes
Imunization	% pop immunized
Suicide	Number of Suicides/pop
WorkingHours	Working-Hours/person
Literacy	Literacy/person
Corruption	% of corruption
GovHealth	Gov Spend with health
FoodSupply	% of Food/person
Unemployment	% of Unemployment
TotalHealth	Pop Spend with health
Alcohol	% of Alcohol/person

Here you can check and analyse a correlation graph between the variables.

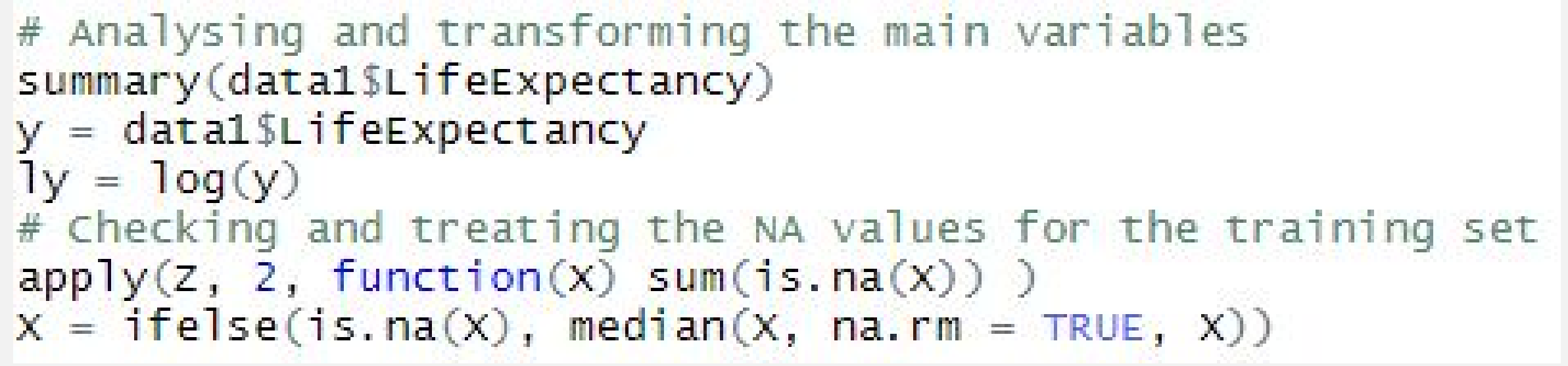


Basic sanitation, food supply, basic water, poverty, corruption and GINI (negative correlation) are the variables that contain a higher correlation with Life Expectancy.

We will analyse these variables on further analysis because they are good candidates for the final model.

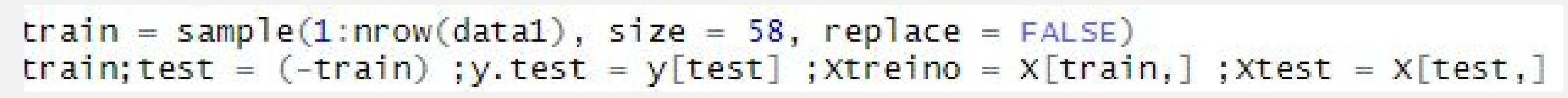
Methodology

Life Expectancy will be converted to a logarithm scale because there are non-linear relationships in the data set and the chances of producing errors may also be skewed negatively. There is 5% of missing data. If we delete the countries that have NaN values, the data set will lose its quality. The missing values were filled by the median, as it is the technique that doesn't interfere with the results.



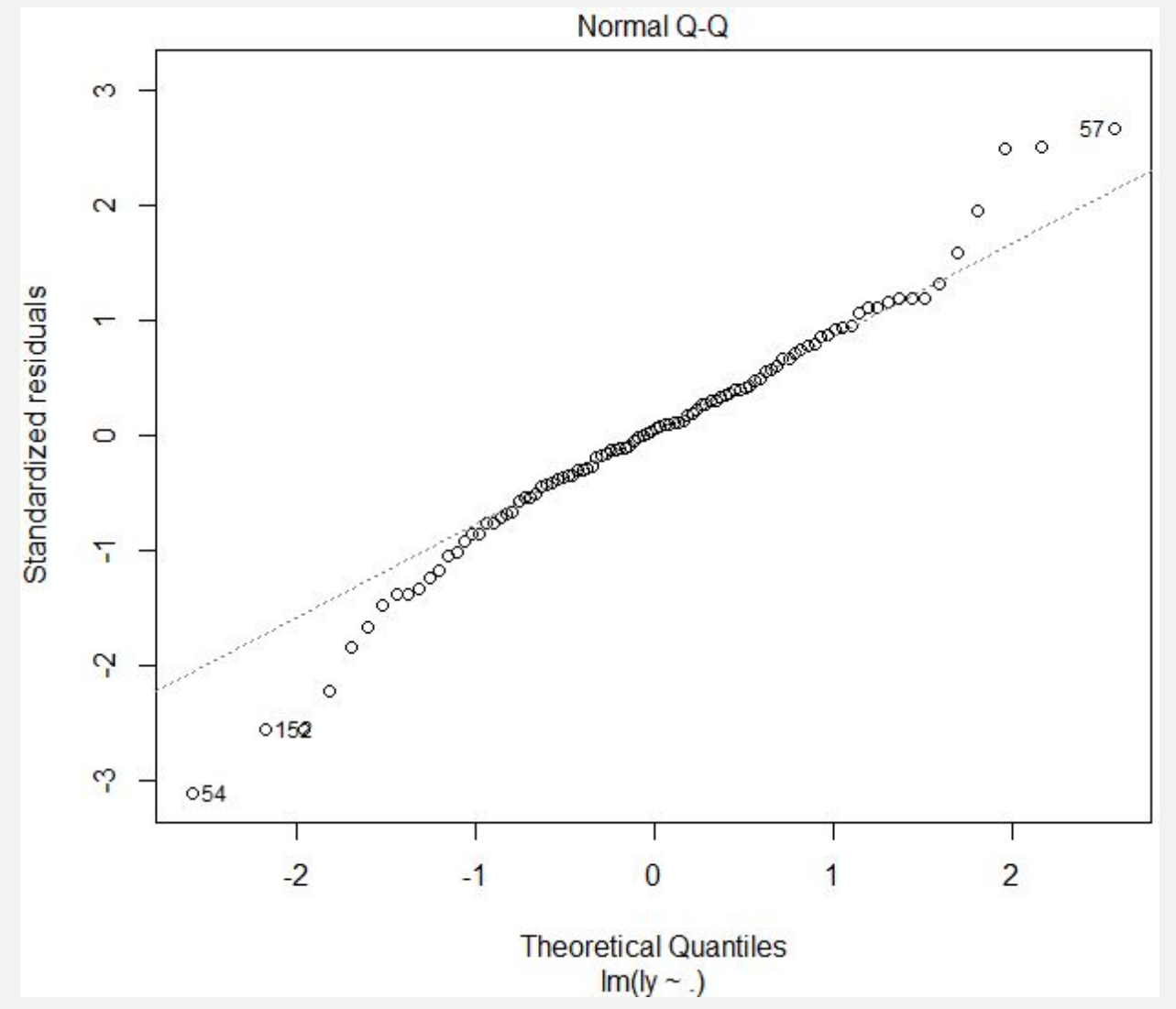
```
# Analysing and transforming the main variables
summary(data1$LifeExpectancy)
y = data1$LifeExpectancy
ly = log(y)
# Checking and treating the NA values for the training set
apply(z, 2, function(X) sum(is.na(X)) )
X = ifelse(is.na(X), median(X, na.rm = TRUE, X))
```

The training set size is equal to 30% of the population. We made sure that the training set contains heterogeneous observations.



```
train = sample(1:nrow(data1), size = 58, replace = FALSE)
train;test = (-train) ;y.test = y[test] ;Xtreino = X[train,] ;Xtest = X[test,]
```

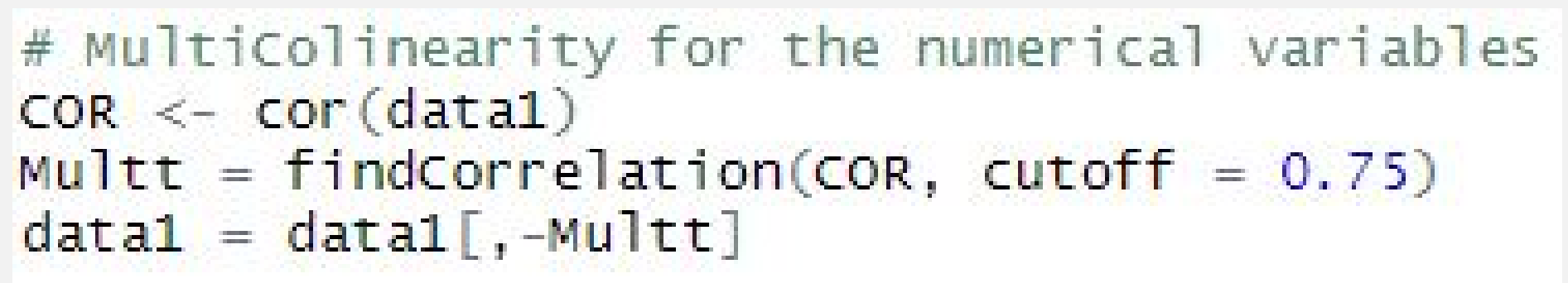
The next step is to check the normality on the training set.



When we analyse the standardized residuals in each theoretical quantile, we can see that there are values outside the qqline, so we can't assume normality on the data set.

The heteroscedasticity was tested using the Breusch Pagan Test, at 95% of significance level, we can say that the null hyphotesis is not rejected, in other words, we don't have a problem with heteroscedasticity.

It is also important to check if any predictors are correlated with each other to avoid multicolinearity and guarantee that the model will not suffer of instability.



```
# MultiColinearity for the numerical variables
COR <- cor(data1)
Multt = findCorrelation(COR, cutoff = 0.75)
data1 = data1[,-Multt]
```

Implementation

1. The **Stepwise Regressions** (Backward, Forward and both) will be applied.

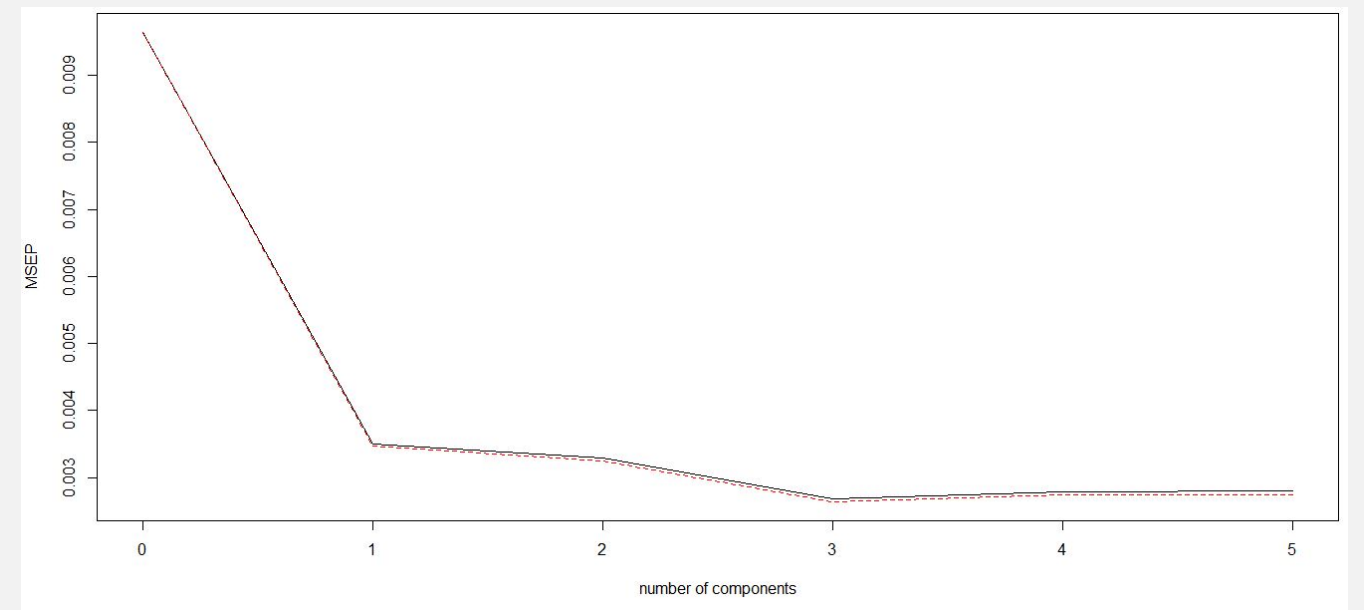
	R^2	N° Vars
Backward	0.82	4
Foward	0.81	4
Both	0.83	5

When we consider the R^2, we see that the Both stepwise regression gives better estimates. After the analysis, considering 95% of confidence, the main variables are Poverty, Basic Sanitation, Basic Water, GINI and Alcohol Consumption. And the final model is:

$$LifeExpec = 4.2 - 0.03Poverty + 0.04Sanitation$$

+0.03BasicWater - 0.02GINI -0.01 AlcoholComnsum (1)

2. **Principal Components Analysis**, we will use the Mean squared prediction error to determine how many components.



To performe Principal Component Analysis, we will use the mean squaredprediction error, in order to determine how many components. In this case we chose 2 principal components, who explain 72% of the total variance.

Conclusion

After the analysis, we can say that the most important variables to increase Life Expectancy are Poverty, BasicSanitation, Basic Water, GINI and alcoholConsum; in other words, these variables have the largest impact on life expectancy.

For example, in the final model(1), when a government increases basic water access by 3%, the Life Expectancy increases 1%.

References

- Bendel, R.B. and Afifi, A.A., 1977. Comparison of stopping rules in