University of Essex

Department of Mathematical Sciences

MA-317: Modelling Experimental Data

Analysis of the Determinants of Life Expectancy Across the World

Group G:

1900396

1908015

1901094

1900716

1901197

Professors: Dr Joseph Bailey and Dr Stella Hadjiantoni

Date of submission: 11 December 2019

Abstract

In this report we investigate the determinants of life expectancy in 2016 using data from the World Development Indicators (WDI). To this aim, a set of linear models are proposed in order to explore the main determinants of life expectancy across countries and an ANOVA analysis to find differences between geographic. We conduct an exploratory statistic analysis and highlight the importance of dealing with the most common issues present in data sets like the one analyzed here. Namely, missing values and collinearity.

Keywords: life expectancy, missing values, collinearity, linear regression, ANOVA

Contents

Word Count: 3,366

1. Introduction

In this report, the determinants of life expectancy are explored taking many country-level aggregated factors into consideration. The relevance of this analysis is that understanding can give an insight into many other well-being and degree of development indicators. In this way, public policies impact can be assessed.

To this extent, using World Bank's *World Development Indicators* data, we propose a set of linear models that can be used to explore the main determinants of life expectancy across countries and, at the same time, use them to predict next years' tendencies.

Along the report, we conduct exploratory statistic analysis and highlight the importance of dealing with the most common issues present in data sets like the one analyzed here. The rest of the report is organized as follows: in section 2, a *descriptive statistics analysis* of the data is conducted, giving special attention to the relation with the variable of interest. Next, several procedures to deal with*missing data* are deployed. Then, we test for *collinearity* between explanatory variables. In section 3.1, four different linear models are proposed to identify the determinants of the life expectancy and one-way ANOVA analysis across continents is conducted. Finally, section 4 concludes.

2. Preliminary Analysis

In this section we describe the data and provide descriptive statistics. Additionally, we propose three different methods to deal with missing data based on the distribution characteristics of each variable, and highlight the disadvantages of conducting a complete case analysis. Finally, we test for collinearity between some variables based on different categories.

2.1. Dataset and Descriptive Statistics

The dataset was obtained from the World Bank's *World Development Indicators* (WDI) compilation. This data comes from officially-recognized international sources. The dataset contains 22 variables, being life expectancy the variable of interest, with a total number of 217 observations corresponding to recognized countries for 2016. Additionally, there are 47 observations of regional and economic aggregations such as; European Union and OECD members.

Table 1 shows a description of all the variables. For life expectancy, there are 199 observations (i.e. 18 missing values), with a minimum of 51 years, a maximum of 84 years and a average life expectancy of 72.3 years with a standard deviation of 7 years. A definition for each variable can be seen in appendix A.

We can observe that there are some variables (for instance, unsafe water mortality, population growth, secondary education, health expenditure per capita, GDP per capita, and mobile subscriptions) have very high dispersion. This can represent an issue, as their scale and variability may affect the correlation magnitude with life expectancy. Therefore, they are log transformed for the rest of the analysis to make the relationship between the variables and life expectancy close to being symmetric. Also, it would help in straightening out the data and improve the regression model.

Table 1: Descriptive statistics

	Mean	Median	Standard Deviation	Min.	Max.	NA.s
Life Expectancy	72.30	73.84	7.78	51.59	84.23	18
Electricity	84.86	99.99	25.42	9.30	100.00	2
Adjusted Income	355.73	29.78	1450.69	0.16	15985.10	45
Children out of school	6.27	1.82	9.21	0.00	42.62	85
Primary education exp.	31.31	30.57	10.66	12.70	64.06	146
PPP	341.76	211.35	331.19	88.90	950.17	211
Unsafe water mortality	12.50	1.30	20.82	0.10	101.00	34
Adult literacy rate	90.21	94.65	14.70	22.31	99.99	183
Population growth	1.29	1.14	1.22	-3.07	4.85	1
Population total	34.26	6.42	134.80	0.01	1378.66	1
Primary completion	92.10	97.06	15.71	40.87	131.02	86
Secondary ed.	6.37	6.00	0.92	4.00	9.00	13
Secondary ed. teachers	209.94	42.12	721.75	0.04	6219.58	94
Health exp.	6.74	6.27	2.99	1.75	23.29	31
Health exp. per capita	1426.02	801.76	1694.91	29.91	9869.74	33
Unemployment	8.34	6.29	6.21	0.15	27.47	106
Youth Unemployment	19.39	15.96	12.58	0.49	54.31	115
Rural population	39.68	38.81	24.07	0.00	87.61	3
Adolescent fertility rate	48.09	39.26	40.57	0.29	189.38	23
GDP per capita	20698.04	13247.65	21828.44	743.90	123573.63	24
Mobile subscriptions	107.21	110.14	40.37	14.25	321.45	16
Internet users	51.32	54.00	28.96	1.18	98.24	13

For the rest of the analysis we focus on only those variables that are highly correlated with life expectancy (see figure *scatter*). In this way, we choose the ten variables with the highest correlation coefficients, their corresponding boxplots are displayed in figure *boxplot*. This was so, as to maintain a simpler analysis. These variables are: electricity, unsafe water mortality, population growth, primary completion, health expenditure per capita, rural population, adolescent fertility rate, internet users, and mobile subscriptions.

We can see in figure 1 that some distributions are very skewed, e.g. electricity, unsafe water mortality and health expenditure per capita. Additionally, some have very high variance such as; internet users.

The relationship between life expectancy and GDP per capita in logarithm form was however analyzed with more detail, as we would expect it to be one of the principal determinants of life expectancy. Consequently, from the scatter plot in figure B.3 (see appendix B) we will expect that *log(GDP pc)* should explain, at least, 67.9% of variance of life expectancy.

2.2. Handling Missing Data

In resolving missing values which mostly occur in statistical analysis, different methods can be used. Amongst these include the *complete case analysis*. Here, observations with missing values for the explanatory variables are omitted, leaving only complete cases. Nevertheless, this could lead to

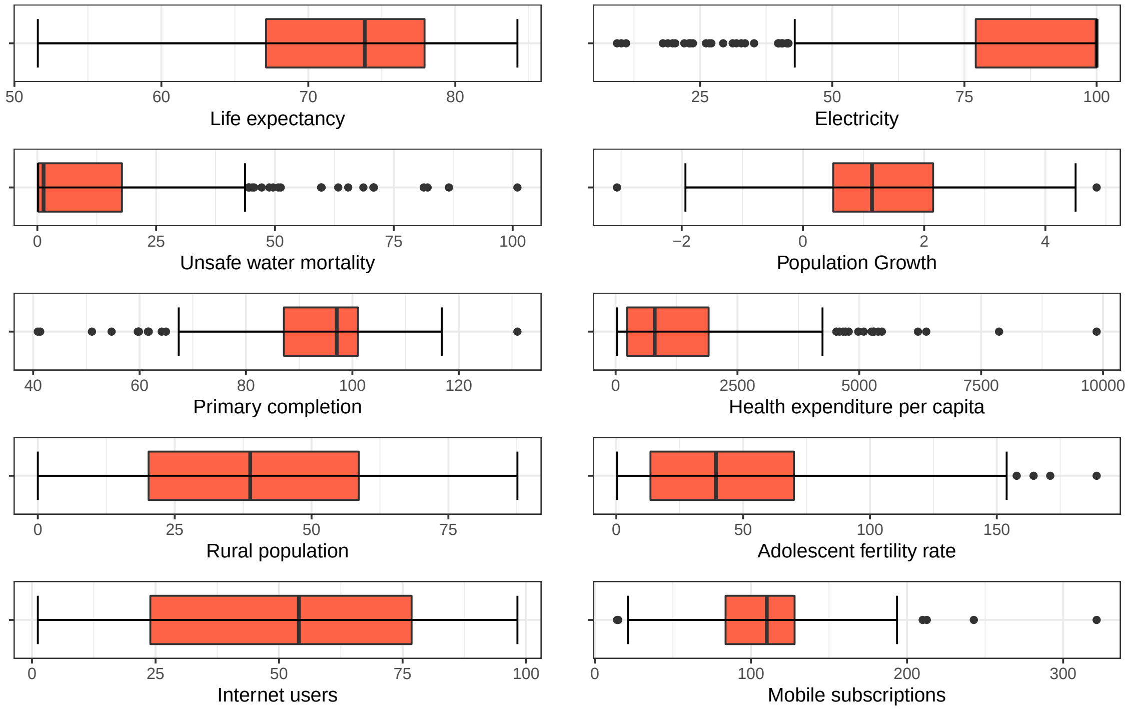


Figure 1: Distribution of variables of interest

biased estimators if the data is not missing completely at random (MCAR). However, even when this alternative is likely to be unbiased, it discards most of the information contained in the variability of the data. Thereby, is the least effective method in resolving missing values [1]. In this regard, keeping only strictly complete cases drops the number of observations to 0. Since there were no countries with complete information. For this reason, the conditions for complete cases were relaxed.

Out of a total of 217 countries, PPP, Literacy rate and Primary expenditure had 211, 183 and 146 missing values respectively (table 1). Thus, they were removed to relax the conditions. This then left us with 48 observations. We then assumed this as a complete case (i.e. excluding PPP, literacy rate and Primary education expenditure).

In the figure 2, the probability density function (PDF) of life expectancy under the *assumed complete case* was then plotted alongside with the PDF of life expectancy considering all observations (black and red curves, respectively).

From the density functions we can infer that countries with lower life expectancy are underrepresented as the distribution is skewed towards the left (i.e. its mass is concentrated in countries with higher life expectancy). Moreover, if we keep only those with complete observations, many of this observations are lost, making our analysis biased as the missingness appears to be not at random.

As we shown in figure B.3, life expectancy is closely correlated with the level of economic development, so we could expect that in lower income countries statistical data monitoring is less efficient.

To avoid the bias that would result from a complete case analysis, different predictor variables and

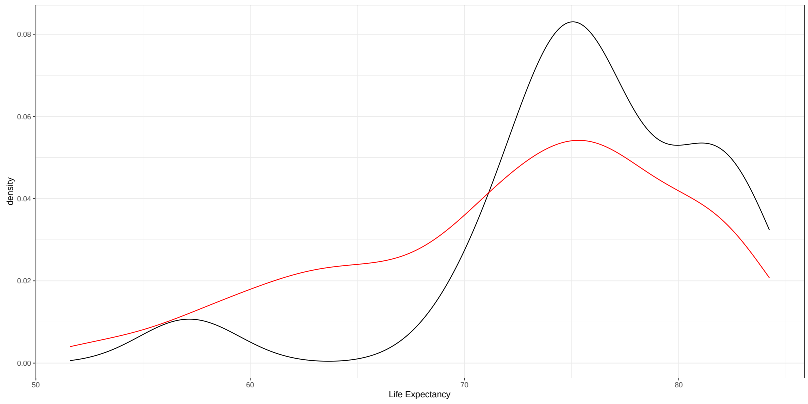


Figure 2: Probability Density Function for Life Expectancy

their features gathered from table 1 gave an insight on what method of imputation is the more optimal.

The imputation methods deployed include: i) Mean Imputation of missing values; ii) Median Imputation of missing values; iii) Linear Regression Method to perform deterministic imputation

Section 2.1 showed that the predictor variables and number of missing values. The criteria used to select the imputation method was the number of missing values. Therefore, *mean imputation* was used to deal with those variables with less than 100 missing values and symmetric distributions (i.e. similar median and mean values). In contrast, in the case of variables with less than 100 missing values but skewed distributions (i.e. large difference between mean and median values), *median imputation* was performed, as using the mean would drastically change their distribution.

Lastly, from the table 1, values within the range of 100-200 missing values were imputed with the *linear regression method* for deterministic imputation. As there are many non-observed data points, implementing the two previous methods would result in a very concentrated distribution around the mean (i.e. very high kurtosis), then most of the information provided by these variables would be lost.

Public Private Partnership Investment had a total of 211 NA's (that is, missing data) out of 217 countries. This is rather insignificant as there are too many missing values to help determine which method of imputation is to be deployed. Hence, inference would be not statistically reliable, therefore this variable was omitted for the rest of the analysis.

After imputation, the PDF of all imputed predictor variables were plotted to see the difference before and after imputation was made (see figure B.2 in appendix B).The red and black functions represent the distribution before and after imputation, respectively. Thus, in general, there were no major changes in their distributions.

2.3. Collinearity

In regression analysis, it can be the case that an explanatory variable could be expressed as a linear combination of the rest. In this case, we can say we are in the presence of collinearity. This issue signifies a problem, analytically the matrix X^tX, becomes singular (i.e. it is not invertible) and the parameters of the model can not be derived. In practice, they significantly increase the variance of the estimators. For this reasons, we conducted an analysis in order to determine if there is collinearity in our dataset.

There are several methods to establish collinearity. One of those consists in conducting a correlation test, to determine if the relation between two given variables is statistically significant. Another method consists in estimating the Variance Inflation Factor (VIF), which is computed from the R_j^2 resulting from regressing variable X_j against the reamining ones, or a subset that we suspect may be correlated with it. VIF can be defined as follows:

$$VIF = \frac{1}{1 - R_j^2} \tag{2.1}$$

There is no analytical rule on which is the correct threshold for this factor. However, some authors propose a maximum value 5 [2, 3]. In the present analysis, we set our threshold to this value. That is, if VIF> 5 we can say that the variables involved are linearly correlated.

Table 2: Results of Collinearity Analysis

	Dependent variable:					
	log(GDP per capita)					log(Water Mort.)
	(1)	(2)	(3)	(4)	(5)	
Rural Population	-0.006^{***}	-0.032^{***}				
	(0.001)	(0.002)				
log(Health Exp. pc)	0.684^{***}		0.807^{***}			0.009^{***}
	(0.049)		(0.029)			(0.001)
Electricity	0.003^{**}			0.032^{***}		
	(0.002)			(0.002)		
Internet Users	0.003^{*}				0.033^{***}	
	(0.002)				(0.002)	
Health Exp.	-0.103^{***}					
	(0.011)					
Adjusted Income	0.009					
	(0.016)					
Unemployment	-0.011^{**}					
	(0.005)					
log(Mobile Subs.)	0.115					
	(0.078)					
Constant	4.702^{***}	10.647^{***}	4.088^{***}	6.619^{***}	7.677^{***}	4.129^{***}
	(0.494)	(0.107)	(0.191)	(0.181)	(0.088)	(0.051)
Observations	217	217	217	217	217	217
Correlation test	0.654	-0.690	0.887	0.735	0.830	-0.834
R^2	0.889	0.476	0.787	0.540	0.688	0.334
VIF	9.008	1.909	4.688	2.173	3.206	1.502

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

Collinearity is tested within two categories: economic and health. We chose this categories as we would expect the variables in each to be correlated either because they are measuring the same thing or there is a third confounder non-observed establishing a mechanism of relation between them.

As seen in table 2, rural population, electricity and internet users have VIF values of 1.90, 2.17 and 3.21, respectively, on GDP pc. Therefore, the relation is not strong enough to establish collinearity. In the same way, the value of VIF is roughly 1.50 in the health category, which is much lower than 5.

In contrast, for health expenditure per capita, the value of VIF is approximately 4.69 (very close to the threshold) which imply a potential collinearity problem with GDP pc (this relation can be explained by the government expenditure multiplier). Finally, model 1 shows that GDP pc appears to be highly correlated with a set of selected variables, with a VIF equal 9.01. Hence, it can be said we are in the presence of collinearity.

As this issue increases the variance of the linear model estimators (i.e. high variance), there are three widely used methods to deal with it: i) Ignore collinearity if it is not strong enough; ii) Use an alternative estimation method such as ridge regression [4]; iii) Implementing the variable selection technique to discard the correlated variable.

3. Analysis

Life expectancy at birth can be explained by several factors specific to each country. Namely, quality and coverage of health services, other public services like security, infrastructure and education, eating patterns, genetics, and even geographical factors (e.g. local outbreaks, risk of natural disasters, etc.) [5]. In this way, a great number of studies establish a correlation between life expectancy and a set of economic indicators that can help to explain its evolution through time and therefore, the source of variations across countries [5, 6, 7, 8].

In this section, we try to identify the determinants of life expectancy at an aggregated country level. The results will help to understand the role of economic development and how public policies affect life quality. To achieve this, a linear regression and ANOVA analyses are conducted.

3.1. Regression Analysis

We propose a multivariate linear model to explore which variables in our data set are good predictors of life expectancy. One of the advantages of using this type of models is the parameters are easy to interpret and, at the same time, it can be used to model non-linear relations between explanatory variables and the dependent variables by transforming the data, maintaining the linearity of the parameters (e.g. log, polynomial or exponential-transformations).

From the analysis in section 2.1 (see figure B.1 in the appendix) we decide to initially consider only those variables with the highest correlation with life expectancy in order to keep the analysis parsimonious. The criteria used to determine relevance is the *Pearson correlation coefficient* to be greater than $|0.5|$ (the correlation matrix is shown in figure B.5 in the appendix).

In section 2.3 it was found that *log(GDP pc)* is highly correlated with other variables in the dataset. The reason why we tested collinearity between these is because we would typically use those variables to explain the GDP per capita [9]. In this context, including the variable *GDP per capita* would be redundant to the analysis, because it adds no additional explanatory power to the model and, in fact, would increase the variance of the estimators as shown by the VIF in table 2.

For this reason, two different saturated models are proposed. The first one excludes *GDP pc*, and the second one includes this variable and a series of controls.

Saturated Model

$$\textit{life expectancy} = X\beta + \epsilon \tag{3.1}$$

where *life expectancy* $\sim N(X\beta, \sigma^2)$ is the dependent variable, $\epsilon \sim N(0, \sigma^2)$ a random error term, β the vector of parameters to be estimated and X a vector of covariates that includes: *electricity*, *log(water mortality)*, *population growth*, *log(adjusted income)*, *primary completion*, *health expenditure pc*, *health expenditure pc*^{2}, *rural population*, *adolescent fertility rate*, *internet users* and *mobile subscriptions*.[1]

As the relation between *life expectancy* and *health expenditure per capita* appears to be non-linear we decide to include a polynomial term of second degree for this variable (see figure B.1 in the appendix). The relation between these two variables is positive at a diminishing rate (concave shape), for this reason we would expect the linear term to be positive and the squared term to be negative.

Log-log model

$$\textit{log(life expectancy)} = \beta_0 + \beta_1 \textit{log(GDP pc)} + \beta_2 \textit{Gini Index} + \beta_j \textit{contintent}_j + \epsilon \tag{3.2}$$

The dependent variable and error term have the same assumptions as in the previous model, in this case the explanatory variables are: *log(GDP pc)*, *Gini index* and a continent indicator (Africa is the omitted group[2]). The parameters of a *log-log* model can be interpreted as the elasticity of the dependent against the explanatory variable.[3]

3.1.1. Results

In order to select the variables that lead to the smaller expected squared error, we perform *stepwise* variable selection. As there is no hard rule on which information criterion is the best, both the *Akaike Information Criterion* (AIC) and *Bayesian Information Criterion* were considered.

After using the *step* command in R in both directions, the results are shown in table 3. Model 1 shows the best model starting from the saturated model. All of the variables are statistically significant and *log(water mortality)* is the one with the highest effect on life expectancy. Nonetheless, as life

[1]A description of each variable is given in appendix A.

[2]Every continent parameter is interpreted as the difference with respect to the omitted group.

[3]For small changes, shows the percentage change of the dependent variable when the explanatory changes by 1%.

expectancy is estimated from the mortality rates at birth of the corresponding cohort, as if they would keep constant throughout the life of the cohort individuals, including the water mortality rate is trivial [10].

For this reason, we omit this variable and perform *stepwise* variable selection again. Model 2 is the one selected based on the AIC and model 3 on the BIC.

In model 2 we observe that the principal determinant of life expectancy is the health expenditure per capita. In this way, life expectancy could be used as an indicator of the public health policy efficacy, however higher mortality would in fact lead to an increase in health expenditure. Other significant explanatory variables are electricity coverage and adolescent fertility. According to the World Health Organization, complications during pregnancy and childbirth are the leading cause of death for teenage girls [11]. Additionally, the access to electricity services and life expectancy correlation could be explained by a hidden variable that reflects the living standards quality.

Model 3 reflects the fact that BIC penalizes more heavily in the number of parameters to estimate, in this way the best model is the one without the squared term.

Finally, in model 4 all the parameters are statistically significant. As mentioned before, the *GDP pc* parameter can be interpreted as an elasticity, in this way a 1% increase in the mean GDP per capita would lead to an increase of 0.05% in the mean life expectancy. But, it is important to note that this correlation does not imply causality, since the relation may be endogenous (i.e. they are mutually correlated). Regarding the continent indicators, every continent has a positive correlation with life expectancy compared to Africa (the omitted group) (see figure B.6).

In order to assess linear regression assumptions fulfillment, a discussion is presented in the appendix C.

So as to determine what specification is the best to predict new unseen observations, cross-validation techniques could be implemented (i.e. randomly splitting the data into test and training sets). Another solution is testing our models with different years observations, or by predicting the life expectancy missing values in our data set and validate these predictions with a secondary source.

3.2. ANOVA

A One-Way ANOVA (Analysis of Variance) is a statistical technique by which we can test if three or more means are equal, i.e. if one differs significantly among three or more levels of a factor.

The benefits of using this method are: it is easy to estimate and can be manually computed using simple algebra rather than complex matrix calculations. It can control the overall Type I error rate and provide an overall test of equality of group means. If normality assumption is true then the test is more powerful.

Table 3: Results of Regression Analyses

	Dependent variable:			
	Life Expectancy			log(Life Expectancy)
	(1)	(2)	(3)	(4)
Electricity	0.101^{***}	0.120^{***}	0.110^{***}	
	(0.017)	(0.016)	(0.015)	
log(Water Mort.)	-0.835^{***}			
	(0.259)			
Health Exp. pc	0.629^{***}	1.751^{***}	0.705^{***}	
	(0.209)	(0.556)	(0.213)	
Health Exp. pc^2		-0.136^{**}		
		(0.067)		
Adolscent Fertility	-0.028^{***}	-0.036^{***}	-0.041^{***}	
	(0.010)	(0.009)	(0.009)	
Internet Users	0.060^{***}	0.074^{***}	0.084^{***}	
	(0.018)	(0.018)	(0.016)	
Mobile Coverage	-1.179^{*}	-1.049		
	(0.690)	(0.699)		
log(GDP pc)				0.053^{***}
				(0.005)
Gini Index				-0.003^{***}
				(0.001)
Asia				0.084^{***}
				(0.013)
Europe				0.103^{***}
				(0.016)
North America				0.145^{***}
				(0.019)
Oceania				0.091^{***}
				(0.018)
South America				0.105^{***}
				(0.014)
Constant	67.143^{***}	63.250^{***}	59.834^{***}	3.811^{***}
	(3.048)	(2.859)	(1.464)	(0.043)
Observations	195	195	195	195
R^2	0.834	0.828	0.823	0.772
AIC	**460.81**	**467.05**	469.37	-1,126.7
BIC	484.34	490.58	**486.18**	-1,100.5

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

In this section ANOVA analysis is conducted to test if there is a statistical difference of mean life expectancy between the 5 different continents. The M49 classification from the United Nations is used.

The results are shown in table 4, the five continents average of life expectancy differed significantly on anxiety level, $F_{(5,189)} = 2e - 16$,p<0.05. As a consequence, this difference can be a reflect of economic and political trends across regions. The p value on the table is 2e-16 which is less than 0.05 indicates that two or more groups have significantly different means.

3.3. Define Null and Alternative Hypotheses

Life expectancy of $\mu_{Asia} = \mu_{Oceania} = \mu_{NorthAmerica} = \mu_{SouthAmerica} = \mu_{Europe}$

Not all of μ are equal. Alpha: $\alpha = 0.05$

3.4. Results

Table 4: ANOVA test

	Df	Sum sq	Mean sq	F value	$Pr > F$
Continents	5	7016	1403.2	57.92	$< 2e - 16$
Residuals	189	4579	24.2		

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05'.' 0.1 ' ' 1**
18 observations deleted due to missingness

If F is greater than $F_{(0.05,5,189)} = 2.261892$ reject the null hypothesis.

The five continents average of life expectancy differed significantly on anxiety level, $F_{(5,189)} = 2e - 16$,p<0.05.

The p value on the table is 2e-16 which is less than 0.05 indicates that two or more groups have significantly different means.

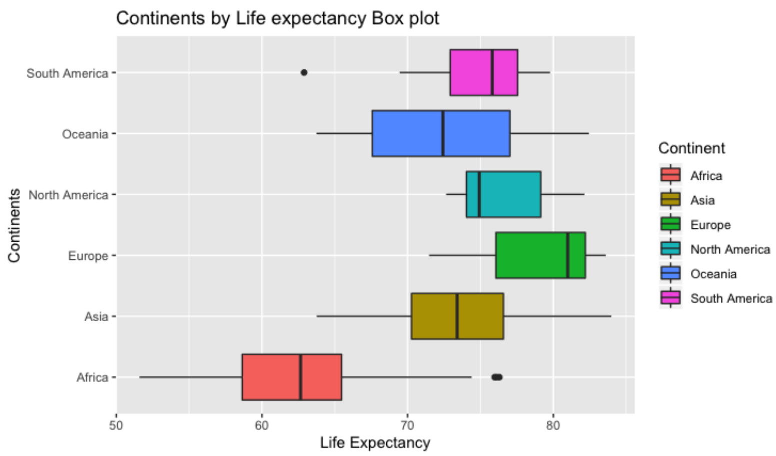


Figure 3: This box chart shows that there is notable different in each sample. Africa's life expectancy is lower than others continents and have a huge differ life expectancy between max and min. The life expectancy in Asia and Oceania are near normal distribution. The box plot of North America is left-skewed. In contrast, Europe and South Americas' are right-skewed.

A discussion on the assumptions of ANOVA is presented in appendix D.

4. Discussion

Understanding the determinants of life expectancy could be useful to explore the extent to which its difference across countries and continents is explained by environmental random factors and by economic and political ones. This in turn, could be used by international organizations to assess the effect and impact of such policies. Namely, if public expenditure is being allocated efficiently, how important is education to improve people's well-being, among others. In this way, helping into the design of better plans and social programs.

However, this type of data is prone to have several statistical issues. For instance, the quality of the information collection techniques varies substantially across countries. In fact, those with lower a lower level of development tend to have incomplete data. As a consequence, it is important to correctly identify, evaluate and correctly deal with problems like: missing data, collinearity, unbalanced categories, high variance variables, etc.

Throughout this report a statistical preliminary analysis was conducted, including: descriptive satistics to identify the properties of the data and possible issues, methods to deal with missing data and collinearity test. Besides, a series of lineal models were proposed aimed to better predict and understand life expectancy, as well as an ANOVA analysis to test difference across continents.

It was found that the most important factors are those related to health services and sanitary factors, and the degree of access to basic services as electricity services and education. Moreover, there is a significant difference across regions. It is important to note that these relations does not imply causality, as it may be endogenous. To this end, further analysis through identification methods like instrumental variables could be explored.

Contributions: All members contributed to the elaboration of the report. However, each one focused on a specific task:

- **1908015**: Task 1
- **1901197**: Task 2
- **1900716**: Task 3
- **1900396**: Task 4
- **1901094**: Task 5

References

[1] Sterne AC, White I, Carlin J, Spratt M, Royston P, Kenward M, and Carpenter A, Woodand J. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338: b2393, 2009.

[2] Rogerson PA. Data reduction: factor analysis and cluster analysis, 2001.

[3] Shah MH and Samdani S. Impact of trade liberalization on fdi inflows to d-8 countries. *Shah, MH, & Samdani, S.(2015). Impact of trade liberalization on FDI inflows to D-8 countries. Global Management Journal for Academic & Corporate Studies (GMJACS)*, 5(1):30–37, 2015.

[4] Draper NR and Smith H. *Applied regression analysis*, volume 326. John Wiley & Sons, 1998.

[5] Rodgers G. B. Income and inequality as determinants of mortality: An international cross-section analysis. *Population Studies*, 33(2):343–351, 1979. ISSN 00324728.

[6] Wilkinson R. G. Income distribution and life expectancy. *BMJ*, 304(6820):165–168, 1992. ISSN 0959-8138.

[7] Messias E. Income inequality, illiteracy rate, and life expectancy in brazil. *American journal of public health*, 93:1294–6, 09 2003.

[8] Shaw J., William H., and Ronald V. The determinants of life expectancy: An analysis of the oecd health data. *Southern Economic Journal*, 71(4):768–783, 2005. ISSN 00384038.

[9] Parkin D., McGuire A., and Yule B. Aggregate health care expenditures and national income: is health care a luxury good? *Journal of health economics*, 6(2):109–127, 1987.

[10] OECD. *Health at a Glance 2019*. 2019.

[11] World Health Organization. Global health estimates 2015: deaths by cause, age, sex, by country and by region, 2000–2015, 2016. Geneva.

Appendix

A. Glossary

Life Expectancy: Life expectancy at birth, total (years)

Electricity: Access to electricity (% of population)

Adjusted income: Adjusted net national income (current US$)

Children out of school: Children out of school (% of primary school age)

Primary education expenditure: Expenditure on primary education (% of government expenditure on education)

PPP: Public private partnerships investment in water and sanitation (current US$)

Unsafe water mortality: Mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene (per 100,000 population)

Adult literacy rate: Literacy rate, adult total (% of people ages 15 and above)

Population Growth: Population growth (annual %)

Population total: Population, total

Primary completion: Primary completion rate, total (% of relevant age group)

Secondary ed.: Secondary education, duration (years)

Secondary ed. Teacher: Secondary education, teachers

Health expenditure: Current health expenditure (% of GDP)

Health expenditure per capita: Current health expenditure per capita, PPP (current international $)

Unemployment: Unemployment, total (% of total labor force) (national estimate)

Youth unemployment: Unemployment, youth total (% of total labor force ages 15-24) (national estimate)

Rural population: Rural population (% of total population)

Adolescent fertility rate: Adolescent fertility rate (births per 1,000 women ages 15-19)

GDP per capita: GDP per capita, PPP (current international $)

Mobile subscriptions: Mobile cellular subscriptions (per 100 people)

Internet users: Individuals using the Internet (% of population)

B. Additional Plots

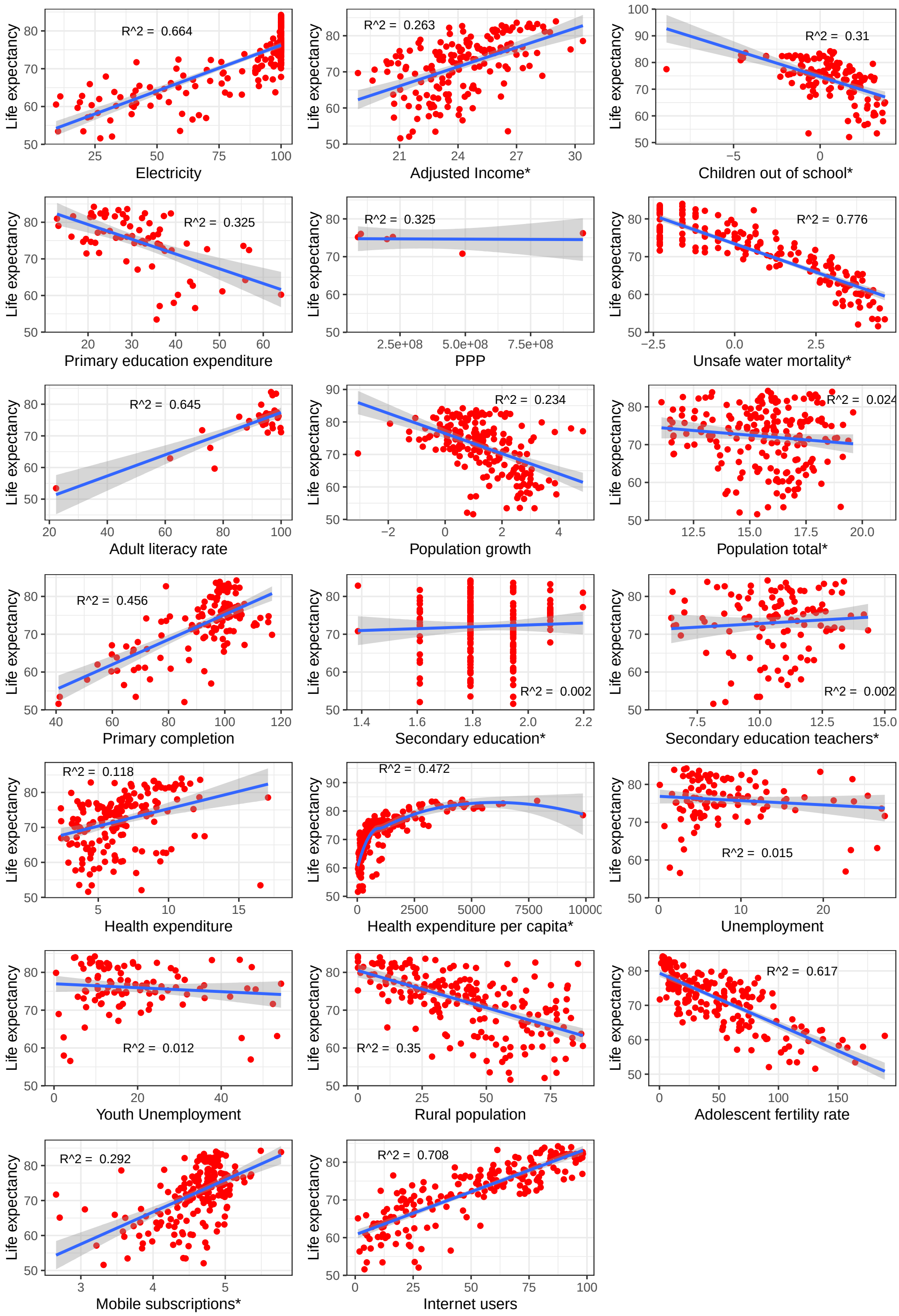


Figure B.1: Life Expectancy against all the Variables

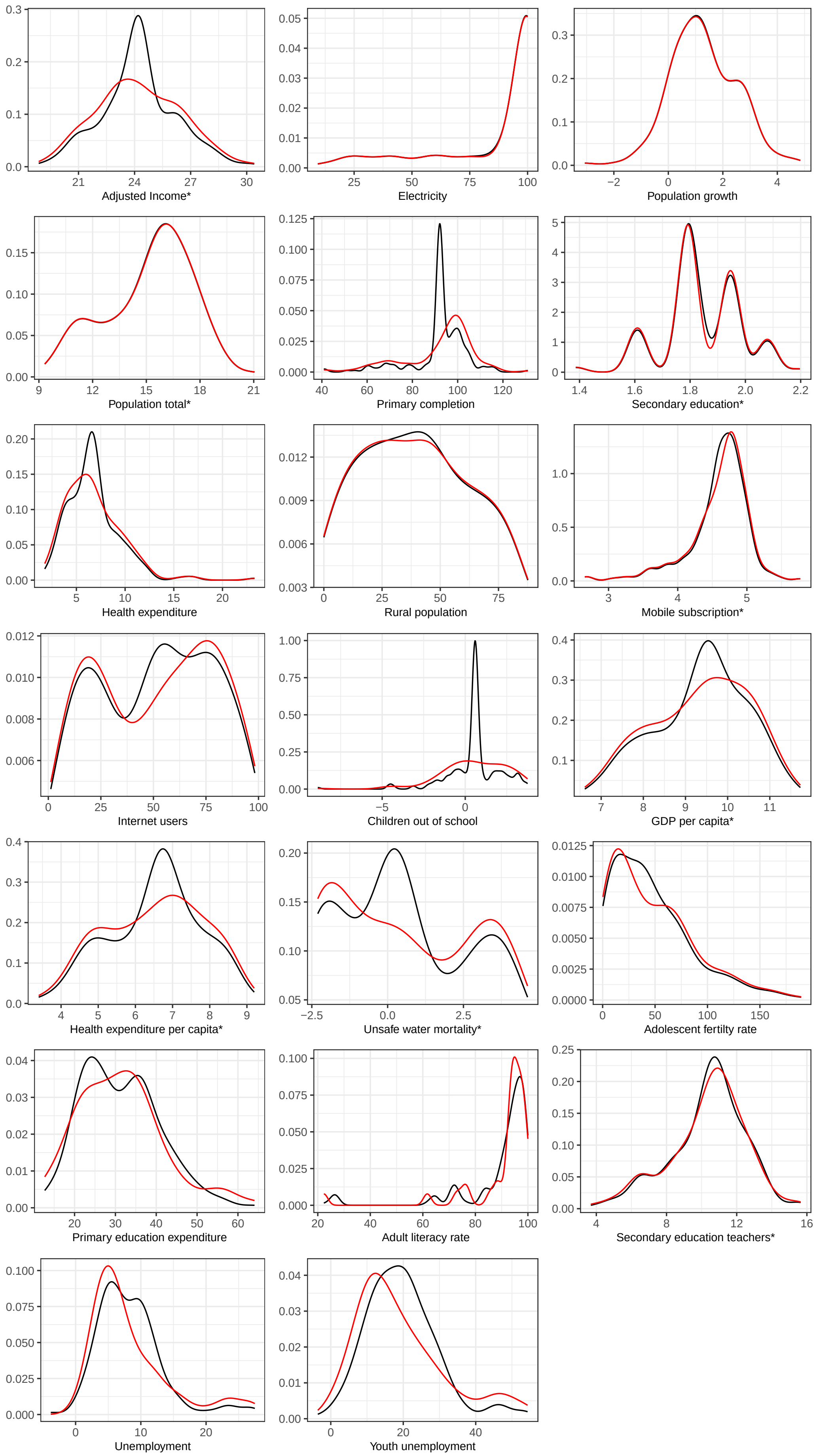


Figure B.2: Probability Density Function for Variables

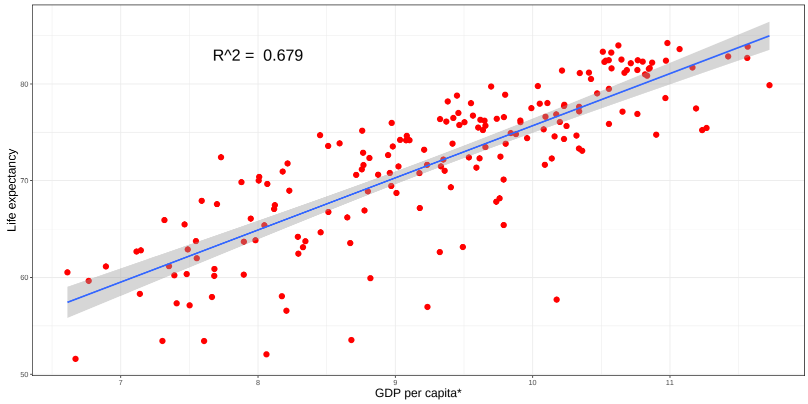


Figure B.3: Correlation between Life Expectancy and log(GDP pc)

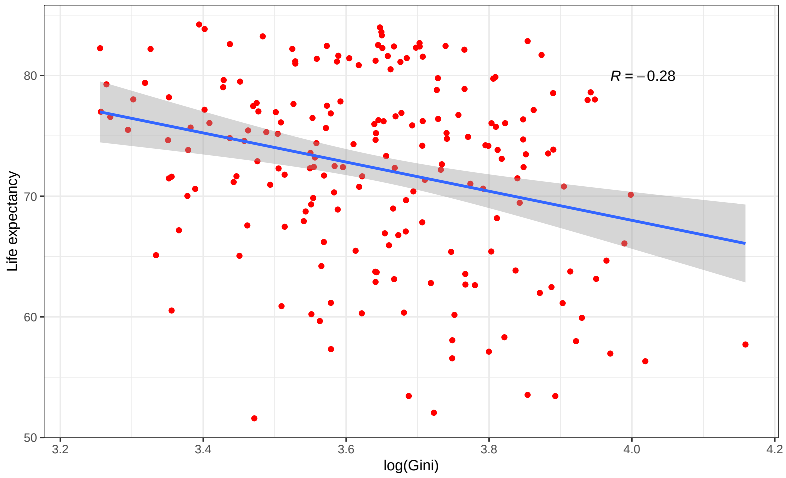


Figure B.4: Gini Index vs Life Expectancy

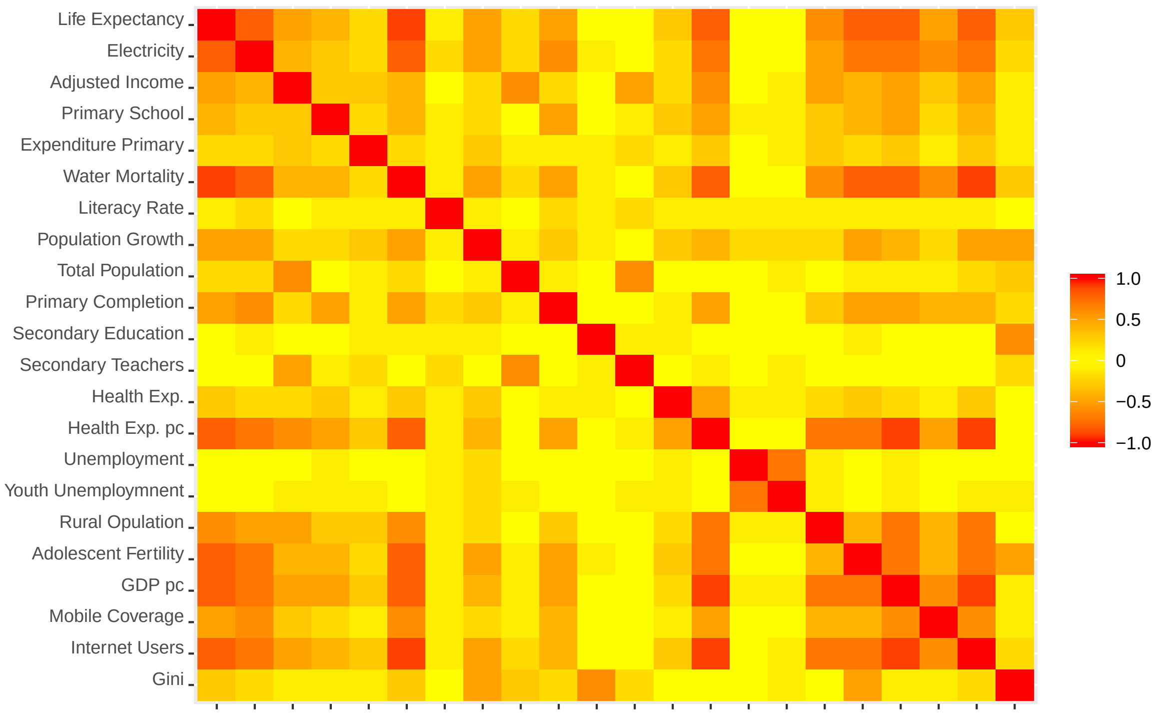


Figure B.5: Correlation Matrix for the Dataset

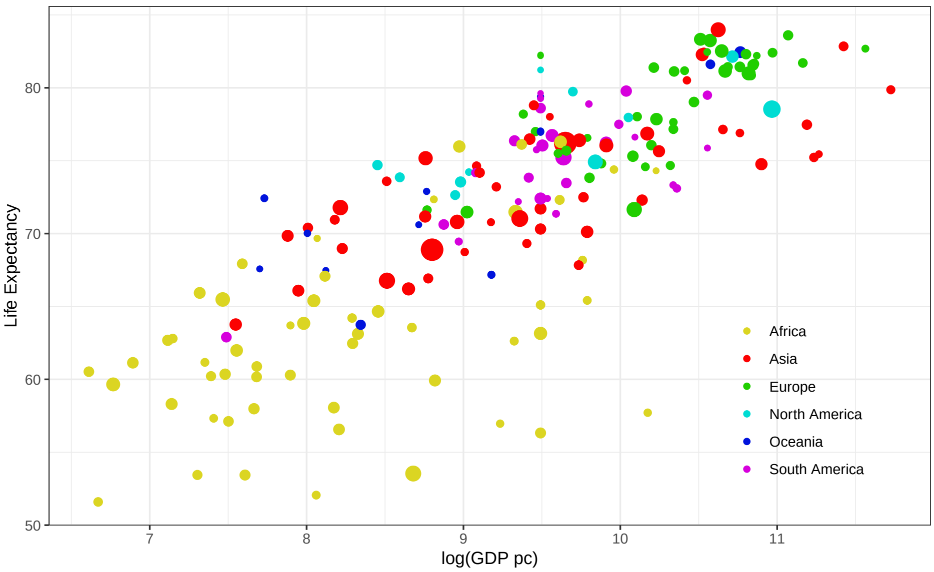


Figure B.6: Life Expectancy by Contnitnet

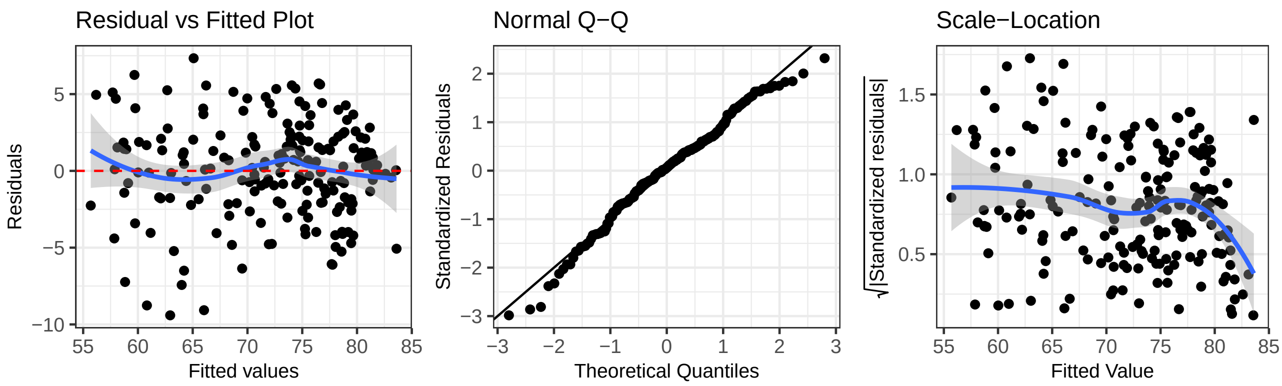
C. Discussing the Assumptions of Linear Regression Analysis

The *residual vs fitted plot* shows the linearity assumption between dependent and explanatory variables, we would expect the errors to be randomly and equally distributed around zero across all the range of life expectancy. In all the models this assumption appears to hold for those values around mean life expectancy. However, the lower age range seem to be underrepresented in the sample. Another reason may be that the determinants of life expectancy are different along the distribution. To solve this issue a quantile regression could be performed, in this way we would be able to find how the weights of each explanatory variable change across the distribution.

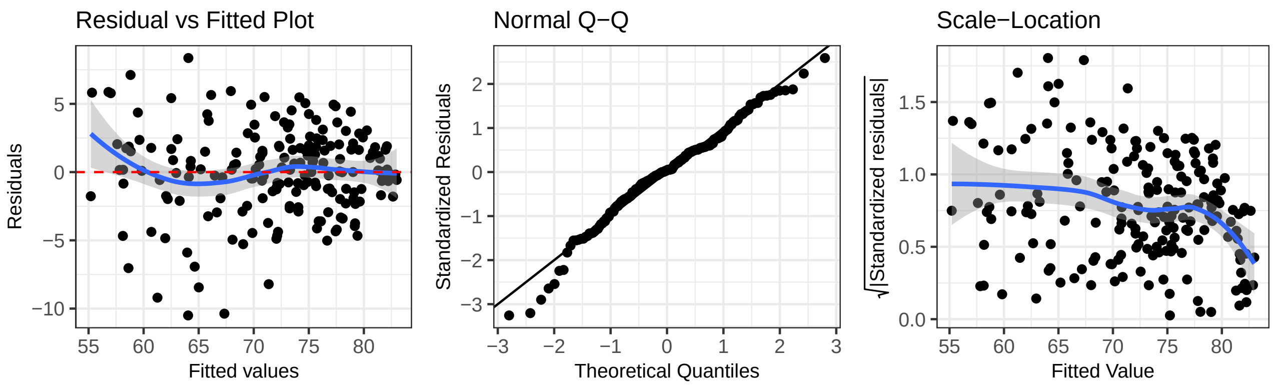
The *Q-Q plot* plot is used to assess visually the normality assumption. It appears to fulfill in general, but, as in the previous paragraph, distribution tails are not well-behaved. The reason could be that, as life expectancy and national income are positively related, lower income countries may have poor quality data.

The third plot *Scale-Location plot* is used to test homoscedasticity assumption (constant variance across the sample)[4]. Again, the residuals behave erratically towards the extremes of the distribution. Two solutions to make our estimations consistent are *Huber-White standard errors* or *Boostrap standard errors* (not discussed in this report).

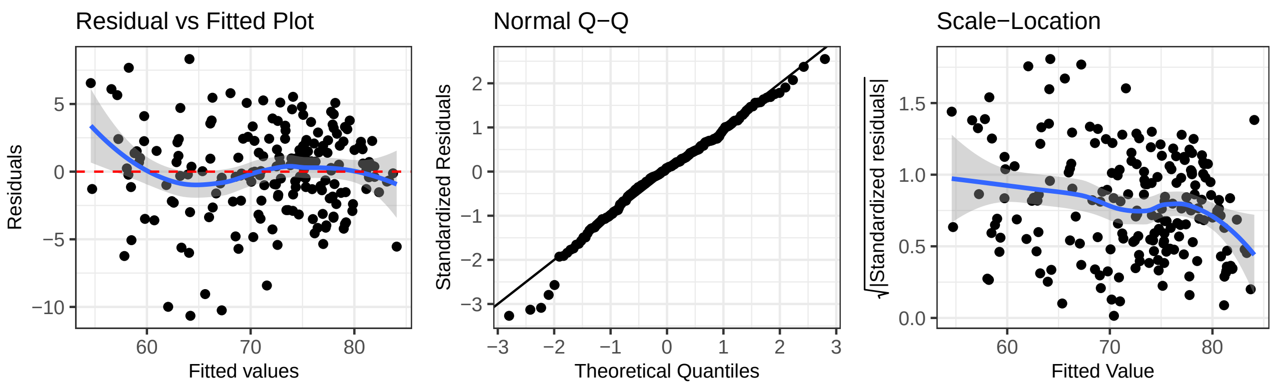
[4]The Barlett test is used to test statistically equal variance assumption.



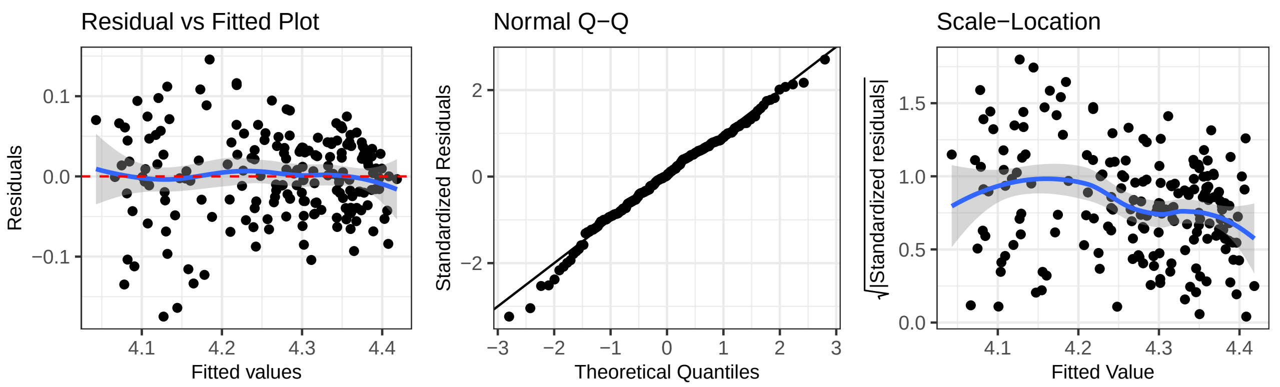
(a) Model 1



(b) Model 2



(c) Model 3



(d) Model 4

Figure C.1: Diagnostic Plots for Linear Regression

D. Discussing the Assumptions of ANOVA

D.1. Bartlett test

Bartlett's test allows you to compare the variance of two or more samples to determine whether they are drawn from populations with equal variance. It is suitable for normally distributed data. The test has the null hypothesis that the variances are equal and the alterntive hypothesis that they are not equal.

Data: Life expectancy by Continent

Bartlett's K-squared = 19.743, df = 5, p-value = 0.001397

From the output we can see that the p-value of 0.001397 is less than the significance level of 0.05. This means we can reject the null hypothesis.

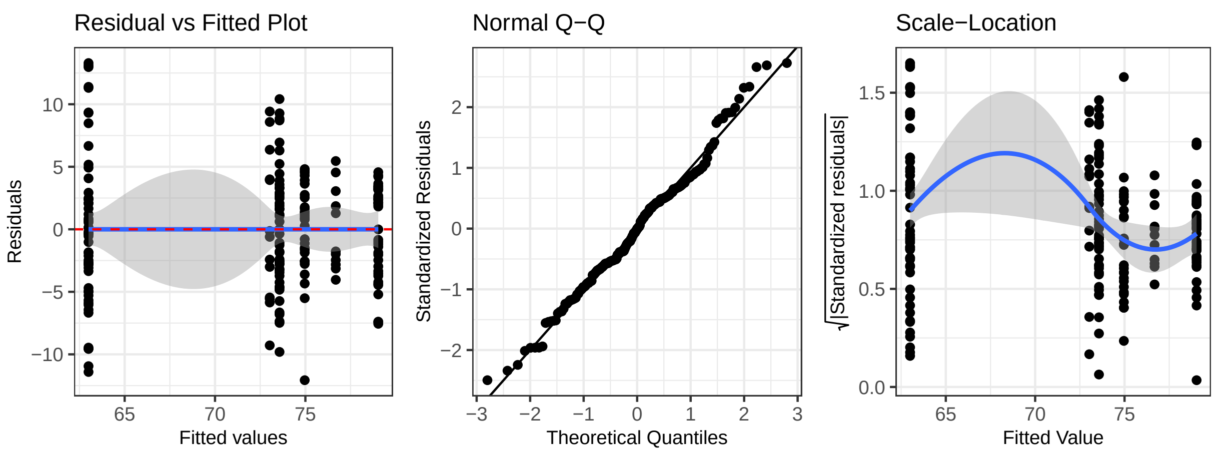


Figure D.1: This plot shows the residuals (errors) on the y-axis and the fitted values (predicted values) on the x-axis. We can see that the points local in two group under 65 and over 72.5 and the residuals are not distributed equally above and blew zero. The red line is flat, then the relationship between the independent and dependent variable is linear.

Figure D.2: The Normal QQ plot indicates that the independent variable is approximately normal, since the observations approximately line-up with the Theoretically-derived normal value.

Figure D.3: The Scale-Location plot shows if residuals are spread equally along the ranges of predictors. This is how we can check the assumption of equal variance.

E. R code

```r
#######################
####### MA 317 #########
##### Group Project ####
## Clenaning the Data ##
######################

#loading libraries
library(ggplot2)
library(pastecs)
library(psych)

#set wd
wd=setwd("D:/Documentos/Essex/Modelling Experimental Data/group project")

#reading the data set
data=read.csv("LifeExpectancy1.csv", header=T)

#asigning different names
names(data)=c("country", "code", "life_exp", "electricity", "adj_income", "primary_school", "exp_primary_ed", "ppp", "mortality_water", "literacy_rate", "pop_growth", "pop_total", "primary_completion", "secondary_education_duration", "sec_ed_teachers", "health_exp", "health_exp_pc", "unemployment_total", "unemployment_youth", "rural_pop", "adolscent_fert", "GDP_pc", "mobile_coverage", "internet_users")

#Creating a list of variables to be transformed to log form
lista=list("adj_income","primary_school","mortality_water","pop_total", "secondary_education_duration", "GDP_pc", "mobile_coverage", "sec_ed_teachers", "health_exp_pc")

#loop to iterate throught the list
for (l in lista){
  print(l)
  data[l]=log(data[l])
}

#save the data
write.csv(data, "LifeExpectancy1_log.csv", row.names = F)

#######################
####### MA 317 #########
##### Group Project ####
#######Task 1#########
#Descriptive Satatistics#
######################
#model of life expectancy with all variables
model1 <- lm(life_exp ~ electricity, data=countries1)
model2 <- lm(life_exp ~ adj_income, data=countries1)
model3 <- lm(life_exp ~ primary_school, data=countries1)
model4 <- lm(life_exp ~ exp_primary_ed, data=countries1)
model5 <- lm(life_exp ~ ppp, data=countries1)
model6 <- lm(life_exp ~ mortality_water, data=countries1)
```

```
model7 <- lm(life_exp ~ literacy_rate, data=countries1)
model8 <- lm(life_exp ~ pop_growth, data=countries1)
model9 <- lm(life_exp ~ pop_total, data=countries1)
model10 <- lm(life_exp ~ primary_completion, data=countries1)
model11<- lm(life_exp ~ secondary_education_duration, data=countries1)
model12 <- lm(life_exp ~ sec_ed_teachers, data=countries1)
model13 <- lm(life_exp ~ health_exp, data=countries1)
model14 <- lm(life_exp ~ health_exp_pc, data=countries1)
model15 <- lm(life_exp ~ unemployment_total, data=countries1)
model16 <- lm(life_exp ~ unemployment_youth, data=countries1)
model17 <- lm(life_exp ~ rural_pop, data=countries1)
model18 <- lm(life_exp ~ adolscent_fert, data=countries1)
model19 <- lm(life_exp ~ GDP_pc, data=countries1)
model20 <- lm(life_exp ~ mobile_coverage, data=countries1)
model21 <- lm(life_exp ~ internet_users, data=countries1)
model19
#make text of r square for scatter plots
e1 <- paste("R^2 = ", round(summary(model1)$r.squared, 3))
e2 <- paste("R^2 = ", round(summary(model2)$r.squared, 3))
e3 <- paste("R^2 = ", round(summary(model3)$r.squared, 3))
e4 <- paste("R^2 = ", round(summary(model4)$r.squared, 3))
e5 <- paste("R^2 = ", round(summary(model5)$r.squared, 3))
e6 <- paste("R^2 = ", round(summary(model6)$r.squared, 3))
e7 <- paste("R^2 = ", round(summary(model7)$r.squared, 3))
e8 <- paste("R^2 = ", round(summary(model8)$r.squared, 3))
e9 <- paste("R^2 = ", round(summary(model9)$r.squared, 3))
e10 <- paste("R^2 = ", round(summary(model10)$r.squared, 3))
e11 <- paste("R^2 = ", round(summary(model11)$r.squared, 3))
e12 <- paste("R^2 = ", round(summary(model12)$r.squared, 3))
e13 <- paste("R^2 = ", round(summary(model13)$r.squared, 3))
e14 <- paste("R^2 = ", round(summary(model14)$r.squared, 3))
e15 <- paste("R^2 = ", round(summary(model15)$r.squared, 3))
e16 <- paste("R^2 = ", round(summary(model16)$r.squared, 3))
e17 <- paste("R^2 = ", round(summary(model17)$r.squared, 3))
e18 <- paste("R^2 = ", round(summary(model18)$r.squared, 3))
e19 <- paste("R^2 = ", round(summary(model19)$r.squared, 3))
e20 <- paste("R^2 = ", round(summary(model20)$r.squared, 3))
e21 <- paste("R^2 = ", round(summary(model21)$r.squared, 3))

#scatter plot with linear regression
gp1 = ggplot(countries1)+geom_point(aes(electricity, life_exp),
↪ colour="red")+theme_bw()+geom_smooth(aes(electricity, life_exp), method = "lm") +
↪ xlab("Electricity") + ylab("Life expectancy")+annotate(geom="text", x = 50, y= 80,
↪ label=e1,color="black", size= 3)
gp2 = ggplot(countries1)+geom_point(aes(adj_income, life_exp),
↪ colour="red")+theme_bw()+geom_smooth(aes(adj_income, life_exp), method = "lm") + xlab("Adjusted
↪ Income*") + ylab("Life expectancy")+annotate(geom="text", x=21, y=83, label=e2,color="black",
↪ size= 3)
```

```r
gp3 = ggplot(countries1)+geom_point(aes(primary_school, life_exp), colour="red")+theme_bw()+geom_smooth(aes(primary_school, life_exp), method = "lm") + xlab("Children out of school*") + ylab("Life expectancy")+annotate(geom="text", x=1, y=90, label=e3,color="black", size= 3)
gp4 = ggplot(countries1)+geom_point(aes(exp_primary_ed, life_exp), colour="red")+theme_bw()+geom_smooth(aes(exp_primary_ed, life_exp), method = "lm") + xlab("Primary education expenditure")+ ylab("Life expectancy") +annotate(geom="text", x=50, y=80, label=e4,color="black", size= 3)
gp5 = ggplot(countries1)+geom_point(aes(ppp, life_exp), colour="red")+theme_bw()+geom_smooth(aes(ppp, life_exp), method = "lm") + xlab("PPP")+ ylab("Life expectancy")+annotate(geom="text", x=2.5e+08, y=80, label=e4,color="black", size= 3)
gp6 = ggplot(countries1)+geom_point(aes(mortality_water, life_exp), colour="red")+theme_bw()+geom_smooth(aes(mortality_water, life_exp), method = "lm") + xlab("Unsafe water mortality*") + ylab("Life expectancy")+annotate(geom="text", x=3, y=80, label=e6,color="black", size= 3)
gp7 = ggplot(countries1)+geom_point(aes(literacy_rate, life_exp), colour="red")+theme_bw()+geom_smooth(aes(literacy_rate, life_exp), method = "lm") + xlab("Adult literacy rate") + ylab("Life expectancy")+annotate(geom="text", x=60, y=80, label=e7,color="black", size= 3)
gp8 = ggplot(countries1)+geom_point(aes(pop_growth, life_exp), colour="red")+theme_bw()+geom_smooth(aes(pop_growth, life_exp), method = "lm") + xlab("Population growth") + ylab("Life expectancy")+annotate(geom="text", x=3, y=87, label=e8,color="black", size= 3)
gp9 = ggplot(countries1)+geom_point(aes(pop_total, life_exp), colour="red")+theme_bw()+geom_smooth(aes(pop_total, life_exp), method = "lm") + xlab("Population total*") + ylab("Life expectancy") + xlim(11,21)+annotate(geom="text", x=20, y=82, label=e9,color="black", size= 3)
gp10 = ggplot(countries1)+geom_point(aes(primary_completion, life_exp), colour="red")+theme_bw()+geom_smooth(aes(primary_completion, life_exp), method = "lm") + xlab("Primary completion") + ylab("Life expectancy")+ xlim(40,120)+annotate(geom="text", x=60, y=79, label=e10,color="black", size= 3)
gp11 = ggplot(countries1)+geom_point(aes(secondary_education_duration, life_exp), colour="red")+theme_bw()+geom_smooth(aes(secondary_education_duration, life_exp), method = "lm") + xlab("Secondary education*") + ylab("Life expectancy")+annotate(geom="text", x=2.1, y=55, label=e11,color="black", size= 3)
gp12 = ggplot(countries1)+geom_point(aes(sec_ed_teachers, life_exp), colour="red")+theme_bw()+geom_smooth(aes(sec_ed_teachers, life_exp), method = "lm") + xlab("Secondary education teachers*") + ylab("Life expectancy")+annotate(geom="text", x=14, y=55, label=e12,color="black", size= 3) + xlim(6,15)
gp13 = ggplot(countries1)+geom_point(aes(health_exp, life_exp), colour="red")+theme_bw()+geom_smooth(aes(health_exp, life_exp), method = "lm") + xlab("Health expenditure") + ylab("Life expectancy")+ xlim(2,18)+annotate(geom="text", x=5, y=86, label=e13,color="black", size= 3)
gp14 = ggplot(countries1)+geom_point(aes(health_exp_pc, life_exp), colour="red")+theme_bw()+geom_smooth(aes(health_exp_pc, life_exp)) + xlab("Health expenditure per capita*") + ylab("Life expectancy")+annotate(geom="text", x=2500, y=95, label=e14,color="black", size= 3)
gp15 = ggplot(countries1)+geom_point(aes(unemployment_total, life_exp), colour="red")+theme_bw()+geom_smooth(aes(unemployment_total, life_exp), method = "lm") + xlab("Unemployment") + ylab("Life expectancy")+annotate(geom="text", x=12, y=62, label=e15,color="black", size= 3)
```

```
gp16 = ggplot(countries1)+geom_point(aes(unemployment_youth, life_exp), colour="red")+theme_bw()+geom_smooth(aes(unemployment_youth, life_exp), method = "lm") + xlab("Youth Unemployment") + ylab("Life expectancy")+annotate(geom="text", x=25, y=60, label=e16,color="black", size= 3)
gp17 = ggplot(countries1)+geom_point(aes(rural_pop, life_exp), colour="red")+theme_bw()+geom_smooth(aes(rural_pop, life_exp), method = "lm") + xlab("Rural population") + ylab("Life expectancy")+annotate(geom="text", x=12, y=60, label=e17,color="black", size= 3)
gp18 = ggplot(countries1)+geom_point(aes(adolscent_fert, life_exp), colour="red")+theme_bw()+geom_smooth(aes(adolscent_fert, life_exp), method = "lm") + xlab("Adolescent fertility rate") + ylab("Life expectancy")+annotate(geom="text", x=120, y=80, label=e18,color="black", size= 3)
gp19 = ggplot(countries1)+geom_point(aes(GDP_pc, life_exp), colour="red", size = 3)+theme_bw()+theme(axis.title=element_text(size=15))+geom_smooth(aes(GDP_pc, life_exp), method = "lm") + xlab("GDP per capita*") + ylab("Life expectancy")+annotate(geom="text", x=8, y=83, label=e19,color="black", size= 7)
gp20 = ggplot(countries1)+geom_point(aes(mobile_coverage, life_exp), colour="red")+theme_bw()+geom_smooth(aes(mobile_coverage, life_exp), method = "lm") + xlab("Mobile subscriptions*") + ylab("Life expectancy")+annotate(geom="text", x=3.2, y=82, label=e20,color="black", size= 3)
gp21 = ggplot(countries1)+geom_point(aes(internet_users, life_exp), colour="red")+theme_bw()+geom_smooth(aes(internet_users, life_exp), method = "lm") + xlab("Internet users") + ylab("Life expectancy")+annotate(geom="text", x=25, y=82, label=e21,color="black", size= 3)
#appendix
plg1 =  plot_grid(gp1, gp2,gp3,gp4,gp5,gp6,gp7,gp8,gp9, ncol=3)
#show output 6 scatterplot
plg2 =  plot_grid(gp10, gp11,gp12,gp13,gp14,gp15,gp16,gp17,gp18, ncol=3)
#show output 6 scatterplot
plg3 =  plot_grid(gp20, gp21, ncol=3, nrow=3)
#show output 2 scatterplot

options(scipen=100)
options(digits=3)
head(countries)
#Des. stat question 1
#Adjusted Income in billion
countries$adj_income = countries$adj_income/1000000000
#PPP in million
countries$ppp = countries$ppp/1000000
#population total in million
countries$pop_total = countries$pop_total/1000000
#secondary education teachers in thousand
countries$sec_ed_teachers = countries$sec_ed_teachers/1000
#exclude Country and Code column
table=stat.desc(countries[,-c(1,2)], basic=F)
#transpose the table
table=t(table[-5,])
#extract only SD column
SD = data.frame(table[,5])
#summary for every variable in transpose
```

```r
stattable = t(summary(countries$life_exp))
stattable1 = t(summary(countries$electricity))
stattable2 = t(summary(countries$adj_income))
stattable3 = t(summary(countries$primary_school))
stattable4 = t(summary(countries$exp_primary_ed))
stattable5 = t(summary(countries$ppp))
stattable6 = t(summary(countries$mortality_water))
stattable7 = t(summary(countries$literacy_rate))
stattable8 = t(summary(countries$pop_growth))
stattable9 = t(summary(countries$pop_total))
stattable10 = t(summary(countries$primary_completion))
stattable11 = t(summary(countries$secondary_education_duration))
stattable12 = t(summary(countries$sec_ed_teachers))
stattable13 = t(summary(countries$health_exp))
stattable14 = t(summary(countries$health_exp_pc))
stattable15 = t(summary(countries$unemployment_total))
stattable16 = t(summary(countries$unemployment_youth))
stattable17 = t(summary(countries$rural_pop))
stattable18 = t(summary(countries$adolscent_fert))
stattable19 = t(summary(countries$GDP_pc))
stattable20 = t(summary(countries$mobile_coverage))
stattable21 = t(summary(countries$internet_users))
#row combine for all summary
rb = rbind(stattable, stattable1,stattable2,stattable3, stattable4,stattable5,stattable6,
stattable7,stattable8, stattable9,stattable10, stattable11,stattable12,stattable13,stattable14,
stattable15,stattable16, stattable17, stattable18, stattable19,stattable20,stattable21)
#make it to be data frame
rbdf = data.frame(rb)
#column combine with Standard Deviation
Dstattable = cbind(SD,rbdf)
#delete column 1st.qu and 3rd.qu
DFdes = Dstattable[,-c(3,6)]
#rename the columns
colnames(DFdes) = c("Standard Deviation","Min.","Median","Mean","Max.","NA.s")
#reorder the columns
DFdes1 = DFdes[,c(4,3,1,2,5,6)]
#des.stat of all variables
DFdes1
#make the table to the latex format
kable(DFdes1, format="latex", digits=2, booktabs=TRUE)
#variables which have most significant r square
countries2 = select(countries, c(life_exp,electricity,literacy_rate, primary_completion,health_exp,
    health_exp_pc,GDP_pc,
mobile_coverage,internet_users, exp_primary_ed,mortality_water, rural_pop,pop_growth,adj_income,
adolscent_fert))
#geom boxplot
attach(countries2)
bb1 = ggplot(countries2, aes(x= "",y=life_exp))+geom_boxplot(fill = "#FF6347")+ylab("Life
    expectancy")+xlab("")+
stat_boxplot(geom ='errorbar') + coord_flip() + theme_bw()
```

```
bb2 = ggplot(countries2, aes(x= "",y=electricity))+geom_boxplot(fill = "#FF6347")+ylab("Electricity")+xlab("")+
stat_boxplot(geom ='errorbar') + coord_flip() + theme_bw()
bb3 = ggplot(countries2, aes(x= "",y=literacy_rate))+geom_boxplot(fill = "#FF6347")+ylab("Adult literacy rate")+xlab("")+
stat_boxplot(geom ='errorbar') + coord_flip() + theme_bw()
bb4 = ggplot(countries2, aes(x= "",y=primary_completion))+geom_boxplot(fill = "#FF6347")+ylab("Primary completion")+xlab("")+
stat_boxplot(geom ='errorbar') + coord_flip() + theme_bw()
bb5 = ggplot(countries2, aes(x= "",y=health_exp))+geom_boxplot(fill = "#FF6347")+ylab("Health expenditure")+xlab("")+
stat_boxplot(geom ='errorbar') + coord_flip() + theme_bw()
bb6 = ggplot(countries2, aes(x= "",y=health_exp_pc))+geom_boxplot(fill = "#FF6347")+ylab("Health expenditure per capita")+xlab("")+
stat_boxplot(geom ='errorbar') + coord_flip() + theme_bw()
bb7 = ggplot(countries2, aes(x= "",y=GDP_pc))+geom_boxplot(fill = "#FF6347" , width = .3)+ylab("GDP per capita")+xlab("")+
stat_boxplot(geom ='errorbar', width = .3) + coord_flip() + theme_bw()
bb8 = ggplot(countries2, aes(x= "",y=mobile_coverage))+geom_boxplot(fill = "#FF6347")+ylab("Mobile subscriptions")+xlab("")+
stat_boxplot(geom ='errorbar') + coord_flip() + theme_bw()
bb9 = ggplot(countries2, aes(x= "",y=internet_users))+geom_boxplot(fill = "#FF6347")+ylab("Internet users")+xlab("")+
stat_boxplot(geom ='errorbar') + coord_flip() + theme_bw()
bb10 = ggplot(countries2, aes(x= "",y=exp_primary_ed))+geom_boxplot(fill = "#FF6347")+ylab("Primary education expenditure")+xlab("")+
stat_boxplot(geom ='errorbar') + coord_flip() + theme_bw()
bb11 = ggplot(countries2, aes(x= "",y=mortality_water))+geom_boxplot(fill = "#FF6347")+ylab("Unsafe water mortality")+xlab("")+
stat_boxplot(geom ='errorbar') + coord_flip() + theme_bw()
bb12 = ggplot(countries2, aes(x= "",y=rural_pop))+geom_boxplot(fill = "#FF6347")+ylab("Rural population")+xlab("")+
stat_boxplot(geom ='errorbar') + coord_flip() + theme_bw()
bb13 = ggplot(countries2, aes(x= "",y=pop_growth))+geom_boxplot(fill = "#FF6347")+ylab("Population Growth")+xlab("")+
stat_boxplot(geom ='errorbar') + coord_flip() + theme_bw()
bb14 = ggplot(countries2, aes(x= "",y=adolscent_fert))+geom_boxplot(fill = "#FF6347")+ylab("Adolescent fertility rate")+xlab("")+
stat_boxplot(geom ='errorbar') + coord_flip() + theme_bw()
#variables
bbmain = plot_grid(bb1,bb2, bb11,bb13,bb4,bb6,bb12,bb14,bb9,bb8, ncol=2)

########################
####### MA 317 #########
##### Group Project ####
####### Task 2 #########
##### Imputation#####
#######################
library(ggpubr)
library(ggplot2)
library(grid)
```

```r
library(gridExtra)
library(cowplot)
setwd('/Users/preciousakinyele/Downloads')
data=read.csv("LifeExpectancy1_log.csv", header=T)
head(data)
countries=data[1:217,]

#USING MEAN IMPUTATION
#Mean Imputation for Log(Adjusted Income)
countries$adj_income[is.na(countries$adj_income)]<- mean(countries$adj_income, na.rm=TRUE)

#Mean Imputation for Access to Electricity
countries$electricity[is.na(countries$electricity)]<- mean(countries$electricity, na.rm = TRUE)

#Mean Imputation for Population growth
countries$pop_growth[is.na(countries$pop_growth)]<- mean(countries$pop_growth, na.rm = TRUE)

#Mean Imputation for Log(Population total)
countries$pop_total[is.na(countries$pop_total)]<- mean(countries$pop_total, na.rm = TRUE)

#Mean Imputation for Primary Completion
countries$primary_completion[is.na(countries$primary_completion)]<-mean(countries$primary_completion,
↪  na.rm = TRUE)

#Mean Imputation for Log(Secondary Education Duration)
countries$secondary_education_duration[is.na(countries$secondary_education_duration)]<-
↪  mean(countries$secondary_education_duration, na.rm = TRUE)

#Mean Imputation for Health Expenditure
countries$health_exp[is.na(countries$health_exp)]<-mean(countries$health_exp, na.rm = TRUE)

#Mean Imputation for Rural Population
countries$rural_pop[is.na(countries$rural_pop)]<- mean(countries$rural_pop, na.rm = TRUE)

#Mean Imputation for Log(Mobile Cellular Subscription)
countries$mobile_coverage[is.na(countries$mobile_coverage)]<- mean(countries$mobile_coverage, na.rm
↪  = TRUE)

#Mean Imputation for Internet Users
countries$internet_users[is.na(countries$internet_users)]<- mean(countries$internet_users, na.rm =
↪  TRUE)

#USING MEDIAN IMPUTATION DUE TO A LARGE DIFFERENCE IN MEAN AND MEDIAN. RESULTING FROM OUTLIERS
#Median imputation for Children out of Primary school
sum.primary=summary((countries$primary_school))[3]
sum.primary
countries$primary_school[is.na(countries$primary_school)]<- sum.primary

#Median imputation for Log(GDP per capita)
sum.gdp=summary((countries$GDP_pc))[3]
sum.gdp
```

```
countries$GDP_pc[is.na(countries$GDP_pc)]<- sum.gdp

#Median imputation for Log(Health expenditure per capita)
sum.health_pc=summary((countries$health_exp_pc))[3]
sum.health_pc
countries$health_exp_pc[is.na(countries$health_exp_pc)]<- sum.health_pc

#Median imputation for Log(Mortality rate due to unsafe water)
sum.mortarlity_water=summary((countries$mortality_water))[3]
sum.mortarlity_water
countries$mortality_water[is.na(countries$mortality_water)]<- sum.mortarlity_water

#Median imputation for Adolescent Fertility rate
sum.adolscent_fert=summary((countries$adolscent_fert))[3]
sum.adolscent_fert
countries$adolscent_fert[is.na(countries$adolscent_fert)]<- sum.adolscent_fert

#LINEAR REGRESSION METHOD FOR DETERMINISTIC IMPUTATION
#Regression method for Primary education expenditure,
lm_exp<-
↪  (lm(countries$exp_primary_ed~countries$electricity+countries$adj_income+countries$primary_school+countries$mo
countries$health_exp_pc+countries$rural_pop+countries$adolscent_fert+countries$GDP_pc+countries$mobile_coverage+c
summary(lm_exp)
pred1 <- predict(lm_exp)
impute <- function (a, a.impute){ ifelse (is.na(a), a.impute, a)}
countries$exp_primary_ed<-impute(countries$exp_primary_ed,pred1)

#Regression method for Adult literacy rate
#LITERACY RATE HAS A LOT OF MISSING VALUES, NEVERTHELESS, HERE IS THE LINEAR REGRESSION FOR IT
lm_litrate<-(lm(countries$literacy_rate~countries$electricity+countries$adj_income+countries$primary_school+count
pred2 <- predict(lm_litrate)
countries$literacy_rate <-impute(countries$literacy_rate,pred2)


#Regression method for Log(Secondary education teachers)
lm_seced<-(lm(countries$sec_ed_teachers~countries$electricity+countries$adj_income+countries$primary_school+count
pred3 <- predict(lm_seced)
countries$sec_ed_teachers<-impute(countries$sec_ed_teachers,pred3)


#Regression method for Total Unemployment
lm_total.unemp<-(lm(countries$unemployment_total~countries$electricity+countries$adj_income+countries$primary_sch
pred4 <- predict(lm_total.unemp)
countries$unemployment_total<-impute(countries$unemployment_total,pred4)

#Regression method Total Youth Unemployment
lm_total.youth<-(lm(countries$unemployment_youth~countries$electricity+countries$adj_income+countries$primary_sch
pred5 <- predict(lm_total.youth)
countries$unemployment_youth <- impute(countries$unemployment_youth,pred5)

#creating a csv file with imputations
```

```
write.csv(countries, "imputation.csv", row.names = F)

#PDF PLOTS
#Plotting the Probability Density Function for Log (Adjusted Income) estimator variable
countries$a = countries$adj_income
countries$adj_income[is.na(countries$adj_income)]<- mean(countries$adj_income, na.rm=TRUE)
gp1 = ggplot()+geom_line(aes(x=na.omit(countries$adj_income) , y= stat(density)), stat =
↪ 'density')+theme_bw()+theme(axis.title.y = element_blank(),axis.title.x = element_text(size =
↪ 9))+ geom_line(aes(countries$a),stat = 'density', color='red')+xlab("Adjusted Income*")

#Plotting the Probability Density Function for Access to Electricity estimator variable
countries$b= countries$electricity
countries$electricity[is.na(countries$electricity)]<- mean(countries$electricity, na.rm = TRUE)
gp2 = ggplot()+geom_line(aes(x=na.omit(countries$electricity), y= stat(density)), stat =
↪ 'density')+theme_bw()+
theme(axis.title.y = element_blank(),axis.title.x = element_text(size =
↪ 9))+geom_line(aes(countries$b),stat = 'density', color='red')+xlab("Electricity")

#Plotting the Probability Density Function for Log Population growth estimator variable
countries$c = countries$pop_growth
countries$pop_growth[is.na(countries$pop_growth)]<- mean(countries$pop_growth, na.rm = TRUE)
gp3 = ggplot()+geom_line(aes(x=na.omit(countries$pop_growth), y= stat(density)), stat =
↪ 'density')+theme_bw()+
theme(axis.title.y = element_blank(),axis.title.x = element_text(size =
↪ 9))+geom_line(aes(countries$c),stat = 'density', color='red')+ xlab("Population growth")

#Plotting the Probability Density Function for Log (Total population) estimator variable
countries$d = countries$pop_total
countries$pop_total[is.na(countries$pop_total)]<- mean(countries$pop_total, na.rm = TRUE)
gp4 = ggplot()+geom_line(aes(x=na.omit(countries$pop_total), y= stat(density)), stat =
↪ 'density')+theme_bw()+
theme(axis.title.y = element_blank(),axis.title.x = element_text(size =
↪ 9))+geom_line(aes(countries$d),stat = 'density', color='red')+xlab("Population total*")

#Plotting the Probability Density Function for Primary Completion estimator variable
countries$e = countries$primary_completion
countries$primary_completion[is.na(countries$primary_completion)]<-mean(countries$primary_completion,
↪ na.rm = TRUE)
gp5 = ggplot()+geom_line(aes(x=na.omit(countries$primary_completion), y= stat(density)), stat =
↪ 'density')+theme_bw()+
theme(axis.title.y = element_blank(),axis.title.x = element_text(size =
↪ 9))+geom_line(aes(countries$e),stat = 'density', color='red')+xlab("Primary completion")

#Plotting the Probability Density Function for Log (Secondary Education) estimator variable
countries$f = countries$secondary_education_duration
countries$secondary_education_duration[is.na(countries$secondary_education_duration)]<-
↪ mean(countries$secondary_education_duration, na.rm = TRUE)
gp6 = ggplot()+geom_line(aes(x=na.omit(countries$secondary_education_duration), y= stat(density)),
↪ stat ='density')+theme_bw()+
theme(axis.title.y = element_blank(),axis.title.x = element_text(size =
↪ 9))+geom_line(aes(countries$f),stat = 'density', color='red')+xlab("Secondary education*")
```

```

#Plotting the Probability Density Function for Health Expenditure estimator variable
countries$g = countries$health_exp
countries$health_exp[is.na(countries$health_exp)]<-mean(countries$health_exp, na.rm = TRUE)
gp7 = ggplot()+geom_line(aes(x=na.omit(countries$health_exp), y= stat(density)), stat = 'density')+theme_bw()+
theme(axis.title.y = element_blank(),axis.title.x = element_text(size = 9))+geom_line(aes(countries$g),stat = 'density', color='red')+xlab("Health expenditure")

#Plotting the Probability Density Function for Rural Population estimator variable
countries$h = countries$rural_pop
countries$rural_pop[is.na(countries$rural_pop)]<- mean(countries$rural_pop, na.rm = TRUE)
gp8 = ggplot()+geom_line(aes(x=na.omit(countries$rural_pop), y= stat(density)), stat = 'density')+theme_bw()+
theme(axis.title.y = element_blank(),axis.title.x = element_text(size = 9))+geom_line(aes(countries$h),stat = 'density', color='red')+xlab("Rural population")

#Plotting the Probability Density Function for Log (Mobile subscription) estimator variable denoted as Mobile Subscription*
countries$i = countries$mobile_coverage
countries$mobile_coverage[is.na(countries$mobile_coverage)]<- mean(countries$mobile_coverage, na.rm = TRUE)
gp9 = ggplot()+geom_line(aes(x=na.omit(countries$mobile_coverage), y= stat(density)), stat = 'density')+theme_bw()+
theme(axis.title.y = element_blank(),axis.title.x = element_text(size = 9))+geom_line(aes(countries$i),stat = 'density', color='red')+xlab("Mobile subscription*")

#Plotting the Probability Density Function for Internet Users estimator variable
countries$j = countries$internet_users
countries$internet_users[is.na(countries$internet_users)]<- mean(countries$internet_users, na.rm = TRUE)
gp10 = ggplot()+geom_line(aes(x=na.omit(countries$internet_users), y= stat(density)), stat = 'density')+theme_bw()+
theme(axis.title.y = element_blank(),axis.title.x = element_text(size = 9))+geom_line(aes(countries$j),stat = 'density', color='red')+xlab("Internet users")

#Plotting the Probability Density Function for Children out of primary school estimator variable
countries$k = countries$primary_school
sum.primary=summary((countries$primary_school))[3]
countries$primary_school[is.na(countries$primary_school)]<- sum.primary
gp11 = ggplot()+geom_line(aes(x=na.omit(countries$primary_school), y= stat(density)), stat = 'density')+theme_bw()+
theme(axis.title.y = element_blank(),axis.title.x = element_text(size = 9))+geom_line(aes(countries$k),stat = 'density', color='red')+xlab("Children out of school")

#Plotting the Probability Density Function for Log (GDP per capita) estimator variable denoted as GP per capita*
countries$l = countries$GDP_pc
sum.gdp=summary((countries$GDP_pc))[3]
countries$GDP_pc[is.na(countries$GDP_pc)]<- sum.gdp
```

```
gp12 = ggplot()+geom_line(aes(x=na.omit(countries$GDP_pc), y= stat(density)), stat = 'density')+theme_bw()+
theme(axis.title.y = element_blank(),axis.title.x = element_text(size = 9))+geom_line(aes(countries$l),stat = 'density', color='red')+xlab("GDP per capita*")

#Plotting the Probability Density Function for Log (Health expenditure per capita) estimator variable denoted as Health expenditure per capita*
countries$m = countries$health_exp_pc
sum.health_pc=summary((countries$health_exp_pc))[3]
countries$health_exp_pc[is.na(countries$health_exp_pc)]<- sum.health_pc
gp13 = ggplot()+geom_line(aes(x=na.omit(countries$health_exp_pc), y= stat(density)), stat = 'density')+theme_bw()+
theme(axis.title.y = element_blank(),axis.title.x = element_text(size = 9))+geom_line(aes(countries$m),stat = 'density', color='red')+ xlab("Health expenditure per capita*")

#Plotting the Probability Density Function for Log (Mortality rate due to unsafe water) estimator variable denoted as Unsafe water mortality*
countries$n = countries$mortality_water
sum.mortarlity_water=summary((countries$mortality_water))[3]
countries$mortality_water[is.na(countries$mortality_water)]<- sum.mortarlity_water
gp14 = ggplot()+geom_line(aes(x=na.omit(countries$mortality_water), y= stat(density)), stat = 'density')+theme_bw()+
theme(axis.title.y = element_blank(),axis.title.x = element_text(size = 9))+geom_line(aes(countries$n),stat = 'density', color='red')+xlab("Unsafe water mortality*")

#Plotting the Probability Density Function for Adolescent Fertility rate estimator variable
countries$o = countries$adolscent_fert
sum.adolscent_fert=summary((countries$adolscent_fert))[3]
countries$adolscent_fert[is.na(countries$adolscent_fert)]<- sum.adolscent_fert
gp15 = ggplot()+geom_line(aes(x=na.omit(countries$adolscent_fert), y= stat(density)), stat = 'density')+theme_bw()+
theme(axis.title.y = element_blank(),axis.title.x = element_text(size = 9))+geom_line(aes(countries$o),stat = 'density', color='red')+xlab("Adolescent fertilty rate ")

#Plotting the Probability Density Function for Primary education expenditure estimator variable
countries$p = countries$exp_primary_ed
lm_exp<-(lm(countries$exp_primary_ed~countries$electricity+countries$adj_income+countries$primary_school+countries$mo
pred1 <- predict(lm_exp)
impute <- function (a, a.impute){ ifelse (is.na(a), a.impute, a)}
countries$exp_primary_ed <-+ impute(countries$exp_primary_ed,pred1)
gp16 = ggplot()+geom_line(aes(x=na.omit(countries$exp_primary_ed), y= stat(density)), stat = 'density')+theme_bw()+
theme(axis.title.y = element_blank(),axis.title.x = element_text(size = 9))+geom_line(aes(countries$p),stat = 'density', color='red')+xlab("Primary education expenditure")

#Plotting the Probability Density Function for Adult literacy rate estimator variable
countries$q = countries$literacy_rate
lm_litrate<-(lm(countries$literacy_rate~countries$electricity+countries$adj_income+countries$primary_school+count
```

```
pred2 <- predict(lm_litrate)
impute <- function (a, a.impute){ ifelse (is.na(a), a.impute, a)}
countries$literacy_rate<-+impute(countries$literacy_rate,pred2)

gp17 = ggplot()+geom_line(aes(x=na.omit(countries$literacy_rate), y= stat(density)), stat =
↪ 'density')+theme_bw()+
theme(axis.title.y = element_blank(),axis.title.x = element_text(size =
↪ 9))+geom_line(aes(countries$q),stat = 'density', color='red')+xlab("Adult literacy rate")

#Plotting the Probability Density Function for Log (secondary education teachers) estimator variable
↪ denoted as Secondary education teachers*
countries$r = countries$sec_ed_teachers
lm_seced<-(lm(countries$sec_ed_teachers~countries$electricity+countries$adj_income+countries$primary_school+count
pred3 <- predict(lm_seced)
impute <- function (a, a.impute){ ifelse (is.na(a), a.impute, a)}
countries$sec_ed_teachers<-+impute(countries$sec_ed_teachers,pred3)
gp18 = ggplot()+geom_line(aes(x=na.omit(countries$sec_ed_teachers), y= stat(density)), stat =
↪ 'density')+theme_bw()+
theme(axis.title.y = element_blank(),axis.title.x = element_text(size =
↪ 9))+geom_line(aes(countries$r),stat = 'density', color='red')+xlab("Secondary education
↪ teachers*")

#Plotting the Probability Density Function for Total Unemployment estimator variable
countries$s = countries$unemployment_total
lm_total.unemp<-(lm(countries$unemployment_total~countries$electricity+countries$adj_income+countries$primary_sch
pred4 <- predict(lm_total.unemp)
impute <- function (a, a.impute){ ifelse (is.na(a), a.impute, a)}
countries$unemployment_total<-+impute(countries$unemployment_total,pred4)
gp19 = ggplot()+geom_line(aes(x=na.omit(countries$unemployment_total), y= stat(density)), stat =
↪ 'density')+theme_bw()+
theme(axis.title.y = element_blank(),axis.title.x = element_text(size = 9))+
↪ geom_line(aes(countries$s),stat = 'density', color='red')+xlab("Unemployment")

#Plotting the Probability Density Function for Total Youth Unemployment estimator variable
countries$t = countries$unemployment_youth
lm_youth.total<-(lm(countries$unemployment_youth~countries$electricity+countries$adj_income+countries$primary_sch
pred5 <- predict(lm_youth.total)
impute <- function (a, a.impute){ ifelse (is.na(a), a.impute, a)}
countries$unemployment_youth <-+ impute(countries$unemployment_youth,pred5)
gp20 = ggplot()+geom_line(aes(x=na.omit(countries$unemployment_youth), y= stat(density)), stat =
↪ 'density')+theme_bw()+
theme(axis.title.y = element_blank(),axis.title.x = element_text(size = 9)) +
↪ geom_line(aes(countries$t),stat = 'density', color='red')+xlab("Youth unemployment")
gp20

#A PDF plot of all the estimator variables from gp1 to gp9 agaisnt density in grids
rplots1= plot_grid(gp1, gp2, gp3, gp4, gp5, gp6, gp7, gp8, gp9, ncol = 3)
rplots1

#A PDF plot of all the estimator variables from gp10 to gp18 agaisnt density in grids
rplots2= plot_grid(gp10, gp11, gp12, gp13, gp14, gp15, gp16, gp17, gp18, ncol=3)
```

```r
rplots2

#A PDF plot of all the estimator variables from gp19 to gp20 agaisnt density in grids
rplots3= plot_grid(gp19, gp20, ncol=3, nrow =3)
rplots3

#MISSING DATA IMPUTATION METHODS JUSTIFICATION

#reload the csv file for justification
data=read.csv("LifeExpectancy1_log.csv", header=T)

#To check if there are any countries with complete estimator variables.
countries %>% filter(complete.cases(.))

#to check the countries with life expectancies before carrying out the complete case
countries$life_exp #211

#Since there are no countries with complete estimator variables, therefore, conditions for complete
#cases were relaxed. We then assumed a complete case by
#removing ppp, literacy rate and Primary education expenditure
b2= countries %>% filter(c(!is.na(countries$electricity)& !is.na(countries$adj_income)& !is.na(countries$primary_school)& !is.na(countries$pop_growth)& !is.na(countries$pop_total)&
!is.na(countries$primary_completion)&!is.na(countries$secondary_education_duration)&!is.na(countries$sec_ed_teach !is.na(countries$adolscent_fert)&!is.na(countries$unemployment_youth)&!is.na(countries$mortality_water)&!is.r !is.na(countries$mobile_coverage)&!is.na(countries$internet_users)))

#to check number of countries that have the assumed complete case.
count(b2)

#to check number of countries with life expectancy under the assumed complete case.
b2$life_exp #48

#to plot the graph of life expectancy under the assumed complete case, as well as life expectancy
#having all the predictor variables.
ggplot()+geom_line(data=b2,aes(x= b2$life_exp), stat = 'density')+theme_bw()+
geom_line(data=countries,aes(x= countries$life_exp), stat = 'density', color='red')+xlab('Life Expectancy')
aspect_ratio<-2
height <-7
ggsave( height = 7, width = 7*aspect_ratio, "completecase.pdf")


######################
####### MA 317 #########
##### Group Project ####
####### Task 4 #########
##### Linear Model#####
#####################

#Cleaning workingspace
rm(list=ls())
```

```r

#loading libraries
library(data.table)
library(ggplot2)
library(ggpubr)
library(stargazer)
library(quantreg)
library(kableExtra)


#Setting wd
wd=setwd("D:/Documentos/Essex/Modelling Experimental Data/Group Project")
countries=read.csv("imputation.csv", header=T) #Read file
head(countries)   #check if it's the correct data
attach(countries) #attach variables to ws

#4. Suggest a model which explains life expectancies in 2016. Justify you answer. Can this model be used
#to predict life expectancies of other countries which have not provided in 2016 data on life expectancy?
#adding a variable GINI (Income inequality)
gini=fread("WDIData.csv", header=T)  #read data
gini=gini[gini$`Indicator Name`=="GINI index (World Bank estimate)",c(1,2,61)]  #keeping just year of interest

countries=merge(countries, gini[,c(1,3)], by.x="country", by.y="Country Name")  #merging with country data
names(countries)[names(countries)=="2016"]="gini"
attach(countries) #attach variables to ws
rm(gini)

#we analyse the gini data
summary(gini)

#perform regression imputation
model_gini=summary(lm(gini~countries$electricity+countries$adj_income+countries$primary_school+countries$mortality_water+countries$pop_growth+countries$pop_total+ countries$primary_completion+countries$secondary_education_duration+countries$health_exp+countries$health_exp_pc+ countries$rural_pop+countries$adolscent_fert+countries$GDP_pc+countries$mobile_coverage+countries$internet_users))

X=matrix(1,nrow(countries))

X=cbind(X, countries$electricity,countries$adj_income,countries$primary_school,countries$mortality_water,countries$pop_growth, countries$pop_total, countries$primary_completion, countries$secondary_education_duration, countries$health_exp, countries$health_exp_pc, countries$rural_pop,countries$adolscent_fert, countries$GDP_pc, countries$mobile_coverage, countries$internet_users)

countries$gini_i=X%*%as.matrix(model_gini$coefficients[,1])

```

```r
#impute <- function (a, a.impute){ ifelse (is.na(a), a.impute, a)}
#countries$gini_i <- impute(countries$gini,model_gini)
attach(countries) #attach variables to ws

#summarizing the new imputed var
summary(countries$gini_i)
#comparing density function before &after imputation
countries$gini_i=ifelse(countries$gini_i<25,mean(countries$gini_i),countries$gini_i) #min value
↪  before imp

ggplot()+geom_line(data=countries, aes(gini), stat="density")+theme_bw()+
↪  geom_line(data=as.data.frame(gini_i), aes(gini_i), stat="density", color="red")

#plotting correlation with life expectancy
gp1 = ggplot(countries)+geom_point(aes(log(gini_i), life_exp), colour="red")+
↪  theme_bw()+geom_smooth(aes(log(gini_i), life_exp), method = "lm") + xlab("log(Gini)") +
↪  ylab("Life expectancy")+stat_cor(aes(log(gini_i), life_exp,label = ..r.label..),method =
↪  "pearson", label.x = 3.97, label.y = 80)

gp1
ggsave("gini.pdf")
countries$gini=NULL

#Defining the saturated model with only those variables with high correlation (meausred
#by the correlation coefficient) to Life Expectancy
    #Y: Life Expectancy
    #matrix X: Electricity, water, mortality,  population growth, adjusted income
    #primary completion rate, health expenditure pc
    #and squared (see scatterplot), rural population, adolescent fertilty rate, internet users,
    #mobile coverage

sub=countries[,-c(1,2,8,26)]
corr <- round(cor(na.omit(sub)), 1)
corr=melt(corr)

n=rev(list("Life Expectancy","Electricity","Adjusted Income", "Primary School", "Expenditure
↪  Primary","Water Mortality","Literacy Rate", "Population Growth", "Total Population", "Primary
↪  Completion", "Secondary Education", "Secondary Teachers", "Health Exp.","Health Exp. pc",
↪  "Unemployment", "Youth Unemploymnent", "Rural Opulation", "Adolescent Fertility", "GDP pc",
↪  "Mobile Coverage", "Internet Users", "Gini"))

ggplot(data = corr, aes(x=Var1,ordered(Var2, levels =     rev(sort(unique(Var2)))), fill=value)) +
geom_tile()+ theme(axis.title.x = element_blank(),axis.text.y = element_text(vjust=0),
↪  axis.title.y = element_blank(),axis.text.x=element_blank(), legend.title = element_blank()) +
↪  scale_fill_gradient2(low="red", mid="yellow", high="red",
↪  midpoint=0,labels=c("-1.0", "-0.5","0","0.5","1.0"),breaks=c(-1,-.5,0,.5,1),limits=c(-1,1))+
scale_y_discrete(labels=n) ggsave("matrix_corr.pdf")

#From section 1 we know there is a non-linear correlation between life expectancy and
# health exp pc in levels. We add an extra squared term to capture this relation.

```

```r
countries$health_pc_levels=exp(health_exp_pc)/1000
attach(countries)

#We perform stepwise variable selection with both AIC and BIC

best1=step(lm(life_exp~electricity +mortality_water+ pop_growth+primary_completion+health_pc_levels +
I(health_pc_levels^2)+ rural_pop+adolscent_fert +internet_users+mobile_coverage+adj_income), direction = "both")

aic1=best1$anova$AIC[length(best1$anova$AIC)]
extractAIC(best1, scale=0, k=log(nrow(countries)))

best1_1=step(lm(life_exp~electricity+ mortality_water+ pop_growth+primary_completion+health_pc_levels +
I(health_pc_levels^2)+ rural_pop +adolscent_fert+internet_users+mobile_coverage+adj_income), direction = "both", k = log(nrow(countries)))

bic1=best1_1$anova$AIC[length(best1_1$anova$AIC)]

best2=step(lm(life_exp~electricity+ pop_growth+primary_completion+health_pc_levels+ I(health_pc_levels^2)+rural_pop+adolscent_fert+ internet_users+mobile_coverage+ adj_income), direction = "both")

aic2=best2$anova$AIC[length(best2$anova$AIC)]
extractAIC(best2, scale=0, k=log(nrow(countries)))

best2_2=step(lm(life_exp~electricity+ pop_growth+primary_completion+health_pc_levels+
I(health_pc_levels^2)+ rural_pop+adolscent_fert+ internet_users+mobile_coverage+adj_income), direction = "both", k = log(nrow(countries)))

bic2=best2_2$anova$AIC[length(best2_2$anova$AIC)]
extractAIC(best2_2, scale=0)

######################
#Summarize best models
#####################

diagPlot<-function(model){
  p1<-ggplot(model, aes(.fitted, .resid))+geom_point()
  p1<-p1+stat_smooth(method="loess")+geom_hline(yintercept=0, col="red", linetype="dashed")
  p1<-p1+xlab("Fitted values")+ylab("Residuals")
  p1<-p1+ggtitle("Residual vs Fitted Plot")+theme_bw()+theme(title = element_text(size=9))

  p2<-ggplot(model, aes(qqnorm(.stdresid)[[1]], .stdresid))+geom_point(na.rm = TRUE)
  p2<-p2+geom_abline()+xlab("Theoretical Quantiles")+ylab("Standardized Residuals")
  p2<-p2+ggtitle("Normal Q-Q")+theme_bw()+theme(title = element_text(size=9))

  p3<-ggplot(model, aes(.fitted, sqrt(abs(.stdresid))))+geom_point(na.rm=TRUE)
  p3<-p3+stat_smooth(method="loess", na.rm = TRUE)+xlab("Fitted Value")
  p3<-p3+ylab(expression(sqrt("|Standardized residuals|")))
```

```
  p3<-p3+ggtitle("Scale-Location")+theme_bw()+theme(title = element_text(size=9))
  p3
  return(ggarrange(p1, p2, p3, ncol=3, nrow=2))
}

#############
#####Model 1
#############
summary(best1)

#diagnostic plots
diagPlot(best1)
ggsave("best1.pdf")

#############
#####Model 2
#############
summary(best1_1)

#diagnostic plots
diagPlot(best1_1)
ggsave("best1_1.pdf")

#############
#####Model 3
#############
summary(best2)

#diagnostic plots
diagPlot(best2)
ggsave("best2.pdf")

############
#####Model 4
#############
summary(best2_2)

#diagnostic plots
diagPlot(best2_2)
ggsave("best2_2.pdf")


#####################
#Model with GDP
#####################
#we will test a model using GDP as explanatory variable and
#a additional covariates, namely: GINI, and Continent indicator
continets=read.csv("continents.csv", header=T)  #load continents names
continents2=read.csv("WDICountry.csv", header=T)  #different classification
continents2=continents2[,c(1,8)] #select coiumn1 and column8 on the table

```

```r
#rename variables
names(continets)[3]="code"
names(continets)[1]="cont"
names(continents2)[1]="code"

#adding continent columns to data frame
countries<-merge(countries,continets[,c(1,3)],by="code") #merge Continent
countries<-merge(countries,continents2,by="code")
attach(countries)


countries$cont <- as.character(countries$cont) #change factor to character
countries$Region <- as.character(countries$Region)  #change factor to character

#renaming continents
countries$Region[countries$Region=="Latin America & Caribbean"]="South America"
#replace Americas to South and North Americas
countries$cont <- ifelse(countries$cont =="Americas",countries$Region,countries$cont)
countries$cont[countries$country=="Greenland"]="Europe"

#using "geographic north" criteria
centro=list("Guatemala", "Mexico", "Belize", "Honduras", "El Salvador", "Costa Rica",
    "Panama","Nicaragua")

for (c in centro){
  countries$cont[countries$country==c]="North America"
}

countries$Region=NULL   #deleting unused variables

#c=unique(cont)

#for (v in c) {
 # countries[,v]=0
  #countries[v]=ifelse(countries["cont"]==v,1,0)
#}

attach(countries)

####MODEL

model_gdp=step(lm(log(life_exp)~GDP_pc+gini_i+cont), direction = "both",
    k=log(length(na.omit(life_exp))))

#diagnostic
summary(model_gdp)

diagPlot(model_gdp)

ggsave("gdp.pdf")

```

```
#scatterplot by continent
ggplot(countries)+geom_point(aes(GDP_pc, life_exp, color=cont, size=exp(pop_total)^0.5))+theme_bw()+
xlab("log(GDP pc)")+ylab("Life Expectancy")+ guides(size = FALSE)+theme(legend.position =
↪  c(0.85,0.25),legend.title=element_blank(),legend.background=element_blank(),
↪  legend.key=element_blank(),)+scale_colour_manual(values = c("#DBD522", "#FB0101", "#20CE00",
↪  "#00DCD3","#0013DC","#D600DC"))

ggsave("scattercont.pdf")  #ggplot is saved

###################################
##Table with all models
##################################

models=list(best1,best2,best2_2,model_gdp)
stargazer(models, align=T, keep.stat = c("n", "rsq", "adj.rsq", "aic","bic"),no.space=T,
dep.var.labels = c("Life Expectancy", "log(Life Expectancy"), covariate.labels = c("Electricity",
↪  "log(Water Mort.)", "Health Exp. pc", "Health Exp. pc", "Adolscent Fertility", "Internet
↪  Users","Mobile Coverage", "GDP pc", "Gini Index", "Asia", "Europe", "North America", "Oceania",
↪  "South America", "Constant"))

######################
####### MA 317 #########
##### Group Project ####
####### Task 5 #########
#####ANOVA#####
######################

setwd("/Users/weiyu/Desktop/R")
Life_E=read.csv("imputation.csv", header=T)
names(Life_E)[names(Life_E)=="country"]="Country.Name" #change the column name to Country.Name
names(Life_E)[names(Life_E)=="code"]="Country.Code" #change the column name to Country.Code
Continets=read.csv("continents.csv", header=T)
names(Continets)[names(Continets)=="ISO.alpha3.Code"]="Country.Code" #change the column name
data2=read.csv("WDICountry.csv", header=T)
data2=data2[,c(1,8)] #select coiumn1 and column8 on the table
head(data)
library(ggplot2)
library(patecs)
library(dobson)
library(knitr)
library(kableExtra)

countries=Life_E[1:217,] #select coiumn 1 and column 217 on the table

countries<-merge(countries,Continets,by="Country.Code") #merge Continent
countries<-merge(countries,data2,by="Country.Code")  #merge Continent from WDICountry

countries1<-countries[,c(1,2,3,25,28)] #select column 1 to column 3, column 25 and 28
names(countries1)[names(countries1)=="Region.Name"]="Continent"

countries1$Continent <- as.character(countries1$Continent) #change factor to character
```

```r
countries1$Region <- as.character(countries1$Region)

countries1$Region[countries1$Region=="Latin America & Caribbean"]="South America" #change Latin America to South America
countries1$Continent <- ifelse(countries1$Continent =="Americas",countries1$Region,countries1$Continent) #replace Americas to South and North Americas

mod1=glm(life_exp~1, family = "gaussian",data=countries1)

(aov(countries1$life_exp~countries1$Continent))
summary(aov(countries1$life_exp~countries1$Continent)) #produce an ANOVA Test table
#make box plot
box1 <- ggplot(data = countries1, aes(x=countries1$Continent, y=countries1$life_exp,fill=Continent))+
    geom_boxplot()+coord_flip()+theme_bw()
  labs(x="Continents", y="Life Expectancy", title="Continents by Life expectancy Box plot ")

plot(aov(countries1$life_exp~countries1$Continent)) # test

bartlett.test(countries1$life_exp~countries1$Continent, countries1) #Run Bartlett test

#combine residual vs Fitted Plot,Normal Q-Q and Scale location graphs to one PDF file.
library(ggpubr)

diagPlot<-function(model){
  p1<-ggplot(model, aes(.fitted, .resid))+geom_point()
  p1<-p1+stat_smooth(method="loess")+geom_hline(yintercept=0, col="red", linetype="dashed")
  p1<-p1+xlab("Fitted values")+ylab("Residuals")
  p1<-p1+ggtitle("Residual vs Fitted Plot")+theme_bw()+theme(title = element_text(size=9))

  p2<-ggplot(model, aes(qqnorm(.stdresid)[[1]], .stdresid))+geom_point(na.rm = TRUE)
  p2<-p2+geom_abline()+xlab("Theoretical Quantiles")+ylab("Standardized Residuals")
  p2<-p2+ggtitle("Normal Q-Q")+theme_bw()+theme(title = element_text(size=9))

  p3<-ggplot(model, aes(.fitted, sqrt(abs(.stdresid))))+geom_point(na.rm=TRUE)
  p3<-p3+stat_smooth(method="loess", na.rm = TRUE)+xlab("Fitted Value")
  p3<-p3+ylab(expression(sqrt("|Standardized residuals|")))
  p3<-p3+ggtitle("Scale-Location")+theme_bw()+theme(title = element_text(size=9))
  p3
  return(ggarrange(p1, p2, p3, ncol=3, nrow=2))
}
diagPlot(an)
ggsave("anovaresults.pdf")
```