University of Washington
Department of Industrial and Systems Engineering
CET 521/IND E 546

FINAL PROJECT

SIMAN NING
ROBIN HALFVORDSSON
CHRISTOPHER SALAZAR

03-18-2020

Contents

1. Abstract

Life expectancy varies across countries and it is speculated what factors that contribute to the discrepancy. This research takes a global scope of 13 countries including developed countries in Europe, Asia, North America and South America to understand how economy, urbanization and environmental factors might interact with life expectancy. World Bank data used as data source, for its authority, cleanliness and efficient data extraction tools. Data availability for certain variables was also a main criteria for the selection of countries. Linear regression, logistic and multinomial logistic models are implemented to analyze the economy, urbanization and environmental factors. The results suggests that economy variables are generally positively related to life expectancy (GDP per capita and exported good as a percentage of GDP). This means wealthier countries and countries that have more international trades are likely to have longer life expectancy. As for rural development, more arable land per capita for a country seems to indicate a shorter life expectancy. However, our result is only limited to countries that may not have serious food crisis. For environmental indicator, the renewable energy ratio has a negative impact on life expectancy, this result might need further investigation on the countries we select. Public health indicators like immunization and physicians density also has a positive relation with life expectancy. Overall, the policy implication of our research is to promote economy, urbanization and public health to improve life expectancy.

2. Introduction

What should countries do to improve the citizen's life expectancy? It is noticed that life expectancy varies across countries. What could explain for the disparity? Indicators such as access to healthcare, exercise, crime rates, and hygiene all make intuitive sense and have evidence based studies that influence the life expectancy of a population. Although these indicators are well known, some initial exploratory analysis on other less evident factors revealed some results worth exploring. This project seeks to explore several indicators(economy, urbanization and environment) that affect life expectancy on a global scale and within certain groupings of countries. This section will introduce the data source and some general modeling techniques that will be used for this study.

2.1 Data Source

Given the global scope of this study, data will be gather from the world bank data organization (https://www.worldbank.org/). This data repository has a vast amount of indicators, including life expectancy, across several years for most countries in the world. The size of this data set will enable exploration of various predictors as different models are developed to explain certain aspects of life expectancy. A clear advantage of using this data set is the R package that has been developed to extract, filter, and organize world bank data sets that the user specifies. This will reduce the amount of data wrangling required prior to enacted modeling.

2.2 Modeling Techniques

Linear, binary logistic and multinomial logistic regression will be the main modeling tools used for this study, but we will also recommend other possible analysis techniques that could be used for future work related to this project. Linear regression is used as a natural choice of the continuous dependent variable. Binary logistic model is used to understand what factors will make life expectancy exceed the mean life expectancy of countries (74 years). Multinomial logistic model is chosen to understand the categorization of countries.

3. Exploratory Analysis

The study begins by exploring the life expectancy variable to see how it can be explained by regression models. The variables and countries that are well suited for this study will be discussed in this section.

3.1 Life Expectancy Response

The life expectancy variable will be the primary response. In general, it is known that global life expectancy has increased significantly within the last 50 years. Particularly, global life expectancy reached 68 years between 2005-2010, a 21-year rise since 1950- 1955, according to a United Nations report [5]. This rising trend can be verified by Figure 3.1, which shows a time-series scatter plot of the average life expectancy over time for specific countries which will be discussed in the subsequent section.

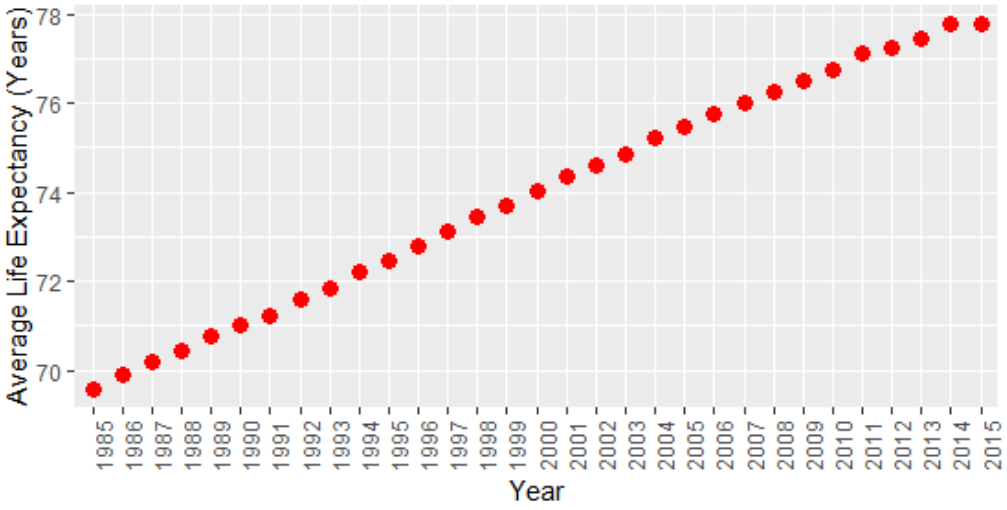

Figure 3.1: Average Life Expectancy

As can been seen, life expectancy has been linearly increasing from 1985 to 2015 for the countries that will in this study. The distribution of life expectancy can also viewed via a histogram with Figure 3.2.

Figure 3.2: Distribution of Life Expectancy

The shape of the histogram appears to have a normal distribution with a mean of approximately 75 years old, with a slight skew to the right. The left tail indicates that the variance may be large around the mean. This indicates that the the collected data for life expectancy is reasonably balanced and may uncover certain characteristics among several variables.

These simple figures omit many details that the report aims to uncover. However, they give a glimpse on how life expectancy behaves across 30 years of data collection. These initial inferences will be used to discuss the rationale for which countries are included in this study.

3.2 Countries

Selecting the countries for this study is based on finding a reasonable distribution of life expectancy within countries. One way this can be handled is by graphing box-plots for various countries. Figure 3.3 shows a violin plot superimposed with boxplots for each country that will be used in this study.

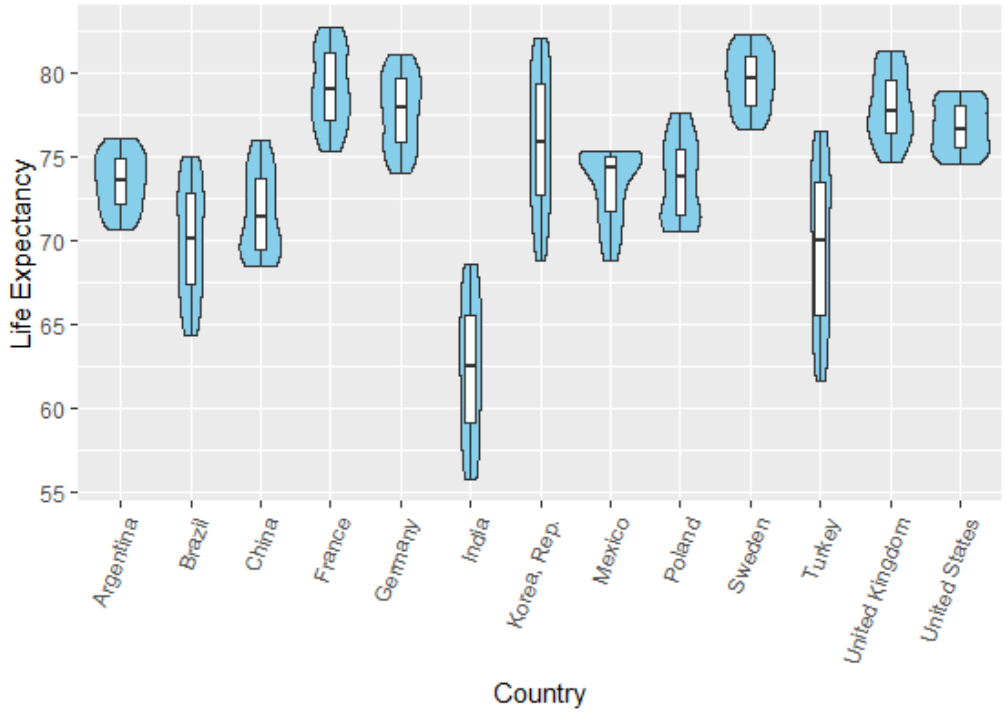

Figure 3.3: Violin and Boxplots for each Country of Life Expectancy

It should be noted that several iterations for different countries were conducted to ultimately these produce a reasonably balanced data set. That is, countries like Sweden, Germany, and France exhibit life expectancy with means of around 77-80 years old. Other countries show life expectancy of around 63-70 years old like India, Turkey and Brazil. This figure also shows the magnitude of variability for each country with some exhibiting high spreads (i.e. India, Turkey, South Korea) and others with tighter variances (i.e. Sweden, USA, Argentina). Additionally, the overlaid violin plots show the kernal density distribution for each country. For example, Mexico shows a skew that is weighted towards the higher range of life expectancy while the USA appears to show a more uniform distribution.

Figure 3.1's time series scatter plot can be broken down by country which is the result of Figure 3.4.

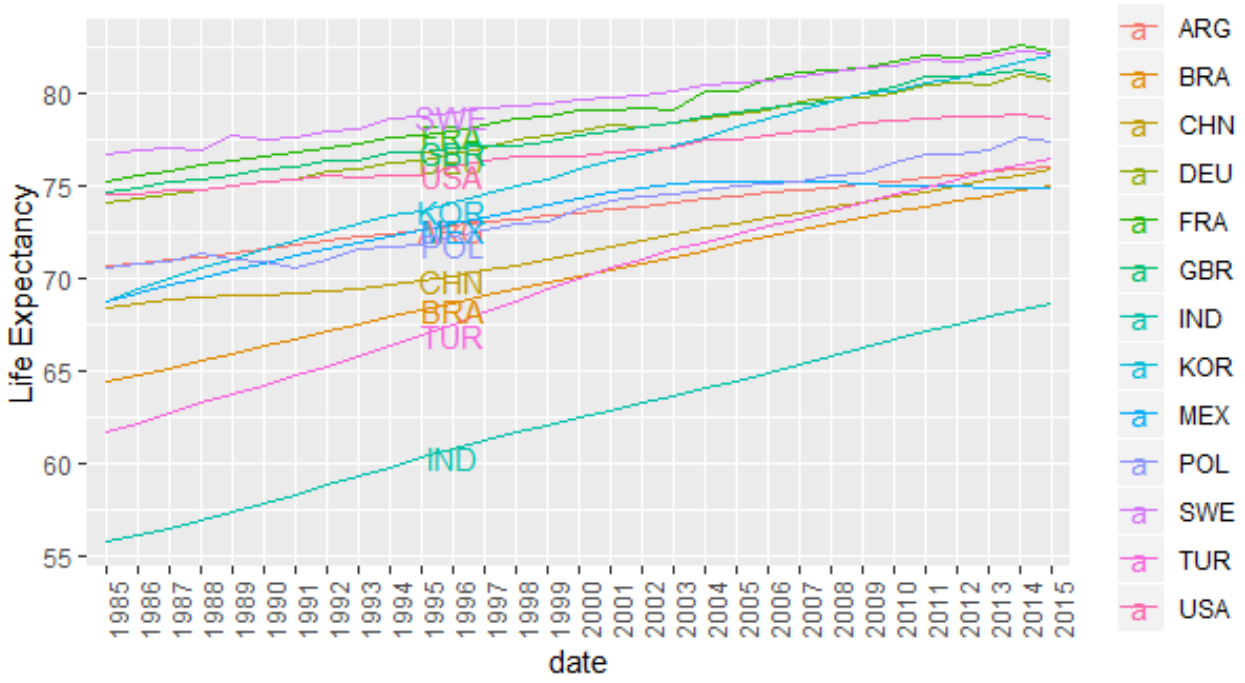

Figure 3.4: Scatter plot by Country (Source: World Bank Data)

There are a few interesting trends that can be observed from the plot. First, the relative high life expectancy countries (Sweden, France, Germany) exhibit a relatively slow growth over the 30 year span. Whereas countries like Brazil, Turkey and India show a low starting life expectancy and has grown quite quickly over this period of time. There are also the other countries that have some combination of these extremities.

3.3 Variables/Predictors

Prior to exploring what predictors to include in this study, a brief literature review was conducted to help guide the scope of the project. The variables are separated in to three major categories: Economic growth, Urbanization, and Environmental factors.

3.3.1 Economic Growth

Economic growth has some clear influence on life expectancy in countries. In fact, in a study of China from 1991-2012 that looked at mortality rates, income and exposure to air pollution, there was a strong positive relationship between GDP and life expectancy[3]. Particularly, it was found that this positive relationship was mostly due to decreased deaths at birth. Acknowledging this trend, this study will build on the notion that economic growth variables will affect countries differently and examine if similar tendencies occur.

3.3.2 Urbanization

Urbanization is a vast definition when characterizing it towards each country. Simply put, urbanization is a direct contrast rural living and the characteristics associated with it. For example, children have a higher chance of surviving to adulthood in urban areas, but the potential benefits of urbanization have been unevenly

exploited around the world [1]. The urban environment favors many of its inhabitants, but especially the rich and not always the poor[1]. These excerpts give a basis for choosing certain urbanization variables and how these studies can be expanded.

3.3.3 Environmental Factors

Environmental factors might have a more intuitive relationship with life expectancy. This is verified through a recent US study that reports strong evidence of an association between recent further reductions in fine-particulate air pollution and improvements in life expectancy in the United States, especially in densely populated urban areas[2]. This major indicator lends itself as a baseline to compare to other countries and how their life expectancy metric is affected.

3.3.4 Predictors

Upon conducting the brief aforementioned research, the following table summarizes a majority of the variables that we will be explored. Other additional variables will be introduced in a later section.

Variable	Description
Export of goods (% of GDP)	Exports of goods and services represent the value of all goods and other market services provided to the rest of the world.
GDP per Capita	GDP per capita is gross domestic product divided by midyear population.
Population growth (annual %)	Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship.
Physicians (per 1,000 people)	Physicians include generalist and specialist medical practitioners.
Total Fertility Rate (births per woman)	Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year.
Arable land (hectares per person)	Arable land (hectares per person) includes land defined by the FAO as land under temporary crops, temporary meadows for mowing or for pasture, land under market or kitchen gardens, and land temporarily fallow.
Immunization of people DPT % of the population	Child immunization, DPT, measures the percentage of children ages 12-23 months who received DPT vaccinations before 12 months or at any time before the survey.
CO2 emission metric tone per capita	Carbon dioxide emissions are those stemming from the burning of fossil fuels and the manufacture of cement.

It should be noted that many other variables were considered, however, upon several modeling iterations, this list was reduced to prevent multicollinearity among the independent variables. These variables can be recovered from the world bank data with the following link: https://data.worldbank.org/.

4. Linear Regression

Since life expectancy is a continuous variable, performing linear regression models is a natural candidate for the chosen predictive variables. Two models are present in the this section; a grand linear regression model and a truncated linear regression.

4.1 1st Model

Given the macroscopic nature of this study, the first model aims to discover which indicators influence life expectancy on a global scale. A few front end matters require a review of the independent variables. Upon plotting several histograms of each predictor, two variables required a transformation to reflect a favorable normal distribution. Figure 4.1 shows the transformation of GDP per capita variable.

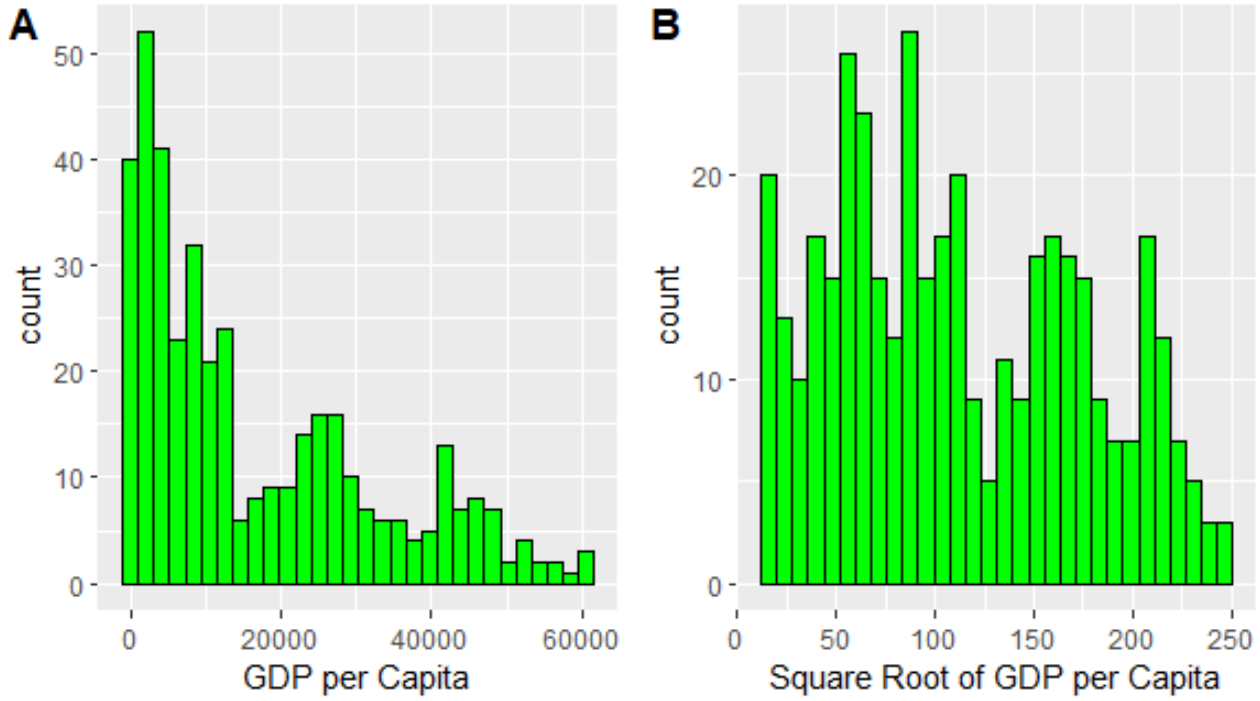

Figure 4.1: GDP per Capita Transformation

Although this is not a perfect bell curve shape, the square root of GDP per Capita provides a transformation variable that reflects a normal distribution. This ensures that the ordinary least squares assumption for model adequacy is not violated.

The other variable that required a transformation was the arable land per capita. This can be seen in Figure 4.2 which follows the same procedure as GDP per Capita.

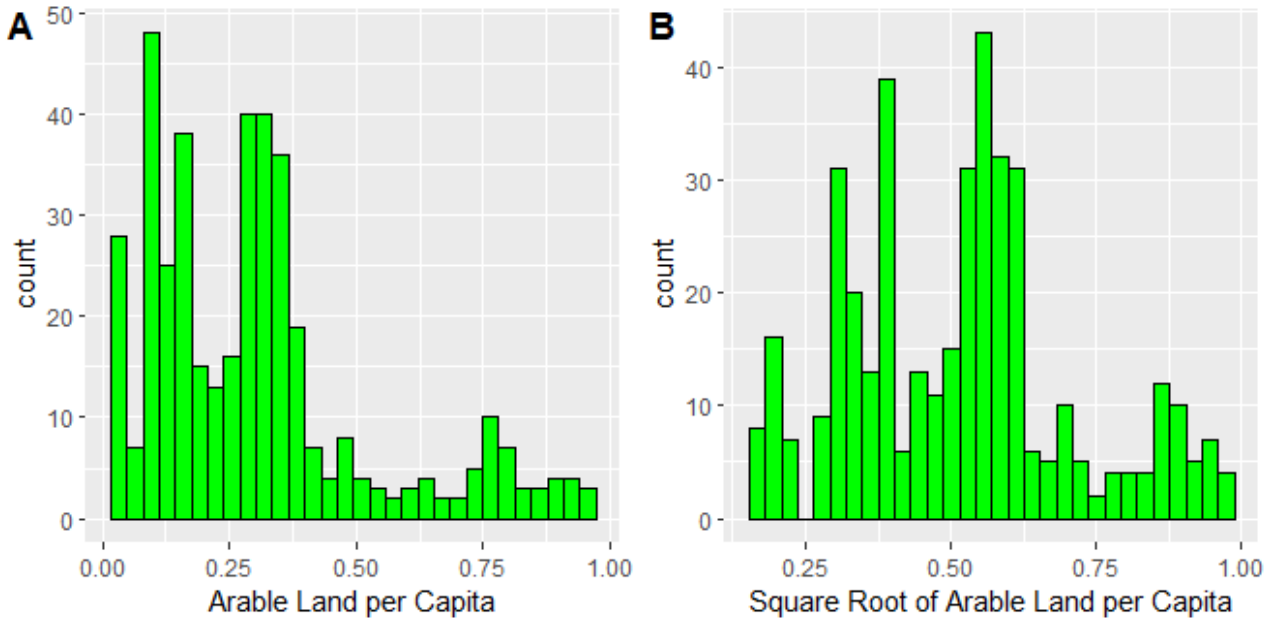

Figure 4.2: Arable Land per Capita Transformation

Upon the transformation of these variables, the first linear regression model can be implemented. The table below shows the results.

Table 4.1: Linear Regression #1 Model Results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	59.71	1.10	54.21	0.00
Exported Goods (% of GDP)	0.058	0.012	5.00	0.00
Square Root of GDP per Capita	0.035	0.003	12.05	0.00
Population Growth (%)	0.600	0.267	2.25	0.025
Physicians per 1,000 people	1.110	0.197	5.63	0.00
Fertility Rate	-1.48	0.300	-4.93	0.00
Square Root of Arable land per Capita	-1.84	0.711	-2.58	0.010
Immunization	0.110	0.009	11.91	0.00
CO2 emission per capita	0.056	0.029	1.92	0.056
Adj $R^2 = 0.925$				
N = 403				

The results of this model will be examined in the discussion section.

4.1.1 Model Adequacy

The validity of this model depends on how well it adheres to an ordinary least squares assumption. First, an examination of the residual plots are discussed below.

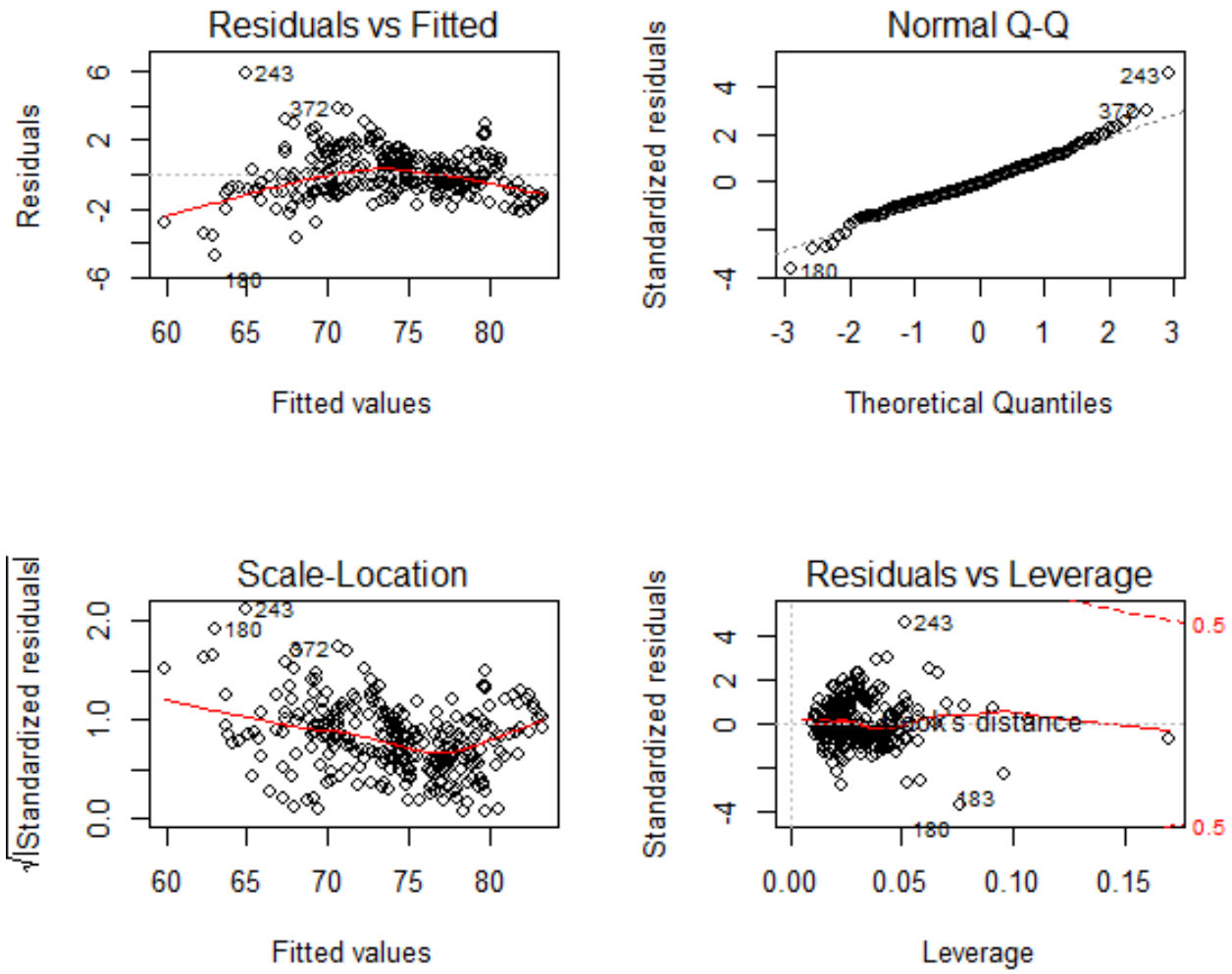

Figure 4.3: Residual Plots of Linear Regression Model #1

Figure 4.3 displays four subplots to exam the model adequacy of the first linear regression model. First, it is seen that residuals plots exhibit a slight trend between the 60-70 years old fitted values. While this is not desirable, this may be explained by the variability of life expectancy as shown in Figure 3.2. The histogram shows a long 'left' tail which may contribute to the trend in the residual plot. The QQ plot shows a favorable trend as most standardized residual values fall along the linear comparison of a theoretical normal distribution. Thus, it can be stated that the errors follow a normal distribution. Lastly, it can be seen that there are a few highlighted outliers. However, given that this model utilizes over 400 observations. These outliers, which are not outside Cook's distance will not likely affect the results of the model.

The model adequacy can be further examined by considering the variance inflation factors. This, along a with confidence intervals are provided in the subsequent table.

Table 4.2: Linear Regression #1 Confidence Int. and VIF Results

	2.5 %	97.5 %	VIF
(Intercept)	57.54	61.87	
Exported Goods (% of GDP)	0.03	0.08	2.70
Square Root of GDP per Capita	0.03	0.04	5.31
Population Growth (%)	0.07	1.12	4.07
Physicians per 1,000 people	0.72	1.50	4.91
Fertility Rate	-2.07	-0.89	5.19
Square Root of Arable land per Capita	-3.24	-0.44	2.89
Immunization	0.09	0.13	1.88
CO2 emission per capita	-0.00	0.11	3.31

Part of this study's variable reduction process was to reduce the variance inflation factor to a values preferably below 5. Although some variables such as fertility rate and square root of GDP per capita showed signs of multicollinearity, previous VIF's were much higher. Thus, it is acknowledged that these variables may be related in some manner.

4.1.2 Model Interpretation

Referring back to table 4.1, it is shown that nearly all the variables show statistical significance at the 95% confidence level with the exception of CO2 emission per capita with a p-value of 0.056. Even though the p-value is marginally not significant, it could still be used to explain some characteristics of life expectancy. Overall, this model says that a country's increase of units for exported goods, square root of GDP per capita, population growth, physicians per 1,000 people and immunization will lead to an increase in life expectancy. Whereas a country's increase in fertility rate or square root of arable land per capita will lead to a decrease in life expectancy. Most of the variables that provide a positive correlation with life expectancy can be grouped into economic growth or urbanization. An increase in arable land can be seen as rural development which is shown to have an indirect relationship with life expectancy.

The increase in fertility rate may not be as clear of a predictor of life expectancy. It should noted that all of the data is dependent on a time series over the last 30 years. Figure 4.4 shows the fertility rate for the last 30 years of each country.

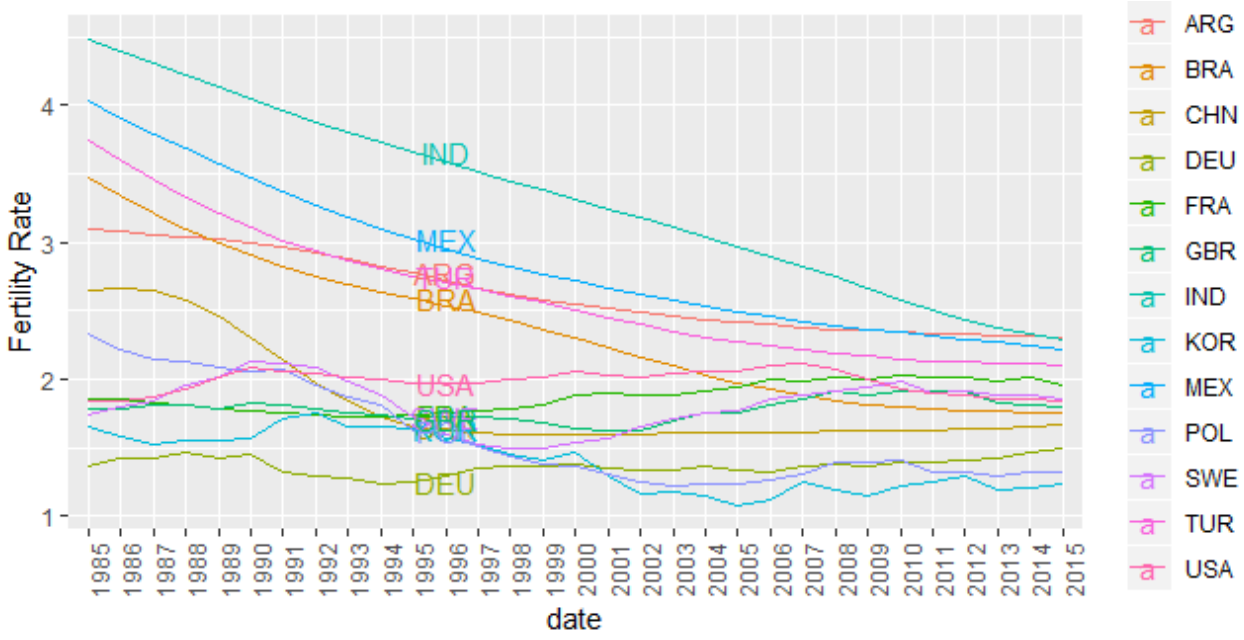

Figure 4.4: Fertility Rate over Time

It is shown that fertility rates have decreased steadily over time. That is, the number of births per woman has declined and could be a factor for an increase of life expectancy overall. These results will be examined further in the discussion section.

4.2 2nd Model

Part of the issue with some indicators is the lack of data points available before certain years. For this reason, the second model truncates the span of years so that other indicators with data from more recent years can provide insight on life expectancy. Particularly, this second regression model span 2005 to 2015. The additional independent variables for this model are shown in the table below:

Variable	Description
Urban Population Growth (Annual %)	Urban population refers to people living in urban areas as defined by national statistical offices.
Employment in agriculture, Female (% of female employment)	Employment is defined as persons of working age who were engaged in any activity to produce goods or provide services for pay or profit. The agriculture sector consists of activities in agriculture, hunting, forestry and fishing.
Renewable energy usage	Renewable energy consumption is the share of renewable energy in total final energy consumption.
Gini index	Gini index measures the extent to which the distribution of income (or, in some cases, consumption expenditure) among individuals or households within an economy deviates from a perfectly equal distribution.Thus a Gini index of 0 represents perfect equality, while an index of 100 implies perfect inequality.

As in the first linear regression model, two variables required certain transformations to ensure the predictor behaved closer to a normal distribution. This includes taking the log of the GDP per capita variable and the square root of the gini index. The results of this model are shown in the table below:

Table 4.3: Linear Regression #2 Model Results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	39.2595	3.2843	11.95	0.0000
Urban Population Growth Rate	0.4532	0.1513	2.99	0.0035
log of GDP per Capita	3.0350	0.1855	16.36	0.0000
Female agriculture employment %	-0.0262	0.0110	-2.40	0.0184
Renewable energy %	-0.0308	0.0076	-4.06	0.0001
Immunization	0.0994	0.0295	3.37	0.0011
Gini index over 32	-1.1205	0.3176	-3.53	0.0006
Adj $R^2 = 0.90$				
N = 143				

The results of this table will be examined in the discussion section.

4.2.1 Model Adequacy

Below are the residual plots used to examine the model adequacy.

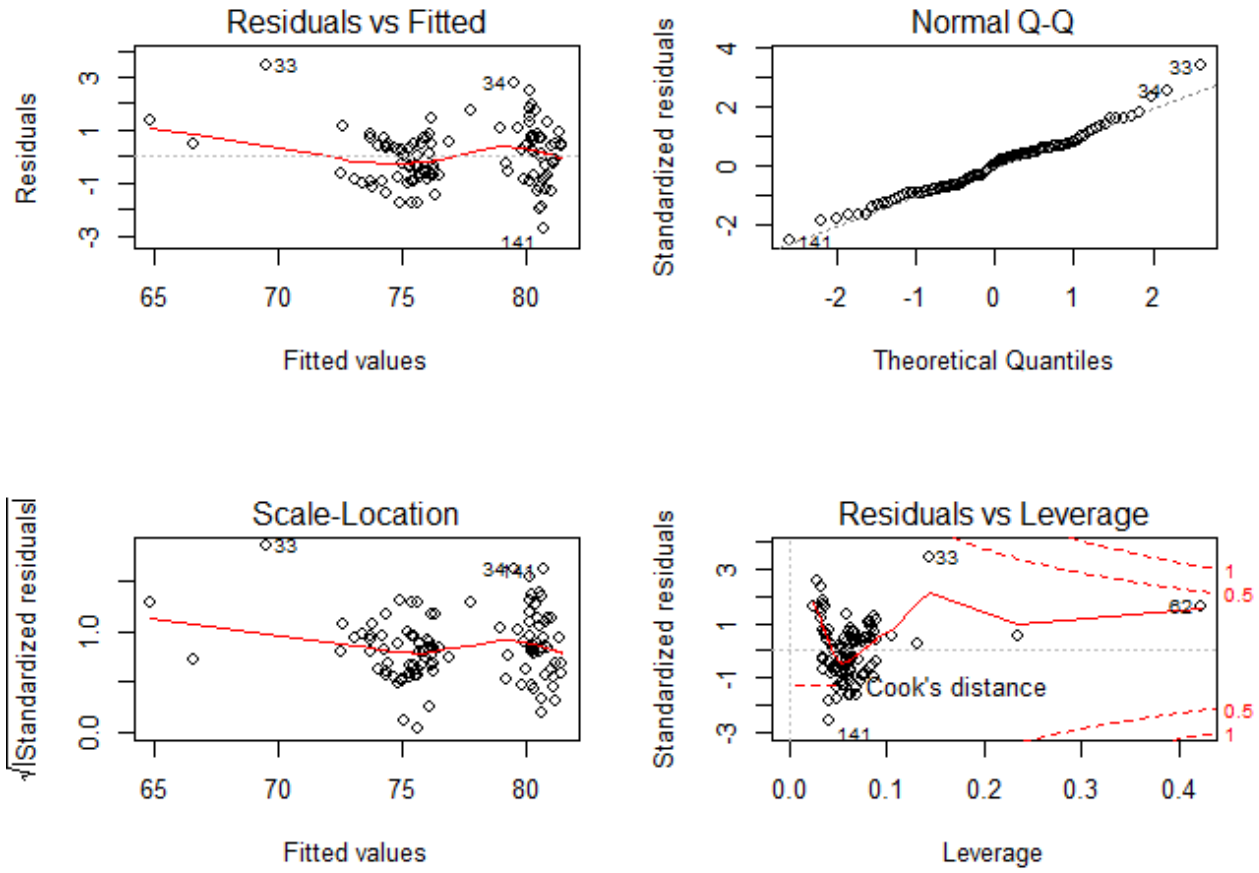

Figure 4.5: Residual Plot for Linear Regression Model #2

It is noted that the residual and standardized residual plots show mostly scattered points without any obvious trends. However, the QQ plot does begin to deviate from the at the tails. This is likely attributed the departure of normal distribution from some of our variables as the data has been truncated. For example, Figure 4.6 shows the distribution of the urban population growth rate.

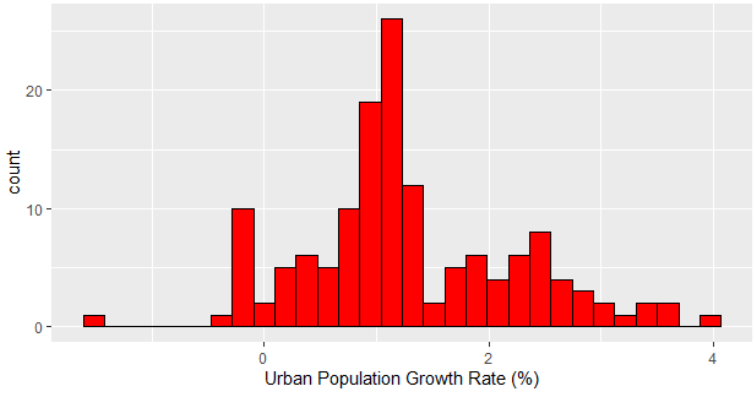

Figure 4.6: Distribution of Urban Population Growth Rate

This figure shows that a normal distribution is not as preserved. The model variance inflation factor is also considered in the following table:

Table 4.4: Linear Regression #2 Confidence Int. and VIF Results

	2.5 %	97.5 %	VIF
(Intercept)	32.74	45.78	
Urban Population Growth (Annual %)	0.15	0.75	1.78
log of GDP per Capita	2.67	3.40	2.54
Female agriculture employment %	-0.05	-0.00	2.46
Renewable energy %	-0.05	-0.02	1.21
Immunization	0.04	0.16	1.13
Gini index over 32	-1.75	-0.49	1.63

All of the variance inflation factors are below 5 indicating that there are no signs of multicollinearity.

4.2.2 Model Interpretation

Table 4.3 provides the results of the second regression model. It shows that all of the variables are statistically significant. In particular, an increase in urban population growth, log of GDP per capita and immunizations correspond to an increase in life expectancy. This is in contrast to an increase in female agriculture employment, renewable energy % and the square root of Gini index that corresponds to a decrease in life expectancy. As mentioned in the first linear regression model, urban population growth rate, GDP per capita and immunization combine economy growth and urbanization variables which also show a positive correlation life expectancy. The increase in female agriculture employment is a direct contrast to urbanization, so it is expected that it has an indirect relationship to life expectancy.

The gini index for this compressed data set has a range of 26 to 56. This variable was re-coded to have a binary outcome of being over 32. The reason this is skewed toward 26 is to try and see the how strong the influence is of a relatively low gini index. This model states that having a gini index over 32 (more towards inequality) will lead to a lower life expectancy.

One interesting finding, is the indirect relation of renewable energy to life expectancy. This model suggest that having a lower percentage of renewable energy will lead to a longer life expectancy. This relationship will be looked at further in the discussion section of this document.

5. Logistic Regression

Another model that will help with the characterization of factors that affect life expectancy is a logistic regression. As in the last section, a macroscopic model will be used to draw some general inferences.

5.1 Binary Life Expectancy Model

Prior to modeling, a binary dependent variable must be constructed. In this case, the binary outcome for life expectancy will be regarded as over or under 74. This was chosen simply as the mean of all the observations for life expectancy. After removing all null values for this particular model, it was found that 113 observations were under 74 and 185 were over. This ensures a reasonable distribution of life expectancy among independent variables to examine. in this analysis, no new variables were introduced that are different from the linear regression section. For reference of variables, see sections 3.3.4 and 4.1.1. Below is a table of results for both the model and the variance inflation factor.

Table 5.1: Logistic Regression, Confidence Int. and VIF Results

	Estimate	Std. Error	z value	Pr(>|z|)	2.5 %	97.5 %	VIF
(Intercept)	-31.325	6.837	-4.58	0.000	-47.13	-19.97	
Immunization	0.080	0.040	1.98	0.047	0.00	0.17	1.18
Export of Goods (% of GDP)	0.375	0.091	4.10	0.000	0.00	0.17	2.43
Square Root of GDP capita	0.135	0.028	4.86	0.000	0.09	0.20	2.39
Physicians per 1000	2.255	0.769	2.93	0.003	0.89	3.96	1.50

Null deviance: 395.546 on 297 degrees of freedom
Residual deviance: 59.438 on 293 degrees of freedom
AIC: 69.438

To supplement this model, a predictive logistic graph of the economic variables are displayed on Figure 5.1 below.

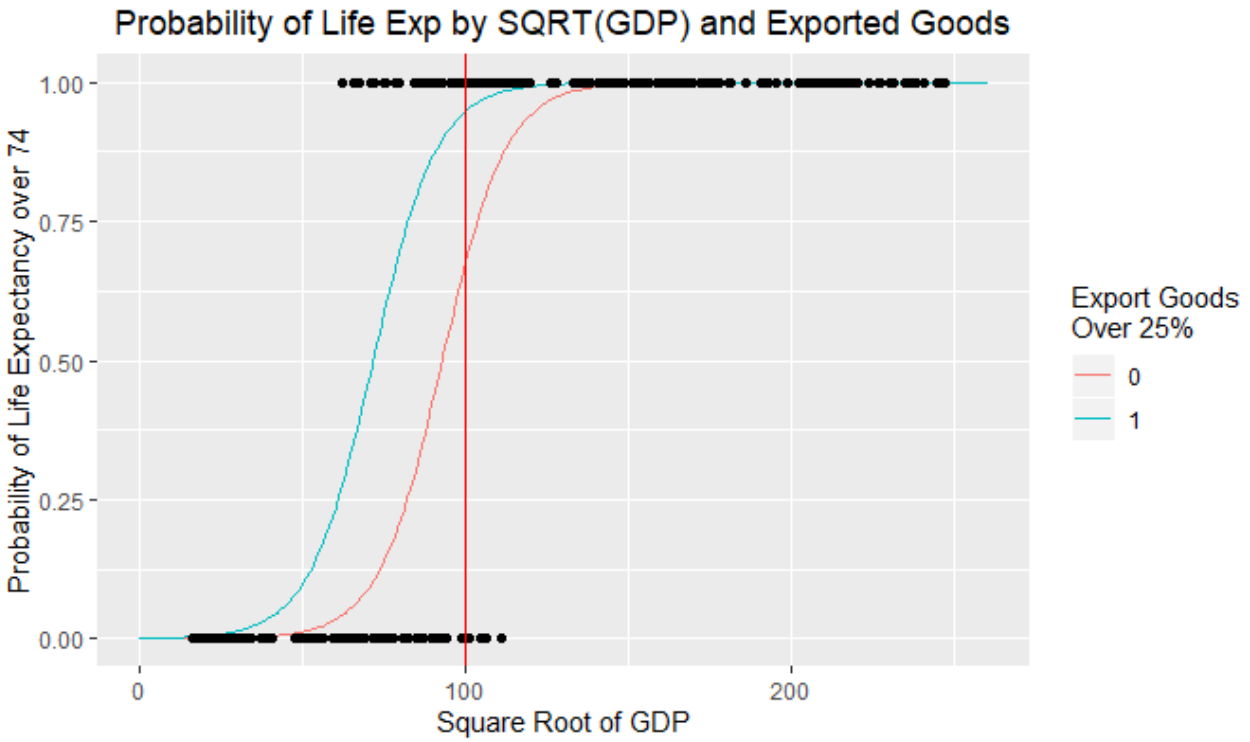

Figure 5.1: Logistic Predictive Plot

5.2 Model Interpretation

First, Table 5.1 suggests that an increase in any of the independent variables will increase the probability that the life expectancy will surpass 74. That is, the more immunizations and physicians per 1000 people there are in a given country (or countries), the more likely the life expectancy will be over 74. The same can be said about the other economic variables. Additionally, since the variance inflation factors from table 5.1 were all under 5, there is no evidence to suggest multicollinearity.

The economic factors can be more closely examined by Figure 5.1. First the "steepness" of the curve suggest that this is a good fit. This is supported by the relatively low residual deviance from Table 5.1. This figure gives the relation between the square root of GDP, export of goods (%) and what the likelihood is of having a life expectancy over 74. For instance, having a Square Root of GDP value of 100 and exporting goods over 25%(indicated by the blue curve) suggest that there is a 95% probability of having life expectancy over 74. In contrast, the same value for square root of GDP and exporting goods under 25% (indicated by the red curve) suggest there is a 67% probability of having life expectancy under 74. The 25% threshold is chosen for the mean of export of goods. This is quite a significant difference that should be considered and further studied.

6. Multinomial Logit

One final way that this study proposes to view life expectancy, is by the grouping of certain countries. Since there is a potential to have more than two groups of countries, this presents a proper opportunity to apply a multinomial logit model.

6.1 Country Groupings

Figure 3.4 shows the life expectancy of each country over time. Looking a little more closely, there are a few groupings that can be categorized based on the life expectancy at the start of 1985. For instance, Brazil, Turkey and India have relatively low life expectancy in 1985. In contrast, the United States, United Kingdom, France, Germany and Sweden have much higher life expectancy in 1985. The rest of the countries are somewhere in the middle. Therefore, the groupings will be based the life expectancy on 1985.

Table 6.1: Country Groupings

Countries	Group #
Brazil, Turkey, India	1
Argentina, South Korea, Mexico, China, Poland	2
United States, United Kingdom, France, Germany, Sweden	3

It could also be noticed in Figure 3.4 that group 1 has a relatively steeper slope that groups 2 and 3, with group 3 having the most shallow increase to life expectancy. This is likely due to a life expectancy plateau.

6.2 Multinomial Logit Model

Upon finalizing the groupings, a multinomial logit model can be performed with the groupings as the dependent variable. Below are the results of the regression model.

Table 6.2: Multinomial Logit Model Results

	Estimate	Std. Error	z-value	Pr(>|z|)
2:(intercept)	2.15	3.49	0.62	0.54
3:(intercept)	7.16	5.21	1.38	0.17
2:Square Root GDP capita	-0.12	0.03	-4.13	0.00
3:Square Root GDP capita	0.12	0.05	2.44	0.01
2:Immunization	0.09	0.04	2.09	0.04
3:Immunization	-0.31	0.11	-2.82	0.00
2:Energy Renewable %	-0.38	0.09	-4.31	0.00
3:Energy Renewable %	-0.02	0.07	-0.29	0.77
2:Physicians per 1000	4.14	1.47	2.81	0.00
3:Physicians per 1000	3.12	1.68	1.86	0.06

AIC = 65.93
Residual deviance: 45.9372 on 640 degrees of freedom
Log-likelihood: -22.9686 on 640 degrees of freedom

6.3 Model Interpretation

This multinomial logit model aims to compare groupings of countries based on life expectancy against certain indicators. As shown in Table 6.2, the reference level is set as group 1. Only the significant results will be discussed. First, an increase in physicians per 1000 people will likely place a country into group 2 compared to group 1. Second, an increase in renewable energy % will likely place a country into group 1 rather than group 2. This may go against intuition as it would expected that more environmental measures will increase life expectancy. This will be discussed further in a later section.

An increase in the square root of GDP per capita indicates that a country will more likely be placed into group 1 than into group 2. However the increase in the same GDP per capita value will likely place a country into group 3 when compared to group 1. This seems counter intuitive as one might expect that higher values of GDP per capita would result to being placed in a country grouping with higher life expectancy.

Lastly, an increase in immunization will likely place a country into group 2 compared to group 1. However, that same increase in immunization totals will likely place a country into group 1 compared to group 3. Again, this goes against intuition, indicated that this model may need to be refined.

7. Discussion

This section discusses the results of each model thoroughly and attempts to make connections the literature review that was conducted for this study. Limitations and future studies will also be discussed.

7.1 Economy, Urbanization and Environment

As discussed in earlier sections, the variables used in this study were chosen on a basis of major categories. This includes economic variables such as GDP and exportation of goods. Urbanization variables include urban population growth, gini index, employment in agriculture, etc. Lastly, environmental variables include renewable energy percentages and CO2 emmisions per capita. It should be noted that Figure 3.1 shows an almost perfect positive linear relationship of time and average life expectancy. Given that many of the variables also exhibit some linear trend (positive or negative), it is expected to find a strong relationship with life expectancy. This is the reason why the linear regression models have such high adjusted R^2 values. The following subsections investigate the major variables.

7.1.1 Economic Factors

In general, linear regression model #1 suggested than an increase in the exported good as a percentage of GDP and an increase in the square root of GDP per capita results in an increase of life expectancy. Visually, this can be seen by Figure 7.1 when plotting the regression line for life expectancy vs. square root of GDP per capita.

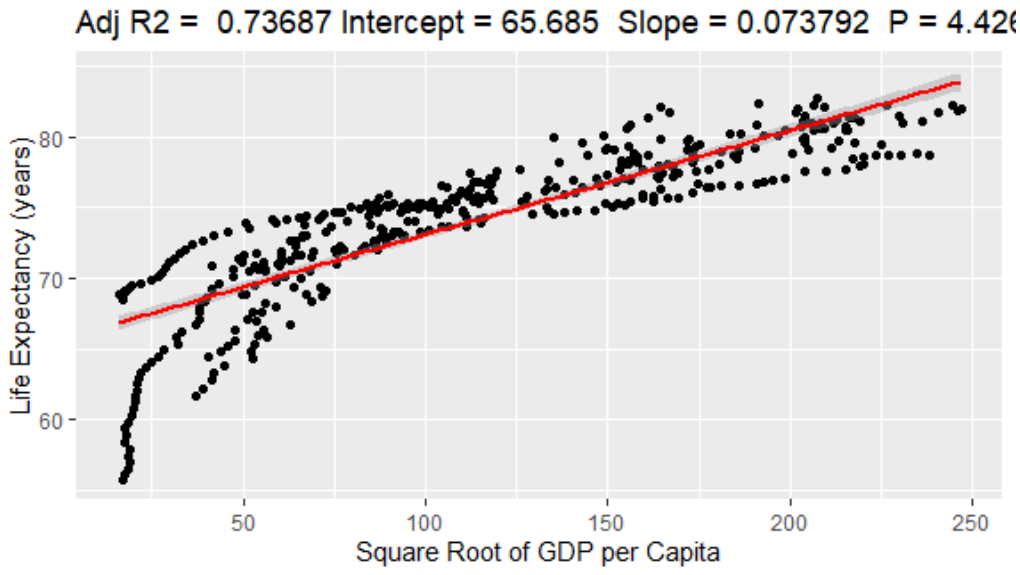

Figure 7.1: Linear relation of SQRT(GDP per Capita) with Life Exp.

Notice the relationship is mostly linear with the exception of values below 50 of the SQRT(GDP per Capita), which look to have some curvature. Also, the adjusted R^2 values are lower than the grand model from section 4. This reaffirms the use of additional variables in the regression model to reduce the overall error term relative to the use of a single independent variable. The effect of economic variables is further confirmed with the logistic model indicating that an increase in the square root of GDP per capita will increase the probability of having a life expectancy over the mean of 74 years old. The results of economic factors is partially confirmed by a study that finds income disparities and illiteracy rates are negatively associated with life expectancy, while

GDP per capita is positively associated with life expectancy through a cross-section study of Brazil states and federal capital [8]. Furthermore, Kennelly et al. found that GDP per capita income and the proportion of health expenditure financed by the government are both positively associated with better health outcomes when trying to explore the relationship between social capital and population health [4].

7.1.2 Urbanization

Similar to economic factors, it appears that an overall increase to urbanization in countries will lead to an overall increase in life expectancy. However, upon conducting the literature review, the impact of urbanization on life expectancy has contradictory results. Singh and Siahpush, in a 1969–2009 U.S. county-level mortality study, finds that life expectancy is inversely related to levels of rurality [9]. Aside from that, the life expectancy gap between rural and urban areas grow wider over time. However, a study carried out in the United Kingdom by Kyte and Wells shows that life expectancy in rural areas is higher than in urban areas [6]. Vast inequalities and deprivation contribute negatively to life expectancy. Although sparsity alone doesn't have a significant impact on life expectancy alone, deprivation accounts for more variation of life expectancy in less sparse areas. The different conclusions might present the complexity of rural and urban development in a different context, confounding with income inequality and other factors. In terms of rural employment, there is a study focusing on female life expectancy conducted by Williamson and Boehmer. The study shows that women's participation in agriculture has a negative impact on their life expectancy [10]. The complexity of this interpretation can be viewed by Figure 7.1 linear regression plot.

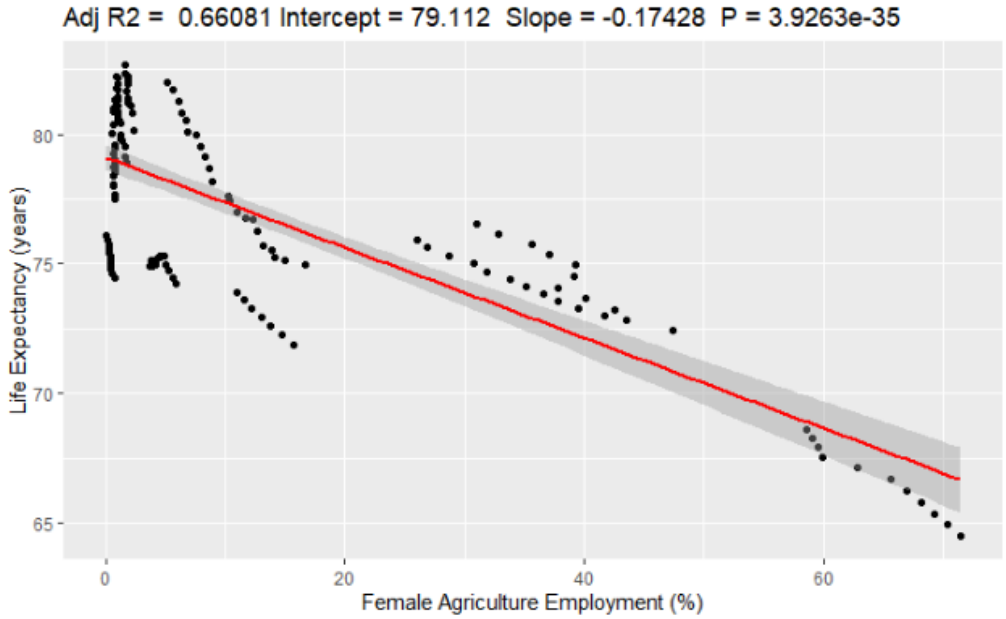

Figure 7.2: Linear relation of Female Employment in Agriculture (%) with Life Exp.

Several clusters that do not match the trend line. Overall, female employment in agriculture is negatively correlated with life expectancy.

Another seemingly intuitive variable that was studied, was the amount of physicians per 1000 people. As expected both the linear regression models and logistic models pointed to an increase in life expectancy (or probability of higher life expectancy) as the physician variable increased. This aligns with a seperate study that claims that expansion of medical care and adult literacy likely increases life expectancy [7].

Ultimately, the models from this study give evidence that urbanization in countries will increase life expectancy. However, some of the literature suggest the evidence is contradictory depending on the variables or particular country. This shows the drawback of generalizing all countries into a grand model.

7.1.3 Environmental

The environmental variables were among to the most challenging to find a relationship with life expectancy. This is mostly due to the lack of data points for each country over a span of thirty years. One surprising result came from the second linear regression model that suggested that a higher percentage of renewable energy

has a negative effect on life expectancy. One possible explanation for this is that the speed of urbanization (using non-renewable resources) has been faster than the generation of renewable resources. Figure 7.3 provides insight on this by looking at India as a particular case.

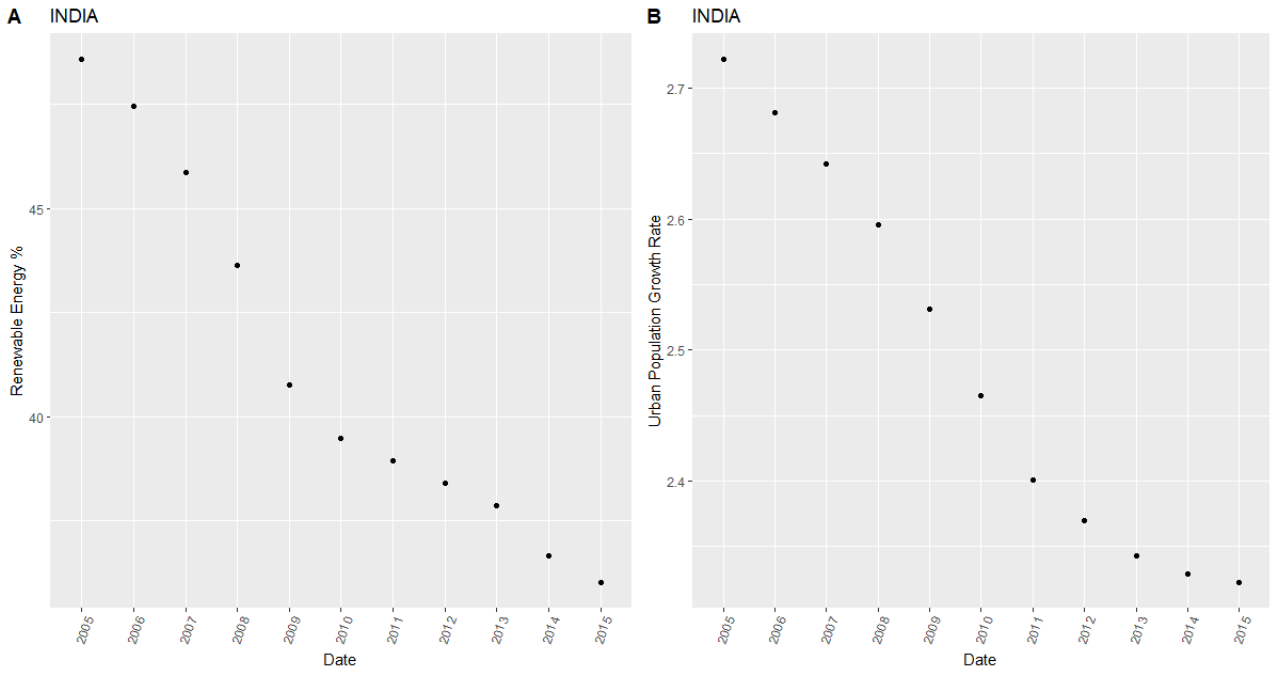

Figure 7.3: Comparative scatter plots for India

It is shown that even though urban growth rate has been declining over time, the rate is still positive indicating that India is constantly growing in urban areas. This increase will likely show a decrease in renewable energy percentage if the rate of renewable sources usage does not also increase. This is only a possible explanation which would need to be verified or supported by documentation.

7.2 Countries

Section 6's multinomial logit model was an attempt to group countries to see what indicators contribute their placing within certain groups. This was partially motivated by the findings of Mazudar that suggests the determining factor for life expectancy likely varies for different income groups of countries[7]. Unfortunately, there were some contradictory results within the groupings. For example, an increase in the square root of GDP per capita indicates that a country will more likely be placed into group 1 than into group 2. However the increase in the same GDP per capita value will likely place a country into group 3 when compared to group 1. The previous linear and logistic regression models suggested that an increase in GDP per capita will always increase life expectancy. This shows the possibility that the groupings may be more similar than intended. Recall that the groupings were based on the life expectancy starting on 1985. However, referring back to Figure 3.4, it is shown that overtime, countries may overlap in life expectancy. This overlap may contribute to the contradictory results. If done differently, a separate study may group in a more clear, robust fashion as oppose to a simple visual grouping. Cluster analysis may be a good candidate for this.

7.3 Limitations

One omission that was not thoroughly explored was the factor that time had in these results. It was used to motivate some part of the analysis, but ultimately this study looked at the relation of indicators without a time factor. In reality, time may play a significant role in the history of each country and may influence the explanation of certain results. For example, 2008 was the start of a global recession which may have impacted some of the indicators (GDP, growth, etc.). Any shift in indicator values will likely be revealed by looking at it's time series. This could unveil a hidden parameter that influences that particular indicator. A separate study may look to incorporate time in a more rigid fashion.

Perhaps another factor that could drastically influence life expectancy, is the mortality rate of under five year olds. A study found that China's life expectancy from 1991 to 2012 increased by 6.4 years. However, this was due to the decrease deaths of under five year old children [3]. Thus, it seems that this could be a hidden factor that may supplement the explanation life expectancy.

Lastly, it should be noted that this study has a macroscopic lens applied to it. That is, the results arrive characterize a wide, but fairly shallow set of explanations. Each country has a complex set of cultures, governments and geopolitical establishments that significantly determine the direction of their respective nations. The disagreement between some of the literature and this report's findings regarding urbanization is a good example of this. This study only aims to identify some possible factors for life expectancy so that future studies can investigate more thoroughly.

7.4 Future Studies

Future studies may take look at this report to use as a starting point to investigate particular countries or factors they are interested in. Another approach to this study would be to conduct a cluster analysis or PCA given the potential size of indicators. This may help in finding certain groupings to analyze backed by a more robust process. It should be noted that African countries were not explored due to the lack of data provided the world bank data. It may be beneficial to deviate towards other sources of data.

7.5 Conclusions

This study's objective was to characterize the factors that influence life expectancy from a global standpoint. In general, it was found that countries moving towards urbanization and increasing their economic prowess will lead to a higher life expectancy. Environmental influences were not as clear, and should be part of a separate study to arrive to more conclusive result. Readers of this study can use this to jumpstart branching research on other countries or inform themselves towards research policy changes to increase life expectancy.

Bibliography

[1] Christopher Dye. Health and urban living. *Science*, 2008.

[2] Correia Andrew et al. The effect of air pollution control on life expectancy in the united states: An analysis of 545 us counties for the period 2000 to 2007. *US National Library of Medicine*, 2014.

[3] Ebenstein Avraham et al. Growth, pollution, and life expectancy: China from 1991-2012. *The American Economic Review*, 105(5):226–231, 2015.

[4] O'Shea E. Garvey E Kennelly, B. Social capital, life expectancy and mortality: a cross-national examination. social science medicine. *Social Science Medicine*, 56:2367–2377.

[5] W. Kondro. Global life expectancy rises. *Canadian Medical Association. Journal 182(8)*, 2010.

[6] C. Kyte L., Wells. Variations in life expectancy between rural and urban areas of england, 2001–07. *Health Statistics Quarterly*, 46:27–52.

[7] K. Mazumdar. Improvements in life expectancy: 1960-1995 an exploratory analysis. *Social Indicators Research*, 55(3):303–328.

[8] E Messias. Income inequality, illiteracy rate, and life expectancy in brazil. *American Journal of Public Health*, 94:1294–1296.

[9] Siahpush M. Singh GK. Widening rural–urban disparities in life expectancy, u.s., 1969–2009. *American Journal of Preventive Medicine*, 46(2):e19–e29.

[10] Boehmer U. Williamson JB. Female life expectancy, gender stratification, health status, and level of economic development: A cross-national study of less developed countries. *Social Science Medicine*, 45(2):305–317.

Appendix A: Resources

R code for this report will be in this section.

```
---
title: "world_bank_indi_4"
author: "Ning-siman, Chris Salazar"
date: "2020/3/8"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```


```{r Initial Data Set-up }

library(wbstats)
library(tidyverse)
#reference https://cran.r-project.org/web/packages/wbstats/vignettes/
Using_the_wbstats_package.html
#str(wb_cachelist)
new_cache <- wbcache()
##get life expectancy
wbsearch(pattern = "life expectancy at birth, total")

# Countries chosen for study
filt1 <- c("United States" , "Mexico", "Sweden", "Argentina" , "Poland" ,
"Turkey", "China", "India","Korea, Rep.", "United Kingdom", "Germany",
"France", "Brazil")

# Indicator codes for world bank data communication
indi <- c("SP.DYN.LE00.IN", #life expectancy
               "NY.ADJ.SVNG.GN.ZS", #Adjusted Net Savings
                "NY.ADJ.NNTY.KD", #Adjusted net national income (US Adjusted)
                "BN.CAB.XOKA.CD",  #Current account balance
           "GC.XPN.TOTL.GD.ZS", # Expense (% of GDP)
          "NE.EXP.GNFS.ZS", # Export of goods (% of GDP)
          "NY.GDP.MKTP.CD", # GDP
          "NY.GDP.MKTP.KD.ZG", # GDP Growth (annual %)
          "NY.GDP.PCAP.CD", # GDP per Cap
          "NY.GNS.ICTR.ZS", # Gross Savings (% of GDP)
          "NV.IND.TOTL.ZS", # Industry (% of GDP)
          "DT.ODA.ODAT.CD", # Net official development assistance received
(current US$)
          "GC.REV.XGRT.GD.ZS", # Revenue, excluding grants (% of GDP)
          "CM.MKT.TRAD.GD.ZS", # Stocks traded, total value (% of GDP)
          "IE.PPI.WATR.CD", # Investments in water/sanitation
          "EG.USE.ELEC.KH.PC", # Electric power consumption
          "TX.VAL.MRCH.CD.WT", # Merchandise Exports
          "IC.TAX.TOTL.CP.ZS", #Total tax and contribution rate (% of profit)

          ##Siman indicator
          "SP.DYN.TFRT.IN",#total fertility rate-(births per woman
         "SH.MED.PHYS.ZS", #Physicians (per 1,000 people
         "NY.ADJ.NNTY.PC.KD.ZG" , #Adjusted net national income per capita
(annual % growth)
         "SP.POP.GROW", #Population growth (annual %)
         "SP.URB.GROW", #urban population growth rate
```
```

```
      "AG.LND.ARBL.ZS", #Arable land (% of land area)
      "AG.LND.ARBL.HA.PC", #Arable land (hectares per person)
"SL.AGR.EMPL.FE.ZS", #Employment in agriculture, female (% of female
employment) (modeled ILO estimate)
"AG.CON.FERT.ZS", #Fertilizer consumption (kilograms per hectare of arable
land)
"AG.LND.CROP.ZS", #Permanent cropland (% of land area)
#"SP.RUR.TOTL", #Rural population
"SP.RUR.TOTL.ZS", #Rural population (% of total population)

#Robin
"EN.ATM.PM25.MC.M3", #"Air pollution mean aunnual exposure"
"SI.POV.GINI",  #Gini index
"EG.FEC.RNEW.ZS",   #"Renewable energy usage"
"SL.UEM.TOTL.ZS", #"Unemployment of the total workforce"
"SH.IMM.IDPT", # "Immunization of people DPT % of the population"
"EN.ATM.CO2E.SF.ZS",#"CO2 emissions of total solid fuel emissions"
"EN.ATM.CO2E.PP.GD", #"CO2 emissions (kg per PPP $of GDP)"
"EN.ATM.CO2E.PC"     #"CO2 emission metric tone per capita"
)

df <- wb(indicator =indi, startdate = 1985, enddate = 2015) %>% filter(country
%in% filt1) #%>%  select(1:5)
which(duplicated(df))

# Organize Countries and indicators
df2 <- pivot_wider(df, names_from = c(indicator,indicatorID), values_from =
value)

df2_country<- pivot_wider(df, names_from =c(iso2c, iso3c, country),
values_from = "value" )

```

Organize Data

```{r RenameV, echo=TRUE}
df3 <- df2
# df3$date <- as.integer(df3$date)

#Rename Variables
names(df3) <- c("iso3c" ,  "date","iso2c",  "country",
                #Chris
                "Life_expect", 'AdjustedNNS','AdjustedNNI', 'CurrAccBal',
'EXPENSE', 'ExpGoods', 'GDP', 'GDPGrowth',
'GDPperCAP', 'GrowthSav', 'IndValAdd', 'Net_Develop_Assistance', 'Rev',
'StockTr', 'InvWater', 'ElecPw', 'MerchExp', 'TaxContribution',
#Siman
"Fertility_rate","Physicians_1000", "AdjustedNNI_growth", "Pop_growth","Urb_pop_growth",
"Arable_land_pct","Arable_land_capita","female_agri_employment_pct","Fertilizer_consumptio
, "Permanent_cropland",  "rural_pop_pct",

#Robin
"PM25_mean_exposure", "Gini_index","Renewable_energy_consumption_pct" ,
"Unemployment", "Immunization","CO2_from_solid_fuel", "CO2_kg_PPP"
,"CO2_ton_capita" )
```

```

*This is the best fit lm model*
```{r best fit, echo = TRUE}
library(car)
library(xtable)
lm_int2 <- lm(Life_expect~ExpGoods + sqrt(GDPperCAP)
     +Pop_growth+ Physicians_1000 + Fertility_rate + # AdjustedNNI_growth +
     # female_agri_employment_pct + #Fertilizer_consumption +
Permanent_cropland +
     #rural_pop_totl + rural_pop_pct
    + sqrt(Arable_land_capita) +#+ Gini_index + PM25_mean_exposure
    Immunization  + CO2_ton_capita

#   #+Fertilizer_consumption*Arable_land_capita #0.27
#   # + Urb_pop_growth*GDPperCAP #0.53
# # + PM25_mean_exposure*GDPperCAP
# # + Gini_index*GDPperCAP
# + Renewable_energy_consumption_pct*AdjustedNNI_growth
# + Pop_growth*AdjustedNNI_growth
# # + Fertility_rate*AdjustedNNI_growth
# + CO2_ton_capita* AdjustedNNI_growth
# # + CO2_kg_PPP* GDPperCAP ##potential!!
# # + Arable_land_capita*AdjustedNNI_growth
# # + Arable_land_capita*Fertility_rate
# + CO2_from_solid_fuel*AdjustedNNI_growth
# # + CO2_from_solid_fuel*GDPGrowth
# # + CO2_from_solid_fuel*GDP
# # +Fertilizer_consumption*GDPperCAP
# # +log(MerchExp)*AdjustedNNI_growth
# + Permanent_cropland *AdjustedNNI_growth
# # + female_agri_employment_pct*Arable_land_capita
# # +rural_pop_pct*AdjustedNNI_growth #siginifcant
, data = df3)

summary(lm_int2)
vif(lm_int2)
par(mfrow=c(2,2))
plot(lm_int2)

xtable(summary(lm_int2))

ConT = confint(lm_int2)
xtable(ConT)

####Best fit model to factor recent changes---2005-2015
df4 <- subset(df3, date %in% c("2005","2006", "2007", "2008", "2009", "2010",
"2011",  "2012", "2013", "2014", "2015"))

lm_20051 <- lm(Life_expect~ # Pop_growth +  Physicians_1000 +
     +Urb_pop_growth
    +log(GDPperCAP)+ female_agri_employment_pct
    + Renewable_energy_consumption_pct+
    Immunization  + sqrt(Gini_index ), data = df4)
```
```

```
summary(lm_20051)
vif(lm_20051)
plot(lm_20051)

fit = lm(Life_expect ~ sqrt(GDPperCAP), data = df3)

# LM Plot for SQUARE ROOT of GDP per Capita
ggplot(df3, aes(x = sqrt(GDPperCAP), y = Life_expect)) +
  geom_point() +
  stat_smooth(method = "lm", col = "red") + ylab('Life Expectancy (years)') +
xlab('Square Root of GDP per Capita') + labs(title = paste("Adj R2 =
",signif(summary(fit)$adj.r.squared, 5),
                     "Intercept =",signif(fit$coef[[1]],5 ),
                     " Slope =",signif(fit$coef[[2]], 5),
                     " P =",signif(summary(fit)$coef[2,4], 5)))

fit = lm(Life_expect ~ female_agri_employment_pct, data = df4)

# LM Plot for SQUARE ROOT of GDP per Capita
ggplot(df4, aes(x = female_agri_employment_pct, y = Life_expect)) +
  geom_point() +
  stat_smooth(method = "lm", col = "red") + ylab('Life Expectancy (years)') +
xlab('Female Agriculture Employment (%)') + labs(title = paste("Adj R2 =
",signif(summary(fit)$adj.r.squared, 5),
                     "Intercept =",signif(fit$coef[[1]],5 ),
                     " Slope =",signif(fit$coef[[2]], 5),
                     " P =",signif(summary(fit)$coef[2,4], 5)))


#Import the cowplot library
library(cowplot)

Renewable = df4 %>% filter(df4$iso3c == 'IND') %>% ggplot(aes(x=date, y =
Renewable_energy_consumption_pct)) + geom_point() + xlab('Date') +
ylab('Renewable Energy %')+theme(axis.text.x = element_text(angle = 70, hjust
= 1))+ggtitle('INDIA')

Urban_Pop = df4 %>% filter(df4$iso3c == 'IND') %>% ggplot(aes(x=date, y =
Urb_pop_growth)) + geom_point() + xlab('Date') + ylab('Urban Population Growth
Rate')+theme(axis.text.x = element_text(angle = 70, hjust = 1))
+ggtitle('INDIA')

# Additional Plots for analysis
plot_grid(Renewable, Urban_Pop, labels = "AUTO")
```


Exploratory Data Vizualization
```{r viz}

##visualization of life expectancy
coplot(Life_expect ~ date|country, type="l", data=df3)
#life expectancy
```

```
ggplot(df3, aes(x= date, y =Life_expect, colour = iso3c, group_by(country)))
+geom_line()+geom_text(data = subset(df3, date == "1995"),aes(label =
country),nudge_x = 1,nudge_y = 0.3, check_overlap = FALSE)

#female_agri_employment_pct
ggplot(df3, aes(x= date, y =female_agri_employment_pct, colour = iso3c,
group_by(country)))+geom_line()+geom_text(data = subset(df3, date ==
"1995"),aes(label = country),nudge_x = 1,nudge_y = 0.3, check_overlap = FALSE)

#sqrt GDP per capita
ggplot(df3, aes(x= date, y = sqrt(GDPperCAP), colour = iso3c,
group_by(country)))+geom_line()+geom_text(data = subset(df3, date ==
"1995"),aes(label = country),nudge_x = 1,nudge_y = 0.3, check_overlap = FALSE)
#log(MerchExp)
ggplot(df3, aes(x= date, y = log(MerchExp), colour = iso3c,
group_by(country)))+geom_line()+geom_text(data = subset(df3, date ==
"1995"),aes(label = country),nudge_x = 1,nudge_y = 0.3, check_overlap = FALSE)

#Gini--> Gini index tells something, but too much missing values-->2003/4/5
starts all
ggplot(df3, aes(x= date, y = Gini_index, colour = iso3c, group_by(country)))
+geom_line()+geom_text(data = subset(df3, date == "2005"),aes(label =
country),nudge_x = 1,nudge_y = 0.3, check_overlap = FALSE)

#Renewable_energy_consumption_pct---> must combine with development
ggplot(df3, aes(x= date, y = Renewable_energy_consumption_pct, colour = iso3c,
group_by(country)))+geom_line()+geom_text(data = subset(df3, date ==
"2005"),aes(label = country),nudge_x = 1,nudge_y = 0.3, check_overlap = FALSE)

#df3$CO2_kg_PPP
ggplot(df3, aes(x= date, y = CO2_kg_PPP, colour = iso3c, group_by(country)))
+geom_line()+geom_text(data = subset(df3, date == "2005"),aes(label =
country),nudge_x = 1,nudge_y = 0, check_overlap = FALSE)

#Immunization
ggplot(df3, aes(x= date, y = Immunization, colour = iso3c, group_by(country)))
+geom_line()+geom_text(data = subset(df3, date == "2005"),aes(label =
country),nudge_x = 1,nudge_y = 0.3, check_overlap = FALSE)

#Immunization--not very explainable??? from graphic point
# ggplot(df3, aes(x= date, y = inverse_immu, colour = iso3c,
group_by(country)))+geom_line()+geom_text(data = subset(df3, date ==
"2005"),aes(label = country),nudge_x = 1,nudge_y = 0.3, check_overlap = FALSE)


#Permanent_cropland--Turkey, India, China, Poland, increase; Korea rep
decrease and increase, France decrease, others hold constant
ggplot(df3, aes(x= date, y = Permanent_cropland, colour = iso3c,
group_by(country)))+geom_line()+geom_text(data = subset(df3, date ==
"2005"),aes(label = country),nudge_x = 1,nudge_y = 0.3, check_overlap = FALSE)


#sqrt(Fertility_rate)
ggplot(df3, aes(x= date, y = sqrt(Fertility_rate), colour = iso3c,
group_by(country)))+geom_line()+geom_text(data = subset(df3, date ==
"2005"),aes(label = country),nudge_x = 1,nudge_y = 0, check_overlap = FALSE)
```

```
#Physicians_1000
ggplot(df3, aes(x= date, y = Physicians_1000, colour = iso3c,
group_by(country)))+geom_line()+geom_text(data = subset(df3, date ==
"2005"),aes(label = country),nudge_x = 1,nudge_y = 0, check_overlap = FALSE)


#AdjustedNNI_growth
ggplot(df3, aes(x= date, y = AdjustedNNI_growth, colour = iso3c,
group_by(country)))+geom_line()+geom_text(data = subset(df3, date ==
"2005"),aes(label = country),nudge_x = 1,nudge_y = 0, check_overlap = FALSE)

#Pop_growth
ggplot(df3, aes(x= date, y =Pop_growth, colour = iso3c, group_by(country)))
+geom_line()+geom_text(data = subset(df3, date == "2005"),aes(label =
country),nudge_x = 1,nudge_y = 0, check_overlap = FALSE)

#Urb_pop_growth
ggplot(df3, aes(x= date, y =Urb_pop_growth, colour = iso3c,
group_by(country)))+geom_line()+geom_text(data = subset(df3, date ==
"2005"),aes(label = country),nudge_x = 1,nudge_y = 0, check_overlap = FALSE)


#Arable_land_pct
ggplot(df3, aes(x= date, y =Arable_land_pct, colour = iso3c,
group_by(country)))+geom_line()+geom_text(data = subset(df3, date ==
"2005"),aes(label = country),nudge_x = 1,nudge_y = 0, check_overlap = FALSE)

#Arable_land_capita
ggplot(df3, aes(x= date, y =Arable_land_capita, colour = iso3c,
group_by(country)))+geom_line()+geom_text(data = subset(df3, date ==
"2005"),aes(label = country),nudge_x = 1,nudge_y = 0, check_overlap = FALSE)

# Fertilizer_consumption ---too many missing values, not started until 2002?
ggplot(df3, aes(x= date, y =Fertilizer_consumption, colour = iso3c,
group_by(country)))+geom_line()+geom_text(data = subset(df3, date ==
"2005"),aes(label = country),nudge_x = 1,nudge_y = 0, check_overlap = FALSE)

# rural_pop_pct
ggplot(df3, aes(x= date, y =rural_pop_pct, colour = iso3c, group_by(country)))
+geom_line()+geom_text(data = subset(df3, date == "2005"),aes(label =
country),nudge_x = 1,nudge_y = 0, check_overlap = FALSE)


#PM25_mean_exposure--->not until 2010
ggplot(df3, aes(x= date, y =PM25_mean_exposure, colour = iso3c,
group_by(country)))+geom_line()+geom_text(data = subset(df3, date ==
"2005"),aes(label = country),nudge_x = 1,nudge_y = 0, check_overlap = FALSE)

#CO2_from_solid_fuel
ggplot(df3, aes(x= date, y =CO2_from_solid_fuel, colour = iso3c,
group_by(country)))+geom_line()+geom_text(data = subset(df3, date ==
"2005"),aes(label = country),nudge_x = 1,nudge_y = 0, check_overlap = FALSE)

#CO2 kg ppp
```

```
ggplot(df3, aes(x= date, y =CO2_kg_PPP, colour = iso3c, group_by(country)))
+geom_line()+geom_text(data = subset(df3, date == "2005"),aes(label =
country),nudge_x = 1,nudge_y = 0, check_overlap = FALSE)
```

```{r Global Plot}

Mean_life = aggregate(df3$Life_expect, list(df3$date), mean)
# Mean Life expectancy time series plot
Mean_life %>% ggplot(aes(Group.1, x)) + geom_point(colour = "red", size = 3)+
xlab('Year') + ylab("Life Expectancy (Years)")+theme(axis.text.x =
element_text(angle = 90, hjust = 1))

df3 %>% ggplot(aes(Life_expect)) + geom_histogram(color = "black", fill =
"red") + xlab('Life Expectancy (years)')

df4 = df3 %>% group_by(country)

# Boxplot and Violin Plot
ggplot(df4, aes(y = Life_expect, x =country)) + geom_violin(fill = 'skyblue')+
geom_boxplot(width=0.15)+theme(axis.text.x = element_text(angle = 70, hjust =
1)) + ylab("Life Expectancy") + xlab('Country')

#life expectancy
ggplot(df4, aes(x= date, y =Life_expect, colour = iso3c))+geom_line(aes(group
= iso3c))+theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
    scale_fill_manual(name = 'Country') + ylab('Life Expectancy') +
geom_text(data = subset(df3, date == "1995"),aes(label = iso3c),nudge_x =
1,nudge_y = 0, check_overlap = FALSE)

#Import the cowplot library
library(cowplot)

Age_Plot = df3 %>% ggplot(aes(x=Arable_land_capita)) +
geom_histogram(color="black", fill="green") + xlab('Arable Land per Capita')

Sqrt_Plot = df3 %>% ggplot(aes(x=sqrt(Arable_land_capita))) +
geom_histogram(color="black", fill="green") + xlab('Square Root of Arable Land
per Capita')

# Arable Land Transformation
plot_grid(Age_Plot, Sqrt_Plot, labels = "AUTO")
```

```
####
#### Logistic Regression and Multinomial Regression Code
####



library(ggplot2)
library(tidyverse)
library(car)
library(ggpmisc)
library(VGAM)
library(nnet)
library(mlogit)
## Write to csv in relevant file

## Group by data
wd <- "C:/Users/RTG/Desktop/UW/Winter Quarter/Inferential data analysis/
Project"
setwd(wd)

## Seperate script for visualizing data in project

# Read in datafile
#original.df = read.csv("Pivot_table_countires_by_year.csv", header = TRUE)
original.df = read.csv("Mydf.csv", header = TRUE)

## Rename categories
# Life expactancy [Years]
original.df <- original.df%>%rename( Life_expectancy =
Life.expectancy.at.birth..total..years._SP.DYN.LE00.IN)
# date = year
original.df <- original.df%>%rename( Year = date)
# Renewable energy
original.df <- original.df%>%rename( Energy_Renewable =
Renewable.energy.consumption....of.total.final.energy.consumption._EG.FEC.RNEW.ZS
)
# Labor unenplyment
original.df <- original.df%>%rename( Labor_unenployment =
Unemployment..total....of.total.labor.force...modeled.ILO.estimate._SL.UEM.TOTL.ZS
)
# Imunization
original.df <- original.df%>%rename( Immun =
Immunization..DPT....of.children.ages.12.23.months._SH.IMM.IDPT )
# GDP per capita
original.df <- original.df%>%rename( GDPcapita =
GDP.per.capita..current.US.._NY.GDP.PCAP.CD )
# Physicians per thousands
original.df <- original.df%>%rename( Physicians = Physicians..per.
1.000.people._SH.MED.PHYS.ZS)
# Export of Goods
original.df <- original.df%>%rename( Export_of_Goods =
Exports.of.goods.and.services....of.GDP._NE.EXP.GNFS.ZS)


### The grand logistic model
```

```
# Take the mean
age_mean <- mean(original.df$Life_expectancy)

# Assign 1 to life expectancy above the mean and 0 below the mean
original.df$age_logit <- ifelse(original.df$Life_expectancy > age_mean,1,0)

# Select data from the original data file
grand_logit.df <- original.df %>%
  select(country, Year, iso3c, Life_expectancy,age_logit ,
           Immun, GDPcapita, Physicians, Export_of_Goods  )

# Cleaning the data from NA values
grand_logit.df <-
grand_logit.df[complete.cases(grand_logit.df$Life_expectancy),]
grand_logit.df <- grand_logit.df[complete.cases(grand_logit.df$Immun),]
grand_logit.df <- grand_logit.df[complete.cases(grand_logit.df$GDPcapita),]
grand_logit.df <-
grand_logit.df[complete.cases(grand_logit.df$Export_of_Goods),]
grand_logit.df <- grand_logit.df[complete.cases(grand_logit.df$Physicians),]

# Squaring the GDP capita
grand_logit.df$GDPcapita_sq <- sqrt(grand_logit.df$GDPcapita)

# Lost observations
lost_obs <- length(original.df$Life_expectancy) -
length(grand_logit.df$Life_expectancy)

# Modeling grand logistic regression model
y_grand <- glm(age_logit ~ Immun + Export_of_Goods + GDPcapita_sq +Physicians
                , family = binomial(), data = grand_logit.df)

summary(y_grand)

# Variance influence factor
vif(y_grand)

# Confidence interval
confint(y_grand)

#New Variable for exported goods
grand_logit.df$Export_Gret_25 = ifelse(grand_logit.df$Export_of_Goods > 25,
1,0)


# Plot logit curve
logit_10_a = glm (formula = age_logit ~  Export_Gret_25 + GDPcapita_sq,
                  family = binomial( ) , data = grand_logit.df)

plot_data = expand.grid(GDPcapita_sq = seq(0,260,1), Export_Gret_25= (0:1))
plot_data$preds = plogis(predict(logit_10_a,newdata = plot_data))

ggplot()+geom_line(data= plot_data, aes(x = GDPcapita_sq, y = preds, color =
as.factor(Export_Gret_25)))+
  geom_point(data =grand_logit.df, aes(x = GDPcapita_sq, y = age_logit)) +
ggtitle("Probability of Life Exp by SQRT(GDP) and Exported Goods")+
```

```
  ylab('Probability of Life Expectancy over 74') +  labs(color ="Export
Goods\nOver 25%")+
  theme(plot.title = element_text(hjust = 0.5)) + xlab('Square Root of GDP')

#####
##### Multinomial Logit Model
#####

## Introduce a Multinominal logit model
multi_nominal.df <- original.df
multi_nominal.df$Life_hier <- as.character(multi_nominal.df$country)

# Recoding Low start
multi_nominal.df$Life_hier[ multi_nominal.df$Life_hier == "Brazil"] = 1
multi_nominal.df$Life_hier[ multi_nominal.df$Life_hier == "Turkey"] = 1
multi_nominal.df$Life_hier[ multi_nominal.df$Life_hier == "India"] = 1

# Recoding middle start
multi_nominal.df$Life_hier[ multi_nominal.df$Life_hier == "Argentina"] = 2
multi_nominal.df$Life_hier[ multi_nominal.df$Life_hier == "Korea, Rep."] = 2
multi_nominal.df$Life_hier[ multi_nominal.df$Life_hier == "Mexico"] = 2
multi_nominal.df$Life_hier[ multi_nominal.df$Life_hier == "China"] = 2
multi_nominal.df$Life_hier[ multi_nominal.df$Life_hier == "Poland"] = 2

# Recoding high start
multi_nominal.df$Life_hier[ multi_nominal.df$Life_hier == "United Kingdom"] =
3
multi_nominal.df$Life_hier[ multi_nominal.df$Life_hier == "United States"] = 3
multi_nominal.df$Life_hier[ multi_nominal.df$Life_hier == "France"] = 3
multi_nominal.df$Life_hier[ multi_nominal.df$Life_hier == "Germany"] = 3
multi_nominal.df$Life_hier[ multi_nominal.df$Life_hier == "Sweden"] = 3

# Factor the code
multi_nominal.df$Life_hier <- as.factor(multi_nominal.df$Life_hier)

# Squaring the GDP capita
multi_nominal.df$GDPcapita_sq <- sqrt(multi_nominal.df$GDPcapita)

## MNL Models
# VGAM
mlogitvgam <- vglm(Life_hier ~ GDPcapita_sq + Immun + Energy_Renewable +
                     Physicians, family = multinomial(),
                   data = multi_nominal.df)
# Akaike information criterion
AIC(mlogitvgam)

summary(mlogitvgam)

# Logliklihood
logLik(mlogitvgam)

# NNET
mlogitnnet <- multinom(Life_hier ~ GDPcapita_sq + Immun + Energy_Renewable
                       + Physicians, data = multi_nominal.df, reflevel = "1")
# Confidence interval
confint(mlogitnnet)
```

```
summary(mlogitnnet)

# Calculate z-score and p-value
z_score <- summary(mlogitnnet)$coefficients/summary(mlogitnnet)
$standard.errors
p_value <- (1 - pnorm(abs(z_score),0,1))*2
p_value
```