Life Expectancy Analysis and Prediction

Xinyue Li (Lydia), Yuqun Tang (Azura), Xuehan Zhao

Abstract

This paper focuses on some potential factors that might affect the life expectancy at birth. A prediction about the lifespan of people in a particular country will also be performed based on our model. Statistical methods such as model selection, ANOVA F test, multiple regression, and model diagnostics are used in this analysis. Based on the results, health expenditure per capita, the percentage of population with access to improved water source, child immunization against measles, and the continent the country belongs to clearly have some impacts on the life expectancy at birth.

Introduction

"Life expectancy in the United States has dropped again following last year's decline, which marked the first downturn in more than two decades." --- From CNN [1]

Some countries, such as the United States, spend 2-5 times more on healthcare than other developed countries while having shorter life expectancies. Does this imply that higher expenditure on healthcare do not necessarily lead to longer life expectancy? Our group is curious about this phenomenon and is interested in figuring out some potential factors that might affect the life expectancy of people and predicting how long a baby born in a particular country can expect to live.

Description

Background Information

We began the project by coming up with some factors that we think may have large impacts on life expectancy. We have considered factors from economics, environment, healthcare, diseases prevention, and food aspects. Then, we searched to see whether there are available datasets for these variables and manually combined them together.

Our dataset has 201 observations in total, which includes almost all countries around the world plus some Special Administrative Regions. All of the data were collected from the United Nations' *UN Data* [2], *The World Bank* [3], and *The Food and Agriculture Organization of the United Nations (FAO)* [4] databases. In order to make reasonable comparisons, all the data collected are for the same year 2014.

We also did literature review on past papers and similar datasets. For example, in the *Applied Linear Regression* [5] textbook, the author uses UN11 dataset to discover the effect of ppgdp, which is the gross national product per person in U.S dollar, and fertility, which is defined as the birth rate per 1000 females, on life expectancy. This dataset contains 199 observations across the world in 2009. The author concludes that life expectancy increases as log(ppgdp) increases, and decreases as fertility increases. Meanwhile, another research collected data for 175 countries, and grouped them according to the geographic position and income level for over 16 years (1995-2010) [6]. It applies a panel data analysis to estimate life expectancy by a function of health expenditure and concluded that there is a significant relationship between health expenditure and life expectancy. Comparing with those literature, which mainly discuss the impacts of one or two specific factors on life expectancy within a specific region or using time-series analysis, our group collect the dataset for the same year and take the whole world into account with possible factors coming from different aspects.

Variable Description

Our dependent variable is life expectancy at birth (measured in years). For independent variables, we chose GDP per capita (in US dollar), improved water source (percentage of population with access to clean water), PM2.5 air pollution (population exposed to levels exceeding WHO guideline value, in % of total), health expenditure per capita (in US dollar), child immunization against measles (% of children ages 12-23 months), and gross per capita Food Production Index Number (2004-2006=100). Moreover, we added a 6-level dummy variable region (Asia (1), Europe (2), North America (3), South America (4), Africa (5), and Oceania (6)) to categorize our data. Since countries in the same continent share similar

characteristics, we hope to use this variable region to control for factors such as food preferences, culture and habits which are hard to summarize in data but may affect life expectancy. The detailed meaning and definition of each variable is explained in the Table 1 in Appendix.

Summary Statistics

Table 2: Summary Statistics and Missing Values

	Exp	Health	Water	GDP	Measles	P.M 2.5	Food
Min	48.90	13.67	31.70	148.80	44.00	0.00	32.43
1st Quantile	66.00	88.58	81.75	1843.00	84.00	99.79	93.06
Median	74.00	358.60	95.35	6318.00	93.00	100.00	101.30
Mean	71.61	1164.00	88.28	17360.00	87.42	86.80	104.10
3rd Quantile	77.60	1100.00	99.60	21160.00	97.00	100.00	116.60
Max	89.69	9674.00	100.00	187600.00	99.00	100.40	159.90
NA's	0	18	9	3	16	15	12

The summary statistics and missing values for the variables are reported in Table 2. For outliers, there are some countries with unexpected high health expenditures, very small percentages of population with access to improved water source, or high GDP (Details see Table 3 in Appendix). Table 4 in Appendix shows that relatively high correlation exists between many pairs of regressors, especially between GDP and health expenditure. The distributions of raw data are shown in the Appendix (Figure 1). All the quantitative variables we considered are skewed except food and life expectancy. Specifically, PM2.5, measles and water are extremely left-skewed. GDP and health are extremely right-skewed.

Statistical Analysis

After having a general idea about the distributions of the variables and their relationship with the life expectancy, the response variable, we transform them. For right-skewed variables, we stretch out smaller x values and compress larger values by changing GDP and health to log(GDP) and log(health).

However, the right-skewed variables are in percentage. So instead of ascending the ladder, we do the following changes for each one: since the range of measles is from 40% to 100%, we transform it into a categorical variable by splitting it into 3 levels (0,50], (50,75], and (75, 100]; PM2.5 contains lots of small and large values, so we take the logit of PM2.5 in order to shrink the range; We use p = 9 to transform water. Details of transformations and the scatterplots of transformed variables are shown in Figure 2 and Figure 3 respectively in Appendix. We also check collinearity among independent variables before fitting a multiple regression. Below is the overall regression model:

$$Model\ (1) \qquad Life\ Expectancy = \beta_0 + \beta_1 \times log(health) + \beta_2 \times \frac{water^{.9}-1}{.9} + \beta_3 \times log(GDP) + \beta_4 logit(PM2.5) + \beta_5 food + \gamma_1 M_1 + \gamma_2 M_2 + \delta_1 R_1 + \delta_2 R_2 + \delta_3 R_3 + \delta_4 R_4 + \delta_5 R_5 + ... + \varepsilon$$

(Note: 1. M_1 = (50,75], M_2 = (75,100], R_1 = Europe, R_2 = North America, R_3 = South America, R_4 = Africa, R_5 = Oceania; 2. "……" represents all interaction terms in Model (1))

In this research, the 201 countries and regions are separated into training and test groups (by using validation set approach). We then apply the best model we get from the training group to the test group and compare the predicted values with the true data to see the accuracy of our model. To begin with, we use model selection (AIC, BIC, Cp and Adjusted R^2 as criteria) to determine the best model that explains the possible factors affecting the life expectancy. Then we test if the interaction terms in that best model are important by using ANOVA F-test. Next we apply the model we get to the test group and calculate the Mean Squared Error (MSE) for this model. In the end, by measuring the influence of some unusual observations, we want to better understand our dataset and which observations may affect the linear regression a lot.

Results

Collinearity

Table 4 in Appendix shows that relatively high correlation exists between many pairs of regressors, especially between GDP and health expenditure. So, we doubt that there could be a

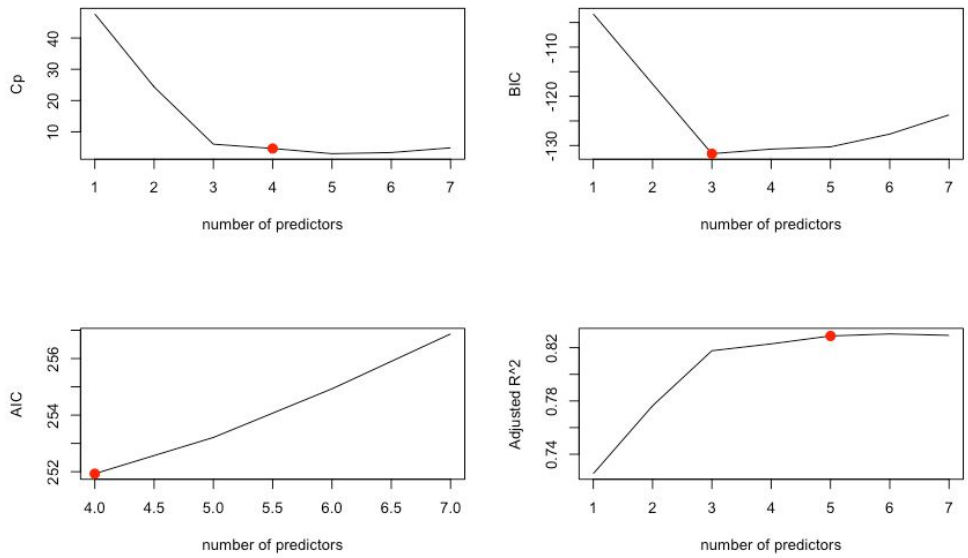
multicollinearity problem in the analysis. A check of Variance Inflation Factor (VIF) values in Table 5 below shows that the VIF values of log(health) and log(GDP) are greater than 5, indicating the presence of multicollinearity.

Table 5: VIF Results

Variable	GVIF	Df	GVIF^(1/(2*Df))	Variable	GVIF	Df	GVIF^(1/(2*Df))
log(health)	16.01	1	4.00	logit(pm2.5)	2.34	1	1.53
(water^9-1)/9	3.43	1	1.85	food	1.22	1	1.10
log(GDP)	13.12	1	3.622	region	5.99	5	1.20
measles	1.29	2	1.07				

Model selection

Figure 4: Model Selection



Then we decide to do the model selection to see whether we can drop some variables. The following four criteria are used to select the number of predictors: AIC, BIC, Cp and adjusted R^2. We want to find the model that leads to smallest AIC, smallest BIC, Cp that is closest to number of variables

in the model plus one relatively stable adjusted R^2. From Figure 4 above, the best model would be the one including four predictors: health, water, measles(75,100] and regoin5. For the predictor measles(75,100] and region5, we think that if one level of measles and region is significant, the variables measles and region should be considered significant. Therefore, we use health, water, measles and region to fit the regression. We then check the VIF values again. With all VIF < 5, there is no serious multicollinearity problem.

Interaction terms

Now, we know that health, water, measles, and region are the useful variables to explain the variation in life expectancy. Before performing a multiple regression, it is necessary to check if any interaction terms are needed. The models are:

Model (2): Full model

$$\begin{aligned} Life\ Expectancy\ &= \beta_0 + \beta_1 \times log(health) + \beta_2 \times \tfrac{water^9 - 1}{9} + \gamma_1 M_1 + \gamma_2 M_2 + \delta_1 R_1 \\ &\quad + \delta_2 R_2 + \delta_3 R_3 + \delta_4 R_4 + \delta_5 R_5 + all\ interaction\ terms + \varepsilon \end{aligned}$$

Model (3): Reduced model

$$\begin{aligned} Life\ Expectancy\ &= \beta_0 + \beta_1 \times log(health) + \beta_2 \times \tfrac{water^9 - 1}{9} + \gamma_1 M_1 + \gamma_2 M_2 + \delta_1 R_1 \\ &\quad + \delta_2 R_2 + \delta_3 R_3 + \delta_4 R_4 + \delta_5 R_5 + \varepsilon \end{aligned}$$

(Note: M_1 = (50,75], M_2 = (75,100], R_1 = Europe, R_2 = North America, R_3 = South America, R_4 = Africa, R_5 = Oceania)

The hypotheses are: H_0 : All the coefficients of interaction terms are zero

H_1 : At least one of them is not zero

We construct Model (2), which includes all the two-way and higher-way interaction terms, as the full model and Model (3), which does not have any interaction terms, as the reduced model. Then, an ANOVA F-test for reduced and full models are performed, and the p-value is 0.8388, which is larger than 0.05. Thus, we fail to reject the null hypothesis that all the coefficients of the interaction terms are zero. As a result, we decide to drop all the interaction terms and continue our analysis with Model (3).

Fit regression

Fitting Model (3) gives the coefficient estimates as shown in Table 6 below.

Table 6: Regression Results

	Estimate	Std. Error	t value
(Intercept)	49.37***	4.64	10.64
log(health)	2.19***	0.42	5.19
(water^9-1)/9	0.00***	0.00	3.59
M_1 = measles(50,75]	5.84	3.91	1.49
M_2 = measles(75,100]	7.61*	3.77	2.02
R_1 = region2	-1.51	1.26	-1.20
R_2 = region3	-1.13	1.40	-0.81
R_3 = region4	-2.70	1.46	-1.85
R_4 = region5	-6.02***	1.32	-4.55
R_5 = region6	-1.60	1.54	-1.03

Notes: 1. significance codes for p-values: 0 '***' 0.001 '**' 0.01 '*' 0.05; 2. Adjusted R-squared: 0.8246

The results in Table 6 give the regression equation as:

$$LifeExpectancy = 49.37 + 2.19log(health) + 6.22 * 10^{-17}\frac{water^9 - 1}{9}$$
$$+\begin{cases} 0, & measles \in (0, 50] \\ 5.84, & measles \in (50, 75] \\ 7.61, & measles \in (75, 100] \end{cases} + \begin{cases} 0, & Asia \\ -1.51, & Europe \\ -1.13, & NorthAmerica \\ -2.70, & SouthAmerica \\ -6.02, & Afica \\ -1.60, & Oceania \end{cases}$$

From Table 6, it is worth noticing that for a baby born in Asia, if the log of health is 0 (which means health expenditure per capita is \$1), water is 1% and the percentage of children got immunized against measles is between 0 to 50% (inclusive) in the country, he/she is expected to live about 49.37 years on average. In addition, the slope estimate of log(health), which is 2.19, is significant. By algebraic manipulation, we know that doubling the health expenditure per capita (log(2 × health)) will increase the life expectancy by about $2.19 \times log(2) \approx 1.52$ years on average, holding everything else constant.

Besides, the slope estimate of $\frac{water^9-1}{9}$ is significant. But, due to the complex data transformation, it is hard to interpret the exact effect of variable water on life expectancy. Instead, we could look at the effect plot of water, shown in Figure 5 in Appendix. In this plot, it could be noticed that the life expectancy does not change much when water is below about 75%. As soon as water reaches 75%, the life expectancy increases dramatically as water increases. Another significant result is 7.61, the coefficient estimate of M_2, which means 75% to 100% (include 100%) of children (ages 12-23 months) are immunized against measles. So, the estimation means that on average, a baby born in the country with 75% to 100% (include 100%) of children (ages 12-23 months) immunized against measles is expected to live about 7.61 years longer than one born in the country with 0 to 50% (inclusive) of children immunized against measles (the base category), holding everything else fixed. The last significant result is -6.02, the coefficient estimate of R_4, which is Africa. This means, a baby born in Africa is expected to live about 6.02 years less than a baby born in Asia (the base category) on average, holding everything else constant.

Therefore, from Table 6, we can conclude that health expenditure per capita, percentage of population with access to improved water source, child immunization against measles and the continent the country belongs to clearly have some impacts on the life expectancy at birth. Moreover, note that the adjusted R^2 is 0.8246, which is a very high value. This indicates that about 82.46% of variation in life expectancy is explained by log(health), $\frac{water^9-1}{9}$, measles and region.

Lastly, it is also worth attention that the coefficient estimate of Europe is -1.51. This means a baby born in Europe is expected to live 1.51 years less than a baby born in Asia (the base category) on average, holding other variables constant. Although this estimate is not significant, this result seems to be inconsistent with the generally higher life expectancy of Europe (than Asia) shown in the boxplot for variable region in Figure 1 in Appendix. A possible explanation is that in the randomly picked training set, nearly all the Asian countries with long life expectancies are selected into the set and many European countries with long life expectancies are not selected.

MSE & Prediction

Next, we apply the regression to the test set to calculate the Mean Squared Error (Test Error) of our model. The MSE is about 14.03 (Test Error: When we are giving a new country that is not in our training set, we will expect 14.03 years off the true life expectancy). Using the model, the expected life expectancy for the United Kingdom is 80.52 while the true value is 81.10; the expected life expectancy for Thailand is 74.53 while the true value is 74.40; the expected life expectancy for Ethiopia is 56.42 while the true value is 64.00. (Note: these three countries are in the test set.) So, our model does a good job in the predictions for developed and developing countries. However, the error of predicting some least developed countries in Africa, such as Ethiopia, is relatively large. A possible explanation comes from the way of splitting training and test sets. We randomly select half of data to be the training set, which happens to include only 37% (=18) of African countries. Since we do not have too much information on the African countries when fitting the regression, it makes sense to have large errors when predicting African countries in the test set.

Diagnostics

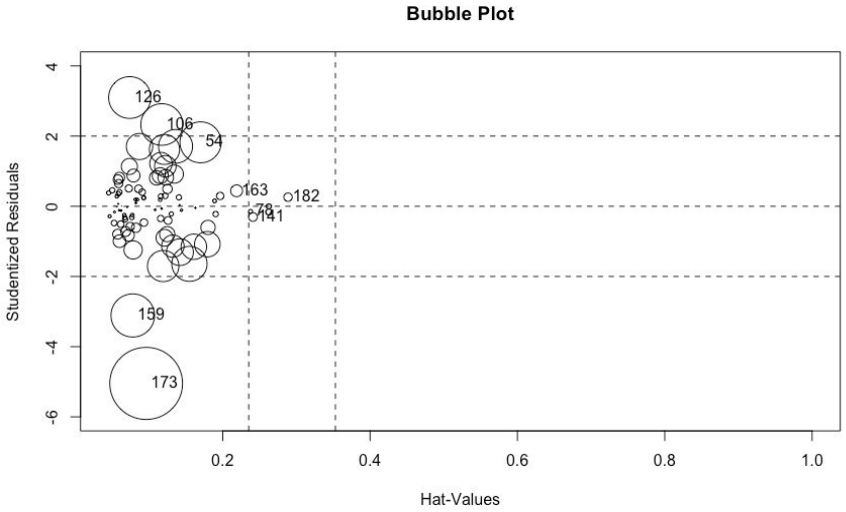


Figure 6: Bubble Plot of h_i(*hat values*), $E^{\star}{}_i$, D_i

From Figure 6 above, observations 173 (Swaziland), 159 (Sierra Leone), 126 (Morocco) and 106 (Libya) are beyond ± 2 on the Studentized-residual scale. So we reject the null hypothesis that those observations are not an outlier. Observations 173 (Swaziland), 159 (Sierra Leone) have high cook's distances. Observations 182 (Tonga), 78 (Haiti) and 141 (Papua New Guinea) have high leverages.

Table 7: Influence Analysis Summary

Influence Measures	Largest	Second largest
Cook's Distance	$D_{173} = 0.20$	$D_{159} = 0.07$
DFBETAS	$D^{\star}_{1,173} = -0.81$	$D^{\star}_{1,106} = 0.52$
	$D^{\star}_{2,173} = 0.73$	$D^{\star}_{2,106} = -0.44$
	$D^{\star}_{3,173} = -0.41$	$D^{\star}_{3,106} = 0.24$
	$D^{\star}_{4,173} = -0.59$	$D^{\star}_{4,106} = 0.31$
	$D^{\star}_{5,173} = 0.20$	$D^{\star}_{5,97} = -0.19$
	$D^{\star}_{6,90} = 0.36$	$D^{\star}_{6,191} = -0.27$
	$D^{\star}_{7,77} = 0.58$	$D^{\star}_{7,54} = 0.58$
	$D^{\star}_{8,173} = 0.75$	$D^{\star}_{8,126} = 0.58$
	$D^{\star}_{9,121} = -0.23$	$D^{\star}_{9,163} = 0.18$

Overall, observations 173 (Swaziland, the country with the lowest life expectancy) and 106 (Libya) appear to have the most influence on the results. Observation 173 (Swaziland) influences the intercepts for Measles in (0,50] and Region = Asia, Measles in (50, 75] and Region = Asia, Measles in (75, 100] and Region = Asia, Measles in (0,50] and Region = South America. It also influences the slopes β_1 *and* β_2. Observation 106 (Libya) influences the intercepts for Measles in (0,50] and Region = Asia, Measles in (50, 75] and Region = Asia. It also influences the slopes β_1 *and* β_2.

Discussion

There are some limitations in our study on life expectancy. Firstly, although we have transformed the highly skewed data, some data still exhibit non-symmetric distributions, especially pm2.5. These unusual values might affect the regression results, weakening our conclusions.

In addition, there is an omitted variable bias in the analysis. Although many control variables are included in Model (3), there are still lots of other important factors that may have impacts on the life expectancy but are not included due to lack of information, such as climate, lifestyle, dietary preferences, physical exercises, and even genetics. Also, there might be some lurking factors affecting the life expectancy that scientists have not yet discovered. These omitted variables could weaken the models, predictions and conclusions discussed above.

Moreover, there are measurement errors in the data collection process. Collecting data on factors such as the clean water source and PM 2.5 air pollution in a country, especially a large one, is never an easy task due to many practical reasons. When using instruments to measure a quantity, there could be measurement errors. The data collection techniques may also have limitations and could lead to biases. These measurement errors would lead to violation of the assumption (A5) that there is no randomness in explanatory variables and to larger uncertainty in the analysis.

Lastly, there are many missing data which reduce the size of the dataset to 173 observations. These missing values would reduce the sample size because the observations with missing data are omitted during regression analysis. Thus, the uncertainty in the analysis could be large and the conclusions could be weakened.

Future studies could include more explanatory variables to control for many other factors that could have large impacts on life expectancy. Also, instead of just using data for 2014, multiple years of data could be collected to form a panel dataset with which those unobserved characteristics that do not

change (or change slowly) over time could be accounted for. In this case, researchers could decide to use Fixed Effect (FE) or Random Effect (RE) models for analysis. In addition, future studies could just focus on one factor that we found useful. For example, researchers could conduct a deep analysis on the effect of health expenditure by dividing the spending into sub-categories, such as public spending on hospitals and spending on health insurance funds, to investigate and compare the effect of each component of health expenditure on the lifespan of people. This would give more detailed advice for policy makers.

Conclusion

The project aims to find the factors that affect the life expectancy at birth. Based on the statistical analysis above, we conclude that health expenditure per capita, percentage of population with access to improved water source, percentage of children (age 12-23 months) got immunized against measles and the continent the country belongs to clearly have some impacts on the life expectancy at birth.

References

[1]Tinker, Ben. "US life expectancy drops for second year in a row." 21 December 2017. 10 Mar 2018. <https://www.cnn.com/2017/12/21/health/us-life-expectancy-study/index.html>.

[2]*UN Data*. Vers. v0.14.6. 2017. United Nations Statistics Division. 10 Mar 2018. <http://data.un.org/>.

[3]*DataBank World Development Indicator*. 2017. The World Bank Group. 10 Mar 2018. <http://databank.worldbank.org/data/reports.aspx?source=2&series=SP.POP.TOTL.FE.ZS&country=>.

[4]*Food and Agriculture Organization of the United Nations*. 2017. 10 Mar 2018. <http://www.fao.org/faostat/en/#data/QI>.

[5]UN11." Weisberg, Stanford. *Applied Linear Regression*. Minneapolis: University of Minnesota, 2014. 9 Mar 2018.

[6]Jaba, Elisabeta. "The relationship between life expectancy at birth and health expenditures estimated by a cross-country and time-series analysis." *Procedia Economics and Finance* 15 (2014): 108-114. 9 Mar 2018. <https://www.sciencedirect.com/science/article/pii/S2212567114004547>.

Appendix

Table 1: Definitions of Variables

	Full name	Unit	Definition/Meaning
exp	Life expectancy	years	Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.
health	Government expenditure on health per capita	US dollar	Total health expenditure is the sum of public and private health expenditures as a ratio of total population. It covers the provision of health services (preventive and curative), family planning activities, nutrition activities, and emergency aid designated for health but does not include provision of water and sanitation.
water	Improved water source (% of population with access to)	%	Access to an improved water source refers to the percentage of the population using an improved drinking water source. The improved drinking water source includes piped water on premises (piped household water connection located inside the user's dwelling, plot or yard), and other improved drinking water sources (public taps or standpipes, tube wells or boreholes, protected dug wells, protected springs, and rainwater collection).
GDP	GDP per capita	US dollar	GDP per capita is gross domestic product divided by midyear population. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources.
measles	Child immunization against measles (% of children ages 12-23 months)	%	Child immunization, measles, measures the percentage of children ages 12-23 months who received the measles vaccination before 12 months or at any time before the survey. A child is considered adequately immunized against measles after receiving one dose of vaccine.
pm2.5	PM 2.5 air pollution (population exposed to levels exceeding WHO guideline value, in % of total)	%	Percent of population exposed to ambient concentrations of PM2.5 that exceed the WHO guideline value is defined as the portion of a country's population living in places where mean annual concentrations of PM2.5 are greater than 10 micrograms per cubic meter, the guideline value recommended by the World Health Organization as the lower end of the range of concentrations over which adverse health effects due to PM2.5 exposure have been observed.
food	Gross per capita Food Production Index Number		The FAO indices of agricultural production show the relative level of the aggregate volume of agricultural production for each year in comparison with the base period 2004-2006. They are based on the sum of price-weighted quantities of different agricultural commodities produced after deductions of quantities used as seed and feed weighted in a similar manner. The resulting aggregate represents, therefore,

			disposable production for any use except as seed and feed.
region	Region		A dummy variable. Includes 6 levels for the 6 continents around the world: Asia (1), Europe (2), North America (3), South America (4), Africa (5), and Oceania (6)

Note: Definitions/Meanings are from the *World Bank* [3] and the *FAO* [4] databases

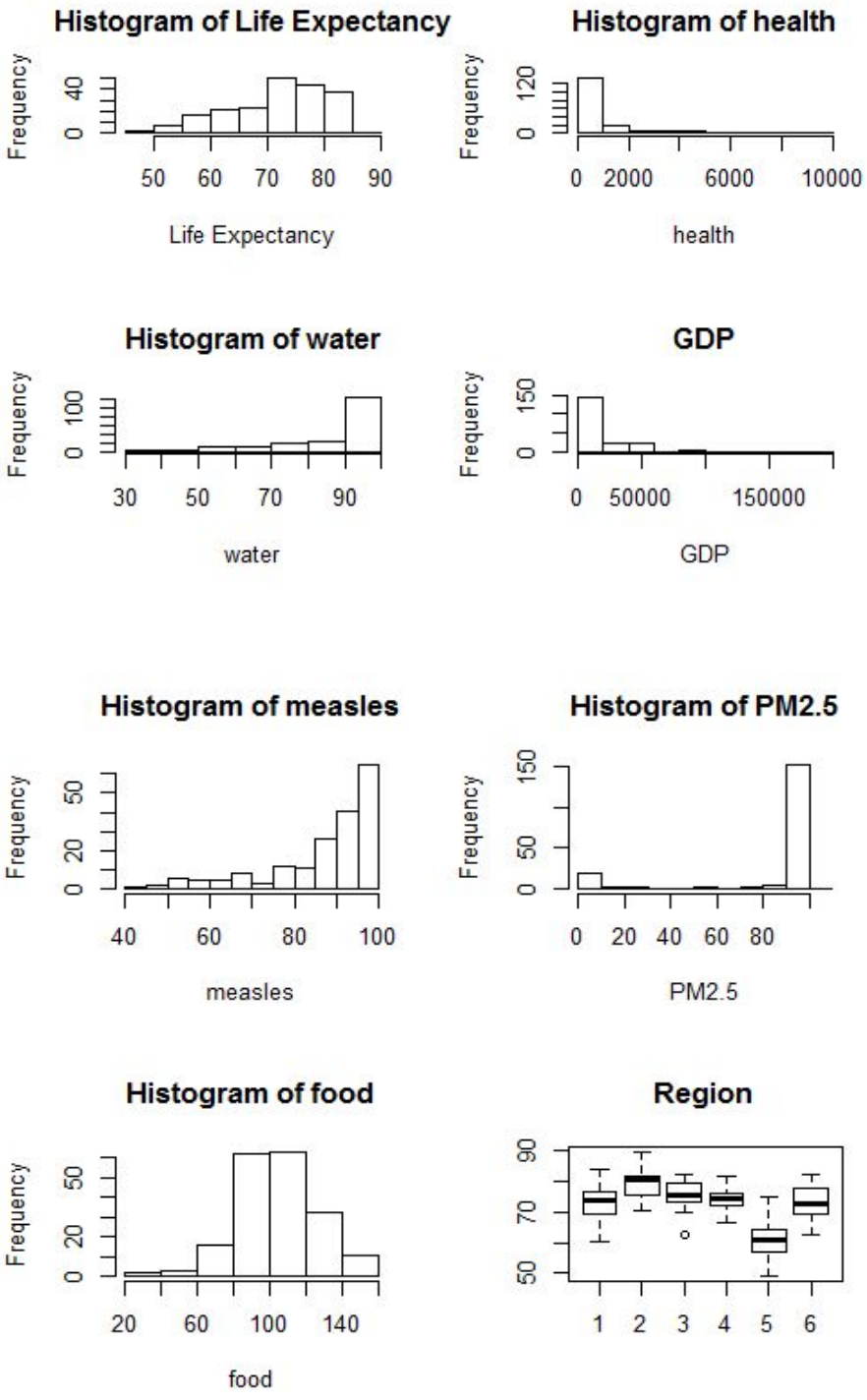
Table 3: Unusual Values

Exp	None
Health	26 outliers, the biggest two are Switzerland (9673.52) and Norway (9522.22)
Water	8 outliers, the smallest two are Somalia (31.70) and Papua New Guinea (40.00)
GDP	18 outliers, the biggest two are Monaco (187649.80) and Liechtenstein (178722.80)
Measles	18 outliers, the smallest two are Equatorial Guinea (44.00) and Somalia (46.00)
P.M 2.5	45 outliers, contains 15 zeros
Food	7 outliers, the biggest one is Lao PDR (159.92) and the smallest one is Hong Kong (32.43)

Table 4: Correlation

	Health	Water	GDP	P.M 2.5	Food
Health	1.00	0.81	0.95	-0.31	-0.16
Water	0.81	1.00	0.78	-0.21	-0.15
GDP	0.95	0.78	1.00	-0.27	-0.19
P.M 2.5	-0.31	-0.21	-0.27	1.00	0.09
Food	-0.16	-0.15	-0.19	0.09	1.00

Figure 1: Informative Plots



For boxplot: Asia(1), Europe(2), North America(3), South America(4), Africa(5), and Oceania(6)

Figure 2: Transformation Process

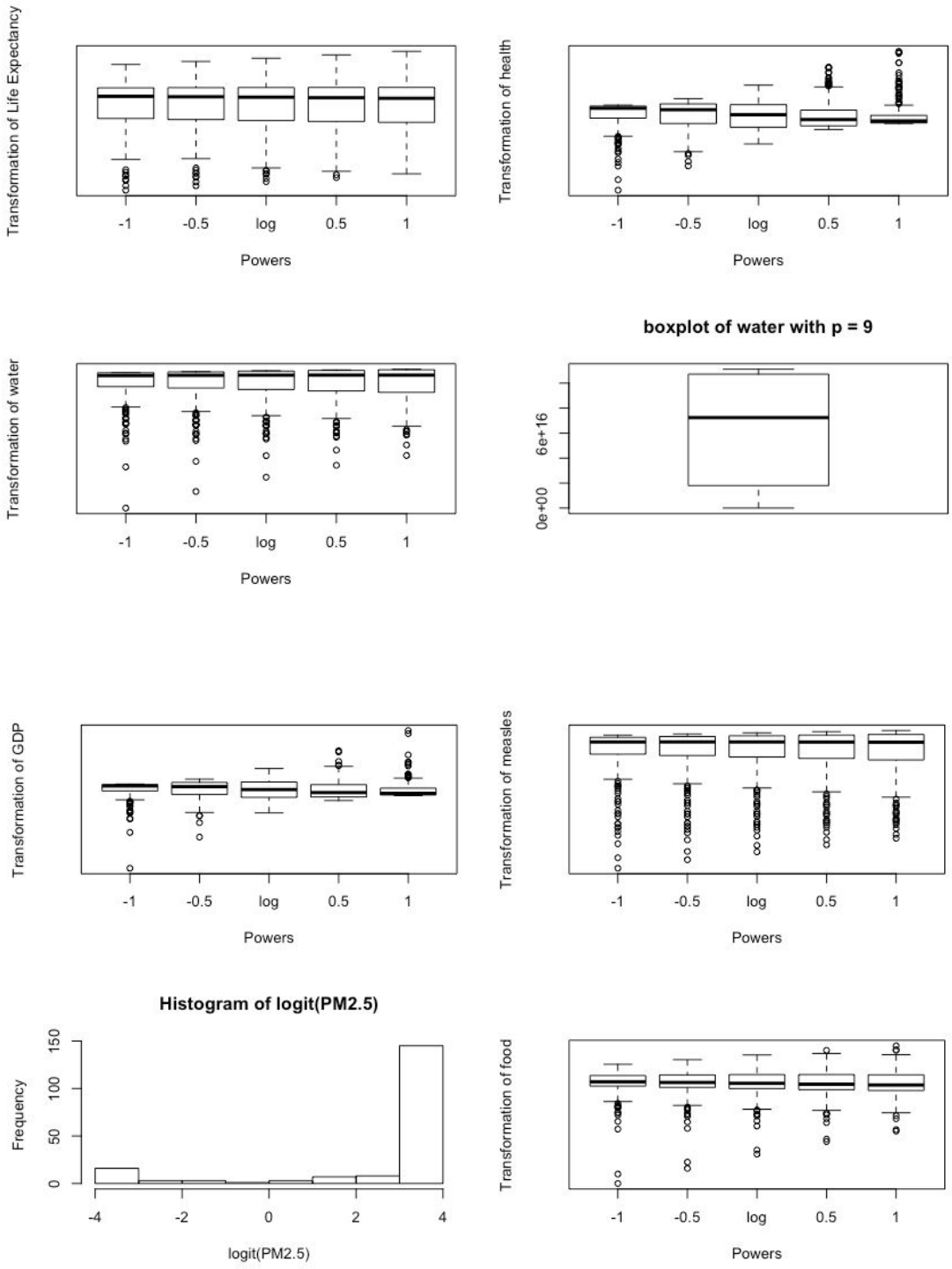
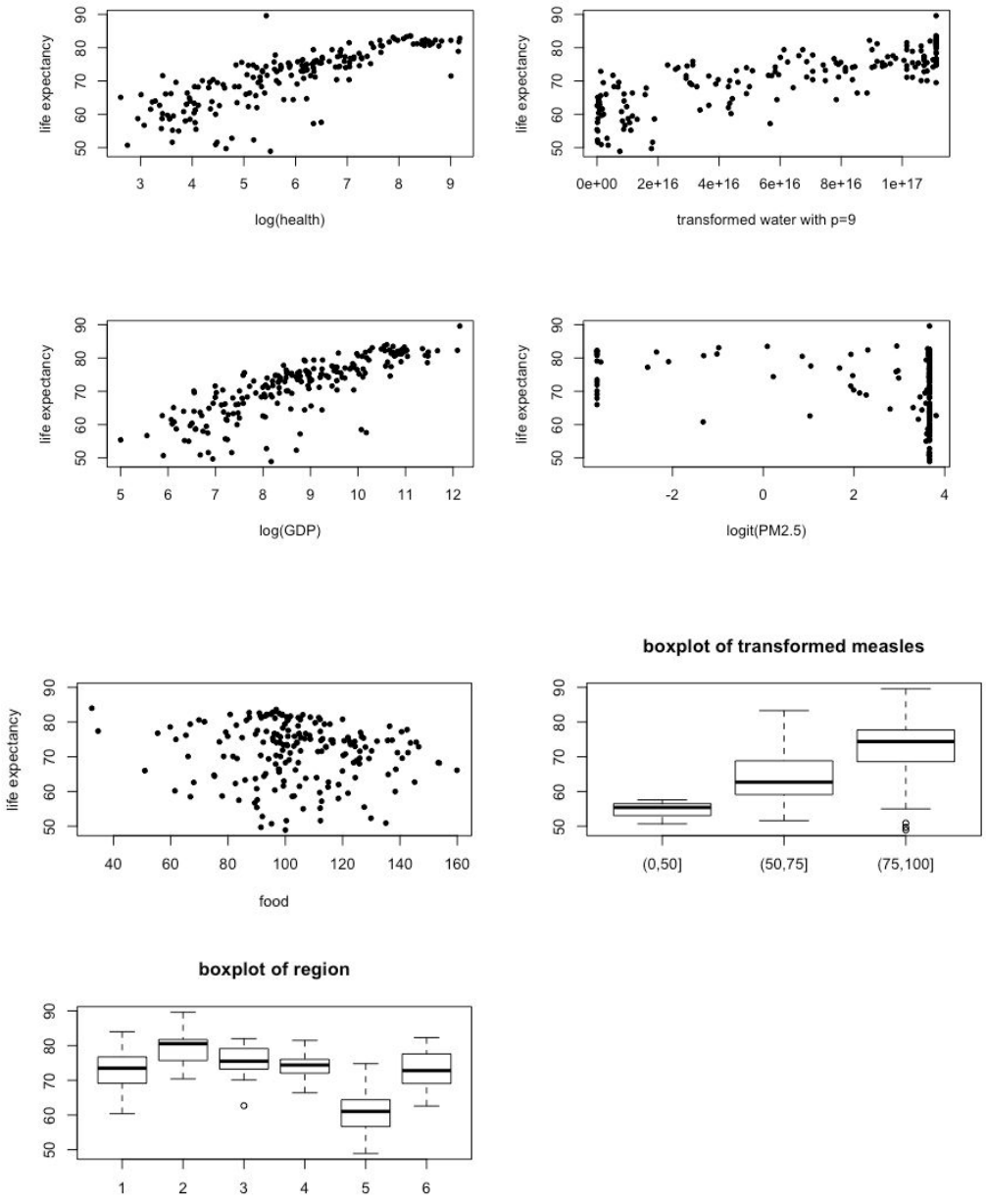


Figure 3: Scatterplots after Transformations



Note: In boxplot of region, Asia(1), Europe(2), North America(3), South America(4), Africa(5), and Oceania(6)

Figure 5: Effect Plot for Water