Regression Analysis
Group Project Final Report

Regression Analysis on Influential Factors of Life Expectancy

Jianlong LI Draco 1830026061
Ziwei CHEN Patrick 1830026017
Delin WANG Dylan 1830001082
Runjie SHEN Jay 1830026098
Xingming CHEN Carl 1830026012

Contents

I. Title

Regression Analysis on Influential Factors of Life Expectancy

II. Group Members with the Corresponding Contributions

Jianlong LI Draco

Research on multiple regression analysis.

Research on data with repeated observations of *developed countries*.

Report compiling.

Ziwei CHEN Patrick

Research on the relation between *life expectancy* and *schooling.*

Residual analysis on the multiple regression analysis.

Report compiling.

Delin WANG Dylan

PPT making.

Research on modeling the *life expectancy* and *GDP per capita*.

Runjie SHEN

Research on the relation between *life expectancy* and *BMI* value.

Residual analysis on the multiple regression analysis.

Collecting corresponding materials on the topic.

Xingming CHEN

Research on data with repeated observations of *developing countries*.

Data cleaning and selecting.

III. Abstract

In this report we analyze and present several models for the life expectancy in different countries, and examine some major factors associated with life expectancy. We obtained the data from the World Health Organization (WHO) on life expectancy with the corresponding data about years, schooling, medical condition score, etc. We model these individually by analyzing recent data since 2000, try to figure out if they have anything to do with the life expectancy and see whether there is a strong relationship. Then use a multiple regression model to analyze how important these factors are for the influence of life expectancy. We use our model to analyze what we can do to help increase the life expectancy in different countries in the years to come, like drink less alcohol, popularize higher education and so on.

IV. Introduction

Nowadays more and more people are paying attention to our life expectancy, especially the elderly ones. Life expectancy is defined statistically as the mean number of years remaining for an individual or a group of people at a given age. According to WHO, dramatic gains in life expectancy have been made globally since 2000, but there still exists major inequalities among different countries. What factors will affect our life expectancy? Different people in different countries have various ideas and points of view.

Our team gives some reasonable speculations and analysis on such problem. Our data is a dataset about life expectancy from WHO (World Health Organization) and we conduct some research on the influential factors of life expectancy. We also used a data set containing GDP per capita from World Bank database. Since there are many missing values and extreme values for some variables, we did some simple data cleaning operation first and selected 25 countries with highest representative and about 400 data in total from the data set and conduct the regression analysis.

Variable explanations:

Y: Life Expectancy (years)

$\mathbf{X_1}$: GDP per capita (GDP/Population) (USD$)

$\mathbf{X_2}$: Medical and public health condition score (0 is really bad and 100 is really good)

$\mathbf{X_3}$: BMI (Body Mass Index is a measure of body fat based on height and weight that applies to adult men and women)

$\mathbf{X_4}$: Mortality in special cases like war, riots, natural disasters, diseases spreading, famine, etc. (‰)

$\mathbf{X_5}$: Alcohol (The proportion of citizen in a certain country who often have alcoholic drinks) (%)

$\mathbf{X_6}$: Schooling (The proportion of entering university of a country) (%)

$\mathbf{X_7}$: Thinness proportion in 1-19 years (Juvenile and adolescent malnutrition) (%)

V. Methodology

The data we collected is from WHO (World Health Organization) and World Bank data (GDP per capita part). Our research aims at building a regression model to measure the life expectancy of human based on different factors. Begins with the simple linear regression analysis to multiple linear regression analysis, we go deeper to conduct the research. We mainly used Microsoft Excel, R programming language and MATLAB as tools to help us analyze the data. These tools are of high efficiency and convenience and free of any charge from the user. We have various analysis on simple linear regression and also multiple linear regression, which can draw some conclusions about these factors and life expectancy. But since the data we used is kind of common data and what we analyze are quantifiable factors. So our regression model maybe can't accurately fit all the situations of all the countries in the world. But it is a general model and can reflect many things.

VI. Data Analysis

Simple Linear Regression Analysis

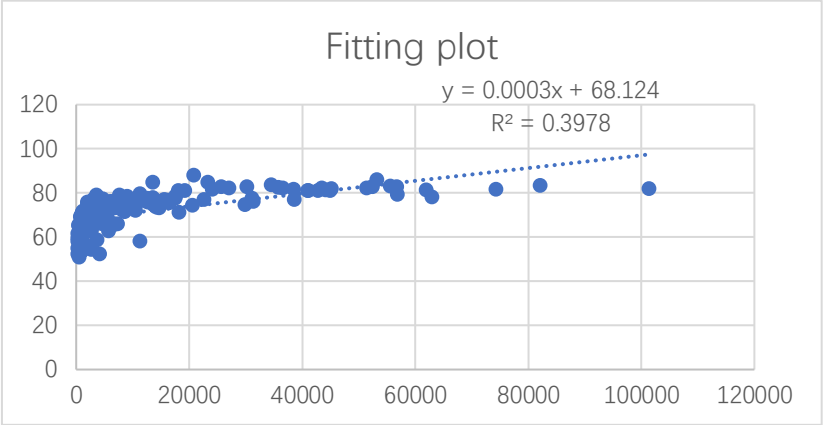
Regression Analysis Between Life Expectancy and GDP Per Capita

We select 178 countries' data of life expectancy in 2015 and their GDP per capita in 2015. We assume life expectancy as dependent variable Y and GDP per capita as X_1. Firstly, we assume the relation between the Y and X_1 is linear, and we take a simple linear regression analysis on it and get the model as well as the corresponding fitting curve.

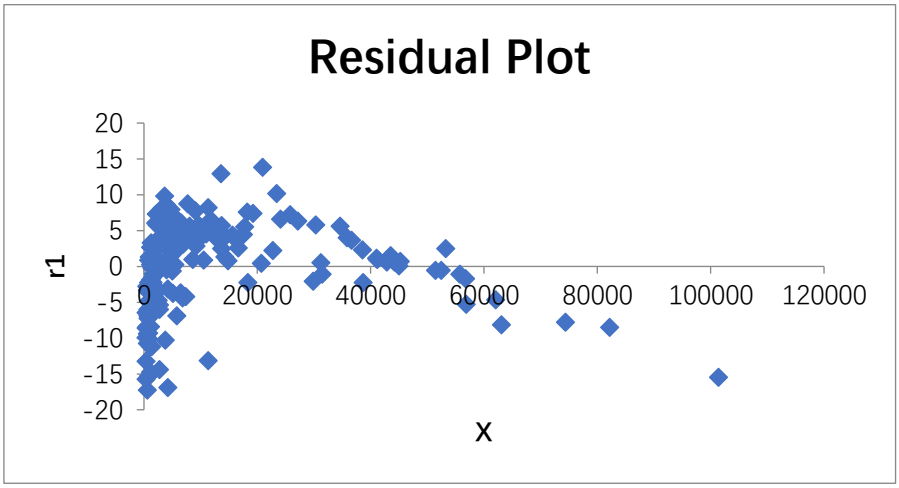
The linear function we get is

$$Y = 0.0003X_1 + 68.124.$$

The R^2 of this model is 0.3978, which is far from 1. Therefore, the fitting is bad and the plot we get is also not satisfying.



Below is the residual plot of X_1. The shape of the residual plot is like a left-opening megaphone, which means the relation between Y and X_1 is nonlinear.



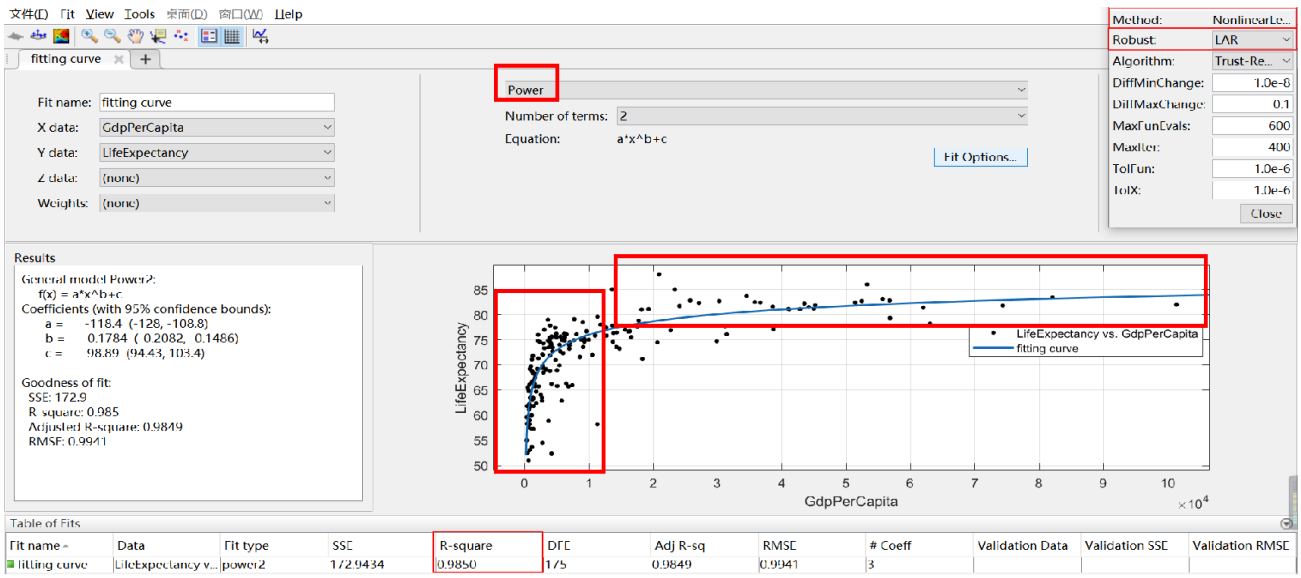
Therefore, we reject the model of simple linear regression and take a nonlinear regression analysis between Y and X_1.

We use the model of power and the open the Robust Estimate, we get the model with a power function:

$$f(x) = -118.4x^{-0.1784} + 98.89$$

The coefficient of determination R square is 0.9850, which is very close to 1. Therefore, the fitting is good, and the model we get is also good. The relation between Y and X_1 is nonlinear.

The following is the fitting curve we get using MATLAB.



According to the fitting curve, when X_1 is small, Y increases rapidly with the growth of X_1. When X_1 getting large, the increase rate of Y slows down.

We also conduct some predictions below.

Situation 1:

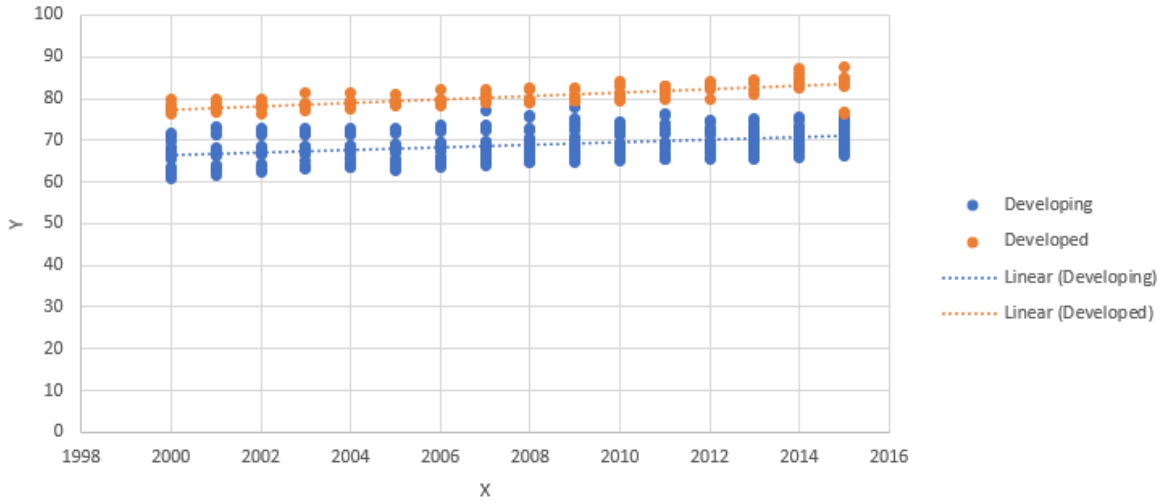
If *GDP Per Capita* increases from $\$ 0.5 \times 10^4$ to $\$ 1 \times 10^4$ (rose by \$5000), *Life Expectancy* increases from 72.99 years old to 75.99 years old. (rose by 3 years old)

Situation 2:

If *GDP Per Capita* increases from $\$ 2 \times 10^4$ to $\$ 2.5 \times 10^4$ (rose by \$5000), *Life Expectancy* increases from 78.66 years old to 79.45 years old. (rose by 0.79 years old)

The *GDP Per Capita* in situation 1 and 2 both increases by 5000 USD, but *Life Expectancy* in situation 1 increases in a larger degree than situation 2. Our prediction confirms our assumption and our model. In this point of view, as many developing countries' economy booming, their citizen's life expectancy will grow in a more significant and appreciable amount.

Regression Analysis of Life Expectancy in Different Years Between Developing and Developed Countries



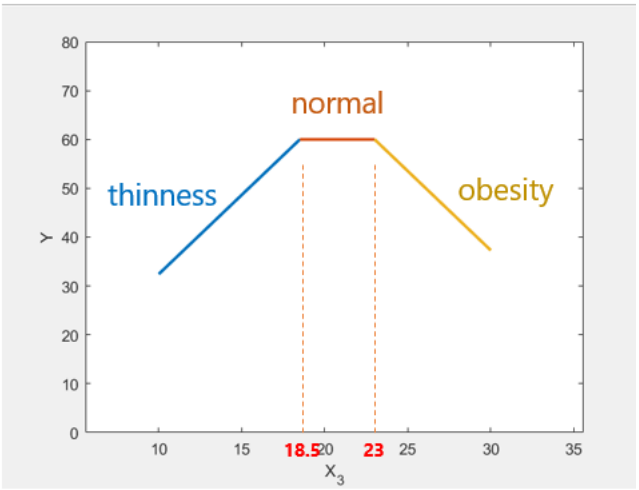
For using different years as X and life expectancy as the dependent variable, we try to analyze the life expectancy in different years. First, before making the analysis, we have to screen the data, for the life expectancy in different countries is very different. After finding the median for developed and developing countries correspondingly, we screened out some data that are concentrating and surrounding the median, and used them to draw the scatter diagram. After doing these, with the coefficient and the scatter diagram, we can see the years and the life expectancy is in a positive linear relation. We can observe that developed countries' life expectancy is generally higher than the developing ones and we can also see the trend for both developing countries and developed countries are increasing for life expectancy as time goes by.

Regression Analysis Between Life Expectancy and BMI Value

Let's go on with Body Mass Index (BMI) which means a measure of body fat based on height and weight that applies to adult men and women. It is calculated by dividing the weight in kilograms by the square of the height in meters. And we assume it to be X_3 here.

$$BMI = \frac{kg}{m^2}$$

According to the normal range, the BMI is $18.5 \leq BMI < 23$, so we use this to do subsection research. As being too fat or too thin is not good for our health, so before the research we assumed that the relation model between BMI and life expectancy maybe as follow. It probably means that a person will live longer if his BMI is in this normal range.



However, our assumption seems incorrect. Below are the outputs. The regression statistics, ANOVA tables and regression equations for each interval are as follows.

X_3 < 18.5:

SUMMARY OUTPUT

Regression statistics	
Multiple R	0.563277
R Square	0.317281
Adjusted R Square	0.300213
Standard error	5.299898
Observation	40

ANOVA Table

Source	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	522.1531	522.1531	17.65982	0.000103
Residual	38	1123.557	29.56729		
Total	39	1645.71			

	Coefficients	Standard error	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	78.30751	5.971824	13.11283	4.55E-16	66.23801	90.37702
X3	-1.59573	0.370108	-4.31153	0.000103	-2.34375	-0.84772

$Y = 78.30751 - 1.59573\ X_3$

18.5 ≤ X_3 < 23:

SUMMARY OUTPUT

Regression statistics	
Multiple R	0.426187
R Square	0.181635
Adjusted R Square	0.160652
Standard error	9.395667
Observation	40

ANOVA Table

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	764.1412	764.1412	8.43407	0.005463
Residual	38	3442.864	90.60168		
Total	39	4207.005			

	Coefficients	Standard error	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-9.2962	22.76078	-0.40843	0.685192	-55.3342	36.74183
X3	3.219937	1.09443	2.942112	0.005463	1.006242	5.433631

$$Y = -9.2962 + 3.219937\ X_3$$

X_3 > 23:

SUMMARY OUTPUT

Regression statistics	
Multiple R	0.569289
R Square	0.32409
Adjusted R Square	0.319729
Standard error	9.203157
Observation	200

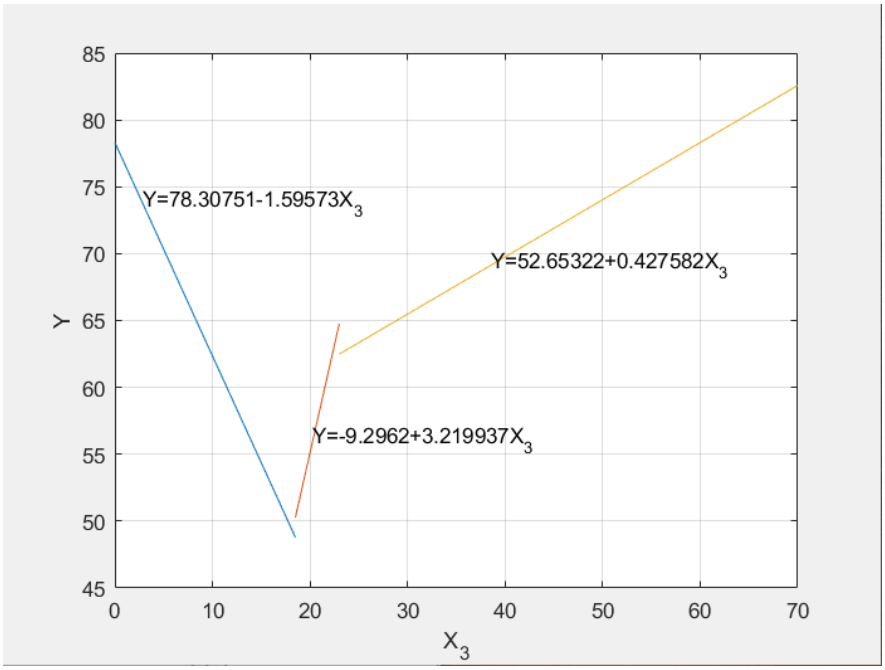
ANOVA Table

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	6294.792	6294.792	75.75884	7.27E-15
Residual	158	13128.2	83.08987		
Total	159	19423			

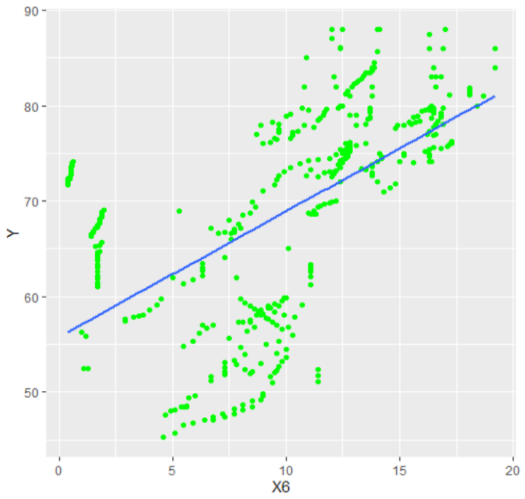
	Coefficients	Standard error	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	52.65322	2.36924	22.22367	4.7E-50	47.97305	57.33338
X3	0.427582	0.049598	8.620925	7.27E-15	0.329606	0.525558

$$Y = 52.65322 + 0.427582\ X_3$$

Then we put these 3 fitting functions into Rectangular Coordinates and we found that in the range of $X_3 < 18.5$, life expectancy will show a decreasing trend with the increase of X_3. And in the range of $X_3 > 18.5$, life expectancy will increase as BMI increases. But don't worry, this doesn't mean that you should gain weight and make your BMI bigger in order to live longer. Don't forget that in these 3 models, their R-Squares are very small and not very closed to 1. So the fitting of models are not good. Then we can draw a conclusion that there is no obvious linear relationship between BMI and life expectancy.



Regression Analysis Between Life Expectancy and Schooling




```
Call:
lm(formula = Y ~ X6, data = LEDX6)

Residuals:
    Min      1Q  Median      3Q     Max
-19.684  -7.116   1.679   6.514  17.504

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.76641    1.04684   53.27   <2e-16 ***
X6           1.31733    0.09447   13.95   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.957 on 398 degrees of freedom
Multiple R-squared:  0.3282,    Adjusted R-squared:  0.3265
F-statistic: 194.5 on 1 and 398 DF,  p-value: < 2.2e-16
```

We try to analyze the influence of schooling on life expectancy. Surprisingly, it turns out that the two are linearly related. From this graph, we can find that the two are linearly correlated positively, so we will put schooling into the multiple regression model later.

Probably, with higher education, people have more knowledge about health and pay more attention to their health care. They possibly have more money and time to deal with health issues. What's more, with higher education, they are more likely to live and work in better conditions rather than places that are dirty or dangerous. That is the most possible reason why the two are correlated.

Multiple Regression Analysis

Since X_3 (BMI value) we mentioned before it is not fitting simple linear regression, so we try to use quadratic form to conduct the multiple linear regression analysis thereafter.

```
> reg1

Call:
lm(formula = Y ~ X2 + X3 * X3 + X4 + X5 + X6 + X7, data = LED)

Residuals:
     Min       1Q   Median       3Q      Max
-12.2427  -1.4674   0.0128   0.9904   8.4062

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 73.604842   1.638963  44.909  < 2e-16 ***
X2           0.087916   0.019578   4.490 9.35e-06 ***
X3           0.069561   0.013854   5.021 7.80e-07 ***
X4          -0.069720   0.002401 -29.042  < 2e-16 ***
X5          -0.041824   0.044951  -0.930  0.35272
X6           0.117117   0.036174   3.238  0.00131 **
X7          -0.023848   0.042933  -0.555  0.57889
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.354 on 393 degrees of freedom
Multiple R-squared:  0.9542,    Adjusted R-squared:  0.9535
F-statistic:  1364 on 6 and 393 DF,  p-value: < 2.2e-16
```

We can see

$$R^2 = 0.9542$$

$$R^2_{adj} = 0.9535$$

```
> LED=read.csv(file.choose(),header=T)
> LED2=lm(Y~X2+X3*X3+X4+X5+X6+X7,data=LED)
> reg1=summary(LED2)
> anova(LED2)
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq   F value    Pr(>F)
X2          1  38833   38833 7007.1144 < 2.2e-16 ***
X3          1   1100    1100  198.4830 < 2.2e-16 ***
X4          1   5353    5353  965.9624 < 2.2e-16 ***
X5          1      1       1    0.2368 0.6268059
X6          1     66      66   11.8557 0.0006367 ***
X7          1      2       2    0.3085 0.5788905
Residuals 393   2178       6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Above is the ANOVA table. We can retain all the X except X_5 and X_7 since the $F < F_{1,393}$ and the p value is also bigger than 0.05. Therefore, we cannot reject the null hypothesis H_0 for X_5 and X_7.

Following we make further analysis on the backward elimination.

```
> max.model = lm(Y ~ X2+X3*X3+X4+X5+X6+X7, data=LED)
> smallest <- formula(lm(Y~1,LED))
> bwd.model = step(max.model, direction='backward', scope=smallest)
Start:  AIC=691.87
Y ~ X2 + X3 * X3 + X4 + X5 + X6 + X7

       Df Sum of Sq    RSS     AIC
- X7    1       1.7 2179.7  690.19
- X5    1       4.8 2182.8  690.75
<none>              2178.0  691.87
- X6    1      58.1 2236.1  700.40
- X2    1     111.7 2289.7  709.89
- X3    1     139.7 2317.7  714.74
- X4    1    4674.1 6852.1 1148.34

Step:  AIC=690.19
Y ~ X2 + X3 + X4 + X5 + X6

       Df Sum of Sq    RSS     AIC
- X5    1       5.2 2184.9  689.14
<none>              2179.7  690.19
- X6    1      65.7 2245.4  700.07
- X2    1     135.6 2315.3  712.33
- X3    1     175.7 2355.4  719.20
- X4    1    4687.9 6867.6 1147.24


Step:  AIC=689.14
Y ~ X2 + X3 + X4 + X6

       Df Sum of Sq    RSS     AIC
<none>              2184.9  689.14
- X6    1      61.8 2246.7  698.30
- X2    1     149.5 2334.4  713.61
- X3    1     177.6 2362.4  718.39
- X4    1    5415.0 7599.8 1185.77
> summary(bwd.model)

Call:
lm(formula = Y ~ X2 + X3 + X4 + X6, data = LED)

Residuals:
     Min       1Q   Median       3Q      Max
-12.2205  -1.4691  -0.1168   1.1000   8.5483

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 73.68879    1.17688  62.614  < 2e-16 ***
X2           0.08225    0.01582   5.199 3.22e-07 ***
X3           0.07277    0.01284   5.666 2.82e-08 ***
X4          -0.07040    0.00225 -31.288  < 2e-16 ***
X6           0.10753    0.03216   3.343 0.000907 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.352 on 395 degrees of freedom
Multiple R-squared:  0.954,     Adjusted R-squared:  0.9536
F-statistic:  2050 on 4 and 395 DF,  p-value: < 2.2e-16
```

Therefore, after the backward elimination we eliminated X_5 and X_7. X_5 represents alcohol and X_7 represents thinness proportion in juvenile and adolescent period, which is not fitting our model here through the examination of F value, p value and through backward elimination. We obtained a new multiple linear regression function:

$$Y = 73.69 + 0.082X_2 + 0.073X_3^2 - 0.07X_4 + 0.108X_6$$

with $R^2 = 0.954$ and $R_{adj}^2 = 0.9536$.

From the multiple linear regression function above we can see that the medical and public health condition score (X_2) and the university entry proportion of a country (X_6) has a positive relation on the life expectancy (Y). In addition, the mortality (X_4) has negative relation upon the life expectancy (Y).

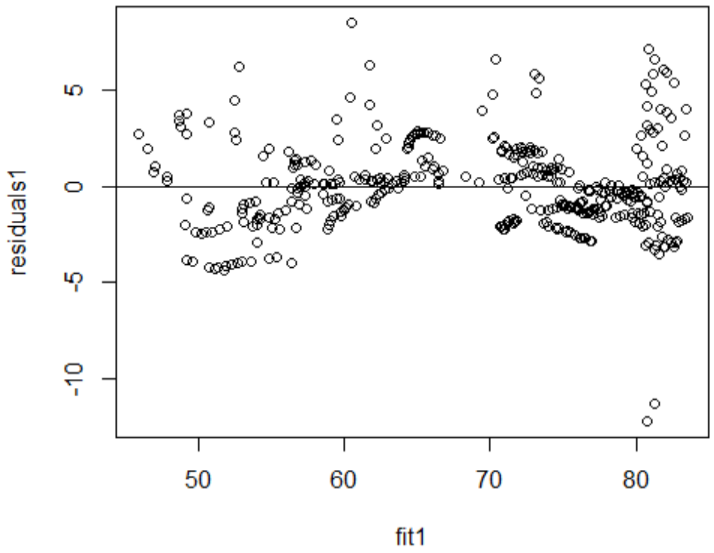
```
> anova(LED3)
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq  F value    Pr(>F)
X2          1  38833   38833 7020.547 < 2.2e-16 ***
X3          1   1100    1100  198.863 < 2.2e-16 ***
X4          1   5353    5353  967.814 < 2.2e-16 ***
X6          1     62      62   11.178 0.0009067 ***
Residuals 395   2185       6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Above is the ANOVA table of our new multiple regression model. All the p values are smaller than 0.05 and F > $F_{1,\ 395}$ which means we reject all the null hypothesis and the variables are valid.

```
> LED3=lm(Y~X2+X3*X3+X4+X5+X6+X7,data=LED)
>
> X=cbind(1,LED$X2,LED$X3,LED$X4,LED$X5,LED$X6,LED$X7)
> H=X%*%solve(t(X)%*%X)%*%t(X)
> Hjj=diag(H)
> PRESS=sum((LED3$residuals/(1-Hjj))^2)
>
> V=cbind(1,LED$X2,LED$X3,LED$X4,LED$X6)
> H=V%*%solve(t(V)%*%V)%*%t(V)
> Hjj=diag(H)
> PRESS2=sum((LED3$residuals/(1-Hjj))^2)
> PRESS
[1] 2271.573
> PRESS2
[1] 2248.133
```

We can also see the PRESS comparison. Without X_5 and X_7, the PRESS decreases and indicates that the model can better do the prediction. Alcohol (X_5) should have negative relation of life expectancy in our common sense. Since some Islam countries don't have alcoholic drinks that much and their life expectancy are supposed

to be high. But there exist some Islam countries like Yemen with low alcoholic level but also low life expectancy since they are always at war. So X_5 can't fit the model quite well in this point of view. But it doesn't mean that the alcohol won't affect our life expectancy. From the full model before we can observe that the coefficient of X_5 is negative, which means it exert a negative effect on our life expectancy. For X_7 represent thinness proportion in juvenile and adolescent period. The lower the proportion is, the higher the dependent variable Y it should be. That's correct in our common sense. But in some western countries, the thinness proportion in juvenile and adolescent did in a low level, but their life expectancy is not that high since the obesity in their society also plays a significant role. In our model the X_3(BMI) can also show such trend and it can fit the model better. So we eliminate X_7 in our multiple regression model.



Then we consider to have a residual plot. We can see there exists a few outliers but all the points are generally surrounding the y = 0 line, which indicates it is a normal one and the model do not need to do any further transformation.

```
> library(car)
> LED5=lm(Y~X2+X3*X3+X4+X6,data=LED)
> outlierTest(LED5)
     rstudent unadjusted p-value Bonferroni p
295 -5.473489         7.8680e-08   3.1472e-05
292 -5.045499         6.9199e-07   2.7680e-04
```

There exist many outliers but these two are the biggest ones shown by R.

We can see that there are two outliers that are clearly away from the line. There could be two reasons to explain these outliers. First, the extreme performance of the variation.

The data is true and maybe something unusual happens in the real world, so there are few numbers in extreme value. Secondly, due to the specific testing method and way of counting number, or because of some mistakes happens in the process of collecting the data set. So, we carefully examine these two outliers in the data set and find where the two outliers are.

Here is what we found. The two outliers are at the 293th and 296th row as the image shows:

290	71.1	60.4	44.9	166	2	9	3.3	Kyrgyzstar
291	78	59.7	43.9	170	2	8.9	3.3	Kyrgyzstar
292	77	59.2	43	174	2	8.7	3.3	Kyrgyzstar
293	69.9	58.7	42.2	18	0.01	8.5	3.2	Kyrgyzstar
294	69.4	58.2	41.4	188	2.13	8.65	3.2	Kyrgyzstar
295	68.8	57.7	40.7	199	2.39	8.4	3.2	Kyrgyzstar
296	68.5	57.4	40.1	21	2.48	8.1	3.3	Kyrgyzstar
297	67.6	56.7	39.5	217	2.53	7.9	3.3	Kyrgyzstar
298	67.2	54	39	229	2.62	8	3.3	Kyrgyzstar
299	66.7	54.3	38.5	234	2.73	7.7	3.3	Kyrgyzstar
300	66.9	54.6	38	224	2.77	7.7	3.4	Kyrgyzstar
301	67.1	54.9	37.5	218	2.81	7.7	3.4	Kyrgyzstar
302	66.6	55.2	37	217	3.28	7.6	3.5	Kyrgyzstar
303	66.7	53.6	36.5	215	3.31	7.3	3.5	Kyrgyzstar
304	67.2	53.4	36	217	3.41	6.7	3.6	Kyrgyzstar
305	66.6	52.4	35.6	225	3.52	7	3.6	Kyrgyzstar

We can see the column where these outliers in represent the mortality in special cases of Kyrgyzstan. We also compared them to the mortality of other countries but could not find any number that are way off the line like these two, so probably these are caused by mistakes in collecting the data. Therefore, we deleted the outliers and replace them with numbers that reasonably fit. (*multiple LED4.csv* is the one with data correction upon these two)

Though there are only two outliers, that does not mean the model is very perfect, because we can see still many data is far away from the regression line but were not counted as outliers.

```
> LED=read.csv(file.choose(),header=T)
> LED4=lm(Y~X2+X3*X3+X4+X6,data=LED)
> regl=summary(LED4)
> regl

Call:
lm(formula = Y ~ X2 + X3 * X3 + X4 + X6, data = LED)

Residuals:
   Min     1Q Median     3Q    Max
-4.145 -1.530 -0.296  1.154  8.451

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 76.270203   1.151045  66.262  < 2e-16 ***
X2           0.059463   0.015107   3.936 9.78e-05 ***
X3           0.069106   0.012047   5.736 1.93e-08 ***
X4          -0.075351   0.002204 -34.192  < 2e-16 ***
X6           0.105770   0.030130   3.511 0.000499 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.204 on 395 degrees of freedom
Multiple R-squared:  0.9596,    Adjusted R-squared:  0.9592
F-statistic:  2347 on 4 and 395 DF,  p-value: < 2.2e-16
```

We conduct the multiple regression analysis again with the adjusted data table (LED4) and obtained the following regression function:

$$Y = 76.27 + 0.06X_2 + 0.069X_3^2 - 0.075X_4 + 0.106X_6$$

with $R^2 = 0.9596$ and $R_{adj}^2 = 0.9592$.

Then we conduct the residual analysis again.

```
> outlierTest(LED4)
  rstudent unadjusted p-value Bonferroni p
7  3.91479          0.0001066      0.042641
```

From R programming we can see there still exist one big outlier. That's the 7th data in our data set. (Djibouti in 2009) It shows the life expectancy in Djibouti improved a lot in 2009. By researching about that we found Djibouti started to work with the United Nations High Commissioner for Refugees (UNHCR) to give refugees a special ID cards and provide help for their survival. This kind of act legalize the refugees in their country and greatly improve their living condition. This is probably the reason that the country will increase their live expectancy in that year.

```
> confint(LED4,level=0.95)
                   2.5 %      97.5 %
(Intercept) 74.00726266 78.53314276
X2           0.02976349  0.08916330
X3           0.04542144  0.09278958
X4          -0.07968376 -0.07101872
X6           0.04653605  0.16500459
```

Above is the confident interval of corresponding β in our new multiple regression function.

```
> vif(LED4)
      X2       X3       X4       X6
6.703901 3.777437 4.909426 1.679666
```

Calculate the VIF value of the multiple linear regression model. All of the VIF are smaller than 10, thus indicates there's no obvious multicollinearity between these variables. The model is generally good.

```
> lm.predict<-lm(Y~X2+X3*X3+X4+X6,data=LED)
> newdata<-data.frame(X2=75,X3=21,X4=10,X6=20)
> predict(lm.predict,newdata)
       1
83.54307
```

We also conduct a prediction using multiple regression model with medical and public health condition scores to 75, BMI value equals to 21, mortality in special cases is 10 ‰ and university entry proportion is 20%. By our model the life expectancy is 83.54 years old. That's a kind of high standard for a country.

VII. Conclusion

When it comes to simple regression model for each independent variable, we figure out that there is a positive relationship between GDP, schooling and life expectancy while GDP and life expectancy shows a non-linear positive relationship. However, the same cannot be said for BMI. We have also made a multiple regression model, showing that how these different factors jointly influence the life expectancy of the country. We adjust the BMI variable to a quadratic form into our multiple regression model to have a better fitting.

The world has made a great effort on reducing the sufferings and premature deaths that arise from preventable and treatable diseases in order to help people gain a longer life expectancy. But the gains have been uneven. Many developing countries are still suffering a lot from different diseases and other crisis that the life expectancy of their country is staggering short. Supporting countries to move towards universal health coverage based on strong primary care is the best thing we can do to make sure no-one is left behind.

Overall, this is all the results that our group has found. We put a great effort into the project and had some valuable feedbacks and findings by our research.

VIII. References

Laursen, T. M., Munk-Olsen, T., & Vestergaard, M. (2012). Life expectancy and cardiovascular mortality in persons with schizophrenia. *Current Opinion in Psychiatry*, 25(2), 83–88. doi: 10.1097/yco.0b013e32835035ca

Diehr, P., Omeara, E. S., Fitzpatrick, A., Newman, A. B., Kuller, L., & Burke, G. (2008). Weight, Mortality, Years of Healthy Life, and Active Life Expectancy in Older Adults. *Journal of the American Geriatrics Society*, 56(1), 76–83. doi: 10.1111/j.1532-5415.2007.01500.x

Galor, O., & Moav, O. (2005). Natural Selection and the Evolution of Life Expectancy. *SSRN Electronic Journal*. doi: 10.2139/ssrn.563741

Santrock, John (2007). Life Expectancy. A Topical Approach to: Life-Span Development (pp. 128–132). New York, New York: The McGraw-Hill Companies, Inc.