Statistically Analysis of Life Expectancy of Countries

Junhui Li, Bin Li

1 Data Source and initial Regression

1.1 Data Source

We get the data sources from the World Bank, it has different datasets which include data in various aspects at country level. Although some of the data sets have updated to 2014, the most completed datasets are all updated to 2013. Therefore. We choose the data in the year in 2013 as our data sources.

We want to examine the life-expectancy on the country-level, Life Expectancy could be impacted by several factors, after we examined the data we have, we divide the factors into several categories: population, economics, facilities, environment and disease.

1.2 Initial Regression

World Bank has various categories of data sets, after researching online about factors of life expectancy, we exclude those who are insufficient or clearly unrelated. And finally, we ended up with 8 indicators which are fitted into 6 categories that we think are important:

For population, we choose urban population rate as the factor;
For economics, we choose the GDP per capita and the Health Expenditure rate in total GDP
For facilities, we choose the Sanitation Facilities rate
For environment, we choose the PM2.5 rate and clean water resources rate
For disease, we choose the Tuberculosis rate
For health, we choose Fertility per woman

Initially, we choose linear regression as our basic model:

life_expectancy~gdp_per_capita+fertility_per_woman+pm2.5+tuberculosis+urban_population+forest_area+health_expenditure_rate+water_resources_rate+Improved_sanitation_facilities_rate

Then we run it in R to come up the summary table (See Figure 1). Based on that, we could observe each coefficients of all 8 indicators and determine how good our regression is to explain the date.

```
Call:
lm(formula = life_expectancy ~ gdp_per_capita + fertility_per_woman +
    pm2.5 + tuberculosis + urban_population + forest_area + health_expenditure_rate +
    water_resources_rate + Improved_sanitation_facilities_rate,
    data = lifeExpectancy)

Residuals:
     Min       1Q   Median       3Q      Max
-17.1366  -1.5457   0.3088   1.7266  11.0280

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        7.451e+01  4.798e+00  15.529  < 2e-16 ***
gdp_per_capita                     5.286e-05  1.879e-05   2.813  0.00559 **
fertility_per_woman               -3.120e+00  3.537e-01  -8.822 3.02e-15 ***
pm2.5                             -1.815e-02  9.990e-03  -1.817  0.07131 .
tuberculosis                      -2.056e-02  2.284e-03  -9.001 1.06e-15 ***
urban_population                   3.781e-02  1.665e-02   2.271  0.02457 *
forest_area                       -1.002e-02  1.312e-02  -0.764  0.44617
health_expenditure_rate            1.076e-01  9.651e-02   1.115  0.26686
water_resources_rate               2.521e-02  4.170e-02   0.605  0.54635
Improved_sanitation_facilities_rate  3.619e-02  1.966e-02   1.841  0.06766 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.312 on 147 degrees of freedom
Multiple R-squared:  0.879,     Adjusted R-squared:  0.8716
F-statistic: 118.7 on 9 and 147 DF,  p-value: < 2.2e-16
```

Figure 1 Summary of initial regression

2 Adjusted Regression and exclude the outliers

Through our observation of each coefficient's p-value, it's obvious that at 10% significance, factors as forest_area(0.44617>0.1), health_expenditure (0.26686>0.1) and water_resources(0.54635>0.1) are not individually significant.

2.1 BP Test

However, at the beginning we choose linear regression as our model simply relying on our own experience. In case of the problem as heteroscedasticity existing in our regression, we also tried model as semi-log and log-log to see which one is better. Accordingly we run the BP test in R for all of three models to see if they have heteroscedasticity problem.

First, we run the BP test for the linear regression we just conduct. Since the p-value is 0.2362>0.05, we cannot reject the null of no heteroscedasticity in our linear regression at significance of 5%

```
> library(lmtest)
> bptest(lifeExpectancy.lm)

        studentized Breusch-Pagan test

data:  lifeExpectancy.lm
BP = 11.61, df = 9, p-value = 0.2362
```

Figure 2 BP test of linear regression

Then, we change our regression to the semi-log model which is

log(life_expectancy)~gdp_per_capita+fertility_per_woman+pm2.5+tuberculosis+urban_population+forest_area+health_expenditure_rate+water_resources_rate+Improved_sanitation_facilities_rate

We run the BP test in R and the p-value is 0.09481>0.05 which means we cannot reject the null of no heteroscedasticity in our linear regression at significance of 5%

```
> bptest(lifeExpectancy.redlm)

        studentized Breusch-Pagan test

data:  lifeExpectancy.redlm
BP = 14.862, df = 9, p-value = 0.09481
```

Figure 3 BP test of semi-log regression

Finally, we conduct the log-log model as following:

log(life_expectancy)~log(gdp_per_capita)+fertility_per_woman+pm2.5+tuberculosis+urban_population+forest_area+health_expenditure_rate+water_resources_rate+Improved_sanitation_facilities_rate

We run the BP test again and the p-value is 0.08303>0.05 which means we cannot reject the null of no heteroscedasticity in our linear regression at significance of 5%

```
> bptest(life_expectancy.log)

        studentized Breusch-Pagan test

data:  life_expectancy.log
BP = 15.3, df = 9, p-value = 0.08303
```

Figure 4 BP test of log-log regression

Among all three models, the linear model obtains the highest p-value which implies that we have serious heteroscedasticity problem existing in our linear model.

2.2 Individually Significance

The issue of heteroscedasticity might be resulted from three variables which seems not significantly correlated with life_expectancy. To verify our assumption, we plot the chart between life_expectancy and variables as forest_area, health_expenditure rate, water_resource_rate (shown as three charts below). It's clear that they are not individually significant with life_expectancy.

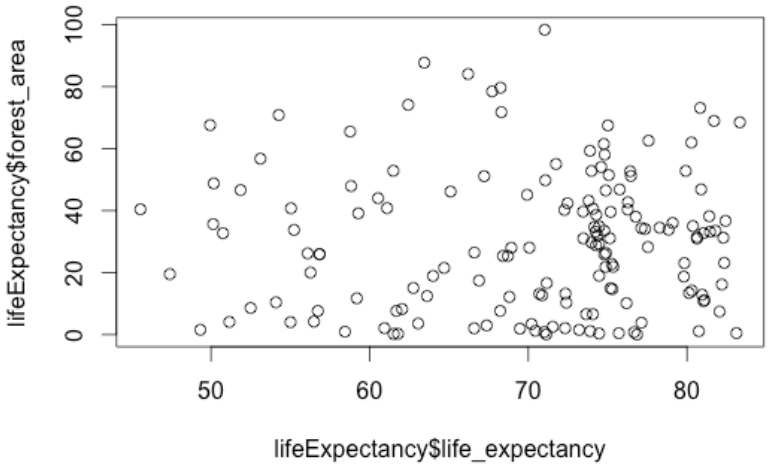

Figure 5 Scatter plot chart between lifeExpectancy and forest_area

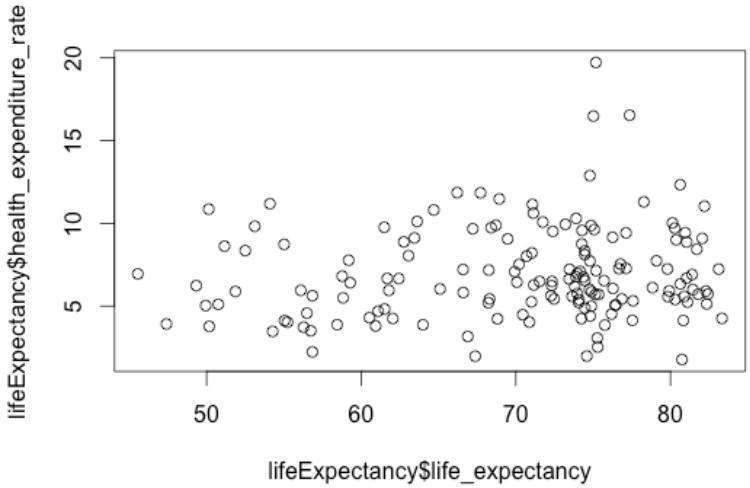

Figure 6 Scatter plot chart between lifeExpectancy and health_expenditure_rate

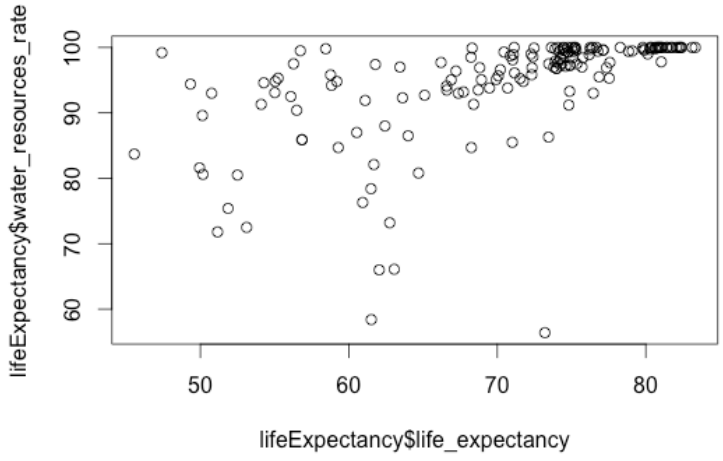

Figure 7 Scatter plot chart between lifeExpectancy and water_resource_rate

2.3 Jointly Significance

Although we come to the conclusion that variables as forest-area, health expenditure rate, water resource rate are not individually significant to life_expectancy. We still want to examine that whether they are jointly significant or not before omitting them. We use anova function in R to test if these three variables are jointly significant (results shown below)

```
> anova(lifeExpectancy.lm,lifeExpectancy.redlm)
Analysis of Variance Table

Model 1: life_expectancy ~ gdp_per_capita + fertility_per_woman + pm2.5 +
    tuberculosis + urban_population + forest_area + health_expenditure_rate +
    water_resources_rate + Improved_sanitation_facilities_rate
Model 2: life_expectancy ~ gdp_per_capita + fertility_per_woman + pm2.5 +
    tuberculosis + urban_population + Improved_sanitation_facilities_rate
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    147 1612.9
2    150 1634.8 -3   -21.914 0.6657 0.5744
```

Figure 8 Anova test results

Since the p-value is 0.5744>0.05, we could conclude that at any reasonable significance level, forest_area and water_resources and health_expenditure are not jointly significant.

2.4 Adjusted Linear Regression

After omitting the 3 unrelated variables as forest-area, health expenditure rate, water resource, we come to the adjusted regression as below:

life_expectancy~gdp_per_capita+fertility_per_woman+pm2.5+tuberculosis+urban_population+Improved_sanitation_facilities_rate

And we run the regression in R and get the summary table as below

```
Call:
lm(formula = life_expectancy ~ gdp_per_capita + fertility_per_woman +
    pm2.5 + tuberculosis + urban_population + Improved_sanitation_facilities_rate,
    data = lifeExpectancy)

Residuals:
     Min       1Q   Median       3Q      Max
-17.1248  -1.6272   0.2437   1.7832  11.3006

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      7.698e+01  2.369e+00  32.494  < 2e-16 ***
gdp_per_capita                   5.329e-05  1.862e-05   2.861  0.00483 **
fertility_per_woman             -3.125e+00  3.318e-01  -9.419  < 2e-16 ***
pm2.5                           -1.722e-02  9.316e-03  -1.849  0.06644 .
tuberculosis                    -2.071e-02  2.259e-03  -9.169 3.44e-16 ***
urban_population                 3.809e-02  1.658e-02   2.297  0.02298 *
Improved_sanitation_facilities_rate  4.044e-02  1.901e-02   2.127  0.03506 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.301 on 150 degrees of freedom
Multiple R-squared:  0.8774,    Adjusted R-squared:  0.8725
F-statistic: 178.9 on 6 and 150 DF,  p-value: < 2.2e-16
```

Figure 9 Summary table of adjusted linear regression

In the summary table, the adjusted R-squared is 0.8725 which is close to 1. High R-stats are often assumed to imply that the estimated regression is good at predicting. Intuitively, we could say that around 87.25% of the data points could be explained by this regression. But still, we want to examine the residuals of our regressions to see if it is mis-specified.

Based on that, we plot the residuals of the regression in R as below

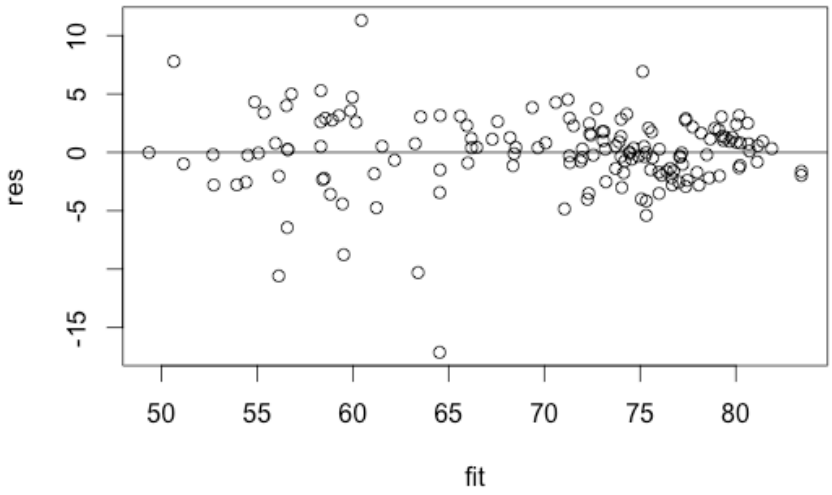

Figure 10 Scatter plot chart of residuals of adjusted regression

From the plot we could see that the residuals has a mean of zero and are plotted closely to the zero, mainly in the area of [-5,5]. Therefore, we could say that our adjusted regression is not mis-specified and not heavily influenced by outliers.

So, we have adjusted our regression model of the Life Expectancy, which is:

LifeExpectancy =76.98 + 0.001 GDP_per_Capita -3.125Fertility_per_woman -0.01722PM2.5 -0.021 tuberculosis+0.038urban_population + 0.04 Improved_sanitation_facilities_rate

For the regression model, we could see that gdp per capita, urban population rate and improved sanitation facility rate are the most statistically significant three, they are all significant at 5% level of confidence. But before we jump into conclusion that increase the amount of these three would be helpful the most in improving the life expectancy in country level, we need to examine the economic significance of these variables.

For the economic significance, the formula is S.D LHS/Coeffecient*S.D.RHS = 0.1, if the result could bigger or equal to 0.1, then it is economic significant.

The standard error of the life expectancy is 3.040415, therefore the coeffecicent multiply the standard error of the coefficient should be equal or greater then 0.304 to be economic significance. However, only the variable fertility per women is

economic significance.

So statistically speaking, the gdp per capita, urban population rate and improved sanitation facility rate are the most significant one, while fertility per women is the only variable in this regression that is economic significant. So increasing the fertility per women would be an effective way to improve the life expectancy.

3 Analysis and Prediction

Before we going into deeper analysis, we calculate the mean, standard error and 95% confidence interval of the life expectancy, the result is as follows:
The mean of the Life Expectancy among these countries is: 70.18761
The standard deviation of it is: 9.244121
The 95 confidence interval of the Life Expectancy among the countries is: [51.9221, 88.45312]

Among the previous discussion about adjusting the linear regression, we left out three countries, which is Costa Rica, Croatia and Cyprus. Since we already have over 100 countries' datasets, so taking three out of them won't make a big error in initializing and adjusting the regression. These three data sets would be used to test how good our linear regression is at predicting.

The stats of the three countries is as follows:

Country Name	Life Expectancy	GDP_Per Capita	Fertility_Per Woman	Pm2.5_Air Pollution	Tuberculosis	Urban_Population_Rate	Improved_Sanitation_Facility_Rate
Costa Rica	79.92100	10461.5768	1.795000	36.5618470	11.0	74.95600	94.50000
Croatia	77.12682927	13597.92145	1.51	97.92124214	13	58.359	97.1
Cyprus	79.80431707	27910.61983	1.461	99.78369236	5.8	67.133	100

Table 1 Dataset of Costa Rica, Croatia, Cyprus

3.1 Prediction of Costa Rica

For Costa Rica, we run the prediction function in R which give us following results:

```
$fit
      1
77.7452

$se.fit
[1] 0.6271964

$df
[1] 150

$residual.scale
[1] 3.301309
```

Figure 11 Prediction results of Costa Rica

The fitted value is 77.7452.

The prediction standard deviation is $\sqrt{0.6271964^2 + 3.301309^2}$= 3.36036

Hence, 95% confidence interval is: [71.15902, 84.33138]. The actual life_expectancy of Costa Rica is 79.92100 which falls into the 95% confidence interval.

3.2 Prediction of Croatia

For Croatia, we run the prediction function in R which give us following results:

```
$fit
      1
77.1776

$se.fit
[1] 0.4446837

$df
[1] 150

$residual.scale
[1] 3.301309
```

Figure 12 Prediction results of Croatia

The fitted value is 77.1776

Prediction standard deviation is $\sqrt{0.4446837^2 + 3.301309^2}$=3.331124

Hence, 95% confidence interval is: [70.64872, 85.19536]. The actual life_expectancy of Croatia is 77.12682927 which falls into the 95% confidence interval

3.3 Prediction of Cyprus

For Cyprus, we run the prediction function in R which give us following results:

```
$fit
       1
78.66197

$se.fit
[1] 0.4615883

$df
[1] 150

$residual.scale
[1] 3.301309
```

Figure 13 Prediction results of Cyprus

Fitted value is 78.66197

Prediction standard deviation is $\sqrt{0.4615883^2 + 3.301309^2}$= 3.333422

Hence, 95% confidence interval is: [72.12858, 83.70648]. The actual life_expectancy of Cyprus is 79.80431707 which falls into the 95% confidence interval

From all three tests, these three countries' actual life_expectancy all fall into their 95% prediction interval, suggesting that our adjusted the regression is doing a great job in predicting life expectancy given the value of those impact factors.

4 Does the Life Expectancy of Countries differ in different continent different?

We are wondering whether the geography plays an important factor in life expectancy, say, if we born on different continents, can we expect to have the same life expectancy?

In order to test this hypothesis, we do the data cleaning and reorganizing first, we categorize countries into different continents as Asia, Africa, Europe, North America and South America.

Then we set up the null hypothesis and calculate the mean and standard deviation of the life expectancy in each continent and get the table below:

H0: The mean Life expectancy have no significant difference between two continents.

H1: The mean Life expectancy have significant difference between two continents.

Continent	Asia	Africa	Europe	South America	North America
Mean	71.87557	59.45172	77.73508	73.33554	74.73569
S.D	6.220824	7.627472	3.889678	3.998621	4.581752
N	36	46	39	11	14

Table 2 Summary of life_expectancy in each continent

The data from different continents should be independent from each other. Moreover, the sample size of each continent is no larger than 50. Based on these two prerequisites, we decide to use t-distribution instead of normal distribution to calculate the p-value.

$$T\text{-statistics:(Mean1-Mean2)/}\sqrt{\left(\frac{S.D1^2}{N1}\right)+\left(\frac{S.D2^2}{N2}\right)}$$

For degrees of freedom, because the population variance of different continent is different, and the sample size of each continent is less than 50, we decide to use the following formula to calculate the degrees of freedom.

$$df = \frac{\left(S_{Y_1}^2/N_1 + S_{Y_2}^2/N_2\right)^2}{\left(S_{Y_1}^2/N_1\right)^2/(N_1-1) + \left(S_{Y_2}^2/N_2\right)^2/(N_2-1)}$$

S_{Y1} stands for the standard deviation of the sample Y_1, N_1 stands for the sample size of sample Y_1.

With these two fomula, we calculate p-value for each comparison between two continents based on t-value and degrees of freedom

For Example:

$$T\text{-Stats(Asia|Europe)} = (77.06517-71.58626)/\sqrt{\left(\frac{5.326994^2}{43}\right)+\left(\frac{5.835904^2}{43}\right)}=4.5469$$

The degree of freedom s(Asia|Europe): 83.31026

*Hence, the p-value : (1-Pt(4.5469,83))*2= 1.836722e-05 <0.01*

Since the p-value is lower than 0.01, the null hypothesis that “the mean Life expectancy have no significant difference between different continents” can be rejected at any significance level. So the mean life expectancy of Asia and Europe are not the same at any reasonable significance level.

Then we run this process to other combination between two continents, and come to the table below:

	t-value	Degrees of freedom	p-value
Asia|Africa	8.12223	79.84378	4.658718e-12 < 0.05
Asia|Europe	4.844555	57.87754	9.83414e-06 < 0,05
Asia| S America	0.9181465	26.17065	0.3669815
Asia|N America	1.78256	32.17807	0.08414858
Africa|Europe	14.08492	69.13676	0 < 0.05
Africa|N America	9.192865	36.64693	4.320122e-11 < 0.05
Africa| S America	8.420951	29.93658	2.130332e-09 < 0.05
Europe|N America	2.183238	20.13575	0.04110688 < 0.05
Europe|S America	3.242079	15.75482	0.005105395 < 0.05
S America|N America	0.8147842	22.69527	0.4235527

Table 3 Summary of p-value between all combinations of two continents

At 5% confidence significance level, we can reject the null hypothesis that two continent share the same expect life expectancy between two continents. Then we have the conclusion as:

1) Life Expectancy in Africa is significantly lower than life expectancy in Europe, North America, South America, Asia

2) Life Expectancy in Europe is significantly higher than life expectancy in Asia, South America, North America

Intuitively, if we are born in Asia, we may have a different life expectancy if we move to Africa, Europe. But we should expect the same life expectancy as those born in North America at 5% confidence significance.

5 Conclusion

To determine factors which will influence life expectancy of countries, we choose variables like urban population, GDP per capita, Health Expenditure, Sanitation Facilities, PM2.5, clean water resources, Tuberculosis, Fertility per woman. Among these 8 indicators, three indicators as forest area, health expenditure and clean water resources are neither individually significant nor jointly significant to life expectancy.

To increase life expectancy, counties could apply following ways:

1) Increase GDP per capita: increase GDP per capita by 1, life expectancy will

increase by 0.001

2) Discourage families to have babies: decrease fertility per women by 1, life expectancy will increase by 3.125
3) Decrease air pollution(pm2.5): decrease pm2.5 rate by 1, life expectancy will increase by 0.01722
4) Decrease tuberculosis: decrease tuberculosis rate by 1, life expectancy will increase by 0.021
5) Encourage urban population: increase urban population rate by 1, life expectancy will increase by 0.038
6) Improve sanitation facilities rate: increase improved sanitation facilities rate by 1, life expectancy will increase by 0.04

However, among all indicators above, factors as GDP per capita, urban population rate and improved sanitation facilities rate are statistically significant but fertility per women is the only variable that is economic significant. Therefore, the most effective way of expending life expectancy is to increase the fertility per women. And secondary effective way is to increase GDP per capita, urban population rate and improved sanitation facilities rate.

Moreover, the mean of life expectancy between certain continents has significant difference. Conclusively, life Expectancy in Africa is significantly lower than life expectancy in Europe, North America, South America, Asia. Life Expectancy in Europe is significantly higher than life expectancy in Asia, South America, North America. Africa as the continent obtains the lowest life expectancy compared with other continents, it should consider to increase its life expectancy by the 6 solutions we offered above.