Statistical Analysis on Factors Influencing Life Expectancy

Kratika Sharma, Xiaokang Wu, Manasvini Hothur, Steven Wang

12/06/20

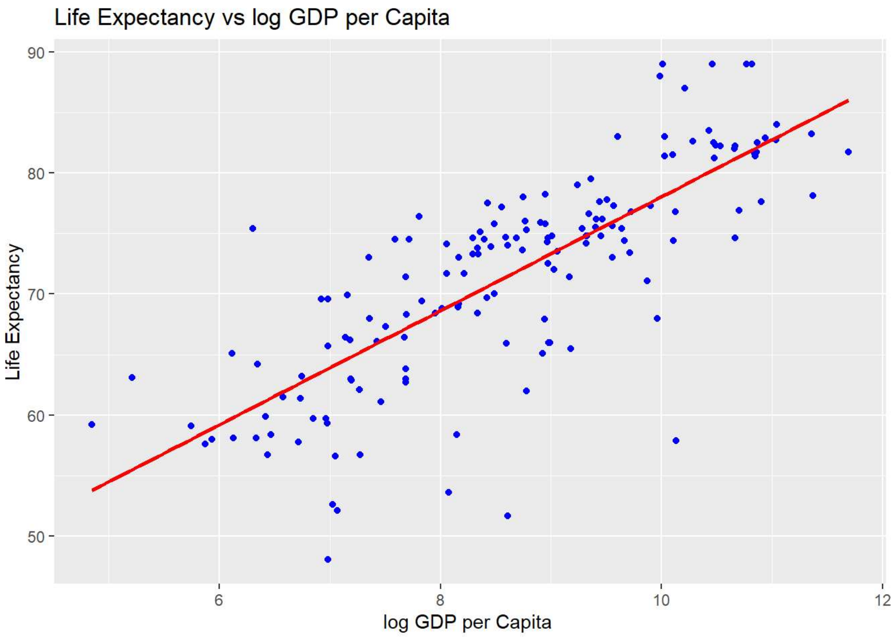
Summary

In this analysis we wanted to explore variables that affect life expectancy. Our dataset (retrieved from WHO.int), contains various information regarding a country's GDP per capita, Health Expenditures, Mortality, Schooling etc. It is widely accepted that countries with a higher GDP per capita tend to have higher life expectancies. Through our analysis we will investigate factors that impact life expectancies to different magnitudes. In further tests, we will investigate how these variables affect less developed countries' life expectancies. What factors are the best determinants of life expectancy?

```
Country Summary
==============================================================
Statistic                     N    Mean     Median  St. Dev.
--------------------------------------------------------------
Life Expectancy              154   71.61    73.70     8.77
GDP per Capita               154 13,836.98 6,105.27 19,266.01
Health Expenditure per Capita 154 1,101.01   388.05  1,839.29
Adult Mortality              154  149.32    132.5    108.76
Education (years)            154   13.00     13.1     2.94
--------------------------------------------------------------
```

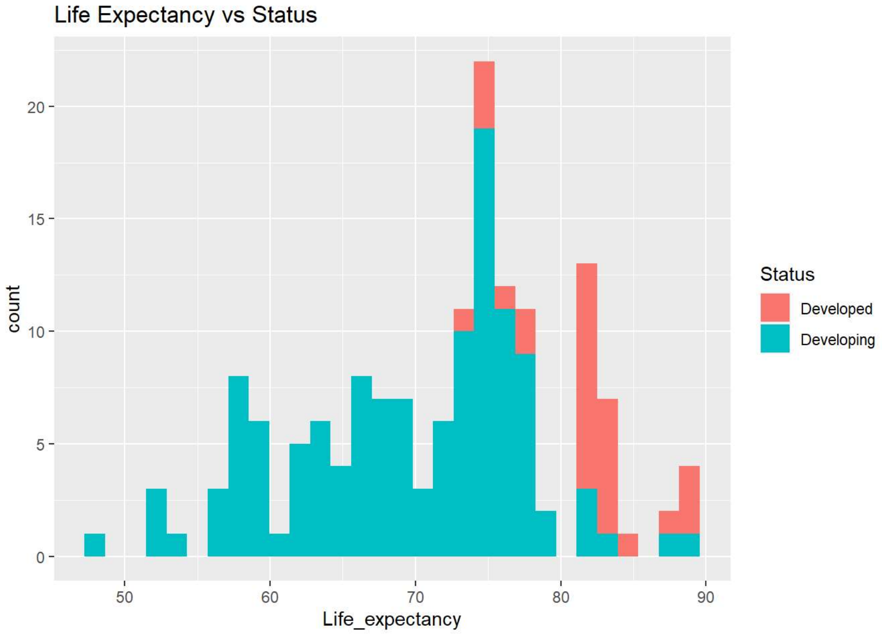
Descriptive Table

In our sample, there are 154 observations. Based on the table of descriptive statistics, the average life expectancy in this sample is 71.61 years; the average health expenditure is 1,101 per capita (in USD $);the average GDP is 13,837 per capita (in USD $); and the average adult mortality rate is 149 per 10,000. Note that for GDP per capita, Health Expenditure and Adult Mortality all have a median lower than the mean, indicating that the distributions of these variables are positively skewed (right-skewed). Life Expectancy and Education have a normal distribution (although it is not perfectly symmetrical). The standard deviation for all the variables are relatively large, indicating a large spread among the data for each variable.



We believe that a nonlinear function would be a better fit for these two variables. We decided to plot Life Expectancy against the logarithm of GDP per capita in the graph above. This nonlinear model is better suited to explain the relationship between these two variables.

Looking at the graph, we observe a positive correlation between life expectancy and GDP per capita - as log GDP per capita increases, so will Life Expectancy. We will explore the exact amount through our regression tests. The lack of universal access to health care in the United States undoubtedly increases mortality and reduces life expectancy: more GDP and health expenditure tends to increase life expectancy.



This graph illustrates the variation of life expectancies between developed and developing countries. Here, we notice that most developed countries have life expectancies all above 70+ years. However, for developing countries, there is much more variation in life expectancies. The plot of the life expectancy distribution group demonstrates that there is a strong relationship between life expectancy and a country's development status.

```
Regression of Life Expectancy against GDP and other variables
===========================================================================
                                       Dependent variable:
                      -----------------------------------------------------
                                         Life_expectancy
                         (1)      (2)     (3)     (4)      (5)       (6)
---------------------------------------------------------------------------
GDP                    0.0003
                      (0.00005)

logGDP                           1.498   1.390   0.964    1.038     1.048
                                (0.659) (0.592) (0.408)  (0.399)   (0.394)

logPopulation                    0.246   0.191   0.343    0.480     0.473
                                (0.196) (0.179) (0.145)  (0.159)   (0.158)

logHealth_expenditure            3.404   1.737   1.502    1.423     1.322
                                (0.601) (0.671) (0.436)  (0.433)   (0.453)

Schooling                                1.107   0.567    0.501     0.485
                                        (0.212) (0.151)  (0.154)   (0.156)

infant_deaths                                   -0.003   -0.023    -0.023
                                                (0.004)  (0.012)   (0.012)

Adult_Mortality                                 -0.035   -0.034    -0.035
                                                (0.005)  (0.004)   (0.005)

I(infant_deaths2)                                        0.00002   0.00002
                                                        (0.00001) (0.00001)

firstworld                                                          0.687
                                                                   (0.891)

Constant               67.862   34.937  32.077  47.066   45.829    46.578
                       (0.789)  (4.630) (4.418) (4.118)  (4.076)   (4.282)

---------------------------------------------------------------------------
Observations             153      153     153     153      153       153
R2                      0.381    0.711   0.750   0.862    0.866     0.867
Adjusted R2             0.377    0.705   0.743   0.857    0.860     0.859
Residual Std. Error     6.888    4.737   4.421   3.304    3.269     3.274
F Statistic            92.917   122.264 111.067 152.352  133.963   116.963
===========================================================================
Note:                                                                    NA
```

Regression 1 :

- Our first regression is a simple regression. We find that an increase in GDP by $1 per capita, increases life expectancy by 0.0003 years. Also, the estimate is statistically insignificant.
- According to economic theory, increase in life expectancy is a key indicator to gauge the economic development of a country. The increase in life expectancy is accompanied with the increase in Gross Domestic Product (GDP) per capita income. When a country's economic output — its GDP per capita— is higher than expected, mortality rates are also higher than expected.
- Looking at the adjusted R2, GDP per capita explains 37.7% of variation in life expectancy. Since the adjusted R2 value is low, it would be a good idea to explore other variables that might influence a country's life expectancy.

  Statistically insignificant: GDP

Regression 2 :

- Here we take the regression of Life Expectancy on log GDP, log population and log Health Expenditure. An increase in GDP per capita by one percent increases life expectancy by 0.015 years holding everything else constant.
- Similarly, increasing log Healthcare Expenditure per capita by 1% increases life expectancy by 0.017. There is an upward omitted variable bias for log GDP of about 0.108.
- Healthcare spending is usually accompanied by a country's GDP per Capita - The higher the GDP of a country, the more that country will spend on Healthcare.
- Uncontrolled population growth for developing countries may have a negative effect on life expectancy. When populations grow uncontrollably, there are more demands for goods and healthcare services while facing shortages of resources. This in turn must be accompanied by an economy that has low unemployment rate, low inflation and high income.
- Log GDP and log Healthcare_expenditures are statistically significant. However, log Population is statistically insignificant.
- The adjusted R2 improved and now our regressors explain 70.5% of the variation in life expectancy.

Statistically insignificant: log population

Statistically significant: log GDP, log health_expenditure, Schooling

Regression 3 :

- We are adding Schooling as a control variable. The coefficient for log GDP changes from 1.498 to 1.390. By adding Schooling, we can see that there was an upward omitted variable bias for log GDP of 0.108. Holding everything constant, an increase in log GDP per capita by 1% increases life expectancy by 0.0139 years.
- An increase of average years of education by 1 year increases life expectancy by 0.567 years, holding everything constant.
- Education is a strong indicator of a country's development status. Having a good infrastructure can have positive effects on life expectancy.
- The adjusted R2 improved here and now our regressors explain 74.3% of the variation in life expectancy.

Statistically insignificant: log population

Statistically significant: log GDP, log health_expenditure, Schooling

Regression 4 :

- We are adding infant_deaths and adult_mortality as control variables here. The coefficient on log GDP changes from 1.390 to 0.964, implying that we had an upward omitted variable bias of 0.426.
- For the added variables, an increase of infant and adult mortality by 1 (per 10,000) will decrease life expectancy by 0.003 and 0.035 years respectively, holding everything constant.
- Infant death rates is a leading indicator of the level of child health and overall development in countries. Similarly adult mortality rate is an important marker about the overall health of a society. A large number of people dying early can cancel out people living to old age, thus bringing down the average lifespan.
- The adjusted R2 improved here and now our regressors explain 85.7% of the variation in life expectancy.

Statistically insignificant: infant_deaths

Statistically significant: log GDP, log population, log health_expenditure, schooling, adult_mortality

Regression 5 :

- We are adding infant_deaths^2 as control variable. The coefficient of log GDP changes from 0.964 to 1.038 which means we had a downward bias.
- When adding a quadratic variable, when increasing infant deaths from 1 unit to 2 units (2 deaths per 1,000), there is a decreases in life expectancy by 0.023 years, holding everything constant (-0.023(2-1) - .00002((2^2. - 1^2) = -0.02306).
- The adjusted R2 improved here and now our regressors explain 86.0% of the variation in life expectancy.

Statistically insignificant: infant_deaths

Statistically significant: log GDP, log population, log health_expenditure, schooling, adult_mortality, infant_deaths^2

Regression 6 :

- Lastly, we added the dummy variable "firstworld", which takes a value 1 if a country is developed and 0 otherwise. The coefficient of log GDP changes from 1.038 to 1.048 which means we had downward bias.
- Increase in life expectancy can be attributed to a number of factors, including rising living standards, improved lifestyle and better education, as well as greater access to quality health services. All these factors are used to classify whether a country is developed or not.
- The adjusted R2 here is 0.859 meaning that our regressors now explain 85.9% of the variation in life expectancy.
- Since the adjusted R2 decreased, a country's development status might not be the best variable to investigate. Other variables such as Schooling, Mortality, Health Expenditure and GDP are more important when looking at life expectancy.

Statistically insignificant: infant death, firstworld

Statistically significant: log GDP, log population, log health expenditure, infant deaths^2, schooling, and adult mortality

```
Linear hypothesis test

Hypothesis:
Schooling = 0
infant_deaths = 0
Adult_Mortality = 0

Model 1: restricted model
Model 2: Life_expectancy ~ logGDP + logPopulation + logHealth_expenditure +
    Schooling + infant_deaths + Adult_Mortality

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)    
1    149                        
2    146  3 29.356 6.525e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-Test 1: Regression 2 and Regression 4

H0: coefficient on Schooling = 0 Infant Death = 0 and coefficient on Adult Mortality = 0 (Model 1 better than Model 2) Ha: coefficient on Schooling And/Or Infant Death And/Or coefficient on Adult Mortality different than 0 (Model 2 better than Model 1)

F value is 29.356, so we can reject the null at 1% critical level or less. Thus, Model 2 is the better model 1. Regression 4 is better than Regression 2. Schooling, Infant Deaths and Adult Mortality should be included in the regression.

```
Linear hypothesis test

Hypothesis:
I(infant_deaths^2) = 0

Model 1: restricted model
Model 2: Life_expectancy ~ logGDP + logPopulation + logHealth_expenditure +
    Schooling + infant_deaths + Adult_Mortality + I(infant_deaths^2)

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F  Pr(>F)  
1    146                    
2    145  1 3.4838 0.06399 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-Test 2: Regression 4 and Regression 5

H0: coefficient on Infant death square = 0 (Model 1 better than Model 2)

Ha: coefficient on Infant death square different than 0 (Model 2 better than Model 1)

Here we find out that F>3 and P value is less than 0.01 so we can reject the null hypothesis at 1% or less, which means the Model 2 is better than Model 1. The quadratic model (Regression 5) is better than the linear model (Regression 4).

USA Life Expectancy Prediction

```
    1 
79.89 
```

Above is the predicted life expectancy for USA

US life expectancy prediction is 79.89 years, which is reasonable given that US has high GDP and high schooling years. The prediction result is also consistent with the number provided by a reliable organization - World Bank.

```
Regressions for Life Expectancy on GDP and other variables for countries with GDP <4,000 USD per Capita
====================================================
                             Dependent variable:    
                        ----------------------------
                               Life_expectancy      
                           (1)       (2)       (3)  
----------------------------------------------------
log(GDP)                  1.947     1.607     2.152  
                         (0.691)   (0.747)   (0.771) 
                                                    
log(Population)           0.628     0.500     0.748  
                         (0.345)   (0.331)   (0.316) 
                                                    
log(Health_expenditure)   0.505     0.933     0.730  
                         (0.836)   (1.235)   (0.910) 
                                                    
Schooling                 0.561     0.586            
                         (0.232)   (0.311)           
                                                    
infant_deaths            -0.025    -0.021    -0.031  
                         (0.016)   (0.014)   (0.014) 
                                                    
Adult_Mortality          -0.032    -0.024    -0.032  
                         (0.006)   (0.010)   (0.006) 
                                                    
I(infant_deaths2)        0.00002   0.00002   0.00003 
                        (0.00002) (0.00001) (0.00001)
                                                    
HIV.AIDS                           -0.643            
                                   (0.464)           
                                                    
thinness_youth                      0.036            
                                   (0.122)           
                                                    
GoodEducation                                 2.563  
                                             (1.859) 
                                                    
Constant                 39.863    41.186    41.334  
                         (8.473)   (8.521)   (7.878) 
                                                    
----------------------------------------------------
Observations               61        59        61    
R2                        0.746     0.760     0.737  
Adjusted R2               0.713     0.715     0.703  
Residual Std. Error       3.539     3.540     3.602  
F Statistic              22.284    17.206    21.250  
====================================================
Note:                                             NA
```

Alternate Regression 1 :

- For our alternate regressions, we are looking at a subset of countries with GDP per capita of < $4,000 per Capita. For these countries, an increase of GDP per Capita by 1% will increase life expectancy by 0.019 years, holding other variables constant.
- Comparing these coefficients to our regression 5 in our first regression table, we can see that for countries with GDP per capita <$4,000, increasing GDP per capita will have a much higher impact on life expectancy (coefficient log GDP values 1.947 vs 1.038).
- There is also an increase for the coefficient on schooling. An increase of schooling by 1 year will increase life expectancy by 0.561 years, holding everything constant.
- However, unlike our baseline regression 5, log Population and log Health Expenditure are now statistically insignificant. The coefficients for infant deaths and adult mortality remains mostly unchanged. For infant_deaths^2, Increasing infant deaths from 1 to 2 units (2 deaths per 1,000) decreases life expectancy by 0.025 years, holding everything constant (-0.025(2-1) - .00002((2^2. - 1^2) = -0.025).
- In this alternative regression, all variables explain about 71.3% of the variation in life expectancy.

Statistically insignificant: log(Population), log(Health_Expenditure), infant_deaths, infant_deaths^2

Statistically significant: log(GDP), Schooling, Adult_Mortality

Alternate Regression 2 :

- Here we explore health and disease control variables HIV.AIDS and thinness_youth. When we add the variables HIV.AIDS and thinness_youth, we see that an increase of GDP per capita by 1% increases life expectancy by 0.016 years. There is an upward variable bias of 0.34 for Log GDP. Also Log Health_Expenditure experienced a downward bias of 0.43.
- Having a high rate of AIDS (or other diseases) usually has a negative effect on life expectancy for poorer countries. An increase of 1 unit for the rate of HIV/AIDS will decrease life expectancy by 0.643 years.
- Thinness youth does not seem to affect life expectancy by much. In this regression, we notice that health-related variables (HIV.AIDS, thinness_youth and log Health_Expenditure) are all statistically insignificant. GDP is still statistically significant though.
- For poorer countries, if their goal is to maximize life expectancy, it might be better to focus on increasing GDP per Capita. The adjusted R2 is about the same as for alternative regression 1.

Statistically insignificant: log health expenditure, Schooling, infant_death, thinness_youth, HIV.AIDS, infant_deaths^2

Statistically significant: log GDP, Adult_Mortality

Alternate Regression 3 :

- We have added a dummy variable here called “GoodEducation”, which assumes a value of 1 when a country has greater than 13 years average education and 0 otherwise. In this regression, an increase of GDP per capita of 1% will increase Life Expectancy by 0.022 years.
- When comparing Regression 1 and Regression 3, log GDP suffers from downward omitted variable bias of 0.2. For countries with good education (>13 years), those life expectancies are 2.563 more years than countries with average education years less than 13 years.
- Interestingly, the dummy variable GoodEducation is not statistically significant. Similar to the previous regression, log GDP is still statistically significant.
- A higher level of education is usually associated with higher income and greater wealth, which can lead to better health. However, from our regression, for poor countries, having good education is not a guarantee to good health and long life expectancy.
- Again, for less developed countries, focusing on improving GDP per Capita might still be the best way to improve the status of a country and its life expectancy. All regressors explain about 70.3% of the variation in life expectancy.

Statistically insignificant: log Health Expenditure, Good Education

Statistically significant: log GDP, log Population, infant_deaths, Adult_Mortality, infant_deaths^2

Internal Validity:

- Log GDP per capita, our main variable of interest, is statistically significant for all our regressions and the other non-control factors that we took into consideration such as schooling, log population, and log health expenditure are also significant.

- We strove to reduce omitted variable bias in all our regressions by adding new variables. We used different functional forms in our regressions (quadratic, log) to try and eliminate the chance of using the wrong functional form. Since we have not collected the data ourselves, we cannot gauge errors-in-variable bias and sample selection bias.

External Validity:

- Our predicted value of the US life expectancy is 79.89 years, which is very close to the number provided by the World Bank - 78.58 years.
- In general, our graphs showed trends that were consistent with our regression results. For example, in our graphs, we have shown that developed countries with higher GDP per capita have higher life expectancies; this result is displayed in our regressions and predicted by us.

Conclusion :

- Based on our graphs and regressions, we can conclude that there is a positive correlation between life expectancy and GDP per capita.
- When we add other variables into our regression, we notice that the models become increasingly more fitting. Nonetheless, the correlation between life expectancy and GDP per capita remains positive.
- Even after adding multiple regressors, GDP per capita is still the strongest determinant for life expectancy. Of our six regressions, we concluded that the best regression model is regression five.
- One reason for this is regression five has the highest adjusted R2 value (0.860).
- To confirm our findings, we performed two F-tests. The first F-test allowed us to confirm that the unrestricted model (Regression 4) is better than our restricted model (Regression 2).
- The second F-test allowed us to confirm that the unrestricted model (Regression 5) is better than our restricted model (Regression 4).
- For poorer countries, they will experience an even greater impact when increasing GDP per capita. We found that Health Expenditures and Schooling were statistically insignificant. Although those variables are important for development, it would be better for less developed countries to focus on increasing GDP per capita if they wanted to increase life expectancy