Analysing the affects of Economic on people's Life Expectancy

Tony Chen (1004265298)

2020/12/17

Abstract

We are interested about the the affects of economic on people's life expectancy, so we look at the data set related to life expectancy, health factors for 193 countries has been collected and its corresponding economic data.We used a given set of attributes to build a multiple linear regression model to analyze people's life expectancy. And use the scatter plots to incorporate some causal inference of people's life expectancy. People can use the result to help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

Keywords

Keywords: Life Expectancy, gross domestic product (GDP), Income Composition, Developing Country, Developed Country.

Introduction

By studying history and humanity, anthropologists and evolutionary biologists have found that before the 1800s, humans rarely exceeded the age of 50 (Finch, 2009). However, over time, as the world developed and transitioned from societies of high mortality and fertility to those of low morality and fertility, extending the time that people lived became a desirable goal.

Today, through medical progress, improvements in technologies, emerging economies, and various other socioeconomic factors such as education availability and better living conditions, life expectancy, which is defined as a statistical measure of the average number of years an individual is expected to live, has seen a global rise. Specifically, in 2016, statistics showed that 72 years was the average life expectancy for an infant at birth, (WHO, n.d.).

The purpose of this study is to create a predictive model based on life expectancy data which consists of several socioeconomic predictors that researchers utilized to predict life expectancy among 193 countries between the years of 2000 and 2015. This analysis focuses on how economic factors influence life expectancy and which one is the most significant regarding how it affects life expectancy.

Data

We obtained the data set "Life Expectancy (WHO)" from the KAGGLE website. "The data-set related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website" (Rajarshi,2017).The target population for this data is all countries in the world, excluding less known countries like Vanuatu, Tonga, Togo, Cabo Verde etc. Since "finding all data for these countries was difficult and hence, it was decided that we exclude these countries from the final model data-set."(Rajarshi,2017).

This data set contains 20 predictor variables relating to socioeconomic and health factors, of which we selected four that focus on economic influences. These five selected variables were: gross domestic product (GDP), percentage expenditure, total expenditure, income composition of resources, and status.

Therefore we use life expectancy and those 5 selected variables from the data set to do the analysis(Table 1).Those variables were selected because they contain more valid responses and likely to have relationships with life expectancy.We removed 'NA' in those variable since some countries has missing data in those variable, so we want to use this way to remove those countries avoid error. Also We perform this analysis on the data relating to the year 2014 instead of 2015, our original target of analysis, as the data for 2015 had a significant amount of missing data for some of the economic influences we were studying.

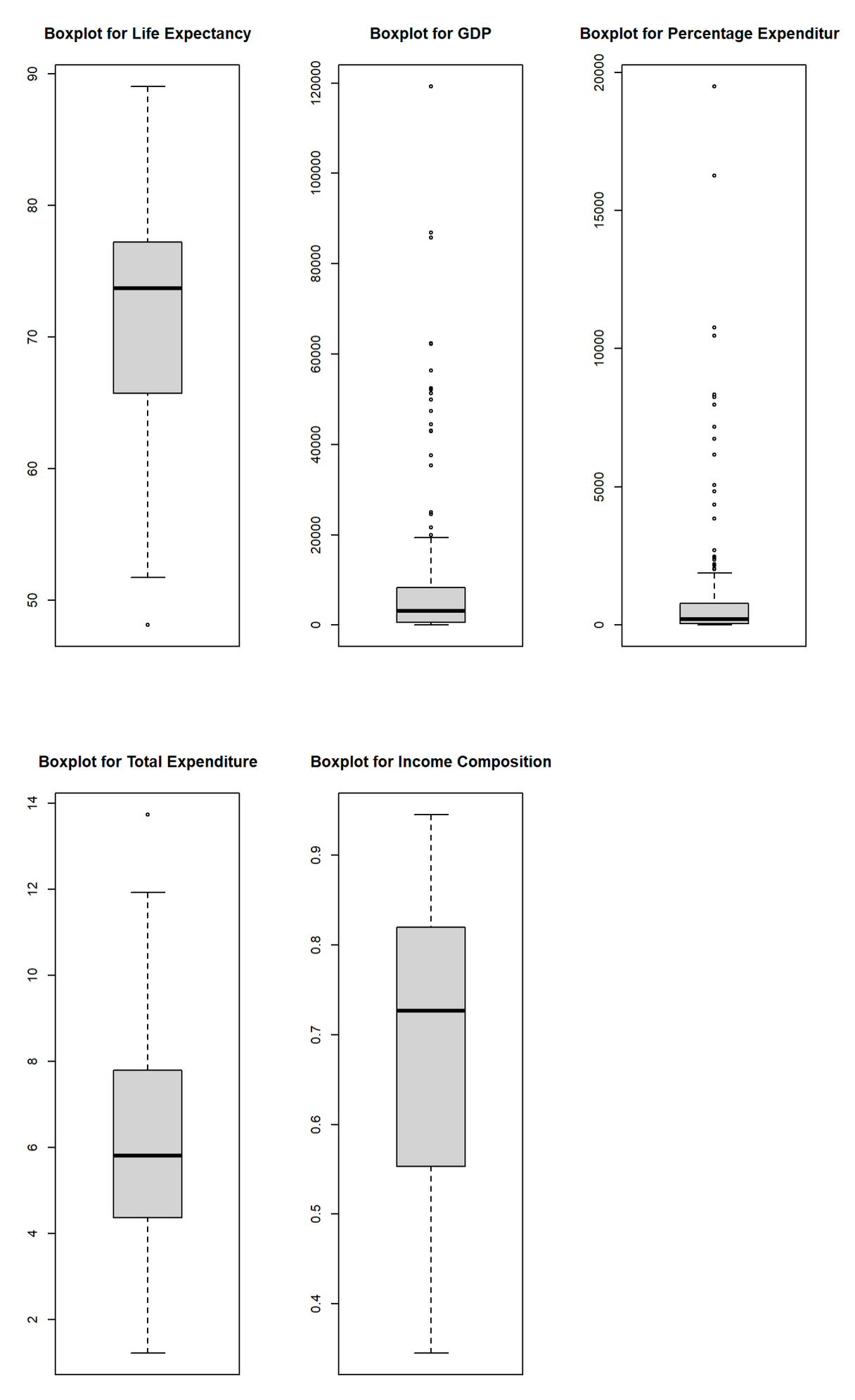
'life_expectancy' is a numerical variable, which measures the average age individuals from a specific country lives for in years. 'percentage_expenditure' is a numerical variable, measures expenditure on health as a percentage of Gross Domestic Product per capital. 'total_expenditure' is a numerical variable, measures general government expenditure on health as a percentage of total government expenditure. 'gdp' is a numerical variable. 'income_composition' is a numerical variable, Human Development Index in terms of income composition of resources (index ranging from 0 to 1). 'status' is a categorical variable, shows the countries is either developed country or developing country.

Here is the table of first six lines of data set:

Table 1

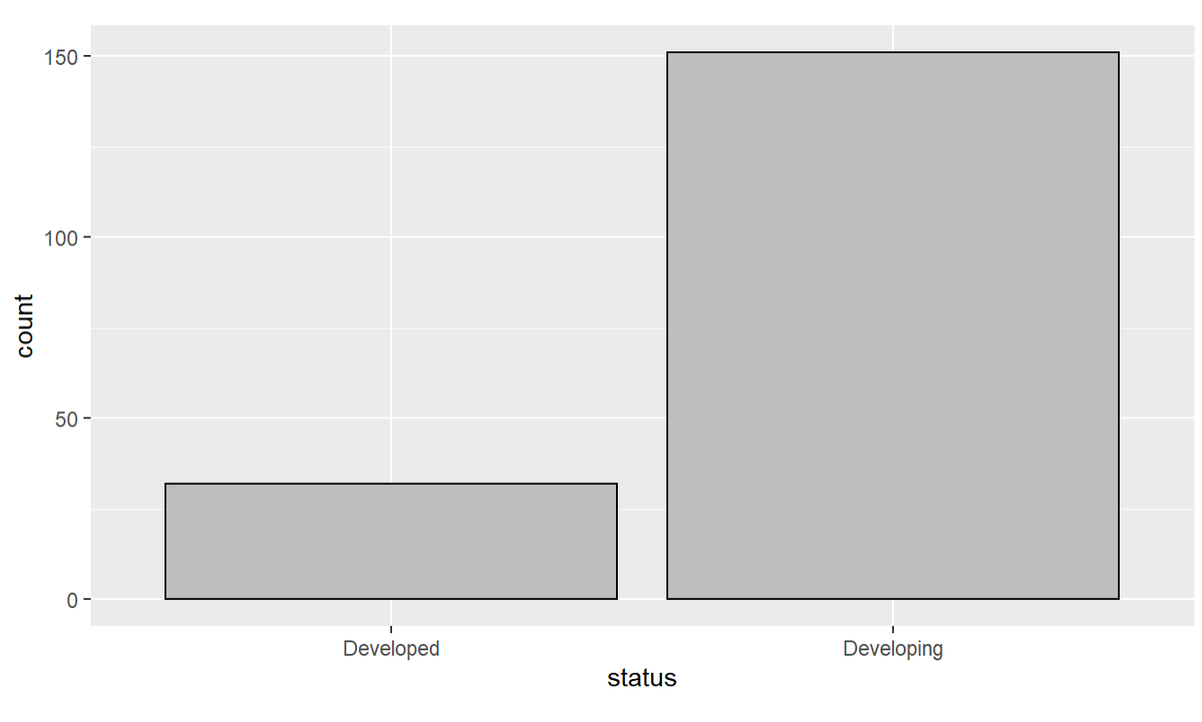
	year	life_expectancy	percentage_expenditure	total_expenditure	gdp	income_composition	status
2	2014	59.9	73.52358	8.18	612.6965	0.476	Developing
18	2014	77.5	428.74907	5.88	4575.7638	0.761	Developing
34	2014	75.4	54.23732	7.21	547.8517	0.741	Developing
50	2014	51.7	23.96561	3.31	479.3122	0.527	Developing
66	2014	76.2	2422.99977	5.54	12888.2967	0.782	Developing
82	2014	76.2	847.37175	4.79	12245.2565	0.825	Developing

The following are the plots of the raw data:



According to the 'Boxplot for the GDP' and percentage_expenditure, the trend for the distribution of both plot are serious right-skewed. Also life expectancy is our dependent variable which measures the average age individuals from a specific country lives for in years. By according to the 'Boxplot for the Life Expectancy', we examined the normality of this response variable and found that it was relatively normally distributed. Thus, we found no need to transform the life expectancy variable for our models.

The following are the bar plot for only one categorical variable. It is the cumulative frequency graph.



Model

We continue our analysis by building a multiple linear regression model to predict life expectancy using gross domestic product (GDP), percentage expenditure, total expenditure, and income composition of resources. We choose this model because the predictor life expectancy is numeric and we believe there exists a linear relationship between explanatory variables and life expectancy. All the explanatory variables are numeric, since we don't want to use status to be explanatory variable in our model. the reason is because the definition of developed country is a country has a mature and sophisticated economy, usually measured by gross domestic product (GDP) and/or average income per resident. So we believe that 'status' has high correlation with other variable, and will affect our model's result.

Using the data selected, we build a linear regression model with the following formula:

$$y_i = \beta_0 + \beta_1 x_{percentage_expenditure,i} + \beta_2 x_{total_expenditure,i} + \beta_3 x_{gdp,i} + \beta_4 x_{income_composition,i} + \epsilon_i$$

β_0 is the coefficient of intercept. Other βs are the coefficients of corresponding variables. $x_{percentage_expenditure,i}$ represents the percentage of GDP on health, $x_{total_expenditure,i}$ is percentage of total government expenditure on health. $x_{income_composition,i}$ is Human Development Index in terms of income composition of resources, and $x_{gdp,i}$ is country's GDP in 2014. Finally ϵ_i is the error of our model.

The results of code ran by R is:

```
##
## Call:
## lm(formula = life_expectancy ~ percentage_expenditure + total_expenditure +
##     gdp + income_composition)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0181  -1.9232   0.3252   1.8264   9.1964
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.563e+01  1.529e+00  23.300   <2e-16 ***
## percentage_expenditure  1.733e-04  2.746e-04   0.631   0.5289
## total_expenditure       2.832e-01  1.219e-01   2.324   0.0215 *
## gdp                    -2.717e-05  4.219e-05  -0.644   0.5206
## income_composition      4.958e+01  2.300e+00  21.555   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.695 on 149 degrees of freedom
## Multiple R-squared:  0.8271, Adjusted R-squared:  0.8224
## F-statistic: 178.1 on 4 and 149 DF,  p-value: < 2.2e-16
```

```
## [1] 0.8270513
```

From the table above, we notice that most p-values of these variables are small, but GDP and percentage_expenditure's p-values are too high. Which means the corresponding coefficients of GDP and percentage_expenditure are not significant, or they has high correlation to each other. And from the last line, R-squared is computed as 0.8270513, representing that 82.71% of variations can be explained by the model.

Overall, the performance is not too good. Even though our r-squared is high, but GDP and percentage_expenditure's p-values are too high, which means there is evidence to prove that the coefficients of GDP and percentage_expenditure are 0.

So we create three more model:

Remove GDP

$$y_i = \beta_0 + \beta_2 x_{percentage_expenditure,i} + \beta_3 x_{total_expenditure,i} + \beta_4 x_{income_composition,i} + \epsilon_i$$

```
##
## Call:
## lm(formula = life_expectancy ~ percentage_expenditure + total_expenditure +
##     income_composition)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2035  -1.9368   0.1684   2.0461   9.2563
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.575e+01  1.515e+00  23.591  < 2e-16 ***
## percentage_expenditure 1.531e-05  1.230e-04   0.124  0.90111
## total_expenditure      3.067e-01  1.160e-01   2.643  0.00909 **
## income_composition     4.908e+01  2.160e+00  22.720  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.688 on 150 degrees of freedom
## Multiple R-squared:  0.8266, Adjusted R-squared:  0.8231
## F-statistic: 238.3 on 3 and 150 DF,  p-value: < 2.2e-16
```

```
## [1] 0.82657
```

Remove percentage expenditure

$$y_i = \beta_0 + \beta_2 x_{total_expenditure,i} + \beta_3 x_{gdp,i} + \beta_4 x_{income_composition,i} + \epsilon_i$$

```
##
## Call:
## lm(formula = life_expectancy ~ total_expenditure + gdp + income_composition)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1722  -1.8727   0.3331   1.9982   9.1663
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.557e+01  1.523e+00  23.354  < 2e-16 ***
## total_expenditure   3.074e-01  1.155e-01   2.662  0.00861 **
## gdp                -3.369e-06  1.890e-05  -0.178  0.85872
## income_composition  4.941e+01  2.279e+00  21.679  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.688 on 150 degrees of freedom
## Multiple R-squared:  0.8266, Adjusted R-squared:  0.8231
## F-statistic: 238.3 on 3 and 150 DF,  p-value: < 2.2e-16
```

```
## [1] 0.8265889
```

Remove percentage expenditure & GDP

$$y_i = \beta_0 + \beta_2 x_{total_expenditure,i} + \beta_4 x_{income_composition,i} + \epsilon_i$$

```
##
## Call:
## lm(formula = life_expectancy ~ total_expenditure + income_composition)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1999  -1.9075   0.2171   2.0266   9.2201
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         35.6737     1.3964  25.547  < 2e-16 ***
## total_expenditure    0.3083     0.1150   2.681  0.00816 **
## income_composition  49.1968     1.9415  25.339  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.676 on 151 degrees of freedom
## Multiple R-squared:  0.8266, Adjusted R-squared:  0.8243
## F-statistic: 359.8 on 2 and 151 DF,  p-value: < 2.2e-16
```

```
## [1] 0.8265521
```

To decide whether removing these two factors would have negative impacts on the model's prediction ability, we performed three models that remove either or both of percentage expenditure and GDP. The summary for model held that removing percentage expenditure would not influence the goodness of fit of the model, removing GDP would not influence the goodness of fit of the model and removing both percentage expenditure and GDP would not influence the goodness of fit of the model respectively. which further suggests that removing those two factors would have no significant effect on the model's prediction ability.

Additionally, we also conducted an AIC analysis on every combination of linear models that could be made for the 4 predictor variables (Table 2). That is, for every possible model that can be generated with a combination of the 4 predictor variables, we found their AIC value and compared their results. With this investigation, we found that the model with the lowest AIC value involved total expenditure and income composition of resources as predictor variables. This means the best model selection based on AIC removed both GDP and percentage expenditure as predictor variables. This coincides with our hypothesis that the best model would remove percentage expenditure and GDP from the initial model which in turn further confirms that removing these two variables would provide an appropriate model.

Table 2

	df <dbl>	AIC <dbl>
model1	6	846.5039
model2	5	844.9318
model3	5	844.9151
model4	4	842.9477

4 rows

Results

Table 3

status <chr>	mean(life_expectancy) <dbl>
Developed	81.47500
Developing	69.41984

2 rows

We also use the 'group_by' find out the mean life expectancy for Developed country and Developing country (Table 3). And we find out mean life expectancy for Developed country is 81 years old and Developing country is 69 years old, this meaning that the Developed country usually has higher life expectancy than the Developing country.

Figure 1

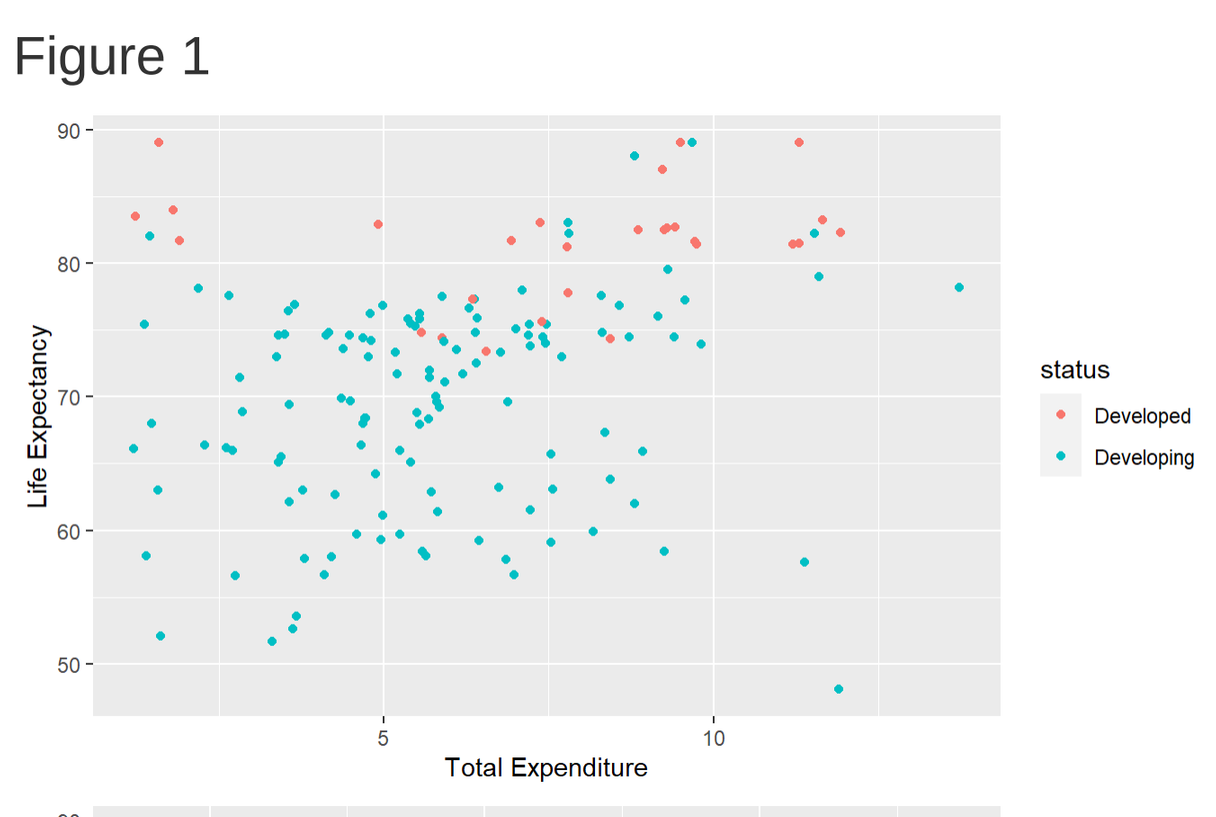
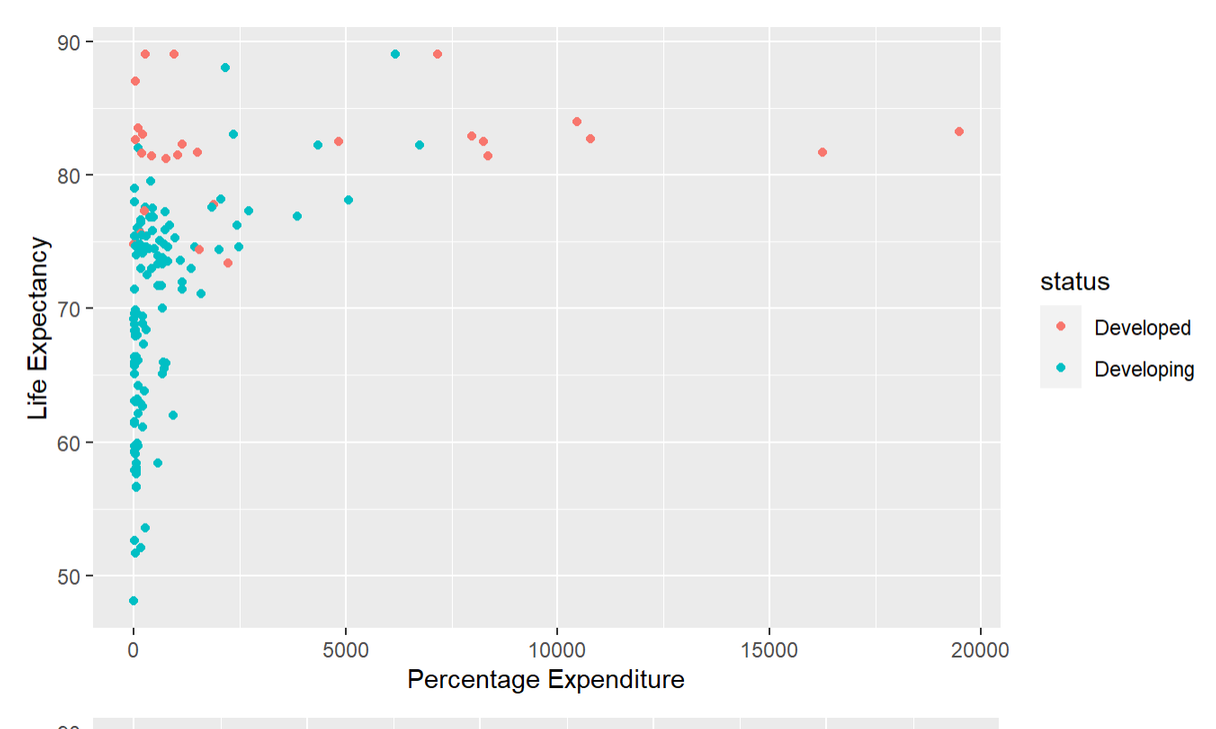
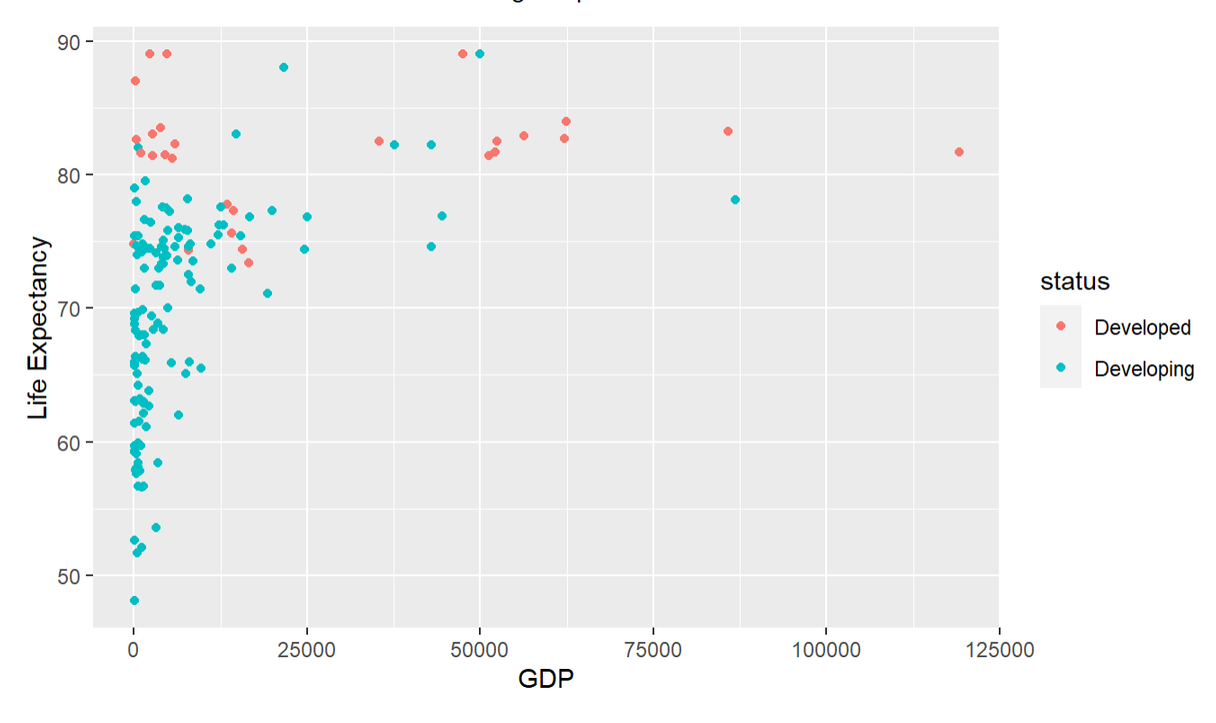


Figure 2

By examining the scatter plots of the response variable life expectancy against both income composition of resources and total expenditure individually (Figure 1&2), we can see that the income composition plot indicates a linear relationship and the total expenditure plot has a relatively random spread that is still positively correlated. This suggests linearity in the relations between life expectancy and the final model's predictor variables. And in Figure 2, we find out all Developed country has high Income Composition of Resources.

Figure 3 & 4





More over, when we check the Figure 3 & 4, we can see that there is not strong liner relationship between Life Expectancy and both GDP and Percentage Expenditure. However we also can find out that when GDP and Percentage Expenditure are high, there is not low Life Expectancy.

Discussion

Summary

Overall, we find out that people in the Developed country will have a higher life expectancy, and the main factor that affect their life expectancy is the income composition of resources for their country. Total expenditure also has a positively correlated with life expectancy, but it is not as strong as income composition of resources shows. For GDP and percentage expenditure, they are also has a small affect to the life expectancy, since when GDP and percentage expenditure are high, there is not low life expectancy. However, it is not necessary to have a high GDP and percentage expenditure, because there also has high life expectancy in low GDP and percentage expenditure. ## Conclusions Regarding our conclusions, from our study, our model showcases that income composition of resources has a greater influence on life expectancy in comparison to total expenditure. This means that the primary variable which affects life expectancy among the 193 countries based on data from 2014 is income composition of resources, or in other words, how productively a country utilizes its resources. Additionally, our model also indicates that if both predictor variables increase, life expectancy is also expected to increase. That is, if a country spends a higher percentage of their government spending on health and allocates their resources in a manner that improves the human development index, life expectancy is predicted to increase.

Moreover, we can see that there exist a causal inference in this data set, which instrument is the country status, treatment is income composition of resources, and outcome is income composition of resources. From World Population Review website, we know that "most developed countries have an HDI score of 0.8 or above. These countries have stable governments, widespread education, health care, high life expectancy, and growing, powerful economies"(World Population Review). And high HDI meaning it will have a high income composition of resources. Also, high income composition of resources meaning that government spending on health and allocates their resources in a manner that improves the human development index, life expectancy is predicted to increase.

Weakness & Next Steps

In general, due to the data set, we can only take into consideration the general economic factors of each country that are present for our model. In addition, the data set contained incomplete data with missing values scattered throughout each variable. This missing data limited our usability of the data set as we initially set to tackle life expectancy data concerning the year 2015, but we were unable to create a sufficient model with over 90% of data missing for one of the economic variables we were interested in. Any missing data we found in the 2014 data was omitted in our model, which in extreme cases, can possibly skew some of our results. The calculated life expectancy might also be biased, since some countries the census might not cover every region of the country. Moreover, in a real world scenario, economic factors are greatly associated with other factors like political, technological, or social influences, which were not included in our model, but could make a significant difference on the life expectancy in cases such as when two countries have similar economic factor readings.

In order to solve our weakness, we will need to doing more survey to get more accurate and diverse economic factors, since there has more economic factors may affect our life expectancy in real life. Moreover, we can also change the topic, which mean we neeed to use not only the economic factors to analysing our life expectancy. We can also use different type of area's factors, in order to understant more about how to let people have more higher life expectancy.

References

1.Finch, C. E. (2009). Evolution of the human lifespan and diseases of aging: Roles of infection, inflammation, and nutrition. Proceedings of the National Academy of Sciences of the United States of America (PNAS), 107, 1718-1724. https://doi.org/10.1073/pnas.0909606106

2.Rajarshi, K. (2017). Life expectancy (WHO): Statistical analysis on factors influencing life expectancy [Data set]. Retrieved from https://www.kaggle.com/kumarajarshi/life-expectancy-who

3.World Health Organization (n.d.). Global Health Observatory (GHO) data: Life expectancy. Retrieved from https://www.who.int/gho/mortality_burden_disease/life_tables/situation_trends/en/

4.World Population Review. Human Development Index (HDI) by Country 2020. Retrieved from https://worldpopulationreview.com/country-rankings/hdi-by-country

Appendix

The link of repository: https://github.com/tonychen1999/STA304-Final-Project-Tony-Chen-.git