**252: Life Expectancy Analytics**

Your submissions:

* Report\_Group number.pdf
* R Codes\_Group number.txt
* R Outputs\_Group number.pdf

For example, Report\_Group 200.pdf, R codes\_Group 200.txt, R outputs\_Group 200.pdf

Notes

* The deadline is in the noon, not midnight
* No extension to the deadline
* Follow the given template
* Each team can only submit one copy by a single member, just list all of your members in the report

|  |  |  |  |
| --- | --- | --- | --- |
| First Name | Last Name | Email (hawk.iit.edu) | Student ID |
| Minguk | Kim | mkim105@hawk.iit.edu | A20437179 |
| Boyun | Jang | bjang7@hawk.iit.edu | A20437298 |

Table of Contents

[**1. Introduction** 2](#_Toc529694108)

[**2. Data** 2](#_Toc529694109)

[**3. Problems to be Solved** 2](#_Toc529694110)

[**4. Solutions** 2](#_Toc529694111)

[**5. Experiments and Results** 2](#_Toc529694112)

[5.1. Methods and Process 2](#_Toc529694113)

[5.2. Evaluations and Results 2](#_Toc529694114)

[5.3. Findings 2](#_Toc529694115)

[**6. Conclusions and Future Work** 2](#_Toc529694116)

[6.1. Conclusions 2](#_Toc529694117)

[6.2. Limitations 2](#_Toc529694118)

[6.3. Potential Improvements or Future Work 2](#_Toc529694119)

# **Introduction**

Life expectancy is one of the factors that measure human development index (HDI) in each country besides human education level and living standard. It is used to describe the quality of life in a particular area. The difference in life expectancy is also cited to demonstrate the need for medical care and the improvement of social support.

In order to prove the necessity of medical and social support for each country, we are going to analyze the factors of life expectancy such as the national mortality rate, economic factors, social factors and other health-related factors.

We will present what kind of improvements are needed to improve the life expectancy of the population by predicting the vulnerable diseases and environments that affect the mortality rate in a specific area in order to improve the mortality rate.

For the result, our main goal is find the factors which are needed to improve the life expectancy.

# **2. Data**

Data sets related to life expectancy and health factors in 193 countries were collected from the same WHO data store website. Economic data was collected on the UN website. Only representative elements were selected from all categories of health-related factors. The data set consists of 22 columns and 2938 rows, which means 20 prediction variables.

|  |  |  |
| --- | --- | --- |
| Variable | Description | Data type |
| Country | Country | Nomial |
| Year | Year | Discrete |
| Status | Developed or Developing status | Binary |
| Life expectancy | Life Expectancy in age | Continous |
| Adult Mortality | Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population) | Discrete |
| Infant deaths | Number of Infant Deaths per 1000 population | Discrete |
| Alcohol | Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol) | Continuous |
| percentage expenditure | Expenditure on health as a percentage of Gross Domestic Product per capita(%) | Continuous |
| Hepatitis B | Hepatitis B (HepB) immunization coverage among 1-year-olds (%) | Discrete |
| Measles | Measles - number of reported cases per 1000 population | Discrete |
| BMI | Average Body Mass Index of entire population | Continuous |
| under-five deaths | Number of under-five deaths per 1000 population | Discrete |
| Polio | Polio (Pol3) immunization coverage among 1-year-olds (%) | Discrete |
| Total expenditure | General government expenditure on health as a percentage of total government expenditure (%) | Continuous |
| Diphtheria | Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%) | Discrete |
| HIV/AIDS | Deaths per 1 000 live births HIV/AIDS (0-4 years) | Continuous |
| GDP | Gross Domestic Product per capita (in USD) | Continuous |
| Population | Population of the country | Continuous |
| thinness 1-19 years | Prevalence of thinness among children and adolescents for Age 10 to 19 (% ) | Continuous |
| thinness 5-9 years | Prevalence of thinness among children for Age 5 to 9(%) | Continuous |
| Income composition of resources | Human Development Index in terms of income composition of resources (index ranging from 0 to 1) | Continuous |
| Schooling | Number of years of Schooling(years) | Continuous |

Source: <https://www.kaggle.com/kumarajarshi/life-expectancy-who>

# **3. Problems to be solved**

The main goal of this project is to predict if the various data elements we collect will affect the average life expectancy.

The following objectives are based on these key objectives.

1. Predict what vulnerable diseases and environments will affect the mortality rate in a given area in order to improve anticipated life expectancy.
2. Analyze the data and discover patterns that can increase life expectancy. For example, we can analyze if increasing medical expenditure or increasing the compulsory schooling period can affect positively the average life expectancy.

# **4. Solutions**

List the solutions for the problems in part 3.

Make sure your solutions can solve the problems in part 3 one by one

If you are going to build predictive models, clearly indicate the dependent and independent variables

The main goal of this project is to predict factors that affect life expectancy and life expectancy. We used multiple linear regressions as a predictive model.

When we had built a model, we found a polynomial and interaction term during the process. Therefore, for the accurate model, we built 3 models by using pure (row) variables, polynomial model and polynomial + interaction term. Finally, we compared among these three models for finding best model.

* Dependent and Independent variables
  + - **Dependent** :

Life Expectancy

* + - **Independent** :

Country

Year

Status

Adult Mortality

Infant deaths

Alcohol

percentage expenditure

Hepatitis B

Measles

BMI

under-five deaths

Polio

Total expenditure

Diphtheria

HIV/AIDS

GDP

Population

thinness 1-19 years

thinness 5-9 years

Income composition of resources

Schooling

# **5. Experiments and Results**

Data size is relatively small(2939), so we use N-Fold evaluation.

To confirm the qualification of the created model, F-test and residual analysis will be performed to check the accuracy. We will analyze the data as the most appropriate model for problem solving through confidence interval and multiple criteria such as p-value test, adjust R-squared test and RMSE verification.

## 5.1. Methods and Process **[input necessary code],[모델비교할 때 AIC말고 에러만비교]**

Solve the problems your proposed one by one

Give the necessary codes, snapshots and explanations

Before start building linear models, we have to preprocess the data first. Understand the data set first, find the missing values and fill these values as the average of each country. Since we will analyze country as continent, separate the counties as 5 continents (Asia, Africa, America, Oceania, and Europe). We’ll convert nominal variables to discrete variables. Country and status columns are nominal; convert these columns to dummy variables.

First of all, we’ll examine linear relationship between X and Y variables via checking the correlation by transforming Y. Also, confirm the transformed X variables which have a small correlation with Y.

Second, build the models by feature selection. Test the VIF to model. If the VIF test value is larger than 4, remove the value for building models. We will perform feature selection methods through backward/forward/both stepwise function and best subset function.

Third, multicollinearity problem can occur because of many variables in our data set. Therefore, we can check the polynomial and interaction term through each variable of the graph. We will identify correlations between x variables and compare variables that have highly correlated with other variables.

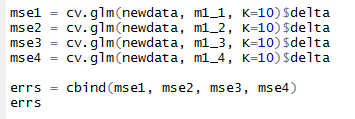
For the fourth, since data size is small we’ll use N-fold evaluation, so we don’t need to check F-test. We are going to verify model is qualified or not. Constant variance, linearity relationship and distribution of residuals will be checked for validation.

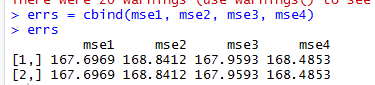
Finally, we will also perform residual analysis and evaluate model by N-fold cross validations. Also, for the accurate model, find the influential points and remove it.

## 5.2. Evaluations and Results **[Compare 3 models]**

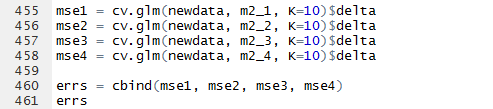
Given a same problem, you may have several solutions or build several models

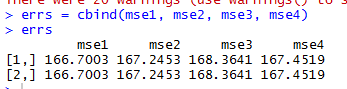
Evaluate your solutions based on selected metrics and compare them





위의 사진들은 linear regression에서 Both, forward, Backward elimination, best subset을 쓴 모델들이다. 보는 것과 같이 m1\_1이 가장 에러가 낮아서 m1\_1을 best model로 선택했다.





위의 사진들은 polynomial linear regression에서 Both, forward, Backward elimination, best subset을 쓴 모델들이다. 보는 것과 같이 m1\_1이 가장 에러가 낮아서 m2\_1을 best model로 선택했다.

각각의 모델(일반/폴리/폴리+인터로부터의 method3개로부터의 모델들)들을 보여주고 평가한다.

## 5.3. Findings **[Final best model]**

Provide the summary of your findings, explanations, conclusions

위의 단계로부터 최적의 모델은 –라는 결론을 얻었다.

결론적으로는 각 x variable에는 폴리노미얼이나 인터엑션 텀이 존재하지 않았다고 볼 수 있으며, 모델을 향상시키기위해 인플루엔셜 포인트를 제거하였다.

우리의 컨클루젼은 밑의 단계에서 언급하겠다.

From the above steps, we conclude that the optimal model is m1\_1\_new.

In conclusion, there is no polynomial variables or interaction term in the optimal model.

We were also able to improve the model by removing Influential Points.

Our conclusions will be addressed in the following steps.

# **6. Conclusions and Future Work**

## 6.1. Conclusions

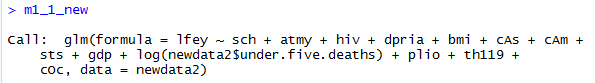
A short summary of your whole project and conclusions, such as what you want to, why you want to do so, which solutions you use, and which findings or final results you get finally.

**Life Expectancy** = 5.746e+01+

(1.003e+00\***Schooling**)+(-1.844e-02\***Adult.Mortality**) + (-4.094e-01\***HIV.AIDS**) +

(3.460e-02\***Diphtheria**) + (3.565e-02\***BMI**)+ (-3.633e+00\*(**Status**=0)) + (5.132e-05\***GDP**) +

(-2.801e-01\*log(**Under.Five.Deaths**)) + (2.157e-02\***Polio**) + (-9.838e-02\***thinness..1.19.years**)



As you can see,[which factors affect life expectancy or not]

세계 데이터를 기준으로 했을 때, 기대 수명에 가장 큰 영향을 끼치는 요소로써는

Schooling, Status이 있다. 교육의 햇수가 증가할수록, 기대 수명 또한 대략 1년 정도 증가하며, 개발도상국이라면 선진국에 비해 대략 -3.6년 정도 짧다는 것을 알 수 있었다.

따라서 기대 수명을 늘리기 위해서는 우선적으로 교육의 발전과 developed country가 되기위한 인프라 형성 등이 매우 중요하다고 유추할 수 있다.

With respect to world data, the most significant factor in life expectancy is Schooling, Status. As the number of years of education increases, life expectancy also increases by about one year. It was also found that developing countries are about 3.6 years shorter than developed countries.

Therefore, in order to increase the life expectancy, it can be inferred that the development of education and the formation of infrastructure to become a developed country are very important.

## 6.2. Limitations

Introduce the limitations of your work

우리의 work의 한계는

어떤 x variable이 기대 수명에 영향을 끼치는지는 알 수 있으나 그 변수에 대한 디테일한 해석이 어렵다는 점이다. . 예를 들어 식에서는 BMI가 높을수록 기대수명도 높아지는 것으로 나와있는데, 우리는 BMI가 높으면 성인병 등에 의해 기대수명이 낮아질 것이라 예측했다. 따라서 해당 요소에 대한 배경지식이 없이는 variable에 대한 해석이 힘든다는 단점이 있다.

전세계의 데이터를 기준으로 하였기 때문에 각각의 지역, 국가에 대해서는 그 변수가 달라질 수 있다. 따라서 데이터 분류를 하고 진행했었어야 더욱 신뢰도 있는 식을 얻었을 것이라고 생각한다.

Our limitation is that we can see which x variable affects life expectancy, but it is difficult to interpret it in detail.

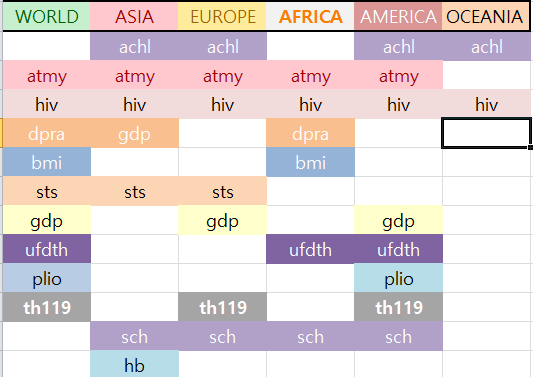
For example, we predicted that high BMI would lower life expectancy due to adult disease or other factors. However, the higher the BMI, the higher the life expectancy.

Therefore, it is difficult to interpret variable without background knowledge.

Secondly, because the data are based on world data, the variables may be different for each region and country. Therefore, I think that if we classified the data by region and proceeded, we would have got a more reliable expression.

## 6.3. Potential Improvements or Future Work

Introduce and discuss possible methods to improve or extend your work in the future



위의 표는 world부터 world를 5가지 분류로 나누어서 linear regression을 진행시켜서 해당 y variable에 영향을 주는 변수들을 정리한 표이다. limitations에서 언급한 것과 같이 world를 베이스로 하여 linear regression을 predict한다면 각 지역별로 환경과 질병중요도, 사망률 등이 다르기 때문에, world를 베이스로 한 linear regression으로부터 기대 수명을 증가시키기 위한 정보를 얻고자 한다면 신뢰도가 낮을 수 밖에 없을 것이다.

따라서 각 지역 또는 국가 별로 식을 만들어 예측을 한다면 훨씬 신뢰할 수 있는 정보를 얻을 수 있을 것 이다.

The table above is a table summarizing the variables affecting the y variable after dividing the world and the world into five categories and proceeding with linear regression.

As mentioned in limitations, if we predict linear regression based on world, environment, disease importance and mortality rate are different in each region. Therefore, if you want to obtain information to increase your life expectancy from a world-based linear regression, this will be unreliable.

Therefore, if you make a prediction by making a linear regression for each region or country, you will get much more reliable information.