UNIVERSITI MALAYA

WIH2001 Data Analytics

Case Study

Name: Amir Firdaus Bin Abdul Hadi

Matric Number: 17204620/2

Lecturer: Professor Madya Ts. Dr. Sri Devi A/p Ravana

Title: Factors affecting Life Expectancy

Table of Contents

1 Introduction

1.1 Problem Statement

i. Does status of a country affect life expectancy?

There are a lot of factors that seems to effect life expectancy of a person, and one of them is the country's status; developing and developed country. And I hypothesize that developed country have longer life expectancy than developing country due to more developed health structure and expenditure towards health sector.

H0: People in developed countries have longer life expectancy than developing countries.

Ha: People in developed countries does not have longer life expectancy than developing countries.

ii. Does Human Development Index in terms of Income Composition of Resources (ICOR) affect life expectancy?

The composition of the total income of a population group or a geographic area refers to the relative share of each income source or group of sources, expressed as a percentage of the aggregate total income of that group or area. This means that if a country utilizes its resources productively, it might make the people live longer than expected.

H0: Countries with high ICOR values have longer life expectancy than countries with lower ICOR values.
Ha: Countries with high ICOR values does not have longer life expectancy than countries with lower ICOR values.

iii. Which linear regression models (GLM, SVM, PCA) has better accuracy for Life Expectancy?

1.2 Description

The dataset chosen is called Life Expectancy by WHO taken from Kaggle dataset. The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries. The data-sets are made available to public for the purpose of health data analysis. The data-set related to life expectancy, health factors for 193 countries have been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website. Among all categories of health-related factors only those critical factors were chosen which are more representative. Therefore, in this project the dataset has considered data from year 2000-2015 for 193 countries for further analysis. The dataset consists of 22 Columns and 2938 rows which meant 20 predicting variables.

All predicting variables was then divided into several broad categories:Immunization related factors, Mortality factors, Economical factors and Social factors:

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

This dataset is important to answer the problem statements above by using variables status, income composition of resources and also life expectancy in further analysis.

2 Data Analytics

2.1 Methodology

For data pre-processing, using “tidyverse” library, rows with missing values are dropped. New data frame is created by selecting only numeric variables for further analysis. Then, to identify which variables significantly correlate with target variable (life expectancy), “ggplot2” and “corrplot” libraries are used to create correlation matrix and removing the variables which are uncorrelated to life expectancy which are between -0.2 and 0.2 indicating weak correlation. The datasets then split into 2 new datasets by status; developing and developed countries.

In this case study of factors affecting life expectancy, the methodology used for data exploration is by descriptive and inferential analysis. For descriptive analysis, “psych” library will be used to study the measures of central tendency, dispersion and skewness of the dataset. For inferential analysis, using “caret” library, linear regression and multiple linear regression are used. ANOVA test is done to find the significance variable in the model. Linear model that are used are generalized linear model, principal component regression and support vector machine. Metric used is RMSE with 10-fold cross validation to choose the best models.

2.2 Code, Results and Explanation

2.2.1 Data Cleaning

In the dataset, referring to appendix 1, we found that there are a lot of missing values across the table. Hence, decided to drop the rows with missing values in order to draw more accurate conclusion from the analysis.

Then, preparation of the data is done by removing non-numeric variables such as country, year and status for linear regression model and correlation matrix. From the correlation matrix in figure 3, uncorrelated variables will be truncated and only used variables that have correlation to life expectancy

2.2.2 Measures of central tendency and dispersion

```
> describe(life)
                                vars    n        mean          sd    median   trimmed       mad
Country*                           1 1649       67.19       38.89     68.00     67.23     50.41
Year                               2 1649     2007.84        4.09   2008.00   2007.97      4.45
Status*                            3 1649        1.85        0.35      2.00      1.94      0.00
Life.expectancy                    4 1649       69.30        8.80     71.70     69.91      7.56
Adult.Mortality                    5 1649      168.22      125.31    148.00    153.60    109.71
infant.deaths                      6 1649       32.55      120.85      3.00     10.62      4.45
Alcohol                            7 1649        4.53        4.03      3.79      4.12      4.74
percentage.expenditure             8 1649      698.97     1759.23    145.10    274.40    198.60
Hepatitis.B                        9 1649       79.22       25.60     89.00     85.04     11.86
Measles                           10 1649     2224.49    10085.80     15.00    276.04     22.24
BMI                               11 1649       38.13       19.75     43.70     38.89     23.43
under.five.deaths                 12 1649       44.22      162.90      4.00     14.62      5.93
Polio                             13 1649       83.56       22.45     93.00     88.94      7.41
Total.expenditure                 14 1649        5.96        2.30      5.84      5.93      2.25
Diphtheria                        15 1649       84.16       21.58     92.00     89.21      8.90
HIV.AIDS                          16 1649        1.98        6.03      0.10      0.50      0.00
GDP                               17 1649     5566.03    11475.90   1592.57   2779.47   2077.51
Population                        18 1649 14653625.89 70460393.40 1419631.00 4330096.73 2062822.92
thinness..1.19.years              19 1649        4.85        4.60      3.00      4.07      3.11
thinness.5.9.years                20 1649        4.91        4.65      3.20      4.13      3.41
Income.composition.of.resources   21 1649        0.63        0.18      0.67      0.65      0.16
Schooling                         22 1649       12.12        2.80     12.30     12.18      2.82
```

	min	max	range	skew	kurtosis	se
Country*	1.00	1.330000e+02	1.320000e+02	-0.02	-1.21	0.96
Year	2000.00	2.015000e+03	1.500000e+01	-0.20	-1.06	0.10
Status*	1.00	2.000000e+00	1.000000e+00	-1.99	1.98	0.01
Life.expectancy	44.00	8.900000e+01	4.500000e+01	-0.63	0.03	0.22
Adult.Mortality	1.00	7.230000e+02	7.220000e+02	1.27	2.38	3.09
infant.deaths	0.00	1.600000e+03	1.600000e+03	8.46	84.95	2.98
Alcohol	0.01	1.787000e+01	1.786000e+01	0.66	-0.60	0.10
percentage.expenditure	0.00	1.896135e+04	1.896135e+04	4.97	30.85	43.32
Hepatitis.B	2.00	9.900000e+01	9.700000e+01	-1.79	2.21	0.63
Measles	0.00	1.314410e+05	1.314410e+05	7.94	74.35	248.37
BMI	2.00	7.710000e+01	7.510000e+01	-0.23	-1.27	0.49
under.five.deaths	0.00	2.100000e+03	2.100000e+03	8.33	81.83	4.01
Polio	3.00	9.900000e+01	9.600000e+01	-2.36	5.03	0.55
Total.expenditure	0.74	1.439000e+01	1.365000e+01	0.21	-0.02	0.06
Diphtheria	2.00	9.900000e+01	9.700000e+01	-2.48	5.82	0.53
HIV.AIDS	0.10	5.060000e+01	5.050000e+01	4.97	27.63	0.15
GDP	1.68	1.191727e+05	1.191711e+05	4.51	27.89	282.60
Population	34.00	1.293859e+09	1.293859e+09	14.16	229.18	1735140.86
thinness..1.19.years	0.10	2.720000e+01	2.710000e+01	1.82	4.13	0.11
thinness.5.9.years	0.10	2.820000e+01	2.810000e+01	1.86	4.49	0.11
Income.composition.of.resources	0.00	9.400000e-01	9.400000e-01	-1.15	2.05	0.00
Schooling	4.20	2.070000e+01	1.650000e+01	-0.13	0.04	0.07

Figure 1: Central tendency, dispersion, skewness

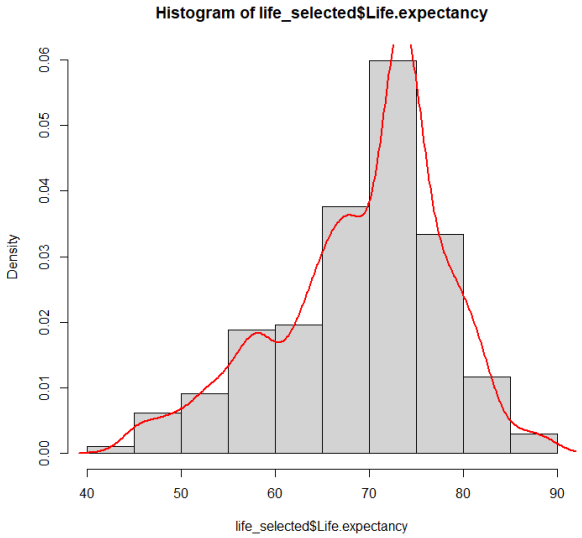

Figure 2: Life Expectancy Histogram

Based on figure above, we can see the summary of the central tendency for each variable in terms of mean, median and mode and also the dispersion of the data. We can see that some variables are very skewed and not normalised such as population, infant deaths and under five deaths. And from the histogram, that on average, people have a life expectancy around 70 – 75 years old. And maximum life expectancy of 89 years old and minimum of 44 years old.

2.2.3 Correlation Matrix

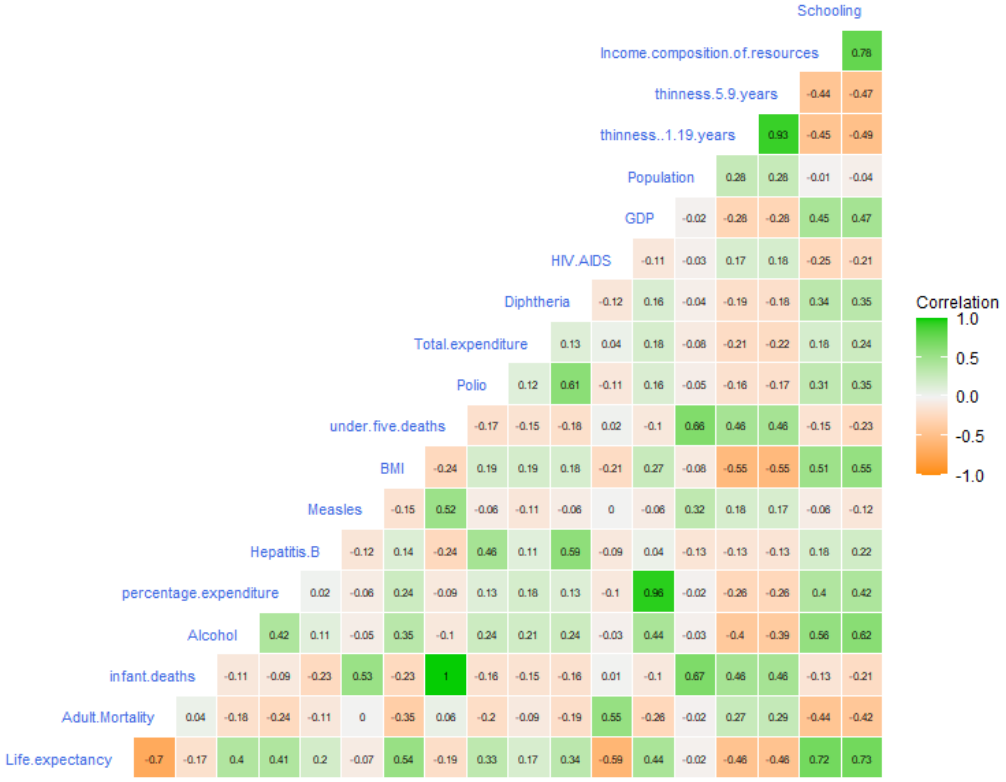

Figure 3: Correlation Matrix

In figure 4, we can see the correlation values of predicting variables to target variables (Life Expectancy), where darker green colour depicts stronger positive correlation while darker orange depicts stronger negative correlation. “Adult mortality” and “HIV/AIDS” have strong negative correlation while “income composition of resources” and “schooling” has strong positive correlation. Variables with low correlation values between -0.2 to 0.2 are truncated as part of feature selection; “population”, “measles”, “infant deaths”, “under-five deaths” and “total expenditure”.

2.2.4 Problem Statement 1: Does status of a country affect life expectancy?

Status	Average Life Expectancy (years)
Developed	78.69
Developing	67.69

Table 1: Average Life Expectancy based on status

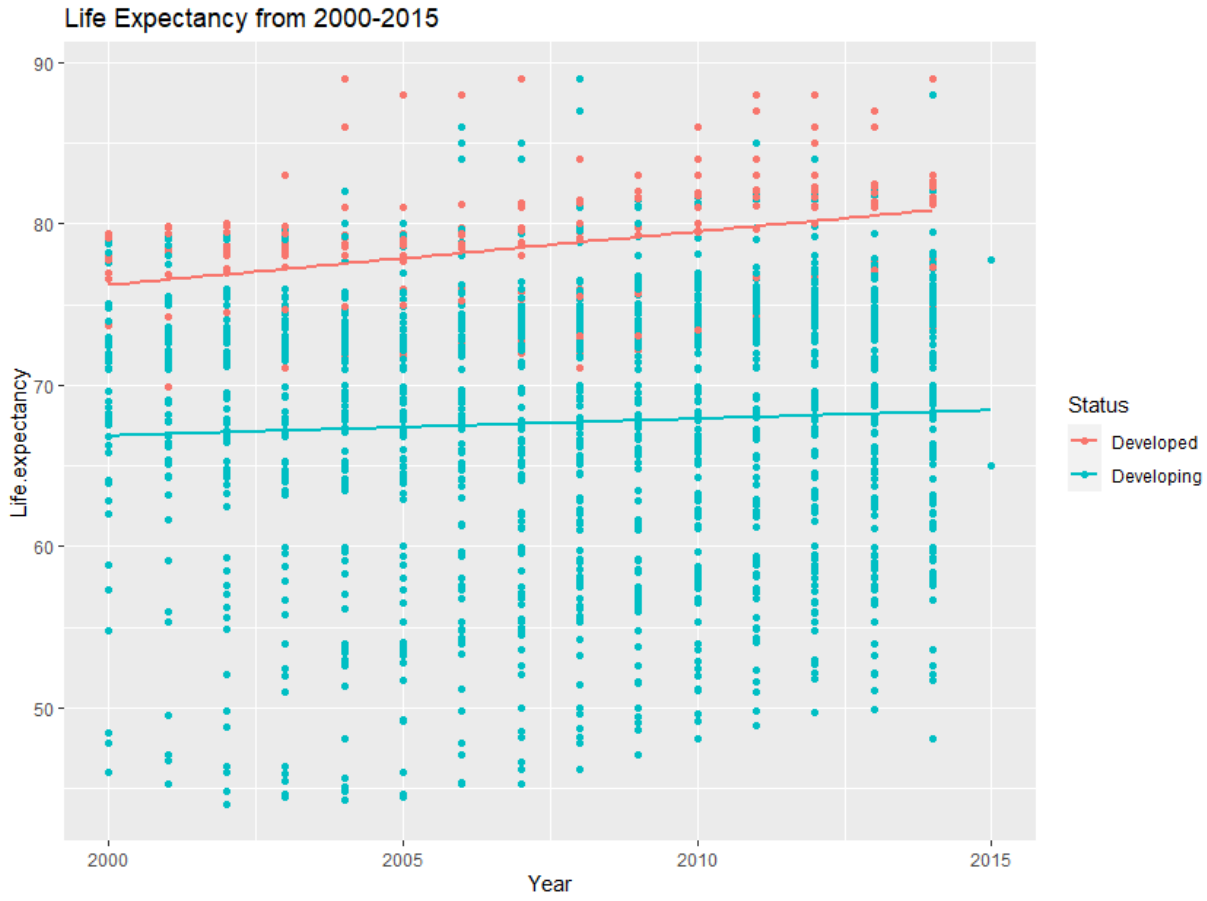

Figure 4: Scatterplot of life expectancy with linear model

Based on table 1, it is evident that on average, developed countries have longer life expectancy (79 years) than people in developing countries (68 years). In figure 4, pink coloured dots and line depicts developed countries have higher increasing linear line than developing countries. Developed data are scattered more on the upper part of the plot while developing data on the lower side. This proof that developed countries do live longer than people in developing countries.

H0: People in developed countries have longer life expectancy than developing countries. Null hypothesis accepted.

2.2.5 Problem Statement 2: Does Income Composition of Resources (ICOR) affect life expectancy?

```
> #lm model for all countries
> life_model <- lm(formula = Life.expectancy ~., data = life_selected)
> summary(life_model)

Call:
lm(formula = Life.expectancy ~ ., data = life_selected)

Residuals:
     Min       1Q   Median       3Q      Max
-17.2956  -2.1063   0.0486   2.3496  12.0706

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.235e+01  7.137e-01  73.341  < 2e-16 ***
Adult.Mortality                -1.785e-02  9.623e-04 -18.551  < 2e-16 ***
Alcohol                        -1.120e-01  3.059e-02  -3.662 0.000259 ***
percentage.expenditure          3.964e-04  1.848e-04   2.145 0.032105 *
Hepatitis.B                    -4.581e-03  4.495e-03  -1.019 0.308283
BMI                             3.545e-02  6.130e-03   5.784 8.74e-09 ***
Polio                           1.106e-02  5.286e-03   2.093 0.036526 *
Diphtheria                      2.058e-02  6.052e-03   3.401 0.000686 ***
HIV.AIDS                       -4.345e-01  1.821e-02 -23.857  < 2e-16 ***
GDP                             1.108e-05  2.910e-05   0.381 0.703547
thinness..1.19.years           -3.439e-02  5.404e-02  -0.636 0.524591
thinness.5.9.years             -3.779e-02  5.324e-02  -0.710 0.477881
Income.composition.of.resources  1.018e+01  8.499e-01  11.977  < 2e-16 ***
Schooling                       9.297e-01  6.025e-02  15.431  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.691 on 1635 degrees of freedom
Multiple R-squared:  0.8254,    Adjusted R-squared:  0.824
F-statistic: 594.4 on 13 and 1635 DF,  p-value: < 2.2e-16
```

Figure 5: Multiple Linear Regression for Life Expectancy

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

Response: Life.expectancy
                                 Df Sum Sq Mean Sq   F value    Pr(>F)
Adult.Mortality                   1  62941   62941 4620.8903 < 2.2e-16 ***
Alcohol                           1  10272   10272  754.1405 < 2.2e-16 ***
percentage.expenditure            1   2890    2890  212.2018 < 2.2e-16 ***
Hepatitis.B                       1   1493    1493  109.6400 < 2.2e-16 ***
BMI                               1   6179    6179  453.6072 < 2.2e-16 ***
Polio                             1   1016    1016   74.5952 < 2.2e-16 ***
Diphtheria                        1    776     776   56.9608 7.344e-14 ***
HIV.AIDS                          1   8448    8448  620.2433 < 2.2e-16 ***
GDP                               1    353     353   25.9287 3.953e-07 ***
thinness..1.19.years              1    528     528   38.7912 5.985e-10 ***
thinness.5.9.years                1      4       4    0.3211     0.571
Income.composition.of.resources   1   7115    7115  522.3413 < 2.2e-16 ***
Schooling                         1   3243    3243  238.1116 < 2.2e-16 ***
Residuals                      1635  22270      14
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6: Anova test for the linear model

	Country	Average ICOR	Average Life Expectancy
1	Netherlands	0.919	81.3
2	Australia	0.918	81.9
3	Ireland	0.905	83.4
4	Sweden	0.904	81.9
5	Canada	0.896	82.2

Table 2: Top 5 Countries with highest ICOR values

	Country	Average ICOR	Average Life Expectancy
1	Bhutan	0.156	65.9
2	Turkmenistan	0.208	64.6
3	Eritrea	0.243	61.1
4	Niger	0.327	61.4
5	Burundi	0.342	56.0

Table 3: Top 5 Countries with lowest ICOR values

Based on figure 5 of linear regression and ANOVA test in figure 6, we can see that Income Composition of Resources (ICOR) has significant values to life expectancy with p-value < 0.05. Also, in figure 3 of correlation matrix, ICOR has correlation value of 0.72 which is the second highest next to schooling. This strong positive correlation of ICOR and also significant value in the ANOVA test shows that if a country utilizes its resources productively, it is more likely to see its citizens live longer than expected.

Looking at table 2 and 3, top 5 countries with highest ICOR values have more than 80 years in life expectancy while top 5 countries with lowest ICOR values have below than 66 years in life expectancy. H0: Countries with high ICOR values have longer life expectancy than countries with lower ICOR values. Null hypothesis is accepted.

Other than ICOR, adult mortality, alcohol, BMI, diphtheria, HIV/AIDS, and schooling also shows significant values towards life expectancy.

2.2.6 Testing different linear regression model

GLM

GLM models allow us to build a linear relationship between the response and predictors, even though their underlying relationship is not linear. This is made possible by using a link function, which links the response variable to a linear model. Unlike Linear Regression models, the error distribution of the response variable need not be normally distributed. The errors in the response variable are assumed to follow an exponential family of distribution (i.e., normal, binomial, Poisson, or gamma distributions). Since we are trying to generalize a linear regression model that can also be applied in these cases, the name Generalized Linear Models.

SVM

Support Vector Regression is a supervised learning algorithm that is used to predict discrete or continuous values. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.

PCR

Principal component regression (PCR) is a regression analysis technique that is based on principal component analysis (PCA). More specifically, PCR is used for estimating the unknown regression coefficients in a standard linear regression model. In PCR, instead of regressing the dependent variable on the explanatory variables directly, the principal components of the explanatory variables are used as regressors.

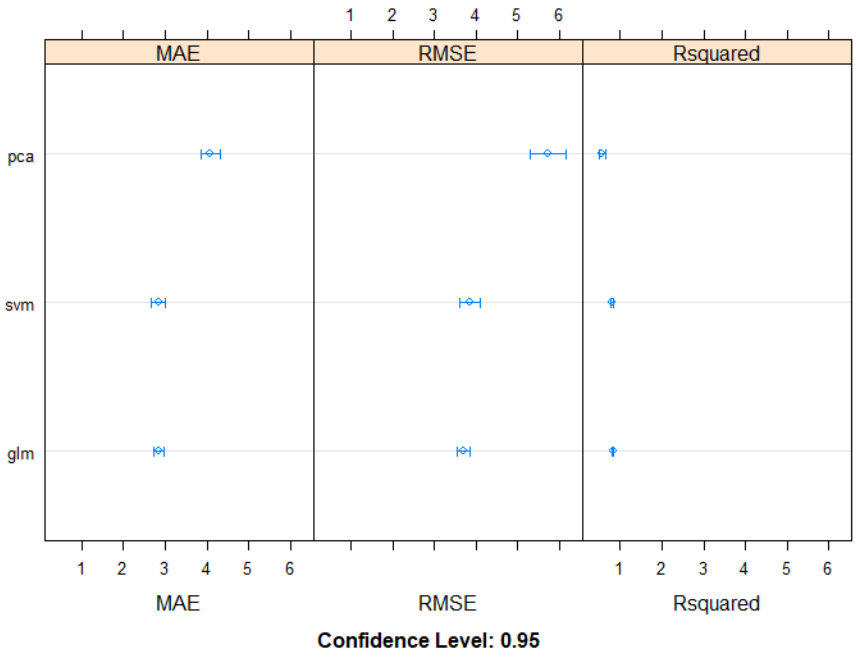

Figure 9: Dotplot for linear model results

	Models	RMSE
1	Generalized Linear Model	3.705
2	Support Vector Machine	3.854
3	Principal Component Regression	5.730

Table 4: RMSE of different linear regression models

Based on figure 9 and table 4, GLM has the lowest RMSE than SVM and PCA. Hence, with lowest mean squared error in GLM, GLM is the most accurate model for life expectancy linear regression model.

3 Conclusion

3.1 Summary of findings

After doing data analysis on Life Expectancy datasheet by WHO, the results showed that:

- Variable “population”, “measles”, “infant deaths”, “under-five deaths” and “total expenditure” has no correlation to Life Expectancy.
- Variable “income composition of resources” and “schooling” has strong positive correlation to Life Expectancy. The increase of this variable’s values will increase life expectancy of people.
- Variable “adult mortality rate” and “HIV/AIDS” has strong negative correlation to Life Expectancy. The increase of this variable’s values will decrease life expectancy of people.
- People in developed countries live longer than people in developing countries by 16%. But both have shown an increase in life expectancy over the years.
- ICOR values do have a significant effect on Life Expectancy and countries with high ICOR values have longer life expectancy.
- Generalized Linear Model is the best performed linear regression model in this analysis with RMSE of 3.705.

3.2 Limitation of study

The limitation of this study is the data is collected up until 2015 only. Lack of new data might reduce the relevance of this study. The data set provided also have a lot of missing and inaccurate data that can affect the outcome of the analysis. Other than that, there should be more relevant variables to be put in the dataset to provide more insights on life expectancy.

3.3 Things to improve

Things that can be improve for this analysis is to clean the data in more details. This is to make sure that all values used are within correct range and possible. Other than that, the dataset needs to cover more variables on lifestyle like food consumption, exercise etc. to make it more relevant. Last but not least, other linear regression models should be tested to find more better performing models. And feature selection should be done more in depths to remove unnecessary features that can increase the accuracy of the models.

4 Reference

15 types of regression in data science. (n.d.). Retrieved from https://www.listendata.com/2018/03/regression-analysis.html

Ganiyu, M. (2021, March 30). How different factors have an influence on your life expectancy? Retrieved from https://towardsdatascience.com/how-different-factors-have-an-influence-on-your-life-expectancy-7b807b04f33e

Ggcorr function. (n.d.). Retrieved from https://www.rdocumentation.org/packages/GGally/versions/1.5.0/topics/ggcorr

Ggcorr: Correlation matrixes with ggplot2. (2015, September 11). Retrieved from https://briatte.github.io/ggcorr/

How to make scatter plot with regression line with ggplot2 in R? (2020, May 4). Retrieved from https://datavizpyr.com/scatter-plot-with-regression-line-with-ggplot2-in-r/

KUMARRAJARSHI. (2019). Life expectancy (WHO). Retrieved from https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?datasetId=12603&language=R&outputs=Visualization

Life expectancy. (2005, October 21). Retrieved from https://ourworldindata.org/life-expectancy

5 Appendix

Appendix 1: Missing values in dataset

```
> sapply(life, function(x) sum(is.na(x)))
                         Country                             Year
                               0                                0
                          Status                  Life.expectancy
                               0                               10
                 Adult.Mortality                     infant.deaths
                              10                                0
                         Alcohol           percentage.expenditure
                             194                                0
                     Hepatitis.B                          Measles
                             553                                0
                             BMI                 under.five.deaths
                              34                                0
                           Polio               Total.expenditure
                              19                              226
                      Diphtheria                         HIV.AIDS
                              19                                0
                             GDP                       Population
                             448                              652
           thinness..1.19.years                thinness.5.9.years
                              34                               34
Income.composition.of.resources                         Schooling
                             167                              163
```

Appendix 2: Link to R-Code, csv file and plot:

https://drive.google.com/drive/folders/19YqSsZMKBd0GOusa0X3huTB2cfQAQp0l?usp=sharing

Appendix 3: R-Code

```
#loading library
library(psych)
library(tidyverse)
library(corrplot)
library(ggplot2)
library(GGally)
library(caret)
library(caTools)

#reading csv file
life <- read.csv("Life Expectancy Data.csv")

#descriptive analysis

#measures of central tendency, dispersion and skewness
describe(life)
sapply(life, function(x) sum(is.na(x)))

#removing na values
life <- life %>% drop_na()

#select only numeric variables for correlation matrix
life_selected <- life %>% select(-Country, -Year, -Status)
life_num <- life_selected %>% select_if(is.numeric)

#histogram of life expectancy with density
hist(life_selected$Life.expectancy, prob = TRUE)
lines(density(life_selected$Life.expectancy), lwd = 2, col ="red")
boxplot(life_selected$Life.expectancy)
```

```
#correlation matrix
ggcorr(life_num,
       label = T,
       label_size = 2,
       label_round = 2,
       hjust = 1,
       size = 3,
       color = "royalblue",
       layout.exp = 5,
       low = "darkorange",
       mid = "gray95",
       high = "green3",
       name = "Correlation")

#Removing the variables which are uncorrelated to life expectancy
#-0.2 > x > 0.2
life_selected <- life_selected %>%
  select(-Population, -Measles, -infant.deaths,
         -under.five.deaths, -Total.expenditure)

#splits data set into 2 list by different Status (developed/developing).
life_split<-split(life_selected,life$Status)

#Separates data set based on Developed/Developing
life_developing<-life_split$Developing
life_developed<-life_split$Developed

#average life expectancy for developing and developed
mean_life <- aggregate(life$Life.expectancy, list(life$Status), FUN=mean)
```

```
print(mean_life)
#how Life Expectancy has changed for both developing and developed countries from 2000-2015
ggplot(data=life,
       mapping=aes(Year,Life.expectancy,color=Status))+
  geom_point()+geom_smooth(method="lm",se=FALSE)+
  labs(title="Life Expectancy from 2000-2015")

icor <- life %>%
  group_by(Country) %>%
  summarise(avg_life = mean(Life.expectancy),
            avg_icor = mean(Income.composition.of.resources))
icor <- icor %>% arrange(desc(avg_icor))

#inferential statistic - linear regression/anova

#lm for developing countries
life_developing_model <- lm(formula = Life.expectancy ~., data = life_developing)
summary(life_developing_model)

#lm for developed countries
life_developed_model <- lm(formula = Life.expectancy ~., data = life_developed)
summary(life_developed_model)

#lm model for all countries
life_model <- lm(formula = Life.expectancy ~., data = life_selected)
summary(life_model)

anova(life_model)

#testing different linear models (lm, glm, pca, svm)
```

```
trainControl <- trainControl(method = "cv", number = 10)
metric <- "RMSE"

set.seed(7)
fit.lm <-train(Life.expectancy ~.,
               data=life_selected, method="lm", metric=metric,trControl=trainControl)

set.seed(7)
fit.glm <-train(Life.expectancy ~.,
                data=life_selected, method="glm", metric=metric,trControl=trainControl)

set.seed(7)
fit.pca <-train(Life.expectancy ~.,
                data=life_selected, method="pcr", metric=metric,trControl=trainControl)

set.seed(7)
fit.svm <-train(Life.expectancy ~.,
                data=life_selected, method="svmLinear",
metric=metric,trControl=trainControl)

#summarize the results
results <- resamples(list( glm = fit.glm, pca = fit.pca,
                           svm = fit.svm))
summary(results)
dotplot(results)

#printing best model glm/lm
print(fit.glm)
```