LIFE EXPECTANCY ANALYSIS

Team 80

Word Count 1145

University of Essex

Table of Contents

INTRODUCTION

Life expectancy has significant indicator of development of a country as well as whole human civilization. Our aim is to analyze the given data and find the best predictor model which will predict the life expectancy and find out which indicators on the dataset have most influence on life expectancy.

EXPLANATORY DATA ANALYSIS

To understand the given world development indicator expectancy of life datasets. Therefore, to conduct the following process an explanatory data analysis is conducted. In the report, a descriptive statistical analysis process is conducted to elaborate and explain the analysis of the datasets, the data-based information is reflected using a few tools. A descriptive abstract is developed to do the comparative study of the child mortality value with the hemisphere. To perform the concerned analysis the use of R studio has been done (Ali, and et.al, 2022). Multiple data segments in the world development indicator in the datasheets related to the child mortality stats consist of a very significant number of metrics, and the presence of the omitted variable in the data is also identified.

DESCRIPTIVE STATISTICS

Various mathematical variables are highlighted by considering the use of statistical analysis. The concerned data set analysis consists of a mathematical feature known as life expectancy. Based on the different continents' lowest, the confidence interval, mean, and R-value of the developed statistical calculation have been marked. The following is the segmentation of the summer analysis that was utilized.

```
$Africa
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  53.28   60.72   63.98   64.11   66.96   76.88

$Asia
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  64.83   71.13   74.29   74.62   77.59   85.08

$`Australia/Oceania`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  64.50   70.69   72.93   73.34   75.48   82.90

$Europe
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  71.83   75.40   78.85   78.62   82.34   83.70

$`North America`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  64.00   72.93   74.48   75.41   78.72   82.05

$`South America`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  69.91   71.97   76.27   75.09   77.08   80.18
```

Figure 1 descriptive stats

From the above breakdown structure, it can be determined the average expectancy of life in the cited continents. It has been observed that the African continent has an average life expectancy of around 64.11 years, along with a maximum age of 76.8 and the least age of 53.28. the life expectancy values of the third quartile and the first quadrant are also shown in the above-created table. The above table highlights the complex lifetime period value reflecting that Europe continent has the highest infant mortality rate compared to all the cited continents. The average life expectancy of the European landmass is 78.62 years (Salvatore, 2021). It has been identified that the African continent has the lower average life expectancy whereas the Asian peninsula region has the greatest life expectancy with an average measure of 85.08 years. Although the average life expectancy in Asia is 74.62 years.

IMPUTATION OF MISSING VALUE

The cited columns containing various sorts of data for the recommended set of data contain a high rate of omission in the data. the following picture represents the dataset's entirely null value.

Country.Name	Country.Code	Continent	SP.DYN.LE00.IN	EG.ELC.ACCS.ZS	NY.ADJ.NNTY.KD.ZG	NY.ADJ.NNTY.PC.KD.ZG
0	0	0	19	1	79	79
SH.HIV.INCD.14	SE.PRM.UNER	SE.PRM.CUAT.ZS	SE.TER.CUAT.BA.ZS	SP.DYN.IMRT.IN	SE.PRM.CMPT.ZS	SE.ADT.LITR.ZS
127	99	181	179	24	89	192
FR.INR.RINR	SP.POP.GROW	EN.POP.DNST	SP.POP.TOTL	SH.XPD.CHEX.PC.CD	SH.XPD.CHEX.GD.ZS	SL.UEM.TOTL.NE.ZS
104	1	1	1	31	31	96
NY.GDP.MKTP.KD.ZG	NY.GDP.PCAP.CD	SP.DYN.CBRT.IN	EG.FEC.RNEW.ZS	SH.HIV.INCD	SH.H2O.SMDW.ZS	SI.POV.LMIC
14	12	13	217	88	89	195
SE.COM.DURS						
19						

Figure 2 NA values in the data

For the provided data column, the great value of the number that is not able to identify is 217. As per the concerned case, not considering the rows with filtered data will not prove to be an intelligent decision taken by the person responsible. Although all the data can be erased. Hence, it is very important that all the null values are computed properly (Jaeger, and et.al 2021). The data column's average value is utilized to fix the null figures in the simplest manner. to make the swapping of the data possible it is very crucial that the datasets are formed with utmost symmetricity (Cahan, and et.al, 2022). The alternates for the numerical attributes are median and mode values. Still exchanging the median and mode value with the null value will skew the entire data that has been developed. There are extremes in the data-sets it is customary to exchange the median with a negative value. Since every data column in the selected dataset has null values, all of the data columns have been changed. Before computing the statistical data, the null values are removed using the average values of the data column.

FITTING THE MODEL

In the previous section, it has been established that the null values are replace with mean values. After imputation, data still have 28 columns and 217 rows. In the data, we try the both forward selection and backward elimination on data to fit the best model. After forward selection, we left with 10 features namely, SP.DYN.IMRT.IN, SP.DYN.CBRT.IN, SP.POP.GROW, SH.H2O.SMDW.ZS, SH.HIV.INCD, EG.ELC.ACCS.ZS, SH.XPD.CHEX.GD.ZS, EN.POP.DNST, SH.HIV.INCD.14 and SE.PRM.CUAT.ZS. Newly affected HIV adult have a negative impact on life expectancy. Having primary education and safe drinking water also have a great impact on the life expectancy. The better water the population of a country have, the more life expectancy expected. Access to electricity also have a positive impact on the population. The

people of a country tend to have more life span if they have access to electricity. Unsurprisingly, expense on health structure also have a beneficial effect on life expectancy. Surprisingly, newly affected children with HIV have a positive impact on the population. But, it reject the null hypothesis. It can be assumed that some of the unwanted features are present due to lack of collinearity testing. We also tried backward elimination process. However the result was not satisfactory because AIC is infinity. So, for now, forward selection process is only used.

P - Value

P – value less than 1 and closer to 0 in a negative exponential form. So, it can be said that null hypothesis is rejected.

Residual Standard Error

From the best model, the residual standad error is 2.575 with 206 degree freedom.

R Squared

From the best model, Multiple R-squared is 0.8757 and Adjusted R-squared is 0.8697

VIF

SP.DYN.IMRT.IN	SP.DYN.CBRT.IN	SP.POP.GROW	SH.H2O.SMDW.ZS	SH.HIV.INCD	EG.ELC.ACCS.ZS
4.093981	6.527991	2.496344	1.735122	2.126599	3.321617
SH.XPD.CHEX.GD.ZS	EN.POP.DNST	SH.HIV.INCD.14	SE.PRM.CUAT.ZS		
1.180165	1.060365	2.274987	1.096343		

Figure 3 VIF

From VIF values, it can be said all the predictor variables are correlated as all the values are greater than 1. Birth rate, mortality rate and electricity access have the biggest VIF values. Now, it becomes clear why HIV has a positive impact on the population. It is due to correlation.

Pairwise Correlation

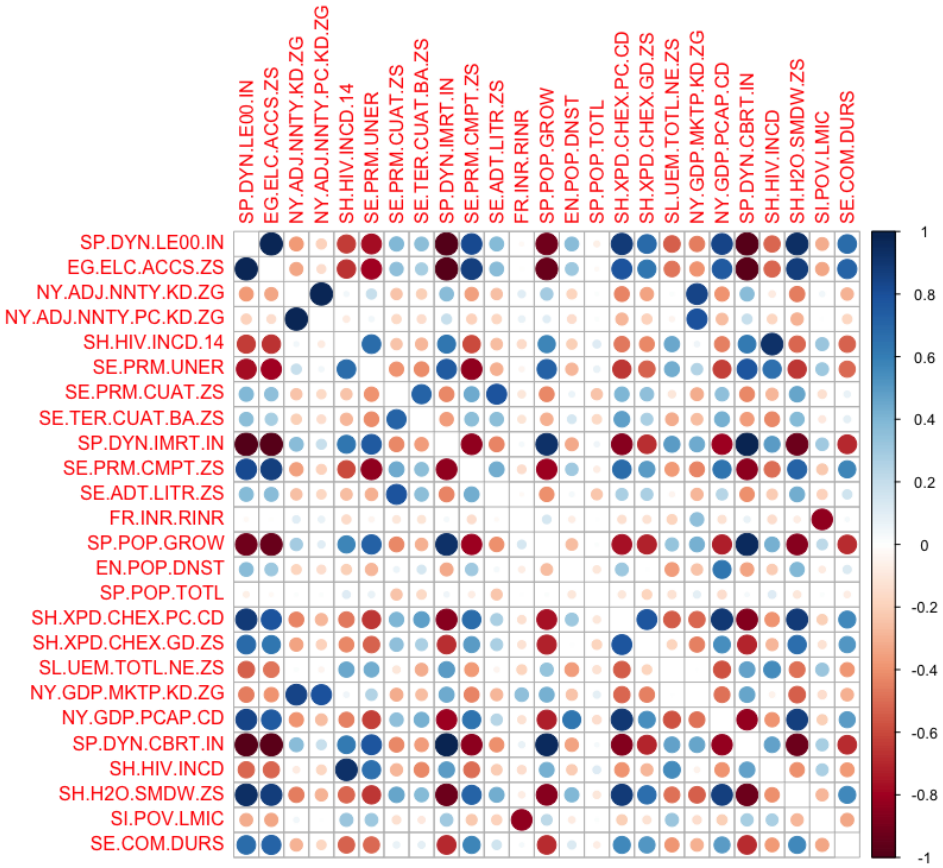

Figure 4 Pairwise Correlation

From pairwise correlation table, correlation between models become clearer. For highly correlated predictors, we could just keep one variable to achieve same effect with the model with less predictors. As a result, we would get a simpler model.

Anova Test

```
Analysis of Variance Table

Model 1: SP.DYN.LE00.IN ~ SP.DYN.IMRT.IN + SP.DYN.CBRT.IN + SP.POP.GROW +
    SH.H2O.SMDW.ZS + SH.HIV.INCD + EG.ELC.ACCS.ZS + SH.XPD.CHEX.GD.ZS +
    EN.POP.DNST + SH.HIV.INCD.14 + SE.PRM.CUAT.ZS
Model 2: SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + SE.PRM.CUAT.ZS + SP.DYN.IMRT.IN +
    SP.POP.GROW + EN.POP.DNST + SH.XPD.CHEX.GD.ZS + SP.DYN.CBRT.IN +
    SH.HIV.INCD + SH.H2O.SMDW.ZS
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    206 1366.2
2    207 1389.3 -1   -23.121 3.4863 0.0633 .
```

Figure 5 Anova Test

After removing SH.HIV.INCD.14 Children (ages 0-14) newly infected with HIV , model become smaller but with same result. Anova test show that, we fail to reject the null hypothesis at any reasonable significance level. So, the smaller model is preferred.

LIMITATION & IMPROVEMENT

For best results, checking collinearity to eradicate predictors was important. Due to absence of one of our members we had to skip this process. Another improvement would be to try multiple imputation and compare result with mean imputation.

CONCLUSION

From the data analysis, imputation and model fitting procedures, it can be observed that life expectancy depends on the 9 indicators on the dataset. Among those, electricity access and safely managed drinking water has great positive impact on the life expectancy. On the other hand, children out primary education and newly infected HIV patients have adverse effect on life expectancy.

REFERENCES

Ali, S.A.G., Al-Fayyadh, H.R.D., Mohammed, S.H. and Ahmed, S.R., 2022, June. A Descriptive Statistical Analysis of Overweight and Obesity Using Big Data. In *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (pp. 1-6). IEEE.

Salvatore, D., 2021. Schaums outline of theory and problems of statistics and econometrics.

Jaeger, B.C., Cantor, R., Sthanam, V., Xie, R., Kirklin, J.K. and Rudraraju, R., 2021. Improving outcome predictions for patients receiving mechanical circulatory support by optimizing imputation of missing values. *Circulation: Cardiovascular Quality and Outcomes*, *14*(9), p.e007071.

Cahan, E., Bai, J. and Ng, S., 2022. Factor-based imputation of missing values and covariances in panel data of large dimensions. *Journal of Econometrics*.

APPENDIX

```
install.packages("tidyverse")

install.packages("GGally")

install.packages("imputeTS")

install.packages("dplyr")


library(tidyverse)

library(GGally)

library(imputeTS)

library(dplyr)


# Accessing the Data set using read.csv()

exp_t=read.csv("Life_Expectancy_Data1.csv")

# Preview of the imported data set

view(exp_t)

print(dim(exp_t))

# Performing the Explanatory Analysis

is.null(exp_t)

is.na(exp_t)

colSums(is.na(exp_t))

df = summary(exp_t$SP.DYN.LE00.IN)

print(df)

par(2,2)
```

```
hist(exp_t$SP.DYN.LE00.IN)

hist(exp_t$EG.ELC.ACCS.ZS)

hist(exp_t$NY.ADJ.NNTY.KD.ZG)

hist(exp_t$SP.POP.GROW)

# Imputation of the NA values with Mean

exp_clean <- na_mean(exp_t)

# The data column 25 has only NA values and hence, this data column has been removed by using the below code

exp_new = subset(exp_clean, select = -c(EG.FEC.RNEW.ZS) )

# Checking again for any missing value

anyNA(exp_new)

summary(exp_new$SP.DYN.LE00.IN)

tapply(exp_new$SP.DYN.LE00.IN, exp_new$Continent, summary)

hist(exp_new$SP.DYN.LE00.IN)
```

```
md.pattern(exp_new)

dim(exp_new)

# Forward feature selection

model1<-lm(SP.DYN.LE00.IN~1,data=exp_new)

step1<-step(model1,scope=~EG.ELC.ACCS.ZS + NY.ADJ.NNTY.KD.ZG + NY.ADJ.NNTY.PC.KD.ZG + SH.HIV.INCD.14 +

          SE.PRM.UNER + SE.PRM.CUAT.ZS + SE.TER.CUAT.BA.ZS + SP.DYN.IMRT.IN + SE.PRM.CMPT.ZS +

          SE.ADT.LITR.ZS + FR.INR.RINR + SP.POP.GROW + EN.POP.DNST + SP.POP.TOTL +

          SH.XPD.CHEX.GD.ZS + SL.UEM.TOTL.NE.ZS + NY.GDP.MKTP.KD.ZG  +

          SP.DYN.CBRT.IN  + SH.HIV.INCD + SH.H2O.SMDW.ZS +

          SI.POV.LMIC + SE.COM.DURS,

         method='forward')

summary(step1)


# Backward feature selection

model2<-lm(SP.DYN.LE00.IN~.,data=exp_new)

step2<-step(model2,method="backward")

summary(step2)
```

```
# Correlation

install.packages("corrplot")

library(corrplot)

install.packages("faraway")

library("faraway")

vif(step1)


X<-exp_new[,-4]

exp_new.corr<-cor(cor(exp_new[, unlist(lapply(exp_new, is.numeric))]))

exp_new.corr

corrplot(exp_new.corr, lower.col = "black", number.cex = .7)


# Picking Smaller Model

model3<-lm(SP.DYN.LE00.IN~EG.ELC.ACCS.ZS +

           SE.PRM.CUAT.ZS +  SP.DYN.IMRT.IN +

           SP.POP.GROW + EN.POP.DNST +

           SH.XPD.CHEX.GD.ZS +

           SP.DYN.CBRT.IN  + SH.HIV.INCD + SH.H2O.SMDW.ZS ,data=exp_new)

summary(model3)


anova(step1, model3)
```

CONTRIBUITION OF MEMBERS

Task 1: Shashi Ranjan Choudhary (2200755)
Task 3: Rifat Monzur (2200367)