3/26/2021

Ademayowa Omole, Hamid Kabir, Dushyanth Bommana, Shagufta Khan, Midhun Molekudiyil, Sruthi Belaganti

University of Essex

MA317 Group 17 Coursework

2018 Life Expectancy Investigation

**Contents**

Abstract…………………………………………………………………………………………………………………………………………….3

Introduction………………………………………………………………………………………………………………………………………4

Question 1…………………………………………………………………………………………………………………………………………4

Question 2…………………………………………………………………………………………………………………………………………5

Question 3…………………………………………………………………………………………………………………………………………6

Question 4…………………………………………………………………………………………………………………………………………7

Question 5……………………………………………………………………………………………………………………………………….11

Appendix …………………………………………………………………………………………………………………………………………14

Word Count: 2842

**Abstract**

Life expectancy at birth refers to the average number of years an infant is predicted to live in the future if mortality rates at the time of birth remain the same. The impact of socio-economic growth on life expectancy is examined in this analysis based on various factors like mortality rate, total population, health expenditure, income etc. There are many factors which contribute to life expectancy and mortality but the most significant ones according to the world bank are expenditure on primary education, literacy rate, health expenditure, mortality rate and employment to population ratio. The problem statement requires to predict the life expectancy of individuals in a country using linear models. The multiple linear regressor built on the data in the analysis predicted the life expectancies with a root mean square error of 1.42.

**Statement of Contributions:**

Question 1: Ademayowa Omole

Question 2: Ademayowa Omole & Hamid Kabir

Question 3: Shagufta Khan

Question 4a: Sruthi Belaganti

Question 4b: Dushyanth Bommana

Question 4c: Dushyanth Bommana

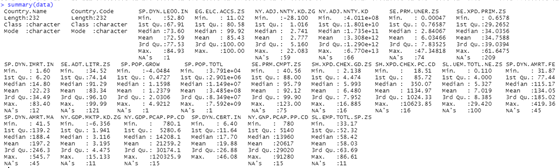
Question 5: Midhun Molekudiyil

Merging and Editing: Ademayowa Omole

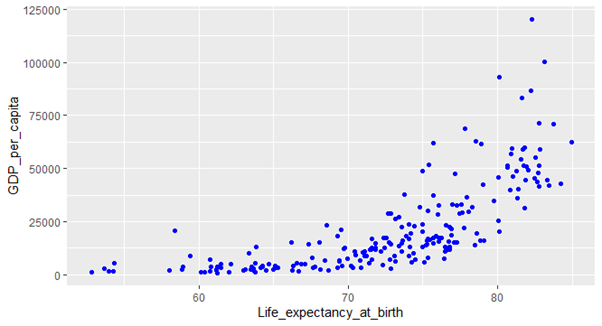
Introduction

All the countries in the world attempt to reduce the mortality rate of their people and improve the health conditions. Life expectancy has been improving globally in recent years, but the rate has differed by country. Life expectancy in developed countries increased steadily in the recent years, owing primarily to the decrease in infectious diseases due to the introduction of antibiotics.

1.

By using the str() function, we can see that the dataset we are working with contains 232 observations representing different countries and regions around the world. There are 21 different world development indicator variables taken from a primary World Bank database which can be seen by using colnames(). For a quick initial glance, we can use the head() and tail() functions which show us the first and last few rows of the data respectively, we can see that the data contains missing values. We then run the summary function to show summary statistics for each column which is shown in figure 1. For the life expectancy variable we want to predict, we can see the mean across all countries is 72.59. A scatter plot is useful to visualise trends in some of the indicator variables (figure 2). We compared life expectancy with GDP per capita as it is the most widely used predictor variable and found a positive relationship between the 2 variables.

Figure



Figure

2.

A complete case analysis gets rid of all data for a predictor variable that has one or more missing values. The extent to which a complete case analysis is an appropriate method to deal with missing values depends on the randomness of said missing values. For example, if data are missing completely at random (MCAR)  meaning that the propensity for a data point to be missing is completely random then a complete case analysis is unbiased.  Therefore a complete case analysis would be appropriate as the remaining complete cases represent a smaller but random sample from the population. However if data is missing not at random(MNAR) then a complete case analysis can produce bias in a model.

Dealing with Missing data is very common in research studies. This could be due to many reasons such as data entry errors or data collection problems. But ignoring these cases can lead to invalid and misleading implications being drawn. We worked to provide guidance on how to deal with missing values and the best ways of analysing a dataset that is incomplete. As the data has 893 null values which needs to be considered.

Strategy 1: Deletion of the missing values

We can do this by two types, List wise and pairwise. First we will use a function to delete the missing values from our data set. This is a bad strategy in order to deal with missing data because we delete any row of the data that has a missing value in any column therefore it is difficult to know what has been taken out. As for the deletion of data, the advantage of this type is simplicity, but it reduces the power of the model because it reduces the sample size so we are not going with method.

Strategy 2: Filling the missing values in selected columns with previous or next observation.

Filling the missing values in selected columns with lower or next entry**.** The  null values are inconsistently missing by position in the columns,so using the direction is equal to up down to fill the missing values in that direction.

Strategy 3: Fill missing values with zero.

Strategy 4: Mean Imputation

In the whole data set we have two categorical variables "Country Name" and "Country Code'' and have no null values. So therefore these two columns must be removed  in order to do the mean imputation  then later merge it back in the data set. Taking the mean of the specific column and replacing it with null values. So eventually merging the two data sets to the final data set for modelling.

3.

# Data Analysis and Pre-processing:

We are provided with life expectancy dataset that contains 232 observations (row) and 23 features (columns). There are 2 categorical columns i.e. Country Name and Country Code and 21 numeric columns. I have converted categorical columns to numeric columns.

The number of null values in each column can be shown as

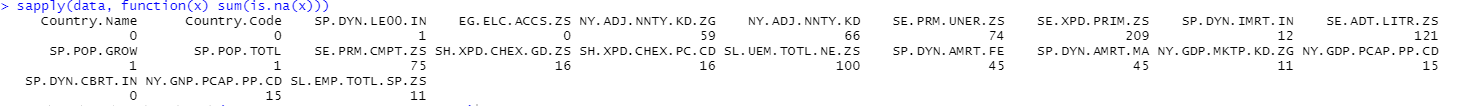


Figure 3

Since, there were large number of null values in some columns, I decided to remove them as they would provide very little or no information at the time of training. So, I retained only those columns that have 85% of the non-null values and I was left with 14 columns in the dataset.

Then I replaced the null values in all columns using mean imputation.

# Collinearity:

Collinearity occurs when variables in the dataset are highly corelated (if not all, at least some variables are correlated with other variables). Simply stated as - when variables are linear combination of other variables in dataset. Model developed using such dataset may not provide as accurate results because we are missing out on relevant features/information.

According to Wikipedia, “Collinearity is a linear association between two explanatory variables. Two variables are perfectly collinear if there is an exact linear relationship between them. For example, X1 and X2 are perfectly collinear if there exist parameters λ0 and λ1 such that, for all observations i, we have

X2i = λ0 + λ1 \* X1i

Collinearity is bad at explaining the relationship between the response and the predictors. I computed the correlation between variables in R and plotted a scatter plot of all pairs of variables in our dataset using pairs method from faraway package.

The correlation plot of variables is as follows:

Chart, bubble chart

Description automatically generated

Figure 4

In order to find collinearity, I trained a linear model and split the dataset into training and testing samples with 80/20 ratio. With all the predictor variables, we have found out an RMSE of 2.46 and R-squared of 0.859 on test dataset. Then, I calculated the Variance Inflation Factor for our linear model. A variance inflation factor(VIF) detects multicollinearity in regression analysis. Multicollinearity is when there’s correlation between predictors (i.e. independent variables) in a model; it’s presence can adversely affect our regression results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

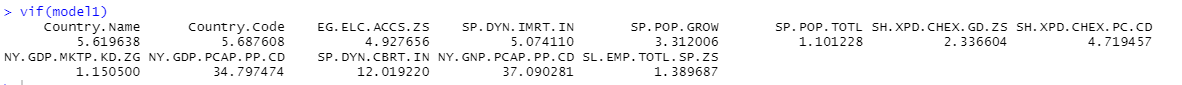


Figure 5

We could see that Country.Name, Country.Code, NY.GDP.PCAP.PP.CD, SP.DYN.CBRT.IN, NY.GNP.PCAP.PP.CD and SL.EMP.TOTL.PP.ZS has value of VIF greater than 5. Correlation increase the variation of the estimates and make the model much harder to interpret. We removed the predictor variables with VIF greater than 5.

4.

Regression is a technique which is used to predict a target variable using a single or a set of independent variables. Thus, the relation between target and predictor variables can be analysed using regression. Simple linear regression is used when there is a linear relation between the dependent variable and a single independent variable. In the given data set, there are more than one independent variables. So, Multiple linear regression strategies, also called multivariate regression models can be used to predict the life expectancy of the people in various countries depending on the indicator variables of the world bank.

The linear regression line for n explanatory variables x1, x2, ..., xn is defined to be :

Y= β0+ β1X1+ β2X2+…. + βnXn.

This line describes how the target Y changes with the explanatory variables. There are some assumptions of a multiple linear regression like:

·       The line of best fit passing through the data points is a straight line.

·       The data follows a normal distribution.

The given dataset requires to analyse and predict the life expectancy of the individuals in different countries based on several factors like total population, mortality rate, literacy rate, employment, health expenditure etc. Therefore, a multiple linear regressor could be built by training the given data and predictions could be made on the life expectancy of the people in the countries which are not listed in the given dataset. Based on this analysis, it would be more convenient for a country to identify the factors which are resulting in a lower life expectancy and effectively improve the circumstances.

The model to assess the life expectancy built, followed the below procedure.

·       The dataset is loaded into R using read.csv () function. The libraries like dplyr, mice which are useful in building a regression model are loaded into R.

·       The data set is checked for missing values. Missing values cause certain problems on the decisions taken on analysing the data. However, excluding missing data from the evaluation causes loss of data and results in inaccurate conclusions. The ideal way to deal with missing values is to impute them with certain values which complete the dataset without disturbing the original structure.

·       In this analysis, MICE (Multivariate Imputation by Chained Equations) imputation technique is used to replace the missing data. This feature automatically detects the columns with missing values and imputes them accordingly using methods such as predictive mean matching, logistic regression, depending on type of the variable.

·       When the data is free of all the missing values, correlation among the variables is calculated using corrplot (). Correlation coefficient which ranges between 1 and -1 gives the relation between two variables. If the value is close to +1 or -1, it means that two variables are highly correlated and thus one variable can be removed from the analysis.

·       In the given dataset, the variables whose correlation coefficient is +/-0.8 are considered as highly correlated.

·       The correlation coefficient between male and female mortality rates is 0.94. Hence, male mortality rate is not considered in building a model.

·       Likewise infant mortality rate, GNP per capita, birth rate crude are highly correlated with female mortality rate , GDP per capita and access to electricity respectively.

·       Therefore, the former variables are ignored from the analysis.

·       Multiple linear regressor model is built on the remaining variables after eliminating the correlated variables using lm () function.

·       Summary of the model provides information about parameters like F-statistic, p- value and adjusted r squared which help in analysing the model. The adjusted R squared value is used to compare the goodness of the fit of the model. For the model built in our analysis, the adjusted R squared value is 0.92.

·       We also get the significant variables in the summary of the model. The model can be further simplified by removing the least contributing variables one at a time and checking the adjusted r squared each time.

·       In the analysis, adjusted net national income, population growth, unemployment, GDP growth are identified as least significant variables and can be eliminated from the model.

·       The adjusted R squared after removing the insignificant variables is 0.92. It can be observed that there is no change after removing the less contributing variables and the model is further simplified.

·       We can now apply this model to analyse and predict the life expectancies.

We can evaluate the model built on the given data set as follows:

·       The given dataset is divided into training and test sets using split() function of the catools package in R.

·       The usual split ratio is 80% of the data will be training set and the remaining 20% will be the test set.

·       The model is trained using the regression model on training data and in the summary of the model, the adjusted r squared value is 0.90.

·       The predictions are made using the regressor test set and the predictions are evaluated using RMSE(Root Mean Squared error) from the Metrics package in R.

·       The RMSE is considered a good metric for numerical predictions. The value obtained in the analysis is 1.42. Small values of RMSE denote good fit.

The same model can be used to predict the life expectancy of the citizens of any country. In the life expectancy data 2 file , there are 11 countries for which the life expectancy is to be predicted. This dataset also contains missing values which are imputed using mice package and by replacing certain missing observations with mean. The dataset is further simplified by eliminating the least significant variables.

The predictions are made using the regressor and the results are copied to a new excel file Datapredict\_2.xlsx.

**5.**

One-way ANOVA

The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of two or more independent (unrelated) groups (although you tend to only see it used when there are a minimum of three, rather than two groups). For example, you could use a one-way ANOVA to understand whether exam performance differed based on test anxiety levels amongst students, dividing students into three independent groups (e.g., low, medium, and high-stressed students). Also, it is important to realize that the one-way ANOVA is an omnibustest statistic and cannot tell you which specific groups were statistically significantly different from each other; it only tells you that at least two groups were different. Since you may have three, four, five or more groups in your study design, determining which of these groups differ from each other is important.

 Assumptions

When you choose to analyse your data using a one-way ANOVA, part of the process involves checking to make sure that the data you want to analyse can actually be analysed using a one-way ANOVA. You need to do this because it is only appropriate to use a one-way ANOVA if your data "passes" six assumptions that are required for a one-way ANOVA to give you a valid result.

 Assumption #1: Your dependent variable should be measured at the interval or **ratio** level (i.e., they are continuous).

Assumption #2: Your independent variable should consist of two or morecategorical**,** independent groups**.** Typically, a one-way ANOVA is used when you have three or morecategorical, independent groups, but it can be used for just two groups (but an independent-samples t-test is more commonly used for two groups).

Assumption #3: You should have independence of observations**,** which means that there is no relationship between the observations in each group or between the groups themselves.

Assumption #4: There should be no significant outliers**.** Outliers are simply single data points within your data that do not follow the usual pattern.

Assumption #5: Your dependent variable should be approximately normally distributed for each category of the independent variable**.**

Assumption #6: There needs to be homogeneity of variances. You can test this assumption in SPSS Statistics using Levene's test for homogeneity of variances.

 Benefits

One-way ANOVA is used when the researcher is comparing multiple groups (more than two) because it can control the overall Type I error rate.

Advantages:

·   It provides the overall test of equality of group means

·   It can control the overall type I error rate (i.e. false positive finding)

·   It is a parametric test so it is more powerful if normality assumptions hold true.

Programming side explanation:

We are using “countrycode” package to map countries to the continents. The countrycode function can convert to and from several different country coding schemes. It uses regular expressions to convert country names (e.g. Sri Lanka) into any of those coding schemes, or into standardized country names in several languages. It can create variables with the name of the continent and/or several regional groupings to which each country belongs.  “sourcevar” is the vector which contains the codes or country names to be converted.

While summarising the data with summarise() function after mapping country on the basics of continents with the mean values of Life expectancy at birth, total (years), we will get the result as :

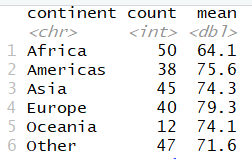


Figure 6

Finally we will conduct the anova test with formula “ Life expectancy at birth, total (years) ~ continent “ and while taking the summary of it we will get the result as:

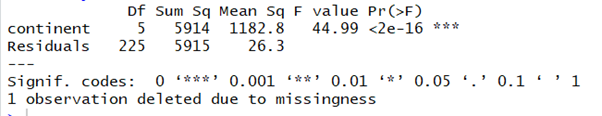


Figure 7

As the p-value is less than the significance level 0.05, We can conclude that there are significant differences between the continents highlighted with "\*" in the model summary.

**Appendix**

head(data)

summary(data)

tail(data)

str(data)

colnames(data)

describe(data)

#ggplot(data = data) +

  geom\_point(mapping = aes(x = Life\_expectancy\_at\_birth, y = GDP\_per\_capita), color="blue")

library(tidyverse)

library(zoo)

library(gridExtra)

library(knitr)

 # Load the csv file

df <- read.csv("https://raw.githubusercontent.com/sruthireddy1482/MA317\_GRP17/main/LifeExpectancyData1.csv", header = T)

head(df)

dim(df)# check the dimension of data frame.

str(df)

summary(df)

head(df)# structure of the data

# The sum and the percentage of missing values

sum(is.na(df))

mean(is.na(df))

 # Looking for missing values in the whole Data set

missing\_val <- sapply(df, function(x) sum(is.na(x)))

print(missing\_val)

# Strategy 1

.# Drop all null values from our data set

drop <- na.omit(df)

drop

# Strategy 2: Filling the missing values in selected columns with up or previous entry

df <- fill(df, SP.DYN.LE00.IN:SL.EMP.TOTL.SP.ZS, .direction = "up")

head(df)

summary(df)

df <- fill(df, SP.DYN.LE00.IN:SL.EMP.TOTL.SP.ZS, .direction = "down")

head(df)

summary(df)

 # Strategy 3:

# To fill missing values with zero

df[is.na(df)] <- 0

summary(df)

subset1 <- subset(df, select = c(3:23))

head(subset1)

dim(subset1)

# Taking the mean of the specific column and it replace it with null values

subset1 <- na.aggregate(subset1)

head(subset1)

summary(subset1)

dim(subset1)

 subset2 <- subset(df, select = c("Country.Name","Country.Code"))

head(subset2)

dim(subset2)

 # So eventually merging the two data set to final data set for modeling

final\_dataset <- cbind(subset2, subset1)

head(final\_dataset)

dim(final\_dataset)

##### Reading dataset with column names

data <- read.csv("https://raw.githubusercontent.com/sruthireddy1482/MA317\_GRP17/main/LifeExpectancyData1.csv", header = TRUE)

###### Loading tidyr library for data manipulation

library(tidyr)

###### Checking the dimensions of our dataset

nrow(data)

ncol(data)

###### Checking the column types i.e. categorical/numeric

str(data)

###### Checking for null values in all columns

sapply(data, function(x) sum(is.na(x)))

###### Summary of dataset

summary(data)

###### Converting categorical columns to numeric

countrycode <- transform(data$Country.Code,id=as.numeric(factor(data$Country.Code)))

data$Country.Code <- countrycode$id

countryname <- transform(data$Country.Name,id=as.numeric(factor(data$Country.Name)))

data$Country.Name <- countryname$id

###### Removing columns that contain more than 15% null values in dataset.

data <- data[, which(colMeans(!is.na(data)) > 0.85)]

###### Checking the dimensions of dataset

nrow(data)

ncol(data)

###### Mean imputation for null values in columns

data$SP.DYN.IMRT.IN[is.na(data$SP.DYN.IMRT.IN)] <- mean(data$SP.DYN.IMRT.IN, na.rm = TRUE)

data$SP.POP.GROW[is.na(data$SP.POP.GROW)] <- mean(data$SP.POP.GROW, na.rm = TRUE)

data$SP.POP.TOTL[is.na(data$SP.POP.TOTL)] <- mean(data$SP.POP.TOTL, na.rm = TRUE)

data$SH.XPD.CHEX.GD.ZS[is.na(data$SH.XPD.CHEX.GD.ZS)] <- mean(data$SH.XPD.CHEX.GD.ZS, na.rm = TRUE)

data$SH.XPD.CHEX.PC.CD[is.na(data$SH.XPD.CHEX.PC.CD)] <- mean(data$SH.XPD.CHEX.PC.CD, na.rm = TRUE)

data$NY.GDP.MKTP.KD.ZG[is.na(data$NY.GDP.MKTP.KD.ZG)] <- mean(data$NY.GDP.MKTP.KD.ZG, na.rm = TRUE)

data$NY.GDP.PCAP.PP.CD[is.na(data$NY.GDP.PCAP.PP.CD)] <- mean(data$NY.GDP.PCAP.PP.CD, na.rm = TRUE)

data$NY.GNP.PCAP.PP.CD[is.na(data$NY.GNP.PCAP.PP.CD)] <- mean(data$NY.GNP.PCAP.PP.CD, na.rm = TRUE)

data$SL.EMP.TOTL.SP.ZS[is.na(data$SL.EMP.TOTL.SP.ZS)] <- mean(data$SL.EMP.TOTL.SP.ZS, na.rm = TRUE)

##### Removing rows with null values in target variable i.e. SP.DYN.LE00.IN

data<-data[-which(is.na(data$SP.DYN.LE00.IN)),]

##### correlation in dataset

cc = cor(data)

corrplot(cc)

round(cc, 2)

##### Loading required packages for model training and data evaluation

install.packages("caret")

install.packages("faraway")

library(caret)

##### Splitting dataset into training and testing sets with 80% training dataset and 20% test dataset.

set.seed(123)

training.samples <- data$SP.DYN.LE00.IN %>% createDataPartition(p = 0.8, list = FALSE)

train.data <- data[training.samples, ]

test.data <- data[-training.samples, ]

###### Fitting a linear model on our dataset with all predictor variables

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

###### Predictions on test dataset and evaluation

predictions <- model1 %>% predict(test.data)

RMSE(predictions, test.data$SP.DYN.LE00.IN)

R2(predictions, test.data$SP.DYN.LE00.IN)

###### Checking correlation between variables using scatter plot

library(faraway)

pairs(data, col = "dodgerblue")

###### Computing VIF values for our predictor variables in our linear model using car package

library(car)

vif(model1)

summary(model1)

###### Dropping columns from the dataset with VIF values greater than 5

drops <- c("Country.Name","Country.Code","NY.GDP.PCAP.PP.CD","SP.DYN.CBRT.IN","NY.GNP.PCAP.PP.CD")

train.data<-train.data[ , !(names(train.data) %in% drops)]

###### Fit model again with collinearity removed from the dataset

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

predictions <- model2 %>% predict(test.data)

RMSE(predictions, test.data$SP.DYN.LE00.IN)

R2(predictions, test.data$SP.DYN.LE00.IN)

###### Checking the effect on model

summary(model2)

library(dplyr)

library(ggplot2)

library(mice)

library(purrr)

library(corrplot)

library(psych)

library(caTools)

library(readxl)

library(ModelMetrics)

library(openxlsx)

library(writexl)

set.seed(123)

data\_4=read.csv("https://raw.githubusercontent.com/sruthireddy1482/MA317\_GRP17/main/LifeExpectancyData1.csv")

View(data\_4)

map(data\_4, ~sum(is.na(.)))

#imputing all the missing values using mice

impute= mice(data\_4[,5:23],method = "cart")

print(impute)

#impute$imp$NY.ADJ.NNTY.KD.ZG

new\_data\_4=complete(impute,3)

new\_data\_4

#replacing missing columns with imputed data

data\_4[,5:23]=new\_data\_4[]

#replacing 1 missing value of life expectency using mean of the column

data\_4$SP.DYN.LE00.IN[is.na(data\_4$SP.DYN.LE00.IN)]<-mean(data\_4$SP.DYN.LE00.IN,na.rm=TRUE)

sum(is.na(data\_4))

M<-cor(data\_4[,4:23])

corrplot(M,method = "color",number.cex=0.5,type="upper")

#checking correlation b/w mortality rate male and female

corr.test(data\_4$SP.DYN.AMRT.FE,data\_4$SP.DYN.AMRT.MA)

#the value is 0.93 - highly correlated so removing male variable from the data

#checking correlation b/w mortality rate female and infant

corr.test(data\_4$SP.DYN.AMRT.FE,data\_4$SP.DYN.IMRT.IN)

#the value is 0.87 - highly correlated so removing infant variable from the data

#checking correlation b/w gdp per capita and gnp per capita

corr.test(data\_4$NY.GDP.PCAP.PP.CD,data\_4$NY.GNP.PCAP.PP.CD)

#the value is 0.99 - highly correlated so removing gnp per capita from the data

#checking correlation b/w access to electricity and birth rate crude

corr.test(data\_4$EG.ELC.ACCS.ZS,data\_4$SP.DYN.CBRT.IN)

#the value is -0.83 - highly correlated so removing birth rate crude from the data

final\_data\_4=data\_4[,-c(18,9,22,21)]

View(final\_data\_4)

#building multiple linear regressor model

model=lm(SP.DYN.LE00.IN~EG.ELC.ACCS.ZS+SE.PRM.UNER.ZS+SE.XPD.PRIM.ZS+SE.ADT.LITR.ZS+           SP.POP.GROW+SE.PRM.CMPT.ZS+SH.XPD.CHEX.GD.ZS+SH.XPD.CHEX.PC.CD+           SP.DYN.AMRT.FE+NY.GDP.PCAP.PP.CD+NY.ADJ.NNTY.KD+SL.EMP.TOTL.SP.ZS+          SP.POP.TOTL+SL.UEM.TOTL.NE.ZS+NY.ADJ.NNTY.KD.ZG+NY.GDP.MKTP.KD.ZG,

         data=final\_data\_4)

summary(model)

anova(model)

#removing least significant variable one each time and check the adjusted r square value

model1=lm(SP.DYN.LE00.IN~EG.ELC.ACCS.ZS+SE.PRM.UNER.ZS+SE.XPD.PRIM.ZS+SE.ADT.LITR.ZS+

           SE.PRM.CMPT.ZS+SH.XPD.CHEX.GD.ZS+SH.XPD.CHEX.PC.CD+

           SP.DYN.AMRT.FE+SL.EMP.TOTL.SP.ZS+

           NY.ADJ.NNTY.KD.ZG,

         data=final\_data\_4)

summary(model1)

anova(model1)

#summary(model1)$coefficient

#confint(model1)

# Hence we have removed all the less signifiant variables from the model

#4b

split\_4= sample.split(final\_data\_4$SP.DYN.LE00.IN, SplitRatio = 0.8)

training\_set\_4=subset(final\_data\_4,split\_4==TRUE)

test\_set\_4=subset(final\_data\_4,split\_4==FALSE)

#fitting multiple linear regressor to the training set

regressor\_4=lm(SP.DYN.LE00.IN~EG.ELC.ACCS.ZS+SE.PRM.UNER.ZS+SE.XPD.PRIM.ZS+SE.ADT.LITR.ZS+

               SE.PRM.CMPT.ZS+SH.XPD.CHEX.GD.ZS+SH.XPD.CHEX.PC.CD+

               SP.DYN.AMRT.FE+SL.EMP.TOTL.SP.ZS+

               NY.ADJ.NNTY.KD.ZG,

             data=training\_set\_4)

summary(regressor\_4)

#predicting on test set

y\_pred\_4= predict(regressor\_4,newdata = test\_set\_4)

y\_pred\_4

#evaluating the model

rmse(test\_set\_4$SP.DYN.LE00.IN,y\_pred\_4)

#4c

set.seed(123)

unseen\_data\_4=read.xlsx("https://github.com/sruthireddy1482/MA317\_GRP17/raw/main/LifeExpectancyData2.xlsx")

unseen\_data\_4[, 4:22] <- sapply(unseen\_data\_4[,4:22], as.numeric)

#imputing missing values using mice

map(unseen\_data\_4, ~sum(is.na(.)))

impute1\_4= mice(unseen\_data\_4[,4:22],method = "cart")

print(impute1\_4)

new\_data1\_4=complete(impute1\_4,3)

new\_data1\_4

unseen\_data\_4[,4:22]=new\_data1\_4[]

unseen\_data\_4$SE.XPD.PRIM.ZS[is.na(unseen\_data\_4$SE.XPD.PRIM.ZS)]<-mean(unseen\_data\_4$SE.XPD.PRIM.ZS,na.rm=TRUE)

unseen\_data\_4$NY.GNP.PCAP.PP.CD[is.na(unseen\_data\_4$NY.GNP.PCAP.PP.CD)]<-mean(unseen\_data\_4$NY.GNP.PCAP.PP.CD,na.rm=TRUE)

unseen\_data\_4$SE.ADT.LITR.ZS[is.na(unseen\_data\_4$SE.ADT.LITR.ZS)]<-mean(unseen\_data\_4$SE.ADT.LITR.ZS,na.rm=TRUE)

unseen\_data\_4$SE.PRM.CMPT.ZS[is.na(unseen\_data\_4$SE.PRM.CMPT.ZS)]<-mean(unseen\_data\_4$SE.PRM.CMPT.ZS,na.rm=TRUE)

map(unseen\_data\_4, ~sum(is.na(.)))

sum(is.na(unseen\_data\_4))

View(unseen\_data\_4)

result\_4=unseen\_data\_4%>%

  select(c("Country.Name","Country.Code","EG.ELC.ACCS.ZS","NY.ADJ.NNTY.KD.ZG",

          "NY.ADJ.NNTY.KD","SE.PRM.UNER.ZS","SE.XPD.PRIM.ZS","SE.ADT.LITR.ZS",

           "SP.POP.GROW","SP.POP.TOTL","SE.PRM.CMPT.ZS","SH.XPD.CHEX.GD.ZS",

           "SH.XPD.CHEX.PC.CD","SL.UEM.TOTL.NE.ZS","SP.DYN.AMRT.FE",

           "NY.GDP.MKTP.KD.ZG","NY.GDP.PCAP.PP.CD","SL.EMP.TOTL.SP.ZS"

           ))

#removing all insignificant variables in unseen data

life\_expectancy\_4= predict(regressor\_4,newdata = result\_4)

life\_expectancy\_4

final\_excel\_4=cbind(result\_4,life\_expectancy\_4)

#write\_xlsx(final\_excel\_4,"D:\\MA317\\Datapredict\_4.xlsx")

data <- read.csv("https://raw.githubusercontent.com/sruthireddy1482/MA317\_GRP17/main/LifeExpectancyData1.csv", header=TRUE)

data$Country.Name

install.packages("countrycode")

library(countrycode)

data$continent <- countrycode(sourcevar = data[,"Country.Name"],

origin = "country.name" ,

destination ="continent")

data$continent

data$continent <- as.character(data$continent)

data$continent[is.na(data$continent)] <- "Other"

data$continent

library(dplyr)

group\_by(data, continent) %>%

summarise( count =n(),

mean =mean(SP.DYN.LE00.IN, na.rm = TRUE),

)

res.aov <- aov(SP.DYN.LE00.IN ~ continent, data = data)

summary(res.aov)