3/28/2022

University of Essex

MA317 Group Coursework

2021-22 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

**Contribution list:**

Question 1: Shashank Reddy

Question 2: Prasanth Arasavilli

Question 3: Siddharth Bharadwaj

Question 4: Deepika Venigalla

Question 5: Prudhvi Krishna Mullapudi

Merging and editing: Prasanth Arasavilli, Deepika Venigalla

**Abstract:**

Life expectancy at birth is the average life expectancy of an infant in the future if the mortality rate at birth remains constant. The impact of socioeconomic growth on life expectancy was examined in this analysis based on various factors such as mortality, total population, healthcare expenditure, income, etc. There are many factors that affect life expectancy and mortality, but the most important is World Bank spending including primary education, literacy, health spending, mortality, and employment-to-population ratios. The problem statement asks to use a linear model to predict the life expectancy of individuals in a country given various factors that are affecting the mortality rate in the country. We were given a dataset with missing data and asked to build a model that helps in assuming the life expectancy of the particular country. We have used multiple imputation method to fill the missing data. And by calculating collinearity and AIC values we have removed parameters and calculated the AIC score for full model, reduced model and best model and compared the performance

**Introduction**:

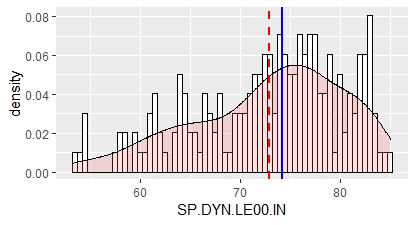
Life expectancy is a key factor to assume for any country. The assumption is helpful to compare what are the factors a particular country is lacking w.r.t to others to maintain a good Life expectancy rate. Assumptions are made that the other factors are directly affecting the Life expectancy we have created a linear model which studies the relation of each parameter with the Life expectancy. As we observe in the well-developed countries the Life expectancy rate has increased with time and that is due to the increase in few parameters life health expenditure, GDP, hospitality.

**Question** 1:

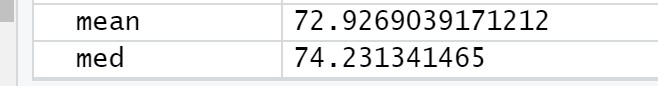
We can see that the dataset we are working with contains 218 observations representing different countries and regions around the world. Installing all required libraries to perform descriptive statistics. There are 26 different world development indicator variables taken from a primary World Bank database. 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 then run the summary function to show summary statistics for each column. Removing the columns having more than 75% missing data as they have more than 75% missing values and they won't contribute to the model. Plotting density function and histogram of the response variable. Histogram with mean, median and density curve.

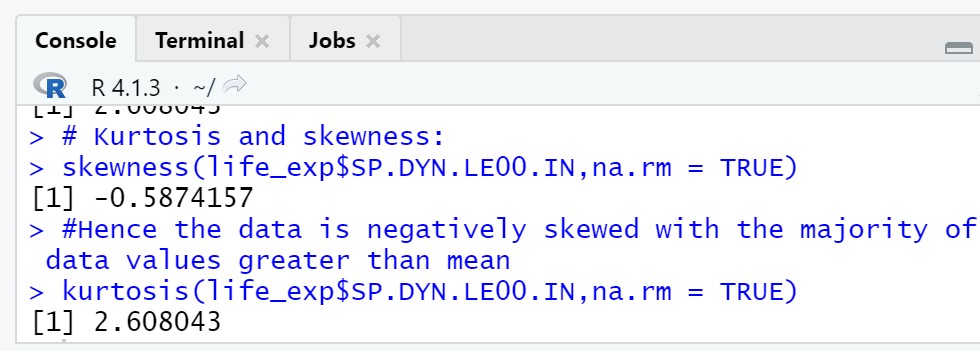
We can see that the mean value is the red dotted line in the graph and blue line is the median of the target variable.



The Mean and the Median values of life\_expectancy from the above graph is mentioned below.



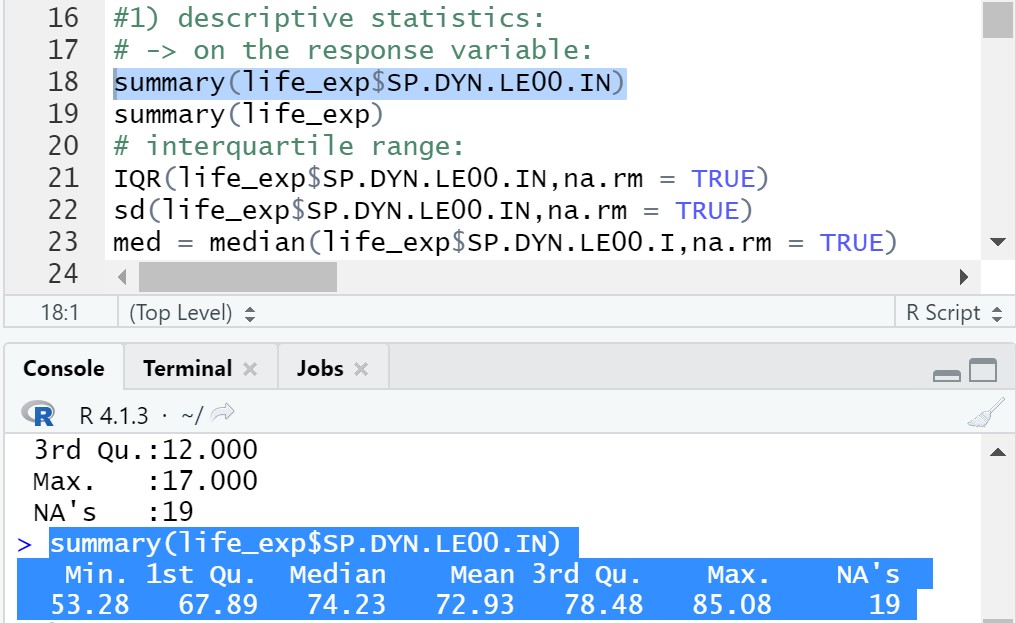
The characteristic of a frequency distribution that ascertains its symmetry about the mean is called skewness.



kurtosis is defined as the parameter of relative sharpness of the peak of the probability distribution curve. It ascertains the way observations are clustered around the centre of the distribution. It is used to indicate the flatness or peakedness of the frequency distribution curve and measures the tails or outliers of the distribution.

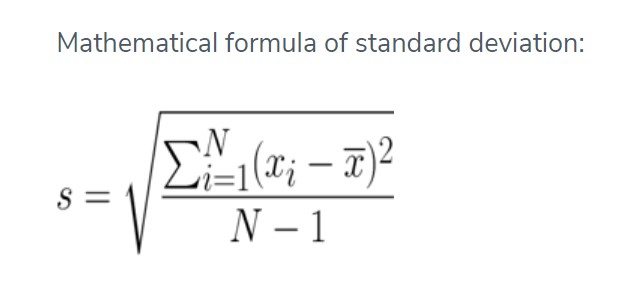
The data is negatively skewed with the majority of data values greater than mean values. As the coefficient of kurtosis is less than 3 the distribution is platykurtic.

Summary of the target column(life\_expectancy):



The interquartile range of a data set is the**difference between the values that fall at the 25% and 75% points when the data points are placed in numerical order. The IQR of life\_exp (SP.DYN.LE00.IN) is 10.58627.**

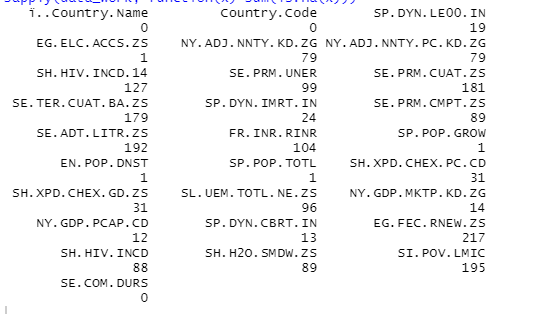
**Standard deviation: The standard deviation of life\_exp (SP.DYN.LE00.IN) is 7.470629.**



**Question 2:**

**Handling missing values using Imputation methods**:

As we analysed the data we have observed few columns having majority of nulls. By using the ‘sapply’ function in the R function we have listed down the count of nulls in each column.



We have total rows of 217 rows in the input data. AS we see the column EG.FEC.RNEW.ZS has all nulls hence we have the removed the column. SI.POV.LMIC, SE.ADT.LITR.ZS has more than 90% of nulls and can be removed directly.

It is difficult to analyse the model with nulls hence we have to impute the data and fill the nulls. We have different imputation methods like mean imputation, max-min imputation and replace the values but this has a disadvantage of having the same values in multiple rows which can mislead the final prediction.

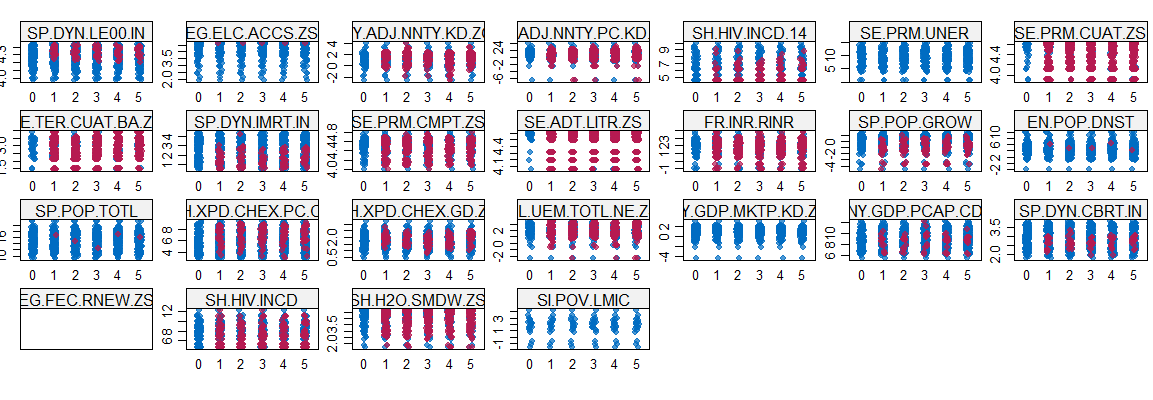
In our data we have different datatypes for columns like strings, numerical and categorial columns. To calculate the missing values, we need only numerical values hence we are dropping “Country Name”, “Country column” and “SE.COM.DURS” and “Continent” columns from the data before performing imputation. Also, for few columns we have very high values lite gdp total and most of the columns have singular digits. Hence to bring them down to a smaller and similar scale we apply log to all the columns

**Multiple Imputation**:

Above we have examined the strategy of eliminating cases with missing values and analysing just full cases, which is a simple approach to implement but has the drawback of discarding a lot of good data. The estimators' bias is reduced, and the dataset's variability is reduced. The multiple imputation approach attempts to address the shortcomings of the first two methods.

This method estimates probable values based on the distribution of seen data, i.e., it tries to forecast what the observed values will be. As a result, more information than just the mean or a single predicted value is employed, and uncertainty is included in, resulting in estimators that are roughly unbiased. It is possible to calculate standard errors and so undertake statistical inference when the variability of the data is taken into consideration.

We have a package called mice in R which is helpful in predicting missing variables. The mice package performs multiple imputation by assuming a conditional distribution of a specific predictor variable given the other predictor variables. Then it draws a possible value from that distribution to replace-impute the missing values of that predictor variable.

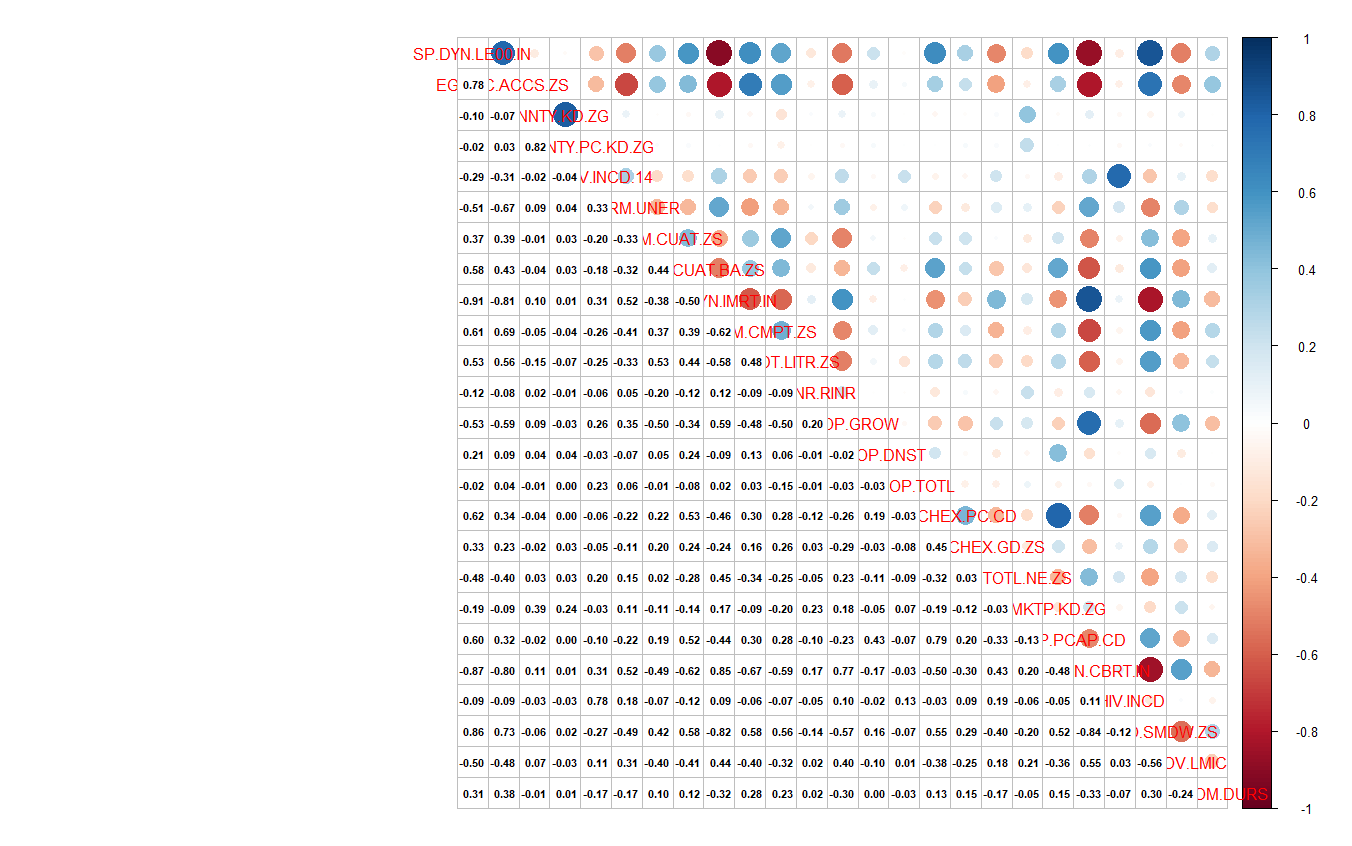


Above are the plots of the imputed data for each column using the mice package. As we can observe the data is randomly distributed and not having the repeated values as in simple imputation. Hence we are using this imputed data to fit the model and then choose the best model

**Question 3:**

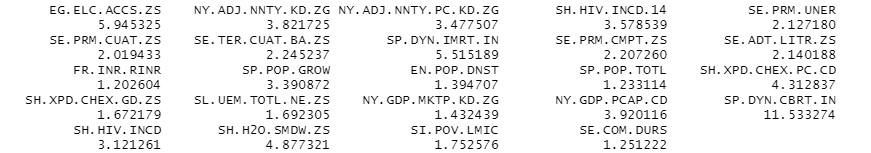
**Multi Collinearity**

As the term collinearity itself suggests, it deals with the correlation between the independent features. It increases the variance of the estimators and hence reduces the adequacy of the model. When we deal with the datasets which contain a significant number of independent variables the problem of multi-collinearity among them may arise. It becomes really important to deal with it. This can be measured with the help of the variance inflation factor (VIF) which will let us know the degree of correlation among the independent features. We can also see the correlation among features with the help of the correlation heatmap.



**Figure No. 3.1:** Correlation Heatmap

VIF has a range of 1 to infinity, if the VIF value ranges between 1 to 5 it means the feature has a low correlation. VIF ranging from 5 to 10 has a strong correlation and VIF values above 10 has are highly correlated. We have selected a VIF score of 10 as the threshold, therefore, we need to remove those features having a VIF score of more than 10. The below table shows all features’ VIF scores and hence we will remove the feature above 10 scores i.e. the birth rate (SP.DYN.CBRT.IN).



Feature having VIF score below 10 will be further be used in model building in the 4th part of the assignment.

**Question 4:**

Simple linear regression is used when there is a linear relation between the dependent variable and a one independent variable. In the data set given, there are multiple 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 response Y changes with the predictor variables X1, X2…Xn. 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 variation of life expectancy of the people in the countries based on the predictor variables the given dataset can be observed. 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.

When we built a model using all the columns in the given data set the efficiency of the model might be decreased because not all columns are important to build a linear model for optimal results.

**Using Cp**

So, to consider optimal columns we have several ways to do it one among that is calculating the min\_Cp and corresponding columns related to Cp. In our case we considered this method and the min Cp value is ?

When we calculate AIC for reduced and full models. The value should be less for the best model to fit in this case the value of AIC for full model is -861.7854 and value of reduced model is -834.7846.

The columns selected for reduced model are EG.ELC.ACCS.ZS, NY.ADJ.NNTY.KD.ZG, NY.ADJ.NNTY.PC.KD.ZG, SH.HIV.INCD.14, SP.DYN.IMRT.IN,SE.PRM.CMPT.ZS, 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,NY.GDP.PCAP.CD, SP.DYN.CBRT.IN,SH.HIV.INCD,SH.H2O.SMDW.ZS

Also when calculate ols\_mallows\_cp between full and reduced model it should be less than the number of parameters selected. In our case ols\_mallows\_cp is 47.13.

From the above two criterias we can say that full model is best fit for this dataset.

**Stepwise Selection**

We can select the best model parameters using this technique. Below is the algorithm for this.

This method will select the parameters with optimal model. By eliminating least significant values from the group one at time we will provide the most optimal parameters.

Below are columns considered after the feature selection.

NY.ADJ.NNTY.KD.ZG, NY.ADJ.NNTY.PC.KD.ZG, SH.HIV.INCD.14, SE.PRM.CUAT.ZS, SE.TER.CUAT.BA.ZS, SP.DYN.IMRT.IN, FR.INR.RINR, SP.POP.GROW, EN.POP.DNST, SP.POP.TOTL, SP.DYN.CBRT.IN, SH.H2O.SMDW.ZS

AIC for this model is -847.9295 which is equal to full model. But, if we consider ols\_mallows\_cp value is approximately 30 which is higher than the parameters selected. So, we consider that this reduced model is also not a best fit.

Based on both methods we consider full model i.e considering all the columns in imputed data gives the best fit.

**Question 5:**

**Analysis of Differences in average Life Expectancies across the Continents**

The “Life Expectancy data1” dataset that is imputed using the mice library is investigated to understand if there is any difference in the average life expectancies across the continents. There is one explanatory categorical variable which is Continent and one continuous response variable which is Average Life Expectancy. This means that the factor level is one and hence One – way Analysis of Variance (ANOVA) is chosen for analysis.

Model used: average\_life\_expectancy ~ as.factor(continent)

Here, the following hypothesis was tested:

Ho (Null Hypothesis): All the Average responses(life expectancies) are equal (vs)

H1 (Alternative Hypothesis): Average responses may differ

Analysing the group means of all the continents, it is understood that life expectancy is least for Africa and highest for Europe. For all the other continents, it appears to be similar with a difference of 1-2 years between all possible pairs.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Group Means of Life Expectancy (years)** | | | | | |
| Africa | Asia | Australia/Oceania | Europe | North America | South America |
| 64.11014 | 74.61739 | 73.44988 | 79.39339 | 76.24878 | 75.091 |

The boxplot of Life Expectancy vs Continent shows that there are 2 and 1 outliers for Africa and North America continents. Also, the distribution appears to be symmetric for Africa, Asia symmetric for others.

Classic ANOVA method assumes that the residuals of the model are normally fitted and variances of all the continents are equal. The normality assumption was tested using the **Shapiro test** and was found that the assumption is valid at 10% confidence.

and Australia and non-Chart, box and whisker chart

Description automatically generated

But, when the equal variance of continents was tested using **Leneve test**, the p-value of 0.02 obtained rejects the hypothesis.

So, instead of classic ANOVA, **Welch test** was used.

|  |  |  |  |
| --- | --- | --- | --- |
| data: lifeexpectancy.data$lifeexpectancy and factor(lifeexpectancy.data$Continent) | | | |
| F = 48.393 | num df = 5.000 | denom df = 65.227 | p-value < 2.2e-16 |

Analysing the table, it can be interrupted that Null Hypothesis can be rejected at 0% significance level. i.e., There is strong evidence that there is at least one response mean that is different.

**Multiple Comparisons**:

To understand the which specific mean is different, pairwise comparison is performed. These tests are t-tests with adjustments in p-value to account for multiple testing.For this Bonferroni and Tukey’s post-hoc test is used.

Using **Bonferroni** test, the following conclusions were made:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Africa | Asia | Australia/Oceania | Europe | North America |
| Asia | < 2e-16 | - | - | - | - |
| Australia/Oceania | 2.20E-10 | 1 | - | - | - |
| Europe | < 2e-16 | 4.10E-05 | 0.00019 | - | - |
| North America | < 2e-16 | 1 | 0.7129 | 0.06959 | - |
| South America | 4.50E-10 | 1 | 1 | 0.10635 | 1 |

* The p-values of 1 for the pairs of NA-SA, AUS-SA, AS-SA, AS-NA, AS-AUS and 0.71 for NA-AUS shows that the life expectancies of these pairs are statistically similar.
* The p-value of 0.106 shows that there is weak evidence to reject the hypothesis that means of Europe and South America are equal.
* The p-value of 0.06 that there is small evidence of difference in means for EUR and NA.
* All the other pairs have very small p-values, which means there is strong evidence that their means are different.

Using the below summary table of **Tukey** test, the following conclusions were made:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Diff | Lwr | Upr | p adj |
| Asia-Africa | 10.50725 | 7.739989 | 13.27451 | 0 |
| Australia/Oceania-Africa | 9.339738 | 5.578741 | 13.10074 | 0 |
| Europe-Africa | 15.28325 | 12.48621 | 18.08028 | 0 |
| North America-Africa | 12.13863 | 9.051751 | 15.22552 | 0 |
| South America-Africa | 10.98086 | 6.480989 | 15.48072 | 0 |
| Australia/Oceania-Asia | -1.16751 | -4.96746 | 2.63244 | 0.9500903 |
| Europe-Asia | 4.775999 | 1.9268 | 7.625197 | 0.0000399 |
| North America-Asia | 1.631385 | -1.50284 | 4.765612 | 0.6665657 |
| South America-Asia | 0.473607 | -4.05887 | 5.006083 | 0.9996681 |
| Europe-Australia/Oceania | 5.94351 | 2.121822 | 9.765198 | 0.0001815 |
| North America-Australia/Oceania | 2.798896 | -1.23976 | 6.837553 | 0.3496781 |
| South America-Australia/Oceania | 1.641118 | -3.55799 | 6.840224 | 0.944182 |
| North America-Europe | -3.14461 | -6.30516 | 0.015933 | 0.0520136 |
| South America-Europe | -4.30239 | -8.85311 | 0.248323 | 0.0756989 |
| South America-North America | -1.15778 | -5.89217 | 3.576612 | 0.9813866 |

* The difference column shows that Average difference in life expectancies is highest for Europe-Africa and least for South America-Asia.
* The lwr and upr columns can also be seen that shows respective boundaries of differences of each combination.
* The p adj column shows the p value for hypothesis analysis.

Using the below **Tukey**’s plot, the interpretations can be made as:

Chart

Description automatically generated with low confidence

For the pairs whose extended lines of confidence intervals passes through the zero point, the difference is not statistically significant and if not, then there is evidence for difference.

**Conclusion**:

We have replaced missing values with imputation and examined the linear models with all the variables (full model) and also model with reduced variables based on the collinearity. Also beat model-based model based on the Cp and AIC values. We have observed that the full model has the best score of the three and no parameters can be removed. Also, we have observed the difference of mean Life expectancy across continents and by the pair-wise comparison we have observed Europe has highest mean life expectancy and Africa the least.

**Appendix:**

library(dplyr)

library(tidyverse)

library(naniar)

library(ggcorrplot)

library(RColorBrewer)

#descriptive statistics

library(pastecs)

library(moments)

library(Hmisc)

library(finalfit)

library(DMwR2)

life\_exp <- read.csv(file = "C:/Users/91960/OneDrive/Desktop/Life\_Expectancy\_Data1.csv")

dim(life\_exp)

head(life\_exp)

tail(life\_exp)

#1) descriptive statistics:

# -> on the response variable:

summary(life\_exp$SP.DYN.LE00.IN)

summary(life\_exp)

# interquartile range:

IQR(life\_exp$SP.DYN.LE00.IN,na.rm = TRUE)

sd(life\_exp$SP.DYN.LE00.IN,na.rm = TRUE)

med = median(life\_exp$SP.DYN.LE00.I,na.rm = TRUE)

mean = mean(life\_exp$SP.DYN.LE00.I,na.rm = TRUE)

# Plotting density function and histogram of the response variable.

# Histogram with mean,median and density curve.

k = ggplot(life\_exp, aes(x=SP.DYN.LE00.IN)) +

geom\_histogram(aes(y=..density..), # Histogram with density instead of count on y-axis

binwidth=.5,

colour="black", fill="white") +

geom\_density(alpha=.2, fill="#FF6666") + geom\_vline(aes(xintercept=med),

color="blue", size=1)

k+geom\_vline(aes(xintercept=mean),

color="red", linetype="dashed", size=1)

# Kurtosis and skewness:

skewness(life\_exp$SP.DYN.LE00.IN,na.rm = TRUE)

#Hence the data is negatively skewed with the majority of data values greater than mean

kurtosis(life\_exp$SP.DYN.LE00.IN,na.rm = TRUE)

#As the coefficient of kurtosis is less than 3 the distribution is platykurtic.

# dimension of the dataset

dim(life\_exp)

mean(life\_exp$SP.DYN.LE00.IN)

#variables of the dataset

names(life\_exp)

#summary of the data is given here

summary(life\_exp)

#checking nan's

life\_exp %>%miss\_var\_summary()

# Removing the columns having more than 75% missing data. as they having more than 75% missnig values they won't contribute to the model.

threshold<-0.75 #for a 75% cut-off

life\_exp <- life\_exp %>% select(where(~mean(is.na(.))< threshold))

head(life\_exp)

#range of life expectancy

range(life\_exp$SP.DYN.LE00.IN,na.rm=TRUE)

dim(life\_exp)

library(car) # package must be installed first

qqPlot(life\_exp$SP.DYN.LE00.IN)

#changing the names to understandable names

names(life\_exp)[which(names(life\_exp) == "SP.DYN.LE00.IN")] <- "Life expectancy at birth"

#Giving understandable column names

colnames(life\_exp)

names(life\_exp)

life\_exp <- life\_exp

head(life\_exp)

#Task-2

data <- read.csv(file = "C:/Users/91960/OneDrive/Desktop/Life\_Expectancy\_Data1.csv")

library(mice)

dim(data)

library(tidyverse)

data %>% slice(218:266)#contains null for all the columns except one

data\_work<-data %>% slice(1:217)

library(mice)

md.pattern(data\_work)

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

names(data\_work)

cols <- c("data\_work$SP.DYN.LE00.IN","data\_work$EG.ELC.ACCS.ZS","data\_work$NY.ADJ.NNTY.KD.ZG","data\_work$NY.ADJ.NNTY.PC.KD.ZG","data\_work$SH.HIV.INCD.14","data\_work$SE.PRM.UNER","data\_work$SE.PRM.CUAT.ZS","data\_work$SE.TER.CUAT.BA.ZS","data\_work$SP.DYN.IMRT.IN","data\_work$SE.PRM.CMPT.ZS","data\_work$SE.ADT.LITR.ZS","data\_work$FR.INR.RINR","data\_work$SP.POP.GROW","data\_work$EN.POP.DNST","data\_work$SP.POP.TOTL","data\_work$SH.XPD.CHEX.PC.CD","data\_work$SH.XPD.CHEX.GD.ZS","data\_work$SL.UEM.TOTL.NE.ZS","data\_work$NY.GDP.MKTP.KD.ZG","data\_work$NY.GDP.PCAP.CD","data\_work$SP.DYN.CBRT.IN","data\_work$EG.FEC.RNEW.ZS","data\_work$SH.HIV.INCD","data\_work$SH.H2O.SMDW.ZS","data\_work$SI.POV.LMIC")

cols

df2 <- data\_work[1:2]

#df2 = data.frame()

df2.names = c("Country\_Name", "Country\_code","SE.COM.DURS")

data.frame(df2, stringsAsFactors = TRUE)

df2$Country.Name<-data\_work$Country.Name

df2$SE.COM.DURS<-data\_work$SE.COM.DURS

df2

data\_work$Country.Name

drop <- c("Country.Name", "Continent","Country.Code","SE.COM.DURS")

data\_work = data\_work[,!(names(data\_work) %in% drop)]

names(data\_work)

data\_work = log(data\_work)

imputations <- mice(data\_work,method = 'pmm')

print(imputations)

new\_data\_4=complete(imputations)

new\_data\_4

sapply(new\_data\_4, function(x) sum(is.na(x)))

drop\_final <- c("EG.FEC.RNEW.ZS", "SI.POV.LMIC","SE.PRM.UNER","NY.GDP.MKTP.KD.ZG")

new\_data\_4 = new\_data\_4[,!(names(new\_data\_4) %in% drop\_final)]

sapply(new\_data\_4, function(x) sum(is.na(x)))

complete(imputations)

stripplot(imputations, pch = 20, cex = 1.2)

#Task 3

library("faraway")

dataset\_linear\_reg\_model<-lm(new\_data\_4$SP.DYN.LE00.IN~.,data=new\_data\_4)

corr<-cor(new\_data\_4)

round(corr,digits=4)

vif\_score<-vif(new\_data\_4)

vif\_score

drop\_new<-c("SP.DYN.CBRT.IN")

final\_data=new\_data\_4[,!names(new\_data\_4)%in%drop\_new]

View(final\_data)

library(corrplot)

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

#Task 4

#build full model using all columns

full.model<-lm(SP.DYN.LE00.IN ~ .,data=new\_data\_4)

summary(full.model)

library(olsrr)

#Calculate min cp from all models

X <- full.model.cp$x

y <-new\_data\_4$SP.DYN.LE00.IN

library(leaps)

all.models <- leaps(X, y, int = FALSE, strictly.compatible = FALSE, method="Cp")

#Plot all cp

plot(all.models$size, all.models$Cp, log="y", xlab="|M|", ylab=expression(C[p]),ylim=c(1,200))

lines(all.models$size, all.models$size)

#evaluate min cp and consider columns rakted to it

min.cp <- all.models$Cp == min(all.models$Cp) #this finds the smallest C\_p value

min(all.models$Cp) #gives the min C\_p value

min.cp <- all.models$which[min.cp, ] #this finds the corresponding model with the smallest C\_p

min.cp

#built reduced model from the features obtained

reduced.model<-lm(SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS+NY.ADJ.NNTY.KD.ZG+NY.ADJ.NNTY.PC.KD.ZG+

SH.HIV.INCD.14+

SP.DYN.IMRT.IN+SE.PRM.CMPT.ZS+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+NY.GDP.PCAP.CD+

SP.DYN.CBRT.IN+SH.HIV.INCD+SH.H2O.SMDW.ZS, data = new\_data\_4)

#reduced.model <- lm(SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS+NY.ADJ.NNTY.KD.ZG+NY.ADJ.NNTY.PC.KD.ZG+

# SH.HIV.INCD.14+

# SP.DYN.IMRT.IN+SE.PRM.CMPT.ZS+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+NY.GDP.PCAP.CD+

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

# ,data=new\_data\_4)

#standardise full and reduced model

stdres\_fullmodel<-rstandard(full.model)

stdres\_reducedmodel<-rstandard(reduced.model)

#Plot the standardised values

plot(full.model$fitted.values,stdres\_fullmodel,pch=16,

ylab="Standardized Residuals",xlab="fitted y",ylim=c(-3,3),main="Full model")

abline(h=0)

abline(h=2,lty=2)

abline(h=-2,lty=2)

#qq plot for full model

qqnorm(stdres\_fullmodel, ylab="Standardized Residuals",

xlab="Normal Scores", main="QQ Plot for Full model" )

qqline(stdres\_fullmodel)

#qq plot for reduced model

qqnorm(stdres\_reducedmodel, ylab="Standardized Residuals",

xlab="Normal Scores", main="QQ Plot for Reduced model")

qqline(stdres\_reducedmodel)

#Plot full and reduced model

plot(full.model)

plot(reduced.model)

#AIC values for full and reduced models

AIC(reduced.model)

AIC(full.model)

library(olsrr)

ols\_mallows\_cp(reduced.model,full.model)

#full.model.cp <- lm(SP.DYN.LE00.IN ~ ., data = new\_data\_4,x=TRUE)

#summary(best.model.cp)

#AIC(best.model.cp)

#AIC(full.model.cp)

#stepwise selection

initial\_model <- glm(SP.DYN.LE00.IN ~ . , data = new\_data\_4)

selection <- step(initial\_model ,

scope = list(lower = ~ 1 ) ,

data = new\_data\_4 ,

direction = "both")

selection

eval <- selection$coefficients

eval

summary(selection)

#reduced model based on the stepwise feature selection

stepwise\_reduced.model <- lm(formula = SP.DYN.LE00.IN ~ NY.ADJ.NNTY.KD.ZG+ NY.ADJ.NNTY.PC.KD.ZG+

SH.HIV.INCD.14+SE.PRM.CUAT.ZS+SE.TER.CUAT.BA.ZS+SP.DYN.IMRT.IN+

FR.INR.RINR+SP.POP.GROW+EN.POP.DNST+SP.POP.TOTL+SP.DYN.CBRT.IN+

SH.H2O.SMDW.ZS,data = new\_data\_4)

#Standardised full model and reduced model

stdres\_Stepwise\_reducedmodel<-rstandard(stepwise\_reduced.model)

AIC(stepwise\_reduced.model)

ols\_mallows\_cp(stepwise\_reduced.model,full.model)

#Task -5

#Reading the imputed data from csv file after the working directory is changed

#Reading the data from csv file after the working directory is changed

lifeexpectancy.data<-read.csv(file = "C:/Users/91960/OneDrive/Desktop/Life\_Expectancy\_Data1.csv")

dim(lifeexpectancy.data)

#Perform imputation for NA values using MICE

imputations<-mice(lifeexpectancy.data,method = "cart")

lifeexpectancy.data<-complete(imputations,1)

#Renaming the SP.DYN.LE00.IN column to average\_life\_expectancy

names(lifeexpectancy.data)[3]<-"average\_life\_expectancy"

#calculating the group means of continents

group.means<-tapply(lifeexpectancy.data$average\_life\_expectancy,lifeexpectancy.data$Continent,mean)

group.means

#boxplot of average\_life\_expectancy vs continent

boxplot(lifeexpectancy.data$average\_life\_expectancy~lifeexpectancy.data$Continent,main='Comparing the Life Expectancies from all the continents', xlab='Continent', col="light gray", ylab = "Life Expectancy (years)",)

#Shapiro-Wilk normality test

anova1way<-aov(lifeexpectancy.data$average\_life\_expectancy~as.factor(lifeexpectancy.data$Continent),data=lifeexpectancy.data)

lifeexpectancy.data$residuals<-anova1way$residuals

shapiro.test(lifeexpectancy.data$residuals)

#Levene's Test for Homogeneity of Variance

library(car)

leveneTest(lifeexpectancy.data$average\_life\_expectancy~factor(lifeexpectancy.data$Continent))

#One-way analysis of means (not assuming equal variances) using Welch test

data.welchtest<-oneway.test(lifeexpectancy.data$average\_life\_expectancy~factor(lifeexpectancy.data$Continent),data=lifeexpectancy.data)

data.welchtest

#Pairwise comparsions

#Bonferroni post-hoc test

cat("Bonferroni post-hoc test","\n")

pairwise.t.test(lifeexpectancy.data$average\_life\_expectancy,lifeexpectancy.data$Continent,p.adj="bonferroni")

#Tukey post-hoc test

cat("/n","Tukey post-hoc test","\n")

tukey.data<-TukeyHSD(anova1way)

tukey.data

plot(tukey.data)