STAT 530: Applied Regression Analysis

Fall 2022

Project Report

Predicting Life Expectancy using Multiple Linear Regression model

Team

Arun Reddy Yeduguru

Hemanth Reddy Kesamreddy

Yaswanth Reddy Soma

OVERVIEW:

In this project we perform analysis on various factors affecting the Life Expectancy. This will help in suggesting a country which factor should be given importance in order to efficiently improve the life expectancy of its population.

DATA DESCRIPTION:

The Life Expectancy Dataset is taken from the Kaggle using below URL:

https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

It has Life Expectancy records from year 2000-2015 for 193 countries. The Dataset contains 2938 rows and 22 columns. The 22 columns include 20 numerical variables and 2 categorical variables.

Categorical Variables: Country, Status (Developed or Developing status)

Numerical Variables: Life Expectancy (Target Variable), Year, Adult Mortality, Infant Deaths, Alcohol, Percentage Expenditure, Hepatitis B, Measles, BMI, under-five deaths, Polio, Total expenditure, Diphtheria, HIV/AIDS, GDP, Population, thinness 1-19 years, thinness 5-9 years, Income composition of resources, Schooling

INSTALLING AND IMPORTING REQUIRED LIBRARIES:

```
#install.packages("dplyr")
#install.packages("car")
#install.packages("leaps")
#options(scipen=999)

library(dplyr)
library(car)
library(leaps)
```

LOADING THE DATASET AND CHECKING THE DIMENSIONS OF THE DATASET:

```
> setwd('C:\\Users\\arunr\\Downloads')
> data = read.csv("Life Expectancy Data.csv",header=T)
> dim(data)
[1] 2938   22
```

We can see that the data contains 2938 rows and 22 columns in which 20 columns are numerical variables and 2 columns are categorical variables

CHECKING FOR THE NULL VALUES AND REMOVING THE ROWS CONTAINING NULL VALUES:

```
> dim(data)
[1] 2938   22
> sum(is.na(data))
[1] 2563
> data = na.omit(data)
> dim(data)
[1] 1649   22
```

The data contains 2563 null values. After removing the null values, the data contains 1649 rows and 22 columns

CHECKING FOR CORRELATIONS BETWEEN ALL NUMERICAL VARIABLES USING CORRELATION MATRIX AND SCATTER PLOTS:

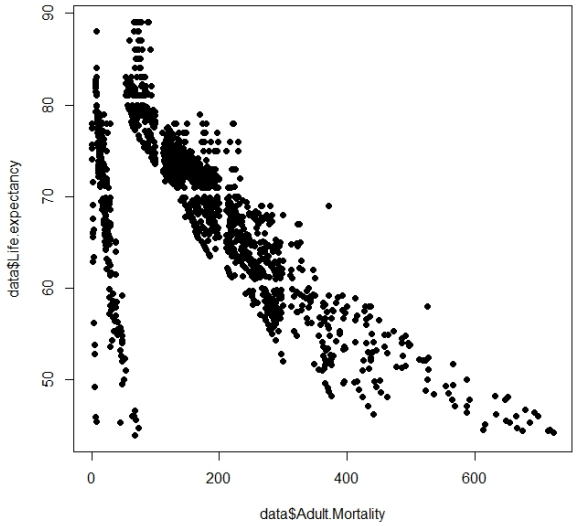
```
> data_numeric = select_if(data, is.numeric)
> dim(data_numeric)
[1] 1649   20
> cor_data = cor(data_numeric)
```

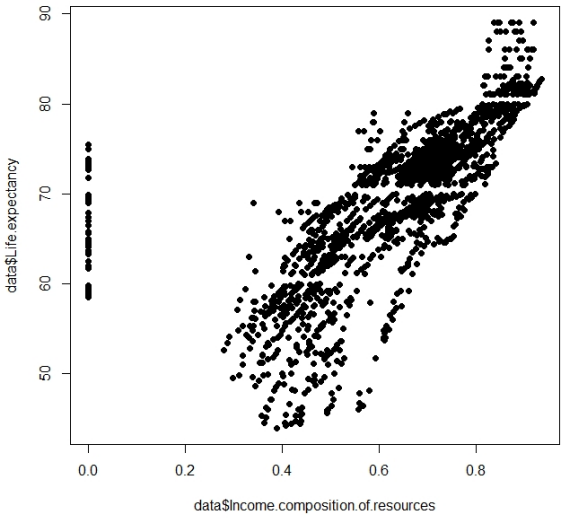
	Year	Life.expectancy	Adult.Mortality	infant.deaths	Alcohol	percentage.expenditure	Hepatitis.B	Measles	BMI	under.five.deaths
Year	1.000000000	0.05077103	-0.037091782	0.008029128	-0.11336476	0.06955347	0.11489709	-0.053822046	0.005739061	0.01047859
Life.expectancy	0.050771035	1.00000000	-0.702523062	-0.169073804	0.40271832	0.40963082	0.19993528	-0.068881222	0.542041588	-0.19226530
Adult.Mortality	-0.037091782	-0.70252306	1.000000000	0.042450237	-0.17553509	-0.23760989	-0.10522544	-0.003966685	-0.351542478	0.06036503
infant.deaths	0.008029128	-0.16907380	0.042450237	1.000000000	-0.10621692	-0.09076463	-0.23176894	0.532679832	-0.234425154	0.99690562
Alcohol	-0.113364764	0.40271832	-0.175535086	-0.106216917	1.00000000	0.41704736	0.10988939	-0.050110235	0.353396205	-0.10108216
percentage.expenditure	0.069553468	0.40963082	-0.237609890	-0.090764632	0.41704736	1.00000000	0.01676017	-0.063070789	0.242738243	-0.09215806
Hepatitis.B	0.114897092	0.19993528	-0.105225443	-0.231768937	0.10988939	0.01676017	1.00000000	-0.124799993	0.143301786	-0.24076603
Measles	-0.053822046	-0.06888122	-0.003966685	0.532679832	-0.05011023	-0.06307079	-0.12479999	1.000000000	-0.153245464	0.51750556
BMI	0.005739061	0.54204159	-0.351542478	-0.234425154	0.35339621	0.24273824	0.14330179	-0.153245464	1.000000000	-0.24213740
under.five.deaths	0.010478594	-0.19226530	0.060365026	0.996905622	-0.10108216	-0.09215806	-0.24076603	0.517505563	-0.242137398	1.00000000
Polio	-0.016698803	0.32729440	-0.199853000	-0.156928805	0.24031453	0.12862605	0.46333080	-0.057850133	0.186267965	-0.17116419
Total.expenditure	0.059492777	0.17471764	-0.085226535	-0.146951117	0.21488509	0.18387236	0.11332668	-0.113582738	0.189468964	-0.14580310
Diphtheria	0.029640586	0.34133123	-0.191428759	-0.161871004	0.24295143	0.13481324	0.58898993	-0.058605907	0.176294503	-0.17844819
HIV.AIDS	-0.123404990	-0.59223629	0.550690745	0.007711547	-0.02711264	-0.09508499	-0.09480197	-0.003521854	-0.210896746	0.01947593
GDP	0.096421485	0.44132181	-0.255034733	-0.098092020	0.44343279	0.95929886	0.04184950	-0.064767590	0.266113973	-0.10033126
Population	0.012566893	-0.02230498	-0.015011838	0.671758310	-0.02888023	-0.01679214	-0.12972265	0.321946377	-0.081415982	0.65867969
thinness..1.19.years	0.019756611	-0.45783819	0.272230044	0.463415256	-0.40375499	-0.25503460	-0.12940595	0.180641506	-0.547017514	0.46478470
thinness.5.9.years	0.014122422	-0.45750829	0.286722882	0.461907925	-0.38620819	-0.25563544	-0.13325099	0.174946217	-0.554093981	0.46228938
Income.composition.of.resources	0.122891780	0.72108259	-0.442203288	-0.134753863	0.56107433	0.40216974	0.18492097	-0.058277256	0.510504831	-0.14809728
Schooling	0.088731787	0.72763003	-0.421170523	-0.214371900	0.61697481	0.42208845	0.21518159	-0.115660481	0.554843903	-0.22601262

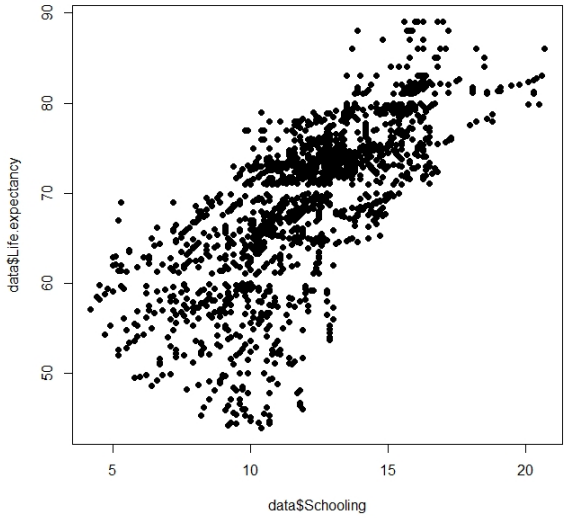
	Polio	Total.expenditure	Diphtheria	HIV.AIDS	GDP	Population	thinness..1.19.years	thinness.5.9.years	Income.composition.of.resources	Schooling
Year	-0.01669880	0.05949278	0.02964059	-0.123404990	0.09642148	0.012566893	0.01975661	0.01412242	0.122891780	0.08873179
Life.expectancy	0.32729440	0.17471764	0.34133123	-0.592236293	0.44132181	-0.022304978	-0.45783819	-0.45750829	0.721082593	0.72763003
Adult.Mortality	-0.19985300	-0.08522653	-0.19142876	0.550690745	-0.25503473	-0.015011838	0.27223004	0.28672288	-0.442203288	-0.42117052
infant.deaths	-0.15692881	-0.14695112	-0.16187100	0.007711547	-0.09809202	0.671758310	0.46341526	0.46190792	-0.134753863	-0.21437190
Alcohol	0.24031453	0.21488509	0.24295143	-0.027112636	0.44343279	-0.028880232	-0.40375499	-0.38620819	0.561074332	0.61697481
percentage.expenditure	0.12862605	0.18387236	0.13481324	-0.095084991	0.95929886	-0.016792141	-0.25503460	-0.25563544	0.402169736	0.42208845
Hepatitis.B	0.46333080	0.11332668	0.58898993	-0.094801971	0.04184950	-0.129722655	-0.12940595	-0.13325099	0.184920970	0.21518159
Measles	-0.05785013	-0.11358274	-0.05860591	-0.003521854	-0.06476759	0.321946377	0.18064151	0.17494622	-0.058277256	-0.11566048
BMI	0.18626797	0.18946896	0.17629450	-0.210896746	0.26611397	-0.081415982	-0.54701751	-0.55409398	0.510504831	0.55484390
under.five.deaths	-0.17116419	-0.14580310	-0.17844819	0.019475927	-0.10033126	0.658679691	0.46478470	0.46228938	-0.148097276	-0.22601262
Polio	1.00000000	0.11976798	0.60924547	-0.107885468	0.15680869	-0.045386572	-0.16406959	-0.17448925	0.314681594	0.35014660
Total.expenditure	0.11976798	1.00000000	0.12991481	0.043100657	0.18037347	-0.079962237	-0.20987232	-0.21786479	0.183653190	0.24378345
Diphtheria	0.60924547	0.12991481	1.00000000	-0.117601074	0.15843774	-0.039897537	-0.18724165	-0.18095238	0.343261772	0.35039793
HIV.AIDS	-0.10788547	0.04310066	-0.11760107	1.000000000	-0.10808060	-0.027800562	0.17259177	0.18314673	-0.248589855	-0.21184020
GDP	0.15680869	0.18037347	0.15843774	-0.108080600	1.00000000	-0.020368964	-0.27749835	-0.27795855	0.446855511	0.46794697
Population	-0.04538657	-0.07996224	-0.03989754	-0.027800562	-0.02036896	1.000000000	0.28252928	0.27791337	-0.008132466	-0.04031242
thinness..1.19.years	-0.16406959	-0.20987232	-0.18724165	0.172591767	-0.27749835	0.282529280	1.00000000	0.92791344	-0.453678854	-0.49119921
thinness.5.9.years	-0.17448925	-0.21786479	-0.18095238	0.183146727	-0.27795855	0.277913374	0.92791344	1.00000000	-0.438483721	-0.47248203
Income.composition.of.resources	0.31468159	0.18365319	0.34326177	-0.248589855	0.44685551	-0.008132466	-0.45367885	-0.43848372	1.000000000	0.78474058
Schooling	0.35014660	0.24378345	0.35039793	-0.211840201	0.46794697	-0.040312419	-0.49119921	-0.47248203	0.784740581	1.00000000

From the correlation matrix we can see that the target variable (Life Expectancy) is highly correlated with 3 numerical variables Adult Mortality, Income composition of resources, Schooling.

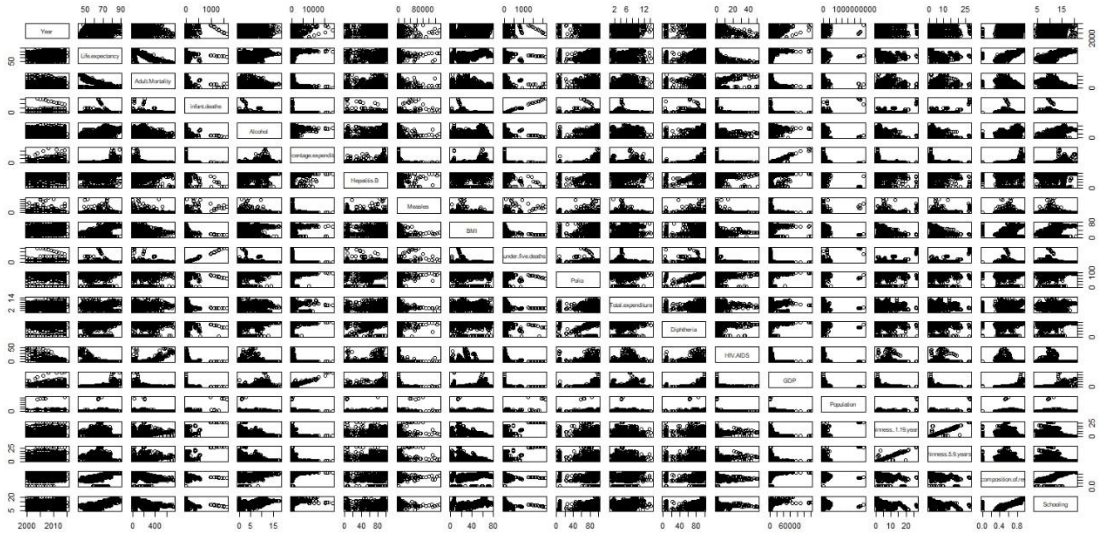
- Correlation between Life Expectancy and Adult Mortality is -0.70252306
- Correlation between Life Expectancy and Income composition of resources is 0.72108259
- Correlation between Life Expectancy and Schooling is 0.72763003







Scatter plot of between all numerical variables:

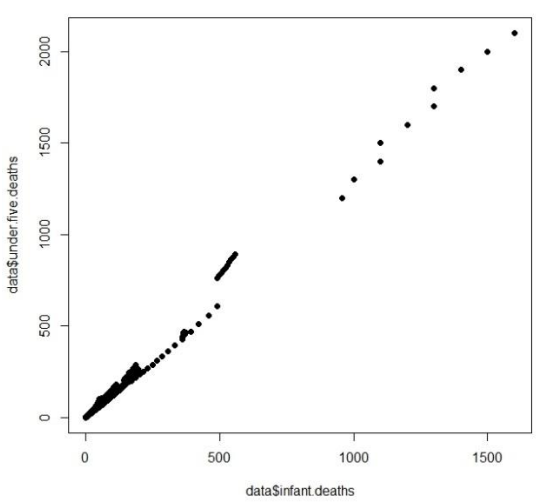


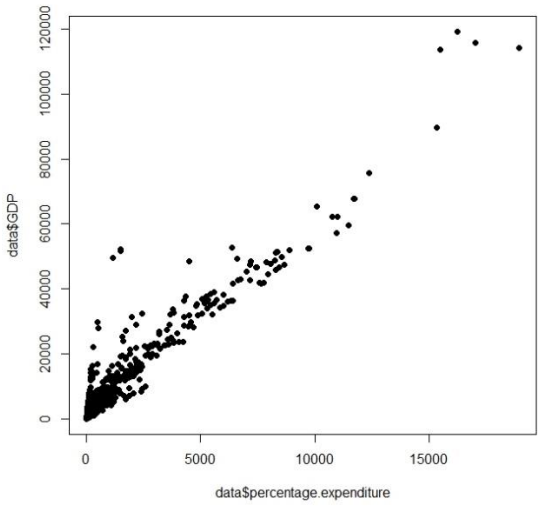
From the correlation matrix and scatter plot of all numerical variables we can see that there exists some multicollinearity between the explanatory variables (i.e., X variables) itself.

CHECKING THE MULTICOLLINEARITY BETWEEN THE EXPLANATORY VARIABLES USING VARIANCE INFLATION FACTOR (VIF):

```
> model1 <- lm(Life.expectancy ~ . , data=data_numeric)
> vif(model1)
                      Year                Adult.Mortality                infant.deaths
                  1.157920                      1.809171                    213.609554
                   Alcohol          percentage.expenditure                   Hepatitis.B
                  2.067310                     12.904426                      1.680406
                   Measles                           BMI               under.five.deaths
                  1.516630                      1.802986                    203.591034
                     Polio             Total.expenditure                     Diphtheria
                  1.722414                      1.124370                      2.094307
                   HIV.AIDS                           GDP                    Population
                  1.500870                     13.649710                      1.943421
 thinness..1.19.years           thinness.5.9.years Income.composition.of.resources
                  7.606109                      7.584832                      3.028945
                Schooling
                  3.538093
```

We are using just the numerical data in the model for finding VIF. From the above image we can see that **VIF for infant deaths and under five deaths is high (i.e., >10)** which indicates that both these are highly correlated, also the **VIF for percentage expenditure and GDP is high (i.e., >10)** which indicates both of these are highly correlated. Let's check this using scatter plots.





Removing one of the infant deaths variable or under five deaths variable based on which variable having less correlation with the life expectancy (target variable)

```
> cor(data$Life.expectancy, data$infant.deaths)
[1] -0.1690738
> cor(data$Life.expectancy, data$under.five.deaths)
[1] -0.1922653
```

We can see that infant deaths variable is less correlated with life expectancy when compared with under five deaths variable

Removing one of the percentage expenditure variable or GDP variable based on which variable having less correlation with the life expectancy (target variable)

```
> cor(data$Life.expectancy, data$percentage.expenditure)
[1] 0.4096308
> cor(data$Life.expectancy, data$GDP)
[1] 0.4413218
```

We can see that percentage expenditure variable is less correlated with life expectancy when compared with GDP variable

Therefore, we remove the infant deaths variable and percentage expenditure variable from the data

```
> data <- subset(data, select = -c(infant.deaths,percentage.expenditure))
> dim(data)
[1] 1649   20
```

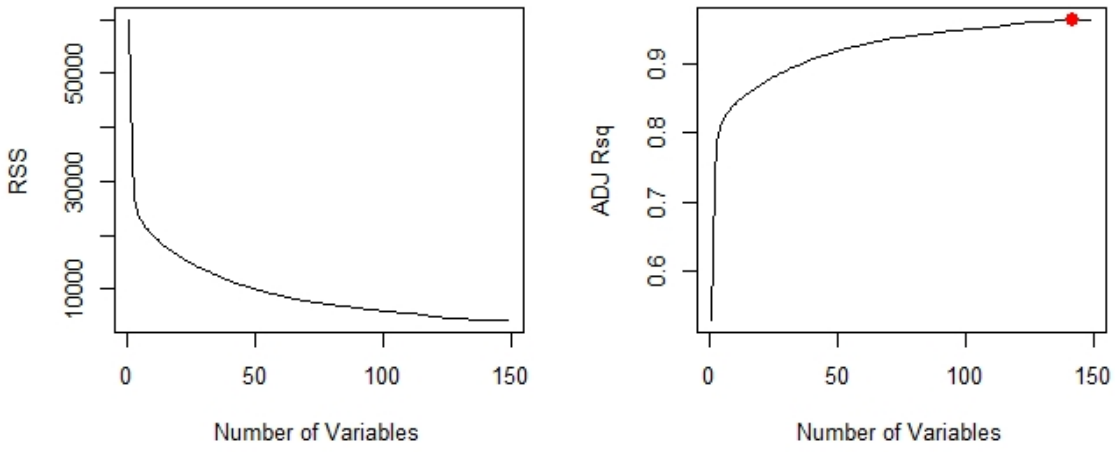
After removing the infant deaths variable and percentage expenditure variable the data contains 20 columns (i.e 18 numerical, 2 categorical) which includes life expectancy (i.e., target variable) as well.

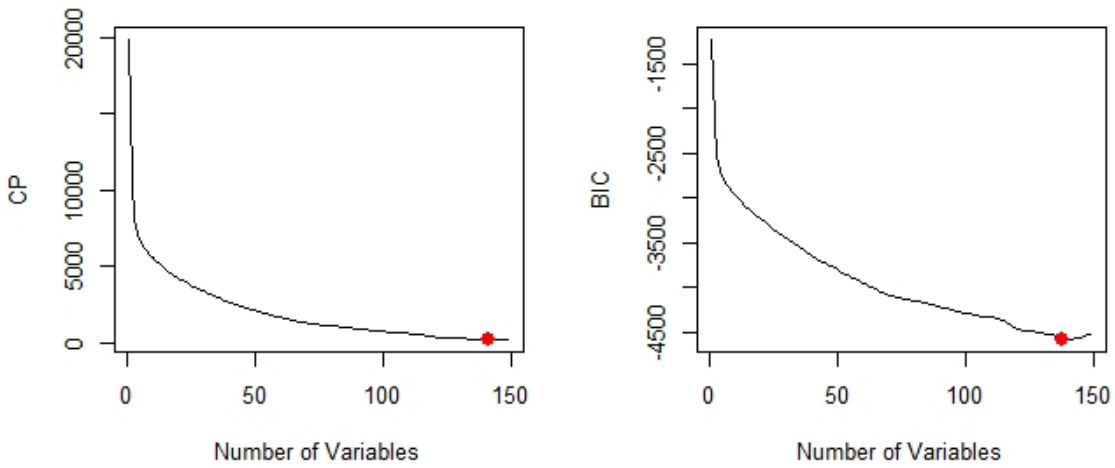
MODEL SELECTION BY FORWARD SELECTION:

```
> regfit.fwd = regsubsets(Life.expectancy ~ . , data=data, nvmax=150, method="forward")
> reg.fwd = summary(regfit.fwd)
```

The nvmax=150 because there is total 150 explanatory variables including the indicator variables of the two categorical variables (i.e., country, status).

```
> par(mfrow=c(2,2))
> reg.summary=reg.fwd;
> plot(reg.summary$rss,xlab="Number of Variables",ylab="RSS",type="l")
> plot(reg.summary$adjr2,xlab="Number of Variables",ylab="ADJ Rsq",type="l")
> which.max(reg.summary$adjr2)
[1] 142
> points(142,reg.summary$adjr2[142],col="red",cex=2,pch=20)
> plot(reg.summary$cp,xlab="Number of Variables",ylab="CP",type="l")
> which.min(reg.summary$cp)
[1] 141
> points(141, reg.summary$cp[141], col="red",cex=2,pch=20)
> which.min(reg.summary$bic)
[1] 138
> plot(reg.summary$bic, xlab="Number of Variables", ylab="BIC",type="l")
> points(138, reg.summary$bic[138], col="red", cex=2, pch=20)
```





From the graphs we can see that the Adjusted R^2 value is high at 142 explanatory variables

```
> coef(regfit.fwd,142)
```

End part of the above code result looks like below:

```
     CountryZimbabwe                 Year              Adult.Mortality
  -8.1540830699929163   0.2839591509432413   -0.0023711990807554
             Alcohol          Hepatitis.B              Measles
  -0.1229663872189872   0.0045080927152851   -0.0000023508533540
   under.five.deaths               Polio                  GDP
  -0.0025257718689371   0.0004668737998839   -0.0000013196685115
          Population     thinness.5.9.years Income.composition.of.resources
   0.0000000004691167   0.0085217973723671    0.4383752953674996
           Schooling      StatusDeveloping
   0.2385763940029021  -21.5578503830794581
```

These 142 explanatory variables include all the indicator variables of Country variable along with Year, Adult Mortality, Alcohol, Hepatitis B, Measles, under five deaths, Polio, GDP, Population, thinness 5-9 years, Income composition of resources, Schooling and Status variables

Fitting the model based on the features obtained from forward selection:

```
> model2 <- lm(Life.expectancy ~ Country+ Year+ Adult.Mortality+ Alcohol+ Hepatitis.B+ Measles
+             + under.five.deaths+ Polio+ GDP+ Population+ thinness.5.9.years+ Income.composition.of.resources
+             + Schooling+ Status, data=data)
> summary(model2)
```

The below image shows the last few lines of summary(model2) output:

```
CountryZambia                    -1.9444059695864   0.8515971134736  -2.283             0.022555 *
CountryZimbabwe                  -8.1008253036621   0.7597218459569 -10.663 < 0.0000000000000002 ***
Year                              0.2867511511712   0.0188424445257  15.218 < 0.0000000000000002 ***
Adult.Mortality                  -0.0023690235466   0.0005978989353  -3.962  0.00007772478383805 ***
Alcohol                          -0.1277572693302   0.0336582340681  -3.796             0.000153 ***
Hepatitis.B                       0.0043100224095   0.0024688885527   1.746             0.081061 .
Measles                          -0.0000024464190   0.0000072100530  -0.339             0.734426
under.five.deaths                -0.0025797239043   0.0018257923685  -1.413             0.157882
Polio                             0.0005554736732   0.0027498237930   0.202             0.839941
GDP                              -0.0000012085788   0.0000063354944  -0.191             0.848737
Population                        0.0000000004602   0.0000000010008   0.460             0.645681
thinness.5.9.years                0.0094404890562   0.0285799708438   0.330             0.741205
Income.composition.of.resources   0.4298337613873   0.6609891776811   0.650             0.515605
Schooling                         0.2208440869874   0.0872042866575   2.532             0.011427 *
StatusDeveloping                               NA                NA      NA                   NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.866 on 1504 degrees of freedom
Multiple R-squared:  0.9589,    Adjusted R-squared:  0.955
F-statistic: 243.9 on 144 and 1504 DF,  p-value: < 0.00000000000000022
```

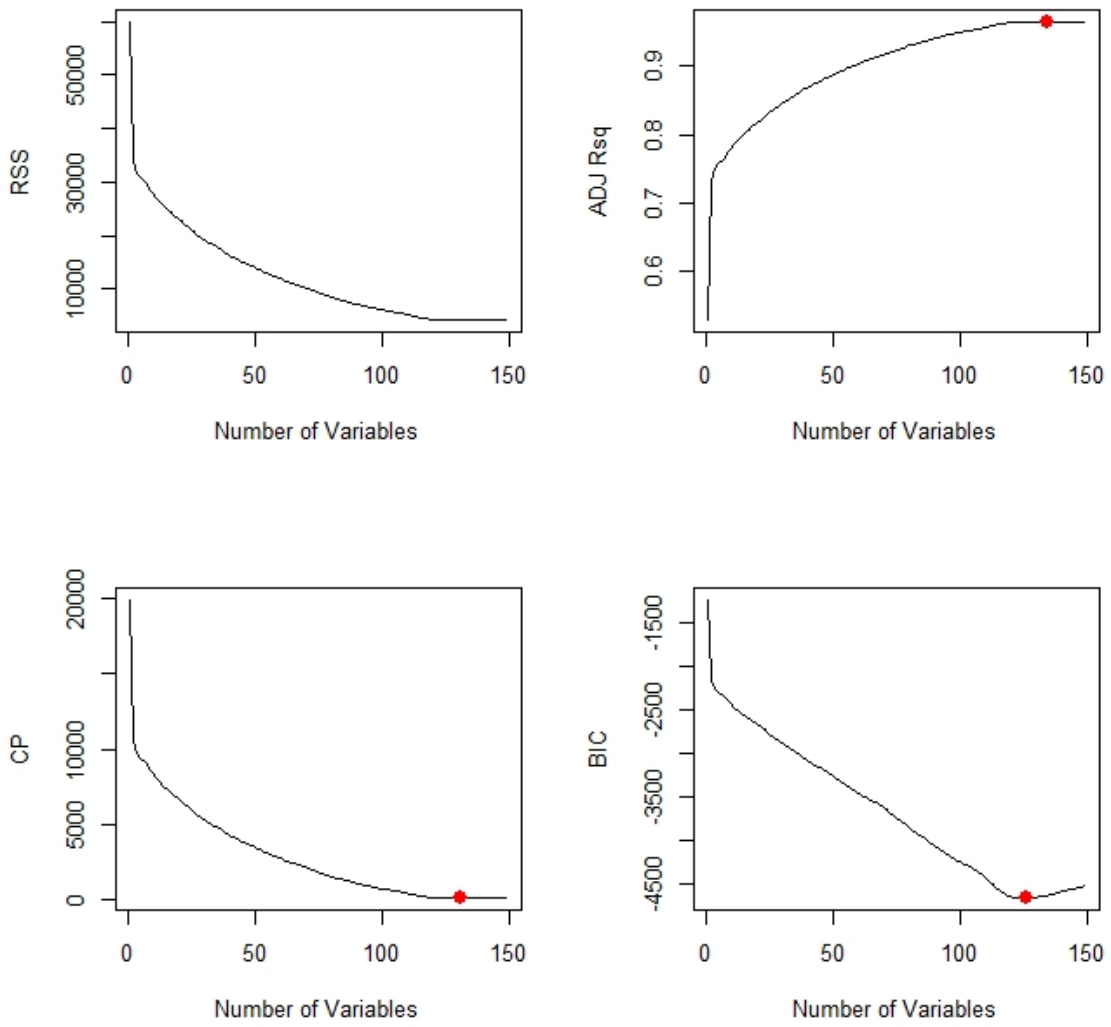
The Adjusted R^2 for this forward selection model is 0.955

MODEL SELECTION BY BACKWARD SELECTION:

```
> regfit.bwd = regsubsets(Life.expectancy ~ . , data=data, nvmax=150, method="backward")
> reg.bwd = summary(regfit.bwd)
```

The nvmax=150 because there is total 150 explanatory variables including the indicator variables of the two categorical variables (i.e., country, status).

```
> par(mfrow=c(2,2))
> reg.summary=reg.bwd;
> plot(reg.summary$rss,xlab="Number of Variables",ylab="RSS",type="l")
> plot(reg.summary$adjr2,xlab="Number of Variables",ylab="ADJ Rsq",type="l")
> which.max(reg.summary$adjr2)
[1] 134
> points(134,reg.summary$adjr2[134],col="red",cex=2,pch=20)
> plot(reg.summary$cp,xlab="Number of Variables",ylab="CP",type="l")
> which.min(reg.summary$cp)
[1] 131
> points(131, reg.summary$cp[131], col="red",cex=2,pch=20)
> which.min(reg.summary$bic)
[1] 126
> plot(reg.summary$bic, xlab="Number of Variables", ylab="BIC",type="l")
> points(126, reg.summary$bic[126], col="red", cex=2, pch=20)
```



From the graphs we can see that the Adjusted R^2 value is high at 134 explanatory variables

```
> coef(regfit.bwd,134)
```

End part of the above code result looks like below:

```
                              Year                            Alcohol                         Hepatitis.B
                      0.295415048020                -0.132677336854                0.004487457222
                           Measles                            Polio                            GDP
                     -0.000004520826                -0.001057049908               -0.000001780426
Income.composition.of.resources                          Schooling                StatusDeveloping
                      0.683927127444                 0.218689862019                0.000000000000
```

These 134 explanatory variables include all the indicator variables of Country variable along with Year, Alcohol, Hepatitis B, Measles, Polio, GDP, Income composition of resources, Schooling and Status variables.

Fitting the model based on the features obtained from backward selection:

```
> model3 <- lm(Life.expectancy ~ Country+ Year+ Alcohol+ Hepatitis.B+ Measles
+              + Polio+ GDP+ Income.composition.of.resources
+              + Schooling+ Status, data=data)
> summary(model3)
```

The below image shows the last few lines of summary(model3) output:

```
CountryZambia                    -2.186224132   0.810133787  -2.699             0.007041 **
CountryZimbabwe                  -8.513677607   0.706430418 -12.052 < 0.0000000000000002 ***
Year                              0.294838908   0.018685464  15.779 < 0.0000000000000002 ***
Alcohol                          -0.127668222   0.033727436  -3.785             0.000160 ***
Hepatitis.B                       0.004473663   0.002465671   1.814             0.069818 .
Measles                          -0.000003216   0.000007118  -0.452             0.651482
Polio                             0.000613963   0.002760895   0.222             0.824050
GDP                              -0.000001735   0.000006362  -0.273             0.785090
Income.composition.of.resources   0.444851310   0.663674794   0.670             0.502779
Schooling                         0.236203180   0.087046331   2.714             0.006733 **
StatusDeveloping                         NA            NA      NA                   NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.874 on 1508 degrees of freedom
Multiple R-squared:  0.9585,    Adjusted R-squared:  0.9546
F-statistic: 248.6 on 140 and 1508 DF,  p-value: < 0.00000000000000022
```

The Adjusted R^2 for this forward selection model is 0.9546

From both the forward selection model(model2) and backward selection model(model3) we choose the forward selection model finally because it has higher Adjusted R^2 value (i.e., 0.955 > 0.9546)

ANALYSIS OF MODEL2

```
> model2 <- lm(Life.expectancy ~ Country+ Year+ Adult.Mortality+ Alcohol+ Hepatitis.B+ Measles
+             + under.five.deaths+ Polio+ GDP+ Population+ thinness.5.9.years+ Income.composition.of.resources
+             + Schooling+ Status, data=data)
> summary(model2)
```

The below image shows the last few lines of summary(model2) output:

```
CountryZambia                    -1.9444059695864    0.8515971134736  -2.283             0.022555 *
CountryZimbabwe                  -8.1008253036621    0.7597218459569 -10.663 < 0.0000000000000002 ***
Year                              0.2867511511712    0.0188424445257  15.218 < 0.0000000000000002 ***
Adult.Mortality                  -0.0023690235466    0.0005978989353  -3.962  0.00007772478383805 ***
Alcohol                          -0.1277572693302    0.0336582340681  -3.796             0.000153 ***
Hepatitis.B                       0.0043100224095    0.0024688885527   1.746             0.081061 .
Measles                          -0.0000024464190    0.0000072100530  -0.339             0.734426
under.five.deaths                -0.0025797239043    0.0018257923685  -1.413             0.157882
Polio                             0.0005554736732    0.0027498237930   0.202             0.839941
GDP                              -0.0000012085788    0.0000063354944  -0.191             0.848737
Population                        0.0000000004602    0.0000000010008   0.460             0.645681
thinness.5.9.years                0.0094404890562    0.0285799708438   0.330             0.741205
Income.composition.of.resources   0.4298337613873    0.6609891776811   0.650             0.515605
Schooling                         0.2208440869874    0.0872042866575   2.532             0.011427 *
StatusDeveloping                               NA                 NA      NA                   NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.866 on 1504 degrees of freedom
Multiple R-squared:  0.9589,    Adjusted R-squared:  0.955
F-statistic: 243.9 on 144 and 1504 DF,  p-value: < 0.00000000000000022
```

The fitted line is:

Life Expectancy = -518.985 + 15.741*CountryAlbania + 13.818*CountryAlgeria + ……

- 8.100*CountryZimbabwe + 0.2867*Year -0.0023*Adult Mortality -0.1277*Alcohol

+ 0.0043*Hepatitis B - 0.000024*Measles – 0.00259*under five deaths

+ 0.00055*Polio - 0.0000012*GDP + 0.0000000004602*Population

+ 0.0094*thinness 5-9 years + 0.4298*Income composition of resources

+ 0.2208*schooling

Here the ……. represents the sum of different country indicator variables multiplied with their respective coefficients obtained from the model

Let’s perform Hypothesis Test on certain explanatory variables to check their significance with the model

Hypothesis test to check the significance of Alcohol in model2:

```
Alcohol                    -0.1277572693302     0.0336582340681  -3.796            0.000153 ***
```

```
> n=nrow(data)
> k=14+1
> qt(1-0.05/2,n-k)
[1] 1.961417
```

Here β_1 represents coefficient of alcohol in the model2

i. hypothesis $\begin{cases} H_0: \beta_1 = 0 \\ H_a: \beta_1 \neq 0 \end{cases}$.

ii. Test statistic: $t_{obs} = \frac{-0.1277572693302-0}{0.0336582340681} = -3.795$

iii. Rejection region: $t > t_{\frac{\alpha}{2}, n-4} = 1.961417 \; or \; t < -t_{\frac{\alpha}{2}, n-4} = -1.961417$

iv. Since t_{obs} is in the rejection region, we reject the H_0. It indicates that Alcohol is significantly linear related to Life Expectancy.

Similarly, we can predict the significance of the other explanatory variables using hypothesis test.

MODEL ADEQUACY FOR MODEL2:

Computing the residuals for the model

Below image shows the first 20 residuals of the model

```
> residuals(model2)
                  1                   2                   3                   4
 4.124968585170249113503 -0.642267324404786843495 -0.357532191779858732339 -0.425717815111092634517
                  5                   6                   7                   8
-0.358021975338406406308 -0.374090248807048109114 -0.185047442122441035961 -0.337663380060163054175
                  9                  10                  11                  12
-0.560649001269311164641 -0.384732903128381154012 -0.066970424563711633548  0.189518537212139681625
                 13                  14                  15                  16
 0.252115654315269943631 -0.403606986860251648608 -0.181671905485137924474 -0.288631177767719293570
                 17                  18                  19                  20
 0.050949999865469941529 -0.121646004753095587914  0.071945433490522792130  0.114943206983261297927
```

Computing the standardized residuals and studentized residuals for the model

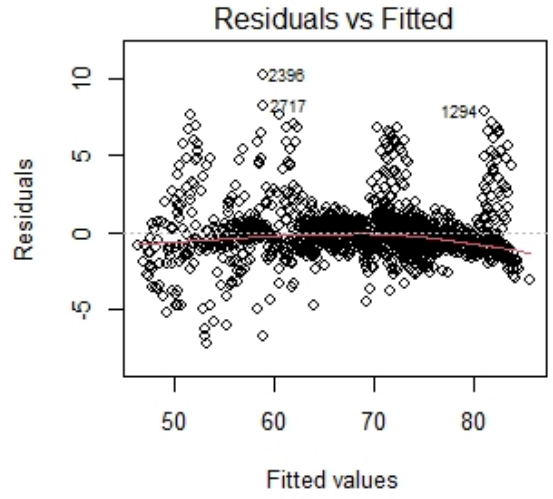
Below image shows the first 21 standardized residuals of the model

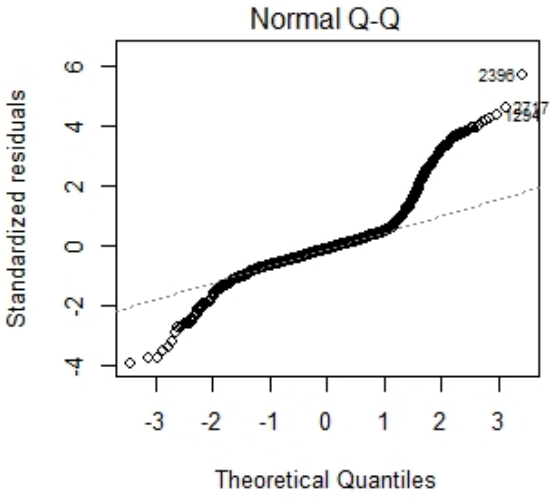
```
> rstandard(model2)
          1           2           3           4           5           6           7
 2.2954281745 -0.3565096092 -0.1984614798 -0.2363094589 -0.1986542148 -0.2074899733 -0.1026207166
          8           9          10          11          12          13          14
-0.1872589107 -0.3109330980 -0.2133507293 -0.0371458105  0.1054934487  0.1401180208 -0.2304945372
         15          16          17          18          19          20          21
-0.1033359131 -0.1644196939  0.0282830221 -0.0675186526  0.0399434295  0.0638295257  0.1882941130
```

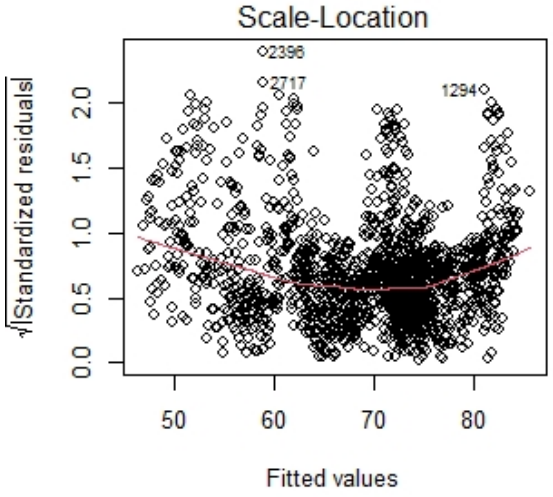
Below image shows the first 21 studentized residuals of the model

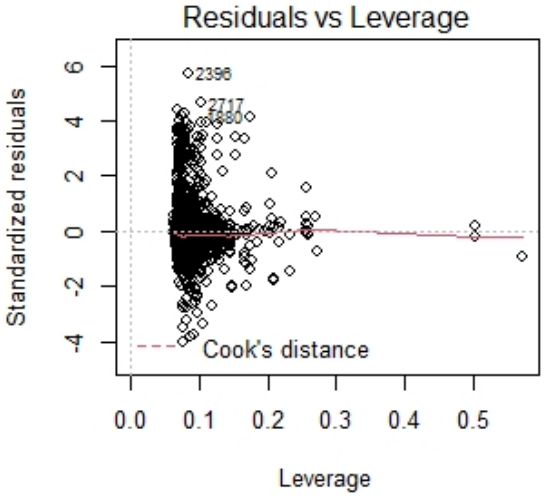
```
> rstudent(model2)
          1           2           3           4           5           6           7
 2.2986950026 -0.3564061288 -0.1983980888 -0.2362352711 -0.1985907673 -0.2074239512 -0.1025869542
          8           9          10          11          12          13          14
-0.1871988290 -0.3108397029 -0.2132830173 -0.0371334765  0.1054587620  0.1400723455 -0.2304219671
         15          16          17          18          19          20          21
-0.1033019205 -0.1643665012  0.0282736255 -0.0674963048  0.0399301695  0.0638083887  0.1882337235
```

SOME IMPORTANT PLOTS:









Analysis of the plots:

The NORMAL Q-Q plot has sharp upward and downward curves at both the extreme. It indicates that the underlying distribution is heavy tailed.

Summary of the Analysis:

1) First, we loaded the data and checked and removed the rows containing null values.
2) Then we checked for multicollinearity between the variables using VIF and removed some of the variables with multicollinearity.
3) Then we performed forward selection and backward elimination methods for finding the best subset of features.
4) After that we have chosen the model obtained from forward selection as it has high adjusted R^2 value when compared with the model obtained after backward elimination.
5) Then we performed Hypothesis Test for checking the significance of the Alcohol on the model obtained after forward selection and found out that alcohol has significant effect on life expectancy.
6) Finally, we performed model adequacy on the model obtained after forward selection by finding the residuals and plotting some important plots.