Code

Life expectancy prediction(Regression Analysis)

Hide

```
#Load data
#check working directory
getwd()
```

```
[1] "E:/"
```

Hide

```
#set working directory
setwd("E:/")
#load file in csv format
data<-read.csv("Life Expectancy Data.csv")
#Explore data
#display first 6 rows
head(data)
```

```
      Country Year     Status Life.expectancy Adult.Mortality infant.deaths Alcohol percentage.e
xpenditure
1 Afghanistan 2015 Developing            65.0             263            62    0.01
71.279624
2 Afghanistan 2014 Developing            59.9             271            64    0.01
73.523582
3 Afghanistan 2013 Developing            59.9             268            66    0.01
73.219243
4 Afghanistan 2012 Developing            59.5             272            69    0.01
78.184215
5 Afghanistan 2011 Developing            59.2             275            71    0.01
  7.097109
6 Afghanistan 2010 Developing            58.8             279            74    0.01
79.679367
  Hepatitis.B Measles  BMI under.five.deaths Polio Total.expenditure Diphtheria HIV.AIDS       G
DP Population
1          65    1154 19.1                83     6              8.16         65      0.1 584.259
21   33736494
2          62     492 18.6                86    58              8.18         62      0.1 612.696
51     327582
3          64     430 18.1                89    62              8.13         64      0.1 631.744
98   31731688
4          67    2787 17.6                93    67              8.52         67      0.1 669.959
00    3696958
5          68    3013 17.2                97    68              7.87         68      0.1  63.537
23    2978599
6          66    1989 16.7               102    66              9.20         66      0.1 553.328
94    2883167
  thinness..1.19.years thinness.5.9.years Income.composition.of.resources Schooling
1                 17.2                17.3                           0.479      10.1
2                 17.5                17.5                           0.476      10.0
3                 17.7                17.7                           0.470       9.9
4                 17.9                18.0                           0.463       9.8
5                 18.2                18.2                           0.454       9.5
6                 18.4                18.4                           0.448       9.2
```

Hide

```
#check structure of data
str(data)
```

```
'data.frame':   2938 obs. of  22 variables:
 $ Country                        : Factor w/ 193 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1
...
 $ Year                           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
 $ Status                         : Factor w/ 2 levels "Developed","Developing": 2 2 2 2 2 2 2 2
2 2 ...
 $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
 $ Adult.Mortality                : int  263 271 268 272 275 279 281 287 295 295 ...
 $ infant.deaths                  : int  62 64 66 69 71 74 77 80 82 84 ...
 $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
 $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
 $ Hepatitis.B                    : int  65 62 64 67 68 66 63 64 63 64 ...
 $ Measles                        : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
 $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
 $ under.five.deaths              : int  83 86 89 93 97 102 106 110 113 116 ...
 $ Polio                          : int  6 58 62 67 68 66 63 64 63 58 ...
 $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
 $ Diphtheria                     : int  65 62 64 67 68 66 63 64 63 58 ...
 $ HIV.AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
 $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
 $ Population                     : num  33736494 327582 31731688 3696958 2978599 ...
 $ thinness..1.19.years           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
 $ thinness.5.9.years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
 $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.
405 ...
 $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
```

Hide

```
#rename variable names
names(data)[c(4,5,6,8,9,12,14,16,19,20,21)]<-c("Life_expectancy","Adult_Mortality","Infant_death
s",
                          "Percentage_expenditure","Hepatitis_B","Under_5_deaths",
                          "Total_expenditure","HIV_AIDS","Thinness_1_to_19_yrs","Thinness_5_to_
9_yrs",
                          "Income_composition_of_resources")
#check summary of data
summary(data)
```

```
               Country          Year             Status     Life_expectancy Adult_Mortality Inf
ant_deaths
 Afghanistan        :  16   Min.   :2000   Developed : 512   Min.   :36.30   Min.   :  1.0   Mi
n.   :   0.0
 Albania            :  16   1st Qu.:2004   Developing:2426   1st Qu.:63.10   1st Qu.: 74.0   1st
Qu.:   0.0
 Algeria            :  16   Median :2008                     Median :72.10   Median :144.0   Med
ian :   3.0
 Angola             :  16   Mean   :2008                     Mean   :69.22   Mean   :164.8   Mea
n   :  30.3
 Antigua and Barbuda:  16   3rd Qu.:2012                     3rd Qu.:75.70   3rd Qu.:228.0   3rd
Qu.:  22.0
 Argentina          :  16   Max.   :2015                     Max.   :89.00   Max.   :723.0   Ma
x.   :1800.0
 (Other)            :2842                                    NA's   :10      NA's   :10

    Alcohol        Percentage_expenditure  Hepatitis_B       Measles              BMI        Und
er_5_deaths
 Min.   : 0.0100   Min.   :    0.000      Min.   : 1.00   Min.   :     0.0   Min.   : 1.00   Mi
n.   :   0.00
 1st Qu.: 0.8775   1st Qu.:    4.685      1st Qu.:77.00   1st Qu.:     0.0   1st Qu.:19.30   1st
Qu.:   0.00
 Median : 3.7550   Median :   64.913      Median :92.00   Median :    17.0   Median :43.50   Med
ian :   4.00
 Mean   : 4.6029   Mean   :  738.251      Mean   :80.94   Mean   :  2419.6   Mean   :38.32   Mea
n   :  42.04
 3rd Qu.: 7.7025   3rd Qu.:  441.534      3rd Qu.:97.00   3rd Qu.:   360.2   3rd Qu.:56.20   3rd
Qu.:  28.00
 Max.   :17.8700   Max.   :19479.912      Max.   :99.00   Max.   :212183.0   Max.   :87.30   Ma
x.   :2500.00
 NA's   :194                              NA's   :553                        NA's   :34

     Polio       Total_expenditure   Diphtheria       HIV_AIDS           GDP              Popula
tion
 Min.   : 3.00   Min.   : 0.370    Min.   : 2.00   Min.   : 0.100   Min.   :     1.68   Min.   :
3.400e+01
 1st Qu.:78.00   1st Qu.: 4.260    1st Qu.:78.00   1st Qu.: 0.100   1st Qu.:   463.94   1st Qu.:
1.958e+05
 Median :93.00   Median : 5.755    Median :93.00   Median : 0.100   Median :  1766.95   Median :
1.387e+06
 Mean   :82.55   Mean   : 5.938    Mean   :82.32   Mean   : 1.742   Mean   :  7483.16   Mean   :
1.275e+07
 3rd Qu.:97.00   3rd Qu.: 7.492    3rd Qu.:97.00   3rd Qu.: 0.800   3rd Qu.:  5910.81   3rd Qu.:
7.420e+06
 Max.   :99.00   Max.   :17.600    Max.   :99.00   Max.   :50.600   Max.   :119172.74   Max.   :
1.294e+09
 NA's   :19      NA's   :226       NA's   :19                       NA's   :448         NA's   :
652
 Thinness_1_to_19_yrs Thinness_5_to_9_yrs Income_composition_of_resources   Schooling
 Min.   : 0.10        Min.   : 0.10       Min.   :0.0000                  Min.   : 0.00
 1st Qu.: 1.60        1st Qu.: 1.50       1st Qu.:0.4930                  1st Qu.:10.10
 Median : 3.30        Median : 3.30       Median :0.6770                  Median :12.30
 Mean   : 4.84        Mean   : 4.87       Mean   :0.6276                  Mean   :11.99
```

```
 3rd Qu.: 7.20        3rd Qu.: 7.20        3rd Qu.:0.7790                  3rd Qu.:14.30
 Max.   :27.70        Max.   :28.60        Max.   :0.9480                  Max.   :20.70
 NA's   :34           NA's   :34           NA's   :167                     NA's   :163
```

Hide

```
#Clean data
#keeping original data safe
data1<-data
#display no. of missing values column wise in decreasing order
sort(colSums(is.na(data1)),decreasing = TRUE)
```

```
                     Population                     Hepatitis_B                             GDP
                            652                             553                             448
             Total_expenditure                         Alcohol Income_composition_of_resources
                            226                             194                             167
                     Schooling                             BMI             Thinness_1_to_19_yrs
                            163                              34                              34
            Thinness_5_to_9_yrs                           Polio                     Diphtheria
                             34                              19                              19
                Life_expectancy                 Adult_Mortality                         Country
                             10                              10                               0
                           Year                          Status                   Infant_deaths
                              0                               0                               0
          Percentage_expenditure                        Measles                  Under_5_deaths
                              0                               0                               0
                        HIV_AIDS
                              0
```

Hide

```
#replace missing values with mean
data1$Life_expectancy[is.na(data1$Life_expectancy)]<-mean(data1$Life_expectancy,na.rm=T)
data1$Adult_Mortality[is.na(data1$Adult_Mortality)]<-mean(data1$Adult_Mortality,na.rm=T)
data1$Alcohol[is.na(data1$Alcohol)]<-mean(data1$Alcohol,na.rm=T)
data1$Hepatitis_B[is.na(data1$Hepatitis_B)]<-mean(data1$Hepatitis_B,na.rm=T)
data1$BMI[is.na(data1$BMI)]<-mean(data1$BMI,na.rm=T)
data1$Polio[is.na(data1$Polio)]<-mean(data1$Polio,na.rm=T)
data1$Total_expenditure[is.na(data1$Total_expenditure)]<-mean(data1$Total_expenditure,na.rm=T)
data1$Diphtheria[is.na(data1$Diphtheria)]<-mean(data1$Diphtheria,na.rm=T)
data1$GDP[is.na(data1$GDP)]<-mean(data1$GDP,na.rm=T)
data1$Population[is.na(data1$Population)]<-mean(data1$Population,na.rm=T)
data1$Thinness_1_to_19_yrs[is.na(data1$Thinness_1_to_19_yrs)]<-mean(data1$Thinness_1_to_19_yrs,na.rm=T)
data1$Thinness_5_to_9_yrs[is.na(data1$Thinness_5_to_9_yrs)]<-mean(data1$Thinness_5_to_9_yrs,na.rm=T)
data1$Income_composition_of_resources[is.na(data1$Income_composition_of_resources)]<-mean(data1$Income_composition_of_resources,na.rm=T)
data1$Schooling[is.na(data1$Schooling)]<-mean(data1$Schooling,na.rm=T)
#check for any missing value column wise
colSums(is.na(data1))
```

```
                     Country                         Year                       Status
                           0                            0                            0
             Life_expectancy               Adult_Mortality                Infant_deaths
                           0                            0                            0
                     Alcohol       Percentage_expenditure                  Hepatitis_B
                           0                            0                            0
                    Measles                          BMI                Under_5_deaths
                           0                            0                            0
                       Polio             Total_expenditure                   Diphtheria
                           0                            0                            0
                    HIV_AIDS                          GDP                   Population
                           0                            0                            0
       Thinness_1_to_19_yrs          Thinness_5_to_9_yrs Income_composition_of_resources
                           0                            0                            0
                  Schooling
                           0
```

Hide

```
#Analyze data
#display dimensions of data
dim(data1)
```

```
[1] 2938   22
```

Hide

```
#list type of each variable
sapply(data1,class)
```

```
                     Country                         Year                       Status
                    "factor"                    "integer"                     "factor"
             Life_expectancy               Adult_Mortality                Infant_deaths
                   "numeric"                    "numeric"                    "integer"
                     Alcohol       Percentage_expenditure                  Hepatitis_B
                   "numeric"                    "numeric"                    "numeric"
                    Measles                          BMI                Under_5_deaths
                   "integer"                    "numeric"                    "integer"
                       Polio             Total_expenditure                   Diphtheria
                   "numeric"                    "numeric"                    "numeric"
                    HIV_AIDS                          GDP                   Population
                   "numeric"                    "numeric"                    "numeric"
       Thinness_1_to_19_yrs          Thinness_5_to_9_yrs Income_composition_of_resources
                   "numeric"                    "numeric"                    "numeric"
                  Schooling
                   "numeric"
```

Hide

```
#checking distribution of variables
table(data1$Status)
```

```
 Developed Developing
       512       2426
```

Hide

```
table(data1$Year)
```

```
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
 183  183  183  183  183  183  183  183  183  183  183  183  183  193  183  183
```

Hide

```
head(table(data1$Country))
```

```
       Afghanistan             Albania             Algeria             Angola Antigua and Barb
uda
                16                  16                  16                  16
16
         Argentina
                16
```

Hide

```
#check correlation between different variables
cor(data1[,4:22])
```

```
                                Life_expectancy Adult_Mortality Infant_deaths     Alcohol
Life_expectancy                       1.0000000     -0.69635931   -0.19653500  0.39159834
Adult_Mortality                      -0.6963593      1.00000000    0.07874713 -0.19040781
Infant_deaths                        -0.1965350      0.07874713    1.00000000 -0.11381227
Alcohol                               0.3915983     -0.19040781   -0.11381227  1.00000000
Percentage_expenditure                0.3817912     -0.24281353   -0.08561222  0.33963429
Hepatitis_B                           0.2037714     -0.13859091   -0.17878339  0.07544715
Measles                              -0.1575738      0.03117404    0.50112834 -0.05105499
BMI                                   0.5592553     -0.38144941   -0.22721997  0.31807030
Under_5_deaths                       -0.2225030      0.09413509    0.99662888 -0.11077713
Polio                                 0.4615738     -0.27269358   -0.17067376  0.21374404
Total_expenditure                     0.2079806     -0.11087454   -0.12656412  0.29489812
Diphtheria                            0.4754184     -0.27301389   -0.17515631  0.21524194
HIV_AIDS                             -0.5564568      0.52372692    0.02523132 -0.04864971
GDP                                   0.4304930     -0.27705292   -0.10710904  0.31859116
Population                           -0.0196377     -0.01250145    0.54852167 -0.03076466
Thinness_1_to_19_yrs                 -0.4721619      0.29986267    0.46559015 -0.41694557
Thinness_5_to_9_yrs                  -0.4666292      0.30536641    0.47122795 -0.40588068
Income_composition_of_resources       0.6924828     -0.44006205   -0.14366278  0.41609923
Schooling                             0.7150663     -0.43510845   -0.19175731  0.49754628
                                Percentage_expenditure Hepatitis_B     Measles        BMI Under_
5_deaths
Life_expectancy                             0.38179117  0.20377144 -0.15757382  0.5592553    -0.
22250302
Adult_Mortality                            -0.24281353 -0.13859091  0.03117404 -0.3814494     0.
09413509
Infant_deaths                              -0.08561222 -0.17878339  0.50112834 -0.2272200     0.
99662888
Alcohol                                     0.33963429  0.07544715 -0.05105499  0.3180703    -0.
11077713
Percentage_expenditure                      1.00000000  0.01167932 -0.05659568  0.2285372    -0.
08785231
Hepatitis_B                                 0.01167932  1.00000000 -0.09031694  0.1349285    -0.
18441262
Measles                                    -0.05659568 -0.09031694  1.00000000 -0.1759253     0.
50780871
BMI                                         0.22853723  0.13492851 -0.17592529  1.0000000    -0.
23758588
Under_5_deaths                             -0.08785231 -0.18441262  0.50780871 -0.2375859     1.
00000000
Polio                                       0.14720343  0.40851923 -0.13614628  0.2821559    -0.
18870311
Total_expenditure                           0.17341415  0.05008430 -0.10456872  0.2318144    -0.
12826939
Diphtheria                                  0.14356978  0.49995767 -0.14186137  0.2810588    -0.
19565055
HIV_AIDS                                   -0.09785682 -0.10240544  0.03089872 -0.2435476     0.
03806151
GDP                                         0.88814032  0.06231758 -0.06805959  0.2766447    -0.
11064032
Population                                 -0.02464822 -0.10981088  0.23624988 -0.0632376     0.
53586402
Thinness_1_to_19_yrs                       -0.25119000 -0.10514361  0.22474217 -0.5320247     0.
```

```
46762640
Thinness_5_to_9_yrs                         -0.25272486 -0.10833424  0.22100716 -0.5389106     0.
47209862
Income_composition_of_resources              0.38037355  0.15099204 -0.11576407  0.4798374    -0.
16153341
Schooling                                    0.38810498  0.17175483 -0.12260854  0.5081055    -0.
20711142
                                     Polio Total_expenditure  Diphtheria    HIV_AIDS         GDP
   Population
Life_expectancy                  0.4615738        0.20798062  0.47541838 -0.55645682  0.43049302
-0.019637702
Adult_Mortality                 -0.2726936       -0.11087454 -0.27301389  0.52372692 -0.27705292
-0.012501453
Infant_deaths                   -0.1706738       -0.12656412 -0.17515631  0.02523132 -0.10710904
  0.548521669
Alcohol                          0.2137440        0.29489812  0.21524194 -0.04864971  0.31859116
-0.030764657
Percentage_expenditure           0.1472034        0.17341415  0.14356978 -0.09785682  0.88814032
-0.024648218
Hepatitis_B                      0.4085192        0.05008430  0.49995767 -0.10240544  0.06231758
-0.109810878
Measles                         -0.1361463       -0.10456872 -0.14186137  0.03089872 -0.06805959
  0.236249877
BMI                              0.2821559        0.23181440  0.28105881 -0.24354758  0.27664474
-0.063237604
Under_5_deaths                  -0.1887031       -0.12826939 -0.19565055  0.03806151 -0.11064032
  0.535864022
Polio                            1.0000000        0.13012949  0.67355332 -0.15948864  0.19398031
-0.034881799
Total_expenditure                0.1301295        1.00000000  0.14559660 -0.00138272  0.12146709
-0.066698201
Diphtheria                       0.6735533        0.14559660  1.00000000 -0.16478684  0.18279456
-0.025457689
HIV_AIDS                        -0.1594886       -0.00138272 -0.16478684  1.00000000 -0.13451379
-0.027318431
GDP                              0.1939803        0.12146709  0.18279456 -0.13451379  1.00000000
-0.025611731
Population                      -0.0348818       -0.06669820 -0.02545769 -0.02731843 -0.02561173
  1.000000000
Thinness_1_to_19_yrs            -0.2199376       -0.26872376 -0.22781986  0.20392213 -0.26774463
  0.236117395
Thinness_5_to_9_yrs             -0.2207101       -0.27524000 -0.22110531  0.20713956 -0.27239955
  0.233940899
Income_composition_of_resources  0.3553976        0.14909473  0.37172915 -0.24745353  0.44031708
-0.007951445
Schooling                        0.3858316        0.21831013  0.38994380 -0.21861975  0.42948916
-0.029464903
                                Thinness_1_to_19_yrs Thinness_5_to_9_yrs Income_composition_of_r
esources
Life_expectancy                           -0.4721619          -0.4666292                     0.6
92482805
Adult_Mortality                            0.2998627           0.3053664                    -0.4
40062048
Infant_deaths                              0.4655902           0.4712279                    -0.1
```

```
43662780
Alcohol                              -0.4169456          -0.4058807                    0.4
16099225
Percentage_expenditure               -0.2511900          -0.2527249                    0.3
80373551
Hepatitis_B                          -0.1051436          -0.1083342                    0.1
50992044
Measles                               0.2247422           0.2210072                   -0.1
15764074
BMI                                  -0.5320247          -0.5389106                    0.4
79837369
Under_5_deaths                        0.4676264           0.4720986                   -0.1
61533413
Polio                                -0.2199376          -0.2207101                    0.3
55397638
Total_expenditure                    -0.2687238          -0.2752400                    0.1
49094726
Diphtheria                           -0.2278199          -0.2211053                    0.3
71729147
HIV_AIDS                              0.2039221           0.2071396                   -0.2
47453534
GDP                                  -0.2677446          -0.2723995                    0.4
40317075
Population                            0.2361174           0.2339409                   -0.0
07951445
Thinness_1_to_19_yrs                  1.0000000           0.9391020                   -0.4
06661733
Thinness_5_to_9_yrs                   0.9391020           1.0000000                   -0.3
95778501
Income_composition_of_resources      -0.4066617          -0.3957785                    1.0
00000000
Schooling                            -0.4461401          -0.4357772                    0.7
96207053
                                Schooling
Life_expectancy                 0.7150663
Adult_Mortality                -0.4351085
Infant_deaths                  -0.1917573
Alcohol                         0.4975463
Percentage_expenditure          0.3881050
Hepatitis_B                     0.1717548
Measles                        -0.1226085
BMI                             0.5081055
Under_5_deaths                 -0.2071114
Polio                           0.3858316
Total_expenditure               0.2183101
Diphtheria                      0.3899438
HIV_AIDS                       -0.2186198
GDP                             0.4294892
Population                     -0.0294649
Thinness_1_to_19_yrs           -0.4461401
Thinness_5_to_9_yrs            -0.4357772
Income_composition_of_resources 0.7962071
Schooling                       1.0000000
```

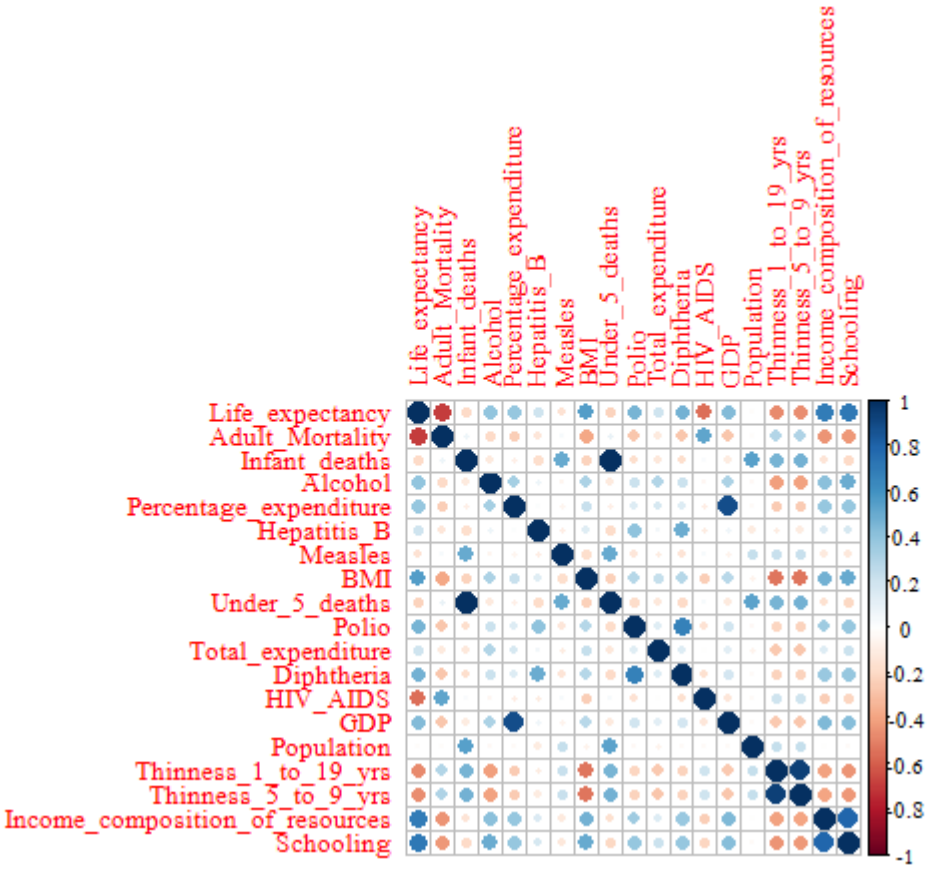
Hide

```
library(corrplot)
```

```
corrplot 0.84 loaded
```

Hide

```
corrplot(cor(data1[,4:22]))
```

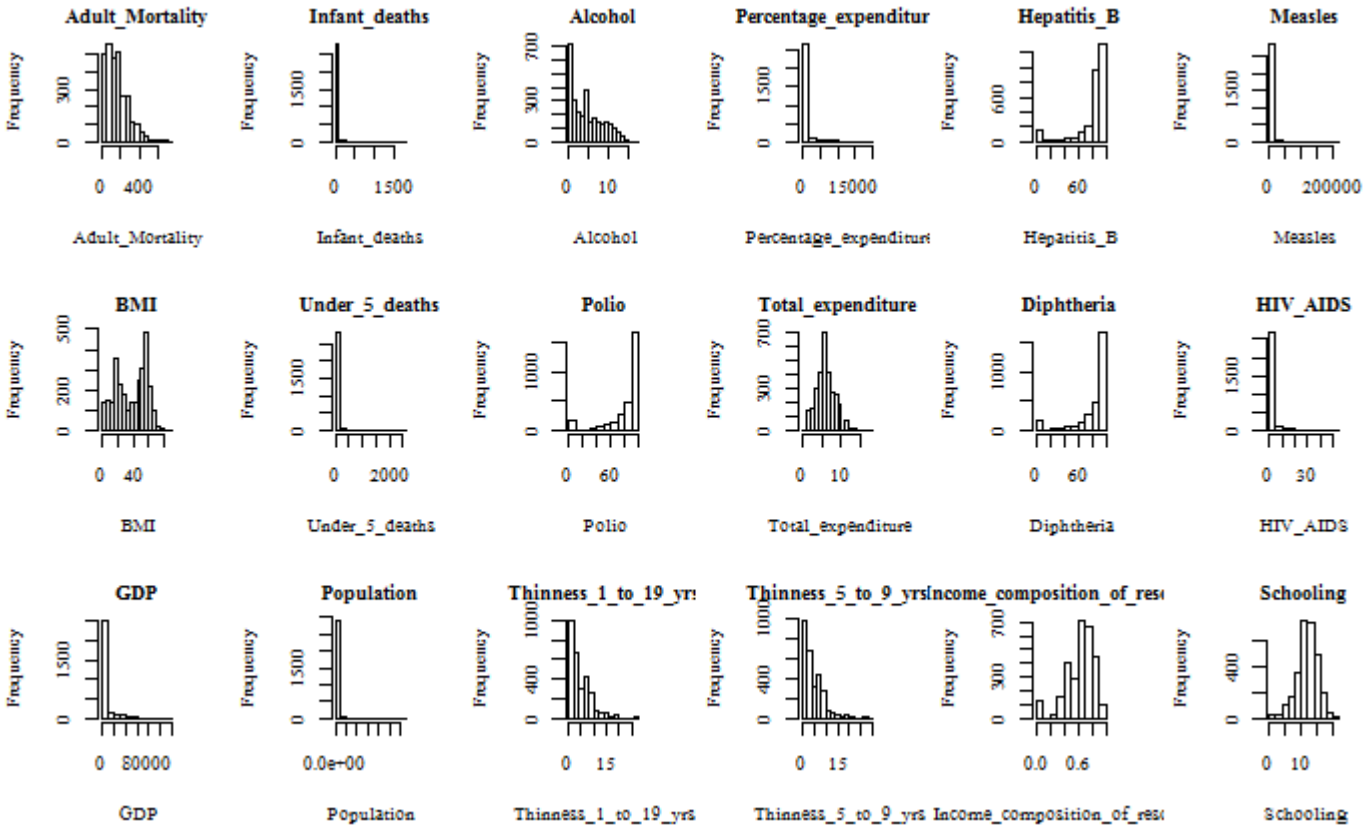


Hide

```
#check distribution of dependent variable i.e. Life.expectancy-histogram
hist(data1[,4], main="Histogram of Life expectancy")
#check distribution of independent variables-density plot
par(mfrow=c(3,6))
```

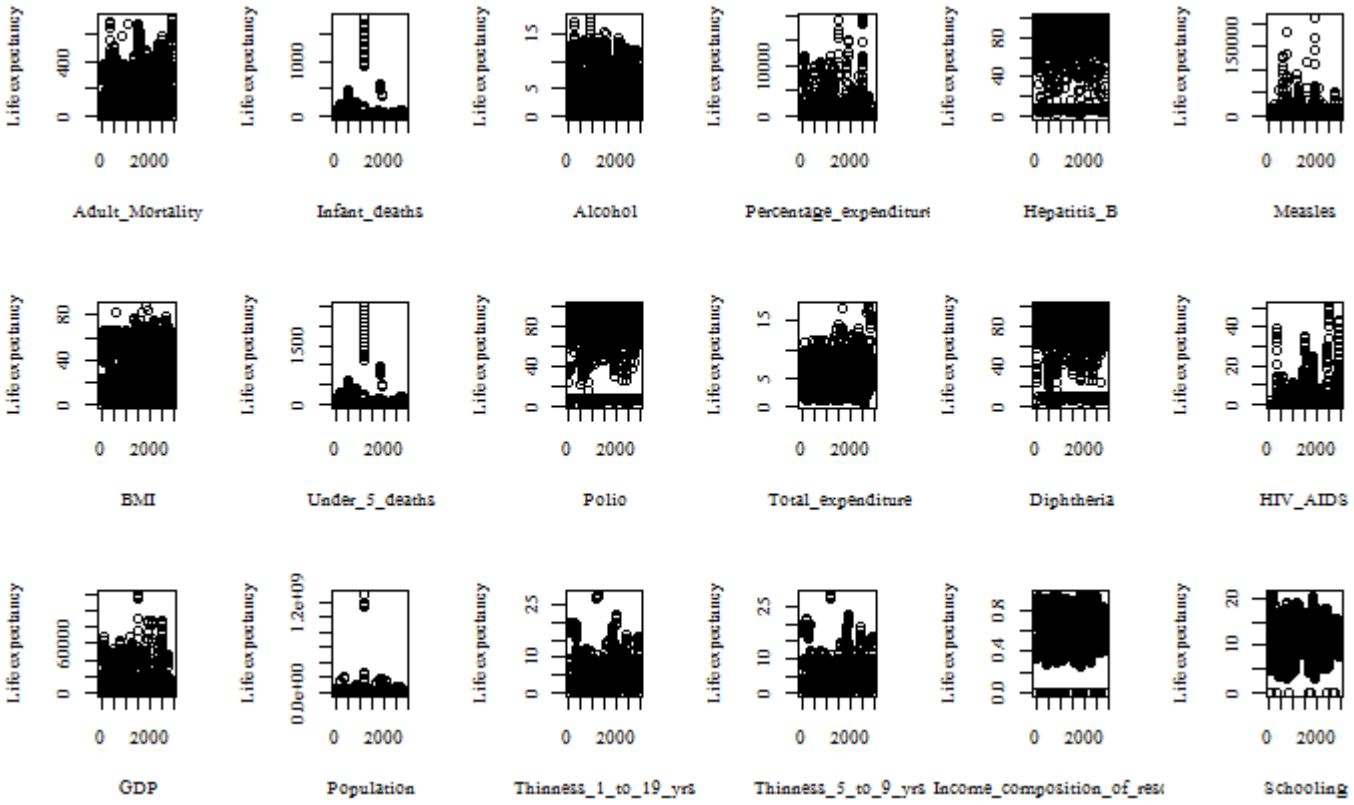
Hide

```
for(i in 5:22){
  hist(data1[,i],main=names(data1)[i],ylab = "Frequency",
       xlab=paste("",names(data1)[i]))
}
```



Hide

```
#Scatterplot of independent vriables against Life expectancy
par(mfrow=c(3,6))
for(i in 5:22){
  plot(data1[,i],data1$Life.expectancy,xlab=paste("",names(data1)[i]),
       ylab="Life expectancy")
}
```



Hide

```
#Remove Year variable
data2<-data1[,c(-2)]
dim(data2)
```

```
[1] 2938   21
```

Hide

```
str(data2)
```

```
'data.frame':   2938 obs. of  21 variables:
 $ Country                        : Factor w/ 193 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1
...
 $ Status                         : Factor w/ 2 levels "Developed","Developing": 2 2 2 2 2 2 2 2
2 2 ...
 $ Life_expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
 $ Adult_Mortality                : num  263 271 268 272 275 279 281 287 295 295 ...
 $ Infant_deaths                  : int  62 64 66 69 71 74 77 80 82 84 ...
 $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
 $ Percentage_expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
 $ Hepatitis_B                    : num  65 62 64 67 68 66 63 64 63 64 ...
 $ Measles                        : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
 $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
 $ Under_5_deaths                 : int  83 86 89 93 97 102 106 110 113 116 ...
 $ Polio                          : num  6 58 62 67 68 66 63 64 63 58 ...
 $ Total_expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
 $ Diphtheria                     : num  65 62 64 67 68 66 63 64 63 58 ...
 $ HIV_AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
 $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
 $ Population                     : num  33736494 327582 31731688 3696958 2978599 ...
 $ Thinness_1_to_19_yrs           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
 $ Thinness_5_to_9_yrs            : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
 $ Income_composition_of_resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.
405 ...
 $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
```

Hide

```
#Do one hot encoding for Status variable
for(i in unique(data2$Status)){
  data2[paste("Status",i)]<-ifelse(data2$Status==i,1,0)
}
dim(data2)
```

```
[1] 2938   23
```

Hide

```
#Remove Status variable
data2<-data2[,c(-2)]
str(data2)
```

```
'data.frame':   2938 obs. of  22 variables:
 $ Country                        : Factor w/ 193 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1
...
 $ Life_expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
 $ Adult_Mortality                : num  263 271 268 272 275 279 281 287 295 295 ...
 $ Infant_deaths                  : int  62 64 66 69 71 74 77 80 82 84 ...
 $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
 $ Percentage_expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
 $ Hepatitis_B                    : num  65 62 64 67 68 66 63 64 63 64 ...
 $ Measles                        : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
 $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
 $ Under_5_deaths                 : int  83 86 89 93 97 102 106 110 113 116 ...
 $ Polio                          : num  6 58 62 67 68 66 63 64 63 58 ...
 $ Total_expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
 $ Diphtheria                     : num  65 62 64 67 68 66 63 64 63 58 ...
 $ HIV_AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
 $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
 $ Population                     : num  33736494 327582 31731688 3696958 2978599 ...
 $ Thinness_1_to_19_yrs           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
 $ Thinness_5_to_9_yrs            : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
 $ Income_composition_of_resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.
405 ...
 $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
 $ Status Developing              : num  1 1 1 1 1 1 1 1 1 1 ...
 $ Status Developed               : num  0 0 0 0 0 0 0 0 0 0 ...
```

Hide

```
#find mean of all variables country wise
library(dplyr)
a<-data2 %>%
  group_by(Country) %>%
  summarise_all(funs(mean))
data3<-a
#Remove Country variable
data4<-data3[,c(-1)]
library(caret)
#normalizing the variables
preproc1 <- preProcess(data4, method = c("range"))
#Applying normalization to data
library(RANN)
data_processed1 <- predict(preproc1, data4)
#Splitting data into train and test set
train<-data_processed1[1:150,]
test<-data_processed1[151:193,]
dim(train)
```

```
[1] 150  21
```

Hide

```
dim(test)
```

```
[1] 43 21
```

Hide

```
#Fit multiple linear regression model, excluding Country variable
model1<-lm(Life_expectancy~.,data=train)
#check summary of model
summary(model1)
```

```

Call:
lm(formula = Life_expectancy ~ ., data = train)

Residuals:
      Min        1Q    Median        3Q       Max
-0.189004 -0.039161  0.001239  0.038598  0.182408

Coefficients: (1 not defined because of singularities)
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     0.389597   0.062946   6.189 7.32e-09 ***
Adult_Mortality                -0.714300   0.063602 -11.231  < 2e-16 ***
Infant_deaths                   1.459105   0.940294   1.552 0.123152
Alcohol                         0.045049   0.028988   1.554 0.122594
Percentage_expenditure          0.148582   0.115726   1.284 0.201457
Hepatitis_B                    -0.088361   0.058309  -1.515 0.132096
Measles                         0.002768   0.074742   0.037 0.970513
BMI                             0.019884   0.046792   0.425 0.671573
Under_5_deaths                 -1.692587   0.862962  -1.961 0.051973 .
Polio                          -0.063351   0.107131  -0.591 0.555317
Total_expenditure               0.090151   0.058342   1.545 0.124725
Diphtheria                      0.344689   0.098005   3.517 0.000602 ***
HIV_AIDS                       -0.158395   0.084702  -1.870 0.063731 .
GDP                            -0.030288   0.093877  -0.323 0.747489
Population                      0.183063   0.188098   0.973 0.332246
Thinness_1_to_19_yrs            0.022758   0.208090   0.109 0.913080
Thinness_5_to_9_yrs            -0.025359   0.212091  -0.120 0.905010
Income_composition_of_resources  0.136092   0.054850   2.481 0.014372 *
Schooling                       0.226947   0.067949   3.340 0.001094 **
`Status Developing`             0.014973   0.022368   0.669 0.504434
`Status Developed`                    NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06591 on 130 degrees of freedom
Multiple R-squared:  0.9338,    Adjusted R-squared:  0.9242
F-statistic: 96.56 on 19 and 130 DF,  p-value: < 2.2e-16
```

Hide

```
#apply model on test data
testing<-predict(model1,test)
```

```
prediction from a rank-deficient fit may be misleading
```

Hide

```
#comapre actual and predicted values of Life_expectancy
head(cbind(test$Life_expectancy,testing))
```

```
             testing
1 0.7644132 0.7685712
2 0.7210021 0.6878274
3 0.0000000 0.2480587
4 0.9708305 0.8480835
5 0.7862045 0.8046964
6 0.9229581 0.9254026
```

Hide

```
#data_processed1
#Fit model by removing Status variable because its not meaningful to add it
#Removing Status variables
data5<-data4[,c(-20,-21)]
#normalizing variables
preproc2 <- preProcess(data5, method = c("range"))
#Applying normalization to data
library(RANN)
data_processed2 <- predict(preproc2, data5)
str(data_processed2)
```

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':   193 obs. of  19 variables:
 $ Life_expectancy                : num  0.3317 0.7974 0.7551 0.0798 0.7946 ...
 $ Adult_Mortality                : num  0.4711 0.0495 0.1683 0.5831 0.2047 ...
 $ Infant_deaths                  : num  0.057255 0.000503 0.014863 0.06128 0 ...
 $ Alcohol                        : num  0.000338 0.374185 0.050956 0.437659 0.597782 ...
 $ Percentage_expenditure         : num  0.00357 0.01972 0.0241 0.01042 0.10218 ...
 $ Hepatitis_B                    : num  0.622 0.989 0.777 0.735 0.98 ...
 $ Measles                        : num  0.03587 0.00081 0.02952 0.05408 0 ...
 $ BMI                            : num  0.126 0.534 0.53 0.156 0.405 ...
 $ Under_5_deaths                 : num  0.059345 0.000517 0.012966 0.073172 0 ...
 $ Polio                          : num  0.438 0.99 0.919 0.412 0.977 ...
 $ Total_expenditure              : num  0.414 0.263 0.181 0.139 0.192 ...
 $ Diphtheria                     : num  0.481 0.99 0.921 0.43 0.992 ...
 $ HIV_AIDS                       : num  0 0 0 0.069077 0.000761 ...
 $ GDP                            : num  0.00356 0.03466 0.04738 0.03213 0.16816 ...
 $ Population                     : num  0.02366 0.00165 0.05137 0.02407 0.03026 ...
 $ Thinness_1_to_19_yrs           : num  0.6104 0.0562 0.222 0.2257 0.1231 ...
 $ Thinness_5_to_9_yrs            : num  0.556 0.0575 0.211 0.2359 0.1176 ...
 $ Income_composition_of_resources: num  0.355 0.723 0.704 0.408 0.446 ...
 $ Schooling                      : num  0.41 0.606 0.634 0.401 0.441 ...
```

Hide

```
#Splitting data into train and test set
train1<-data_processed2[1:150,]
test1<-data_processed2[151:193,]
#check dimensions of train and test set
dim(train1)
```

```
[1] 150  19
```

Hide

```
dim(test1)
```

```
[1] 43 19
```

Hide

```
#Fit multiple linear regression model, excluding Country variable
model2<-lm(Life_expectancy~.,data=train1)
#check summary of model
summary(model2)
```

```
Call:
lm(formula = Life_expectancy ~ ., data = train1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.189048 -0.042524  0.000253  0.038878  0.181167

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      0.409691   0.055210   7.421 1.31e-11 ***
Adult_Mortality                 -0.712452   0.063408 -11.236  < 2e-16 ***
Infant_deaths                    1.426686   0.937065   1.523 0.130293
Alcohol                          0.037534   0.026669   1.407 0.161673
Percentage_expenditure           0.125364   0.110173   1.138 0.257248
Hepatitis_B                     -0.093854   0.057607  -1.629 0.105670
Measles                          0.004329   0.074548   0.058 0.953778
BMI                              0.017764   0.046586   0.381 0.703586
Under_5_deaths                  -1.670435   0.860509  -1.941 0.054379 .
Polio                           -0.065465   0.106858  -0.613 0.541180
Total_expenditure                0.087794   0.058113   1.511 0.133259
Diphtheria                       0.350529   0.097410   3.598 0.000453 ***
HIV_AIDS                        -0.158586   0.084523  -1.876 0.062847 .
GDP                             -0.022813   0.093014  -0.245 0.806638
Population                       0.193621   0.187040   1.035 0.302490
Thinness_1_to_19_yrs             0.025453   0.207612   0.123 0.902613
Thinness_5_to_9_yrs             -0.031617   0.211438  -0.150 0.881364
Income_composition_of_resources  0.133129   0.054556   2.440 0.016015 *
Schooling                        0.227749   0.067796   3.359 0.001024 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06577 on 131 degrees of freedom
Multiple R-squared:  0.9336,    Adjusted R-squared:  0.9245
F-statistic: 102.3 on 18 and 131 DF,  p-value: < 2.2e-16
```

Hide

```
#apply model on test data
testing1<-predict(model2,test1)
#comapre actual and predicted values of Life_expectancy
head(cbind(test$Life.expectancy,testing1))
```

```
Unknown or uninitialised column: 'Life.expectancy'.
```

```
   testing1
1 0.7633643
2 0.6845834
3 0.2477425
4 0.8614448
5 0.8146803
6 0.9321225
```

Hide

```
#combine Counttry name, actual and predicted values of life expectancy
cbind(data3[151:193,"Country"],test$Life_expectancy,testing1)
```

	Country	test$Life_expectancy	testing1
1	Serbia	0.7644132	0.76336433
2	Seychelles	0.7210021	0.68458338
3	Sierra Leone	0.0000000	0.24774249
4	Singapore	0.9708305	0.86144481
5	Slovakia	0.7862045	0.81468035
6	Slovenia	0.9229581	0.93212251
7	Solomon Islands	0.5929993	0.56783949
8	Somalia	0.1978380	0.22893986
9	South Africa	0.3126287	0.22535696
10	South Sudan	0.2131091	0.12477859
11	Spain	0.9871311	0.95781699
12	Sri Lanka	0.7491421	0.70811909
13	Sudan	0.4313658	0.38796168
14	Suriname	0.6580302	0.61462038
15	Swaziland	0.1431023	0.21469339
16	Sweden	0.9994852	0.99991899
17	Switzerland	0.9943377	1.02813451
18	Syrian Arab Republic	0.6791352	0.53346041
19	Tajikistan	0.5640014	0.59934982
20	Thailand	0.7403912	0.69396909
21	The former Yugoslav republic of Macedonia	0.7687028	0.79270102
22	Timor-Leste	0.5118394	0.50239374
23	Togo	0.2896362	0.33277699
24	Tonga	0.7252917	0.67585316
25	Trinidad and Tobago	0.6851407	0.62605139
26	Tunisia	0.7753946	0.89345093
27	Turkey	0.7632121	0.73997445
28	Turkmenistan	0.5080645	0.48213132
29	Tuvalu	0.6345211	0.39532148
30	Uganda	0.2633837	0.32049306
31	Ukraine	0.6540837	0.62107111
32	United Arab Emirates	0.8122855	0.83153111
33	United Kingdom of Great Britain and Northern Ireland	0.9521277	0.83305342
34	United Republic of Tanzania	0.2716198	0.34290518
35	United States of America	0.8771448	0.89621611
36	Uruguay	0.8225806	0.78568305
37	Uzbekistan	0.6017502	0.62441663
38	Vanuatu	0.6938916	0.53951780
39	Venezuela (Bolivarian Republic of)	0.7487989	0.63683861
40	Viet Nam	0.7868909	0.68739612
41	Yemen	0.4873027	0.48024026
42	Zambia	0.2139671	0.25923745
43	Zimbabwe	0.1201098	0.04488484

Hide

```
#denormalized<-(data_processed1$Life.expectancy)*(max(data4$Life.expectancy)-min(data4$Life.expe
ctancy))+min(data4$Life.expectancy)
##Denormalizing the life expectancy values so that we can compare predicted values with original
 values
normalize<-test$Life_expectancy
#denormalization formula for original values
denormalize<-function(x){
  fun<- normalize*(max(x)-min(x))+min(x)
  return (fun)
}
normalize1<-testing1
#denormalization formula for predicted values
denormalize1<-function(x){
  fun1<- normalize1*(max(x)-min(x))+min(x)
  return (fun1)
}
#Denormalized values of Actual and predicted life expectancy
denormalize_original<-denormalize(data4$Life_expectancy)
denormalize_original
```

```
 [1] 73.95625 72.37500 46.11250 81.47500 74.75000 79.73125 67.71250 53.31875 57.50000 53.87500 8
2.06875
[12] 73.40000 61.82500 70.08125 51.32500 82.51875 82.33125 70.85000 66.65625 73.08125 74.11250 6
4.75625
[23] 56.66250 72.53125 71.06875 74.35625 73.91250 64.61875 69.22493 55.70625 69.93750 75.70000 8
0.79375
[34] 56.00625 78.06250 76.07500 68.03125 71.38750 73.38750 74.77500 63.86250 53.90625 50.48750
```

Hide

```
denormalize_predicted<-denormalize1(data4$Life_expectancy)
denormalize_predicted
```

```
       1        2        3        4        5        6        7        8        9       10
11       12
73.91805 71.04845 55.13652 77.49063 75.78723 80.06506 66.79605 54.45163 54.32113 50.65756 81.000
98 71.90574
      13       14       15       16       17       18       19       20       21       22
23       24
60.24400 68.50005 53.93271 82.53455 83.56230 65.54380 67.94382 71.39032 74.98663 64.41219 58.233
90 70.73045
      25       26       27       28       29       30       31       32       33       34
35       36
68.91642 78.65645 73.06607 63.67413 60.51209 57.78646 68.73502 76.40102 76.45647 58.60282 78.757
17 74.73101
      37       38       39       40       41       42       43
68.85688 65.76444 69.30935 71.15090 63.60525 55.55522 47.74743
```

Hide

```
#Combine denormalized values of actual and predicted life expectancy
cbind(data3[151:193,"Country"],denormalize_original,denormalize_predicted)
```

	Country	denormalize_original	denormalized_predicted
1	Serbia	73.95625	73.91805
2	Seychelles	72.37500	71.04845
3	Sierra Leone	46.11250	55.13652
4	Singapore	81.47500	77.49063
5	Slovakia	74.75000	75.78723
6	Slovenia	79.73125	80.06506
7	Solomon Islands	67.71250	66.79605
8	Somalia	53.31875	54.45163
9	South Africa	57.50000	54.32113
10	South Sudan	53.87500	50.65756
11	Spain	82.06875	81.00098
12	Sri Lanka	73.40000	71.90574
13	Sudan	61.82500	60.24400
14	Suriname	70.08125	68.50005
15	Swaziland	51.32500	53.93271
16	Sweden	82.51875	82.53455
17	Switzerland	82.33125	83.56230
18	Syrian Arab Republic	70.85000	65.54380
19	Tajikistan	66.65625	67.94382
20	Thailand	73.08125	71.39032
21	The former Yugoslav republic of Macedonia	74.11250	74.98663
22	Timor-Leste	64.75625	64.41219
23	Togo	56.66250	58.23390
24	Tonga	72.53125	70.73045
25	Trinidad and Tobago	71.06875	68.91642
26	Tunisia	74.35625	78.65626

```
45
27                                          Turkey             73.91250              73.066
07
28                                      Turkmenistan           64.61875              63.674
13
29                                          Tuvalu             69.22493              60.512
09
30                                          Uganda             55.70625              57.786
46
31                                         Ukraine             69.93750              68.735
02
32                                 United Arab Emirates        75.70000              76.401
02
33 United Kingdom of Great Britain and Northern Ireland        80.79375              76.456
47
34                           United Republic of Tanzania       56.00625              58.602
82
35                              United States of America       78.06250              78.757
17
36                                         Uruguay             76.07500              74.731
01
37                                       Uzbekistan            68.03125              68.856
88
38                                         Vanuatu             71.38750              65.764
44
39                    Venezuela (Bolivarian Republic of)       73.38750              69.309
35
40                                       Viet Nam              74.77500              71.150
90
41                                          Yemen              63.86250              63.605
25
42                                          Zambia             53.90625              55.555
22
43                                         Zimbabwe            50.48750              47.747
43
```

Hide

```
#summary of model2
summary(model2)
```

```
Call:
lm(formula = Life_expectancy ~ ., data = train1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.189048 -0.042524  0.000253  0.038878  0.181167

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      0.409691   0.055210   7.421 1.31e-11 ***
Adult_Mortality                 -0.712452   0.063408 -11.236  < 2e-16 ***
Infant_deaths                    1.426686   0.937065   1.523 0.130293
Alcohol                          0.037534   0.026669   1.407 0.161673
Percentage_expenditure           0.125364   0.110173   1.138 0.257248
Hepatitis_B                     -0.093854   0.057607  -1.629 0.105670
Measles                          0.004329   0.074548   0.058 0.953778
BMI                              0.017764   0.046586   0.381 0.703586
Under_5_deaths                  -1.670435   0.860509  -1.941 0.054379 .
Polio                           -0.065465   0.106858  -0.613 0.541180
Total_expenditure                0.087794   0.058113   1.511 0.133259
Diphtheria                       0.350529   0.097410   3.598 0.000453 ***
HIV_AIDS                        -0.158586   0.084523  -1.876 0.062847 .
GDP                             -0.022813   0.093014  -0.245 0.806638
Population                       0.193621   0.187040   1.035 0.302490
Thinness_1_to_19_yrs             0.025453   0.207612   0.123 0.902613
Thinness_5_to_9_yrs             -0.031617   0.211438  -0.150 0.881364
Income_composition_of_resources  0.133129   0.054556   2.440 0.016015 *
Schooling                        0.227749   0.067796   3.359 0.001024 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06577 on 131 degrees of freedom
Multiple R-squared:  0.9336,    Adjusted R-squared:  0.9245
F-statistic: 102.3 on 18 and 131 DF,  p-value: < 2.2e-16
```

Hide

```
#predicted values of Life expectancy by model
head(fitted(model2))
```

```
        1         2         3         4         5         6
0.3204883 0.8459385 0.7553932 0.1841300 0.6763499 0.8563646
```

Hide

```
#difference between actual and predicted values of life expectancy by model
head(residuals(model2))
```

```
           1            2            3            4            5            6
 0.011186398 -0.048580913 -0.000245665 -0.104342791  0.118262339 -0.059006988
```

The multiple R-squared (0.9336) indicates that the model accounts for 93.3 percent of the variance in expectancy. The multiple R-squared is also the correlation between the actual and predicted value.

The residual standard error (0.06) can be thought of as the average error in predicting life expectancy from independent variables using this model.

F-statistic tests whether the predictor variables taken together, predict the response variable.

summary() function provides no information that whether we've satisfied the statistical assumptions underlying the model.

Hide

```
#checking confidence interval of model
confint(model2)
```

```
                                  2.5 %       97.5 %
(Intercept)                    0.30047247  0.518909390
Adult_Mortality               -0.83788825 -0.587014944
Infant_deaths                 -0.42705255  3.280425547
Alcohol                       -0.01522319  0.090291396
Percentage_expenditure        -0.09258503  0.343312442
Hepatitis_B                   -0.20781342  0.020105990
Measles                       -0.14314496  0.151803640
BMI                           -0.07439416  0.109922124
Under_5_deaths                -3.37272625  0.031856347
Polio                         -0.27685631  0.145925996
Total_expenditure             -0.02716661  0.202754754
Diphtheria                     0.15782859  0.543228931
HIV_AIDS                      -0.32579293  0.008621271
GDP                           -0.20681571  0.161190494
Population                    -0.17638893  0.563630817
Thinness_1_to_19_yrs          -0.38525398  0.436159792
Thinness_5_to_9_yrs           -0.44989218  0.386658449
Income_composition_of_resources  0.02520441  0.241054074
Schooling                      0.09363270  0.361864545
```

The results suggest that we can be 95 percent confident that the interval [0.15,0.54] contains the true change in life expectancy for a 1 unit change in Diphtheria.

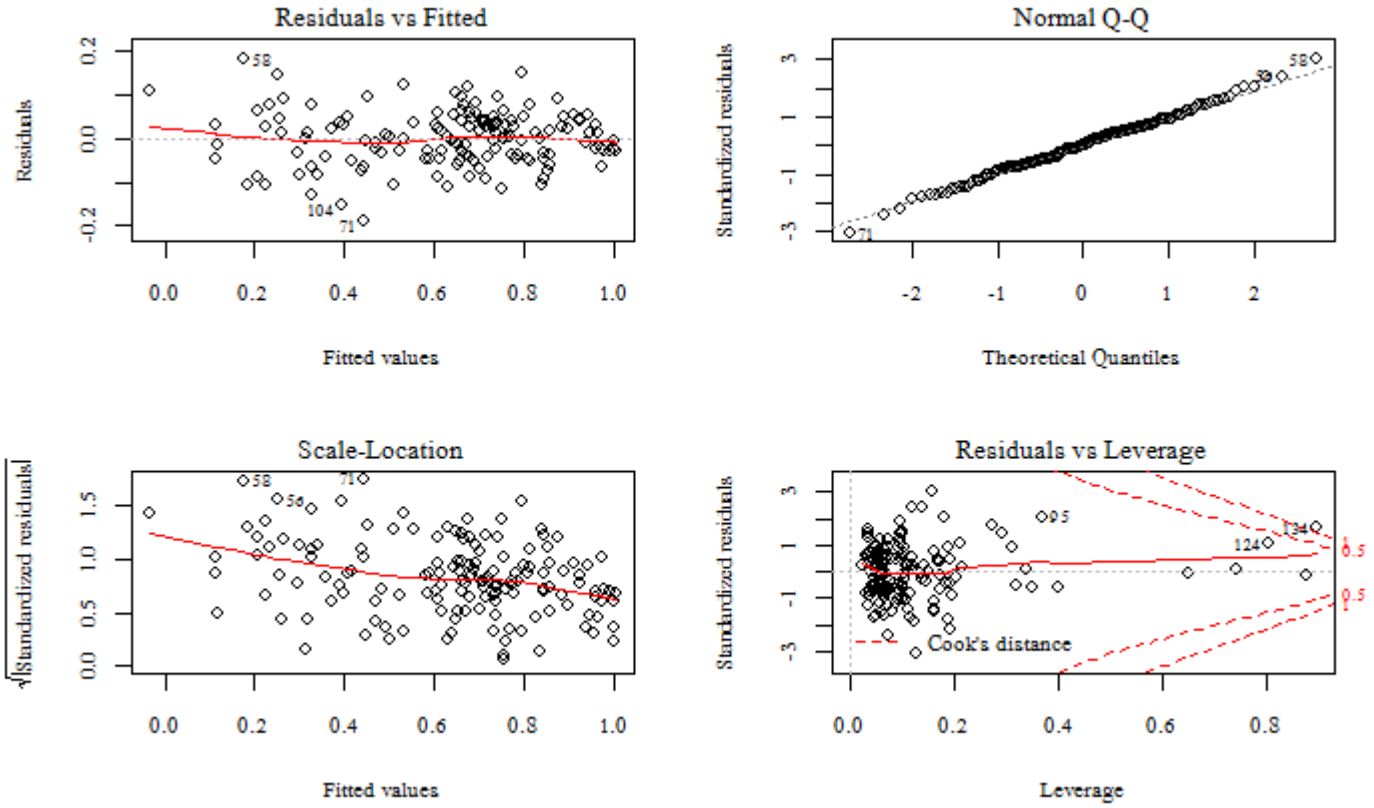
checking whether our model satisfies statistical assumptions:

Terms meaning:

Normality -For fixed values of the independent variables, the dependent variable is normally distributed. Independence(Autocorrelation) - The Yi values are independent of each other. Linearity -The dependent variable is linearly related to the independent variables. Homoscedasticity -The variance of the dependent variable doesn't vary with the levels of theindependent variables. We could call this constant variance, but saying homoscedasticity makes me feel smarter.

Hide

```
#combine 4 plots
par(mfrow=c(2,2))
plot(model2)
```



Normal Q-Q plot (upper right) is a probability plot of the standardized residuals against the values that would be expected under normality.Since we met the normality assumption, the points on this graph fall on the straight 45-degree line.

Since points in the Scale-Location graph (bottom left) has a random band around a horizontal line, homoscedasticity assumption is met.

We can see that in graph 4, residuals vs. leverage, there are few outliers at 124,134 and 95 in predictors value.

To properly interpret the coefficients of the OLS model, we must satisfy a number of statistical assumptions:

The car package provides a number of functions that significantly enhance your ability to fit and evaluate regression models.

gvlma package provides a global test for linear model assumptions.

Hide

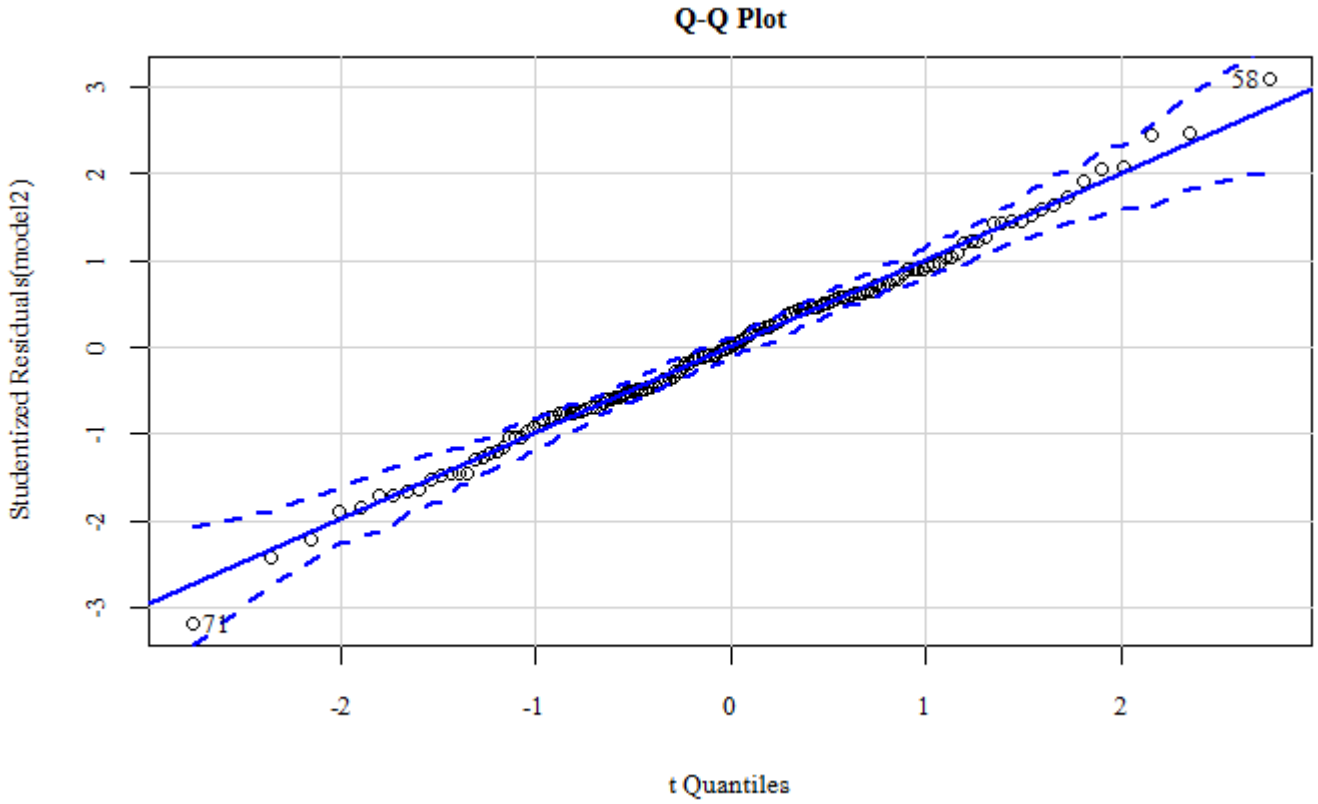
```
library(car)
```

```
Loading required package: carData

Attaching package: <U+393C><U+3E31>car<U+393C><U+3E32>

The following object is masked from <U+393C><U+3E31>package:dplyr<U+393C><U+3E32>:

    recode
```

Hide

```
#NORMALITY

qqPlot(model2, labels=row.names(train), id.method="identify",simulate=TRUE, main="Q-Q Plot")
```

```
[1] 58 71
```



With the exception of points 58 and 71(these are outliers), all the points fall close to the line and are within the confidence envelope, suggesting that we've met the normality assumption fairly well.

Hide

```
#INDEPENDENCE OF ERRORS(AUTOCORRELATION)
durbinWatsonTest(model2)
```

```
 lag Autocorrelation D-W Statistic p-value
   1      0.05732573      1.883105   0.384
 Alternative hypothesis: rho != 0
```

The nonsignificant p-value (p=0.404) suggests a lack of autocorrelation, and conversely an independence of errors. The lag value (1 in this case) indicates that each observation is being compared with the one next to it in the dataset.

Unless we add the option simulate=FALSE, we'll get a slightly different value each time we run the test.

Hide

```
#LINEARITY of predicted values (component plus residual plots or partial residual plots)
crPlots(model2)
```

Nonlinearity in any of these plots suggests that we may not have adequately modeled the functional form of that predictor in the regression. If so, we may need to add curvilinear components such as polynomial terms, transform one or more variables (for example, use log(X) instead of X), or abandon linear regression in favor of some other regression variant.

The component plus residual plots confirm that you've met the linearity assumption. The form of the linear model seems to be appropriate for this dataset.

Hide

```
#Homoscedasticity
ncvTest(model2)
```

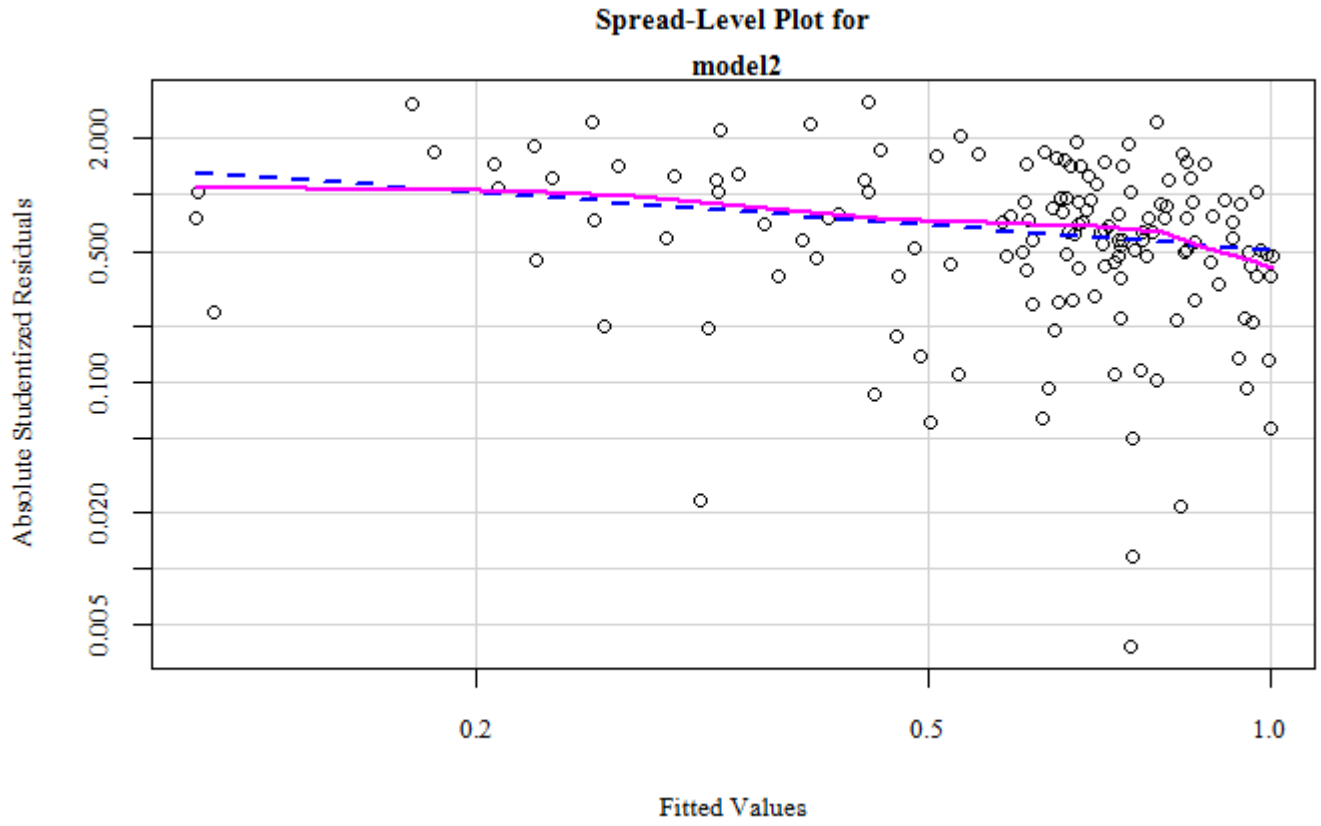
```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 15.71179, Df = 1, p = 7.3763e-05
```

Hide

```
spreadLevelPlot(model2)
```

```

1 negative fitted value removed
```

```

Suggested power transformation:  1.449594
```



The ncvTest() function produces a score test of the hypothesis of constant error variance against the alternative that the error variance changes with the level of the fitted values. A significant result suggests heteroscedasticity (nonconstant error variance).

The spreadLevelPlot() function creates a scatter plot of the absolute standardized residuals versus the fitted values.

In the spread-level plot,the points form a random horizontal band around a horizontal line of best fit. If we'd violated the assumption, we'd expect to see a nonhorizontal line.

suggested power p is that would stabilize the nonconstant error variance. For example, if the plot showed a nonhorizontal trend and the suggested power transformation was 0.5, then using sq root Y rather than Y in the regression equation might lead to a model that satisfies homoscedasticity. If the suggested power was 0, we'd use a log transformation. In the current example, there's no evidence of heteroscedasticity and the suggested power is close to 1 (no transformation required).

Hide

```
library(gvlma)
gvmodel <- gvlma(model2)
summary(gvmodel)
```

```
Call:
lm(formula = Life_expectancy ~ ., data = train1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.189048 -0.042524  0.000253  0.038878  0.181167

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      0.409691   0.055210   7.421 1.31e-11 ***
Adult_Mortality                 -0.712452   0.063408 -11.236  < 2e-16 ***
Infant_deaths                    1.426686   0.937065   1.523 0.130293
Alcohol                          0.037534   0.026669   1.407 0.161673
Percentage_expenditure           0.125364   0.110173   1.138 0.257248
Hepatitis_B                     -0.093854   0.057607  -1.629 0.105670
Measles                          0.004329   0.074548   0.058 0.953778
BMI                              0.017764   0.046586   0.381 0.703586
Under_5_deaths                  -1.670435   0.860509  -1.941 0.054379 .
Polio                           -0.065465   0.106858  -0.613 0.541180
Total_expenditure                0.087794   0.058113   1.511 0.133259
Diphtheria                       0.350529   0.097410   3.598 0.000453 ***
HIV_AIDS                        -0.158586   0.084523  -1.876 0.062847 .
GDP                             -0.022813   0.093014  -0.245 0.806638
Population                       0.193621   0.187040   1.035 0.302490
Thinness_1_to_19_yrs             0.025453   0.207612   0.123 0.902613
Thinness_5_to_9_yrs             -0.031617   0.211438  -0.150 0.881364
Income_composition_of_resources  0.133129   0.054556   2.440 0.016015 *
Schooling                        0.227749   0.067796   3.359 0.001024 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06577 on 131 degrees of freedom
Multiple R-squared:  0.9336,    Adjusted R-squared:  0.9245
F-statistic: 102.3 on 18 and 131 DF,  p-value: < 2.2e-16


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = model2)

                     Value p-value                Decision
Global Stat        2.90828  0.5733 Assumptions acceptable.
Skewness           0.03354  0.8547 Assumptions acceptable.
Kurtosis           0.46565  0.4950 Assumptions acceptable.
Link Function      1.90461  0.1676 Assumptions acceptable.
Heteroscedasticity 0.50449  0.4775 Assumptions acceptable.
```

we can see from the printout (the Global Stat line) that the data meet all the statistical assumptions that go with the OLS regression model (p = 0.573)

gvlma() function performs a global validation of linear model assumptions as well as separate evaluations of skewness, kurtosis, and heteroscedasticity. In other words, it provides a single omnibus (go/no go) test of model assumptions.

Hide

```
#Multicollinearity
vif(model2)
```

```
             Adult_Mortality                 Infant_deaths                       Alcohol
                    3.854441                    263.589659                      2.053672
       Percentage_expenditure                   Hepatitis_B                       Measles
                    8.384375                      2.812541                      3.445196
                         BMI                  Under_5_deaths                         Polio
                    3.484853                    232.309147                     12.238672
           Total_expenditure                    Diphtheria                       HIV_AIDS
                    1.701906                     10.784185                      2.491161
                         GDP                    Population           Thinness_1_to_19_yrs
                    8.328654                      9.354902                     34.437459
       Thinness_5_to_9_yrs Income_composition_of_resources                     Schooling
                   34.646011                      4.876674                      4.072688
```

Hide

```
sqrt(vif(model2))>2
```

```
             Adult_Mortality                 Infant_deaths                       Alcohol
                       FALSE                          TRUE                         FALSE
       Percentage_expenditure                   Hepatitis_B                       Measles
                        TRUE                         FALSE                         FALSE
                         BMI                  Under_5_deaths                         Polio
                       FALSE                          TRUE                          TRUE
           Total_expenditure                    Diphtheria                       HIV_AIDS
                       FALSE                          TRUE                         FALSE
                         GDP                    Population           Thinness_1_to_19_yrs
                        TRUE                          TRUE                          TRUE
       Thinness_5_to_9_yrs Income_composition_of_resources                     Schooling
                        TRUE                          TRUE                          TRUE
```

Multicollinearity can be detected using a statistic called the variance inflation factor (VIF). For any predictor variable, the square root of the VIF indicates the degree to which the confidence interval for that variable's regression parameter is expanded relative to a model with uncorrelated predictors. sq. root vif >2 indicates a multicollinearity problem.

We see that multicollinearity is there in our data. Infant deaths, under 5 deaths, polio, thinness 1 to 19 yrs and thinness 5 to 9 yrs are highly correlated.

Hide

```
#outlier test
outlierTest(model2)
```

```
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
71 -3.178932          0.0018478      0.27716
```

Outliers are observations that aren't predicted well by the model. They have either unusually large positive or negative residuals. Positive residuals indicate that the model is underestimating the response value, while negative residuals indicate an overestimation.

One way to identify outliers is points in the Q-Q plot that lie outside the confidence band are considered outliers. A rough rule of thumb is that standardized residuals that are larger than 2 or less than -2 are worth attention.

The car package also provides a statistical test for outliers. The outlierTest() function reports the Bonferroni adjusted p-value for the largest absolute studentized residual:

this function tests the single largest (positive or negative) residual for significance as an outlier. If it isn't significant, there are no outliers in the dataset. If it is significant, we must delete it and rerun the test to see if others are present.

Comparing models: one including status variable and other without including status variable

Hide

```
#Comparing nested models using the anova() function
anova(model2,model1)
```

```
Analysis of Variance Table

Model 1: Life_expectancy ~ Adult_Mortality + Infant_deaths + Alcohol +
    Percentage_expenditure + Hepatitis_B + Measles + BMI + Under_5_deaths +
    Polio + Total_expenditure + Diphtheria + HIV_AIDS + GDP +
    Population + Thinness_1_to_19_yrs + Thinness_5_to_9_yrs +
    Income_composition_of_resources + Schooling
Model 2: Life_expectancy ~ Adult_Mortality + Infant_deaths + Alcohol +
    Percentage_expenditure + Hepatitis_B + Measles + BMI + Under_5_deaths +
    Polio + Total_expenditure + Diphtheria + HIV_AIDS + GDP +
    Population + Thinness_1_to_19_yrs + Thinness_5_to_9_yrs +
    Income_composition_of_resources + Schooling + `Status Developing` +
    `Status Developed`
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    131 0.56664
2    130 0.56469  1 0.0019464 0.4481 0.5044
```

Because the test is nonsignificant (p = .504), we conclude that status variable don't add to the linear prediction and we're justified in dropping them from our model.

Hide

```
#Comparing models with the AIC-Akaike Information Criterion
AIC(model1,model2)
```

```
       df       AIC
model1 21 -369.6355
model2 20 -371.1194
```

the model with the lowest AIC score is preferred. The absolute values of the AIC scores do not matter. These scores can be negative or positive.The AIC values suggest that the model2 is the better model.

Hide

```
#Variable selection-STEPWISE REGRESSION
library(MASS)
```

```


Attaching package: <U+393C><U+3E31>MASS<U+393C><U+3E32>

The following object is masked from <U+393C><U+3E31>package:dplyr<U+393C><U+3E32>:

    select
```

Hide

```
#Backward stepwise selection (model1 with status variable)
stepAIC(model1, direction="backward")
```

```
Start:  AIC=-797.32
Life_expectancy ~ Adult_Mortality + Infant_deaths + Alcohol + 
    Percentage_expenditure + Hepatitis_B + Measles + BMI + Under_5_deaths + 
    Polio + Total_expenditure + Diphtheria + HIV_AIDS + GDP + 
    Population + Thinness_1_to_19_yrs + Thinness_5_to_9_yrs + 
    Income_composition_of_resources + Schooling + `Status Developing` + 
    `Status Developed`


Step:  AIC=-797.32
Life_expectancy ~ Adult_Mortality + Infant_deaths + Alcohol + 
    Percentage_expenditure + Hepatitis_B + Measles + BMI + Under_5_deaths + 
    Polio + Total_expenditure + Diphtheria + HIV_AIDS + GDP + 
    Population + Thinness_1_to_19_yrs + Thinness_5_to_9_yrs + 
    Income_composition_of_resources + Schooling + `Status Developing`

                                 Df Sum of Sq     RSS     AIC
- Measles                         1   0.00001 0.56470 -799.32
- Thinness_1_to_19_yrs            1   0.00005 0.56474 -799.30
- Thinness_5_to_9_yrs             1   0.00006 0.56475 -799.30
- GDP                             1   0.00045 0.56514 -799.20
- BMI                             1   0.00078 0.56547 -799.11
- Polio                           1   0.00152 0.56621 -798.91
- `Status Developing`             1   0.00195 0.56664 -798.80
- Population                      1   0.00411 0.56880 -798.23
- Percentage_expenditure          1   0.00716 0.57185 -797.43
<none>                                       0.56469 -797.32
- Hepatitis_B                     1   0.00998 0.57467 -796.69
- Total_expenditure               1   0.01037 0.57506 -796.59
- Infant_deaths                   1   0.01046 0.57515 -796.56
- Alcohol                         1   0.01049 0.57518 -796.56
- HIV_AIDS                        1   0.01519 0.57988 -795.34
- Under_5_deaths                  1   0.01671 0.58140 -794.94
- Income_composition_of_resources 1   0.02674 0.59143 -792.38
- Schooling                       1   0.04846 0.61315 -786.97
- Diphtheria                      1   0.05373 0.61842 -785.68
- Adult_Mortality                 1   0.54788 1.11257 -697.59

Step:  AIC=-799.32
Life_expectancy ~ Adult_Mortality + Infant_deaths + Alcohol + 
    Percentage_expenditure + Hepatitis_B + BMI + Under_5_deaths + 
    Polio + Total_expenditure + Diphtheria + HIV_AIDS + GDP + 
    Population + Thinness_1_to_19_yrs + Thinness_5_to_9_yrs + 
    Income_composition_of_resources + Schooling + `Status Developing`

                                 Df Sum of Sq     RSS     AIC
- Thinness_1_to_19_yrs            1   0.00005 0.56474 -801.30
- Thinness_5_to_9_yrs             1   0.00006 0.56476 -801.30
- GDP                             1   0.00045 0.56514 -801.20
- BMI                             1   0.00080 0.56550 -801.10
- Polio                           1   0.00152 0.56622 -800.91
- `Status Developing`             1   0.00195 0.56665 -800.80
- Population                      1   0.00561 0.57031 -799.83
```

```
- Percentage_expenditure           1   0.00719 0.57189 -799.42
<none>                                       0.56470 -799.32
- Hepatitis_B                      1   0.01013 0.57482 -798.65
- Total_expenditure                1   0.01038 0.57508 -798.58
- Alcohol                          1   0.01051 0.57521 -798.55
- Infant_deaths                    1   0.01082 0.57551 -798.47
- HIV_AIDS                         1   0.01522 0.57991 -797.33
- Under_5_deaths                   1   0.01671 0.58140 -796.94
- Income_composition_of_resources  1   0.02684 0.59154 -794.35
- Schooling                        1   0.04881 0.61350 -788.88
- Diphtheria                       1   0.05403 0.61873 -787.61
- Adult_Mortality                  1   0.54801 1.11270 -699.58

Step:  AIC=-801.3
Life_expectancy ~ Adult_Mortality + Infant_deaths + Alcohol +
    Percentage_expenditure + Hepatitis_B + BMI + Under_5_deaths +
    Polio + Total_expenditure + Diphtheria + HIV_AIDS + GDP +
    Population + Thinness_5_to_9_yrs + Income_composition_of_resources +
    Schooling + `Status Developing`

                                 Df Sum of Sq     RSS     AIC
- Thinness_5_to_9_yrs             1   0.00002 0.56476 -803.30
- GDP                             1   0.00042 0.56517 -803.19
- BMI                             1   0.00079 0.56554 -803.09
- Polio                           1   0.00155 0.56630 -802.89
- `Status Developing`             1   0.00197 0.56671 -802.78
- Population                      1   0.00556 0.57031 -801.83
- Percentage_expenditure          1   0.00715 0.57190 -801.41
<none>                                       0.56474 -801.30
- Hepatitis_B                     1   0.01027 0.57501 -800.60
- Alcohol                         1   0.01047 0.57521 -800.55
- Total_expenditure               1   0.01074 0.57548 -800.48
- Infant_deaths                   1   0.01082 0.57557 -800.46
- HIV_AIDS                        1   0.01520 0.57995 -799.32
- Under_5_deaths                  1   0.01667 0.58142 -798.94
- Income_composition_of_resources 1   0.02680 0.59154 -796.35
- Schooling                       1   0.04876 0.61350 -790.88
- Diphtheria                      1   0.05399 0.61873 -789.61
- Adult_Mortality                 1   0.55545 1.12020 -700.57

Step:  AIC=-803.3
Life_expectancy ~ Adult_Mortality + Infant_deaths + Alcohol +
    Percentage_expenditure + Hepatitis_B + BMI + Under_5_deaths +
    Polio + Total_expenditure + Diphtheria + HIV_AIDS + GDP +
    Population + Income_composition_of_resources + Schooling +
    `Status Developing`

                                 Df Sum of Sq     RSS     AIC
- GDP                             1   0.00043 0.56519 -805.19
- BMI                             1   0.00127 0.56603 -804.96
- Polio                           1   0.00156 0.56632 -804.88
- `Status Developing`             1   0.00201 0.56677 -804.76
- Population                      1   0.00556 0.57032 -803.83
- Percentage_expenditure          1   0.00719 0.57195 -803.40
```

```
<none>                                      0.56476 -803.30
- Hepatitis_B                    1   0.01026 0.57502 -802.60
- Infant_deaths                  1   0.01082 0.57558 -802.45
- Total_expenditure              1   0.01082 0.57558 -802.45
- Alcohol                        1   0.01183 0.57659 -802.19
- HIV_AIDS                       1   0.01521 0.57997 -801.31
- Under_5_deaths                 1   0.01666 0.58142 -800.94
- Income_composition_of_resources  1   0.02745 0.59221 -798.18
- Schooling                      1   0.04879 0.61355 -792.87
- Diphtheria                     1   0.05399 0.61875 -791.60
- Adult_Mortality                1   0.55570 1.12046 -702.53

Step:  AIC=-805.19
Life_expectancy ~ Adult_Mortality + Infant_deaths + Alcohol +
    Percentage_expenditure + Hepatitis_B + BMI + Under_5_deaths +
    Polio + Total_expenditure + Diphtheria + HIV_AIDS + Population +
    Income_composition_of_resources + Schooling + `Status Developing`

                                Df Sum of Sq     RSS     AIC
- BMI                            1   0.00112 0.56630 -806.89
- Polio                          1   0.00178 0.56697 -806.71
- `Status Developing`            1   0.00182 0.56701 -806.70
- Population                     1   0.00544 0.57063 -805.75
<none>                                      0.56519 -805.19
- Hepatitis_B                    1   0.01092 0.57610 -804.32
- Infant_deaths                  1   0.01110 0.57629 -804.27
- Alcohol                        1   0.01179 0.57698 -804.09
- Total_expenditure              1   0.01283 0.57801 -803.82
- HIV_AIDS                       1   0.01506 0.58025 -803.24
- Under_5_deaths                 1   0.01699 0.58217 -802.74
- Percentage_expenditure         1   0.01716 0.58234 -802.70
- Income_composition_of_resources  1   0.02713 0.59232 -800.15
- Schooling                      1   0.04926 0.61445 -794.65
- Diphtheria                     1   0.05571 0.62089 -793.08
- Adult_Mortality                1   0.55748 1.12267 -704.24

Step:  AIC=-806.89
Life_expectancy ~ Adult_Mortality + Infant_deaths + Alcohol +
    Percentage_expenditure + Hepatitis_B + Under_5_deaths + Polio +
    Total_expenditure + Diphtheria + HIV_AIDS + Population +
    Income_composition_of_resources + Schooling + `Status Developing`

                                Df Sum of Sq     RSS     AIC
- Polio                          1   0.00167 0.56797 -808.45
- `Status Developing`            1   0.00176 0.56806 -808.42
- Population                     1   0.00618 0.57248 -807.26
<none>                                      0.56630 -806.89
- Infant_deaths                  1   0.01021 0.57651 -806.21
- Alcohol                        1   0.01101 0.57731 -806.00
- Hepatitis_B                    1   0.01155 0.57785 -805.86
- HIV_AIDS                       1   0.01546 0.58176 -804.85
- Under_5_deaths                 1   0.01605 0.58236 -804.70
- Percentage_expenditure         1   0.01621 0.58251 -804.66
- Total_expenditure              1   0.01713 0.58343 -804.42
```

```
- Income_composition_of_resources  1   0.03331 0.59961 -800.32
- Schooling                        1   0.05110 0.61740 -795.93
- Diphtheria                       1   0.05624 0.62255 -794.69
- Adult_Mortality                  1   0.57443 1.14073 -703.84

Step:  AIC=-808.45
Life_expectancy ~ Adult_Mortality + Infant_deaths + Alcohol +
    Percentage_expenditure + Hepatitis_B + Under_5_deaths + Total_expenditure +
    Diphtheria + HIV_AIDS + Population + Income_composition_of_resources +
    Schooling + `Status Developing`

                                  Df Sum of Sq     RSS     AIC
- `Status Developing`              1   0.00181 0.56978 -809.97
- Population                       1   0.00604 0.57401 -808.86
<none>                                         0.56797 -808.45
- Infant_deaths                    1   0.01001 0.57798 -807.83
- Alcohol                          1   0.01096 0.57893 -807.58
- HIV_AIDS                         1   0.01519 0.58316 -806.49
- Percentage_expenditure           1   0.01574 0.58371 -806.35
- Under_5_deaths                   1   0.01581 0.58377 -806.33
- Total_expenditure                1   0.01605 0.58402 -806.27
- Hepatitis_B                      1   0.01620 0.58416 -806.23
- Income_composition_of_resources  1   0.03268 0.60065 -802.06
- Schooling                        1   0.05045 0.61841 -797.68
- Diphtheria                       1   0.14017 0.70813 -777.36
- Adult_Mortality                  1   0.57359 1.14156 -705.74

Step:  AIC=-809.97
Life_expectancy ~ Adult_Mortality + Infant_deaths + Alcohol +
    Percentage_expenditure + Hepatitis_B + Under_5_deaths + Total_expenditure +
    Diphtheria + HIV_AIDS + Population + Income_composition_of_resources +
    Schooling

                                  Df Sum of Sq     RSS     AIC
- Population                       1   0.00657 0.57635 -810.25
<none>                                         0.56978 -809.97
- Alcohol                          1   0.00915 0.57894 -809.58
- Infant_deaths                    1   0.00955 0.57933 -809.48
- Percentage_expenditure           1   0.01398 0.58376 -808.33
- Total_expenditure                1   0.01509 0.58487 -808.05
- HIV_AIDS                         1   0.01526 0.58504 -808.01
- Under_5_deaths                   1   0.01533 0.58512 -807.99
- Hepatitis_B                      1   0.01814 0.58792 -807.27
- Income_composition_of_resources  1   0.03191 0.60169 -803.80
- Schooling                        1   0.05059 0.62037 -799.21
- Diphtheria                       1   0.14514 0.71492 -777.93
- Adult_Mortality                  1   0.57181 1.14159 -707.73

Step:  AIC=-810.25
Life_expectancy ~ Adult_Mortality + Infant_deaths + Alcohol +
    Percentage_expenditure + Hepatitis_B + Under_5_deaths + Total_expenditure +
    Diphtheria + HIV_AIDS + Income_composition_of_resources +
    Schooling
```

```
                                Df Sum of Sq     RSS     AIC
<none>                                       0.57635 -810.25
- Alcohol                        1   0.00931 0.58566 -809.85
- Percentage_expenditure         1   0.01360 0.58995 -808.75
- Total_expenditure              1   0.01453 0.59088 -808.52
- HIV_AIDS                       1   0.01802 0.59437 -807.63
- Hepatitis_B                    1   0.01984 0.59620 -807.17
- Infant_deaths                  1   0.02007 0.59642 -807.12
- Under_5_deaths                 1   0.02266 0.59901 -806.47
- Income_composition_of_resources  1   0.03295 0.60930 -803.91
- Schooling                      1   0.05283 0.62918 -799.10
- Diphtheria                     1   0.15685 0.73321 -776.14
- Adult_Mortality                1   0.56531 1.14166 -709.72

Call:
lm(formula = Life_expectancy ~ Adult_Mortality + Infant_deaths +
    Alcohol + Percentage_expenditure + Hepatitis_B + Under_5_deaths +
    Total_expenditure + Diphtheria + HIV_AIDS + Income_composition_of_resources +
    Schooling, data = train)

Coefficients:
                  (Intercept)                 Adult_Mortality                    Infant_death
s
                      0.40250                        -0.71011                         1.7903
1
                      Alcohol          Percentage_expenditure                      Hepatitis_
B
                      0.03656                         0.09432                        -0.1123
5
                Under_5_deaths               Total_expenditure                       Diphtheri
a
                     -1.87038                         0.09508                         0.3108
0
                     HIV_AIDS  Income_composition_of_resources                       Schoolin
g
                     -0.16994                         0.13978                         0.2340
6
```

Improving model accuracy

by taking into account, multicollinearity and variable section method-stepwise regression, we are deleting few variabes such as: Infant deaths, under 5 deaths,polio, thinness_1to19_yrs and thinness_5_to9_yrs

Hide

```
data_processed3<-data_processed2[,c(-3,-9,-10,-16,-17)]
#Splitting data into train and test set
train2<-data_processed3[1:150,]
test2<-data_processed3[151:193,]
#check dimensions of train and test set
dim(train2)
```

```
[1] 150  14
```

Hide

```
dim(test2)
```

```
[1] 43 14
```

Hide

```
#Fit multiple linear regression model, excluding Country variable
model3<-lm(Life_expectancy~.,data=train2)
#check summary of model
summary(model3)
```

```

Call:
lm(formula = Life_expectancy ~ ., data = train2)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.183894 -0.044177 -0.000536  0.042615  0.157510 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      0.382622   0.049638   7.708 2.40e-12 ***
Adult_Mortality                 -0.712885   0.063051 -11.307  < 2e-16 ***
Alcohol                          0.027736   0.024847   1.116 0.266286    
Percentage_expenditure           0.112276   0.109734   1.023 0.308043    
Hepatitis_B                     -0.115155   0.053580  -2.149 0.033387 *  
Measles                         -0.083396   0.048690  -1.713 0.089033 .  
BMI                              0.009785   0.038362   0.255 0.799053    
Total_expenditure                0.084550   0.057260   1.477 0.142093    
Diphtheria                       0.329985   0.051193   6.446 1.85e-09 ***
HIV_AIDS                        -0.169468   0.084134  -2.014 0.045954 *  
GDP                             -0.025710   0.091913  -0.280 0.780115    
Population                       0.006393   0.071968   0.089 0.929349    
Income_composition_of_resources  0.153317   0.053802   2.850 0.005058 ** 
Schooling                        0.236615   0.067717   3.494 0.000642 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06612 on 136 degrees of freedom
Multiple R-squared:  0.9303,    Adjusted R-squared:  0.9237 
F-statistic: 139.7 on 13 and 136 DF,  p-value: < 2.2e-16
```

Hide

```
#apply model on test data
testing2<-predict(model3,test2)
#comapre actual and predicted values of Life_expectancy
head(cbind(test$Life_expectancy,testing2))
```

```
            testing2
1 0.7644132 0.7678996
2 0.7210021 0.6939607
3 0.0000000 0.2461779
4 0.9708305 0.8680210
5 0.7862045 0.8153240
6 0.9229581 0.9403442
```

Hide

```
#combine Counttry name, actual and predicted values of life expectancy
cbind(data3[151:193,"Country"],test$Life_expectancy,testing2)
```

	Country	test$Life_expectancy	testing2
1	Serbia	0.7644132	0.76789965
2	Seychelles	0.7210021	0.69396067
3	Sierra Leone	0.0000000	0.24617794
4	Singapore	0.9708305	0.86802104
5	Slovakia	0.7862045	0.81532398
6	Slovenia	0.9229581	0.94034424
7	Solomon Islands	0.5929993	0.57380556
8	Somalia	0.1978380	0.22202970
9	South Africa	0.3126287	0.21468486
10	South Sudan	0.2131091	0.11321511
11	Spain	0.9871311	0.94986976
12	Sri Lanka	0.7491421	0.72532967
13	Sudan	0.4313658	0.39206734
14	Suriname	0.6580302	0.60755199
15	Swaziland	0.1431023	0.21080996
16	Sweden	0.9994852	1.00672945
17	Switzerland	0.9943377	1.02159460
18	Syrian Arab Republic	0.6791352	0.53437066
19	Tajikistan	0.5640014	0.60480822
20	Thailand	0.7403912	0.68551932
21	The former Yugoslav republic of Macedonia	0.7687028	0.79208538
22	Timor-Leste	0.5118394	0.50322430
23	Togo	0.2896362	0.32876703
24	Tonga	0.7252917	0.68098107
25	Trinidad and Tobago	0.6851407	0.63598762
26	Tunisia	0.7753946	0.90306039
27	Turkey	0.7632121	0.72316945
28	Turkmenistan	0.5080645	0.47800922
29	Tuvalu	0.6345211	0.36973266
30	Uganda	0.2633837	0.32186127
31	Ukraine	0.6540837	0.62072736
32	United Arab Emirates	0.8122855	0.83517500
33	United Kingdom of Great Britain and Northern Ireland	0.9521277	0.82670265
34	United Republic of Tanzania	0.2716198	0.37343488
35	United States of America	0.8771448	0.88376383
36	Uruguay	0.8225806	0.79222367
37	Uzbekistan	0.6017502	0.63448148
38	Vanuatu	0.6938916	0.53649794
39	Venezuela (Bolivarian Republic of)	0.7487989	0.63402291
40	Viet Nam	0.7868909	0.69441859
41	Yemen	0.4873027	0.47770515
42	Zambia	0.2139671	0.25133311
43	Zimbabwe	0.1201098	0.04072535

Hide

```
#Comparing nested models using the anova() function
anova(model3,model2)
```

```
Analysis of Variance Table

Model 1: Life_expectancy ~ Adult_Mortality + Alcohol + Percentage_expenditure +
    Hepatitis_B + Measles + BMI + Total_expenditure + Diphtheria +
    HIV_AIDS + GDP + Population + Income_composition_of_resources +
    Schooling
Model 2: Life_expectancy ~ Adult_Mortality + Infant_deaths + Alcohol +
    Percentage_expenditure + Hepatitis_B + Measles + BMI + Under_5_deaths +
    Polio + Total_expenditure + Diphtheria + HIV_AIDS + GDP +
    Population + Thinness_1_to_19_yrs + Thinness_5_to_9_yrs +
    Income_composition_of_resources + Schooling
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    136 0.59455
2    131 0.56664  5  0.027912 1.2906 0.2718
```

Hide

```
#Since p-value is non signifacnt, therefore we can say our model3 is better model.
#Comparing models with the AIC-Akaike Information Criterion
AIC(model2,model3)
```

```
       df       AIC
model2 20 -371.1194
model3 15 -373.9066
```

Hide

```
#Also using AIC value, we see that our model3 is better model
#lets check our model by removing GDP and Population variable
str(data_processed3)
```

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':   193 obs. of  14 variables:
 $ Life_expectancy                : num  0.3317 0.7974 0.7551 0.0798 0.7946 ...
 $ Adult_Mortality                : num  0.4711 0.0495 0.1683 0.5831 0.2047 ...
 $ Alcohol                        : num  0.000338 0.374185 0.050956 0.437659 0.597782 ...
 $ Percentage_expenditure         : num  0.00357 0.01972 0.0241 0.01042 0.10218 ...
 $ Hepatitis_B                    : num  0.622 0.989 0.777 0.735 0.98 ...
 $ Measles                        : num  0.03587 0.00081 0.02952 0.05408 0 ...
 $ BMI                            : num  0.126 0.534 0.53 0.156 0.405 ...
 $ Total_expenditure              : num  0.414 0.263 0.181 0.139 0.192 ...
 $ Diphtheria                     : num  0.481 0.99 0.921 0.43 0.992 ...
 $ HIV_AIDS                       : num  0 0 0 0.069077 0.000761 ...
 $ GDP                            : num  0.00356 0.03466 0.04738 0.03213 0.16816 ...
 $ Population                     : num  0.02366 0.00165 0.05137 0.02407 0.03026 ...
 $ Income_composition_of_resources: num  0.355 0.723 0.704 0.408 0.446 ...
 $ Schooling                      : num  0.41 0.606 0.634 0.401 0.441 ...
```

Hide

```
data_processed4<-data_processed3[,c(-11,-12)]
str(data_processed4)
```

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':   193 obs. of  12 variables:
 $ Life_expectancy                : num  0.3317 0.7974 0.7551 0.0798 0.7946 ...
 $ Adult_Mortality                : num  0.4711 0.0495 0.1683 0.5831 0.2047 ...
 $ Alcohol                        : num  0.000338 0.374185 0.050956 0.437659 0.597782 ...
 $ Percentage_expenditure         : num  0.00357 0.01972 0.0241 0.01042 0.10218 ...
 $ Hepatitis_B                    : num  0.622 0.989 0.777 0.735 0.98 ...
 $ Measles                        : num  0.03587 0.00081 0.02952 0.05408 0 ...
 $ BMI                            : num  0.126 0.534 0.53 0.156 0.405 ...
 $ Total_expenditure              : num  0.414 0.263 0.181 0.139 0.192 ...
 $ Diphtheria                     : num  0.481 0.99 0.921 0.43 0.992 ...
 $ HIV_AIDS                       : num  0 0 0 0.069077 0.000761 ...
 $ Income_composition_of_resources: num  0.355 0.723 0.704 0.408 0.446 ...
 $ Schooling                      : num  0.41 0.606 0.634 0.401 0.441 ...
```

Hide

```
#Splitting data into train and test set
train3<-data_processed4[1:150,]
test3<-data_processed4[151:193,]
#check dimensions of train and test set
dim(train3)
```

```
[1] 150  12
```

Hide

```
dim(test3)
```

```
[1] 43 12
```

Hide

```
#Fit multiple linear regression model, excluding Country variable
model4<-lm(Life_expectancy~.,data=train3)
#check summary of model
summary(model4)
```

```
Call:
lm(formula = Life_expectancy ~ ., data = train3)

Residuals:
      Min        1Q    Median        3Q       Max
-0.184240 -0.044913  0.000915  0.042472  0.155312

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                     0.38453    0.04870   7.895 8.09e-13 ***
Adult_Mortality                -0.71367    0.06253 -11.414  < 2e-16 ***
Alcohol                         0.02795    0.02466   1.133 0.259092
Percentage_expenditure          0.08530    0.05371   1.588 0.114513
Hepatitis_B                    -0.11852    0.05133  -2.309 0.022428 *
Measles                        -0.08181    0.04420  -1.851 0.066312 .
BMI                             0.00816    0.03770   0.216 0.828948
Total_expenditure               0.08808    0.05456   1.614 0.108709
Diphtheria                      0.33083    0.05027   6.580 9.01e-10 ***
HIV_AIDS                       -0.16905    0.08326  -2.031 0.044224 *
Income_composition_of_resources  0.15072    0.05258   2.866 0.004806 **
Schooling                       0.23744    0.06718   3.534 0.000557 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06566 on 138 degrees of freedom
Multiple R-squared:  0.9303,    Adjusted R-squared:  0.9247
F-statistic: 167.4 on 11 and 138 DF,  p-value: < 2.2e-16
```

Hide

```
#apply model on test data
testing3<-predict(model4,test3)
#comapre actual and predicted values of Life_expectancy
head(cbind(test$Life_expectancy,testing3))
```

```
            testing3
1 0.7644132 0.7664879
2 0.7210021 0.6935353
3 0.0000000 0.2464121
4 0.9708305 0.8737387
5 0.7862045 0.8165721
6 0.9229581 0.9401714
```

Hide

```
#combine Counttry name, actual and predicted values of life expectancy
cbind(data3[151:193,"Country"],test$Life_expectancy,testing2)
```

	Country	test$Life_expectancy	testing2
1	Serbia	0.7644132	0.76789965
2	Seychelles	0.7210021	0.69396067
3	Sierra Leone	0.0000000	0.24617794
4	Singapore	0.9708305	0.86802104
5	Slovakia	0.7862045	0.81532398
6	Slovenia	0.9229581	0.94034424
7	Solomon Islands	0.5929993	0.57380556
8	Somalia	0.1978380	0.22202970
9	South Africa	0.3126287	0.21468486
10	South Sudan	0.2131091	0.11321511
11	Spain	0.9871311	0.94986976
12	Sri Lanka	0.7491421	0.72532967
13	Sudan	0.4313658	0.39206734
14	Suriname	0.6580302	0.60755199
15	Swaziland	0.1431023	0.21080996
16	Sweden	0.9994852	1.00672945
17	Switzerland	0.9943377	1.02159460
18	Syrian Arab Republic	0.6791352	0.53437066
19	Tajikistan	0.5640014	0.60480822
20	Thailand	0.7403912	0.68551932
21	The former Yugoslav republic of Macedonia	0.7687028	0.79208538
22	Timor-Leste	0.5118394	0.50322430
23	Togo	0.2896362	0.32876703
24	Tonga	0.7252917	0.68098107
25	Trinidad and Tobago	0.6851407	0.63598762
26	Tunisia	0.7753946	0.90306039
27	Turkey	0.7632121	0.72316945
28	Turkmenistan	0.5080645	0.47800922
29	Tuvalu	0.6345211	0.36973266
30	Uganda	0.2633837	0.32186127
31	Ukraine	0.6540837	0.62072736
32	United Arab Emirates	0.8122855	0.83517500
33	United Kingdom of Great Britain and Northern Ireland	0.9521277	0.82670265
34	United Republic of Tanzania	0.2716198	0.37343488
35	United States of America	0.8771448	0.88376383
36	Uruguay	0.8225806	0.79222367
37	Uzbekistan	0.6017502	0.63448148
38	Vanuatu	0.6938916	0.53649794
39	Venezuela (Bolivarian Republic of)	0.7487989	0.63402291
40	Viet Nam	0.7868909	0.69441859
41	Yemen	0.4873027	0.47770515
42	Zambia	0.2139671	0.25133311
43	Zimbabwe	0.1201098	0.04072535

Hide

```
#Comparing nested models using the anova() function
anova(model4,model3)
```

```
Analysis of Variance Table

Model 1: Life_expectancy ~ Adult_Mortality + Alcohol + Percentage_expenditure +
    Hepatitis_B + Measles + BMI + Total_expenditure + Diphtheria +
    HIV_AIDS + Income_composition_of_resources + Schooling
Model 2: Life_expectancy ~ Adult_Mortality + Alcohol + Percentage_expenditure +
    Hepatitis_B + Measles + BMI + Total_expenditure + Diphtheria +
    HIV_AIDS + GDP + Population + Income_composition_of_resources +
    Schooling
  Res.Df     RSS Df  Sum of Sq      F Pr(>F)
1    138 0.59493
2    136 0.59455  2 0.00037874 0.0433 0.9576
```

Hide

```
#Comparing models with the AIC-Akaike Information Criterion
AIC(model3,model4)
```

```
       df       AIC
model3 15 -373.9066
model4 13 -377.8111
```

Hide

```
#We see that GDP and Population is also not a good predictor.
```

So, we got the final accuracy 93.03%