MA317 GROUP COURSEWORK

Submitted

in

Partial Fulfilment

of

the Module

MA 317 MODELLING EXPERIMENTAL DATA

Guided by

Dr. Joseph Bailey

Dr. Stella Hadjiantoni

Submitted by

Dilpa Rao	**-**	**1906319**
Divya Arora	**-**	**1901423**
Rohini Subramaniam	**-**	**1908736**
Srinidhi Karthikeyan	**-**	**1900637**
Viraj Kumar Dewangan	**-**	**1901181**

Date – 12/11/2019

TABLE OF CONTENTS

1. INTRODUCTION:

In this project, we have analyzed a dataset of the World Development Indicators (WDI), derived from a primary World Bank database. We removed those variables from the dataset which had more than 50 percent NA values. We have then performed mean imputation to fill in the missing values for the remaining columns in the data set. Collinearity between the predictor variables was investigated using F- G Test and few variables were removed based on VIF values. We then used multiple linear regression model to regress life expectancy versus the remaining variables based on optimizing R^2 values for the predicted model. Finally, we did an ANOVA for the average life expectancies across the continents to check the hypothesis that mean life expectancies were same across different continents.

The summary statistics and data plots was carried out by Rohini

The task of imputation was done by Divya.

Collinearity investigation and resolution was collectively done by Dilpa.

Srinidhi contributed by doing the Multiple Linear Regression Modelling for the Life Expectancy.

The ANOVA analysis was done by Viraj. Viraj added the continents to the dataset and grouped them to carry out the ANOVA analysis. Processing , formatting and compiling is done by Viraj to produce the final report for group submission.

2. ANALYSIS:

2.1 Data Summary and Plots

The section describe the data available for constructing a model. In this section we use descriptive statistics for univariate. The original dataset contains information on 24 World Development Indicators (WDI) for 269 countries. This data has been derived from a World Bank database.1. The variables are the following.

Dependent variable: Life - Life expectancy at birth,

Predictor variables:

Elec - Access to electricity

nat_income - Adjusted net national income

child_droupouts - Children out of school of primary school age

exp_p_health - Expenditure on primary education of government expenditure

inv_water_san - Public private partnerships investment in water and sanitation mort - Mortality rate attributed to unsafe water

litrate - Literate rate adult

gpop - Population growth

tpop - Population total .

pcomrate - Primary completion rate

secedu - Secondary education

teachers - Secondary education teachers

hexpen - Current health expenditure

heexpenpcc - Current health expenditure per capita

tunemp - Unemployment total of total labor force

unemp - Unemployment youth total of total labor force

rpop - Rural population of total population

fert - Adolescent fertility rate births

gdp - GDP per capita

msub - Mobile cellular subscriptions.

Isub - Individuals using the Internet of population

Graphical Representation

Some variables are chosen to check for graphical representation.

In Figure 1 below, we can see the relationship between log of Income per capita and life expectancy. By looking at the plot, we cannot identify a clear linear relationship between these 2 variables. Further analysis can be carried out to check if there is a linear relationship.

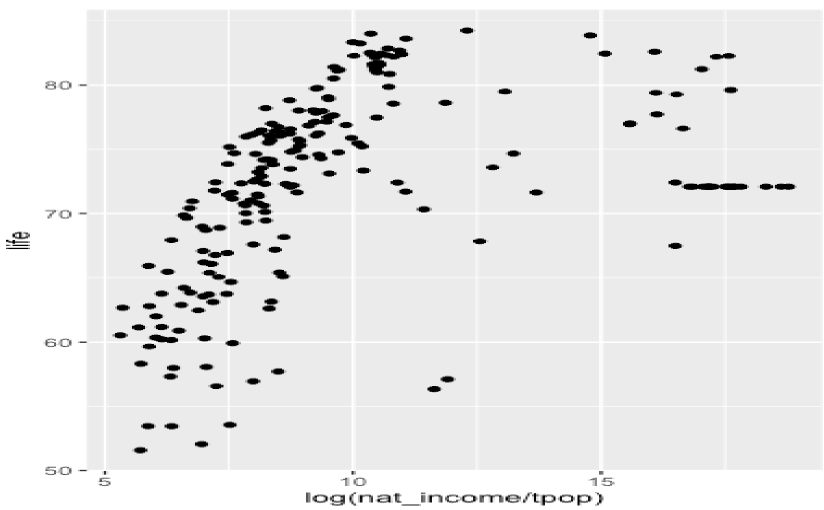


Figure 1- Income per Capita vs Life Expectancy

Figure 2 below shows a scatter diagram which clearly indicates that there is a linear relationship between GDP and Life expectancy with positive slope. Further analysis can be carried out for more information.

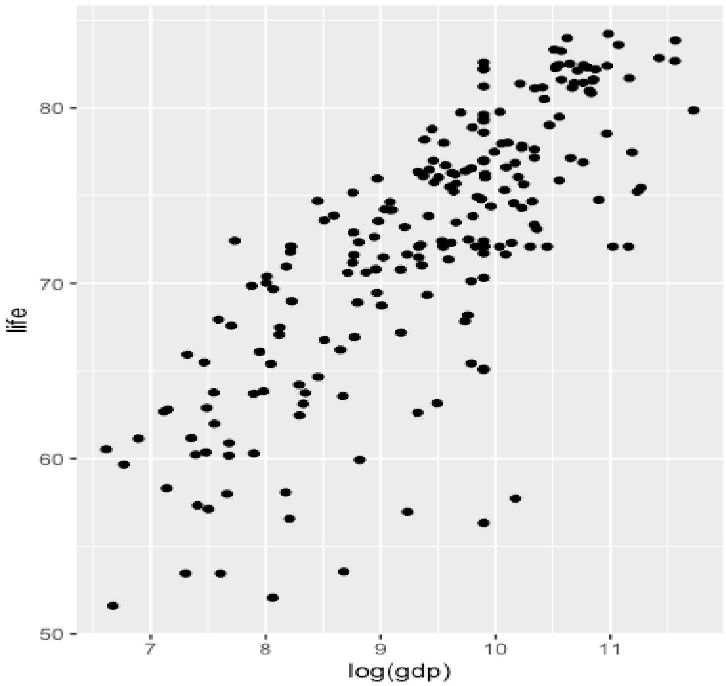


Figure 2- GDP vs Life Expectancy

Figure 3 represents a linear relationship between internet subscription and life expectancy with positive slope. Further analysis will be carried out in the later part of the report.

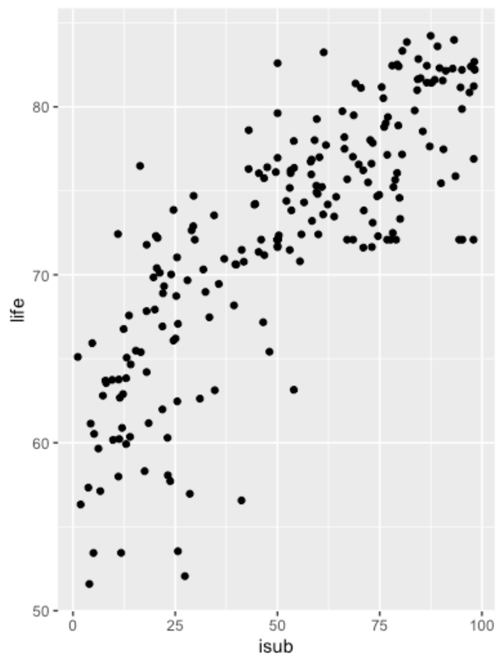


Figure 3 – Internet Subscription vs Life Expectancy

Univariate summary statistics for initial predictive variables

	mean	sd	min	max	quadl.0%	quadl.25%	quadl.50%
life	7.228338e+01	7.448287e+00	5.159300e+01	8.422683e+01	5.159300e+01	6.793000e+01	7.264400e+01
elec	8.485872e+01	2.530521e+01	9.298458e+00	1.000000e+02	9.298458e+00	7.742182e+01	9.998625e+01
nat_income	6.154740e+11	1.388249e+12	1.566043e+08	1.600000e+13	1.566043e+08	1.021743e+10	6.174885e+10
child_droupouts	6.528120e+00	7.182964e+00	1.400000e-04	4.261848e+01	1.400000e-04	1.219400e+00	6.935628e+00
mort	1.254249e+01	1.910715e+01	1.000000e-01	1.010000e+02	1.000000e-01	3.000000e-01	2.900000e+00
gpop	1.288819e+00	1.215480e+00	-3.066274e+00	4.845614e+00	-3.066274e+00	5.004802e-01	1.144029e+00
tpop	3.547986e+07	1.356897e+08	1.122500e+04	1.378665e+09	1.122500e+04	7.713660e+05	6.492164e+06
pcomrate	9.164426e+01	1.219857e+01	4.087417e+01	1.310185e+02	4.087417e+01	9.095416e+01	9.095416e+01
secedu	6.367647e+00	8.908653e-01	4.000000e+00	9.000000e+00	4.000000e+00	6.000000e+00	6.000000e+00
teachers	1.001506e+06	1.057310e+06	4.000000e+01	6.219580e+06	4.000000e+01	3.626000e+04	3.144820e+05
hexpen	6.737924e+00	2.769628e+00	1.749862e+00	2.328730e+01	1.749862e+00	4.981881e+00	6.698909e+00
hexpenpcc	1.423967e+03	1.560084e+03	2.990777e+01	9.869742e+03	2.990777e+01	2.978559e+02	1.071347e+03
rpop	3.969725e+01	2.390397e+01	0.000000e+00	8.761200e+01	0.000000e+00	2.042300e+01	3.935500e+01
fert	4.812793e+01	3.834522e+01	2.882000e-01	1.893790e+02	2.882000e-01	1.540260e+01	4.799420e+01
gdp	2.060898e+04	2.058161e+04	7.439036e+02	1.235736e+05	7.439036e+02	5.833996e+03	1.561315e+04
msub	1.071041e+02	3.884567e+01	1.424865e+01	3.214517e+02	1.424865e+01	8.514868e+01	1.079013e+02
isub	5.124205e+01	2.807780e+01	1.177119e+00	9.824002e+01	1.177119e+00	2.507330e+01	5.219133e+01

3. IMPUTATING MISSING VALUES IN THE DATA SET –

The dataset we have analysed are having Missing values Completely at Random (MCAR) which means that probability of missing values is not related to any of the variables but only concerns with itself. After identifying patterns or reasons behind the missing data we need to understand the distribution of missing data. In our dataset, we have five variables i.e. Children-Education-Expenditure, Water-Sanitation-Expenditure, Literacy Rate, Total Employment, Youth Employment which contains more than fifty percent missing values.

Complete case analysis is not an appropriate method to handle the missing values because it reduces the statistical power by removing a substantial amount of data as well as cannot be able to use all the available information and discards data for any cases that has one or more missing values. In majority of the scenarios while doing this listwise deletion, estimates are biased. This method is preferred when proportion of missing data is less than fifteen percent but, in this case, we can witness that unknown values are much higher than fifteen percent.

We have used two approaches significantly variable deletion and single mean imputation. First we have delete unnecessary rows from the data then undergo variable deletion to remove the columns and used single mean imputation for filling the missing values with approximately normal distributions where all observations are clustered around the mean. This is the best approach as mean is the most preferred approach for estimating the unknowns in any real time scenarios as well.

(Q3) ADDRESSING COLLINEARITY BETWEEN PREDICTOR VARIABLES -

Multicollinearity increases the variance of the estimators and hence reduces the adequacy of the model. The columns of the dataset of World Development Indicators has multicollinearity present between its predictor variables(X). The solution β will be unstable as a small change in the data cause large changes in ***β.*** The easiest way for the detection of multicollinearity is to examine the correlation between each pair of explanatory variables. However, it might not be considered sufficient.

The second easy way for detecting the collinearity is to estimate the multiple regression model. As, the coefficient of determination in the regression of regressor X_j in the model, increases toward unity, the VIF (Variance Inflation Factor) also increases. Therefore, we can use VIF as an indicator of collinearity. VIF > 5 means there is collinearity between the variables, and it will affect the adequacy of the model as shown below. The 'mctest' package provides the Farrar – Glauber test for collinearity, the two functions 'omcdiag' and 'imcdiag' provides the overall and individual checking for multicollinearity.

```
> library(mctest)
> X< data2[,2:17]
> omcdiag(x=X,life)

Call:
omcdiag(x = X, y = life)


Overall Multicollinearity Diagnostics

                       MC Results detection
Determinant |X'X|:         0.0000         1
Farrar Chi-Square:      3051.2708         1
Red Indicator:             0.3961         0
Sum of Lambda Inverse:    81.2136         1
Theil's Method:           -2.3742         0
Condition Number:         72.5540         1

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

> imcdiag(x-X,y- life)

Call:
imcdiag(x = X, y = life)


All Individual Multicollinearity Diagnostics Result

                     VIF    TOL       Wi       Fi Leamer    CVIF Klein   IND1   IND2
elec              5.9543 0.1679  75.6349  81.3913 0.4098  0.2044     0 0.0110 1.2777
nat_income        1.4138 0.7073   6.3180   6.7988 0.8410 -0.0485     0 0.0463 0.4495
child droupouts   2.3502 0.4255  20.6135  22.1823 0.6523 -0.0807     0 0.0279 0.8822
mort              4.5713 0.2188  54.5223  58.6718 0.4677 -0.1569     0 0.0143 1.1997
gpop              1.7780 0.5624  11.8767  12.7806 0.7500 -0.0610     0 0.0368 0.6719
tpop             16.5385 0.0605 237.2214 255.2757 0.2459 -0.5677     1 0.0040 1.4428
pcomrate          2.8635 0.3492  28.4487  30.6139 0.5910 -0.0983     0 0.0229 0.9993
secedu            1.1033 0.9064   1.5769   1.6970 0.9520 -0.0379     0 0.0594 0.1438
teachers         16.6956 0.0599 239.6196 257.8564 0.2447 -0.5731     1 0.0039 1.4436
hexpen            2.4099 0.4150  21.5238  23.1619 0.6442 -0.0827     0 0.0272 0.8984
hexpenpcc         6.3559 0.1573  81.7674  87.9905 0.3967 -0.2182     0 0.0103 1.2940
rpop              2.3214 0.4308  20.1733  21.7086 0.6563 -0.0797     0 0.0282 0.8741
fert              3.4482 0.2900  37.3756  40.2202 0.5385 -0.1184     0 0.0190 1.0903
gdp               5.7723 0.1732  72.8573  78.4023 0.4162 -0.1982     0 0.0113 1.2696
msub              2.1014 0.4759  16.8154  18.0952 0.6898  0.0721     0 0.0312 0.8049
isub              5.5360 0.1806  69.2494  74.5198 0.4250 -0.1900     0 0.0118 1.2582

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
```

```
> X<-data2[,c(3:9,11,13:15,17)]
> cor(X)
                  nat_income child_droupouts        mort        gpop        tpop    pcomrate       secedu      hexpen        rpop        fert         gdp
nat_income       1.000000000     -0.14567844 -0.137189855 -0.15981321  0.241834135  0.09943632 -0.008579045  0.32812405 -0.16423674 -0.19235455  0.199178971
child_droupouts -0.145678439      1.00000000  0.507564279  0.31619775  0.038186972 -0.71133027 -0.028216584 -0.20496432  0.29097041  0.49445020 -0.388424464
mort            -0.137189855      0.50756428  1.000000000  0.49387451  0.008815807 -0.61796970  0.003102033 -0.20495591  0.44423255  0.74258367 -0.414863505
gpop            -0.159813211      0.31619775  0.493874506  1.00000000 -0.014372888 -0.33415617 -0.105380566 -0.25671745  0.24520063  0.52103639 -0.192728306
tpop             0.241834135      0.03818697  0.008815807 -0.01437289  1.000000000 -0.05152410  0.002838835 -0.05140592  0.09756871 -0.04080638 -0.094766939
pcomrate         0.099436319     -0.71133027 -0.617969701 -0.33415617 -0.051524097  1.00000000 -0.022772937  0.14516576 -0.33217235 -0.52796780  0.310898814
secedu          -0.008579045     -0.02821658  0.003102033 -0.10538057  0.002838835 -0.02277294  1.000000000  0.03847139  0.02435122 -0.13725969 -0.003433738
hexpen           0.328124054     -0.20496432 -0.204955914 -0.25671745 -0.051405920  0.14516576  0.038471385  1.00000000 -0.27735369 -0.25536594  0.200165604
rpop            -0.164236741      0.29097041  0.444232549  0.24520063  0.097568712 -0.33217235  0.024351224 -0.27735369  1.00000000  0.42059631 -0.634014780
fert            -0.192354552      0.49445020  0.742583674  0.52103639 -0.040806377 -0.52796780 -0.137259689 -0.25536594  0.42059631  1.00000000 -0.562693181
gdp              0.199178971     -0.38842446 -0.414863505 -0.19272831 -0.094766939  0.31089881 -0.003433738  0.20016560 -0.63401478 -0.56269318  1.000000000
isub             0.203472944     -0.50745136 -0.659194677 -0.48627436 -0.100422698  0.48347013 -0.005291262  0.32597627 -0.66958852 -0.68091037  0.729168204
                        isub
nat_income       0.203472944
child_droupouts -0.507451358
mort            -0.659194677
gpop            -0.486274359
tpop            -0.100422698
pcomrate         0.483470133
secedu          -0.005291262
hexpen           0.325976271
rpop            -0.669588524
fert            -0.680910368
gdp              0.729168204
isub             1.000000000
> omcdiag(x=X,life)

Call:
omcdiag(x = X, y = life)


Overall Multicollinearity Diagnostics

                       MC Results detection
Determinant |X'X|:         0.0031         1
Farrar Chi-Square:      1380.9186         1
Red Indicator:             0.3699         0
Sum of Lambda Inverse:    28.4901         0
Theil's Method:           -3.5589         0
Condition Number:         63.1675         1

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

> imcdiag(x=X,y= life)

Call:
imcdiag(x = X, y = life)

All Individual Multicollinearity Diagnostics Result

                   VIF    TOL      Wi      Fi Leamer    CVIF Klein   IND1   IND2
nat_income      1.2900 0.7752  6.1424  6.7856 0.8805 -0.0688     0 0.0366 0.4692
child_droupouts 2.3136 0.4322 27.8249 30.7387 0.6574 -0.1233     0 0.0204 1.1850
mort            3.2872 0.3042 48.4480 53.5215 0.5515 -0.1752     0 0.0144 1.4522
gpop            1.7054 0.5864 14.9416 16.5063 0.7658 -0.0909     0 0.0277 0.8633
tpop            1.1316 0.8837  2.7878  3.0797 0.9401 -0.0603     0 0.0417 0.2427
pcomrate        2.6455 0.3780 34.8556 38.5057 0.6148 -0.1410     0 0.0178 1.2982
secedu          1.0836 0.9229  1.7700  1.9554 0.9607 -0.0578     0 0.0436 0.1610
hexpen          1.4214 0.7035  8.9256  9.8603 0.8388 -0.0758     0 0.0332 0.6187
rpop            2.2261 0.4492 25.9716 28.6914 0.6702 -0.1187     0 0.0212 1.1495
fert            3.3876 0.2952 50.5728 55.8688 0.5433 -0.1806     0 0.0139 1.4710
gdp             3.2213 0.3104 47.0513 51.9786 0.5572 -0.1717     0 0.0147 1.4392
isub            4.7768 0.2093 79.9993 88.3769 0.4575 -0.2547     0 0.0099 1.6501
```

It is clearly seen that the variables - elec, tpop, teachers, hexpencc, gdp, isub have a VIF value > 5, so we need to exclude some of the variables to solve collinearity. The R output below clearly shows that after removing some variables it solves the problem of collinearity as the VIF is below 5.

There are several remedial measures to deal with collinearity such as Principal Component Regression, Ridge Regression, Stepwise Regression etc. However, in our case we will solve it with the exclusion of variables whose VIF values are above 5.

(Q4) MULTIPLE LINEAR REGRESSION MODEL FOR LIFE EXPECTANCY IN 2016.

The model we are suggesting for the problem is **Multiple Linear Regression (MLR)**, also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.

Justification:

The goal of MLR is to model the linear relationship between the explanatory variables and response variable. Here, in the given problem we have one response variable **life expectancy (Y)** and multiple predictor variables like nat_income, child_droupouts, mort, pcomrate, secedu, hexpen, rpop, fert, gdp, isub .etc (x1,x2,x3…xn).Since we are having the same kind of problem we can use the MLR to solve this problem.

$$\boxed{y = \alpha_{0+}\ \beta_0 X_1 + \beta_1 X_2 + \cdots + \beta_{n-1} X_n + \varepsilon}$$

where,

$i = n$ observations

y = dependent variable

X_i =explanatory variables

β_0 = y-intercept (constant term)

β_i = slope coefficients for each explanatory variable

Ɛ = the model's error term (also known as the residuals)

A Simple linear regression allows us to make predictions about one response variable based on one predictor variable. Linear regression can be used only where are two continuous variables i.e., dependant and independent variable. But the MLR extends this to multiple independent variables. MLR examines how multiple independent variables are related to one dependent variable.

Implementation of the model:

After imputation and correlation matrix the 10 variables that were considered are nat_income, child_droupouts, mort, pcomrate, secedu, hexpen, rpop, fert, gdp, isub.

MLR for the above variables with Life Expectancy as the response variable is computed and the summary is analysed.

```
Coefficients:
                  Estimate Std. Error t value             Pr(>|t|)
(Intercept)      54.066221   5.682463   9.515 < 0.0000000000000002 ***
log(nat_income)   0.231937   0.105488   2.199              0.02901 *
child_droupouts   0.015406   0.042509   0.362              0.71741
mort             -0.143456   0.018819  -7.623    0.000000000000893 ***
pcomrate          0.004870   0.026656   0.183              0.85522
secedu            0.274842   0.253949   1.082              0.28040
hexpen            0.193868   0.086949   2.230              0.02685 *
rpop              0.001939   0.013606   0.143              0.88681
fert             -0.032577   0.009927  -3.282              0.00121 **
log(gdp)          0.834955   0.445445   1.874              0.06229 .
isub              0.082760   0.015544   5.324    0.00000263971970 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.209 on 206 degrees of freedom
Multiple R-squared:  0.8229,    Adjusted R-squared:  0.8143
F-statistic: 95.74 on 10 and 206 DF,  p-value: < 0.00000000000000022
```

The coefficient t-value is a measure of how many standard deviations our coefficient estimate is far away from 0. We want it to be far away from zero as this would indicate we could reject the null hypothesis - that is, we could declare a relationship between speed and distance exist. So, we remove the variables mort and fert whose t values are not far away from 0.

After this we are computing a model with the remaining predictor values, are nat_income, child_droupouts, pcomrate, secedu, hexpen, rpop, gdp, isub.

Then the coefficient of the summary of the model is analysed.

```
Coefficients:
                Estimate Std. Error t value        Pr(>|t|)    
(Intercept)     27.61245    5.91054   4.672 0.000005357853 ***
log(nat_income)  0.08702    0.12883   0.675        0.50013    
child_droupouts  0.02840    0.05260   0.540        0.58986    
pcomrate         0.09825    0.03109   3.161        0.00181 ** 
secedu           0.49677    0.30531   1.627        0.10522    
hexpen           0.31202    0.10617   2.939        0.00367 ** 
rpop             0.01964    0.01649   1.191        0.23492    
log(gdp)         2.21274    0.51415   4.304 0.000025852720 ***
isub             0.12546    0.01863   6.736 0.000000000156 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.981 on 208 degrees of freedom
Multiple R-squared:  0.725,     Adjusted R-squared:  0.7144 
F-statistic: 68.54 on 8 and 208 DF,  p-value: < 0.00000000000000022
```

Our final model is represented as,

life = 27.612+ 0.087 **nat_income**+ 0.0284 **Child_droupouts** + 0.0982 **pcomrate** + 0.49677 **secedu**+ 0.312 **hexpen**+ 0.0196 **rpop** + 0.2127 **gdp** +0.125 **isub**

where,

y=life

β_0=27.612

β_1=0.087 X_1 = nat_income

β_2=0.0284 X_2 = Child_droupouts

β_3=0.0982 X_3 = pcomrate

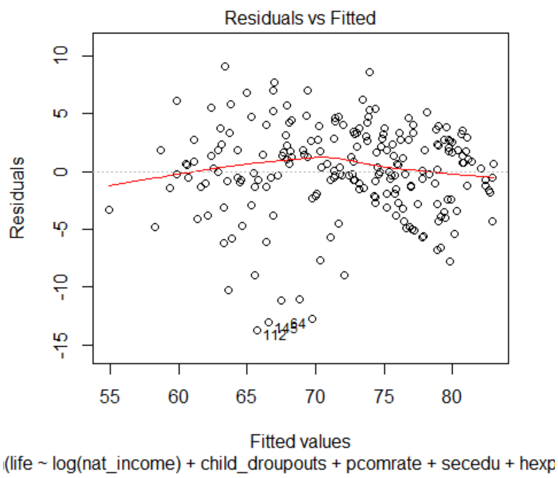
β_4=0.49677 X_4 = secedu

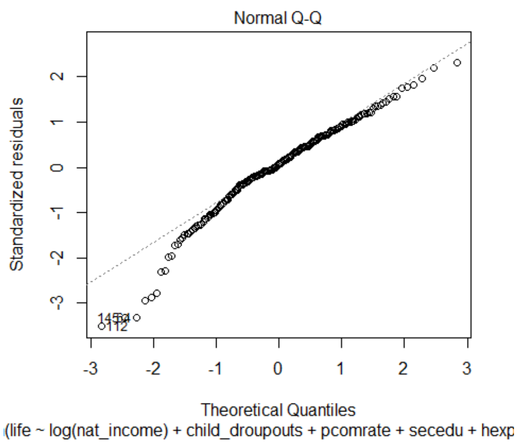
β_5=0.312 X_5 = hexpen

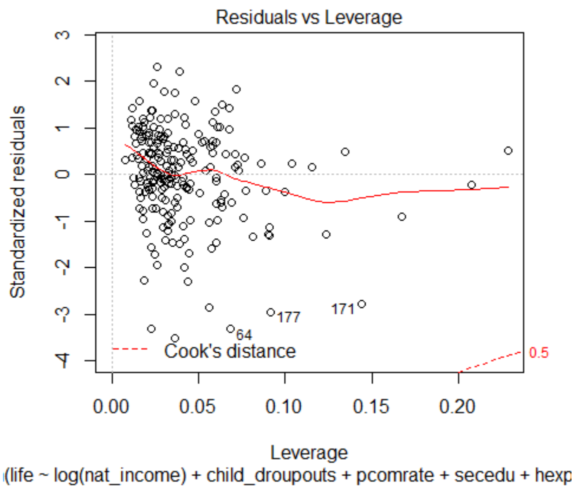
β_6=0.0196 X_6 = rpop

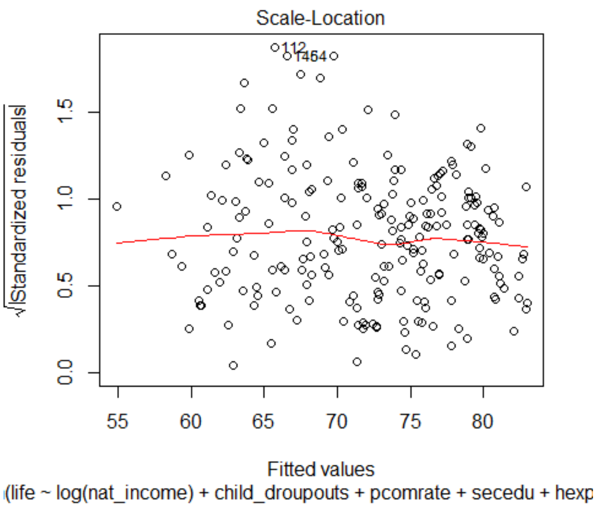
β_7=0.2127 X_7 = gdp

β_8=0.125 X_8 = isub









Model accuracy assessment:

Residual Standard Error (RSE), or sigma:

The RSE estimate gives a measure of error of prediction. The lower the RSE, the more accurate the model (on the data in hand).

Residual standard error: 3.981
sigma(finmodel)/mean(lifeexpec1$Life_Expectancy)

Error rate = 0.05506799 (5%)

The RSE is 3.981corresponding to 5% error rate

Q4.b)

Yes, the model can predict the life expectancy of the life expectancy of the countries that are not given in the table. Because the beta values are already identified so any country with all the predictor variables, we can easily predict the life expectancy of any country.

1. **Q(5) ANOVA to study differences of average life expectancies across continents.**

Below is the screenshot of the ANOVA table with the summary table of the analysis carried on the linear model of life expectancy versus continents grouped into 6 groups as factors.

```
> X <- read.csv("anova.csv")
> analysis <- lm(X$life_expect~as.factor(X$i..continent), data = X)
> anova(analysis)
Analysis of Variance Table

Response: X$life_expect
                           Df Sum Sq Mean Sq F value    Pr(>F)    
as.factor(X$i..continent)   5 7102.7 1420.54  61.351 < 2.2e-16 ***
Residuals                 211 4885.6   23.15                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We found that since probability value is very less, we can reject our null hypothesis that the average life expectancies is same for different continents at the confidence level of 95 % (alpha = 0.05)

Assumptions taken for ANOVA:

1. **Normality –** That each sample is taken from a normally distributed population
2. **Sample Independence –** That each sample has been drawn independently of the other samples
3. **Variance Equality –** That the variance of data in the different groups of continents are the same
4. **Continuous Dependent variable –** here, Average life expectancy should be continuous – that is, measured on a scale which can be subdivided using increments (i.e. years)

Hypotheses of Our One-Way ANOVA:

- The **null hypothesis (H0)** is that there is no difference between the groups of continents and equality between means.

  (Continents have the same average life expectancies.)

- The **alternative hypothesis (H1)** is that there is a difference between the average life expectancy of continent groups.

  (Continents have different average life expectancies)

Advantages of One way ANOVA:

- Provides the overall test of equality of group means.
- Controls the overall type I error rate (i.e. false positive finding).
- As the number of groups increases, the number pair comparisons increases substantially and calculations become overwhelming very quickly. If we test enough pairs, we begin to make observations that are less significant, until we find p values that are insignificant. ANOVA puts all the data into one F number and gives us one P to test the null hypothesis.
- Robust design
- Increases statistical power

2. APPENDIX

7.1 R Code for Data Summary and Plots.

Code for scatter plot

```
install.packages("psych")
install.packages("pastecs")
library(pastecs)
library(psych)
ggplot(finaldata2, aes(x=log(nat_income/tpop),y=life))+ geom_point()
ggplot(finaldata2, aes(x=log(gdp), y=life))+geom_point()
ggplot(finaldata2, aes(x=isub, y=life))+geom_point()

summary(finaldata2)
install.packages("plyr")
library(plyr)
```

code for descriptive statistics

```
t(apply(finaldata2, 2, function(x) c(mean=mean(x), sd=sd(x),min=min(x),max=max(x),quadl=quantile(x))))
```

7.2 R Code for Imputation.

```
library(DMwR)
library(dplyr)

data <- read.csv("C:/Users/HP/Desktop/modelling experimental data corsework1/LifeExpectancyFinal.csv",header=TRUE)
data37=data

#Replacing V1 and V2 columns
lifeexpec=data37[-1,]

#Change the name of data column
names(lifeexpec)[1] <- "Country"
names(lifeexpec)[2] <- "Country_Code"
names(lifeexpec)[3] <- "Continent"
names(lifeexpec)[4] <- "Life_Expectancy"
names(lifeexpec)[5] <- "Electricity_Consumption"
names(lifeexpec)[6] <- "Net_National_Income"
```

```
names(lifeexpec)[7] <- "Illiterate_Childrens"
names(lifeexpec)[8] <- "Children_Education_Expenditure"
names(lifeexpec)[9] <- "Water_Sanitation_Expenditure"
names(lifeexpec)[10] <- "Mortality_Rate"
names(lifeexpec)[11] <- "Literacy_Rate"
names(lifeexpec)[12] <- "Population_Growth"
names(lifeexpec)[13] <- "Total_Population"
names(lifeexpec)[14] <- "Agegroups_Completion_Rate"
names(lifeexpec)[15] <- "Secondary_Education_Duration"
names(lifeexpec)[16] <- "Secondary_Education_Teachers"
names(lifeexpec)[17] <- "Health_Expenditure"
names(lifeexpec)[18] <- "Percapita_Health_Expenditure"
names(lifeexpec)[19] <- "Total_Unemployment"
names(lifeexpec)[20] <- "Youth_Employment"
names(lifeexpec)[21] <- "Rural_Population"
names(lifeexpec)[22] <- "Fertility_Rate"
names(lifeexpec)[23] <- "GDP_Per_Capita"
names(lifeexpec)[24] <- "Mobile_Subscriptions"
names(lifeexpec)[25] <- "Internet_Usage"

#Finding out incomplete cases
lifeexpec[!complete.cases(lifeexpec),]

#counting the rows
nrow(lifeexpec[!complete.cases(lifeexpec),])

#remove unecessary rows from 265 to 269
lifeexpec <- lifeexpec[-c(264,265,266,267,268,269),]

#Variable Deletion
# Remove columns with more than 50 percent missing values
half <- c()
for(i in 1:ncol(lifeexpec))
{
if(length(which(is.na(lifeexpec[,i]))) > 0.5*nrow(lifeexpec)) half <- append(half,i)
}
lifeexpec1 <- lifeexpec[,-half]
```

```
# impute all missing values of columns with mean
for(i in 1:ncol(lifeexpec1))
{
lifeexpec1[is.na(lifeexpec1[,i]), i] <- mean(lifeexpec1[,i], na.rm = TRUE)
}
```

7.3 R Code for Investigating Collinearity.

```
# To test the multicollinearity within the variables using F-G test with mctest package using function omcdiag and imcdiag
#and comparing the the predictor variables with the correlation function.
library(mctest)
X<-data2[,2:17]
omcdiag(x=X,life)
imcdiag(x=X,y= life)
cor(X)

# Checking the VIF factor by excluding various variables
X<-data2[,c(1,2:9,11:17)]
cor(X)
omcdiag(x=X,life)
imcdiag(x=X,y= life)

X<-data2[,c(2:9,11:17)]
cor(X)
omcdiag(x=X,life)
imcdiag(x=X,y= life)

X<-data2[,c(2:9,11:16)]
cor(X)
omcdiag(x = X,life)
imcdiag(x = X,y= life)

X<-data2[,c(3:9,11,13:17)]
cor(X)
omcdiag(x=X,life)
imcdiag(x=X,y= life)
```

```
# final set of predictor after exclusion of variables with VIF > 5
X<-data2[,c(3:9,11,13:15,17)]
cor(X)
omcdiag(x=X,life)
imcdiag(x=X,y= life)
```

7.4 R Code for Multiple Linear Regression Modelling.

```
#code for MLR
library(Hmisc)
library(tidyverse)
library(highcharter)
library(dbplyr)
library(dplyr)
data<-read.csv('LifeExpectancyFinal.csv')
data1=data[4:25]
names(data1)<-
c("life","elec","nat_income","child_droupouts","exp_phealth","inv_waterandsan","mort","litrate","gpop","tpop
","pcomrate","secedu","teachers","hexpen","hexpenpcc","tunemp","yunemp","rpop","fert","gdp","msub","isub
")
sum=summary(data1$life)
names(data1)
column=c(1,2,3,4,7,9,10,11,12,13,14,15,18,19,20,21,22)
data2<-data1[column]
summary(data2)
attach(data2)
d1<-data[,c(1,3)]
#col1=c(1,3,4,5,6,7,10,12,13,14,15,16,17,18,21,22,23,24,25)
#data2<- data[col1]
for(i in 1:ncol(data2)){
  data2[,i]=ifelse(is.na(data2[,i]),
                  ave(data2[,i],FUN=function(y) mean(y, na.rm = TRUE)),
                  data2[,i])
}

d2<-cbind(d1,data2)
finaldata<-d2[1:217,]
finaldata
```

nrow(finaldata)op

ncol(finaldata)

head(finaldata)

names(finaldata)

##Nat_income, child_dropouts, Mort, pcomrate, rpop, fertility, GDP, isub,Sevedu,Hexpen

clonm=c(3,5,6,7,10,11,13,15,16,17,19)

mydat<-finaldata[clonm]

mydat

names(mydat)

library(tidyverse)

#MLR for the variables that are left after correlation

model <- lm(life~log(nat_income)

+child_droupouts+mort+pcomrate+secedu+hexpen+rpop+fert+log(gdp)+isub, data = mydat)

options(scipen=999)

summary(model)

#removing the variables Mortality_Rate,Fertility_Rate because their t values are not far away from 0. So we are neglecting them.

#final model:

finmodel <- lm(life~log(nat_income) +child_droupouts+pcomrate+secedu+hexpen+rpop+log(gdp)+isub, data = mydat)

options(scipen=999)

summary(finmodel)

#plotting the model

plot(finmodel)

#Residual Standard Error

sigma(finmodel)/mean(mydat$life)

#0.05506799

7.5 R Code for ANOVA.

```
> X <- read.csv("anova.csv")
> analysis <- lm(X$life_expect~as.factor(X$ï..continent), data = X)
> anova(analysis)
Analysis of Variance Table

Response: X$life_expect
                           Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(X$ï..continent)   5 7102.7 1420.54  61.351 < 2.2e-16 ***
Residuals                 211 4885.6   23.15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```