Regression Analysis on Life Expectancy – Further Study

Brief Background:

(i) Initially, dataset had 2938 observations for 193 countries on 22 variables from the year 2000-2015.
(ii) Removal of Missing observations by mean of each variables for each country and then dropping the observations for which missing values still existed.
(iii) Created subset of the dataset for the variables - Percentage Expenditure, Total Expenditure, Population, GDP, Income
(iv) Carried out analysis on **full** dataset – **2128** observation for **133** countries on **6** variables.

Dependent Variable: Life Expectancy

Independent Variable: Percentage Expenditure (PE), Total Expenditure (TE), Population (Popl), GDP, Income (I)

Brief Observations:

(i) Life Expectancy has significant correlation with GDP and Income.
(ii) GDP and PE has high correlation about 0.934.

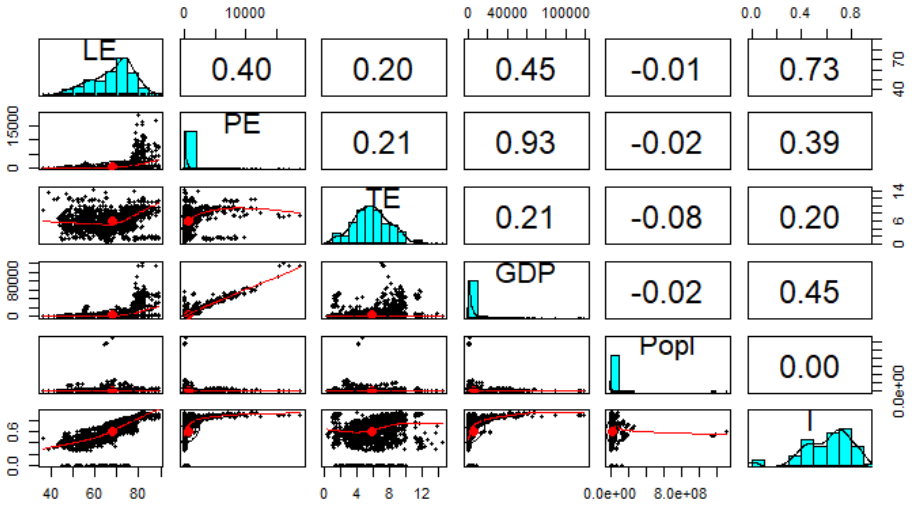

Figure 1- Correlation Plot of all variables

Model Fitting:

- Linear Model on all the variables.
  Observations:
  (i) Population is insignificant
  (ii) Adjusted R^2 = 55.56
  (iii) MSE = 41.0426

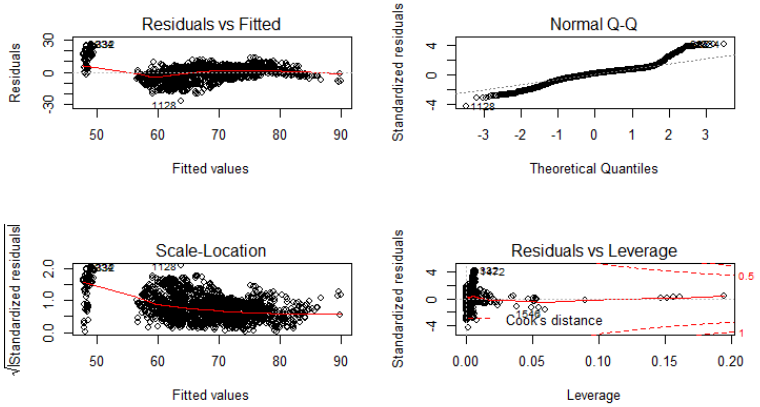

Figure 2- Linear Model Plot

Assumptions:
(i) Homogeneity of Variance – seems to hold
(ii) Normality – Violated

- Check for Multicollinearity –
  VIF values: PE = **7.9**, TE = 1.07, GDP = **8.5**, Population = 1.0068, Income = 1.29

1. Model 1 - **Removed high VIF variables** (PE and GDP)
   Observations:
   (i) Population becomes insignificant
   (ii) Adjusted R^2 = 54.21
   (iii) MSE = 42.29

2. Model 1_1 - Removed variable Population
   Observations:
   (i) Model: LE = 45.87 + 0.26TE + 34.122I
   (ii) Adjusted R^2 = 54.23
   (iii) MSE = 42.27
   (iv) R^2 predicted = 0.54

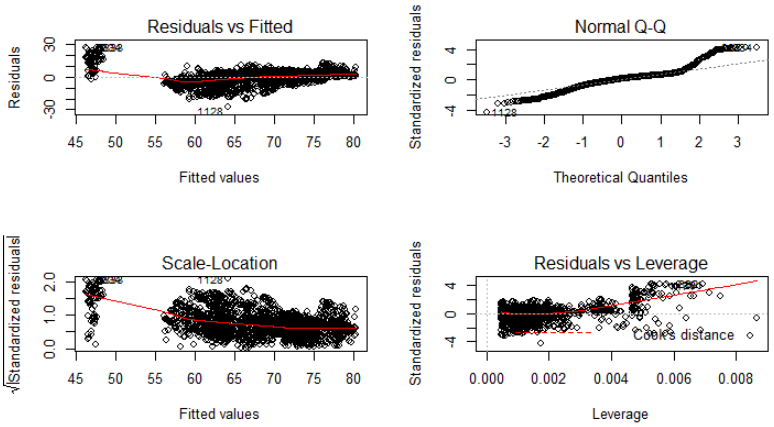

Figure 3- Model 1_1 Plot

3. Model 2 - **Ridge Regression**
   Observations:
   (i) PE, Population insignificant
   (ii) Model: LE = 1.876e-1 + 1.133e-4*GDP + 3.136e01*I
   (iii)Ridge Parameter = **0.001355**

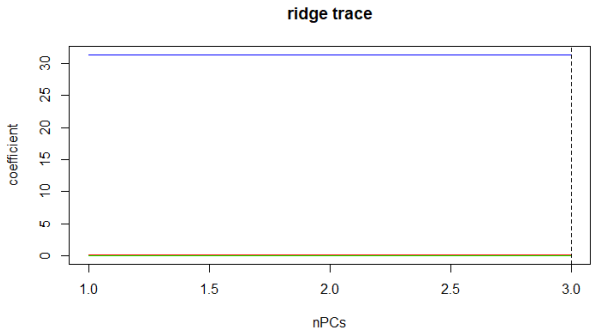

Figure 4 - Ridge Plot

4. Model 4 - **Stepwise Regression**
   Observations:
   (i) AIC = 7907.1
   (ii) Model: **LE = 4.735e1 + 1.873e-1*TE + 1.131e-4*GDP + 3.141e1*I**
   (iii)Adjusted R^2 = **0.556** (highest among all models)
   (iv)MSE = **41.011** (least among all models)
   (v) No further multicollinearity problem
   (vi)R^2 predicted = 55.38%

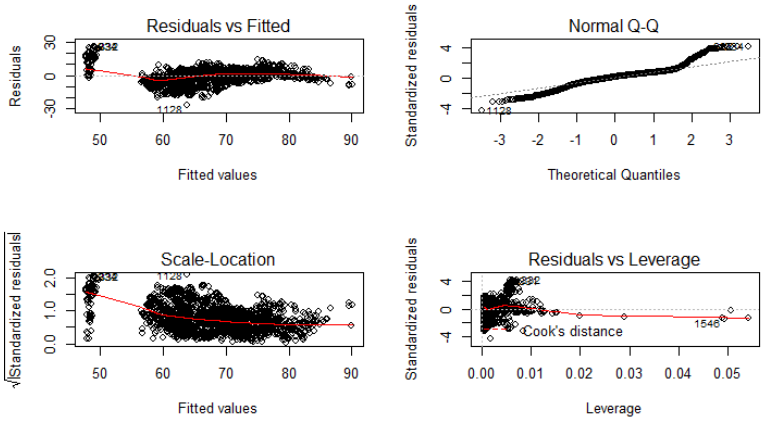

Figure 5 - Step wise Model Plot

- **Final Model** – Model obtained from Stepwise Regression

(i) Checking of assumption of final model - Normality assumption is **violated** from graph and from Shapiro test, p value<<0.05.

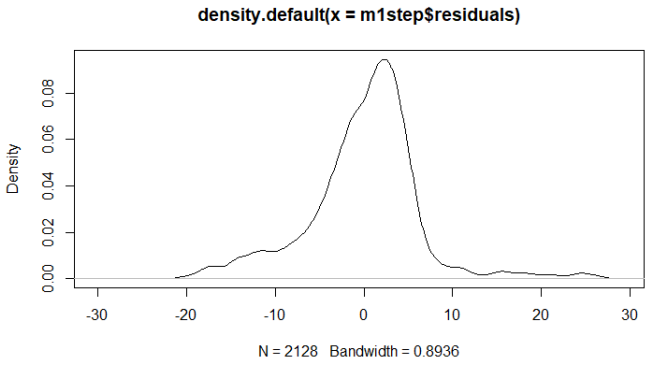

Figure 6- Residual Density Plot

(ii) **Multiple Influence Observations** present from Cook's Distance plot.

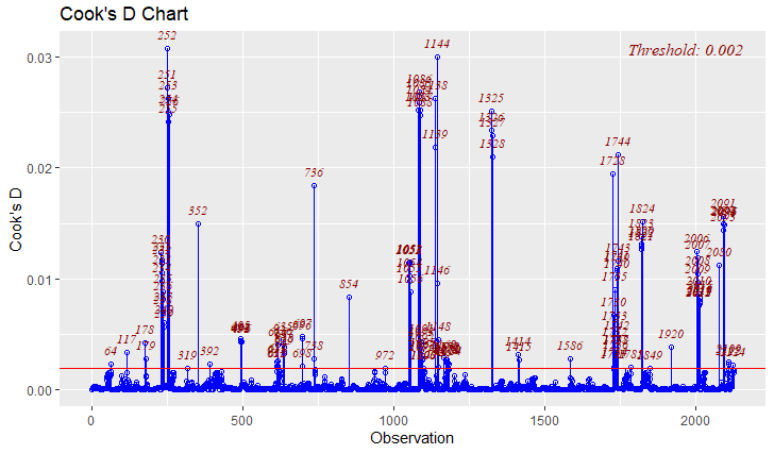

Figure 7 - Cook's D chart

Further Study:

(i) Transformation required for the model to hold Normality Assumption.
(ii) Remove Influence observations and observe the new model.
(iii) Carry out Partial Regression to observe how the independent variables have effect on Life Expectancy.

By Shweta Dutta.