**Linear Regression Assignment**

R Final Project

In this case study, we have considered data from year 2000-2015 for 193 countries for further analysis. The individual data files have been merged together into a single dataset.

Here we focus on Approach, steps, results found and their interpretation, significance of the variables and their business meaning etc.

1. Install required package and calling the same library at the beginning
2. Set the directory where my data is kept
3. Read the data for this project and store it in a variable i.e. DataFrame (taken a replica of the same)
4. Did basic exploration like

Structure: 2938 obs. of 23 variables. Displayed all columns and its datatypes followed by some starting values of each one.

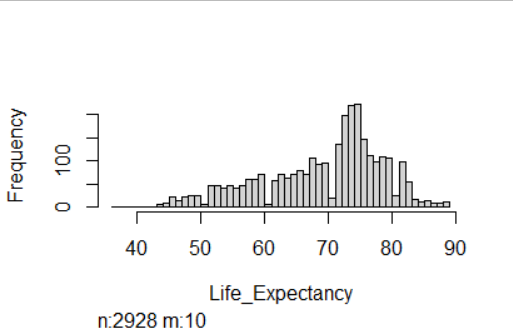
Summary: Obtained mean, median, min, max, 1st and 3rd quartile value for all continuous columns.

Dimension: 2938 23

1. Changed some variables as factor variable such as Year and status.
2. Renaming some column names for better understanding of mine. Those are Under\_five\_Deaths, HIV\_AIDS, Thinness1to19Years, and Thinness5to9Years.
3. Moved towards missing value treatment.

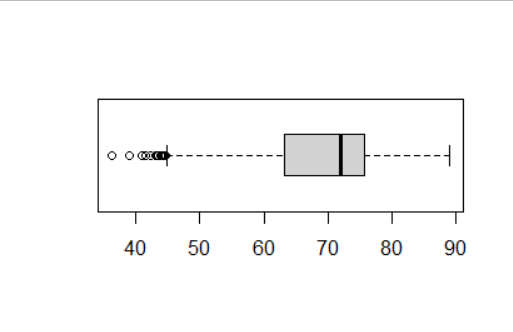
We have found lot more missing values. Our target column is Life\_Expectancy and contains 10 missing values.

First checking its pattern,



We can impute missing values by the mean of it. And did that.

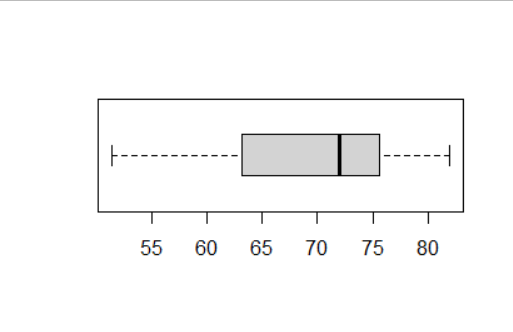
1. Then, moved towards Outlier checking for the target variable (life\_expectency).



We have checked the quantile method and seen the value in every quarterlies.

Treatment: replace by capping with the value 82 and replace by flooring with the value 51.4.

The boxplot after outlier detection:



1. Feature selection: ANOVA Test

It deals with probability significance to predict the target variable. It is only deals with categorical variables. Those are given below

Year: 1.05e-09 – Significant

Status: <2e-16 – Significant

\*\*\* Hence, keeping all columns for modelling.

1. Train test Split

Did splitting on the final data based on Life\_Expectancy variable with 70 and 30% ratio. The variable name for training and testing is original.data and test.data.

Dimensions are 2072 23 and 866 23

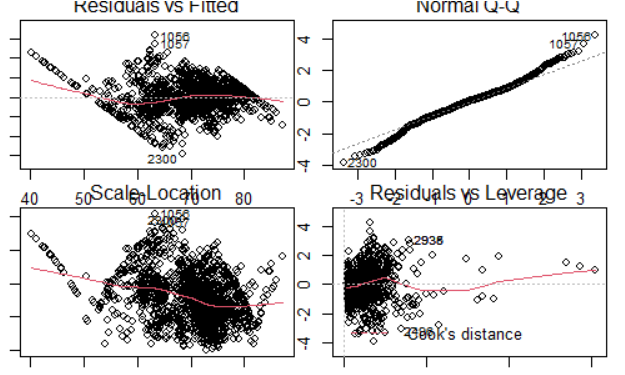
1. Omit other NA values form both the train and test data.
2. Linear Regression Model Building

Took all variables for the first time and one by one removing the insignificant variables based on p values. Significant predictors are

Year (2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014), Adult\_Mortality, BMI, Under\_five\_Deaths, Polio, Total\_Expenditure, Diphtheria, HIV\_AIDS, Per\_Capita\_GDP, Thinness1to19Years, Income\_Composition\_of\_Resources. All variables are significant at least 90% significance.

1. Multicollinearity check. Checked vif of the model and got values under or equal to 1.8 of each variable, which is the very good threshold of it.

Get the predicted or fitted values



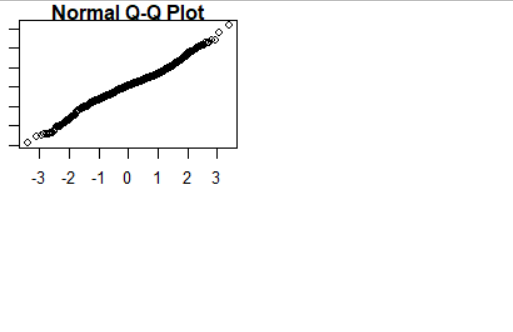
1. MAPE: 0.0427. Very good result
2. Autocorrelation: p value is 0. Since, the p-value is <0.05, we reject H0. It is Positive autocorrelation
3. Homoscedasticity test: p-value < 0.00000000000000022

Null hypothesis accepted and Homoscedasticity exists

1. Normality test:

Tested Anderson-Darling test, Cramer-von Mises test, Lilliefors (Kolmogorov-Smirnov) test, Pearson chi-square test, and Shapiro-Francia test

All p values are <0.000001. QQ plot is given below



1. Variable importance: Among the significant or potential variables, the top three significant predictors are

HIV\_AIDS, Adult\_Mortality, and Per\_Capita\_GDP.

1. Make predictions:

Prediction error, RMSE- 3.58

R-square - 0.7799865

Adj. R-square - 0.7720693

1. Interpretable coefficients of the final model is given below

Year 2007 -0.9222

Year 2008 -0.8008

Year 2009 -0.8450

Year 2010 -0.7206

Year 2011 -1.0960

Year 2012 -1.0153

Year 2013 -0.9366

Year 2014 -0.8716

Adult\_Mortality -0.0183

BMI 0.0455

Under\_five\_Deaths -0.0025

Polio 0.0202

Total\_Expenditure 0.1534

Diphtheria 0.0227

HIV\_AIDS -0.3883

Per\_Capita\_GDP 0.00008

Thinness1to19Years -0.1245

Income\_Composition\_of\_Resources 12.5374

All coefficients have their individual impact on the model. All other related interpretations are presented in PPT file.

Thanks and Regards,

Supriyo Barik