# **Linear Regression Assignment**

# Content

The dataset comes from the Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries. The data-sets are made available to public for the purpose of health data analysis. The data-set related to life expectancy, health factors for 193 countries have been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website. Among all categories of health-related factors only those critical factors were chosen which are more representative. It has been observed that in the past 15 years, there has been a huge development in health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, 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 data-set.

# **Objective**

## **Problem description:**

In this case study you’re provided average life expectancy of people of 193 Countries. You have to predict next year value using linear regression.

***Suggested Approach:-***

1. Understand the variables in the data set, study about the Industry and perceive the business objective.
2. Create hypothesis and validate.
3. Identify the statistical model to use (compare pros / cons of different models before accepting a model to follow) based on the kind of variables.
4. Clean the data set. That is, imputing missing values with the necessary values, detection of outliers and treatment of extreme outliers in the variables with the required method.
5. Run Regression model.
6. Check which variables are significant by looking at p values, conclude them accordingly. Observe variable relationships of predictor variables and target variables.
7. Do different tests like multicollinearity test, Homoscedasticity test, Normality test, serial correlation, MAPE (mean and median error percentage) and model accuracy.
8. Mention goodness of fit statistics (R-square, adjusted R-squares).

***Deliverables***

1. Codes file.
2. Approach, steps, results found and their interpretation, significance of the variables and their business meaning etc. in Word doc.
3. PowerPoint containing Objective, significant variables and results obtained in laymen language, business recommendations (if any).

# Variable Description

This data set contains the following features:

**'Country' :** Name of the nation

**'Year':** The year when the data is collected

**'Status'**: Developed or Developing status

**'Life Expectancy'**: Life Expectancy in age

**'Adult\_Mortality'** : Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)

**'Infant\_Deaths'**: Number of Infant Deaths per 1000 population

**'Alcohol'**: Recorded per capita (15+) consumption (in litres of pure alcohol)

**'Percentage\_Expenditure'**: Expenditure on health as a percentage of Gross Domestic Product per capita(%)

**'Hepatitis\_B'**: Hepatitis B (HepB) immunization coverage among 1-year-olds (%)

**'Measles'**: Number of reported cases per 1000 population

**'BMI'**: Average Body Mass Index of entire population

**'Under-five\_Deaths'**: Number of under-five deaths per 1000 population

**'Polio'**: Polio (Pol3) immunization coverage among 1-year-olds (%)

**'Total\_Expenditure'**: General government expenditure on health as a percentage of total government expenditure (%)

**'Diphtheria'**: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)

**'HIV/AIDS'**: Deaths per 1 000 live births HIV/AIDS (0-4 years)

**'GDP'**: Gross Domestic Product per capita (in USD)

**'Per\_Capita\_GDP'**: GDP per capita is gross domestic product divided by midyear population.

**'Population'**: Population of the country

**'Thinness\_1-1.9\_Years'**: Prevalence of thinness among children and adolescents for Age 1to 1.9 (%)

**'Thinness\_5-9\_Years'**: Prevalence of thinness among children for Age 5 to 9(%)

**'Income\_Composition\_of\_Resources'**: Human Development Index in terms of income composition of resources (index ranging from 0 to 1)

**'Schooling'**: Number of years of Schooling(years)