REGRESSION ANALYSIS PROJECT

by

Vivek Salunkhe (D001)

Nikhil Kesarkar (D016)

- Introduction:

This project primarily aims to predict a continuous dependent variable from a number of independent variables. It signifies the impact of each independent variable over the dependent variable and helps us to understand relationship between them.

- Dataset:

The dataset used to understand Regression Analysis is Life Expectancy by GHO (Global Health Observatory). GHO is a data repository under WHO keeps track of the health status as well as many factors for various countries. This data was collected from WHO and United Nations website. The data consists of various factors or variables that has some relation with Life Expectancy. The data from year 2000-2015 for 193 countries is used for analysis.

- Regression Analysis on Dataset:

The Life Expectancy Dataset includes a total of 22 variables and approximately 3000 observations pertaining to several countries. The major focus for our regression analysis is the variable **"Life Expectancy"** i.e. the dependent variable. We have used **Multiple Regression Analysis** technique for our project as the dependent variable is analyzed over several independent variables.

The following image provides information about the structure of the dataset that is taken into consideration for regression analysis:

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

```
1 Life_Expectancy_Data <- read.csv("C:/Users/vivek/Desktop/STUDY/Datasets/Life_Expectancy_Data.csv")
2
3 str(Life_Expectancy_Data)
```

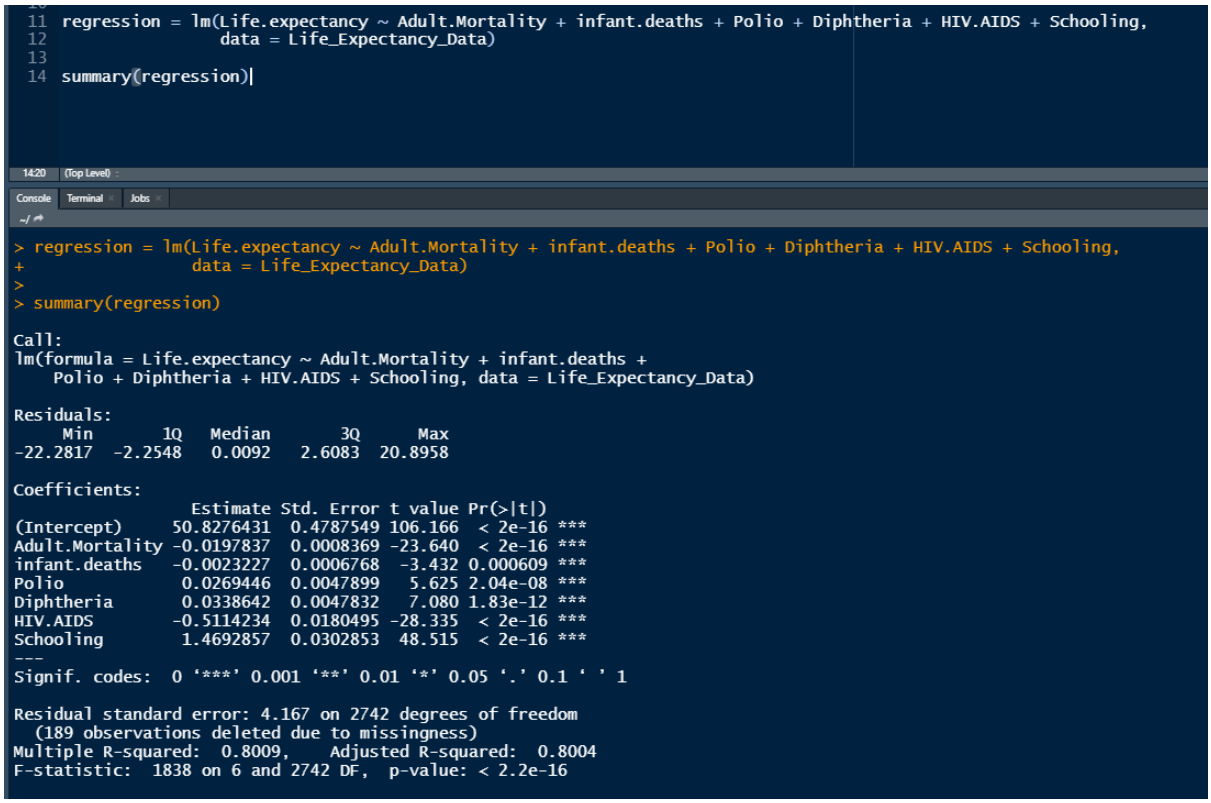
```
> Life_Expectancy_Data <- read.csv("C:/Users/vivek/Desktop/STUDY/Datasets/Life_Expectancy_Data.csv")
> str(Life_Expectancy_Data)
'data.frame':   2938 obs. of  22 variables:
 $ Country                        : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
 $ Year                           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
 $ Status                         : chr  "Developing" "Developing" "Developing" "Developing" ...
 $ 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 ...
>
```

Initially we analyzed our dependent variable Life Expectancy over all the relevant independent variables like Adult Mortality, infant deaths, Alcohol, Hepatitis B, Measles, Polio, Diphtheria, HIV/AIDS, GDP, Population and Schooling (ignoring the extra variables like Country, Year etc.).

```
5 regression = lm(Life.expectancy ~ Adult.Mortality + infant.deaths + Alcohol + Hepatitis.B + Measles + Polio+ Diphtheria
6                 + HIV.AIDS + GDP + Population + Schooling,
7                 data = Life_Expectancy_Data)
8
9 summary(regression)
10
```

```
> regression = lm(Life.expectancy ~ Adult.Mortality + infant.deaths + Alcohol + Hepatitis.B + Measles + Polio+ Diphtheria
+                 + HIV.AIDS + GDP + Population + Schooling,
+                 data = Life_Expectancy_Data)
> summary(regression)

Call:
lm(formula = Life.expectancy ~ Adult.Mortality + infant.deaths +
    Alcohol + Hepatitis.B + Measles + Polio + Diphtheria + HIV.AIDS +
    GDP + Population + Schooling, data = Life_Expectancy_Data)

Residuals:
     Min       1Q   Median       3Q      Max
-19.7919  -2.3546   0.1035   2.5415  12.3637

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.368e+01  6.995e-01  76.747  < 2e-16 ***
Adult.Mortality -2.062e-02  1.006e-03 -20.505  < 2e-16 ***
infant.deaths   -4.406e-03  1.259e-03  -3.499  0.00048 ***
Alcohol         -2.160e-02  3.181e-02  -0.679  0.49719
Hepatitis.B     -6.637e-03  4.798e-03  -1.383  0.16674
Measles          1.369e-05  1.137e-05   1.204  0.22872
Polio            9.303e-03  5.606e-03   1.660  0.09720 .
Diphtheria       2.547e-02  6.380e-03   3.993 6.82e-05 ***
HIV.AIDS        -4.582e-01  1.930e-02 -23.742  < 2e-16 ***
GDP              8.152e-05  9.857e-06   8.270 2.72e-16 ***
Population       2.698e-09  1.880e-09   1.435  0.15152
Schooling        1.430e+00  5.118e-02  27.943  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.928 on 1647 degrees of freedom
  (1279 observations deleted due to missingness)
Multiple R-squared:  0.8015,    Adjusted R-squared:  0.8001
F-statistic: 604.4 on 11 and 1647 DF,  p-value: < 2.2e-16
```

After analyzing our regression model, we eliminated few variables that had least significant impact on the dependent variable. We identified those variables by observing their p-values. Few variables like Alcohol, Hepatitis B, Measles, GDP, Population were having p-values greater than significant code value (α) 0.05. Hence, we refined and developed a new model after eliminating the above independent variables.



```
regression = lm(Life.expectancy ~ Adult.Mortality + infant.deaths + Polio + Diphtheria + HIV.AIDS + Schooling,
                data = Life_Expectancy_Data)

summary(regression)
```

```
> regression = lm(Life.expectancy ~ Adult.Mortality + infant.deaths + Polio + Diphtheria + HIV.AIDS + Schooling,
+                 data = Life_Expectancy_Data)
>
> summary(regression)

Call:
lm(formula = Life.expectancy ~ Adult.Mortality + infant.deaths +
    Polio + Diphtheria + HIV.AIDS + Schooling, data = Life_Expectancy_Data)

Residuals:
     Min       1Q   Median       3Q      Max
-22.2817  -2.2548   0.0092   2.6083  20.8958

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     50.8276431  0.4787549 106.166  < 2e-16 ***
Adult.Mortality -0.0197837  0.0008369 -23.640  < 2e-16 ***
infant.deaths   -0.0023227  0.0006768  -3.432 0.000609 ***
Polio            0.0269446  0.0047899   5.625 2.04e-08 ***
Diphtheria       0.0338642  0.0047832   7.080 1.83e-12 ***
HIV.AIDS        -0.5114234  0.0180495 -28.335  < 2e-16 ***
Schooling        1.4692857  0.0302853  48.515  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.167 on 2742 degrees of freedom
  (189 observations deleted due to missingness)
Multiple R-squared:  0.8009,    Adjusted R-squared:  0.8004
F-statistic:  1838 on 6 and 2742 DF,  p-value: < 2.2e-16
```

Finally, we predicted that our dependent variable **"Life Expectancy"** is significantly impacted by the independent variables Adult Mortality, Infant Deaths, Polio, Diphtheria, HIV/AIDS and Schooling.

The relationship between the independent and dependent variables is made clearer by the regression equation which is as follows:

Life Expectancy = 50.8276431 - 0.0197837(Adult Mortality) - 0.0023227(Infant Deaths) + 0.0269446(Polio) + 0.0338642(Diphtheria) - 0.5114234(HIV/AIDS) + 1.4692857(Schooling)

Conclusion: Life Expectancy can be marginally improved if there is a decrease in the rate of Adult Mortality, Infant Deaths and HIV/AIDS. Polio and Diphtheria vaccines are available thus they have significantly less impact on Life Expectancy. Also, improvement in Schooling (Knowledge) have provided a great positive impact on Life Expectancy of an individual.