**STATISTICS FOR DATA ANALYTICS**

**SUBMITTED BY:  
RASHMIKANT T SHUKLA**

## 1) Multiple Linear Regression

### Objective:

This analysis is an attempt to explore a dataset statistically by applying multiple regression under its various assumption. Finally giving a statistical interpretation of data based on various indicators in the analysis.

### Source of Data:

Data for this analysis is taken from World Health Organization’s (WHO) data repository. WHO is United Nation’s (UN) associate organization, which undertakes public health responsibilities of UN. WHO data repository has data of various health indicators of its member country.

### Context and Background of data:

WHO uses data from various national agency, surveys, its own campaigns and publish them in form of various indicator which are direct or indirect representation of health situation in member country. These indicators carry their own meaning and further information about how they are calculated, from which data they are derived can be found on the website.

For our analysis we have taken 9 indicators from year 2015 and built a dataset. One of them is life expectancy at birth and few indicators which can be contributing in overall life expectancy. So, idea is to explore underlying mathematical linear pattern between life expectancy and other indicators.

### Dataset:

Dataset has 80 entry for 9 Variable(each). Each row is depicting health related indicators for some country. Below table gives the detailed description of variables.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Descriptions** | **Measure Type** |
| Total\_Sanitation | Population using at least basic sanitation services(GHO, no date) | % of Total Population |
| Basic\_Drinking\_Water | Population Using at least basic Drinking Water Services(Global Health Observatory, 2015) | % of Total Population |
| Diarrhoeal\_Diseases | Number of Death by Diarrhoeal Diseases in age 1 to 59 months(WHO, 2016) | Count |
| CHE | Current Health Expenditure (CHE) as percentage of  Gross Domestic Product (GDP ) (WHO, 2015) | % of GDP |
| Medical\_Doctor | Medical Doctors(WHO, 2018) | Count |
| Maternal\_Deaths | Number of Maternal Deaths(World Health Organisation, 2015) | Count |
| DTP3\_Immunization | Diphtheria Tetanus Toxoid and Pertussis Immunization Coverage Estimates | Count |
| BMI | Mean body mass index (BMI) of defined population among Children aged between 5 to 19 years(World Health Organization, no date) | kg/m2 |
| Life\_Expectancy | Life Expectancy at birth(‘GHO | By category | Life expectancy and Healthy life expecancy - Data by WHO region’, no date) | Years |

Table-1.1

All variables are continuous in nature.

Life\_Expectancy is our dependent variable; all other variables are independent variable.

In initial analysis all other variables are taken as independent variable going forward that can change in final model depending on significance of variables.

### Hypothesis:

H0= There is no significant correlation between our chosen Dependent and independent variable.

H1=There is significant correlation between our selected dependent variable and independent variables.

At confidence level of 95%.

### Exploratory Data Analysis:

#### Missing Values and Descriptive Statistics:

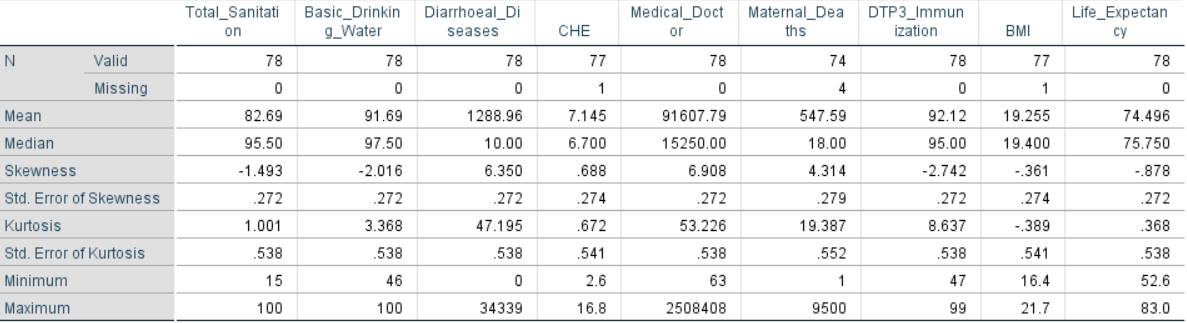


Figure-1.1

#### Missing Value Treatment:

Missing values of CHE and BMI are replaced by their respective mean since these variables are continuous, and their minimum value and maximum value are not drastically different.

While in Maternal\_Death range is quite spread out but approx. 70% observation are between 1 to 110. So missing value are replaced by mean of these 70% observation which comes out to be 22.

#### Checking Correlation between Variables:

Before proceeding with model, we are checking correlation to minimize the effect of multicollinearity and to ensure that our selected independent and dependent variable have considerable correlation.

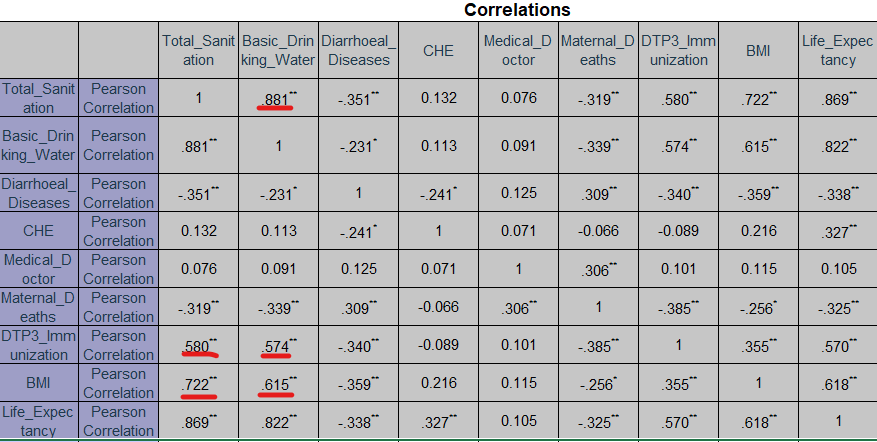
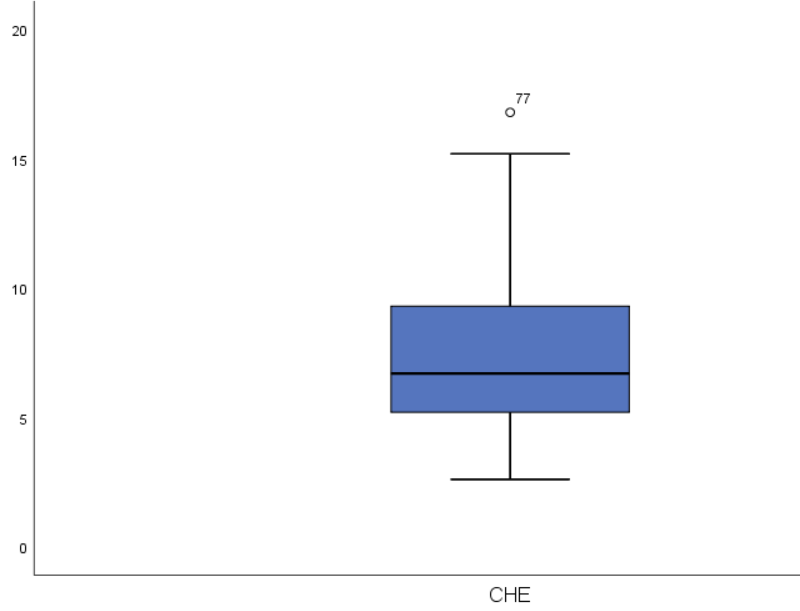
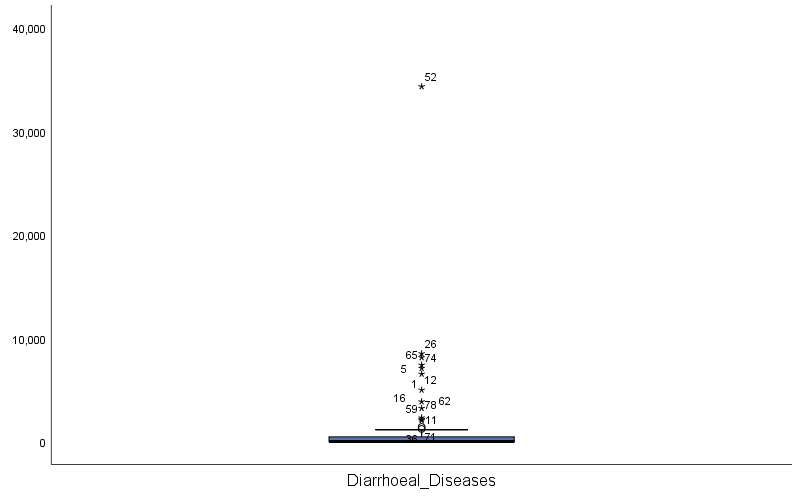
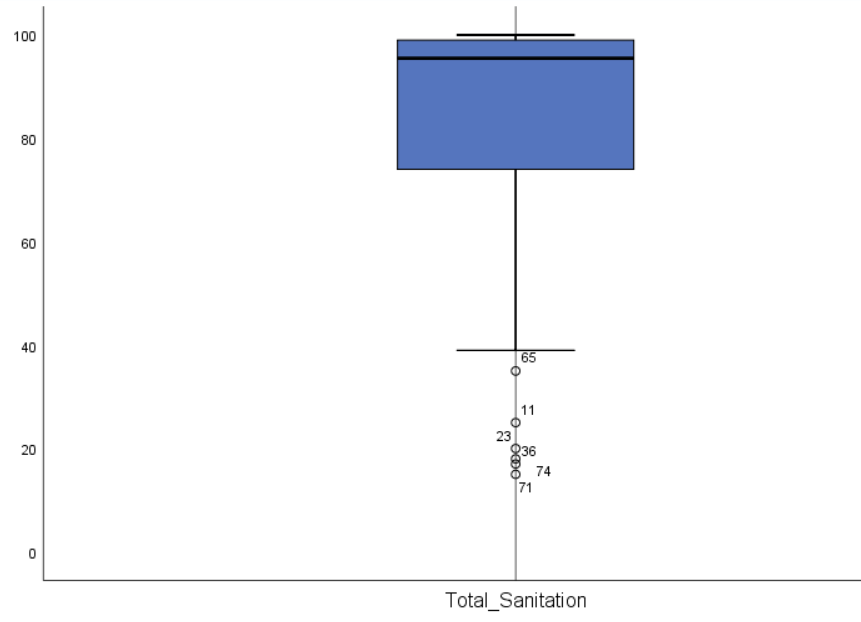


Figure-1.2

From above image we see Total\_Sanitation and Basic\_Drinking\_Water are highly correlated to each other and, to DTP3\_Immunization and BMI Variable so to prevent our model from violating multicollinearity assumption of Linear Multiple Regression, we are dropping Basic\_Drinking\_Water, BMI, DTP3\_Immunization variables from the analysis. Remaining variables have moderate to high correlation with our dependent variable.

#### Checking for Outliers:

We plotted boxplot for remaining variables



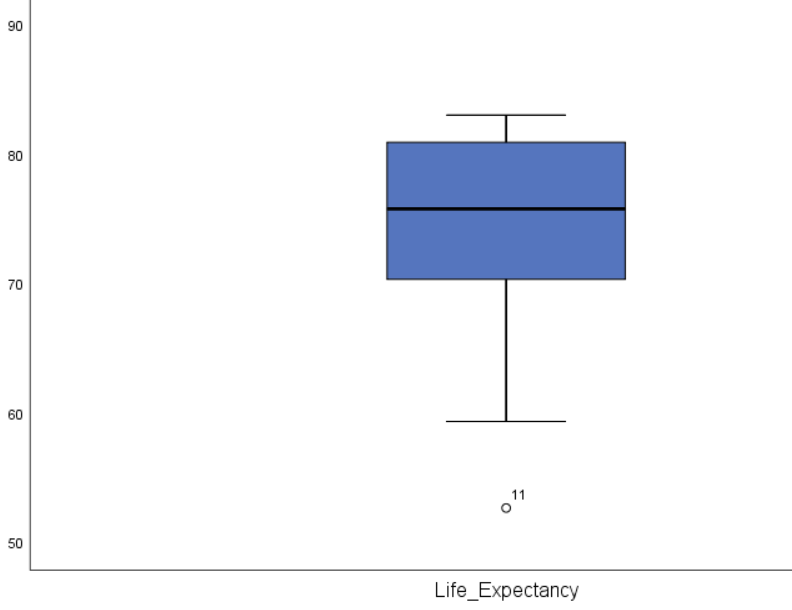
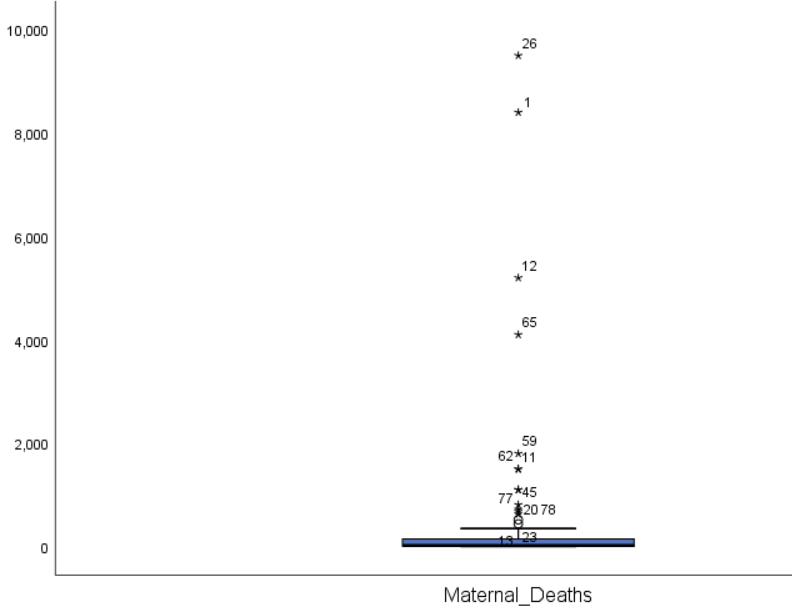
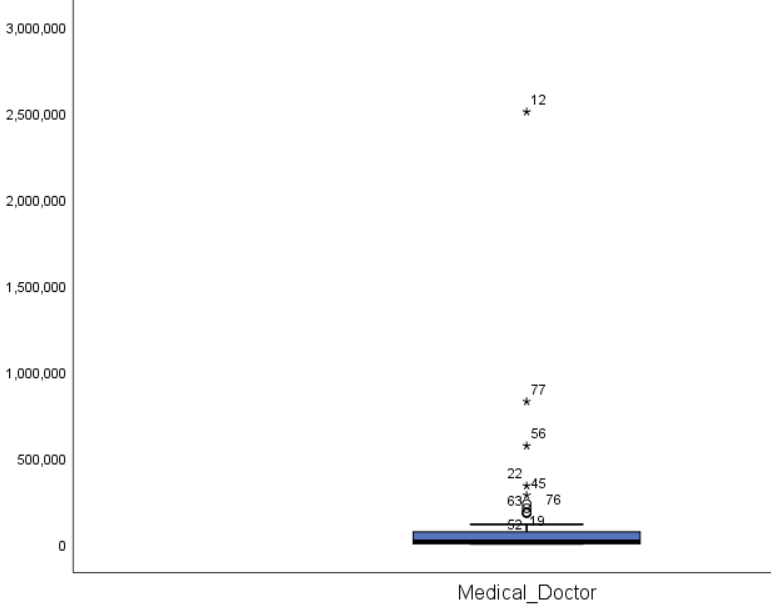


Figure-1.3

From above plots we see that Diarrhoeal\_Diseases, Medical\_Doctor, Maternal\_Deaths have many outliers, so we have used log transformation for Medical\_Doctor and Maternal\_Deaths while square root transformation for Diarrhoeal\_Disease, since it has many observations as “0” that’s why we avoided log transformation.

After Transformation box plots are as below

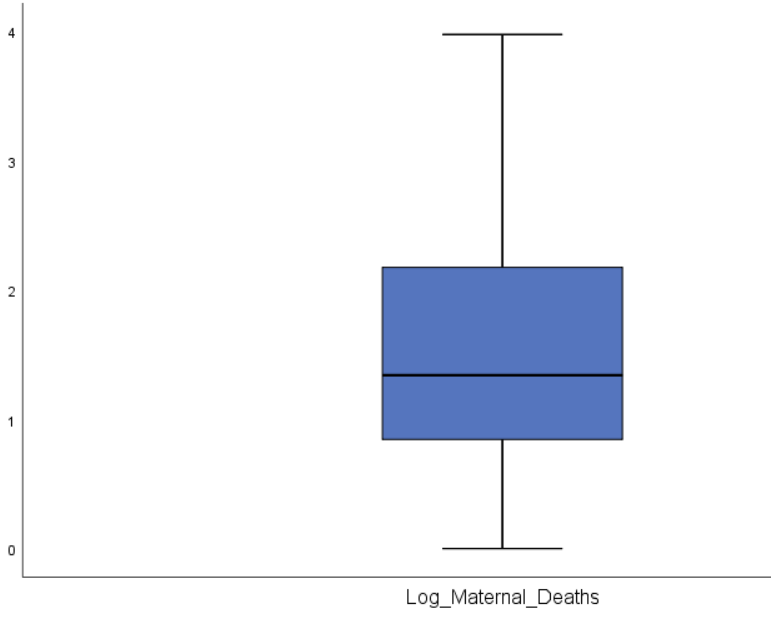
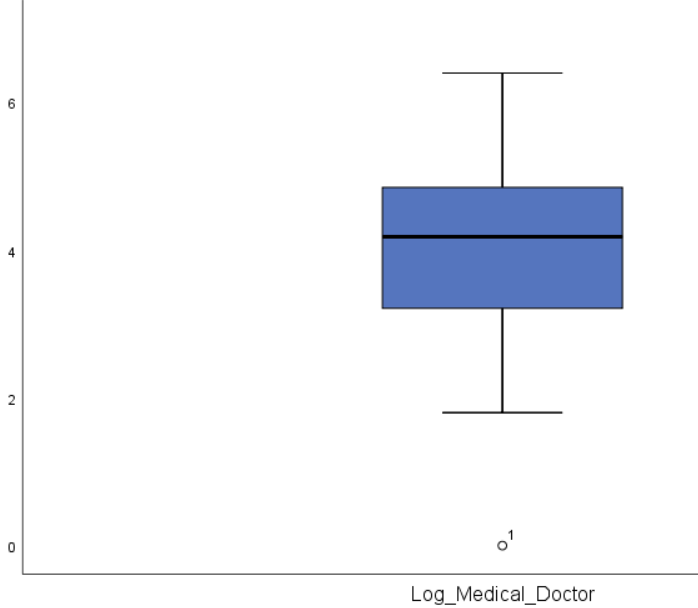
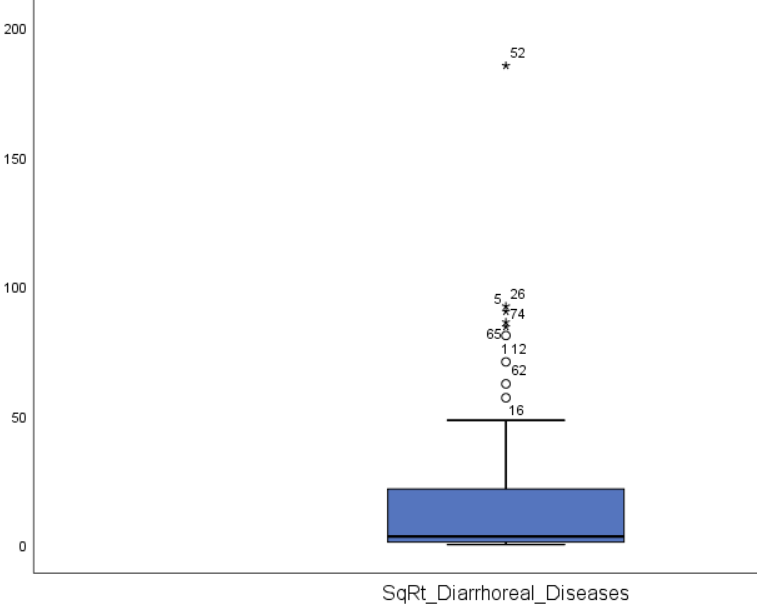


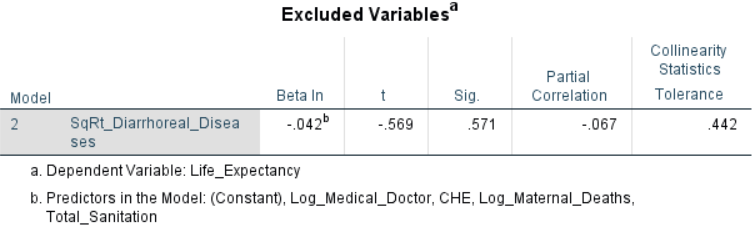
Figure-1.4

So, after transformations only Diarrhoeal\_Diseases have few outliers.

From different boxplots it is also clear that most of the variables are approximately normally distributed.

We are going forward with these variables for model creation.

### Model Creation:

We have used backward selection method for the model. It is one of the variable selection methods.

Diarrhoeal\_Diseases comes out to be non-contributing predictor in the model.

Figure-1.5

So, in final model Log\_Medical\_Doctor,CHE,Log\_Maternal\_Death and Total\_Sanitation are our predictor for response Life\_Expectancy

#### F-Statistics:

This means that at least one or more coefficients in the model are non-zero.

Figure-1.6

#### Model Summary:

This shows that 81.4 % variance in response is explained by our model.

Figure-1.7

#### Model Coefficient:

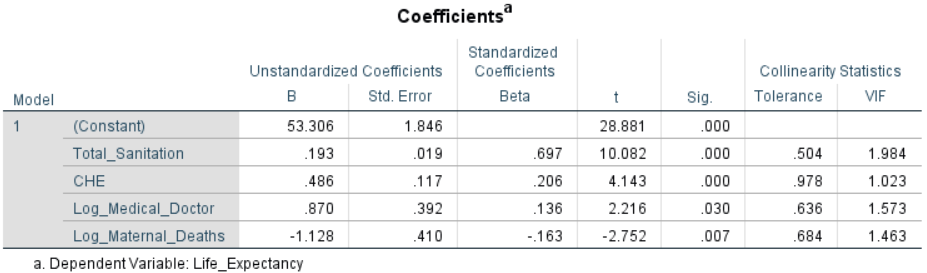


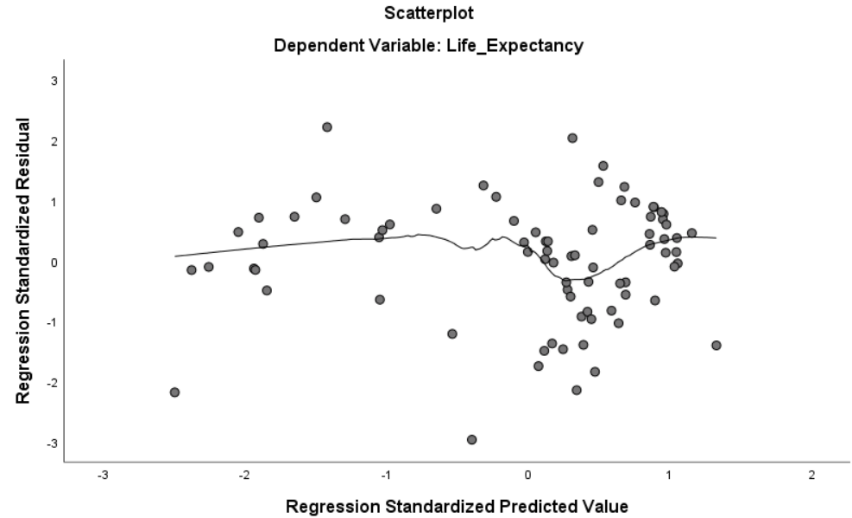
Figure-1.8

All the variable coming to be significant. We can write equation as

Life\_Expectancy=53.306+0.697(Total\_Sanitation)+0.206(CHE)+0.136(Log\_Medical\_Doctor)-0.163(Log\_Maternal\_Death)

### Assumptions:

**Linearity:** One of the assumptions is that predictor and response variable has linear relationship.

There is no pattern present that means this assumption is met.

**Homoscedasticity:** This assumption says that errors have approximately constant variance.

As residual are randomly distributed around 0 we can say this assumption is also met.

Figure-1.9

**Autocorrelation:** One of the assumptions is that Variables do not have any autocorrelation.

This can be tested by Durbin-Watson Test for our model it is 2.085 as value is close to 2. We can say this assumption is also met.

**Normality of the Error Terms:** Linear regression model assume that errors are normally distributed.

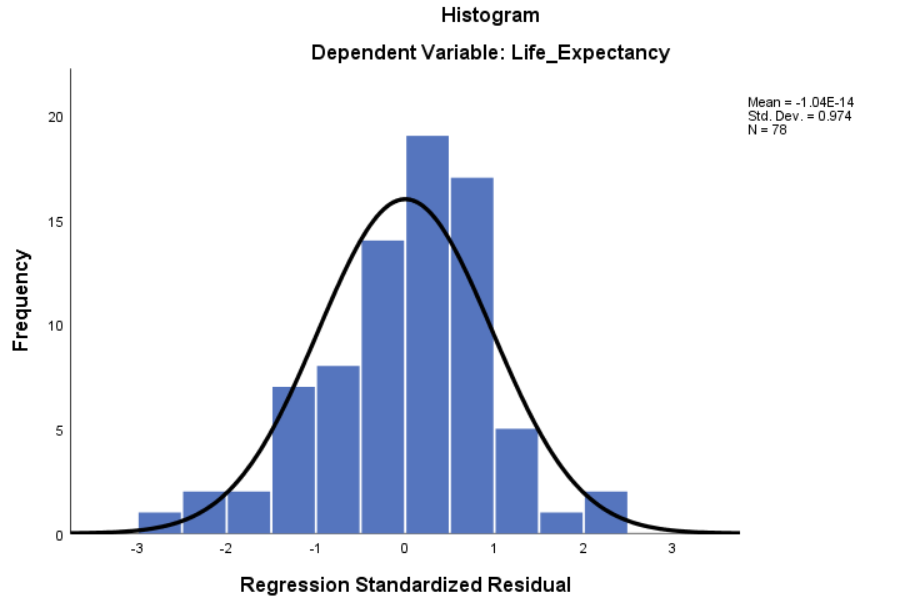


Figure-1.10

This shows the errors are normally distributed.

**Multicollinearity:** In initial phase of analysis we have dropped independent variables which had more correlations between them this same can be confirmed from VIF values. So, we can say there is no multicollinearity is present.

So, after our analysis we accept our alternative hypothesis H1.

### Result:

Multiple regression is applied to predict life expectancy at birth. The model used total sanitation, current health expenditure, number of medical doctors, maternal death as independent variable. Our model does not violate any assumption of linearity, normality, homoscedasticity and multicollinearity. Our model is statistically significant (F-statistics significant), that means model can predict Life expectancy. Our model can explain 81.4 % variance in Life expectancy (Adjusted R Square =0.814). As shown in ------ 4 variables are making statistically significant contribution into the model.

As Total sanitation is increased by 1 standard deviation life expectancy at birth increases by 0.697 std. deviation given all the other predictor are kept constant. Similarly, we can say life expectancy at birth increases by 0.206, 0.136 std. deviation by 1 std deviation change in current health expenditure and log (medical doctor) respectively given that all other predictors are kept constant.

For maternal death relationship is negative that means every 1 std. deviation increase in log (maternal death) decreases life expectancy at birth by 0.163 std. deviation.

## 2) Binary Logistic Regression

### Objective:

Aim of this analysis is to apply binary logistic regression on a dataset and state the result statistically.

### Source of data:

Data for this analysis is derived from the pew research center’s data repository. Pew research center is an American organization which conducts survey on public opinions on different issues.

Background: Our dataset is taken from STEM (Science, Technology, Engineering, Medical fields) survey of pew research center. This STEM survey explores different questions in American working adults (18 and above) like what attracted them to work in STEM field, how it is different from other field, is there a discrimination between STEM professionals based on race or gender, K-12 Education system is proficient enough for pursuing STEM careers, their personal interests in STEM field.

Out of these we have taken few columns with the intention of applying logistic regression to classify them in two distinct groups. Whether they are working in STEM filed or not.

### Dataset:

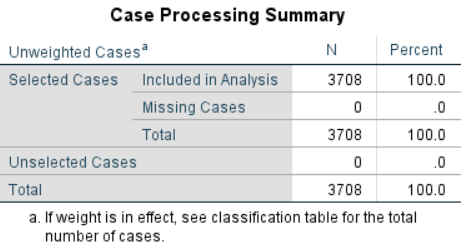
We have created a data set from by taking few of the columns from original file(Pew Research Center, no date). Below is the description of variables taken.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Type** | **Description** | **Levels** |
| SCH8a | Categorical | From k-12 level one liked science class. | 1-Liked 2-Disliked |
| SCH8b | Categorical | From k-12 level one liked maths class. | 1-Liked 2-Disliked |
| FAMSTEM1 | Categorical | Do you have any close family member working in STEM field? | 1-Yes 2-No |
| PPHHHEAD | Categorical | If person is household head | 1-Yes 2-No |
| PPGENDER | Categorical | Gender of person | 1-Female 2-Male |
| ppagecat | Continuous | Age of person | NA |
| HH\_INCOME\_col | Continuous | Income of person | NA |
| WORKTYPE\_FINAL | Categorical | Is person is working in STEM field. | 0-No 1-Yes |

Table-2.1

### Exploratory Data Analysis:

We are using Excel and SPSS for the model creation. Initial check is done in the Excel. There were few cases where respondents have chosen not to answer, we have removed those rows from the dataset.

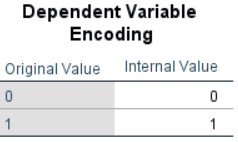


Our final dataset has 3708 rows and no missing values.

Figure-2.1

#### *Dependent Variable:*

In WORKTYPE\_FINAL variable originally had levels as 1-Non-STEM and 2-STEM.



We are considering variable as dependent variable so for easier interpretation in SPSS, we converted them 0-Non-STEM and 1-STEM, this can be seen in the figure (2.2).

Figure-2.2

#### Independent Variables:

Ppagecat and HH\_INCOME\_col variables are continuous in our data set.

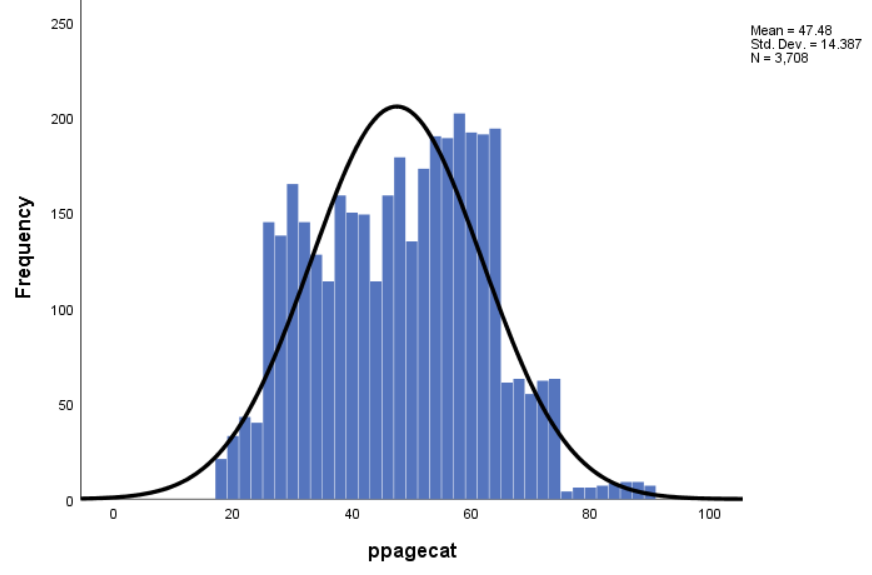
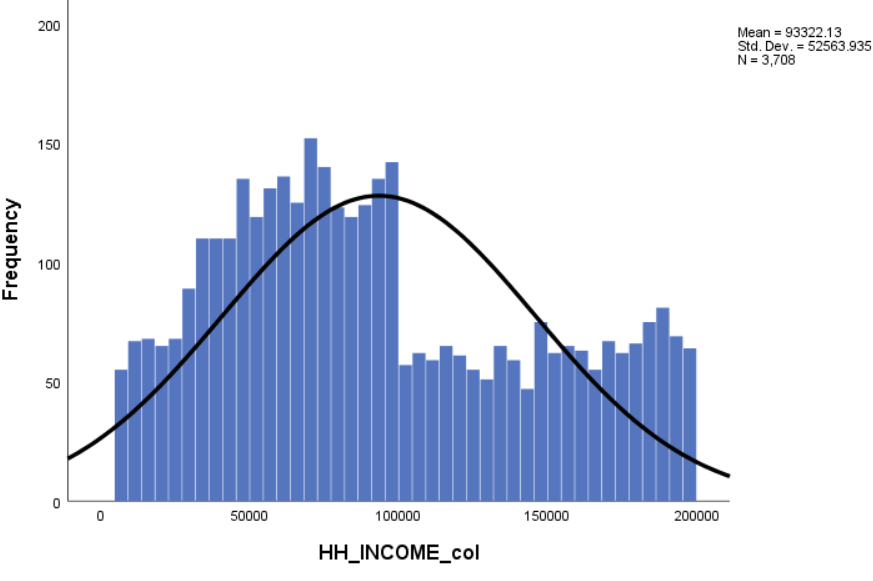
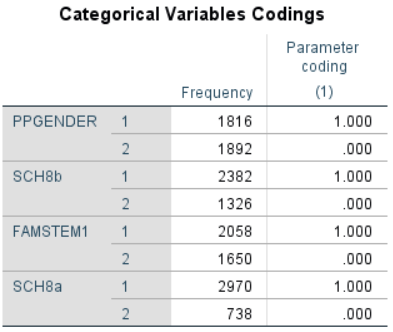
 

Figure-2.3

From above histogram, we can see they are normal and do not have outliers.

Below figure shows our categorical variables and their frequency they all have balanced

distribution except for SCH8a variable.

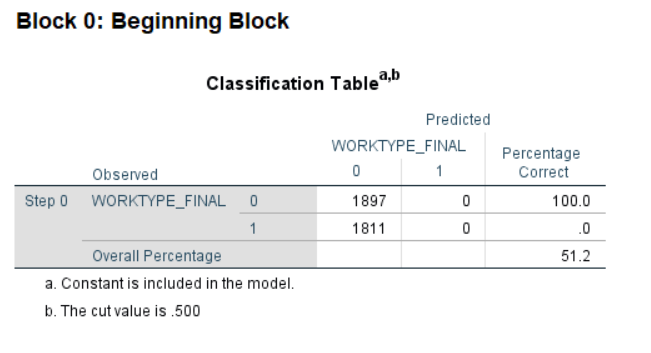
This table is also showing that how SPSS have encoded them for analysis, since all of them has two levels coded as 1 or 2.

Figure-2.4

### Model Creation:

#### Block-0:

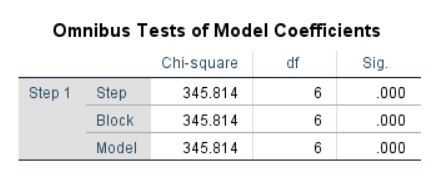
SPSS firstly have given Block 0 of model. This shows if we do not take effect of any variable and try to classify them then what will be the accuracy of it.



Even without considering effect of any variable we can classify 51% of our dependent variable correctly.

Figure-2.5

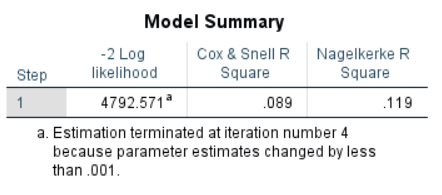
#### Block-1:

Omnibus Test is having somewhat similar interpretation as F-statistics in linear regression model.

it’s coming out to be significant that means our block 1 is making better classification than block 0. Our selected independent variable is contributing in model.

Figure-2.6

#### *Model Summary:*

 Negelkerke R square shows how much variance in dependent variable explained

from the model(Pallant, 2010, p. 174).

Figure-2.7

#### Goodness of Fit Test:

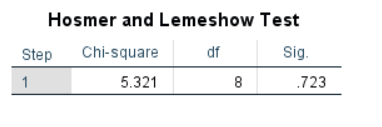
 if Hosmer and Lemeshow test significance >.05 then one can say model is good fit(Pallant, 2010, p. 174)In figure (2.8), we can see Significance is 0.723>0.05 so we can say our model is good fit for the data.

Figure-2.8

#### Accuracy:

Accuracy of model can be checked from classification table. Accuracy of our model is coming out to be 62.6 at cutoff level of 0.5 see figure (2.9).

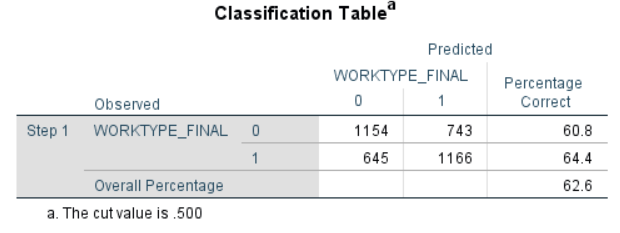
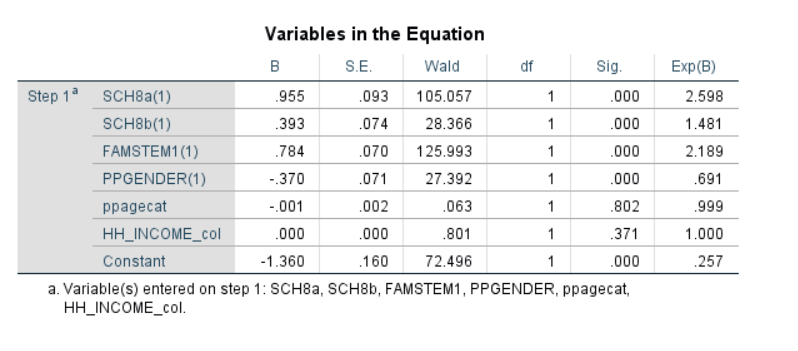


Figure-2.9

#### Model Coefficient:

Figure-2.10

Above figure gives coefficient of independent variable and their significance. Here other than ppagecat and HH\_INCOME\_col all the variables are contributing into the model. In this we are going to use Exp(B) values for interpretation.

### Result:

Binary logistic regression is applied to evaluate the effect of our selected independent variable on the likelihood of person working in STEM related field. Our analysis has 6 independent variables (See Table-2.1). Our final model came as significant since Chi-Square (6, N=3708) is significant (See Figure 2.6). This model explained 11.9% (Negelkerke R Squared Figure 2.7) of variance in person working in STEM field. Hosmer and Lemeshow test (see Figure 2.8) depict that our model is a good fit for used dataset.

From fig. (2.10) we see that HH\_INCOME\_col and ppagecat are not contributing in model as their p value in not significant. Other than this remaining variable is having statistically significant effect in final model.

In our model SCH8a and FAMSTEM are most prominent variables. If person liked science in their k to 12 level school, then their odds of working in STEM field is 2.5 time more than odds of working in non-STEM filed. Similarly, we can interpret one which are having some close family member working in STEM field have 2.189 more odds of working in STEM field.

For Female odds of Working in STEM field is decreases by factor of 0.691 in comparison to male.

## References:

GHO (no date) ‘Basic and safely managed sanitation services. Data by country’, *WHO*. World Health Organization. Available at: http://apps.who.int/gho/data/node.main.WSHSANITATION?lang=en (Accessed: 28 November 2019).

‘GHO | By category | Life expectancy and Healthy life expecancy - Data by WHO region’ (no date) *WHO*. Available at: http://apps.who.int/gho/data/view.main.SDG2016LEXv?lang=en (Accessed: 28 November 2019).

Global Health Observatory (2015) *Basic and safely managed drinking water services - Data by country*, *WHO Statistics*. Available at: http://apps.who.int/gho/data/node.main.WSHSANITATION?lang=en (Accessed: 28 November 2019).

Pallant, J. (2010) *SPSS survival manual : a step by step guide to data analysis using SPSS*. Fourth edition. Maidenhead : Open University Press/McGraw-Hill, 2010. Available at: https://search.library.wisc.edu/catalog/9910095047202121.

Pew Research Center (no date) *Datasets | Pew Research Center*. Available at: https://www.pewresearch.org/science/datasets/ (Accessed: 28 November 2019).

WHO (2015) ‘GHO | By category | Current health expenditure (CHE) as percentage of gross domestic product (GDP) (%) - Data by country’, *WHO*. Available at: http://apps.who.int/gho/data/node.main.GHEDCHEGDPSHA2011?lang=en (Accessed: 28 November 2019).

WHO (2016) ‘GHO | By category | Number of deaths by cause’, *Who*. Available at: http://apps.who.int/gho/data/view.main.ghe1002015-CH3?lang=en (Accessed: 28 November 2019).

WHO (2018) *GHO | By category | Medical doctors*, *World Health Organization*. Available at: http://apps.who.int/gho/data/node.main.HWFGRP\_0020?lang=en (Accessed: 28 November 2019).

World Health Organisation (2015) *GHO | By category | Maternal mortality - Data by country*, *WHO*. Available at: http://apps.who.int/gho/data/node.main.15?lang=en (Accessed: 28 November 2019).

World Health Organization (no date) ‘GHO | By category | Mean body mass index trends, age-standardized (kg/m2) - Estimates by country’, *Global Health Observatory*. Available at: http://apps.who.int/gho/data/node.main.NCDMBMIMEANC?lang=en (Accessed: 28 November 2019).