Analysis for Multiple regression (CA Part 1)

Amit Sahoo,

Student#18188851

Course :  [Statistics for Data Analytics](https://moodle.ncirl.ie/course/view.php?id=1016)

**1.Abstract :-**

Infant mortality is the number of infants(less than the age of 1) death in a year . It is measured in Infant mortality rate which signifies the number of infant’s death per 1000 live births . Protecting the lives of the new born has been a key issue for most the countries over a period of time .Almost all the countries are facing this serious problem due to several environmental factors like pollution , temperature change , civic factors like sanitation , and social factors like gender favouritism , birth spacing , education etc .

**2.Introduction :-**

Infant mortality number relates to a lot of factors predicting a countries failure in the field of health and development . A country with a smaller number of infant deaths is likely to show better result in breathing , heartbeat, umbilical cord pulsation or movement of voluntary muscles among the infants .Whereas countries with more Infant death rate is considered to lagging behind in the field of sanitation ,medical care, nutrition , education .

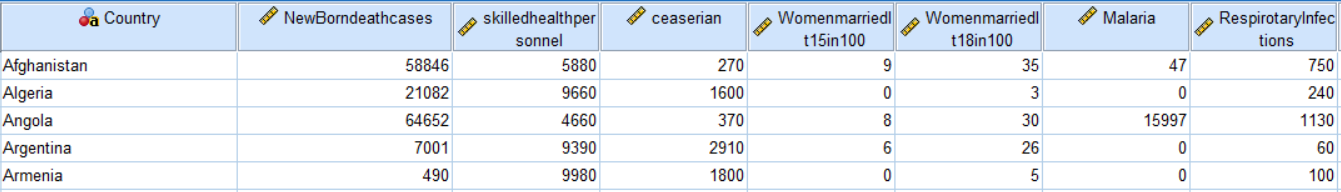
**3.Dataset Description :-**

This dataset has been derived from the WHO(World health organisation) portal(<https://www.who.int/health-topics/>) . In this study we are trying to check if the number of deaths reported in 105 countries ,is being resulted because of the following factors :-

1. Skilled health personal
2. No. of caesarean operated
3. Women married below 15 years(numbers in 100) .
4. Women married before 18yrs and after 15years(numbers in 100)
5. Number of malaria deaths reported .
6. Number of deaths due to respiratory infections .

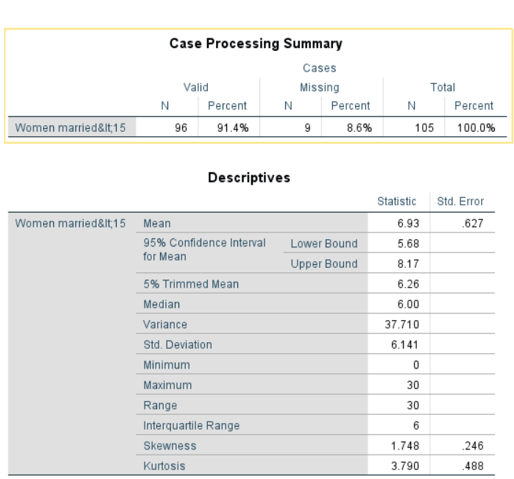
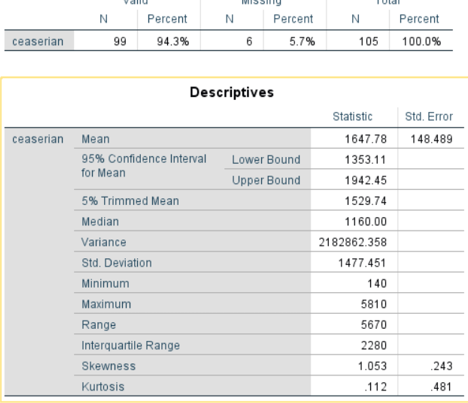
A sample size of 105 has been chosen for the study . The data has been combined from various sources (from the above portal) .

Below are the first five records from the dataset :-



**4.Missing value treatment :-**

Six and Nine missing values were observed(8.6 and 5.7 percent respectively of the total data) in the Caesarean and Womenmarriedlt15 respectively . Since there is no significant difference between mean and the 5%trimmed mean , the missing values are imputed with the 5% trimmed mean in both the above cases .

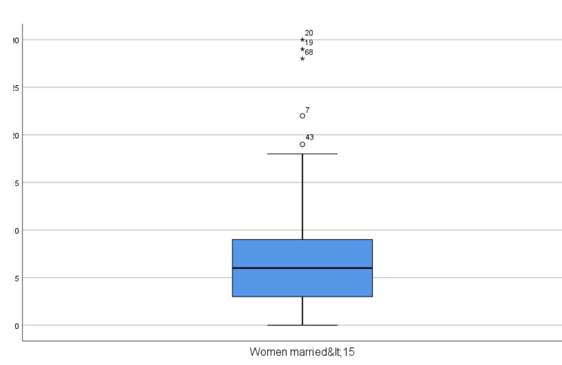
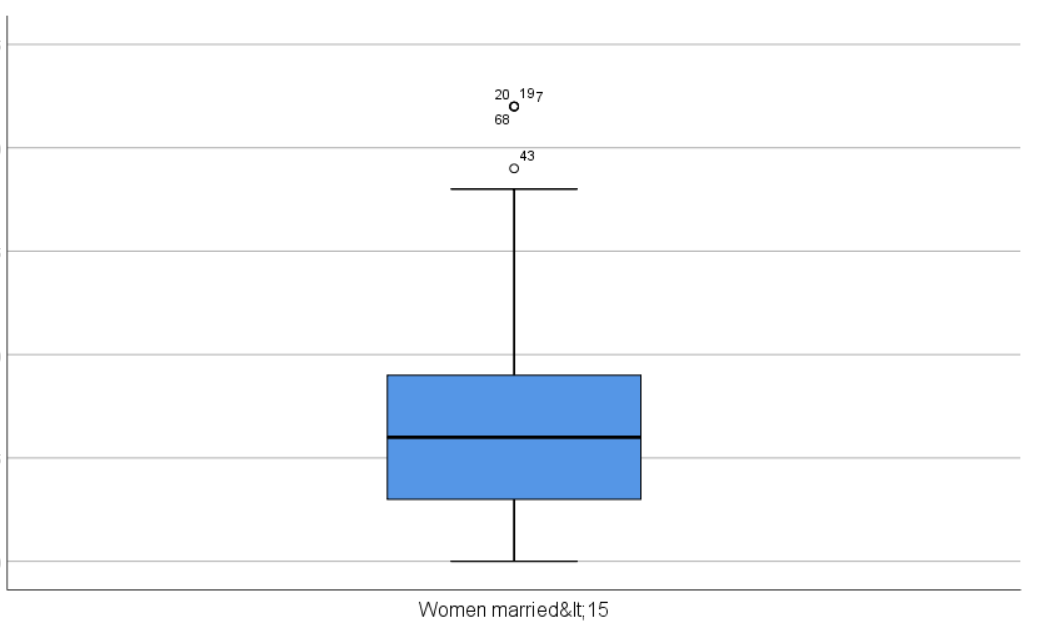


(Descriptive statistics for Number of caesarean ) (Descriptive statistics for Women married below 15)

**5.Outlier Treatment :-**

Based on the box plot analysis , we have observed few outliers in the below column :-

5.1 .Women married below the age of 15

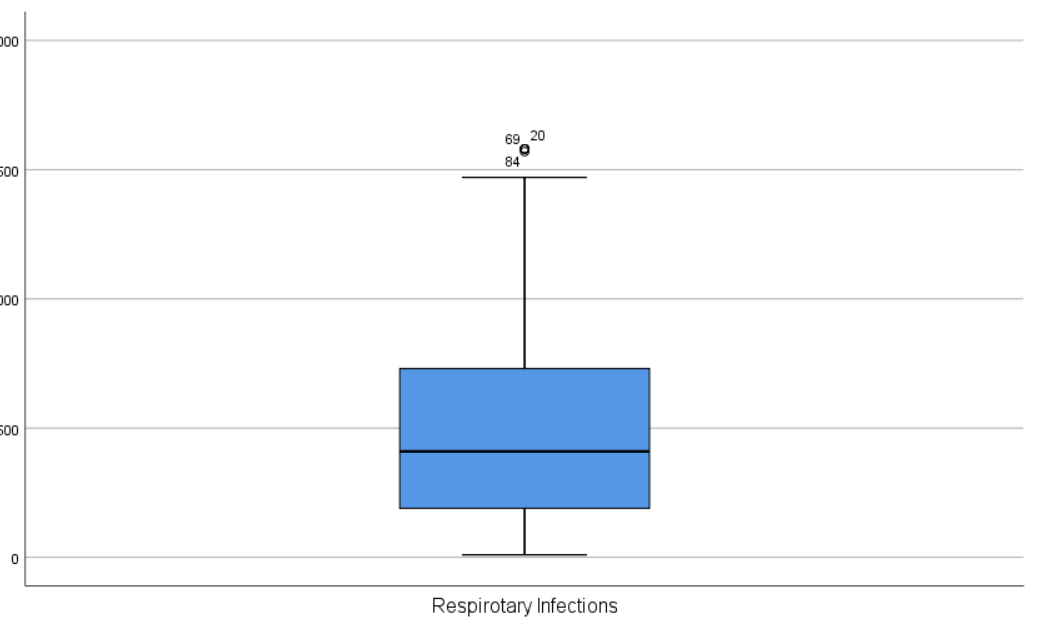
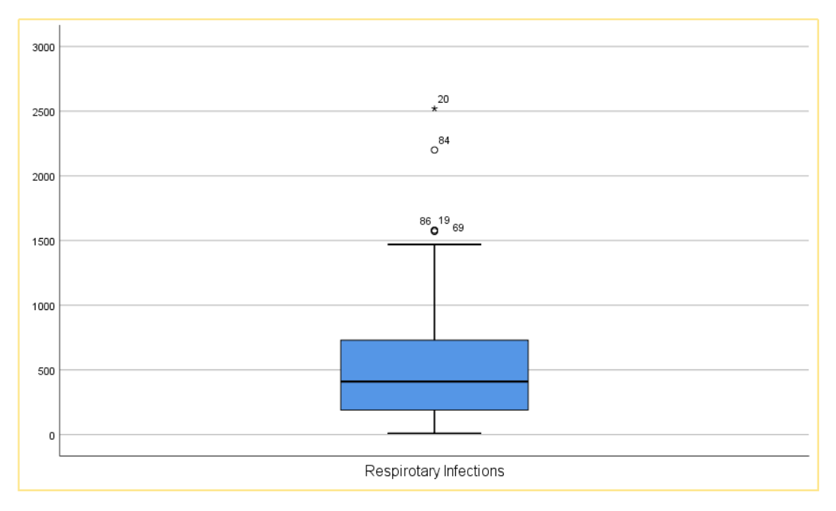
 

(Values before substituation) (Values after substituation)

Extreme outliers for 3 records(marked as ’\*’ )- record no. 20,19,68 has been substituted with the value of 7th record to minimise the overall effect on the complete dataset .

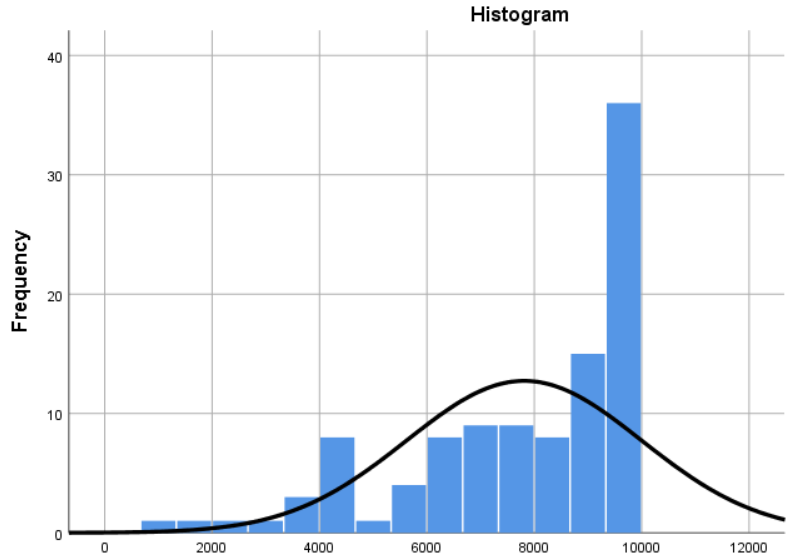
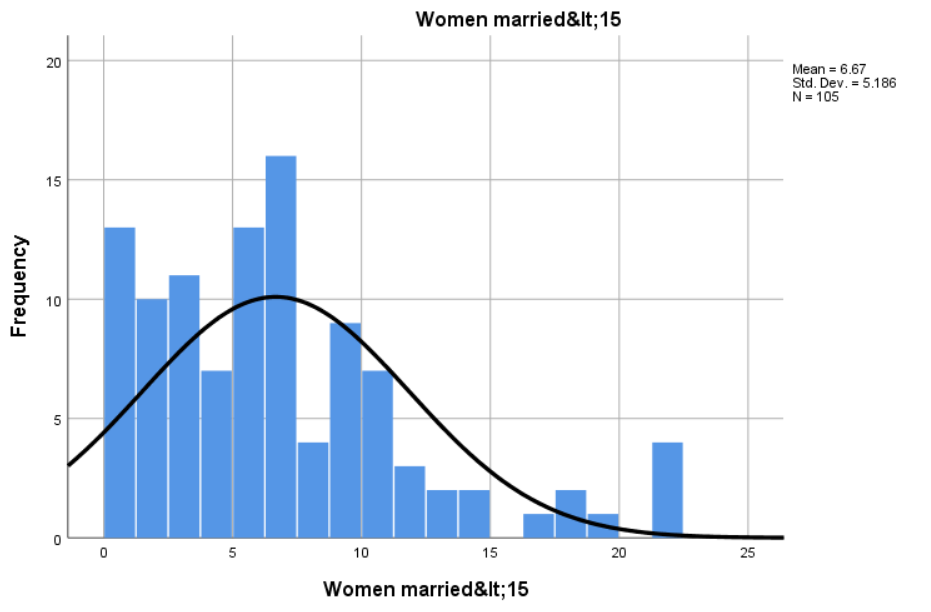
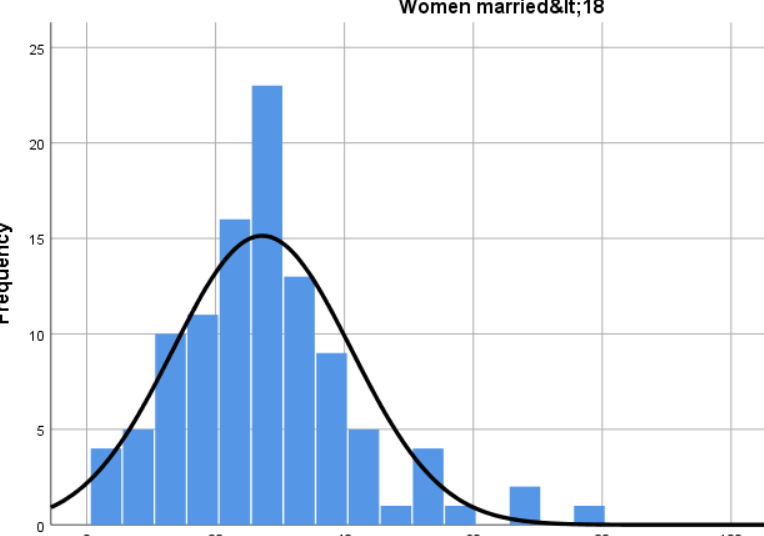
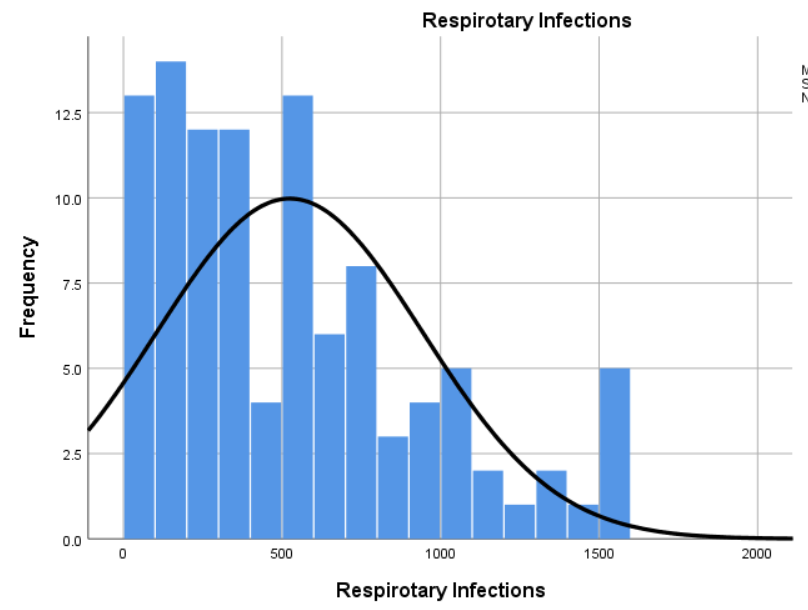
5.2 .No. of death due to Respiratory Infections :-

Similarly , there were two extreme outliers in 20th, 84th record , which has been substituted with values present in 19th row , to minimise the effect on the dataset as whole .



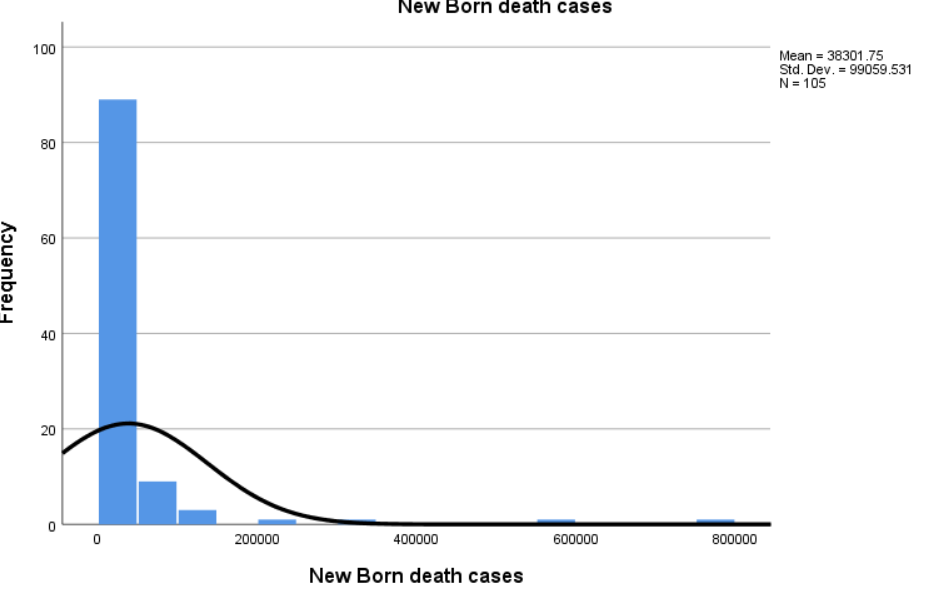
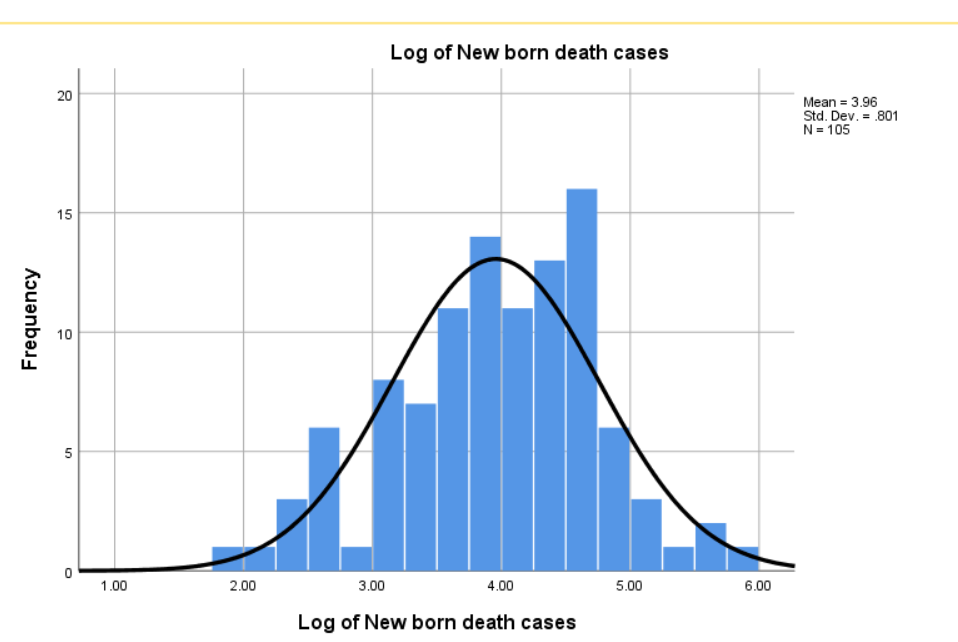
**6.Check for Assumption** (Statistical Consulting, n.d.)

6.1 Normal distribution of all the variables used :- Out the all the variable used in this study , below variables are found to be normally distributed (graphs shown below) :-

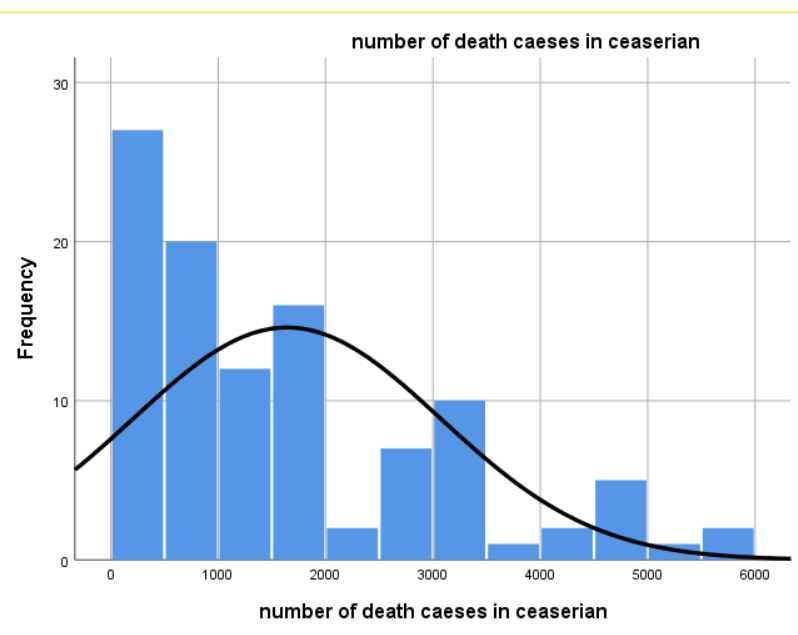
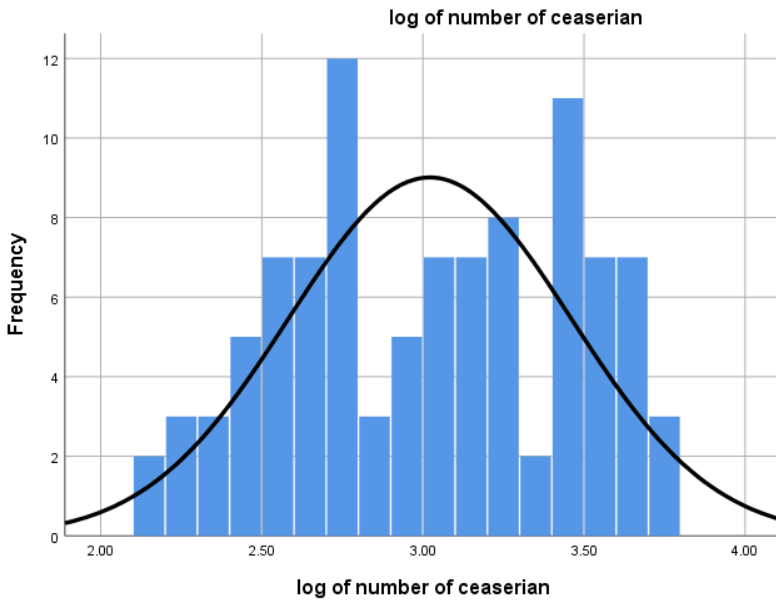
   

(Skilled health personal) (women married under 15) (women married within 15-18) (respiratory infection)

6.2Normality of the log of the variables used :- The below variables were NOT normally distributed . Hence, we took the log of the values to achieve the desired bell-shaped normality distribution .

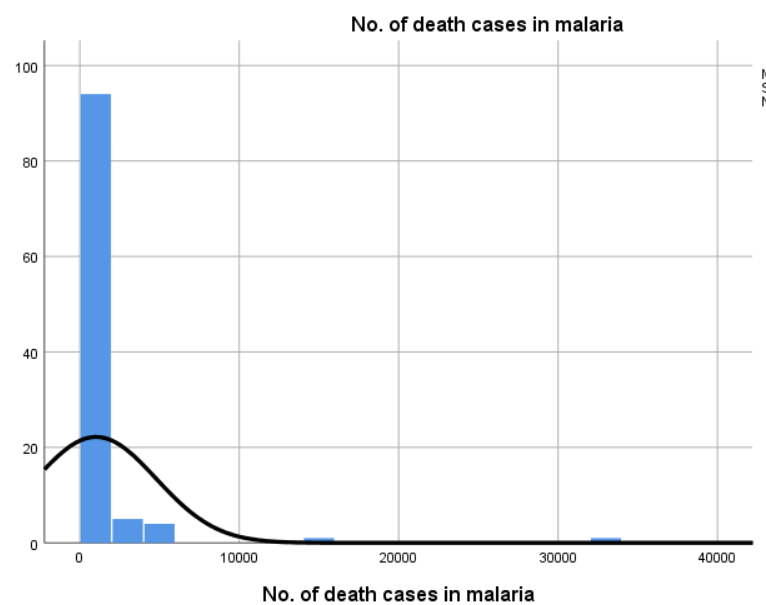
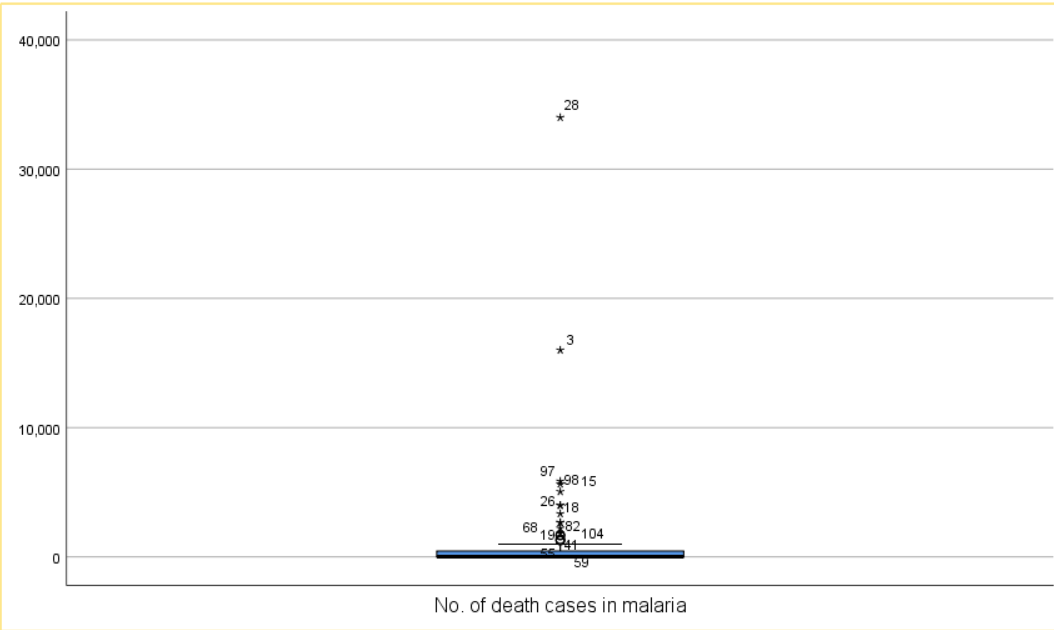
 

(New born death cases) (Log of new born death cases)

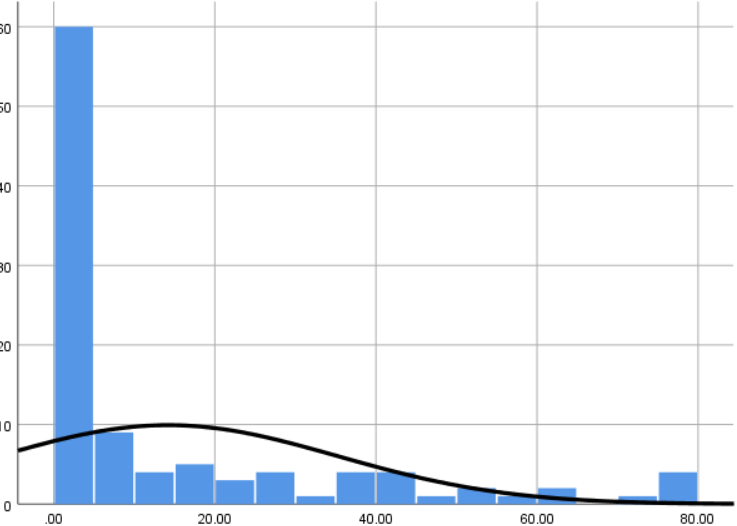
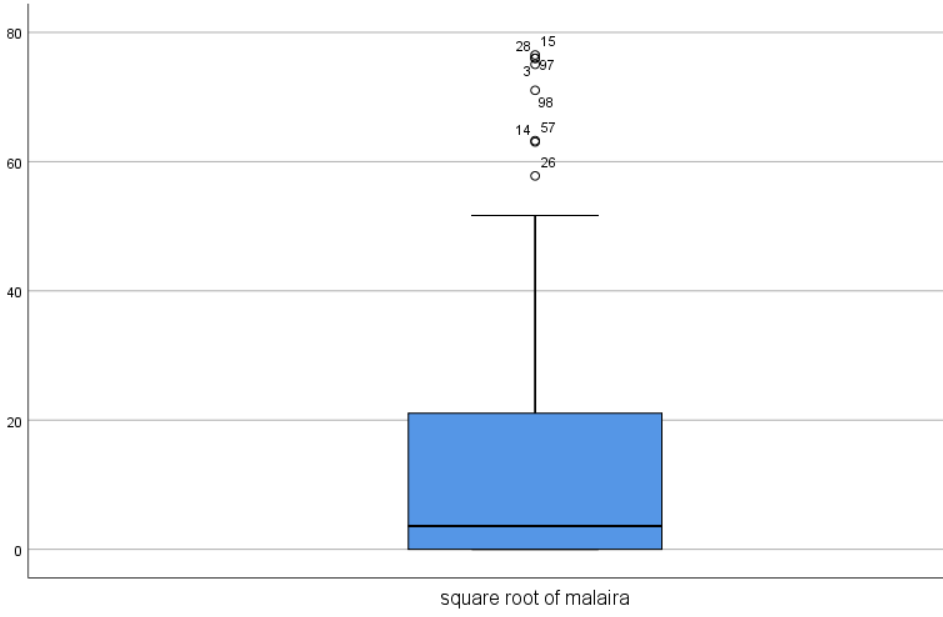
 

(Number of death cases in caesarean ) Log of number of death case in caesarean)

6.3 Number of malaria cases was NOT normal as well . We tried to achieve the normality by taking the LOG10 function , but since most the values were Zero , LOG10 function gave us a lot of NOT DEFINED values . Hence Square Root function was achieved to penalize the higher values and supress the outliers .

(Original Data distribution and outlier check)

(Square root of vlaues of malaria deaths and its outlier check)

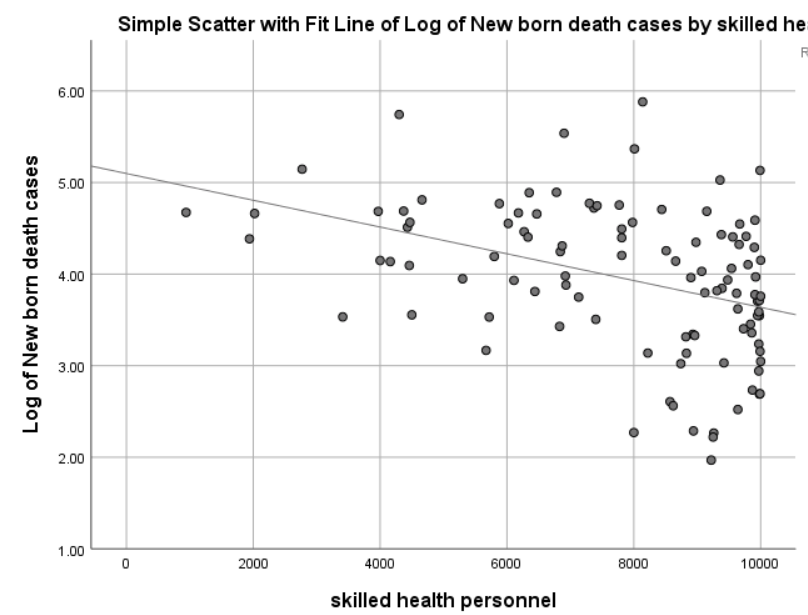
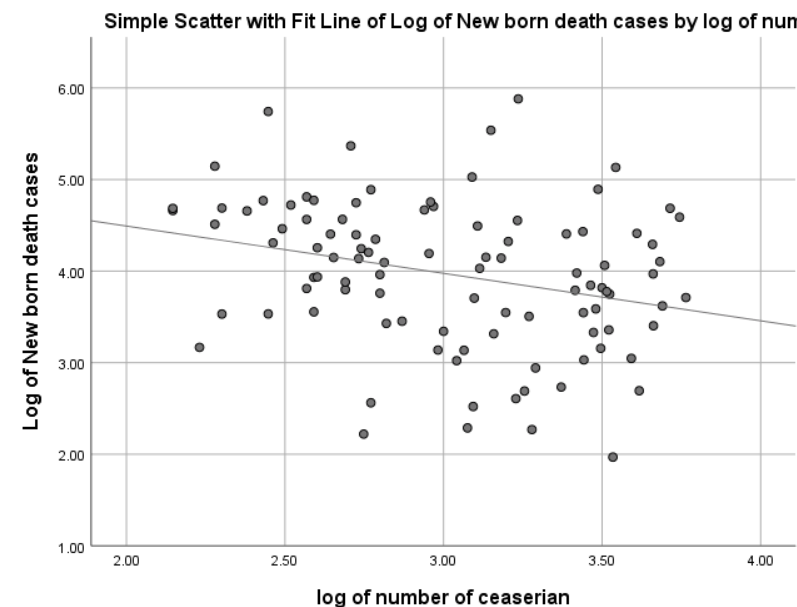
**8.0 Relation of Independent variable with the dependent variables :-**

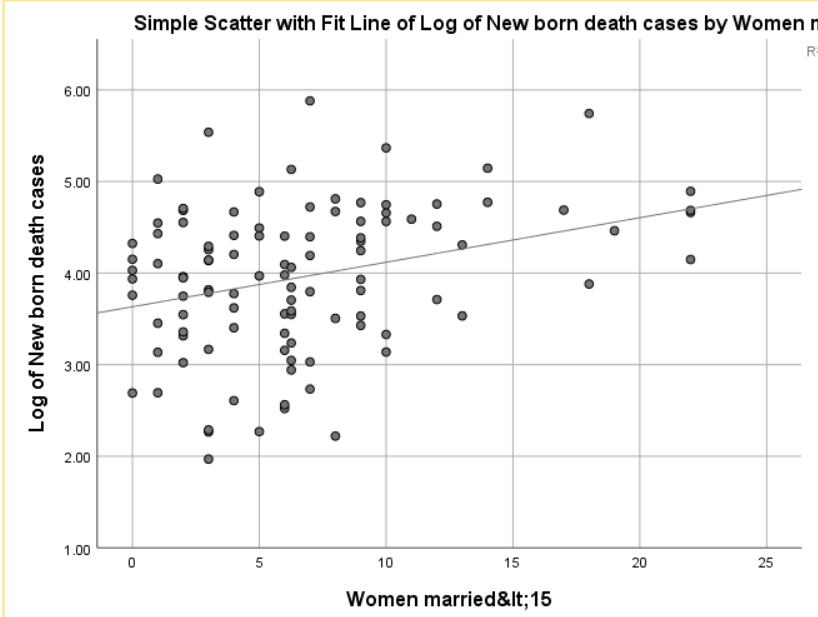
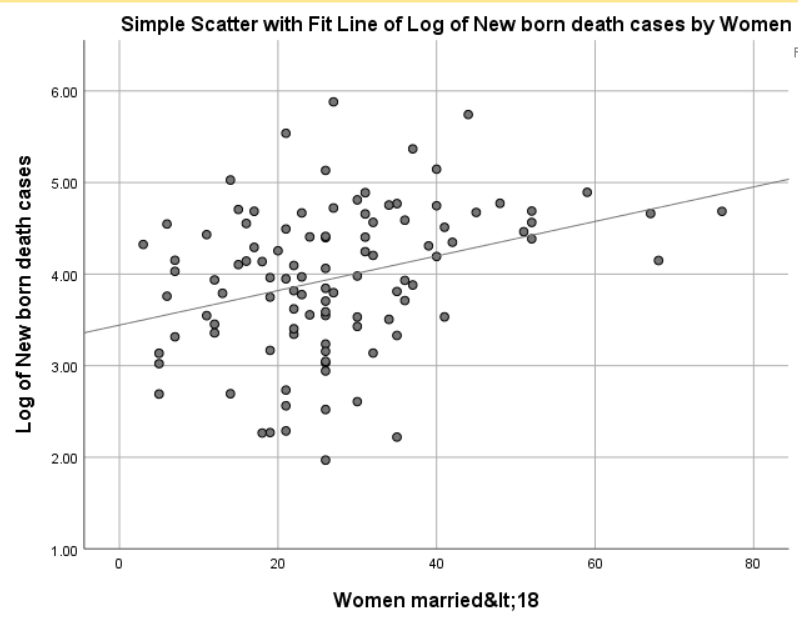
Independent Variable :-

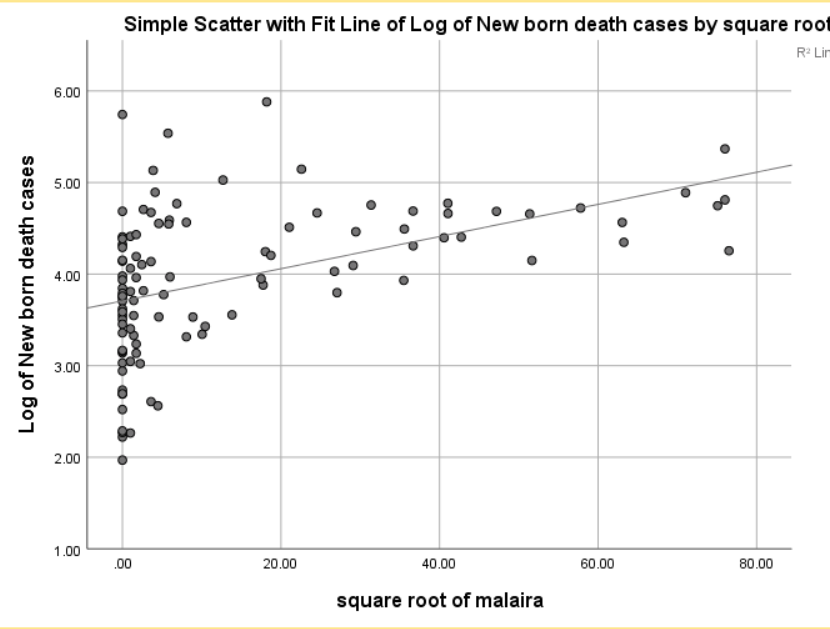
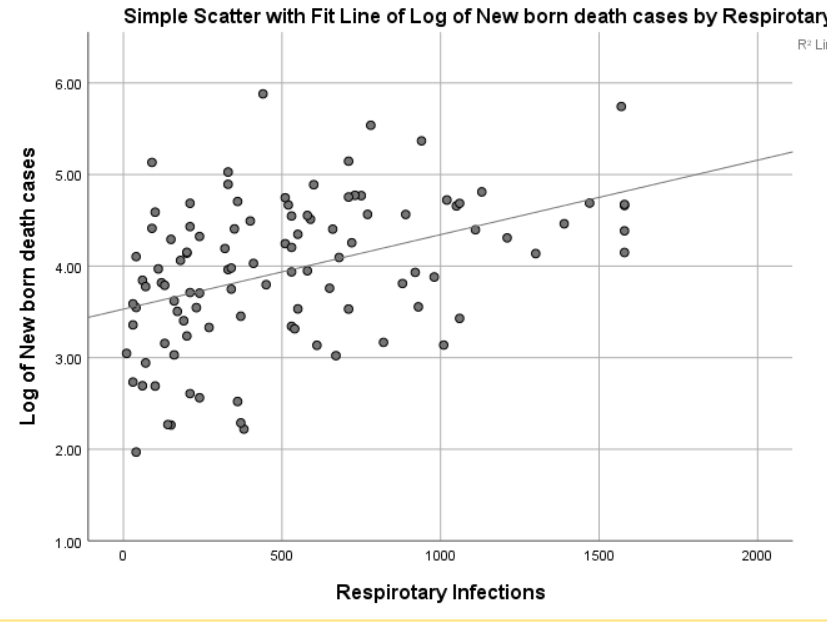
1. Skilled health personal
2. No. of caesarean operated in Log10
3. Women married below 15 years(numbers in 100) .
4. Women married before 18yrs and after 15years(numbers in 100)
5. Number of malaria deaths reported in Square root .
6. Number of deaths due to respiratory infections .

Dependent variable :- Number of infant deaths in Log10 .

The relation between each of the Independent variable and the Dependent variable has been observed on scatter plot individually .

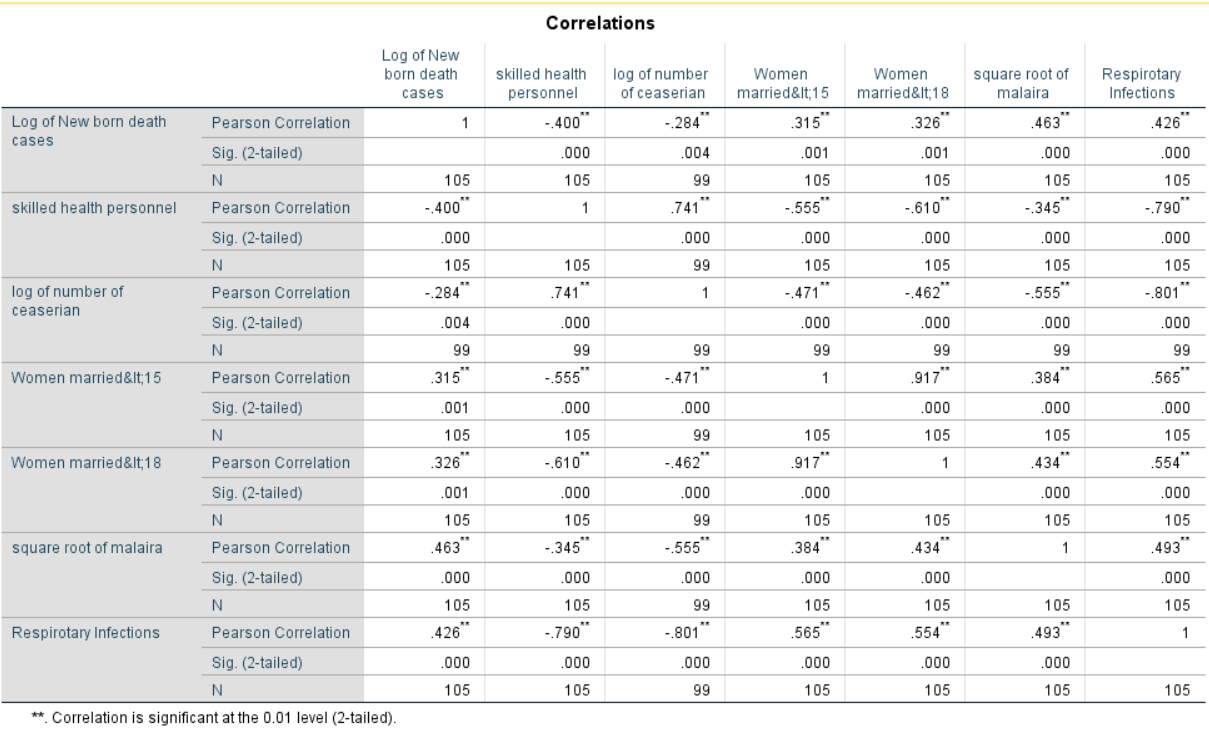
 

9.0 Checking the Collinearity among the independent variable :

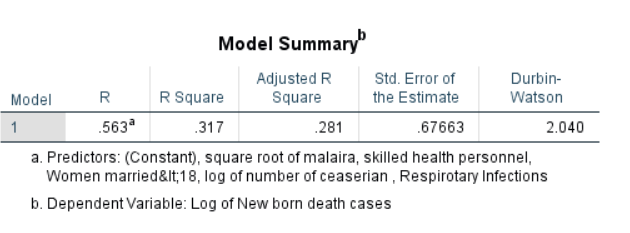
One of the assumptions for linear regression is to remove the collinearity among the variables . Hence Correlation matrix can be useful to check :-

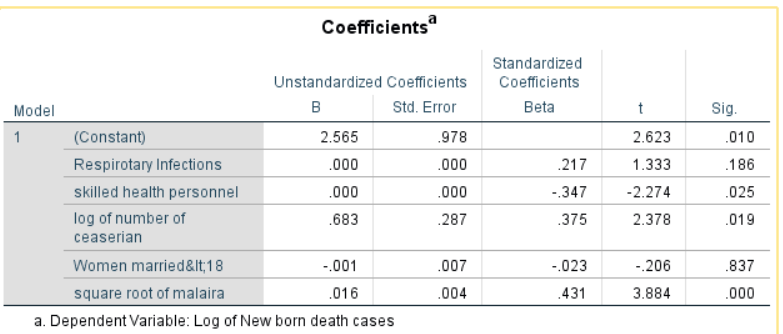


From the above table we can conclude the below information :-

1. There is strong relationship between the number of females married below the age 15 and number of females married below the age of 18(R=.9.17) . Thus, one of these variables can be excluded from the linear regression model .
2. There is also a strong relationship of Skilled health personal and respiratory Infections .

10. Linear regression model :-





Here the unstandardized coefficient B , gives information about an independent variable and the dependent variable . This coefficient is amount of change in the dependent variable , with one unit of change on independent variable , given the other variable remains constant .

Ex :- Log of number of Caesarean’s coefficient is .683 , which implies that for everyone unit in log of number of caesareans , the new born death cases will increase by .683 times .

Similar inference can be drawn for the other variables .

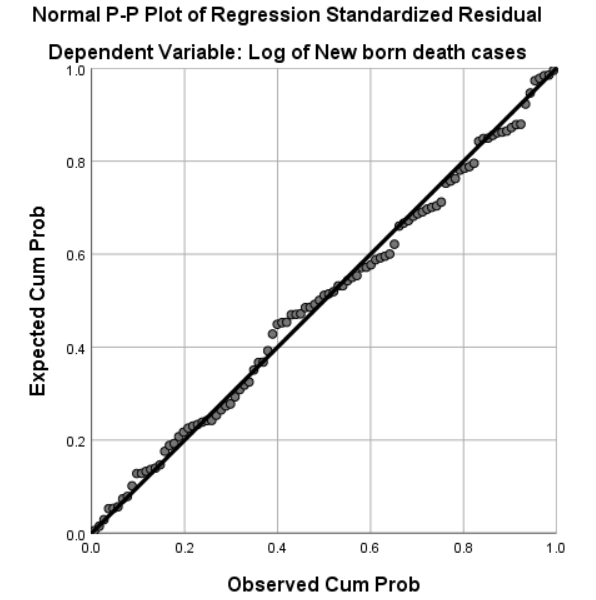
Coefficients for respiratory infections and skilled health personal is negligible , hence both the coefficients need not be added into the equation :-

Y= 2.565 + 0.683(log of number of caesarean) – 0.01(women married&lt18) + 0.16 (square root of malaria )

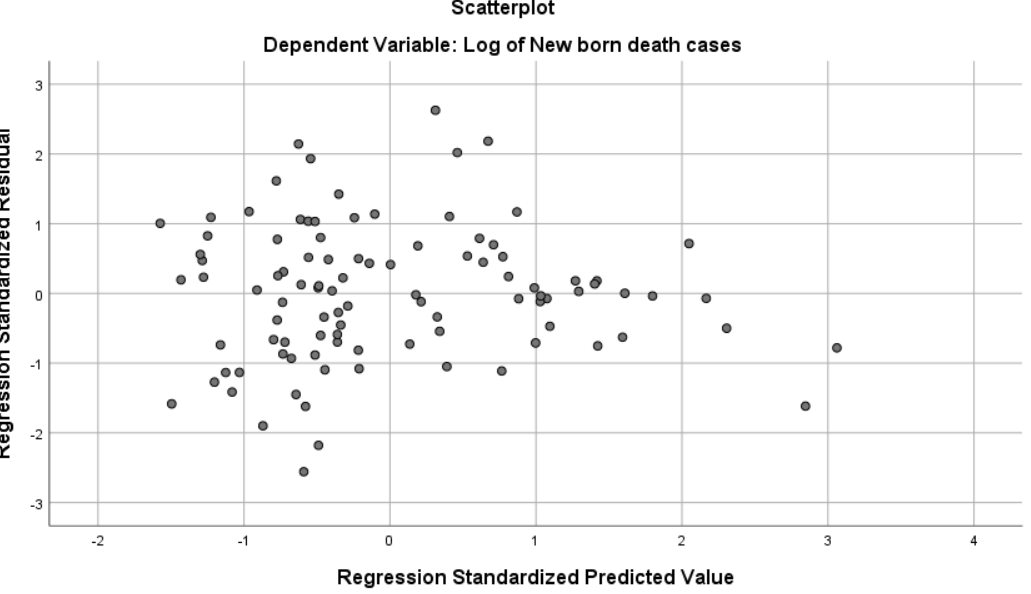
From the given equation , only two variables are statistically significant which contributes mostly towards the model .

Normality check of the residuals :-

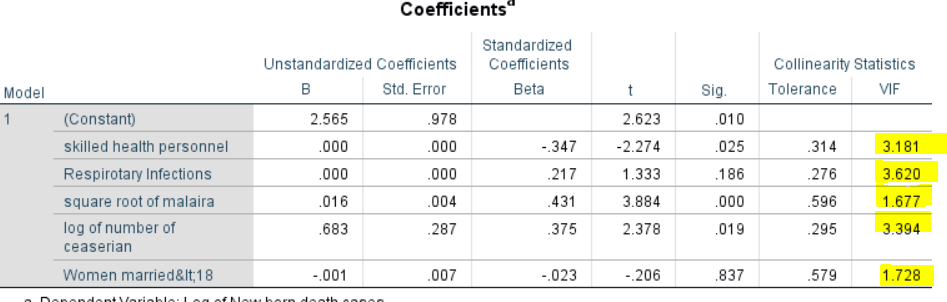
This can be verified in the P-P plot of regression standardized residuals in the below snap :-



Homoscedasticity graph can also viewed to check if the residuals are equally distributed .



Absence of multicollinearity using VIF values :-



All the highlighted value is less than 10 , indicating absence of multicollinearity .

**Bibliography** :-Anon., n.d. *Testing Assumptions of Linear Regression in SPSS.* [Online]   
Available at: https://www.statisticssolutions.com/testing-assumptions-of-linear-regression-in-spss/

Analysis for Binary Logistic regression (CA Part 2)

Amit Sahoo,

Student#18188851

Course :  [Statistics for Data Analytics](https://moodle.ncirl.ie/course/view.php?id=1016)

**Abstract** :-

Local news refers to the coverage of events that happens in a locality . This news might not be of any concern to the members from other locality . It includes subjects like local politics , business , crime , sports etc . This kind of news doesn’t give any sort of information about the whole country or International news . In this globalized era , with rise of computers and mobile , there has been significant decline in the number of people following local news , either in TV , Newspaper , Social media , Radio . The ease of access to any events occurring anywhere in the globe is perhaps the prime factor for this decline .

**Introduction** :-

Capturing location of the user through electronic equipment’s has such as mobile , laptops has increased the local business in the recent years. However, this work intends to find out various factors which determines the popularity of local news in a particular demography . The source of the local news has been kept out of the scope to maintain simplicity and better understanding of the reasons for the decrease in local news popularity .

**Dataset** **description** :-

This dataset has been derived from a survey that was conducted by Pew Research Center's American Trends Panel and a detailed report has been published on <https://www.pewresearch.org/download-datasets/> . In this dataset we need to derive the factors that are statistically significant for the change in the popularity of the local news fan based on of the county . Factors taken into consideration are as follows :-

|  |  |
| --- | --- |
| Attributes | Possible values |
| People Following Local News | 1-Yes, 0-No(Dependent variable) |
| Age | In Numbers |
| Gender | 1-Male,0-Female |
| Highest Education qualification | 1-Graduate,2-In College,3-High School or less |
| Marital status | 0-Living without a partner, 1- Living with Partner |
| Monthly income of the Individual | 1-less than $10,000 , 2- 10,000-20,000 , 3-20,000-40,000, 4-40,000-75,000, 5- greater than 75,000 |
| Internet usage | 0-Doesn’t have internet, 1-have Internet |

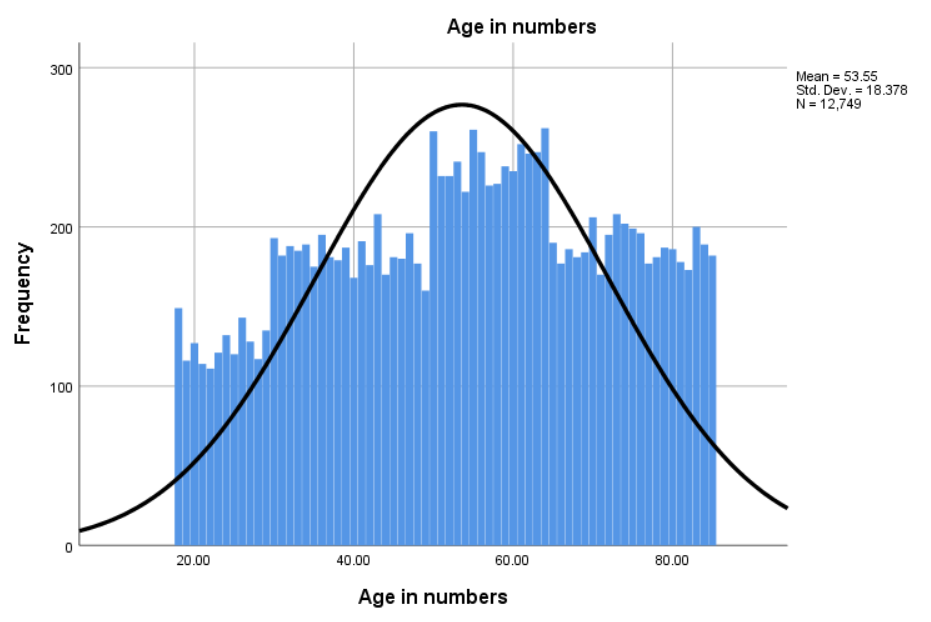
**Assumptions**:-

1.Sample Size :- The sample size(n) of the chosen data set is 12749 . This sample size is big enough for our analysis .

2. Normal Distribution check for continuous variable :-

We have only one continuous variable in this data set – Age .

Age normally distributed , which can also be seen the below graph :-

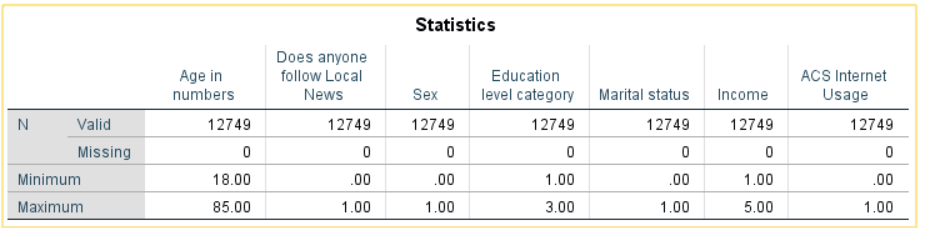


2.Absence of multicollinearity :-

We have only one continuous variable in our dataset . Hence there is no need to check correlation with any other variables as all others are categorical values . Any correlation among the continuous predictors can be checked by Pearson R coefficient .

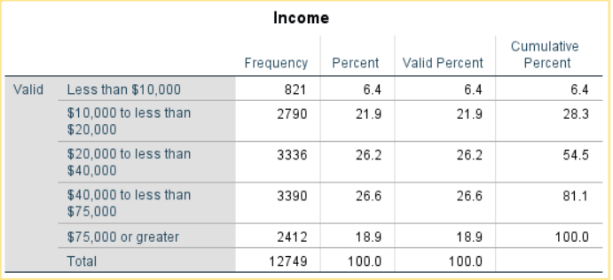
**Analysis** :-

Overall analysis of all the variables used :-



In this dataset 12749 sample has been taken . Fortunately, there was no missing values in the dataset .This can clearly be seen in the above snapshot .

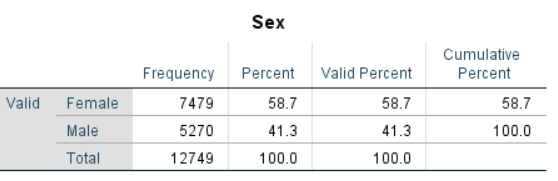
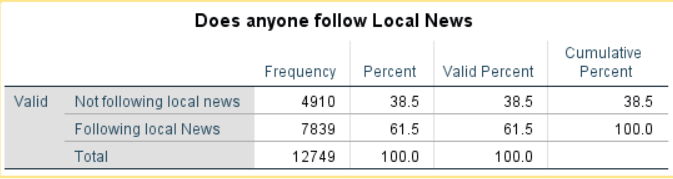
Distribution of each variable over the dataset :-

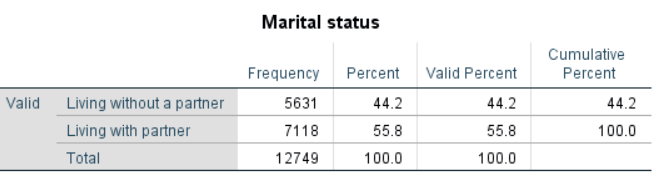
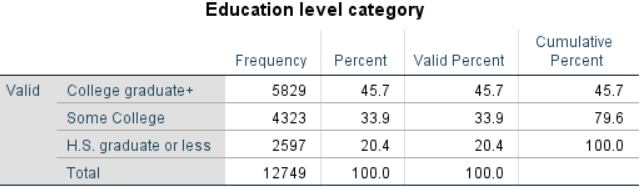


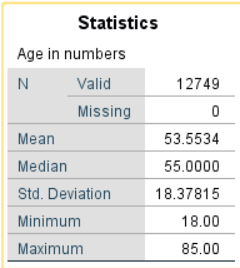
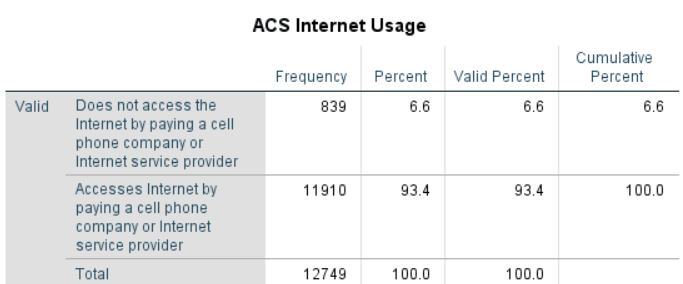
We can see from the adjacent table that people having income less the $10000 constitutes only 6.4percent of the data . People with income between 10,000 and 20,000 constitute 21.9% of data .

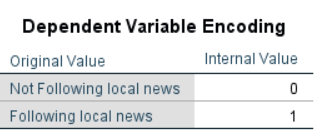
People with income between 20,000 to 40,000 and 40,000 to 75,000 constitute 26.2% and 26.6% respectively . The % of people having income more than 75,000 is only 18.9% .

Similarly, data distribution for other variables can be seen below :-

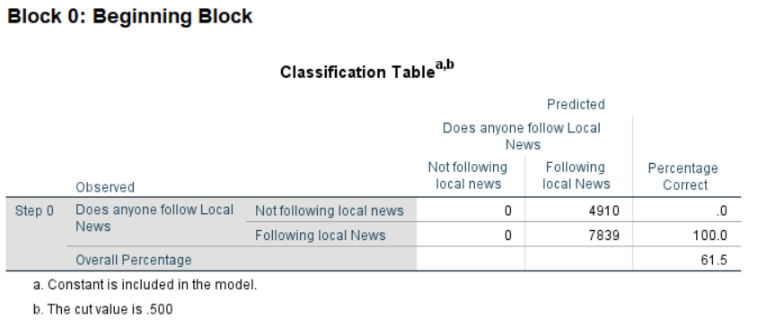






 If the people are following the Local news has been taken as a dependent variable here . Following the local news has been considered as 1 , whereas Not following the local news is considered as 0 . The same is shown in the adjacent photo .

**Block 0 Model analysis:-**



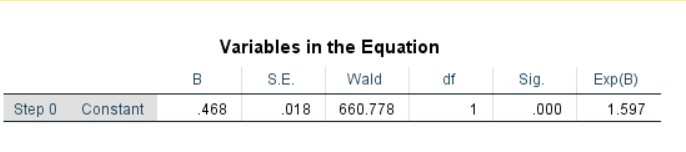
Block 0 Model includes only the intercept(termed as constant in SPSS) with No other predictor attributes in consideration .

This Model has correctly classified 7839 cases which is 61.5% accurate . Probability of Following Local news = 7839\*100/(7839+4910)= 61.5% .

Whereas Not following the local news percentage is 4910\*100/(7839+4910)= 38.5% .

Odds Ratio = Probability of Following Local News/Probability of Not following Local News .

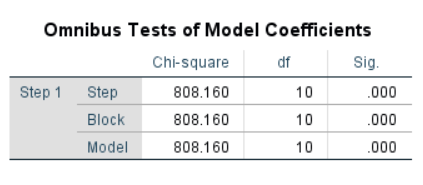
Odds Ratio in this case is 61.5/38.5= 1.59 . The same can be seen in Exp(B) column in the below snap .



In this Model Log(Odds)(this has been explained further below) is .468 (mentioned in B section of the above data) . And the Odds ratio can be calculated from intercept as ‘e’ raised to the power .468 which equals to 1.597(as mentioned in Exp(B)of the above column ) .

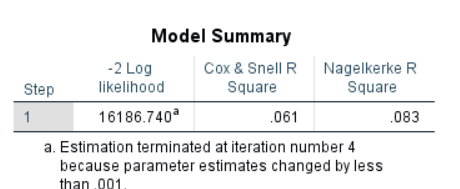
**Block 1 Analysis :-**

In Block 1 , All our predictors are taken into consideration in designing the model . Omnibus Tests of Model Coefficients

shows Chi-square of 808.160 with Degree of Freedom as 10 and highly statistically significant (.000) .

In this case the Null Hypothesis is that predictors adding in Model 1 will fetch the same accuracy as Model Zero mentioned above . Since the Significance level is beyond .001 in this case , we reject the Null the Hypothesis . Therefore, we can Conclude that adding predictors into the model has increased the ability to predict correctly .

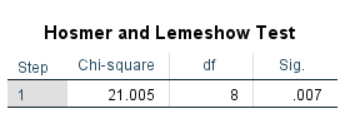
Model Summary :-



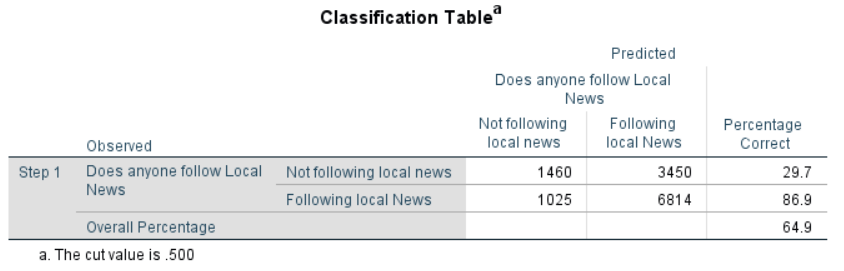
-2 Log Likelihood in Model Summary shows how poor this model can predict the decision . If this value is small , then the model has better ability to predict the decision .

In our case the value is very big(16186.740) , which shows that the model **Doesn’t** have good ability to predict the decision .

Nagelkerke R square which is analogous to R-square in case of multiple regression . The value is .083 in this case . which says that only 8.3% variance in the Dependent variable is explained by the independent variable . Similarly, Cox and Snell R square value is .061 in this case , which suggest that 6.1% of the variance in dependent variable is explained by the Independent variables .This value can never be Zero . Between , Amond Cox & Snell R Square and Nagelkerke Rsquare , the later one is widely used .

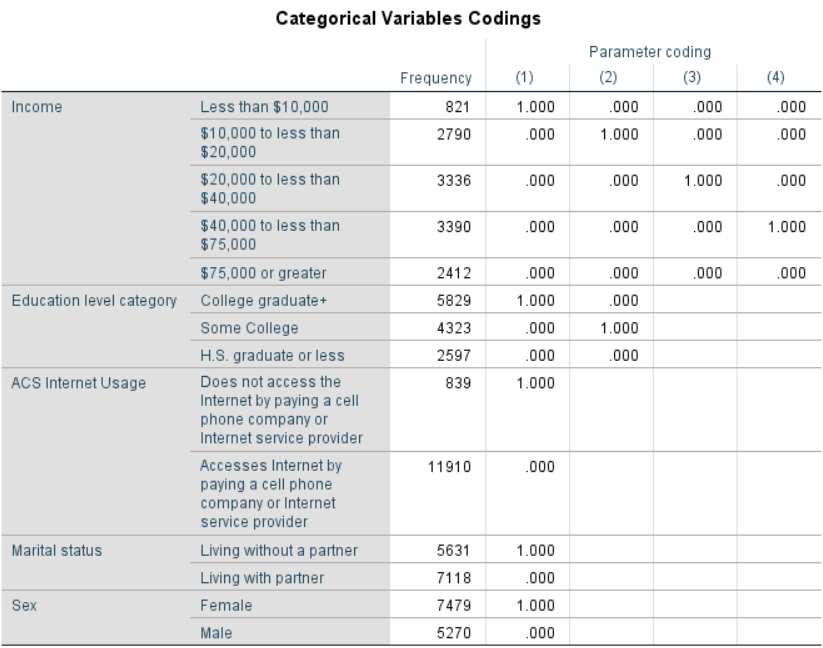


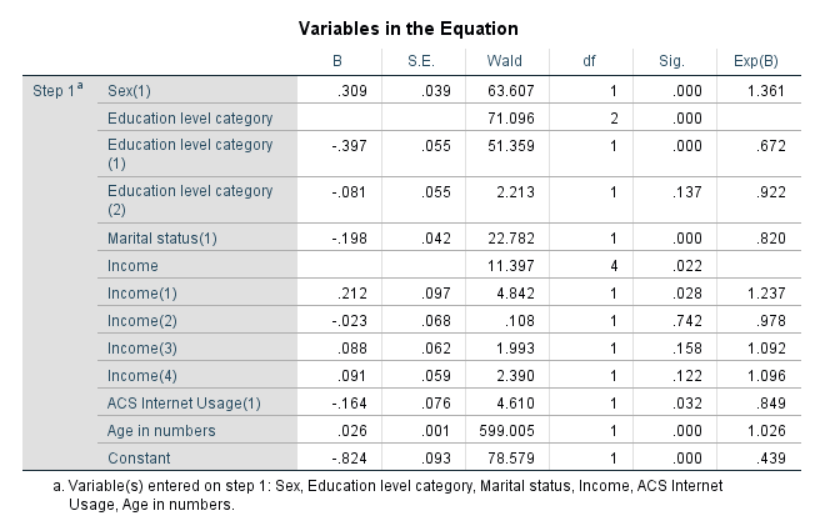
Hosmer and Lemeshow test shows if the model adequately fits the observed data . In this case our null Hypothesis is that the model adequately fits the observed data . Since the Significance is less than .05 , we reject the fact that the model adequately fits the data .

By Looking at the classification table we can observe that the model classifies 64.9% of the data accurately .

Therefore, this model is better that the Model 0 , shown above .

Variable in the equation with categorical coding :-





**Inferences :-**

Age is continuous variable in our case ; the coefficient of Age is .026 signifies that log of odds towards following the local news will increase by .026 units , with each unit increase in the age variable with other parameters as constant .

In this case, the odds ratio (in Exp(B)) for Age in Numbers is 1.026 , this signifies that for each additional unit of increase in age variable , will increase the odds of following the local news by a factor of 1.026 times .

ACS Internet Usage is a categorical value with 0 as No internet usage and 1 as internet usage . In this case the ‘B’ coefficient as -.164 . That means for one-unit change in ACS Internet usage (same as change from ‘NO Internet usage’ to ‘Internet Usage’) , the log of odds of following the local news will decrease by .164 times with other parameters as constant.

The Odds ratio of the ACS internet usage (Exp(B)) is .849 , this shows that changing from ‘No internet usage’ to ‘Internet usage’ , the odds of the event of following the local news will increase by .849 times (as the value is less than 1 , probability decreases) with other parameters as constant.

Categorical variable with multiple categories , such as Education , we have three categories :-

In this case , one-unit change signifies , change from ‘high school graduate or less’ to ‘college graduate+’ or ‘college graduate+’ to ‘some College’ .

The odds ratio for education level category(1) Exp(B) is .672 , which means the odds of an event of following the Local news will increase by .672 times , with the change from ‘high school graduate or less’ to ‘college graduate ’ , given that the other parameters are held constant .

Similarly , the odds ratio for education level category(2) Exp(B) is .922 , which means the odds of an event of following the Local news will increase by .922 times , with the change from ‘college graduate ‘ to ‘Some college’ , given that the other parameters are held constant .

Other inferences can be drawn similarly .

We can write the equation of the model as follows :-

Log(p(x)/1-p(x))= -0.824 + 0.309(gender) -0.397(Education level1) -0.81(Education level2)-0.198(marital status)+0.212(Income1)-0.23(Income2)+0.088(Income3)+0.091(Income4)-0.164(Acs Internet Usage)+0.26(Age)

P(x) is the probability of following the local news .

From the variable equation table above, we can see that Gender , education level(1),marital status(1),Income(1),Acs Internet usage , and Age in numbers are contributing significantly to this model .