**Statistics for Data Analytics– CA2**

**Regression**

1. **Multiple regression:**

**Objective of the study:** To analyze andestimate a multiple regression model using the software SPSS by IBM and compare dependent variable with three independent variables.

**Source of data set:**

The data sets have been taken out of the 3 given websites. The specific URL to each data set has been mentioned below:

1. Gross domestic product (GDP) of all the countries for the year 2015: <http://data.un.org/_Docs/SYB/CSV/SYB61_T13_GDP%20and%20GDP%20Per%20Capita.csv>
2. Percentage of individuals using the internet in all the countries in 2015: <http://data.un.org/_Docs/SYB/CSV/SYB61_T29_Internet%20Usage.csv>
3. Tourist/visitor arrivals (thousands) in all the countries for year 2015: <http://data.un.org/_Docs/SYB/CSV/SYB61_T30_Tourist-Visitors%20Arrival%20and%20Expenditure.csv>
4. Labour force participation and unemployment in the year 2015: http://data.un.org/\_Docs/SYB/CSV/SYB61\_T17\_Labour%20Force%20and%20Unemployment.csv

**Description of the data set:**

The data set contains one dependent variable (variable being predicted) – Gross Domestic Product (GDP) which is dependent over three independent variables (predictor variable) – Internet Usage (Percentage of individuals using the internet), Tourist Arrivals (in thousands) and unemployment rate. The data available in the data sets is huge and for a period of several years. I’ve taken the data only for the year 2015 to make the analysis clear and accurate.

**Description of Analysis:**

The analysis has been done to analyse and visualise the impact of different independent variables on the dependent variable. The GDP column is considered as dependent variable which is assumed to be dependent on the other factors taken into consideration – Internet usage, Tourist Arrivals and Unemployment.

The analysis has been carried out to observe the variable being predicted (dependent variable) over the values of the predictor variable (independent variable). I have used the 95% confidence interval testing observation which may be true or false depending on the values in the datasets stating our procedure to be right or wrong respectively. Following are the tests conducted in SPSS:

1. **b-value:**

The b-value tells us to what degree each predictor (independent variables) affects the outcome if the effects of all other predictors are held constant.

1. **Durbin-Watson Statistic:**

While doing any analysis, auto-correlation can be a significant issue for analysing the data. The Durbin-Watson Statistic is a test for auto-correlation in a dataset. The value of the Durbin-Statistic test will lie between 0 to 4. If the test value is around 2, that means there is no auto-correlation detected in the samples.

R – It is the correlation that is obtained by the relationship between obtained values and the values proposed by the dataset.

R2 – It is the proportion of variation in the dependent variable that is explained by the independent variables. It ranges between 0 to 1 and cannot have negative values.

R2 Adj. - The adjusted R2 gives us some idea of how well our model generalizes. Its value has to be the same as, or very close to, the value of R2.

1. **The F-Test:**

It looks at whether the variance explained by the model is significantly greater than the error within the model. It tells us whether using the regression model is significantly better at predicting values of the outcome than using the mean.

1. **Collinearity Test:**

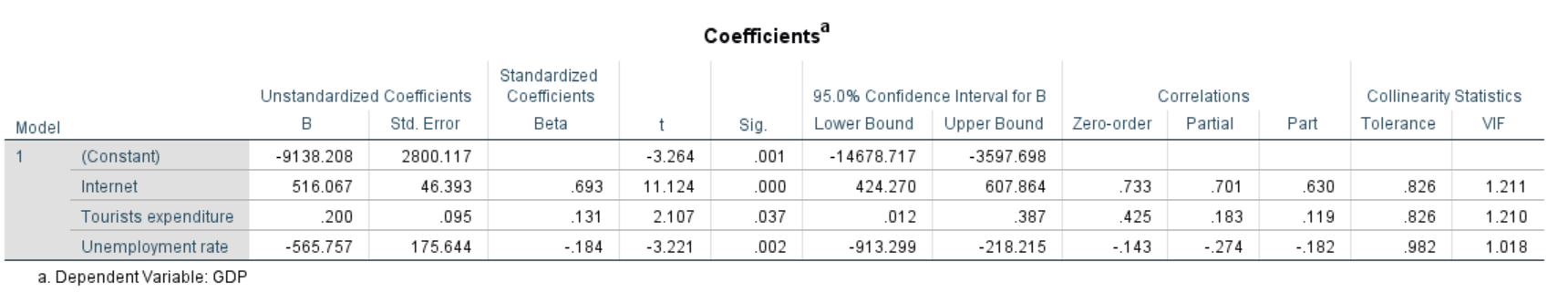
Collinearity means 2 variables are nearly perfect linear combinations of one another. Multicollinearity comes into picture with more than two variables. If there is multicollinearity in the sample, regression estimates are likely to be unstable and will have high standard errors.

**Assumptions:**

1. The first assumption is that the prediction variable is a continuous variable.
2. The second assumption is that the predictor variables are in categorized form and continuous.
3. The third assumption is that there is no auto-correlation in the samples.
4. The fourth assumption is that the variance around the regression line is the same for all values of the predictor variable.

**IBM SPSS Analysis Interpretations**

1. B-value:

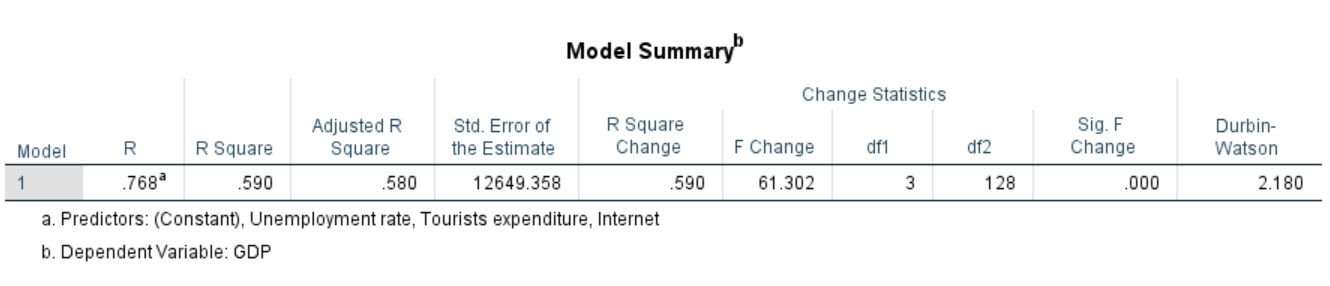


Observations:

To find out which beta value is contributing the most out of each independent variable, we take the highest value in the Standardized beta column. So, Independent variable – Internet makes the strongest contribution to explaining the dependent variable. Hence, to conclude, if there is a one unit change in internet variable, the GDP will be increased by 516.067 USD. If the tourist expenditure is increased by one unit, the GDP will get increased by 0.200 USD. For one unit change in Unemployment rate, the GDP will be decreased by 565.757 USD.

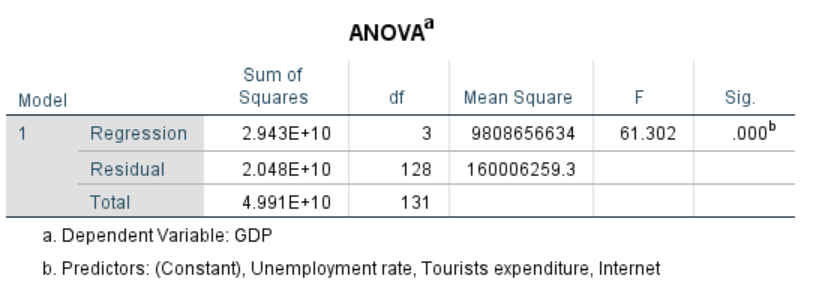
The significance values should be well under 0.05, hence we can contribute the variable is making significant unique contribution to the prediction of the dependent variable.

1. Durbin-Watson Statistic:



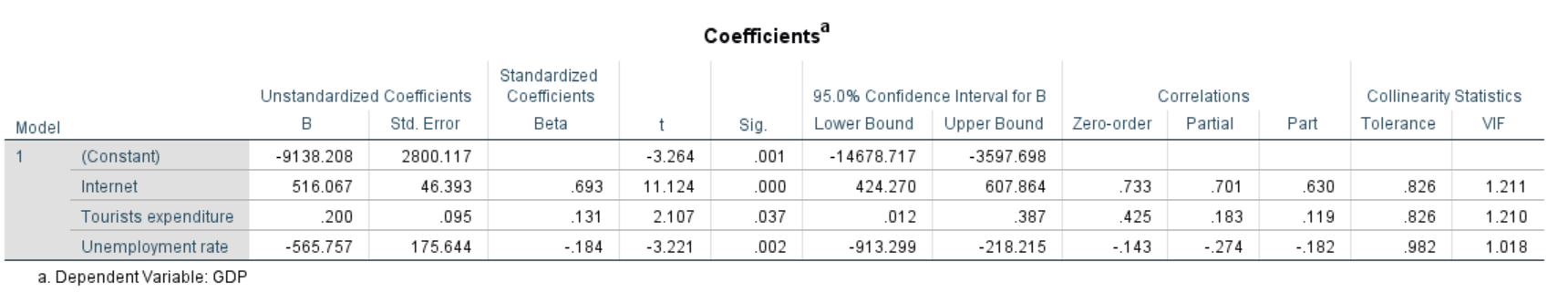
The Durbin-Watson statistic value obtained of the sample at 95% CI level is 2.180 which clearly depicts our observation fits in to data model. Based on this, we can conclude that the predictor and prediction values are continuous in the sample taken. Hence, we can conclude that the GDP of the countries in the sample taken is influenced by the number of internet users, total number of tourists arriving, and the unemployment rates in the respective countries.

1. ANOVA:



The significance value is 0.00 which is lesser than 0.05. Hence, we can conclude that the observations taken are fitting accurately into the model.

1. Collinearity:



To check the collinearity, the Tolerance and VIF values are accountable. If tolerance value is very small (less than 0.10), it indicates that the multiple correlation with other variables is high, suggesting the possibility of multi-collinearity. As per the model that I have used, the tolerance for all the independent variables is above 0.10 (0.826, 0.826 and 0.982), hence, we can conclude that the independent variables are not multi-collinear.

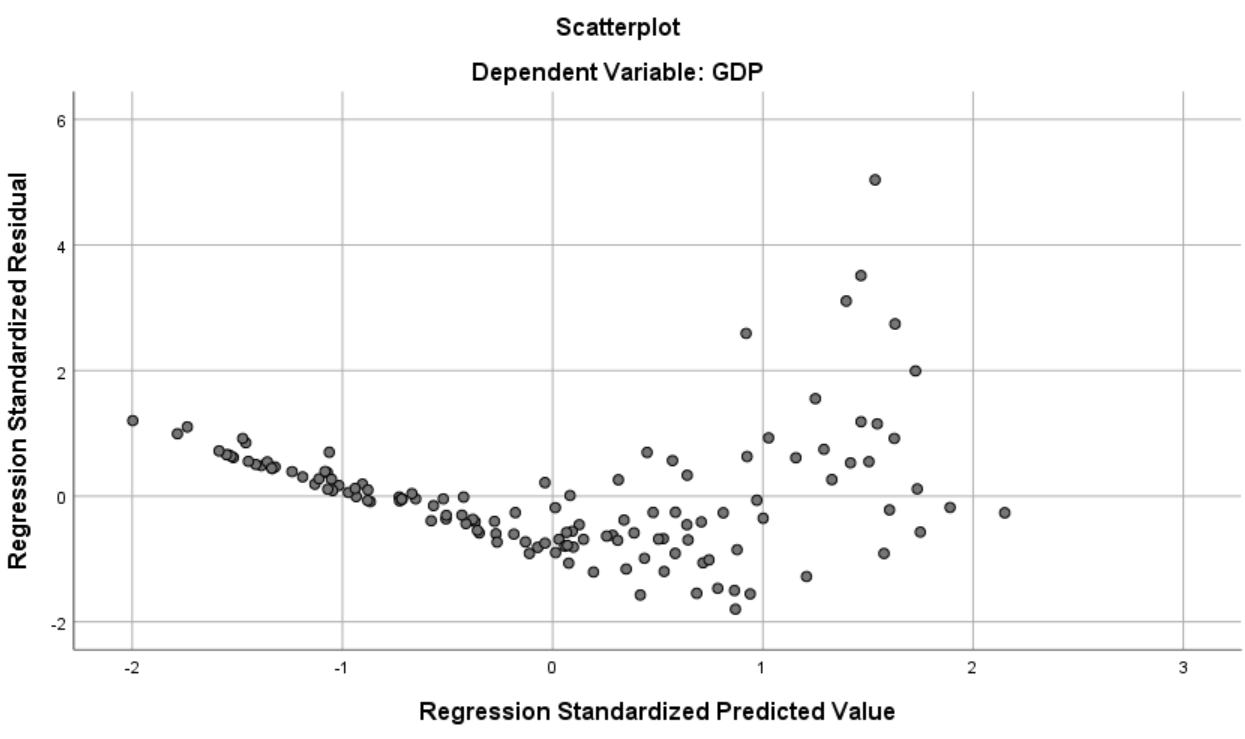
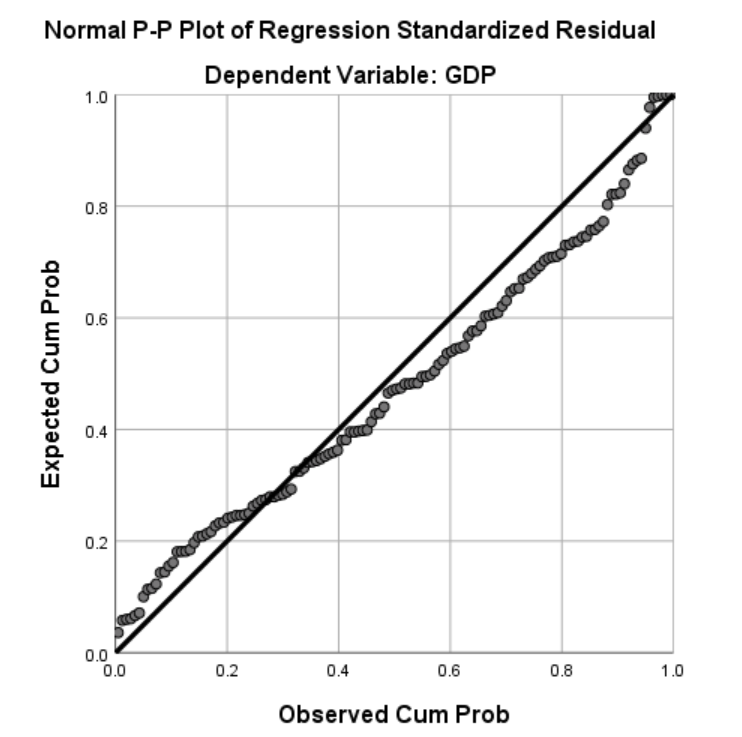
The VIF (Variance Inflation Factor) value should be between 1-5 and should not exceed 10. Hence, we can conclude that all the independent variables are not multi-collinear as their values are within the range.

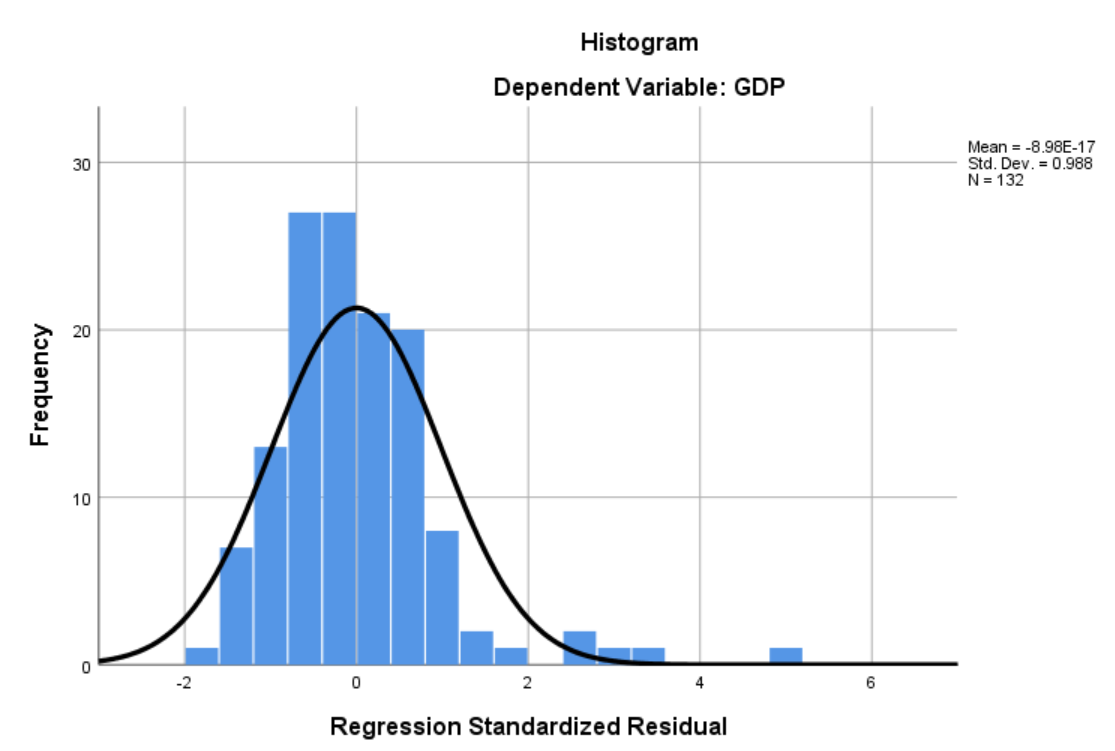
**Conclusion**

The conclusion that can be drawn after doing the above tests are as follows:

* The dependent and the independent variables are closely related to each other.
* The independent variables result in a significant amount of change in the dependent variable.

The charts below justify and prove the results provided above:





1. **Logistic Regression:**

**Objective of the study:** To analyze and estimate a binary logistic regression model using the software SPSS by IBM and compare discrete or dichotomous dependent variable with two or more independent variables.

**Source of data set:**

The data sets have been taken out of the 3 given websites. The specific URL to each data set has been mentioned below:

1. Cancer, age-standardized death rates by country - <http://apps.who.int/gho/data/node.main.A1110?lang=en>
2. YLL score by country - <http://apps.who.int/gho/data/node.main.A1106?lang=en>
3. Alcohol-attributable fractions, all-cause deaths by country - <http://apps.who.int/gho/data/node.main.A1091?lang=en>

**Description of the data set:**

The dataset contains three variables Alcohol – Years of life lost, Male Deaths by Cancer and Female Deaths by Cancer which are recognized as predictor variables by the software also known as independent variables. The logistic regression (binary) is applied only when the dependent variable is dichotomous, i.e., it has only two possible values like true or false, 1 or 0, yes or no etc. Hence, the dependent variable taken in the data set is in binary form, i.e. 0 or 1 for the Male or Female deaths by Alcohol. So, there is definitely a relationship between the dependent and independent variables. We are going to analyze it by putting the sample through different tests.

**Description of Analysis:**

The analysis has been done to analyse and visualise the impact of different independent variables on the dependent variable which is dichotomous in nature. The column – ‘MorFDeaths’ is considered as dichotomous dependent variable which is assumed to be dependent on the other factor taken into consideration – Years of life lost, Male Deaths by Cancer and Female Deaths by Cancer.

I have used the 95% confidence interval testing observation which may be true or false depending on the values in the datasets stating our procedure to be right or wrong respectively. Following are the tests conducted in SPSS:

1. **Cox & Snell and Nagelkerke R2:**

The Cox & Snell R2 and Nagelkerke R2 are analogous to the R2 measure in multiple regression. They are also known as pseudo R2 statistics. They refer to the amount of variation in the dependent variable which is predicted by the predictor variables collectively. The maximum value of this, in theory, is 1.00 if the relationship is perfect; it will be 0.00 if there is no relationship. The Cox & Snell R2 and Nagelkerke R2 are called pseudo-statistics as they appear like R2 but they are actually analogous to it.

1. **Hosmer & Lemshow Test:**

This is a test of model fit. If the sigma value is less than 0.05, it means that the sample does not fit the model. If the value exceeds 0.05, it signifies the sample fits the model perfectly.

1. **Variables in equation table:**

The key values are found in the Variables equation table. The p values (sig) for all the variables indicate how much they contribute significantly to the predictive ability of the model. The B values are the equivalent to β values in multiple regression – sign tells the direction of influence for each variable.

1. **Percentage accuracy in classification:**

In this table, the overall percentage that the variable has correctly achieved the accuracy towards overall factors.

1. **The Omnibus Test:**

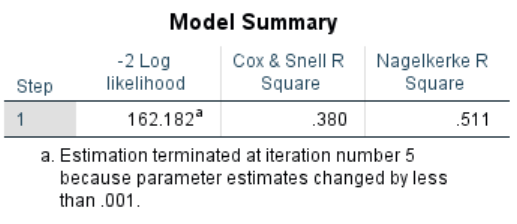
The omnibus test gives us an overall indication of how well the model has performed, over and above Block 0.

**Assumptions:**

1. The prediction variable and the predictor variable should show a relationship between each other and should show a better model after re-estimation.
2. If the model doesn’t show any improvement after re-estimation, a new model should be implemented.
3. There should be a linear relationship between the prediction and predictor variables.

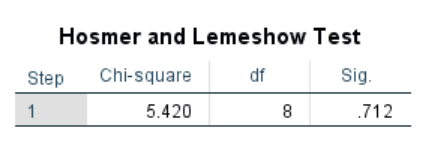
**IBM SPSS Analysis Interpretations**

1. Cox & Snell and Nagelkerke R2:



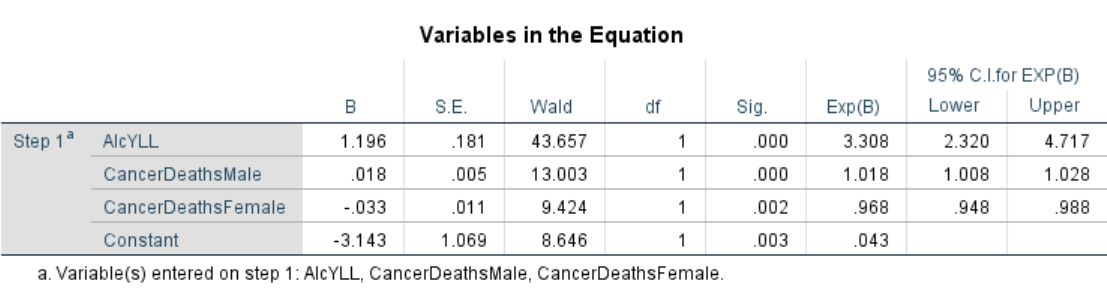
The model summary gives us information about the usefulness of the model. The values Cox & Snell and Nagelkerke R2 values provides us an estimation of the amount of variation in the dependent variable which can be between 0 and 1. The values 0.380 and 0.511 suggest that between 38.0 % and 51.1% of the variability is explained by this set of variables.

1. Hosmer & Lemshow Test:



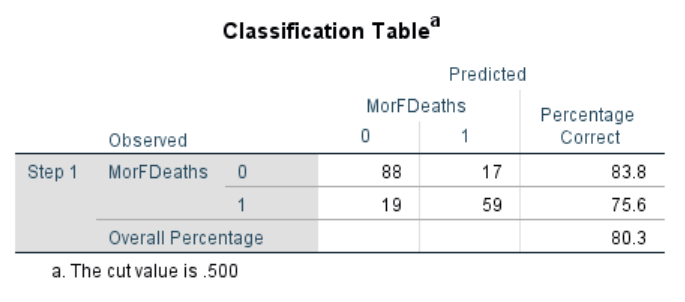
The Chi-square value should be greater than 0.05. For the sample we have taken, the Chi-square is 5.420, hence the data supports the model.

1. Variable in equation table:



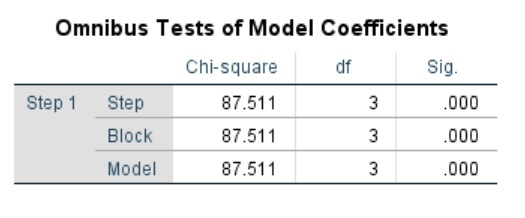
The sigma values of each of the predictor variables is less than 0.05, which clearly signifies that the predictor variables contribute significantly to the predictive ability of the model. In this model, AlcYLL, CancerDeathMale, CancerDeathFemale with the sigma values 0.000, 0.000, and 0.002 respectively contribute significantly to the model.

1. Percentage accuracy in classification:



The above results are without any of the independent variables used in the model. The overall percentage of correctly classified cases is 80.3 %.

1. The Omnibus test:



This test is for model fitness. This test compares the results obtained over and above block 0. This test clearly shows that the model has performed better at block 1 for a degree of freedom = 3 and Chi-square value = 87.511. Also, the significance level is below 0.05, which shows that values are highly significant.

**Conclusion:**

Logistic regression has been performed to analyse the impact of independent (predictor) variables on the dependent (prediction) variable. The model contained three independent variables - Alcohol Years of life lost, Male Deaths by Cancer and Female Deaths by Cancer.

The full model having all the predictors came out to be significant as all the sigma values were less than 0.05, which implies that the model is able to distinguish between male and female deaths by Alcohol.

The variables in equation table tell us that all the independent variables are significantly contributing to the model.

