

# Report on Evaluating Multiple Regression and Logistic Regression Model

## Statistics for Data Analytics

Nandita Sharma

### Description of Data:

To perform multiple regression and logistic regression models, we have extracted data from the “World health organization”. Data is about the “Mortality rate” which states that the probability of dying between the age group 15 and 60 years per thousand population in the year **2013** and what are factors are responsible in different countries. The following are the independent variable:

1. Death due to total NCD(Noncommunication disease) in thousand is taken from “<http://apps.who.int/gho/data/node.main.A860?lang=en>”.
2. Predictor- a percentage of total the population using sanitation services data by country is taken from “<http://apps.who.int/gho/data/node.main.WSHSANITATION?lang=en>”.
3. Another predictor of Consumption of alcohol per capita income in liters have been extracted from “<http://apps.who.int/gho/data/node.main.A1039?lang=en>” which tells amount of alcohol is consumed in various country and its effect on mortality rate.
4. Fourth variable expectancy of life at the age of 60 and above is considered from Health life expectancy (HALE) which state the contribution of this variable to mortality rate.

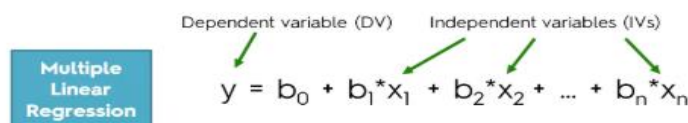
\*Dependent variable or target is “Mortality rate (Dying between 15&60 year)”

### Multiple Regression

#### 1. OBJECTIVE:

The purpose of this dataset is to find how well the parameters like total NCD, Alcohol consumption, percent of sanitation, life expectancy at age 60 help us to predict the Mortality rate by building a multiple regression model. With this, we will able to choose best predictor.

Final main is to come up with the optimum predictors and construct an equation that will forecast the probability of (dying between 15 and 60 year/1000 population).



The diagram shows the equation  $y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$ . A green arrow points from the text "Dependent variable (DV)" to the variable  $y$ . Another green arrow points from the text "Independent variables (IVs)" to the terms  $b_1x_1 + b_2x_2 + \dots + b_nx_n$ . To the left of the equation is a blue box with the text "Multiple Linear Regression".

$Y$  = Predicted variable

$b_0, b_1, b_2, b_n$  = estimated coefficient

$X_1, X_2, X_n$  = predictors

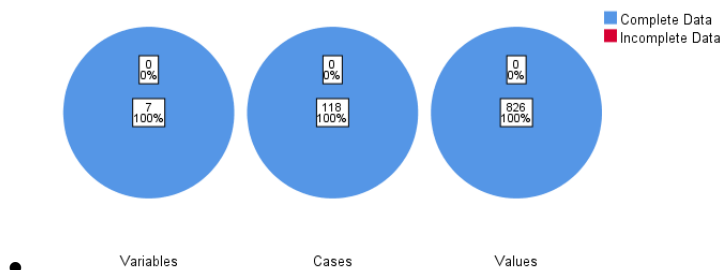
## 2. SAMPLE DATA:

	Country	Year	MortalityrateD yingbetween1 5amp60year	Alcoholrecorded percapita	Lifeexpectancya tage60	TotalNCDDeaths	Populationusingat leastbasicsanitati on services
1	Afghanistan	2013	235	.03	16.2	109.5	38
2	Albania	2013	100	5.06	20.4	24.3	97
3	Algeria	2013	100	.54	21.6	144.0	87
4	Angola	2013	249	8.02	17.1	61.9	45
5	Antigua and Barbuda	2013	127	8.50	19.8	.5	88
6	Argentina	2013	116	8.20	21.4	254.5	94
7	Armenia	2013	121	3.72	19.4	26.3	92
8	Australia	2013	60	9.87	25.5	142.7	100
9	Austria	2013	67	11.60	24.0	74.4	100
10	Azerbaijan	2013	124	2.04	18.6	54.1	89
11	Bahamas	2013	166	8.55	22.4	1.8	94
12	Bahrain	2013	63	2.12	21.2	2.3	100
13	Bangladesh	2013	138	.01	19.2	572.6	43
14	Barbados	2013	101	8.88	19.6	2.7	95
15	Belarus	2013	183	12.37	18.8	107.5	97
16	Belgium	2013	77	10.33	23.7	94.9	99
17	Belize	2013	184	6.17	16.9	1.4	87
18	Benin	2013	250	1.45	17.1	35.1	15
19	Bhutan	2013	219	.19	20.2	3.2	65
20	Bolivia (Plurinational St...	2013	192	3.73	21.2	45.9	53
21	Bosnia and Herzegovina	2013	97	4.43	20.2	36.3	95

## PRE-PROCESSING OF DATA:

- First, integrated different variable country wise and then cleaned not required data from dataset.
- Then, checked for missing value for both independent and dependent variable and observed there is no missing value is present in overall summary ,

Overall Summary of Missing Values



Above graph show there is no missing value is present in dataset, Hence we can consider it.

### 3. TYPES OF VARIABLE:

VARIABLE	TYPE	SCALE
<u>Mortality rate</u> (Dying between 15&60 year)	Dependent	Continuous
Alcohol, recorded per capita (15+)consumption( in L of pure alcohol)	Independent	Continuous
Life expectancy at age 60(year)	Independent	Continuous
Total NCD Deaths (in thousands)	Independent	Continuous
Population using at least basic sanitation services (%)	Independent	Continuous

### 4. ASSUMPTIONS:

Below are following assumption must be taken care to perform multiple regression model and one must check not to violate this assumption:-

- Multicollinearity:** In multicollinearity, must check the correlation between your independent and dependent variable. Preferably, the correlation should be above **0.3**. Also, check the relationship between two independent variables. It should not be above **0.7**. If Pearson correlation  $r$  is greater than 0.7, in such cases avoid one similar parameter.

Correlations						
		log_Mortality_rate	Alcohol, recorded per capita	Life expectancy at age 60	Total NCD Deaths	Population using at least basic sanitation services (%)
Pearson Correlation	log_Mortality_rate	1.000	-.315	-.852	-.075	-.763
	Alcohol, recorded per capita	-.315	1.000	.482	.041	.425
	Life expectancy at age 60	-.852	.482	1.000	.031	.716
	Total NCD Deaths	-.075	.041	.031	1.000	.021
	Population using at least basic sanitation services (%)	-.763	.425	.716	.021	1.000
Sig. (1-tailed)	log_Mortality_rate	.	.000	.000	.158	.000
	Alcohol, recorded per capita	.000	.	.000	.293	.000
	Life expectancy at age 60	.000	.000	.	.340	.000
	Total NCD Deaths	.158	.293	.340	.	.391
	Population using at least basic sanitation services (%)	.000	.000	.000	.391	.
N	log_Mortality_rate	181	181	181	181	181
	Alcohol, recorded per capita	181	181	181	181	181
	Life expectancy at age 60	181	181	181	181	181
	Total NCD Deaths	181	181	181	181	181
	Population using at least basic sanitation services (%)	181	181	181	181	181

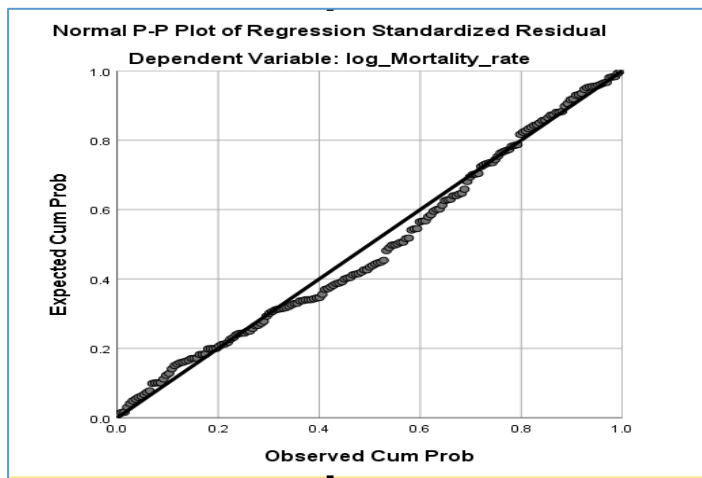
- Alcohol recorded , Life expectancy at 60, population using at least basic sanitation service have a significant effect on Log Mortality rate(dependent variable) as they have strong correlation with each other. (above 0.3)
- Total NCD deaths have **-.075** and **p=0.158** correlation with the target variable which is extremely low means it does not have any significant relation with log Mortality rate. Hence we can remove this independent variable.

- No independent variables are correlated with each other, so we can consider all of them for our model.

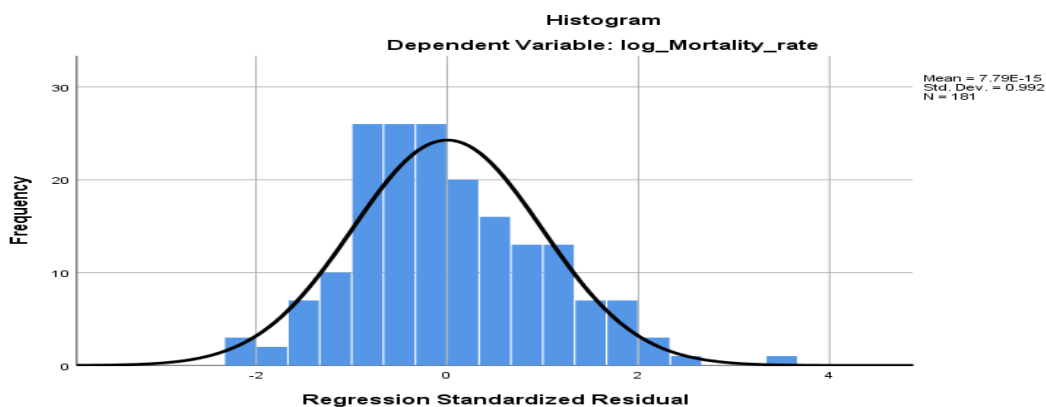
**1.1 VIF:** The second test is Tolerance and VIF (variance inflation factor) test: This will give a clear idea about common variance that exist between two predictors. VIF should be **less than 10**. If VIF value is near to 1, it means there is no chance of multicollinearity. And Tolerance should **not be less than 0.10 or (above 0.10)**.

All three variables have tolerance and VIF value greater than 0.10 and less than 10 respectively. Hence, it not violating multicollinearity assumption. And we can remove NCD parameter because it doesn't have any relation with target variable.

## 2. Normality:



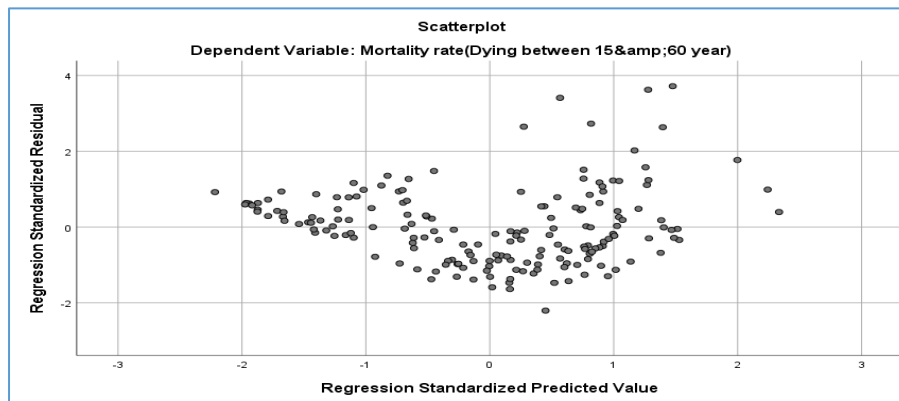
- It is normally distributed and point on the graph is also falling along the straight line making 45 degree line.



- Both Histogram and normality curve are normally distributed. Hence we can conclude that it is not violating our assumption.

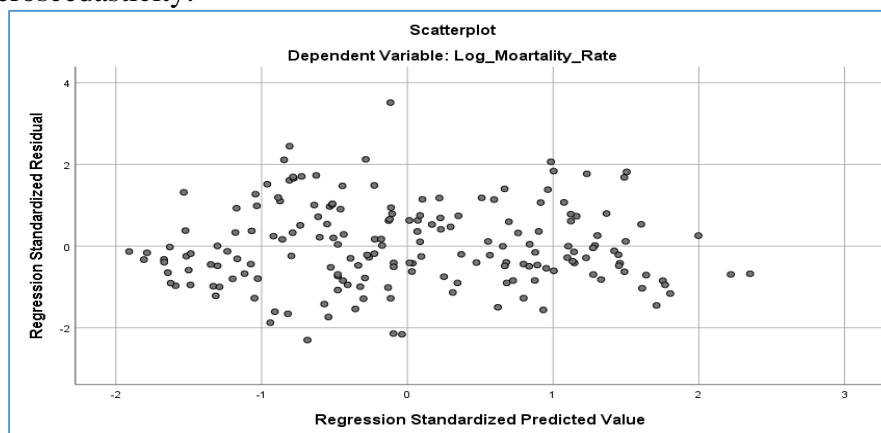
### 3. Check Heteroscedasticity:

**Before:** Multiple regression model show standard errors is biased toward one side or in another word generated an error which does not have constant variance. Hence, it is violating our assumption.



**After:** In order to remove Heteroscedasticity, we used **log of dependent variable** ie. Target variable.

Applied log to Mortality rate who are dying between 15&60 year. We are able to successfully remove Heteroscedasticity.



### 4. NO influential data point:

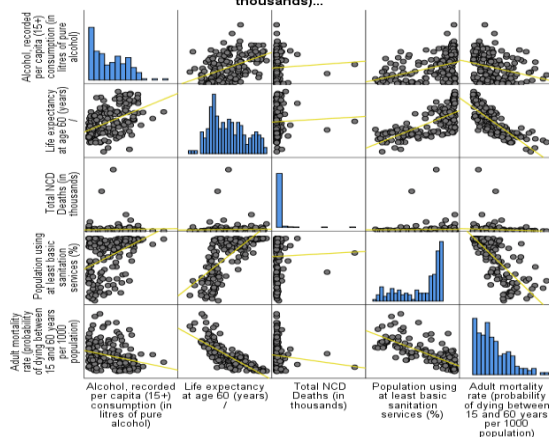
Residuals Statistics <sup>a</sup>					
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	4.0593	6.1449	4.9939	.48953	181
Std. Predicted Value	-1.909	2.351	.000	1.000	181
Standard Error of Predicted Value	.020	.068	.037	.008	181
Adjusted Predicted Value	4.0606	6.1516	4.9942	.48951	181
Residual	-.58064	.88589	.00000	.25016	181
Std. Residual	-2.302	3.512	.000	.992	181
Stud. Residual	-2.324	3.560	-.001	1.002	181
Deleted Residual	-.59210	.91066	-.00028	.25531	181
Stud. Deleted Residual	-2.354	3.685	.001	1.008	181
Mahal. Distance	.106	12.177	2.983	1.728	181
Cook's Distance	.000	.089	.005	.009	181
Centered Leverage Value	.001	.068	.017	.010	181

a. Dependent Variable: log\_Adultmortality

**Cook's Distance:** It is not violating our assumption as observed cook's distance lies under 1. And it also mean it does not have either outliers or have high leverage on our regression estimation.

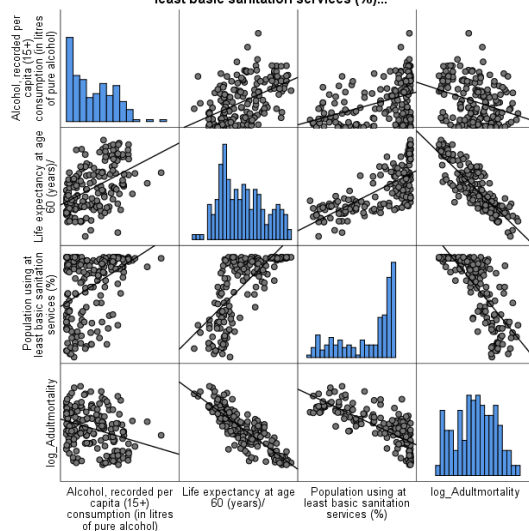
## 5. Check for Linearity :

Scatterplot Matrix Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol),Life expectancy at age 60 (years)/, Total NCD Deaths (in thousands)...



After applying log to target variable and removing total NCD variable we have observed good correlation between independent and dependent variable.

Scatterplot Matrix Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol),Life expectancy at age 60 (years)/, Population using at least basic sanitation services (%)...



## 5. INTERPRETING THE OUTPUT OF MULTIPLE REGRESSION ANALYSIS:

Below are the observed output of table from multiple regression which determine how well the regression model fits the actual data, assuming that no assumption got violated.

### 1. Model\_Summary:

Model Summary <sup>d</sup>										
Mode	R	Adjusted R	Std. Error of	Change Statistics				Sig. F	Durbin-	
1	R	Square	the	R Square	F	df1	df2	Change	Watson	
			Estimate	Change	Change					
1	.852 <sup>a</sup>	.725	.28888	.725	472.840	1	179	.000		
2	.880 <sup>b</sup>	.774	.26298	.048	38.003	1	178	.000		
3	.890 <sup>c</sup>	.793	.25227	.019	16.439	1	177	.000		2.072

- a. Predictors: (Constant), Life expectancy at age 60
- b. Predictors: (Constant), Life expectancy at age 60, Population using at least basic sanitation services (%)
- c. Predictors: (Constant), Life expectancy at age 60, Population using at least basic sanitation services (%), Alcohol, recorded per capita
- d. Dependent Variable: log\_Mortality\_rate

**R:** - It represents the correlation between predictor as Alcohol, recorded per capita (15+) consumption (in liters of pure alcohol), Life expectancy at age 60 (years), Total NCD Deaths (in thousands) with dependent variable ie. Mortality rate. As it is stepwise regression, value 0.890 represents it is good (right) level of the forecast.

**R- Square:-** also known as the coefficients of determination. It tells how much percentage of variance is contributed by each independent variable towards its target variable. In the third level of stepwise, we can say 0.789 which is 78.90% of variance explained by life\_expectancy at age\_60, population using at least\_basic\_sanitation\_service and alcohol, recorded\_per\_capita for Mortality rate. Which means we must consider these three variable for building our model.

**Adjusted R Square:-** Adjusted R square increases only if the new predictor improves the model more than would be expected by chance. It decrease when a predictor improves the model by less than expected by chance. The Adjusted R square value is 78.90% which signifies that it is a good fit.

**Durbin-Watson:-** observed Durbin Watson is 2.072 which means there is positive autocorrelation in residuals or in another word we can say in sample there is **no autocorrelation detected** which is quite good. It has two cases;

1. **0 to 2** (it illustrate no autocorrelation, (+) autocorrelation.
2. **2 to 4** (it illustrate correlation exist)

## 2. ANOVA:

ANOVA <sup>a</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	39.461	1	39.461	472.840	.000 <sup>b</sup>
	Residual	14.938	179	.083		
	Total	54.399	180			
2	Regression	42.089	2	21.044	304.294	.000 <sup>c</sup>
	Residual	12.310	178	.069		
	Total	54.399	180			
3	Regression	43.135	3	14.378	225.939	.000 <sup>d</sup>
	Residual	11.264	177	.064		
	Total	54.399	180			
a. Dependent Variable: log_Mortality_rate						
b. Predictors: (Constant), Life expectancy at age 60						
c. Predictors: (Constant), Life expectancy at age 60, Population using at least basic sanitation services (%)						
d. Predictors: (Constant), Life expectancy at age 60, Population using at least basic sanitation services (%), Alcohol, recorded per capita						

**ANOVA(Analysis of Variance):-** It is used to find out the result which we got by applying model is significant or not. It tells overall how well you're the model is good (right) fit to the data. Here we check F score value and significance value to illustrate that model is statistically significant at predicting target variable or outcomes.

- Observed F-score=472.840 and p=0.00, which means model is significant or we can say the regression model is a good fit.

### 3. Coefficients:

Coefficients <sup>a</sup>											
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	8.094	.144		56.142	.000					
	Life expectancy at age 60	-.159	.007	-.852	-21.745	.000	-.852	-.852	-.852	1.000	1.000
2	(Constant)	7.699	.146		52.728	.000					
	Life expectancy at age 60	-.117	.010	-.626	-12.252	.000	-.852	-.676	-.437	.487	2.054
	Population using at least basic sanitation services (%)	-.006	.001	-.315	-6.165	.000	-.763	-.419	-.220	.487	2.054
3	(Constant)	7.836	.144		54.390	.000					
	Life expectancy at age 60	-.127	.010	-.684	-13.397	.000	-.852	-.710	-.458	.448	2.230
	Population using at least basic sanitation services (%)	-.006	.001	-.341	-6.900	.000	-.763	-.460	-.236	.479	2.089
	Alcohol, recorded per capita	.023	.006	.160	4.055	.000	-.315	.292	.139	.755	1.325

a. Dependent Variable: log\_Mortality\_rate

- Beta= “0.023” state, if Alcohol, recorded per capita (15+) consumption (in liters of pure alcohol) is increasing by 1% then Mortality rate per 1000 population dying between 15 and 60 year age group is increased by 0.023 rate.
- If the percentage of the population using at least basic sanitation service is increase by 1% then the Mortality rate decreases by -0.006 rate.
- Beta= “-0.127” which means that if life expectancy at age 60 increase by one percent then Mortality rate decrease by -0.127 rate.
- Standardized coefficient beta is the same as standard deviation it provides similar information like we get through standard deviation. If percentage of population using at least basic sanitation service increases by 1 standard deviation then the mortality rate decreases by **-0.315**.

**EQUATION:-  $7.836 - 0.127 X_1 - 0.006 X_2 - 0.23 X_3$**

**Summarise :-** we checked the assumption for not violating. Linearity, NO influential data point, Normality, multicollinearity are not going against our assumption. In the case of Heteroscedasticity, we have applied log, so that it won't violate assumption. All predictors are statistically significant to make model as p- value is less than 0.05, stated under evaluation method. Stepwise multiple linear is used which tell in step 3, 89% of variance explained by variables to construct model than step 2.ie 88%.



## LOGISTIC REGRESSION:-

**Objective:** The fundamental aim is to apply logistic Regression model to predict the effect of “Alcohol\_recorded\_per\_capita”, “Life\_expectancy at age 60”, “Total NCD Death”, “Population\_using \_at least\_BasicSanitation\_services” on Logistic Motility rate (target variable).

**Dataset:** Used a similar dataset which we have used for multiple linear regression. Here we have converted or coded **continuous target variable into categorical target variable**. It indicate as follows:

1 → Died      0 → Not Died

	Country	Year	Alcoholrecorded percapita	Lifexpectancya tage60	TotalNCDDeaths	Populationusing atleastbasicsanitation services	LogisticMotilityrate
1	Afghanistan	2013	.03	16.2	109.5	38	1
2	Albania	2013	5.06	20.4	24.3	97	0
3	Algeria	2013	.54	21.6	144.0	87	0
4	Angola	2013	8.02	17.1	61.9	45	1
5	Antigua and Barbuda	2013	8.50	19.8	.5	88	0
6	Argentina	2013	8.20	21.4	254.5	94	0
7	Armenia	2013	3.72	19.4	26.3	92	0
8	Australia	2013	9.87	25.5	142.7	100	0
9	Austria	2013	11.60	24.0	74.4	100	0
10	Azerbaijan	2013	2.04	18.6	54.1	89	0
11	Bahamas	2013	8.55	22.4	1.8	94	0
12	Bahrain	2013	2.12	21.2	2.3	100	0
13	Bangladesh	2013	.01	19.2	572.6	43	0
14	Barbados	2013	8.88	19.6	2.7	95	0
15	Belarus	2013	12.37	18.8	107.5	97	1
16	Belgium	2013	10.33	23.7	94.9	99	0
17	Belize	2013	6.17	16.9	1.4	87	1
18	Benin	2013	1.45	17.1	35.1	15	1
19	Bhutan	2013	.19	20.2	3.2	65	1
20	Bolivia (Plurinational...	2013	3.73	21.2	45.9	53	1
21	Bosnia and Herzego...	2013	4.43	20.2	36.3	95	0

VARIABLE	TYPE	SCALE
<u>Mortality rate</u> (Dying between 15&60 year)	Dependent	Dichotomous
Alcohol, recorded per capita (15+)consumption( in L of pure alcohol)	Independent	Continuous
Life expectancy at age 60(year)	Independent	Continuous
Total NCD Deaths (in thousands)	Independent	Continuous
Population using at least basic sanitation services (%)	Independent	Continuous

### 1. Case Processing Summary:

Case Processing Summary			
Unweighted Cases <sup>a</sup>		N	Percent
Selected Cases	Included in Analysis	181	98.9
	Missing Cases	2	1.1
	Total	183	100.0
Unselected Cases		0	.0
Total		183	100.0
a. If weight is in effect, see classification table for the total number of cases.			

Observed case summary tells 183 cases were consider to build logistic model.

Dependent Variable Encoding	
Original Value	Internal Value
0	0
1	1

Here variable encoding is being done for both dependent

## 2. Block 0: Beginning Block:

Classification Table <sup>a,b</sup>					
	Observed		Predicted		
			LogisticMotalityrate		Percentage Correct
			0	1	
Step 0	LogisticMotalit	0	0	87	.0
	yrate	1	0	94	100.0
	Overall Percentage				51.9
a. Constant is included in the model.					
b. The cut value is .500					

The Above table states that the analysis is done without considering predictor variable. The Observed overall percentage is 51.9%. Which means 51.9% motility rate among data received as the rate of death is higher as compared to 'no death'.

Adding predictors (independent variable) model should improve its accuracy which means that independent variables have given an effect on its dependent variable.

## 4. Omnibus Test:

Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	135.375	3	.000
	Block	135.375	3	.000
	Model	135.375	3	.000

Sig value in omnibus test is observed 0.000, which is less than 0.005. It means selected independent variable are significant to our dependent variable. So we can say our predictors are significant.

## 5. Model Summary:

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	115.274 <sup>a</sup>	.527	.703
a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.			

Cox & snell and nagelkerke R square showing 52.7% to 70.3% of variability explained by independent variable to dependent variable.

## 6. Hosmer and Lemeshow Test:-

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	6.221	8	.622

Value of chi-square for Hosmer and lemeshow test is 6.221 with a significance level of 0.622. Sig value 0.622 which is larger than 0.05. Hence indicate a good fit to the model.

## 7 Variable in the Equation:

Variables in the Equation								
	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Step 1 <sup>a</sup> Alcohol, recorded per capita (15+) consumption (in liters of pure alcohol)	.165	.064	6.714	1	.010	1.179	1.041	1.336
Life expectancy at age 60 (years)/	-.619	.135	20.893	1	.000	.538	.413	.702
Population using at least basic sanitation services (%)	-.072	.019	14.952	1	.000	.931	.898	.965
Constant	17.272	2.880	35.963	1	.000	31702351.317		

a. Variable(s) entered on step 1: Alcohol, recorded per capita (15+) consumption (in liters of pure alcohol), Life expectancy at age 60 (years)/, Population using at least basic sanitation services (%).

- Wald score of life expectancy at age 60(year) is 20.893 which means it has higher contribution to a model as compare to another independent variable.

- All three predictors showing significant contribution to model as their all p-value lies below 0.05.
- B value of life\_expectancy at\_age\_60 and population using at least basic sanitation service is -0.619 and -0.072 respectively. It means both having a negative correlation with the dependent variable.
- Exp(B) of Alcohol, recorded per capita consumption=1.179 which is larger than 1. It means odd of death reports mortality rate increase by a factor of 1.179. Similarly, exp(B) life expectancy at age 60 = 0.538 which is less than 1. It means for every 1 unit of life expectancy the odd death reports mortality rate decrease by a factor of 0.538.

Now from above we can conclude:

$$Y = \frac{e^{17.272 - 0.72 X_1 - 0.619 X_2 - 0.165 X_3}}{1 + e^{17.272 - 0.72 X_1 - 0.619 X_2 - 0.165 X_3}}$$

## 8. Classification Table:

Classification Table <sup>a</sup>					
	Observed		Predicted		
			LogisticMotalityrate		Percentage Correct
			0	1	
Step 1	LogisticMotalit	0	78	9	89.7
	yrate	1	16	78	83.0
	Overall Percentage				86.2
a. The cut value is .500					

- This table calculates improvement when predictors are introduced to the model.
- It tells per case, how well the logistic model can forecast the correct category.
- The percentage of prediction has been increased from 51.9% to 86.2% after involving the independent variable.

**Summarise:** The 4 variable is used to find the likelihood of dependent variable. 52.7% to 70.3% of variances is given by independent variable to predict mortality rate and it has given insight on 86.2 % of case. Hosmer and Lemeshow Test and Omnibus Test suggested a good fit of the model.

## REFERENCES:

PALLANT, J., 2016. *SPSS Survival Manual, A step by step guid to Data Analysis using IBM Spss*. Sixth Edition ed. s.l.:The McGraw Hill, ISBN-10: 033526154X.