Multiple Regression Analysis

Objective:

Multiple regression is used to predicts value of the dependent variable base on the values of number of independent variable. The objective of this project is use of Multiple regression to predict the ratio of deaths of the Females and Males in the world of due to consumption of the Tobacco and alcohol. There are different type of cancer like mouth, throat (pharynx), voice box (larynx), oesophagus (foodpipe), breast, liver, bowel, Number of people died because of this type of cancers. This project use cancer as the dependent variable and tobacco and alcohol consumption as the independent variable for the analysis

2) Specify the number and levels of measurement of all independent/dependent variables in your analysis.

Dataset Description :

Datasets: Global Health Observatory data repository

This project uses one dependent and two independent variable for the analysis

1) Alcohol Dataset
This dataset contains **Percent** of alcohol consumption for the year of 2016 by the male and female in the world and this is used as the Independent variable.

2) Tobacco
This dataset contains **Percent** Tobacco Consumption for the year 2016 by the male and female in the world and this is used as the Independent variable.

3) Cancer
This dataset contains **Percent** of Death of the male and female for the year 2016 because of consumption of alcohol and cancer

Multiple linear regression is use to calculate death of male and female in the world based on consumption of alcohol and tobacco.

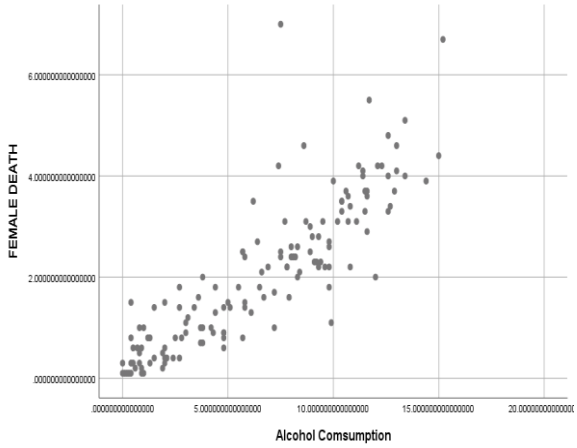
1) Sample size

N>50+8m (m= Number of independent variable)
N=143
In this case N=143 so sample size assumption is not violated.

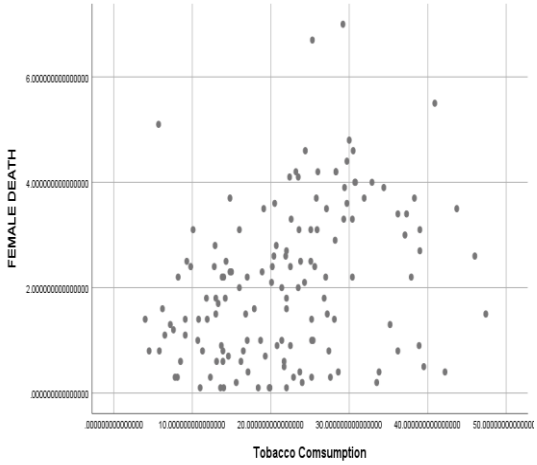
2) **Linearity**

First step is to checks linearity between the independent and dependent variable. Ad in this case dependent variable is female death and independent variable is Alcohol and tobacco consumption. and we need to check outliers visually as regression analysis is very sensitive to outliers.

Scatter plot



Linearity check between Female-death(DV)and Alcohol Consumption (IV)



Linearity check between Female-death(DV)and Tobacco Consumption

First graph shows that there is liner relationship between Female death due to cancer and the consumption of the Alcohol and second graph tells about the linear relationship between Female death and tobacco consumption so there is no need to transform data as assumption of linearity is not violated and the unusual cases or the outliers which impact on the regression model handled while creating regression model through SPSS and std deviation is set to be 3 and outliers beyond 3 will not be consider so this dataset is useful to run multiple Regression.

Initial run of the Regression which includes Descriptive statistic, Correlations, variable entered/removed, model summary, ANOVA, coefficients, collineartiy Diagnostics, Case wise diagnostics

Descriptive statistics:

This is use to describe the given data set. In this dataset total number of variable is are and mean and std deviation of each variable is calculated and number of sample is14.

	Mean	Std. Deviation	N
Female Death	2.070629371	1.455578	143
Alcohol Consumption	6.48951049	4.19382	143
Tobacco Consumption	21.9	9.612118	143

Correlations:
This is use to define strength and direction of the relationship between two variable.

		FEMALE DEATH	Alcohol Consumption	Tobacco Consumption
Pearson Correlation	FEMALE DEATH	1.000	0.862	0.328
	Alcohol Consumption	0.862	1.000	0.281
	Tobacco Consumption	0.328	0.281	1.000
Sig. (1-tailed)	FEMALE DEATH		0.000	0.000
	Alcohol Consumption	0.000		0.000
	Tobacco Consumption	0.000	0.000	
N	FEMALE DEATH	143	143	143
	Alcohol Consumption	143	143	143
	Tobacco Consumption	143	143	143

Assumption of Correlation :
Correlation Between Dependent and independent variable should be > 0.3 (SPSS survival manual)
Correlation Between independent variable should be < 0.7(SPSS survival manual)

According to Assumption
1) Correlation between Female death **(DV)** and Alcohol consumption is 0.8662(**IV**)
2) Correlation between Female death **(DV)** and Tobacco Consumption(**IV**) is 0.328 and
3) Correlation between Alcohol consumption **(IV)** and Tobacco Consumption **(IV)** is 0.328 and

So correlation assumption is not violated.

Variable entered:

This is used to define the number of variable entered in the model

Model	Variables Entered	Variables Removed	Method
1	Tobacco Consumption, Alcohol Consumption		Enter
a. Dependent Variable: FEMALE DEATH			
b. All requested variables entered.			

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.866[a]	0.75	0.747	0.7324767
a. Predictors: (Constant), Tobacco Consumption, Alcohol Consumption				
b. Dependent Variable: FEMALE DEATH				

R is 0.866 this value is generated when tobacco and alcohol consumption is use as predictor And this is simple correlation between the predictor and female death. This value should be >0.5.

R square is use to define the tobacco and alcohol consumption causes 75% of variation in the female death.

Adjusted R Square value (0.75) is almost same as that of R Square (0.747) this shows that if this model derived from the population rather than sample will account for (0.3%) variation in the predicted output.

Std Error is of the estimate is 0.7324.

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	225.744	2	112.872	210.377	.000[b]
	Residual	75.113	140	0.537		
	Total	300.857	142			

Significance of ANOVA should be <0.05. and in this case significance level is 0.00.
So this means independent variable has an impact on dependent variable.

Coefficient :

B values are most important in Coefficient as these values define the amount of degree each predictor affect the outcome. In this case b0 is 0.290 and b1 is 0.014. Beta values are same only expressed ad std deviations

So the death rate increase by 1 std deviation when the alcohol consumption increase by 0.015 std deviation ,Similarly the death rate increase by 1 std deviation when the Tobacco consumption increase by 0.007

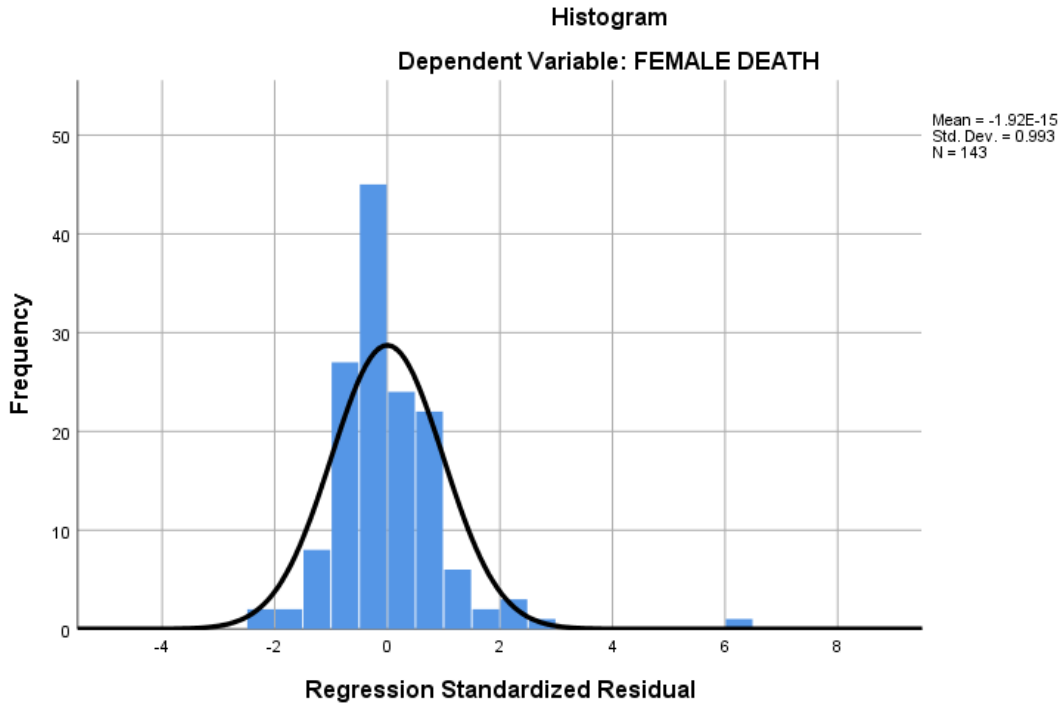
When the significance level for the b values is < 0.5 then then predictor making contribution so in this case values are 0.00 and 0.035 and value of t is 18.963 and 2.173 so the predictor contribution is very High

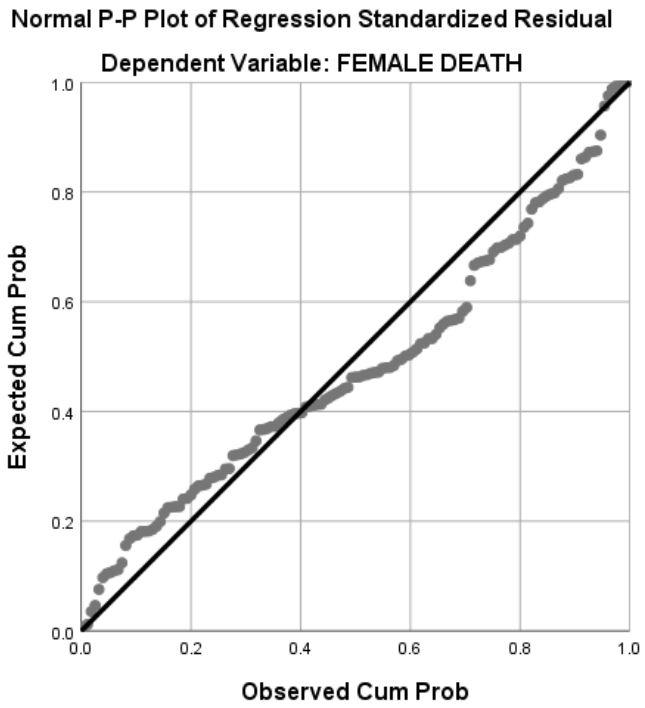
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-0.121	0.164		-0.740	0.460	-0.444	0.202					
	Alcohol Comsumption	0.290	0.015	0.835	18.983	0.000	0.260	0.320	0.862	0.849	0.802	0.921	1.086
	Tobacco Comsumption	0.014	0.007	0.094	2.127	0.035	0.001	0.027	0.328	0.177	0.090	0.921	1.086

Collinearity Diagnosticsa :

Model		Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	Alcohol Comsumption	Tobacco Comsumption
1	1	2.726	1.000	0.02	0.03	0.02
	2	0.190	3.783	0.10	0.96	0.15
	3	0.084	5.714	0.89	0.01	0.83

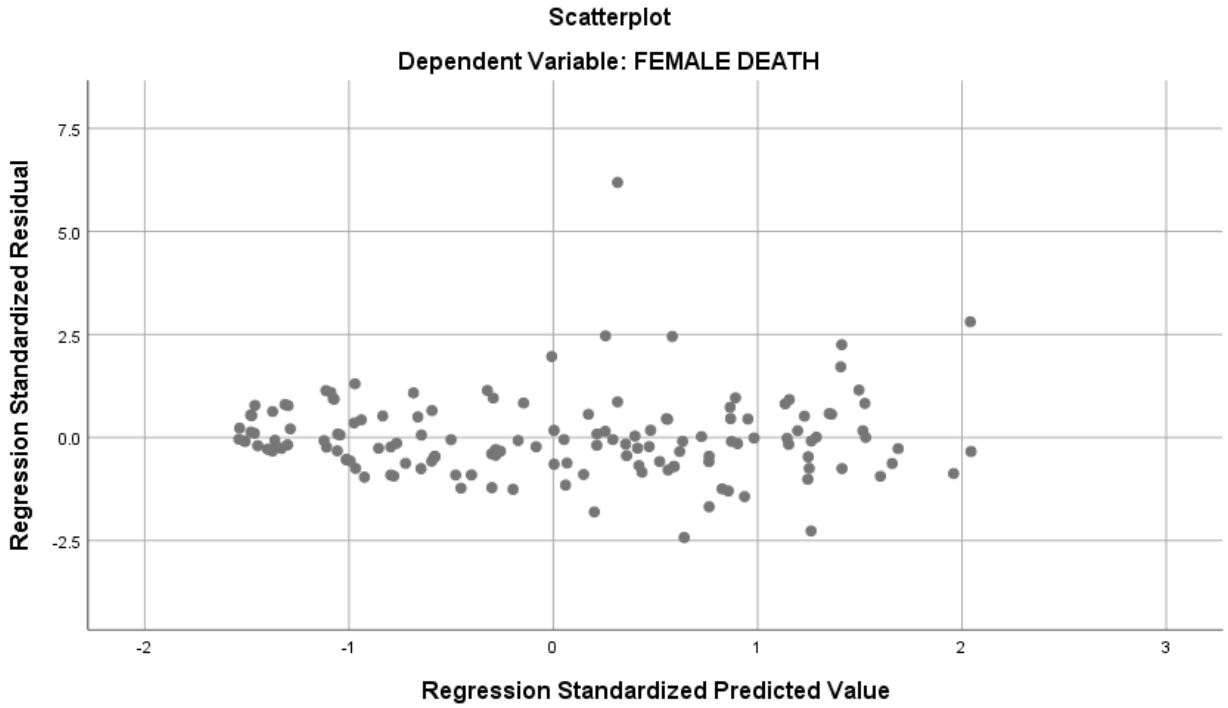
Tolerance is= 1- R2 and this value should be >0.10 in this case the value of tolerance is 0.921
The value of VIF is 1 / tolerance which should be <10. And in our case it is 1.086.





Check Residuals: error residuals are normally distributed and PP plot of the regression std residual is linear in nature

Homoscedasticity



For each value of the predictors the variance of the error term is constant as plotted in above scatter plot.

REFERANCE

Dataset links

World Health Organization (2016) Alcohol Data by country
At : http://apps.who.int/gho/data/view.main.1800 [Accessed 24 November 2018].

World Health Organization (2016) Prevalence of current tobacco use Data by country Available at: http://apps.who.int/gho/data/view.main.GSWCAH20v [Accessed 24 November 2018].

World Health Organization (2016) Alcohol-attributable fractions, cancer deaths by country Available at: http://apps.who.int/gho/data/view.main.53481 [Accessed 24 November 2018].

SPSS SURVIVAL MANUAL : JULLIE PALLANT

LOGISITC RESGRESSION:

OBJECTIVE:

Logistic regression is use to analyse when the dependent variable is dichotomous (binary)and useful for the predictive analysis. I this project dataset contains number of death of male and female because of the different diseases namely Infectious and parasitic diseases, Tuberculosis, Intestinal infectious diseases, Viral hepatitis. This project uses logistic analysis to identify whether male or female is more impacted by the diseases.

DATASET:

In this project total 4 datasets are used and these dataset two common column country and sex these dataset are combine to create one dataset

Below are the link of the datasets.

1) Intestinal infectious diseases, number of deaths, by sex.

2) Viral hepatitis, number of deaths, by sex.

Below is the sample data us for the analysis which contain new variable (Sex).

I have created this variable to represent MALE and FEMALE as 1 and 0.

Sample data

COUNTRY	category	Intestinalinfectiousdiseasesnumberofdeathsbysex	Viralhepatitisnumberofdeathsbysex	SEX
KAZ	MALE	28	26	1
KAZ	FEMALE	34	25	0
AND	FEMALE	0	0	0
AND	MALE	0	1	1
ARM	FEMALE	3	16	0

Level's of measurement of all independent/dependent variables

1) Dependent variable

Sex

2) Independent variable

Intestinal infectious diseases, number of deaths, by sex

Viral hepatitis, number of deaths, by sex

Name	Type	Width	Deci...	Label	Values	Missing	Columns	Align	Measure	Role
COUNTRY	String	3	0	COUNTRY	None	None	9	Left	Nominal	Input
CATEGORY	String	6	0	CATEGORY	None	None	11	Left	Nominal	Input
INTESTINALINF...	Numeric	4	0	INTESTINAL INFEC...	None	None	18	Right	Scale	Input
VIRALHEPATITI...	Numeric	4	0	VIRAL HEPATITIS, ...	None	None	13	Right	Scale	Input
SEX	Numeric	8	0	SEX	None	None	6	Right	Nominal	Input

Logistic regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	70	100.0
	Missing Cases	0	.0
	Total	70	100.0
Unselected Cases		0	.0
Total		70	100.0

a. If weight is in effect, see classification table for the total number of cases.

Total number of applied cases are 70 and these cases and details are confirmed in case processing summary

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

IBM SPSS require Dependent variable value should be in 0 and 1.if the value of dependent value is not 0 and 1 toll will automatically create the value as 0 and 1.In this case I have already converted value into 0 and 1.

Classification Table

Block 0: Beginning Block

Observed			Predicted		
			sex		Percentage Correct
			0	1	
Step 0	sex	0	0	35	0
		1	0	35	100
	Overall Percentage				50

This is result of the analysis without any of the independent variable used in the model and this is used for comparing model with our predicted variable included.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	0.839	2	0.657
	Block	0.839	2	0.657
	Model	0.839	2	0.657

Omnibus Tests of Model gives us how well model perform over the result from the block 0 and this is called goodness of fit test. Performance of the model is dependent upon the significant value which is <0.05.but in this case the value is >0.05 which is 0.657.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3.138	8	.925

This is one of the test which is useful for the checking of the model whther it is working correctly or not. according to Hosmer and lemeshow test poor fit is define as the significance level is less than <0.05 but for the value of 3.138 the significance level is 0.925 which is much better than the 0.05 this indicate support for the model .

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	96.201[a]	.012	.016

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

This summary gives us amount of variation in the dependent variable evaluated from the model so the values are 0.012 and 0.016 i.e 1.2 and 1.6 percent of variability from the set of variable

Classification Table

Observed			Predicted		
			sex		Percentage Correct
			0	1	
Step 1	sex	0	13	22	37.1
		1	5	30	85.7
	Overall Percentage				61.4

This table gives us how well model predict the categorical values. And this is compare with the classification table of the block 0. When the independent variable are consider the model correctly classified 61.4 of the cases overall an improvement 50% in block 0 .In this case 37.5% of the male have the problem with the disease and 85.1% of male candidate have the problem.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I .for EXP(B)	
								Lower	Upper
Step 1[a]	Viral hepatitis, number of deaths, by sex	0.000	0.001	0.218	1	0.641	1.000	0.998	1.002
	Intestinal infectious diseases, number of deaths, by sex	0.001	0.001	0.681	1	0.409	0.999	0.998	1.001
	Constant	0.037	0.271	0.019	1	0.891	1.038		

This table gives us information about the contribution of each of our predictor variable in this case it is Viral hepatitis, and Intestinal infectious. And the value of the significance should be < 0.05 for the variable, as the value is less variable has more impact on the model for the prediction .In this Case the value for significant is > 0.05 so we can say that this insignificant and have less impact on the model prediction ability.

- B values are used in the equation to calculate probability of the case into the category. this will also define the positive and negative relationship between the variable as the value of independent variable increases value of the dependent variable is increase as its positive value increase. Exp (B) column is the odds ratio for the independent variable and the 95% of the confidence interval of the lower and upper bound Odds of a person

having Intestinal infectious 0.99 times higher for someone who reports person who does not have Intestinal infectious problem (all other factors being equal).

Casewise List[a]

a. The case wise plot is not produced because no outliers were found.

Case wise list gives the information about the case in the sample for which sample dose not fit. In this sample there is no outliers so case wise list is not produce.

Conclusion : There are limitations to this model as there are data constraint and limited sample size because of that we are not able to add more variables along with the existing one which could increase the fitness of the logistic regression model.

Dataset links

World Health Organization (2015) Intestinal infectious diseases, number of deaths, by sexAvailable at: https://gateway.euro.who.int/en/indicators/hfamdb_416-deaths-intestinal-infectious-diseases/ [Accessed 24 November 2018].

World Health Organization (2015) Viral hepatitis, number of deaths, by sex . Available at: https://gateway.euro.who.int/en/indicators/hfamdb_821-deaths-viral-hepatitis/ [Accessed 24 November 2018].

SPSS SURVIVAL MANUAL : JULLIE PALLANT