Statistics(CA-2)

SAIGIRISH PALAVARAPU

Student ID-x17170401

Multiple Regression:

Dataset source: Multiple Linear Regression and Logistic Regression

1) https://ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing?file=data/tran_sf_roadse.tsv.gz
2) https://ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing?file=data/tran_sf_roadag.tsv.gz
3) https://ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing?file=data/tran_sf_roadve.tsv.gz
4) https://ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing?file=data/tran_sf_roadro.tsv.gz

I combined all these data sets and taken the columns which are used for multiple and logistic regression. The final data set consists of the four independent variables and one dependent variable that is the total number of accidents and I took the variables which are majorly affecting the dependent variable.

Correlation: This technique describes the strength between the two variables and as well as the direction of the linear relationship between them, these correlation is used in regression analysis like

1)Simple linear regression

2)Multiple linear regression.

Regression:

If we consider statistical modeling, regression analysis is a model to define the relation between two variables which are dependent and independent to each other, and in this regression, we can use one dependent variable with one dependent or one dependent with multiple independent variables which are defined clearly in the below sections.

Simple Linear regression:

In this simple linear regression, we can predict one dependent variable outcome with the help of another independent variable which is continuous. let us consider X and Y are two continuous and value of X depends on Value of Y.to analyze the relation between the two variables statistically we use the Durbin-Watson value, Standard error of estimate in Regression Model.

Equation: X=a + bY

Multiple Linear Regression:

This regression is also represented as same as simple linear, but the only change is the number of independent variables which affect the dependent variable.

In this we use a single dependent variable with multiple independent variables and the equation is represented as below

Equation: X=b0+b1Y+b2Z where x is Dependent and Y, Z is independent of each other.

This process makes it easier to analyze the complex issues which are real-time and even in the laboratory. This multiple regression can be used to predict how well the set of variables able to guess the output.

In my dataset, I wanted to know how far the independent variables are affecting the dependent variable so that am taking dependent as a total number of accidents and independent variable are gender and age.

Objectives Of Multiple Linear Regression:

There are five different types of data sets on road accidents for the year 2016 from ec.europa.eu,

1)Number of road Accidents in 2016 by gender wise

2) Number of road accidents by their age

3) Number of road accidents by type of road

4) Number of road accidents by type of vehicle

5)Total number of accidents in 2016

Variables used:

Dependent and independent variables chosen to analyze the data,

Dependent Variable:

A total number of accidents on the road in the year 2016 by Eurostat which is an authentic government website which focuses on the statistics of people.

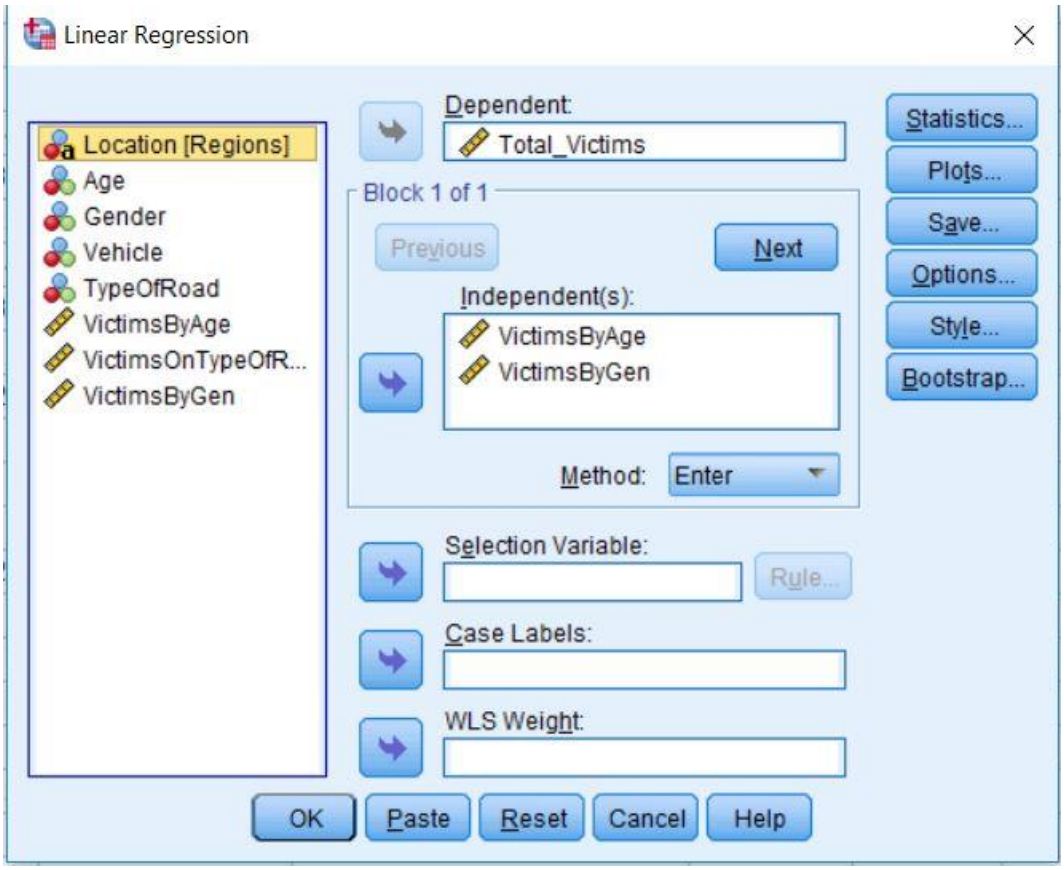
Independent variable:

1)Number of Accidents done on the road by males and females individually

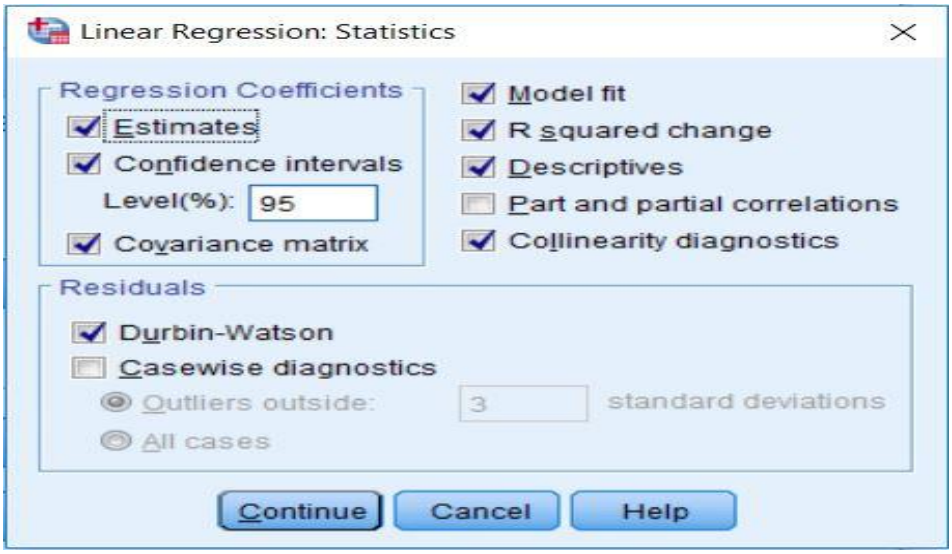
2) Number of accidents done by each age group and the age group is divided into certain frequencies.

Steps underwent in SPSS:

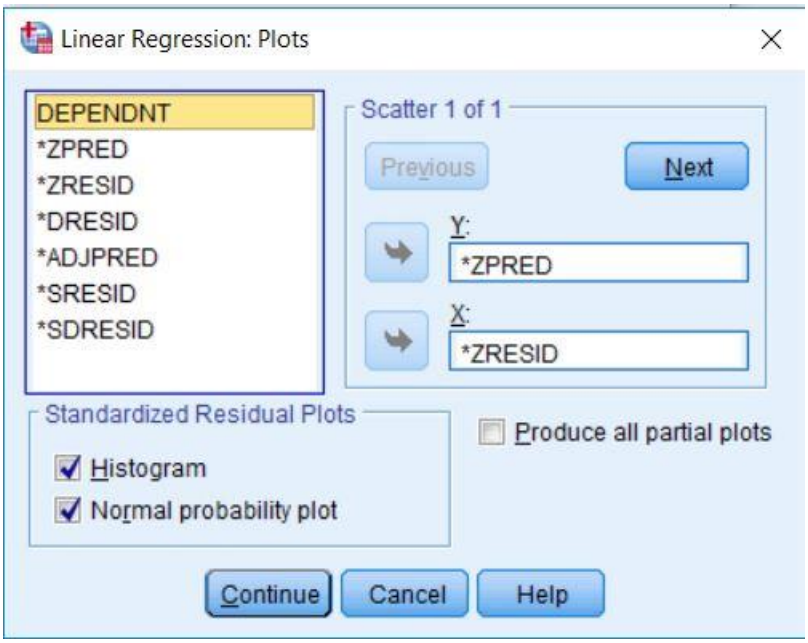
1)Go to analyze ➟ Regression➟ Linear and choose dependent and independent variables.



2)Then to check for Estimates under Regression Coefficients go to statics tab and the select model fit, R Square changes, Descriptive and collinearity diagnostics and in regression coefficients select the estimates confidence interval and in residuals check the box of Durbin-Watson.



3) Under Plots tab, insert *ZPRED to Y-axis and *ZRESID to X-axis and then for plotting the residuals, check Histogram and Normal probability plot and then Click continue.



4)the last step to generate output is to click OK in the Multiple Regression Tab

Explanation of Output:

Descriptive Statistics

	Mean	Std. Deviation	N
Total_Victims	1774.02	2720.369	120
VictimsByAge	461.29	2406.961	120
VictimsByGen	405.37	773.005	120

From the above table, it gives the primary information about the independent variables like Mean standard deviation, and observation in that variable. Here N represents the sample size of the variable.

Correlations:

The correlation table represents the correlation between the dependent and the independent variable and the correlation which is low that is defined as the variables is suitable to use in the regression model as an independent variable.

Correlations

		Total_Victims	VictimsByAge	VictimsByGen
Pearson Correlation	Total_Victims	1.000	.854	.247

Here in the above table regression sometimes represented as negative also.

Model summary:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.922[a]	.850	.842	1080.224	.850	106.950	6	113	.000	2.353

a. Predictors: (Constant), TypeOfRoad, VictimsByAge, VictimsByGen, Age, Gender, Vehicle

b. Dependent Variable: Total_Victims

The model summary is very important when analyzing the dependent and independent variable. In this, it consists of many values like R, R square, Adjusted R square etc.

R square value gives that the combines influence of independent variables on the dependent variable and in this table, it gives R square value as 85% and Adjusted R square will give the value by considering the residuals in the data and we got that as 84.2%, so is no much more difference between the values.

We can find the significance value from last but one column in the above table, we know that the significance value should be less than 0.05 that is 5% to prove that our model is a valid mode and here we got 0.0003 which is less than 0.5, so we have a statistical evidence to prove our model is valid.

The value for Durbin-Watson is d=2.353 and for an ideal process the value should be in between 1.5 and 2.5 so it proves that there is no first order linear correlation in between members of series in multiple linear regression.

ANOVA:

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	748790691.799	6	124798448.633	106.950	3.106E-44
	Residual	131857812.972	113	1166883.301		
	Total	880648504.771	119			

a. Dependent Variable: Total_Victims

b. Predictors: (Constant), TypeOfRoad, VictimsByAge, VictimsByGen, Age, Gender, Vehicle

The ANOVA table is obtained as a result to the F-test. It is a test that states if R square

obtained from the model summary table is significantly greater than 0. Since the p-value is less than

As a result of F-Test, we got ANOVA table. From this test, we can get the sum of squares, degrees of freedom and the significance value. If R square from the model table is greater than zero and p-value is less than 0.05 and the regression model is significant.

The whole model is represented as,

F (6, 113)=106,p;0.05,R square=0.85.

Coefficients:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	883.978	121.874		7.253	.000	642.613	1125.344		
	VictimsByAge	.988	.044	.874	22.353	.000	.900	1.076	.996	1.004
	VictimsByGen	1.071	.138	.304	7.784	.000	.799	1.344	.996	1.004

a. Dependent Variable: Total_Victims

Unstandardized coefficients will explain that if a one-unit change of dependent variable then there is .98 units changes in the victims by Age. for example, if the total victims are changed by one unit then the victims by age will be changed to .98 units.

If the significance value that is p-value is less than 0.05 then the variable is showing a significant difference on the dependent variable, the present table we got the two values which are less than 0.05 but not equal to 0.

The 95% confidence interval gives the lower bound and the upper bound of the variable.

Collinearity test gives the two values they are tolerance and variance influence factor(VIF), in linear regression shows the level of difference in the regression approximates are raised because of the multicollinearity. If VIF is above 10 then the multicollinearity is difficult.

The Beta value is also consisting of the same interpretation that if a unit change of dependent value then the beta value consists of .87 units of change for victims by age.

Collinearity Diagnostics:

Collinearity Diagnostics

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	VictimsByAge	VictimsByGen
1	1	1.514	1.000	.24	.06	.22
	2	.979	1.244	.00	.84	.11
	3	.507	1.728	.76	.10	.67

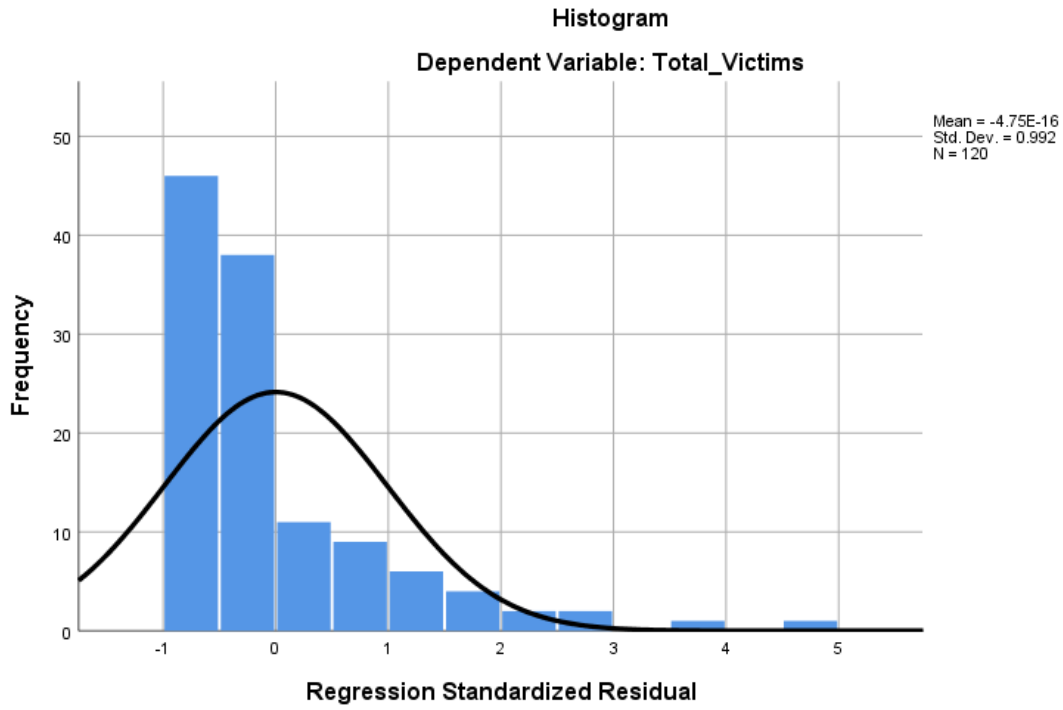
a. Dependent Variable: Total_Victims

Residuals Statistics

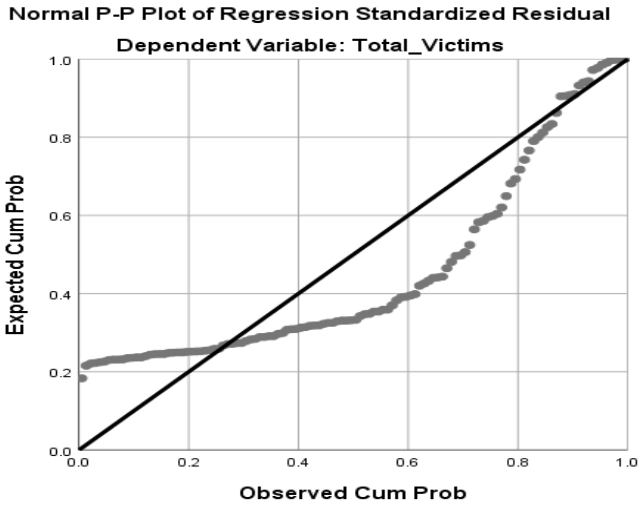
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	883.98	26219.77	1774.02	2466.153	120
Residual	-1045.444	5425.022	.000	1148.258	120
Std. Predicted Value	-.361	9.913	.000	1.000	120
Std. Residual	-.903	4.685	.000	.992	120

a. Dependent Variable: Total_Victims

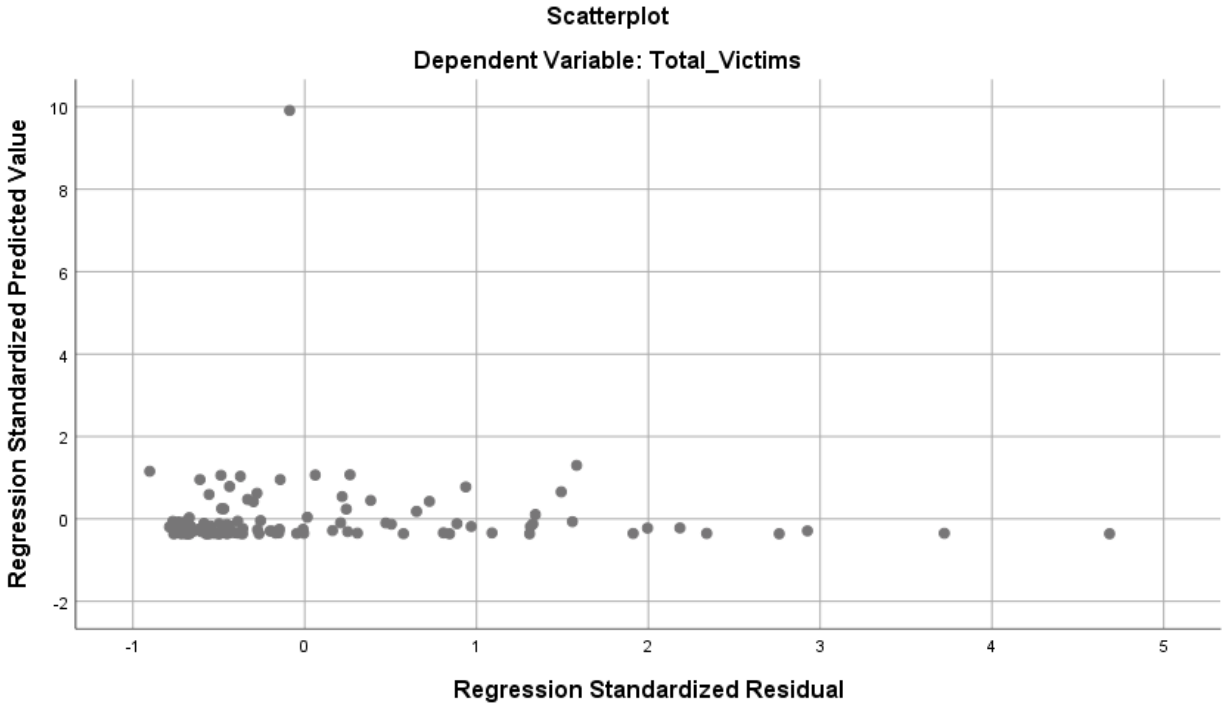
Diagrammatical representation of the data:



The table which used to look up for the histogram is the residual table and the histogram shows the normal curve which the data is normally distributed. The mean of the curve is 1774 and the standard deviation is 2466 for the population of 120.



From the P-P plot of regression standard residuals upon total victims is observed that the dotted line and the solid line, there should not be much difference between those two and the closer they are the output significance is greater.



Conclusion:

The conclusion states that the two independent variables are impacting the dependent variables up to 80% approximately and a dependent variable is a total number of victims and the independent's variables are victims by age and the victims by gender.

These two are the independent variables have more impact on the total number of accidents occurred in 2016. Road accidents.

Logistic Regression:

The logistic Regression is used when the dependent variable is a categorical variable which is dichotomous. The dependent variable which we use in the data must be continuous and which have the normal distribution and the independent or the predictor variable can be a categorical or continuous variable.

Details of Example:

To explain the logistic regression, I will be using real-time data on road accidents in 2016 in Europe. These data are taken from the links which I have mentioned above.

To make sense when the interpretation of results obtained we need to code the variables in the variable view section. I have coded the gender variable like the male is Equal to 1 and the female is 0.

Summary:

Question: The Researcher wants to know that if an accident happened on the roads of Europe then is there any impact of victims on their gender.

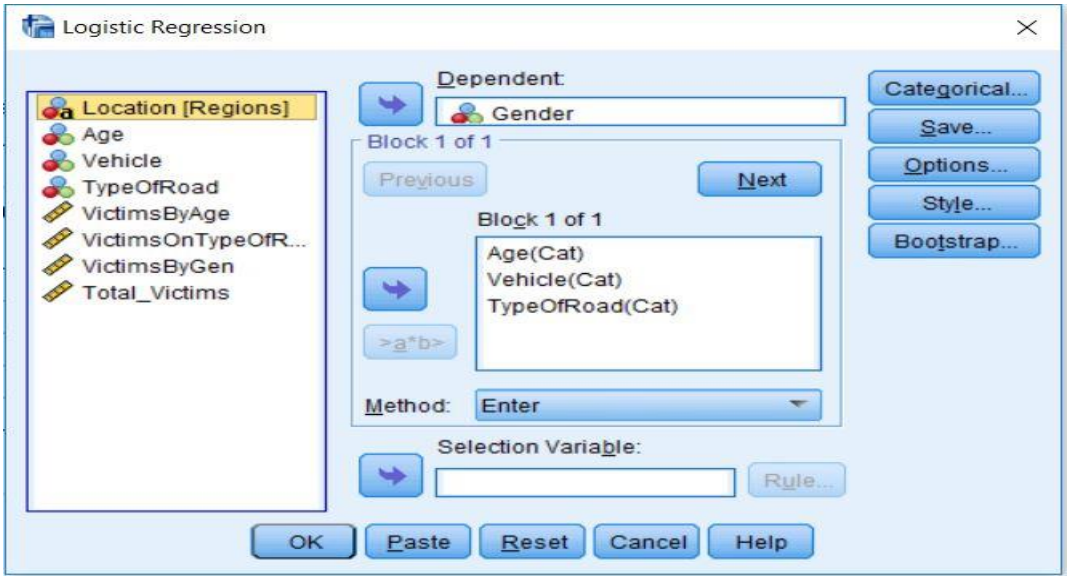
- One categorical dependent variable which dichotomous either male or female in other words either 1 or zero and
- Three categorical variables which are the type of road, age, and the vehicle they driving will these factors impact on gender.

The logistic regression will allow you to know how well the predictor variable or the independent variable set by you will explain your dependent or the categorical variable.

Procedure In SPSS:

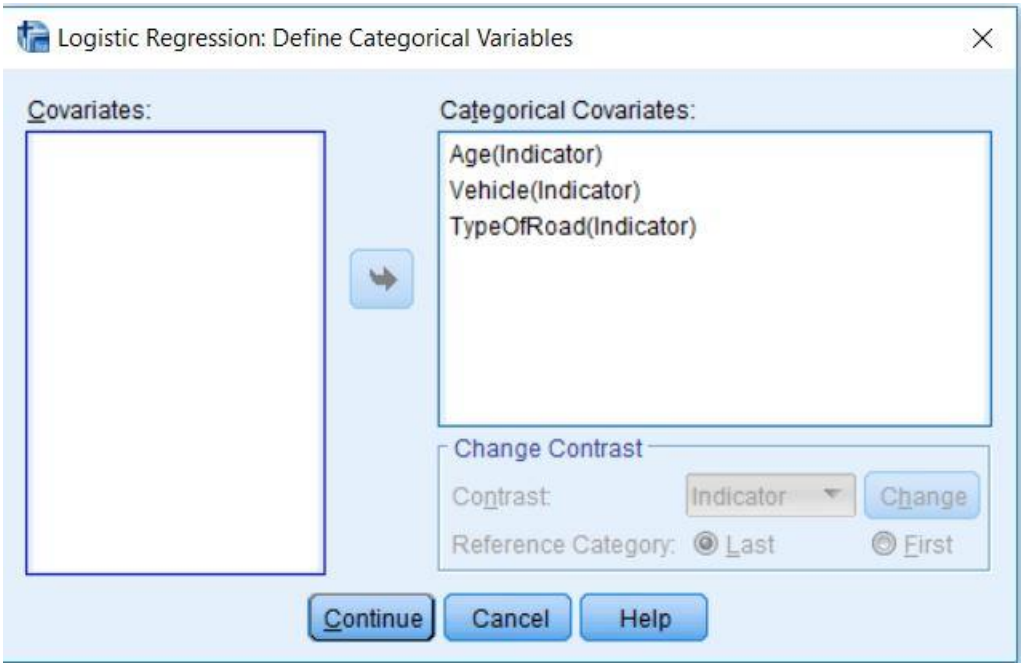
The process to obtain the logistic regression in spss is simple but the interpretation takes time.

1)go to the menu of the top of the system screen then click on analyze and go to Regression, a then binary logistic.

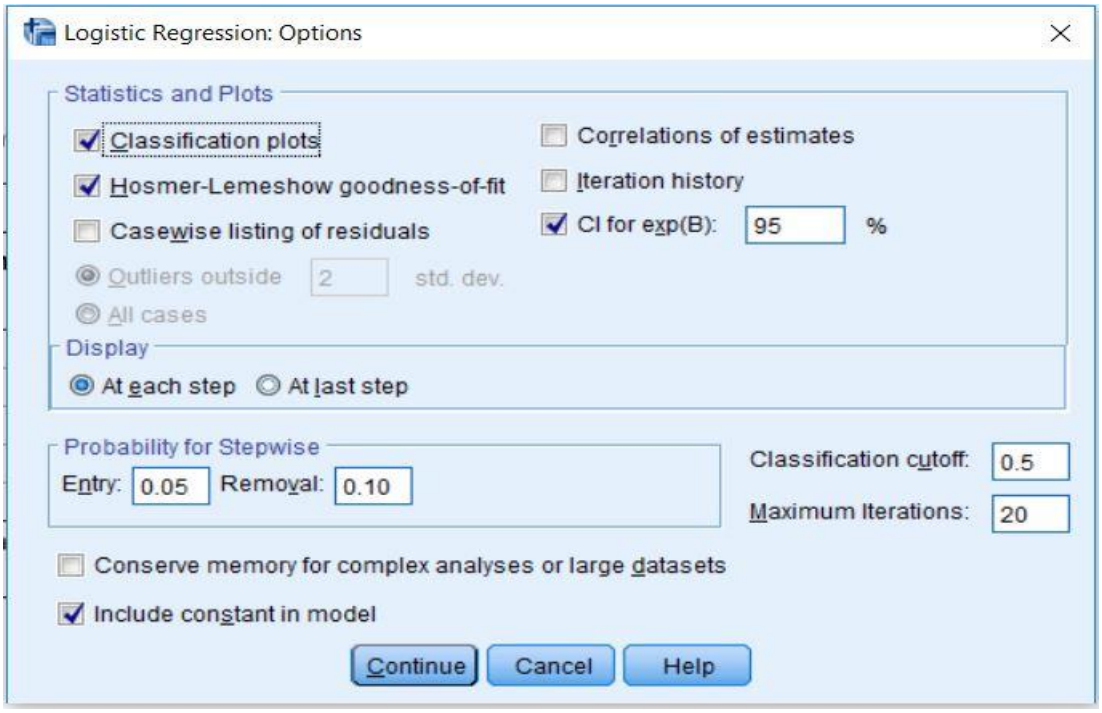


2)choose your dependent variable which is categorical and move towards dependents tab.

3)In my dataset, my independent variables have categorical variables so go to the categorical section and move the variables to the categorical covariates section and then click continue.



4)for analyzing the data further go to the options section and then click the check boxes of classification plots, Hosmer-Lemeshow goodness of fit and CI(Confidence interval) for exponential(B) which is default 95% which shown below.



Then click on continue and the ok, the results will be obtained. As the spss will give us a massive amount of information about the data on logistic regression, am going to explain the key concepts in the below section.

Interpretation of output:

1) **Dependent Variable Encoding:**

Dependent Variable Encoding

Original Value	Internal Value
FEMALE	0
MALE	1

The dependent variable encoding tells that which variable is coded with 1's and 0's. if the variables are coded with other numbers then the spss will convert to them automatically.

2)Classification table:

Classification Table[a,b]

			Predicted		
			victim		
	Observed		KIL	INJ	Percentage Correct
Step 0	victim	KIL	0	33	.0
		INJ	0	66	100.0
	Overall Percentage				66.7

a. Constant is included in the model.

b. The cut value is .500

The table above tells the information about a number of males and females that are considered while obtaining the results of the project and their percentages and how well the model is predicting. the overall prediction was 66.7% in block 0.

3)**Categorical Variable Coding:**

Categorical Variables Codings

			Parameter coding		
		Frequency	(1)	(2)	(3)
Vehicle	ANI_RD	11	1.000	.000	.000
	BIKE	31	.000	1.000	.000
	BUS	31	.000	.000	1.000
	BUS	54	.000	.000	.000
Age	65+	33	1.000	.000	.000
	<65	32	.000	1.000	.000
	15-17	31	.000	.000	1.000
	18-24	31	.000	.000	.000
TypeOfRoad	RD_URB	30	1.000	.000	
	MWAY	58	.000	1.000	
	RD_RUR	39	.000	.000	

The categorical variable coding table gives how many categories are there in one categorical variable and including its frequencies.

4)**Variables Not in the equation:**

Variables not in the Equation[a]

			Score	df	Sig.
Step 0	Variables	Age	64.601	3	.000
		Age(1)	43.375	1	.000
		Age(2)	21.802	1	.000
		Age(3)	4.463	1	.035
		Vehicle	31.529	3	.000
		Vehicle(1)	14.790	1	.000
		Vehicle(2)	6.391	1	.011
		Vehicle(3)	16.953	1	.000
		TypeOfRoad	28.135	2	.000
		TypeOfRoad(1)	1.072	1	.301
		TypeOfRoad(2)	25.030	1	.000

a. Residual Chi-Squares are not computed because of redundancies.

The above table which gives the information about the categories in the categorical variable individually with their degrees of freedom and its significance value. we know that if sig value or the p-value < 0.05 then the variables are significant so from the above table we conclude that the variables which we are using are significant.

5)**OmniBus Test:**

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	105.729	7	.000
	Block	105.729	7	.000
	Model	105.729	7	.000

In the omnibus test, we observe that the model improves in block 1 then in block 0. The chi-square value of the test is 105.7 at 7 degrees of freedom and the p-value or the sig value is not equal to zero but less than 0.05so that we can say that the model is good to fit for logistic regression.

6)Model Summary:

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	86.771[a]	.327	.455

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

“model summary gives us information about the usefulness of the obtained model. the Cox & Snell R square and the Nagelkerke R Square values provide an indication of the amount of variation in the dependent variable explained by the model (0<value<1)” (Julie.p,2016)
From the table above, we can say the R values obtained is Pseudo R values rather than actual R values. The Nagelkerke R Square value is .455 and the Cox & Snell R square value is .327 from the obtained output.

7)Homer and Lemeshow test:

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.000	2	1.000

This test is more reliable for the model fit available in spss and it is different from the omnibus test and even interpreted differently. In this test, if the significant value or the p-value is less than 0.05 then the model is considered to be the poor fit to the model.
In our table, the p-value is 1.0 so according to Homer and Lemeshow test the model is reliable.

8)Classification table in Block 1

Classification Table

			Predicted		
			victim		Percentage
	Observed		KIL	INJ	Correct
Step 1	victim	KIL	17	16	51.5
		INJ	13	53	80.3
	Overall Percentage				70.7

a. The cut value is .500

This table tells us that how well the model is able to predict the correct category for each and every case in the depend variable.so by using the above model we can predict the killed victims up to 51% and injured up to 80% and overall the model predicted approximately 71%, which is an improvement when comparing to block 0.

9)Contingency Table for Hosmer and Lemeshow test:

Contingency Table for Hosmer and Lemeshow Test

		victim = KIL		victim = INJ		
		Observed	Expected	Observed	Expected	Total
Step 1	1	17	17.000	13	13.000	30
	2	16	16.000	17	17.000	33
	3	0	.000	30	30.000	30
	4	0	.000	6	6.000	6

10)Variables in the equation:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Accident			.000	2	1.000	
	Accident(1)	-21.142	23205.767	.000	1	.999	.000
	Accident(2)	.000	24338.064	.000	1	1.000	1.000
	pers_inv			.000	2	1.000	
	pers_inv(2)	-21.471	33628.106	.000	1	.999	.000
	pers_inv(3)	.000	24338.382	.000	1	1.000	1.000
	Constant	21.203	23205.767	.000	1	.999	1615484950

a. Variable(s) entered on step 1: Accident, pers_inv.

The above table is the most important table when considering the logistic regression. it estimates the values of Beta, Standard error, Wald value, degrees of freedom, significant value, the exponential of B.

If we consider Beta value if a unit increase in the dependent variable value then there is a decrease at 21.142 units in the Accident (1) and also decrease of 21.471 in the pers_inv variable.

If we consider the significant value or the p-value, no value in the table is less than 0.05 so the dependent variables which are used are not providing the statistical evidence for the model to be statistically significant.

Conclusion:

By considering only two independent variables it is proved that the victims of road accidents do not depend on the gender and it's not enough to take only two categorical variables in this logistic Regression.

The model may be significant if we consider more variables which are independent to the dependent variables.