

Statistics (CA- 2)
Student ID: x18110096

Data Source:

I have used combined 4 datasets sourced from Eurostat and performed Multiple Regression and Logistic Regression Analysis.

Link:

<https://ec.europa.eu/eurostat/data/database>

Data Navigation tree:

Database by theme > Transport > Multimodal Data > Transport safety > Rail transport safety.

- 1) Rail accidents by type of accident:
- 2) Rail accidents victims by type of accident:
- 3) Rail accidents involving the transport of dangerous goods
- 4) Suicides involving railways

Data Description:

- The rail accident data are provided to Eurostat by the European Railway Agency (ERA). The ERA manages and is responsible for the entire data collection. The Eurostat data constitute a part of the data collected by ERA and are part of the so-called Common Safety Indicators (CSIs). In Eurobase, the following data are available:
 - Number of rail accidents by type of accident.
 - Number of rail accident victims by type of accident
 - Number of rail accidents involving the transport of dangerous goods
 - Number of suicides involving railways.

Statistical Concept behind the data:

- The Common Safety Indicators (CSIs) is set of indicators allowing to measure railway safety performance of Member States and provides an objective evidence of the development over time within Member States. Accidents data are also used for estimating national reference values, setting common safety targets and assessing their achievement.

Unit of measure:

- The unit of measure is either the number of accidents, or the number of persons that have been injured or killed in railway accidents. For the table on suicides, the unit is the number of persons.
- 'Accident' in this respect means an unwanted or unintended sudden event or a specific chain of such events which have harmful consequences; accidents are divided into the following categories: collisions, derailments, level-crossing accidents, accidents to persons caused by rolling stock in motion, fires and others;
- 'Killed person' means any person killed immediately or dying within 30 days as a result of an accident, excluding suicides.
- 'Seriously injured person' means any person injured who was hospitalised for more than 24 hours as a result of an accident, excluding attempted suicides.
- A separate table is available with information on the number of suicides involving railways.

Correlation

It is an analysis to describe the strength and direction of the linear relationship between two variables.

Multiple Regression

Objective of Multiple Regression :

The objective of multiple regression analyses is to predict the total number of deaths due to rail accident with help of Accidents while moving goods and Victims on type of accident.

Level of measurement :

Variable Name	Variable Type
Total number of deaths due to rail accident	Dependent
Accidents while moving goods	Independent
Victims on type of accident	Independent

Correlation table:

Correlations				
		TotalVictims	AccedientsWhileMovingGoods	VictimsOnTypeOfAccidents
Pearson Correlation	TotalVictims	1.000	.796	.444
	AccedientsWhileMovingGoods	.796	1.000	.460
	VictimsOnTypeOfAccidents	.444	.460	1.000
Sig. (1-tailed)	TotalVictims	.	.000	.000
	AccedientsWhileMovingGoods	.000	.	.000
	VictimsOnTypeOfAccidents	.000	.000	.
N	TotalVictims	3108	99	2760
	AccedientsWhileMovingGoods	99	99	99
	VictimsOnTypeOfAccidents	2760	99	2760

- This table is used to check correlations between the variables in the model.
- This table provides information that whether the independent variables show relationship with the independent variable which be above 0.3.
- In this case, the statistics in above table indicates that the dependent variable Total victims is strongly correlated to the independent variables Accidents while moving goods and victims on type of accident.
- Also, according to Tabachnick and Fidell the variables Accidents while moving goods and victim on type of accident has 0.46 correlation which is less than 0.7, therefore variable will be retained

Tolerance and VIF:

Coefficients ^a								
		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	−9.642	4.925		−1.958	.053		
	AccedientsWhileMovingGoods	27.368	2.509	.751	10.910	.000	.788	1.268
	VictimsOnTypeOfAccidents	.208	.144	.099	1.438	.154	.788	1.268

a. Dependent Variable: TotalVictims

- The above table indicates that the tolerance value for both the variables is greater than 0.10, this indicates that there is absence of multi collinearity.
- VIF(Variance Inflation Factor) which is 1.268 is inverse of Tolerance is less than 10 which indicates absence of multicollinearity.

Model Summary:

Model Summary ^b									
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	Change Statistics			
						F Change	df1	df2	Sig. F Change
1	.801 ^a	.642	.634	46.864	.642	85.941	2	96	.000
a. Predictors: (Constant), VictimsOnTypeOfAccidents, AccedientsWhileMovingGoods									
b. Dependent Variable: TotalVictims									

- The R square value is 0.642.
- The adjusted R value is **0.634** which is nearby to R Square value.
- As the model is not too big so we consider using R-square value.
- This model explains that **63.4%** of the variance in dependent variable Total Victims.

Comparing significance value from Anova table:

ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	377493.945	2	188746.972	85.941	.000 ^b
	Residual	210838.806	96	2196.238		
	Total	588332.751	98			
a. Dependent Variable: TotalVictims						
b. Predictors: (Constant), VictimsOnTypeOfAccidents, AccedientsWhileMovingGoods						

- This table is formed as a result of F-test.
- By looking at the above anova table, it can be seen that the regression model is statistically significant, because the significance value 0.00 is less than **0.05**.

Coefficients table:

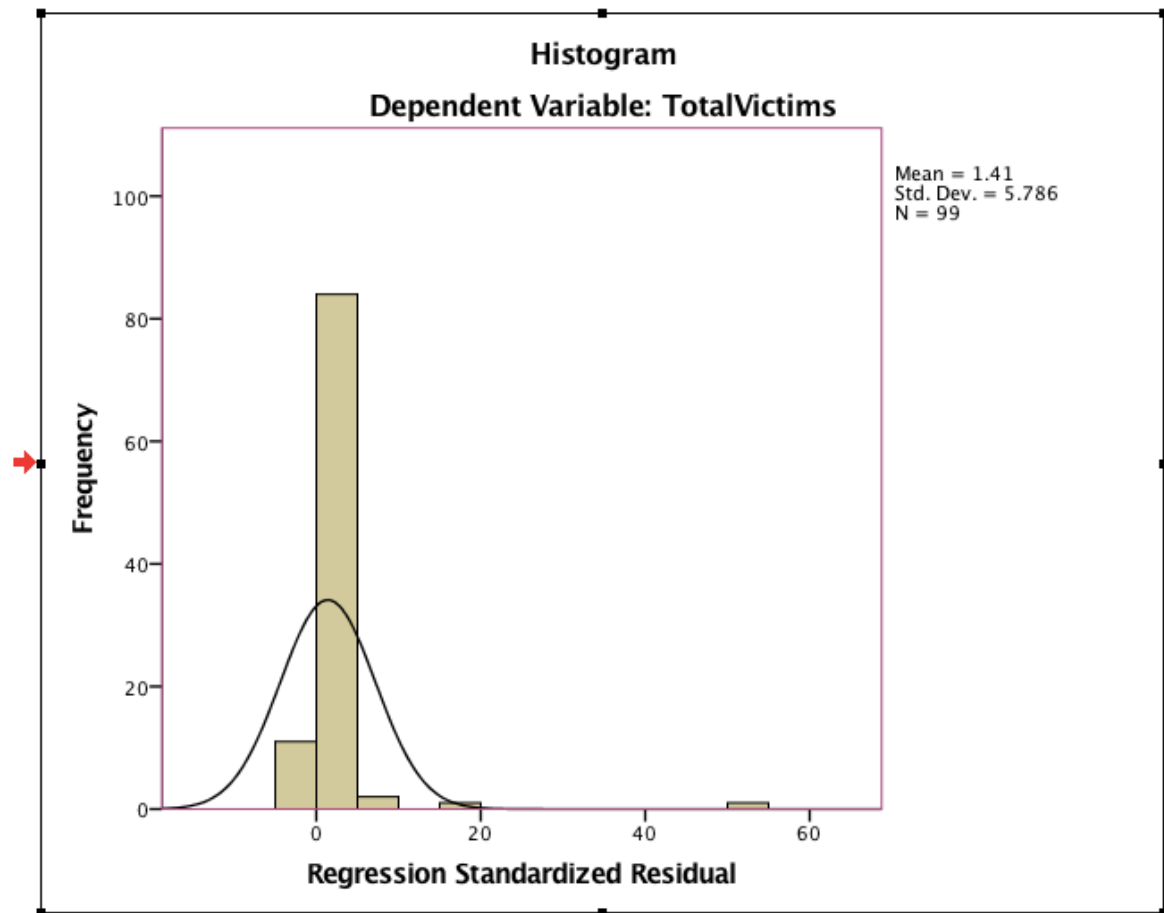
Coefficients ^a								
		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-9.642	4.925		-1.958	.053		
	AccedientsWhileMovingGoods	27.368	2.509	.751	10.910	.000	.788	1.268
	VictimsOnTypeOfAccidents	.208	.144	.099	1.438	.154	.788	1.268
a. Dependent Variable: TotalVictims								

- Following observations have been made by looking at the above table:
- **Standardized coefficient:**
 - 1) In Standardized coefficients the Beta column we can see that the independent variable Accidents while moving goods beta value is 0.751 and Victims on type of accident is 0.99.
 - 2) This shows that independent variable Accident while moving goods makes strongest unique contribution to dependent variable Total Victims to the equation.
 - 3) While the other independent value victim on type of accident which has Beta value 0.99 contributes less to unique contribution.
- **Unstandardized coefficient:**
 - 1) Here in unstandardized coefficient beta value for Accidents while moving goods predictor variable is 27.36 and for Victims on type of accidents is 0.208.
 - 2) It shows that a unit change in Accidents while moving goods predictor variable will reflect a change of 27.36 in the dependent variable Total victims.
 - 3) It shows that a unit change in victim on type of accident predictor variable will reflect a change of 0.208 in the dependent variable Total victims.
- The P-value of the independent variable Accidents while moving goods is 0.000 which is less than 0.05 that means it significantly contributes to the variance in the dependent variable.
- The P-value of the independent variable Victim on type of accidents is 0.154 which is higher than 0.05 that means it contributes less to the variance in the dependent variable.

Residual Analysis:

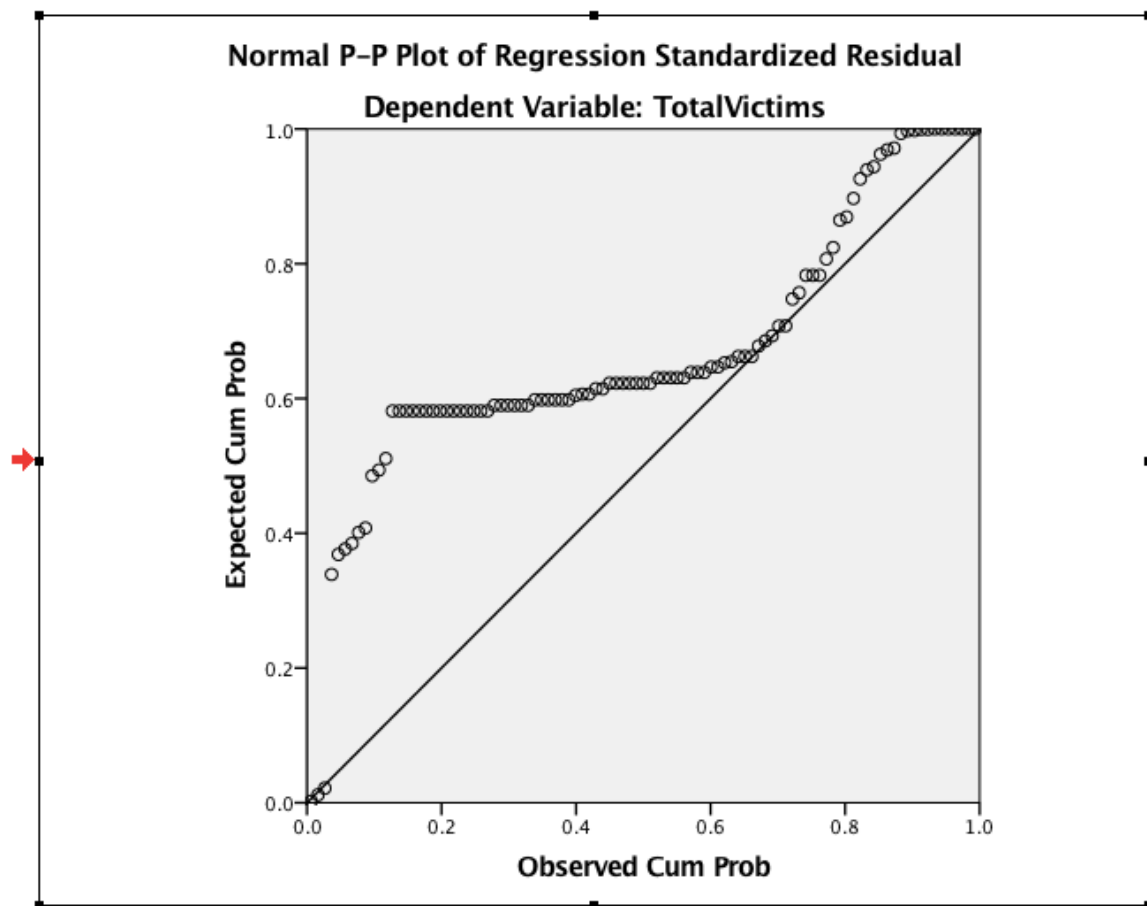
Residuals Statistics ^a					
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-9.64	429.90	8.12	58.290	99
Std. Predicted Value	-.303	6.779	-.017	.939	99
Standard Error of Predicted Value	4.858	38.629	5.809	4.109	99
Adjusted Predicted Value	-5007.06	236.66	-47.54	505.034	99
→ Residual	-132.508	2565.097	65.909	271.156	99
Std. Residual	-2.827	54.735	1.406	5.786	99
Stud. Residual	-3.104	96.675	1.838	9.853	99
Deleted Residual	-159.662	8002.059	121.575	806.369	99
Stud. Deleted Residual	-3.255	8.710	.781	1.687	97
Mahal. Distance	.063	65.596	1.261	6.969	99
Cook's Distance	.000	6603.247	66.777	663.643	99
Centered Leverage Value	.001	.669	.013	.071	99
a. Dependent Variable: TotalVictims					

- The Residual Statistics is a good way to make interpretation about the data using histogram.



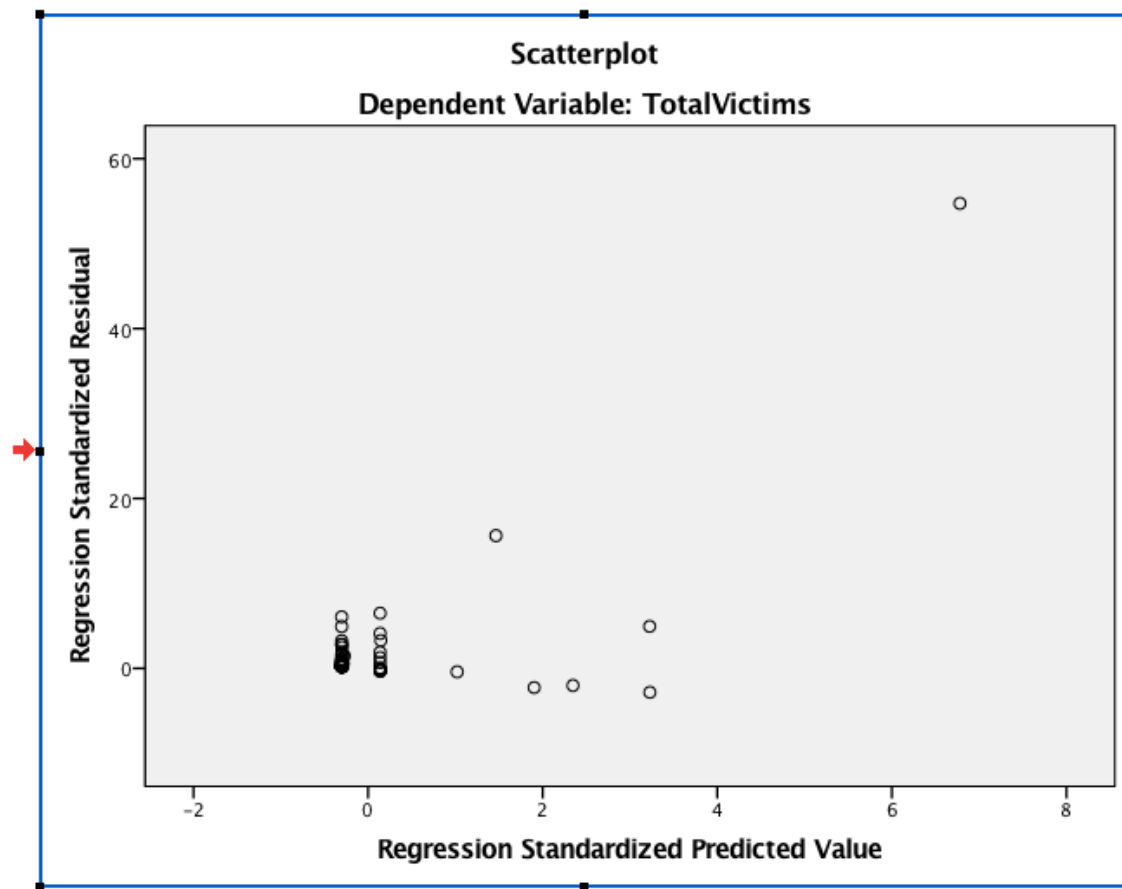
- The histogram shows the normal distribution of data. The values show negative skewed distribution.
- In the above histogram Frequency on Y-axis and Regression standardized residual is on X-axis. The mean is 1.41, standard deviation is 5.786 and N=99.

Normal P-P Plot of Regression Standardized Residual:



- The normal P-P plot indicated that the dotted line will lie in a straight diagonal fashion from bottom left to top right along with the line.
- But, here it is not normally distributed, it has a negative skewed distribution.

Scatter Plot:



- It can be seen there is positive correlation with one outlier.

Conclusion:

- To conclude, a unit change in Accidents while moving goods predictor variable will reflect a change of 27.36 in the dependent variable Total victims and a unit change in victim on type of accident predictor variable will reflect a change of 0.208 in the dependent variable Total victims.
- Accident while moving goods independent variable will produce significant variance in the dependent variable.
- And, the independent variable victim on type of accident cannot be used to predict the values because the P-value is greater than 0.05.

Logistics Regression

In logistic regression the dependent variable is categorical and the predictor (independent) variables can be either categorical or continuous, or mix of both in the same model.

Objective :

Victim(Killed or Injured) is dependent and Type of accidents and person involved are independent variables.

Whether the victim(Killed or Injured) depends on type of accidents and the person involved

Level of measurement:

Variable Name	Variable Type
Victim	Dependent
Type of accidents	Independent
Person involved	Independent

The Categorical Variables encoding:

1)Victim :

Killed - 0

Injured - 1

2) Person involved:

EMP - 1

LV_USR - 2

OTH - 3

PAS - 4

3) Type of accident:

COLLIS - 1

COLLIS_X_LVLCROS - 2

DERAIL – 3

LVLCROS – 4

OTH – 5

RSTK_FIRE – 6

RSTK_MOT – 7


Hypothesis:

H0: The null hypothesis states that the predictor variables affect the outcome of Victim(Killed or Injured)

H1: The alternate hypothesis in this scenario will be the predictor variables do not affect the outcome of the Victim(Killed or Injured).

SPSS Output:

1) Dependent variable:




Original Value	Internal Value
KIL	0
INJ	1

- Encoding of dependent variable Number of victim killed during an accident is 0 and Number of victim injured during an accident is 1.

2) Classification table:

Classification Table^{a,b}



Observed			Predicted		Percentage Correct
			victim KIL	INJ	
Step 0	victim	KIL	1658	0	100.0
		INJ	1450	0	.0
	Overall Percentage				53.3

a. Constant is included in the model.

b. The cut value is .500

- The above model states the model can be taken into account as null hypothesis as it shows the overall ability of the model to predict without involving the predictor variables.
- The ability of the model for prediction without the predictors is calculated as 53.3.

3) Variables in the equation:

The null model $B0 \ln(p/p-1) = -.134$ where p is the probability.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.134	.036	13.899	1	.000	.875

4) Variables not in equation:

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	TypeOfAccidents	17.721	1	.000
		pers_inv	.053	1	.818
	Overall Statistics		17.777	2	.000

- The independent variable Type of Accidents has significance value 0.00 which is less than 0.05, it indicates that it is statistically significant.
- The independent variable pers_inv (person involved) has significance level 0.818 which is greater than 0.05 which in block 1 will be modified to a better significance value.
- Further, test are performed to compare above mentioned null hypothesis and derived to see that better model can be made by using the predictor variables to do the analysis.

Block 1:

5) Omnibus test:

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	17.792	2	.000
	Block	17.792	2	.000
	Model	17.792	2	.000

- Here in omnibus test, we check that whether the block 1 model has been improved or not with respect to block 0 mode.
- The significance level is 0.00 which is less than 0.05 and the chi-square value is 17.792 at 2 degrees of freedom which is resulting in good predictive model for logistic regression analysis.

6) Model Summary:

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	4276.880 ^a	.006	.008
a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.			

- Nagelkerke R square statistics signifies that 8 percent variance in the outcome of the dependent variable is explained by the independent variables.

7) Hosmer and Lemeshow Test:

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	27.319	8	.001

- The hosmer and Lemeshow test tells the predictive capacity of the model.
- A P-value greater than or equal to 0.05 will describe the model good for predicting the dependent variable as whether the victim is killed or injured.
- Here it is less than 0.05 shows significant P-value, which tells that it is not a good fit.

8) Contingency Table:

Contingency Table for Hosmer and Lemeshow Test

		victim = KIL		victim = INJ		Total
		Observed	Expected	Observed	Expected	
Step 1	1	168	161.784	108	114.216	276
	2	183	151.159	79	110.841	262
	3	189	210.928	183	161.072	372
	4	173	172.981	135	135.019	308
	5	165	180.549	165	149.451	330
	6	129	136.754	129	121.246	258
	7	162	167.767	162	156.233	324
	8	162	162.773	162	161.227	324
	9	162	157.778	162	166.222	324
	10	165	155.527	165	174.473	330

- The Hosmer and Lemeshow contingency test table can be used to predict if the victim is killed or injured.
- The 10th Step of above table shows that the observation for victim (killed) is 165 and its expected value is 155.527 which is not close and also for victim (injured) is 165 and its expected value is 174.473 which again is not close enough.
- Therefore, the generated model for prediction cannot be termed as a good significantly good.

9) Classification Table:

Classification Table ^a					
Observed			Predicted		Percentage Correct
			victim		
			KIL	INJ	
Step 1	victim	KIL	1268	390	76.5
		INJ	1060	390	26.9
		Overall Percentage			

a. The cut value is .500

- The classification table can be used to derive the information using the predictors or independent variables for how good is the model is in terms of predicting the dependent variable.
- It shows 53.3 percent which shows that it has slightly crossed the 50% mark, which is 3% higher than the 50% mark. This will be used to predict whether the victim is injured or killed.

10)Variables in Equations.

		Variables in the Equation					
		B	S.E.	Wald	df	Sig.	Exp(B)
→ Step 1 ^a	TypeOfAccidents	.076	.018	17.665	1	.000	1.079
	pers_inv	-.006	.024	.056	1	.813	.994
	Constant	-.404	.100	16.251	1	.000	.667
a. Variable(s) entered on step 1: TypeOfAccidents, pers_inv.							

- The above variables in the equation table shows the real beta coefficient value for type of accidents and person involved independent variable and the odds ratio with respect to the variables.
- If the Exp(B) value for a predictor variable is on a higher side than 1, the variable is likely to predict the outcome variable. It indicates that it is directly proportion to likeliness of predicting the dependent variable.
- In above table, it can be seen that both the independent variables Type of accidents and Person have Exp(B) values as 1.079 and 0.994
- The prediction that the victims is killed or injured can be determined from the type of accident variable as the Exp(B) value is 1.079 which is greater than 1.
- The Exp(B) value for Person involved variable in an accident is 0.994 which is less than 1. Therefore, the dependent variable cannot be predicted from person involved variable.

Conclusion:

On examining all the test for logistic regression model, it can be concluded that the fact or quality of being alike of victim is killed or injured in an accident can be predicted from Type of accident independent variable but not from person involved independent variable.

References:

Pallant, J. (2013). SPSS survival manual. 5th ed. Buckingham: Open University Press.