Proceedings of IOE Graduate Conference, 2017
Volume: 5 ISSN: 2350-8914 (Online), 2350-8906 (Print)

Using Logistic Regression to Estimate the Influence of Crash Factors on Road Crash Severity in Kathmandu Valley

Rajeeb Shakya [a], Anil Marsani [b]

[a, b] *Department of Civil Engineering, Pulchowk Campus, IOE, TU, Nepal*
Corresponding Email: [a] rajeebs4@gmail.com, [b] anilmarsani@ioe.edu.np

Abstract

There are various factors which are related to Road Traffic Crashes (RTCs). In this study, Logistic Regression is used to estimate the severity of factors related to RTCs in Kathmandu Valley. The dependent variable is the Crash Severity (Fatal or Non-Fatal). The independent variables are crash cause, vehicle type, age & sex of the driver at fault, age of the injured personnel, time of crash, collision type, location of the crash and injured type. Data are obtained from Nepal Traffic Police records for the past five years. Because of the binary nature of the dependent variable, logistic regression was found suitable. Of the nine independent variables, three variables were found to be significantly associated with the outcome of the dependent variable namely age of the driver at fault, age of the injured personnel and time of the crash. A statistical interpretation of these significant variables in terms of odds and odd ratio concept is done in the analysis part. Further the association between the time of the crash and the traffic volume is also checked in the analysis. The findings show that logistic regression as used in this thesis is a promising tool in providing meaningful interpretation that can be used for future safety improvement policies in Kathmandu Valley.

Keywords

Logistic regression, Road Traffic Crashes

1. Introduction

The number of vehicles on the road is increasing day by day and alongside this increase, the number of road crashes and deaths is also on the increase. Every year approximately 1.25 million people around the world lose their lives and between 20-50 million more people suffer non-fatal injuries due to road traffic crashes. Unless any action is taken, road traffic injuries are predicted to be the 5th leading cause of death by 2030 (World Health Organization, 2010). In 2002, the overall global road injuries rate was 19 per 100000 people, with 90% of the cases in low and mid-income countries like Nepal. According to the report published in 2002 by World Bank "Cities on the move", nearly 0.5 million people die and 15 million people are injured in urban road traffic accidents in developing countries each year, at a direct economic cost between 1 and 2% of worldwide Gross Domestic Product.

Kathmandu is one of the fast growing cities among the developing countries in South Asia. The population according to the 2011 census is approximately 2.5 million. The population density in the valley is 5140 people per sq. km. With such a high population density, the growth of vehicles is bound to increase day by day. This rapid urbanization is showing its toll on the valley traffic condition with increased congestion, pollution and crashes. Further, the road space in Kathmandu valley is just around 7-8%. With such low road space and such high volume of vehicles, immense care needs to be taken to minimize the number of Road Traffic Crashes (RTCs).

In the year 2068/69, the total number of RTCs was 5096 (148[1] 396[2] 3317[3]). This number counted to 4770 (148[1] 246[2] 3431[3]), 4672 (143[1] 229[2] 3481[3]), 4999 (133[1] 233[2] 3642[3]) in the years 69/70, 70/71, 71/72 that followed. There is no significant improvement in these numbers

[1] Number of deaths

[2] Number of serious injuries

[3] Number of minor injuries

over the past four years.

Logistic Regression is used when the dependent variable is binary in nature. It is a powerful statistical tool which can be used to predict the effect of the factors that are related to the crashes.

2. Literature Review

Many researches have been done around the world using Regression Analysis. Among the different regression analysis methods, the most commonly used is the Conventional Regression Analysis either linear or Non- linear when the dependent variable is continuous in nature. However, logistic regression analysis seems to be a promising tool especially when the dependent variable is binary in nature (i.e. it can take only two values) and the same is used in this paper to predict the severity of the factors that are involved in the RTCs.

In the logistic regression model developed By Ali. S. Al-Ghamdi (2000) to study the accidents in Riyadh, Location and Crash Cause were found to be significantly associated with the outcome of the crash being fatal or non-fatal. The study showed that the odds of being in a fatal crash at a non-intersection location are 2.64 times higher than those at an intersection. The study also showed that the odds that a crash will be fatal because of running a red light (RRL) are 2.72 times higher than for a non-RRL crash.[1]

Mohadeseh Khalili & Alireza Pakgohar (2013) investigated the impact of Road Defects on the severity factor of the road crashes according to the vehicle movability situation after the accident. The results from this research shows that most important factors reducing the safety on the suburban roads in Iran is “Insufficient road width” pertaining to frequency and “Level difference between road & shoulder” pertaining to crash severity. [2]

Wiredu Sampson & Richard Tawiah (2015) developed a logistic model which revealed that the significant predictors of severe crashes were mainly overcrowding, driver discipline on roads, driver fatigue and design & condition of roads with estimated odds ratio of 2.42, 3.83, 10.51 and 12.06 respectively. Other variables including speed, drunk driving, Not using Helmets, mechanical failure, over speeding and indiscriminate use of roads by pedestrians were not found to be significant predictors of severe crashes. [3]

Mahmoud Saffarzadeh et.al (2012) developed a binary logistic regression regression model analysis of forensic medicine data from the Khorasan Razavi province, Iran and found that the pedestrian’s age, vehicle involved in the accident and accident location had a significant impact on the probability of the death at the scene. [4]

Sareh Bahrololoom et.al (2011) carried out the research that studied the factors that increase the likelihood of hit and run in crashes that include at least one cyclists in Victoria, Australia. The results showed that crash time, bicyclist’s age and gender, helmet use (for bicyclists), other road user’s intent, bicyclist’s intent, traffic control (other road user’s approach), traffic control (bicyclist’s approach) and crash severity are significant variables in the Binary Logistic Regression model. [5]

Murat Karacasu et.al (2013) studied the causes of the traffic accidents in Eskisehir, Turkey using logistic regression and discriminant analysis and found that traffic sign, pavement type, vehicle type, purpose, education and primary fault were determined to be significant variables. [6]

Boakye Agyemang et.al (2013) carried out the Regression Analysis of Road Traffic Accidents and Population Growth in Ghana and found that population growth is 72.9% accountable for the changes in accidents in Ghana. Further, the model was used to forecast the Road Traffic Accidents in the near future. [7]

Gentiana Qirjako et.al (2008) studied the factors associated with fatal road crashes in Tirana, Albania and found that Younger Age (OR: 3.97, 95% CI: 2.28-6.91), High Speed (OR: 2.54, 95% CI: 1.62-3.98) and especially Alcohol Consumption (OR: 6.15, 95% CI: 3.54-10.66) were strong and significant predictors of fatal crashes. The study also showed that fatal crashes were more prevalent on Intercity Roads (OR: 4.25, 95% CI: 3.11-5.82) and involved especially Vans and Trucks (OR: 4.12, 95% CI: 2.34-7.24). [8]

S. Renuraj et.al (2014) carried out the logistic regression analysis to study factors influencing Traffic Accidents in Jaffna. The study showed that Type of Vehicle and Age were found to be significant in influencing the accident severity. [9]

3. Theoretical Background of Logistic Regression

Logistic Regression is a very powerful statistical tool when the analysis contains dependent (response) variable that is binary or dichotomous in nature (i.e. it can take only two values). The independent variables can be either continuous or categorical in nature. The response variable thus takes the value 0 or 1. The linear regression equation is in the form of

$$E(Y/x) = \beta_0 + \beta_i \cdot x_i$$

Where, $E(Y/x)$i the expected value of Y given x. β_0 is the value of Y when x equals zero and β_i are the model parameters. The above equation is linear and Y can take any values from 0 to infinity. The likelihood function is the function which when maximized or minimized the predicted values of the dependent variable tends closer to the actual values. In linear regression, the likelihood function is the sum of least square of the difference of the predicted and actual values. It can be written in the following form.

$$LL(x_i) = \Sigma_0^n (Y_{pi} - Y_i)^2$$

Where $Y_(pi)$ are the predicted (modeled) values and Y_i are the actual (observed) values. The values of β_i can be determined by minimizing the Likelihood function. This process of computing the parameters is called the Sum of Least Squares. In logistic regression, however the dependent variable Y can take only two values (0 or 1). To make this possible, the above equation needs to be transformed as mentioned below.

$$\pi(x) = \frac{e^{\beta_0 + \beta_i x_i}}{1 + e^{\beta_0 + \beta_i x_i}}$$

Where $\pi(x)$ is used instead of $E(Y/x)$ in logistic regression to simplify notation. By doing this, the value of dependent variable is limited to take values in-between 0 or 1 including 0 and 1. The above form of $\pi(x)$ can be transformed into linear form which is called Logit transformation as shown below.

$$g(x) = ln\frac{\pi(x_i)}{1 - \pi(x_i} = \beta_0 + \beta_i x_i$$

This transformation is important considering the fact that the right hand side of the equation can now be completely treated in the same way as in linear analysis, the dependent variable now being equal to $g(x)$. The exponent of the term, $g(x)$ also called Odd, is described in the later section (See Development of Logistic Model in the Methodology section). However the Sum of Least Squares cannot be used to predict the parameters in logistic regression, the reason being the fact that the dependent variable is binary in nature and not continuous. A convenient way to express the contribution to the likelihood function for the pair (x_i, y_i) is through the term,

$$\tau(x_i) = \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}$$

Since (x_i) values are assumed to be independent, the product for the terms given in the foregoing equation gives the likelihood function as follows.

$$L(\beta) = \prod_{i=1}^{n} \tau(x_i)$$

It is easier to work mathematically with the log of the equation which gives the likelihood expression,

$$L(\beta) = ln[l(\beta)]$$

$$= \Sigma_{i=1}^n (y_i ln[\pi(x_i)] + (1 - y_i) ln[1 - \pi(x_i)]$$

Maximizing the above function with respect to β and setting the resulting expression to zero will give the values of β_i values.

There is a statistic which is used to check the significance of the variables in the model which is called Deviance. It is the ratio of the likelihood of the current model to the likelihood of the saturated model multiplied by minus 2. Saturated model is the one which contains as many parameters as there are data points and the current model is the one that contains only the variable under question.

$$D = -2ln\frac{Likelihood\,of\,the\,current\,model}{Likelihood\,of\,the\,saturated\,model}$$

For the purpose of assessing the significance of an independent variable, the value of D should be compared with and without the independent variable in the model. It can be obtained as follows

$$G = D(for\,the\,model\,without\,the\,variable) -$$

$$D(for\,the\,model\,with\,the\,variable)$$

The G-values will follow the Chi-square (χ^2) distribution with one degree of freedom. The critical values of the Chi-Square distribution can be easily obtained from the statistic tables (Not mentioned in this paper). If the change in the Chi-Square values is greater than the critical value, then the change is significant and the variable under question is significant and if it is below the critical value, then the change is just random and the variable under question is insignificant.

Another important statistic called P-value is the region outside the confidence interval of the normal distribution of the predicted value of the coefficient of the independent variable. For 95% confidence interval, the p-value should be less than 0.05.

4. Methodology

4.1 Data Description

The data for the regression analysis of this model is collected from the traffic police records from the year 2068/69 to 2072/73. Only the serious crashes occurring inside the Kathmandu valley are used. Minor crashes are not used in this analysis. Serious crashes are to be filtered manually from the thousands of crash records that are recorded in the traffic police records. A Software called SPSS-21 (Statistical Package for Social Sciences) is used to carry out the logistic regression analysis for our purpose of study.

The dependent variable is “Crash Severity” which is coded as 0 if it results in at least one injury and no fatality and 1 if it results at least one fatality.

There are 8 independent variables in this analysis. All of the independent variables are categorical in nature i.e. they cannot be put in order in terms of their magnitude except two variables which are “Age of the driver at fault” and “Age of the injured personnel”. The categorical variables need to be coded in a different way which is defined in the section below. The list of variables is summarized in the Table 1 given below.

The categorical variables need to be interpreted in a different way from the continuous variables. The two continuous variables “Age of the driver at fault” and “Age of the injured personnel” are measured in terms of number of years. A collection of design variables (Also called Dummy variables) are needed to define the categorical variables. One way of coding the dummy variables is to have (k-1) design variables for k levels of nominal scale of that categorical variable. An example for this coding is given in the Table 2 below. The independent variable “Injured Type” is taken as an example. As can be seen in Table 2, it has 6 categories namely “Bike Driver”, “Pedestrian”, “Bike Passenger “, “Other Vehicle Passenger”, “Multiple Injuries” and “4W Driver”. Thus it needs 5 dummy variables D_1, D_2, D_3, D_4 and D_5. One variable is set as a base variable and all the other dummy variables are calculated relative to this base variable. Here for example if the crash involves “Bike Driver” (which is a base variable) as the affected party, then all 5 dummy variables D_1, D_2, D_3, D_4 and D_5 are set equal to zero. If the crash involves Pedestrian as the affected party, then D_1 is set to 1 and D_2, D_3, D_4 and D_5 are set to zero.

And, if the affected party includes “Bike Passenger”, then D_2 is set equal to 1, and D_1, D_3, D_4 and D_5 set to zero. Similarly if the affected party includes “Other Vehicle Passenger”, then D_3 is set equal to 1, and $D1$, D_2, D_4 and D_5 set to zero. If there are multiple injuries, then D_4 is set equal to 1 and D_1, D_2, D_3 and D_5 are set to zero. And finally if the affected party is “4W Driver”, then D_5 is set equal to 1 and D_1, D_2, D_3 and D_4 are set equal to zero. The coding technique is same for all the remaining categorical variables. However these codings are automatically done by the software (SPSS).

4.2 Development of Logistic Model

There are many ways of developing a Logistic model. The mostly used techniques are Forward Selection process and Backward Selection Process. In the Forward Selection process, first the regression analysis is carried out with only one independent variable and its significance is checked at 95%. However due to the possibility of omission of the significant variables from the complete model, the critical significance level is kept at 75% (P-value 25%) initially (Hosmer & Lemeshow, 2000)[10]. And then the remaining variables are added to the model one by one until we are left with only significant variables at 95% confidence level.

The backward selection process is carried out with all the independent variables in the model with no interaction between the variables also called as the saturated model.

Table 1: List of variables

No.	Description	Coded Values	Abbreviation
1	Crash Severity	0 = Non fatal	CRA_SEV
		1 = Fatal	
2	Crash Cause	1 = Alcohol Consumtion	CRA_CAUSE
		2 = Negligence of the driver	
		3 = Overtaking	
		4 = Overspeeding	
		5 = Mechanical Failure	
		6 = Road Condition	
3	Collision Type	1 = Head on	COLL_TYPE
		2 = Right Angled	
		3 = Side Swipe	
		4 = Rear End	
		5 = Out of control	
		6 = Pedestrian Hit	
		7 = Collision with Fixed Objects	
4	Vehicle Type	1 = Bus	VEH_TYPE
		2 = Bike	
		3 = Car	
		4 = Minibus	
		5 = Truck	
		6 = Cycle	
5	Age of the driver at fault	years	AGE_FAULT
6	Age of the Injured	years	AGE_INJ
7	Gender	0 = Male	GENDER
		1 = Female	
8	Location	1 = Intersection	LOC
		2 = Turning	
		3 = Straight Section	
9	Injured type	1 = Bike Driver	INJ_TYPE
		2 = Pedestrian	
		3= Bike Passenger	
		4 = Other Vehicle Passenger1	
		5 = Multiple Injuries	
		6 = 4W Driver	

No.	Description	Coded Values	Abbreviation
10	Time of Crash	1 = Morning Off Peak (4-8)	CRA_TIME
		2 = Morning Peak (8-12)	
		3 = Day off Peak (12-16)	
		4 = Day Peak (16-20)	
		5 = Evening Off Peak (20-24)	
		6 = Night (24-4)	

Table 2: Coding of Categorical Variables

INJ_TYPE	Design Variables				
	D1	D2	D3	D4	D5
Driver	0	0	0	0	0
Pedestrian	1	0	0	0	0
Bike Passenger	0	1	0	0	0
Other vehicle Passenger	0	0	1	0	0
Multiple Injuries	0	0	0	1	0
4W Driver	0	0	0	0	1

Their P-values are checked. Those with P-values greater than 25% are rejected at the initial stage until the model is left with only significant variables at 95% confidence level. The change in deviance i.e. the G-value is also observed to interpret the significance of the variable.

In our model, the backward selection process is used.

The logistic model will be in the following form.

$$ln\frac{P(fatal)}{1-P(fatal))} = \beta_0+\beta_{1i}*(CRA_{CAUSE})_i$$

$$+\beta_{2i}*(COLL_{TYPE})_i+\beta_{3i}*(VEH_{TYPE})_i+\beta_4*(AGE_{FAULT})$$

$$+\beta_5*(AGE_{INJ})+\beta_{6i}*(GENDER)_i+\beta_{7i}*(LOC)_i$$

$$+\beta_{8i}*(INJ_{TYPE})_i+\beta_{9i}*(CRA_{TIME})_i$$

However one should keep it mind that all the independent variables may not be significant and the final model will contain only those variables that are significant.

The term $\frac{P(fatal)}{1-P(fatal)}$ is called the odd. As can be seen from the earlier equation $ln\frac{\pi(x_i)}{1-\pi(x_i} = \beta_0+\beta_i x_i$, this equation can also be written in the following form.

$$\frac{P}{1-P} = e^{\beta_{x_i}+\beta_0}$$

The odd ratio can be defined as the ratio of the odds for every unit increase in the value of the independent variable.

Odd Ratio, $OR(a+1,a) = \frac{P/((1-P))_{x=a+1}}{P/(1-P))_{x=a}} = e^{B_i}$

Hence, the Odd Ratio is the exponent of the coefficient of the independent variable under consideration.

The odd ratio is a useful tool in the model interpretation process.

If the odd ratio is greater than 1, then the odd of the success (fatal in this case) for certain value of the independent variable under consideration is greater than the odd of the success for the unit increase in the value of that independent variable.

If the odd ratio is less than 1, then the odd of the success (fatal in this case) for certain value of the independent variable under consideration is less than the odd of the success for the unit increase in the value of that independent variable.

4.3 Entry of Data in SPSS

There are a total of 12 columns in the SPSS entry sheet. The 1st column records the date of the crash. The 2nd column records whether the crash recorded is fatal or non-fatal. “1” means the case is fatal and “0” means that it is non-fatal. This is the dependent variable. The independent variable starts from the 3rd column and ends on the 12th column. The 3rd column records the cause of the crash. There are 6 causes that can be recorded as shown in Table 1. The 4th column records the collision type. There are 7 types of collision as shown on Table 1.

The 5th column records the vehicle type at fault. There are 6 vehicle types as shown in Table 1. The 6th and 7th column records the age of the driver at fault and the age of the injured personnel. The 8th column records whether the driver at fault is male or female "0" for male and "1" for female. The 9th column records the location of the crash. There can be 3 categories of the location of the crash as shown in the Table 1. The 10th column records the injured type. There can be 6 categories of the injured type as shown in the Table 1. The 11th column records the time of the crash. It has 6 categories as shown in the Table 1. The last and the 12th column is for the remarks if any.

4.4 Backward Selection Method

In this case, the total number of crash cases studied to develop a logit model were obtained from the fiscal year 2068/69, 2069/70 and 2070/71. There were a total of 504 number of serious crash cases. Out of these 504 accident cases, only 476 cases were valid. The 28 invalid cases were due to the unknown ages of either the driver at fault or of that person injured. These data are recorded as "999" during the entry process.

The first run of the analysis is done in the SPSS (Statistical Packages for Social Science) to test the significance of the independent variable at 95%.

The four independent variables "Crash Cause", "Vehicle Type", "Gender of the driver at fault" and "Location of the Crash" were not significant at 95% so these four variables are eliminated and the selection process is carried out with the remaining variables. The variables "Age of the Injured" and "Crash time" are significant at 95% as can be seen in figure 1.

The second run of analysis is done as can be seen in figure 2 The variable "Age of the Injured", "Age of the driver at fault" and "Time of crash" appeared to be significant. The variable "Injured Type" appeared to be insignificant and removed for the next selection process.

In the third run of analysis, the variables "Age of the Injured", "Age of the driver at fault" and "Time of crash" still appeared to be significant as can be seen in figure 3 The variable "Collision Type" is removed in this step.

In the fourth and final run of the analysis in figure 4, all the remaining three variables "Age of the Injured", "Age of the driver at fault" and "Time of crash" appeared to be significant at 95% confidence Interval.

And thus the final logistic model of our case is in the following form:

$$ln\frac{p}{1-p} = -0.832 + 0.025 * Age_{fault} + 0.021 * Age_{Injured}$$
$$-1.425 * Crash_{TIME(4-8)} - 1.189 * Crash_{TIME(8-12)} - 1.057$$
$$* Crash_{TIME(12-16)} - 1.076 * Crash_{TIME(16-20)} - 1.344$$
$$* Crash_{TIME(20-24)}) - 0 * Crash_{TIME(24-4)}$$

The value of "p" in the above equation gives the probability that the crash will be fatal.

5. Analysis of the Results

5.1 Validation of the model

The final model needs to be validated against the data that were not used to develop the model. Hence for the validation process, the data from the fiscal year 2071/72 and 2072/73 were taken.

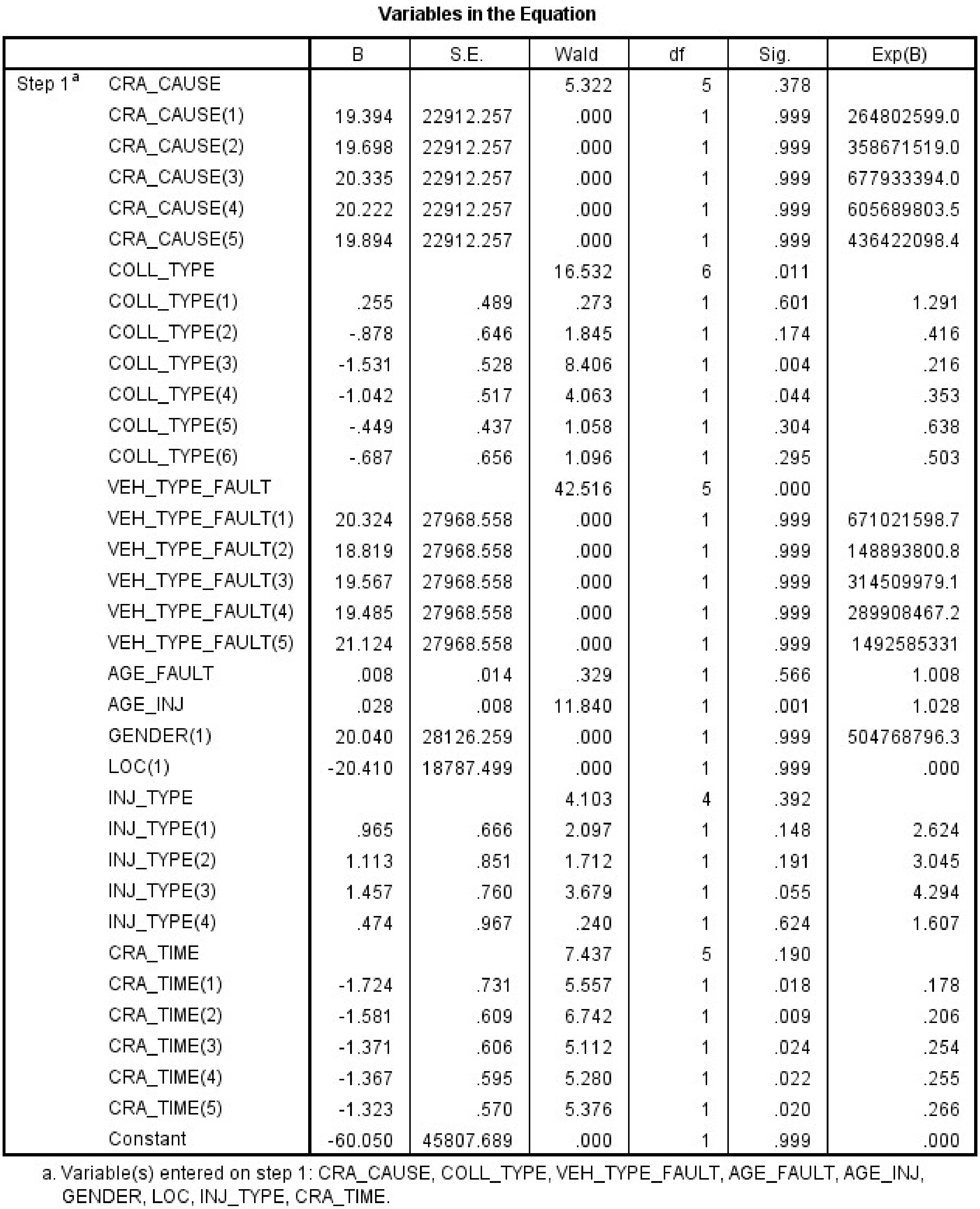
The value of "p" in the above equation determines the probability of the crash being fatal. A cut point of 0.31 is used to separate the fatal crashes from the non-fatal. The reason behind using this value of cut point is because it maximizes the accuracy of this model. This means that any crash case with value of "p" below 0.31 is termed as a non-fatal case and any crash case with value of "p" above 0.31 is termed as a fatal case.

A total of 304 cases were used out of which 217 cases were successful and 87 cases failed which means that the accuracy of the model is 71.38 %.

5.2 Model Interpretation

The exponent of the co-efficient of the variables gives the odd ratio as defined earlier in the "Development of the Logistic Model" Section.

As can be seen in the final result in figure 4, the exponent of the variable "Age of the driver at fault" is 1.025. Similarly that of the variable "Age of the injured personnel" is 1.021. Similarly the exponent of the coefficient of the categories of the variable "Time of crash" is respectively 0.241, 0.305, 0.348, 0.341 and 0.261 respectively.

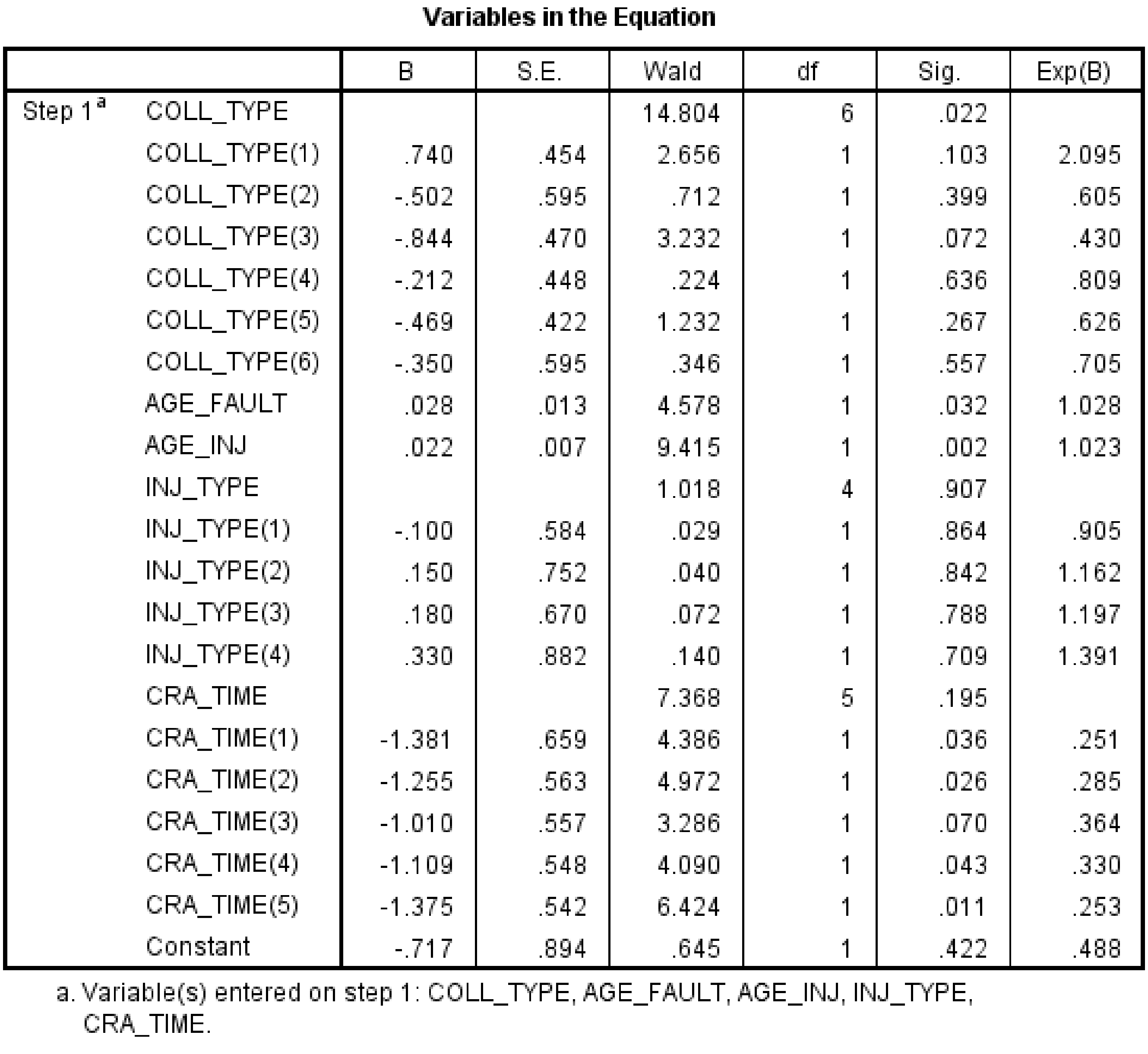


Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 [a]	CRA_CAUSE			5.322	5	.378	
	CRA_CAUSE(1)	19.394	22912.257	.000	1	.999	264802599.0
	CRA_CAUSE(2)	19.698	22912.257	.000	1	.999	358671519.0
	CRA_CAUSE(3)	20.335	22912.257	.000	1	.999	677933394.0
	CRA_CAUSE(4)	20.222	22912.257	.000	1	.999	605689803.5
	CRA_CAUSE(5)	19.894	22912.257	.000	1	.999	436422098.4
	COLL_TYPE			16.532	6	.011	
	COLL_TYPE(1)	.255	.489	.273	1	.601	1.291
	COLL_TYPE(2)	-.878	.646	1.845	1	.174	.416
	COLL_TYPE(3)	-1.531	.528	8.406	1	.004	.216
	COLL_TYPE(4)	-1.042	.517	4.063	1	.044	.353
	COLL_TYPE(5)	-.449	.437	1.058	1	.304	.638
	COLL_TYPE(6)	-.687	.656	1.096	1	.295	.503
	VEH_TYPE_FAULT			42.516	5	.000	
	VEH_TYPE_FAULT(1)	20.324	27968.558	.000	1	.999	671021598.7
	VEH_TYPE_FAULT(2)	18.819	27968.558	.000	1	.999	148893800.8
	VEH_TYPE_FAULT(3)	19.567	27968.558	.000	1	.999	314509979.1
	VEH_TYPE_FAULT(4)	19.485	27968.558	.000	1	.999	289908467.2
	VEH_TYPE_FAULT(5)	21.124	27968.558	.000	1	.999	1492585331
	AGE_FAULT	.008	.014	.329	1	.566	1.008
	AGE_INJ	.028	.008	11.840	1	.001	1.028
	GENDER(1)	20.040	28126.259	.000	1	.999	504768796.3
	LOC(1)	-20.410	18787.499	.000	1	.999	.000
	INJ_TYPE			4.103	4	.392	
	INJ_TYPE(1)	.965	.666	2.097	1	.148	2.624
	INJ_TYPE(2)	1.113	.851	1.712	1	.191	3.045
	INJ_TYPE(3)	1.457	.760	3.679	1	.055	4.294
	INJ_TYPE(4)	.474	.967	.240	1	.624	1.607
	CRA_TIME			7.437	5	.190	
	CRA_TIME(1)	-1.724	.731	5.557	1	.018	.178
	CRA_TIME(2)	-1.581	.609	6.742	1	.009	.206
	CRA_TIME(3)	-1.371	.606	5.112	1	.024	.254
	CRA_TIME(4)	-1.367	.595	5.280	1	.022	.255
	CRA_TIME(5)	-1.323	.570	5.376	1	.020	.266
	Constant	-60.050	45807.689	.000	1	.999	.000

a. Variable(s) entered on step 1: CRA_CAUSE, COLL_TYPE, VEH_TYPE_FAULT, AGE_FAULT, AGE_INJ, GENDER, LOC, INJ_TYPE, CRA_TIME.

Figure 1: Figure of the table from SPSS showing first run of the analysis

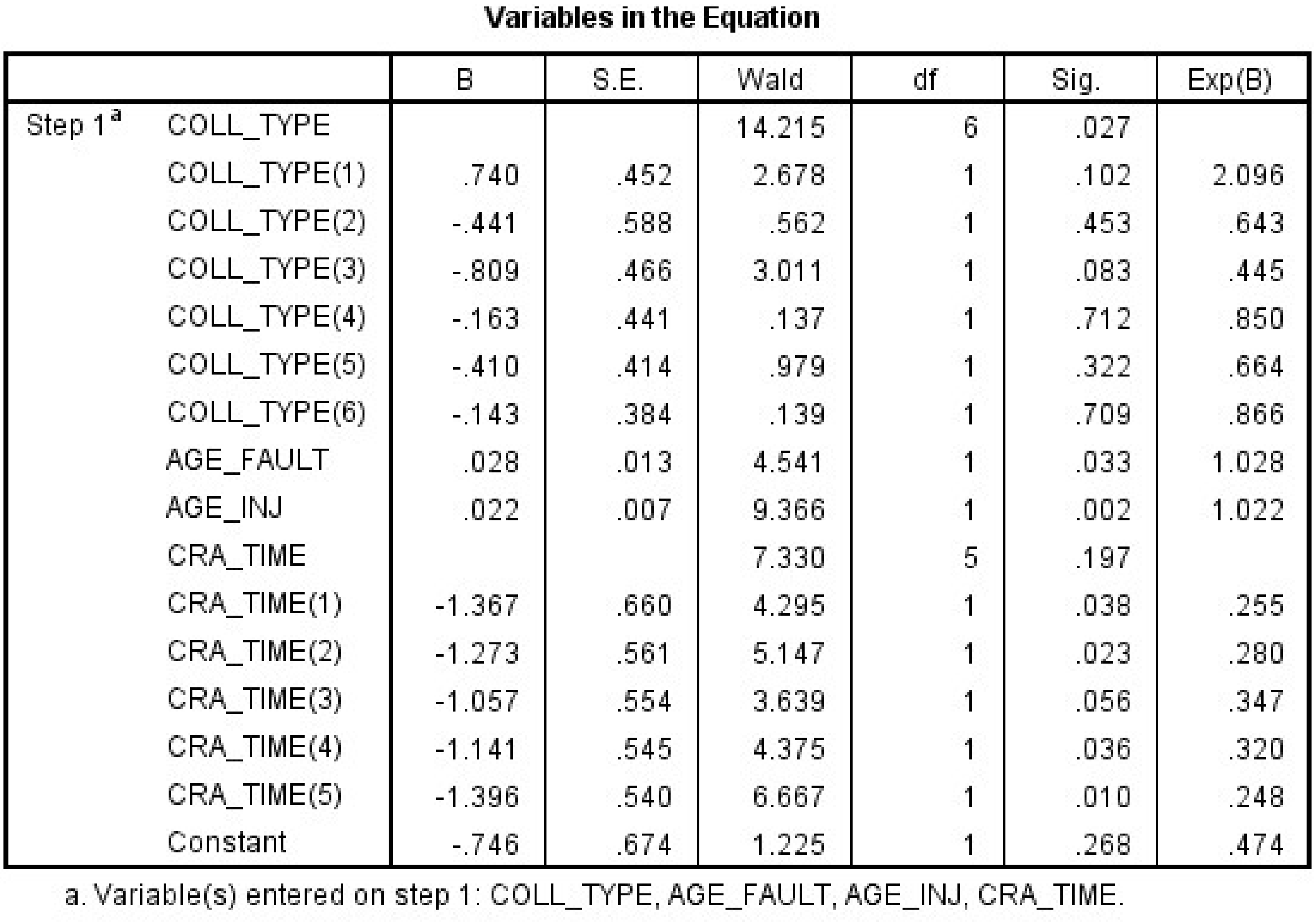


Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	COLL_TYPE			14.804	6	.022	
	COLL_TYPE(1)	.740	.454	2.656	1	.103	2.095
	COLL_TYPE(2)	-.502	.595	.712	1	.399	.605
	COLL_TYPE(3)	-.844	.470	3.232	1	.072	.430
	COLL_TYPE(4)	-.212	.448	.224	1	.636	.809
	COLL_TYPE(5)	-.469	.422	1.232	1	.267	.626
	COLL_TYPE(6)	-.350	.595	.346	1	.557	.705
	AGE_FAULT	.028	.013	4.578	1	.032	1.028
	AGE_INJ	.022	.007	9.415	1	.002	1.023
	INJ_TYPE			1.018	4	.907	
	INJ_TYPE(1)	-.100	.584	.029	1	.864	.905
	INJ_TYPE(2)	.150	.752	.040	1	.842	1.162
	INJ_TYPE(3)	.180	.670	.072	1	.788	1.197
	INJ_TYPE(4)	.330	.882	.140	1	.709	1.391
	CRA_TIME			7.368	5	.195	
	CRA_TIME(1)	-1.381	.659	4.386	1	.036	.251
	CRA_TIME(2)	-1.255	.563	4.972	1	.026	.285
	CRA_TIME(3)	-1.010	.557	3.286	1	.070	.364
	CRA_TIME(4)	-1.109	.548	4.090	1	.043	.330
	CRA_TIME(5)	-1.375	.542	6.424	1	.011	.253
	Constant	-.717	.894	.645	1	.422	.488

a. Variable(s) entered on step 1: COLL_TYPE, AGE_FAULT, AGE_INJ, INJ_TYPE, CRA_TIME.

Figure 2: Figure of the table from SPSS showing second run of the analysis

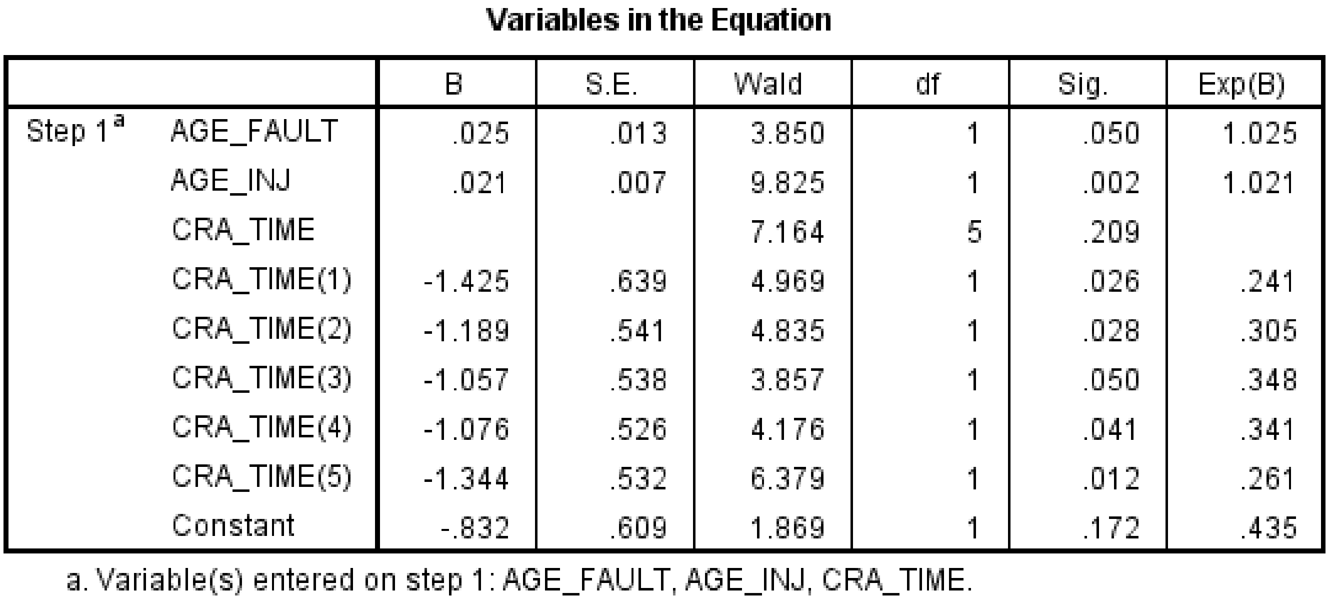


Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	COLL_TYPE			14.215	6	.027	
	COLL_TYPE(1)	.740	.452	2.678	1	.102	2.096
	COLL_TYPE(2)	-.441	.588	.562	1	.453	.643
	COLL_TYPE(3)	-.809	.466	3.011	1	.083	.445
	COLL_TYPE(4)	-.163	.441	.137	1	.712	.850
	COLL_TYPE(5)	-.410	.414	.979	1	.322	.664
	COLL_TYPE(6)	-.143	.384	.139	1	.709	.866
	AGE_FAULT	.028	.013	4.541	1	.033	1.028
	AGE_INJ	.022	.007	9.366	1	.002	1.022
	CRA_TIME			7.330	5	.197	
	CRA_TIME(1)	-1.367	.660	4.295	1	.038	.255
	CRA_TIME(2)	-1.273	.561	5.147	1	.023	.280
	CRA_TIME(3)	-1.057	.554	3.639	1	.056	.347
	CRA_TIME(4)	-1.141	.545	4.375	1	.036	.320
	CRA_TIME(5)	-1.396	.540	6.667	1	.010	.248
	Constant	-.746	.674	1.225	1	.268	.474

a. Variable(s) entered on step 1: COLL_TYPE, AGE_FAULT, AGE_INJ, CRA_TIME.

Figure 3: Figure of the table from SPSS showing third run of the analysis



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE_FAULT	.025	.013	3.850	1	.050	1.025
	AGE_INJ	.021	.007	9.825	1	.002	1.021
	CRA_TIME			7.164	5	.209	
	CRA_TIME(1)	-1.425	.639	4.969	1	.026	.241
	CRA_TIME(2)	-1.189	.541	4.835	1	.028	.305
	CRA_TIME(3)	-1.057	.538	3.857	1	.050	.348
	CRA_TIME(4)	-1.076	.526	4.176	1	.041	.341
	CRA_TIME(5)	-1.344	.532	6.379	1	.012	.261
	Constant	-.832	.609	1.869	1	.172	.435

a. Variable(s) entered on step 1: AGE_FAULT, AGE_INJ, CRA_TIME.

Figure 4: Figure of the table from SPSS showing fourth and final run of the analysis

5.2.1 Age Effect

Effect of the Age of the driver at fault

The exponent of the coefficient of the variable "Age of the driver at fault" is 1.025 which means that for a given "Age of the injured personnel" and given "Time of crash", the ratio of the odds of the crash being fatal for a unit increase in the "Age of the driver at fault" is 1.025. This means that for every unit increase in the age of the driver at fault, the odd of the crash being fatal increases by 2.5% for a given "Age of the injured personnel" and given "Time of crash".

Effect of the Age of the Injured Personnel

Similarly, the exponent of the coefficient of the variable "Age of the injured personnel" is 1.021 which means that for a given "Age of the driver at fault" and given "Time of crash", the ratio of the odds of the crash being fatal for a unit increase in the "Age of the injured personnel" is 1.021. This means that for every unit increase in the age of the injured personnel, the odd of the crash being fatal increases by 2.1% for a given "Age of the driver at fault" and given "Time of crash".

5.2.2 Effect of the time of Crash

Part-I

Since this is the categorical variable and it has 6 categories, one category is used as a base category. And the other 5 dummy categories are stated relative to this base category as defined in the "Methodology" section above.

In this case, the last category i.e. "Night (24-4)" is used as the base model. All other categories are represented relative to this category for the given "Age of the driver at fault" and "Age of the injured personnel".

a) The exponent of the coefficient of the dummy variable $CRA_{TIME(1)}$ i.e. Morning off peak (4-8) is 0.241 which means that the odd of a crash being fatal in the time "Morning off peak (4-8)" is 75.9% less than that of the crash in the time "Night (24-4)".

b) The exponent of the coefficient of the dummy variable $CRA_{TIME(2)}$ i.e. Morning Peak (8-12) is 0.305 which means that the odd of a crash being fatal in the time "Morning Peak (8-12)" is 69.5% less than that of the crash in the time "Night (24-4)".

c) The exponent of the coefficient of the dummy variable $CRA_{TIME(3)}$ i.e. Day Off Peak (12-16) is 0.348 which means that the odd of a crash being fatal in the time "Day Off Peak (12-16)" is 65.3% less than that of the crash in the time "Night (24-4)".

d) The exponent of the coefficient of the dummy variable $CRA_{TIME(4)}$ i.e. Day Peak (16-20) is 0.341 which means that the odd of a crash being fatal in the time "Day Peak (16-20)" is 65.9% less than that of the crash in the time "Night (24-4)".

e) The exponent of the coefficient of the dummy variable $CRA_{TIME(5)}$ i.e. Evening Off Peak (16-20) is 0.261 which means that the odd of a crash being fatal in the time "Evening Off Peak (16-20)" is 73.9% less than that of the crash in the time "Night (24-4)".

Part II

Associating "Time of the crash" with the "Traffic Volume"

From our results, the time of the crash is significantly associated with the outcome of the crash severity being fatal or non-fatal. This analysis is further carried out to check whether there is any significant association between the time of the crash and the traffic volume data.

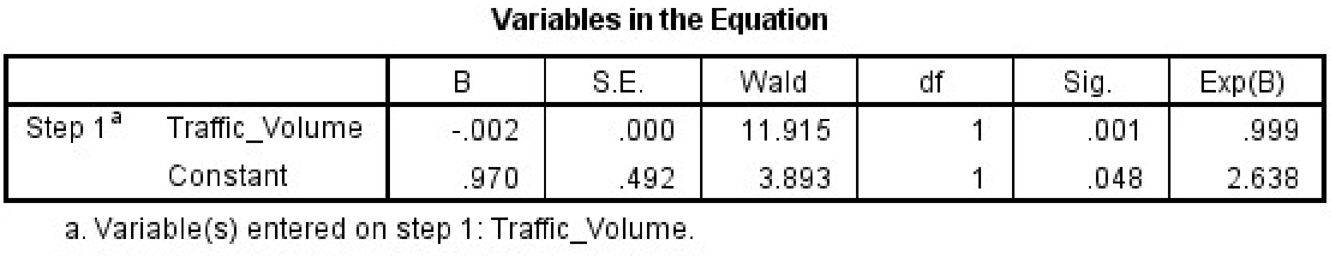
There are altogether 160 stations set across the major corridor roads by Departments of Roads all over the country. Among them, 24 stations lie inside Kathmandu valley. It provides the traffic volume along with the vehicle types at one hour interval during the whole time of the day. The traffic volume data for the fiscal year 2011/12, 2012/13, 2014/15, 2015/16 and 2016/17 can be obtained from the online site of the Department of Roads (http: //ssrn.aviyaan.com/traffic_ controller/get _ summary). Each station gives the traffic volume data for the consecutive three days of a particular fiscal year. Hence the traffic volume needs to be the average of these three days. Also to mention, that there is no traffic volume data for the fiscal year 2013/14 (2070/71) and hence the crash cases of the year 2070/71 cannot be analyzed. The traffic volume for our analysis needs to be interpolated to get the traffic volume at a specific time of the day. Linear interpolation is used in this case.

The details of the 24 stations can be seen in the table 3.

The 504 crash cases that were used to develop the

Table 3: 24 monitored stations inside Kathmandu Valley

S No	*Station No*	*Location*	*Link Name*	*Number of Crashes*
1	58	Satdobato South (Chapagaun)	Satdobato-Sunakothi	4
2	59	Satdobato Junction South	Satdobato-Karmanus Bridge	1
3	60	Ring Road (Manohara Bridge)	Gwarko-Manohara River(Balkumari)	4
4	61	Ring Road (Balkhu East)	Balkhu - Ekantakuna (KTM Ringroad)	2
5	62	Kharipati	Bhaktapur-Army camp	
6	63	Hanumante Bridge	Sallaghari-Hanumante Culvert	7
7	64	Manohara Bridge	Koteshwar-Manohara bridge	1
8	65	Ring Road (Sinamangal)	Tinkune - Sinamangal - Gaushala	8
9	66	Chabahil East	Pipal Bot-Sankhu	
10	67	Jorpati North	Jorpati-Sundarijal	
11	68	Ring Road (Narayan Gopal Chowk)	Sankhapark - Maharajganj (KTM Ringroad)	
12	69	Gangalal Hospital North	Maharajgunj-Bansbari	3
13	70	Balaju Bypass North	Balaju bypas-Nagarjun	
14	71	Ring Road (Banasthali)	Balaju Junction - Banasthali - Swoyambhu	13
15	72	T.U. Gate	Balkhu-Chovar	7
16	73	Taudaha	Chovar-Chhaimale	
17	74	Nagdhunga	Peepalmod-Nagdhunga	
18	154	Narayan Gopal Chowk West	Maharajganj - Balaju Bypass Junction	8
19	155	Narayan Gopal Chowk South	Lainchaur-Maharajgunj	5
20	156	Sitapaila South	Kalimati-Bahiti	1
21	157	Kalanki	Kalanki to Balkhu	3
22	158	Gwarko East	Gwarko-Lubhu-Lankuri Bhanjyang	1
23	159	Byasi Chowk North	Byasi (Bhaktapur)-Changunarayan	
24	160	Satdobato North	Satdobato- Gwarko (KTM Ringroad)	4
			Total	**72**



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Traffic_Volume	-.002	.000	11.915	1	.001	.999
	Constant	.970	.492	3.893	1	.048	2.638

a. Variable(s) entered on step 1: Traffic_Volume.

Figure 5: Figure of the table from SPSS showing the final run of the analysis of part II of 5. Analysis section

model occurred all over the Kathmandu Valley. The traffic volume data of all the places in the Kathmandu Valley cannot be obtained. Hence only those crashes that occurred along the corridor monitored by the 24 stations provided in table 3 can be analyzed. As can be seen in table 3, there are only 72 accident cases within the original 504 cases that occurred along the monitored corridor.

The analysis of the 72 crash cases in the SPSS includes Crash Severity as the dependent variable which is either fatal or non-fatal. Only the significant independent variables from the initial model are taken for this analysis. They include “Age of the driver at fault”, “Age of the injured personnel”, “Time of crash” and newly added “Traffic volume”. The traffic volume data between two different time needs to be interpolated to get the traffic volume at the time of the crash.

The analysis is done in SPSS in the same manner as previously. Only one variable ”Traffic volume” was found to be significant. The reason behind the insignificance of the previously significantly associated independent variables (“Age of the driver at fault”, “Age of the injured personnel” and “Time of Crash”) is because of the reduced number of crash cases from 504 to 72. The final result can be seen in figure 5 of the table from SPSS.

The exponent of the coefficient of the independent variable ‘Traffic volume” is 0.998 (0.999 in the table but 0.998 as calculated in excel) which means that the odd of the crash to be fatal decreases by 0.2% for every increase in the traffic volume by 1 PCU/hour. If the traffic volume is increased by 100 PCU/hour, the odd of the crash to be fatal decreases by 18.27%. If expressed in the form of the equation, it can be stated as below:

$$ln\frac{p}{1-p} = 0.970 - 0.002 * Traffic_{Volume}$$

However it should be kept in mind that this equation is applicable only for the crashes occurring on the given 24 station-monitored corridors. To be more specific, the 72 crashes occurred in 16 monitored corridors out of the 24 corridors as can be seen in Table 3. Hence this model will apply effectively in these 16 corridors.

5.2.3 Interpretation on insignificant variables

The independent variables like “Vehicle type”, “Crash cause”, “Gender of the driver at fault” and “Location of the crash” which were eliminated in the first run of the analysis were far from being significant.

The first reason behind the independent variable “Vehicle type” not being significant may be due to the fact that the nature of the traffic in Kathmandu valley is homogeneous with all kinds of vehicle occupying the same right of way. With this homogenous nature of traffic, all kinds of vehicles are involved in both fatal and non-fatal crashes without any kind of significant relation between vehicle type at fault and the severity of the crash. The second reason may be due to the under reporting from the traffic authority. The actual size of the road crash problem may be greater than that shown by the official crash data recorded by the traffic authority.

“Crash cause” not being significant may be due to the fact that 319 crash cases out of the total 476 crash cases were accounted for the negligence of the driver. This implies that the 67% of the crashes were caused due to negligence of the driver. Hence there seems to be lack of interest shown by the traffic authority to write down all the details which led to the crash.

Only 2 crash cases involved female out of the 476 crash cases. This tiny number of crash cases involving female is the very reason behind the independent variable “Gender of the driver at fault” not being significant.

Similarly for the independent variable “Location of the crash”, it was difficult to pin point the exact location of the crash with only the descriptive nature of the crash data being available. The data also did not include any kind of sketch of the location.

6. Conclusion

The result of the model interpretation shows that

- Both the increasing ages of the driver at fault and that of the injured personnel is contributing towards the increasing number of the fatal crashes in Kathmandu Valley.
- The chances of fatal crashes in Kathmandu Valley are the highest in the Night time (24-4) followed by the “Day Off Peak (12-16)”, “Day Peak

(16-20)", "Morning Peak (8-12)", "Evening Off Peak (16-20)" and "Morning off peak (4-8)" respectively.

- For the crashes occurring along the 16 monitored corridors inside Kathmandu valley as given in table 3, the increase in the traffic volume results in the decreasing number of fatal crashes. The increase in the traffic volume results in the decrease in the speed of the vehicles ultimately causing the reduction in the number of fatal crashes.

Acknowledgments

The authors are grateful to Pulchowk Engineering College, Pulchowk, Lalitpur for providing with the opportunity to proceed with the research and guidance. Authors would also like to thank Metropolitan Traffic Police Division for providing with the required data. Authors would also like to express gratitude to friends and family who have been through the research periods.

References

[1] S. AL-Ghamdi Ali. Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis and Prevention 34*, pages 729–741, 2002.

[2] Khalili Mohodeseh and Pakgohar Alireza. Logistic Regression Approach in Road Defects Impact on Accident Severity. *Journal of Emerfing Technologies in Wen Intelligence*, 05(02).

[3] Sampson Wiredu and Tawiah Richard. Exploring the predictors of Accident Severity in Urban Ghana. *Developing Country Studies, ISSN 2224-607X (Paper),*, 5(14), 2015.

[4] Saffarzadeh Mahmoud, Dovom Zangooei Mehdi, and Nadim Navid. An analysis of Pedestrian Fatal Accident Severity Using a Binary Logistic Regression Model. *ITE Journal*.

[5] Bahrololoom Sareh, Moridpour Sara, and Tay Richard. A logistic Regression Model for Hit and Run Bicycle Crashes in Victoria, Australia. 2011.

[6] Karacasu Murat, Ergul Baris, and Yavuz Altin Arzu. Estimating the causes of traffic accidents using logistic regression and discriminant analysis. *International Journal of Injury Control and Safety Promotion*, 21(4):305–312, 2014.

[7] Agyemang Boakye, Abledu Dr.G.K., and Semevoh Reuben. Regression Analysis of Road Traffic Accidents and Population Growth in Ghana. *International Journal of Business and Social Research (IJBSR)*, 3(10).

[8] Oirjaka Gentiana, Burazeri Genc, Hysa Bajram, and Roshi Enver. Factors associated with Fatal Traffic Accidents in Tirana, Albania: Cross Sectional Study. 2008.

[9] S Renuraj, N Varathan, and N Satkunananthan. Factors Influencing Traffic Accidents in Jaffna. *Sri Lankan Journal of Applied Statistics*, 16(2).

[10] H. David and Lemeshow Stanley. *Applied Logistic Regression, Second Edition*.