International Journal of Mechanical, Industrial and Aerospace Sciences
ISSN: 2517-9950
Vol:8, No:2, 2014

Prediction of Road Accidents in Qatar by 2022

M. Abou-Amouna, A. Radwan, L. Al-kuwari, A. Hammuda, K. Al-Khalifa

***Abstract*—**There is growing concern over increasing incidences of road accidents and consequent loss of human life in Qatar. In light to the future planned event in Qatar, World Cup 2022; Qatar should put into consideration the future deaths caused by road accidents, and past trends should be considered to give a reasonable picture of what may happen in the future. Qatar roads should be arranged and paved in a way that accommodate high capacity of the population in that time, since then there will be a huge number of visitors from the world. Qatar should also consider the risk issues of road accidents raised in that period, and plan to maintain high level to safety strategies. According to the increase in the number of road accidents in Qatar from 1995 until 2012, an analysis of elements affecting and causing road accidents will be effectively studied. This paper aims to identify and criticize the factors that have high effect on causing road accidents in the state of Qatar, and predict the total number of road accidents in Qatar 2022. Alternative methods are discussed and the most applicable ones according to the previous researches are selected for further studies. The methods that satisfy the existing case in Qatar were the multiple linear regression model (MLR) and artificial neutral network (ANN). Those methods are analyzed and their findings are compared. We conclude that by using MLR the number of accidents in 2022 will become 355,226 accidents, and by using ANN 216,264 accidents. We conclude that MLR gave better results than ANN because the artificial neutral network doesn't fit data with large range varieties.

***Keywords*—**Road Safety, Prediction, Accident, Model, Qatar.

I. Introduction

IN the world, 1.2 million people die because of traffic accidents each year. For that reason, most of the developed countries design and implement deferent strategies and different scales to reduce the road accidents by using education, and engineering. To have safe roads and to reduce the scale of road accidents, we need to know the safety level of the road and the most affecting variables in the road.

The number of road accidents can be predicted by Accident Prediction Model which is a mathematical formula describing the relation between the safety level of existing roads (i.e. crashes, victims, injured, fatalities) and variables that explain this level (road length, width, traffic volume). Traffic volumes and road lengths (km) are the most important descriptive variables in an accident prediction model.Accident prediction models are command to research organizations to develop basic accident prediction models for applicable road type.

In 2010, Qatar had the world's highest GDP per capita. The ongoing increases in production and exports of liquefied natural gas, oil, petrochemicals, and related industries are the main reason for this rapid growth. Qatar National Vision 2030 builds a bridge from the present to the future. It aims to provide a high standard of living for all people in Qatar for generations to come. Qatar is involved in a program of infrastructure investment to upgrade roads, utilities, railways and related services in response to the country's growing economy and population.

The discovery of oil around the middle of the last century has changed many aspects of life in Qatar. There was an increase in immigration and population, with a corresponding increase in vehicle numbers accompanied by rapidly expanding road construction programs. Unfortunately patterns of behavior did not change so rapidly, with behaviors rooted in traditional cultures mixed together with the Western culture. The result has been a large increase in the number of road traffic accidents with casualties and fatalities creating a serious public health problem. This problem drastically needs targeted research in order to identify methods of reducing accidents and fatalities. Road safety has been given high priority for some years in Qatar through extensive safety and awareness campaigns and more aggressive law enforcement, which over the last three years has helped reduce fatalities to their lowest mark in two decades. But road accidents involving at least one vehicle and resulting in damage and injury have been increasing for the last five years as population growth has led to more congested roads. While the 2008 fatality rate of 15.9 deaths per 100,000 people is a fair national result for Qatar, it is still markedly higher than the average for high-income countries [8].

Many applicable methodologies can be used to predict the number of road accident however the most applicable ones that best fit the existing data are the multiple linear regression and the artificial Neural Network model. The models are implemented and the findings are analyzed.

II. Three Main Factors Modeling

In order to be able to build a model that best fits the collected input data, the entire data should normalized. The data points of each factor have different ranges that don't coincide with any other range factors. Standardization had to be implemented to achieve data points with unified ranges. All data ranges lie between -1 and 1. The data of each factor were first sorted ascending independently, the smallest rank had a value of -1 and the largest had a value of 1. Each data point among the smallest to the largest value was subtracted from the least value, divided by the whole range average, and the ratio was subtracted from 1. The obtained values are larger than -1 but less than 1. All factors data points have similar ranges. Minitab software package is used to insert the

M.Abou-Amouna and K. Al Khalifa* are with the Qatar Road Safety Studies Center, Qatar University, Qatar (*corresponding author to provide phone: 00974-4403-4338; fax: 00974-4403- 4302; e-mail: alkhalifa@ qu.edu.qa).

A. Radwan, L. Al-kuwari, and A. Hammuda are with Mechanical and Industrial Engineering Department , Qatar University , Qatar (phone: 00974-4403-4303; fax: 00974-4403- 4101; e-mail: hamouda@ qu.edu.qa).

standardized data points of the three main factors (i.e. population estimate, number of driving licenses, and number of vehicles). A command of building a linear regression model was requested to construct a prediction model that best fit the entire data points. The attained model was used to validate and verify the output results of the original system [1].

III. UsingThree Main Factors to Predict the Number of Road Accidents

The three main factors (i.e. Number of vehicles, population estimate, and number of driving licenses) selected based on the previous studies were standardized and implemented in Minitab. The study uses cross-sectional data for 16 observed data points (i.e. recorded data from year 1995-2010. We have one dependent variable, number of road accidents, and three independent variables, number of vehicles, population estimate, and number of driving licenses.

Fig. 1 shows the output for this regression analysis. The Minitab gives the regression equation, as well as the coefficients, T statistic for each coefficient, and the corresponding P-value by which the significance level of each coefficient can be evaluated. A notice is that, while the F statistic level of significance was at the 100% level (indicating the overall model is significant), the Adjusted R-square value of 96.7% indicates the model accounts for about 97% of the response variable variation [4].

Regression Analysis: RA versus NV, PE, NDL

The regression equation is
RA = 0.113 + 0.111 NV + 1.91 PE - 0.982 NDL

Predictor	Coef	SE Coef	T	P
Constant	0.11306	0.04104	2.75	0.017
NV	0.1106	0.2022	0.55	0.594
PE	1.9054	0.3341	5.70	0.000
NDL	-0.9819	0.3740	-2.63	0.022

S = 0.138735 R-Sq = 97.3% R-Sq(adj) = 96.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	8.4835	2.8278	146.92	0.000
Residual Error	12	0.2310	0.0192		
Total	15	8.7144			

Source	DF	Seq SS
NV	1	7.1958
PE	1	1.1550
NDL	1	0.1327

Unusual Observations

Obs	NV	RA	Fit	SE Fit	Residual	St Resid
13	0.14	0.6881	0.3827	0.0536	0.3054	2.39R

R denotes an observation with a large standardized residual.

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	0.9700	0.1805	(0.5768, 1.3633)	(0.4741, 1.4660)XX

XX denotes a point that is an extreme outlier in the predictors.

Fig. 1 Output results of manipulating linear regression to three factors

IV. Prediction of Number of Road Accidents BasedOn the Classification of the Main Factors

The concepts and principles developed in dealing with simple linear regression discussed above may be extended to deal with several explanatory variables, Therefore the study is expanded to include eight factors, which are number of vehicle driving license for female and male, truck driving license for male, construction driving license for male, motorbikes driving license for male, number of vehicles, and population estimate for female and male.

The decision-making process for a hypothesis test can be based on the probability value (p-value) for the given test. If the p-value is less than or equal to a predetermined level of significance (α-level), then reject the null hypothesis and claim support for the alternative hypothesis. If the p-value is greater than the α-level, you fail to reject the null hypothesis and cannot claim support for the alternative hypothesis.

In the following ANOVA table in Fig. 2, the p-value (0.000) of the model provides sufficient evidence that the factors affecting the number of road accidents are different for at least one of them when α is 0.05. In the individual 95% confidence intervals table, notice that none of the intervals overlap, which supports the theory that the means are statistically different.

The ANOVA table does not have the F-critical value obtained from the F table, but we do have the P-value, which is 0.000, therefore less than the critical value of 0.05. So we must reject the null hypothesis and conclude that at least one independent variable is correlated with the dependent variable. The P-values are the results of hypotheses testing for every individual coefficient. The tests will help determine if the variable whose coefficient is being tested is significant in the model, i.e., if it must be kept or deleted from the model. The P-value is compared to the α level, which in general is equal to 0.05. If the P-value is less than 0.05, we are in the rejection zone and what conclude that the variable is significant and reject the null hypothesis. Otherwise, the null hypothesis cannot be rejecting. In the obtained results from Minitab displayed below, all the P-values are greater than the 0.05 except for “DLCM, NV, PEF, and PEM” which are 0.02, 0.022, 0.022, and 0.001 respectively. So those factors are the mostly independent variables that are significantly correlated with the dependent factor “Number of Road Accidents". The coefficient of determination is defined as the proportion in the variation of the response variable that is explained by the independent factor. But taking into account sample sizes and the degrees of freedom of independent factors it is recommended to assure that the coefficient of determination is not inflated. The formula for the adjusted coefficient of determination is determined by the following equation. From the obtained results by Minitab, the coefficient of determination is 99.4%, this means that about 99% of the original uncertainty is described by the built model.

$$AdjR^2 = 1 - \left|(1 - R^2)\frac{n-1}{n-1-k}\right|$$

where k is the number of independent factors, which is eight.

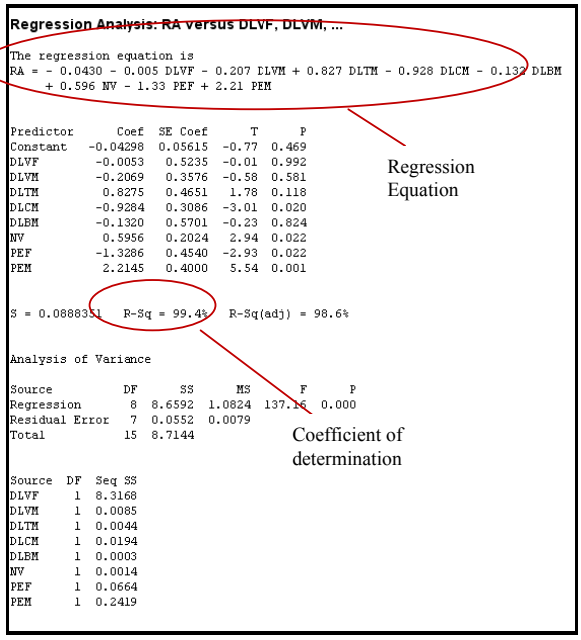

Fig. 2 Output results of manipulating linear regression to eight factors

The obtained results from the Minitab validate the model, since it gives an output results matching and reasonably close enough to the real output. The predicted value from implementing the regression model for year 2011 is 194693 road accidents, in which the real observed value was 216380 road accidents in the same year. This perceives of approximately 90% accuracy in the model system is reflecting reality and operating as the original system. Therefore, this obtained model is verified and used to predict the number of road accidents in year 2022. The following sections illustrate the analysis and prediction of each factor used to predict the number of road accidents in 2022 [5], [2].

V. Factors Correlation with the Number of Road Accidents

The following is a graphical display of each independent factor with the number of road accidents in the specified time period. Each factor is plot versus the number of road accidents, while all other factors are held constant. The graphs are displayed in one plot for the ease of interface.

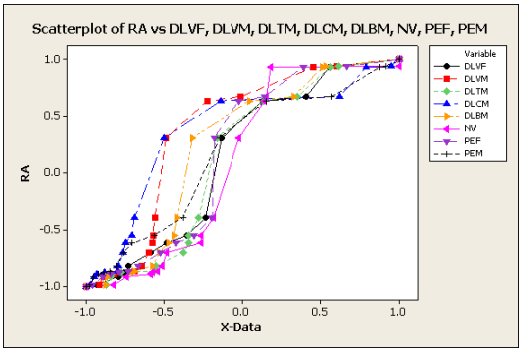

Fig. 3 Factors correlation with the number of road accidents

From the graph, all factors have positive correlation with the number of road accidents. As the value of the independent factor increases, the number of road accidents in a specified period also increases. This positive, direct proportion doesn't follow a constant increasing trend each with a different relative ratio.

VI. Predicting the Factors Based On the Total Population Estimate

The factors are interred correlated with others; a certain trend of one factor is proportional to another. In this section we have predicted the factors based on the total population estimate. The observed values of the total population estimate and each factor were standardized; the ratio was calculated (the standardized value of the factor over the standardized value of population estimate) and plotted against the sequential number of years.

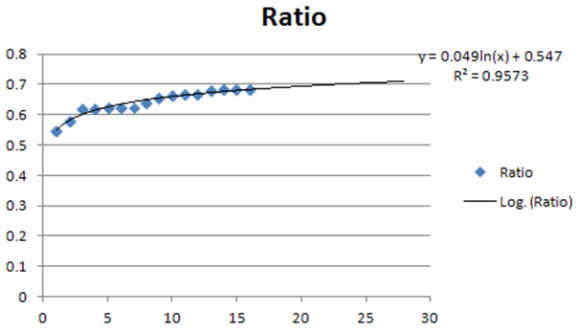

Fig. 4 Obtained results of plotting the ratio versus the std. DLVF

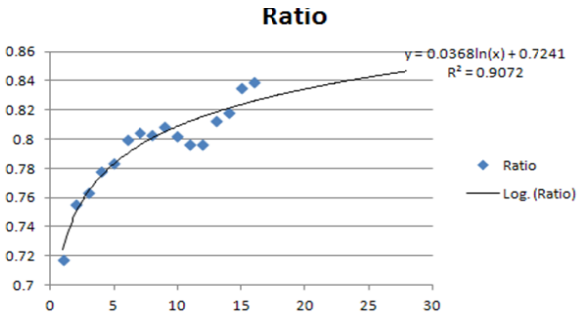

Fig. 5 Obtained results of plotting the ratio versus the std. DLVM

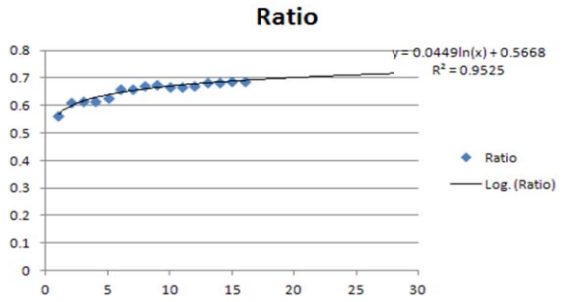

Fig. 6 Obtained results of plotting the ratio versus the std. DLTM

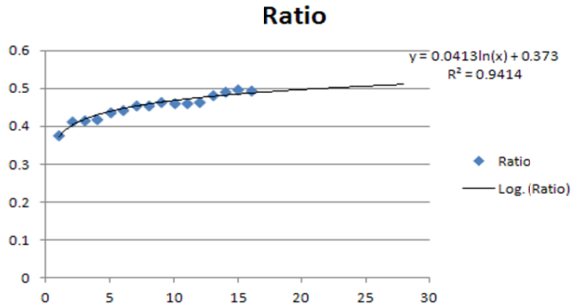

Fig. 7 Obtained results of plotting the ratio versus the std. DLCM

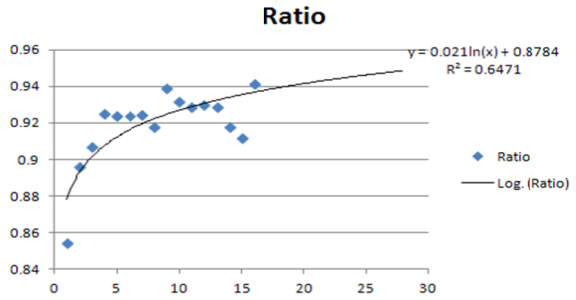

Fig. 8 Obtained results of plotting the ratio versus the std. DLBM

VII. Prediction of Road Accidents in Qatar 2022

The following output displays the obtained results; the built model indicates high value of the r-square, which in terms illustrates that the model is strong. Another good indication of the model validity is the model p-value, which is very small (i.e. approaching zero).

Regression Analysis: RA versus DLVF, DLVM, DLTM, DLCM, DLBM, NV, PE

The regression equation is
RA = - 0.0104 - 0.694 DLVF - 0.924 DLVM - 0.118 DLTM - 0.419 DLCM + 0.83 DLBM
+ 0.092 NV + 2.21 PE

Predictor	Coef	SE Coef	T	P
Constant	-0.01039	0.06598	-0.16	0.878
DLVF	-0.6945	0.8496	-0.82	0.435
DLVM	-0.9243	0.4565	-2.02	0.074
DLTM	-0.1178	0.7155	-0.16	0.873
DLCM	-0.4194	0.4247	-0.99	0.349
DLBM	0.834	1.085	0.77	0.462
NV	0.0915	0.2588	0.35	0.732
PE	2.2112	0.7407	2.99	0.015

S = 0.141605 R-Sq = 98.2% R-Sq(adj) = 96.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	7	9.6208	1.3744	68.54	0.000
Residual Error	9	0.1805	0.0201		
Total	16	9.8013			

Source	DF	Seq SS
DLVF	1	9.2421
DLVM	1	0.0766
DLTM	1	0.0000
DLCM	1	0.1050
DLBM	1	0.0137
NV	1	0.0047
PE	1	0.1787

Fig. 9 RA vs DLVF DLVM DLTM DLCM DLBM NV PE (a)

A prediction interval refers to a specific point. The aim of this part is to predict the fitted value, with 95% confidence, for the number of road accidents in2022. This is done as an extra option of the previous analysis, producing additional output. The fitted value obtained by Minitab which is 2.5979 was converted backward with the same equation used for standardization, and the number of road accidents in 2022 was predicted to be 355,226 road accidents.

Obs	DLVF	RA	Fit	SE Fit	Residual	St Resid
11	-0.24	-0.3948	-0.1292	0.0854	-0.2656	-2.35R

R denotes an observation with a large standardized residual.

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	2.5979	0.3942	(1.7060, 3.4897)	(1.6503, 3.5455)XX

XX denotes a point that is an extreme outlier in the predictors.

Values of Predictors for New Observations

New Obs	DLVF	DLVM	DLTM	DLCM	DLBM	NV	PE
1	1.92	1.38	1.38	1.15	1.23	2.21	2.09

Fig. 10 Observations

Plotting the residuals against the fitted values is one of the standard techniques introduced to check the assumption of homogeneity of variance. It is illustrated here with thedataset. The standardized residuals were used rather than raw residuals for these model checking plots, as they are easier to interpret. The Normality of error may be examined by plotting a histogram of residuals or a Normal order plot. Normal Order plots allow a more quantitative assessment of the Normality of the distribution of residuals. This method will first of all be described with some normally distributed residuals. The residuals in following histogram plot appear to be normally distributed.The residuals can also be plotted against the fitted values. This technique is used as a means of checking for homogeneity of variance. The residuals show increasing and decreasing variance with the fitted values. The plots indicate consistency with the earlier assumptions.

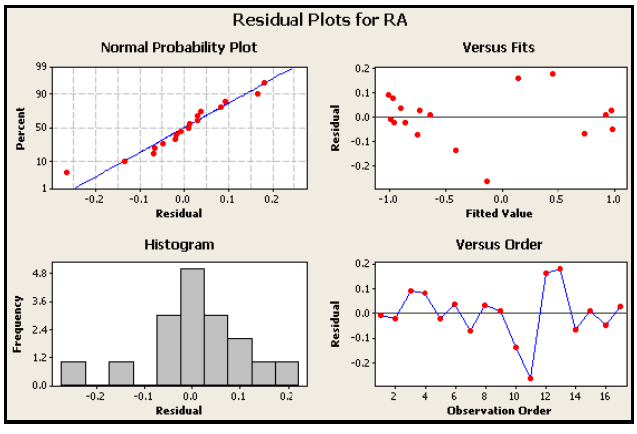

Fig. 11 Residual Plots for RA (a)

VIII. Prediction of Road Accidents in Qatar 2022 by Using Neural Network

Starting with building a model and training the Neural Network: The idea of the Neural Network is that it needs to see enough values to be trained the more data given to the Neural Network the better results it will achieve. An objective function is specified which is a measure of how closely the outputs of the network match the target outputs in the training set of data [3], [7].

The output values of the network after training were close to the observed values of road accidents over the past 16 years as shown in the figure below.

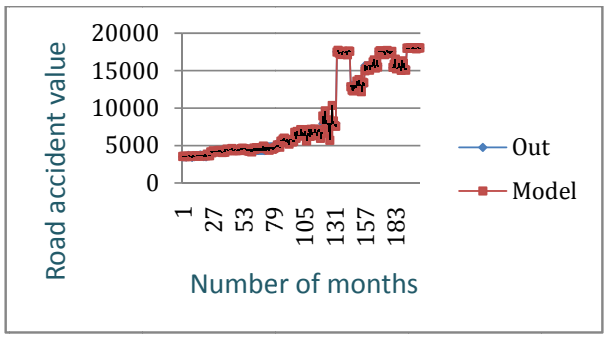

Fig. 12 Fit of output values of ANN to the observed values

After training the network Neural Works Predict (Test) automatically integrates all the components required to effectively solve prediction and classification problems.

The Training Complete dialog shown in the figure above represents a summary of network attributes and performance statistics for the model when it is run with the Train and Test sets.The specific content of the Training Complete dialog box depends on the type of model; however, for the prediction model built for this project, the following information is displayed.

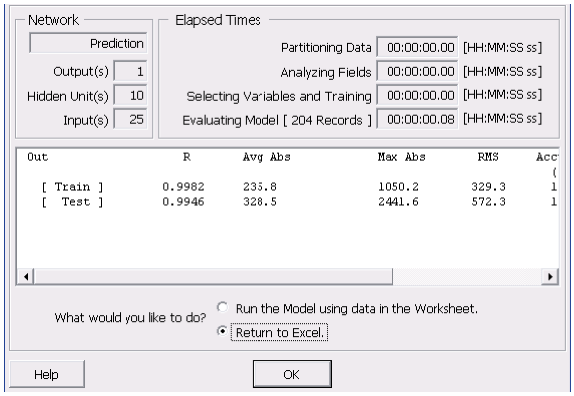

Fig.13 Normal probability plot of regression model

The most important and useful metrics for a prediction model are usually the R (person R) value. The R value indicates how close one data series is to another. In this case the data series are the actual output values and the corresponding predicted output values generated by the model. A large R value indicates a higher correlation which means a better model. Comparing the Train and Test sets highlighted in the above figure, the relatively small differences between values (0.9982 and 0.9946) suggests that the model generalizes well and that it is likely to make accurate predictions when it processes new data. Hidden Units are the total number of hidden processing elements that the network may contain. The number of hidden points in the Training dialog box are acceptable since it's not very high to extremes and also not very low.

The reason for building any model is so that output values can be accurately predicted when a new data record is processed by the model. Using the monthly value of each factor in 2022 the predicted value equals 18022.04.

The predicted value is the value for each month, multiplying the value by 12 gives 216264 which is the value of road accidents in 2022. This value is close to the value of the network output in 2011 which equals 216540. The reason for this is that the test record shown in the above table contains values that are larger than the largest corresponding value in the training data. Internally, the Predict engine does not scale extreme values; the model can accommodate some values that are outside the range of training data, but not values that are extreme. Input6 provides the greatest relative influence on the model as the highest value in the trained data was 149345, so the effect of an outlier is magnified. We conclude that the ANN needs to be trained with values close to those needed to predicted, it could be more applicable to use the ANN to predict for 2011 since the range of the values will not vary much than those used to train the network (i.e. 2010).

IX. Comparing Neural Network and Linear Regression

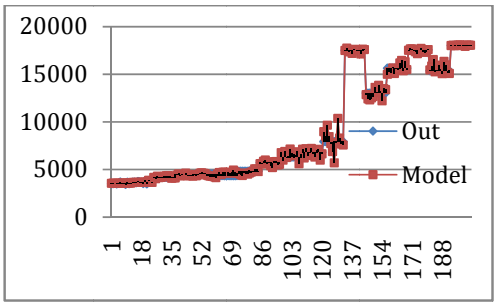

Fig.14 Fit of output data of ANN to the observed values

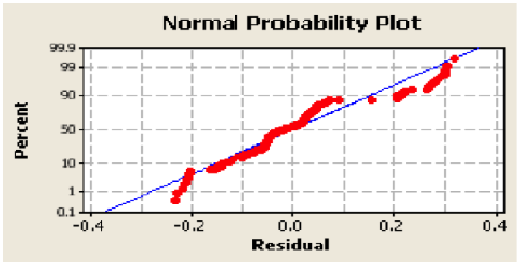

Fig.15 Normal probability plot of regression model

The two figures illustrate the accuracy of the prediction of each method. It's clearly obvious that the Neural Network predicted values are closer to the observed values; however, that doesn't mean that the regression model is not able to accurately predict. Unlike the regression model the Neural Network is able to follow the trend of the data since it is not linear [6].

The following table shows the R value for each of the methods, the higher the value of R the better the model is.

TABLE I
R VALUE FOR BOTH METHODS

Method	Neural Network	Linear Regression
value ofR	0.9946	0.975

It is clear that the Neural Network model is better in terms of R value; therefore, it gives better result for road accidents in 2011, the value equals 216540 which is very close to the observed value in 2011 (216380), on the other hand the regression model predicted value for 2011 equals 194232 ,although the number is not far from the observed value ,the Neural Network is a better method to predict for 2011 since it follows the trend of the data; however, the Neural Network will always give values close to the values used to train the network ,therefore it's more preferable to use the regression model to predict for 2022.

X. CONCLUSION

Traffic accidents are greatly concern for all members of society, and became one of the most important problems that drain the material and human resources, traffic accident usually caused by damage and injuries, ranging from minor to property and vehicles serious to result in death or permanent disability.

The main objective in this study was to predict the number of road accidents in year 2022. To sum up what is done in this report, first different studies were done about traffic safety and prediction of road accidents from different countries. The methods used to predict road traffic accidents are multiple liner regression models and the artificial neural network. Then the used methods were explained in details with input and outputs parameters and the interaction between the data. Finally findings and analysis were described in details in order to check whether the erected model by the collected input data gives an output results that significantly match the observed real life outcomes.

REFERENCES

[1] Melvyn Hirst "Building and Operating a Forecasting Model:The Regression Analysis Approach" University of Warwick, 2007 .

[2] Fajaruddin Mustakim, Ismail Yusof, Ismail Rahman, Abdul Aziz Abdul Samad, Nor Esnizah Binti Mohd Salleh, "Blackspot Study and Accident Prediction Model Using Multiple LinerRegression" First International Conference on Construction In Developing Countries (ICCIDC–I), August 4-5, 2008, Karachi,, Pakistan.

[3] Fu Huilin, ZhouYucai, "The Traffic Accident Prediction Based On Neural Network," Digital Manufacturing and Automation (ICDMA), 2011 Second International Conference on Digital Manufacturing & Automation, vol., no., pp.1349-1350, 5-7 Aug. 2011.

[4] Keay, Kevin and Simmonds, Ian, „Road Accidents and Rainfall in a Large Australian City", Journal of Accident Analysis & Prevention, May 2006, Vol. 38 Issue 3, p445-454,

[5] Luis Felipe Rodricuez"Accident Prediction Models for UnsignalizedIntersections: A Master of Applied Science from The University of British Columbia", April 1998

[6] Mehmet MetinKunt, ImanAghayan, and NimaNoii, "Prediction for Traffic Accident Severity: Comparing the Artificial Neural Network, Genetic Algorithm, Combined Genetic Algorithm and Pattern Search Methods", online 2011 Vilnius Gediminas Technical University (VGTU) Press Technika Volume 26(4): 353–366.

[7] Yu Rujun; Liu Xiuqing, "Study on Traffic Accidents Prediction Model Based on RBF Neural Network," Information Engineering and Computer Science (ICIECS), 2010 2nd International Conference on Information Engineering and Computer Science, vol., no., pp.1-4, 25-26 Dec. 2010.

[8] General Sectorial for Development Planning, "Qatar National Vision2030".