DEVELOPMENT OF CRASH PREDICTION MODEL USING MULTIPLE REGRESSION ANALYSIS

Harshit Gupta[1], Dr. Siddhartha Rokade[2]
[1]PG Student, [2]Assistant Professor,
Department of Civil Engineering, Maulana Azad National Institute of Technology, Bhopal

Abstract
The purpose of the study is to develop a model for prediction of crashes in urban medium size cities. In this paper Crash prediction model (CPM) is developed using multiple regression analysis. A model is a simplified representation of a real world process. It should be representative in the sense that it should contain the salient features of the phenomena under study. In general, one of the objectives in modeling is to have a simple model to explain a complex phenomenon.
Keywords: Crash prediction model (CPM), Multiple regression Analysis, Medium size city

1. INTRODUCTION

Road accidents are increasing every year. The total number of road accidents increased marginally from 4,86,476 in 2013 to 4,89,400 in 2014 (MoRT&H,2015). This is probably an underestimate. The actual numbers of injuries are much higher than official statics, as not all injuries are reported to the police stations. Road crashes are an outcome of various factors, some of which are the length of road network, vehicle population, human population and enforcement of road safety regulations etc.

Bhopal city is selected for development of CPM. About 3500 traffic crashes occurred every year in Bhopal city. Crash Prediction Models (CPMs) is developed to know the expected number of crashes on a road in a certain time frame. Crash frequency depends on Segment length, the traffic flow and several risk factors. No. of accidents is called the dependent (variable). The other components like traffic volume, carriage way width are incorporated in the CPM as independent model variables, also called predictors. SPSS software is used for regression analysis.

2. DESCRIPTION OF DATA

To develop a crash prediction model independent and dependent variables are needed. The selection of appropriate variables for model development is an important step as these variables significantly influence the accuracy and validity of model. The collected data was divided into two parts that is dependent data and independent data.

a) Dependent data - Road accident data
b) Independent data - Factors affecting the crashes like segment length, carriageway width, median width etc. These variables are selected on the basis former literatures and data of these variables are collected from the selected segments.

Traffic volume and road geometric design characteristics are related to accidents. In many research results show that changing in geometric design characteristics and traffic volume accident rate change.

Traffic volume is also an important parameter. Traffic volume on selected segment is counted manual at peak hour. Peak hour of each segment is decided on the basis of expert opinion.

2.1. Collection of Accident Data

A 6 year (2011-2015 & 2016) Accident data was collected in form the police department of the Bhopal city. The data is digitized in three categories; accident details, vehicle details, and victim details (include Gender and age of victim).

Different type of land uses are selected for the study like Shops, Block of flats, Industrial, Residential, neighborhood, scattered housing, open spaces, Market areas (CBDs) etc.

2.2. Road characteristics data

The road characteristics data are collected in the year 2016 The road characteristics data segment length, road width, traffic volume, median width, no. access road and pedestrian volume to be incorporated in the model analysis. Due to limitations, the road design elements like vertical alignment, horizontal alignment, lighting, road markings, and haptic feedbacks attributes could not be included in the model analyses. Environmental factors, vehicular factor, human factor are also not included in this analysis.

A total of 76 segments were selected for development of accident prediction model. The sections that were considered included the following

1. The length of the access roads varied from 100 m to 1400 m with a mean of 413 m and total length of segments is 32,210 meters
2. The Peak hour traffic volume varied from 1810 to 4560 PCU with mean of 3144.
3. The number of crashes over 6 years varied from 10 to 39 crashes in a section with mean of 20. Total number of crashes was 1558.
4. The no. of access road varied from 1 to 8 with the mean of 2.75
5. The pedestrian volume varied from 13 to 345 with the mean of 145

3. MODEL DEVELOPMENT

Multiple regression is an extension of simple linear regression. If the independent variables are two or more in numbers, then the analysis is known as multiple linear regression.The main idea of multiple linear regression method is to build correlation analysis between dependent and independent variables. The function will be the following form:

$$\mathbf{Y=B_0+B_1X_1+B_2X_2+B_3X_1+\ldots+B_mX_m}$$

Where,

Y= Dependent variable

B_0= regression constant,

B_1, B_2, B_3...B_m = Regression coefficients of the respective m independent variables.

X_1, X_2, X_3, ……..X_m = Independent variables.

A statistic test of the model was necessary, including determination coefficient test (R^2 test), significance test of regression coefficient (t-test), and significance test of regression equation (F test). If the significant test of regression equation failed, it was possible that important factors were missed during the selection of independent variables, or the relationship between independent and dependent variables was nonlinear, in which situation the model should be rebuilt (Feng et al., 2016). The multiple linear regression analysis makes several key assumptions such as linear relationship, multicollinearity, normality, independence and homoscedasticity between the model variables. Finally these assumptions should also be examined for developed model.

It assumes that, there is a linear relationship between the dependent variable and each of independent variables and the dependent variable and independent variable collectively.

3.1. Correlation Analysis

Bivariate correlation analysis is done by SPSS. Pearson correlation coefficient is selected for analysis. Correlation is a unit free measure of relationship between two variables and take value in [-1, +1], where r is close to +1(-1), there is strong and positive (negative) relationship and a correlation coefficient of 0 indicates that there is no linear relationship between the two variables. It measures only linear relationship.

In table 1 seems that the correlation test value of independent variable with dependent variable. Sign (**) indicates a Correlation is significant at the 0.01 level where the sign (*) Indicate a moderate Correlation is significant at the 0.05 level.

Table 1 shows parameters having positive signs with traffic accident. It is presented that variable has positive correlation which means that variables are perfectly related in a positive linear sense if the correlation value is positive and negative linear sense when correlation value is negative.

Table 1: Correlation Value of Independent Variable (X) With Dependent Variable(Y)

Sr. No.	Independent Variable	Symbol	Correlation Test value
1.	Segment length	SL	0.295**
2.	Carriageway Width	CW	0.425**
3.	Pedestrian Volume	PV	0.249*
4.	Traffic Volume	TV	0.440**
5.	No. of Access Road	AP	0.232*

3.2. Multicollinearity analyses

In statistics, multicollinearity or collinearity is a phenomenon in which two or more predictor variables (Independent variables) in a MLRM are highly correlated. This means, one can be linearly predicted from the others with a substantial degree of accuracy. Multicollinearity poses a problem only for multiple regression analyses. Collinearity is between -1 to +1. Perfect collinearity means, two predictors that are perfectly correlated have a correlation coefficient is 1). If there is perfect collinearity between predictors it becomes impossible to obtain unique regression coefficients because there are an infinite number of combinations of coefficients that would work equally well. Put simply, if we have two predictors that are perfectly correlated, then the values of b for each variable are interchangeable. (Field 2005)

One way of identifying multicollinearity is to scan a correlation matrix of all of the predictor variables and see if any correlate very highly (by very highly mean correlations of above .80 or .90). SPSS produces various collinearity diagnostics, one of which is the variance inflation factor (VIF). The VIF indicates whether a predictor has a strong linear relationship with the other predictor(s).

Table 2: Correlation Matrix of Selected Independent Variable

Predictor	TV	AL	AP	PV	CW
TV	1	.052	.111	.089	.483**
SL	.052	1	.481**	-.416**	-.178
AP	.111	.481**	1	-.338**	-.043
PV	.089	-.416**	-.338**	1	.203
CW	.483**	-.178	-.043	.203	1

In this research correlation matrix is used to detect the problem of multicollinearity. The values of correlation between the independent variables are larger than |0.5| indicate the multicollinearity otherwise there is no multicollinearity. Table 2 shows the correlation values of selected independent variables with each other. It seems that the multicollinearity is absent.

In the above section it is presented that there is no coloration between selected variable. If collinearity present than we remove variables with multicollinearity. All variables are selected for model formulation.

3.3. Model Formulation by Multiple linear regression analysis using SPSS

Parameters which are used in this study for development of model are identified using literature reviews. The parameter include as an input for model development are:

- Segment Length i.e. Length of the road segment were an accident occurred
- Traffic Volume i.e. Peak hour traffic volume (PCU/hour) on carriageway
- Carriageway width i.e. Width of road on which a vehicles are not restricted by any physical barriers to move laterally.
- Access Road i.e. No. of access road at road segment.
- Pedestrian volume i.e. Total number of pedestrian on the road
- Vehicle Accident i.e. Accident of vehicle on the selected segments.

I. Model Summary: Table 3 shows the model summary.

The "R" column represents the value of R, the multiple correlation coefficients.

The "R Square" column represents the R^2 value (also called the coefficient of determination).

Table 3: Model Summary of MLR

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.686[a]	.471	.433	4.789

It is the proportion of variance in the dependent variable that can be explained by the independent variable.

II Selection of Best Model

The most sensitive approach to selecting a subset of important variables in a complex linear model is to compare all possible subsets. This procedure simply fits all the possible regression models (i.e., all possible combinations of predictors) and chooses the best one (or more than one) based on the criteria described below.

1. High R, R^2 and Adjusted R^2 value
2. Low significance F value and Minimum standard error

3. Low p value for the coefficients of independent variables and y intercept

III. Formulation of Model Equation

Segment length, traffic volume and pedestrian volume have high value in comparison of other parameters. Regression coefficients of respective variables SL, TV, PV are in three decimals. To obtain a proper regression coefficient SL, TV, PV are divided by respectively 10, 100 and 10. The Coefficients part of the output gives us the values that we need in order to write the regression equation. The accident Prediction model for selected road segments using the multiple regression analysis is presented in equation 5.1

TC = -2.179 + 1.164(CW) + 0.0114(SL) + 0.525(AP) + 0.0312(PV) + 0.00122(TV)

Where,

-2.179 = regression constant or intercept

1.164, 0.0114, 0.525, 0.0312 and 0.00122 = regression coefficients of their respective variables.

CW = Width of road on which a vehicles are not restricted by any physical barriers to move laterally

SL = Segment Length

AP = No. of access road

PV = Pedestrian Volume

TV = Peak hour Traffic Volume in PCU/hour

3.4. Validation and Assumptions of MLR Model

ANOVA, F-test, T-test and normality check in multiple linear regression.

1. ANOVA:

Analysis of Variance (ANOVA) is a statistical method used to test differences between two or more means. Analysis of Variance (ANOVA) consists of calculations that provide information about levels of variability within a regression model and form a basis for tests of significance. The ANOVA calculations for multiple regression are nearly identical to the calculations for simple linear regression, except that the degrees of freedom are adjusted to reflect the number of explanatory variables included in the model.

Table 4 shows the output of the ANOVA analysis. Result of ANOVA test is f (5, 70) = 12.451 which is at the significance level 0.000 (P-value = .000), P value is less than 0.005. It means that the regression equation is significant at 95 % confidence limit in levels of variability within a regression model.

Table 4: ANOVA, Multiple Linear Regression Analysis in SPSS

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	1427.68	5	285.538	12.451	.000[b]
Residual	1605.31	70	22.933		
Total	3033.00	75			

2. Examining Normality

The normal probability plot shows the theoretical percentiles of a normal distribution versus the actual percentiles of the standard residuals with the same variance and mean. If the collected data are normally distributed, then the observed values (the dots on the chart) should fall closely along the straight line (meaning that the observed values are the same as you would expect to get from a normally distributed data set). In broad, the more closely the points are clustered about the 45-degree line, the stronger the indication supporting the normality assumption. Any substantial curvature in the plot is evidence that the residuals have not come from a normal distribution

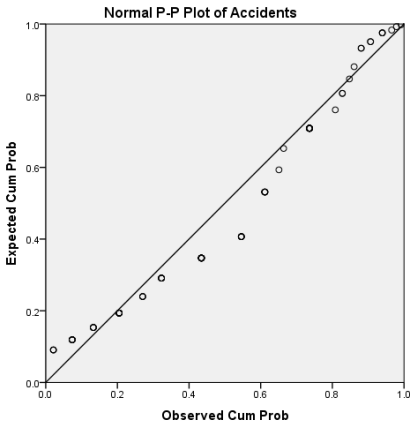

Figure 5.6: Normal P-P Plot of Regression Residual

In multiple regression analysis the accurate relationship between variable can be estimate if the relationships between variable are linear. If the relationship between dependent and independent variable is not a linear relationship, the results of the regression analysis will underestimate the true relationship. A preferable method of detection linearity is examination of residual plots between dependent and independent. The first step of assessment of assumption begins with initial descriptive plots of dependent Variable versus each of the explanatory variables, separately.

4. CONCLUSION

This research developed a crash prediction model by multiple linear regression analysis. Predictor variable carriage way widths, segment length, traffic volume, pedestrian volume, no. of access road are selected for the study. The risk is increasing as the traffic volume, pedestrian volume, carriage-way width, segment length and no. of access road increases. Previous study shows that presence of median decrease the risk of accident.

High R^2 value is needed for the model selection. Low R^2 value shows either some predictor variables are missing or the relation between dependent and independent variable is not linear or not follow normal distribution.

The accident prediction models can be used to decrease the risk on the road and safety performance. In terms of future work, it is suggested to taking other predictor variables such as horizontal and vertical alignment, median width, sidewalk width, environmental factors etc. CPM can be developed by regression analysis such as poisson regression, negative binomial regression, zero inflated regression model etc.

REFERENCES

[1] Feng S., Li Z., Ci Y. and Zhang G. (2016), *Risk factors affecting fatal bus accident severity: Their impact on different types of bus drivers*, Accident Analysis & Prevention, *86*, pp.29-39.

[2] MORT&H, (2015), *Road Accident Statistics*, Ministry of Road Transport and Highways, Government of India, New Delhi

[3] Field A. (2005), *Discovering statistics using SPSS*, SAGE Publications Ltd ISBN 978-1-84787-906-6

[4] IBM Software Group (2015), "*IBM SPSS Advanced Statistics 22*" Available from: http://library.uvm.edu/services/statistics/SPSS22Manuals/IBM%20SPSS%20Advanced%20Statistics.pdf

[5] Ranjitkar P and Chengye P. (2013), *Modelling motorway accidents using negative binomial regression*, Eastern Asia Society for Transportation Studies Vol. 9.

[6] Sawalha Z. and Sayed T. (2003), *Statistical issues in traffic accident modeling*, Canadian Journal of Civil Engineering, 33(9), P1115-1124

[7] Prajapati, P., Tiwari, G., Evaluating Safety of Urban Arterial Roads of Medium Sized Indian City. Proceedings of the Eastern Asia Society for Transportation Studies, Vol.9, 2013

[8] WHO (2015), *Global status report on road safety 2015*