Traffic Accident Analysis Using Structural Equation Modeling – A Review

Pankaj Prajapati
Associate Professor,
Deptt. of Civil Engineering,
Faculty of Technology & Engineering,
The Maharaja Sayajirao University of Baroda,
Vadodara, Gujrat, India

Abstract

There has been tremendous growth of both road network and road traffic in India in last few decades. While it is good for the economic and social development of the country, it has brought in its wake the problem of road accidents resulting in injuries and fatalities to road users and its own social negative externalities apart from human suffering. As automobiles transportation continues to increase around the world, bicyclists, pedestrians and motorcyclists, also known as vulnerable road users (VRUs), will become more susceptible to traffic accidents.

For improving traffic safety, there are many traditional methods like Poisson regression model, negative binomial regression, multinomial logistic regression, log linear model etc. this all methods are part of generalized linear regression. It is applied to assess various factors such as road, environment, vehicle, and driver.

In accident analysis, structural equation model (SEM) is another reliable and widely accepted in transportation field compared to other traditional methods. SEM is adopted to capture the complex relationships among variables because the model can handle relationships among endogenous and exogenous variables simultaneously and furthermore it can include latent variables in the model which are unobserved variables and represent one-dimensional concept in their purest form. The observed variables contain random or systematic error. Latent variables specified by linear combination of the observed variables. The linear combination are weighted averages. Hence, regression, path analysis, factor analysis and canonical correlation analysis are all special case of SEM. SEM consists of a set of equation that are specified by direct links between variables and it can be called “the simultaneous equation”.

Keywords: Poisson Regression; Binary Logistic Regression; Multinomial Logistic Regression; Log-Linear Model; Generalized Linear Model; Structural Equation Model; Latent Variable

Introduction

Transportation network is a heart of a nation and transport services are considered as growth engine of economy. There is a great social and economical loss to the society because of road accidents. There are many traditional methods to evaluate the road accidents, but now it is a time to study these accidents with new method like Structural Equation Model (SEM). Road network is the backbone of any country, acting as an indicator for the economic development of that country. The boom in trade, commerce, and industry depends directly on the growth of roads of a country. More the length of roads, more the prosperity of the nation. The prosperity brigades of a nation normally comprise of intelligentsia, hard labour, infrastructure available and lastly smooth functioning of its roads. However, with the positive qualities, the by-product of transportation is pollution and accidents.

Accident is the word causing spine chill in people who get associated with this in the course of their lifetime by any fate or haste. The accidents on roads are pronouncing and disastrous. The loss to the nation due to the ever-increasing accidents is untold, eating into the economics of the nation, to a larger chunk. The gravity of the situation is such that, the loss of limbs is directly proportional to the increase in the vehicle population. The term “safety” implies that no accidents are acceptable. However, accidents do happen and are caused due to the presence of several accident causative/contributory factors.

Aim of the Study

This paper aimed for the detailed review of traditional methods as well as SEM to ascertain parameters those are responsible for accident size. This will assists administrators to take necessary steps to address the problem to reduce the accident severity and size both.

Global Road Accident Scenario

World Health Organisation has cited road accident statistics in "Global Status Report on Road Safety 2018" (WHO 2018). There were 1.35 million people lost their lives globally in road traffic accidents in 2016 and thus road traffic injuries became the 8th leading cause of death. The rate of road traffic death per 100,000 population varies from 9.3 to 26.6 for high-income group countries and low-income group countries, which shows how road safety aspect is neglected in low-income group countries. Pedestrians, two wheeler users, and bicyclists are considered as vulnerable road users as they are directly exposed and come in direct contact with the impacting vehicle or obstacle during a collision resulting in severe injuries and fatality. The proportion of pedestrians and cyclists as victims is 26% while of motorised two- and three-wheelers is 28%. (WHO, 2018)

National Road accident scenario

Road network in India, of about 55 lakh km as of March 2015 is one of the largest in the world. The country's road network consists of National Highways, State Highways, Districts roads, Rural and Village roads. During the calendar year 2016, the total number of road accidents is reported at 4,80,652 causing injuries to 4,94,624 persons and claiming 1,50,785 lives in the country. This would translate, on an average, into 1317 accidents and 413 accident deaths taking place on Indian roads every day; or 55 accidents and 17 deaths every hour. National highway contribute 29.6 per cent of total road accidents and 34.5 per cent of total number of persons killed. The State Highways accounted for 25.3 per cent of total accidents and 27.9 per cent of the total number of persons killed in road accident in 2016. Nearly, 60 percent of total road accident take place during nights though the night traffic is hardly 15 percent of 24 hours volume.

The number of accidents and persons killed is going to be increasing from the year 2005 to 2016. In the year 2005 fatality rate is 21.6 and it is increase up to 31.4 in 2016. It is observed that number of registered motor vehicle is increasing, even though accident rate have decreased during this period. This may be due to increasing traffic leads to the reduction of speed, introduction of new roads, improved quality of road, increased width of roads, installation of traffic signs and signals, channelizing island at intersection, introducing medians on roads and improvement in geometrics of roads.

It is found that the proportion of two-wheelers, four wheelers, buses, goods vehicles and other vehicles has remained approximately constant from the year 2005 to 2016 and its value is 70%, 13%, 1%, 5% and 11% respectively. (MoRT&H, 2016)

Regional Road Accident Scenario

Gujarat is one of the most industrially developed and agriculturally advanced fertile state of India. So as the road length in Gujarat has increased from 47,426 km in 1981 to 67,065 km in 1991 to 79,619 km in 2011. With increased in road length, the total number of registered vehicles in Gujarat has increased from 10,28,90,560 in 2007-2008 to 23,28,64,180 in 2017-2018. Due to that the rate of accidents in Gujarat is 12.4 accidents per 10000 vehicles. Gujarat has 38 fatality rate of 100 accident average and it is increased since 2007. The fatality rate of 2007 is 21 of 100 accident average. (RTO, 2017)

In 2016, total number of registered accident is 1046 in Vadodara city of Gujarat, India. In that injury accident is 654 and fatal accident is 203 recorded. In 2016, 878 person is injured and 214 person are killed in road accident in Vadodara. The severity index is 20.5 per number of person killed per 100 accidents in Vadodara city. (MoRT&H, 2016)

Accident Analysis Using Different Types of Methods

There are many statistical method used in analysis of traffic accident. But basic technique which are used in analysis is generalized linear model in which many methods are included such as logistic regression, binary logistic regression, multinomial logistic regression, poisson regression, log-linear model and SEM.

Generalized Linear Model

Ackaah and Salifu (2011) conducted research and their main objective of the study was to develop a prediction model for road traffic crashes occuring on the rural sections of he highways in the ashanti region of Ghana. The main objective of the study was to develop a prediction model for road traffic crashes occurring on the rural sections of the high ways. The model was developed for all injury crashes occurring on selected rural highways in the Region over the three (3) year period 2005–2007. Data was collected from76 rural high way sections and each section varied between 0.8km and 6.7km. Data collected for each section comprised injury crash data, traffic flow and speed data, and roadway characteristics and road geometry data. The Generalised Linear Model (GLM) with Negative Binomial (NB) error structure was used to estimate the model parameters. Two types of models, the 'core' model which included key exposure variables only and the 'full' model which included a wider range of variables were developed.

'Core Model'- exposure variables only.

$$E(Y) = a_0 L^{a1} Q^{a2}$$

'Full Model' – product of the powers of the exposure variables multiplied by an exponential function incorporating the remaining explanatory variables.

$$E(Y) = a_0 L^{a1} Q^{a2} exp \sum j b_j x_j$$

Where,

E(Y)	predicted crash frequency,
L	section length (km),
Q	ADT (per day)
X_J	is any variable additional to L and Q, and
exp	exponential function, e=2.7183
a_0, a_1, a_2, b_j	are the model parameters.

After the statistical analysis the data was found to be over dispersed as the variance of the data was greater than the mean. Initial modelling using

Poisson error structure also showed that the estimated dispersion parameter (Φ) defined as:

$$\Phi = \frac{Pearson\,\varkappa^2}{(N-p)}$$

Where N is the total number of sections and p is the number of parameters in the model was greater than one (1) indicating that the data set was over-dispersed. That means Poisson distribution is not capable of explaining the true distribution underlying the crash frequency. Generalised Linear Model (GLM) was used to estimate the model coefficients using the STATA software package and assuming a Negative Binomial error distribution.

Logistic regression model

Milton et.al (2008) proposed that many transportation agencies use accident frequencies, and statistical models of accidents frequencies, as a basis for prioritizing highway Safety improvements. However, the use of accident severities in safety programming has been often been limited to the locational assessment of accident fatalities, with little or no emphasis being placed on the full severity distribution of accidents (property damage only, possible injury, injury)—which is needed to fully assess the benefits of competing safety-improvement projects. In this paper we demonstrate a modeling approach that can be used to better understand the injury-severity distributions of accidents on highway segments, and the effect that traffic, highway and weather characteristics have on these distributions. The approach we use allows for the possibility that estimated model parameters can vary randomly across roadway segments to account for unobserved effects potentially relating to roadway characteristics, environmental factors, and driver behavior. Using highway-injury data from Washington State, a mixed (random parameters) logit model is estimated. Estimation findings indicate that volume-related variables such as average daily traffic per lane, average daily truck traffic, truck percentage, interchanges per mile and weather effects such as snowfall are best modeled as random-parameters—while roadway characteristics such as the number of horizontal curves, number of grade breaks per mile and pavement friction are best modeled as fixed parameters. Our results show that the mixed logit model has considerable promise as a methodological tool in highway safety programming.

The application of the mixed logit model (also called the random parameters logit model) is undertaken by considering injury-severity proportions for individual roadway segments. Severity is defined as the resulting injury level of the most severely injured person in the observed accident. To develop the modeling approach, a severity function determining the pro- portion of injury severities (of all reported accidents per year) on a roadway segment is defined as:

$$S_{in} = \beta_i X_{in} + \varepsilon_{in}$$

where $S_{in}$is a severity function determining the injury-severity category iproportion (property damage only, possible injury, evident injury, disabling injury and fatality) on roadway segment n; X_{in} is a vector of explanatory variables (weather, geometric, pavement, roadside and traffic variables); βiis a vector of estimable parameters; and $\varepsilon_{in}$is error term. If ε_{in}'s are assumed to be generalized extreme value distributed, McFadden (1981) has shown that the multinomial logit model results such that:

$$P_n(i) = \frac{EXP[\beta_i X_{in}]}{\sum_I EXP[\beta_i X_{in}]}$$

Where $P_n(i)$ is the proportion of injury-severity category i (from the set of all injury-severity categories I) on roadway segment n. To generalize this to allow for parameter variations across roadway segments (variations in β), a mixing distribution is introduced giving injury-severity proportions (see Train, 2003):

$$P_n(i) = \int \frac{EXP[\beta_i X_{in}]}{\sum_I EXP[\beta_i X_{in}]} f(\beta/\varphi)d\beta$$

Where f(β|ϕ) is the density function of βwith ϕ referring to a vector of parameters of the density function (mean and variance), and all other terms are as previously defined. Eq. (3) is the formulation for the mixed logit model. For model estimation, βcan now account for segment-specific variations of the effect of X on injury-severity proportions, with the density function f(β|ϕ) used to determine β. Mixed logit proportions are then a weighted average for different values of βacross roadway segments where some elements of the vector βmay be fixed and some may be randomly distributed. If the parameters are random, the mixed logit weights are determined by the density function f(β|ϕ). Most studies have used a continuous form of this density function in model estimation (such as a normal distribution). The difficulties associated with modeling accident-injury severities have led many highway agencies to focus on safety-improvement programs that deal primarily with accident frequency. Where efforts have addressed accident-injury severity, the approach has generally been to identify locations that have an abnormally high number of fatalities. The ability to understand and address the accident injury-severity potential in a multivariate context (understanding how multiple factors affect injury-severity distributions) is a priority for many transportation agencies. The modeling approach presented in this paper offers methodological flexibility that can be used as a basis for safety programs to move beyond simple accident-frequency and observed-fatality approaches. By using a combination of frequency models and the proposed mixed logit model to determine severity proportions, agencies can gain a much better understanding of the effect that possible safety enhancements will have on overall roadway safety.

Kelvin (2004) has conducted a study using single-vehicle traffic accidents occurred in Hong Kong during 1999 to 2000. Several factors considered in the study were district, human, vehicle, safety, environmental and site factors. Stepwise logistic regression models were developed to analyse the effect of these factors. The results shown district board, gender of driver, vehicle age, time of accident and street light conditions as significant factors for private vehicles. The seat-belt usage and weekdays were derived as significant factors for goods vehicles.

Poisson regression model

Cheygye and Ranjitkar(2013) studied and investigate motorway safety by developing accident prediction models that link accident frequencies to their non-behavioral contributing factors, including traffic conditions, geometric and operational characteristics of road, and weather conditions. They used a sample of accidents occurred from 2004 through 2010 on a 74 km long section of auckland motorway. A number of accidents prediction model were developed and assessed for their predictive ability using negative binomial regression model under three categories: first for the whole of the motorway, second for rural and unban motorway segments separately.

In this modeling form, let y_i be random variable that represents the number of accidents occuring at a given motorway segment I during a given time intervel, where i=1,2…..,n, and y_i is a non-negative integer, in a poisson regression model, y_i follows the possion probability law, which takes the following form:

$$P(y_i) = \frac{e^{(-\lambda)}\lambda_i^{y_i}}{y_i!}$$

Where $P(y_i)$ is the probability of segment i experiencing y_i accidents over one year and λ_i is the possion parameter for segment i, which is equal to the expected number of accident per year on segment i, i.e. the mean of accident frequency, $E(y_i)$. the possion regression model commonly assumes the log-linear relationship between possion parameter λ_i and explanatory variables

$$\lambda_i = E(y_i) = e^{\beta X_i}$$

Where X_i is a vector of explanatory variables, such as traffic, road and environmental characteristics of segment i, and β is a vector of unknown regression coefficents which can be estimated by the method of standarad maximum likelihood.

The corresponding likelihood function is:

$$L(\lambda_i) = \prod \frac{\Gamma((1/\alpha) + y_i!)}{\Gamma(1/\alpha)\, y_i!} \left(\frac{1/\alpha}{(1/\alpha) + \lambda_i}\right)^{1/a} \left(\frac{\lambda_i}{(1/\alpha) + \lambda_i}\right)^{\lambda_i}$$

The function is maximized to obtain coefficient estimates for β and α.

A number of negative binomial regression models are developed to estimate accident frequancy as a function of traffic conditions, geometric and operational characteristics of road, and weather condition. The functional form is transformed into an explict strucure as follow:

$$E(y_i) = e^{\beta_0 + LnL + \beta_1 \times LnAADTperlane + \sum \beta_i \times x_i}$$

Where $E(y_i)$ = expected accident frequency in segment i,

L = segment length,
β_0 = intercept,
β_1 and β_i = model coefficents, and
x_i = independent variables in addition to L and LnAADTperlane.

They gave result that the safetly impacts of different non-behavioral contributing factors, in which segments length, AADT per lane and the number of lanes always have the most profound effects on accidents frequency. The validation tools were applied to examine the ability of models to predict accidents.

Motorcyclists are the most crash-prone road-user group in many Asian countries including India. Statistics of accident on Indian roads reveals that motorcycles accounted for the highest share in total road accidents in 2011. They examined the effect of road geometry and traffic variables on motorcycle crashes using a statistical technique called as zero inflated negative binomial regression. The independent variables selected fo their study includes access density (AD) , annual average daily traffic (AADT) , heavy vehicle percentage (HVPER) , speed variation from model speed (VARMSP) , standard deviation of speed (STDSP), and shoulder width deficiency (SWDEF). Accident per year per km (ACCR) is taken as dependent variable. They collected accident data between 2005-09 over a stretch of 100 km of road length which was used for modeling.

Model to predict frequancy of zero accidents:

$$P(y=0) = \sum * [p_0 + (1-p_0) * \left(\frac{1}{1+\alpha * \lambda_i}\right)^{\frac{1}{\alpha}}$$

Model to predict frequancy of accidents other than zero:

$$P(y) = \sum * [(1-p_0) * \frac{\Gamma\left(\frac{1}{a} + y_i\right)}{\Gamma\left(\frac{1}{a}\right)\Gamma(y_i+1)} \left(\frac{1}{1+\alpha * \lambda_i}\right)^{\frac{1}{a}} \left(\frac{a * \lambda^i}{1+\alpha * \lambda^i}\right)^{y_i}]$$

Where ε is exposure term = total no. of road sections α is overdispersion paramter.

They observed that shoulder width deficiency, percentage of heavy vehicles in traffic and speed variations have significant impact on safety of motorcyclist.Motorcycle accident is greatly influenced by shoulder width deficiency on a particular road section. Reducing the shoulder width deficiency by 1m may reduce the accidents by about 24%. Every 2% increase in heavy vehicle traffic may increase the accidents by about 28%. Similarly speed variation is also a influential factor affecting accidents

Sharma and landge (Sharma et. al 2013A) were developed stochastic regression models using crash data collected during 2005-09 over a stretch of 100 km of road length. Their study presents the research work aiming to correlate the road traffic crash rate with road geometry and traffic characteristics for crashes involving heavy vehicles on national highway number 6(NH-6), one of the busy rural roads in central India. Zero inflated negative binomial (ZINB) regression method has been used to model the occurrence of road traffic crashes by authors. They have been used akaike information criterion (AIC) to measure the relative goodness of fit. The independent variables selected in this study were shoulder width (SW), lane width (LW), access density (AD), spot speed (SS) and annual average daily traffic

(AADT). Their model demonstrates that access density, lane width and shoulder width are important parameters affecting the traffic safety of the selected highway. (Sharma & Landge, 2013(B))

The negative binomial regression is extension of Poisson regression and frequently used to predict road accident studies. Ackaah and Salifu (2011) conducted a study using negative binomial regression to predict road accidents for two-lane rural highways. Using three years (2205-2007) crash data on 76 rural highway sections with traffic flow and road geometry, accident prediction model has been developed. The significant parameters observed were traffic flow, road segment length, junction density, type of terrain, and village settlements around road.

Structural equation model(SEM) in transportation field

Hamdar et.al (2008) developed a quantitative intersection aggressiveness propensity index (API). The index was intended to capture the overall propensity for aggressive driving to be experienced at a given signalized intersection. The index was a latent quantity that could be estimated from observed environmental, situational and driving behaviour variables using SEM techniques. The exogenous variables were number of heavy vehicles, number of pedestrians, traffic volume, average queue length, percent grade, number of lanes, number of left turn lanes and so forth. In addition, SEM is frequently adopted in travel value and behaviour field.

ChungandAhn (2002) developed SEM that presented relationships among socio-demographics, activity participation (i.e., time use), and travel behaviour for each day during a week in a developing country. It was tentatively concluded that there were similar relationships between socio-demographics and travel behaviours in developing and developed countries. It was also confirmed that activity patterns were significantly different on weekdays and weekends. Furthermore, during weekdays there were day-to-day variations in the patterns of activity participation and travel behaviour.

Chung and Lee (2002) constructed an SEM to estimate aggregated automobile demand with data from Korea. The results indicated that both the number of driver’s license holders and total road length had a statistically significant effect on automobile demands. In addition, several other determinants of the endogenous variables were found such as average household size, economically active population, personal transportation expenditure, urbanized area, and population density.

Lu and Pas (1999) described the development, estimation and interpretation of a model relating socio-demographics, activity participation (time use) and travel behaviour. Activity participation (time allocated to a number of activity types) and travel behaviour were endogenous to the model. They reported the relationships between in-home and out-of-home activity participation and travel behaviour.

Introduction of SEM

Byrne (2016) stated that Structural Equation Modeling (SEM) is a statistical methodology that takes a confirmatory approach to analysis of the structural theory.SEM uses various types of models to depict relationships among observed variables, with providing a quantitative test of a theoretical model hypothesized by the researcher.Juan et. al stated that SEM methodology is a powerful multivariate analysis technique allowing the modeling of a phenomenon in which a set of relationships between observed and unobserved variables are established .

Two goals in SEM are,

1. To understand the patterns of correlation/covariance among a set of variables and
2. To explain as much of their variance as possible with the model specified (Kline, 1998)

General statistical modeling technique used to establish relationships among variables.

In a confirmatory techniqueTests models that are conceptually derived.Combination of factor analysis and multiple regression can simultaneously test measurement and structural relationships. A structural equations model (SEM) is adopted to capture the complex relationships among variables because the model can handle complex relationship among endogenous and exogenous variables simultaneously and furthermore, it can include latent variables in the models. SEM allows for multiple dependent variables, while linear regression allows only single dependent variable.SEM allows variables to correlate, whereas regression adjusts for other variables in model.SEM accounts for measurement error, whereas regression assumes perfect measurement.

SEM in traffic accident

Wang and Qin conducted a research on Use of Structural Equation Modeling to Measure Severity of Single-Vehicle Crashes and they said that Injury severity and vehicle damage are two of the main indicators of the level of crash severity. Other factors, such as driver characteristics, roadway conditions, highway geometry, environmental factors, vehicle type, and roadside objects, may also be directly or indirectly related to crash severity. All these factors interact in such complicated ways that it is often difficult to identify their interrelationships. The aim of this study was to examine the relationships between these contributors and the severity of single-vehicle crashes. In this study, the number of latent variables were defined by the understanding of collision force, kinetic energy, and mechanical process of a collision, as well as statistical goodness of fit that was based on available data. Three SEM models (one with one latent variable, one with two, and one with three) representing the hypothesized relationships between collision force, speed of a vehicle, and severity of a crash were developed and evaluated in an attempt to unravel the relationships between exogenous factors and severity of single-vehicle crashes. On the basis of goodness of fit and model predictive power, the model with two latent variables outperformed the other two. Additional insights about model selection were provided through the development and comparison of the three models.

The comparison shows that the SEM model with two latent variables (speed and force) had the best statistical goodness of fit and the most statistically significant variables at a significance level of 5%. The SEM results revealed that vehicle speed can positively influence collision force, and both vehicle speed and collision force can significantly increase injury severity and vehicle damage. Males are more likely to drive faster than females, and older drivers tend to drive slower than younger drivers, although this variable is not significant in any of the three at the 5% level. When compared with normal roadway conditions, adverse surface and lighting characteristics decrease both injury severity and vehicle damage because vehicle speed is reduced. In addition, the crash severity of heavy vehicles may be decreased because of their slower traveling speed, but it can also be increased because of vehicle weight. The authors anticipate that the results of this study can unravel complex relationships between injury severity, vehicle damage, and contributing factors via different SEMs and offer additional insights about the model choices for safety analysis.

Lee et. al (2008) conducted a research on analysis of traffic accident size for Korean highway using structural equation models and they said that Accident size can be expressed as the number of involved vehicles, the number of damaged vehicles, the number of deaths and/or the number of injured. Accident size is the one of the important indices to measure the level of safety of transportation facilities. Factors such as road geometric condition, driver characteristic and vehicle type may be related to traffic accident size. However, all these factors interact in complicate ways so that the interrelationships among the variables are not easily identified. A structural equation model is adopted to capture the complex relationships among variables because the model can handle complex relationships among endogenous and exogenous variables simultaneously and furthermore it can include latent variables in the model. In this study, they used 2649 accident data occurred on highways in Korea and estimate relationship among exogenous factors and traffic accident size. The model suggests that road factors, driver factors and environment factors are strongly related to the accident size.

In this study, they used 2649 accident data occurred on highways of Korea and estimate relationships among exogenous factors and traffic accident size. In modeling process, they were create exogenous latent variables such as “road factors”, “driver factors” and “environment factors” to identify latent relationships to an endogenous variable “accident size”. While vehicle factors such as the number of involved vehicles and the number of damaged vehicles in accidents can describe some aspects of accidents, unexplained aspects of accidents still exist. Hence, a new statistic “accident size” was adopted in this study, which could be described in terms of the number of deaths and injured persons as well as the number of damaged vehicles and the number of vehicles involved in accidents.

The data used in this study were 2880 complete accident records during the year 2005, which are collected by Korean Expressway Corporation. Each accident record has various and rich information such as the accident location (where the accident took place), pavement type, horizontal alignment, vertical alignment, vehicle type, driver’s gender, driver’s age, road surface condition, the day (week end or weekday), weather condition, day or night time, the number of deaths, the number of injured persons, the number of involved vehicles, and the number of damaged vehicles. After eliminating missing and erroneous data, 2649 accident data are utilized in this search.

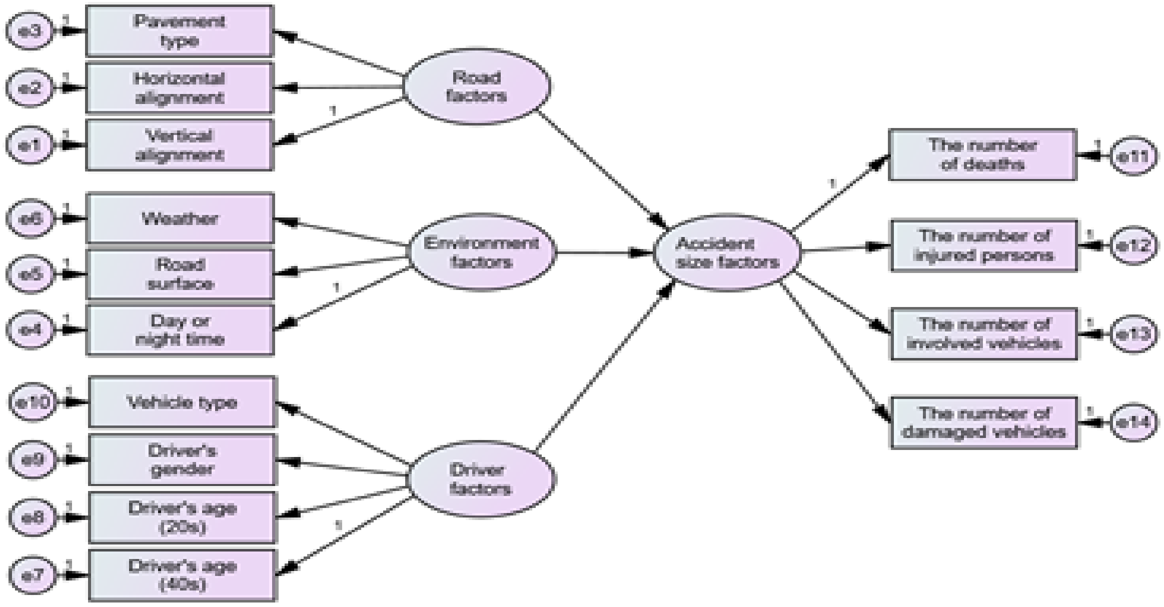
In their initial model, 15 observed variables are used and they are split to four groups: road, environment, driver and accident size group. Road group includes the accident location, pavement

Type, horizontal and vertical alignment characteristics. Environment group has road surface condition, weekends or weekdays, weather condition and daytime or night time. The driver group consists of vehicle type, driver’s gender and their age. Accident size group is explained by the number of deaths, the number of injured persons, the number of vehicles involved and the number of damaged vehicles.

The final model specification is derived using a two-stage development process. At the first stage, we conduct factor analysis to classify observed variables into several groups. Factor analysis is often used to analyse the correlations among several variables in order to estimate and to describe the number of fundamental dimensions that underlie the observed data. Those fundamental dimensions (factors) can be latent variables in SEM. At the second stage, authors estimated the polychoric correlations matrix of observed variables and finally develop a SEM having the best-fit statistic.

WLS estimation method is employed because distributions of variables do not have multivariate normality. In the figure, the numbers in the arrows are parameters estimated and standard error and t-value and they conclude that the SEM illustrates positive or negative effects of each variable on the accident size. According to the SEM model, the total effect of road factors on accident size is higher, so that accident size tends to increase when road factors have higher values. Road factors increase in case of pavement of concrete, straight and level/downward slope. The operating speed is slower in poor (curve or upward slope) section of roads.

Figure 1: Final structural equation model of traffic accident size



According to the Model the estimated coefficient of environment factors is a positive value. This result indicates that poor weather and wet road surface contribute to decrease accident size. In case of driver factors auto vehicle and female drivers contribute to decrease accident size. For the analysis purpose the estimated coefficient are all taken as standardized solutions, so they can conclude that the major factors influencing on the accident size is road factors. These results do not mean that the poor sections (curve or slope) and environments (nasty weather or night time) decrease accidents. Said in another way, obviously under poor designed sections and environment conditions, accident occurs more frequently. However, once accident occurred in straight sections, the accident size is higher than curve or up-slope ones. Factors increasing the accident size are related to operating speed and driver's carelessness. Among three exogenous latent variables (road, environment and driver factors), the effect of road factor on accident size is highest. In order to decrease the traffic accident size handling the road factor is more effective than handling driver and environment factors. It can be a positive result to traffic engineers because as they can handle 'road factors', they hardly manage 'driver and environment factors'.

Validation of model

Model fit determines the degree to which the sample variance–covariance data fit the structural equation model. Model-fit criteria commonly used are chi-square (*χ2),* the goodness-of-fit index (GFI), the adjusted goodness-of-fit index (AGFI), and the root-mean-square residual index (RMR).These criteria are based on differences between the observed (original, *S)* and model-implied (reproduced, Σ) variance–covariance matrices.A significant χ2 value relative to the degrees of freedom indicates that the observed and implied variance–covariance matrices differ.Statistical significance indicates the probability that this difference is due to sampling variation.The chi-square test of model fit can lead to erroneous conclusions regarding analysis outcomes. (Byrne, 2016)

Chi-Square (*χ2*)

The *χ2 model-fit criterion is sensitive to sample* size because as sample size increases (generally above 200), the *χ2 statistic* has a tendency to indicate a significant probability level. In contrast, as sample size decreases (generally below 100), the *χ2* statistic indicates non-significantprobability levels. The chi-square statistic is therefore affected by sample size

Goodness of fit (GFI), Adjusted Goodness of fit (AGFI)

The goodness-of-fit index (GFI) is based on the ratio of the sum of the squared differences between the observed and reproduced matrices to the observed variances, thus allowing for scale. The GFI measures the amount of variance and covariance in *S that is predicted by the reproduced* matrix Σ. Value vary from 0 (poor fit) to 1 (perfect fit)The adjusted goodness-of-fit index (AGFI) is adjusted for the degrees of freedom of a model relative to the number of variables.The GFI and AGFI indices can be used to compare the fit of two different models with the same data or compare the fit of a single model using different data, such as separate datasets for males and females, for example, or examine measurement invariance in group models. The Value vary from 0 (poor fit) to 1 (perfect fit)

Root Mean Residual (RMR)

The RMR index uses the square root of the mean-squared differences between matrix elements in *S and Σ.* Because it has no defined acceptable level, it is best used to compare the fit of two different models with the same data. Value vary with the sample size.

Akaike Information Criterion (AIC)

The AIC measure is used to compare models with differing numbers of latent variables.The first AIC is positive and the second AIC is negative, but either AIC value close to zero indicates a more parsimonious model. Compare value of alternate model.

Normed Fit Index (NFI) and Comparative Fit Index (CFI)

The NFI is a measure that rescales chi-square into a 0 (no fit) to 1.0 (perfect fit) range (Bentler&Bonett, 1980). It is used to compare a restricted model with a full model using a baseline null model

The comparative fit index (CFI) measures the improvement in no centrality in going from model MltoMk(the theoretical model*)*

RMSEA: (Root Mean Square Error of Approximation)

MacCallum, Browne, and Sugawara (1996) provided a different approach to testing model-fit using the root mean square error of approximation (RMSEA).Their approach also emphasized confidence intervals around RMSEA, rather than a single point estimate, so they suggested null and alternative values for RMSEA (exact fit: Ho =0.00 versus Ha =0.05; Close fit: Ho= 0.05 versus Ha =0.08; and Not close fit: Ho = 0.05 versus Ha = 0.10); researchers can also select their own. The MacCallum et al. (1996) method tests power, given exact fit (Ho; RMSEA= 0), close fit (Ho, RMSEA ≤ 0.05), or not close fit (Ho, RMSEA ≥ 0.05).

Conclusion

Structural Equation Modeling is a flexible and powerful statistical methodology used to examine the relationships between measured variables and latent constructs.

In SEM we can create exogenous latent variables such as “road factors”, “driver factors” and “environment factors” to identify latent relationships to an endogenous variable “accident size”.

There are differences and similarities between “traditional” statistical techniques and SEM. With SEM techniques, models are specified a priori, measurement error is specified explicitly, and models are tested for acceptable fit with chi-square and several fit indices. SEM gives you the power not available with “traditional” statistical procedures. You are challenged to design and plan research where SEM is an appropriate analysis tool.

References

Ackaah, W., & Salifu, M. (2011). Crash prediction model for two lane rural highways in Ashanti region of Ghana. International Association of Traffic and Safety Sciences(IATSS) Research 35, 34-40.

Byrne, B. M. (2016). Structural equation modeling with AMOS. New York: Routledge.

Chengye, P., &Ranjitkar, P. (2013). Modeling Motorway Accidents using Negative Binomial Regression. Proceedings of the Eastern Asia Society for Transportation Studies, Vol.9.

Chung, J. H., & Ahn., Y. (2002). Structural equation models of day-to-day activity participation and travel behaviour in a developing country. Transportation Research Record 1807, 109-118.

Chung, J. H., & Lee, D. (2002). Structural model of automobile demand in Korea. Transportation Research Record 1807, 87-91.

Hamdar, S. H., Mahmassani, H. S., & Chen, R. B. (2008). Aggressiveness propensity index for driving behavior at signalized intersections. Accident Analysis and Prevention 40 (1), 315–326.

Kelvin, Y. K. (2004). Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong. Accident Analysis and Prevention 36, 333–340.

Kline, R. B. (1998). Software Review: Software Programs for Structural Equation Modeling: Amos, EQS, and LISREL. Journal of Psychoeducational Assessment, 343-364.

Lee, J.-Y., Chung, J.-H., & Son, B. (2008). Analysis of traffic accident size for korean highway using structural equation models. Accident analysis and Prevention 40, 1955-1963.

Lu, X., & Pas, E. I. (1999). Socio-demographics activity participation and travel behavior. Transportation Research Part A 33, 1-18.

Milton, J. C., Shankar, V., & Mannering, F. L. (2008). Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accident Analysis and Prevention 40 (1), 260–266.

MoRT&H. (2016). Road Accident in India.

RTO, G. (2017). Commissionerate of Transport. Retrieved from Department of Port and Transport, Government of Gujarat: http://rtogujarat.gov.in

Sharma, A. K., & Landge, V. S. (2013(B)). Zero Inflated Negative Binomial for Modeling heavy Vehicle crash rate on Indian rural highway. International Journal of Advances in Engineering & Technology, 292 - 301.

Sharma, A. K., Landge, V. S., & Deshpande, N. V. (2013(A)). Modeling Motorcycle Accident on Rural Highways. International Journal of Chemical, Environmental & Biological Sciences(IJCEBS), 313-317.

WHO, (2018), Global status report on road safety 2018, World Health Organization . Retrieved from www.who.int/violence_injury_prevention/road_safety_status/2018/en/