Factors Affecting the Frequency of Road Accidents: A Case Study on Saudi Arabia

香港公開大學

THE OPEN UNIVERSITY OF HONG KONG

BSc (Hons) in Data Science
STAT S461F
Data Science Project 2020-2021

By

WONG, Ho Sum	**12311942**
LAI, Tsun Wai	**12332716**
HO, Wai Lok	**11933364**
TSUI, Tak Yu	**12325817**

Supervised By
Dr. Tahir Mahmood

Department of Technology
School of Science and Technology
The Open University of Hong Kong

Table of Contents

Abstract 4

Chapter 1 5

1.1 Introduction 5

1.2 Aim and Objective 6

1.3 Overview of Report 6

Chapter 2 7

Literature Review 7

Chapter 3 21

Methodologies 21

3.1 Data Pre-processing Procedure 21

3.2 Poisson Distribution 22

3.3 Poisson Regression 22

3.4 Negative Binomial Distribution 23

3.5 Negative Binomial Regression 23

3.6 COM-Poisson Distribution 24

3.7 COM-Poisson Regression with GLM 26

3.8 Performance Measures 27

3.8.1 Log-likelihood 27

3.8.2 Akaike Information Criterion (AIC) 27

3.8.3 Bayesian Information Criterion (BIC) 28

Chapter 4 29

4.1 Result and Discussion 29

4.2 Data Description 29

4.3 Univariate Data Description 31

4.4 Bivariate Data Description 37

4.5 Regression Analysis 41

4.5.1 Distribution Fitting 41

4.5.2 All Possible Regression 43

4.5.3 Model Selection 45

Chapter 5 47

Conclusion 47

Recommendation, Limitation and Future Studies 48

Appendix 50

Reference 64

Abstract

In this project, the aim and objective is to identify the influential factors that leading to the frequencies of card accident. A Saudi Arabia crash dataset from 2016 to 2019 measured form Ministry of Transport are used under study. We did a comprehensive literature review to study how the researchers performed the crash data modelling in the road safety assessment. After that, we adopted Poisson Regression, Negative Binomial Regression, COM-Poisson Regression as our candidate models. Also, we introduced the methodologies of the data pre-processing procedure, regression techniques and the corresponding probability distributions, as well as the performance measures (AIC, BIC, Log-likelihood). Besides, the data description, descriptive and explanatory analyses are provided. In the regression analysis, we performed distribution fitting to find the best fitting distribution of the response variable which is the number of daily car accidents. We found the response variable best fits to COM-Poisson distribution and selected COM-Poisson Regression as our regression technique. Then, we performed All Possible Regression, and two models are selected. The one covariate model indicates the explanatory variable road geometric details having the best explanatory effect. The full model contains at least one explanatory variable from different categories which are informative sign, roadway factors, and weather factors. Therefore, informative sign roadway factors, and weather factors are the influential factors leading to the frequencies of car accidents. Finally, recommendation, limitation and future studies are provided.

Key Words: Saudi Arabia, road accidents, crash data modelling, influential factors

Chapter 1

1.1 Introduction

Nowadays, road accident is a very common type of accident which is happening around the world, especially in Africa and Middle East. It is a very serious problem that kills and injuries many people every year. The ARAB news (2017) indicated that in 2016, over 9000 people were killed by car accident in Saudi Arabia. It represented that on average, over 25 people died due to car accident every day. Besides, according to the World Life Expectancy (2018), it collected data from World Health Organization in 2018. It could be found that the countries of Africa monopolized the first place to thirty first. Also, Saudi Arabia is on the thirty second, placing the first place in the Middle East. It reveals that Africa and Saudi Arabia has a severe road traffic accident problem. It must have an improvement to reduce the frequencies of traffic accident. There are many researchers aim to enhance the road safety monitoring by finding out the significant factors that leading to the car accidents. Similarly, in our research, we aim at exploring the influential factors leading to road accident in Saudi Arabia. Therefore, we can give some insight and suggestions to the authorities in terms of the roadway design, maintenance, and monitoring.

Many road safety assessment researches noted that traditional Multiple Linear Regression is insufficient to model the crash data due to its lack of discrete assumption on the dependent variable and dissatisfying performance. For crash data modelling, the excess zero problem and the data dispersion is the major consideration. To tackle the problem, there should be some reliable statistical methods to test the existence of excess zero problem and the data dispersion (under-dispersion, equi-dispersion or over-dispersion). For the manipulation of the excess zero problem, different zero-inflated models should be adopted, for example Zero-inflated Poisson model and Zero-inflated Negative Binomial model. For the manipulation of data dispersion, Poisson model, Negative Binomial, and COM-Poisson are usually used to model the equi-dispersed and over-dispersed count data, respectively. In few words, these two problems will be the challenges for crash data modelling in the research.

1.2 Aim and Objective

In this project, the aim and objective are to identify the influential factors that leading the frequencies of car accidents.

1.3 Overview of Report

This report consists of five chapters. Firstly, Chapter 1 states the introduction, aims and objectives of the project. Chapter 2 is the literature review, mainly focusing on the road safety assessment researches. Chapter 3 mentions the methodologies which include the data pre-processing procedures, the candidate regression techniques and corresponding probability distributions. Chapter 4 presents a detailed data description, descriptive and explanatory analysis as well as the regression analysis. Lastly, chapter 5 provides some conclusion, recommendation, limitations and future studies.

Chapter 2

Literature Review

Traditionally, most researchers are using Poisson and Negative Binomial (NB) Regression to model crash data. Based on different cases, such as the vehicle type, another model may have better performance than them. Joshua and Garber (1990) estimated multiple linear and Poisson Regression to identify the characteristics and the contribution of the traffic and roadway geometric variables in large truck accidents. The crash data is all accidents on Virginia highways between 1984 and 1986, reported by the Virginia Department of Transportation. The roadways from the sites were further divided into three environments (subgroups) with respect to the number of the lane(s). From the results, Poisson Regression models sufficiently explain the relationship between the large truck accidents and the traffic and roadway geometric variables. Furthermore, they had investigated the significant variable on traffic and roadway geometric factors on each environment analysis.

Miaou (1994) considers the three regression models in his study: Poisson regression, zero-inflated Poisson (ZIP) regression, and NB regression to study the relationship between truck accidents and the geometric design of road sections. Comparing the above models, the study uses the maximum likelihood (ML) method to estimate based on the regression parameters. In response to the degree of dispersion, the researcher chose the best model for this paper. And Maher and Summersgill had found the reason why multiple linear regression models do not appropriate in accident modeling. Maher and Summersgill (1996) reported the process of continuing TRL (Transport Research Laboratory) junction accident studies by giving examples of the types of models which have been developed, with the effect of design on junction safety. They found that GLMs with Poisson error structure are far more appropriate than conventional multiple linear regression models, based on least squares. They also confirmed the advantages of using GLMs with Poisson error structure from other authors' findings. Furthermore, they extended their findings by describing the use of the Poisson and NB models have been successfully overcome various technical problem. They found that GLMs will produce robust and reliable results with some considerations like low mean value problem, over dispersion, and data disaggregation over time.

Similarly, Poch and Mannering (1996) estimated a Negative Binomial regression of the frequency of accidents at the intersection between geometric and traffic-related elements. The dataset used is seven-year accident histories from 63 intersections in Bellevue, Washington. They found Negative Binomial regression modeling with intersection approach accidents to identify significant traffic and geometric elements that increase or decrease accident frequency. First, this knowledge can be applied during the development of new intersection construction or intersection rebuilding programs. Second, it is useful for the determination of high intersection accident location reviews.

In the late 1990s, most of the researcher are using NB and Poisson in accidents data modeling. But some of the researchers still developed other regression models. Anderson, Bauer, Harwood, and Fitzpatrick (1999) developed Poisson, Negative Binomial, and Lognormal regression to assess the significance of five geometric variables in the crash prediction for rural two-lane highway alignments. They are speed reduction on a horizontal curve relative to the preceding tangent or curve, average radius, a maximum radius to the minimum radius, the average rate of vertical curvature, and the ratio of individual curve radius to the average radius. It is found that all of them are statistically significant as expected. Therefore, they are suggested as the candidate measures for the evaluation of the geometric design.

In the conclusion of the 1990s, most of the researchers started using Negative Binomial and Poisson Regressions. Some of them tried to consider other regression models, but Negative Binomial and Poisson Regression usually have the best performance in accident data modeling.

In the 2000s, the researchers started using NB regression to identify the important variable. Abdel-Aty and Radwan (2000) estimated Negative Binomial regression to predict the accident frequencies in State Road 50 (which supposed to be homogenous among segments) in Central Florida, from 1992 to 1994, with traffic and roadway variables. First, they identify the important variables in the model. Second, the research takes account of the demographic characteristics of the drivers. It has been shown that female drivers experience accident more frequently under heavy traffic volume, reduced median width, narrow lane width, and the larger number of lanes. Male drivers tend to involve in accidents while speeding. Younger and old drivers have the same problem under heavy traffic volume and reduced shoulder and median widths. Besides, Wood (2002) examined the goodness of fit testing on generalized linear accident models (GLMs). Given the fact of the low mean value problem on the goodness of fit testing for GLMs, he presented a resolution and demonstrated the resolution on a dataset of injury accidents on 392 approaches to intersections in New Zealand from 1980 to 1991. At the computational level and theoretical level, the resolution involves a practical grouping technique and recognition, respectively. Furthermore, he introduced a self-contained description of GLMs with a first-principles explanation of the iterative fitting process. The GLMs includes Poisson and Negative Binomial models.

Some of the data may not occur every day, so it is an issue we need to consider in crash data with daily bases modeling. Kumara and Chin (2003) revealed that the accident might not occur every day; sometimes it may be zero cases per day. Therefore, it may have a lot of day showing that it is zero. In regard to this problem, the researchers use zero-inflated negative binominal to assign the probability to the accident outcome. The researchers use Singapore as an example to show the result.

For some count data modeling in crash data, usually, they will use COM-Poisson regression because of the overdispersion and underdispersion problem. Famoye, Wulu, and Singh (2004) found that the generalized Poisson regression has the best performance base on the test for the dispersion and the goodness of fit to the crash data. And the research target is drivers aged 65 or above. Similarly, Shmueli, Minka, Kadane, Borle, and Boatwright (2005) introduced COM-Poisson as a flexible distribution that can account for overdispersion or underdispersion. They described three methods for estimating the parameters of the COM-Poisson distribution. They are fast simple weighted least squares method, maximum likelihood, and Bayesian. They also explored two sets of real world data (the first consists of quarterly sales of a well-known brand of a particular articleof clothing at stores of a large national retailer, the second is lengths of words in a Hungarian dictionary. Even at the quarterly level, sales of a particular) that do no follow the Poisson distribution and evaluate the flexibility and elegance of the COM-Poisson distribution. Note that the first and second dataset are over-dispersed and under-dispersed, respectively.

Some researcher still considers other regression for modeling crash data. Lord, Washington, and Ivan (2005) found that zero-inflated Poisson (ZIP) and Negative Binomial models (ZINB) have better performance of excess zeros frequently observed in crash count data by comparing Binomial, Poisson, NB, ZIP and ZINB. Chang (2005) revealed that Poisson and Negative Binomial had been used to analyzed vehicles accidents for the past few decades. This analysis mainly compares that the Negative Binomial and artificial neural network (ANN) method. The difference between the two methods is that whether pre-defined underlying relationship between dependent and independent variable. 1997–1998 accident data for the National Freeway 1 in Taiwan has chosen by the authors to analyze. By the comparison of prediction performance, ANN has been chosen to demonstrate.

The dispersion parameter is an important parameter for modeling . Lord (2006) found that when the data exhibit over dispersion Poisson-gamma model is a very common choice for transportation safety modeling. And the result shows that a low mean problem (LMP) combined with a small sample size can gravely affect the estimation of the dispersion parameter. So, if the sample size and the sample mean decrease, the dispersion parameter's probability will become more unreliable. El-Basyouny and Sayed (2006) conducted a comparative analysis between the traditional Negative Binomial (TNB) and the modified Negative Binomial (MNB). The research used a dataset with crash data (including traffic volume and geometric factors) from 382 segments in Vancouver and Richmond, British Columbia, Canada. Both models perform well while MNB fits the data better. Also, it is more beneficial to investigate the accident-prone with the use of MNB. Moreover, given the high significance of the parameters estimated of the variance function from MNB model, the variability of the dispersion parameter is justified. Also, 100 locations of accident-prone are identified by two models. They found the investigation of the accident-prone depending on the existence of deviant sites.

Lord, Washington, and Ivan (2007) wrote additional points that were not covered in the previous paper (Lord et al., 2005) and showed zero-inflated models (ZIs) for modeling highway safety data is problematic, ranging from the statistical fit fallacy to logic problems. The additional information focuses on the rationality of the application of ZIs. ZI should not be used on highway entities' crashes modeling, especially when other alternatives (for example, small area statistics and extreme value models) can provide better single-state analysis performances. Besides, Xie, Lord, and Zhang (2007) evaluated the Bayesian neural network application (BNN) on motor vehicle crashes prediction. They had a comparative analysis on the performance between back-propagation neural network (BPNN), BNN, and Negative Binomial (NB), using crash data from 88 highway sections on rural two-lane frontage roads, Texas. The mean absolute deviation (MAD) and mean squared prediction error (MSPE) are used as the performance metrics throughout the analysis. From the results, NB provides an inferior performance than BPNN and BNN. BPNN sometimes provides an excellent prediction (approximately the same as the actual value), while its average performance is inferior to BNN's. This indicates BNN has better non-linear approximation and generalization abilities than BPNN. Furthermore, they carried out a sensitivity analysis to prove that neural networks can contribute to developing accident modification factors (AMFs). Also, they proposed to widely apply BNNs to the crash dataset with small sample mean values and sample size.

Yuanchang and Yunlong (2008) have proposed the recent crash frequency primarily based on generalized linear models. It has assumed that the logarithm of expected crash frequency and other explanatory variable is a linear relationship. But there are some explanatory variables; such a linear assumption may be invalid. Thus, the authors use a Negative Binomial generalized additive model to compare the Negative Binomial generalized linear model. Finally, they choose the Negative Binomial generalized additive model by comparison. On the other side, the report from Montella, Colantuoni, and Lamberti (2008)is finding the best prediction models for total crashes and severe crashes to check whether the variables have a relationship with road safety. A general linear regression and negative-binomial regression will be used to fit the model. Data collected from 2001 to 2005 on Motorway A16 between Naples and Canosa in Italy. The variables that will be considered have been filtered by a stepwise-forward procedure such as traffic volume and composition, vertical alignment etc. Based on the stepwise procedure, two prediction models have been formed. Montella et al. (2008) indicated that the two models are very similar. The same sign, such as operating speed reduction, year effect, and length of the tangent preceding the curve, positively relates to road safety.

Meanwhile, X. Li, Lord, Zhang, and Xie (2008) evaluated the Support Vector Machine (SVM) application in vehicle crashes prediction. They conducted a comparative analysis between SVM, the traditional Negative Binomial (NB) model and the Bayesian neural network

(BPNN) model, which is developed from the previous study (Xie et al., 2007). The models are developed from a dataset from rural frontage roads in Texas. The results show that SVM does not have the over-fitting problem and performs better and similar to NB and BPNN. Also, they suggested implementing SVM for fast implementation reason, if the sole objective of the study is to predict the crash frequencies.

Park and Lord (2009) estimated seven models with the finite mixtures of Poisson or Negative Binomial (NB) regression models, using vehicle crash data on signalized 4-legged intersections in Toronto Ont. They found the standard NB and one of the NB mixtures models performed similarly. Also, the NB mixtures model's parameter estimate captured some significant characteristics from the data that the standard NB model does not capture. Furthermore, they had evidence to indicate the dataset is generated from two distinct sub-populations, where each population has its own coefficients and degrees of over-dispersion. Also, they proposed that the standard NB model can provide an insufficient fit if the data were generated by multiple component finite mixtures of Poisson regression models. Apart from the NB model, Malyshkina, Mannering, and Tarko (2009) developed two-state (fewer frequencies and more frequencies states) Markov switching models to predict the roadway accident frequencies on interstate highway segments, Indiana, from 1995 to 1999. Bayesian inference methods and Markov Chain Monte Carlo simulations are used for model estimation. In the 2000s, most of the researcher started to compare NB with another model such as BPNN, SVM and ZIs. And NB still has a good performance in modeling crash data. Therefore, there are some research is comparing traditional NB and GLM-based NB.

Sellers and Shmueli (2010)found that Poisson regression is very useful for handling count data, but real-life data often under dispersed and overdispersed and not conductive to Poisson regression. And COM-Poisson regression generalizes Poisson and logistic regression model, and it can also fit for count data with under dispersed and overdispersed. On the other side, the GLM approach takes advantage of exponential family properties, so COM-Poisson regression with GLM base will better than traditional Poisson regression. Besides, the report from Rakha, Arafeh, Abdel-Salam, Guo, and Flintsch (2010) investigated a linear regression model that is satisfied to predict car crash. Data background is collected from Highway Traffic Records Information System Crash Database for year 2001 to 2005, and it contains 186 road sections in Virginia. Rakha et al. (2010)indicated that a least-square LRM approach could use to predict car crash data with some conclusion. For example, the number should be zero when the exposure is zero when using the exponential function.

Usually, crash data will have an overdispersion or underdispersion problem. And some of the researchers try to find out which model can have better handling with this problem in crash data. Lord, Geedipally, and Guikema (2010)used the Conway-Maxwell-Poisson GLM to evaluate the traffic crash data analysis and use the Poisson-gamma model to show over-dispersion. Besides, the researchers' output result shows the signs which is under-dispersion, and they found that the best model for this data set is the COM-Poisson model.

Similarly, Lord and Mannering (2010) had a detailed review of the key issues associated with crash-frequency data and the strengths and weakness of the various methodological approaches. The intent is to provide contemporary thinking in the crash frequency-analysis field and show the evolution of methodological approaches over the years.

The key issues are in terms of data characteristics (over-dispersion, under-dispersion, time-varying explanatory variables, low sample-means and size, crash-type correlation, under-reporting of crashes, omitted-variables bias, and issues related to functional form and fixed parameters). The stream of methodological innovation has introduced some great statistical approaches. Random-parameter models, finite mixture models, Markov switching models and others improve the understanding of the factors that affect the frequency of crashes. Besides, Lord et al. (2010) examined alternative modelling for capturing heterogeneity in crash count models using regression models (the mixtures of Poisson or NB regression models).

These models were compared to the standard NB regression model using the data collected at signalized 4-legged intersections in Toronto, Ontario. The results of this study show that the dataset seemed to be generated from two distinct sub-populations, each having different regression coefficients and degrees of over-dispersion. Also, it is found that the standard NB regression model can be misleading when the data were generated by two or more component finite mixtures of Poisson regression models. Therefore, it is recommended that transportation safety analysts use this type of model before the traditional NB model, especially when they are suspected of belonging to different groups.

In parallel, Haleem, Abdel-Aty, and Santos (2010)discuss in this report found out that the multivariate adaptive regression splines (MARS) technique and Negative Binominal can be fitted and compared with the use of extensive data collected on unsignalized intersections in Florida. MARS could predict crashes almost like the traditional NB models by the studies. The MARS with a technique (random forest) can be explored. Therefore, it is suitable to use for counting accident and examine the relationship between the significant factors and the frequency of accidents. Haleem, Abdel-Aty, and Mackie (2010)estimated a Negative Binomial regression of the frequency of accidents as a crash prediction model as it is over-dispersion. The data from 433 unsignalized intersections in Orange County, Florida, were collected and used in the analysis. It used four Bayesian-structure models to examine a non-informative prior with a log-gamma likelihood function, a non-informative prior with an NB likelihood function, an informative prior with an NB likelihood function, and an informative prior with a log-gamma likelihood function. Also, it has used the Akaike information criterion. (AIC), mean absolute deviance (MAD), mean square prediction error (MSPE) and overall prediction accuracy to compare NB and Bayesian model.

Cafiso, Di Graziano, Di Silvestro, La Cava, and Persaud (2010) had a road safety analysis on a sample of 168.20 km of two-lane rural roads, in Italy, with four groups of variables (exposurem geometric, consistency, and context). The Generalized Linear Modeling technique with a Negative Binomial error structure is utilized to develop 19 models. Three models are selected as the recommended models. The first recommended model contains only the exposure variables. They are available from the network, and it facilitates the network screening analysis. Model 2 and 3 have the best statistical fit and contain at least one variable from each group of the four variables. Thus, it is sufficiently interpretable to evaluate the safety performance and alternative safety improvement policies.

Highway crash data analysis is one of the important topics in transportation analysis, Akın (2011)used the Negative Binomial and Poisson regression model to analyze the highway crash data, which measure the factors that affect the accident severity on the highway. By this report, the analyst had listed out the parameters which affect the crash frequently; the findings on these two models are similar, but the estimations had shown a bit different between the two models.

Srinivas Reddy Geedipally, Dominique Lord and Soma Sekhar Dhavala(Geedipally, Lord, & Dhavala, 2012) found that the Negative Binomial Lindley GLM model can have better performance than the Negative Binomial model and zero-inflated model when the datasets containing a large number of zero and long tail. Besides, it can also have better performance when the crash data are overdispersed. Apart from the modeling performance, some of the research found some important factors affecting road safety. Obaidat and Ramadan (2012) report the main factors that affect traffic accidents in Hazardous location by occurring logarithmic and linear models. Data is collected from Great Amman Municipality, Traffic Institute, Police Traffic Department, and field studies. The author used stepwise regression to find out the best regression model to achieve goals. As a result, Obaidat and Ramadan (2012) claimed that for effecting urban roads, geometrical, behavioural, traffic condition, and environment are the most important things to consider. For hazardous locations, speed, road surface type, traffic properties, degree of curvature, horizontal and vertical curves, lighting conditions should be considered.

Most of the studies had reported that GLM has a better performance compared with the traditional method. Z. Li, Wang, Liu, Bigham, and Ragland (2013) found that GLM is a common technique for crashing data modeling at the county level, but it can not capture the spatial heterogeneity that exists in the relationship between crash count and explanatory variable over counties. And Geographically Weighted Poisson Regression have more useful in capturing the spatially non-stationary relationships at county level crash data between crashes and predicting factor.

On the other side, the report from Mustakim, Abdullah, Yunus, Abdullah, Mat Deris, et al. (2014) formed an accident prediction model for unsignalized junction in Malaysia Rural Roadways. Models are developed by using multiple regression that has a relationship with accident rate, road geometry and traffic situation. After formed the models, Mustakim, Abdullah, Yunus, Abdullah, Deris, et al. (2014) concluded that building road median and opening signalized at hazardous access point will substantially decrease road accident. Besides, increasing the number of motorcycle crossing and speed, reducing the gap will increase the number of accident. Besides, as for the analysis conducted by Gupta, Sim, and Ong (2014), they found out that Conway-Maxwell Poisson Regression can fit a better result than Generalized Poisson Regression. One regression shows the result is over-dispersion, and the other is under-dispersion, so this indicates a conclusion for the researchers to find a better model.

Similarly, the report discussed different regression models that people have used for analyzing traffic accident past through decades and explore their pros and cons, according to the data collected from Global Status Report on Road Safety 2015 and Open Government Data Platform India. Basu and Saha (2017) indicated that some countries have an increasing trend regarding road accident. And considering the characteristics of the different regression model, they formed a table to compare the models. It can be found that more than half models are suitable to analyze over-dispersion data, such as the Negative Binomial and

Conway-Maxwell-Poisson model. Poisson regression cannot handle over and under dispersion data. And Farag and Hashim (2017) utilized Poisson or Negative Binomial Regression with Generalized Linear Modelling technique to evaluate the safety performance of the roundabouts in Oman. They developed crash prediction models using traffic, geometric and crash three years data for 60 approaches from 15 roundabouts from Salalah city, Dhofar governorate, Sultanate of Oman. They found the most significant variables are the circulating width, the 85th percentile speed (at specific approach), the entry angle, the percentage between circulating width and the inscribed diameter with a positive sign.

Shaon, Qin, Shirazi, Lord, and Geedipally (2018) proposed the application of the random parameters (RPNB-L) generalized linear model (GLM) for crash prediction with the implementation of a Negative Binomial Lindley (NB-L) model with varied site coefficients. Two datasets (338 rural interstate roadway segments, in Indiana, from 1995 to 1999; crash data from South Dakota Department of Transportation, 2008 to 2012) with high dispersion and heavy tail are used. The results were compared to Negative Binomial, RPNB, and fixed parameters NB-L models. It is shown that NB-L, RPNB-L outperforms fixed and random parameters NB with Generalized Linear Modeling (GLM). Therefore, both the fixed and random parameters of NB-L GLMs offer a reliable approach to the traditional NB GLMs for over-dispersed crash analysis.

In addition to applying the random parameters (RPNB-L) GLM, the report from Naghawi (2018)discussed the reasons that lead to a severe road crash by using data collected from Jordan Traffic Institute that recorded in 2014. The author used two ways to accomplish the report’s goal: using descriptive analysis to quantify the factors of road crash severity and Negative Binomial regression to predict the factors. After these two steps, Naghawi (2018)indicated few factors that will increase the road crash severity. First, the younger driver may increase the severity of road crash. Second, the road's environment will affect the severity; severer accident may happen in the morning rather than evening and weekday rather than the weekend. Also, a driver or car's faults may make severe crash appear such as a broken brake or old car will increase the probability of road crash. The Negative Binomial regression formed a model with twenty variables that may affect a severe road crash. The author can use a bunch of estimators to find out the relationship. It is helpful when analyzing a situation with many variables. The report from Kamaluddin et al. (2018) produced a traffic safety analysis based on underreporting registered crashes. Data is collected from three databases: ScienceDirect, Scopus and Transport Research International Documentation for a systematic literature search. After analysing self-reported crash studies, Kamaluddin, Andersen, Larsen, Meltofte, and Várhelyi (2018) indicated few things that can improve the bias conclusions. For example, self-reported crashes can be more common in developing countries, reports about crashes can be more focused on pedestrians, motorcyclists etc. Besides, Hou, Tarko, and Meng (2018) conducted research on identifying significant factors of crash frequencies. They developed a random-effects Negative Binomial (RENB), a random parameter Negative Binomial (RPNB) and a Negative Binomial (NB) model to predict the crash frequencies of eight freeways from Liaoning Province and Guangdong Province, China. The freeways data are mainly from 2008 to 2012. In objective to identify some new factors (for example freeway interchanges and service area segments), RENB and RPNB, with random effects and random parameters were utilized and compared. They found since RENB and RPNB can consider the unobserved heterogeneity across groups or observations, they outperform NB model. Furthermore, the RPNB performs slightly superior to RENB in the goodness-of-fit testing. Also, a comprehensive discussion on the significance on group of factors (exposure, traffic conditions, freeway design, pavement conditions, and weather conditions) was made.

In addition, to use Poisson or NB to modeling crash data, AlKheder and Al-Rashidi (2020) proposed a Bayesian hierarchical model for forecasting the crash frequencies at 143 road sites in Abu Dhabi. They utilized a standard Markov chain Monte Carlo method to generate a predictive summary for the past year's seven sites (2008, 2011) and future year

(2012). Also, the predicted frequencies are close to the observed one and within the 95% confidence interval of the observed frequencies. Furthermore, it is found that the model is useful to future crash prediction in the Gulf Cooperation Council countries. The policymaker from GCC can improve traffic safety by decreasing road traffic accidents and road traffic injuries. On the other side, Jamal, Mahmood, Riaz, and Al-Ahmadi (2021) proposed the use of statistical monitoring methods for real-time highway safety surveillance. The candidate data count models are Poisson, Negative Binomial, Conway-Maxwell-Poisson. The dataset used is the inter-cities rural highways vehicles crash data under the jurisdiction from the ministry of transport, Riyadh, Saudi Arabia, from 2017 to 2019. It is found that Conway-Maxwell-Poisson is the best model. Furthermore, they identified that road type and road surface conditions highly influence the crash frequencies.

In summary, in the 1990s, the researcher found out that Poisson and NB have good performance in modeling crash data. In the 2000s, they started to improve modeling accident data, so they compare NB and Poisson with another regression model. In the 2010s, Sellers and Shmueli (2010) found out that COM-Poisson regression has better performance than traditional Poisson regression. Also, most of them will be using Poisson, NB, COM-Poisson, but still have some researcher like AlKheder and Al-Rashidi (2020) will use the Bayesian hierarchical model for forecasting the crash frequencies.

Chapter 3

Methodologies

3.1 Data Pre-processing Procedure

In the dataset, all the input attributes are categorical. Each observation from the dataset indicates one case of a crash accident at a fixed time. To group the dataset from a case-to-case basis into a daily basis, we need to draw special attention. It is because if the classical method, which is dummy variables, is used, then the number of input attributes will unnecessarily inflate a lot, and the model complexity will be unreasonably increased. In the project, we calculated the weighted averages based on the input attributes to generate the daily crash data. Before introducing the corresponding index function, a clarification of the definition of "label" should be provided. "Label" has a different meaning in the aspect of Statistics and Data Mining. In the following content, "label" refers to the numerical representation for the outcome of a categorical variable.

$$IV_i = \frac{\sum_{j=1}^{L} l_{ij} n_{ij}}{\sum_{j=1}^{L} n_{ij}} \; . \tag{1}$$

where i is the indexed days, L is number of labels from the categorical variable IV, l_{ij} is the jth label on indexed day i, n_{ij} is the total number of l_{ij}. For example, categorical variable road type has 3 labels which are 1, 2, 3, and they represent divided highway, expressway, and single highway, respectively. Suppose there are 10, 5, 7 vehicle accidents that occurred on the indexed day 1. Then, the indexed value of road type on day 1 is derived as follows.

$$IV_1 = \frac{1\times10+2\times5+3\times7}{10+5+7} \approx 1.8636. \tag{2}$$

3.2 Poisson Distribution

The Poisson distribution is used to model the number of occurrence(s) of an event within a given time interval. The probability mass function (PMF) is in the following.

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!} \tag{3}$$

y is a count number; λindicates the average number of the event(s) that occurred in the given time interval. The Poisson distribution assumes the mean and the variance are the same. Therefore, the first two central moments are

$$E(Y) = \lambda \tag{4}$$

$$var(Y) = \lambda \qquad (5).$$

3.3 Poisson Regression

In Poisson regression, the expected number of counts follows a Poisson distribution, where the expected count for $\hat{y}_i$, for i=1, …, n, is a function of p covariates X_{ij}, j=1, …, p. Therefore, we have the following equations.

$$y_i \sim Poi(\lambda_i) \tag{6}$$

$$\widehat{\lambda_\iota} = exp\left(\beta_0 X_{i0} + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}\right) \tag{7}$$

where the β_j's are the estimated regression coefficients corresponding to j=1, …, p (X_{i0} is a vector of 1's for the slope intercept model).

Given the Poisson model is heteroscedastic, the coefficients are estimated with maximum likelihood methods. The likelihood function is defined as follow.

$$L(\beta) = \prod_i \frac{exp(-exp(\beta X_i))(exp(\beta X_i))^{y_i}}{y_i!} \quad (8)$$

If the model has a perfect fit to the data, then the maximum likelihood value is zero. If the mean of count is not equal to the variance, then the data indicates under/over dispersion. Poisson regression cannot handle the under/over dispersion well. Oh, Washington, and Nam (2006) mentioned that crash data is commonly to be over-dispersed. Special attention may need to be drawn if traditional Poisson regression is used to fit the crash data.

3.4 Negative Binomial Distribution

A non-negative random variable Y of the Negative Binomial distribution takes the following PMF:

$$\mathrm{P(Y = y)} = \left(\frac{\mu}{\mu + \alpha}\right)^{y} (1 + \frac{\mu}{\alpha})^{-\alpha} \frac{\Gamma\,(y+\alpha)}{y!\,\Gamma\,(\alpha)} \quad (9)$$

where μ = E(Y), v is the dispersion parameter and letting α =$\frac{1}{v}$, and Γ is the gamma function.

$$\text{Also, } \mathrm{Var(Y)} = \mu + v\mu^2 \quad (10)$$

3.5 Negative Binomial Regression

Negative Binomial regression is the generalization of Poisson regression since when v = 0, the Negative Binomial model reduces the Poisson model. Negative Binomial (NB) regression is a common statistical method to model overdispersed data. The NB model takes a relationship between the expected count frequencies at $\hat{y}_i$, for i = 1, …, n and the p covariates Xij, j=1, …, p.

$$\mathrm{y}_i \sim Poi(\lambda_i) \quad (11)$$

$$\hat{\lambda}_i = \exp\left(\beta_0 X_{i0} + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i\right) = \exp\left(\sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i\right) \quad (12)$$

where exp(ε_i)~Gamma(1, v^2), v is a dummy variable indicating v^2 is a constant variance. The error term allows the over-dispersion of the variance, as expressed from the next equation.

$$\mathrm{var}(y_i) = \mathrm{E}(y_i) + \mathrm{vE}(\mathrm{y}_i)^2 \quad (13)$$

,where v is the dispersion parameter. Combining Eq. (11) and the gamma heterogeneity from (13), we have the following equation.

$$y_i \sim NB(\lambda_i, \mathrm{E}(y_i) + \mathrm{vE}(\mathrm{y}_i)^2) \quad (14)$$

The greater value of v, the greater variability of the data over the mean $\hat{\lambda}_i$. Similarly, the coefficients βj's are estimated by the maximum log-likelihood method (logeL(β)).

3.6 COM-Poisson Distribution

Conway and Maxwell (1962) first introduced Conway-Maxwell-Poisson (COM-Poisson) distribution for modeling queues and service rates. COM-Poisson distribution is a generalization of the Poisson distribution, and its probability mass function (PMF) is as follows.

$$\mathrm{P(Y = y)} = \frac{\lambda^{y}}{(y!)^{v} Z(\lambda,\mathrm{v})} \quad (15)$$

$$\mathrm{Z}(\lambda,\mathrm{v}) = \sum_{n=0}^{\infty} \frac{\lambda^{n}}{(n!)^{v}} \quad (16)$$

Y is a count number; λ is a rate parameter; v is the dispersion parameter; Z(λ, v) is the normalizing factor. The COM-Poisson is used for modelling over data with under-dispersion (v>1), equi-dispersion (v=1), and over-dispersion (v<1). Some of the well-known distributions are the special cases of Com-Poisson distribution. If v=1, then the COM-Poisson distribution reduces into Poisson distribution. If v=0, then the COM-Poisson distribution turns into the geometric distribution. If v → ∞, then it becomes the Bernoulli distribution in the limit. This indicates COM-Poisson distribution has great flexibility for data count modelling.

Note that the COM-Poisson distribution does not have its closed-form expressions to express its first two central moments. Shmueli et al. (2005) used an asymptotic expression for Z in Eq. (15) to derive an approximation. Thus the moments have the following forms:

$$E(Y) = \frac{\partial logZ}{\partial log\lambda} \approx \lambda^{\frac{1}{v}} + \frac{1}{2v} - \frac{1}{2} \quad (17)$$

$$var(Y) = \frac{\partial^2 logZ}{\partial log^2 \lambda} \approx \frac{1}{v}\lambda^{1/v} \tag{18}$$

Please pay attention that the approximations may not be accurate for v>1 or $\lambda^{1/v}$ (Shmueli et al., 2005).

However, the COM-Poisson has some limitations associated with Generalized Linear Modelling (GLM). Guikema and Goffelt (2008) mentioned that neither λ nor v could provide a clear rate parameter. That is, λ can provide a reliable approximation to the mean when v is closed to one, while it provides unreliability when v is small (i.e. over-dispersion). Therefore, Guikema and Goffelt (2008) proposed a re-parameterization of the COM-Poisson distribution in order to have a reliable rate parameter. The refined parameterization is as follows.

$$P(Y = y) = \frac{1}{S(\mu,v)}(\frac{\mu^y}{y!})^v \tag{19}$$

$$S(\mu, v) = \sum_{n=0}^{\infty}(\frac{\mu^n}{n!})^v \tag{20}$$

We substitute $\mu=\lambda^{1/v}$ in the approximation of Eqs. (17), (18) and have a new moments' formulation and approximations, where asymptotic approximations from Shmueli et al. (2005) are used.

$$E(Y) = \frac{1}{v}\frac{\partial logS}{\partial log\mu} \approx \mu + \frac{1}{2v} - \frac{1}{2} \tag{21}$$

$$var(Y) = \frac{1}{v^2}\frac{\partial^2 logS}{\partial log^2\mu} \approx \frac{\mu}{v} \tag{22}$$

The approximations are more reliable when μ>10. Lord, Guikema, and Geedipally (2008) commented that the new formulation leads to higher efficiency in COM-Poisson GLM development. The clear rate parameter offers a clear direction for the construction of the rate link function, with the ease of interpretation with a range of values of the dispersion parameter. Besides, the dispersion parameter enables the usage of a second link function to capture the varying dispersion from one measurement to another measurement.

3.7 COM-Poisson Regression with GLM

Guikema and Goffelt (2008) estimated a COM-Poisson GLM framework for count data modelling. The framework makes use of two link functions such that the mean and the variance are dependent on the covariates. Suppose x_i and z_j are covariates with p covariates and q covariates used in the rate link function and dispersion link function, respectively. Note that p and q are not necessary to be equal. The rate link function and the dispersion link function are as follows.

$$ln(\mu) = \beta_0 + \sum_{i=1}^{p} \beta_i \, x_i \tag{23}$$

$$ln(v) = \gamma_0 + \sum_{j=1}^{q} \gamma_j \, z_j \tag{24}$$

Lord et al. (2008) commented that the GLM above is greatly flexible and interpretable. The GLM can be used to model the count data with under/equi/over dispersion. If the data indicated intermingled variation of dispersion, the two link functions could still provide reliable modelling. For the crash accident modelling, the crash data are usually from many roadways and are suspected of belonging to groups (heterogeneous sites). The dispersion may be varied among groups. The Com-Poisson regression with GLM is possibly useful to capture the variation. Furthermore, the covariate values are allowed to contribute to the variance. The parameters are directly linked to the mean or the variance. Therefore, the covariates and the regression parameter estimates are contributed to the mean and variance of the predicted values.

3.8 Performance Measures

This section will provide the details of performance measures such as Log-likelihood, AIC and BIC, which are used for distribution fitting and model selection in the project.

3.8.1 Log-likelihood

Likelihood function measures the goodness of fit of a statistical model with some values of parameters. Among a number of statistical models, the model with the greatest likelihood value is considered as the model best fits the sample data. To be more precise, this refers to the model with the combination of parameter values maximizing the probability of the occurrence of the sample data obtained. Log-likelihood is simply the logarithm of the likelihood value. The log-likelihood estimation varies among different statistical models, so that it is difficult to mention the estimators used for every case. Since the statistical models developed in the project are some popular statistical distributions and regression models, all the parameters follow a probability distribution. Therefore, the branch of log-likelihood estimation used is maximum likelihood estimation.

3.8.2 Akaike Information Criterion (AIC)

AIC is a measure of goodness of fit of a statistical model and has the following form:

$$\mathrm{AIC} = -2\ln\mathrm{L} + 2\mathrm{p} \tag{25}$$

Note that L is the maximized likelihood function value, and p is the number of parameters in the statistical model. The first term measures the bias and the second term adds a penalty to the number of parameters used. Thus, AIC takes a balance of the bias and the model complexity. The model has the lowest value of AIC, indicating it is the best model among other models.

3.8.3 Bayesian Information Criterion (BIC)

BIC is similar to AIC since it can be regarded as a multivariate function with the maximized function value and the number of parameters as the input.

$$\mathrm{BIC} = \mathrm{p} \times \ln(\mathrm{n}) - 2\ln\,(\mathrm{L}) \tag{26}$$

Note that L, p, n, are the maximized likelihood function value, the number of parameters in the statistical model, and the number of observation from the sample, respectively. Comparing to AIC, BIC adds a more severe penalty to the number of parameter in use. Also, the candidate model with the lowest value of BIC is the best model.

Chapter 4

4.1 Result and Discussion

In this report, we will provide a descriptive and exploratory analysis of the data. For the first part below, we will provide boxplots and pie charts to explain the univariate data analysis. As for the bivariate data analysis, we will be using the scatter plots to explain. In the regression analysis, we will talk about Distribution Fitting, All Possible Regression as well as Model Selection.

4.2 Data Description

The dataset we have chosen from Saudi Arabia is from October 2016 to the end of 2019 contains 48551 observations and 23 variables. The data are measured from Ministry of Transport. In order to group the dataset from a case-to-case basis into a daily basis, we sum the number of the individual crash accident on an indexed day i as the response y_i. For the numerical predictor, we simply calculate the frequency-based average. For the categorical predictor, the manipulation is done by using the index measure reported in the section 3.1. Our variable Y is the number of accidents per day, and there are 6 explanatory variables X, which are On-scene Road Markings, Road cats-eyes, Road Geometric Details, Road Surface Condition, Road Type as well as Weather. In each variable X, there are different code corresponds to a different situation.

In Weather Condition, “0” is “not available”, “1” is “dusty”, “2” is“fog”, “3” is “shiny”, “4” is “others” as well as “5” as “rainy”.

In road geometric details, “0” is “not available”, “1” is “tangent”, “2” is “horizontal curve”, “3” is “vertical curve” and “4” is “intersection”.

For the road type, the dataset is separated three types which are Divided Highway, Expressway, and single way. There are 3 code representing each type which are “1”, “2” and “3” respectively.

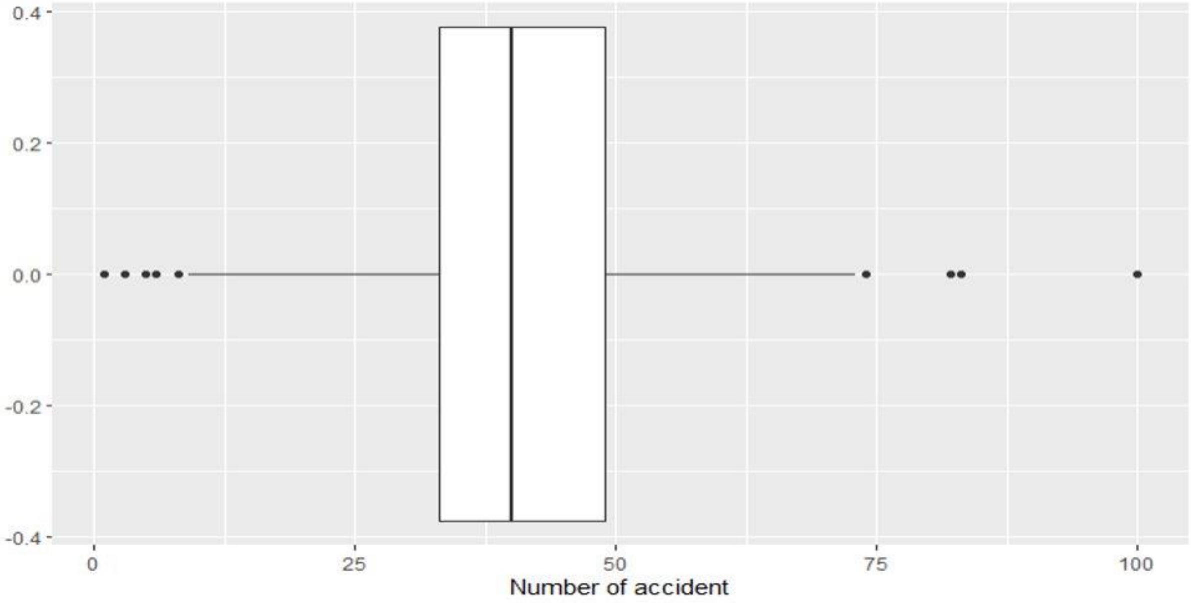
As for road surface condition, there are 8 codes representing different conditions, which are 0 to 7, 0) not available, (1) cracks and fossils, (2) dry, (3) good, (4) maintenance on the road, (5) others, (6) sand and (7) wet.

There are only two conditions for the On-scene Road Markings: road markings present and Road markings absent, respectively 1 and 0. Besides, Road cats eyes have similar situations, which also have only two conditions: road catseyes present and road catseyes absent, respectively 1 and 0.

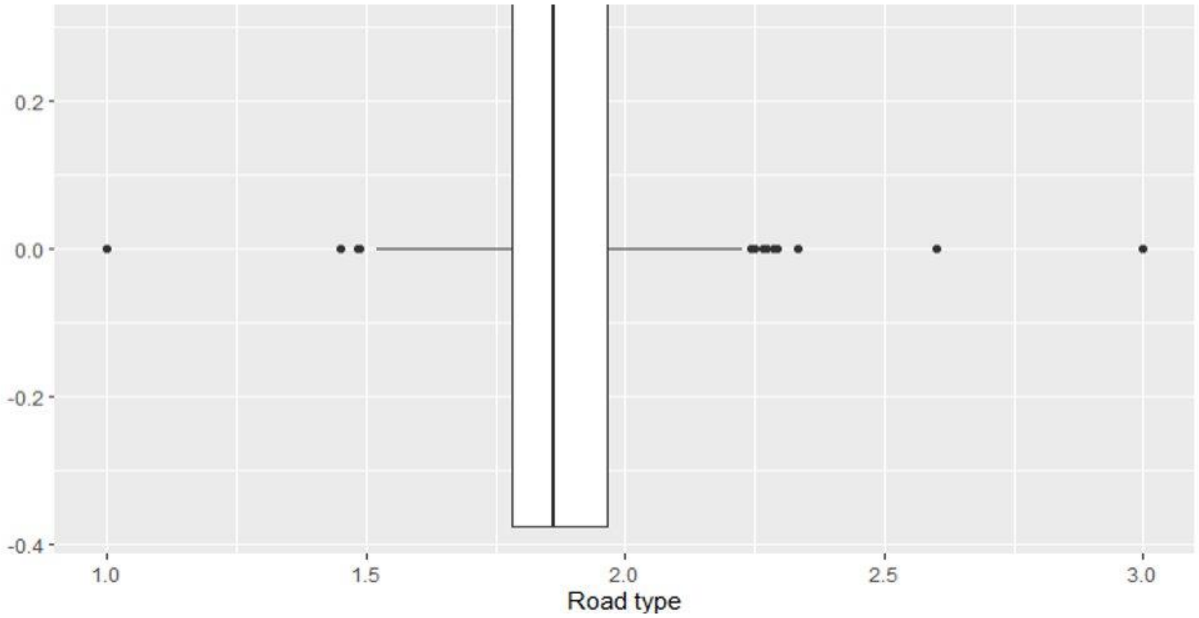
However, there are some missing values in the dataset. In order to handle this problem, we define that all of the missing values to be zero, which means it is not available. Using this method, we found that there are 1187 rows that have missing data. After removing this variable, we only have 5 missing value, 3 rows from accident cause and 2 from road surface condition.

4.3 Univariate Data Description

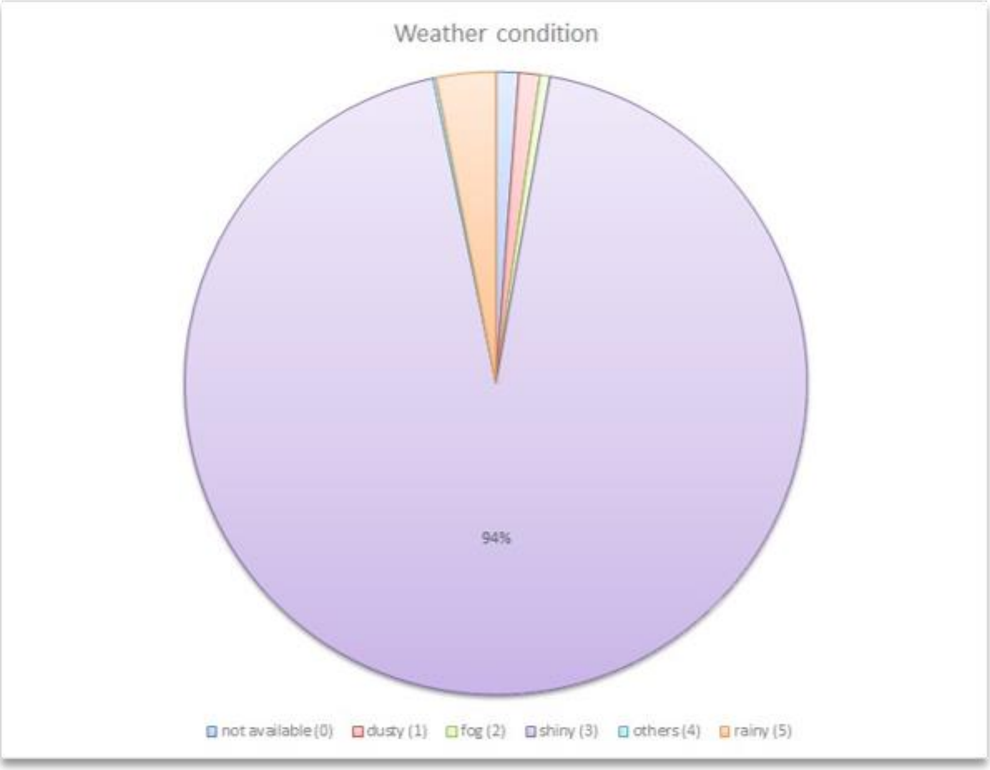
The boxplot of the number of accidents shows us that the middle half of the data between 25 to 50, which means that there are 50% occur 25 to 50 accidents daily. The minimum is near to 0, and the maximum is near 100, so it reveals that it is a high risk to happen a large number of accidents daily.



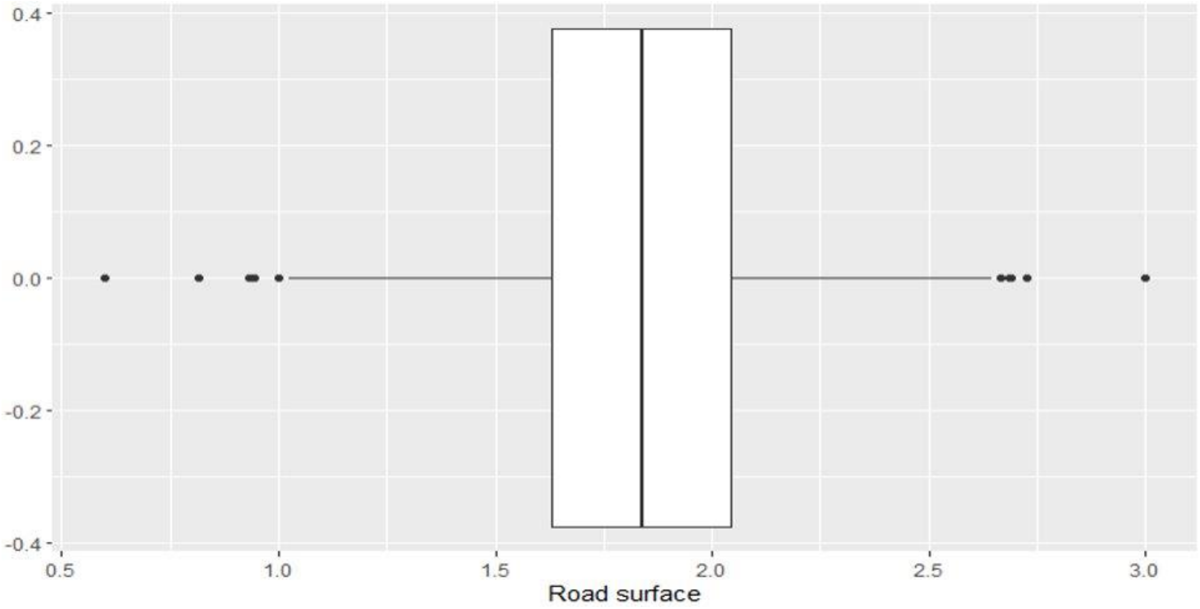
Also, the boxplot of road type demonstrates that the middle half of the data between 1.75 to 2, it can conclude that the road type has 50% of approximate data equals to 2 which means 50% opportunity the accident will occur on“expressway”. The minimum and maximum are respectively 1 and 3, so the divided highway and single highway also have opportunity happen accidents.



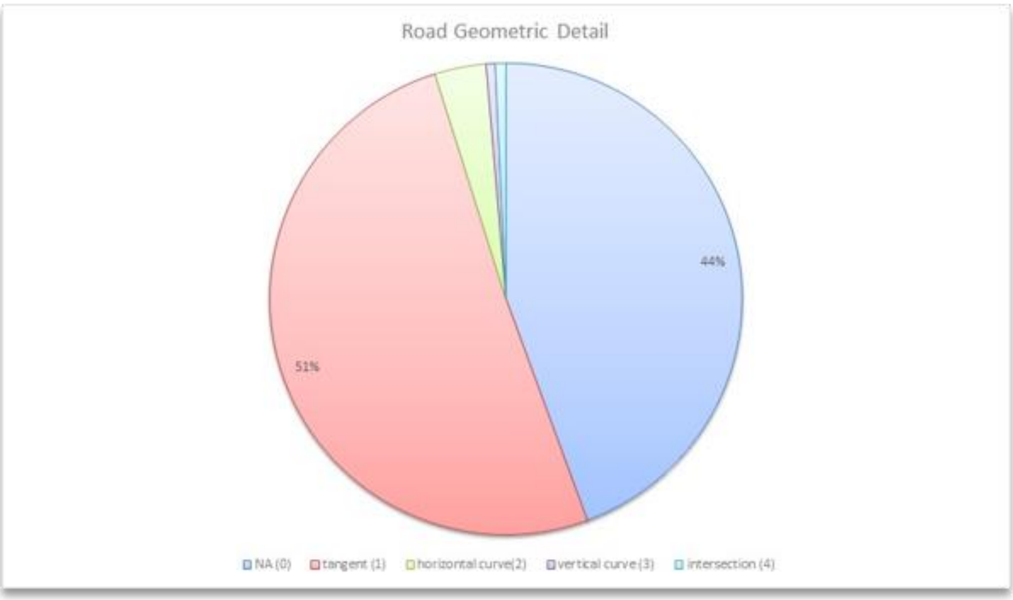
The pie chart of weather shows that 94% of weather is 3 which is “shiny". As a result of the pie chart, most accidents may occur when the weather is shiny. The remaining parts of the pie chart are not available, dusty, fog, others, and rainy.



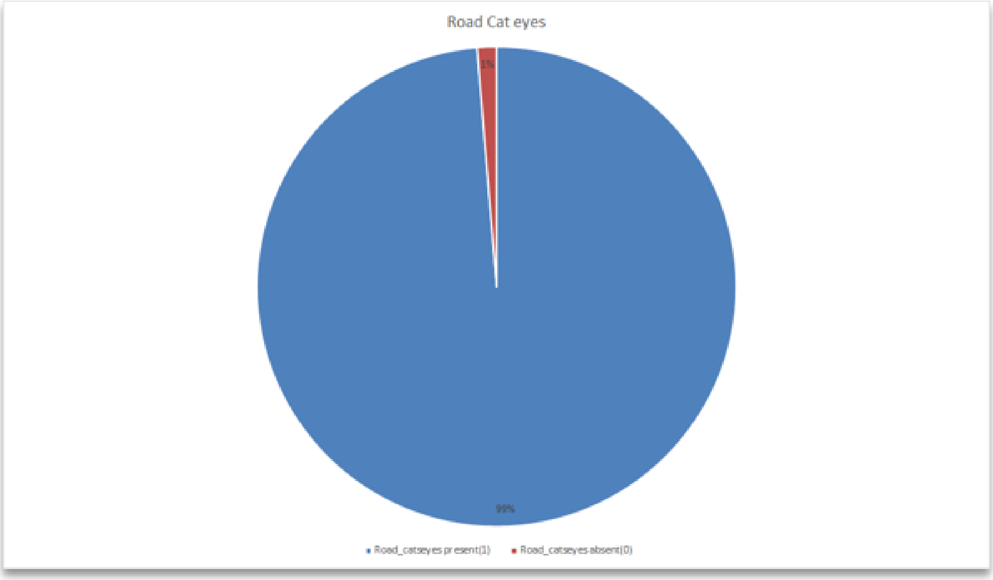
The boxplot of the road surface shows us that the middle half of the data is near 1.5 to 2, which means that 50% of accidents occur when the road surface is "dry". The minimum is near to 0.5, and the maximum is near to 3; it indicates that accidents commonly occur when the road surface is "cracks and fossils", "dry" or "good".

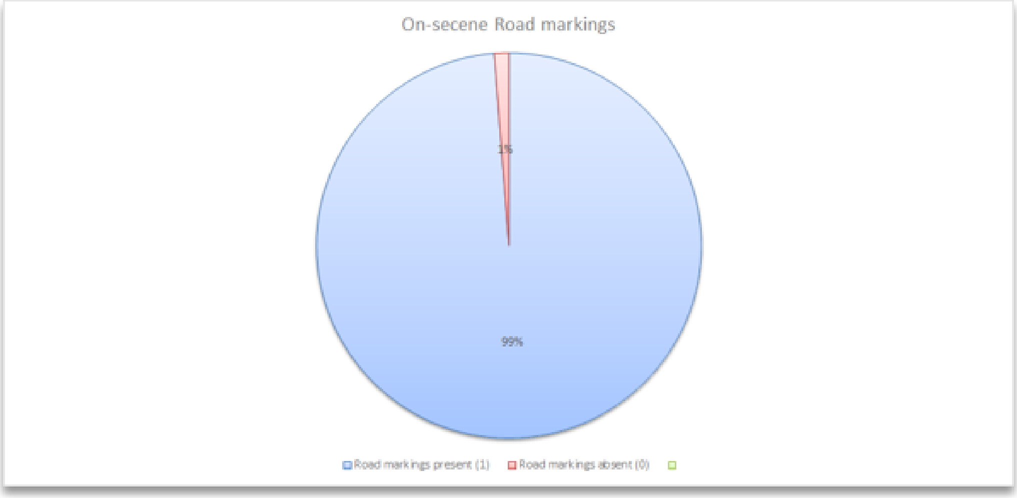


The pie chart of the road geometric details shows us that 51% is tangent and 44% is not available, the remaining part are horizontal curve, vertical curve, and intersection. The most possible weather when accidents occur trends to not available and tangent.



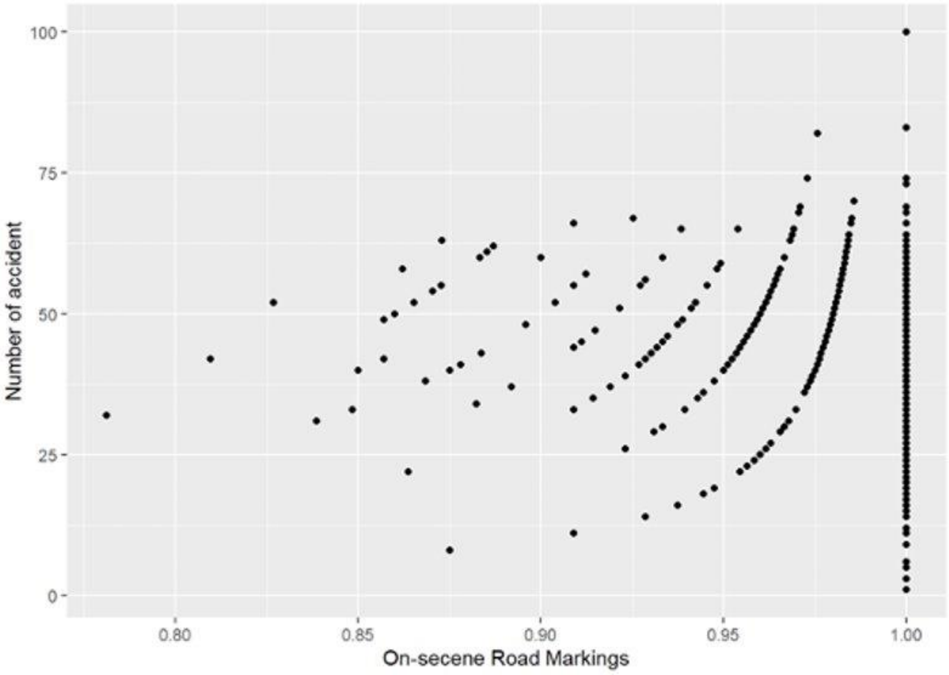
The pie chart of Road Cat Eyes showed that 99% of the data is 1, only 1% is zero, which means that the situation mainly on present road cat eyes. Similarly, the pie chart of On-scene Road markings showed that 99% of the data is 1, only 1% is zero, which means that the situation mainly on present road markings.



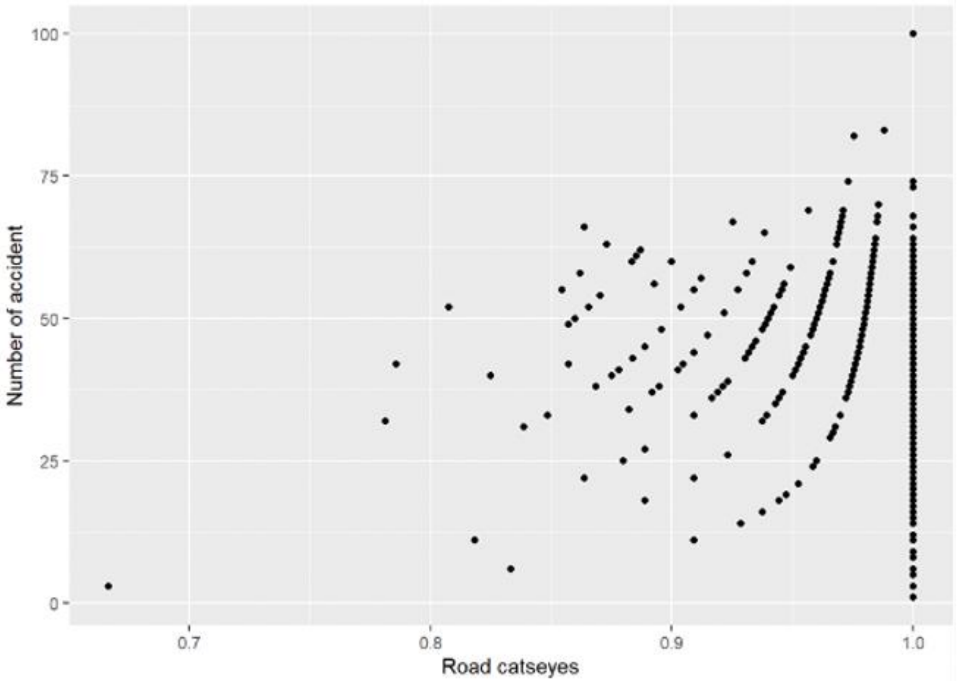


4.4 Bivariate Data Description

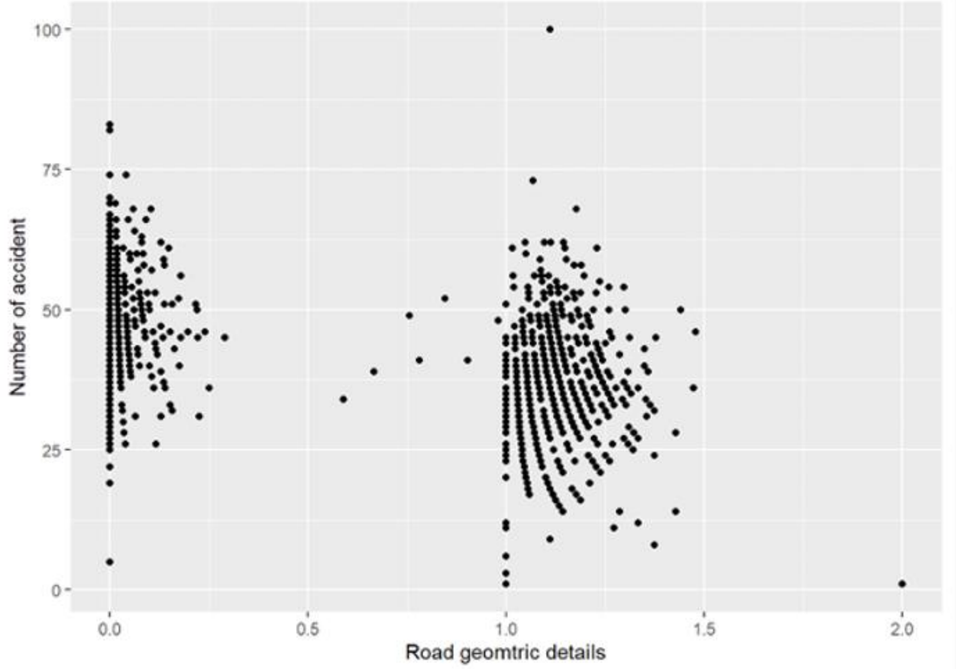
The scatterplot of on-scene road markings shows that all of the points near to 1, only a few points below 0.9. As the on-scene road markings only have two conditions, which are 0 and 1, so the results should be trended to 1. That means the on-scene road markings present when accidents occur. The correlogram shows us that the correlation of On-scene Road Markings between the number of accidents is –0.154.



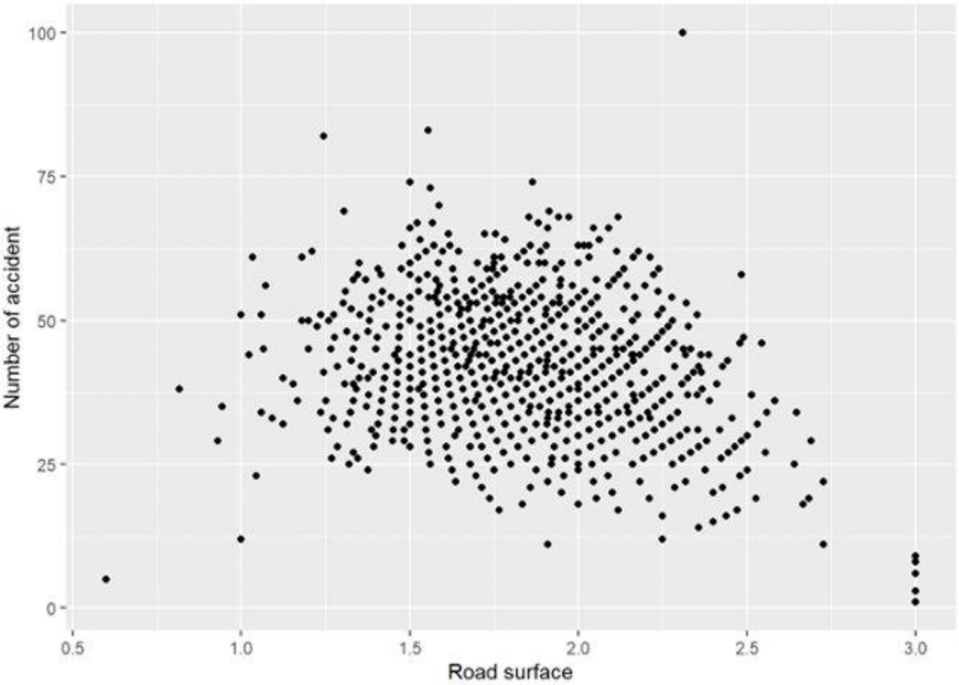
The scatterplot of road cat eyes shows that all of the points near 1, only a few points below than 0.9. As the road cat eyes only have two conditions, which are 0 and 1, so the results should be near to 1. That means the road cat eyes present when accidents occur. The correlogram shows us that the correlation of Road Cat Eyes between the number of accidents is –0.096.



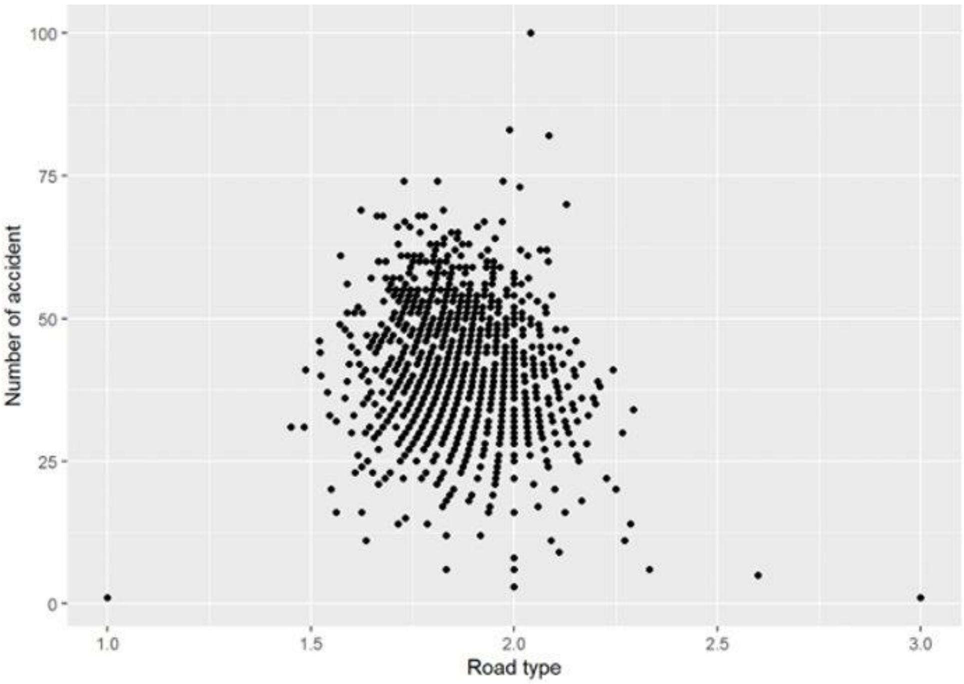
The scatterplot of road geometric details shows us that most of the points distributed at 0 to 0.5 and 1 to 1.5, which are “not available” and “tangent. Only one point reach to 2, which is a horizontal curve. The correlogram shows us that the correlation of road geometric details between the number of accidents is -0.477.



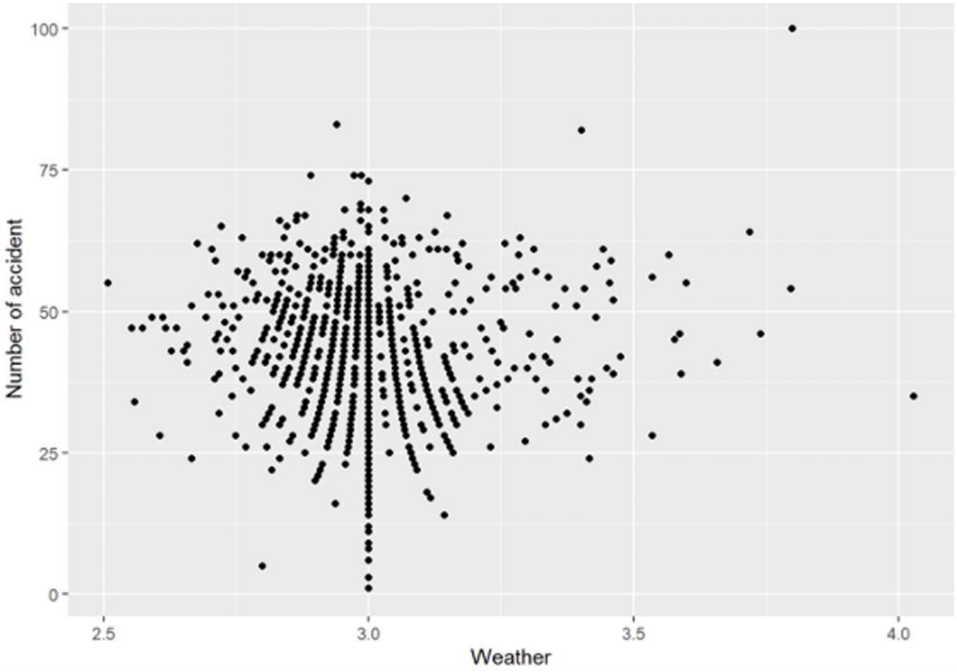
The scatterplot of road surface shows that all the points within 0.5 to 3, which are “cracks and fossils”, “dry” and “good” and the most apparent is 1 to 2. It indicates that most of the accidents occur when the road surface either “cracks and fossils” or “dry”. The correlogram shows us that the correlation of road surface between the number of accidents is -0.268.



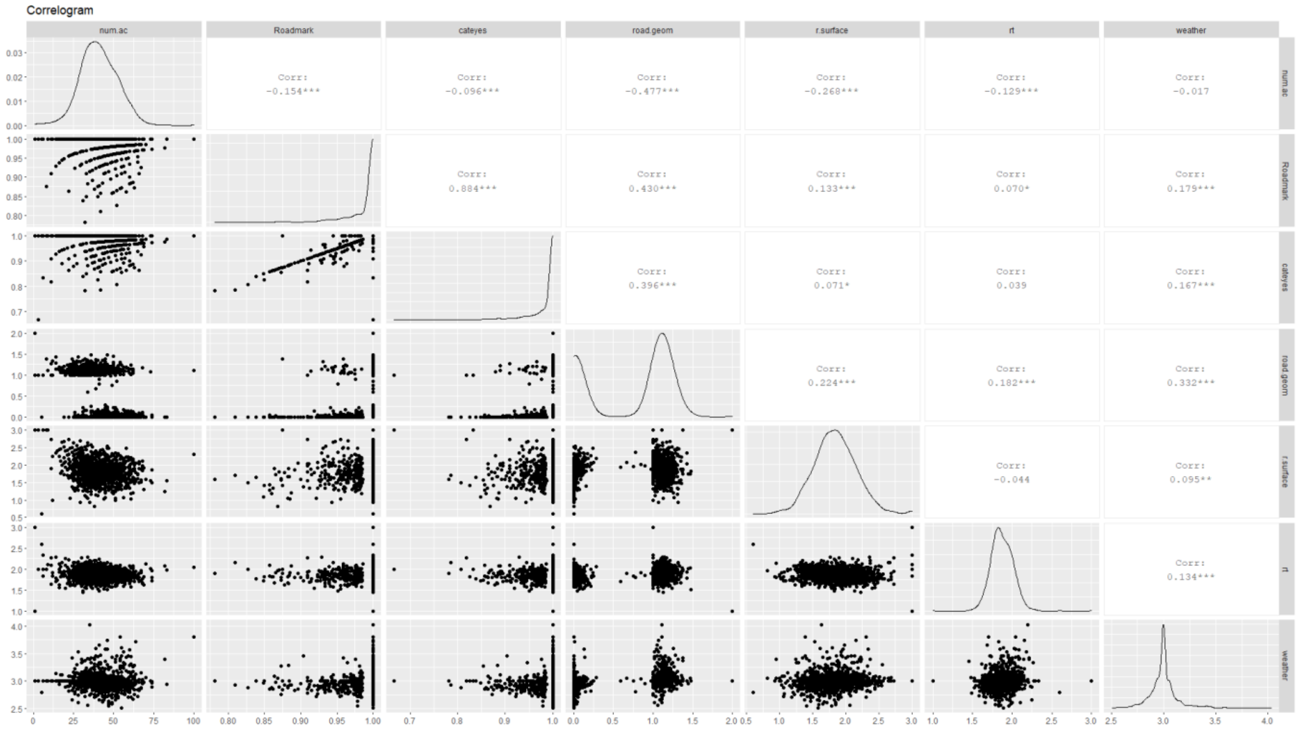
The scatterplot of road type demonstrates that most of the points present at 1.5 to 2, which indicates that most of the accident occurs when the road type is the expressway. There are few points that will reach 1 or 3; it also means that divided highway and single highway have accidents, but it is not as much as an expressway. The correlogram shows us that the correlation of road type between the number of accidents is -0.129.



The scatterplot of weather demonstrates that most of the points present at 2.5 to 3.5, which indicates that most of the accident occurs when the weather is shiny. There are few points out of this range; it also means that there are accidents that occur when weather is "other" as it is near to 4. The correlogram shows us that the correlation of weather between the number of accidents is -0.017.



We further categorize all the dependent variables into three factor categories. On-secene road markings and road catseyes are categorized into informative sign. Road geometric details, road surface condition, and road type are categorize into roadway factors. Weather condition is the weather factor.



4.5 Regression Analysis

In this study, since the crash data shows there is some car accident happening every day. Obviously, there is no excess zero problem from the data. The major consideration of the adoption of the candidate models is the data dispersion. We adopt Poisson Regression, Negative Binomial, and COM-Poisson as the candidate models.

4.5.1 Distribution Fitting

The three candidate count models, which are Poisson, Negative Binomial, and COM-Poisson distribution, are used to find the best fitting distribution of the response variable.

Distribution fitting is an effective and efficient technique for model selection, regardless of the number of candidate models.

The fitting distribution result is shown in table 1. Taking Poisson distribution as an example, the response best fits a Poisson distribution with lambda estimate = 40.9023. Also, the goodness of fit result with AIC, BIC, log-likelihood are shown. From the results, comparing to the Poisson distribution, the response fits better in a Negative Binomial distribution. Since the log-likelihood for COM-Poisson distribution cannot be determined, we just compare COM-Poisson distribution with Negative Binomial Distribution in terms of AIC and BIC. We can see the AIC and BIC for COM-Poisson distribution are the lowerest, and hence the response variable is better fitted to a COM-Poisson distribution. The estimate of the dispersion parameter equals to 0.28 which is much lower than 1 indicating that the crash data are significally over-dispersed. Therefore, for the regression implementation, we would only implement the models with COM-Poisson regression.

Table 1: Distribution Fitting to the Dependent Variable

Goodness of Fit			
	Poisson	**Negative Binomial**	**COM-Poisson**
Parameter(s)	$\hat{\lambda} = 40.90$	$\hat{v} = 15.42$	$\hat{\lambda} = 3.71$ $\hat{v} = 0.28$
Log Likelihood	-5367.89	-4652.01	N.A.
AIC	10737.78	9308.03	9241.2094
BIC	10742.86	9318.19	9251.3678

4.5.2 All Possible Regression

All Possible Regression is selected to be the technique for feature selection. The number of predictors for regression is six, which is not too many. All Possible Regression compares the performance of the number of $2^6 – 1 = 63$ combinations of COM-Poisson models. Note that the combination of predictor(s) of each fitting contributes to the mean the two link functions for mean and dispersion parameter. That is, in our fitting, the number of p covariate(s) and q covariate(s), mentioning from equation (23) and (24), chapter 3.7 COM-Poisson Regression with GLM, are always the same.

Another popular feature selection technique is Best Subset Selection which votes the best variable k model for further decision. Compared to All Possible Regression, Best Subset Selection cannot allow multiple models with the same number of predictors for the final model selection. Hence, some potential model with good performance may be omitted. Therefore, in the case of a few predictors, All Possible Regression leads to a wiser model selection.

The regression results are as follow.

Table 2: All Possible Regression Results

Goodness of Fit results				
p	**Formula**[1]	**AIC**	**BIC**	**Log-likelihood**
1	Y = X1	10651.53	10666.77	-5322.77
1	Y = X2	10706.12	10721.36	-5350.06
1	Y = X3	9850.658	9865.895	-4922.33
1	Y = X4	10455.77	10471.01	-5224.88
1	Y = X5	10675.82	10691.06	-5334.91
1	Y = X6	10740.65	10755.89	-5367.32
2	Y = X1 + X2	10621.92	10642.23	-5306.96

[1] Y is the number of daily accident and X1 to X6 are the index values of the presence/absence of on-secene road marking and cateyes, road geometric details, road surface condition, road type, and weather condition, respectively.

2	Y = X1 + X3	9840.834	9861.15	-4916.42
2	Y = X1 + X4	10406.4	10426.71	-5199.2
2	Y = X1 + X5	10597.37	10617.68	-5294.68
2	Y = X1 + X6	10653.14	10673.45	-5322.57
2	Y = X2 + X3	9811.759	9832.076	-4901.88
2	Y = X2 + X4	10436.3	10456.61	-5214.15
2	Y = X2 + X5	10645.55	10665.86	-5318.77
2	Y = X2 + X6	10708.11	10728.43	-5350.06
2	Y = X3 + X4	9744.916	9765.233	-4868.46
2	Y = X3 + X5	9845.488	9865.805	-4918.74
2	Y = X3 + X6	9765.276	9785.593	-4878.64
2	Y = X4 + X5	10374.48	10394.79	-5183.24
2	Y = X4 + X6	10457.48	10477.8	-5224.74
2	Y = X5 + X6	10677.82	10698.13	-5334.91
3	Y = X1 + X2 + X3	9792.165	9817.561	-4891.08
3	Y = X1 + X2 + X4	10387.35	10412.75	-5188.67
3	Y = X1 + X2 + X5	10571.35	10596.74	-5280.67
3	Y = X1 + X2 + X6	10623.66	10649.06	-5306.83
3	Y = X1 + X3 + X4	9730.082	9755.478	-4860.04
3	Y = X1 + X3 + X5	9835.779	9861.175	-4912.89
3	Y = X1 + X3 + X6	9758.226	9783.622	-4874.11
3	Y = X1 + X4 + X5	10335.27	10360.67	-5162.64
3	Y = X1 + X4 + X6	10405.35	10430.75	-5197.68
3	Y = X1 + X5 + X6	10596.91	10622.3	-5293.45
3	Y = X2 + X3 + X4	9706.252	9731.648	-4848.13
3	Y = X2 + X3 + X5	9809.666	9835.062	-4899.83
3	Y = X2 + X3 + X6	9733.472	9758.868	-4861.74
3	Y = X2 + X4 + X5	10359.3	10384.7	-5174.65
3	Y = X2 + X4 + X6	10436.62	10462.02	-5213.31
3	Y = X2 + X5 + X6	10646.63	10672.03	-5318.32
3	Y = X3 + X4 + X5	9733.222	9758.618	-4861.61
3	Y = X3 + X4 + X6	9654.376	9679.772	-4822.19
3	Y = X3 + X5 + X6	9755.785	9781.181	-4872.89
3	Y = X4 + X5 + X6	10373.02	10398.42	-5181.51
4	Y = X1 + X2 + X3 + X4	9692.21	9722.685	-4840.11
4	Y = X1 + X2 + X3 + X5	9788.855	9819.33	-4888.43
4	Y = X1 + X2 + X3 + X6	9710.092	9740.567	-4849.05

4	Y = X1 + X2 + X4 + X5	10319.66	10350.13	-5153.83
4	Y = X1 + X2 + X4 + X6	10386.7	10417.17	-5187.35
4	Y = X1 + X2 + X5 + X6	10571.32	10601.8	-5279.66
4	Y = X1 + X3 + X4 + X5	9718.31	9748.785	-4853.16
4	Y = X1 + X3 + X4 + X6	9642.633	9673.108	-4815.32
4	Y = X1 + X3 + X5 + X6	9748.922	9779.397	-4868.46
4	Y = X1 + X4 + X5 + X6	10328.96	10359.44	-5158.48
4	Y = X2 + X3 + X4 + X5	9695.59	9726.065	-4841.8
4	Y = X2 + X3 + X4 + X6	9620.483	9650.958	-4804.24
4	Y = X2 + X3 + X5 + X6	9725.172	9755.647	-4856.59
4	Y = X2 + X4 + X5 + X6	10354.94	10385.41	-5171.47
4	Y = X3 + X4 + X5 + X6	9636.44	9666.915	-4812.22
5	Y = X1 + X2 + X3 + X4 + X5	9682.898	9718.452	-4834.45
5	Y = X1 + X2 + X3 + X4 + X6	9605.337	9640.891	-4795.67
5	Y = X1 + X2 + X3 + X5 + X6	9703.035	9738.59	-4844.52
5	Y = X1 + X2 + X4 + X5 + X6	10314.09	10349.64	-5150.04
5	Y = X1 + X3 + X4 + X5 + X6	9624.692	9660.246	-4805.35
5	Y = X2 + X3 + X4 + X5 + X6	9603.844	9639.398	-4794.92
6	Y = X1 + X2 + X3 + X4 + X5 + X6	9590.372	9631.006	-4787.19

4.5.3 Model Selection

From the result, two models are selected to identify the significant factors leading to the frequencies of daily car accidents. The model performance information is available in Table 3 as below. Model one contains only one predictor, which is road geometric detail. Model one has the best performance among all the one predictor models and has the similar performance to the best model which is the full model. As a result, we can conclude road geometric detail provides the best explanatory effect.

Moreover, model two, which is the full model, is selected due to the performance and interpretability. Try to come up with the performance first, it has the minimum of the AIC, BIC and maximum of Log -likelihood among other models. For the discussion of the balance between accuracy and model complexity, we can take a look at the BIC. BIC reflects the performance of a model with the consideration of accuracy and model complexity and penalizes severely to the latter one. From the result, the full model still has the minimum of BIC, which indicates performance gain surpassing the increase of model complexity. In terms of the interpretability, the full model contains at least one predictor from different categories, which are informative sign, roadway factors, and weather factors. Therefore, we can conclude that informative sign, roadway factors, and weather factors are influential to the road safety assessment.

Table 3: The selected models from All Possible Regression

Goodness of Fit				
Model	**Predictor(s)**	**AIC**	**BIC**	**Log Likelihood**
Model 1	road.geom	9850.6576	9865.8951	-4922.3288
Model 2	roadmark, cateyes, road.geom, r.surface, rt, weather	9590.3723	9631.0058	-4787.1861

Chapter 5

Conclusion

This study began by introducing the serious car accidents problem in the Kingdom of Saudi Arabia. Then we tried to provide some suggestions to the road safety assessment by finding out the factors affecting the frequencies of car accidents. A comprehensive review on crash data from 1980 to the present suggested that Negative Binomial Regression, Poisson Regression, and Conway-Maxwell Poisson Regression can have a reliable performance on crash data. These three models are useful for count data modelling. In the descriptive and explanatory analysis, we provided some boxplots, pie charts and scatterplots to have an univariate and bivariate data description. Before conducting the modelling, we used the index function to convert the predictors from categorical to numerical data. Besides, we determined a distribution fitting between Poisson, Negative Binomial, and COM-Poisson distribution in order to find out the best fitting distribution of the response variable. As for the result, Conway-Maxwell Poisson Distribution is the best distribution for the data. We used

Conway-Maxwell Poisson Regression to select the best model by using All Possible Regression. We had 63 combinations in total. We found out the full model has the best performance. The full model contains at least one predictor from different categories, which are informative sign, roadway factors, and weather factors. Therefore, informative sign, roadway factors, and weather factors are the influential factors leading to the frequencies of the car accident. Moreover, by comparing the model with one predictor, we realized that “Road Geometric Detail” has the best explanatory effect.

Recommendation, Limitation and Future Studies

In this project, it has been shown that informative sign, roadway factors, and weather factors are influential to the frequencies of car accidents. The authorities should take account of these factors on roadway design and maintenance. For example, they should consider the relationship between the road geometries and the road safety, the configuration of the informative signs, the scalability of the drainage design to deal with the adverse weather.

Also, we will extend our study on some other crash datasets. The datasets are more appropriate from the Middle East or some developing countries which are sharing the similar roadway characteristics with the dataset in this study. Therefore, we can have a more comprehensive and reliable results on the identification of the influential factors.

In addition, the index function in use cannot provide some relatively robust indexes. The magnitude of the index value is not much meaningful since the codes of the categorical variables are randomly assigned. For example, if there are very high frequencies of the codes on the both ends, then the index value would tend to be the code approaching to the midpoint. Then, we would conclude that the daily accident tends to happen when the code is near to the midpoint. Also, interchange of a pair of codes would influence the index value too. We will try to find a robust index function to well handle the categorical variables in the future studies.

Besides, it is suggested to have a measure on the historical traffic flow on the dataset in study. Some popular measures of the traffic flow are average daily traffic (ADT) and hourly volume (VH), especially ADT has been the common explanatory variable in many road safety assessment.

For more about the future studies, we will try to introduce more appropriate modelling techniques in the crash data study. The techniques from machine learning are in our considerations. Some researchers used Artificial Neural Network and Support Vector Machine on the crash data assessment and had been shown the reliable prediction result. However, the machine learning models are usually with less interpretability than the regression models. Therefore, we will pay attention to the models adoption with the objective of identification of the influential factors.

Appendix

Stage 1 Data Preparation

```
#library we may use
 library(dplyr)
 library(tidyr)
library(ggplot2)
 library(epiDisplay)
library(splines2)
library(gam)
library(readxl)
 library(GeneCycle)
 library(scales)
 library(tidyverse)
 library(sqldf)
 library(GGally)
df$Accident_date <- as.Date(df$Accident_date)
 Num_accident <- count(df, vars = df$Accident_date)
 sum(is.na(df))
sum(is.na(df$no_of_vehicles_involved))
##Road type
## splitting and count data by factor
rt1 <- sqldf('select Accident_date, COUNT(Road_type) as rt1 from df where Road_type=1 group by Accident_date')
```

```
rt2 <- sqldf('select Accident_date, COUNT(Road_type) as rt2 from df where Road_type=2 group by Accident_date')
rt3 <- sqldf('select Accident_date, COUNT(Road_type) as rt3 from df where Road_type=3 group by Accident_date')

##merge counted data
rt <- merge(rt1,rt2,by='Accident_date', all=TRUE)
rt <- merge(rt,rt3,by='Accident_date',all=TRUE)

##We define null variable to zero
rt[is.na(rt)] <- 0

## calculate the index
for(i in 1:nrow(rt)){
   rt$rtindex[i] <- (rt$rt1[i] * 1 + rt$rt2[i] * 2 + rt$rt3[i] * 3)/(rt$rt1[i]+rt$rt2[i]+rt$rt3[i])
}
##Weather condition
 wc0 <- sqldf('select Accident_date, COUNT(Weather_Status) as wc0 from df where Weather_Status=0 group by Accident_date')
 wc1 <- sqldf('select Accident_date, COUNT(Weather_Status) as wc1 from df where Weather_Status=1 group by Accident_date')
 wc2 <- sqldf('select Accident_date, COUNT(Weather_Status) as wc2 from df where Weather_Status=2 group by Accident_date')
 wc3 <- sqldf('select Accident_date, COUNT(Weather_Status) as wc3 from df where Weather_Status=3 group by Accident_date')
 wc4 <- sqldf('select Accident_date, COUNT(Weather_Status) as wc4 from df where
Weather_Status=4 group by Accident_date')
 wc5 <- sqldf('select Accident_date, COUNT(Weather_Status) as wc5 from df where
Weather_Status=5 group by Accident_date')

##merge counted data
 wc <- merge(wc0,wc1,by='Accident_date', all=TRUE)
 wc <- merge(wc,wc2,by='Accident_date', all=TRUE)
 wc <- merge(wc,wc3,by='Accident_date', all=TRUE)
 wc <- merge(wc,wc4,by='Accident_date', all=TRUE)
 wc <- merge(wc,wc5,by='Accident_date', all=TRUE)

wc[is.na(wc)] <- 0

## calculate the index
 for(i in 1:nrow(wc)){
   wc$wcindex[i] <- (wc$wc0[i] * 0 + wc$wc1[i] * 1 + wc$wc2[i] * 2 + wc$wc3[i] * 3 +
wc$wc4[i] * 4 + wc$wc5[i] *
5)/(wc$wc0[i]+wc$wc1[i]+wc$wc2[i]+wc$wc3[i]+wc$wc4[i]+wc$wc5[i])
 }
## road surface condition
 rsc0 <- sqldf('select Accident_date, COUNT(road_surface_condition) as rsc0 from df where
road_surface_condition=0 group by Accident_date')
 rsc1 <- sqldf('select Accident_date, COUNT(road_surface_condition) as rsc1 from df where
road_surface_condition=1 group by Accident_date')
 rsc2 <- sqldf('select Accident_date, COUNT(road_surface_condition) as rsc2 from df where
road_surface_condition=2 group by Accident_date')
```

```
  rsc3 <- sqldf('select Accident_date, COUNT(road_surface_condition) as rsc3 from df where
road_surface_condition=3 group by Accident_date')
  rsc4 <- sqldf('select Accident_date, COUNT(road_surface_condition) as rsc4 from df where
road_surface_condition=4 group by Accident_date')
  rsc5 <- sqldf('select Accident_date, COUNT(road_surface_condition) as rsc5 from df where
road_surface_condition=5 group by Accident_date')
  rsc6 <- sqldf('select Accident_date, COUNT(road_surface_condition) as rsc6 from df where
road_surface_condition=6 group by Accident_date')
  rsc7 <- sqldf('select Accident_date, COUNT(road_surface_condition) as rsc7 from df where
road_surface_condition=7 group by Accident_date')

##merge counted data
  rsc <- merge(rsc0,rsc1,by='Accident_date', all=TRUE)
  rsc <- merge(rsc,rsc2,by='Accident_date',all=TRUE)
  rsc <- merge(rsc,rsc3,by='Accident_date',all=TRUE)
  rsc <- merge(rsc,rsc4,by='Accident_date',all=TRUE)
  rsc <- merge(rsc,rsc5,by='Accident_date',all=TRUE)
  rsc <- merge(rsc,rsc6,by='Accident_date',all=TRUE)
  rsc <- merge(rsc,rsc7,by='Accident_date',all=TRUE)
  rsc[is.na(rsc)] <- 0
  ## calculate the index
  for(i in 1:nrow(rsc)){
    rsc$rscindex[i] <- (rsc$rsc0[i] * 0 + rsc$rsc1[i] * 1 + rsc$rsc2[i] * 2 + rsc$rsc3[i] * 3 + +
rsc$rsc4[i] * 4 + rsc$rsc5[i] * 5 + rsc$rsc6[i] * 6 + rsc$rsc7[i] *
7)/(rsc$rsc0[i]+rsc$rsc1[i]+rsc$rsc2[i]+rsc$rsc3[i]+rsc$rsc4[i]+rsc$rsc5[i]+rsc$rsc6[i]+rsc$r
sc7[i])
  }
## On-scene Road markings
  Osrm1 <- sqldf('select Accident_date, COUNT(Road_markings) as Osrm1 from df where
Road_markings=1 group by Accident_date')
  Osrm0 <- sqldf('select Accident_date, COUNT(Road_markings) as Osrm0 from df where
Road_markings=0 group by Accident_date')

##merge counted data
  Osrm <- merge(Osrm1,Osrm0,by='Accident_date', all=TRUE)

Osrm[is.na(Osrm)] <- 0
  ## calculate the index
  for(i in 1:nrow(Osrm)){
    Osrm$Osrmindex[i] <- (Osrm$Osrm1[i] * 1 + Osrm$Osrm0[i] *
0)/(Osrm$Osrm1[i]+Osrm$Osrm0[i])
  }
## Road_catseyes
  Rc1 <- sqldf('select Accident_date, COUNT(Road_catseyes) as Rc1 from df where
Road_catseyes=1 group by Accident_date')
  Rc0 <- sqldf('select Accident_date, COUNT(Road_catseyes) as Rc0 from df where
Road_catseyes=0 group by Accident_date')

##merge counted data
  Rc <- merge(Rc1,Rc0,by='Accident_date', all=TRUE)
```

```
Rc[is.na(Rc)] <- 0
 ## calculate the index
 for(i in 1:nrow(Rc)){
   Rc$Rcindex[i] <- (Rc$Rc1[i] * 1 + Rc$Rc0[i] * 0)/(Rc$Rc1[i]+Rc$Rc0[i])
 }
## Road geometric details
 rgd0 <- sqldf('select Accident_date, COUNT(Road_geometric_details) as rgd0 from df where Road_geometric_details=0 group by Accident_date')
 rgd1 <- sqldf('select Accident_date, COUNT(Road_geometric_details) as rgd1 from df where Road_geometric_details=1 group by Accident_date')
 rgd2 <- sqldf('select Accident_date, COUNT(Road_geometric_details) as rgd2 from df where Road_geometric_details=2 group by Accident_date')
 rgd3 <- sqldf('select Accident_date, COUNT(Road_geometric_details) as rgd3 from df where Road_geometric_details=3 group by Accident_date')
 rgd4 <- sqldf('select Accident_date, COUNT(Road_geometric_details) as rgd4 from df where Road_geometric_details=4 group by Accident_date')
 rgd <- merge(rgd0,rgd1,by='Accident_date', all=TRUE)
 rgd <- merge(rgd,rgd2,by='Accident_date',all=TRUE)
 rgd <- merge(rgd,rgd3,by='Accident_date',all=TRUE)
 rgd <- merge(rgd,rgd4,by='Accident_date',all=TRUE)
 rgd[is.na(rgd)] <- 0
 ## calculate the index
 for(i in 1:nrow(rgd)){
   rgd$rgdindex[i] <- (rgd$rgd0[i] * 0 + rgd$rgd1[i] * 1 + rgd$rgd2[i] * 2   +   rgd$rgd3[i] * 3 +   rgd$rgd4[i] * 4)/(rgd$rgd0[i]+rgd$rgd1[i]+rgd$rgd2[i]+rgd$rgd3[i]+rgd$rgd4[i])
 }
```

```
num.ac <- Num_accident$n
 Roadmark <- Osrm$Osrmindex
 cateyes <- Rc$Rcindex
 road.geom <- rgd$rgdindex
 r.surface <- rsc$rscindex
 rt<- rt$rtindex
 weather <- wc$wcindex
 ## calculate the index
 index_df <- data.frame(num.ac, Roadmark, cateyes, road.geom, r.surface, rt, weather)#put all the index in one data frame
 #head(index_df, n=10)
write.csv(index_df,"E:\\FYP\\index_df.csv")
```

Stage 2 Fit distribution

```
library(MASS)
 library(COMPoissonReg)
 library(fitdistrplus)
 library(mpcmp)
 library(kableExtra)
 library(glm2)

## read data
 d = data.frame(read.csv("E:\\FYP\\index_df.csv"))
d = d[, c(2:12)]
```

```
## fit the distribution
 pois.fit<-fitdist(d$num.ac,"pois")

fit.pois=rbind(as.numeric(pois.fit$estimate),as.numeric(pois.fit$sd),NA,NA,pois.fit$loglik,pois.fit$aic,pois.fit$bic)

nb.fit<-fitdist(d$num.ac,"nbinom")

fit.nb=rbind(as.numeric(nb.fit$estimate[2]),as.numeric(nb.fit$sd[2]),as.numeric(nb.fit$estimate[1]),as.numeric(nb.fit$sd[1]),nb.fit$loglik,nb.fit$aic,nb.fit$bic)

comp.fit<-glm.cmp(d$num.ac ~ 1)
 sum.comp <- summary(comp.fit)

fit.comp=rbind(summary(comp.fit)$coefficients[1],summary(comp.fit)$coefficients[2],as.numeric(summary(comp.fit)$nu),NA,NA,sum.comp$aic,sum.comp$bic)



results=data.frame(cbind(fit.pois,fit.nb, fit.comp))
 rownames(results)=c("par1","SE.par1","par2","SE.par2","Loklik","AIC","BIC")
 colnames(results)=c("Poisson","Negative Binomial", "COM-Poisson")
 kable(results, caption = "Goodness of fit based on the simulated COM-Poisson distributed data set")
comp.fit.12<-glm.cmp(d$num.ac ~ d$ac.cause+d$ac.type)
 fit.comp.12=rbind(NA,NA,sum.comp$aic,sum.comp$bic)
```

```
results=data.frame(cbind(fit.comp.12))

 rownames(results)=c("R^2","Loklik","AIC","BIC")

 colnames(results)=c("COM-Poisson")

 kable(results, caption = "COM-Poisson")
```

Stage 3 All Possible Regression

```
install.packages("stringr")
install.packages("COMPoissonReg")
install.packages("dplyr")

library(stringr)
library(COMPoissonReg)
library(dplyr)

# read the data

d = data.frame(read.csv("C:\\Users\\HWLOK\\Desktop\\OU
sem4\\FYP_new\\index_df.csv"))
d = d[, c(2, 6:10, 12)]
colnames(d) = c("Y", str_c("X", 1:6)) # rename the column names

# generate a data frame containing all the possible predictor combinations.


predictors = str_c("X", 1:6)
m = matrix(NA, nrow=6, ncol=1)
for (a in 1:6) {
  temp = as.matrix(combn(predictors, a))
  if (nrow(temp) < 6) {
    dummy = matrix(NA, nrow = 6 - nrow(temp), ncol = ncol(temp))
    temp = rbind(temp, dummy)
  }
  m = cbind(m, temp)
}
m = as.data.frame(t(m[,-1]))
print(dim(m))
all(duplicated(m)==FALSE) # there is no duplicated record => all the predictor combinations
are distinct
# write.csv(m,"C:\\Users\\HWLOK\\Desktop\\OU sem4\\FYP_new\\predictors.csv",
row.names = FALSE)

options(COMPoissonReg.optim.method = "SANN")

# fit all possible regression

m2 = matrix(NA, nrow=63, ncol=7)
for(i in 1:nrow(m)) {
  cat("fitting model", i, "\n")
  f = m[i, ]
  f = f[!is.na(f)]
  cmp.out = glm.cmp(reformulate(f, response = "Y"), data=d)
  m2[i, 1] = length(coef(cmp.out)) - 2
  m2[i, 2] = toString(reformulate(f, response = "Y"))
  # y.hat = predict(cmp.out, newdata=d)
  # m2[i, 3] = (sum((y.hat-mean(d$Y))^2)/sum(((d$Y-mean(d$Y))^2)))
  # m2[i, 4] = 1 - (1-as.numeric(m2[i, 3]) * (nrow(d)-1)) / (nrow(d) - as.numeric(m2[i, 1]) -
1)
  s = summary(cmp.out)
  m2[i, 5] = s$aic
  m2[i, 6] = s$bic
  m2[i, 7] = cmp.out$loglik
}

colnames(m2) = c("p", "formula", "R-Squared", "Adj. R-Squared", "AIC", "BIC", "Log
Likelihood")

# write.csv(m2, "C:\\Users\\HWLOK\\Desktop\\OU sem4\FYP_new\\SANN.csv",
row.names = FALSE)

# model selection
```

```
model = data.frame(read.csv("C:\\Users\\HWLOK\\Desktop\\OU
sem4\\FYP_new\\SANN.csv"))

which(model$AIC == min(model$AIC))
which(model$BIC == min(model$BIC))
which(model$Log.Likelihood == max(model$Log.Likelihood))
```

Stage 4 Descriptive and Exploratory Analysis

```
library(dplyr)

library(tidyr)

library(ggplot2)

 library(epiDisplay)

library(splines2)

library(gam)

library(readxl)

 library(GeneCycle)

 library(scales)

 library(tidyverse)

 library(sqldf)

 library(GGally)


index_df= data.frame(read.csv("E:\\FYP\\index_df.csv"))


## Correlogram

 cor.plot = ggpairs(index_df, title= "Correlogram")

 cor.plot


##Boxplot
```

```
ggplot()+
  geom_boxplot(aes(x=index_df$num.ac))+
  xlab("Number of accident")


ggplot()+
  geom_boxplot(aes(x=index_df$r.surface))+
  xlab("Road surface")
ggplot()+
  geom_boxplot(aes(x=index_df$rt))+
  xlab("Road type")
hist(index_df$road.geom,xlab = "Road Geometric", main = "Frequency of Index road
Geometric detail",col = "red")
hist(index_df$weather,xlab = "Weather", main = "Frequency of Index Weather",col = "red")
hist(index_df$Roadmark,xlab = "On-Scene Roadmark", main = "Frequency of Index On
Scene Roadmark",col = "red")
hist(index_df$cateyes,xlab = "Road Cat eyes", main = "Frequency of Index Road Cat eyes",col
= "red") ##scatter plot
 ggplot()+
   geom_point(aes(y=index_df$num.ac,x=index_df$ac.cause))+
   ylab("Number of accident")+
   xlab("Accident Cause")
 ggplot()+
   geom_point(aes(y=index_df$num.ac,x=index_df$ac.type))+
   ylab("Number of accident")+
   xlab("Accident type")
 ggplot()+
   geom_point(aes(y=index_df$num.ac,x=index_df$dam.RT))+
   ylab("Number of accident")+
   xlab("Damage of road type")
 ggplot()+
   geom_point(aes(y=index_df$num.ac,x=index_df$Roadmark))+
   ylab("Number of accident")+
   xlab("On-secene Road Markings")
 ggplot()+
   geom_point(aes(y=index_df$num.ac,x=index_df$cateyes))+
   ylab("Number of accident")+
   xlab("Road catseyes")
 ggplot()+
   geom_point(aes(y=index_df$num.ac,x=index_df$road.geom))+
   ylab("Number of accident")+
   xlab("Road geomtric details")
 ggplot()+
   geom_point(aes(y=index_df$num.ac,x=index_df$r.surface))+
   ylab("Number of accident")+
   xlab("Road surface")
 ggplot()+
   geom_point(aes(y=index_df$num.ac,x=index_df$rt))+
   ylab("Number of accident")+
   xlab("Road type")
 ggplot()+
   geom_point(aes(y=index_df$num.ac,x=index_df$Vehicle.t))+
   ylab("Number of accident")+
   xlab("Vehicle type")
 ggplot()+
   geom_point(aes(y=index_df$num.ac,x=index_df$weather))+
   ylab("Number of accident")+
   xlab("Weather")
```

Reference

Abdel-Aty, M. A., & Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention, 32*(5), 633-642.

Akın, D. (2011). *Analysis of highway crash data by Negative Binomial and Poisson regression models.* Paper presented at the Second International Symposium on Computing in Science and Engineering, Kusadasi, Izmir, Turkey.

AlKheder, S., & Al-Rashidi, M. (2020). Bayesian hierarchical statistics for traffic safety modelling and forecasting. *International journal of injury control and safety promotion, 27*(2), 99-111.

Anderson, I. B., Bauer, K. M., Harwood, D. W., & Fitzpatrick, K. (1999). Relationship to safety of geometric design consistency measures for rural two-lane highways. *Transportation Research Record, 1658*(1), 43-51.

Arab News. (2017, May 11). *Car accidents kill over 9,000 people in 2016*. https://www.arabnews.com/node/1097886/saudi-arabia

Basu, S., & Saha, P. (2017). Regression models of highway traffic crashes: a review of recent research and future research needs. *Procedia engineering, 187*, 59-66.

Cafiso, S., Di Graziano, A., Di Silvestro, G., La Cava, G., & Persaud, B. (2010). Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables. *Accident Analysis & Prevention, 42*(4), 1072-1079.

Chang, L.-Y. (2005). Analysis of freeway accident frequencies: Negative Binomial regression versus artificial neural network. *Safety science, 43*(8), 541-557.

Conway, R. W., & Maxwell, W. L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering, 12*(2), 132-136.

El-Basyouny, K., & Sayed, T. (2006). Comparison of two Negative Binomial regression techniques in developing accident prediction models. *Transportation Research Record, 1950*(1), 9-16.

World Life Expectancy. (2018). *World Health Rankings*. World Health Rankings. https://www.worldlifeexpectancy.com/saudi-arabia-road-traffic-accidents

Famoye, F., Wulu, J. T., & Singh, K. P. (2004). On the generalized Poisson regression model with an application to accident data. *Journal of Data Science, 2*(3), 287-295.

Farag, S. G., & Hashim, I. H. (2017). Safety performance appraisal at roundabouts: Case study of Salalah City in Oman. *Journal of Transportation Safety & Security, 9*(sup1), 67-82.

Geedipally, S. R., Lord, D., & Dhavala, S. S. (2012). The Negative Binomial-Lindley generalized linear model: Characteristics and application using crash data. *Accident Analysis & Prevention, 45*, 258-265.

Guikema, S. D., & Goffelt, J. P. (2008). A flexible count data regression model for risk analysis. *Risk Analysis: An International Journal, 28*(1), 213-223.

Gupta, R. C., Sim, S., & Ong, S. (2014). Analysis of discrete data by Conway–Maxwell Poisson distribution. *AStA Advances in Statistical Analysis, 98*(4), 327-343.

Haleem, K., Abdel-Aty, M., & Mackie, K. (2010). Using a reliability process to reduce uncertainty in predicting crashes at unsignalized intersections. *Accident Analysis & Prevention, 42*(2), 654-666.

Haleem, K., Abdel-Aty, M., & Santos, J. (2010). Multiple applications of multivariate adaptive regression splines technique to predict rear-end crashes at unsignalized intersections. *Transportation Research Record, 2165*(1), 33-41.

Hou, Q., Tarko, A. P., & Meng, X. (2018). Investigating factors of crash frequency with random effects and random parameters models: New insights from Chinese freeway study. *Accident Analysis & Prevention, 120*, 1-12.

Jamal, A., Mahmood, T., Riaz, M., & Al-Ahmadi, H. M. (2021). GLM-Based Flexible Monitoring Methods: An Application to Real-Time Highway Safety Surveillance. *Symmetry, 13*(2), 362.

Joshua, S. C., & Garber, N. J. (1990). Estimating truck accident rate and involvements using linear and Poisson regression models. *Transportation planning and Technology, 15*(1), 41-58.

Kamaluddin, N. A., Andersen, C. S., Larsen, M. K., Meltofte, K. R., & Várhelyi, A. (2018). Self-reporting traffic crashes–a systematic literature review. *European transport research review, 10*(2), 1-18.

Kumara, S., & Chin, H. C. (2003). Modeling accident occurrence at signalized tee intersections with special emphasis on excess zeros. *Traffic injury prevention, 4*(1), 53-57.

Li, X., Lord, D., Zhang, Y., & Xie, Y. (2008). Predicting motor vehicle crashes using support vector machine models. *Accident Analysis & Prevention, 40*(4), 1611-1618.

Li, Z., Wang, W., Liu, P., Bigham, J. M., & Ragland, D. R. (2013). Using geographically weighted Poisson regression for county-level crash modeling in California. *Safety science, 58*, 89-97.

Lord, D. (2006). Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accident Analysis & Prevention, 38*(4), 751-766.

Lord, D., Geedipally, S. R., & Guikema, S. D. (2010). Extension of the application of Conway-Maxwell-Poisson models: Analyzing traffic crash data exhibiting underdispersion. *Risk Analysis: An International Journal, 30*(8), 1268-1276.

Lord, D., Guikema, S. D., & Geedipally, S. R. (2008). Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis & Prevention, 40*(3), 1123-1134.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation research part A: policy and practice, 44*(5), 291-305.

Lord, D., Washington, S., & Ivan, J. N. (2007). Further notes on the application of zero-inflated models in highway safety. *Accident Analysis & Prevention, 39*(1), 53-57.

Lord, D., Washington, S. P., & Ivan, J. N. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention, 37*(1), 35-46.

Maher, M. J., & Summersgill, I. (1996). A comprehensive methodology for the fitting of predictive accident models. *Accident Analysis & Prevention, 28*(3), 281-296.

Malyshkina, N. V., Mannering, F. L., & Tarko, A. P. (2009). Markov switching Negative Binomial models: an application to vehicle accident frequencies. *Accident Analysis & Prevention, 41*(2), 217-226.

Miaou, S.-P. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus Negative Binomial regressions. *Accident Analysis & Prevention, 26*(4), 471-482.

Montella, A., Colantuoni, L., & Lamberti, R. (2008). Crash prediction models for rural motorways. *Transportation Research Record, 2083*(1), 180-189.

Mustakim, F., Abdullah, A. H., Yunus, R., Abdullah, M. E., Deris, M. b. M., & Ahmad4, S. K. (2014). Accident Prediction Models for Unsignalised Junction in

Malaysia Rural Roadways.

Mustakim, F., Abdullah, A. H., Yunus, R., Abdullah, M. E., Mat Deris, M., & Ahmad Khalid, S. K. (2014). Accident prediction models for unsignalised junction in Malaysia rural roadways.

Naghawi, H. (2018). Negative Binomial regression model for road crash severity prediction. *Modern Applied Science, 12*(4), 38.

Obaidat, M. T., & Ramadan, T. M. (2012). Traffic accidents at hazardous locations of urban roads. *Jordan Journal of Civil Engineering, 159*(700), 1-12.

Oh, J., Washington, S. P., & Nam, D. (2006). Accident prediction model for railway-highway interfaces. *Accident Analysis & Prevention, 38*(2), 346-356.

Park, B.-J., & Lord, D. (2009). Application of finite mixture models for vehicle crash data analysis. *Accident Analysis & Prevention, 41*(4), 683-691.

Poch, M., & Mannering, F. (1996). Negative Binomial analysis of intersection-accident frequencies. *Journal of transportation engineering, 122*(2), 105-113.

Rakha, H., Arafeh, M., Abdel-Salam, A., Guo, F., & Flintsch, A. (2010). Linear regression crash prediction models: issues and proposed solutions. *Efficient Transportation and Pavement Systems: Characterization, Mechanisms, Simulation and Modeling*, 241-256.

Sellers, K. F., & Shmueli, G. (2010). A flexible regression model for count data. *The Annals of Applied Statistics, 4*(2), 943-961.

Shaon, M. R. R., Qin, X., Shirazi, M., Lord, D., & Geedipally, S. R. (2018). Developing a random parameters Negative Binomial-lindley model to analyze highly over-dispersed crash count data. *Analytic methods in accident research, 18*, 33-44.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., & Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 54*(1), 127-142.

Wood, G. (2002). Generalised linear accident models and goodness of fit testing. *Accident Analysis & Prevention, 34*(4), 417-427.

Xie, Y., Lord, D., & Zhang, Y. (2007). Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis. *Accident Analysis & Prevention, 39*(5), 922-933.