Regression Analysis for Estimation of the Influencing Factors on Road Accident Injuries in Oman

Poisson Regression Model and Poisson's Alternative Models,

Submitted in Fulfilment of the Requirements of
the Master by Research Degree in Statistics

Author
Rahma Ahmed Al-Jabri
Department of Mathematical Sciences
University of Essex
20/09/2015

Abstract

Road safety programs use statistical models to predict the occurrence of accidents and casualties and to identify the influencing factors that affect their occurrence. They are also used to identify the causes of an accident and the hazardous locations where more accidents happen (the hot spots or black spots). Causal factors could depend on human behaviour, road geometries, traffic volumes, weather, or the interactions among these. For decision makers, it is very important to understand road patterns and behaviours to apply road safety improvements and road maintenance activities efficiently. Statistical modelling of road safety is conducted by taking the data of past accidents and the attributes of many sites and using them to produce the best prediction models. The objective is to discover the relationship between a function of the dependent variable (e.g., expected number of accidents at a certain point), $E(Y_i) = \lambda_i$, in relation to number of covariates, $X_{i1}, X_{i2}, X_{i3}, ...X_{ik}$ that are assumed to have an effect on the dependent variable Y_i. It is a standard practice in road safety research to model accident counts Y_i as Poisson distributed random variables that $Y_i \sim Pois(\lambda_i)$ corresponds to a random distribution of the accidents over time and space. Accident data have often been shown to exhibit overdispersion, which make it essential to use alternatives of Poisson to model such data. In this research, we apply the Poisson regression model and its alternatives in addition to the binary and ordered probit logistic regression model.

Key words: Accidents injuries, Poisson regression, Over/Underdispersion, Quasipoisson models, Negative binomial models, Hurdle models, Zero-inflated models

Acknowledgements

This research was sponsored by the Directorate of Technical Colleges in Oman, Ministry of Manpower. I express my gratitude to all the concerned people who supported and facilitated the scholarship processing. I also thank Dr. Abdullah Al-Maniri from the Research Council in Oman who helped in obtaining the data from the Royal Oman Police. I also like to show gratitude to Professor Edward Codling, the board memeber, for his guidance and support. I particularly thank my supervisor, Dr. Aris Perperoglou, for his comments on earlier versions of the manuscript, although any errors are my own and should not tarnish the reputations of these esteemed persons. Last but not least, my gratitude is for the head of the Department of Mathematical Sciences, Professor Abdellah Salhi, for sharing his pearls of wisdom with us during the course of this research.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Traffic accidents in Oman have gradually developed to be a serious issue that is insisting for more focus. The huge social and economic losses caused by these accidents are a real burden to the welfare of citizens as well as to the overall development of Oman. According to Al-Lamki (2010),'Oman was ranked 5^{th} in the list of countries with the highest road accident rates in the world according to the statistics of World Health Organization as quoted by Oman Tribune on 24^{th} Feb, 2010'. Accordingly, the government in Oman has been setting many traffic policies and interventions to control the situation. It also has recently launched the road-safety research program under the responsibility of the research council in Oman to encourage and support research in the area. Road traffic accidents(RTA) are, in fact, a global problem that has caught the attention of many countries. According to Al-Maniri(2013), RTA was ranked as the eleventh leading cause of deaths and the ninth leading cause of disability around the globe in 2002. He wrote that every year, approximately 1.2 million people die because of RTA with a global mortality rate of 19 per 100,000 people. Most developed countries have already established research and interventions that are effective in reducing the dimensions of the problem; however, in the low- and middle-income countries, the research and interventions to the problem are still in the baseline (Lawrence et al., 1992; Elvik, 1995; Wayatt et al., 1996; DiGuiseppi et al., 1997; VÃd'gverket, 2006). On March 10, 2010, the United Nations held a general assembly that adopted a text proclaiming the Decade of Action for

Road Safety (2011-2020) to reduce traffic-related deaths and injuries. The assembly was scheduled to discuss a new resolution on road safety. The key components of the initiative for the decade are to include governmental technical assistance, road traffic education, road safety curriculum development, helmets for kids, safe routes to school, research and evaluation, and setting up non-profit helmet assembly plants that employ the physically disabled. According to Al-Lamki (2010), the Moscow Declaration showed concern that more than 90 percent of RTA occur in low- and middle-income countries. The annual cost of these deaths and injuries run to over 65 billion US dollars. The declaration stated that by the year 2020, without appropriate action, road traffic deaths will become one of the leading causes of death in low- and middle-income countries. Most of the developed countries with declining death rates have been using Haddon's Matrix that was developed by William Haddon in 1970. Haddon's Matrix analyses injury by looking at certain factors which, when simplified, are 'Host or Road User Factors, Vector or Vehicle Factors, Physical or Road Factors, and Socio-economic Environmental Factors' in the horizontal axis of the matrix and 'Pre-Event (Crash), during the Event and Post-event' in the vertical axis (Haddon, 1980).

The countries of the Gulf Cooperation Council (GCC) in particular suffer a real growing problem of RTA with a true dearth of research in the road safety field. In the GCC countries, the growth of economy, development of infrastructure, and motorisation over the last four decades have resulted in a massive increase in automobile usage and ownership and consequently in an increase in the RTA. In Oman, road accidents have indeed become a major concern for families and communities (Al-Qareeni, 2008). Unfortunately, Oman has the highest fatality rate (23.7/100,000 pop.), and despite this fact, very little has been done to establish the baseline facts of the problem. According to Islam and Al-Hadhrami(2012), no comprehensive work has been undertaken on level trends and determinants of RTA and its causality in Oman because of the scarcity of data in the past. According to Al-Lamki(2010), the fatality rate in Oman has now reached 30 per 100,000 people, 127 per 100,000 vehicles, and 111 per 100,000 licensed drivers, compared with 14, 17, and 21, respectively, in the United States. She said that road safety education in Oman is needed for adults as well as prelicence age groups, and this should be made a priority in conjunction with speed management.

She stated that there is a need to work harder at changing drivers' behaviours and attitudes towards the risks associated with high speed. In their study, Al-Ismaily and Probert (1998), stated that there are on average 230 vehicles per 1,000 people in Oman, higher than many middle-income countries. Motorisation level showed increasing trend levels in Oman, and between 2000 and 2009, it increased by 26 percent. Road construction programmes have increased in parallel with other development programmes in Oman. For example, in the 1960s, Oman had only 10 kilometres of paved roads, which increased to more than 25,000 in 2009 according to the statistics of the MoNE, 2010.

In Oman, the most common cause of road deaths is excessive speeding, which caused 57 percent of all deaths in 2007. Overtaking comes next, followed by drivers' carelessness, improper acts by drivers, and vehicle condition. Thus, four out of the top five causes of death in Oman are road user errors, and they constituted 89 percent of the causes of road deaths in 2007. This does not include other road user related factors, such as tiredness and alcohol, which constitute a total of only 2 percent of the causes of road deaths in Oman (Al-Lamki, 2010; Al-Maniri,2012). Between 2000 and 2009, the population of Oman increased by 21.6 percent, with a mean annual increase by approximately 2.0 percent. On the other hand, the automobile fleet in the country increased by 52.4 percent, with a mean annual increase by 4.3 percent between 2000 and 2009. Reporting of RTA related data by royal oman police (ROP) is thought to be of high coverage because of the enforcement of a law that car insurance companies, garages,and repair establishments could not accept a vehicle involved in an accident for insurance claim and repair if a police report is not produced. Similar traffic laws exist in other Arabian Gulf countries (El-Sadig et al., 2002; Ziyad and Akhtar, 2011). They suggested that Oman is currently in an era where it needs to establish different public transportation alternatives as the only modes available now are only the public-shared taxis and buses. Oman envisions a 200-kilometre railway track for trains containing goods between the industrial cities of Sohar and Barka. The unavailability of public transport causes inconvenience, and excessive dependence on private cars leads to heavy traffic, a large number of accidents, and high individual expenditure on transport. Belwal, R. and Belwal, S.(2010) stated that the Sultanate

of Oman is marked by the second highest death toll from traffic accidents in the world coming after Libya, which is reported as the worse performing nation in this respect. In October 2011, Oman experienced 670 traffic accidents in which 110 persons were killed and 903 seriously injured. Given the size of the population, these numbers signal very high casualty rates. (OECD countries with populations that are three times as large may experience similar numbers in any given period.)

This research is aimed to analyse the accident related factors that influence the occurrence of human casualties in Omani road. We want to study these factors in order to be able to identify the magnitude and the direction of the effect of each factor on road accidents injuries. We have the number of injuries per accident as our dependent variable which is a discrete count variable. We start analyses with the assumption that the number of injuries per accidents follows a Poisson distribution given that they occur among large number of trails and include many zeros in the process. We also want to identify the causes of an accident and the hazardous locations where more accidents happen (the hot spots or black spots). Causal factors of an accident could depend on human behaviour, road geometries, traffic volumes, weather, or the interactions among these. For decision makers, it is very important to understand road patterns and behaviours to apply road safety improvements and road maintenance activities efficiently. Statistical modelling of road safety is conducted by taking the data of past accidents and the attributes of many sites and using them to produce the best prediction models. The objective is to discover the relationship between a function of the dependent variable (e.g., expected number of accidents at a certain point), $E(Y_i) = \lambda_i$, in relation to number of covariates, $X_{i1}, X_{i2}, X_{i3}, ...X_{ik}$ that are assumed to have an effect on the dependent variable Y_i. It is a standard practice in road safety research to model accident counts Y_i as Poisson distributed random variables that $Y_i \sim Pois(\lambda_i)$ corresponds to a random distribution of the accidents over time and space. Accident data have often been shown to exhibit overdispersion, which make it essential to use alternatives of Poisson to model such data. In this research, we apply the Poisson regression model and its alternatives in addition to the binary and ordered probit logistic regression model.

1.1 Source of RTA Data in Oman

The main sources of traffic accident data in Oman is the Department of Statistics at the Royal Oman Police (ROP). Information about the crash, persons, and vehicles are recorded in three separate databases, including variables such as time, location, day, date, number of deaths, number of injuries, cause of the accident, and other variables. Data collected include a summary narrative of the accident, a detailed scaled scene diagram, and information on accident events. Recently, with the cooperation of the research council, the work is progressing to compile the three databases to attain a full view of information on accidents easily, accurately, and comprehensively. The aim is to develop a common and efficient platform on which researchers and regional and international educational institutes can obtain information about the traffic situation in Oman. However, this process may cause some lack in the data and issues with reliability until it is completely finalised and a report of limitations is available. Some variables, such as using seatbelts and mobiles, also seem to be not correctly recorded. That's mostly because there is no clear mechanism on how these should be collected. Other sources of accident data are patient records in the Ministry of Health and information about roads and road geometry from the Ministry of Transportation. For this research, we have obtained permission to use data of accidents from the Statistics Department at ROP with support from the research council in Oman. We also received published booklets of tabulated data for accidents in different years.

Chapter 2

Descriptive Statistics of the Data

The statistics in Table 1 of RTA for the years 2002-2012 that were published by ROP during the GCC traffic week 2013 showed that there is a gradual increase with fluctuations in the number of accidents through the period, whereas the toll of accident injuries and deaths have rapidly increased. Figure 1 shows the trend of deaths in RTA during the period more clearly. This came along with a rapid increase in the number of speeding offences through the period according to the same source. The statistics published indicate that males constitute 85 percent of the deaths and 73 percent of the injuries. These statistics also revealed that around 80 percent of deaths and injuries are drivers and passengers, whereas pedestrians make 22 percent of deaths and only 6 percent of injuries. The statistics showed that overspeeding is the main reason for fatal accidents. The next cause is negligence and then overtaking. This means that 90 percent of RTA in Oman are caused by wrong driver behaviours in the road. The highest portion of RTA casualties was in the category of young people as 47 percent of the fatalities are in the age category 26-50 and 32 percent in the age category 16-25. For injuries, statistics showed that 46 percent are in the age category 26-50 and 37 percent in the age category 16-25. This indicates that more than 80 percent of RTA casualties were in the category of young people. The same source showed a comparison between the traffic in the years 2010-2011 and 2011-2012 in terms of number of total accidents, deaths, and injuries. It is found that there was an increase in the latter year of 2, 15, and 3 percent respectively. The numbers are really

high if we compare them with countries that have bigger populations and heavier traffic volumes, such as the UK. The reported road casualties in Great Britain as released in November, 2012 by the Department of Transport showed that in the traffic year ending June 2012, 1,790 people were killed in reported road accidents, a 6 percent drop from the traffic year ending June 2011 (1,901). Overall, for the year ending June 2012, there were 148,100 reported injury accidents and 199,740 reported casualties of any severity (slight injuries, serious injuries and fatalities), falls of 3 percent and 4 percent respectively from the year ending June 2011. In comparison, motor vehicle traffic levels have risen by 0.1 percent in the year ending June 2012. The ratio of Oman's population (3,314,001 by 2012) to the UK's population (63,887,988 by 2012) is 5.2 percent while the ratio of the RTA in Oman to that of the UK in the same year is 63.81 percent.

Year	Total	Deaths	Injuries
2002	9107	580	7907
2003	10197	578	6735
2004	9460	637	6636
2005	9247	689	6658
2006	9869	681	7548
2007	8816	798	8531
2008	7982	951	10558
2009	7253	953	9783
2010	7571	820	10066
2011	7719	1056	11437
2012	8209	1139	11618
Average	8675.455	807.45	8861.55
St.Dev	941.19	185.18	1821.22

Table 2.1: Deaths, injuries and total of accidents from 2002 to 2012

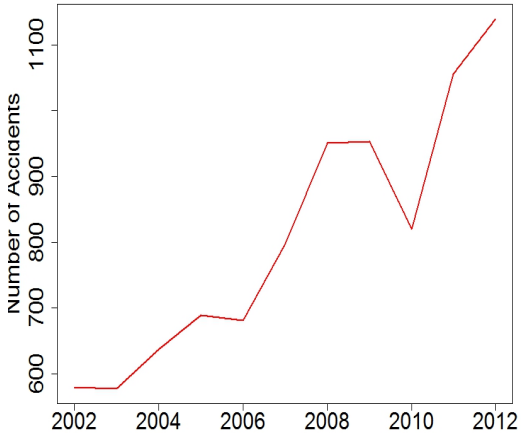


Figure 2.1: Deaths in Road Accidents from 2002 to 2012

Statistics in Table 2 show a sharp increase in the number of offences, and according to the source, most of these are speeding offences. Driving tests doubled during the period with an average of 340,523 and a standard deviation of 80,846. The total licences also doubled with an average of 753,245 and a standard deviation of 172,338. In the same way, the number of new vehicles increased with an average of 108,045 and a standard deviation of 48,863. Vehicle inspections also increased

during the period with an average of 29,944 and a standard deviation of 14,379.

Table 2.2: Statistics of driving tests, driving licences, new vehicles, vehicle inspections and offences during the period 2002-2012

Year	Driv.Tests	Licences	New Vehicle	Vehic.Insp	Offence
2002	229363	541752	39376	38752	266792
2003	233401	567953	42561	38451	409081
2004	250400	578808	57130	36923	452267
2005	269188	620025	73421	41704	476221
2006	322808	667917	104891	52089	1433862
2007	336723	718697	136516	43229	1589895
2008	393796	777741	177441	21343	2067173
2009	408721	840002	127001	13971	2070347
2010	410824	909978	120662	13446	2205623
2011	436480	989279	137968	14556	2529634
2012	454052	1073538	171532	14927	3239953
Average	340523.27	753244.55	108045.36	29944.64	1521895.27
St. Dev	80846.97	172338.04	46588.72	13709.47	958048.37

2.1 The Population

Research studies are initiated by setting questions on issues that are of great relevance to specific groups of individuals known as research population. A research population is the collection of individuals or elements that is relevant to the main focus of a scientific query. To get information about this population accurately, we need to study information of every element. However, commonly the large sizes of populations, make it difficult for researchers to test every individual in the population because it is too expensive and time consuming. This is the reason why researchers rely on sampling techniques. A research population is also known as a well defined collection of individuals or objects known to have similar characteristics. All individuals or objects within a certain population usually have a common, binding characteristic or trait. The **target population** of this study includes all the accidents that happen in Oman's roads with or without casualties. The individual element of the population is an accident on Oman's roads. However, as some accidents were not reported for different reasons,**the accessible population** is used, which is every accident that happens in Oman's roads that is reported in sufficient details to ROP.

2.2 Types of Variables

In a research study, the measurable attributes of interest that varies for the elements in the population are called **variables**. Mostly, the variables can be described as discrete or continuous. Discrete variables can assume only certain values: fixed and countable. Continuous variables, on the other hand, can take an infinite number of values. Examples of discrete variables include number of patients, students, accidents, citizens, sex, income level, and treatment group. Common examples of continuous variables include age, height, weight, grades, blood pressure, and time. Discrete variables are divided into two categories: nominal(unordered) and ordinal(ordered). Nominal variables take values, such as yes/no, female/male, or treatment A/B/C, where the order of the categories is not important. A nominal variable that takes only two possible values is called binary. Ordinal variables take naturally ordered values, such as Statistics course (I, II, III) or education level (less than high school, high school, college, graduate school). Ordering among the categories is meaningful, but spacing between categories may be arbitrary. The variables at focus in this study are time, road type, description of the road where the accident happened, gender, age, nationality of the driver, the licence type, the type and the cause of accident, vehicle type, the number of involved vehicles, and the number of involved persons. The majority of these variables are categorical variables that simply indicate the existence of a certain condition, such as the road type at the accident location. Table 3 summarises the variables of the study with basic descriptive statistics.

Table 2.3: Descriptive statistics of the variables

abbrev	variable namme	n observ.	median	mad	min	max	skew	kurtosis
injuryc	injury count	24191	1	1.48	0	6	1.55	2.29
injdgre	injury severity level	24187	5	0.00	1	5	-1.31	1.04
dead no	death count	24187	0	0.00	0	8	5.40	46.00
year	year of accident	24192	2010	1.48	2009	2012	0.20	-1.04
time1	time of accident in hour	24192	14	7.41	1	24	-0.32	-0.71
day	day of accident	24192	4	2.97	1	7	-0.03	-1.26
month	month of accidet	24187	6	4.45	1	12	0.16	-1.22
roadtyp	road type	24192	1	0.00	1	3	0.72	-0.71
acctyp	accident type	24192	2	1.48	1	5	0.40	-1.33
loctndsc	accident location description	24128	1	0.00	1	4	1.88	1.93
age	age of driver	24043	28	8.90	10	96	1.30	1.86
cause	cause of accident	24160	1	0.00	1	8	2.20	4.78
gender	gender of driver	24186	1	0.00	1	2	2.57	4.62
nationlty	nationality of driver	24192	1	0.00	1	2	1.74	1.03
climcond	weather condition	24192	1	0.00	1	2	6.60	41.54
licens	license availability	24192	1	0.00	1	2	4.47	18.02
vehctyp	vehicle type	24192	1	0.00	1	6	1.47	0.91
hrmdtl	harm detail	24192	2	1.48	1	5	0.86	0.68
prsns	persons count	24192	2	1.48	1	49	5.28	66.96
vhcls	vehicle counts	24192	1	0.00	1	37	6.83	267.69

This research is aimed to analyse human road casualties in Oman by observing the number of injuries in an accident, the injury degree, and the number of deaths. Before proceeding to further analysis, we perform descriptive statistics of the variables included in the study. Table 4 shows the distribution of RTA by the number of injury cases, and it can be seen that 26.737 percent of RTA resulted in no injury, whereas 73.263 percent included at least one injury case. Figure 2 illustrate more the distribution of the accidents by number of injuries per accident.

Table 2.4: Distribution of RTA during 2009-2012 by cases of injury

Injury Cases/per accident	Frequency	Cum.Freq.	Percentage	Cum.Percent.
0	6468	6468	26.737	26.737
1	10365	16833	42.847	69.584
2	3573	20406	14.770	84.354
3	1633	22039	6.750	91.104
4	930	22969	3.844	94.949
5 or more	1223	24192	5.051	100.000

In the data set, the persons' degree of injury in RTA is classified into five categories: fatal injury, severe injury, moderate injury, slight injury,and no injury. Table 5 shows that in 49.50 percent of the accidents, an injury occurred. Of these, 25.25 percent are accidents with slight injuries. The

fatal and severe accidents represent 8.40 percent of the total accidents, and 14.839 percent are accidents with moderate injury.

Table 2.5: Distribution of RTA during 2009-2012 by injury degree

Degree of Injury/per accident	Frequency	Cum.Freq.	Percentage	Cum.Percent.
Fatal	1091	1091	4.511	4.511
Severe	940	2031	3.886	8.401
Moderate	3589	5620	14.839	23.236
Slight	6109	11729	25.257	48.493
No Harm	12458	24187	51.507	100.000

Figure 3 shows the distribution of RTA by the number of death cases per accident. As can be seen in the data, 89.713 percent of the RTA did not involve death cases while 8.744 percent included one death case. Only 1.542 percent of the accidents included two or more cases. Table 6 illustrate the levels of the variables in the research and the frequency of the accident per level.

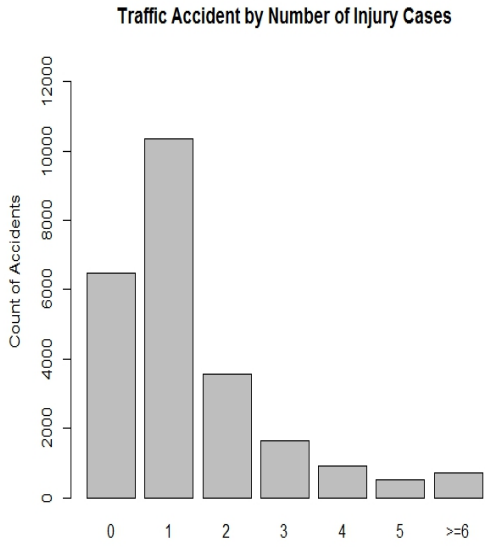


Figure 2.2: Distribution of RTA during 2009-2012 by number of injury cases

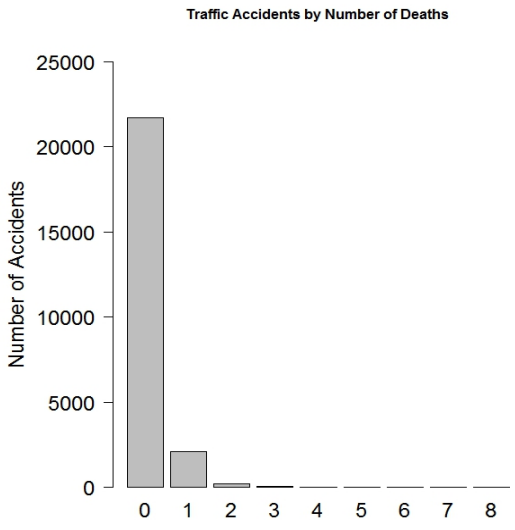


Figure 2.3: Distribution of RTA during 2009âĂŞ2012 by deaths cases

Table 2.6: Definition of the Response/Exploratory Variables in the Study

Number	Variable	Variable Definition	Levels' Code/Value	
1	age	Age of Driver's at fault	Years	
2	time1	Time of Accidents	Hours	
3	day	Day of Accidents	1=Sat→7=Fri	
4	month	Month of the Accident	1=Jan→12=Dec	
5	gender	Gender of Driver	1= Male	21641
			2= Female	2545
6	nationality	Nationality	1=Omani	20036
			2= Non-Omani	4156
7	licens	license status	1=Licensed	23139
			2=Unlicensed	1053
8	roadtyp	Type of Road	1=Main	14571
			2=Sub	9164
			3=Unpaved	457
9	location	Location Description	1=Straight	19062
			2=Roundabout	1291
			3=Intersection	1603
			4=Others	2172
10	cimcond	Climate Condition	1=Normal	23672
			2=Ubnormal	520
11	acctyp	Type of Accident	1=Vehicle Collision	10483
			2=Run-Over (person or animal)	3167
			3=Over-Turn	3897
			4=Fixed Object Collision	5575
			5=Motorcycle/Bicycle	1070
12	cause	Cause of Accident	1=Over-speeding	12398
			2=carelessness	7458
			3=Safe Distance	1425
			4=Overtaking	1133
			5=Fatigue/alcohol	638
			6=Climate Condition	217
			7=Vehicle	671
			8=Road	220
13	vehtyp	Vehicle Type	1=Saloon	15308
			2=Pick up	2648
			3=Four wheel	2750
			4=Bi/Motorcycle	738
			5=Heavy	2018
			6=Other	730
14	hrmdtl	Vehicle Harm Detail	1=Severe	7194
			2=Moderate	10924
			3=Slight	4469
			4=No harm	1194
			5=Not Specified	411

Chapter 3

Generalized Linear Models (GLMs)

In road safety programs, statistical modelling is conducted by taking the data of past accidents and the attributes of many sites and using them to produce the best of prediction models. The objective is to discover the relationship between a function of the dependent variable (e.g., expected number of accidents at a certain point, i), $E(Y_i) = \lambda_i$, in relation to the number of covariates, $X_{i1}, X_{i2}, X_{i3}, ...X_{ik}$, that are assumed to have an effect on the dependent variable Y_i. It is a standard practice in road safety research to model accident counts or accident casualties, Y_i as Poisson distributed random variables that $Y_i \sim Pois(\lambda_i)$ corresponds to a random distribution of the accidents over time and space. When modelling count data, (e.g. the number of occurrences of an event in a fixed period and when the outcome variable is a count with a low arithmetic mean (typically <10), standard ordinary least squares regression may produce biased results. Accident data have often been shown to exhibit overdispersion, which make it more appropriate to use alternatives of Poisson to model such data. Poisson and its alternative models belong to the general linear model(GLM) that describes a linear model that has the stochastic component with a non-normal distribution of errors.

3.1 General Linear Models(OLS)

The GLM is a linear model that relates the response y to several predictors and can be written in the form

$$y_i = x_i\beta + \epsilon_i, \quad where \quad \epsilon_i \sim N(0, \sigma^2) \tag{3.1}$$

where $i = 1, 2, ...n$; Y_i is a dependent variable, X_i is a vector of k independent variables, vector β is a vector that represents linear parameter estimates to be computed, and the vector ϵ_i are zero-mean stochastic errors. Generally, the component of the normal linear regression model (OLS) can be distinguished in two parts:

1. **Stochastic or random component** The Y_i are usually assumed to have independent normal distributions with $E(Y_i) = \mu_i$, with constant variance σ^2, or $Y_i \sim N(\mu_i, \sigma_i^2) iid.$

2. **Deterministic or systematic component** This specifies the explanatory variable or the independent variables for the model: $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$ The covariates x_i are combined linearly with the coefficients to form the linear predictor$\eta_i = X_i\beta$.

We view the two probability distributions that are most commonly used to model the stochastic component of linear models. The normal or Gaussian distribution is most familiar because it is used with GLMs. The binomial distribution describes the stochastic component for logistic regression models with two outcomes.

The Normal Distribution

The normal distribution has two parameters, namely, the mean (μ) and the variance (σ^2), with a probability density function (PDF) as we can see in Figure 4 and Figure 5.

$$f(y; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\} \quad -\infty < y < \infty \tag{3.2}$$

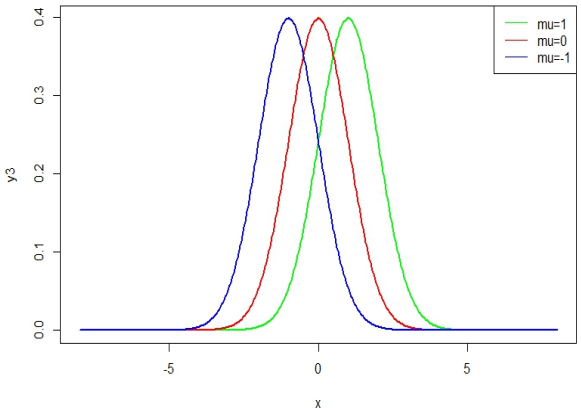


Figure 3.1: Normal distribution with different means (mu)

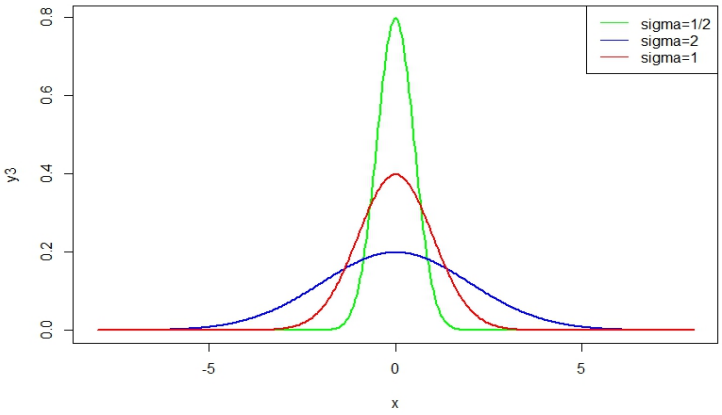


Figure 3.2: Normal distribution with different values for variance ($sigma^2$).

The Binomial Distribution

The binomial distribution is based on a Bernoulli trial, which is a random experiment in which there are only two possible outcomes: success (S) and failure (F). We conduct the Bernoulli trial and let

$$y_i = \begin{cases} 1 & \text{if the } i^{th} \text{ outcome is S} \\ 0 & \text{if the } i^{th} \text{ outcome is F} \end{cases}$$

If the probability of 'success' is π, then the probability of 'failure' must be $(1-\pi)$ and the probability mass function (PMF) is

$$f(y_i;\pi) = \pi^{y_i}(1-\pi)^{(1-y_i)} \quad where \quad y = 0,1 \quad i = 1,2,...,n. \tag{3.3}$$

The binomial distribution has two parameters which are

1. the sample size(n), which is the number of sampling units per experimental unit,

2. the probability of the success outcome of the response event (π),

and the probability mass function PMF is

$$f(y_i;\pi) = \binom{n}{y_i}\pi^{y_i}(1-\pi)^{(n-y_i)} \ \ where \ \ y = 0, 1, ... \ \ i = 1, 2, ..., n. \tag{3.4}$$

As we can see in Figure 6, the sample size must be a positive integer and the probability can only have a value between zero and one. The two values of the parameters are not constrained by each other's value. The response variable can be the proportion or the number (out of the n sampling units) of successes for a given experimental unit. It could also be a binary response of an experiment with two outcomes, such as yes/no or success/failure that is usually coded with 0/1. The response variable, the number of successes, has a mean $(n\pi)$ and a variance $[n\pi(1-\pi)]$.

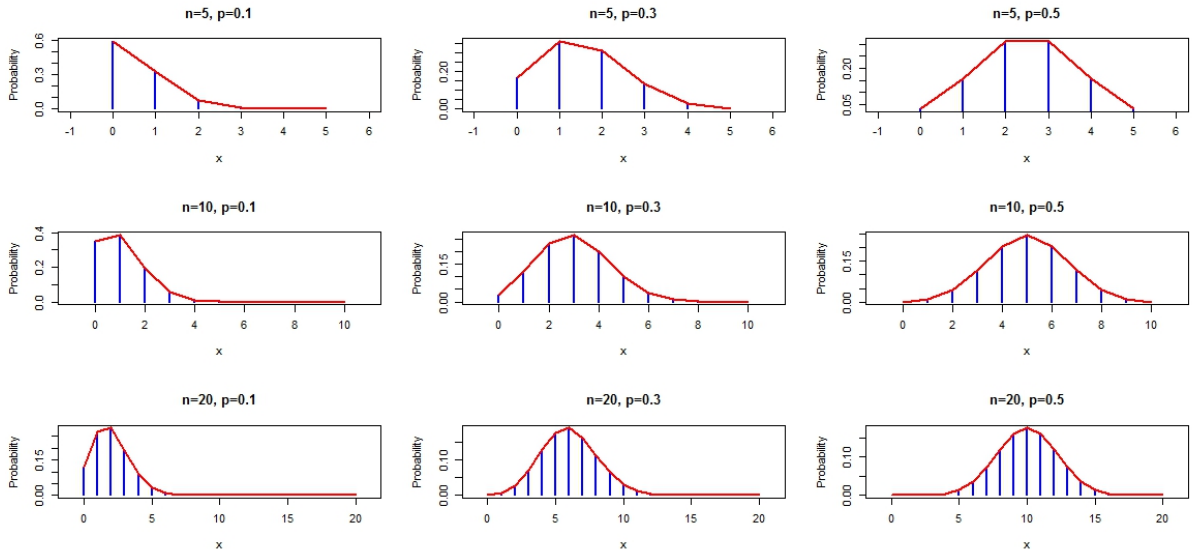


Figure 3.3: The binomial distribution for various values of π and n

The normal distribution is continuous and symmetric with no restrictions on the possible values of the response variable. On the other hand, the binomial distribution is discrete and asymmetric except when $\pi = 0.5$, and the response variable is limited to the integer values between zero and the sample size inclusive. The sample size is an explicit parameter of the binomial distribution. Binomial data are often approximated by a normal distribution when both success and failure mean

counts, $n\pi$ and $n(1-\pi)$, are greater than five. The binomial distribution is reasonably symmetric and multivalued when this is the case.

3.2 Generalized Linear Models (GLM)

The traditional linear model, $y_i = x_i\beta + \epsilon_i$, which we have just introduced above, is not suitable for modelling the data that is not following normal distribution. Furthermore, the discrete response variables, either count or categorical, can't assume normality by their nature and consequently generalised linear models(GLMs) were developed to model the event count models based on Poisson, binomial and beta-binomial distributions (King, 1989), and Winkelmann and Zimmermann(1994). The generalised linear model extends OLS such that $g(\cdot)$ is a link function that maps the relation between the non-normal stochastic component y with the systematic part of linear predictors $\eta_i = x_i\beta$ where $y_i \sim EF(\lambda_i, \phi)$. EF$(\lambda_i, \phi)$ is an exponential family distribution and ϕ is a known or unknown scale parameter.

Assume that $y_1, y_2, ..., y_n$ are n independent observations of a random variable Y_i. In the GLM, we assume that Y_i has a normal distribution with mean μ_i and variance σ^2, and we assume that the expected value μ_i is a linear function of k predictors that take values $x_i = (x_{i1}, x_{i2}, ..., x_{ik})$ for the i^{th} case, so that

$$\mu_i = x_i\beta \tag{3.5}$$

where β is a vector of unknown parameters. We will generalise this in two steps, dealing with the stochastic and systematic components of the model.

The Exponential Family

If the observations are assumed to be coming from a distribution of the exponential family, their probability density function can be written in the form

$$f(y_i) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\theta)} + c(y_i, \phi)\right\} \tag{3.6}$$

where θ is the natural parameter of location,ϕ is the dispersion parameter, and $c(y;\phi)$ are unknown functions. The natural parameter θ relates to the mean, and the scale parameter ϕ relates to the variance of the exponential family distribution members. One of its properties is that if y follows a distribution from the exponential family, we can write $var(y) = V(\mu_i)$, where V is a known variance function of $\mu_i = E(y_i)$ and ϕ is a scale parameter. Here,θ_i and ϕ are location and scale parameters, and $a_i(\phi), b(\theta_i)$, and $c(y_i, \theta)$ are known functions. In the models considered here, the function $a_i(\phi)$ has the form $a_i(\phi) = \phi/w_i$, where w_i is a known prior weight, usually one. If Y_i has a distribution in the exponential family then it has mean and variance

$$\begin{aligned} E(Y_i) &= \mu_i = b^{'}(\theta_i) \\ var(Y_i) &= \sigma_i^2 = b^{''}(\theta_i) \end{aligned} \tag{3.7}$$

where $b^{'}(\theta_i)$ and $b^{''}(\theta_i)$ are the first and second derivatives of $b(\theta_i)$. The exponential family includes as special cases the normal, binomial, Poisson, exponential, gamma, and inverse Gaussian distributions.

The Link Function

The second element of the generalisation introduces a $one - to - one$ continuous differentiable transformation function $g(\mu_i)$ that is called the link function. The link function plays a role in GLMs such that it maps the relation between the non-normal response y and the linear predictor $\eta = \beta X$, such that

$$\eta_i = g(\mu_i) = x_i\beta \tag{3.8}$$

Here we assume that the transformed mean follows a linear model where the quantity η_i is called the linear predictor, and since the link function is $one - to - one$, we can invert it to obtain

$$\mu_i = g^{-1} = (x_i\beta) \tag{3.9}$$

Here, we do not transform the response y_i but rather its expected value μ_i. The standard linear model(LM) we have described earlier is a generalised linear model with normal errors and identity link. When the link function makes the linear predictor η_i the same as the canonical parameter θ_i, we say that we have a canonical link. The identity is the canonical link for the normal distribution. We will see that the 'logit' is the canonical link for the binomial distribution and the 'log' is the canonical link for the Poisson distribution as mentioned in Table 7.

Table 3.1: Examples of common link functions based on the error distribution

Error distribution	Default link	Used for
Normal(Gaussian)	Identity	normally distributed error
Binomial	Logit	proportions or binary (0,1) data
Poisson	Log	counts (many zeros, various integers)

Maximum Likelihood Estimation

Likelihood is the basic concept when using the maximum likelihood method of fitting and testing models. For discrete data, it is derived from the probability function of the assumed distribution, such as the binomial distribution, that predicts the probability of obtaining specific data values given known values of the parameters. The generalised linear models can all be fitted to data using the same algorithm, a form of iteratively reweighed least squares. Given a trial estimate of the parameters, we calculate the estimated linear predictor $\hat{\eta_i} = x_i\hat{\beta}$ and use that to obtain the fitted values $\hat{\mu_i} = g^{-1}(\hat{\eta_i})$. Using these quantities, we calculate the working dependent variable, the derivative of the link function evaluated at the trial estimate

$$z_i = \hat{\eta_i} + (y_i - \hat{\mu_i})\frac{d\eta_i}{d\mu_i}, \tag{3.10}$$

which is the derivative of the link function evaluated at the trial estimate. Next, we calculate the iterative weights

$$w_i = \pi / b^{''}(\theta_i) \left[\frac{d\eta_i}{d\mu_i}^2 \right], \tag{3.11}$$

where $b^{''}(\theta_i)$ is the second derivative of $b(\theta_i)$ evaluated at the trial estimate, and we have assumed that $a_i(\theta)$ has the usual form ϕ/w_i. This weight is inversely proportional to the variance of the working dependent variable z_i given the current estimates of the parameters, with proportionality factor ϕ. We obtain an improved estimate of ϕ regressing the working dependent variable z_i on the predictors x_i using the weights w_i, that is, we calculate the weighted least-squares estimate

$$\hat{\beta} = (X^{'} W X)^{-1} X^{'} W z, \tag{3.12}$$

where $\mathbf{X}$ is the model design matrix,$\mathbf{W}$is a diagonal matrix of weights with entries w_i, and z is a response vector with entries z_i. The procedure is repeated until successive estimates change by less than a specified small amount (see McCullagh and Nelder, 1989).

Newton-Raphson Method

A linear equation can be estimated numerically and the most popular method for doing this is the Newton-Raphson method.

$$\beta^{(1)} = \beta^{(0)} + [-l^{''}(\beta^{(0)})]^{-1} . l^{'}(\beta^{(0)}) \tag{3.13}$$

Let μ be a column vector of length N with elements $\mu_i = n_i \pi_i$. Note that each element of μ can also be written as $\mu_i = E(y_i)$, the expected value of y_i. Using matrix multiplication, we can show that

$$l'(\beta) = X^{'}(y - \mu) \tag{3.14}$$

is a column vector of length $K + 1$ whose elements are $\dfrac{\partial l(\beta)}{\partial \beta_k}$, as derived in Equation 7. Now, Equation 9 becomes

$$l''(\beta) = -X^{'}WX \tag{3.15}$$

is a $K+1 \times K+1$ square matrix with elements $\frac{\partial^2 l(\beta)}{\partial(\beta)\partial(\beta)}$. Now,Equation 8 can be written as

$$\beta^{(1)} = \beta^{(0)} + [X^{'}WX]^{-1}.X^{'}(y-\mu) \tag{3.16}$$

Tests of Hypotheses

In GLMs, testing the hypothesis of the model's goodness of fit is performed through measuring the Wald tests, likelihood ratio tests, and the deviance statistic.

Waldtests follow immediately from the fact that the information matrix for generalised linear models is given by

$$i(\beta) = \phi^{-1}X^{'}WX \tag{3.17}$$

which is used to calculate standard errors of the estimates, confidence intervals, test statistics, and other statistics of the model that can be derived using the usual likelihood theory. The large sample distribution of the maximum likelihood estimator $\hat{\beta}$ is a multivariate normal with mean β and variance-covariance matrix $(X^{'}WX)^{-1}\phi$

$$\hat{\beta} \sim N_p(\beta, (X^{'}WX)^{-1}\phi) \tag{3.18}$$

and use a z-test to test the significance of a coefficient. Specifically, we test

$$H_0 : \beta_j = 0 \quad versus \quad H_1 : \beta_j \neq 0 \tag{3.19}$$

using the test statistic

$$z_i = \frac{\hat{\beta}_j}{\sqrt{\phi}(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}_{jj}} \tag{3.20}$$

which is asymptotically $N \sim (0,1)$ under H_0

Standard Errors The estimates $\hat{\beta}$ have the usual properties of maximum likelihood estimators. In particular, $\hat{\beta}$ is asymptotically

$$N(\beta, i^{-1}) \tag{3.21}$$

where

$$i(\beta) = \phi^{-1} X^{'} W X \tag{3.22}$$

Standard errors for the β_j may therefore be calculated as the square roots of the diagonal elements of

$$\hat{cov}(\hat{\beta}) = \phi(X^{'}\hat{W}X)^{-1} \tag{3.23}$$

in which $(X^{'}\hat{W}X)^{-1}$ is a by-product of the final IWLS iteration. If ϕ is unknown, an estimate is required.

Likelihood Ratio Tests and The Deviance We will show how the likelihood ratio criterion for comparing any two nested models, say $\omega_1 \subset \omega_2$, can be constructed in terms of a statistic called the deviance and an unknown scale parameter ϕ. Consider first comparing a model of interest ω with a saturated model Ω that provides a separate parameter for each observation. Let $\hat{\mu_i}$ denote the fitted values under ω and let $\hat{\theta}_i$ denote the corresponding estimates of the canonical parameters. Similarly, let $\bar{\mu_O} = y_i$ and $\tilde{\theta}_i$ denote the corresponding estimates under ω. The likelihood ratio criterion to compare these two models in the exponential family has the form

$$2log\lambda = 2\sum_{i=1}^{n} \frac{y_i(\tilde{\theta}_i - \hat{\theta}_i)}{a_i(\phi)} \tag{3.24}$$

Assume as usual that $a_i(\phi) = \phi/w_i$ for known prior weights w_i. Then we can write the likelihood-ratio criterion as follows:

$$D(y, \hat{\mu}) = 2\sum p_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) + b(\tilde{\theta}_i)] \tag{3.25}$$

The numerator of this expression does not depend on unknown parameters and is called the deviance. The likelihood ratio criterion $2logL$ is the deviance divided by the scale parameter ϕ, and is called the scaled deviance. Recall that for the normal distribution we had $\theta_i = \mu_i$, $b(\theta_i) = \frac{1}{2}\theta_i^2$, and $a_i(\phi) = \sigma^2$, so the prior weights are w_i. Thus, the deviance is, the residual sum of squares.

$$D(y, \hat{\mu}) = 2\sum(y_i - \hat{\mu_i}) - \frac{1}{2}y_i^2 + \frac{1}{2}\hat{\mu_i^2} \tag{3.26}$$

$$= 2\sum\left\{\frac{1}{2}y_i^2 - y_i\hat{\mu}_i^2 + \frac{1}{2}\hat{\mu_i}^2\right\} \tag{3.27}$$

$$= 2\sum(y_i - \hat{\mu_i})^2 \tag{3.28}$$

Let us now return to the comparison of two nested models ω_1, with $k1$ parameters, and ω_2, with k_2 parameters, such that $\omega_1 \in \omega_2$ and $k_2 > k_1$. The log of the ratio of maximised likelihoods under the two models can be written as a difference of deviances, since the maximised log-likelihood under the saturated model cancels out. Thus, we have

$$-2log\lambda = \frac{D(w_1) - D(w_2)}{\phi} \tag{3.29}$$

The scale parameter ϕ is either known or estimated using the larger model ω_2.

Large sample theory tells us that the asymptotic distribution of this criterion under the usual regularity conditions is χ_{ν^2} with $\nu = k_2 - k_1$ degrees of freedom. In the linear model with normal errors, we estimate the unknown scale parameter ϕ using the residual sum of squares of the larger model, so the criterion becomes

$$-2log\lambda = \frac{RSS(\omega_1) - RSS(\omega_2)}{RSS(\omega_2)/(n - k_2)} \tag{3.30}$$

In large samples, the approximate distribution of this criterion is χ_i^ν with $\nu = k_2 - k_1$ degrees of freedom. Under normality, however, we have an exact result: dividing the criterion by $k_2 - k_1$

we obtain an F with $k_2 - k_1$ and $n - k_2$ degrees of freedom. Note that as $n \longrightarrow \infty$ the degrees of freedom in the denominator approach one and the function converges to $(k_2 - k_1)\chi^2$, so the asymptotic and exact criteria become equivalent.

Chapter 4

Poisson Regression Analysis for Injury Data

In the literature of road safety modelling, the process of an accident's occurrence is assumed to follow Bernoulli distribution with unequal probability of independent events. The process is also known as Poisson trials which define the number of n trials in which the probability of success π_i varies from trial to trial. According to this approach, a vehicle that enters the road will either have an accident or will not, such that an accident represents a 'success', while a no-accident represents a 'failure'. According to Feller(1968), count data that arise from Poisson trials do not follow a standard distribution. However, the mean and variance of these trials share similar characteristics to the binomial distribution when the number of trials n and the expected value $E(y)$ are fixed. In most cases, these assumptions do not hold for accident data since n is not known with certainty but is an estimated value and varies for each site and the probability of an accident occurrence π_i also varies from one vehicle to another. The probability density function of the Bernoulli distribution is

$$f(y_i;\pi) = \pi^{y_i}(1-\pi)^{(1-y_i)} \quad where \quad y = 0,1 \quad i = 1,2,...,n \tag{4.1}$$

where 1 represents an accident occurrence and 0 represents a no-accident occurrence. The

process results in what are called Bernoulli trials, and when summing these trials of Bernoulli processes,it gives the binomial distribution $B(n, \pi)$ where$\pi$is the average probability of accident occurrence, n is the number of vehicles that enter the road at a specific point of time, area, or segment of the road, and y is the number of accidents at that point.

$$f(y_i;\pi) = \binom{n}{y_i}\pi^{y_i}(1-\pi)^{(n-y_i)} \ \ where \ \ y = 0, 1, ... \ \ i = 1, 2, ..., n \tag{4.2}$$

According to **The Law of Rare Events**, the total number of events follows, approximately, Poisson distribution if an event occurs in any given trial is small. More formally, let $Y_{n,\pi}$ denote the total number of successes in a large number n of independent Bernoulli trials, with the success probability π of each trial being very small

$$lim_{n\to\infty}\left[\binom{n}{y}\binom{\lambda/n}{y}\left(1-\frac{\lambda}{n}\right)^{n-y}\right] = \frac{\lambda^y e^{-\lambda}}{y!}, \tag{4.3}$$

Given the large number of vehicles that pass through a road segment, an accident represents an event that has a low probability to happen. Therefore, the Binomial distribution can be approximated by Poisson distribution where the mean of this distribution is $\lambda = np$ (Olkin et al., 1980). Therefore, Poisson probability distribution is considered the standard discrete distribution used for modelling accidents occurrence data.

$$f(Y = y_i;\lambda) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \ \ where \ \ y = 0, 1, 2, ... \ \ i = 1, 2, ..., n \tag{4.4}$$

4.1 Literature Review

Road accident data encounters many examples of count variables like the number of accidents at a certain point or time, number of casualties, number of deaths, number of injuries, number of vehicles, and number of people involved in an accident and others. Therefore, Poisson distribution has been used as a standard discrete distribution for modelling count data in road safety modelling. It was derived by Poisson(1837) as a limiting case of the binomial distribution. An old classical ap-

plication of the model was to model the number of soldiers in a Prussian army who died from being kicked by mules. Two other early applications of the Poisson regression were discussed by Cochrance (1940) and Jorgenson (1961). Nevertheless, real data have often been shown to exhibit overdispersion, which violates the equal mean-variance restriction of Poisson model. Poisson-gamma/negative binomial(NB) models are usually preferred over Poisson regression models in road safety modelling. The negative binomial distribution is a standard generalisation of the Poisson regression that was derived by Greenwood and Yule(1920)and by Eggenberger and Polya (1923). However, in some cases, real life data shows characteristics of underdispersion where negative binomial models by structure are not suppose to handle. Models like the Restricted Generalised Poisson by Consul and Famoye (1992) and the COM-Poisson distribution that was reintroduced by Sellers and Shemuli (2010) have been used by statisticians to model count data that are characterised by either over- or underdispersion. Another limitation of the Poisson model is that it has an assumption that there is the possibility of a zero counts even if there are no such records in the data. Zero-inflated or zero-altered probability models were applied to capture the apparent 'excess' zeroes that are commonly arise in crash data. Johnson and Kotz (1968) defined a modified Poisson distribution known as Poisson with added zeroes that explicitly accounted for excess zeroes in the data. The 'generalised linear models', of which Poisson regression is a special case, were first introduced by Nelder and Wedderburn (1972) and detailed in McCullagh and Nelder (1983, 1989). The book by Hardin and Hilbe (2012) is a latest good textbook to be referred to for GLMs and extensions. Barbour (1992) and Kingman (1993) provide clear and direct introductions to thePoisson process. Another good introductory textbook of thePoisson process and renewal theory is that written by Taylor and Karlin (1998), while Feller (1971) presents a more advanced treatment of the model. Another comprehensive reference on Poisson and related distributions with analyses is Haight (1967).

According to a review paper of statistical techniques that was conducted by Lord and Mannering (2010), the range of statistical models commonly applied for road accident data included binomial, Poisson, Poisson-gamma (or negative binomial), zero-inflated Poisson and negative binomial models(ZIP and ZINB), multinomial probability models, and many other statistical analysis tools. Jones

and Jorgensen (2002) introduced the potential of a recently developed form of regression models, known as multilevel models, for quantifying the various influences on casualty outcomes. The benefits of using multilevel models to analyse accident data are discussed along with the limitations of traditional regression modelling approaches. Lord et al.(2005, 2007) provided defensible guidance on how to appropriate model crash data, clarifying a collective reluctance to adopt zero-inflated(ZI) models for modelling highway safety data. We first examine the motor vehicle crash process using theoretical principles and a basic understanding of the crash process. It is shown that the fundamental crash process follows a Bernoulli trial with unequal probability of independent events, also known as Poisson trials. It then presents two critical and relevant issues: the maximising statistical fit fallacy and logic problems with the ZI model in highway safety modelling. Famoye et al.(2004) used the generalised Poisson regression(GPR) model for identifying the relationship between the number of accidents and some covariates. They found that based on the test for the dispersion parameter and the goodness-of-fit measure for the accident data, the GPR model performs as good as or better than the other regression models.

The problem of overdispersion is central to all GLMs for discrete responses. Overdispersion in discrete-response models occurs when the variance of the response is greater than the nominal variance. It is generally caused by positive correlation between responses or by an excess variation between response probabilities or counts. The problem with overdispersion is that it may cause underestimation of standard errors of the estimated coefficient vector. A variable may appear to be a significant predictor when it is not. We can recognise possible overdispersion by observing that the value of the Pearson χ^2 or deviance divided by the degrees of freedom $(n-p)$ is larger than one. The quotient of either is called the dispersion. Small amounts of overdispersion are usually of little concern; however, if the dispersion statistic is greater than 2.0, then an adjustment to the standard errors may be required. There is a distinct difference between true overdispersion and apparent overdispersion. Outward indicators, such as a large dispersion statistic, may be only a sign of apparent overdispersion. Apparent overdispersion may arise for different reasons: the model omits important explanatory predictors, the data contain outliers, the model fails to

include enough interaction terms, and a predictor needs to be transformed (to the log or some other scale). The data may be overdispersed if the value of the estimated dispersion after fitting is greater than expected or under-dispersed if the value less than expected. For overdispersion, the simplest remedy is to assume a multiplicative factor in the usual implied variance. As such, the resulting covariance matrix will be inflated, and the estimated dispersion parameter may result from model misspecification rather than overdispersion, indicating that the model should be assessed for appropriateness by the researcher.

4.2 Application and illustration (Injuries Data)

We select a sample of 5,000 accidents from the dataset to avoid the exaggeration of positive significance. We then fit the Poisson model using the glm() function and store it in the object(allP) and use the Anova() function from the car package to test for the overall significance of the included variables. As in Table 8, the analysis of deviance shows that mostly all the variables are significant except for the year, driver's licence status, climate condition, and number of vehicles. We used the step() function with the option 'both', forward and backward selection to obtain the best model. It excluded the nonsignificant variables and provided a model that confirms the result given by the Anova() function. We then used the anova() function in R to test the overall significance of the effect of the removed variables by comparing the deviance of the full model with the deviance of the model chosen by the step() function. The test confirms again that the excluded variables are not statistically significant. The null of the test is that the coefficients of the removed variables are equal to zero. According to the p-value from the test, $0.535 > 0.05$, we cannot reject the null hypothesis, and the result is that the model selected by the step() function is correct. We also used the step() function to check a model with interaction. We want to test the hypothesis that the full model adds explanatory value over the reduced model. That hypothesis is $H_0 : \beta = 0$.

Table 4.1: Analysis of Deviance Table

The Model	Resid.	DfResid.	Dev	Df	Deviance	Pr(>Chi)
1	Full model	9929	6058.2			
2	Best Model	9906	6036.5	23	21.76	0.535

Exposure and Offset

In LM and GLM modelling, an 'offset' is "a quantitative variable whose regression coefficient is known to be 1"(McCullough and Nelder, 1983). In a Poisson regression, the offset is most often used to include exposure time, the Poisson model being for log rate. We model the number of injuries y_i as Poisson(μ_i), where

$$\mu_i = exposure_i = \lambda_i,$$

so that λ_i is the mean injuries rate per accident. Because of the varying exposures(number of persons), we should believe that λ_i, not μ_i, is related to the covariates. Under a log link, we have

$$log\mu_i = \mu_i = o_i + x_i\beta$$

where $o_i = logexposure_i$, o_i is called an offset. An offset is a covariate in the linear predictor whose coefficient is not estimated but assumed to be equal to one. Offsets are very common in Poisson regression because exposure often varies from one observation to the next. For example, suppose y_i is the number of injuries in an accident i; it would be sensible to use the log-population(persons) as an offset to adjust for the effect of varying number of persons in an accident. An offset is a term to be added to a linear predictor, such as in a generalised linear model, with known coefficient 1 rather than an estimated coefficient. The way this works is that the mean value parameter is $n\lambda$. The link function we are using is 'log' (the default for the Poisson family), which makes the linear predictor the same as the canonical parameter, η. Thus we see that $log(exposure_i)$ is just a known constant additive term in the linear predictor. The way R handles such a term in the linear predictor that does not contain an unknown parameter to fit is as an 'offset'. Since the variable n in the math formula is the variable prsns in R, the 'offset' is log(prsns). Thus, the model being fit

is that the linear predictor value of the i^{th} case is

$$\eta_i = log(n_i) + \sum_{i=1}^{k} X_j = d_{ij}\beta_j$$

where the β_j are the regression coefficients and d_{ij} the value of the j_{th} dummy variable for the $i_t h$ case.

It is recommended to use robust standard errors for the parameter estimates to control the mild violation of the distribution assumption that the variance equals the mean.(For further details, see Cameron and Trivedi, 2009.) We use R package 'sandwich' below to obtain the robust standard errors and calculated the p-values accordingly. Together with the z-scores (Wald test) and p-values, we have also calculated the 95% confidence interval using the parameter estimates and their robust standard errors. Deviance residuals are approximately normally distributed if the model is specified correctly. In our example, it shows a little bit of skewness since the median is not quite zero.

The typical Poisson regression model expresses the log outcome rate as a linear function of a set of predictors. The β coefficients are interpreted as increasing or decreasing the log odds ratio of an event, and exp β (the odds multiplier) are used as the odds ratio for a unit increase or decrease in the explanatory variable.

$$Log(injuries) = \beta_0 + \beta_1 X1 + ... + \beta_k X_k$$

by exponentiating we get the model's equation

$$injuries = e^{\beta_0} + e^{\beta_1 X1} + ... + e^{\beta_k X_k}$$

The information on deviance is also provided. We can use the residual deviance to perform a goodness-of-fit test for the overall model. The residual deviance is the difference between the deviance of the current model and the maximum deviance of the ideal model where the predicted values are identical to the observed. Therefore, if the residual difference is small enough, the

goodness of fit test will not be significant, indicating that the model fits the data. In our model, the residual difference is significant evidenced with $p - value < 0.001$. The model doesn't seem to fit the data well because the goodness-of-fit, chi-squared test is not statistically significant. We need to determine if there are omitted predictor variables, if our linearity assumption holds, and/or if there is an issue of overdispersion. Figure 7 shows that the model is reasonably fitting the data but is not satisfactory.

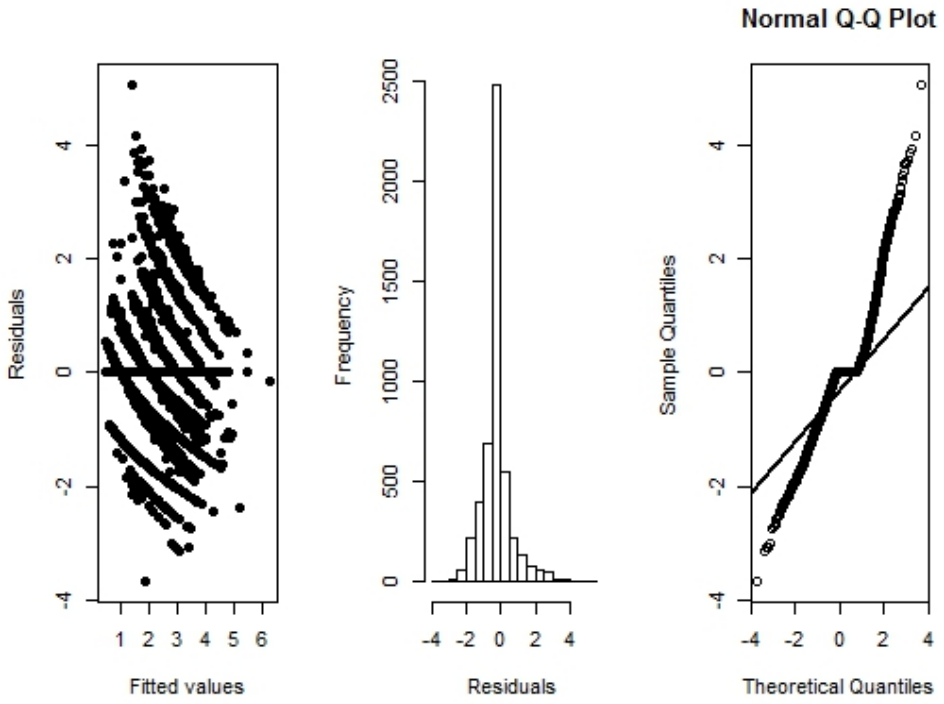


Figure 4.1: Poisson model diagnostic statistics

Over/Underdispersion Test

The standard Poisson GLM models the (conditional) mean $E[y] = \mu$ with the assumption that it is equal to the variance $VAR[y] = \mu$. However, in real-life data, this assumption rarely holds and therefore, overdispersion test is used to test this assumption of equidispersion against the alternative that the variance is of the form

$$VAR[y] = \mu + \alpha \times (\mu)$$

According to this test, if $\alpha > 0$ that is a proof of overdisersion and when $\alpha < 0$ then an underdispersion case is present. The coefficient alpha can be estimated by using OLS regression and tested with the corresponding $t(orz)$ statistic which is asymptotically standard normal under the null hypothesis. The trafo function is specified in two forms:$(\mu) = \mu^2$ corresponds to a negative binomial (NB) model with quadratic variance function (NB2), while $(\mu) = \mu$ corresponds to a NB model with linear variance function (called NB1 by Cameron and Trivedi, 2005) or quasi-Poisson model with dispersion parameter

$$VAR[y] = (1 + \alpha) \times \mu = dispersion \times \mu$$

The simple principle behind this is that in a Poisson model, the mean, $E(Y) = \mu$, and the variance, $Var(y) = \mu$, are equal. The test simply tests this assumption as a null hypothesis against an alternative where $Var(y) = \mu + c \times f(\mu)$ where the constant $c < 0$ means underdispersion and $c > 0$ means overdispersion. The resulting test is equivalent to testing $H_0 : c = 0$ vs. $H_1 : c0$ and the test statistic used is a t statistic which is asymptotically standard normal under the null. Here, we clearly see that there is evidence of underdispersion since c is estimated to be -0.482, an evidence against the assumption of equidispersion (i.e.,$c = 0$). Using $trafo = 1$ will actually do a test of $H_0 : c\times = 1$vs. $H_1 : c * 1$ with $c* = c + 1$, which of course has the same result as the other test apart from the test statistic being shifted by one. The reason for this, though, is that the latter corresponds to the common parametrisation in a quasi-Poisson model. Both tests indicate that there is a problem with overdispersion presence as can be seen in Table 9.

Model Coefficient Interpretation

In Table 10, we look to the model selected after removing and testing some of the basic regression aspects that are mentioned above like the interactions and dispersion.

Table 4.2: Overdispersion test

H_a:true dispersion is greater than 1	sample estimates: dispersion = 0.523 z = -46.6912, p-value = 1
H_a: true alpha is greater than 0	sample estimates: alpha = -0.477 z = -46.6912, p-value = 1

$$log(\hat{\mu_i}) = -57.037 + 0.028X_{runover} - 0.265X_{overturn} + ...$$

Since the estimate of $\beta > 0$, the increase in years result in an increase in the expected number of injuries resulted from accidents as $exp(0.028) = 1.028$. More specifically, for a one-unit increase in years, the number of injuries will increase by a multiple of 0.028.

The coefficient of the indicator variables like accident type are explained in a different way. The indicator variable acctypRun-over compares between acctyp = 'Run-over' and the reference group acctyp = 'two-vehc collision'. The coefficient of run-over type accidents is -0.265. Exponentiating the value gives $exp(-0.265) = 0.767$, and this means that the expected log count for a run-over accident is ≈ 0.767. Hence, the relative risk of having an injury case in a run-over accident is less by 23.3% than if the accident type is a collision with another vehicle. Similarly, the coefficient of turn-over type accidents is 0.190, and exponentiating the value gives $exp(0.190) = 1.209$. This is the log count of the relative risk of having an injury case in a run-over accident, which is less by 20.9% than if the accident type is a collision with another vehicle. The coefficient of collision with fixed object type accidents is , and exponentiating the value gives $exp(-0.074) = 0.929$. This is the log count of the relative risk of having an injury case in a fixed object accident, which is less by 7.1% than if the accident type is a collision with another vehicle. The coefficient of collision with fixed object type accidents is , and exponentiating the value gives $exp(-0.085) = 0.919$. This is the log count of the relative risk of having an injury case in a fixed object accident, which is less by 8.1% than if the accident type is a collision with another vehicle. The indicator variable Location discription, abbreviated loctnRoundabout, compares between loctn = 'Roundabout' and the refer-

ence group loctn = 'Straight Road'. The coefficient of collision with fixed object type accidents is , and exponentiating the value gives $exp(-0.079) = 0.924$. This is the log count of the relative risk of having an injury case in a roundabout accident, which is less by 7.6% than if the accident type is a straight road. The coefficient of triangle accidents type is 0.077, and exponentiating the value gives $exp(0.077) = 1.080$. This is the log count of the relative risk of having an injury case in a triangle accident, which is more by 8.0% than if the accident type is a straight road. The coefficient of other location type of accidents is -0.013, and exponentiating the value gives $exp(-0.013) = 0.987$. This is the log count of the relative risk of having an injury case in an other locations accident, which is less by 1.3% than if the accident location is a straight road.

The indicator variable cause of accident abbreviated cause, compares between cause = 'Carelessness' and the reference group cause = 'High-speed'. The coefficient of carelessness accidents type is 0.032 and exponentiating the value gives $exp(0.032) = 1.033$. This is the log count of the relative risk of having an injury case in a safe distance accident, which is higher by 3.3% than if the cause of the accident is high-speed. The coefficient of safe distant accidents is -0.049, and exponentiating the value gives $exp(-0.049) = 0.952$. This is the log count of the relative risk of having an injury case in a safedistant accident, which is less by 4.8% than if the accident type is a high-speed accident. The coefficient of overtaking accidents is 0.070 and exponentiating the value gives $exp(0.070) = 1.072$. This is the log count of the relative risk of having an injury case in a overtaking accident, which is more by 7.2% than if the accident type is a high-speed accident. The coefficient of fatigue/alcohol accidents is -0.597, and exponentiating the value gives $exp(-0.597) = 0.550$. This is the log count of the relative risk of having an injury case in a fatigue/alcohol accident, which is less by 45% than if the accident type is a high-speed accident. The coefficient of other accidents is -0.037, and exponentiating the value gives $exp(-0.037) = 0.963$. This is the log count of the relative risk of having an injury case in an other accident, which is less by 3.7% than if the accident type is a high-speed accident.

The coefficient of gender family accidents is 0.088, and exponentiating the value gives $exp(0.088) =$

1.092. This is the log count of the relative risk of having an injury case in a female accident, which is more by 9.2% than if the accident type is a male accident. The coefficient of non-Omani driver accidents is -0.059, and exponentiating the value gives $exp(-0.059) = 1.092$. This is the log count of the relative risk of having an injury case in a 0.943 accident, which is more by 5.7% than if the accident involves an Omani driver.

The indicator variable vehicle type, abbreviated vehctyp, compares between vehctyp= 'Four-weel' and the reference group vehicle = 'Saloon'. The coefficient of vehicle type is 0.032, and exponentiating the value gives $exp(-0.044) = 0.957$. This is the log count of the relative risk of having an injury case in a four-wheel accident, which is more by 4.3% than if the accident vehicle is a small saloon. The coefficient of vehicle type is 0.038, and exponentiating the value gives $exp(0.038) = 1.039$. This is the log count of the relative risk of having an injury case in a bicycle or motorcycle accident, which is more by 3.9% than if the accident vehicle is small saloon. The coefficient of vehicle type is -0.034 and exponentiating the value gives $exp(-0.34) = 0.967$. This is the log count of the relative risk of having an injury case in a heavy vehicle accident, which is more by 3.3% than if the accident vehicle is a small saloon. The coefficient of vehicle type is -0.254 and exponentiating the value gives $exp(-0.254) = 0.775$. This is the log count of the relative risk of having an injury case in a four-wheel accident, which is less by 22.5% than if the accident vehicle is small saloon.

The indicator variable vehicle harm level, abbreviated hrmdtl, compares between hrmdtl= 'moderate' and the reference group vehicle = 'sever. The coefficient of vehicle type is -0.011, and exponentiating the value gives $exp(-0.011) = 0.890$. This is the log count of the relative risk of having an injury case in an accident where the harm to the vehicle is moderate, which is less by 11% than if the accident vehicle has severe harm. The coefficient of vehicle type is -0.263, and exponentiating the value gives $exp(-0.263) = 0.769$. This is the log count of the relative risk of having an injury case in an accident where the harm to the vehicle is slight, which is less by 23.1% than if the accident vehicle has severe harm. The coefficient of vehicle type is -0.379, and exponentiating the value gives $exp(-0.379) = 0.685$. This is the log count of the relative risk of having an injury case

in an accident where there is no harm to the vehicle, which is less by 31.5% than if the accident vehicle has severe harm. The accidents appear to result in less injuries if the accidents involve more vehicles. The coefficient of number of vehicles, abbreviated vhcls, is -0.097, and exponentiating the value gives $exp(-0.097) = 0.907$. This means that the log count of the relative risk of having an injury case in an accident decreases by a factor of 0.093.

Table 4.3: Poisson Regression results using rubust S.Error

	Estimate	Odds Ratio	Robust SE	Pr(>|z|)	95% LL	95% OR LL	95% UL	95% OR UL
(Intercept)	−57.037	0	12.472	0.00000	−81.482	0	−32.591	0
year	0.028	1.029	0.006	0.00001	0.016	1.016	0.040	1.041
acctypRun-Over	−0.265	0.767	0.034	0	−0.333	0.717	−0.198	0.820
acctypOver-Turn	0.190	1.209	0.030	0	0.132	1.141	0.248	1.281
acctypFixed Object Collision	−0.074	0.929	0.030	0.014	−0.133	0.875	−0.015	0.985
acctypMotorcycle/Bicycle	−0.085	0.919	0.035	0.015	−0.154	0.858	−0.016	0.984
loctndscRoundabout	−0.079	0.924	0.030	0.008	−0.138	0.871	−0.020	0.980
loctndscTriangle	0.077	1.080	0.022	0.0004	0.034	1.035	0.120	1.128
loctndscOther	−0.013	0.987	0.022	0.544	−0.056	0.945	0.030	1.030
causecarelessness	0.032	1.033	0.018	0.072	−0.003	0.997	0.068	1.070
causesafedist.	−0.049	0.952	0.038	0.195	−0.124	0.884	0.025	1.026
causeovertaking	0.070	1.072	0.027	0.009	0.017	1.017	0.122	1.130
causefatigue/alcohol	−0.597	0.550	0.088	0	−0.770	0.463	−0.424	0.654
causeother	−0.037	0.963	0.030	0.212	−0.096	0.908	0.021	1.022
genderfemle	0.088	1.092	0.017	0.00000	0.055	1.056	0.121	1.129
nationalitynon-omani	−0.059	0.943	0.016	0.0003	−0.091	0.913	−0.027	0.974
vehctypFour Wheel	−0.044	0.957	0.019	0.021	−0.082	0.922	−0.007	0.993
vehctypBi/Motorcycle	0.038	1.039	0.040	0.347	−0.041	0.960	0.117	1.124
vehctypHeavy	−0.034	0.967	0.024	0.150	−0.080	0.923	0.012	1.012
vehctypOther	−0.254	0.775	0.074	0.001	−0.399	0.671	−0.110	0.896
hrmdtlModerate	−0.116	0.890	0.015	0	−0.145	0.865	−0.087	0.917
hrmdtlSlight	−0.263	0.769	0.022	0	−0.307	0.736	−0.220	0.803
hrmdtlNo harm	−0.379	0.685	0.036	0	−0.450	0.638	−0.307	0.735
vhcls	−0.097	0.907	0.018	0.00000	−0.133	0.876	−0.062	0.940

4.3 Poisson's Alternative Models (Injury data)

Real-life count datasets are mostly characterized by overdispersed and excess of zeros. Based on the result above the standard poison is not providing a best fit of the data. We therefore, employ other count models with different distributional assumptions which are generalizations and extensions to Poisson. In addition to applying Poisson, we here, in brief introduce quassipoisson, negative binomial, Poisson logit hurdle, and negative binomial logit hurdle model and zero-inflated Poisson and negative binomial. We fit all these models to our dataset and compare their performance in providing a best fit and best explaining the effects of different factors of an accident in the occurrence of injuries. While these GLMs all have the same mean function, the zero-augmentation also employs the same mean function for the count part.

Quasi-Poisson Model

Quasi-poisson model works by using the same mean regression function and the variance function of the Poisson GLM. It differs from Poisson in that it accounts for overdispersion through leaving the dispersion parameter ϕ unrestricted rather than assuming ϕ equals to 1 and estimate it from the data. This method gives the same coefficient estimates as the standard Poisson model but inference is adjusted for overipersion. In other words, the Quasi-poisson estimating function is the same as of the Poisson model and do not correspond to models with specified likelihoods. In R, glm function is also used to fit Quasipoisson regression models by setting family=quasipoisson.

Negative Binomial Model

Negative binomial is developed to model the overdispersed data with the assumption that unexplained variability is present is present among observations that have the same predicted value. This unexplained variability between observations result in larger variance than assumed by Poisson. The conditional mean of the NB given the predictors should be equal to that given by standard Poisson model while the conditional variance will be larger in the negative binomial model. The function of the variance for the negative binomial is given by $\mu + \alpha \times \mu^2$ rather than μ as in Poisson

regression. The α parameter is a measure of overdispersion, and when it is equals zero there is no overdispersion, and the negative binomial reduces to standard Poisson. If the value of α is greater than zero is an indication of overdispersion where larger values indicate more overdispersion. The interpretation of regression coefficients for the negative binomial model is identical to that for the standard Poisson model. The parametrization of its probability density function is

$$P(Y_i = y_i|\mathcal{X}_\rangle) = \frac{\theta^\theta \mu_i^{y_i}{}_i \Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)(\mu_i + \theta)^{\theta + y_i}} \tag{4.5}$$

with mean μ, shape parameter θ and $Gamma$ is the gamma function. This parametrization is of NB1 type and thus is another special case of the GLMs framework. It also has $\phi = 1$ but with variance $V(\mu) = \mu + \frac{\mu}{\theta}$.

Zero-Inflated Models

Zero-inflated count models were derived to model the excess of zeros and to deal with overdipsersion (Mullahy 1986; Lambert 1992). They are based on the assumption that the data are a mixture produced by two separate data generating processes: one generates counts with either Poisson or negative binomial distribution and the other generates only zeros with binomial distribution. Each observation is generated in a two possible data generation process; the result of a Bernoulli trial determines which process is used. For observation i, process 1 is chosen with probability π_i and process 2 with probability $1-\pi_i$. Process 1 generates only zero counts, whereas process 2 $f(z(y_i, x_i)$, generates counts from either a Poisson or a negative binomial model.

$$y_i \sim \begin{cases} 0 \quad \text{with probability } \pi_i \\ f(y_i|X_i) \quad \text{with probability 1-}\pi_i \end{cases} \tag{4.6}$$

The probability of $Y_i = y_i|\mathcal{X}$ is

$$P(Y_i|y_i|x, z) = \begin{cases} \pi(\gamma^{'} z_i) + 1 - \pi(\gamma^{'} z_i) f_{zero}(0, x) \quad \text{if} \quad \text{y}_i = 0 \\ 1 - \pi(\gamma^{'} z_i) f_{count}(y_i, x) \quad \text{if} \quad \text{y}_i > 0 \end{cases} \tag{4.7}$$

The probability π depends on the characteristics of observation i and π_i is a function of $z'\gamma$, where z' is the zero-inflated vector of covariates and *gamma* is the zero-inflated vector of coefficients in the model. The scalar product $z_i^{'}$ to the probability π_i are related through the link function, that is specified as either the logistic function (the logit function) or the standard normal cumulative distribution function (the probit function). The zero-inflated Poisson model (ZIP) has a mean and variance as

$$\begin{aligned} E(y_i|X_i, Z_i) &= \mu_i(1-\pi_i) \\ V(y_i|X_i, Z_i) &= \mu_i(1-\pi_i)(1+\mu(\pi_i) \end{aligned} \tag{4.8}$$

and the zero-inflated negative binomial (ZINB) mean and variance are

$$\begin{aligned} E(y_i|X_i, Z_i) &= \mu_i(1-\pi_i) \\ V(y_i|X_i, Z_i) &= \mu_i(1-\pi_i)(1+\mu(\pi_i+\alpha)) \end{aligned} \tag{4.9}$$

Hurdle Models

Hurdle models, originally developed by Mullahy(1986) is a class of models that is extension to Poisson which is derived to capture both overdispered and excess of zeros. They are mixture models with two component; one part is a truncated count component for positive counts that is usually Poisson or negative binomial. The binomial model with a censored count distribution.

$$P(Y_i|y_i|x,z) = \begin{cases} \pi(\gamma^{'} z_i) + 1 - \pi(\gamma^{'} z_i) f_{zero}(0,x) & \text{if} \quad \mathrm{y}_i = 0 \\ 1 - \pi(\gamma^{'} z_i) f_{count}(y_i, x) & \text{if} \quad \mathrm{y}_i > 0 \end{cases} \tag{4.10}$$

Maximum likelihood is used to estimate the hurdle model parameters which provides an advantage of allowing specification of the likelihood of the count and relationship of the model is given by

$$log(\mu_i) = x_i\beta + log(1 - f_{zero}(0, z, \gamma)) - log(1 - f_{count}(0, x_i, \beta)) \tag{4.11}$$

If regressors $x_i = z_i$ are used in the same count model in both components such that $f_{count} =$

$f-zero$ then a set of the null hypothesis $\beta = \gamma$ then tests whether hurdle is needed or not. Writing the model formula as $y = x_1 + x_2$ describes the count regression relationship of y_i and x_i and also indicates that the same set of regressors is used for the zero hurdle component different set of regressor, for example, $y \sim x_1+x_2z_1+z_2+z_3$. This is read as that the count data model $y \sim x_1+x_2$ is conditional on (-) the zero hurdle model $y \sim z_1 + z_2 + z_3$

4.3.1 Result's Comparison (Injury Data)

Zero-augmented models, hurdle and zero-inflated, are both built to deal with overdispersion and excess zeros. Especially, these two characteristics occur commonly in counts in real life datasets better than their classical counterparts. Using cross-sectional data of the accident injures in Oman, we compare the performance of these models to obtain a best estimate of the influencing factors that affect injuries occurrence in an accident. We fitted and illustrated standard Poisson regression model to get a first view of the relationship between the dependent variable (i.e. injuries in an accident) an the accident related factors. We obtained the coefficient estimates along with associated partial Wald tests. All coefficient estimates are highly significant with acctyp, cause, gender having larger Wald statistic values compared to other covariates. If overdispersion is present in the dataset, the Wald tests might be inaccurate as a result of misspecification of the likelihood. Therefore, we calculate the sandwich standard error and are more reasonable. we want to compare this result of Poisson with more powerful models which deals with overdipsersion and excess zeros problems to see which model is producing better fit for our data. The result in Table 11 support the findings from Poisson model about the factor significance.

Table 4.4: Regression analysis of injuries using different Poisson's alternative models

	Dependent variable:				
	injuryc				
	Poisson	*negative binomial*	*glm: quasipoisson link = log*	*hurdle Poisson*	*hurdle negative binomial*
	(1)	(2)	(3)	(4)	(5)
year	0.011**	0.011**	0.012**		
	(0.006)	(0.005)	(0.006)		

time1	0.00004***	0.00004***	0.00004***		
	(0.00001)	(0.00001)	(0.00001)		
daySun	−0.004	−0.004	−0.004	−0.0004	0.003
	(0.020)	(0.020)	(0.021)	(0.034)	(0.029)
dayMon	−0.023	−0.023	−0.023	−0.039	−0.036
	(0.020)	(0.021)	(0.021)	(0.035)	(0.029)
dayTue	−0.008	−0.008	−0.008	−0.023	−0.020
	(0.020)	(0.020)	(0.021)	(0.035)	(0.029)
dayWed	0.029	00.029	0.029	0.032	0.028
	(0.020)	(0.020)	(0.020)	(0.034)	(0.028)
dayThu	0.109***	0.109***	0.109***	0.141***	0.1356***
	(0.109)	(0.019)	(0.019)	(0.032)	(0.027)
dayFri	0.108***	0.108***	0.108***	0.148***	0.138***
	(0.109)	(0.020)	(0.020)	(0.033)	(0.028)
monthFeb	0.004	0.004	0.005	0.001	−0.004
	(0.023)	(0.024)	(0.024)	(0.041)	(0.034)
monthMar	−0.026	−0.026	−0.025	−0.025	−0.024
	(0.022)	(0.023)	(0.023)	(0.039)	(0.032)
monthApr	0.037	0.037	0.038	0.050	0.046
	(0.024)	(0.024)	(0.025)	(0.041)	(0.034)
monthMay	0.020	0.020	0.021	0.046	0.040
	(0.025)	(0.025)	(0.025)	(0.046)	(0.040)
monthJun	−0.028	−0.028	−0.028	−.033	−0.027
	(0.025)	(0.026)	(0.026)	(0.044)	(0.036)
monthJul	0.005	0.005	0.005	0.002	0.003
	(0.024)	(0.025)	(0.025)	(0.042)	(0.035)
monthAug	−0.020	−0.020	−0.020	.011	0.008
	(0.025)	(0.025)	(0.026)	(0.043)	(0.035)
monthSep	0.020	0.020	0.020	0.032	0.027
	(0.025)	(0.025)	(0.026)	(0.043)	(0.036)
monthOct	−0.013	−0.013	−0.012	−0.024	−0.025
	(0.026)	(0.026)	(0.026)	(0.044)	(0.037)
monthNov	−0.013	−0.013	−0.012	−0.024	−0.025
	(0.026)	(0.026)	(0.027)	(0.045)	(0.037)
monthDec	−0.013	−0.013	−0.011	−0.016	−0.021
	(0.025)	(0.025)	(0.026)	(0.043)	(0.035)
roadSub	−0.089***	−0.089***	−0.088***	−0.155***	−0.142***
	(0.012)	(0.012)	(0.012)	(0.021)	(0.017)
roadPaved	−0.119***	−0.119***	−0.119***	−0.232***	−0.209***
	(0.044)	(0.044)	(0.045)	(0.075)	(0.063)
acctypRu-Over	−0.492***	−0.492***	*0.493***	−0.974***	−0.930***
	(0.027)	(0.027)	(0.027)	(0.062)	(0.056)
acctypOver-Turn	0.067***	0.067***	0.068***	0.108***	0.145***
	(0.019)	(0.019)	(0.019)	(0.037)	(0.030)
acctypFixed Object Collision	−0.013	−0.013	−0.012	0.035	0.022
	(0.021)	(0.022)	(0.022)	(0.040)	(0.032)

acctypMotorcycle/Bicycle	-0.430^{***}	-0.430^{***}	-0.429^{***}	-1.018^{***}	-0.974^{***}
	(0.045)	(0.045)	(0.045)	(0.092)	(0.085)
loctndscRoundabout	-0.180^{***}	-0.180^{***}	-0.180^{***}	-0.236^{***}	-0.221^{***}
	(0.030)	(0.030)	(0.030)	(0.052)	(0.045)
loctndscTangle	0.202^{***}	0.202^{***}	0.202^{***}	0.220^{***}	0.201^{***}
	(0.017)	(0.017)	(0.017)	(0.029)	(0.023)
loctndscOther	0.028	0.028	0.028	0.032	0.029
	(0.017)	(0.018)	(0.018)	(0.030)	(0.025)
age	0.002^{***}	0.002^{***}	0.002^{***}	0.002^{***}	0.002^{***}
	(0.0005)	(0.0005)	(0.0005)	(0.001)	(0.001)
causecarelessness	0.024^{*}	0.024	0.024	0.036	0.040^{*}
	(0.014)	(0.015)	(0.015)	(0.025)	(0.021)
causesafedist.	-0.049^{**}	-0.049^{**}	-0.049^{**}	-0.089^{**}	-0.074^{**}
	(0.022)	(0.022)	(0.022)	(0.038)	(0.032)
causefatigue/alcohol	-0.446^{***}	-0.446^{***}	-0.447^{***}	-0.187^{**}	-0.185^{**}
	(0.052)	(0.053)	(0.053)	(0.090)	(0.076)
causeother	0.053^{**}	-0.053^{**}	0.053^{**}	0.069	0.072^{**}
	(0.025)	(0.026)	(0.026)	(0.043)	(0.035)
genderfemale	0.107^{***}	0.107^{***}	0.107^{***}	0.117^{***}	0.115^{***}
	(0.016)	(0.016)	(0.017)	(0.027)	(0.023)
nationalitynon-omani	-0.091^{***}	-0.091^{***}	-0.092^{***}	-0.084^{***}	-0.073^{***}
	(0.016)	(0.016)	(0.016)	(0.027)	(0.023)
vehctypFour Wheel	0.055^{***}	0.055^{***}	0.054^{***}	0.096^{***}	0.090^{***}
	(0.016)	(0.016)	(0.016)	(0.028)	(0.023)
vehctypBi/Motorcycle	-0.116^{**}	-0.116^{**}	-0.114^{**}	-0.214^{*}	-0.226^{*}
	(0.056)	(0.057)	(0.057)	(0.110)	(0.100)
vehctypHeavy	-0.058^{***}	-0.058^{**}	-0.057^{**}	-0.061	-0.072^{**}
	(0.022)	(0.023)	(0.023)	(0.038)	(0.032)
vehctypOther	0.398^{***}	0.398^{***}	0.397^{***}	0.489^{***}	0.457^{***}
	(0.023)	(0.023)	(0.024)	(0.044)	(0.035)
hrmdtlModerate	-0.219^{***}	-0.219^{***}	-0.220^{***}	-0.254^{***}	-0.237^{***}
	(0.012)	(0.012)	(0.012)	(0.020)	(0.016)
hrmdtlSlight	-0.527^{***}	-0.527^{***}	-0.527^{***}	-0.813^{***}	-0.763^{***}
	(0.019)	(0.019)	(0.019)	(0.035)	(0.030)
hrmdtlNo harm	-0.699^{***}	-0.699^{***}	-0.697^{***}	-1.517^{***}	-1.451^{***}
	(0.038)	(0.039)	(0.039)	(0.0.097)	(0.093)
vhcls				0.111^{***}	0.094^{***}
				(0.017)	(0.012)
Constant	-22.027^{*}	$-22,027^{*}$	-22.368^{*}	0.437^{***}	0.556^{***}
	(11.245)	(11.393)	(11.569)	(0.065)	(0.051)
Observations	24,080	24,080	24,080	16,162	16,162
Log Likelihood	-28,228,660		$-28,218,550$	$-24,152,510$	$-24,303,490$
θ			$44,134^{***}(9.754)$		
Akaike Inf. Crit.	56,545,320		56,525,100		

Note: *p<0.1; **p<0.05; ***p<0.01

Chapter 5

Conclusion

Many studies have documented the application of crash-severity models to explore the relationship between accident severity and its contributing factors, such as driver characteristics, vehicle characteristics, roadway conditions, and weather factors. Lord et al.(2010) conducted a study about quantifying the sample size requirements for crash-severity models. They found that similar to many count data models, small datasets could significantly influence the model performance. Using the data of 24,192 traffic accidents that involve all types of accidents in different areas in Oman, we used Poisson regression and alternative count models to explore the relationship between injuries that resulted from accidents and their contributing factors. We found that given the high variation on the data, the results from fitting different random samples drawn from the full dataset agree with prior expectations in that small sample sizes significantly affect the development of crash-severity models. As they concluded, we find that further research is essential to generalise sample size requirements for developing the different models applied for crash data, which may be partly dependent upon the characteristics of the data. For example, in our analysis, we found that the selected datasets could be overdispersed, and it could be underdispersed according to the overdispersion test explained in the context. The results produced by the different models show a reasonable statistical fit. Here, we applied regression models that are basic in analysing count data and have been used widely in road safety research. We present thoroughly the analysis of the data

using standard Poisson regression and interpreted the model's results and its goodness of fit. We also fitted quasi-Poisson, negative binomial, ZIP, and ZINB regression models for comparison. This Poisson family of regression models provides improved and easy-to-implement analyses of count data. The objective of the study is to provide a demonstration of a model that can be used to assess the most important factors contributing to the severity of traffic accidents in Oman. Based on traffic police accident data, 18 explanatory variables were used in the model development process. Using the measure of deviance and the Wald statistic, the variables of interest were subjected to statistical testing. All the variables appear to be significant in the model when using the full dataset or large samples. However, as we reduce the sample size, we find that the most significant factors are accident type, location of the accident, accident cause, gender, vehicle harm details, and number of vehicles. The rest of the variables' significance changed when different samples were drawn.

Regardless of which factors are more significant, the results of this study provided valuable information on how the collision types, road specifications, time, weather, and driver characteristics are related to the occurrence of injuries in accidents. The most important variables in predicting the occurrence of injuries in an accident are accident type and cause of accident. More advanced analysis tools could be more efficient to confirm these findings. The finding of the study might lead to a greater focus on road accident sites, such as intersections and roundabouts, which should help decision makers in the government to focus their safety improvements more cost-effectively. However, not only the relative danger as expressed by the odds ratio, but also the absolute density of accidents with regard to location should be taken into account to develop cost-effective strategies. The odds presented in this study can be used to help establish priorities for programs to reduce serious accidents. It is important to note that the odds described in this study were computed with no consideration for traffic exposure or the data that are not available or difficult to obtain in Oman. However, the findings of this study can be a guidance for future studies. Some research limitations arise in this study since police report data were analysed. Even though there are a lot of independent variables available in the database, the database contains missing values and

unreliable data, and some important variables cannot be further investigated, such as information about road geometry, speed of motorcycles, helmet, and alcohol use, as well as mobile use as causes of an accident. In future studies, more explanatory variables that might be available from other sources in Oman should be considered.

Looking at each factor separately, we find that the cause of accidents in Oman is the high speed. The government has updated the rules of the limits in speed and the fines for breaking them. They have also set many fixed radars in the main cities' highways. However, in the long-distance lines, they still use the temporary hidden radars which are still not sufficient to reduce the high speeding and monitor the behaviour of the drivers on long distances lanes. The other major causes are also related to wrong conduct of drivers or passengers, and these could be tackled by focusing on educating drivers on the risks of not following road rules. Increasing the police patrols around the country is an important way to keep the public aware of road conduct. The road, vehicle, and weather do not appear to be significant causes of the accidents in Oman; however, more investigation for the interactions with other factors could reveal further information about the contribution of these factors in accident severity. It can be summed up that in Oman, road accidents are mainly related to human factors. Again, these should be tacked with all possible preventions,such as educating the public, increasing police patrols, and improving the fines and punishment system.

Our result says that the highest frequency of accidents occurs in straight roads, triangles (intersections), and roundabouts. The road type does not appear to be a major cause of accident, but it must be related to human conduct, especially in a road with two opposite directions, in triangles (intersections), and in roundabouts. For injuries, they occur more in accidents that happen in the triangles (intersections) than in straight roads, while roundabout accidents result in less injuries cause by accidents in straight roads. Male accidents constitute more than 90 percent of the total accidents. We can justify this by the fact that males go out more than females and tend to take more risks in the road. The percentage indicates that male drivers are more likely to have RTA. The percentage of expatriate people in Oman is 44.2 percent which indicates that Omani driver

is more likely to be involved in a road accident. The age that the driver is more likely to have an accident is in the range of 20âĂŞ30. More specifically, the mean average of a driver who gets involved in an accident in Oman is thirty. Since all drivers should be licenced, the 4.3 percent of the nonlicenced accident drivers should be given attention.

Our study shows that there is an increasing trend of RTA in Oman over the years that goes along with increasing trend in licences and vehicles. The distribution of accident by hour of the day is parallel to the volume of traffic in the relative time. The distribution of the accidents is not equal throughout the days of the week, but the percentages are close. Most of the accidents happen with small passenger vehicles and involves one vehicle. More than 50 percent of the accidents involve a single vehicle. The population of heavy vehicles may indicate that heavy vehicles affect the occurrence of RTA. Separating the accidents by number of involved vehicles and analysing separately gives different results of significant factors contributing to the occurrence of injuries.

Accident type, description of location, cause of accident, gender of the driver, and the vehicle harm degree are the most significant variables to injury occurrence. The other variables should be investigated more in interaction with other variables to confirm their effect in accident severity. In accident type, we found that the run-over and collision with fixed objects are almost half risky in affecting the injury severity than two-vehicle collision with fixed objects, but overturn accidents and motorcycle accidents result in more severe injuries. compared to the two-vehicle collision, overturn accidents are 2.301 and bicycle and motorcycle accidents are 3.356. However,, in the model it appears that only fixed objects and bicycles and motorcycles are the significant, but not all the segments of the road are significant in explaining the occurrence of severe accidents. We classified the segments of the road according to the frequency of accidents on these segment to straight road, triangles(intersections), roundabouts, and the rest of the accident locations compiled in one group as others. We found that triangles (intersections) are 2.64 more risky in causing sever accidents than straight roads, while the roundabouts are 0.612 less risky than having an accident in straight road. For the cause of accident, we know from our descriptive statistics and literature review that

high speed is the major cause of accidents in Oman. In our model, we find only the alcohol and sugar in the blood is significant in explaining the occurrence of injury case. Having an accident with the cause that the driver was having alcohol or high sugar level in his blood seems to be less sever than an accident that is caused by the high speed, as it is making 0.217 of the risk that high-speed accident make in causing human casualties. The risk that high-speed accident make in causing human casualties. The rest of the causes are not significant in our model, but we can see from their coefficient that overtaking is 1.824 of the high speed accident risk. The safe distance is 1.353, and the other causes are 1.508. Though the statistics showed that a male driver is more likely to have an accident than a female driver, the model showed that an accident caused by a female driver results in sever injuries by 2.44 compared to an accident caused by a male driver. The harm in the vehicle is naturally significant in explaining the severity of the accidents, and the less harm to the vehicle there is, the less sever injuries are caused.

Bibliography

References

[1] Cameron, A.C. and Trivedi, P.K, *Microeconometrics: Methods and Applications. Cambridge*. Cambridge University Press. 2005

[2] Conway, R. W.; Maxwell, W. L., *A queuing model with state dependent service rates.* Journal of Industrial Engineering 12: 132136, 1962.

[3] Anderson, T., Kernel density estimation and K-means clustering to profile road accident hotspots, Accident Analysis and Prevention. *Journal of Industrial Engineering* 41: 359âĂŞ364, 2009.

[4] Barbour, A.D., Holst,L and Janson, S., *Poisson Approximation*. Oxford University Press 1992.

[5] Hardin,J. and Hilbe,J. *Generalized Linear Models and Extensions*. Stata Press 2012.

[6] Marshall,A. and Olkin,I. *Life Distribution;structure of nonparametric,semiparametric and parametric Families*. Springer 2007.

[7] Lidsey,J., *Modelling Frequency and Count Data*. Oxford University Press 1995.

[8] Venables,W. and Ripley,B. *Modern Applied statistics with S-Plus*. Splinger-Verlag 1994.

[9] McCullagh,P. and Nelder,J. *Generalized Linear Models*. Chapman and Hall 1983.

[10] McCulloch,C. and Searle,S. *Generalized, Linear and Mixed Models*. Wiley Inter-Science 2001.

[11] Madsen,H. and Thyregod, P. *Introduction to General and Generalized Linear Models.* CRC Press 2011.

[12] Fahrmeir,L., Kneib,T., Lang,S. and Marx,B. *Regression.* Springer 2013.

[13] Fahrmeir,L., Kneib,T., Lang,S. and Marx,B. *Kernel density estimation and K-means clustering to profile road accident hotspots.* Accident Analysis and Prevention 41 (2009) 359-364 2009.

[14] Bauer, K.M., and D.W. Harwood,04 *Statistical Models of At-Grade Intersection Accidents.* Report No. FHWA-RD-99-094, Research, Development,, Federal Highway Administration, McLean, VA, 2000.

[15] Madsen,H. and Thyregod, P. *Modelling for Identifying Accident-Prone Spots: Bayesian Approach with a Poisson Mixture Model.* Journal of Civil Engineering,16(3):441-449 2012.

[16] Royal Oman Police, *Facts and Figures.* GCC Trafic Week 2013 2012.

[17] Elvik, R. *Assessing causality in multivariate accident models.* Accident Analysis and Prevention,46 253-264 2011.

[18] Hauer E. *Overdispersion in modelling accidents on road sections and in Empirical Bayes estimation.* Accident Analysis and Prevention,33 799-808 2001.

[19] Jianming Ma Kara M. Kockelman *Bayesian Multivariate Poisson Regression for Models of Injury Count, by Severity.* Transportation Research Record No. 1950:24-34, 2006.

[20] King E.C *A test for the equality of two regression curves based on kernel smoothers.* Ph.D. Dissertation, Dept. of Statist., Texas A M Univ, College Station, TX 1989.

[21] Lord,D. *Modelling 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, Vol. 38(4), pp. 751-766 2006.

[22] Poch, M and Mannering, F. *Negative Binomial Analysis Intersection-Accidents Frequencies.* Journa of transportation Engineering, 105-113. 1996.

[23] Conway, R. W.; Maxwell, W. L., *A queuing model with state dependent service rates.* Journal of Industrial Engineering 12: 132136, 1962.

[24] Shmueli G., et al. *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 127142.[1], 2005.

[25] Guikema, S.D. and J.P. Coffelt, *A Flexible Count Data Regression Model for Risk Analysis* Risk Analysis,28 (1), 213-223. doi:10.1111/j.1539-6924.2008.01014.x, 2008.

[26] Lord, D., S.D. Guikema, and S.R. Geedipally, *Application of the ConwayâĂŞMaxwellâĂŞPoisson Generalized Linear Model for Analyzing Motor Vehicle Crashes* Accident Analysis Prevention, 40 (3), 1123-1134. doi:10.1016/j.aap.2007.12.003, 2008.

[27] Lord, D., S.D. Guikema, and S.R. Geedipally, *Extension of the Application of Conway-Maxwell-Poisson Models: Analyzing Traffic Crash Data Exhibiting Under-Dispersion* Risk Analysis, 30 (8), 1268-1276. doi:10.1111/j.1539-6924.2010.01417.x, 2010.

[28] Sellers, K. S. and Shmueli, G., *A Flexible Regression Model for Count Data* Annals of Applied Statistics, 4 (2), 943-961, 2010.

Appendix A

Logistic Regression Analysis of Injury Data

Many studies were focused at identifying the most probable factors that affect accident severity and estimating the statistically significant effects of these factors. Some factors, such as accident location, type, and time; collision type; and age and nationality of the driver at fault, his licence status, and vehicle type, are analysed to see if they have an impact on higher potential for serious injury or death. Such analysis is useful for decision makers to evaluate the effect of the factor on accident occurrence and severity and set policies and interventions accordingly. It examines how one or more independent variables influence the dependent variable by examining the relationship between these variables and the log odds of the dichotomous outcome by calculating changes in the log odds of the response as opposed to the response variable itself. The odds ratio is the ratio of two odds that is used to measure the relationship between two variables such that π is the probability of the outcome 'success' of the event and $(1-\pi)$ is the probability of the opposite outcome of the event 'failure'. In logistic regression, the log odds ratio provides a simple description of the probabilistic relationship of the

variables and the outcome. Logistic regression is categorised into two: binomial/binary and multinomial. Binary logistic regression is used when the response variable is dichotomous and the explanatory variables are continuous or categorical variables.

In this chapter, we apply logistic regression to estimate the probability of the occurrence of an injury case in an accident having the dependent variable in binary (dichotomous) form. We start the analysis of the injury data as a binary dependent variable to examine the relationship between the occurrence of injury cases in an accident and some factors related to driver, vehicle, or road characteristics. The dependent variable here, the occurrence of injury cases in an accident, is classified as 1 if one or more injury cases are recorded and 0 if no injury case is recorded, and the independent variables are mostly categorical and few are continuous. Fitting the binary model to the injury data showed that the accident type, accident location, cause of accident, gender, nationality, vehicle type, and harm vehicle level are significant factors in explaining the occurrence of the injury case. Mathematically, the principle that underlies the logistic regression is calculating the logit transformation of the dependent variable Y at some value of X-the natural logarithm of an odds ratio. For example, in 2×2 contingency table, considering an instance in which the distribution of the dichotomous outcome variable occurrence of an injury case (yes/no) and the predictor variable the gender of the driver. If we want to assess the probability of having injury given the driver is male relative to having a female driver, we calculate $\pi/(1-\pi)$, the ratio of the two probabilities. If the value is 1:1, then the probability of having injury is equal between the two cases. If the value is more than 1, then the male driver has more probability of having a severe accident than females, and if the value is less than 1, then the female driver has more probability of having a severe accident than males.

A.1 Theoretical Concepts of Binary Logistic Regression(LR)

Binary logistic regression is a method for fitting a regression curve, $y = f(x)$, when y consists of proportions or probabilities, or binary coded data (0,1; failure,success). It works by fitting models to data using an S-shaped function called logistic function.

$$y_i = \begin{cases} 1\text{if the } i^{th} \text{ accident involved at least one injury case} \\ 0\text{if no injury recorded} \end{cases}$$

Logistic regression is linear regression that uses the logit transform of y, where y is the proportion (or probability) of success $Pr(Y = 1|X = x)$ at each value of X. Traditional least squares regression(OLS) is not suitable as neither the normality nor the homoscedasticity[1] assumption will be met. It is used to analyse the effect of potential risk factors that significantly influence the probability of the outcome of the event $y = 1$, here, occurrence of injury case in an accident.

A.1.1 Odds ratios

Logistic regression model estimates **the odds ratios (OR)** at $(1-\alpha)$ percent confidence intervals (CI) as a determinant of which variables should be included (commonly $\alpha = 0.05$). Where the odds ratios means the ratio between the success and failure cases.

$$odds_i = \frac{\pi_i}{1 - \pi_i}$$

If the probability of the success is one half, the odds are one-to-one (1:1). If the probability is one third, the odds are one-to-two (1:2).

[1] Heteroscedasticity (the violation of homoscedasticity) is present when the size of the error term differs across values of an independent variable. The impact of violating the assumption of homoscedasticity is a matter of degree, increasing as heteroscedasticity increases.

A.1.2 Logit transformation

In logistic regression, the probabilities π_i depend on a vector of observed covariates x_i. We cannot let π_i be a linear function of the covariate directly, η_i,

$$\eta_i = x_i\beta$$

where β is a vector of regression coefficients as in the linear probability model. However, logistic regression fits b_0 and b_1, the regression coefficients (which were 0 and 1, respectively, for the graph above). Note that the curve is not linear; however, the point of the logit transform is to make it linear.

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{i=1}^{N} x_i\beta \qquad i = 1, 2, ..., N$$

We discuss the stochastic structure y of the data in terms of the Bernoulli and Binomial distributions, and the systematic structure $X\beta$ in terms of the logit transformation or what is called the generalised linear model with Binomial response and link logit. Figure 8 shows the a logistic curve with logit link.

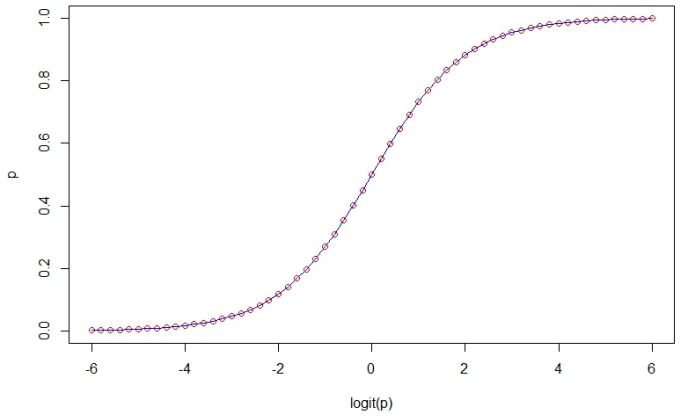


Figure A.1: The logistic curve :π = exp(logit)/[1 + exp(logit)]

So the logistic function is

$$E(y_i) = \pi_i = exp(logit)/[1 + exp(logit)]$$

where π_i is the probability of success given some predictors or it is the expected value for the response variable $y = 1$ given some predictors x_i for some i_{th} observation.

$$E(y_i|x_i) = \pi_i = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} = \frac{1}{1 + e^{-x_i\beta}}$$

A.1.3 Assumptions of Logistic Regression

Logistic regression assumes linear relationship between the logit of independent variables and the dependent variable but the relationship between the dependent and independent variables doesn't have to be linear. The sample size should be large as the reliability of the estimation declines when there are only few cases. In logistic regression, independent variables are not linear functions of each other. Normal distribution is not necessary or assumed for the dependent variable. Homoscedasticity is not necessary for each level of the independent variables. Normally distributed description of errors are not assumed. The independent variables need not be interval level.

A.1.4 Linear Probability Model

In the linear probability model, the expected value of the dependent variable, Y_i, is defined as a linear function of some independent variables, X_i, such that:

$$E[y_i] = x_i\beta$$

For a binary independent variable, the expected value of Y_i is defined as

$$E[y] = 1 \times Pr(y = 1) + 0 \times Pr(y = 0) = Pr(y = 1)$$

$$Pr(y = 1|x_i) = x_i\beta = \pi_i$$

The coefficient of this model illustrates how a one-unit change in X affects the probability of the success outcome, $Pr(y = 1)$. However, there are many reasons that make the linear probability model not suitable for modelling binary data, among these, the main two reasons are:

[28] **The unbounded predicted values**: $x_i\beta$ can take on values greater than one and less than zero.

[28] **The conditional heteroskedasticty**: The variance of residuals is related to the value of x. Specifically,

$$var(y = 1) = E[y = 1](1 - E[y = 1]) = x_i\beta(1 - x_i\beta)$$

This indicates that the variance of Y_i depends on the values of X_i and β and is, therefore, heteroskedastic by construction.

A.1.5 Logit Model

If the dependent variable is binary (dichotomous) and follows a Binomial distribution, $Y \sim Bin(n_i, \pi_i)$, the outcome values are: $Pr(Y = 1|X = x) = E[Y|X = x] = \pi_i$ is the probability of success and $Pr(Y = 0|X = x) = 1 - \pi_i$ is the probability of failure. Taking the logit transformation, we get a linear function of the predictors that defines the systematic part of the model.

$$log\left(\frac{\pi_i}{1 - \pi_i}\right) = \sum_{i=1}^{N} x_i\beta \qquad i = 1, 2, ..., N$$

where x_i is a vector of covariates and β is a vector of regression coefficients. The likelihood of the Binomial distribution $f(y|\beta)$ is

$$L(\beta|y) = \prod_{i=1}^{N} \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{(n_i - y_i)}$$

We can verify by direct calculation that the expected value and variance of y_i are

$$E(y_i) = \mu_i = n_i \pi_i,$$

$$var(y_i) = \sigma_i^2 = n_i \pi_i (1-\pi_i)$$

It can be noticed that the mean and variance depend on the underlying probability π_i. The factor that affects the probability of the response will affect the mean and also the variance of the observation. There is no π_i in the factorial terms so they are just constants and ignored. Note that since $a^{x-y} = a^x/a^y$, when we rearrange the equation, it becomes

$$\prod_{i=1}^{N} \left(\frac{\pi_i}{1-\pi_i} \right)^{y_i} (1-\pi_i)^{(n_i)}$$

Taking the exponent of both sides of this equation gives the odds of the $i_t h$ unit

$$\left(\frac{\pi_i}{1-\pi_i} \right)^{y_i} = e^{x_i \beta}$$

Solving for π_i we get

$$\pi_i = \frac{e^{x_i \beta}}{1 + e^{x_i \beta}}$$

The cumulative standard logistic is

$$Pr(y_i = 1|x_i) = \frac{e^{x_i \beta}}{1 + e^{x_i \beta}} = \frac{1}{1 + e^{-x_i \beta}} = \Lambda(x_i \beta)$$

and the following log-likelihood function

$$ln\mathcal{L} = \sum_{i=1}^{N} \{y_i ln[\Lambda(x_i\beta)] + (1-y_i) ln[1-\Lambda(x_i\beta]\}$$

$$= \sum_{i=1}^{N} \left\{ y_i ln\left[\frac{1}{1+e^{-x_i\beta}}\right] + (1-y_i) ln\left[1 - \frac{1}{1+e^{-x_i\beta}}\right]\right\}$$

Thus, differentiating with respect to each β_k gives

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{N} y_i x_i - n_i . \frac{1}{1+e^{x_i\beta}} . \frac{\partial}{\partial \beta} \left(1 + e^{x_i\beta}\right)$$

$$= \sum_{i=1}^{N} y_i x_i - n_i . \frac{1}{1+e^{x_i\beta}} . e^{x_i\beta} . \frac{\partial}{\partial} \sum_{i=1}^{N} x_i\beta$$

$$= \sum_{i=1}^{N} y_i x_i - n_i . \frac{1}{1+e^{x_i\beta}} . e^{x_i\beta} . \sum_{i=1}^{N} x_i$$

$$= \sum_{i=1}^{N} y_i x_i - n_i \pi_i x_i$$

The estimates of β can be found by setting each of the $K+1$ equations equal to zero and solving for each β_k. The solution of each equation, if it exists, is a critical point-either a maximum or minimum. The critical point will be a maximum if the matrix of second partial derivatives is negative definite-meaning that every element of the matrix is less than zero. This matrix also forms the variance-covariance matrix of the parameter estimates. It is formed by differentiating each element of β, denoted by β. The general form of the second partial derivatives is the Hessian matrix

$$H = \frac{\partial^2 ln\mathcal{L}}{\partial\beta\partial\beta'} = -\sum_{i=1}^{n} \Lambda(x_i\beta)[1-\Lambda(x_i\beta)] x_i x_i^{'}$$

The matrix of the second partial derivatives is

$$\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta'} = \frac{\partial}{\partial\beta'}\sum_{i=1}^{N} y_i x_i - n_i\pi_i x_i$$

$$= \frac{\partial}{\partial\beta'}\sum_{i=1}^{N} n_i\pi_i x_i$$

$$= -\sum_{i=1}^{N} n_i x_i \frac{\partial}{\partial\beta'}\left(\frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right)$$

A.1.6 Probit Model

The cumulative standard normal is

$$Pr(y_i = 1|x_i) = \int_{-\infty}^{x_i\beta} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - x_i\beta)^2}{2\sigma^2}} dx = \Phi(x_i\beta)$$

where $x_i\beta$ is just a linear function of some sort. The integral doesn't have a closed form solution, which is why we normally abbreviate it as $\Phi(x_i\beta)$. Substituting this in for g(.) gives the following likelihood function

$$\mathcal{L} = \prod_{i=1}^{N} [\Phi((x_i\beta)]^{y_i} [1 - \Phi(x_i\beta)]^{1-y_i}$$

and the following log-likelihood function

$$ln\mathcal{L} = \sum_{i=1}^{N} \{y_i ln[\Phi(x_i\beta)] + (1 - y_i) ln[1 - \Phi((x_i\beta)]\}$$

Because of the symmetry of the normal density, we can express $1-\Phi((x_i\beta)$ as $\Phi(-x_i\beta)$. This means that we can express the log-likelihood function as

$$ln\mathcal{L} = \sum_{i=1}^{N} y_i ln[\Phi(x_i\beta)] + (1 - y_i) ln[\Phi(-x_i\beta)]$$

The log-likelihood function is non-linear, so there is no closed form solution for β. However, numerical maximisation is easy since the log-likelihood is globally concave.

A.2 Application and illustration (Injury Data)

For this study, we classify the accidents according to the occurrence of an injury case in the accident. We code an accident with no injury as 0 and an accident with one or more injury as 1. The definition of the variables is obtained from a coding system supplied with the data set. The majority of these variables are categorical variables that simply indicate the existence of a certain condition, such as the road type at the accident location. The analyses here are aimed to identify the factors that might have an effect on the accident severity. Thus, we summarise from the data 16 variables detailed in Table 6 earlier in Poisson analysis. Table 12 and Figure 9 in this chapter show again the distribution of RTA by number of injury cases and we can see that 26.737 percent of accident result in no injury while 73.263 percent included one injury case or more.

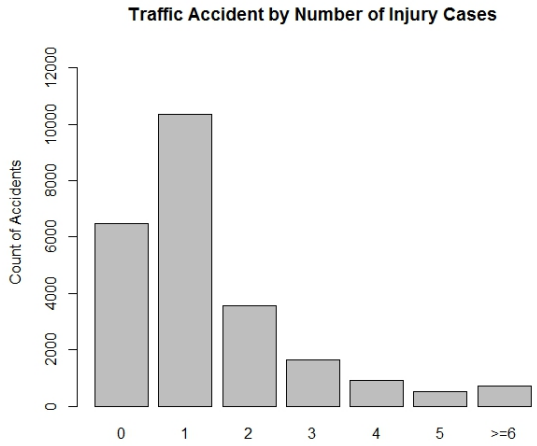


Figure A.2: Distribution of RTA during 2009-2012 by injury cases

Table A.1: Distribution of RTA during 2009-2012 by number of cases of injury

Injury Cases	Frequency	Cum.Freq.	Percentage	Cum.Percent.	
0	6468	16833	26.737	26.737	
1	10365	10365	42.847	69.584	
2	3573	20406	14.770	84.354	
3	1633	22039	6.750	91.104	
4	930	22969	3.844	94.949	
5 or more	1223	24192	5.051	100.000	
injury case	Binary Dependent Variable	injury Number	(0,1)		
Min.0	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	0.00	1.00	0.73	1.00	1.00

Recall that logistic regression model is useful to investigate the relationship between the binary outcome variable and the predictor variables. It models the logit-transformed probability as a linear relationship with the predictor variables. Let the occurrence of an injury case y be the binary outcome variable indicating no/yes as $0/1$ and $\pi(x_i)$ be the probability of y to be yes $= 1$, $\pi(x_i) = Prob(y_i = 1)$. Let $x_1, .., x_k$ be the set of predictor variables, then the logistic regression of y on $x_1, ..., x_k$ estimates parameter values for $\beta_0, \beta_1, ..., \beta_k$ via maximum likelihood method of the following equation.

$$logit[\pi(x_i)] = log\frac{\pi(x_i)}{1-\pi(x_i)} = \alpha + \beta x_i$$

In terms of probabilities, the equation above is translated into

$$P(y = 1|x_i) = \pi(x_i) = \frac{e^{\beta x_i}}{1 + e^{\beta x_i}}$$

Modelling different samples of 5,000 accidents that are taken from the injury data with binary logit model gave almost the same result about the significant factors for road injury. Stepwise selection was followed to check the best model, which is a combination of the forward and backward methods. Given the big number of covariates tested, we started with including all the variables with no interactions and removing the non significant variables. After every step, we checked to see if a variable that has been dropped should be added back into the model. The model is tested by checking the deviance and the Wald (W) statistic to eliminate the variables that were not significant.

We used the step function in R with no interaction to select the best model and the **anova** function to test the overall significance of each variable. It gives information on how much deviance the variable adds to the model. From the W values, it appears that the variables *acctyp*, *loctndsc*, *cause*, *gender*, *nationality*, *vehctyp* and *hrmdtl* have significant effect. In the other hand, for the variables *day*, *month*, *age*, *licens*, *roadtyp* and *climcond*, the results indicate that they are not adding useful information to the variability in the response variable and should be removed. We work to test the hypothesis with the null hypothesis

$$H_0 : \beta_j = 0$$

The GLM function in R calculates the Wald-test (z) based on the large sample distribution of the maximum likelihood estimate, which is approximately normal with mean β and variance-covariance matrix $\hat{var}(\hat{\beta})$ to calculate the βs. We test the significance of a single coefficient by calculating the ratio of the estimate to its standard error.

$$z = \frac{\hat{\beta}_i}{\sqrt{\hat{var}(\beta_j)}}$$

We regress z on the covariates calculating the weighted least squares estimate

$$\hat{\beta} = (X'WX)^{-1}X^{'}W_z$$

where W is the diagonal matrix of weights with entries

$$w_{ii} = \hat{\mu}_i(n_i - \hat{\mu}_i)/n_i$$

The resulting estimate is consistent and its large-sample variance is given by the variance-covariance matrix

$$var(\hat{\beta}) = (X^{'}WX)^{-1}$$

Interpretation of Coefficients

In OLS, β equals the change in Y with one unit change in X, but in logistic regression, this interpretation is not suitable since the linear predictor $x\beta$ is not directly estimating $Pr(Y)$. Instead, we have to translate the relation of the response and the linear predictor, $x\beta$, using the exponent function. When we do that, we have a type of 'coefficient' that is interpreted differently. This coefficient is called the odds ratio. **The odds ratio** is equal to $\exp(\beta)$, or sometimes written as e^{β}. Here, $\pi(x)$ is the overall probability of having an injury case ($inj = 1$) given x. The indicator variables have a slightly different interpretation. For example, having a run-over accident, versus a two-vehicle collision accident changes the log odds of injury by 0.480. Having an overturn accident versus a two-vehicle collision changes the log odds of injury by 1.145. Having a fixed object collision versus a two-vehicle collision changes the log odds of injury by -0.270. Having a motorcycle or bicycle accident versus a two-vehicle collision changes the log odds of injury by 0.566.

For the effect of **accident location**, having an accident at a roundabout versus an accident in a straight road changes the log odds of injury by -0.163 not significant at the 10% level. Having an accident in a triangle versus an accident in a straight road changes the log odds of injury by 0.457 significant at the 5% level. Having an accident in any other location is not significant. The **cause of the accident** appeared to be significant in explaining the variation of injury severity. The model shows that an accident caused by carelessness versus an accident caused by high-speed changes the log odds of injury by 0.308. An accident caused by not leaving safe distance versus an accident caused by high-speed changes the log odds of injury by 0.256. An accident caused by overtaking versus an accident caused by high-speed changes the log odds of injury by 0.542. An accident caused by fatigue or alcohol effect versus an accident caused by high-speed changes the log odds of injury by -1.185. Other causes

of accidents, such as road, vehicle, and climate condition, versus an accident caused by high-speed change the log odds of injury by -0.097. For the **driver characteristics**, we find that having an accident with a female driver versus an accident with a male driver changes the log odds of injury by 0.646. The age and nationality of the driver appears to be not significant here. The harm level of the vehicle is highly significant. Having an accident with moderate harm versus an accident with severe vehicle harm changes the log odds of injury by -0.752. Having an accident with slight harm versus an accident with server vehicle harm changes the log odds of injury by -0.841. Having an accident with no harm versus an accident with server vehicle harm changes the log odds of injury by -1.100.

We calculated the confidence intervals for the coefficient estimates. Note that for logistic models, confidence intervals are based on the profiled log-likelihood function; however, we can also get CIs based on just the standard errors by using the default method. The Wald (z) test can be used to calculate a confidence interval for β_j. We can state with $100(1-\alpha)\%$ confidence that the true parameter lies in the interval with boundaries

$$\hat{\beta} \pm z_{1-\alpha/2}\sqrt{\hat{var}(\hat{\beta})} \tag{A.1}$$

where $z_{1-\alpha/2}$ is the normal critical value for a two-sided test of size α. Confidence intervals for effect in the logit scale can be translated into confidence interval for odds ratios by exponentiating the boundaries. The coefficients are the logs of the odds ratios of the probabilities of injury, so we exponentiate and interpret them as odds ratios. We do the same with confidence intervals of the coefficient and get the odds ratios by exponentiating them and putting them all in one table. Now it should be easier to interpret the coefficients in relation to the probability of the injury. If the OR is exactly one, the two levels of the variable have equal effect on explaining the variation on the response variable. If the OR is less than one, the level has less effect in the

variation of the response variable, and if OR is more than one, the level has more effect on the response variable. Now we can say that for a one-unit increase in the run-over accidents, the odds of having an injury case (versus not having injury case) increases by a factor of 0.669. This means that a run-over accident is 33.1% less likely to cause an injury case compared with a two-vehicle collision. We interpret the rest of the coefficients in the same way.

We tested for an overall effect of the accident type using the wald.test function of the **aod** package. The wald-test function refers to the coefficients by their order in the model. In the wald-test function, b provides the coefficients, and Sigma provides the variance-covariance matrix of the error terms, and terms in the function tells R which terms in the model are to be tested. In this case, terms 2, 3, 4, and 5, are the four terms for the levels of the accident type. The chi-squared test statistic of 43.7, with four degrees of freedom that is associated with a p-value less than 0.001, indicates that we reject the null that $H_0 : \beta_j = 0$, and we can claim that the overall effect of accident type is statistically highly significant. The location description has a chi-squared test statistic of 9.1, with three degrees of freedom that is associated with a p-value of 0.028, indicating that the overall effect of the location description is statistically significant. The cause of the accident has a chi-squared test statistic of 13.9, with five degrees of freedom that is associated with a p-value of 0.19, indicating that the overall effect of the cause is statistically significant. The gender of the driver has a chi-squared test statistic of 9.0, with one degree of freedom that is associated with a p-value of 0.0027, indicating that the overall effect of the gender of the driver has a significant effect on the occurrence of an injury case. The harm in the vehicle has chi-squared test statistic of 15.6, with three degrees of freedom that is associated with a p-value of 0.0014, indicating that the overall effect of the harm in the vehicle is statistically significant.

We test additional hypotheses about the differences in the coefficients for the different

levels of the variables. Below we test that the coefficient for run-over accidents is equal to the coefficient for overturn accidents $H_0 : \beta_{run-over} = \beta_{over-turn}$. The first line of the code below creates a vector l that defines the test we want to perform. We want to test the difference of the terms for run-over and overturn accidents (i.e., the second and third terms in the model). To contrast these two terms, we multiply one of them by 1 and the other by -1. The other terms in the model are not involved in the test, so they are multiplied by 0. The second line of code below uses L=l to tell R that we wish to base the test on the vector l instead of using the Terms option. The chi-squared test statistic of 11.6, with one degree of freedom is associated with a p-value of 0.00065, indicating that the difference between the coefficient for run-over accidents and the coefficient for overturn accidents is not zero and is statistically significant. We run the test for the rest of the coefficients, and we get the same result that the differences between the coefficients of levels of variables are not zero and there are significant differences between them. Table 13 gives the results of running a binary logistic regression on the data to see if we get results that supports our findings in the previous chapters when applying Poisson and alternatives.

Table A.2: Binary logistic regression results using robust S.Error

	Estimate	Odds Ratio	Robust SE	Pr(>|z|)	LL	OR LL	UL	OR UL
(Intercept)	0.146	1.157	0.312	0.640	−0.465	0.628	0.756	2.130
year2010	0.092	1.097	0.086	0.283	−0.076	0.927	0.261	1.298
year2011	0.423	1.526	0.090	0.00000	0.246	1.279	0.600	1.822
year2012	0.677	1.967	0.144	0.00000	0.395	1.484	0.959	2.608
acctypRun-Over	0.480	1.615	0.189	0.011	0.109	1.115	0.850	2.340
acctypOver-Turn	1.145	3.144	0.198	0	0.757	2.131	1.534	4.638
acctypFixed Object Collision	−0.270	0.763	0.176	0.124	−0.614	0.541	0.074	1.077
acctypMotorcycle/Bicycle	0.566	1.761	0.248	0.023	0.079	1.083	1.052	2.864
loctndscRoundabout	−0.163	0.850	0.144	0.259	−0.446	0.640	0.120	1.128
loctndscTriangle	0.457	1.580	0.172	0.008	0.119	1.127	0.795	2.215
loctndscOther	0.028	1.029	0.125	0.823	−0.218	0.804	0.274	1.315
causecarelessness	0.308	1.360	0.097	0.002	0.118	1.125	0.498	1.645
causesafedist.	0.256	1.291	0.171	0.136	−0.080	0.923	0.591	1.807
causeovertaking	0.542	1.719	0.207	0.009	0.136	1.145	0.948	2.581
causefatigue/alcohol	−1.185	0.306	0.218	0.00000	−1.613	0.199	−0.758	0.468
causeother	−0.097	0.907	0.161	0.545	−0.413	0.662	0.218	1.243
genderfemle	0.646	1.907	0.134	0.00000	0.384	1.468	0.908	2.479
nationalitynon-omani	−0.362	0.696	0.090	0.0001	−0.539	0.583	−0.185	0.831
vehctypFour Wheel	−0.219	0.803	0.103	0.033	−0.422	0.656	−0.017	0.983
vehctypBi/Motorcycle	0.986	2.681	0.317	0.002	0.364	1.439	1.608	4.994
vehctypHeavy	−0.083	0.920	0.126	0.511	−0.331	0.718	0.165	1.179
vehctypOther	0.234	1.263	0.216	0.280	−0.190	0.827	0.658	1.930
hrmdtlModerate	−0.752	0.472	0.091	0	−0.930	0.394	−0.573	0.564
hrmdtlSlight	−0.841	0.431	0.121	0	−1.079	0.340	−0.604	0.547
hrmdtlNo harm	−1.100	0.333	0.175	0	−1.443	0.236	−0.756	0.470
vhcls	0.695	2.004	0.145	0.00000	0.412	1.510	0.978	2.660
Observations	4,976							
LogLikelihood	-2,538.693							
AkaikeInf.Crit.	5,129.387							
Note:	*p<0.1; **p<0.05; ***p<0.01							

Deviance. The residual sum of squares (here it's the deviance ratios) and the coefficient estimates are the same as the ones given by the linear models **lm** function. For our model, the deviance is 5,077.387. In ordinary least squares OLS regression, the primary measure of model goodness-of-fit is R^2, which is an indicator of the percentage of variance in the dependent variable explained by the model. With logistic regression, instead of R^2, the statistic for the overall goodness-of-fit of the model, we have deviance instead. We use chi-square as a measure of our model fit similarly. It is the fit of the observed values (Y) to the expected values ($\hat{Y}$). The larger the ('deviance') the difference of the observed values from the expected values, the poorer the fit of the model. So, we want a small deviance if possible. As we include more variables to the model, the deviance should get smaller, indicating an improvement in the fit. The deviance is a measure of discrepancy between observed and fitted values and it is given by

$$D = 2\sum \left\{ y_i log(\frac{y_i}{\hat{\mu - i}}) + (n_i - y_i) log(\frac{n_i - y_i}{n_i - \hat{\mu_i}}) \right\} \tag{A.2}$$

where y_i is the observed and $\hat{\mu_i}$ is the fitted value for the $i - th$ observation.
We can test the goodness of fit of the model by looking to the ratios given below the coefficients, including the null deviance (4739.102) and deviance residuals and the AIC. We also can use the anova analysis to measure the model's goodness-of-fit by testing the significance of the overall model. We test the hypothesis that the model with predictors fits significantly better than the null model- a model with only the intercept. The test statistic is the difference between the residual deviance for the model with predictors and the null model. The test statistic has chi-squared distribution with degrees of freedom equal to the differences in degrees of freedom between the current and the null model (the number of predictor variables in the model). We calculated below the difference in deviance for the two models,

```
\begin{verbatim}
 B1$deviance
```

```
[1] 5077.387
> glm(injuryb~1)$deviance
[1] 4739.102
> (glm(injuryb~1)$deviance-B1$deviance)/
glm(injuryb~1)$deviance
[1] -0.07138158
```

```
> ## change in deviance
> with(B1, null.deviance - deviance)
[1] 626.1778
> ## change in degrees of freedom
> with(B1, df.null - df.residual)
[1] 25
> ## chi square test p-value
> with(B1, pchisq(null.deviance - deviance,
df.null - df.residual, lower.tail = FALSE))
[1] 4.04921e-116
```

```
> logLik(B1)
'log Lik.' -2538.693 (df=26)
```

The degrees of freedom for the difference between the two models is equal to the number of predictor variables in the models, and can be obtained using:
In OLS regression, we find the best fitting line by minimising the squared residuals. In logistic regression, a different approach is used-that is Maximum Likelihood (ML). ML is a way of finding the smallest possible deviance between the observed and predicted values using calculus precisely. ML tries different iterations in which it tries different

solutions until it gets the smallest possible deviance or best fit and provides a final value for the deviance. The deviance statistic is called âĂŞ2LL, and it can be thought of as a chi-square value. The likelihood ratio test, G, is a chi-square difference test using the 'null' or intercept-only model. Instead of using the deviance (âĂŞ2LL) to judge the overall fit of a model, however, another statistic is usually used to compare the fit of the model with and without the predictor(s). The difference between these two deviance values is often referred to as G for goodness of fit.

$$G = \chi^2 = D(null) - D(with/predictors)$$

or, using the using Cohen et al. notation,

$$G = \chi^2 = D_{null} - D_k = -2LL_{null} - (-2LL_k)$$

Where D_{null} is the deviance for the intercept-only model and D_k is the deviance for the model containing k number of predictors. Another equivalent formula is

$$G = \chi^2 = -2ln\left(\frac{L_{null}}{L_k}\right)$$

The chi-square of -516.38 with 17 degrees of freedom and the corresponding p-value of less than 0.0001 tells us that our model fits significantly better than the null model. This is sometimes called a likelihood ratio test (the deviance residual is -2 log likelihood). We plot the basic residual versus the predictors and versus the linear predictor which are Pearson residuals versus each of the predictors. Instead of plotting residuals against fitted values, however, the residual Plots function plots residuals against the estimated linear predictor. Each panel in the graph by default includes a smooth fit rather than a quadratic fit. In the binary regression, the plots of Pearson residuals or deviance residuals are strongly patterned, particularly the plot against the linear predictor, where the residuals can take on only two values, depending on whether the response is equal to

zero or one. Gender can take on two values, and so the residuals can take on four values for each value of the other factors. Even in this extreme case, however, a well-fitting model should have the conditional mean function in any residual plot be constant as we move across the plot. The fitted smooth helps us learn about the conditional mean function, and neither of the smooths shown is especially curved. The residuals for the variables are shown as a boxplot because all of these are factors. Unfortunately, it's not easy to interpret the boxplots because of the discreteness in the distribution of the residuals. Figures 10-15 shows residual plots of the injury data with different factors.

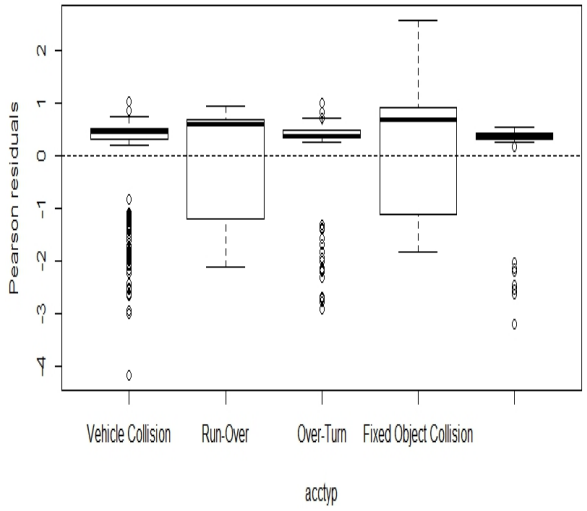


Figure A.3: Residuals plot (accident type)

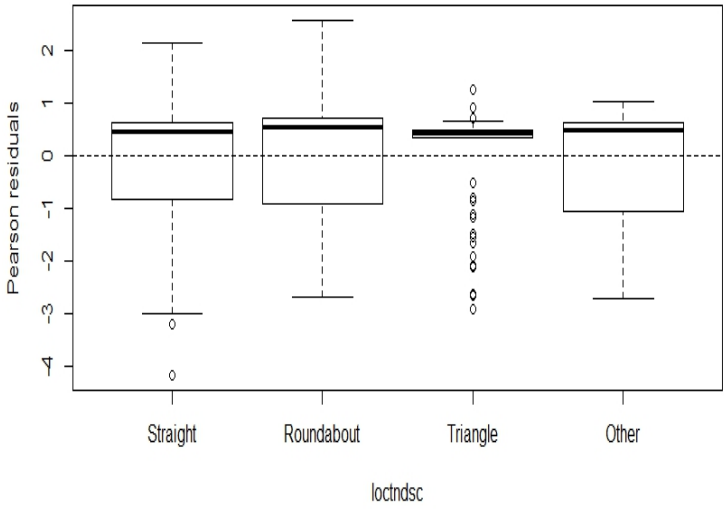


Figure A.4: Residuals plot of the injury data(location description)

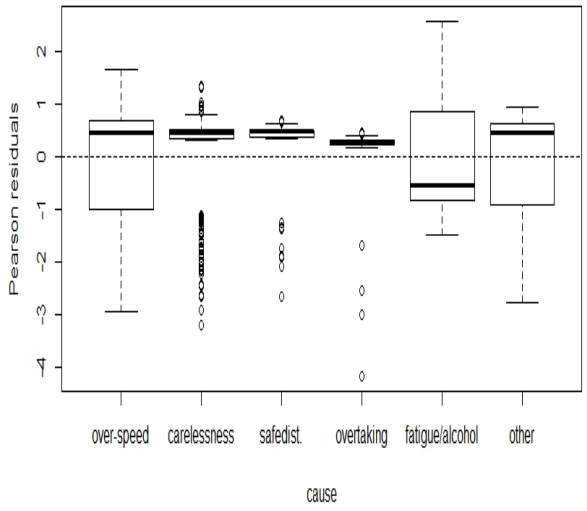


Figure A.5: Residuals plot (accident cause)

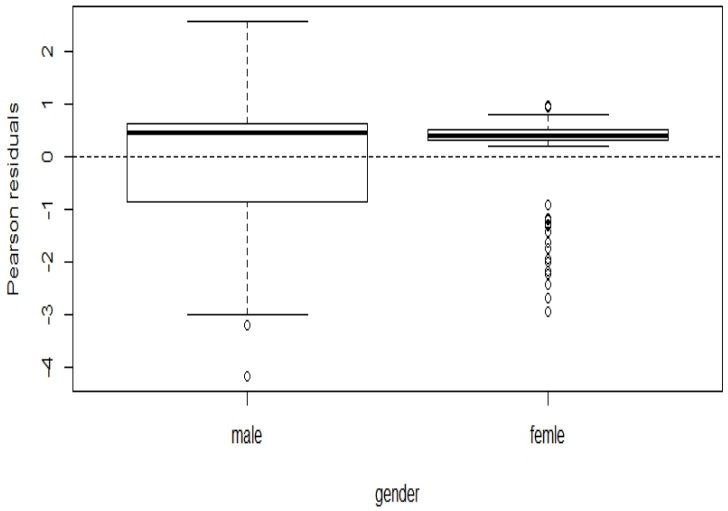


Figure A.6: Residuals plot (driver's gender)

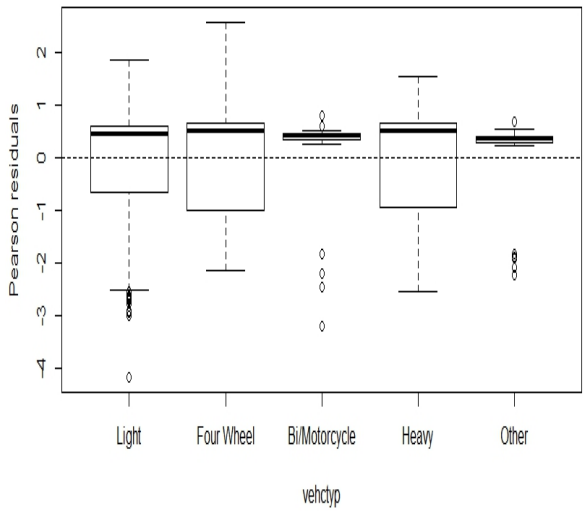


Figure A.7: Residuals plot of the injury data(vehicle type)

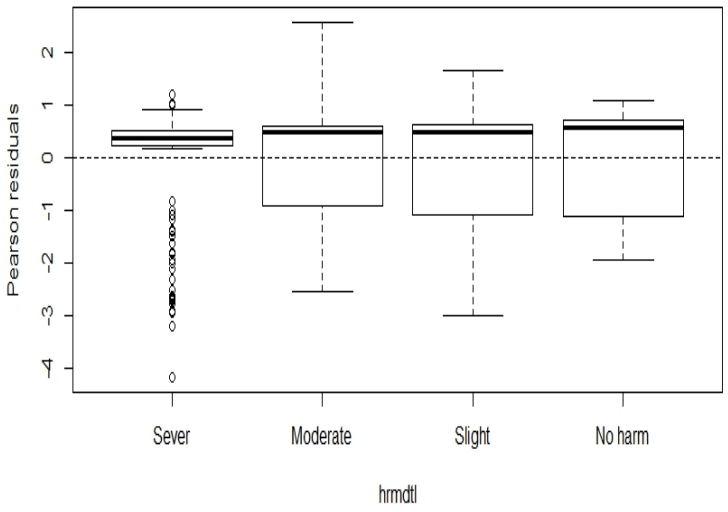


Figure A.8: Residuals plot (vehicle harm)

We calculate Cook's distance for GLMs approximately using

$$D = \frac{e^2_{PS_i}}{k+1} \times \frac{h_i}{1-h_i}$$

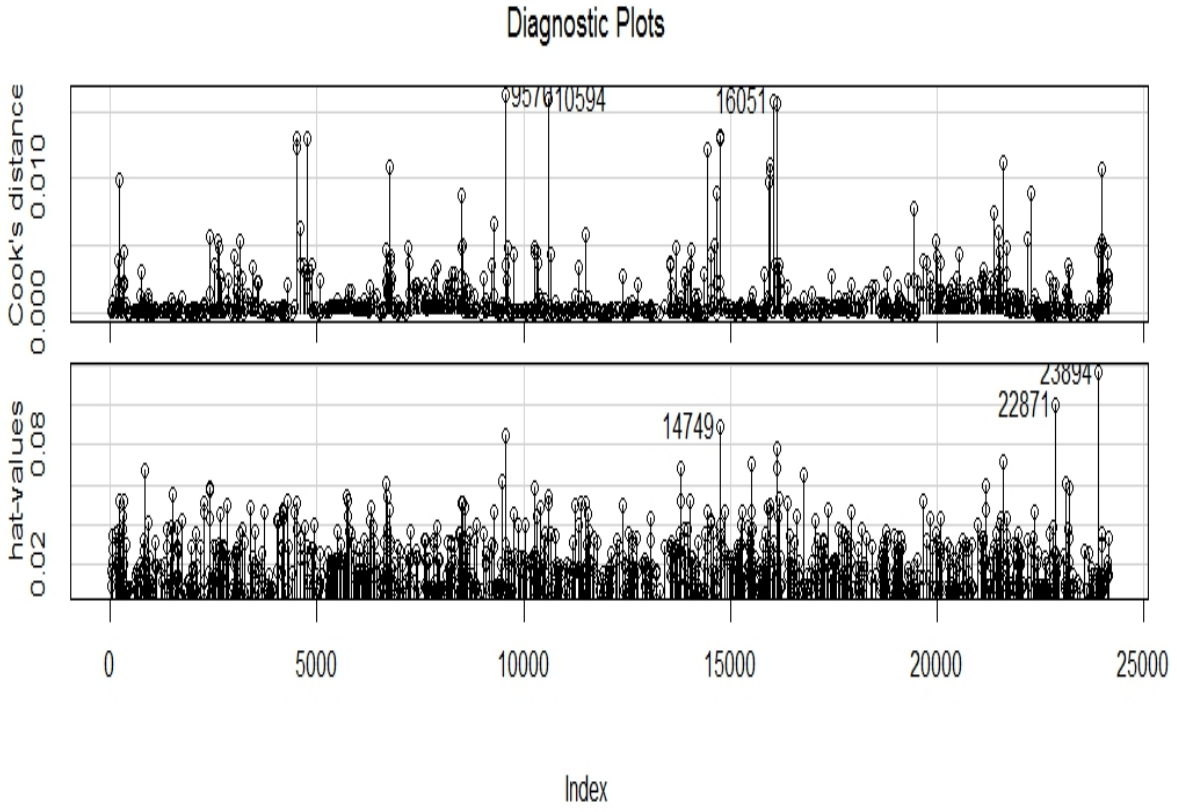


Figure A.9: Cook's distances and hat-values

In Figure 16 we have tow diagnostics to check for outliers; the Cook's distances and the hat-values. Generally, deleting one of the observations should not change the model's

estimates, however, sometimes deleting an observation changes the coefficient value of some variables. we can see the values that are returned by the Cook's distance calculation. Exact values of df β_{ij} and df βs_{ij} could be found directly from the final iteration of the IRWLS procedure. Figure 9 shows index plots of Cook's distances and hat-values, There is no indication of outliers.

Appendix B

Ordered Probit Model

In road safety management, the regulations and policies of road maintenance are adjusted based on the knowledge of the factors that influence the accident frequency and severity. Several research studies have been conducted over the years to identify these factors and to explain their influence on road traffic accidents. These factors commonly categorised into: driver attributes, vehicle attributes, road characteristics, and accident characteristics. The objectives of this research is to investigate the factors associated with injury severity level of the accidents in Oman. For this purpose, the ordered probit model is applied to a real dataset, which comprises 24,192 records related to all types of road accidents that took place all over Oman, between January, 2009 and April, 2012. Researches have applied a variety of statistical techniques including ordered logit model, generalised ordered logit model, and multivariate ordered-response probit model. However, the ordinal probit model was found to be better in recognising the increasing severity and the categorical nature of the ordinal independent variable. Furthermore, The model attracts researchers for being parsimonious in the number of parameters which makes it easy for interpretation than the other mentioned models.

B.1 Literature Review

Various number of studies have documented the application of the logistic regression to analyse the injury severity levels of traffic accident. Such multinomial dependent variables are ordinal by nature and other discrete choice models such as logit model mostly fail to account for the ordinal nature of such response variables. In particular, the ordered probit model have shown a robustness to handle such data in an efficient and parsimonious way. Xiea, Y., et, al (2012) analysed the injury severities of single-vehicle crashes on rural roads using a latent class logit model. The model has the advantage of not restricting the coefficients of each explanatory variable in different severity levels which helps to identify the effect of the explanatory variable on different severity levels. Weissa, H., Kaplanb, S., Pratob, C. (2014) developed a mixed logit model to account for heterogeneity and heteroscedasticity in the propensity to injury severity outcomes and for correlation between serious and fatal injuries. The model provided a better fit than a binary and a generalized ordered logit. They applied their analysis using a dataset of single-vehicle and two-vehicle crashes in New Zeeland which included at least one 15-24 year-old driver between 2002 and 2011. Their result showed that (1) seatbelt non-use, inexperience and alcohol use were the deadliest behavioural factors in single-vehicle crashes, while (2) fatigue, reckless driving and seatbelt non-use were the deadliest factors in two-vehicle crashes.

Another application of ordered model on injury severity was conducted by Garridoa, R., et. Al (2014). They used the model to examine the contributing factors to the injury severity of the occupants of the involved vehicles in road accidents in Coimbra. His findings suggest that (1) light-vehicles travelling at (2) two-way roads, and on (3)dry road surfaces result in more severe injuries than those who travel in (1) heavy-vehicles, at (2) one-way roads, and on (3) wet road surfaces. They also found that the (1) driver's seat seems to be safer than other positions in the involved vehicle, (2) urban

areas seem to experience less serious accidents than rural areas, and (3) women are more likely to face serious or fatal injuries than men. Obeng, K., Rokonuzzaman, R., (2013) applied ordered logit model for pedestrian injury severity from crashes at signalized intersections in a medium-size city. His findings is that (1) vehicle type, gender, land-use, speed limit, traffic volume, the presence of side-walks and visual obstruction significantly explain pedestrian injury severity in vehicle pedestrian crashes at signalized intersections. He also found that (2) females are remarkably involved in these crashes, (2) side-walks increase the probability of a pedestrian sustaining a serious injury while (3) passenger cars, sport utility vehicles and pick-ups are associated with less severe pedestrian injuries.

Abdel-Aty, M., (2003) has also conducted a study in the same context in which he applied ordered probit model. He analysed the driver injury severity in accidents at roadway sections, signalized intersections, and toll plazas in Central Florida. The three models showed that injury severity level was significantly affected by (1) driver's age, gender, seat-belt use, point of impact, speed, and vehicle type. While some variables like the driver's violation was significant only in the case of signalized intersections and alcohol, lighting conditions, and the existence of a horizontal curve were significant in the roadway sections only. Similarly, Chimba, D., et. Al, (2012) applied multinomial Logit model to analyse the influencing roadway features, traffic characteristics and environmental conditions on bicycle crash injury severities. The model has advantage for its flexibility in quantifying the effect of the independent variables for each injury severity categories. He found that severity of bicycle crashes increases with increase in vehicles per lane, number of lanes, bicyclist alcohol or drug use, routes with 35-45 mph posted speed limits, riding along curved or sloped road sections, when bicyclists approach or cross a signalized intersection, and at driveways. Also, the (1) routes with a high percentage of trucks, (2) roadway sections with curb and gutter,(3) cloudy or foggy weather and (4) obstructed vision were found to

have high probability of severe injury. Segments with wider lanes, wide median and wide shoulders were found to have low likelihood of severe bicycle injury severities. Limited lighting locations was found to be associated with incapacitating injury and fatal crashes, indicating that insufficient visibility can potentially lead to severe crashes.

Kockelman, K., Kweon, Y., (2002) used ordered probit model to examine the different probability of injury levels when applying the model to (1) all crash types, (2) two-vehicle crashes, (3) and single-vehicle crashes. The results suggest that (4) pick-ups and sport vehicles are less safe than passenger cars under single-vehicle crash conditions. In two-vehicle crashes, however, (5) these vehicle types are associated with less severe injuries for their drivers and more severe injuries for occupants of their collision partners. Other findings is that (4) males and younger drivers in newer vehicles at lower speeds sustain less severe injuries. Khan, G., Bill, A., Noyce, D., (2015) studied the feasibility of using GUIDE Classification Tree method to analyse the severity of CMCs to discover if any additional information could be revealed. Additionally, the effects of variable types (continuous or discrete), misclassification costs, and tree pruning characteristics on models results were also explored. showing that the GUIDE Classification Trees revealed new variables (median width and traffic volume) that affect CMC severity and provided useful insight on the data. The results of this research suggest that the use of Classification Tree analysis should at least be considered in conjunction with regression-based crash models to better understand factors affecting crashes. Classification Tree models were able to reveal additional information about the dependent variable and offer advantages with respect to multicollinearity and variable redundancy issues.

Mamdoohi, A., et. al (2014) used a binary model to estimate the severity of accidents in Tehran urban which can be used in road safety planning. Human characteristics and collision attributes were employed to act as surrogates for point of impact. Results

indicate that wearing seat belt decreases the probability of accidents resulting in injury. Furthermore, road misconduct, as a human reason of an accident, results in the most severe accidents compared to other human reasons. In the other hand, as a consequence of accidents caused by other non-human reasons, property damage only was found to be the most probable outcome. Finally, drivers involved in front to front collision types were most prone to injury. Other factors in decreasing order are: front to rear, front to side, other types of collision, rear to side, and side to side. A review study that was conducted by Mujalli, R., and Ona, J. (2011) investigated 19 modelling techniques used in injury severity analysis of traffic accidents that involved a 4-wheeled vehicle. They compared between the models performance based on seven criteria which are modelling method, number of observations, number of covariates, area type, features, injury level and model fit. Their conclusion was that it is not possible to recommend a method as the best one. Each modelling technique has its own limitations and characteristics, awareness of which will help analysts to decide the best method to be used in each particular modelling advantages and disadvantages. However, their general conclusions is that in most cases the results of model' fits are found to be satisfactory, though not excellent; in the case of data mining models, accuracy improves with balanced datasets; and no correlation was found between the number of accident records and the number of analysed variables.

B.2 Methodology: Ordered Probit Model

In here, the dependent variable, injury-severity of an accident has five discrete ordinal levels: fatal injury, severe injury, moderate injury, slight injury and no injury. The ordered probit model recognizes the ordinal (increasing severity) and categorical nature of such independent variable and is also much easier to interpret than the counterparts models because of its structure which provides a parsimonious number of parameters. The model is usually built around the notion of the latent underlying injury risk

propensity occurring from a road accident that determines the observed ordinal injury severity level. Suppose the injury-severity level is a count variable y_i that takes values $0, 1, 2, ..., m$. Define the $m + 1$ indicator variable

$$d_{ij} = \begin{cases} 1 & \text{if } y_i = j; \\ 0 & \text{if } y_i \neq j. \end{cases}$$

Also define the corresponding probabilities $Pr[d_{ij} = 1] = P_{ij}, \quad j = 0, ..., m$, where p_{ij} may depend on regressors and parameters. Then the density function for the i_{th} observation can be written

$$f(y_i) = f(d_{i0}, d_{i1}, ..., d_{im}) = \prod_{j=0}^{m} P_{ij} d_{ij} \tag{B.1}$$

and the log-likelihood function is

$$lnL = \sum_{i=1}^{n} \sum_{j=0}^{m} d_{ij} ln P_{ij} \tag{B.2}$$

Now, the ordered probit model latent(unobserved) random variable is

$$y_i^* = \beta X_i^{'} + \epsilon_i \tag{B.3}$$

where y^* is an unobserved latent and continuous variable measuring the injury severity resulted from accident i. β denotes a row vector of parameters to be estimated; X_i is a column vector of observed explanatory variables; ϵ_i is a random error term which is assumed to follow a standard normal distribution, $\epsilon \sim N[0, 1]$. The observed discrete data variable y_i is generated from the unobserved y_i^* in the following way $y_i = j$ if

$$\alpha_j < y_i^* \leq \alpha_{j+1}, \tag{B.4}$$

where $j = 0, ..., m$ and $\alpha_0 = -\infty$ and $\alpha_{m+1} = \infty$ it follows that

$$P_{ij} = Pr[\alpha_j < y_i^* \leq \alpha_{j+1}]$$

$$= Pr[\alpha_j - \beta X_i^{'} < \epsilon_i \leq \alpha_{j+1} - \beta X_i^{'}$$

$$= \Phi[(\alpha_{j+1} - \beta X_i^{'}) - \Phi(\alpha_j - \beta X_i^{'})]$$

where $\Phi(.)$ is the standard normal cdf, $j = 0, 1, 2, ..., m$ and $\alpha_{m+1} = \infty$, The log-likelihood function with probabilities is

$$lnL = \sum_{i=1}^{n} \sum_{j=0}^{m} d_{ij} ln[\Phi(\alpha_{j+1} - X_i^{'}\beta) - \Phi(\alpha_j - X_i^{'}\beta)] \tag{B.5}$$

The observed and discrete injury-severity variable, Y_i, is given as

$$y_i = \begin{cases} 0, & \text{if } -\infty \leq y_i^* \leq \mu_1 \text{ (no injury)} \\ 1, & \text{if } \mu_1 < y_i^* \leq \mu_2 \text{ (slight)} \\ 2, & \text{if } \mu_2 < y_i^* \leq \mu_3 \text{ (moderate)} \\ 3, & \text{if } \mu_3 < y_i^* \leq \mu_4 \text{ (sever)} \\ 4, & \text{if } \mu_4 < y_i^* \leq \infty \text{ (fatal).} \end{cases}$$

Estimation of β and $\alpha_1, ..., \alpha_m$ by maximum likelihood is straightforward. Identification requires a normalization, such as 0, for one of $\alpha_1, ..., \alpha_m$ or for the intercept term in β. So for given X_i the predicted probabilities of the five injury severity levels sustained in accident i can be illustrated as

$$P_i(0) = Pr(Y_i = 0) = Pr(y_i^* \leq \mu_1) = Pr(\beta X_i + \epsilon_i \leq \mu_1)$$

$$= Pr(\epsilon_i \leq \mu_1 - \beta X_i) = \Phi(\mu_1 - \beta X_i),$$

$$P_i(1) = Pr(Y_i = 1) = Pr(y_i^* \leq \mu) = Pr(\beta X_i + \epsilon_i \leq \mu_1)$$

$$= Pr(\epsilon_i \leq \mu_1 - \beta X_i) = \Phi(\mu_2 - \beta X_i) - \Phi(\mu_1 - \beta X_i),$$

$$P_i(2) = Pr(Y_i = 2) = Pr(y_i^* \leq \mu) = Pr(\beta X_i + \epsilon_i \leq \mu_1)$$

$$= Pr(\epsilon_i \leq \mu_1 - \beta X_i) = \Phi(\mu_3 - \beta X_i) - \Phi(\mu_2 - \beta X_i),$$

$$P_i(3) = Pr(Y_i = 3) = Pr(y_i^* \leq \mu) = Pr(\beta X_i + \epsilon_i \leq \mu_1)$$

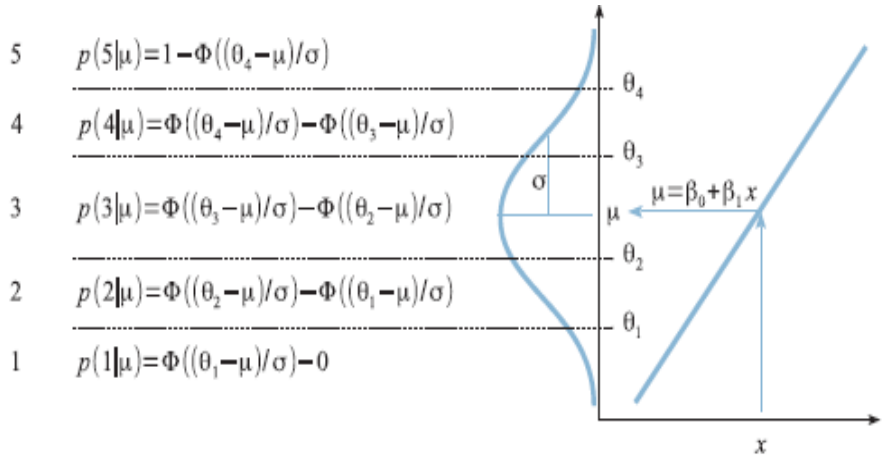
$$= Pr(\epsilon_i \leq \mu_1 - \beta X_i) = \Phi(\mu_4 - \beta X_i) - \Phi(\mu_3 - \beta X_i),$$

$$P_i(4) = Pr(Y_i = 4) = Pr(y_i^* \leq \mu) = Pr(\beta X_i + \epsilon_i \leq \mu_1)$$

$$= Pr(\epsilon_i \leq \mu_1 - \beta X_i) = 1 - \Phi(\mu_4 - \beta X_i),$$

Interpretation of parameters: Positive signs suggest greater injury severity as an increase in the value of corresponding variables while negative signs indicate the opposite results. The influence of certain variable on the probabilities of injury severity cannot be adequately interpreted through directly viewing only the estimated parameter, since a negative parameter may in fact lead to an increase in probability. It is therefore more helpful to examine the marginal effect of each variable on the probabilities of different accident injury severity levels. Figure 17 shows the cumulative-normal regression curve where in the right side shows metric predictor variable mapped to metric underlying variable, as in simple linear regression and the left side shows a mapping from metric

underlying to observed ordinal variable.



]Right side shows metric predictor variable mapped to metric underlying variable, as in simple linear regression. Left side shows mapping from metric underlying to observed ordinal variable. Copyright(c) by John Kruschke and Elsevier.

Figure B.1: Cumulative-normal regression

B.3 Data

The dataset used in the study was extracted from police records of crashes reported between 2009 to 2012 and involved all intersection accidents in Oman. Table 14 gives further descriptive statistics and more details about the research variables. The crash injury severity in five categories: fatal, sever, Moderate, slight and no injury. A dataset involving information about 24,192 accidents was analysed in which the fatal accidents were (4.5 %), the sever accident were (3.9%), the moderate accidents were (14.8%), the slight injury accident represented (25.3%) and the accidents with no injury were (51.5%). Figure 18 shows the distribution of accidents by driver characteristics and Figure 19 shows the accident distribution by vehicle characteristics.

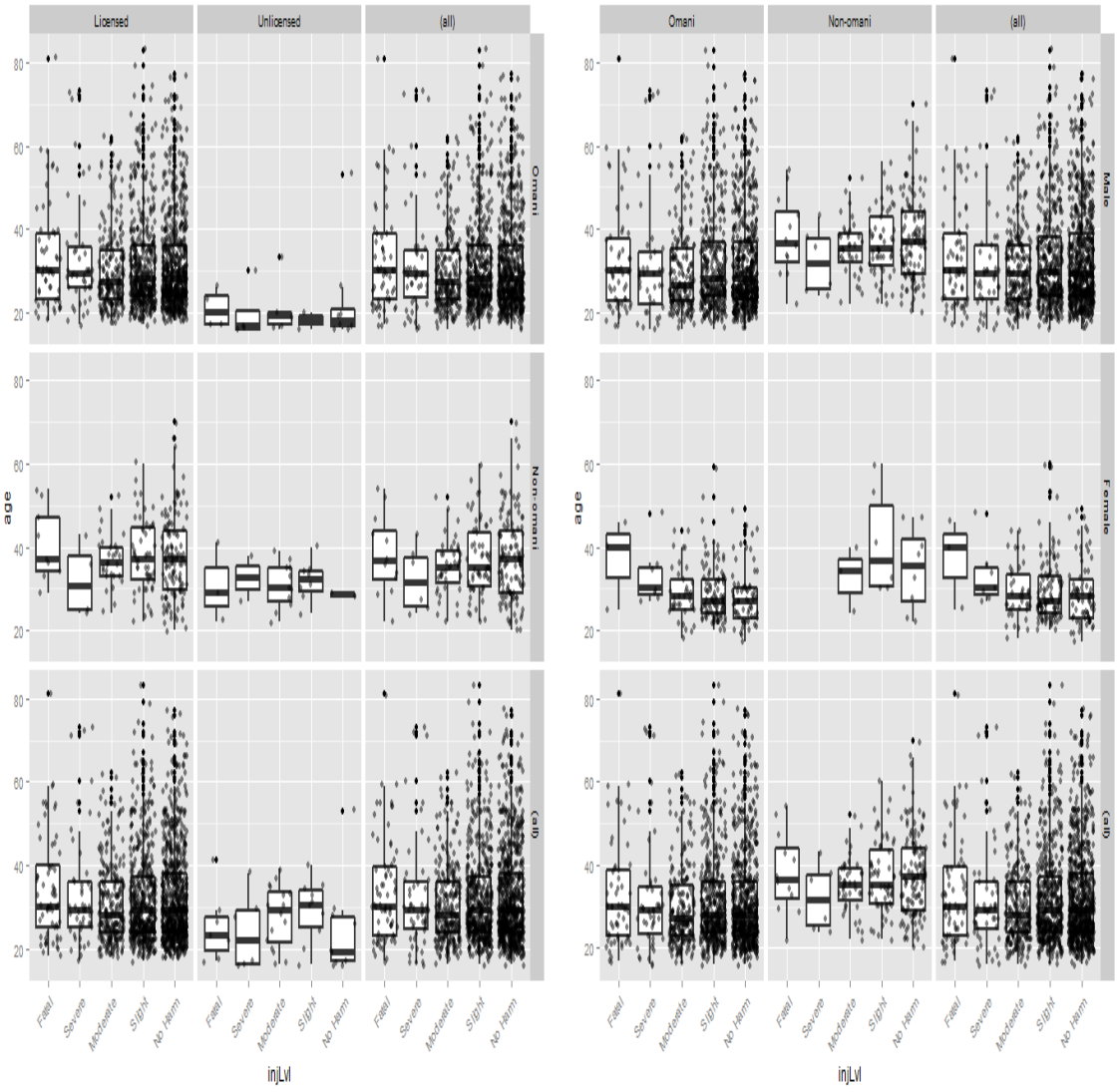


Figure B.2: Distrib. of accidents by driver characteristics

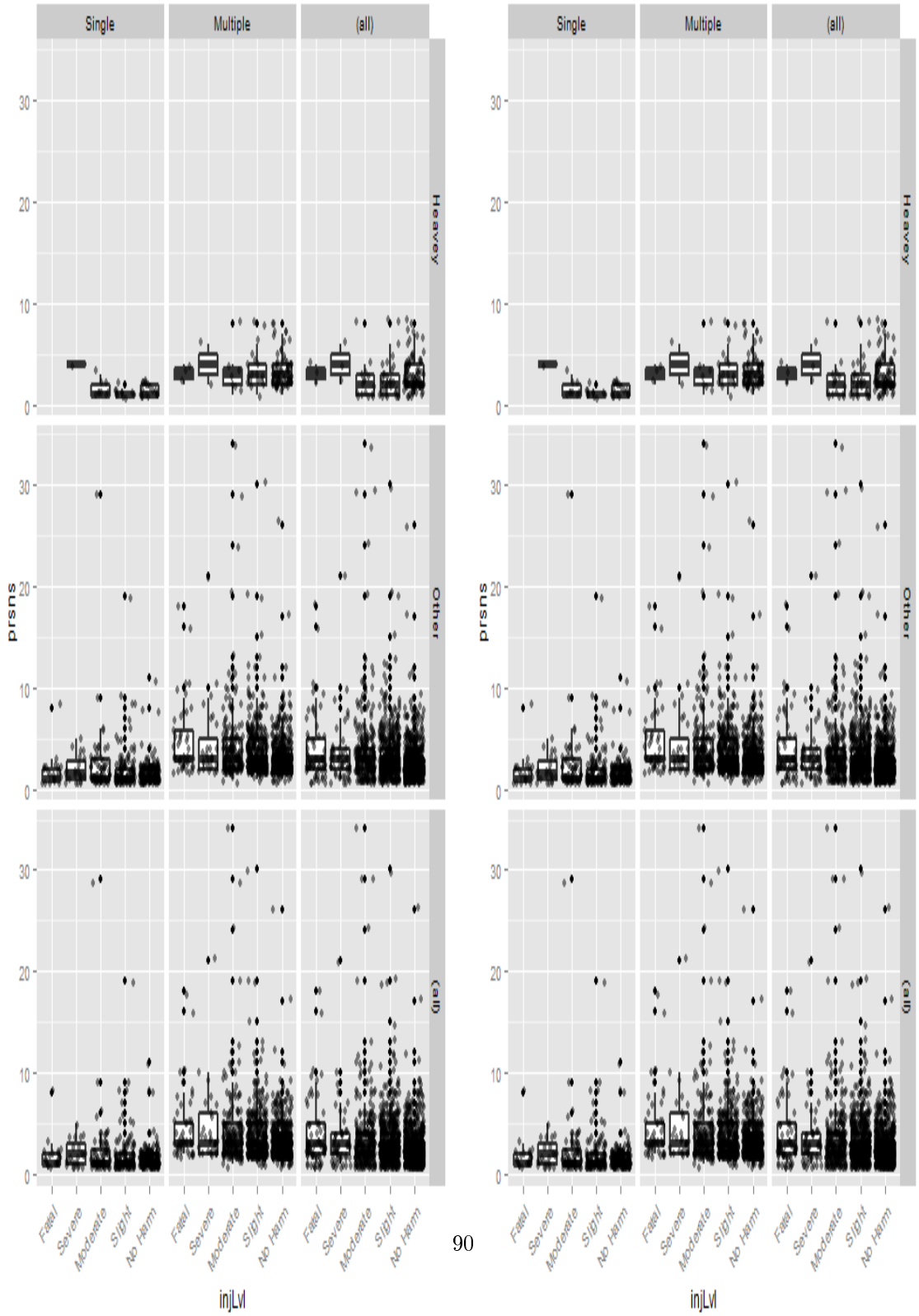


Figure B.3: Distrib. of accidents by vehicle characteristics

Table B.1: Description of research variables

Variable	Description	Mean	SD
Dependent variable(injLvl)	5 crash severity categories		
Fatal	1 if severity level is fatal 0 otherwise	0.045	0.912
Severe	1 if severity level is sever 0 otherwise	0.039	0.924
Moderate	1 if severity level is moderate 0 otherwise	0.148	0.726
Slight	1 if severity level is slight 0 otherwise	0.253	0.558
No Harm	1 if severity level is no Harm 0 otherwise	0.515	0.236
Accident attributes			
Time	time of accident	1371.322	608.031
Type of Accident			
VclColsion	1 if accident type is vehicle collision 0 otherwise	0.434	0.566
Runover (person or animal)	1 if accident type is run-over 0 otherwise	0.131	0.869
Overturn	1 if accident type is over-turn 0 otherwise	0.161	0.839
FxdObjctCol	1 if accident type is fixed object collision 0 otherwise	0.230	0.770
Mot/Bicycle	1 if accident type is vehicle collision 0 otherwise	0.044	0.956
Cause			
High speed	1 if accident cause is high speed 0 otherwise	0.513	0.487
Wrong Conduct	1 if accident cause is wrong conduct 0 otherwise	0.235	0.766
Carelessness	1 if accident cause is carelessness 0 otherwise	0.074	0.926
Fatigue	1 if accident cause is fatigue 0 otherwise	0.026	0.974
Overtaking	1 if accident cause is overtaking 0 otherwise	0.047	0.954
Climcond	1 if accident cause is climate condition 0 otherwise	0.009	0.992
safedist.	1 if accident cause is safe distance 0 otherwise	0.059	0.941
Vehicle	1 if accident cause is vehicle 0 otherwise	0.0277	0.972
Road	1 if accident cause is road 0 otherwise	0.009	0.991
Driver Characteristics			
Gender			
Male	1 if driver's gender was male 0 otherwise	0.895	0.011
Female	1 if driver's gender was male 0 otherwise	0.105	0.801
Nationality			
Omani	1 if driver's nationality was Omani 0 otherwise	0.828	0.029
Non-Omani	1 if driver's nationality was Non-Omani 0 otherwise	0.172	0.686
License Status			
Licensed	1 if driver's nationality was licensed 0 otherwise	0.956	0.044
Unlicensed	1 if driver's nationality was unlicensed 0 otherwise	0.044	0.956
Age			
age	age of the driver	30.884	10.6262
Vehicle Characteristics			
Vehicle type			
Saloon	1 if the vehicle type is saloon 0 otherwise	0.633	0.135
Pick up	1 if the vehicle type is pick up 0 otherwise	0.114	0.786
Four wheel	1 if the vehicle type is four wheel 0 otherwise	0.109	0.793
Truck	1 if the vehicle type is truck 0 otherwise	0.027	0.946
Bus	1 if the vehicle type is bus 0 otherwise	0.083	0.840
Bi/Motorcycle	1 if the vehicle type is bi/motorcycle 0 otherwise	0.031	0.940
Others	1 if the vehicle type is others 0 otherwise	0.003	0.994
Vehicle involved count			
Vehicle	Number of involved vehicle	1.545	0.715
Road Characteristics			
Road Type			
Main	1 if the road type is main 0 otherwise	0.602	0.158
Sub	1 if the road type is sub 0 otherwise	0.379	0.386
Unpaved	1 if the road type is unpaved 0 otherwise	0.0189	0.963
Accident Location			
Straight	1 if the accident location is straight 0 otherwise	0.790	0.044
Side	1 if the accident location is side 0 otherwise	0.033	0.935
Intersection	1 if the accident location is intersection 0 otherwise	0.066	0.872
Roundabout	1 if the accident location is roundabout 0 otherwise	0.053	0.896
Signal	1 if the accident location is signal 0 otherwise	0.020	0.960
Others	1 if the accident location is other 0 otherwise	0.036	0.929
Weather Condition			
Normal	1 if the weather condition is normal 0 otherwise	0.978	0.021
Abnormal	1 if the weather condition is abnormal 0 otherwise	0.022	0.979

B.4 Injury-severity analysis using ordered probit model

This section presents an analysis of the injury-severity level of the accidents in Oman using ordered probit model. We apply the probit model to a dataset that consist of all accidents that happened around Oman during the period January, 2009 to April, 2012. We fit the ordered logistic regression model with the link "probit" using the polr command from the MASS package in R language to estimate an ordered proportional odds logistic regression. The regression output includes the coefficient of the covariates, standard errors, and t-test. It also includes the estimate of the intercepts μ_m where the latent variable y^* is cut to make the five groups that we observe in the data, which are sometimes called cut points. The latent variable is continuous measure of injury severity faced by driver in a crash i. The coefficients are bit confusing in interpretation because they are scaled in terms of logs. These coefficients of order probit model are interpreted similar as in the binary logistic regressions using odds ratios and are called proportional odds ratios. To get the OR of the estimates and confidence intervals, we exponentiate the coefficient values. In the output , we also get the residual deviance, $2Loglikelihood$ of the model as well as the AIC for the model comparison. We get confidence intervals for the parameter estimates by profiling the likelihood function or by using the $\beta \pm percentile \times SE(\beta)$, where SE is the standard error, t-testis the ratio of the coefficient to its standard error. If the 95% CI does not cross 0, the parameter estimate is statistically significant.

An important assumptions of the ordinal probit regression is the parallel regression assumption which indicates that the relationship between each pair of outcome groups is the same. In other words, the coefficients that describe the relationship between, the lowest level versus all higher levels of the response variable are the same as those that describe the relationship between the next lowest level and all higher severity levels,

etc. Because the relationship between all pairs of groups is assumed to be the same, there is only one group of coefficients. Thus, in order to assess the appropriateness of the fitted model, we need to evaluate whether the proportional odds assumption holds with the null hypothesis that the sets of coefficients are the same. In Figure 4, the values displayed are (linear) predictions produced by our probit model when we regress the dependent variable(injLvl) on our independent variables one each time. The parallel slopes assumption is examined through running binary logistic regressions with varying cutpoints on the dependent variable to check the equality of coefficients across cutpoints.

Our dependent variable has 5 levels, labelled 1, 2, 3, 4, 5. We graph the probability that y is greater than or equal to a given value for each level of y. We use the predicted logits to test the proportional odds assumption using one predictor (x) variable at a time, where the outcome groups (severity levels) are defined by either injLvl >= 2 and injLvl >= 3. **If the difference** between predicted logits for varying levels of a predictor, say gender, are the same whether the outcome is defined by injLvl >= 2 or injLvl >=3, then we can be confident that **the proportional odds assumption holds**. For example **if the difference between** logits for gender = 0 (female) and gender = 1 (male) is the same when the outcome is injLvl >= 2 as the difference when the outcome is injLvl >= 3, then the proportional odds assumption likely holds. We calculate the log odds of being greater than or equal to each value of the target variable. For gender, we would say that for a one unit increase in gender, i.e., going from 0 (male) to 1 (female), the odds of "fatal" accident versus "Sever" or other severity levels combined are -0.208 greater, given that all of the other variables in the model are held constant. Likewise, the odds "fatal" or other severity level versus "no harm" accident is -0.208 times greater, given that all of the other variables in the model are held constant. For age (and other continuous variables), the interpretation is that when a driver age moves 1 unit, the odds of moving from "no harm" accident to "slight" or

other severity levels (or from the lower and middle categories to the high category) are multiplied by 0.005.

Table B.2: units of ordered logits (ordered log odds)

		N	Y>=1	Y>=2	Y>=3	Y>=4	Y>=5
timhour	[100,1100)	7011	Inf	2.928193	2.2524747	1.0742278	0.00142633
	[1100,1500)	5706	Inf	3.325690	2.5824436	1.3604262	0.06030569
	[1500,2000)	6366	Inf	3.075227	2.4798124	1.2433959	0.10124883
	[2000,2400]	4991	Inf	2.921685	2.2661069	1.1263209	0.08219411
acctyp	Vehicle Collision	10440	Inf	2.864684	2.2132918	1.0111112	-0.14931896
	Run-Over	3144	Inf	3.436617	2.6732990	1.3779659	0.27396545
	Over-Turn	3877	Inf	3.178591	2.4748899	1.2974237	0.12862688
	Fixed Object	5549	Inf	3.270898	2.6244753	1.4398663	0.31285322
	Motor/Bicycle	1064	Inf	2.638051	2.0532757	1.0103197	-0.10536052
cause	Overspeed	12352	Inf	2.869629	2.2073403	1.0281696	-0.09527927
	Wrong conduct	5646	Inf	3.724515	2.9719983	1.4787378	0.21119598
	Carelessness	1792	Inf	4.007333	3.2027464	1.9979375	1.05445500
	Fatigue	634	Inf	3.074675	2.6458370	1.6991667	0.76015789
	Overtaking	1131	Inf	2.041220	1.3984964	0.4812252	-0.55302198
	Climcond	216	Inf	3.396424	2.5257286	1.2527630	-0.22314355
	safedist.	1416	Inf	4.606598	3.2870841	1.6664482	0.26998978
	Vehicle	668	Inf	2.262094	1.7847908	0.7590053	-0.61115221
	Road	219	Inf	2.297573	1.8024548	0.8625653	-0.57219457
roadtyp	Main	14496	Inf	2.750020	2.1312305	0.9886528	-0.18179039
	Sub	9126	Inf	3.911688	2.9971138	1.5914006	0.45106306
	Unpaved	452	Inf	2.332144	1.9358345	0.9950716	-0.02655023
climate	Normal	23556	Inf	3.070989	2.4029073	1.2037337	0.06726937
	Abnormal	518	Inf	2.328338	1.7760817	0.8053592	-0.34309478
gender	Male	21544	Inf	2.978372	2.3170174	1.1735480	0.10816326
	Female	2530	Inf	3.967075	3.2329431	1.3813610	-0.36772478
national	Omani	19940	Inf	3.109229	2.4329807	1.2120529	0.04353747
	Non-omani	4134	Inf	2.795435	2.1796100	1.1109062	0.13081030
age	[1,25)	7270	Inf	2.990258	2.3242620	1.1480096	-0.07376107
	[25,29)	5340	Inf	3.288868	2.5136561	1.2085227	0.03296179
	[29,37)	5884	Inf	3.123501	2.3853536	1.1781550	0.05916069
	[37,96]	5580	Inf	2.860875	2.3518661	1.2600357	0.25586713
licns	Licensed	23025	Inf	3.178054	2.5122274	1.2964294	0.12062379
	Unlicensed	1049	Inf	1.650023	0.9056382	-0.4961083	-1.55439876
vhcls	1	12880	Inf	2.984378	2.3233429	1.1400365	0.03571808
	2	9669	Inf	3.148195	2.4657907	1.2477834	0.07180662
	[3,37]	1525	Inf	3.004720	2.4334707	1.3298733	0.16694403
Overall		24074	Inf	3.048621	2.3855428	1.1942960	0.05850301

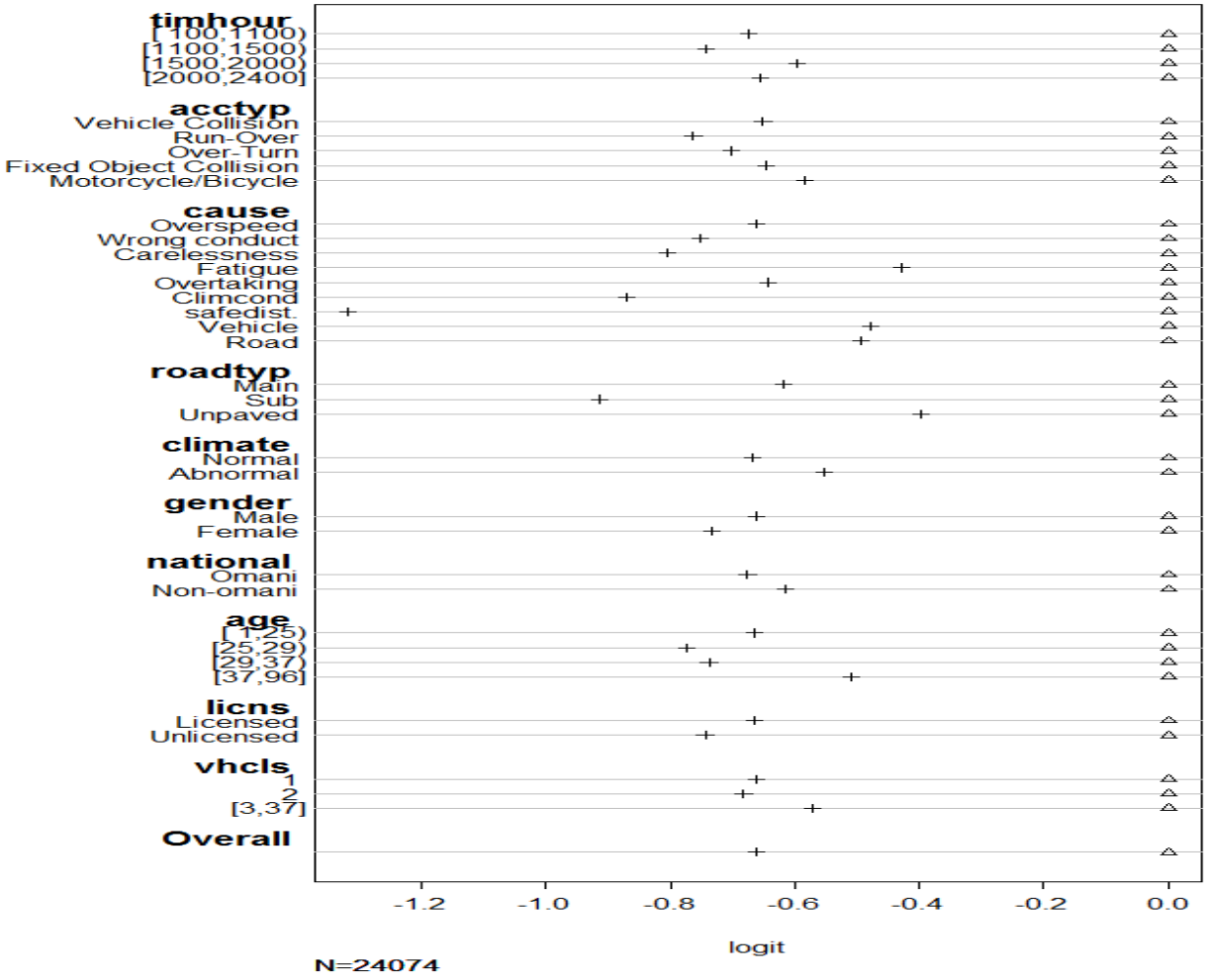


Figure B.4: Test of the proportional odds assumption

In the output in Table 15 the estimates are given in units of ordered logits (ordered log odds). The coefficients with positive signs indicate increase in injury severity level as they increase while negative signs indicate decrease in injury severity level as they increase. For continuous variables like age of driver, time in hours and count of vehicles, the severity level will increase as the value of the variable increase and decrease as the variable value increase in case the coefficient sign is negative. In

the other hand, for indicative variables like gender, we would say that the severity increases if we have female (1) accident compared to having male (0) accident in case of positive sign coefficient and decrease if the coefficient is negative. Similarly, we for a factor with more than two levels like cause of accident we would say that if the cause is high speed equal 1 and 0 otherwise, the severity would increase in case of positive sign coefficient and decrease in case of negative sign coefficient.

So our results showed that for the **type of accident** factor, only if the accident type is motor/bicycle the logits of the injury severity decreases by -0.217 as we go from fatal to sever or any of the severity levels having all the covariates held constant. For the other accident types, the severity of the accident increase when moving between the levels of severity. For the **cause of accident** if the accident cause is fatigue, overtaking and vehicle the severity of accident decrease given the rest of the covariates are held constant. The other causes increase the severity of the accident. For the **road type** factor, the severity of the accident increase in the sub roads and unpaved than the main roads. for the **weather condition** the severity of the accident appears to be decreasing in the abnormal weather conditions. for the **gender of the driver**, the severity of the accident decrease if the driver is female. The models showed that severity increase if the **driver nationality** is not Omani. For **the age** of the driver the severity increase with increase in the age. For the **license status**, the severity of the accident decrease if the driver is unlicensed. The models showed that the severity of the accident decrease if the **vehicle type** is not heavy. The models showed that the increase in **number of vehicles** involved reduces the severity of the accidents.

Table B.3: Ordered probit model: influencing factors on injury severity levels

Variable	Estimate	Std. Error	t-value	95%CI	p-value
acctypRun-Over	0.264	0.024	11.002	(0.217, 0.312)	< 0.0001
acctypOver-Turn	0.174	0.022	7.919	(0.131, 0.217)	< 0.0001
acctypFixed Object Collision	0.271	0.020	13.357	(0.231, 0.311)	< 0.0001
acctypMotorcycle/Bicycle	0.069	0.036	1.905	$(-0.004, 0.139)$	0.057
causeWrong conduct	0.192	0.021	9.200	(0.151, 0.233)	< 0.0001
causeCarelessness	0.654	0.034	19.448	(0.588, 0.719)	< 0.0001
causeFatigue	0.407	0.050	8.103	(0.308, 0.505)	< 0.0001
causeOvertaking	-0.343	0.035	-9.681	$(-0.412, -0.273)$	< 0.0001
causeClimcond	0.313	0.094	3.331	(0.129, 0.497)	0.001
causesafedist.	0.353	0.035	10.110	(0.285, 0.422)	< 0.0001
causeVehicle	-0.191	0.043	-4.403	$(-0.278, -0.108)$	< 0.0001
causeRoad	-0.163	0.075	-2.176	$(-0.311, -0.018)$	0.030
roadtypSub	0.398	0.016	24.736	(0.366, 0.429)	< 0.0001
roadtypUnpaved	0.100	0.055	1.805	$(-0.010, 0.207)$	0.071
locationSide	-0.354	0.040	-8.939	$(-0.432, -0.277)$	< 0.0001
locationIntersection	-0.050	0.030	-1.669	$(-0.108, 0.009)$	0.095
locationRoundabout	0.354	0.035	10.190	(0.284, 0.421)	< 0.0001
locationSignal	0.298	0.054	5.475	(0.190, 0.403)	< 0.0001
locationOthers	0.125	0.041	3.084	(0.045, 0.204)	0.002
climateAbnormal	-0.266	0.061	-4.381	$(-0.385, -0.147)$	< 0.0001
genderFemale	-0.208	0.024	-8.746	$(-0.254, -0.160)$	< 0.0001
age	0.002	0.001	2.401	(0.00003, 0.003)	0.016
licnsUnlicensed	-1.138	0.035	-32.697	$(-1.215, -1.076)$	< 0.0001
heavyVclOther	-0.157	0.028	-5.636	$(-0.206, -0.095)$	< 0.0001
vhcls	-0.038	0.012	-3.050	$(-0.062, -0.013)$	0.002
No Harm |Slight	-1.666	0.042	-39.994		< 0.0001
Slight |Moderate	-1.320	0.041	-32.449		< 0.0001
Moderate |Severe	-0.617	0.040	-15.459		< 0.0001
Severe |Fatal	0.135	0.040	3.397		0.001
Observations:	24,074				
Res.Dev.:	56582.05				
AIC:	56640.05				

B.5 Modelling injuries by number of vehicles involved

In Table 17, we present analysis of the injury data by the count of vehicles involved. We sort the data to three sets; accidents with single-vehicle, accidents with two-vehicles and accidents with multiple-vehicles. We Compare the results of these groups by the result when modelling the full dataset of all accidents. The severity of the accident appears to be influenced by factors differently for each group of data. The results shows that most of the factors are significant in explaining the severity level of the

accident injuries though there are cases where some variables coefficient equals 0 such as the coefficient of the age in the two- and multiple-vehicle accidents. Having a large dataset, most of the variables appears significant when modelling injuries using the full dataset of the 24,192 accidents. However, looking to the t-value of each covariates enable us to see the size of the effect. The first and second groups still have large sample sizes that is enough to get reasonable estimate of the factors coefficients. Yet, the third group estimates could be affected by the sample size since that is less than 2000 observations, though it is still acceptable size for such analysis especially that sorting the data should reduce the variability that was present in the full data set.

Comparing the results of the three groups of accidents, the **time of the accident** appears to be not significant in explaining the severity of the accident though it appears that injury-severity increases as the time goes on through the day but the coefficient was almost 0 and was eliminated by the step function. The **type of accident** is highly significant in single-vehicle accident more than in the two-vehicle while in multiple-vehicle accidents, the accident type have 0 coefficients and hence does not seem to give information about the injury-severity in the multiple-vehicle accidents. The **cause of accident** appears to be highly significant in the single- and two-vehicle accidents. For the **road type** factor, is highly significant in explaining the severity of single- and two-vehicle accidents. The severity of the accident seems to decrease in the abnormal **weather conditions**. For the **gender of the driver**, in both models the severity of the accident decreases if the driver is female. The models showed that severity increase if the **driver nationality** is not Omani. For **the age** of the driver, the severity increases with the increase in the age. For the **license status**, the severity of the accident decreases if the driver is unlicensed. The models showed that the severity of the accident decreases if the **vehicle type** is not heavy. All the models showed that the increase in **number of vehicles** involved reduces the severity of the accidents.

B.6 Conclusion

In this study, the ordered probit model was used to investigate the influence of different accident attributes on the injury severity faced by road users. This research models traffic injury severity in Oman which is a developing country using a four year accident dataset from the Royal Oman Police records, it identifies factors related to the environment, roadway, driver and accident characteristics that contributed to the injury severity level. We find the the findings of applying Poisson and it's alternatives and the binary logistic are more consistent with many of the ones published in of the former studies than the result here using the probit model of the accident severity levels. This could be a result of mistake in data preparation for running the model. However, this research in general should provide important transport safety inferences regarding the accidents in developing countries. Especially, that little research is focused on this particular problem in GCC. It also provides a foundation to compare and contrast the role of different factors. We applied the ordered probit model using different subsets selected from the full dataset to investigate the accuracy of using the model and compared the results looking to different dimensions of the problem. Apparently the model fits the data reasonably well but the marginal effects can be misleading, mainly when the explanatory variable is a categorical variable. Comparing the performance of the model using the full dataset with modelling a smaller sampled datasets, we found that the model still provided results that were reasonably consistent in most cases. Though it must be as large sample datasets usually contain enough information to fit an ordered probit model. Although there is already extensive and successful applications of more advanced techniques in the literature such as Bayesian models in many fields, the simple ordered model found to be efficient and parsimonious and still attracts researches to apply for injury severity data analysis.

Table B.4: Ordered Probit Model: different samples from the model

	Dependent variable:					
	injLvl					
	(full dataset)	(sample1)	(sample2)	(sample3)	(sample4)	(sample5)
acctypRun-Over	0.264^{***}	0.152^{*}	0.188^{**}	0.219^{**}	0.375^{***}	0.307^{***}
	(0.024)	(0.084)	(0.083)	(0.087)	(0.086)	(0.083)
acctypOver-Turn	0.174^{***}	0.215^{***}	0.174^{**}	0.273^{***}	0.164^{**}	0.235^{***}
	(0.022)	(0.075)	(0.076)	(0.077)	(0.077)	(0.077)
acctypFixed Object Collision	0.271^{***}	0.200^{***}	0.190^{***}	0.336^{***}	0.301^{***}	0.288^{***}
	(0.020)	(0.069)	(0.070)	(0.072)	(0.070)	(0.070)
acctypMotorcycle/Bicycle	0.069^{*}	0.095	0.108	0.263^{**}	0.283^{**}	-0.043
	(0.036)	(0.139)	(0.121)	(0.125)	(0.133)	(0.128)
causeWrong conduct	0.192^{***}	0.231^{***}	0.281^{***}	0.288^{***}	0.157^{**}	0.230^{***}
	(0.021)	(0.071)	(0.066)	(0.075)	(0.073)	(0.072)
causeCarelessness	0.654^{***}	0.610^{***}	0.683^{***}	0.698^{***}	0.487^{***}	0.658^{***}
	(0.034)	(0.116)	(0.111)	(0.120)	(0.104)	(0.117)
causeFatigue	0.407^{***}	0.502^{***}	0.582^{***}	0.723^{***}	0.771^{***}	0.468^{**}
	(0.050)	(0.177)	(0.181)	(0.177)	(0.185)	(0.188)
causeOvertaking	-0.343^{***}	-0.416^{***}	-0.061	-0.311^{**}	-0.169	-0.236^{*}
	(0.035)	(0.120)	(0.127)	(0.127)	(0.133)	(0.130)
causeClimcond	0.313^{***}	-0.196	0.783^{**}	0.278	-0.317	0.209
	(0.094)	(0.293)	(0.355)	(0.338)	(0.318)	(0.236)
causesafedist.	0.353^{***}	0.463^{***}	0.407^{***}	0.442^{***}	0.366^{***}	0.415^{***}
	(0.035)	(0.121)	(0.116)	(0.119)	(0.130)	(0.131)
causeVehicle	-0.191^{***}	-0.253^{*}	-0.166	-0.288^{**}	0.088	-0.083
	(0.043)	(0.141)	(0.157)	(0.132)	(0.147)	(0.144)
causeRoad	-0.163^{**}	-0.179	-0.116	0.072	0.189	-0.314
	(0.075)	(0.248)	(0.302)	(0.247)	(0.356)	(0.248)
roadtypSub	0.398^{***}	0.371^{***}	0.459^{***}	0.362^{***}	0.359^{***}	0.422^{***}
	(0.016)	(0.055)	(0.056)	(0.057)	(0.056)	(0.056)
roadtypUnpaved	0.100^{*}	0.075	0.294	-0.063	0.327^{*}	0.214
	(0.055)	(0.183)	(0.210)	(0.166)	(0.196)	(0.204)
locationSide	-0.354^{***}	-0.262^{**}	-0.520^{***}	-0.404^{***}		-0.273^{**}
	(0.040)	(0.131)	(0.142)	(0.146)		(0.135)
locationIntersection	-0.050^{*}	-0.086	-0.275^{***}	0.005		-0.048
	(0.030)	(0.100)	(0.103)	(0.107)		(0.105)
locationRoundabout	0.354^{***}	0.358^{***}	0.263^{**}	0.428^{***}		0.425^{***}
	(0.035)	(0.122)	(0.118)	(0.115)		(0.126)
locationSignal	0.298^{***}	0.480^{**}	0.217	0.350^{*}		0.484^{**}
	(0.054)	(0.225)	(0.176)	(0.197)		(0.196)
locationOthers	0.125^{***}	0.084	0.010	0.083		0.181
	(0.041)	(0.147)	(0.129)	(0.135)		(0.158)
climateAbnormal	-0.266^{***}		-0.442^{**}	-0.408^{*}		
	(0.061)		(0.208)	(0.223)		
genderFemale	-0.208^{***}	-0.276^{***}	-0.250^{***}	-0.160^{*}	-0.243^{***}	-0.164^{**}
	(0.024)	(0.082)	(0.081)	(0.082)	(0.081)	(0.083)
age	0.002^{**}					0.005^{*}
	(0.001)					(0.002)
nationalNon-omani				0.113		
				(0.072)		
licnsUnlicensed	-1.138^{***}	-0.940^{***}	-1.160^{***}	-1.022^{***}	-1.170^{***}	-1.184^{***}
	(0.035)	(0.130)	(0.118)	(0.137)	(0.111)	(0.111)
heavyVclOther	-0.157^{***}	-0.272^{***}	-0.146	-0.263^{***}	-0.199^{**}	
	(0.028)	(0.102)	(0.094)	(0.091)	(0.100)	
vhcls	-0.038^{***}	-0.130^{***}		-0.079^{*}	-0.069	-0.067
	(0.012)	(0.044)		(0.047)	(0.049)	(0.045)
Observations	24,074	2,000	2,000	2,000	2,000	2,000

Note: *p<0.1; **p<0.05; ***p<0.01

Table B.5: Ordered Probit Model: Injury-severity-level of accidents by number of vehicle involved

	Dependent variable:			
	injLvl			
	(all data)	(single-vehicle)	(two-vehicle)	(multiple vehicle)
timhour	0.0001***	0.00005***	0.0001***	
	(0.00001)	(0.00002)	(0.00002)	
acctypRun-Over (person or animal)	0.256***	0.439***	0.002	
	(0.024)	(0.035)	(0.034)	
acctypOver-Turn	0.165***	0.249***	−0.014	
	(0.022)	(0.030)	(0.034)	
acctypFixed Object Collision	0.256***	0.670***	−0.102***	
	(0.020)	(0.032)	(0.029)	
acctypMotorcycle/Bicycle	0.068*	0.095**	0.014	
	(0.036)	(0.042)	(0.016)	
causeWrong conduct	0.165***	0.179***	0.148***	0.287***
	(0.019)	(0.044)	(0.026)	(0.078)
causeCarelessness	0.643***	1.034***	0.225***	0.457***
	(0.033)	(0.048)	(0.049)	(0.170)
causeFatigue	0.413***	0.510***	0.315***	0.128
	(0.005)	(0.009)	(0.013)	(0.211)
causeOvertaking	−0.374***	0.162***	−0.425***	−0.457***
	(0.033)	(0.001)	(0.040)	(0.106)
causeClimcond	0.296***	0.219***	0.548***	−0.053
	(0.002)	(0.007)	(0.004)	(0.277)
causesafedist.	0.315***	0.422***	0.129***	0.535***
	(0.032)	(0.001)	(0.040)	(0.095)
causeVehicle	−0.193***	−0.142***	−0.229***	0.075
	(0.018)	(0.046)	(0.003)	(0.247)
causeRoad	−0.166***	−0.065***	−0.450***	0.445
	(0.002)	(0.005)	(0.001)	(0.468)
climateAbnormal	−0.267***	−0.215***	−0.394***	
	(0.003)	(0.011)	(0.007)	
roadtypSub	0.398***	0.427***	0.349***	0.401***
	(0.016)	(0.022)	(0.025)	(0.067)
roadtypUnpaved	0.105***	0.068***	0.054***	0.368
	(0.003)	(0.017)	(0.001)	(0.553)
locationSide	−0.355***	−0.505***	−0.117**	−0.247
	(0.036)	(0.052)	(0.057)	(0.167)
locationIntersection	−0.053*	0.011	−0.067*	−0.091
	(0.029)	(0.052)	(0.038)	(0.101)
locationRoundabout	0.355***	0.302***	0.435***	0.480***
	(0.034)	(0.045)	(0.055)	(0.159)
locationSignal	0.290***	0.338***	0.279***	0.215
	(0.004)	(0.002)	(0.019)	(0.144)
locationOthers	0.123***	0.237***	0.027	−0.066
	(0.037)	(0.016)	(0.057)	(0.139)
age	0.002***	0.003***		
	(0.001)	(0.001)		
genderFemale	−0.208***	−0.223***	−0.201***	−0.148
	(0.024)	(0.035)	(0.035)	(0.091)
nationalNon-omani			0.064**	0.213***
			(0.032)	(0.082)
licnsUnlicensed	−1.148***	−0.862***	−1.351***	−0.953***
	(0.034)	(0.023)	(0.047)	(0.200)
heavyVclOther	−0.157***		−0.264***	
	(0.022)		(0.034)	
Observations	24,074	12,880	9,669	1,525
Note:				*p<0.1; **p<0.05; ***p<0.01

Bibliography

[1] Mujalli,R., On, J., *Injury severity models for motor vehicle accidents: a review* . ice publishing,Transport (Impact Factor: 0.32). 10/2013; 166(5):255-270. DOI: 10.1680/tran.11.00026 2011.

[2] Zhang, Y., et al, *Exploring Driver Injury Severity at Intersection: An Ordered Probit Analysis.* Hindawi Publishing Corporation, Advances in Mechanical Engineering, Article ID 567124, 2014.

[3] Weiss, H., et al, *Analysis of factors associated with injury severity in crashes involving young New Zealand drivers.* Volume 65, April 2014, Pages 142-155, 2014.

[4] Cameron, C. and Trivedi, P., *Regression Analysis of Count Data.* 2nd edition, Econometric Society Monograph No.53, Cambridge University Press, 1998 (566 pages.), 2013.

[5] Garridoa, R., Bastosa, A., Elvasc, J., 13 *Prediction of road accident severity using the ordered probit model.* Transportation Research Procedia 3(214-223), 2014.

[6] Obeng, K., Rokonuzzaman, M., *Pedestrian Injury Severity in Automobile Crashes.* Open Journal of Safety Science and Technology, 3, 9-17, 2013.

[7] Bajracharya, S., *Intersection Accident Severity Analysis using Ordered Probit Model.* Proceedings of the Eastern Asia Society for Transportation Studies, Vol.9, 2013.

[8] Weissa, H., Kaplanb, S., Pratob, C., *Analysis of factors associated with injury severity in crashes involving young New Zealand drivers.* Accident Analysis and Prevention 65, 142-155, 2014.

[9] Donnell CJ, Connor DH, *Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice.* Accident Analysis and Prevention. 1996 Nov;28(6):739-53, 1996.

[10] Al-Ghamdi, A. *Using logistic regression to estimate the influence of accident factors on accident severity.* Accident Analysis and Prevention 34 729-741, 2002.

[11] Kockelman, K., Kweon, Y., *Driver injury severity: an application of ordered probit models* Accident Analysis and Prevention. Volume 34, Issue 3, May 2002, Pages 313-321, 2002.

[12] Quddus, M., Noland, R., Chin, H., *An analysis of motorcycle injury and vehicle damage severity using ordered probit models.* Journal of Safety Research - J SAFETY RES , vol. 33, no. 4, pp. 445-462, 2002, 2002.

[13] Abdel-Aty, M., *Analysis of driver injury severity levels at multiple locations using ordered probit models* Journal of Safety Research Volume 34, Issue 5, 2003, Pages 597-603, 2003.

[14] Xuesong Wang, Mohamed Abdel-Aty, *Analysis of left-turn crash injury severity by conflicting pattern using partial proportional odds models.* Accident Analysis and Prevention 40 (2008) 1674-1682 2008.

[15] Rebecca C. Graya, Mohammed A. Quddusb, Andrew Evansa, *Injury severity analysis of accidents involving young male drivers in Great Britain* Journal of Safety Research Volume 39, Issue 5, 2008, Pages 483-495, 2008.

[16] Duboisa, S., Bedarda, M., Weaverb, B., *The association between opioid analgesics*

and unsafe driving actions preceding fatal crashes. Accident Analysis and Prevention. Volume 42, Issue 1, January 2010, Pages 30-37, 2010.

[17] Lord, D., Mannering, F., *The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives.* Volume 44, Issue 5, June 2010, Pages 291-305 2010.

[18] Bandyopadhyaya, R., Mitra, S., *Modelling Severity Level in Multi-vehicle Collision on Indian Highways* Procedia - Social and Behavioral Sciences Volume 104, 2 December 2013, Pages 1011-1019, 2013.

[19] Ma, J. and Kockelman, K., *Bayesian Multivariate Poisson Regression for Models of Injury Count, by Severity* Transportation Resaerch Board (TRB), 2006.

[20] Huang H., et al. *Bayesian Hierarchical Analysis on Crash Prediction Models*, Transportation Research Board (TRB),Capital Hilton, Washington DC., 2008.

[21] Kim et al. *Bayesian Hierarchical Poisson Regression Models: An Application to Driving Study With Kinematic Events.* Journal of the American Statistical Association., 2013.