ANALYSIS OF CRASH SEVERITY USING HIERARCHICAL BINOMIAL LOGIT MODEL

VU VIET HUNG
(B.Sc. in CIVIL Eng., HCMUT)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF ENGINEERING

DEPARTMENT OF CIVIL ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2009

ACKNOWLEDGEMENTS

I would like to express my deep and sincere thanks and gratefulness to my supervisor, Associate Professor Chin Hoong Chor for his invaluable advice, patient guidance, exceptional support and encouragement throughout the course of this research work.

I gratefully acknowledge the National University of Singapore for giving me a chance to study and do a research.

Special thanks are extended to Mdm. Theresa, Mdm. Chong Wei Leng and Mr. Foo for their kind assistance during this study period.

My heartfelt thanks and appreciation goes to my colleagues and friends namely, Ms. Tuyen, Mr. Ashim, Mr. Shimul, Ms. Sophia, Mr. Habibur, Ms. Duong, Mr. Thanh and Ms. Qui for their nice company, help, and cooperation thereby making my stay in Singapore, during my research period, a memorable experience.

Finally, the author wishes to dedicate this work to his parents and his sisters for the many years of endless love and care.

Vu Viet Hung

National University of Singapore

August 2009

SUMMARY

Crash severity is a concern in traffic safety. To propose efficient safety strategies to reduce accident severity, the relationship between injury severity and risk factors should be insightfully established. The purpose of this study is to identify the effects of factors of time, road features, and vehicle and driver characteristics on crash injury. This study on the severity of accidents at signalized intersections is investigated because the numbers of these crashes are the highest of total accidents and result in a variety of injured drivers.

To establish the relationship between injury severity and the risk factors and to solve multilevel data structures in the dataset, hierarchical binomial logit model is selected for the study. The reported accident data in Singapore from year 2003 to 2007 are used to calibrate the model. From twenty-two pre-selected variables, the significant factors in both fixed and random part are identified by using 95% Bayesian Credible Interval (BCI). In addition, Deviance Information Criterion (DIC) is also employed to find the suitable model.

The result indicates that ten variables are identified as significant factors. Crashes at night, with high speed limit or at intersection with presence of red light camera vitally increase the severity while a variable, wet road surface, reduces the injury. Vehicle movement also significantly affects the crash severity. This study also finds that Honda manufacture is safer than other vehicle makes. With driver characteristics, driver gender and age are also associated with crash severity, while involvement of offending party positively affects crash severity.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ILLUSTRATIONS

AIC	Akaike Information Criterion
BCI	Bayesian Credible Interval
BIC	Bayesian Information Criterion
BL	Binary Logit Model
DIC	Deviance Information Criterion
GLMs	Generalized Linear Regression Models
GVE	Generalized Extreme Value
HBL	Hierarchical Binomial Logit Model
IIA	Independence of Irrelevant Alternatives
MCMC	Markov Chain Monto Caelo algorithm
O.R.	Odds Ratio
S.D.	Standard Deviation

LIST OF SYMBOLS

β	A vector of coefficients; β_0 is the intercept; β_i is the coefficient for $\mathbf{x}_i$
β_{0j}	The intercept term of j^{th} crash in individual level model of HBL
β_{pj}	The p^{th} regression coefficients j^{th} crash in individual level model of HBL
γ_{00}	The intercept term for regressing β_{0j} in the crash level model of HBL
γ_{p0}	The intercept term for regressing β_{pj} in the crash level model of HBL
γ_{0q}	The q^{th} regression coefficients for regressing β_{0j} in the crash level model of HBL
γ_{pq}	The q^{th} regression coefficients for regressing β_{pj} in the crash level model of HBL
ε	Random error term in the ordered logit/probit model
$\Phi(.)$	The cumulative distribution function for the standard normal distribution
π_i	The probability of Y_i=1 in Binomial distribution
τ_M	The threshold or cut point for the ordered logit/probit model
τ_0^2	The variance of random effects U_{0j}
τ_p^2	The variance of random effects U_{pj}
$\sum_{i=1}^{n}(.)$	Summation of a given function from 1 to n observation
i	The index for observation individual

$\mathrm{Logit}(\pi_i)$	$\mathrm{Log}(\pi_i/1-\pi_i)$
N	The total number of observation
p	Probability of success in Bernoulli trial
$\mathrm{Probit}(\pi_i)$	The inverse of the cumulative standard normal distribution (π_i)
U_{0j}	Within-crash random effects of β_{0j}
U_{pj}	Within-crash random effects of β_{pj}
$\mathbf{X}_i$	A row vector of independent variables for the i^{th} observation; the i^{th} row of x
$\mathbf{X}_{pij}$	The p^{th} covariate for i^{th} driver-vehicle unit in the j^{th} crash in level 1
$\mathbf{Y}_{ij}$	Binary severity variable for the i^{th} driver-vehicle unit in the j^{th} crash
y^*	The latent dependent variable
$\mathbf{Z}_{qj}$	The q^{th} covariate of the j^{th} crash in level 2

CHAPTER 1: INTRODUCTION

1.1 RESEARCH BACKGROUND

Road systems both satisfy transportation demand and provide transportation supply efficiently. Road safety is one of the most important concerns of transportation supply. Therefore, reducing crash frequency and severity not only ameliorates safety but also saves a lot of money as well as improves transportation. To propose efficient safety strategies, several studies have been trying to fully identify how accident severity varies. In Singapore, although crash severity decreases, based on some studies' findings such as Quddus et al. (2002) and Rifaat and Chin (2005), accident rate and severity are still high in recent years. For instance, accident data show that the numbers of drivers are 2661, 2923, 2255, 2516, and 2933 from year 2003 to 2007, respectively. Thus, clearly understanding the relationship between the injury severity and risk factors is necessary for developing safety countermeasures.

Statistical models have been developed for road safety and applied for predictions of accident severity in specific situations. Firstly, several researchers have improved crash severity prediction models in order to take into account the severity levels. For example, some studies have applied some generalized linear models (GLMs) to classify nominal categories. Binary probit or logit models have been employed when the severity levels are classified as two levels: injury and non-injury. In addition, multinomial probit and logit have been used in order to explore the important factors affecting severity, categorized as multinomial states. On the other hand, one of the most common models used for categorizing the severity levels is ordered probit or logit model. The advantage of this model is to take into account the ordered nature of severity levels from the lowest severity to the highest severity such as no injury, possible injury, evident injury, disabling injury, and fatal. Secondly, other studies have examined and focused on specific effects, such as driver age and gender, vehicle type, mass, and size, collision type and others, on degree of severity. For instance, Islam and Mannering (2006); Lonczak et al. (2007); Ulfarsson and Mannering (2004) separated driver gender and driver age to evaluate how difference between male and female affects severity and examine how different age groups influence fault and crash injury. In addition, Gray et al. (2008) and Yannis et al. (2005) concentrated on young (or old) drivers to find countermeasures that reduce the severity of specific groups. On the other hand, vehicle type, mass, and size have been studied by several researchers (Chang and Mannering 1999; Evans and Frick 1992; Evans and Frick 1993; Fredette et al. 2008; Islam and Mannering 2006; Khorashadi et al. 2005; Kim et al. 2007b; Langley et al. 2000; Savolainen and Mannering 2007; Ulfarsson and Mannering 2004) because they are directly associated with the increase of severity. Moreover, a series of studies (Kim et al. 2007a; Kockelman and Kweon 2002; Pai ; Pai and Saleh 2008a; Pai and Saleh 2008b; Preusser et al. 1995; Wang and Abdel-Aty 2008) have centered on evaluating the relationship between severity and crash types. Last, but not least, previous studies (Abdel-Aty 2003; Abdel-Aty and Keller 2005; Huang et al. 2008; Kim et al. 2007a; Milton et al. 2008; Obeng 2007; Pai and Saleh 2008a) have also investigated severity of accident at specific locations. All of the studies mentioned above provided us with the knowledge to both understand various severities and suggest efficient countermeasures so that accident severity is decreased.

Selection of suitable statistical models is dependent on some assumptions made in these models. It also depends on how accident data confirm these assumptions. For example, generalized linear regression models (GLMs) that are used for predicting severity assume that all samples in the dataset are independent of one another. However, when this assumption is violated, the estimation of parameters and standard errors is incorrect. As a result, conclusions that the factors are significant are not correct. In fact, Jones and Jørgensen (2003) clearly explored the existence of dependence between samples such as samples of vehicle. Casualties within the same vehicle would have the same probability of survival. However, in reality, some casualties are killed and others are survived even though all of them travel in the same vehicle. Therefore, the assumption of independence may not hold true. The model without overcoming this problem, especially when there is clearly an existence of dependence between samples, would lead to inaccurate estimates of parameters and standard errors. Although some previous researches (Huang et al. 2008; Jones and Jørgensen 2003; Kim et al. 2007a) developed approaches to solve this problem which is also called multilevel data, these models are not fully developed; thus, resulting in the fact that some conclusions are incorrect. Therefore, this study continues to improve the hierarchical models with the purpose of better and more clearly taking into account the impacts of risk factors on crash severity at signalized intersection in Singapore.

1.2 OBJECTIVE AND SCOPE OF THIS STUDY

The main purpose of this study is to examine how accident severity is affected by risk factors. The severity of road accidents at signalized intersections is chosen in this analysis. This is because the numbers of collisions at signalized intersections are the highest (20% of total accidents) and the numbers of drivers and vehicles increase from 2003 to 2007, based on accident data provided by Traffic Police in Singapore.

In order to obtain this objective, the hierarchical logit model with random slope effects has been developed for analyzing occupant severity. Moreover, accident data are used to explore the relationship between the crash severity and several factors such as general factors, road features, and vehicle and casualty characteristics. The model calibration and validation are then estimated to prove the appropriateness of hierarchical logit model compared with another model.

1.3 OUTLINE OF THE THESIS

The organization of this thesis contains five chapters and is presented as follows.

Chapter 1 provides the research background in which the limitations of statistical models are identified. The objective and scope of this study are also mentioned in this chapter. The outline demonstrates the organization of this thesis.

Chapter 2 presents the literature reviews of the severity models in recent year. The problem of statistical models is also identified.

Chapter 3 describes the formulation and assessment of the hierarchical logit model.

Chapter 4 demonstrates the application of hierarchical logit model for crash severity at intersections. The parameter estimation, model calibration and validation, and explanation of significant covariates are also given in this chapter.

Finally, conclusions of analyzing severity are discussed in Chapter 5. Besides, research contributions and recommendations are presented.

CHAPTER 2: REVIEW OF CRASH SEVERITY MODELS

2.1 INTRODUCTION

Reducing accident severity is a target of traffic safety. Before proposing countermeasures to improve road safety, experts and engineers have to establish the relationships between risk factors and the crash severity or crash frequency. Therefore, a number of researchers have been interested in developing and improving statistical approaches in order to clearly and correctly explore how the response variables are dependent on the explanatory variables, such as road features, traffic factors, and vehicle and driver characteristics. In addition to using count models such as Poison and Negative binomial models to predict accident frequency, generalized linear regression models (GLMs) have been broadly employed for investigating crash severity. Since the injury severity variable is discrete, sporadic and nominal, at least three types of GLMs: binary logit/probit models, multinomial logit/probit model, and ordered logit/probit models are suitable for taking into account the severity level. Previous studies (such as Factor et al. 2008; Obeng 2007; Pai 2009 and Simoncic 2001) successfully used binary logit/probit models to overcome the severity levels, which are categorized as less and high injury, and find several risk factors that significantly influence the severity. On the other hand, when data contain the severity variables classified as more than two states and nominal categories, multinomial logit/probit models are employed so that estimates of parameters, standard errors, and significances are more accurate. Some researchers such as De Lapparent (2006); Kim et al. (2007b); Savolainen and Mannering (2007); Shankar and Mannering (1996); Simoncic (2001); Ulfarsson and Mannering (2004) did some of these studies.

Moreover, a lot of accident data commonly contain crash severity that is ranked from the lowest severity to the highest severity. Consequently, several studies (Abdel-Aty 2003; Kockelman and Kweon 2002; Lee and Abdel-Aty 2005; O'Donnell and Connor 1996; Pai and Saleh 2008a; Pai and Saleh 2008b; Quddus et al. 2002; Rifaat and Chin 2005; Zajac and Ivan 2003) employed ordered logit and probit models to explain and overcome the ordinary outcomes of the severity.

This chapter presents a literature review of GLMs. In addition, mathematical formulations, general forms, assumptions, and limitations of GLMs such as binary, multinomial, and ordered logit/probit models are provided in this chapter. Based on the information, a potential problem is also identified.

2.2 REVIEW OF STATISTICAL MODELS

2.2.1 BINARY LOGIT AND PROBIT MODEL

In the studies of accident severity, logit and probit models are appropriate to investigate the fact that crash severity is a binomial or multinomial outcome. Binary logit and probit models are employed when the response variable has two states such as injury or non-injury, hit-and-run or not-hit-and-run crash, or at-fault or not-at fault case. In these models which are applied for predicting the injury, the crash severity is a binomial distribution. So, the response variable Y_i for the i^{th} observation can take one of two values: Y_i= 0 or 1, where Y_i=1 presents the first state such as injury and Y_i= presents the other state: non-injury. The probability of Y_i is denoted by $\pi_i = \Pr(Y_i = 1)$. The logit transformation of the probability π_i of a crash being injured is given by

$$\mathrm{Logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) \tag{2.1}$$

Besides, the logit transformation is linked to the linear predictor, presented as follows

$$\mathrm{Logit}(\pi_i) = \beta X_i \tag{2.2}$$

Thus, the logit models are obtained and given by

$$\mathrm{Log}\left(\frac{\pi_i}{1-\pi_i}\right) = \beta X_i \tag{2.3}$$

Based on Equation (2.3), the probability π_i of a crash being injured is solved by

$$\pi_i = \Pr(Y_i = 1) = \frac{\exp(\beta X_i)}{1-\exp(\beta X_i)} \tag{2.4}$$

where, X_i is a vector of explanatory such as road features, traffic factors, and vehicle and driver characteristics which may have influences on crash severity. Besides, β is the coefficient regression vector of the independent variables, presenting how each independent variable affects the increase or decrease of injury.

Binary probit models are similar to binary logit models. The difference between them is the error distribution. In the binary logit models, the errors are assumed to have a standard logistic distribution with mean 0 and variance $\frac{\pi}{3}$, while the errors in binary probit models have an assumption that the error distribution has mean 0 and variance 1. Therefore, the establishment of the probit models is the same as that of the logit model and described as follows.

The probit transformation of the probability π_i is given by inverse of standard cumulative normal distribution function and written as

$$\mathrm{Probit}(\pi_i) = \Phi^{-1}(\pi_i) \tag{2.5}$$

where $\Phi(.)$ is the cumulative distribution function of standard normal distribution. In addition, the probit transformation is linked to the linear predictor, described as

$$\mathrm{Probit}(\pi_i) = \beta X_i \tag{2.6}$$

Consequently, the probit models are obtained and given by

$$\Phi^{-1}(\pi_i) = \beta X_i \tag{2.7}$$

Based on Equation (2.7), the probability π_i of a crash being injured is solved by

$$\pi_i = \Pr(Y_i = 1) = \Phi(\beta X_i) \tag{2.8}$$

where the explanations of β, X_i and $\Phi(.)$ are mentioned above.

Both binary logit and probit model have been broadly used in traffic safety. For instance, Simoncic (2001), who applied binary logit model to overcome injury severity of collisions between a pedestrian, bicycle or motorcycle and a car, found that some variables, including no use of protective devices, older age, intoxication of pedestrians, cyclists, motorcyclists or car divers, and accidents at night, on motorway or at weekend significantly influence the increase of participants' injury. Moreover, Haque et al. (2009) identified time factors, road features (such as wet surface, lane position, and speed limit) and driver-vehicle characteristics (such as driver age and license, and vehicle capacity and registration) that contribute to the fault of motorcyclist in crashes at specific locations by applying binary logit model. Furthermore, Tay et al. (2008) employed a logit model to analyze hit-and-run accidents on which the roadway, environmental, vehicle, crash, and driver characteristics have influences.

Although binary logit and probit models have little difference on the error distribution, binary logit models are always chosen in previous studies. This is because the probability density function (pdf) and cumulative distribution function (cdf) of logit models are simpler than those of probit models. Especially, it is easy for the logit model to interpret log-odds ratio which probit models cannot estimate. Due to the advantages of logit models, the following sections focus on demonstrating multinomial logit and ordered logit models.

2.2.2 MULTINOMIAL LOGIT MODEL

Multinomial logit models can be thought of as an extension of the binary logit models. For the multinomial response variable, multinomial logit models are most frequently chosen in order to analyze the crash severity because accident datasets contain multiple severity levels and binary logit models are unable to solve more than two levels of severity. Another reason is that multinomial logit models' mathematical structure and estimation are simple and easy respectively. MacFadden (1973) demonstrated the multinomial logit models as the most widely-used discrete choice model. This discrete choice model is based in the principle that an individual chooses the outcome that maximizes the utility gained from that choice. Based on this principle and assumption that the error term is generalized extreme value (GVE) distributed, MacFadden (1981) derived the simple multinomial logit model. The final formulation of the models is written as

$$\pi_i(y_i = j) = \frac{\exp(\beta_j X_i)}{\sum_J \exp(\beta_j X_i)} \tag{2.9}$$

where $\pi_i(y_i = j)$ is the probability of individual i having alternative j in a set of possible choice categories J. X_i is a vector of measurable characteristics that determine alternative j. β_j is a vector of statistically estimable coefficients.

However, the multinomial logit model has the limitation of independence of irrelevant alternatives (IIA) (Ben-Akiva and Lerman 1985), such that the odd of m versus n $(m, n \in 1..J)$ is not affected by other alternatives, i.e.

$$\frac{\pi_i(y_i = m)}{\pi_i(y_i = n)} = \exp(X_i[\beta_m - \beta_n]) \tag{2.10}$$

This expression is only a function of the respective utilities of alternatives m and n, and is not affected by the introduction/removal of other alternatives. This analytical feature implies that the relative shares of the two given alternatives are independent of the composition of the alternative set.

The limitation of independence of irrelevant alternatives in multinomial logit model was also identified by Chang and Mannering (1999); Lee and Mannering (2002); Shankar et al. (1996) in their studies on accident severity. Shankar et al. (1996) classified severity of an accident to be one of five discrete categories: property damage, possible injury, evident injury, disabling injury and fatality. However, according to them, property damage and possible injury accidents may share unobserved effects such as internal injury or effects associated with lower-severity accidents. However, the basic assumption in the derivation of the multinomial logit model is that error terms or disturbances are independent from one accident severity category to another. Shankar et al. (1996) suggested that if some severity categories share unobserved effects (i.e. have correlated disturbances), the model derivation assumptions are violated and serious specification errors will result.

On the other hand, according to Long (1997), a significant advantage of multinomial probit models is that the errors can be correlated across choices, which eliminates the IIA restriction. However, computational difficulties make multinomial probit models impractical.

2.2.3 ORDERED LOGIT MODEL

According to Long (1997), when the response variable is ordinal in nature and models for nominal variables are used, there will be loss of efficiency due to information being ignored. Therefore, multinomial logit model cannot handle ordinal dependent variables. One way to deal with this problem is to use ordered logit models instead of multinomial logit ones. Ordered logit models are usually motivated in a latent (i.e., unobserved) variables framework. The general form of the model is given by

$$y_i^* = x_i\beta + \varepsilon_i \tag{2.11}$$

where, y_i^* is a latent, unobservable and continuous dependent variable; x_i is a row vector of observed non-random explanatory variables; β is a vector of unknown parameter; ε_i is the random error term which is assumed to be logistically distributed.

According to Long (1997), ordered logit models can be derived from a measurement model in which a latent variable y_i^* ranging from $-\infty$ to $+\infty$ is mapped to an observed ordinal variable y. The discrete response variable y is thought of as providing incomplete information about an underlying y_i^* according to the measurement equation:

$$y_i = \begin{cases} 1 \quad \text{if } \tau_0 \le y_i^* < \tau_1 \quad \text{(the lowest injury)} \\ \ldots \\ m \quad \text{if } \tau_{m-1} \le y_i^* < \tau_m \\ \ldots \\ M \quad \text{if } \tau_{M-1} \le y_i^* < \tau_M \ \text{(the highest injury)} \end{cases} \tag{2.12}$$

where, the threshold values $\tau's$ are unknown parameters to be estimated. The extreme categories, 1 and M, are defined by open-ended intervals with $\tau_0 = -\infty$ and $\tau_M = +\infty$. The mapping from the latent variable to the observed categories is illustrated in Figure 2.1 below:

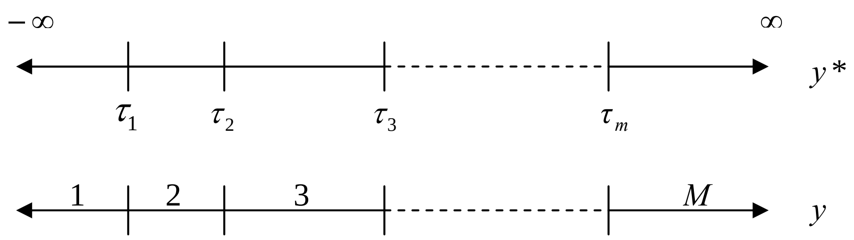

Figure 2.1 Mapping of latent variable to observed variable

Since the distribution of ε_i is specified as standard logit distribution with mean 0 and variance $\frac{\pi}{3}$, the probabilities of observing a value of y given x_i can be computed. The final formulation of the probabilities of observing value of y=m given x_i is described as follows

$$\Pr(y_i = m|x_i) = F(\tau_m - x_i\beta) + F(\tau_{m-1} - x_i\beta) \qquad (2.13)$$

where, $F(.)$ is the cumulative distribution function of standard logistical distribution; x_i, β, and τ_m are mentioned above.

Since accident data usually contain severity levels that are ordered from the lowest to the highest severity such as slight injury, serious injury, and fatality, the ordered logit and probit models are most commonly applied. These models are also proved to be appropriate for analyzing road accidents by several previous studies. For example, O'Donnell and Connor (1996) used two models of multiple choice; the ordered logit and probit models, to examine how variations of road-user attributes result in variations in the probability of motor vehicle accident severity. In this study, several factors that significantly affected injury include driver's characteristics such as the age, seating position, and blood alcohol level, vehicle features such as vehicle type and make, and others such as type of collision. This study also indicated that the results from the ordered probit and ordered logit models are similar. Moreover, Quddus et al. (2002) indentified that time factor such as driving at weekends and time of day, road factors including location, traffic type, surveillance camera, road surface, and lane of nature, driver's factors consisting of nationality, at-fault drivers, gender, and age group, vehicle's features such as engine capacity and headlight not turned on during daytime, and the collision types contribute to both various motorcycle injury and vehicle damage severity by using the ordered probit models. Furthermore, Kockelman and Kweon (2002) employed the ordered probit models for all crash types, two-vehicle crashes, and single-vehicle crashes to estimate the probability of crash severity. The results analyzed from an application for all crash types showed the significances of gender, violator and alcohol, vehicle type as well as crash type on the severity level. On the other hand, some variables, including the same factor in all crash type case and other factors such as age, are found to importantly affect injury severity in two-vehicle crashes and single-vehicle crashes. Besides, driver severity levels at multiple locations, such as roadway sections, signalized intersections, and toll plazas, are solved by Abdel-Aty (2003), using the ordered probit models. The findings indicated that driver's age, gender, seat belt use, and vehicle speed and type are significant on all of the locations. This study also found other variables that have effects on injury in specific cases. For example, while a driver's violation influences injury severity at signalized intersections, alcohol, lighting conditions, and horizontal curves contribute to the likelihood of injury at roadway sections, and vehicle equipped with Electronic Toll Collection has an effect on the probability of injury. In addition to studies mentioned above, the ordered logit and probit models have been applied by several other researchers (Abdel-Aty and Keller 2005; Gray et al. 2008; Lee and Abdel-Aty 2005; Pai and Saleh 2008b; Rifaat and Chin 2005; Zajac and Ivan 2003) to deal with the injury severity of overall and specific crashes at signalized intersections, young male drivers, vehicle-pedestrian crashes at intersections, various motorcycle crash types at T junctions, single-vehicle crashes, and motor vehicle-pedestrian collisions, respectively. Based on several above-mentioned applications of the ordered approaches, it is worth mentioning that these approaches contributed good explanations about ordinal discrete measure of severity levels to appropriately modeling and solve the crash severity.

However, ordered logit and probit models still have some limitations. Eluru et al. (2008) gave a good example to explain a problem of the ordered model. In this paper, the crash severity was categorized as the ordinal response variable including no injury, possible injury, non- incapacitating injury, incapacitating injury, and fatal injury. The ordered models were applied to compute the threshold values which were fixed across five crash groups. However, this did not correctly describe the fact that the effects of some independent variables may have no difference between two crash groups. This can lead to inconsistent estimates of the effects of variables. Besides, other studies such as Jones and Jørgensen (2003) found that accident data are multilevel. This means that dependence between samples such as samples of vehicles exists, which these ordered approaches cannot model and handle in order to solve the effects of risk factors on the crash severity.

2.3 IDENTIFIED PROBLEM

Although a number of studies on traffic safety have proved that the GLMs including the binary logit/probit models, multinomial logit/probit models and ordered logit/probit approaches are useful for modeling crash severity, they are incapable of investigating dependences between different observations. In fact, accident data contain some independent variables that are ranked in levels of a hierarchy. For instance, among group factors affecting accident severity, vehicles' and driver's characteristics such as vehicle registration, vehicle movement, age and gender may be the lowest level of the hierarchy of crash injury. In addition, the features of crashes have higher levels because the same crash may have different effects on the severity of drivers. A hierarchy of crash severity is presented in Figure 2.2. The fact that the predictors are classified from the lowest to the highest levels of a hierarchy leads to an assumption of independence of different samples to be invalid. Consequently, the GLMs are likely to produce poorly estimated parameters and standard errors (Skinner et al. 1989). Specially, the problem with the estimation of standard errors is very serious when intra-class correlation, by which the degree of resemblance between individual casualties belonging to the same crashes can be expressed, is very large; thus, resulting in the fact that the null hypothesis of parameters' significances may be incorrectly concluded.

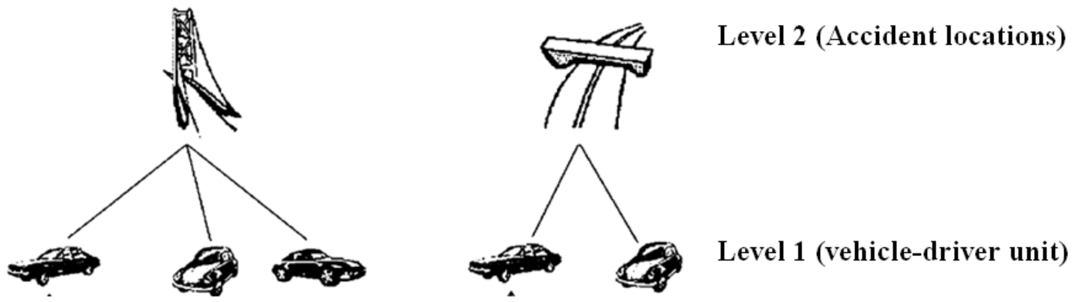

Figure 2.2: A hierarchy of severity at level 1, within accident locations at level 2

Moreover, although hierarchical severity models have been developed in traffic safety by some researchers (Huang et al. 2008; Jones and Jørgensen 2003; Kim et al. 2007a) in order to solve multilevel data, these studies have not employed a full model. An assumption in these studies is that only the random intercept effect exists. However, according to Snijders and Bosker (1999), omitting some variables which are random slope effects may have influences on the estimated standard errors of the other variables. Hence, statistical models are needed to be improved so that the estimates of standard errors are more accurate; meaning that prediction of the accident severity is better.

2.4 SUMMARY

This chapter provides a critical review of the GLMs including binary logit/probit models, multinomial logit/probit approaches, and ordered logit/probit models. In each statistical model, the probabilistic formulations of accident severity are established to find the impacts of a variety of possible independent variables, such as time factors, road features, environmental factors, and vehicle-driver characteristics as well, on crash severity. Furthermore, applications and limitations of each statistical model are identified on the purpose of assisting researchers to predict the severity more accurately.

In addition, potential problems are realized in this chapter. One of the most fundamental problems is that multilevel structure of accident data contains dependence between different observations, which the GLMs have troubles handling and solving. Another problem is that hierarchical binomial logit models to deal with the previous problem have not been fully developed. Hence, all of them can result in incorrect estimates of standard errors.

In the rest of this thesis, full formulations of the hierarchical binomial logit models are developed to overcome multilevel data structures and predict accident severity, by using Singapore accident data at signalized intersections

CHAPTER 3: DEVELOPMENT OF HIERARCHICAL BINOMIAL LOGIT MODEL WITH RANDOM SLOPE EFFECTS FOR CRASH SEVERITY

3.1 INTRODUCTION

Accident severity is a concern in traffic safety because both much money and time are spent in taking care of victims and the society loses human resource. Therefore, reducing crash severity is a necessary focus. To develop and propose safety countermeasures in an effective manner, we need to insightfully understand the relationship between crash severity and risk factors. Data analysis techniques are powerful tools for establishing this relationship. Consequently, several statistical models have been developed for about two decades in order to examine the impacts of risk factors on the accident severity.

Generalized linear regression models (GLMs) including logit/probit models and ordered discrete choice models are widely used for predicting the crash severity in order to solve problems where some dependent variables such as severity in accident data are discrete response variables. Some studies have employed binary logit models for solving specific accidents. For instance, while Factor et al. (2008); Pai ; Simoncic (2001) applied these models for predicting motorcycle injury severity, Obeng (2007) used these models to solve crash injury at signalized intersection. The binary logit models are also used in other fields of accidents such as effects of risk factors on red-light-running crashes (Porter and England 2000), influences of roadway, environmental, vehicle, crash, and driver characteristics on hit-and-run crashes (Tay et al. 2008), and impacts of time factors, road features, and vehicle-driver characteristics on the fault of motorcyclists in crashes at specific locations. Moreover, other researchers have used multinomial logit models to take into account injury severity classified as a multinomial category. While De Lapparent (2006); Savolainen and Mannering (2007); Shankar and Mannering (1996) focused on studying motorcyclist injury via the multinomial logit models, Lee and Mannering (2002) tried to establish the connection between road feature and severity of run-of-roadway crashes and Kim et al. (2007b) examined how risk factors affect the bicyclist injury in bicycle-motor vehicle crashes. Furthermore, ordered logit/probit models are widely applied for investigating crash severity that is ranked from the lowest to the highest injury. For example, O'Donnell and Connor (1996); Pai and Saleh (2008a); Pai and Saleh (2008b); Quddus et al. (2002) analyzed motorcycle accident severity by using ordered probit models. On the other hand, Kockelman and Kweon (2002) applied ordered probit models for the risk of different injury severity with all crash types, two-vehicle crashes, and single-vehicle crashes, while Gray et al. (2008) centered their study on predicting injury severity of young male drivers.

However, the models previously mentioned only yield accurate estimations of parameters and standard errors when assumptions, that all predictors are independent and that different observations are independent, are satisfied. Some studies such as Jones and Jørgensen (2003); Kim et al. (2007a) found that the correlation between individuals involved in the same cluster such as occupants in the same vehicle or driver-vehicle in the same crash is available. Specially, when this correlation is strongly significant, the generalized linear regression models (GLMs) are insufficiently powerful to correctly deal with this problem which is also called multilevel data structure.

According to Goldstein (2003); Snijders and Bosker (1999), one of statistical techniques which can solve multilevel data is hierarchical models. The most important is, when hierarchical models are applied, that hierarchy is available and identified in the dataset. In traffic safety studies on accident severity, Jones and Jørgensen (2003) insightfully explained that probabilities of severity of occupants in the same vehicle are different, which the techniques used in most past studies cannot model. Thus, this study introduced a developed form of regression models, multilevel logit models, to analyze individual severity. In addition, after multilevel accident data are identified, a number of researchers have focus on applying hierarchical logit models for predicting drivers' injury and vehicles' damage. For instance, Kim et al. (2007a) use hierarchical binomial logit models to predict crash severity of different crash types at rural intersections, while Huang et al. (2008) found the impacts of risk factors on severity of drivers' injury and vehicles' damage in crashes at signalized intersections by using a Bayesian hierarchical analysis.

Although they are successful when employing hierarchical binomial logit models for the investigation of individual severity, several studies used these models with a simple assumption that only random intercept effects exist instead of using both random intercept and random slope effects. According to Snijders and Bosker (1999), refraining from using random slopes may yield invalid statistical tests. This is because if some variables have a random slope, then omitting this feature from models could affect the estimated standard errors of the other variables. Therefore, this study develops the full hierarchical binomial logit models to predict crash severity at signalized intersections in Singapore.

In the rest of this chapter, the formulation of hierarchical binomial logit (HBL) models is established. In addition, model evaluation, deviance information criterion (DIC), is presented. Pre-selection of predictors is then summarized .The hierarchical binomial logit (HBL) models with these covariates are applied in next chapter to identify the significant factors that increase or decrease accident severity at signalized intersections.

3.2 MODEL SPECIFICATION

3.2.1 HIERARCHICAL BINOMIAL LOGIT MODEL

Some previous studies have found the existence of within-crash correlation of drivers' severity. Models without solving this correlation might yield incorrect parameter and inaccurate standard error estimations. Thus, conclusions of significant variables may not be precise. To investigate accident data which are multilevel, some studies (Huang et al. 2008; Jones and Jørgensen 2003; Kim et al. 2008) used hierarchical binomial logistics models to explain severity correlations between driver-vehicle units involved in the same crash. However, random slope effects still are ignored. This may yield incorrect or biased estimates of parameters in both the fixed part and the random part. To deal with this problem, a full model is developed, thus resulting in the fact the cross-level interactions between covariates are specified and estimated. In the individual-level model (level 1), the response Y_{ij} for the i^{th} driver-vehicle unit in the j^{th} crash takes one of two values: $Y_{ij}=1$ in case of high severity, otherwise, $Y_{ij}=0$. The probability of Y_{ij} is denoted by $\pi_{ij} = \Pr(Y_{ij} = 1)$. The logistics model is presented as follows.

$$\operatorname{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_{oj} + \sum_{p=1}^{P}\beta_{pj}X_{pij} \tag{3.1}$$

where: X_{pij} is the pth covariate at the individual-level for the ith driver-vehicle unit in the jth crash such as vehicle registration, type of driving license, nationality, age and gender. Besides, β_{0j} and β_{pj} are the intercept and the regression coefficients, respectively. Both of them in Eq. (3.1) vary with the different crash (level 2) and are presented as the follows.

$$\beta_{0j} = \gamma_{00} + \sum_{q=1}^{Q}\gamma_{0q}Z_{qj} + U_{0j} \tag{3.2}$$

$$\beta_{pj} = \gamma_{p0} + \sum_{q=1}^{Q}\gamma_{pq}Z_{qj} + U_{pj} \tag{3.3}$$

where: γ is the parameter. Z_{qj} is the qth covariate at the crash-level, depending only on the crash j, rather than on the driver-vehicle unit i. According to this definition, the Z_{qj} covariates in road traffic consist of time factors, road features, and environmental factors. Random effects (U_{0j} and U_{pj}) are also included to permit the potential random variations across the crash. The random slopes are addressed in this study. Therefore, the combined model is yielded by substituting Eqs. (3.2) and (3.3) with Eq. (3.1) and is presented as follows:

$$\operatorname{logit}(\pi_{ij}) = \gamma_{00} + \sum_{q=1}^{Q}\gamma_{0q}Z_{qj} + \sum_{p=1}^{P}\gamma_{p0}X_{pij} + \sum_{p=1}^{P}\sum_{q=1}^{Q}\gamma_{pq}Z_{qj}X_{pij} + U_{0j} + \sum_{q=1}^{Q}U_{pj}X_{pij} \tag{3.4}$$

It is assumed that U_{pj} is independent of the level-one residuals R_{ij} and that R_{ij} has a normal distribution with zero mean and variance of $\frac{\pi^2}{3}$.It is also assumed that the random effects (U_{pj}) have a multivariate normal distribution with zero mean and a constant covariance matrix, as suggested by Snijders and Bosker (1999). This matrix is presented as follows.

Var (U_{hj}) = τ_h^2 (h=0,…,p)

Cov (U_{hj},U_{kj}) = τ_{hk}^2 (h,k=0,…,p)

In the fixed part of coefficient estimation, the exponential of effect coefficients, $\exp(\gamma)$, is computed to gain Odds Ratio (O.R.) estimates in the hierarchical binomial logit model. The purpose of Odds Ratio (O.R.) is to interpret that a unit increase variable X_{pij} or Z_{qj} will reduce/increase the odds of severity by multiplicative effect of $\exp(\gamma)$. For the category in the model, where dummy variables are used, $\exp(\gamma_a - \gamma_b)$ presents the odds ratios between these two categorical variables. In this case, the parameter makes sense when one category is compared with another.

3.2.2 ESTIMATION

There are several methods available for estimating regression coefficients and random effects. One of convenient methods is known as empirical Bayes estimation which produces so-called posterior means. Several previous studies such as (De Lapparent 2006; Washington et al. 2005) have used empirical Bayes estimation in transportation applications. Besides, Winbugs and application of this software (Spiegelhalter et al. 2003b) are available and easy to model empirical Bayes estimation. Thus, this study employs empirical Bayes estimation and Winbugs software to estimate regression coefficients and random intercept and slope effects. To obtain posterior means, strong prior information is needed to input to the model. According to Winbugs guide, to easily reach convergence, prior distributions of all regression coefficients should be normal distributions (0, 1000) and prior distributions of all variances in random part should be gamma distribution (0.001, 0.001) in this study. In Winbugs software, each of three chains of iterations for estimating posterior means produces a trace plot. Convergence has been achieved if all the chains appear to be overlapping one another. After convergence has been achieved, the Markov Chain Monto Caeclo (MCMC) simulation should be run for a further number of iterations to obtain samples that can be used for posterior inference. The more samples the simulation has, the more accurate will be the posterior estimates. One way to assess the accuracy of the posterior estimates is by calculating the Monte Carlo error for each parameter. As a rule of thumb, the simulation should be run until the Monte Carlo error for each parameter of interest is less than about 5% of the sample standard deviation.

3.3 MODEL EVALUATION

3.3.1 BAYESIAN CREDIBLE INTERVAL (BCI) AND DEVIANCE INFORMATION CRITERION (DIC)

The important step of model evaluation is to examine which the variables in the model are significant and evaluate which models are better. While Bayesian Credible Interval

(BCI) is used to find the significance of the variables, Deviance information criterion (DIC) is employed to compare two models.

3.3.1.1 Bayesian Credible Interval (BCI)

In this study, Empirical Bayes estimation is employed to compute the posterior mean, standard deviation, and BCI. According to Bolstad (2007), 95% BCI is computed for each covariate to examine whether each coefficient is significant or not. The parameter, which has 95% BIC containing 0, is insignificant. Then, the model is run again, where the insignificant variables are dropped, to find the final group containing all of the significant variables. In addition, the significance of variables in the random part is evaluated using the same method.

3.3.1.2 Deviance information criterion (DIC)

To ensure that the hierarchical binomial logit model is more accurate than the binary logit model, the later is also estimated, where the covariates in both the two models are the same and there is no random effect in binary logit model. So, the formulation of binary logit model is given by

$$\log it(\pi_{ij}) = \gamma_{00} + \sum_{q=1}^{Q} \gamma_{0q} Z_{qj} + \sum_{p=1}^{P} \gamma_{p0} X_{pij} + \sum_{p=1}^{P} \sum_{q=1}^{Q} \gamma_{pq} Z_{qj} X_{pij} \tag{3.5}$$

where: X_{pij} is the p^{th} covariate at the individual-level for the i^{th} driver-vehicle unit in the j^{th} crash, γ is the parameter and Z_{qj} is the q^{th} covariate at the crash-level.

For model comparison, Deviance Information Criterion (DIC), proposed by Spiegelhalter et al. (2003a), is calculated in both two models. Basically, DIC is intended as the traditional model comparison criteria such as Akaike's Information Criterion (AIC). Therefore, to easily understand DIC, a review of previous model comparison criteria is necessary.

First of all model comparison uses a measure of fit, called the deviance statistic (G^2), and complexity, called degree of freedom, to examine which models are better. The formulation of the deviance statistic (G^2) is given by

$$G^2 = -(\log L_c - \log L_f) \tag{3.6}$$

where L_c denotes the likelihood of current model and L_f denotes the likelihood of estimated from the full (or saturated) model.

Since increasing complexity is accompanied by a better fit, models are compared by trading off these two quantities. In addition, following early work of Akaike (1973), proposals are often based on minimizing a measure of expected loss (Akaike's Information Criterion, AIC) on a future replicate data set as follows:

$$\mathrm{AIC}(b) = -(\log L_c) + 2b \tag{3.7}$$

where b is a number of variables in the model. After AICs of all models are calculated, according to Joshua and Garber (1990), the minimum AIC indicates the selected model.

The second model comparison is Bayesian information criterion (BIC) statistic. Exactly, when samples are much large, Raftery (1986); Raftery (1995) found the use of the G^2 statistic as a good-of-fit measure may not be enough powerful to choose the better model when two models are compared. Therefore, a new criterion, Bayesian information criterion (BIC) statistic, is proposed to solve this problem. The BIC index provides an approximation to $-2\log(\text{transformed Bayes factor})$, which may be considered as the ratio in likelihood between one model (M_0) and another model (M_1). The basic idea is to compare the relative plausibility of two models instead of finding the absolute deviation of observed data from a specific model. However, the statistical methods for computing the Bayes factor are complicated. Many studies have found the BIC statistic, proposed by Raftery (1986); Raftery (1995), is useful. The formulation of the BIC statistic is given by

$$\mathrm{BIC} = G^2 - \mathrm{DF} \times \log(n) \tag{3.8}$$

where the G^2 statistic is mentioned above, DF denotes a number of degree of freedom, and n denotes a number of observations.

Both AIC and BIC expects the specification of the number of parameter in each model. However, Gelfand and Dey (1994) suggested that observations in complex hierarchical models may be outnumbered and that model comparison using AIC or BIC cannot be directly used. Therefore, Deviance information criterion (DIC) is proposed to improve comparison between two models that contain multilevel data structures.

Final model comparison reviewed in this chapter is Deviance information criterion (DIC). Spiegelhalter et al. (2003a) proposed Bayesian measures of complexity and fit that can combine traditional model comparison. The purpose of Bayesian measures is to identify models that have the best explanation of observed data with the expectation that they are to minimize uncertainty about observations generated in the same way. The formulation is given by:

$$\mathrm{DIC} = D(\bar{\theta}) + 2p_D = \overline{D(\theta)} + p_D \qquad (3.9)$$

where $D(\theta)$ is termed as 'Bayesian deviance', in general given by

$$D(\theta) = -2\log\{p(y|\theta)\} + 2\log\{f(y)\} \qquad (3.10)$$

and, more specifically, for members of exponential family with $E(Y) = \mu(\theta)$ we shall use the saturated deviance $D(\theta)$ which is obtained by setting $f(y) = p\{(y|\mu(\theta) = y\}$

p_D is motivated as a complexity measure for effective number of parameters in a model, as the difference between the posterior mean of deviance and the deviance at the posterior estimates of the parameters of interest. It is given as

$$p_D = \overline{D(\theta)} - D(\bar{\theta}) \qquad (3.11)$$

This is also called "mean deviance minus the deviance of the means". $D(\bar{\theta})$ is regarded as classical estimate of fit given by MCMC simulation. The posterior mean deviance $\overline{D(\theta)}$ can be taken as a Bayesian measure of fit or "adequacy". The DIC is formed by the sum of classical estimate of fit and twice the effective number of parameters (p_D). We also can consider DIC as a Bayesian measure of fit or adequacy, penalized by an additional complexity term p_D. This is a reason that explains why DIC is intended as generalization of Akaike's Information Criterion (AIC). In summary, this method, DIC is also applied in this study to choose the fittest model between hierarchical binomial logit model and binary logit model.

3.4 PRE-SELECTION OF VARIABLES IN ACCIDENT DATASET

To apply the model for predicting crash severity, it is necessary to pre-select risk factors including time-related factors, road and environmental features, crash factors, and vehicles and drivers' characteristics. One way to choose variables is to examine previous researches. Besides, in accident data, some variables which relevantly affect drivers' injury are also considered in this study. On the other hand, categorizing independent variables is also based on similar studies on predicting crash severity. The description of predictors will be presented in the next chapter.

Accident data in Singapore contain three types including general accident information, vehicle and driver related information, and pedestrian information, each of which depicts different factors involved in accident. Therefore, based on previous studies and Singapore accident data, risk factors are selected to have effects on accident severity in

Singapore condition. Table 3.1 shows the selected variables in this study and reasons why these variables are considered. Finally, 22 factors that may be associated with drivers' injury have been selected from general accident information, vehicle and driver related information.

Table 3.1: Risk factors related to crash severity at signalized intersections in Singapore

	Variables	References of other studies	Selected variables for the study	Reasons
GENERAL ACCIDENT INFORMATION				
	accident severity at SI (A dependent variable)			Accidents occurring at signalized intersections consist of 20% of total accidents.
Time related factors	Year of accident	(Gray et al. 2008; Lee and Mannering 2002; Pai and Saleh 2008b; Quddus et al. 2002)	Y	New safety strategies are suggested in each year. This variable may present the efficiency of the strategies
	Month of accident	(Gray et al. 2008; Pai and Saleh 2008b; Quddus et al. 2002)	N	This variable presents seasons in year. It is dangerous to drive in winter. But seasons is not clear in Singapore.
	Day of accident	(Gray et al. 2008; Huang et al. 2008; Lee and Mannering 2002; Pai and Saleh 2008b; Quddus et al. 2002)	Y	Traffic volume may affect vehicle's speed. The higher speed, the more serious injury severity.
	Time of accident	(Chang and Mannering 1999; Gray et al. 2008; Huang et al. 2008; O'Donnell and Connor 1996; Pai and Saleh 2008b; Quddus et al. 2002; Zhang	Y	

		et al. 2000)		
Location related factors	Intersection type	(Huang et al. 2008; Quddus et al. 2002; Zhang et al. 2000)	Y	
Road features	Lane nature	(Huang et al. 2008; Quddus et al. 2002)	Y	Vehicle's position may present its directions such as turning left or right, or going straight. This may affect vehicle's speed.
	Street lighting	(Abdel-Aty 2003; Gray et al. 2008; Huang et al. 2008; Pai and Saleh 2008b; Quddus et al. 2002)	Y	This variable affects driver's visibility influencing the reduction of speed.
	Road speed limit	(Abdel-Aty 2003; Gray et al. 2008; Huang et al. 2008; Pai and Saleh 2008b; Quddus et al. 2002; Shankar and Mannering 1996)	Y	
	Road surface	(Gray et al. 2008; Huang et al. 2008; Quddus et al. 2002; Shankar and Mannering 1996)	Y	When the road is wet or weather is not good, drivers tend to reduce speed to control their vehicles. This may lead to less harmful.
	Weather condition	(Huang et al. 2008; Pai and Saleh 2008b; Quddus et al. 2002)	Y	
Crash related factors	Movement type	(Chang and Mannering 1999; Huang et al. 2008; O'Donnell and Connor 1996; Pai and Saleh 2008b; Quddus et al. 2002; Wong et al. 2007; Zhang et al. 2000)	Y	Head on collisions are more injured than other collisions: U turn or left turn etc because speed is also affected by movement type.
Other factors	Type of warning signs	(Pai and Saleh 2008b)	N	Signals may reminder drivers that a risk of accident may occur. But almost all observations

				are “not applicable”
	Pedestrian involvement	(Huang et al. 2008)		
	Safe drive zone in use		Y	Users may drive carefully and reduce vehicle’s speed because they know there is high population density in this area.
	Red light camera	(Huang et al. 2008; Quddus et al. 2002)	Y	These variables are to curb red-light running and driver’s fault. This may relieve severities
	Speed camera within 200m		Y	
	Hit & run	(Johnson 1997)	Y	Notification and emergency are delayed.
VEHICLE-DRIVER INFORMATION				
Vehicles factors	Vehicle registration number		N	
	Countries’ vehicle registration		Y	Different countries have different standard of vehicle maintenance, different training.
	Type of vehicle	(Abdel-Aty 2003; Chang and Mannering 1999; Huang et al. 2008; Pai and Saleh 2008b)	Y	Vehicle’s weight and speed produce energy when accidents occur. The more energy, the more severity.
	Vehicle make code		Y	Vehicle’s maintenance, engine, mass, and size affect injury severity
Driver factors	Child seat offence		N	96% of observations are not applicable
	Child injured		N	99% of observations are not applicable
	Driver	(Abdel-Aty 2003)	N	99% of

	belted			observations are use of the belt and not applicable.
	Type of driving license		Y	Licenses present driver's skills and training.
	Driver nationality	(Gray et al. 2008; Quddus et al. 2002)	Y	Different nationality may have different habits and behavior.
	Driver likely at fault	(Pai and Saleh 2008b; Porter and England 2000)	Y	Offending party affects driving ability of drivers. Driver's fault increase conflict with other vehicles.
	Age	(Abdel-Aty 2003; Gray et al. 2008; Huang et al. 2008; Quddus et al. 2002)	Y	These variables may present driver's experience, and immaturity
	Gender	(Abdel-Aty 2003; Gray et al. 2008; Huang et al. 2008; Quddus et al. 2002)	Y	

Note: Y denotes the selected variables and N denotes the unselected variables

3.5 SUMMARY

This chapter presents the formulation of full hierarchical binomial logit models. In addition, model evaluation including BCI and DIC is introduced to examine the significance of variables in the fixed part and random part and to select the best model between hierarchical binomial logit model and binary logit model, respectively. Pre-selection of variables is also prepared in this chapter so that application of hierarchical binomial logit model for crash severity at signalized intersections in Singapore will be illustrated and validated in the next chapter.

CHAPTER 4: APPLICATION OF HIERARCHICAL BINOMIAL LOGIT MODEL FOR ACCIDENT SEVERITY AT SIGNALIZED INTERSECTIONS

4.1 INTRODUCTION

Based on the proposed model and Singapore accident data, this chapter describes the application of hierarchical binomial logit model for solving injury severity of crashes at signalized intersections in Singapore. In this application, a description of dataset for predicting severity and model evaluation for validating the methodology are also summarized. The result of this study indicates factors that importantly influence crash severity, each of which will be discussed in detail. Finally, the summary of this study is given.

4.2 ACCIDENT DATA

For this study, accident data in Singapore from year 2003 to 2007 are used. This study focuses on investigating injury severity of accidents occurring at signalized intersections because the numbers of these crashes and vehicle-driver units are the highest in the dataset. In fact, based on data collection, 6991 crashes occur at signalized intersections, accounting for 20% of total accidents. Besides, the data show 13289 driver-vehicle units involved in these crashes, of which 5.1% cause fatal and serious injury and 94.9% cause slight and no injury.

In the hierarchical binomial logit model, a binary dependent variable refers to crash severity. The dependent variable (Y_{ij}) can take the value 0 or 1. If an accident has fatal or serious injury, it is called higher severity and Y_{ij} is equal to 1. Meanwhile, if an accident has slight or no injury, it is considered as less severe and Y_{ij} is equal to 0.

In addition to severity levels, independent variables which may have influences on accident severity are selected from Singapore accident data. Based on pre-selection of these variables presented in the previous chapter, there are 22 variables coded for each intersection accident. The definitions of covariates, together with their mean and standard deviation (S.D.) are presented in Table 4.1. According to Agresti (1996), an ordinal explanatory variable is treated as quantitative with conditions that statistical models fit well and have a single parameter rather than several ones. Therefore, to better analyze injury severity of accidents, all of the variables are split into groups of dummy variables based on previous and similar traffic safety researches. In addition, Greene (1993) suggested that continuous variables have been scaled (by dividing by N) to have their means lying between 0 and 1. This is because dummy variables have means between 0 and 1, and models are almost never correctly estimable if the continuous variables are of very different magnitudes (Greene 1993). This is also because the choice of a continuous variable's score has effect only on the results, where observations in each category are very unbalanced (Agresti 1996). Thus, time trend variable is categorized as year 2003=0.2, year 2004=0.4, year 2005=0.6, year 2006=0.8, and year 2007=1.

A correlation matrix for the explanatory variables, which may be associated with severity level, is checked to avoid multi-collinearity as well as wrong signs in the estimated coefficients. For the highly correlated variables, only the most significant variable is kept in the analysis. For example, weather condition is removed due to high correlation with road surface. Finally, the total covariates in the level 2: the crash level, used in analysis are *Time trend, Day of week, Time of day, Intersection type, Lane nature, Night time indicator, Road surface, Road speed limit, Safe drive zone, Presence of RLC, Speed camera within 200m, Hit & Run, and Pedestrian involvement*. In addition, covariates in the vehicle-driver level are *Vehicle movement, Registration, Driver nationality, Vehicle manufacture, Type of driving license, Involvement of offending party, Driver age, and Driver gender*.

Table 4.1: Covariates used in the model

Explanatory Covariates	**Description of the variables**	**Two-wheel vehicle**		**Light vehicle**		**Heavy vehicle**	
		Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
I.GENERAL							
1.Time trend	Year (Assuming 2003=0.2 to 2007=1.0)	0.601	0.287	0.602	0.292	0.605	0.284
2.Day of week	If accident at weekend=1, otherwise=0	0.281	0.449	0.334	0.472	0.278	0.448
3.Time of day							
- Peak time period (7am – 10am or 5pm – 8pm)	If accident at peak period =1, otherwise=0	0.296	0.456	0.255	0.436	0.320	0.467
II. ROAD CHARACTERISTICS							
4.Intersection type							
- X intersection	If accident at X intersection=0, otherwise=1	0.742	0.438	0.759	0.428	0.716	0.451
- Y/T intersection	If accident at Y/T intersection=0, otherwise=1	0.246	0.431	0.231	0.422	0.274	0.446
- Others	If accident at other intersections=0, otherwise=1	0.012	0.110	0.010	0.100	0.010	0.099
5.Lane nature							
- Left lane	If accident at left lane=1, otherwise=0	0.191	0.393	0.159	0.366	0.251	0.434
- Centre lane	If accident at centre lane=1, otherwise=0	0.291	0.454	0.318	0.466	0.264	0.441
- Right lane	If accident at right lane=1, otherwise=0	0.304	0.460	0.303	0.460	0.266	0.442
- Others	If accident at others=1, otherwise=0	0.196	0.397	0.202	0.402	0.190	0.392
6.Night time indicator							
- Night time	If accident in night time = 1, otherwise 0	0.678	0.467	0.685	0.464	0.622	0.485
7.Road surface	If dry = 0, otherwise 1	0.118	0.323	0.120	0.325	0.136	0.343
8.Weather condition	If fine = 0, otherwise 1	0.081	0.273	0.094	0.292	0.110	0.313
9.Road speed limit							
- <=50 km/h	If road speed limit is less than 50 km/h=1, otherwise=0	0.853	0.355	0.854	0.353	0.876	0.329
- 60 km/h	If road speed limit is 60 km/h=1, otherwise=0	0.105	0.307	0.107	0.310	0.085	0.278
- 70 km/h	If road speed limit is 70 km/h=1, otherwise=0	0.040	0.196	0.035	0.185	0.032	0.176
III. OTHER CHARACTERISTICS							
10.Safe drive zone in use	If Yes=1, otherwise=0	0.006	0.077	0.005	0.072	0.005	0.072
11.Presence of RLC	If RLC is present=1, otherwise=0	0.060	0.237	0.059	0.235	0.046	0.209
12.Speed camera within 200m	If speed camera within 200m C is present=1, otherwise=0	0.005	0.067	0.005	0.068	0.005	0.069
13.Hit & Run	If the offending vehicle hit and run away=1, otherwise=0	0.023	0.150	0.014	0.117	0.018	0.131
14.Pedestrian involvement	If pedestrian involved =1, otherwise=0	0.021	0.155	0.051	0.220	0.027	0.173
IV. VEHICLE CHARACTERISTICS							
15.Registration	If country's registration is Singapore=0, otherwise=1	0.102	0.303	0.022	0.146	0.172	0.377
16.Vehicle make code							
- HONDA	If a vehicle is HONDA =1, otherwise=0	0.518	0.500	0.510	0.500	-	-
- YAMAHA	If a vehicle is YAMAHA =1, otherwise=0	0.226	0.419	-	-	-	-
- SUZUKI	If a vehicle is SUZUKI =1, otherwise=0	0.157	0.364			-	-
- SYM	If a vehicle is SYM =1, otherwise=0	0.028	0.164	-	-	-	-
- KAWASAKI	If a vehicle is KAWASAKI =1, otherwise=0	0.029	0.167	-	-	-	-
- VESPA	If a vehicle is VESPA =1, otherwise=0	0.022	0.148	-	-	-	-
- TOYOTA	If a vehicle is TOYOTA =1, otherwise=0	-	-	0.213	0.409	0.570	0.495
- NISSAN	If a vehicle is NISSAN =1, otherwise=0	-	-	0.082	0.274	0.062	0.242
- HUYNDAI	If a vehicle is HUYNDAI =1, otherwise=0	-	-	0.066	0.249	0.039	0.194
- MITSHUBITSHI	If a vehicle is MITSHUBITSHI =1, otherwise=0	-	-	0.065	0.247	0.095	0.293
- MERCEDES BENZ	If a vehicle is MERCEDES BENZ =1, otherwise=0	-	-	0.013	0.115	0.009	0.092
- MAZDA	If a vehicle is MAZDA =1, otherwise=0	-	-	0.011	0.102	-	-
- B.M.W	If a vehicle is B.M.W =1, otherwise=0	-	-	0.008	0.090	-	-
- PROTON	If a vehicle is PROTON =1, otherwise=0	-	-	0.006	0.077	-	-
- RENAULT	If a vehicle is VESPA =1, otherwise=0	-	-	0.004	0.063	-	-
- FORD	If a vehicle is RENAULT =1, otherwise=0	-	-	-	-	0.057	0.231
- VOLVO	If a vehicle is VOLVO =1, otherwise=0	-	-	-	-	0.029	0.166
- ISUZU	If a vehicle is ISUZU =1, otherwise=0	-	-	-	-	0.005	0.069
- FIAT	If a vehicle is FIAT =1, otherwise=0	-	-	-	-	0.007	0.084
- OTHERS	If a vehicle is others =1, otherwise=0	0.010	0.099	0.022	0.146	0.128	0.334
17.Vehicle movement							
- Single vehicle self-skidded	If single vehicle self-skidded =1, otherwise=0	0.078	0.268	0.011	0.102	0.011	0.104

- Single vehicle against stationary objective or pedestrian	If single vehicle against stationary objective or pedestrian =1, otherwise=0	0.026	0.159	0.025	0.156	0.033	0.178
- Between moving vehicle(s) and stationary vehicle	If between moving vehicle(s) and stationary vehicle=1, otherwise=0	0.872	0.335	0.907	0.291	0.846	0.361
- Between moving vehicles	If between moving vehicle=1, otherwise=0	0.021	0.143	0.055	0.228	0.087	0.283
- Other movements	If other movements=1, otherwise=0	0.003	0.056	0.002	0.047	0.020	0.142
V. DRIVER CHARACTERISTICS							
18.Type of driving license	If driver license is Qualified Driving License-normal=0, otherwise=1	0.068	0.252	0.103	0.304	0.205	0.404
19.Nationality	If driver nationality is Singapore, =0, otherwise=1	0.163	0.370	0.065	0.246	0.222	0.416
20.Involvement of offending party	If driver is likely at fault=1, otherwise=0	0.496	0.500	0.657	0.475	0.320	0.467
21.Age							
- 0 – 25	If age <=25, =1, otherwise=0	0.396	0.489	0.102	0.303	0.126	0.332
- 26 – 45	If age within 26-45=1, otherwise=0	0.399	0.490	0.488	0.500	0.478	0.500
- 46 – 65	If age within 46-65=1, otherwise=0	0.180	0.384	0.377	0.485	0.368	0.482
- 66 – 100	If age > 66=1, otherwise=0	0.025	0.157	0.032	0.177	0.028	0.164
22.Gender	If gender is female =1, otherwise=0	0.036	0.187	0.161	0.368	0.029	0.169
Observations	N=13288						

4.3 MODEL CALIBRATION AND VALIDATION

4.3.1 MODEL CALIBRATION

At the beginning, the hierarchical binomial logit model is run with the 21 covariates from the dataset. Empirical Bayes estimation is employed to compute posterior mean, standard deviation, and Bayesian Credible Interval (BCI). According to Bolstad (2007), 95% BCI is computed for each covariate to examine whether each coefficient is significant or not. The covariate, which has 95% BCI containing 0, is eliminated. In addition, Winbugs software is used to estimate regression coefficients and random effects. Each of three chains of iterations produces a trace plot. Convergence has been achieved if all the chains appear to be overlapping one another. After convergence has been achieved, the MCMC simulation should be run for a further number of iterations to obtain samples that can be used for posterior inference. The more samples the simulation has, the more accurate posterior estimates will be. One way to assess the accuracy of the posterior estimates is by calculating the Monte Carlo error for each parameter. As a rule of thumb, the simulation should be run until the Monte Carlo error for each parameter of interest is less than about 5% of the sample standard deviation. In this study, trace plots with a good degree of mixing, produced from three chains of 40,000 iterations, indicate that estimation of coefficients are convergent. Then, 5% of the sample standard deviation of Monte Carlo error for each coefficient is obtained after next 10,000 iterations. The means and 95% BCI of estimated random effects and regression coefficients are monitored and presented in the Table 4.3.

In addition, the hierarchical binomial logit model is employed for each type of vehicles. This is because although vehicles are produced from the same manufactures, different types of vehicle have different influences on crash severity. Moreover, this study deals with accident severity at signalized intersections with all vehicles including two-wheel vehicles, light vehicles and heavy vehicles. Consequently, the HBL model is separated into 3 models to evaluate crash severity with two-wheel vehicles, light vehicles and heavy vehicles.

Table 4.2: Estimate of Deviance Information Criterion (DIC)

Deviance Information Criterion (DIC)	**Two-wheel vehicles**	**Light vehicles**	**Heavy vehicles**
- Hierarchical binomial logit model	1074.540	1521.640	946.803
- Binary logit model	1237.290	1606.970	1011.110

Table 4.3: Estimate of fixed part and random part

Explanatory Covariates in the HBL model	Two-wheel vehicles				Light vehicles				Heavy vehicles			
	Mean	95% BCI		OR	Mean	95% BCI		OR	Mean	95% BCI		OR
FIXED EFFECTS												
Day of week (relative to weekday)	-	-	-	-	-	-	-	-	0.472	0.116	0.830	1.604
Night time indicator (relative to daytime)												
- Night time	0.720	0.235	1.246	2.054	0.617	0.306	0.940	1.853	0.712	0.396	1.107	2.038
Road surface (relative to wet road surface)	-0.522	-1.021	-0.019	0.593	-0.436	-0.849	-0.034	0.646	-	-	-	-
Road speed limit (relative to Speed limit which is less than 50 km/h)												
- Speed limit is 60 km/h	0.980	0.405	1.555	2.665	0.550	0.144	0.947	1.734	-	-	-	-
- Speed limit is 70 km/h	0.525	-0.377	1.382	1.690	0.434	-0.240	1.035	1.544	-	-	-	-
Presence of RLC(relative to no red light camera)	1.099	0.436	1.776	3.001	0.387	0.036	0.721	1.472	-	-	-	-
Vehicle make code (relative to Honda manufacture)												
- YAMAHA	1.129	0.725	1.526	3.093	-	-	-	-	-	-	-	-
- SUZUKI	0.560	0.044	1.068	1.750	0.603	-0.454	1.548	1.828	-	-	-	-
- SYM	1.426	0.646	2.132	4.162	-	-	-	-	-	-	-	-
- KAWASAKI	1.376	0.537	2.115	3.959	-	-	-	-	-	-	-	-
- VESPA	0.578	-0.507	1.524	1.783	-	-	-	-	-	-	-	-
- TOYOTA	-	-	-	-	0.479	0.112	0.845	1.615	-	-	-	-
- NISSAN	-	-	-	-	0.572	0.056	1.059	1.771	-	-	-	-
- HUYNDAI	-	-	-	-	0.769	0.243	1.257	2.358	-	-	-	-
- MITSHUBITSHI	-	-	-	-	-0.040	-1.041	0.833	0.961	-	-	-	-
- MERCEDES BENZ	-	-	-	-	0.945	0.252	1.590	2.573	-	-	-	-
- MAZDA	-	-	-	-	0.627	-0.447	1.580	1.872	-	-	-	-
- B.M.W	-	-	-	-	0.166	-1.148	1.301	1.180	-	-	-	-
- PROTON	-	-	-	-	-0.191	-1.660	1.100	0.826	-	-	-	-
- RENAULT	-	-	-	-	0.021	-1.539	1.379	1.021	-	-	-	-
- FORD	-	-	-	-	0.814	-0.604	2.084	2.256	-	-	-	-
- VOLVO	-	-	-	-	-	-	-	-	-	-	-	-
- ISUZU	-	-	-	-	-	-	-	-	-	-	-	-
- FIAT	-	-	-	-	-	-	-	-	-	-	-	-
- OTHERS	0.475	-1.423	1.881	1.608	0.029	-1.074	0.977	1.030	-	-	-	-
Vehicle movement (relative to crashes between moving vehicle and stationary vehicle)												
- Single vehicle self-skidded	-1.357	-2.429	-0.396	0.257	0.047	-0.431	0.516	1.048	-0.077	-0.664	0.502	0.925
- Single vehicle against stationary objective or pedestrian	-0.647	-1.922	0.513	0.524	-	-	-	-	-0.264	-0.821	0.285	0.768
- Between moving vehicles	1.172	0.190	2.123	3.228	0.371	0.005	0.739	1.449	0.660	0.234	1.079	1.934
- Other movements	0.055	-1.663	1.705	1.056	0.137	-0.375	0.654	1.147	-0.103	-0.676	0.468	0.902
Involvement of offending party (relative to non-offending)	-	-	-	-	0.692	0.245	1.197	1.997	-	-	-	-
Driver age (relative to age 26 – 45)												
- 0 – 25	0.111	-0.259	0.489	1.118	-	-	-	-	-0.546	-1.018	-0.088	0.579
- 46 – 65	-0.008	-0.528	0.474	0.992	-	-	-	-	-0.185	-0.602	0.184	0.831
- 66 – 100	1.160	0.316	1.943	3.190	-	-	-	-	-0.007	-0.574	0.548	0.993
Driver gender (relative to male)	-	-	-	-	-0.535	-0.934	-0.155	0.586	-0.948	-2.260	-0.200	0.388
RANDOM EFFECTS												
Between-crash variance (τ_0^2)	2.942	0.467	5.897		0.700	0.295	1.303		1.579	0.636	2.745	
Involvement of offending party (τ_1^2)	-	-	-		1.506	0.465	2.886		-	-	-	
Gender (τ_2^2)	-	-	-		0.608	0.225	1.254		0.926	0.212	2.885	
Age (τ_3^2)												
- 0 – 25 (τ_4^2)	1.576	0.275	4.787		-	-	-		0.574	0.177	1.478	
- 46 – 65 (τ_5^2)	1.024	0.214	3.561		-	-	-		0.839	0.208	2.257	
- 65 – 100 (τ_6^2)	2.544	0.275	10.870		-	-	-		1.440	0.246	5.217	

4.3.2 MODEL VALIDATION

Model evaluation using Deviance Information Criterion (DIC) is also presented in Table 4.2. The model that has the minimum DIC is selected as the best. The result shows that in all three models with two-wheel vehicles, light vehicles, and heavy vehicles, the DIC values for hierarchical binomial logit modes (1074.540; 1521.640; and 946.803) are less than those in binary logit models (1237.290; 1606.970; and 1011.11), respectively. This means that the use of hierarchical binomial logit model in all of three cases is more suitable than that of binary logit models.

In addition, 95% BCI of estimated random effects indicates existence of random intercept effects in all three models. Besides, random slope effects are also identified. For example, while the age variable has random slope in the two-wheel-vehicle model, there are two random slope effects: involvement in party variable and gender variable in light-vehicle model and three random slope effects: age variable, gender variable and vehicle registration variable in heavy-vehicle model.

4.4 DISCUSSION OF SIGNIFICANT RISK FACTORS

From the hierarchical binomial logit model, the effects of the covariates are presented in Table 4.1. In the final model, 10 variables are significant with 95% BCI which does not contain 0. They are: 1)Day of week, 2)Night time indicator, 3)Road surface, 4)Road speed limit, 5)Presence of RLC, 6)Vehicle make code, 7)Vehicle movement, 8)Involvement of offending party, 9)Age, and 10)Gender. The interpretations of these significant covariates are discussed in the following.

Day of week

Day of week is categorized into 2 groups: crash occurrence at weekend or on weekday. This covariate is found to significantly affect the crash severity involved in only heavy vehicles. The parameter is positive (0.472, 95% BCI (0.116; 0.830), OR 1.640), indicating that crashes at the weekend have 64.0% higher odds of high crash severity than those on weekdays. This finding is similar to a study of Chang and Mannering (1999) who found that truck-involved crash severity both increases at weekends and is higher than non-truck-involved crashes. This may be reasonable because lower traffic volume at the weekend may lead to the increase of vehicle speed. The fact that heavy-vehicle drivers may drive fast to finish their work as soon as possible at weekend to take a rest significantly increases casualties' injury. Meanwhile, light vehicles and two-wheel vehicles do not affect the severity because drivers may carefully control their vehicles and there are a few two-wheel vehicles at weekend.

Night time indicator

Night time indicator covariate has two categories including day time and night time. The finding indicates this covariate is found to be significant in all of the three vehicle types. Crashes in night time have 105.4%, 85.3% and 103.8% higher of odd ratio of the severity than those in day time with two-wheel vehicles, light vehicles, and heavy vehicles, respectively. This result is consistent with Simoncic (2001) finding that crashes at night are more seriously severe than those during day time. The reasons are that driver visibility in night time may be less than that in day time and that speeding and alcohol use increase severity at night. Among three models, crashes associated with two-wheel vehicles have the highest increase of severity because two-wheel vehicles may not been clearly seen by other vehicles.

Road surface

Wet road surface is identified as a significant factor that has effects on the crash severity associated with two-wheel vehicles and light vehicles instead of heavy ones. The analysis described above shows that the coefficient of two-wheel-related accidents is -0.522 (95%BCI (-1.202; -0.019)) and that of heavy vehicle-related ones is -0.436 (95%BCI (-0.849; -0.034)). Occupants in two-wheel vehicles and light vehicles have a decrease of severity in odds ratio by 40.7% and 35.4%, respectively, when compared with those involved in crashes on dry-road surface. Some studies (Quddus et al. 2002; Rifaat and Chin 2005) also found the same result that accident severity decreases on the wet road surface. According to statistics about Singapore weather, the rain is often heavy so that driver visibility may reduce; thus, drivers are inclined to reduce their speed during the bad surface. So, the fact wet road surface decrease crash severity may be reasonable.

Road speed limit

The finding indicates that speed limit covariate significantly influences the crash severity related to two-wheel vehicles and light vehicles. Compared with those where speed limit is less than 50 km/h, the crashes on roads, in which speed limit is 60 km/h, increase the severities by 166.5% and 73.4% with two-wheel vehicles and light vehicles, respectively. Zhang et al. (2000) also found that the odds of fatality in crashes occurring in zones with higher speed are higher than those in crashes occurring in zones with lower speed. The higher their speed is, the more difficult drivers are able to stop. Therefore, drivers are more likely to have fault in controlling their vehicles, resulting in more serious severity.

Presence of Red Light Camera

The result shows that the presence of Red Light Camera is associated with higher severity by 200.1% and 47.2% with both two-wheel vehicles and light vehicles. This finding is also similar to some studies: (Erke ; Huang et al. 2008; Quddus et al. 2002). The reasons are that many drivers tend to run when light is red. However, they know the existence of RLC, suddenly stopping their vehicles. Specially, two-wheel vehicles are more likely to be skidded when the wheel is suddenly stopped. Besides, Red Light Cameras are often installed at high risk locations. Thus, more information such as drivers' behavior and distraction, when drivers know the existence of RLC at intersections, should be obtained to better understand the effects of this variable on crash severity.

Vehicle movement

Five vehicle-movement categories are single self-skidded, vehicle against stationary or pedestrian, between vehicle and stationary vehicle, between vehicles, and others, where a reference case is a crash between vehicles and stationary vehicles. The finding indicates that movement between vehicles covariate when compared with the base case is positive and significant in 3 types of vehicles: two-wheel vehicles, light vehicles, and heavy vehicles, where their odds ratios are 3.228, 1.449, and 1.934, respectively. This means that vehicle movement between vehicles increases severity. The reasons are that more energy is created when collisions between two vehicles occur from opposite directions and that vehicles have higher speed in the same directions when a signal light allows them to enter across intersections at that time. On the other hand, a self-single vehicle movement is only negatively and significantly affected in two-wheel vehicle case (-1.357, 95% BCI (-2.429; -0.396), OR 0.257). This covariate decreases the odds ratio of severity by 74.3%. In this situation, driver's damage results from skid between drivers and road surface. However, helmet and clothes can protect motorcyclists from the injury. So, the decrease of severity in this case may be reasonable.

Vehicle manufacture

Vehicle make covariate is found to significantly affect the crash severity containing two-wheel vehicles and light vehicles. In two-wheel vehicles, compared with reference case: HONDA, four manufactures, including YAMAHA, SUZUKI, SYM, and KAWASAKI, have significant influences on severity by odds ratio 3.093, 1.750, 4.162 and 3.959, respectively. O'Donnell and Connor (1996) also found that a specific vehicle make increases motorcyclist crash severity among different manufactures. On the other hand, light vehicles are made by HONDA, TOYOTA, NISSAN, HYUNDAI, MITSUBISHI, MERCEDES BENZ, SUZUKI, MAZDA, B.M.W, PROTON, RENAULT, FORD and others, where other makes have a total of less than 10 units. Relative to HONDA, four manufactures which are positively and significantly related to the accident severity are TOYOTA, NISSAN, HYUNDAI, and MERCEDES BENZ, where odds ratios are 1.615, 1.771, 2.358 and 2.573, respectively. This is because the population of Honda two-wheel vehicles and Honda light vehicles has the most increase every year, meaning that vehicles of Honda are always new. The newer vehicles are better maintained and less breakdown. So, the crash severities of Honda decrease in both two-wheel vehicles and light vehicles.

Involvement of offending party

The finding indicates only the crash severities of light vehicles are significantly associated with the at-fault driver covariate. The at-fault drivers have 99.7% higher odds ratio of crash severity than the not-at-fault driver (0.692, 95% BCI (0.245; 1.197), OR 1.997). The reason is that drivers involved in offending party may neither give way to other vehicles nor stop their vehicles when entering on intersections even though the signal light is red. This also provides evidence for educating drivers to keep away from risk-taking maneuvers.

Age

Four age groups are categorized based on the similarities of drivers' behavior and ability to compare the effect of age on severity. The finding shows that the crash severity associated with two-wheel vehicles is highest for the group that is more than 65 (1.160, 95% BCI (0.316; 1.943), OR 3.190). The reasons are that decrease of visual power, deterioration of muscle strength and reaction time may be responsible for an age group of 65 to be associated with severity (Rifaat and Chin 2005) and older drivers have relatively weak risk reacting ability. On the other hand, the finding indicates that the crashes in age group being less than 25 decreases the severity related to heavy vehicles, where the parameter, BCI, and odds ratio are (-0.546, 95% BCI (-1.018; -0.088), OR 0.579), respectively. Young heavy-vehicle drivers are most likely to be in good health and trained. Therefore, the finding may be reasonable.

Gender

The gender variable is classified as 2 cases male and female where the base case is male. The estimations find that the crash severity related to light vehicles and heavy vehicles is significantly affected by this predictor. The female drivers have 41.4% and 61.2% lower odds ratio of crash severity than the male driver in the light-vehicle model and the heavy-vehicle model, respectively. The reasons are that female drivers usually drive more carefully and use new version cars and that female health and ability are improved. This finding is also similar to the study of Chang and Mannering (1999) who found that female drivers decrease crash severity.

4.5 SUMMARY

This study develops hierarchical binomial logit model with both random intercept and slope effect to find the impacts of risk factors on individual severity of occupants involved in crashes at signalized intersections in Singapore. Model evaluation including DIC and BCI is used to ensure that the hierarchical binomial logit model is more suitable than binary logit mode and that there is existence of random intercept and slope effects in hierarchical binomial logit model.

Application of hierarchical binomial logit model for individual severity of occupants involved in crashes at intersections indicates that 10 variables are identified as significant factors by using 95% BCI. These variables include *Day of week, Night time indicator*, *Road surface, Road speed limit, and Presence of RLC* in the level 2. In particular, crashes occurring at night increase accident severity in all 3 situations of vehicle types. Besides, in both 2 cases: two-wheel vehicles and light vehicles, wet road surface reduces the injury severity while high speed limit and presence of red light camera increase the accident injury. In the vehicle-driver level of crash severity, *Vehicle manufacture, Vehicle movement, Involvement of offending party, Age, and Gender* are also identified to be associated with crash severity. For example, with vehicle characteristics, this study finds that Honda manufacture is safer than other vehicle makes in two-wheel vehicle and light vehicle cases. In addition, vehicle movement variable significantly affects all of three models of crash severity. Meanwhile, three driver factors are vitally indentified. Female drivers decrease severity in crashes related to light vehicles and heavy vehicles. Furthermore, age group over 65 related to two-wheel vehicles is also positively associated with occupant severity, while *Involvement of offending party* increases crash severity involved in light vehicles.

In summary, this study solves multilevel data structure which may exist in dataset by using hierarchical techniques and identifies some risk factors which contribute to the injury severity of crashes at signalized intersections.

CHAPTER 5: CONTRIBUTIONS, DISCUSSIONS, RECOMENDATIONS AND CONCLUSIONS

5.1 RESEACH CONTRIBUTIONS

The principal objective of this study is to identify factors affecting severity of crashes at signalized intersections by using the hierarchical binomial logit model with both random intercept and slope effects. In order to achieve this objective, various factors (e.g. general accident characteristics, road conditions, vehicle characteristics and driver characteristics) have been investigated. In addition, this model calculated with Winbugs software establishes the relationship between injury severity and risk factors. Besides, model evaluation including DIC and BCI is applied to assess the suitability of the model. This study uses Singapore accident data to illustrate the application of hierarchical binomial logit model. In the result, 95% BCI in random part indicates the random slope effects (such as *Involvement of offending party, Gender and Age* variables) exist. Furthermore, based on the DIC values of two models in three cases of vehicle types, the finding also shows this model is able to take account for severity correlation of vehicle-driver unit involved in the same crash as well as to improve the estimation of regression coefficients and standard errors (more details of DIC and 95% BCI value in three vehicles are presented in Table 4.2). Finally, the result demonstrates 10 variables (details of parameters are presented in Chapter 4) significantly affect the severity.

5.2 DISCUSSIONS AND RECOMENDATIONS

The hierarchical binomial logit model establishes the relationship between accident severity at signalized intersections and risk factors. The result indicates three groups of factors are important.

First of all, general characteristics including *Day of week* and *Night time indicator* have influences on the crash severity. Accidents occurring at weekend are increasingly severed since drivers have a tendency to speed when a density of vehicle is low. Besides, because of low visibility, alcohol and high speed at night, drivers' reaction which is delayed may increase the severity. Therefore, in order to improve traffic safety, drivers should be alert and not be tempted to increase speed to such an extent that makes it difficult to control the vehicle.

The second group is road factors (such as *Road surface, Road speed limit* and *Presence of RLC*). The wet road surface condition has been found to significantly reduce the severity because drivers carefully control their vehicles on wet surface and across signalized intersection. In addition, road speed limit variable are significant. Drivers tend to run fast on roads which have high speed limit. As a result, it is difficult for drivers to manage vehicle when accidents happen. Therefore, the finding that high road speed limit positively affects the severity is reasonable. On the other hand, the presence of RLC is associated with higher severity. It does not imply that presence of RLC increases the severity level because it is installed at dangerous locations with more severe accidents. Thus, more information such as drivers' behavior and distraction should be obtained so that prediction of severity is more accurate.

Finally, driver-vehicle characteristics consist of five variables. *Vehicle manufacture* and *Vehicle movement* are significant. Accidents between 2 moving vehicles result in the high impact force. So, the finding that crash severity in 3 case studies increases significantly when vehicles are moving is reasonable. On the other hand, the at-fault driver-vehicle unit of *Involvement of offending party* variable has a positive effect on the severity. This provides a more convincible evidence for educating drivers to keep away from risk-taking maneuvers. Furthermore, *Age and Gender* are also identified to be associated with the severity of crashes at signalized intersections. For example, over 65 age group related to two-wheel vehicles is also positively associated with the crash severity because visual and physical ability of older driver is deteriorated. Meanwhile, female drivers decrease severity in crashes related to light vehicles and heavy vehicles due to driving more carefully and soberly. Based on the finding related to driver-vehicle characteristics, public information programs should be developed to encourage all drivers to properly follow traffic legislation.

In summary, this study investigates one problem that multilevel data structures are ignored in traffic safety by using full hierarchical binomial logit model. However, this study still has some limitations such as models and data. For example, this model cannot be able to handle dependent variables that are classified as ordinary variables. Besides, this study only solves multilevel data that contain 2 levels: the severity within crash clusters. Therefore, a new model such as hierarchical ordered logit/probit model with random intercept or both random intercept and slope effects should be developed.

5.3 CONCLUSIONS

In conclusion, the research develops full hierarchical binomial logit model with both random intercept and slope effects in order to investigate multilevel data structures and establish the relationship between the severity and risk factors. This study also finds that some factors such as day of week, night time, road surface, speed limit, present of RLC, vehicle manufacture and movement, involvement of offending party, and driver gender and age are significant influences on crash severity at signalized intersections. The findings of this study give a basis for developing effective countermeasures to improve road safety.

References

Abdel-Aty, M. (2003). "Analysis of driver injury severity levels at multiple locations using ordered probit models." *Journal of Safety Research*, 34(5), 597-603.

Abdel-Aty, M., and Keller, J. (2005). "Exploring the overall and specific crash severity levels at signalized intersections." *Accident Analysis & Prevention*, 37(3), 417-425.

Agresti, A. (1996). "An Introduction to Categorical Data Analysis." *New York: Wiley*.

Akaike, H. (1973). "Information theory and an extension of maximun likelihood principle." *In Proc 2 nd Int. Symp. Information Theory, Budapest: Akademiai*, 267-281.

Ben-Akiva, M., and Lerman, S. R. (1985). "Discrete Choice Analysis: Theory and Application to Travel Demand." Cambridge, MA: MIT Press.

Bolstad, W. M. (2007). "Introduction to Bayesian statistics." *A John Wiley & Sons, Inc, Hoboken, New Jeysey, Publication*.

Chang, L.-Y., and Mannering, F. (1999). "Analysis of injury severity and vehicle occupancy in truck- and non-truck-involved accidents." *Accident Analysis & Prevention*, 31(5), 579-592.

De Lapparent, M. (2006). "Empirical Bayesian analysis of accident severity for motorcyclists in large French urban areas." *Accident Analysis and Prevention*, 38(2), 260-268.

Eluru, N., Bhat, C. R., and Hensher, D. A. (2008). "A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes." *Accident Analysis & Prevention*, 40(3), 1033-1054.

Erke, A. "Red light for red-light cameras?: A meta-analysis of the effects of red-light cameras on crashes." *Accident Analysis & Prevention*, In Press, Corrected Proof.

Evans, L., and Frick, M. C. (1992). "Car Size or Car Mass: Which Has Greater Influence on Fatality Risk?" *American Jounal of Public Health*.

Evans, L., and Frick, M. C. (1993). "Mass ratio and relative driver fatality risk in two-vehicle crashes." *Accident Analysis & Prevention*, 25(2), 213-224.

Factor, R., Mahalel, D., and Yair, G. (2008). "Inter-group differences in road-traffic crash involvement." *Accident Analysis & Prevention*, 40(6), 2000-2007.

Fredette, M., Mambu, L. S., Chouinard, A., and Bellavance, F. (2008). "Safety impacts due to the incompatibility of SUVs, minivans, and pickup trucks in two-vehicle collisions." *Accident Analysis & Prevention*, 40(6), 1987-1995.

Gelfand, A. E., and Dey, D. K. (1994). "Bayesian model choice: asymptotics and exact calculations." *Journal of the Royal Statistical Society, Series B*, 56, 501-514.

Goldstein, H. (2003). "Multilevel Statistical Models." *3 rd Edition. Edward Arnold, London*.

Gray, R. C., Quddus, M. A., and Evans, A. "Injury severity analysis of accidents involving young male drivers in Great Britain." *Journal of Safety Research*, In Press, Corrected Proof.

Gray, R. C., Quddus, M. A., and Evans, A. (2008). "Injury severity analysis of accidents involving young male drivers in Great Britain." *Journal of Safety Research*, 39(5), 483-495.

Greene, W. H. (1993). "Econometric Analysis (2nd ed.)." *New York: Macmillan*.

Haque, M. M., Chin, H. C., and Huang, H. "Modeling fault among motorcyclists involved in crashes." *Accident Analysis & Prevention*, In Press, Corrected Proof.

Huang, H., Chin, H. C., and Haque, M. M. (2008). "Severity of driver injury and vehicle damage in traffic crashes at intersections: A Bayesian hierarchical analysis." *Accident Analysis & Prevention*, 40(1), 45-54.

Islam, S., and Mannering, F. (2006). "Driver aging and its effect on male and female single-vehicle accident injuries: Some additional evidence." *Journal of Safety Research*, 37(3), 267-276.

Johnson, C. D. (1997). "Pedestrian Fatalities on Interstate Highways Characteristics and Countermeasures." *Transportation Research Record*(*1578*), pp.23-29.

Jones, A. P., and Jørgensen, S. H. (2003). "The use of multilevel models for the prediction of road accident outcomes." *Accident Analysis & Prevention*, 35(1), 59-69.

Joshua, S. C., and Garber, N. J. (1990). "Estimating truck accidents rate and involvements using linear and Poission regression models." *Transportation Planning and Technology*, 15(1), 41-58.

Khorashadi, A., Niemeier, D., Shankar, V., and Mannering, F. (2005). "Differences in rural and urban driver-injury severities in accidents involving large-trucks: An exploratory analysis." *Accident Analysis & Prevention*, 37(5), 910-921.

Kim, D.-G., Lee, Y., Washington, S., and Choi, K. (2007a). "Modeling crash outcome probabilities at rural intersections: Application of hierarchical binomial logistic models." *Accident Analysis & Prevention*, 39(1), 125-134.

Kim, J.-K., Kim, S., Ulfarsson, G. F., and Porrello, L. A. (2007b). "Bicyclist injury severities in bicycle-motor vehicle accidents." *Accident Analysis & Prevention*, 39(2), 238-251.

Kim, J.-K., Ulfarsson, G. F., Shankar, V. N., and Kim, S. (2008). "Age and pedestrian injury severity in motor-vehicle crashes: A heteroskedastic logit analysis." *Accident Analysis & Prevention*, 40(5), 1695-1702.

Kockelman, K. M., and Kweon, Y.-J. (2002). "Driver injury severity: An application of ordered probit models." *Accident Analysis and Prevention*, 34(3), 313-321.

Langley, J., Mullin, B., Jackson, R., and Norton, R. (2000). "Motorcycle engine size and risk of moderate to fatal injury from a motorcycle crash." *Accident Analysis and Prevention*, 32(5), 659-663.

Lee, C., and Abdel-Aty, M. (2005). "Comprehensive analysis of vehicle-pedestrian crashes at intersections in Florida." *Accident Analysis & Prevention*, 37(4), 775-786.

Lee, J., and Mannering, F. (2002). "Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis." *Accident Analysis & Prevention*, 34(2), 149-161.

Lonczak, H. S., Neighbors, C., and Donovan, D. M. (2007). "Predicting risky and angry driving as a function of gender." *Accident Analysis & Prevention*, 39(3), 536-545.

Long, J. S. (1997). "Regression Models for Categorical and Limited Dependent Variables." *Thousand Oaks, CA, Sage Publications*.

MacFadden, D. (1973). "Conditional logit analysis of qualitative choice bahavior." *Frontiers of Econometrics* 105-142. New York: Academic Press.

MacFadden, D. (1981). "Econometric model of probabilistic choice." *Structural Analysis of Discrete Data*, 198-272. Cambridge, MA: MIT Press.

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

O'Donnell, C. J., and Connor, D. H. (1996). "Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice." *Accident Analysis and Prevention*, 28(6), 739-753.

Obeng, K. (2007). "Some determinants of possible injuries in crashes at signalized intersections." *Journal of Safety Research*, 38(1), 103-112.

Pai, C.-W. "Motorcyclist injury severity in angle crashes at T-junctions: Identifying significant factors and analysing what made motorists fail to yield to motorcycles." *Safety Science*, In Press, Corrected Proof.

Pai, C.-W., and Saleh, W. (2008a). "Exploring motorcyclist injury severity in approach-turn collisions at T-junctions: Focusing on the effects of driver's failure to yield and junction control measures." *Accident Analysis & Prevention*, 40(2), 479-486.

Pai, C.-W., and Saleh, W. (2008b). "Modelling motorcyclist injury severity by various crash types at T-junctions in the UK." *Safety Science*, 46(8), 1234-1247.

Porter, B. E., and England, K. J. (2000). "Predicting red-light running behavior: a traffic safety study in three urban settings." *Journal of Safety Research*, 31(1), 1-8.

Preusser, D. F., Williams, A. F., and Ulmer, R. G. (1995). "Analysis of fatal motorcycle crashes: crash typing." *Accident Analysis & Prevention*, 27(6), 845-851.

Quddus, M. A., Noland, R. B., and Chin, H. C. (2002). "An analysis of motorcycle injury and vehicle damage severity using ordered probit models." *Journal of Safety Research*, 33(4), 445-462.

Raftery, A. E. (1986). "Choosing models for cross-classifications." *American Sociological Review*, 51, 145-146.

Raftery, A. E. (1995). "Bayesian model selection in social research." *In P.Marsden (Ed.) Sociological Methodology, Washington, DC: The American Sociological Association*, 111-163.

Rifaat, S. M., and Chin, H. C. (2005). "Analysis of severity of single-vehicle crashes in Singapore." *In: TRB 2005 Annual Meeting CD-Rom, Transportation Research Board, National Research Council, Washington D.C.*

Savolainen, P., and Mannering, F. (2007). "Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes." *Accident Analysis & Prevention*, 39(5), 955-963.

Shankar, V., and Mannering, F. (1996). "An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity." *Journal of Safety Research*, 27(3), 183-194.

Shankar, V., Mannering, F., and Barfield, W. (1996). "Statistical analysis of accident severity on rural freeways." *Accident Analysis & Prevention*, 28(3), 391-401.

Simoncic, M. (2001). "Road accidents in Slovenia involving a pedestrian, cyclist or motorcyclist and a car." *Accident Analysis and Prevention*, 33(2), 147-156.

Skinner, C. J., Holt, D., and Smith, T. M. F. (1989). "Analysis of Complex Surveys." *Wiley, Chichester, UK.*

Snijders, T. A. B., and Bosker, R. J. (1999). "Multilevel analysis : an introduction to basic and advanced multilevel modeling." *Thousand Oaks, CA, Sage Publications*.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Linde, V. D. (2003a). "Bayesian measures of model complexity and fit (with discussion)." *Journal of the Royal Statistical Society, Series B*, 64(4), 583-616.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Lunn, D. (2003b). "WinBUGS version 1.4.1 User Manual." *MRC Biostatistics Unit, Cambridge, UK.*

Tay, R., Rifaat, S. M., and Chin, H. C. (2008). "A logistic model of the effects of roadway, environmental, vehicle, crash and driver characteristics on hit-and-run crashes." *Accident Analysis & Prevention*, 40(4), 1330-1336.

Ulfarsson, G. F., and Mannering, F. L. (2004). "Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents." *Accident Analysis & Prevention*, 36(2), 135-147.

Wang, X., and Abdel-Aty, M. (2008). "Analysis of left-turn crash injury severity by conflicting pattern using partial proportional odds models." *Accident Analysis & Prevention*, 40(5), 1674-1682.

Washington, S., Congdon, P., Karlaftis, M., and Mannering, G. (2005). "Bayesian multinomial logit models: exploratory assessment of transportation applications." *In TRB 2005 Annual Meeting CD-ROM, Tranportation Research Board, National Research Council, Washington D.C.*

Wong, S. C., Sze, N. N., and Li, Y. C. (2007). "Contributory factors to traffic crashes at signalized intersections in Hong Kong." *Accident Analysis & Prevention*, 39(6), 1107-1113.

Yannis, G., Golias, J., and Papadimitriou, E. (2005). "Driver age and vehicle engine size effects on fault and severity in young motorcyclists accidents." *Accident Analysis & Prevention*, 37(2), 327-333.

Zajac, S. S., and Ivan, J. N. (2003). "Factors influencing injury severity of motor vehicle-crossing pedestrian crashes in rural Connecticut." *Accident Analysis & Prevention*, 35(3), 369-379.

Zhang, J., Lindsay, J., Clarke, K., Robbins, G., and Mao, Y. (2000). "Factors affecting the severity of motor vehicle traffic crashes involving elderly drivers in Ontario." *Accident Analysis and Prevention*, 32(1), 117-125.