INVESTIGATING THE EFFECTS OF SAMPLE SIZE, MODEL MISSPECIFICATION AND UNDERREPORTING IN CRASH DATA ON THREE COMMONLY USED TRAFFIC CRASH SEVERITY MODELS

A Dissertation

by

FAN YE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2011

Major Subject: Civil Engineering

INVESTIGATING THE EFFECTS OF SAMPLE SIZE, MODEL MISSPECIFICATION AND UNDERREPORTING IN CRASH DATA ON THREE COMMONLY USED TRAFFIC CRASH SEVERITY MODELS

A Dissertation

by

FAN YE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Dominique Lord
Committee Members,	Luca Quadrifoglio
	Yunlong Zhang
	Michael Speed
Head of Department,	John Niedzwecki

May 2011

Major Subject: Civil Engineering

ABSTRACT

Investigating the Effects of Sample Size,

Model Misspecification and Underreporting in Crash Data

on Three Commonly Used Traffic Crash Severity Models. (May 2011)

Fan Ye, B.E., Southeast University, China;

M.S., Southeast University, China

Chair of Advisory Committee: Dr. Dominique Lord

Numerous studies have documented the application of crash severity models to explore the relationship between crash severity and its contributing factors. These studies have shown that a large amount of work was conducted on this topic and usually focused on different types of models. However, only a limited amount of research has compared the performance of different crash severity models. Additionally, three major issues related to the modeling process for crash severity analysis have not been sufficiently explored: sample size, model misspecification and underreporting in crash data. Therefore, in this research, three commonly used traffic crash severity models: multinomial logit model (MNL), ordered probit model (OP) and mixed logit model (ML) were studied in terms of the effects of sample size, model misspecification and underreporting in crash data, via a Monte-Carlo approach using simulated and observed crash data.

The results of sample size effects on the three models are consistent with prior expectations in that small sample sizes significantly affect the development of crash severity models, no matter which model type is used. Furthermore, among the three models, the ML model was found to require the largest sample size, while the OP model required the lowest sample size. In addition, when the sample size is sufficient, the results of model misspecification analysis lead to the following suggestions: in order to decrease

the bias and variability of estimated parameters, logit models should be selected over probit models. Meanwhile, it was suggested to select more general and flexible model such as those allowing randomness in the parameters, i.e., the ML model. Another important finding was that the analysis of the underreported data for the three models showed that none of the three models was immune to this underreporting issue. However, setting data properly could minimize the bias and variability. Furthermore, when the full or partial information about the unreported rates for each severity level is known, treating crash data as outcome-based samples in model estimation, via the Weighted Exogenous Sample Maximum Likelihood Estimator (WESMLE), dramatically improve the estimation for all three models.

to my family: Mom, Dad and Brother

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and deep appreciation to my advisor, Dr. Dominique Lord, for his encouragement and guidance throughout the completion of this dissertation. I also would like to extend my thanks to the committee members, Dr. Yunlong Zhang, Dr. Luca Quadrifoglio and Dr. Michael Speed for their time, suggestions and advise for this dissertation. Special thanks are given to Dr. Yunlong Zhang for his hearty support and advice during my study all these years at Texas A&M University.

I'm greatly indebted to the Texas Transportation Institute (TTI) for providing me an invaluable research experience and supporting me with funding the entire period of my PhD study. I am very thankful to Dr. Paul Carlson and Dr. Jerry Ullman's supervision at TTI on my projects. Especially, I wish to express my deep gratitude and special thanks to Dr. Paul Carlson for his continuous support and help and for giving me many opportunities during my research and study.

I also wish to thank my dear friends, Tara Ramani, Ridwan Quaium, Jon Re, John Lowery, Hien Pham, Lequn Hu, Pei-Fen Kuo, Chung-Wei Shen, Teresa Qu, Lindsay Liggett, Cameron Williams, Byung-Jung Park, Sunil Patil and Rachael Stensrud for their friendship, support and encouragement throughout graduate school which made my studies at Texas A&M University memorable.

Last but not least, I would like to express my deep gratitude to my parents and brother for their endless love, support, and patience.

TABLE OF CONTENTS

CHAPTER Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

Page

CHAPTER I

INTRODUCTION

Motor vehicle crashes are an issue of concern worldwide. In the U.S., traffic crashes bring about more loss to human life (as measured in human-years) than almost any other cause – falling behind only cancer and heart disease based on 2004 data (NHTSA, 2005). Furthermore, they are the leading cause of death for people between the ages of 3 and 34 (NHTSA, 2009). Much progress has been made to improve traffic safety in the U.S. by implementing various traffic safety research programs, with the aim of reducing the number of deaths and serious injuries. For instance, in 2008, the fatality rate per 100 million vehicle miles of travel (VMT) fell to a historic low of 1.27, compared to the number of 1.58 for the 1998 (NHTSA, 2009). However, the number of people involved in traffic crashes is still very high. In 2008, from the total of 5,811,000 police-reported traffic crashes, 37,261 people were killed, 2,346,000 people were injured, and 4,146,000 crashes involved property damage only (NHTSA, 2009). The number of fatalities above indicates that an average of 102 people died each day in motor vehicle crashes which is equal to about one fatality every 14 minutes. The loss caused by traffic crashes is represented by the enormous economic cost, which was estimated to be $230.6 billion in the year of 2000, even without including the cost of delays imposed on other travelers which are also significant. In addition, traffic crashes are also one of the leading causes of death and injury in many other countries. For example, China, with only 1.9% of the world's vehicles, has about 15% of global traffic deaths. Expressed differently, this is translated to the loss of 100,000 lives due to traffic crashes in the year of 2005 (NHTSA, 2005) which is equivalent to 250 Boeing 747 Jumbo jets crashing every year, in China alone.

This dissertation follows the style of Accident Analysis and Prevention.

A traffic crash is usually a consequence of the three elements: driver errors, vehicle characteristics and road environment. Therefore, in order to reduce the frequency of traffic crashes and their resulting injury severity, we need safer drivers, safer vehicles and safer roads. Some driver education programs and law enforcement will help drivers to increase their awareness with regards to safer driving. For safer vehicles, some new technologies used in vehicles can help drivers to avoid crashes or decrease the severity of crashes by controlling the vehicles better. These technologies include backup cameras, anti-lock braking systems, emergency brake assist, etc. In terms of road environment, it is important to improve road environment by incorporating safety information in geometric design of roadways with various traffic facilities. A good geometric design of roadways not only meets the design standard, but is also forgiving of the drivers' mistakes, by providing sufficient clear zone, end treatments of guardrails, flat slope of roadside, etc.

To raise awareness of the importance of road safety, increased attention is being paid to traffic safety analysis using various statistical models. They can be used to extract the information about the impacts of the contributing factors to traffic crashes based on crash records. Among all types of traffic safety analysis, the development and application of crash prediction models is one of the most important aspects. Usually there are two types of crash prediction models: crash count models and crash severity models. Crash count models are generally count regression models (e.g., Poisson and Negative Binomial distributions) which estimate the probability of observing the number of crashes falling into different severity levels. Crash severity models (e.g., various types of logit or probit models) intend to estimate the probability of a crash falling into one of the severity levels conditional on the fact that a crash has occurred. In addition, crash severity models using disaggregate data could capture the relationship between each crash and its contributing factors such as driver characteristics, vehicle characteristics, roadway conditions, and road-environment factors, while crash count models using aggregate data could not offer such a great insight.

As mentioned previously, crash severity models play an important role in traffic crash prediction. Therefore, developing a sound and reliable crash severity model is necessary for traffic safety analysis. The primary goal of this research is to study the three commonly used traffic crash severity models: multinomial logit model (MNL), ordered probit model (OP) and mixed logit model (ML), in terms of the effects of sample size, model misspecification and underreporting in crash data. This would equip researchers with a deeper understanding of the three models and furthermore to develop more sound and reliable crash severity models. The remainder of this chapter consists of three sections. Section 1.1 provides the problem statement. In Section 1.2, specific objectives of this research are provided. The outline of the dissertation is presented in Section 1.3.

1.1 Problem Statement

Crash severity models, such as various logit or probit models, are widely used in safety analyses. Among those, crash severity models can be categorized into two groups: nominal models and ordinal models. However, few research studies have been conducted on comparing different crash severity models, though each model type has its own unique benefits and limitations. So far, there is no consensus on which type of model performs the best. Some researchers prefer to choose nominal models instead of ordinal models because of the limitation of ordered models, which do not have the flexibility to explicitly control interior category probabilities (Washington et al., 2010). Thus, it is necessary to further compare various crash severity models to allow researchers to select the appropriate one for their analysis.

In terms of model estimation for various crash severity models, sample size requirement is the first item to be considered. Similar to count data models, crash severity model can be influenced by the sample size from which they are estimated (see Lord, 2006). As discussed by Lord and Mannering (2010), crash data are often characterized by small sample sizes, which are attributed to the large cost of assembling crash data. Although it

is anticipated that the size of the sample will influence the performance of crash severity models, nobody has so far quantified how the sample size affects the most commonly used crash severity models and consequently provided guidelines on the data size requirements. A few have proposed such guidelines, but only for crash count models (Lord, 2006; Lord and Miranda-Moreno 2008; Park et.al, 2010). Thus, there is a need to examine how the sample size can influence the development of commonly used crash severity models. Providing this information could help safety analysts in their decision to use one model over another given the size of the data.

Furthermore, even when the sample size is sufficient, there could still exist a bias in the estimated results when the specified model is not based on the true one for crash data. Even though we usually have no knowledge of the true model that crash data come from, there will be less bias in the estimation results if we fit the crash data with a model which is less affected by model misspecification. The model misspecification issue for crash severity models is yet to be studied. Therefore, there is a need to analyze how the model misspecification affects on the commonly used crash severity models.

Finally, both the crash count and crash severity models are usually based on police reported crash data and are used for investigating crash occurrences that are related to highway design features, environmental conditions and traffic flow. However, many crashes often go unreported, particularly those associated with lower severity crashes. This results in underreported crash data, which can yield biased estimation when used to predict the probability of crash severity (Hauer and Hakkert, 1989). In other words, for estimating crash severity models, inferences about a population of interest will assuredly be biased if crash data are treated as a random sample coming from the population without considering the different unreported rates of each crash severity level. There are numerous studies that have investigated the level of incompleteness of police crash surveillance systems. However, few studies have deeply investigated underreporting issues in the analysis of traffic crash models (Kumara and Chin, 2005; Ma, 2009;

Yamamoto et al., 2008). Thus, it is necessary to explore the effects of underreporting crash data on the commonly used crash severity models.

1.2 Research Objectives

The primary goal of this research is to study the difference of three commonly used traffic crash severity models: the MNL, OP and ML models, in terms of sample size, model misspecification, and underreporting in crash data. To accomplish this goal, the following objectives are planned to be addressed in this research:

1. Examine the effects of the sample size on the three most commonly used crash severity models, via a Monte-Carlo approach using simulated and observed crash data. The sample size requirements for the three models will be proposed after comparing the bias and variability caused by the small sample size of the data.

2. Compare the bias and variability caused by model misspecification among the three commonly used crash severity models, via a Monte-Carlo approach using simulated data. The results will provide additional information about the differences between these models to help researchers select the appropriate crash severity model to minimize the bias and reduce the variability caused by a model misspecification.

3. Investigate how each of the three commonly used models behaves in terms of the prediction results for various underreporting scenarios. In addition, the outcome-based sampling method (the Weighted Exogenous Sample Maximum Likelihood Estimator, i.e., WESMLE is to be used in the study) will be quantified for crash severity models on how much it can account for specific underreporting conditions with different knowledge of unreported rates. A Monte-Carlo approach based on simulated and observed data will be used in the analysis. Eventually, recommendations will be developed for implementing the three crash severity

models when the full or partial information about the unreported rates for each severity level is known.

1.3 Outline of the Dissertation

The remainder of this dissertation is organized as follows:

Chapter II provides a brief overview of various crash severity models that have been proposed for modeling crash data, and discusses the selection of three models for analysis in this research: the MNL, OP and ML models. Previous research studies related to the comparison of crash severity models are summarized, and the findings indicate that more model comparisons need to be demonstrated in order to provide more information on model selection for crash severity analyses. In addition, the underreporting issue in crash data and crash prediction models are described. In the last section of this chapter, the model estimation methods for underreported crash data are presented.

Chapter III provides the methodologies for the three target crash severity models: MNL, OP, and ML models, as well as the WESMLE used in the model estimation for the crash severity models. The WESMLE method treats crash data as outcome-based samples rather than random samples to take account of the underreporting issue in crash data.

Chapter IV applies the three crash severity models: the MNL, OP and ML models to observed crash data. The crash dataset includes 26,175 single-vehicle crashes involving fixed objects on rural two-way highways. Meanwhile, the estimation results of the three models are compared. Furthermore, the estimation results in this chapter are treated as baselines for the analysis of sample size (in Chapter V) and the underreporting issue (in Chapter VII), since the crash data used in this chapter are assumed as a complete dataset without any underreporting.

Chapter V examines the sample size requirements for the three models: the MNL, OP, and ML models. Using a Monte-Carlo approach based on simulated and observed crash data, the bias and variability caused by small sample sizes of the data for the three models are evaluated. At the end of this chapter, recommended sample sizes for applying the three models are given.

Chapter VI compares the bias and variability caused by model misspecification among the three models (the MNL, OP and ML models) when sufficient sample sizes are used, based on the results obtained from Chapter V (which are assumed as baselines for analysis). A Monte-Carlo approach using simulated data is utilized for analysis. At the end of this chapter, the effects of model misspecification on the three crash severity models are described.

Chapter VII investigates the effects of underreporting in data for the three crash severity models: the MNL, OP, and ML models, via a Monte-Carlo approach using simulated and observed crash data. More specifically, this chapter explores how each of these models performs for various unreported rates of crash severity levels. Furthermore, it quantifies how much the WESMLE method could account for specific underreporting conditions when transportation safety analysts have the full or partial knowledge about the unreported rates for each severity level.

Chapter VIII summarizes the major results found in this research along with the general conclusions and recommendations for future research.

CHAPTER II

BACKGROUND

Motor vehicle crashes are usually categorized into five crash injury severity categories in decreasing order of levels of injury severity, in "KABCO" scale: (K) fatal injury, (A) incapacitating injury, (B) non-incapacitating injury, (C) possible injury, and (O) no injury (NHTSA, 2010). The American National Standard ANSI D16.1-2007 (2007) provides the definition of the five injury levels. A fatal injury is any injury that results in death. An incapacitating injury refers to any injury, other than a fatal injury, which prevents the injured person from walking, driving or normally continuing the activities the person was capable of performing before the injury occurred. A non-incapacitating injury is any injury, other than a fatal injury or an incapacitating injury, which is evident to observers at the scene of the crash in which the injury occurred. A possible injury is any injury reported or claimed which is not a fatal injury, incapacitating injury or non-incapacitating evident injury. A no injury also refers to property-damage-only (PDO).

However, not all crashes satisfying the above definition of motor vehicle crashes are reportable. Reportable crashes should involve injury or significant property damage (Hauer, 2006). A reportable PDO crash includes damages to a vehicle that are estimated to be at least some monetary value specified by law, which is changeable from time to time, country to country, and state to state. For example, in Wisconsin (in 2003) the damage had to exceed $1000 or more to property owned by any person or $200 to government property. In Pennsylvania, a PDO crash is reportable if one of the involved vehicles cannot be driven from the scene of the collision under its own power (Hauer, 2006). All the "underreporting" mentioned in this research refers to the reportable crashes that go unreported.

Many crash severity models are used to predict the probability of each crash severity level once a crash occurs, based on crash data reported by police. In this chapter, various crash severity models that have been used for modeling traffic safety are reviewed, and model comparisons from previous research studies are summarized. In addition, the underreporting issue for crash data and crash prediction models are described, following which the model estimation methods for underreported crash data are presented.

The chapter is divided into four sections. Section 2.1 provides a brief review of crash severity models which have been used so far to show how widely these models are used in crash analysis. Section 2.2 describes the underreporting issue in both crash count models and crash severity models. Section 2.3 presents the model estimation methods for underreported crash data where crash data are treated as outcome-based samples rather than random samples. Section 2.4 summarizes the chapter.

2.1 Crash Severity Models

Discrete outcome models (usually named as discrete choice models in the previous research) in traffic safety study are usually used to explore the relationship between crash severity and its contributing factors such as driver characteristics, vehicle characteristics, roadway conditions, and road-environment factors. Among others, the main discrete outcome models for crash severity analysis can be categorized into two groups: nominal models and ordinal models. Nominal discrete outcome models used for crash severity do not account for the ordinal nature of the severity level of each crash, but relax the limitation of the ordinal discrete outcome models (it will be stated later). Nominal models are widely used for crash severity analysis, including binomial/multinomial logit (MNL) models, nested logit models, mixed logit (ML) models and mixed probit models. In contrast to nominal models, ordinal discrete outcome models account for the ordering of injury-severity outcomes, including ordered logit models, ordered probit (OP) models and ordered mixed logit models. Differences

between these models are a result of different assumed distributions of unobserved factors (error terms) which can have an impact on model prediction results. For instance, the error term of all logit models is identically and independently distributed (IID) Type I extreme value (such as those for the MNL or nested logit models), or can be decomposed into a part that is IID Type I extreme value (such as those for the ML or ordered mixed logit models). Meanwhile, the error term of all probit models is joint normal distributed (such as those for the OP model), or can be decomposed into a part that follows normal distribution (such as those for the mixed probit models).

Table 2.1 provides a list of these models utilized in previous crash-severity studies to show how widely they have been used in crash data analysis. From this table, it can be

TABLE 2.1 Summary of Previous Research on Crash Severity Models

Model Type	Previous Research
Nominal model	
Multinomial Logit Model (MNL)	Lui et al., 1988; Hilakivi et al., 1989; Shibata and Fukuda, 1993; Mannering and Grodsky, 1995; Farmer et al, 1996; James and Kim, 1996; Mercier et al., 1997; McGinnis et al., 1999; Krull et al., 2000; Ossenbruggen et al. 2001; Al-Ghamdi, 2002; Bedard et al, 2002; Dissanayake and Lu, 2002; Toy and Hammitt, 2003; Ulfarsson and Mannering, 2004; Khorashadi et al., 2005; Conroy et al., 2006; Savolainen and Mannering, 2007; Kim et al., 2008
Nested Logit Model	Nassar et al., 1994; Shankar et al., 1996; Chang and Mannering, 1999; Savolainen and Mannering, 2007; Haleem and Abdel-Aty, 2010
Mixed logit model (ML)	Gkritza and Mannering, 2008; Milton et al., 2008; Pai et al., 2009; Kim et al.,2010
Ordinal model	
Ordered logit model	O'Donnell and Connor, 1996; Wang and Kockelman, 2005
Ordered probit model (OP)	Hutchinson, 1986; O'Donnell and Connor, 1996; Klop, 1998; Duncan et al., 1999; Khattak, 1999; Renski et al., 1999; Kockelman and Kweon, 2002; Quddus et al., 2002; Abdel-Aty, 2003; Zajac and Ivan, 2003; Abdel-Aty and Keller, 2005; Garder, 2006; Ma and Kockelman, 2006; Xie et al., 2009; Haleem, 2010
Ordered mixed logit model	Srinivasan, 2002; Eluru et al, 2008

seen that the MNL and OP models are the most prominent types of models used for traffic crash severity analysis. Meanwhile, the ML model is a promising model that has recently been used. Details of each model listed in the table are then discussed further.

2.1.1 Nominal model

This section describes the three nominal discrete outcome models used for crash severity studies.

2.1.1.1 Multinomial Logit Model (MNL)

Table 2.1 indicates that the MNL model is the most widely used one for traffic crash severity studies by many transportation safety researchers. A variation of the MNL model refers to as the multivariate multinomial logit model has also been used for modeling crash severity (Ulfarsson and Mannering, 2004).

The MNL model is derived under the assumption that the unobserved factors are uncorrelated over the alternatives (or outcomes) and have the same variance for all alternatives, also known as the independence from irrelevant alternatives (IIA) assumption. This assumption is the most notable limitation of the MNL model since unobserved factors related to one alternative could be similar to those related to another alternative (Train, 2003). For example, level of alertness may not be a variable included when modeling crash severity and would be considered as an unobserved factor. However, a sleepy driver involved in a fatal crash might have a similar probability of being involved in an incapacitating injury crash; if so, the unobserved factors (level of alertness) affecting fatal crash and incapacitating injury crash would be correlated rather than independent. Despite this limitation, the IIA assumption makes the MNL model very convenient to use which also explains its popularity.

2.1.1.2 Nested Logit Model

In order to release the IIA assumption of the MNL model and to account for the correlation of unobserved factors over alternatives, nested logit models have been used for traffic crash severity analysis.

The nested logit model belongs to the family of Generalized Extreme Value (GEV) models, which allow partial relaxation of the IIA property since IIA holds within nests but not across nests (i.e. unobserved factors have the same correlation for all alternatives within a nest but no correlation for alternatives in different nests). The nested logit model structure is useful when there are similarities between some alternatives. In the previous mentioned sleepy driver example, fatal crashes and incapacitating crashes could be grouped into a nest to account for the possible correlation among unobserved factors. The nested logit model will collapse to the MNL model when the existing correlation is zero.

2.1.1.3 Mixed logit model (ML)

The ML model has attracted considerable attention by traffic safety researchers because of its flexibility in model definition, allowing the unobserved factors to follow any distribution. Thus, the ML model overcomes the IIA limitation of the MNL model. In addition, Train (2003) illustrated that the ML was fully general which could approximate any discrete outcome model. From Table 2.1, it is found that the ML model has been widely used in traffic crash severity analysis in recent years. It has become popular due to the improvement of computer speed and the development of simulation techniques which are necessary for the model estimation.

The ML model can approximate any discrete outcome model since it allows the unobserved factors to follow any specified distribution. It is obtained by decomposing the unobserved factors into two parts, one containing all the correlation by following any

distribution and the other following the IID Type I extreme value distribution (Train, 2003).

2.1.2 Ordinal model

The crash severity categories do have an inherent order, with PDO being the least severe and fatal being the most severe. Some information is to be lost by ignoring the ordinal nature of the five crash severities. As stated by Amemiya (1985) that if ordinal data were fit by a nominal model, the estimated parameters remained consistent but less of efficiency. However, although ordered logit/probit models include the order information of the data, they restrict the effect of explanatory variables on ordered discrete outcome probabilities by using the same coefficient of an explanatory variable among different crash severities. Therefore, it causes the variable either to increase the probability of higher severities with the decrease of the probability of lower severities, or to reduce the probability of higher severities with the increase of the probability of lower severities (this will be described further in Section 3.2). This result may not be realistic because it is possible that some explanatory variables cause an increase in the probability of some outcomes predictions but a decrease in the probability of other outcomes predictions. For instance, inclement weather might lead to an increase in the probability of both highest severity (or severities) and lowest severity (or severities) and reduce the probability of all other severities. Meanwhile, ordered logit/probit models do not have the flexibility to explicitly control interior category probabilities (Washington et al., 2010). The effects on interior categories depend on the thresholds.

The three most commonly used ordinal discrete outcome models in traffic safety studies are ordered logit model, ordered probit (OP) model, and ordered mixed logit model. They are described below.

2.1.2.1 Ordered logit model

The ordered logit model is derived from specifying the MNL model for ordinal data. Standard ordered logit models are not frequently used in crash severity analysis, when compared to, the use of other ordinal models. O'Donnell and Connor (1996) used the ordered logit model to predict the severity of motor vehicle crash injuries. However, there are a couple of variations of ordered logit model used by researchers for crash severity studies. Wang and Kockelman (2005) used the heteroscedastic ordered logit model for studying the contributing factors of occupant injury severity, which parameterizes the variance of the error term as a function of some variables such as speed limit, vehicle type and vehicle curb weight. Wang and Abdel-Aty (2008) used a partial proportional odds model to examine the left-turn crash injury severity, for which some of the coefficients can differ across outcomes, with the relaxation of the parallel-lines assumption for some variables. Eluru et al. (2008) and Yamamoto et. al. (2008) used sequential logit models for crash severity analysis, which relaxed the restrictions imposed by standard ordered logit model allowing separate parameter coefficients for variables.

2.1.2.2 Ordered probit model (OP)

The OP model is derived from specifying the standard probit model for ordinal data. Although the standard probit model is not used for crash severity analysis because of its estimation difficulties, the OP model has been widely used in crash severity analysis over the ordered logit model (as shown in Table 2.1) due to its underlying assumption of normality. A variation of the OP model (a.k.a., bivariate ordered probit model) was also used a few years ago to predict the probability of driver's and passenger's injury severities in collisions with fixed objects (Yamamoto and Shankar, 2000).

2.1.2.3 Ordered mixed logit model

Similar to other ordinal discrete outcome models, the ordered mixed logit model is derived from specifying the ML model for ordinal data. Ordered mixed logit model generalizes the ordered logit model by allowing parameters of each variable for the model to follow any distribution. Srinivasan (2002) extends the ordered logit model by allowing random coefficients of variables in each severity level, and by a Chi-square test, ordered mixed logit model had a better goodness-of-fit (GOF) than ordered logit model. Eluru et al. (2008) developed an ordered mixed logit model to find the contributing factors to the non-motorist injury severity.

2.1.3 Model comparison

Each model type has its own unique benefits and limitations and there is no consensus on which model is the best. Some researchers prefer to choose nominal models instead of ordinal models because of the limitation of ordinal models, which restrict the effect of variables across outcomes (Khorashadi et al., 2005).

From the literature, few researchers have directly examined different types of crash severity models. Abdel-Aty (Abdel-Aty, 2003), as one of the few researchers, compared the MNL, OP and nested logit models in driver injury severity analysis for roadway sections, signalized intersections and toll plazas. The author concluded that the OP model was easy to estimate and performed well in modeling driver injury severity. He also stated that the MNL model did not perform as well as the OP model, which resulted in fewer significant variables for the model and a lower GOF. The nested logit model had fewer significant variables than the OP model, but shared a slight improvement in the GOF. However, the nested logit was more complex than the OP model because the former requires the specification of a nested structure. Thus, he recommended the OP for modeling driver injury severities.

Haleem et al. (2010) compared the binary probit model, OP model and nested logit model in crash severity analysis at three- and four-legged unsignalized intersections in Florida from 2003 to 2006. By comparing the GOF of model estimation, the authors concluded that the binary probit model produces comparable if not better results than the OP model, and meanwhile it was found that the nested logit models did not show any improvement over the probit models. As a result, it was recommended that binary probit models be used for modeling crash severity at unsignalized intersections if the objective is to identify the factors contributing to severe injuries in general. However, it was also noted that binary probit models were not suitable for the analysis of a specific injury category since the crash severity levels were combined into two levels for binary probit models.

A review of the literature reveals that more extensive model comparisons need to be conducted in order to provide more information on model selection in crash severity analysis. This task will be undertaken in this research, particularly for the purpose of analyzing sample size requirements, the effects of model misspecification and underreporting data for crash severity models.

2.2 Underreporting Issue in Crash Data

About twenty years ago, Hauer and Hakkert (1989) raised the issue that not all traffic crashes were reportable and not all reportable crashes were in fact reported[1]. This limited

1 There is another possible issue about the accuracy of crash data that some crashes were incorrectly recorded for their severity levels. For instance, a crash of possible injury (C) was very likely to be categorized as no injury (PDO), and vice-verse. As stated by Winston et al. (2006), reporting errors of misclassifying the severity level of crashes would lead to biased parameter estimates. However, by applying the procedure developed by Hausman et al. (1998) to explore the effects of misclassification of crash severity levels on the model estimation results, they found their data were not subject to systematic misclassification of crash severities. More research studies need to be done on the misclassification issue in crash data.

the ability to manage road safety, since most of the analysis related to road safety was based on reported crashes. Analysis of underreported crash data would lead to a biased estimation of crash severity and thus result in ineffective treatments. Having realized the underreporting issue in crash data, some researchers began to study this topic in a greater detail (Hauer and Hakkert, 1989; James, 1991; Hvoslef, 1994; Stutts and Hunter, 1998; Aptel et al., 1999; Elvik and Mysen, 1999; Alsop and Langley, 2001; Cryer et al., 2001; Dhillon et al., 2001; Rosman, 2001; Amoros et al., 2006; Hauer, 2006; Tsui et al., 2009). These studies reveal that crashes are underreported in all the countries with high levels of motorization. In addition, underreporting issue in crash data is worse in developing countries. The probability of reporting was found to be influenced by crash severity, age of the victim, role of the victim (whether the victim is the driver, the passenger, or etc.), and number of vehicles involved (Hauer and Hakkert, 1989).

Underreported data tend to produce biased estimations in both the crash count models and crash severity models. However, underreporting is more critical in crash severity models because of the different reported rates of various severity categories. Crashes with lower severity such as PDO are more likely to be unreported which leads to the over-representation of crashes with higher severity and under-representation of crashes with lower severity. It is widely accepted that fatal crashes have the highest reported rate and PDO crashes have the lowest reported rate. After reviewing 18 studies in which researchers examined police, hospital and insurance sources for common entries, Hauer and Hakkert (1989) concluded that the unreported rate was 5 percent for fatal injuries, 20 percent for injuries requiring hospitalization, and perhaps 50 percent for all injuries. In a comprehensive meta-analysis, based on 49 studies in 13 countries, Elvik and Myssen (1999) found that 5 percent for fatal injuries, 30 percent for serious injuries, 75 percent for slight injuries, and 90 percent for very slight injuries went unreported. According to Blincoe et al. (2002), up to 25 percent of all minor injuries and almost 50 percent of PDO crashes likely went unreported because most drivers did not wish to involve the police or other authorities due to insurance concerns or legal repercussions; by contrast, fatal injuries were completely reported.

Several studies have investigated the effects of underreporting of crash data in both crash count models and crash severity models, as discussed in Section 2.2.1 and Section 2.2.2 respectively. As a result, some approaches which were developed for outcome-based samples in discrete outcome models could be used to account for underreporting of crash data in traditional crash model studies, which will be described in Section 2.3.

2.2.1 Underreporting in crash count models

Kumara and Chin (2005) pointed out that a reliable crash data system was essential to develop a good representative model; however, the underreporting issue in crash studies might obscure true information about crashes. In their study, the standard Poisson regression model was modified into a Poisson underreporting model, which improved the quality of parameter estimation, while taking into account the effect of underreporting.

Realizing that underreporting in crash data invalidates the assumptions of the standard Poisson regression model, Ma (2009) presented an approach to model underreported traffic crash data with a modified latent Poisson regression model, which was a simple form of the combination of a standard Poisson and a beta distribution. The modified latent Poisson regression model allowed for the fact that the crash data for model estimation might be underreported which differed from previous specifications of count data models. The underreporting Poisson regression model results were found to differ substantially from standard Poisson models that did not consider underreporting.

The two studies discussed here did not verify their estimated results with actual crash data. Model validation is difficult since information on actual crash occurrences and the unreported rates for various crash severities is unknown.

2.2.2 Underreporting in crash severity models

The inconsistent unreported rates among different severity levels lead to biased results, which cause the overestimation of probabilities of higher severity crashes and the underestimation of lower severity crashes, particularly PDO. In addition, underreporting causes biased parameters, which skew the inferences on the effects of key explanatory variables for the prediction models. However, only one study has been found that deals with modeling crash severity with underreported data, as described below.

Yamamoto et al. (2008) investigated the effects of underreporting on parameter estimation for the OP and sequential binary probit models. In the study, the results indicated that the estimates of the explanatory variables and parameter elasticities for both models could be significantly biased if underreporting was not considered. In addition, the researchers regarded traffic crash data as outcome-based samples with unknown population shares of the crash severities, and used a pseudo-likelihood function (Cosslett, 1981a, b) to account for the effects of underreporting on parameter estimation for both models. The population shares of each severity category were estimated for each model, which provided insights on the levels of underreporting in each crash severity. However, the effectiveness and efficiency of the methods were not confirmed, and no analysis was conducted about the model effects of different combinations of unreported rates of each crash severity.

2.3 Model Estimation Methods for Underreported Crash Data

No matter which type of discrete outcome model is used for crash analysis, crash severity models are usually estimated based on random sampling without considering the underreporting in crash data. However, because of the unique underreporting characteristics in crash data, crash data should be treated as outcome-based or endogenous stratified samples. Without considering the underreporting issue for the model, model estimation results will definitely be biased (Yamamoto et al., 2008).

Though it is rare to treat crash data as outcome-based samples, outcome-based samples (a.k.a., choice-based samples) are commonly used in other areas of research. Choice-based samples are usually collected by stratifying to obtain better information about alternatives that are infrequently chosen in the population when a random sampling does not collect enough samples for effective statistical analysis. For example, an analyst may want to analyze a product with a small market share by over-sampling users, while collecting sufficient data with a simple random sampling may require a prohibitively large sample size (Bierlaire et al., 2008).

2.3.1 Sampling strategies

To better understand the outcome-based samples of crash data, different sampling strategies are overviewed. As stated by Bierlaire et al. (2008) and Pendyala (1993), sampling strategies can be classified into two major categories: random sampling and stratified sampling.

A simple random sampling, which is the most frequently used method of sampling, draws independent observations from the population and the probability of being sampled is equal for every unit. A random sample may contain few individuals having certain selection and thus lead to poor estimates of the relevant parameters, unless the sample size is large (Cosslett, 1981b).

A stratified sampling requires dividing the population into groups according to a set of measurable variables and draws a random selection of participants at different rates from each group, called a "stratum". Over-sampling those individuals who select infrequently chosen alternatives allows for precise estimates than could be obtained from a random sample of the same overall sample size (Cosslett, 1981b).

Stratified sampling can be exogenous or endogenous or both. Each of these is described below:

- An exogenous stratified sampling draws independent observations from the groups which are characterized only by exogenous variables or independent variable.
- An endogenous stratified sampling, also called choice-based or outcome-based sampling, draws independent observations from the groups which are characterized by the endogenous variables or dependent variable only.
- An exogenous and endogenous stratified sampling draws independent observations from the groups which are characterized by both the exogenous and endogenous variables.

As mentioned above, crash data for model estimation consists of outcome-based samples since the unreported rates differ according to the crash severity category. As a result, this study is interested in the estimation of discrete outcome models for outcome-based sampling.

The estimation of discrete outcome models is a very difficult task when the sampling strategy is based on the outcome, since the covariate distribution entering the likelihood function cannot be factored out. As stated by Lancaster (1997), unlike the random sampling with the covariate distribution factor out of the likelihood, the likelihood function of outcome-based sampling involves both an unknown finite parameter and an unknown distribution function of the covariates, so the likelihood function does not factor out and the inference problem is semi-parametric.

2.3.2 Methods for treating outcome-based sampling in discrete outcome models

There are several methods that have been developed by economists since 1977 to handle outcome-based samples as stated below.

Seven main methods are in use to handle outcome-based samples in discrete outcome models: Weighted Exogenous Sample Maximum Likelihood Estimator (WESMLE), Conditional Maximum Likelihood Estimator, Full information Maximum Likelihood

Estimator, Weighted Generalized Method of Moments, Bayesian Weighted Exogenous Sample Maximum Likelihood Estimator, Smoothed Maximum Likelihood Estimator, and Weighted Conditional Maximum Likelihood Estimator. One of the methods will be selected for handling the underreported crash data, by treating underreported crash data as outcome-based samples for crash severity models.

2.3.2.1 Weighted Exogenous Sample Maximum Likelihood Estimator (WESMLE)

As long ago as 1977, Manski and Lerman (1977) demonstrated that the maximum likelihood estimator (MLE) is generally asymptotically biased and inconsistent when applied to outcome-based samples (Xie, 1989). As a correction, they developed the WESMLE as a simple method to yield consistent estimates for outcome-based sampling.

The WESMLE is consistent and asymptotically normal, and can be computed easily by modifying the existing maximum likelihood function, but is not fully efficient. The WESMLE is the most widely used estimator, available when the population shares are known. The WESMLE for outcome-based samples is not efficient since its variance-covariance matrices do not asymptotically attain the Cramer-Rao lower bound. However, its efficiency loss relative to more difficult estimation is typically modest which has been supported by its wide use among all the methods. Since the WESMLE was proposed in 1977, various other estimators have appeared which are summarized as below (Xie, 1989).

2.3.2.2 Conditional Maximum Likelihood Estimator

Manski and McFadden (1981) introduced a conditional maximum likelihood method, treating more generally the problems of sample design and estimation of discrete choice models. However, this method also has efficiency problems (Cosslett, 1981b), although Amemiya and Vuong (1987) noted that it is asymptotically more efficient than the WESMLE.

2.3.2.3 Full information Maximum Likelihood Estimator

Cosslett (1981a) proposed a full information maximum likelihood estimator by involving the population proportion of outcomes in the likelihood function for parameters estimation under choice-based or outcome-based sampling. Cosslett's method is efficient, but is very difficult to compute as it may require estimating the joint distribution of the exogenous variables, and it has only been applied by Hsieh et al. (1985). Furthermore, in Cosslett's method, the knowledge of population proportion greatly improves the precision of the estimates, and estimators will have low efficiency without the knowledge of population proportion. Therefore, using Cosslett's method to estimate parameters and unreported rates for crash severity models, as developed by Yomamoto (2008), may not give an efficient and valid result.

2.3.2.4 Weighted Generalized Method of Moments

More recently, Imbens (1992) developed an efficient weighted generalized method of moments (GMM) estimator for choice-based samples. Butler (2000) compared the variances of the WESMLE and weighted generalized GMM estimation and concluded that the latter could be more efficient than the former. However, both methods were less efficient than the full information MLE (Cosslett's method) under choice-based sampling. In addition, simpler to compute than Cosslett's, Imben's estimator cannot be computed as simply as the WESMLE.

2.3.2.5 Bayesian Weighted Exogenous Sample Maximum Likelihood Estimator

Lancaster (1997) provided a Bayesian interpretation of the WESMLE for choice-based samples using a binary probit model. It was concluded that the knowledge of the marginal choice probabilities largely affected the precision of model estimation. However, this conclusion might not be generalized to multinomial choice or models other than probit.

2.3.2.6 Smoothed Maximum Likelihood Estimator

Cosslett (2007) proposed a smoothed likelihood function to construct an efficient estimator for some semi-parametric models that contained unknown density functions together with parametric index functions. A binary choice model from a choice-based sample was estimated using this method to show the efficiency gains from knowledge of population shares.

2.3.2.7 Weighted Conditional Maximum Likelihood Estimator

Bierlaire et al. (2008) proposed the weighted conditional maximum likelihood (WCML) estimator. The WCML is a generalization of the conditional MLE by Manski and McFadden (1981) and the WESMLE by Manski and Lerman (1977).

2.3.2.8 Method Selected for Crash Severity Models

Among all the methods mentioned previously, though not efficient, the WESMLE is consistent and easy to compute which makes it the most widely used method for choice-based samples. LIMDEP (on LIMDEP 9.0, Econometric Modeling Guide, 2007), a software program for estimating discrete outcome models also uses the WESMLE method to account for the outcome-based sampling. In this research, the WESMLE method will be used to estimate the three crash severity models for underreported crash data.

2.4 Chapter Summary

This chapter provided a review of crash severity models that have been used in traffic safety analysis. It was found that the MNL and OP models were the most prominent ones used for traffic crash severity analysis. Meanwhile, the ML is a promising model that has recently been widely used in many areas. Therefore, the three crash severity models: the MNL, OP and ML models were selected for analysis in this research.

Furthermore, each model type has its own unique benefits and limitations, but few research studies have been conducted on comparing different crash severity models. Thus, further model comparisons need to be demonstrated in order to provide more information on model selection in crash severity analysis. In this research, the comparisons among the three crash severity models (the MNL, OP and ML models) are developed, particularly for the purpose of analyzing sample size requirements, the effects of model misspecification and underreporting in crash data for crash severity models.

In terms of the crash data used in the traffic safety studies, such as crash count models and crash severity models, those data are usually based on the police reported crashes. However, not all traffic crashes are reportable and not all reportable crashes are in fact reported. Thus, underreported crash data used for modeling crashes tend to produce biased estimations in both crash count models and crash severity models. In addition, underreporting produces a more critical issue for crash severity models. Reported rates of various severity categories differ that crashes with lower severity are more likely to be unreported than those with higher severity. Therefore, the probability of higher severity crashes such as fatal crashes are probably overestimated and on the other hand the probability of lower severity crashes particularly PDO usually are underestimated. Therefore, crash data should be treated as outcome-based samples rather than random samples from the population, and the WESMLE method was selected to treat the underreported crash data for all three crash severity models since it is consistent and easy to compute.

Before discussing the effects of underreported data on three crash severity models (the MNL, OP and ML models) in Chapter VII, two basic issues in the estimation of crash severity models are demonstrated: sample size requirements and model misspecification effects on the three crash severity models, in Chapter V and VI respectively. The next chapter (Chapter III) describes the methodologies for the three crash severity models, and the WESMLE method for outcome-based sampling in crash severity models.

CHAPTER III

METHODOLOGY

This chapter provides the methodologies by which the three crash severity models (the MNL, OP and ML models) are used to predict the probabilities associated with each crash severity level. In addition, the WESMLE method is introduced in the chapter, which is used in the model estimation for the crash severity models, treating crash data as outcome-based samples rather than random samples to take account of the underreporting issue in crash data.

This chapter contains five sections. Sections 3.1, 3.2, and 3.3 briefly discuss the structure and estimation method of the MNL, OP and ML models respectively. In Section 3.4, WESMLE is described with the comparison of the traditional model estimation method in crash severity models, i.e. Maximum Likelihood Estimator (MLE). Finally, Section 3.5 summarizes the chapter.

3.1 Multinomial Logit Model (MNL)

In the general case of a MNL model for crash severity outcomes, the propensity of crash i towards severity category k is represented by severity propensity function, T_{ki}, as shown in Equation (3.1) (Khorashadi et.al, 2005).

$$T_{ki} = \alpha_k + \beta_k X_{ki} + \varepsilon_{ki} \tag{3.1}$$

Where,

α_k is a constant parameter for crash severity category k; β_k is a vector of the estimable parameters for crash severity category k; k=1, ⋯, C (C=5 in this research), representing all the five severity levels as KABCO;

X_{ki} represents explanatory variables affecting the crash severity for i^{th} individual crash at severity category k (geometric variables, environmental conditions, driver characteristics, etc.);

ε_{ki} is a random error term following the Type I generalized extreme value (i.e., Gumbel) distribution; and,

$i = 1, \cdots, n$ n is the total number of crash events included in the model.

Equation (3.2) shows the formula used to calculate the probability of each crash severity type. Let $P_i(k)$ as the probability of i^{th} individual crash ending in crash severity category k, such that

$$P_i(k) = \frac{\exp(\alpha_k + \beta_k X_{ki})}{\sum_{\forall k} \exp(\alpha_k + \beta_k X_{ki})} \tag{3.2}$$

For the MNL model, the unobserved effects associated with each severity category are independent from each other, which are evident by the fact that β_k varies for different crash severity categories. Another important feature of the MNL model is that it could not account for any correlation among unobserved effects. However, the MNL model is relatively easy to calculate and implement. The MLE method is most commonly used for parameter estimation for the MNL model. The likelihood function of the data is shown in Equation (3.3).

$$L(\beta \mid y) = \prod_{i=1}^{n} \prod_{k=1}^{C} [P_i(k)]^{I(y_i = k)} \tag{3.3}$$

Where,

$I(y_i = k)$ is an indicator function. When $y_i = k$, $I(y_i = k)$=1, otherwise, $I(y_i = k)$ =0.

Since the probabilities for the crash severity categories must satisfy $\sum_{\forall k} P_i(k) = 1$, the equations for the probabilities of five crash severity categories as shown in Equation (3.2) are mutually consistent and one of them is redundant. Thus, we only need to specify four P(k) functions, the other category is set as the baseline. The parameters are normalized to be zero in Equation (3.2) for whichever category is the baseline. In addition, as a fundamental property of the MNL model: the equivalent differences property (Koppelman and Bhat, 2006), the choice of baseline category is arbitrary and does not make any difference for the estimation results. Usually the baseline category is chosen in a manner that facilitates interpretation of the data.

For the MNL model, elasticities are usually calculated to measure the magnitude of the impacts of a variable on the probability of crash severities. The elasticity is computed using Equation (3.4) (Savolainen and Mannering, 2007).

$$E_{x_{kj}}^{P(j)} = \partial \ln P(j) \Big/ \partial \ln x_{kj} = [1 - P(j)]\beta_{kj} x_{kj} \tag{3.4}$$

Where,

β_{kj} is the estimated coefficient of variable x_{kj}.

Elasticity values show how a percentage change in an explanatory variable will affect the probability of each crash severity, and cannot be interpreted for indicator variables that take on values of zero or one. For such indicator variables, pseudo-elasticity is used which gives the average probability change due to a variable change from zero to one.

The pseudo-elasticity of the indicator variable x_j for severity k is given as shown in Equation (3.5) (Yamamoto et al., 2008).

$$E_{x_j}^{P(k)} = \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \frac{P(\mathrm{k} \mid x_{i1,} \cdots, x_{i,\mathrm{j}-1}, 1, x_{i,\mathrm{j}+1}, \cdots, x_{iJ}) - P(\mathrm{k} \mid x_{i1,} \cdots, x_{i,\mathrm{j}-1}, 0, x_{i,\mathrm{j}+1}, \cdots, x_{iJ})}{P(\mathrm{k} \mid x_{i1,} \cdots, x_{i,\mathrm{j}-1}, 0, x_{i,\mathrm{j}+1}, \cdots, x_{iJ})} \quad (3.5)$$

Where,

J is the number of explanatory variables; $\tilde{N}$ is the sample of crashes with $x_{.,\mathrm{j}} = 0$

3.2 Ordered Probit Model (OP)

The OP model uses a latent variable as shown in Equation (3.6) to disaggregate crash severity outcomes instead of using severity propensity function as for the MNL model.

A latent variable z_i is: $z_i = \beta X_i^T + \varepsilon_i$ (3.6)

Where,

$X_i = \{1, x_{i1}, \cdots, x_{ij}, \cdots, x_{im}\}^T$, the input value for the i^{th} individual crash;

x_{ij} is the j^{th} explanatory variable for the i^{th} individual crash;

$\beta = \{\beta_0, \beta_1, \cdots, \beta_j, \cdots, \beta_m\}^T$, the column vector of the coefficients for the explanatory variables;

ε_i is a random error term following standard normal distribution;

$i = 1, \cdots, n$ n is the total number of crash events including in the model;

$j = 1, \cdots, m$ m is the total number of explanatory variables.

The dependent variable y_i is an integer representing crash severity that has C categories (in this research, C=5). Usually, y_i is set as 1 to C, either in an increasing order as OCBAK or in a decreasing order as KABCO. The value of y_i is determined by:

$$y_i = \begin{cases} 1, if \gamma_0 < z_i \le \gamma_1 \\ k, if \gamma_{k-1} < z_i \le \gamma_k \\ C, if \gamma_{C-1} < z_i \le \gamma_C \end{cases} \tag{3.7}$$

Where,

$\gamma = \{\gamma_0, \cdots \gamma_k, \cdots, \gamma_C\}$ are the threshold values for all crash severity categories.

The relationship between these threshold values are subject to the constraint: $-\infty = \gamma_0 < \gamma_1 \le \cdots \le \gamma_k \le \cdots \le \gamma_{C-1} < \gamma_C = +\infty$.

Given the value of x_i, the probability that the crash severity of i^{th} individual crash belongs to each category is

$$\begin{cases} P(y_i = 1) = \Phi(\gamma_1 - X_i^T \beta) \\ P(y_i = k) = \Phi(\gamma_k - X_i^T \beta) - \Phi(\gamma_{k-1} - X_i^T \beta) \\ P(y_i = C) = 1 - \Phi(\gamma_{C-1} - X_i^T \beta) \end{cases} \tag{3.8}$$

Where,

$\Phi(\cdot)$ stands for the cumulative probability function of the standard normal distribution.

From formula (3.8), only C-1 thresholds need to be estimated. The unknown parameters to be estimated are β and γ which are usually determined by the MLE method. The likelihood function of the data is:

$$L(\beta,\gamma \mid y)=\prod_{i=1}^{n}\prod_{k=1}^{C}\left[\Phi(\gamma_k - X_i^T\beta)-\Phi(\gamma_{k-1}-X_i^T\beta)\right]^{I(y_i=k)} \quad (3.9)$$

Where,

$I(y_i=k)$ is an indicator function. When $y_i=k$, $I(y_i=k)$=1, otherwise, $I(y_i=k)$ =0.

The parameters β and γ can be determined by maximizing $L(\beta,\gamma|$y). However, there are several limitations with the MLE method. For example, if the data are not representative of the population then the estimated model may be erroneous, since the parameter estimation results depend completely on the data. In addition, the maximization process is a nonlinear optimization problem, which is not guaranteed to converge to a global optimal solution (Xie et al., 2009).

In terms of the effects of estimated parameters on the probabilities of each crash severity level, only the highest and lowest ordered category (fatal and PDO crashes) have unambiguous effects (Washington et al., 2010). Figure 3.1 illustrates this clearly. In Figure 3.1, the five crash severity levels are categorized in an increasing order as OCBAK. A smaller βX (a positive value of β with a decrease in X or a negative value of β with an increase of X) shows a shift of value $\gamma_k - \beta X$ from left to right, which leads to an unambiguous decrease of the probability of the highest crash level (y=5) and an increase of the lowest level (y=1). However, it is not clear about the corresponding probability changes of the "interior" categories (i.e. y=2, 3, 4) since they could be either decreasing or increasing depending on whether $\gamma_k - \beta X$ is located on the left or right side of the peak. As stated by Washington et al. (2010), the direction of the effects on the interior categories could be attained from their marginal effects.

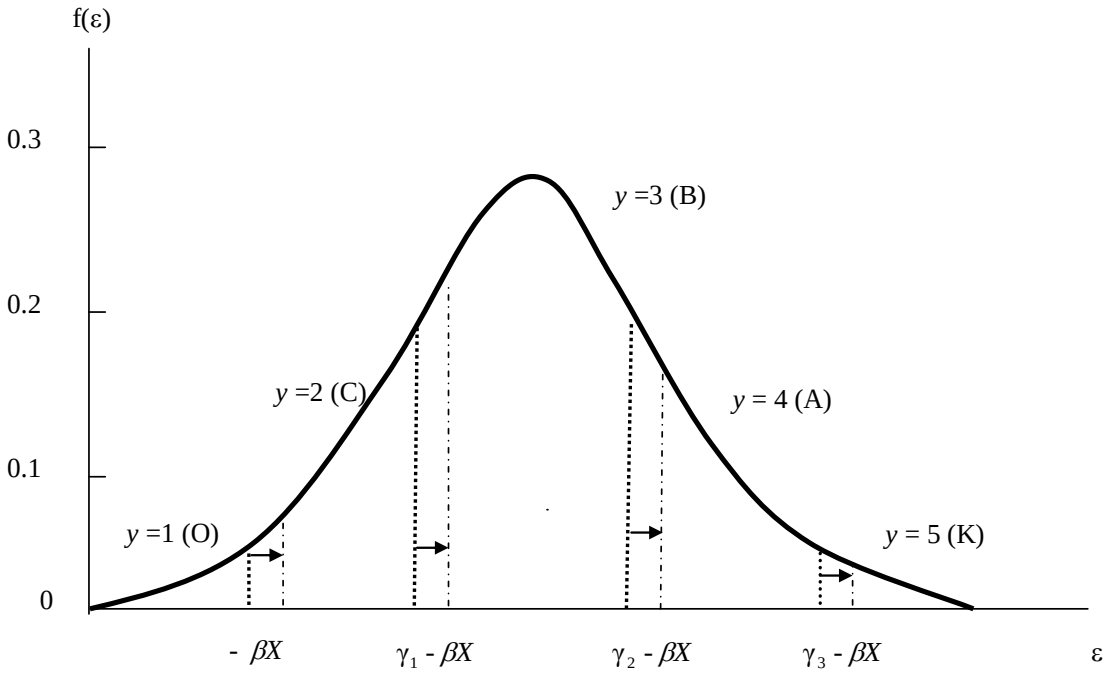

Figure 3.1 An Ordered Probability Model with a Decrease in βX
(Washington et al., 2010)

3.3 Mixed Logit Model (ML)

The ML probabilities are the integrals of standard logit probabilities over a density of parameters (i.e., it is a weighted average of the logit formula evaluated at different value of parameters (β), with the weights given by the density $f(\beta)$).

The ML model shares the same structure of severity propensity function, T_{ki}, as shown in Equation (3.1). Equation (3.10) shows the formula used to calculate the probability of each crash severity level for the ML model.

Let $P_i(k)$ as the probability of i^{th} individual crash ending in crash severity category k, such that:

$$P_i(k) = \int \frac{\exp[\alpha_k + \beta_k X_{ki}]}{\sum_{\forall k} \exp(\alpha_k + \beta_k X_{ki})} f(\beta \mid \theta) d\beta \tag{3.10}$$

Where,

$f(\beta \mid \theta)$ is the density function of β with θ referring to a vector of parameters of the density function (mean and variance).

For the ML model, $f(\beta \mid \theta)$ is usually specified to be continuous. If $f(\beta \mid \theta)$ is discrete, with β taking a finite set of distinct values, then the ML model becomes a latent class model which could be regarded as a simplified version of the ML model. In addition, when $f(\beta \mid \theta)$ is a fixed variable, then the ML model becomes a MNL model. This indicates that the ML model is a generalized one of the MNL model, and thus the ML model has some features of the MNL model such as the equivalent differences property. Usually for the ML model, $f(\beta \mid \theta)$ is considered to use normal, lognormal (which restricts the impact of the estimated parameter to be strictly positive or negative), triangular and uniform distributions.

A simulation-based maximum likelihood method instead of MLE is usually used for parameter estimation for the ML model. The MLE for the ML models is computationally cumbersome due to the required numerical integration of the logit function over the distribution of the random parameters. The most popular simulation approach uses Halton draws, which have been shown to provide a more efficient distribution of draws for numerical integration than purely random draws (Gkritza et al., 2008).

As for the MNL model, elasticities or pseudo-elasticities are usually computed to assess the effect of variables on probabilities of each crash severity level, as shown in Equation (3.4) and (3.5).

3.4 Method for Outcome-based Sampling in Crash Severity Models

The WESMLE is the maximand of the weighted likelihood function where the weights depend upon both the population share of each severity level (the fraction of each severity level in a complete dataset) and the sample share of each severity level (the fraction of each severity level in an underreported dataset). By weighting the observations appropriately, the WESMLE makes the outcome-based samples behave asymptotically as if they were random samples (Xie & Manski, 1989).

The log-likelihood function for a WESMLE, as shown in Equation (3.11), is equivalent to that for the MLE except that each traffic crash is weighted by the ratio of the actual crash severity's population share Q_k, to the sample share H_k which is the severity share in the underreported crash data.

$$\text{Log-likelihood for WESMLE} = \sum_{n=1}^{N} \sum_{k \in C_n} d_{nk} \left(\frac{Q_k}{H_k}\right) \ln P(k \mid x_n, \beta) \qquad (3.11)$$

Where,

N is the number of recorded crashes;

C_n is the set of severity categories from which individual crash *n* belongs to, in the study, C_n =(K,A,B,C,O);

d_{nk} is an indicator variable equal to 1 if individual crash *n* belongs to severity level *k*, and 0 otherwise;

x_n is the vector of contributing factors associated with individual crash *n*;

β is the vector of estimated parameters associated with contributing factors x_n;

$P(\mathrm{k} \mid x_n, \beta)$ is the probability of severity level belonging to k given the contributing factors, x_n, and estimated β.

Another difference between the WESMLE and MLE is the corrected asymptotic covariance matrix for the estimators. For the WESMLE, let H be the Hessian of the (weighted) log-likelihood (i.e., the usual estimator for the variance matrix of the estimators) and let $G'G$ be the summed outer products of the first derivatives of the (weighted) log-likelihood. The covariance matrix of the estimators for the WESMLE is $V = (-H)^{-1} G'G (-H)^{-1}$.

3.5 Chapter Summary

In this chapter, we have provided the fundamental methodologies of three crash severity models (the MNL, OP and ML models) selected for analysis in this research. The methodological approaches for the three models are similar, while the underlying theories of deriving these three models are different. The derivation of the three models from Section 3.1 to 3.3 not only offers the methodological supports for model estimations in the following chapters, but also helps to generate simulated data for three models in the later analysis with the provision of the structures of the three models.

As described in Section 3.1, the MNL model is derived under the assumption that error term (or unobserved factor) ε_{ki} in the propensity function for each severity level is independent and identically distributed (IID) extreme value. The critical assumption of the MNL model is that the error terms are uncorrelated and having the same variance over the severity levels which is the largest limitation of the MNL model, though resulting in convenient computations for the model. As discussed in Section 3.3, the ML model relaxes the limitation of MNL model by allowing the parameters of individual variables to be variant over observations with any types of distribution. However,

introducing random parameters to the ML model makes the model estimation complicated. Both of the MNL model and ML model are nominal, while the derivation of the OP model (ordinal model) was provided in Section 3.2. The main differences between the OP model and two logit models (the MNL and ML models) are at the error terms and parameters. For the OP model, the error terms are standard normal distributed, and the parameter of each variable is fixed to be the same across crash severity levels which leads to the loss of flexibility to explicitly control interior category probabilities (for crash severity A,B, and C in this research). As stated by Washington et al. (2010), though the OP model recognizes the ordering of the crash severities, it loses the flexibility in model specification.

After the investigation of the derivation of the three models, the WESMLE method was introduced in Section 3.4. Unlike the MLE or simulated MLE method in model estimation, the WESMLE method includes the weights of each observation in model estimation rather than directly treating each observation as a random sample. Therefore, the WESMLE method could be used in the estimation of all the three models to take account of the underreporting issue in the crash data, by treating the underreported crashes as outcome-based samples. However, the limitation of the WESMLE in model estimation is the requirement of the actual population shares of each outcome, which are usually not known for crash data. This will be further discussed in Chapter VII when we address the issue of crash underreporting in the three models.

CHAPTER IV

MODEL ESTIMATION

This chapter examines the performances of the three commonly used crash severity models: the MNL, OP and ML models, using a crash dataset which includes 26,175 single-vehicle crashes involving fixed-object on rural two-way highways. Furthermore, based on a comparison of the estimation results of the three models, this chapter explores which model is more appropriate. The estimated results in this chapter are compared with the results from Monte-Carlo analyses for the crash data related to the effects of sample size and underreporting in crash data, discussed in Chapter V and VII respectively.

This chapter is divided in three sections. Section 4.1 briefly discusses the crash data used for model estimation in the three models. In Section 4.2, the estimation results for the MNL, OP and ML models are given respectively, and the comparisons of the estimation results of the three models are summarized as well. Section 4.3 provides a summary of the chapter.

4.1 Crash Data

The primary data sources utilized in this study include four years (from 1998 to 2001) of traffic crash records provided by the Texas Department of Public Safety (TxDPS) and the Texas Department of Transportation (TxDOT) general road inventory. This research investigated the probability of each crash severity for single-vehicle traffic crashes involving fixed objects that occurred on rural two-way highways (excluding those occurring at intersections). There are two reasons for selecting this type of crashes for analysis in the research. Firstly, fixed-object crashes have more fatalities than other

crash types which could avoid the problem of insufficient sample sizes of fatal crashes (Generally, severe or fatal crashes are much fewer than the less severe crashes, such as PDO.). This is especially important when analyzing the effects of underreporting in crash data in Chapter VII, as too low percentage of fatal crashes probably leads to the issue of insufficient sample size for fatal crashes. Based on 2008 data, for the U.S, collisions with fixed objects accounted for 20 percent of all reported crashes, but resulted in 46 percent of all fatal crashes (NHTSA, 2009), which is much higher than that of other crash types. Also, crashes occurred on rural two-way highways have different contributing factors from those on urban highways, as crashes involving single vehicle compared to those involving multiple vehicles. Thus, it is necessary to choose a specific type of crash in order to decrease the variability of data for analysis. In addition, in order to have sufficient sample size for the dataset, it was decided to select single-vehicle crashes involving fixed objects occurred on rural two-way highways, which occurred more often than those on urban highways.

There are 26,175 usable records in the database, which contains a variety of information including weather conditions, roadway and driver's characteristics, vehicle features as well as the crash severity level reported at the time of the crash. The crash severity was classified into five categories: PDO (O), possible injury (C), non-incapacitating injury (B), incapacitating injury (A), and fatal (K). For this dataset, these categories have 11,844 (45.3%), 5,270 (20.1%), 5,807 (22.2%), 2,449 (9.4%), and 805 (3.1%) observations for severity O, C, B, A, and K respectively. There are 26 independent variables used in the empirical analysis, as summarized in Table 4.1.

Table 4.1 Summary Statistics of Variables for Modeling

Variable Type	Description	Mean	St.d
Road condition			
Log(ADT)	Log of average daily traffic	7.597	0.999
Shoulder width	Shoulder width is between 0 and 20ft	4.865	3.264
Lane width	Lane width is between 8ft and 16ft	11.341	1.251
Speed limit	Maximum speed limit is between 30mph and 75mph	58.330	6.935
Curve & level indicator	1=curve, level; 0=otherwise	0.373	0.484
Curve & grade indicator	1=curve, grade; 0=otherwise	0.002	0.048
Curve & hill indicator	1=curve, hill; 0=otherwise	0.002	0.047
Crash information			
Night indicator	1=night;0=day	0.495	0.500
Dark with no light indicator	1=dark with no light; 0=otherwise	0.424	0.494
Dark with light indicator	1=dark with light; 0=otherwise	0.033	0.177
Rain indicator	1=rain; 0=otherwise	0.806	0.395
Snow indicator	1=snow; 0=otherwise	0.005	0.068
Fog indicator	1=fog; 0=otherwise	0.023	0.149
Surface condition indicator	0=good surface(dry); 1=otherwise	0.267	0.442
Driver information			
Vehicle type indicator	1=truck; 0=otherwise	0.474	0.499
Driver gender indicator	1=female; 0=male	0.340	0.474
Driver defect indicator	1=defect (including physical and mental defect); 0=otherwise	0.176	0.381
Restraining device use indicator	1=no restraining device used; 0=otherwise	0.120	0.325
Fatigue indicator	1=fatigued or asleep; 0=otherwise	0.151	0.358
Airbag deploy indicator	1=air bag deployed; 0=otherwise	0.179	0.384
Seat belt use indicator	1=seat belt used; 0=otherwise	0.649	0.477
Fixed-object type information			
Hit pole indictor	1=hit pole; 0=otherwise	0.113	0.317
Hit tree indictor	1=hit tree; 0=otherwise	0.224	0.417
Hit fence indictor	1=hit fence; 0=otherwise	0.261	0.439
Hit bridge indictor	1=hit bridge; 0=otherwise	0.052	0.222
Hit barrier indictor	1=hit barrier; 0=otherwise	0.058	0.233

4.2 Model Estimation Results

Using the data in Section 4.1, the three models were developed, estimating the probabilities for the five crash severity levels conditioned on a crash having occurred. NLOGIT version 4.0 (on NLOGIT 4.0, Reference Guide, 2007), an extension of LIMDEP for estimation of various discrete outcome models, was used for model estimation. The estimation of each model is described as follows.

4.2.1 Analysis for the MNL model

In the procedure for estimating the MNL model, all 26 explanatory variables listed in Table 4.1 were tested for inclusion, and only 10 variables were retained in the estimation result as shown in Table 4.2. The criteria used for variables inclusion are data availability, engineering judgment, and significance level (0.05 was used in this research study). For the five crash severities, fatal (K) was used as the base outcome. Initially, coefficients of a variable in the severity propensity function T_{ki} were specified to be different across all four severity categories (except for fatal, as a base outcome). If no significant difference at a 0.05 significance level was observed among the coefficients in any two of the severity propensity functions, they were set to be equal. Likelihood ratio tests[2] were used to test whether the coefficients of a variable in the four severity propensity functions were significantly different from each other.

---

2 The likelihood ratio test statistic is $\chi^2=2[LL(\beta_R)-LL(\beta_U)]$. $LL(\beta_R)$ is the log-likelihood at convergence of the restricted model when the parameters of a variable in the severity propensity function T_{ki} are restricted to be the same value across some crash severity levels (not necessary across all of the four estimated crash severity levels). $LL(\beta_U)$ is the log-likelihood at convergence of the unrestricted model when there is no restriction of the parameters of a variable which means different values of parameters of a variable in the severity propensity function T_{ki} are used across all the four estimated crash severity levels. The statistic is χ^2 distributed with the degrees of freedom equal to the difference in the numbers of parameters between the restricted and unrestricted model. (See more details in Washington et al., 2010). In the dissertation, when the χ^2-value is larger than the critical value of the test statistic at 0.05 significance level from standard statistical tables, the unrestricted model should be used.

Table 4.2 Estimation Result of the MNL Model

Severity level / Variable	PDO		Possible injury		Non-incapacitating injury		Incapacitating injury	
	Coef.	**t-Ratio**	**Coef.**	**t-Ratio**	**Coef.**	**t-Ratio**	**Coef.**	**t-Ratio**
Constant	4.489	11.99	4.166	11.14	3.816	10.21	3.213	9.32
Road condition								
log(ADT)	0.153	7.48	0.074	3.72	0.074	3.72		
Speed limit	-0.020	-3.82	-0.020	-3.82	-0.020	-3.82	-0.020	-3.82
Crash information								
Night indicator			-0.229	-6.80	-0.153	-5.16	-0.153	-5.16
Dark with light indicator	0.152	2.09						
Rain indicator	-0.933	-7.21	-0.819	-6.22	-0.523	-3.97	-0.394	-2.83
Snow indicator	0.473	2.40						
Driver information								
Driver defect indicator	-1.255	-10.03	-0.280	-3.28	-0.280	-3.28	-0.280	-3.28
Fatigue indicator	0.465	4.59	-0.258	-5.24				
Restraining device used indicator	-2.532	-30.77	-1.987	-23.15	-1.401	-17.53	-0.830	-9.77
Fixed-object type information								
Hit tree indicator	-1.046	-13.20	-0.826	-10.05	-0.612	-7.63	-0.363	-4.24
Log-likelihood at zero = -42127.0								
Log-likelihood at convergence = -33926.2								
Adjusted $\rho^2 = 0.194$								

* Shaded coefficients were made to be the same across respective crash severity categories.

The final estimation results of the MNL model using the four-year traffic crash records are presented in Table 4.2. As shown in this table, 10 variables were retained in the final model: two variables of road condition, four variables of crash information, three variables of driver information and one variable of fixed-object type information. The absolute values of the t-ratios for all 10 variables retained are larger than 1.96, which means that these variables' coefficients are significantly different from zero at the 5% level. The effects of each variable on the crash severity levels were quantified by

elasticity in Table 4.3. Overall model fit[3] is 0.194 which is good given the large amount of variability in crash severity data.

In order to directly explain the magnitude of the impact of a specific variable on crash severity probabilities, elasticities of each variable were calculated. The direct elasticity calculated by Equation (3.3) for the continuous variables log(ADT) and Speed Limit, and pseudo-elasticity calculated by Equation (3.4) for the rest of eight indictor variables are shown in Table 4.3 for the estimated MNL model.

From Table 4.3, for road condition variables, the result shows that a 1% increase in *log(ADT)* results in a 63.3%, 45,2%, and 44.1% increase in the probability of PDO, possible injury, and non-incapacitating injury, respectively. This result is reasonable as more traffic is usually believed to lead to more crashes but of lower crash severity. All the negative values of elasticity for the variable of *speed-limit* show that increasing the speed limit which accordingly results in higher traffic speeds decreases the likelihood of getting involved in PDO, possible injury, non-incapacitating injury, and incapacitating injury crashes relative to fatal crashes. A 1% increase in speed limit results in a 64.5%, 94.1%, 91.6%, and 106.8% decline in the probability of the above four injury severity outcomes respectively. This seems to support that a higher speed leads to a higher risk of drivers being involved in fatalities, once a single-vehicle crash involving fixed-object

---

3 The adjusted rho-squared value (ρ^2), also called the likelihood ratio index, is widely used to describe the overall goodness-of-fit for discrete outcome models (Koppelman and Bhat, 2006). The rho-square with respect to zero is calculated as $\rho^2 = 1 - \frac{LL(\hat{\beta}) - K}{LL(0)}$, where $LL(\hat{\beta})$ represents the log-likelihood for the estimated model, $LL(0)$ represents the log-likelihood with zero coefficients (which results in equal likelihood of occurring each severity type), and K is the number of parameters (degrees of freedom) used for the model. For the MNL model estimation, the estimated parameters (degrees of freedom) is 27, the log-likelihood at zero is -42127.0 and the log-likelihood at convergence is -33926.2, so the adjusted rho-squared for the zero model is 0.194.

occurs on rural two-way highways. Furthermore, the variable of *speed-limit* is elastic or near elastic for each severity level (except the baseline severity level: fatal) since the absolute values of the elasticity are greater than or close to 1.

Table 4.3 Direct Elasticity Estimates of the MNL Model

Severity level / Variable	Elasticity and Pseudo-elasticity in percent (%)			
	PDO	**Possible injury**	**Non-incapacitating injury**	**Incapacitating injury**
Road condition				
log(ADT)	63.3	45.2	44.1	
Speed limit	-64.5	-94.1	-91.6	-106.8
Crash information				
Night indicator		-9.1	-6.0	-7.0
Dark with light indicator	4.1			
Rain indicator	-26.1	-33.4	-22.1	-18.4
Snow indicator	11.6			
Driver information				
Driver defect indicator	-39.4	-11.5	-11.3	-12.8
Fatigue indicator	11.1	-10.5		
Restraining device used indicator	-74.6	-71.4	-54.7	-37.3
Fixed-object type information				
Hit tree indicator	-32.3	-33.8	-24.9	-16.7

Considering the pseudo-elasticity values of the indicator variables and turning first to the *night indicator* variable it was found that drivers driving at night were 9.1%, 6%, and 7% less likely to be involved in possible injury, non-incapacitating injury, and incapacitating injury crashes. This finding reflects that driving at night on rural two-way highways increases the probability of fatal crash once the driver hits a fixed object.

In terms of the pseudo-elasticity values of *Dark with light indicator* and *Snow indicator* variables, both are positive for PDO. It indicates that drivers driving at dark with road

light are 4.1% more likely to be involved in PDO crashes. Similarly, during snow days the probability for drivers to be involved in PDO crashes increases 11.6%. This is consistent with previous findings regarding the effects of snow on crash severities. For instance, a study, which analyzed crash rates in the 48 contiguous states, suggested during snowfall crash severity decreased due to drivers' adjustment of their driving behavior such as driving slower (Eisenberg and Warner 2005). In both of the dark with road light and snowfall situation, the decrease in the probability of fatalities could be due to drivers slowing down and being more cautious, which leads to the decrease in driver and passenger injuries.

The pseudo-elasticity value of *rain indicator* variable indicates a relative increase in the prediction of fatal crashes when it is raining, which results in a 26.1%, 33.4%, 22.1%, and 18.4% decline in the probability of PDO, possible injury, non-incapacitating injury, and incapacitating injury crashes relative to fatal crashes. This seems reasonable since in the rain, vehicles have less friction with the road surface, which increases the probability of their hitting a fixed object at high speeds. Thus, once a crash occurs in the rain, it is more likely to involve fatalities.

The negative values of pseudo-elasticity of *Driver defect indicator* and *Restraining device used indicator* variables for all severity levels reflect that when drivers have physical or mental defects or they do not use any restraining devices, they are less likely to be involved in PDO, possible injury, non-incapacitating injury, and incapacitating injury crashes relative to fatal crashes. For drivers having physical or mental defects, a 39.4%, 11.5%, 11.3%, and 12.8% decrease in the probabilities of the above four injury severity outcomes happens, respectively. For drivers without any restraining devices, the probabilities for being involved in the above four injury severities decrease by 74.6%, 71.4%, 54.7% and 37.3% respectively. Another factor of drivers, *fatigue*, results in an 11.1% increase in PDO crash and a 10.5% reduction in possible injury crash.

The negative values of pseudo-elasticity of *Hit tree indicator* variable show that the probabilities of all crash severities (except for fatal) decrease when drivers hit a tree in crashes. A 32.3%, 33.8%, 24.9% and 16.7% decline in the probability of PDO, possible injury, non-incapacitating injury, and incapacitating injury crashes respectively relative to fatal crashes, which indicates that hitting a tree increases the probability of being involved in a fatal crash.

4.2.2 Analysis for the ML model

The ML model allows for the randomness of parameters of individual variables, and thus in developing the model, it was first assumed that all parameters of the variables included in the model were random. The popular distributions (normal, uniform and lognormal distribution) were tested for random parameters, so numerous combinations were evaluated by modifying the parameter assumptions. Then, the t-test was used to examine their estimated standard deviations for exploring the randomness of each parameter: if their standard deviation was not found statistically different from zero at the 0.05 significance level, they were restricted to be fixed rather than random. Meanwhile, the simulation-based maximum likelihood method was used for the parameter estimation, with Halton draws=200.

The final result of the ML model estimation is shown in Table 4.4, which is the final decision based on the engineering judgment with the consideration of a better GOF. The estimated ML model has 29 estimated parameters (degrees of freedom), the log-likelihood at zero is -42127.0 and the log-likelihood at convergence is -33919.9. The adjusted rho-squared for the zero model is 0.195, almost the same value as that of the MNL model.

Table 4.4 Estimation Result of the ML Model

Severity level / Variable	PDO		Possible injury		Non-incapacitating injury		Incapacitating injury	
	Coef.	**t-Ratio**	**Coef.**	**t-Ratio**	**Coef.**	**t-Ratio**	**Coef.**	**t-Ratio**
Constant	4.43	11.50	4.155	10.84	3.764	9.79	3.235	9.21
Road condition								
log(ADT)	0.167	7.65	0.079	3.73	0.079	3.73		
Speed limit	-0.020	-3.76	-0.020	-3.76	-0.020	-3.76	-0.020	-3.76
Crash information								
Night indicator			-0.238	-6.86	-0.183	-5.21	-0.183	-5.21
Dark with light indicator	0.166	2.01						
Rain indicator	-0.939	-7.10	-0.810	-6.10	-0.997	-4.32	-0.397	-2.844
Std.dev. of distribution					1.568	4.043		
Drive information								
Driver defect indicator	-1.359	-9.75	-0.240	-2.67	-0.240	-2.67	-0.240	-2.67
Fatigue indicator	0.507	4.39	-0.328	-5.71				
Restraining device used indicator	-3.406	-7.57	-2.003	-23.215	-1.252	-11.89	-0.834	-9.79
Std.dev. of distribution	2.220	3.03						
Fixed-object type information								
Hit tree indicator	-1.144	-11.79	-0.857	-10.30	-0.561	-6.28	-0.378	-4.40
Std.dev. of distribution	0.939	2.40						
Log-likelihood at zero = -42127.0								
Log-likelihood at convergence = -33919.9								
adjusted $\rho^2 = 0.195$								

* Shaded coefficients were made to be the same across respective crash severity categories.

As shown in Table 4.4, the ML model estimation has similar results as those from the MNL estimation. Both models have the same significant variables except for the *Snow indicator*, which is no longer significant for the ML model. Meanwhile, all the fixed parameters of the variables in both of the MNL and ML models have similar values with the same signs. However, there exists randomness for the *Rain indicator* in Non-incapacitating injury, the *Restraining device used indicator* in PDO, and the *Hit tree indicator* in PDO. All of the three random parameters have a normal distribution. As the

signs and values of all the significant variables are consistent with the MNL model estimation results, only the three random variables are discussed further.

For the *Rain indicator* variable in non-incapacitating injury, the parameter is estimated to be normally distributed with a mean equal to -0.997 and standard deviation equal to 1.568, which indicates that the parameter sign switches from negative in 74% of the drivers to positive in 26% of the drivers. This implies that under most conditions, rain increases the probability of a non-incapacitating injury crash. The indicator variable *Restraining device used* has a randomness parameter in PDO outcome that is normally distributed with a mean equal to -3.406 and standard deviation equal to 2.220. This gives the parameter being less than zero for 94% of the drivers and greater than zero for 6% of the drivers, which indicates that for the majority of drivers, driving without using restraining devices decreases the likelihood of involving in PDO injuries relative to involving in fatalities once a crash occurs. Finally for the randomness of the *Hit tree indicator* variable in the PDO outcome, it was found that its parameter varies over the sample of drivers. The parameter has a normal distribution with a mean equal to -1.144 and standard deviation equal to 0.939, which results in 89% of the distribution less than zero and 11% greater than zero. Thus, for the majority of drivers, hitting on the tree decreases the likelihood of being involved in a PDO crash relative to being involved in a fatal crash once a crash occurs. As described above, the three random parameters add more flexibility for the ML model, by capturing the variability of each parameter's effect over the sample of drivers involved in crashes.

4.2.3 Analysis for the OP model

In the OP model estimation, all the 26 variables listed in Table 4.1 were available for the model, and the backward selection was used for simplification so that only those significant at the 0.05 significance level were included in the model. In addition, the five crash severities were set in an ascending order as OCBAK. The final result of the OP model is shown in Table 4.5. From this table, more significant variables were included in

the OP model (18 variables total) than the MNL (10 variables total) and ML models (9 variables total).

Table 4.5 Estimation Result of the OP Model

Variable	Coef.	t-Ratio
Constant	0.249	2.83
Road Condition		
log(ADT)	-0.049	-6.90
Speed limit	0.002	2.47
Curve & level indicator	0.062	4.18
Crash Information		
Night indicator	-0.124	-4.42
Dark with no light indicator	0.064	2.28
Fog indicator	0.106	2.29
Surface condition indicator	-0.259	-15.67
Driver Information		
Vehicle type indicator	0.0561	3.83
Driver gender indicator	0.132	8.62
Driver defect indicator	0.398	9.38
Restraining device used indicator	0.802	21.75
Fatigue indicator	-0.173	-3.82
Airbag deploy indicator	0.447	12.62
Seat belt use indicator	-0.128	-3.87
Fixed-object Type Information		
Hit pole indictor	-0.076	-3.19
Hit tree indicator	0.188	10.12
Hit fence indictor	-0.160	-8.83
Hit barrier indictor	-0.090	-2.87
Threshold Parameters		
γ_1	0.561	86.19
γ_2	1.393	139.62
γ_3	2.186	133.17
Log-likelihood at zero = -42127.0		
Log-likelihood at convergence = -33328.9		
Adjusted $\rho^2 = 0.208$		

As discussed in Section 3.2, a positive value of a parameter implies that an increase of the variable value will increase the probability of fatal crash (the highest ordered discrete category in the research) and decrease the probability of PDO crash (the lowest ordered discrete category in the research). Thus, the largest positive parameter value among all variables is the *Restraining device used indicator*, which implies that not wearing restraining devices is the most critical factor that significantly aggravates the average risk of injury in hitting a fixed object crashes on rural two-way highways.

The positive sign for the continuous variable *speed limit* implies that with a higher speed limit (accordingly, a higher traffic speed), the likelihood of a fatal crash will be increased and the likelihood of a PDO crash will be decreased. This finding is consistent with what was found by O'Donnell (O'Donnell, 1996) and with common knowledge regarding the effects of high speed on traffic safety. The negative sign for the continuous variable *log(ADT)* suggests that more traffic on the road results in a higher probability of PDO crash and a lower probability of fatal crash, consistent with the finding from the MNL estimation that more traffic results in more crashes of lower crash severity.

The following dummy variables have positive signs of their parameters: *Curve & level indicator*, *Dark with no light indicator*, *Fog indicator*, *Vehicle type indicator*, *Driver gender indicator*, *Driver defect indicator*, *Airbag deploy indicator*, and *Hit tree indicator*. It indicates that drivers who travel at a curve and level segments, or in a dark with no light circumstance, or at a foggy weather, or in a truck are more likely to be involved in severe or fatal crashes. Meanwhile, female drivers, or drivers who have physical/mental defect have a higher probability to be involved in crashes that result in more severe injury. In addition, when airbags are deployed in a crash, or when drivers hit a tree, the odds of drivers suffering more severe injury also increase. All the findings are intuitive except that when airbags are deployed the average risk of injury increases, since airbags are known to prevent occupant fatalities. However, it is still possible that in a fixed-object crash, the deployment of airbag implies a high impact with the object, which could lead to a high injury severity of the crash. As described by Kent (2003), the

airbag was originally developed to provide protection for unbelted occupants. However, it was also found the airbag deployment significantly increased the abdominal injury risk for belted drivers in frontal crashes (Augenstein, et.al, 1995, Thor and Gabler, 2007). Potential explanations for this include the change in occupant kinematics after the chest and head engage the deployed airbag, or the excessive loading of the abdomen by the airbag itself or the rim of the steering wheel.

Furthermore, the following dummy variables have negative parameters: *Night indicator*, *Surface condition indicator*, *Fatigue indicator*, *Seat belt use indicator*, *Hit pole indictor*, *Hit fence indictor*, and *Hit barrier indictor*. It shows that drivers have a lower probability of suffering fatal and a higher probability of being involved in PDO when drivers travel at night, drive on a bad road surface such as a wet road surface, drivers are fatigued or asleep, drivers wear seat belts, or drivers hit poles/fences /barriers. The estimated parameters for most of these variables have signs and values that are intuitively reasonable. However, the counterintuitive variables are *Night indicator*, and *Surface condition indicator*. These could be explained by the drivers' tendency of driving more cautiously and slowly at night or on a bad road surface, which lowers the severity level of a crash, when it occurs.

4.2.4 Comparison of model results

Based on the outputs of three model estimations, it was found that the ML model is more interpretive than the MNL model, since the ML model includes the randomness associated with parameters for some variables. For the ML model, the involvement of the randomness in the parameters results in the prediction of a mean value and standard deviation for the probability of each severity level rather than a single point probability. Meanwhile, though accounting for the ordinal information of crash severities, the OP model still does not have the interpretive power of the MNL and ML models. The OP model restricts the effects of explanatory variables on ordered discrete outcome probabilities by using the same coefficient for an individual explanatory variable across

different crash severity levels. It causes the variable either to increase the probability of highest severity (fatal in the study) and decrease the probability of lowest severity (PDO in the study), or to decrease the probability of highest severity and increase the probability of lowest severity. This may not be realistic because it is possible that certain explanatory variable can create an increase in the probability for some outcome predictions but decrease the probability for other outcome predictions. For instance, inclement weather could lead to an increase in the probabilities for both the two highest severities (KA) and the lowest severity (O), but reduce the probability of the other severities (BC). In addition, it is not clear what the effect a positive or negative variable parameter has on the probabilities of the "interior" severity levels: A, B, and C, as discussed in Section 3.2.

In terms of the GOF among the three models, the OP model includes more significant variables (18 variables) which results in a slightly higher adjusted rho-squared value (Adjusted ρ^2= 0.208) than those of the MNL and ML models (Adjusted ρ^2=0.194 for the MNL model and Adjusted ρ^2=0.195 for the ML model, almost the same). Since the MNL model is a nested model of the ML model, we can further compare their GOF using a likelihood ratio test, even though both of them have nearly the same adjusted rho-squared value. From the MNL model estimation results, the log-likelihood function at convergence is -33926.2 with 27 estimated parameters (degrees of freedom, including four estimated constant variables), and the log-likelihood function at convergence for the ML model estimation is -33919.9 with 29 estimated parameters (three more randomness in the variables than the MNL model and one less significant variable *Snow indicator*). Thus, the likelihood ratio statistic is 2*(-33919.9-(-33926.2))=12.6 with 2 degrees of freedom, which is larger than the χ^2 table value of 5.99 for the 0.05 level of significance. This indicates that the ML model is statistically better than the MNL model in terms of GOF at the 5% significance level, thereby rejecting the null hypothesis that the MNL model has a better GOF than the ML model.

4.3 Chapter Summary

In this chapter, the three most commonly used crash severity models: the MNL, OP and ML models were applied to a crash dataset that included 26,175 single-vehicle crashes involving fixed objects on rural two-way highways. The major aim of this chapter was to set a baseline for the Monte-Carlo analysis of the effects of sample size and underreporting in data on the three models using crash data, rather than to put insight into contributing factors such as characteristics of traffic, driver, vehicle, and road environment for single-vehicle crashes involving fixed objects on rural two-way highways. For this reason, only a brief description was provided on the estimated parameters of each variable for all the three models in this chapter. A total of 26 variables were examined for each model. Only 10 variables were found to be significant for the MNL model and nine variables for the ML model, while 18 variables were significant for the OP model. The general trends produced by the estimated parameters for the common variables for the three models were similar.

With this crash dataset, it was found that the ML model had a better interpretive power than the MNL model, while the MNL model had a superior interpretive power than the OP model. The OP model had the least interpretive power since it does not have the flexibility to explicitly control the probabilities of interior categories, which depend on the thresholds. Meanwhile, the OP model requires the variable either to increase the probability of highest severity crashes (fatal in the study) and decrease the probability of lowest severity crashes (PDO in the study), or to decrease the probability of highest severity and increase the probability of lowest severity. However, it does not allow the probabilities of both of the highest and lowest severity to increase or decrease. In addition, the OP model had a slightly better GOF than that of the MNL and ML models, while the ML model had a significant better fit than the MNL model at the 5% significance level for the crash dataset used in the study.

The estimated results in this chapter for the three models are set up as the baseline conditions for the analysis of effects of sample size and underreporting in data on observed crash data in Chapters V and VII respectively. That is, the model estimation results for observed crash data in Chapters V and VII will be compared to the estimated results in this chapter in order to quantify the bias and variability of model estimation due to the effects of small sample size and underreported data.

CHAPTER V

MODEL COMPARISONS BY SAMPLE SIZE

As stated in Chapter I, sample size is an important factor to be considered in model selection. In this chapter, simulated data is used as well as the observed crash data described in Chapter IV to examine the effects of sample size for the three crash severity models: the MNL, OP and ML models. Recall that the dataset from Chapter IV includes 26,175 single-vehicle crashes involving fixed objects on rural two-way highways. Intuitively, it would seem that a small sample size in crash severity models lead to erratic results, which limits the ability to estimate the true parameters and results in an inaccurate prediction of the probabilities for each severity outcome. In order to find the differences in sample size requirements for the three models, a Monte-Carlo analysis based on both simulated data and crash data is used to examine the bias associated with different sample sizes for each model type.

This chapter consists of three sections. Section 5.1 presents the analysis of the effects of sample size on the three models based on simulated data. In Section 5.2, the estimation results from the three models are compared for each sample size based on the crash dataset described in Section 4.1. Section 5.3 provides a summary of important conclusions.

5.1 Analysis Based on Simulated Data

Before studying the sample size requirements for the three models using real crash data, a Monte-Carlo analysis based on simulated data is performed. By repeating the sampling to produce estimates more clustered around the true values, the Monte-Carlo simulation is an ideal way to verify the sample size effects on three models. In this case, we create

the data with the knowledge of true values of estimators and true propensity functions. Thus, the bias and variability can be estimated by comparing the model estimation with the true values of estimators for different sample sizes.

5.1.1 Simulation design

All the variables included in a crash severity model are observation-related rather than outcome-related, which means that the variables keep the same values no matter what crash severity the target observed crash is (Khorashadi et.al, 2005). In other words, a variable included in the propensity functions for each severity outcome for an observed crash is identical though its parameters that describe the effects of the variable might differ across each severity level. Thus, an individual covariate in the propensity functions generated in the simulation should be kept the same at all severities for each observation.

Since crash data usually has five severity categories, the number of parameters in the propensity functions is very large. For simplification, one covariate randomly generated from the standard normal distribution was introduced in all three models. In addition, five outcomes were used to replicate the five severity categories.

The three datasets for each model were generated as follows:

For the MNL model, data were simulated based on Equation (3.4). Outcome 5 (denoted as level 5) was set as the baseline whose propensity function was zero and, thus, we only need to design the propensity functions for outcome 1 through outcome 4 (denoted as level 1 to level 4). For the sake of simplification, the parameters of a covariate were kept the same with a value equal to 1 for level 1 to 4, i.e., β_k =1. α_k (constant parameter) was 0, 0.5, 1, 1.5 for level 1 through 4 (level 5 was the baseline with α5 = β5 =0). The covariate x for each level was drawn from a normal distribution with mean equal to -2 and variance equal to 1. The error term for each level was drawn independently from a

Type I extreme value distribution by obtaining draws from the uniform random distribution and applying the following transformation $-\ln[-\ln(u)]$, where u is a random number drawn from a uniform distribution between 0 and 1. This resulted in the following proportions 5.7%, 9.4%, 15.4%, 25.4%, and 44.1% for levels 1 to 5 respectively, which represented their proportions observed in the data. Levels 1 to 5 in the simulation represented the five crash severities from fatal through PDO in a decreasing order (KABCO), since usually the more severe a crash is, the less likely it is expected to occur.

For the OP model, as we did for the MNL model, only one covariate x was designed for the function of latent variable as shown in Equation (3.5). As stated above, the effect of explanatory variables on all severity levels was restricted to be the same by using the identical coefficient for each explanatory variable across different crash severities, and then β was set as 1 for each level. Meanwhile, x was drawn from a normal distribution with mean equal to 2.2 and variance equal to 1, and γ_k (threshold value, as shown in formula (3.6)) was 0, 0.8, 1.5 and 2.4 for levels 1 to 4. The reason to use these parameter values was to keep the population ratios of each level as close as possible to those for the MNL data, for purpose of simplification. The error term was standard normal distributed for each level. Thus, they gave the following proportions: 6.0%, 10.1%, 15.0%, 24.6%, and 44.3% for levels 1 to 5 respectively, which represented the proportions of the five crash severities from fatal to PDO in a decreasing order (KABCO) for the observed crash dataset.

For the ML model, the steps for generating the dataset were very similar to those used in generating the dataset for the MNL model. The only difference was that the independent variable was assumed to have random parameter for level 1, which was assumed to be normally distributed (mean=1, variance=1). The population ratios for each level were 14.1%, 8.7%, 14.3%, 23.6%, and 39.3% for levels 1 to 5, representing the proportions of the five crash severities from fatal to PDO (KABCO).

Table 5.1 summarizes the true values assumed for three models. The parameter values chosen for these three models were based on the assumption that the results would not be affected much by different values of the parameters for all three models. In addition, Table 5.2 summarizes the population ratio of each level for the three models based on the designed parameter values in Table 5.1.

Table 5.1 True Parameter Values in the Simulation of Three Models

Model Parameter		True Values		
		MNL	OP	ML
Constant Parameter*	α1	0	0	0
	α2	0.5	0.8	0.5
	α3	1	1.5	1
	α4	1.5	2.4	1.5
Variable Parameter	β1	1	1	N(1,1)
	β2	1		1
	β3	1		1
	β4	1		1
Sample Size(N)		100, 250, 500, 1,000, 1,500, 2,000, 5,000, 10,000		

*Constant parameter for OP is represented by γ1- γ4, which are the threshold variables for each level in OP model.

Table 5.2 Population Ratios of Each Level for the Three Models

Outcome	Population Ratio		
	MNL	OP	ML
Level 1	5.7%	6.0%	14.1%
Level 2	9.4%	10.1%	8.7%
Level 3	15.4%	15.0%	14.3%
Level 4	25.4%	24.5%	23.6%
Level 5	44.1%	44.3%	39.3%

Datasets of each model were repeatedly drawn 100 times for each sample size according to the designed true parameter values of the model. Therefore, summary statistics such as mean and standard deviation of 100 iterations of each parameter for a model could be calculated. The sample sizes were designed as 100, 250, 500, 1,000, 1,500, 2,000, 5,000, and finally 10,000. In addition, based on the summary statistics from 100 iterations for a designed sample size, the 95% confidence interval of each estimated parameter could be computed, which is to be compared with the true value of the parameter (as shown in Table 5.1). The entire process of the Monte-Carlo analysis on sample size for simulated data is described in a flowchart, as shown in Figure 5.1. Parts of the code in NLOGIT used to carry out the Monte-Carlo simulation for a specific sample size for each model are provided in Appendix A.

5.1.2 Simulation results

This section describes the results for the three models based on the above simulation design.

5.1.2.1 Results of the MNL Model

The graphs in Figure 5.2 show the relationship between 95% confidence intervals for the four estimated constant parameters and the parameters of independent variables for the sample sizes designed in Table 5.1. In each graph, the Y-axis is the parameter estimate, and the X-axis is the sample size. For each sample size, there are two estimates of the parameter, one for the lower-bound and the other for the upper-bound of the 95% confidence interval. Thus, the interval encloses a 95% probability of the real value of each parameter.

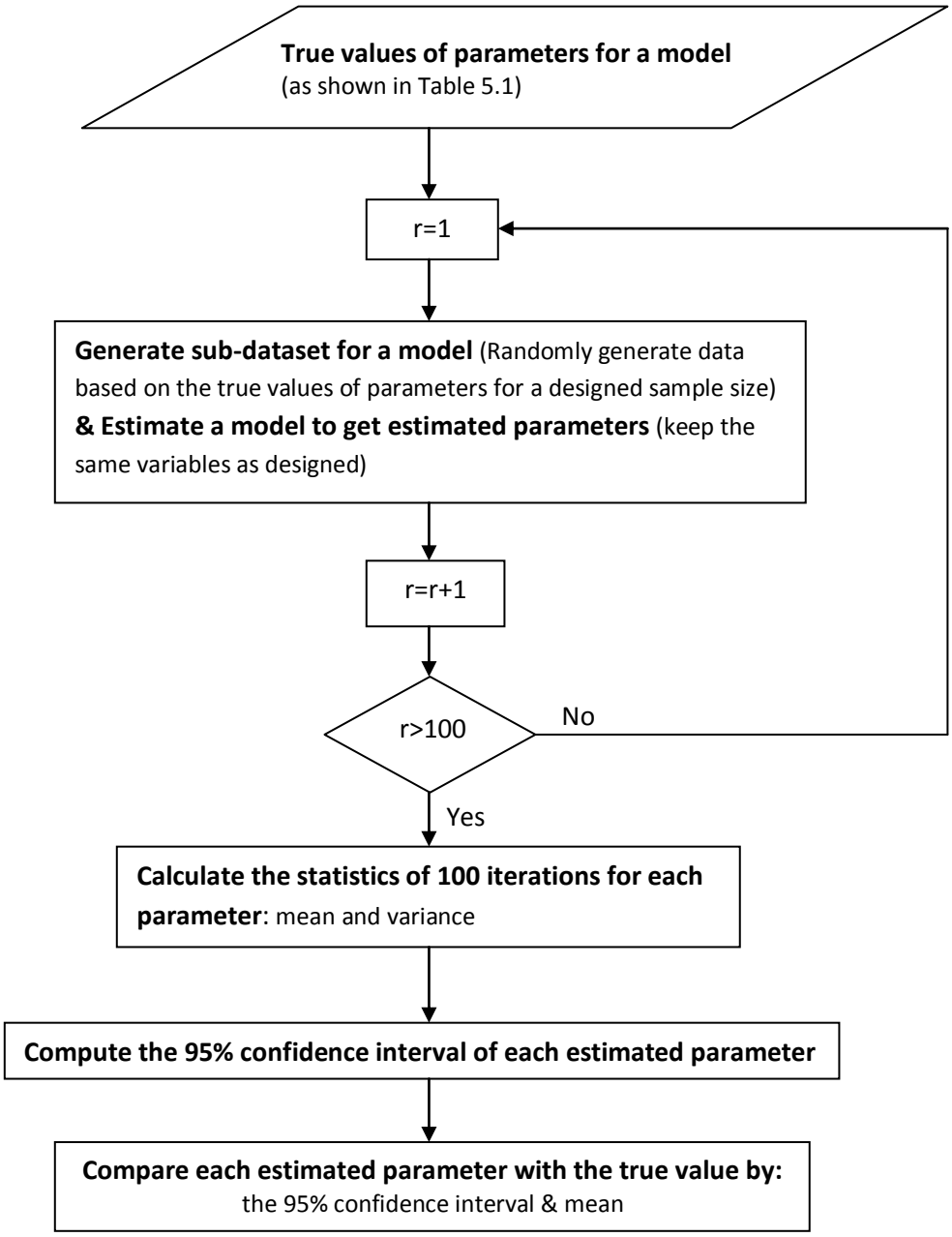

Figure 5.1 Monte-Carlo Analysis on Sample Size for Simulated Data

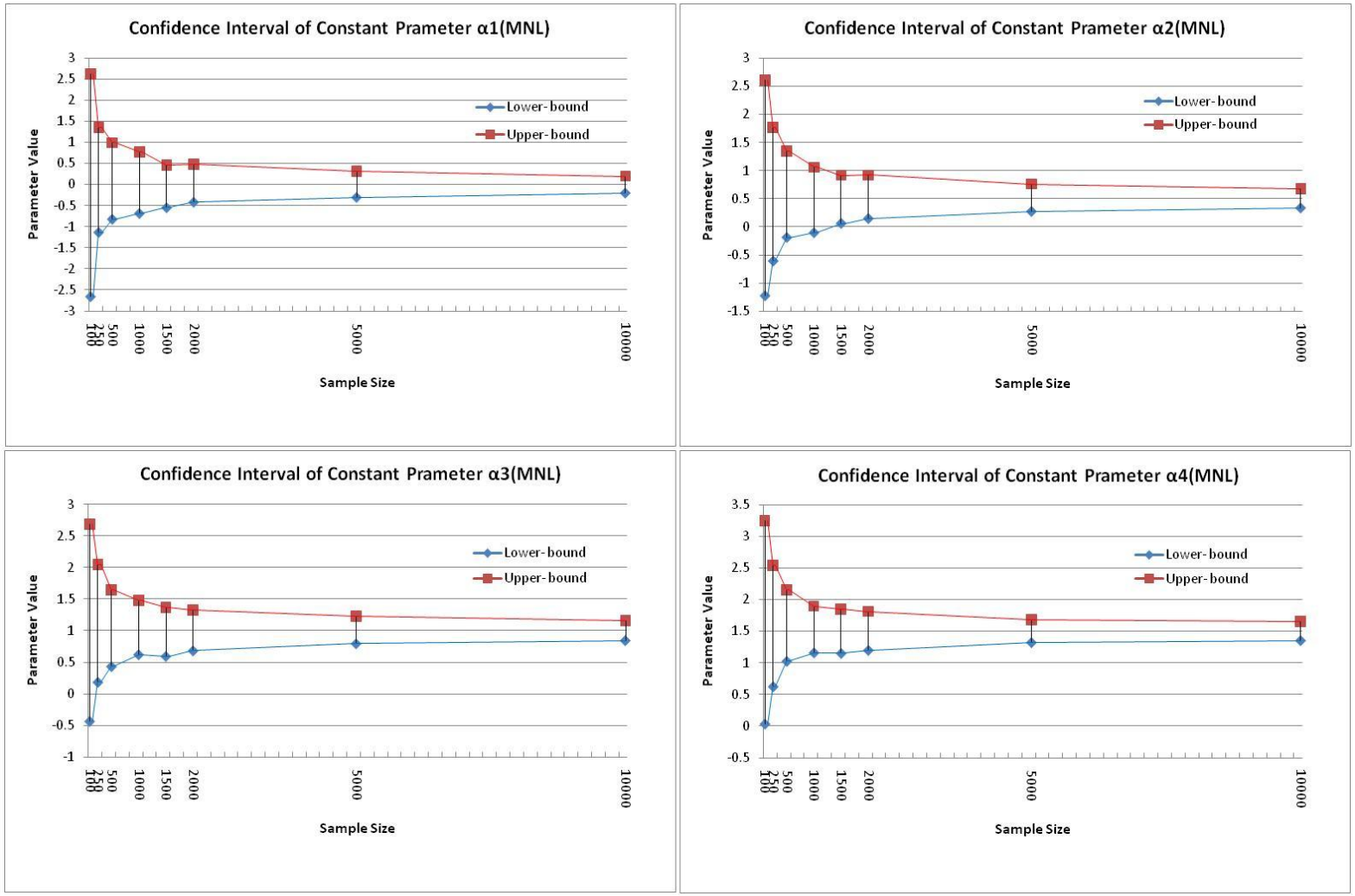

Figure 5.2 Confidence Intervals of the Parameters by Sample Size for the MNL

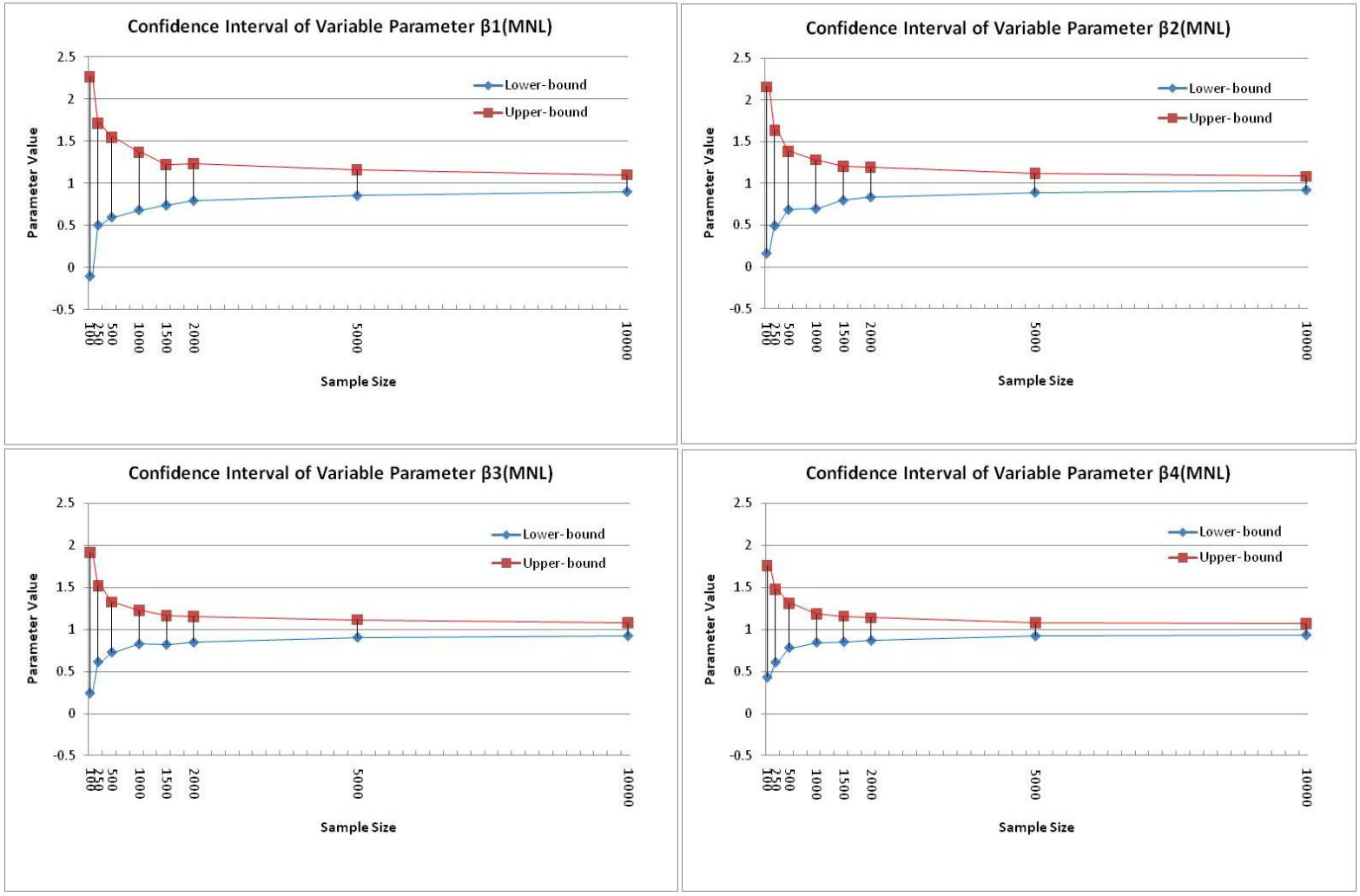

Figure 5.2 Continued

From Figure 5.2, it can be seen that for each parameter, the range of 95% confidence interval becomes narrower as the sample size increases. In addition, as the sample size reaches 2,000, the 95% confidence interval gets narrow and stable around the true value for each parameter. In order to further examine the simulation results, the relationship between the mean value of each parameter and sample size was extracted and is illustrated in Figure 5.3. Figure 5.3 indicates that sample sizes of less than 2,000 are erratic and inconsistent in terms of the ability to find the true parameters. The mean value becomes stable for sample sizes larger than 2,000.

5.1.2.2 Results of the OP Model

As shown in Figures 5.4 and 5.5, larger sample sizes lead to the narrower range of the 95% confidence interval for parameters and closer value of the mean. When compared with the MNL model, the only difference for the OP model is that the stable point arrives at a smaller sample size, which is about half of that for the MNL model (1,000). In other words, as the sample size reaches 1,000, the 95% confidence interval of parameters becomes narrow and stabilizes around the true value. Meanwhile, the mean value stabilizes towards the true value for each parameter, as the sample size increases to 1,000.

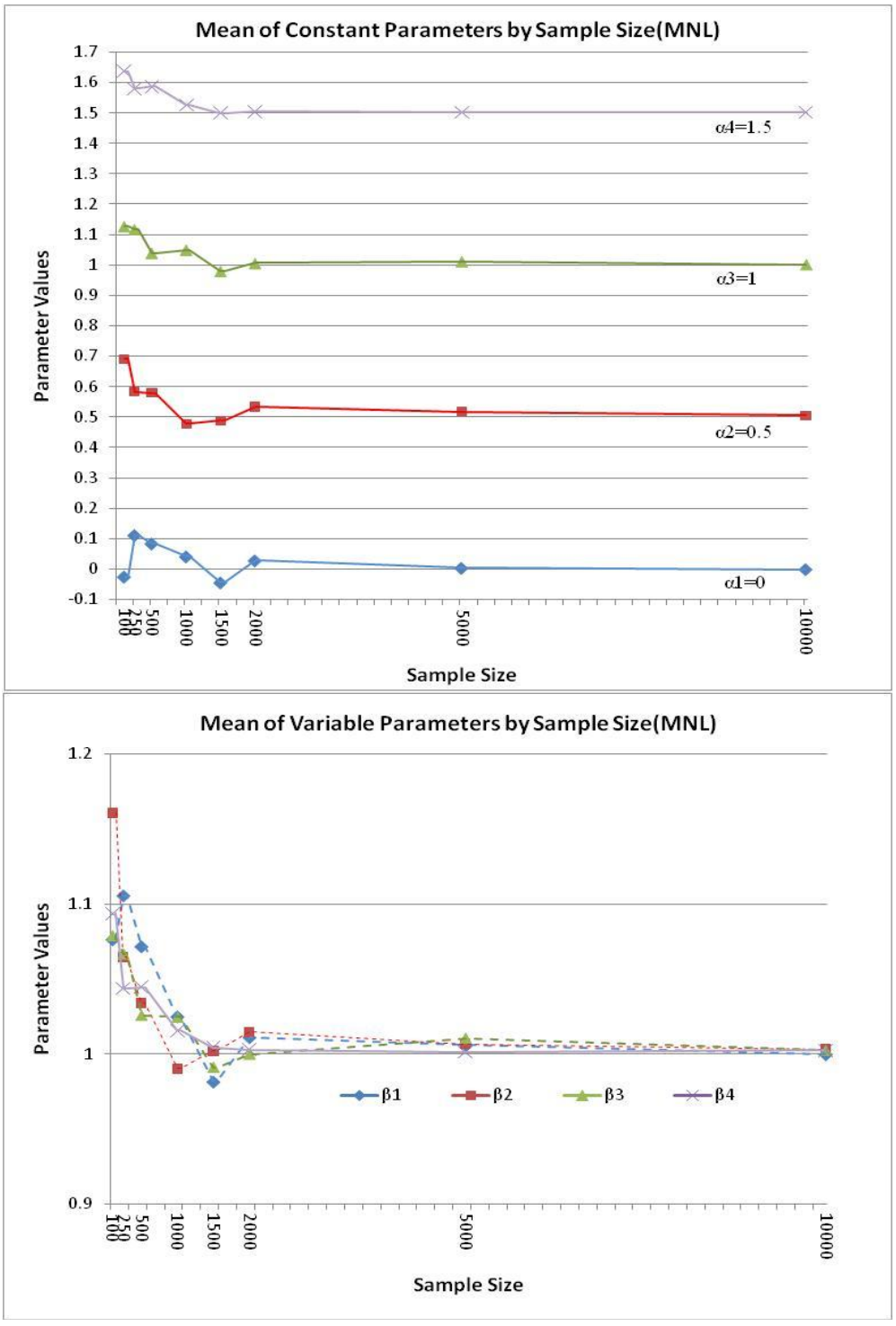

Figure 5.3 Mean of the Parameters by Sample Size for the MNL

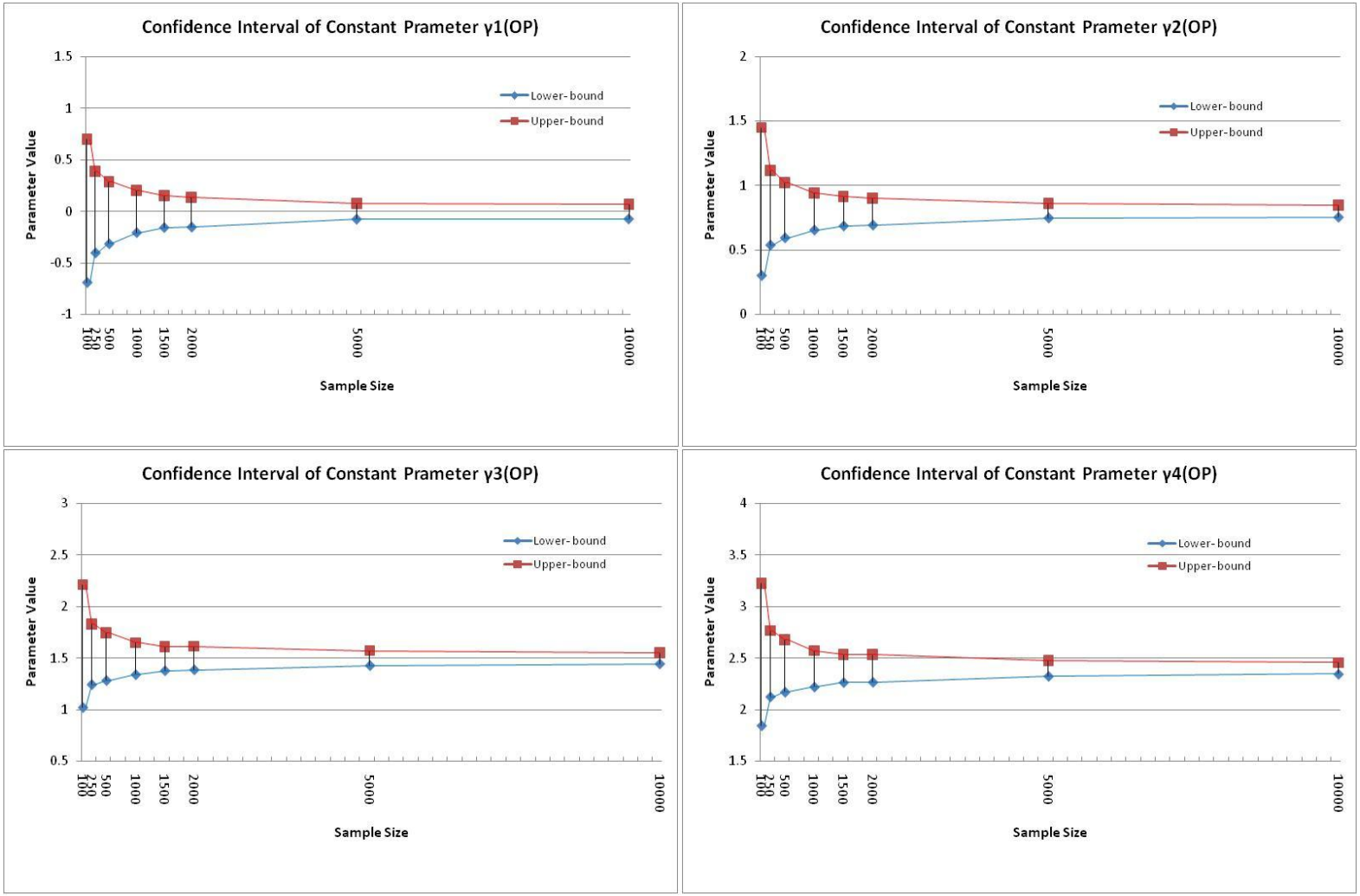

Figure 5.4 Confidence Intervals of the Parameters by Sample Size for the OP

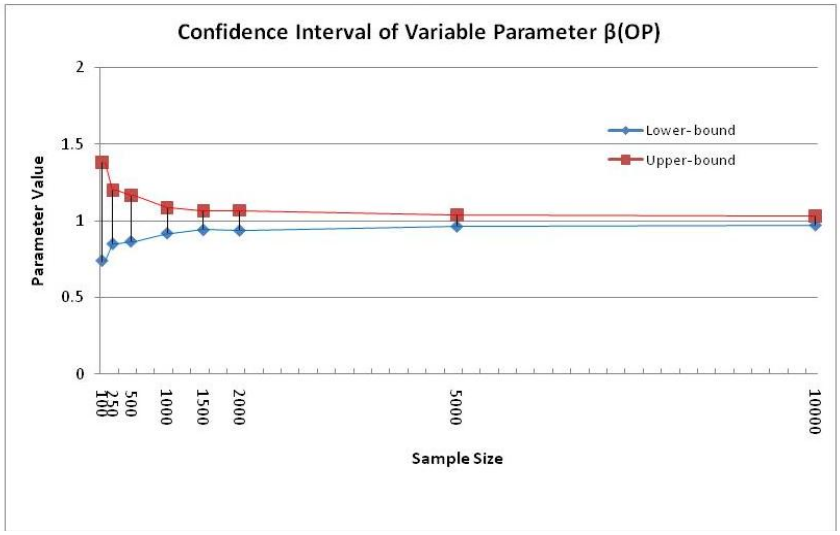

Figure 5.4 Continued.

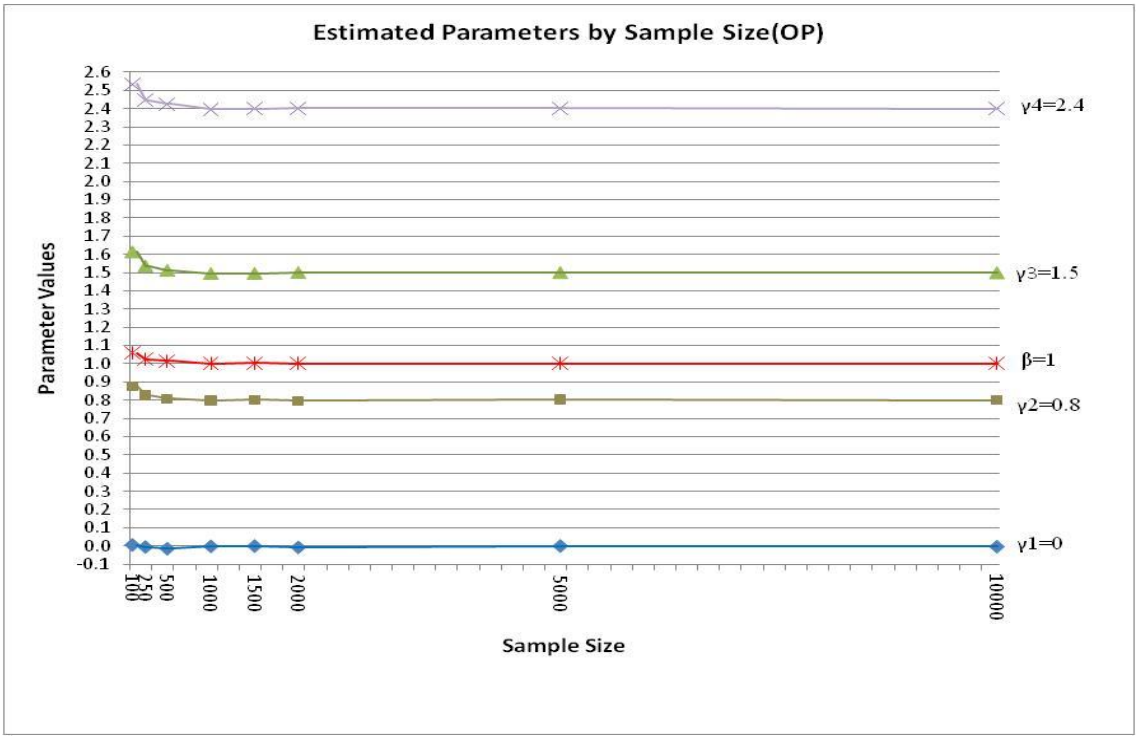

Figure 5.5 Mean of the Variable Parameters by Sample Size for the OP

5.1.2.3 Results of the ML Model

Figures 5.6 and 5.7 demonstrate the relationships between both the 95% confidence intervals for the parameters and the mean value for each parameter as a function of the sample size. Very similar patterns as those observed in the previous figures can be seen in these two figures. However, certain differences can be noticed for β_1. Since the parameter β_1 is random, its estimated value was found to be less stable, especially for small sample sizes. The point of stabilization is around 5,000, which is the largest amongst the three models. Finally, it is anticipated that a larger sample size may be needed for the ML model, if more random parameters are introduced into the model.

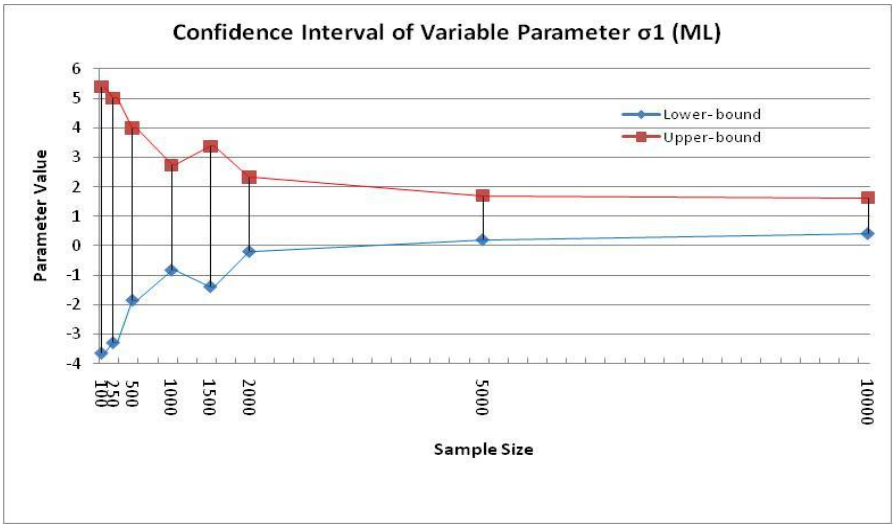

Figure 5.6 Confidence Intervals of the Parameters by Sample Size for the ML

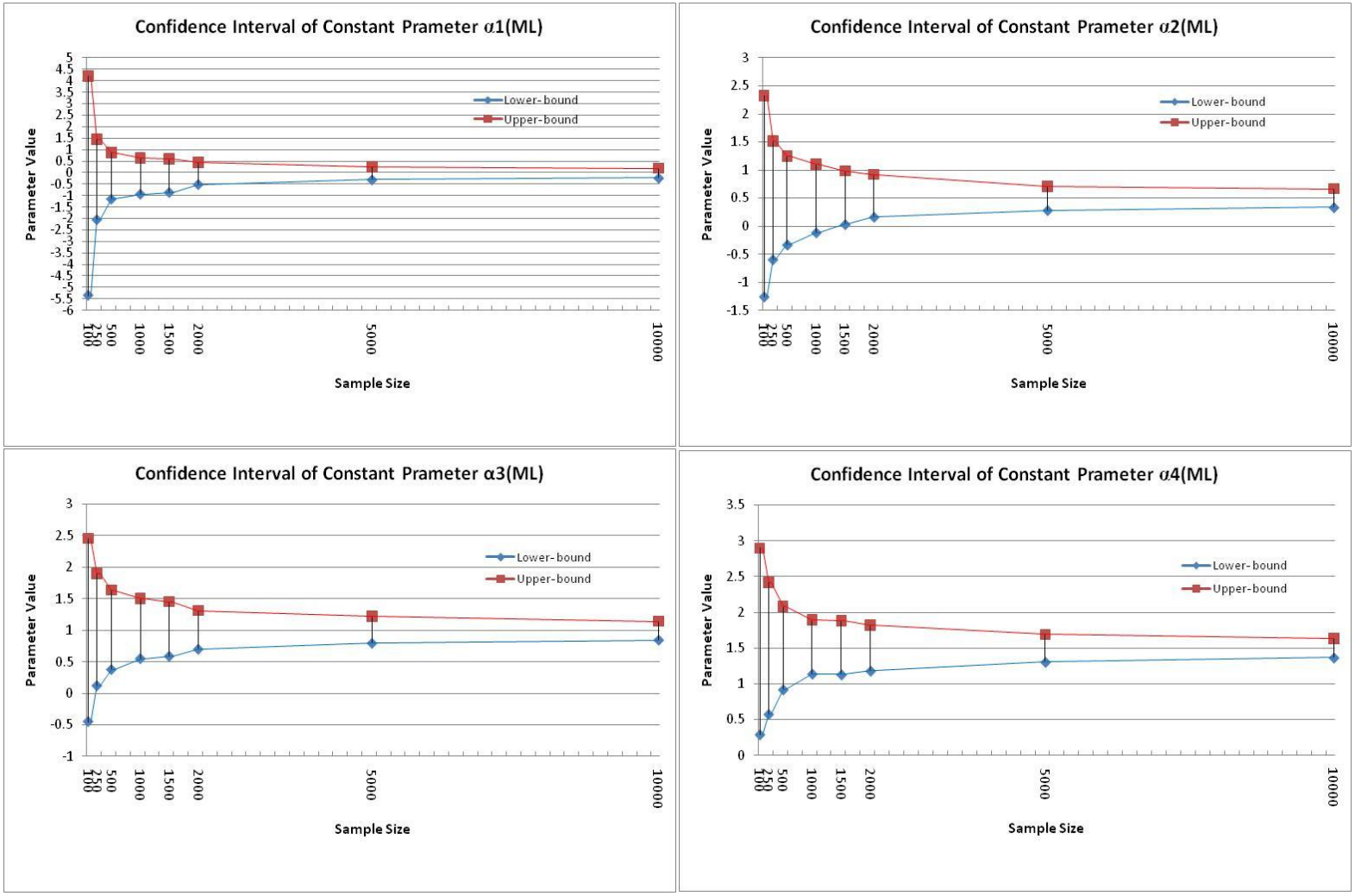

Figure 5.6 Continued I

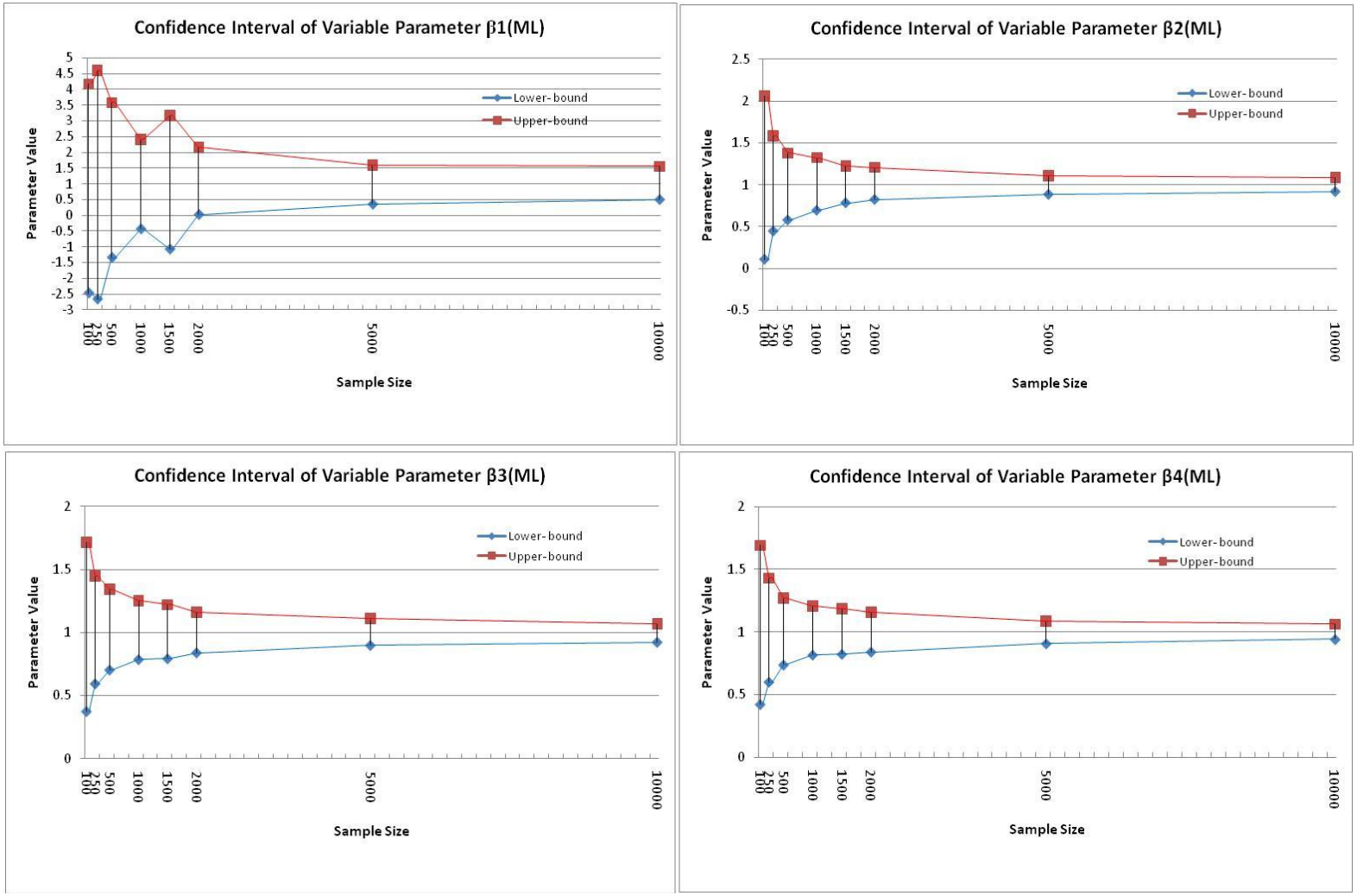

Figure 5.6 Continued II.

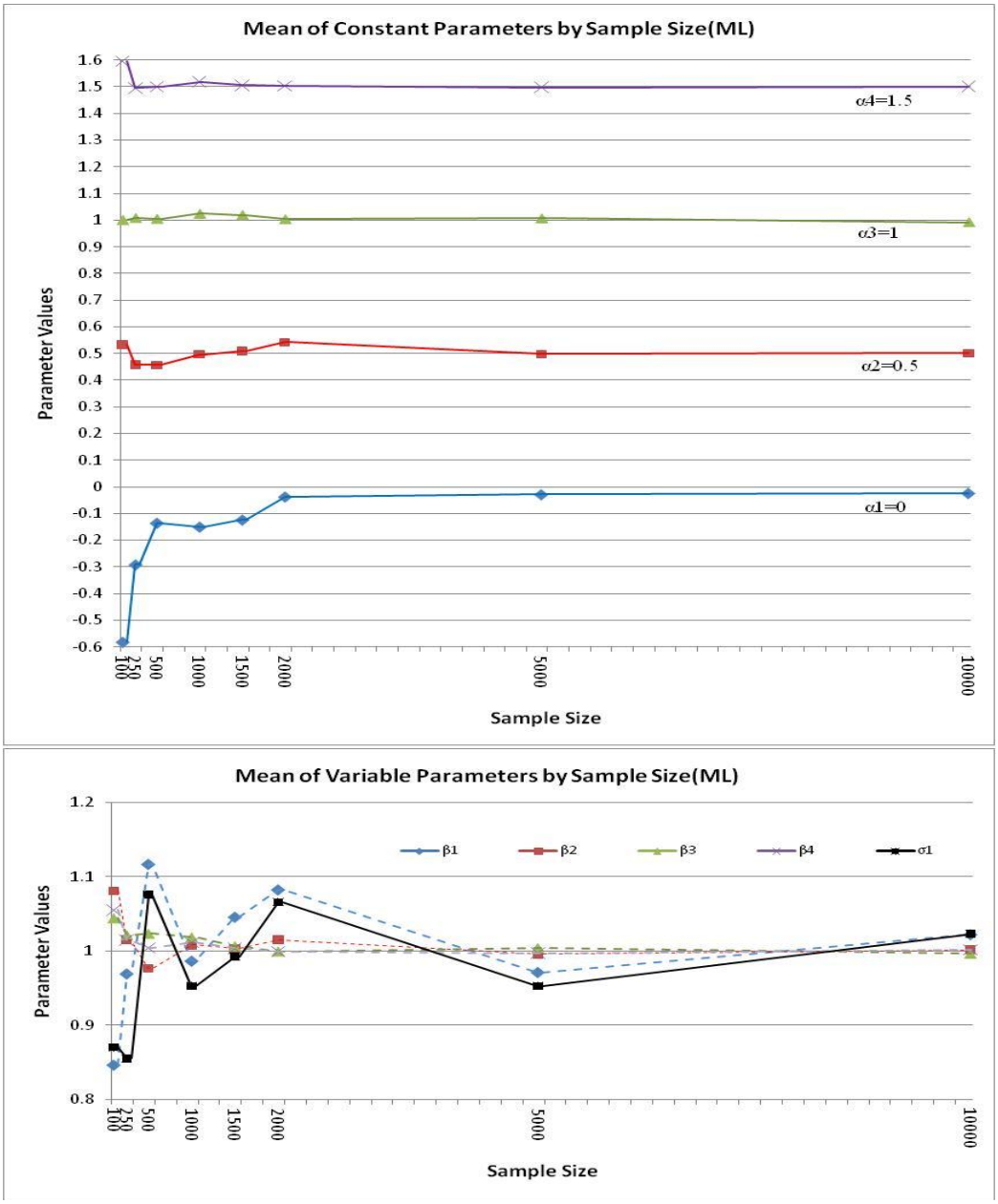

Figure 5.7 Mean of Variable Parameters by Sample Size for the ML

5.1.2.4 Summary Results of the Simulated Data

Although the previously discussed results were based on simulated data, there are still a few findings that could be generalized regarding sample sizes for the three models.

Firstly, crash severity models with sample sizes below 1,000 should not be estimated. In addition, the OP model is the one that requires the least samples (>1000), the ML model is the most demanding on samples (>5000), while the MNL model requirements are intermediate to that of the OP and ML models (>2000).

5.2 Analysis Based on Crash Data

For the sake of simplification, the previous section only included one variable which was assumed to be normally distributed. However, the crash severity data have a large amount of variation which might lead to different sample size requirements for the three models. Therefore, we conducted further analyses of sample size requirements using crash data described in Section 4.1. For this, the models estimated from the full dataset conditions (as estimated in Section 4.2) were set as the baselines. Then, the MNL, ML and OP models were estimated using a stratified sampling method (i.e., crash observations were randomly chosen at a probability based on their severity categories) for different sampling sizes: 100, 500, 2,000, 5,000, 10,000, and 20,000 crashes. The stratified sampling method was used in order to keep the same proportion rates as in the one used for the full dataset: 3.1%, 9.4%, 22.2%, 20.1% and 45.3% for severity of K, A, B, C, O, respectively. In all, 30 random samples were selected for each sample size.

After model estimation results for the 30 estimated models were attained for a specific sample size, the average of the coefficients of each variable was taken. For example, Tables 5.3 through 5.5 are the average values of coefficients and their standard deviation from the 30 model estimations when the 10,000 crash records were used for the MNL, ML and OP models respectively. Then the average values of parameters were compared with those calculated from the baseline conditions (in this case, for the sample size=10,000 for the MNL model, parameters in Table 5.3 were compared with those in Table 4.2). According, the bias, absolute-percentage bias (APB) and root mean square error (RMSE) were attained for each parameter, which will be described in further detail

in following sections. Since there were a large amount of parameters included in each model (27 parameters for the MNL model, 22 parameters for the OP model, and 29 parameters for the ML model), it was impractical to compare the bias based on each parameter as was done in the Monte-Carlo simulation in Section 5.1. Thus, the mean of APB, maximum of APB and total RMSE were estimated as functions of the sample size for each model, and used for comparison.

Table 5.3 Average Estimation Result for the Sample Size=10,000 (MNL)

Severity level / Variable	PDO		Possible injury		Non-incapacitating injury		Incapacitating injury	
	Coef.	**St.d**	**Coef.**	**St.d**	**Coef.**	**St.d**	**Coef.**	**St.d**
Constant	4.277	0.54	3.998	0.51	3.643	0.49	3.088	0.53
Road condition								
log(ADT)	0.165	0.03	0.083	0.02	0.083	0.02		
Speed limit	-0.018	0.01	-0.018	0.01	-0.018	0.01	-0.018	0.01
Crash information								
Night indicator			-0.241	0.04	-0.156	0.04	-0.156	0.04
Dark with light indicator	0.144	0.10						
Rain indicator	-0.945	0.14	-0.834	0.16	-0.531	0.16	-0.386	0.16
Snow indicator	0.481	0.37						
Driver information								
Driver defect indicator	-1.235	0.18	-0.315	0.10	-0.315	0.10	-0.315	0.10
Fatigue indicator	0.403	0.17	-0.242	0.05				
Restraining device used indicator	-2.536	0.08	-1.989	0.12	-1.385	0.07	-0.839	0.08
Fixed-object type information								
Hit tree indicator	-1.028	0.10	-0.820	0.09	-0.599	0.10	-0.342	0.11

* Shaded coefficients were made to be the same across respective crash severity categories.

Table 5.4 Average Estimation Result for the Sample Size=10,000 (ML)

Severity level / Variable	PDO		Possible injury		Non-incapacitating injury		Incapacitating injury	
	Coef.	**St.d**	**Coef.**	**St.d**	**Coef.**	**St.d**	**Coef.**	**St.d**
Constant	4.223	0.56	4.013	0.52	3.612	0.51	3.126	0.55
Road condition								
log(ADT)	0.180	0.03	0.085	0.02	0.085	0.02		
Speed limit	-0.019	0.01	-0.019	0.01	-0.019	0.01	-0.019	0.01
Crash information								
Night indicator			-0.251	0.04	-0.186	0.04	-0.186	0.04
Dark with light indicator	0.152	0.12						
Rain indicator	-0.952	0.14	-0.825	0.17	-1.054	0.26	-0.389	0.16
Std.dev. of distribution					1.606	0.63		
Drive information								
Driver defect indicator	-1.340	0.21	-0.270	0.11	-0.270	0.11	-0.270	0.11
Fatigue indicator	0.445	0.20	-0.314	0.06				
Restraining device used indicator	-.3.65	1.17	-2.007	0.12	-1.227	0.09	-0.844	0.09
Std.dev. of distribution	2.461	1.67						
Fixed-object type information								
Hit tree indicator	-1.145	0.12	-0.852	0.09	-0.541	0.12	-0.357	0.11
Std.dev. of distribution	0.960	0.59						

* Shaded coefficients were made to be the same across respective crash severity categories.

Table 5.5 Average Estimation Result for the Sample Size=10,000 (OP)

Variable	Coef.	St.d
Constant	0.263	0.12
Road Condition		
log(ADT)	-0.052	0.01
Speed limit	0.003	0
Curve & level indicator	0.058	0.02
Crash Information		
Night indicator	-0.111	0.03
Dark with no light indicator	0.055	0.04
Fog indicator	0.099	0.05
Surface condition indicator	-0.258	0.02
Driver Information		
Vehicle type indicator	0.060	0.02
Driver gender indicator	0.132	0.02
Driver defect indicator	0.395	0.05
Restraining device used indicator	0.813	0.04
Fatigue indicator	-0.165	0.06
Airbag deploy indicator	0.455	0.05
Seat belt use indicator	-0.118	0.04
Fixed-object Type Information		
Hit pole indictor	-0.079	0.03
Hit tree indicator	0.184	0.02
Hit fence indictor	-0.164	0.02
Hit barrier indictor	-0.100	0.04
Threshold Parameters		
γ_1	0.562	0.01
γ_2	1.393	0.01
γ_3	2.185	0.02

Based on the 30 estimated models, for each parameter, the bias was calculated as follows, $Bias = E(\hat{\beta}_r) - \beta_{baseline}$ (where r is the number of replications (r=30), and β represents each parameter for the model); the RMSE was calculated using the following equation $RMSE = \sqrt{Bias^2 + Var}$; the APB was computed by dividing the absolute value of bias to the baseline value of each parameter, i.e. $APB = |\beta| / \beta_{baseline}$. Thus, the mean of the APB among all the parameters in a model could be calculated by taking the average of

the APB values of all parameters. Furthermore, the maximum of APB was determined by comparing the APB value of each parameter in a model. Finally, total RMSE was attained by summing up the RMSE value of each parameter for a model. As a summary, the process described above regarding the Monte-Carlo analysis on sample size for crash data is demonstrated in a flowchart, as shown in Figure 5.8. Furthermore, the values of the three criteria (the mean of APB, maximum of APB and total RMSE) for various sample sizes are listed in Table 5.6.

Table 5.6 Three Criteria by Sample Size of Crash Data for the Three Models

Sample Size	Mean of APB			Max of APB			Total RMSE		
	MNL	ML	OP	MNL	ML	OP	MNL	ML	OP
100	5.50E+13	2.10E+11	143%	9.70E+14	2.90E+12	2.10E+01	7.40E+15	1.60E+13	20.7
500	2.00E+14	1.10E+04	25%	4.50E+15	1.10E+05	94%	1.30E+16	1.20E+06	4.5
2000	16%	26%	11%	45%	167%	40%	12.9	28.7	2.2
5000	9%	13%	5%	27%	52%	20%	7.6	13.7	1.2
10000	4%	5%	4%	13%	13%	14%	4.7	8.7	0.7
20000	2%	3%	2%	9%	21%	9%	1.9	3.4	0.4

Note: Smaller values in the table indicate better estimations with less bias.

From Table 5.6, we note the following results:

(1) As expected, all three models show the same tendency indicated as the simulated data: the increase in sample size leads to the reduction in all three criteria (mean of APB, max of APB and total RMSE), improving the accuracy of model estimation.

(2) In terms of the values of all three criteria, the MNL and ML models are more sensitive to small sample sizes than the OP model and this is especially noticeable

for the sample sizes equal to 100 and 500. Nonetheless, for a sample size below 500, all models perform poorly.

(3) Similar to the results shown in the previous section, the ML model needs a lot of data to lower the value of three criteria. Even at 5,000 observations, the mean of APB, max of APB and total RMSE for the ML model is still twice as large as those for the MNL model.

(4) According to the three criteria, the minimum sample size for the OP, MNL and ML models should be 2,000, 5,000 and 10,000, respectively. At that point, the estimated values become very close to the "true" values for all three criteria. In short, these findings are consistent with those found with the simulated data about which models are more affected by the small sample size problem. However, the minimum numbers are larger than the ones proposed in simulation. This may be partly explained by the large variability of crash data and the number of random samples running (30 for each sample size).

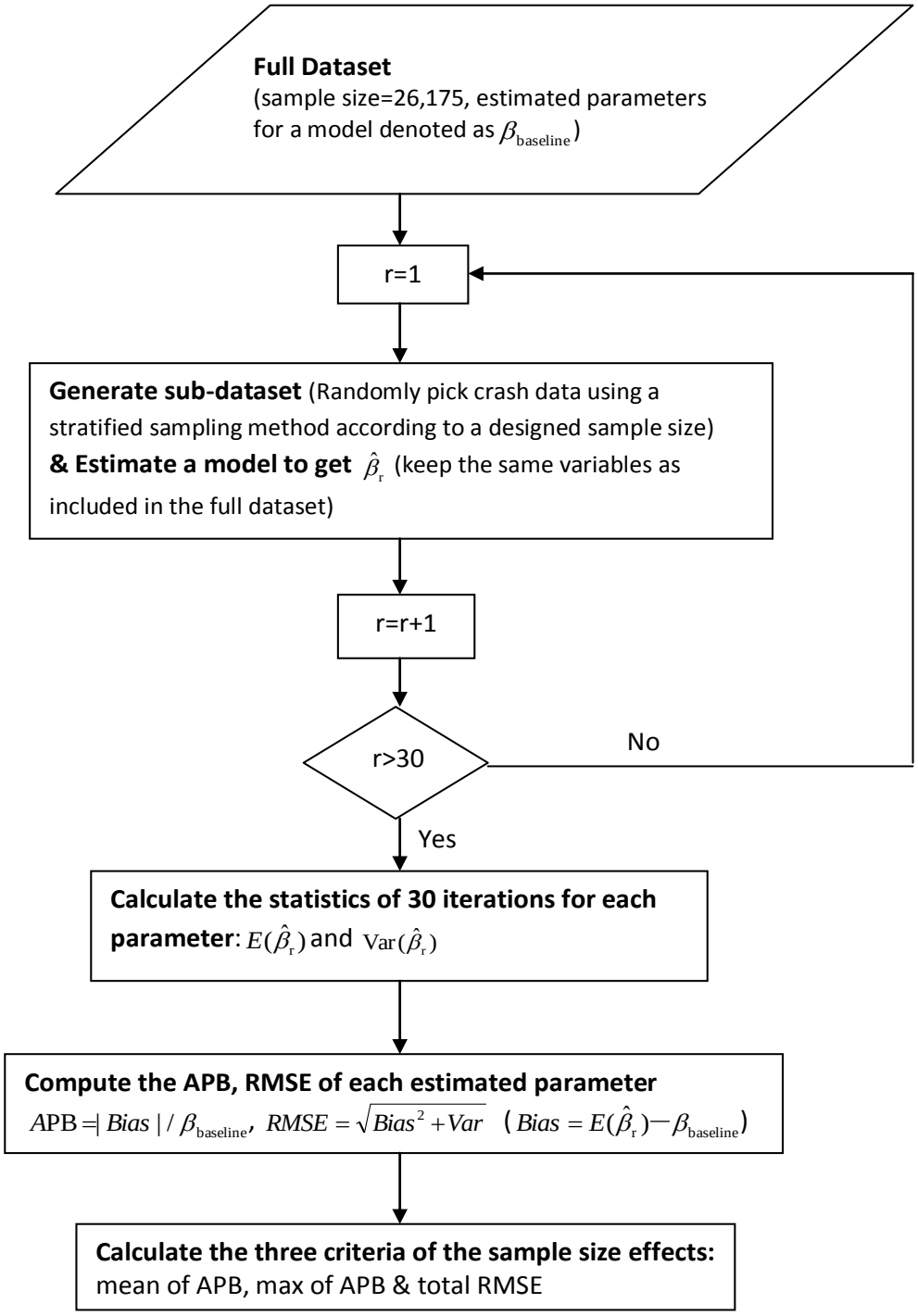

Figure 5.8 Monte-Carlo Analysis on Sample Size for Crash Data

5.3 Chapter Summary

Although there have been a lot of studies that have documented the application of crash severity models, no research has been conducted about quantifying the sample size requirements for crash severity models. Similar to count data models, small data sets could significantly influence model performance for crash severity models. The objective of this chapter consisted in examining and quantifying the effects of different sample sizes on the performance of the three most commonly used crash severity models: the MNL, OP and ML models. The objective of this study was accomplished by using a Monte-Carlo analysis based on simulated data and observed crash data. The sample size investigated varied between 100 and 10,000 observations.

The results from the simulated data and random samples drawn from 26,175 crash records, are consistent with prior expectations in that small sample sizes significantly affect the development of crash severity models, no matter which model type is used. Furthermore, among the three models, the ML model requires the largest sample size, while the OP model requires the lowest sample size. The sample size requirement for the MNL model is intermediate to the OP and ML models. Overall, the recommended absolute minimum numbers of observations for the OP, MNL, and ML models are 1,000, 2,000 and 5,000, respectively. Although those values are recommended guidelines, larger datasets should be sought, as demonstrated by the analysis using observed crash data (larger variability in the crash data or more randomness estimated in the ML model). In order to minimize the bias produced by the insufficient sample size, the sequence of selecting a model among the three ones is OP, MNL and ML as mentioned previously.

CHAPTER VI

COMPARISONS BY MODEL MISSPECIFICATION

In Chapter V, the sample size requirements for the three most commonly used models were investigated. The comparison was accomplished by a Monte-Carlo analysis based on simulated and observed crash data. The analysis results from both data sources indicated that each model had different requirements for sample size in order to get reasonable estimation results. However, even with sufficient sample sizes, a bias of the model estimation results might still exist due to a model misspecification issue. Since the performance of model estimation is expected to depend on the accuracy of the underlying model specification, the purpose of this chapter is to investigate the amount of bias that exists when the specified model is not the true (or exact) one and does not reflect the actual characteristics of the data. When a dataset, which comes from one of the three model structures (the MNL, OP and ML models), is fit by some other model rather than the true one, it is expected to observe a bias in the model estimation. For instance, when using the MNL model to fit the MNL data (MNL data refers to a dataset generated according to a MNL model, such as the dataset simulated in Section 5.1.1), the estimated coefficients would be virtually unbiased. However, when the OP model is fit to the MNL data, it is expected to see some bias in the estimated probability of each crash severity level because the specified model deviates from the real underlying data structure, and vice-versa.

It seems obvious to predict the existence of a bias in model estimation when the model structure is misspecified. However, what is the extent of the bias caused by model misspecification? Do different misspecified models lead to the same amount of bias for a dataset? No research study has been done to address these questions regarding the

misspecification issue among the three models. Therefore, this chapter attempts to answer the above questions based on the simulated datasets designed in Section 5.1.1.

This chapter is divided into three sections. Section 6.1 describes the simulation design and the criteria for estimating the effects of model misspecification on the three models. In Section 6.2, bias due to model misspecification for each of the simulated datasets is compared for the three models. Section 6.3 provides a summary of the chapter.

6.1 Simulation Design and Estimation Criteria

As with the analysis of effects of sample size on the three models in Chapter V, this chapter examines these models, comparing the bias caused by model misspecification, to get a sense of which model has less bias based on various model misspecification scenarios. The results will provide researchers with a model selection criterion among the three models in terms of model misspecification. In order to investigate the models based on model misspecification, it is important to know what model structure the data has. However, it is impossible for researchers to know the true model structure of the observed crash data (model selection among the entire candidate models could only help to find the best fit model for a dataset, rather than the true one). Therefore, rather than observed crash data, simulated data, with the knowledge of true values of estimators and true propensity functions, are used to explore the bias caused by model misspecification.

In terms of the approach of estimating the bias caused by model misspecification, the comparison of individual parameters between the estimated values and the true ones, as used for the sample size effects in Chapter V, is not suitable here. The comparison in this case is across different types of models in terms of model misspecification, and parameters for various models could not be compared directly since each model has different model structures and variables. For instance, when an OP model fits a MNL dataset, the estimated parameters cannot be compared with the true ones directly to get

the bias caused by model misspecification, since the MNL model and the OP model have different parameters due to different model structures. For the OP model, parameters of an individual variable are fixed to be the same across the five crash severity levels (i.e., one variable has only one parameter), while the MNL model has different values of parameters for a variable across the five crash severity levels. Therefore, to overcome the limitations associated with the measurement of bias for parameters, the estimated probabilities and true probabilities of each crash severity level were calculated, based on a specified value of independent variable. The probabilities of the five crash severity levels estimated in each model would well represent the performance of model estimation, and accordingly could be used in the quantification of the bias of model estimations due to model misspecification.

The simulated datasets used for the analysis of model misspecification are the same as these used in Section 5.1 for sample size analysis for the three models. The true values of parameters that were designed for the three models were listed in Table 5.1, excluding the information of sample size. According to the result of Monte-Carlo simulation on sample size in Section 5.1.2, sample size equal to 10,000 was used in this chapter which was guaranteed to be sufficient for all the three models, since they were larger than the required sample sizes (the OP model >1000, the MNL model >2000, and the ML model >5000). In other words, it is assumed that using 10,000 samples in the model estimation would eliminate bias caused by sample size limits for the simulated dataset.

For each of the three models, in order to estimate the probability of each crash severity level, it is necessary to designate values of independent variable x. Three different values of the independent variable x were used for the probability calculation for each model (i.e., x = -2, -1, 0 for the MNL and ML models, and x = 2.2, 1, 0 for the OP model). Recall x was drawn from a normal distribution with mean equal to -2 and variance equal to 1 for the MNL and ML model, and x was drawn from a normal distribution with mean equal to 2.2 and variance equal to 1 for the OP model. It could be

noticed that the three selected x values were the ones across the distribution of x in order to cover different distribution of the probabilities for five crash severity levels.

Each dataset was generated 100 times for each model as designed in Section 5.1.1; the estimated probability of each crash severity level was calculated based on both the estimated parameters from the 100 repetitions and the designated value of the independent variable. Therefore, 100 estimated probabilities of each crash severity level for each model specification were obtained. In addition, summary statistics such as the mean and the standard deviation for these probabilities of each crash severity level can be calculated according to the estimated probabilities. This provides the bias of the probabilities of each level $Bias = E(\hat{\Pr}ob_{\mathrm{r}}) - \Pr ob_{true}$, where r is the number of replications. In addition, absolute-percentage bias (APB) and root mean square error (RMSE) of each probability were also calculated for comparison. The APB was computed by dividing the absolute value of bias of the probability to the true probability value of each level, i.e. $APB = |\Pr ob| / \Pr ob_{true}$. Meanwhile, $RMSE = \sqrt{Bias^2 + Var}$, which combined both of the bias and variability, was used to compare the estimated probabilities for different misspecified models. According to the above calculation, the mean of the APB among all the five probabilities of each crash severity levels for a misspecified model could be calculated by taking the average of the APB values of all the five probabilities. Furthermore, the maximum of APB was found by comparing the APB value of each probability in a model. Lastly, total RMSE was calculated by summing up the RMSE value of the five probabilities in a model. Therefore, the mean of APB, max of APB and total RMSE of the five estimated probabilities were used as three indexes to quantify the effects of model misspecification. The process of the Monte-Carlo analysis on model misspecification for simulated data is also described in a flowchart, as shown in Figure 6.1. Also, part of the code in NLOGIT used to carry out the Monte-Carlo analysis on the simulated data for a model misspecification are provided in Appendix B.

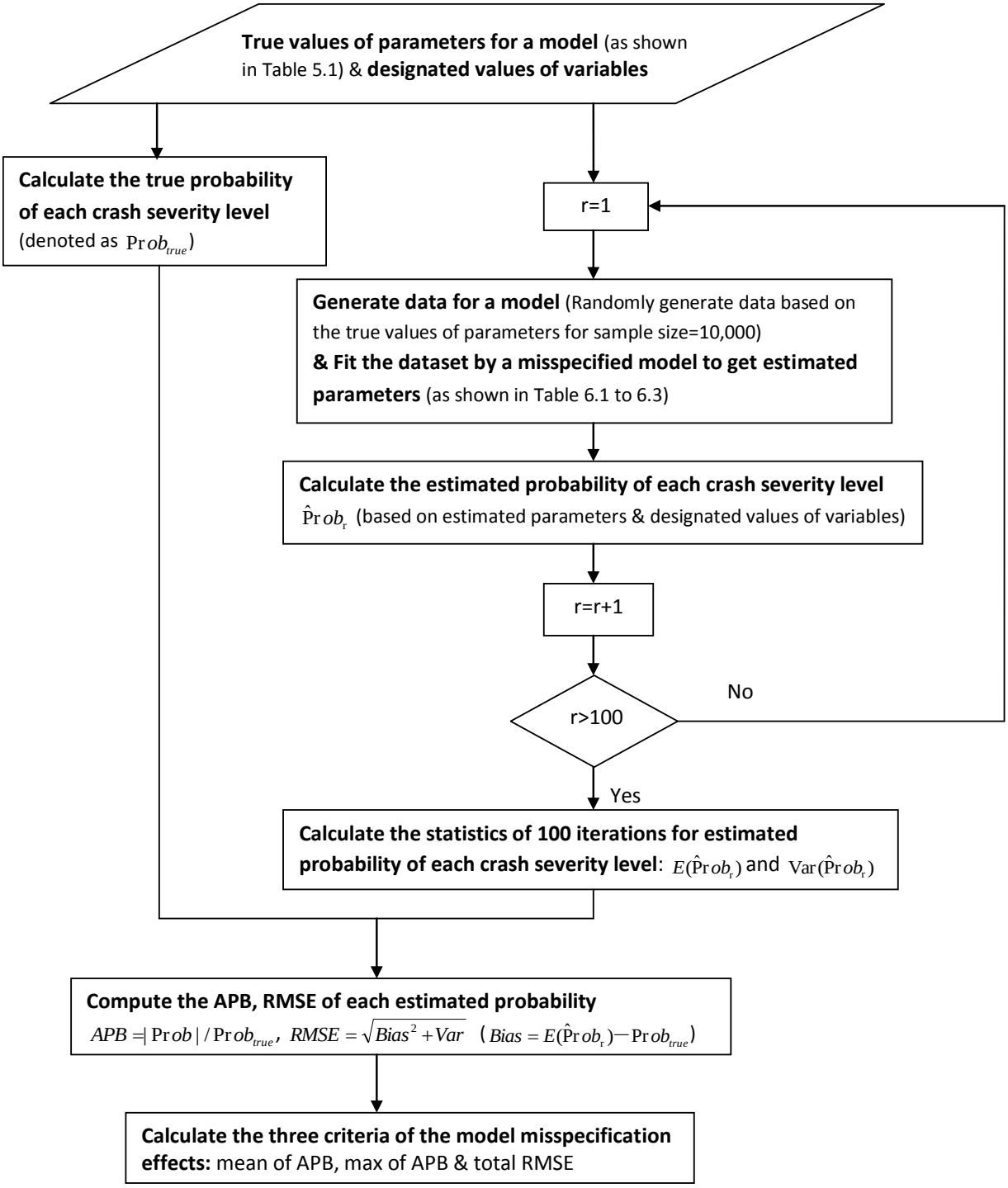

Figure 6.1 Monte-Carlo Analysis on Model Misspecification for Simulated Data

6.2 Simulation Results

In order to compare the bias due to model misspecification among the three models, for each dataset simulated for one model, another particular model (between the rest two models) was used to fit the simulated dataset. Recall the primary assumption that was made before, that the results would not be affected much by different designed values of the parameters. This assumption indicates here that the effects of model misspecification are independent of the datasets used for analysis for each model. Thus, it is reasonable to compare the model misspecification results by three different datasets for each model. Another point that needs to be mentioned is that the estimation of the ML model is more complicated than the other two models because the randomness of the model needs to be designated beforehand. Therefore, different assignments of random parameters and their distribution would lead to different model estimation results. In order to simplify the problem, one parameter β1 (the variable parameter for level 1) was designated as a random parameter, following a normal distribution. The estimation results for the three simulated datasets (the MNL, OP and ML data) are shown in Tables 6.1, 6.2 and 6.3 respectively.

Table 6.1 Model Estimation Results for the OP and ML Models for a MNL Data

Model Parameter		Parameter Values				
		MNL(true)	OP(estimated)		ML(estimated)	
			mean	st.d	mean	st.d
Constant Parameter*	α1	0	0.835	0.027	0.022	0.11
	α2	0.5	0.563	0.015	0.508	0.088
	α3	1	1.112	0.019	1.005	0.08
	α4	1.5	1.836	0.022	1.505	0.077
Variable Parameter	β1	1	-0.427	0.011	N(1.036,0.103)	(0.079,0.116)
	β2	1			1.004	0.041
	β3	1			1.004	0.041
	β4	1			1.003	0.036

* Constant parameters for the OP model are represented by γ1- γ4, which are the threshold variables for each level.

Based on the true and estimated parameters listed in Tables 6.1 to 6.3, the true and estimated probabilities of each level in terms of the three designated independent variables were calculated using Equations (3.2), (3.7) and (3.9) for the MNL, OP and ML models respectively. For the ML model, since some parameters of a variable for certain levels are not fixed, following some types of distribution, the probabilities of each level using Equation (3.9) are not easy to calculate. The probabilities were approximated using a numerical method as proposed by Train [4] (2003).

---

4 The probabilities of each outcome for the ML model is calculated by Equation (3.9):

$$P_i(k) = \int \frac{\exp[\beta_k X_{ki}]}{\sum_{\forall k} \exp(\beta_k X_{ki})} f(\beta \mid \theta) d\beta$$

Where, $\frac{\exp[\beta_k X_{ki}]}{\sum_{\forall k} \exp(\beta_k X_{ki})}$ is the logit function, represented by $L_i(\beta_k)$;

k is the indication for five outcomes, k=1,2,3,4,5.

The procedure of computing the probabilities of each outcome is demonstrated by an example, the calculation of the true probabilities of each outcome for a simulated ML data. As designed, for the ML model β_1 follows a normal distribution N(1,1), rather than a fixed value. Therefore, the probabilities of each outcome are approximated by simulation due to the required numerical integration of the logit function over the distribution of the random parameter in Equation (3.9). The procedure is listed as below:

(1) Randomly draw a value of β_1 from $f(\beta_1)$ which is N(1,1), and β_2 , β_3 , β_4 are fixed value as designed or estimated from each repetition. Label it β^r , with the superscript r=1 referring to the first draw.

(2) Calculate the logit function $L_i(\beta_k^r)$ with this draw.

(3) Repeat step 1 and 2 many times (we use 1000 times in the study), and average the results. The average value of $L_i(\beta_k^r)$ is taken as the simulated probability of outcome k: $\tilde{P}_i(k) = \frac{1}{R}\sum_{r=1}^{R} L_i(\beta_k^r)$, where R is the number of draws (R=1000 here), $\tilde{P}_i(k)$ is an unbiased estimator of $P_i(k)$ by construction.

Table 6.2 Model Estimation Results for the MNL and ML Models for an OP Data

Model Parameter		Parameter Values				
		OP(true)	MNL(estimated)		ML(estimated)	
			mean	st.d	mean	st.d
Constant Parameter*	γ1	0	4.584	0.129	4.666	0.143
	γ2	0.8	4.04	0.11	4.061	0.112
	γ3	1.5	3.397	0.1	3.404	0.1
	γ4	2.4	2.532	0.09	2.533	0.09
Variable Parameter	β1	1	-3.693	0.08	N(-3.882,0.396)	(0.155,0.174)
	β2		-2.69	0.057	-2.703	0.059
	β3		-2.004	0.049	-2.008	0.05
	β4		-1.277	0.037	-1.277	0.037

*Constant parameters for the MNL and ML models are represented by α1- α4.

Table 6.3 Model Estimation Results for the MNL and OP Models for a ML Data

Model Parameter		Parameter Values				
		ML(true)	MNL(estimated)		OP(estimated)	
			mean	st.d	mean	st.d
Constant Parameter*	α1	0	-0.321	0.088	0.631	0.022
	α2	0.5	0.466	0.083	0.324	0.011
	α3	1	0.957	0.075	0.744	0.016
	α4	1.5	1.466	0.068	1.372	0.018
Variable Parameter	β1	N(1,1)	0.308	0.036	-0.236	0.011
	β2	1	0.984	0.042		
	β3	1	0.979	0.037		
	β4	1	0.984	0.03		

*Constant parameters for the OP model are represented by γ1- γ4, which are the threshold variables for each level.

Both of the true probabilities and the estimated ones of all five levels are shown in Tables 6.4, 6.5, and 6.6 for the MNL, OP and ML data respectively. Furthermore, comparing the true probabilities of each level with the estimated ones, the bias, APB and

RMSE of the probabilities of each level were computed, which are also listed in Tables 6.4 to 6.8. In addition to the three indexes of the effects of model misspecification, the mean value of APB, maximum value of APB, and total RMSE are also summarized in Tables 6.7 and 6.8.

Table 6.4 Model Misspecification for the MNL Data

Probability of	True Value	OP model			ML model		
		mean	st.d	bias	mean	st.d	bias
	x = -2						
Level 1	5.8%	4.6%	0.2%	-1.2%	5.8%	0.3%	0.0%
Level 2	9.6%	8.4%	0.2%	-1.1%	9.6%	0.3%	0.0%
Level 3	15.8%	15.2%	0.3%	-0.6%	15.8%	0.4%	0.0%
Level 4	26.0%	27.7%	0.4%	1.7%	26.0%	0.4%	0.0%
Level 5	42.9%	44.2%	0.5%	1.3%	42.9%	0.6%	0.1%
	x = -1						
Level 1	8.0%	10.3%	0.4%	2.4%	7.9%	0.4%	0.0%
Level 2	13.1%	13.9%	0.4%	0.8%	13.2%	0.5%	0.0%
Level 3	21.6%	19.8%	0.4%	-1.8%	21.6%	0.6%	0.0%
Level 4	35.7%	27.7%	0.4%	-8.0%	35.7%	0.7%	0.0%
Level 5	21.6%	28.3%	0.5%	6.7%	21.6%	0.7%	0.0%
	x = 0						
Level 1	9.2%	20.2%	0.8%	11.0%	9.4%	0.8%	0.2%
Level 2	15.2%	19.1%	0.5%	3.9%	15.2%	0.9%	0.0%
Level 3	25.1%	21.6%	0.5%	-3.5%	25.0%	1.1%	0.0%
Level 4	41.3%	23.3%	0.4%	-18.0%	41.2%	1.2%	-0.1%
Level 5	9.2%	15.8%	0.6%	6.6%	9.2%	0.6%	-0.1%

Note: Smaller values in the table indicate better estimations with less bias.

Table 6.5 Model Misspecification for the OP Data

Probability of	True Value	MNL model			ML model		
		mean	st.d	bias	mean	st.d	bias
	x = 2.2						
Level 1	1.4%	1.3%	0.1%	-0.1%	1.4%	0.2%	0.0%
Level 2	6.7%	6.6%	0.3%	0.0%	6.6%	0.3%	-0.1%
Level 3	16.1%	15.8%	0.5%	-0.3%	15.8%	0.5%	-0.4%
Level 4	33.7%	32.9%	0.5%	-0.8%	32.9%	0.5%	-0.8%
Level 5	42.1%	43.4%	0.6%	1.3%	43.4%	0.6%	1.3%
	x = 1						
Level 1	15.9%	16.4%	0.7%	0.6%	15.8%	0.8%	-0.1%
Level 2	26.2%	26.0%	0.8%	-0.2%	26.3%	0.9%	0.1%
Level 3	27.1%	27.1%	0.8%	0.1%	27.3%	0.8%	0.3%
Level 4	22.8%	23.7%	0.8%	0.9%	23.8%	0.8%	1.0%
Level 5	8.1%	6.7%	0.3%	-1.3%	6.8%	0.3%	-1.3%
	x = 0						
Level 1	50.0%	49.4%	2.3%	-0.6%	51.1%	2.6%	1.1%
Level 2	28.8%	28.7%	1.7%	-0.1%	27.9%	1.8%	-0.9%
Level 3	14.5%	15.1%	1.1%	0.6%	14.5%	1.2%	0.0%
Level 4	5.9%	6.4%	0.5%	0.5%	6.1%	0.6%	0.2%
Level 5	0.8%	0.5%	0.1%	-0.3%	0.5%	0.1%	-0.3%

Note: Smaller values in the table indicate better estimations with less bias.

Table 6.6 Model Misspecification for the ML Data

Probability of	True Value		MNL model			OP model		
	mean	st.d	mean	st.d	bias	mean	st.d	bias
	x = -2							
Level 1	14.7%	20.1%	15.2%	0.4%	0.5%	13.5%	0.4%	-1.2%
Level 2	8.7%	2.0%	8.6%	0.3%	-0.1%	8.3%	0.3%	-0.4%
Level 3	14.3%	3.4%	14.2%	0.4%	-0.1%	14.2%	0.4%	-0.1%
Level 4	23.5%	5.5%	23.4%	0.5%	-0.2%	24.6%	0.5%	1.1%
Level 5	38.8%	9.1%	38.6%	0.5%	-0.2%	39.4%	0.4%	0.6%
	x = -1							
Level 1	11.0%	9.8%	11.3%	0.5%	0.3%	19.3%	0.5%	8.3%
Level 2	12.7%	1.4%	12.6%	0.5%	-0.1%	10.1%	0.3%	-2.6%
Level 3	20.9%	2.3%	20.7%	0.6%	-0.2%	15.8%	0.4%	-5.2%
Level 4	34.5%	3.8%	34.2%	0.7%	-0.2%	24.2%	0.5%	-10.3%
Level 5	20.9%	2.3%	21.2%	0.6%	0.2%	30.7%	0.5%	9.8%
	x = 0							
Level 1	9.2%	0.0%	7.1%	0.5%	-2.1%	26.4%	0.7%	17.2%
Level 2	15.2%	0.0%	15.5%	0.9%	0.3%	11.5%	0.4%	-3.7%
Level 3	25.1%	0.0%	25.4%	1.2%	0.3%	16.6%	0.4%	-8.5%
Level 4	41.3%	0.0%	42.2%	1.3%	0.9%	22.5%	0.4%	-18.8%
Level 5	9.2%	0.0%	9.8%	0.5%	0.5%	23.0%	0.7%	13.7%

Note: Smaller values in the table indicate better estimations with less bias.

Table 6.7 APB ($\times 10^{-3}$) of Model Specification for the Three Models

Probability of	MNL data		OP data		ML data	
	OP model	ML model	MNL model	ML model	MNL model	OP model
	x= -2		x= 2.2		x= -2	
Level 1	214	3	91	6	33	79
Level 2	117	0	5	13	6	41
Level 3	38	1	21	23	5	7
Level 4	64	1	25	25	7	45
Level 5	30	1	32	31	5	14
mean APB	93	1	35	20	11	37
max APB	214	3	91	31	33	79
	x= -1		x= 1		x= -1	
Level 1	300	3	36	6	28	757
Level 2	59	2	7	5	7	206
Level 3	85	0	2	10	10	247
Level 4	224	0	38	43	7	299
Level 5	308	1	165	161	11	467
mean APB	195	1	50	45	13	395
max APB	308	3	165	161	28	757
	x= 0					
Level 1	1190	18	13	21	231	1863
Level 2	257	2	4	31	23	240
Level 3	138	2	40	1	14	339
Level 4	437	2	85	35	22	454
Level 5	718	6	382	411	58	1491
mean APB	548	6	105	100	69	878
max APB	1190	18	382	411	231	1863

Note: Smaller values in the table indicate better estimations with less bias.

Table 6.8 RMSE ($\times 10^{-3}$) of Model Specification for the Three Models

Probability	MNL data		OP data		ML data	
	OP model	ML model	MNL model	ML model	MNL model	OP model
	x = -2		x = 2.2		x = -2	
Level 1	13	3	2	2	6	12
Level 2	11	3	3	3	3	5
Level 3	7	4	6	6	4	4
Level 4	17	4	10	10	5	12
Level 5	14	6	15	14	5	7
Total RMSE	62	19	35	35	24	39
	x = -1		x = 1		x = -1	
Level 1	24	4	9	8	6	83
Level 2	8	5	8	9	5	26
Level 3	19	6	8	9	7	52
Level 4	80	7	12	12	8	103
Level 5	67	7	14	13	6	98
Total RMSE	198	27	51	52	31	362
	x = 0					
Level 1	110	8	24	28	22	172
Level 2	39	9	17	20	10	37
Level 3	35	11	13	12	12	85
Level 4	181	12	7	6	16	188
Level 5	66	6	3	3	7	138
Total RMSE	431	46	64	69	67	619

Note: Smaller values in the table indicate better estimations with less bias and variability.

According to Tables 6.4 through 6.8, it can be noted that as value of independent variable x moves farther away from its mean, not only the bias but also the RMSE, total RMSE, mean APB and max APB of probabilities tends to be larger for each simulated dataset in terms of model misspecification. In another word, bias and variability caused by model misspecification were dependent on the value of the independent variable, and the smallest bias and variability were achieved at the mean value of independent variable x. However, no matter what the values of independent variable were, there were some

common tendencies indicated from the above tables for model misspecification. They are listed as follows.

(1) The smallest values of the three indexes existed for the MNL data estimated by the ML model. This is reasonable since the ML model is a generalized version of the MNL model, including randomness in the parameters. That is, among the three models, the ML model has the smallest effect on a model misspecification.

(2) When the ML data was fit by the MNL model, all the three indexes were slightly higher than those from the MNL data when applied by the ML model, though still smaller compared to those from other misspecification scenarios. This is because that the MNL model is a specified version of the ML model. However, keeping the logit structure of the ML data for the estimated model led to the three indexes smaller than those from other misspecification scenarios.

(3) The largest values of the three indexes occurred when the ML data was estimated by the OP model, and the second largest occurred when the MNL data estimated by the OP model. This was probably due to the difference between logit structure of the data and probit specification of the estimated model. In addition, the OP model could not take into account the randomness of parameters for the ML model, which might cause more bias and variability in the model estimation when the ML or MNL data was estimated by the OP model.

(4) When the OP data estimated by the MNL and ML models, the three indexes were much smaller than those from the MNL and ML data when applied by the OP model. Combined with what found in point (3), it seemed that logit specification (both the MNL and ML models) could estimate the probit data (data from the OP model) well to some degree, but not vice-versa. Thus, when applying the OP model to a dataset, it should be kept in mind about whether the true model structure of the data follows an

ordered structure. If not, large bias and variability might exist as the data came from logit structure (the MNL or ML).

What can be concluded from the above findings is that, in terms of model misspecification, without knowing the true underlying structure of the data, logit models (the MNL and ML models) are probably more accurate in model estimation than the OP model. In addition, in terms of the effects of model misspecification, between the MNL and ML models, the ML model is a better choice since it is more general than the MNL model, allowing for randomness in the parameters.

6.3 Chapter Summary

This chapter described the effects of model misspecification for the three models, with an in-depth analysis of their performances, in order to provide traffic safety researchers with another criterion for selecting models based on model misspecification. Simulated datasets described in Section 5.1.1 were used for analysis in this chapter. Due to the lack of the exact characteristics of observed crash data, such as the true underlying model structure of the crash data, the analysis could not be extended to real crash data. By applying a misspecified model to a simulated dataset, the estimated probabilities of each crash severity level were attained based on the estimated parameters and designated values of the independent variable. Accordingly, the bias and variability of the estimated probabilities could be calculated by comparing the estimated probabilities with the true known values. In addition, the mean value of APB, maximum value of APB, and total RMSE of the probabilities were computed as the three indexes of the effects of model misspecification.

The results indicated that the performance of the estimated models depended on the accuracy of the underlying model specification. When the sample size was sufficient, specifying the probit structure to a logit data produced the largest bias and variability.

However, the logit models could estimate the probit data much better than using the probit models to estimate the logit data. Furthermore, for the logit models, misspecifying the ML model to a MNL dataset would probably lead to smaller bias and variability than those from misspecifying the MNL model to a ML dataset. This is because the ML model is a generalized model of the MNL model, allowing randomness in the parameters. Overall, in terms of the effects of model misspecifiction, without knowing the real model structure of a crash dataset, a suggested sequence of selecting crash severity model among the three models is: first the ML model, then the MNL model, and finally the OP model.

CHAPTER VII

MODEL COMPARISONS BY DATA UNDERREPORTING

As discussed in Chapter II, although a significant amount of work has been devoted on developing crash severity models to predict the probabilities of crashes for different severity levels, very few studies have considered the underreporting problem in the modeling process. After comparing the effects of sample size and model misspecification on the three commonly used crash severity models, the effects of data underreporting on the three models are explored in this chapter. Crash data is usually based on police reported crash data. It has been well documented that crashes are often unreported, particularly those associated with lower severity levels. This underreporting issue can yield to significant biases in the inferences about a population of interest if crash data are treated as random samples coming from the population without considering the different unreported rates for each crash severity level.

The primary objective of this chapter is to examine the effects of underreporting for the three models. More specifically, this study investigated how each of these models performs for different unreported rates. A secondary objective consisted in quantifying how the outcome-based sampling method, via the Weighted Exogenous Sample Maximum Likelihood Estimator (WESMLE), could account for specific underreporting conditions when the transportation safety analyst had the full or partial knowledge of unreported rates for different severities. The study objectives were accomplished via a Monte-Carlo approach using simulated and observed crash data.

This chapter consists of three sections. Section 7.1 describes the results for the three types of models by various underreported data generated by simulation. Section 7.2

further presents the modeling results for the three models using observed crash data. Section 7.3 provides the summary of this chapter.

7.1 Analysis Based on Simulated Data

In order to study the effects of underreporting on three models and verify the effectiveness of the WESMLE method for underreported data, a Monte-Carlo approach was developed using simulated and observed crash data. Since simulated data are controllable, we developed a couple of scenarios to simulate various underreporting cases in the dataset.

7.1.1 Simulation design and estimation criteria

The datasets generated in Section 5.1.1 for the three models based on the true parameters were treated as the complete datasets, i.e., the population. The underreported dataset were replicated by randomly eliminating some data according to the designed unreported rates. In order to generate sufficient samples even after the random removal of some data, the original sample size was set to be 50,000. In other words, the complete datasets had 50,000 observations for three models and all the eliminated observations were considered to be the unreported ones (Recall in Chapter VI for the model misspecification study, we used sample size=10,000 which was guaranteed to be sufficient for all the three models. Since in this chapter we randomly eliminate some data to simulate underreporting, we have enlarged the sample size).

In this part of the study, we developed three scenarios to evaluate the change in bias and variability of estimated parameters for different unreported rates for the three models. The objective of Scenario 1 is to verify the effects of the number of unreported observations on the bias and variability of estimated parameters, and the function of WESMLE based on various unreported rates for all three models. The goal of Scenario 2 is to further examine the function of the WESMLE when underreporting exists in all five

levels. The aim of Scenario 3 is to evaluate the function of the WESMLE when using incorrect unreported rates. More details of each scenario are described in Section 7.1.2.

Datasets for each model were repeatedly drawn 100 times for each designated unreported rate, according to the designed true parameter values of the model. Based on the 100 estimated models, the bias of each parameter was calculated as $Bias = E(\hat{\beta}_r) - \beta_{baseline}$, where r was the number of replications (r=100), and β represented each parameter in the model (both constant parameters and variable parameters). The root-mean-square-error (RMSE) of each parameter for a model was calculated using the equation $RMSE = \sqrt{Bias^2 + Var}$, and the total RMSE of all the variable parameters for each model was used to measure the underreporting effects since it comprised both bias and variability. As a summary, the whole process described above about the Monte-Carlo analysis on the effects of underreporting for simulated data is shown in Figure 7.1. Meanwhile, parts of the code in NLOGIT used to carry out the Monte-Carlo analysis on the simulated data for underreporting in data are provided in Appendix C.

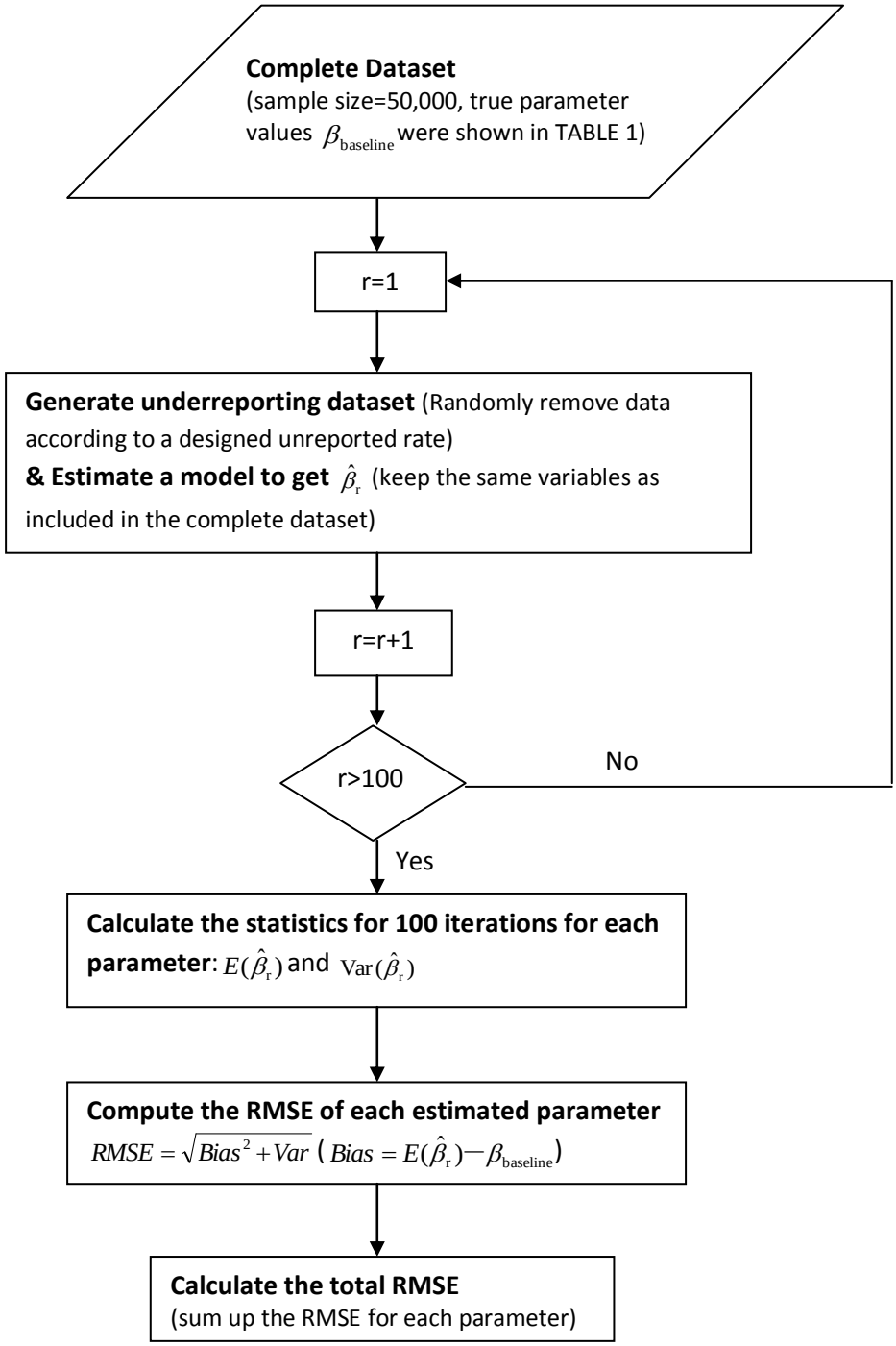

Figure 7.1 Monte-Carlo Analysis on Underreporting for Simulated Data

7.1.2 Simulation results

7.1.2.1 Scenario 1

For Scenario 1, the change in bias and variability with the increase of the unreported rates (five unreported rates 5%, 10%, 20%, 40%, and 80% were simulated for each level) for the three models was examined, in order to quantify the effects of the number of unreported observations. (Note: for the complete datasets, based on the designed data for the MNL and the OP models, the number of observations for the outcomes increased from levels 1 to 5; while for the ML model, the number of observations ranked from low to high: levels 2, 1, 3, 4 and 5, respectively.) In addition, for each underreported dataset, the WESMLE was used to verify whether it could provide a good model estimate based on the known unreported rates.

After 100 repetitions, statistics such as the mean and the standard deviation were calculated. For example, the results from various unreported rates of level 1 or level 5 are listed in Table 7.1 through 7.3 for the MNL, ML and OP models, respectively. In addition, the total RMSEs of each parameter in a model were compared across different unreported rates for each level and for all three models (see Table 7.4).

Table 7.1 Underreporting in Level 1 & 5 for the MNL Model Using Simulated Data

Parameter	Unreported Rates for Level 1									
	5%		10%		20%		40%		80%	
	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)
α1=0	-0.05 (0.04)	0.00 (0.04)	-0.10 (0.04)	0.00 (0.04)	-0.22 (0.04)	0.00 (0.04)	-0.51 (0.05)	0.01 (0.05)	-1.61 (0.08)	0.00 (0.08)
α2=0.5	0.50 (0.04)	0.50 (0.04)	0.50 (0.04)	0.50 (0.04)	0.50 (0.04)	0.50 (0.04)	0.50 (0.04)	0.50 (0.04)	0.50 (0.04)	0.50 (0.04)
α3=1	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)
α4=1.5	1.50 (0.03)	1.50 (0.03)	1.50 (0.03)	1.50 (0.03)	1.50 (0.03)	1.50 (0.03)	1.50 (0.03)	1.50 (0.03)	1.50 (0.03)	1.50 (0.03)
β1=1	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.05)	1.00 (0.05)
β2=1	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)
β3=1	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)
β4=1	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)
Total RMSE	0.23	0.21	0.28	0.21	0.39	0.21	0.68	0.22	1.80	0.28
	Unreported Rates for Level 5									
	5%		10%		20%		40%		80%	
	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)
α1=0	0.05 (0.04)	0.00 (0.04)	0.11 (0.04)	0.00 (0.04)	0.23 (0.04)	0.00 (0.04)	0.51 (0.04)	0.00 (0.04)	1.61 (0.05)	0.00 (0.05)
α2=0.5	0.56 (0.04)	0.50 (0.04)	0.61 (0.04)	0.50 (0.04)	0.73 (0.04)	0.51 (0.04)	1.02 (0.04)	0.50 (0.04)	2.11 (0.05)	0.50 (0.05)
α3=1	1.06 (0.03)	1.00 (0.03)	1.11 (0.03)	1.00 (0.03)	1.23 (0.04)	1.01 (0.04)	1.52 (0.04)	1.00 (0.04)	2.61 (0.05)	1.00 (0.05)
α4=1.5	1.55 (0.03)	1.50 (0.03)	1.61 (0.03)	1.50 (0.03)	1.73 (0.03)	1.50 (0.03)	2.01 (0.03)	1.50 (0.03)	3.11 (0.05)	1.50 (0.05)
β1=1	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)
β2=1	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)
β3=1	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)
β4=1	1.00 (0.02)	1.00 (0.02)	1.00 (0.01)	1.00 (0.01)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)
Total RMSE	0.33	0.21	0.52	0.21	0.99	0.21	2.14	0.23	6.54	0.29

Note: Smaller values of "Total RMSE" in the table indicate better estimations with less bias and variability.

Table 7.2 Underreporting in Level 1 & 5 for the ML Model Using Simulated Data

Parameter	Unreported Rates for Level 1									
	5%		10%		20%		40%		80%	
	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)
$\alpha 1=0$	-0.05 (0.07)	-0.01 (0.07)	-0.10 (0.07)	-0.01 (0.07)	-0.21 (0.08)	-0.01 (0.08)	-0.48 (0.11)	-0.01 (0.09)	-1.58 (0.17)	-0.03 (0.13)
$\alpha 2=0.5$	0.50 (0.06)	0.50 (0.06)	0.51 (0.06)	0.50 (0.06)	0.51 (0.06)	0.50 (0.06)	0.51 (0.06)	0.50 (0.06)	0.51 (0.06)	0.50 (0.06)
$\alpha 3=1$	1.00 (0.05)	1.00 (0.05)	1.00 (0.05)	1.00 (0.05)	1.00 (0.05)	1.00 (0.05)	1.00 (0.05)	1.00 (0.05)	1.01 (0.05)	1.00 (0.05)
$\alpha 4=1.5$	1.51 (0.04)	1.50 (0.04)	1.51 (0.04)	1.50 (0.04)	1.51 (0.04)	1.50 (0.04)	1.51 (0.04)	1.51 (0.04)	1.51 (0.04)	1.50 (0.04)
$\beta 1=1$	0.99 (0.16)	1.00 (0.16)	0.99 (0.16)	1.00 (0.16)	0.97 (0.18)	1.00 (0.18)	0.92 (0.20)	0.99 (0.21)	0.79 (0.26)	1.01 (0.33)
$\beta 2=1$	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)
$\beta 3=1$	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)
$\beta 4=1$	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)
$\sigma 1=1$	0.97 (0.18)	1.00 (0.18)	0.94 (0.17)	1.00 (0.18)	0.89 (0.18)	1.00 (0.20)	0.76 (0.19)	0.99 (0.24)	0.52 (0.20)	0.99 (0.40)
Total RMSE	0.65	0.64	0.70	0.64	0.85	0.69	1.25	0.76	2.67	1.09
	Unreported Rates for Level 5									
	5%		10%		20%		40%		80%	
	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)
$\alpha 1=0$	0.03 (0.07)	-0.01 (0.07)	0.07 (0.07)	-0.01 (0.07)	0.16 (0.07)	-0.02 (0.07)	0.39 (0.08)	-0.01 (0.08)	1.30 (0.08)	-0.02 (0.09)
$\alpha 2=0.5$	0.55 (0.06)	0.50 (0.06)	0.61 (0.06)	0.50 (0.06)	0.72 (0.06)	0.50 (0.06)	1.00 (0.06)	0.50 (0.06)	2.06 (0.08)	0.50 (0.08)
$\alpha 3=1$	1.05 (0.05)	1.00 (0.05)	1.10 (0.05)	1.00 (0.05)	1.22 (0.05)	1.00 (0.05)	1.50 (0.05)	1.00 (0.05)	2.56 (0.06)	0.99 (0.06)
$\alpha 4=1.5$	1.56 (0.04)	1.50 (0.04)	1.61 (0.04)	1.50 (0.04)	1.72 (0.05)	1.50 (0.05)	2.00 (0.05)	1.50 (0.05)	3.06 (0.06)	1.50 (0.06)
$\beta 1=1$	1.00 (0.16)	0.99 (0.16)	1.00 (0.16)	0.99 (0.16)	0.99 (0.16)	0.99 (0.16)	0.93 (0.15)	0.99 (0.16)	0.40 (0.10)	0.99 (0.16)
$\beta 2=1$	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	0.98 (0.04)	1.00 (0.04)
$\beta 3=1$	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.03)	1.00 (0.03)	0.99 (0.03)	1.00 (0.03)	0.97 (0.03)	1.00 (0.03)
$\beta 4=1$	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.03)	0.98 (0.03)	1.00 (0.03)
$\sigma 1=1$	1.01 (0.18)	0.99 (0.18)	1.03 (0.19)	0.99 (0.18)	1.06 (0.19)	0.99 (0.19)	1.08 (0.19)	0.99 (0.19)	0.38 (0.22)	0.99 (0.18)
Total RMSE	0.72	0.65	0.88	0.65	1.29	0.66	2.35	0.69	7.37	0.73

Note: Smaller values of "Total RMSE" in the table indicate better estimations with less bias and variability.

Table 7.3 Underreporting in Level 1 & 5 for the OP Model Using Simulated Data

Parameter	Unreported Rates for Level 1									
	5%		10%		20%		40%		80%	
	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)
α1=0	0.03 (0.01)	0.00 (0.01)	0.06 (0.01)	0.00 (0.01)	0.12 (0.01)	0.00 (0.01)	0.28 (0.01)	0.00 (0.01)	0.83 (0.01)	0.00 (0.02)
α2=0.8	0.82 (0.01)	0.80 (0.01)	0.83 (0.01)	0.80 (0.01)	0.87 (0.01)	0.80 (0.01)	0.97 (0.01)	0.80 (0.01)	1.37 (0.01)	0.80 (0.01)
α3=1.5	1.52 (0.01)	1.50 (0.01)	1.54 (0.01)	1.50 (0.01)	1.59 (0.01)	1.50 (0.01)	1.70 (0.01)	1.50 (0.01)	2.15 (0.01)	1.50 (0.01)
α4=2.4	2.42 (0.01)	2.40 (0.01)	2.44 (0.01)	2.40 (0.01)	2.49 (0.01)	2.40 (0.01)	2.61 (0.01)	2.40 (0.01)	3.07 (0.01)	2.40 (0.01)
β=1	1.00 (0.01)	1.00 (0.01)	0.99 (0.01)	1.00 (0.01)	0.99 (0.01)	1.00 (0.01)	0.97 (0.01)	1.00 (0.01)	0.94 (0.01)	1.00 (0.01)
Total RMSE	0.10	0.06	0.19	0.06	0.39	0.06	0.89	0.06	2.77	0.06
	Unreported Rates for Level 5									
	5%		10%		20%		40%		80%	
	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)	MLE Mean (St.d)	WESMLE Mean (St.d)
α1=0	0.00 (0.01)	0.00 (0.01)	-0.01 (0.01)	0.00 (0.02)	-0.01 (0.01)	0.00 (0.02)	0.00 (0.02)	0.00 (0.02)	0.05 (0.01)	0.00 (0.02)
α2=0.8	0.80 (0.01)	0.80 (0.01)	0.80 (0.01)	0.80 (0.01)	0.81 (0.01)	0.80 (0.01)	0.82 (0.01)	0.80 (0.01)	0.84 (0.01)	0.80 (0.01)
α3=1.5	1.51 (0.01)	1.50 (0.01)	1.51 (0.01)	1.50 (0.01)	1.53 (0.01)	1.50 (0.01)	1.56 (0.01)	1.50 (0.01)	1.64 (0.01)	1.50 (0.01)
α4=2.4	2.42 (0.01)	2.40 (0.01)	2.45 (0.01)	2.40 (0.01)	2.50 (0.01)	2.40 (0.01)	2.62 (0.01)	2.40 (0.01)	3.10 (0.02)	2.40 (0.02)
β=1	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)	0.99 (0.01)	1.00 (0.01)	0.97 (0.01)	1.00 (0.01)	0.89 (0.01)	1.00 (0.01)
Total RMSE	0.07	0.06	0.10	0.06	0.17	0.06	0.35	0.06	1.04	0.07

Note: Smaller values of "Total RMSE" in the table indicate better estimations with less bias and variability.

Table 7.4 Total RMSE for Different Unreported Rate Using Simulated Data

Outcome in Underreporting	Unreported Rate									
	5%	10%	20%	40%	80%	5%	10%	20%	40%	80%
	MLE					WESMLE				
the MNL Model										
Level 1	0.23	0.28	0.39	0.68	1.80	0.21	0.21	0.21	0.22	0.28
Level 2	0.23	0.28	0.39	0.68	1.79	0.21	0.21	0.21	0.22	0.25
Level 3	0.23	0.28	0.39	0.68	1.79	0.20	0.20	0.21	0.21	0.24
Level 4	0.23	0.28	0.40	0.69	1.80	0.21	0.21	0.21	0.21	0.25
Level 5	0.33	0.52	0.99	2.14	6.54	0.21	0.21	0.21	0.23	0.29
the OP Model										
Level 1	0.10	0.19	0.39	0.89	2.77	0.06	0.06	0.06	0.06	0.06
Level 2	0.09	0.15	0.28	0.56	1.25	0.06	0.06	0.06	0.06	0.06
Level 3	0.08	0.12	0.21	0.42	0.91	0.06	0.06	0.06	0.06	0.06
Level 4	0.08	0.13	0.23	0.48	1.12	0.06	0.06	0.06	0.06	0.07
Level 5	0.07	0.10	0.17	0.35	1.04	0.06	0.06	0.06	0.06	0.07
the ML Model										
Level 1	0.65	0.70	0.85	1.25	2.67	0.64	0.64	0.69	0.76	1.09
Level 2	0.63	0.71	0.79	1.10	2.65	0.60	0.62	0.65	0.65	0.85
Level 3	0.66	0.69	0.78	1.13	2.68	0.64	0.60	0.67	0.66	0.77
Level 4	0.67	0.71	0.88	1.15	2.75	0.64	0.63	0.66	0.66	1.01
Level 5	0.72	0.88	1.29	2.35	7.37	0.65	0.65	0.66	0.69	0.73

Note: Smaller values in the table indicate better estimations with less bias and variability.

There are four key findings for Scenario 1:

(1) For all three models, with larger unreported rates, the total RMSE increased using the MLE method. However, when the WESMLE method is used to take account of the underreporting issue, considering the variation caused by the randomness in the ML model, the total RMSE remained relatively constant given the change in unreported rate for the three models.

(2) When the MLE was used for model estimation (i.e., without considering the underreporting issue in the data), the underreported data did not show any clear effects on the total RMSE. Instead, for either the MNL model or the ML model, with the same unreported rate, similar total RMSE values were observed for the parameters from levels 1 to 4, and while a much larger value of total RMSE was found when level 5 contained underreported data. This is reasonable since level 5 was used as the baseline outcome in both the MNL and ML models. The probabilities of other levels (levels 1 through 4) are based on the baseline outcome, so underreporting of baseline outcome would cause more bias in the likelihood function than other levels and accordingly it leads to more bias in the model estimation. It was further verified whether the underreporting of the baseline level would result in more bias in the model estimation using the MLE method. In this case, the MNL model was applied to the simulated data with underreporting in level 1 and level 5 respectively, while setting level 1 as the baseline rather than level 5. After 100 repetitions, statistics such as the mean and the standard deviation were calculated, as shown in Table 7.5. It should be noted that the true values of each parameter were changed as the baseline level was switched from level 5 to level 1, where $\alpha 1 = \beta 1 = 0$ rather than $\alpha 5 = \beta 5 = 0$ (when level 5 was set as a baseline). The true values of parameters for level 1 to 4 could easily be calculated by normalizing $\alpha 1$ and $\beta 1$ to zero for the designed parameter values in Table 5.1, according to the equivalent differences property of the MNL model (Koppelman and Bhat, 2006). In addition, the total RMSEs of parameters were compared with those when level 5 was set as baseline level in the MNL model estimation (see Table 7.1), across different unreported rates. The results are shown in Table 7.6.

Table 7.5 Underreporting in Level 1 & 5 for the MNL Model Using Simulated Data when Baseline Level Changed to Be Level 1 (the MLE method)

Parameter	Unreported Rates for Level 1									
	5%		10%		20%		40%		80%	
	Mean	(St.d)	Mean	(St.d)	Mean	(St.d)	Mean	(St.d)	Mean	(St.d)
α5=0	0.05	(0.04)	0.10	(0.04)	0.22	(0.04)	0.51	(0.05)	1.62	(0.08)
α4=1.5	1.55	(0.04)	1.61	(0.04)	1.72	(0.04)	2.01	(0.05)	3.12	(0.08)
α3=1	1.05	(0.04)	1.11	(0.04)	1.22	(0.05)	1.51	(0.05)	2.62	(0.09)
α2=0.5	0.55	(0.04)	0.61	(0.04)	0.72	(0.04)	1.01	(0.05)	2.12	(0.08)
β5=-1	-1.00	(0.02)	-1.00	(0.02)	-1.00	(0.02)	-1.00	(0.03)	-1.00	(0.05)
β4=0	0.00	(0.02)	0.00	(0.02)	0.00	(0.02)	0.00	(0.03)	0.01	(0.05)
β3=0	0.00	(0.02)	0.00	(0.02)	0.00	(0.02)	0.00	(0.03)	0.01	(0.05)
β2=0	0.00	(0.02)	0.00	(0.02)	0.00	(0.03)	0.00	(0.03)	0.01	(0.05)
Total RMSE	0.35		0.54		1.00		2.16		6.69	
	Unreported Rates for Level 5									
	5%		10%		20%		40%		80%	
	Mean	(St.d)	Mean	(St.d)	Mean	(St.d)	Mean	(St.d)	Mean	(St.d)
α5=0	-0.05	(0.04)	-0.11	(0.04)	-0.22	(0.04)	-0.51	(0.04)	-1.61	(0.05)
α4=1.5	1.50	(0.04)	1.50	(0.04)	1.50	(0.04)	1.50	(0.04)	1.50	(0.04)
α3=1	1.00	(0.04)	1.00	(0.04)	1.00	(0.04)	1.00	(0.04)	1.00	(0.04)
α2=0.5	0.50	(0.04)	0.50	(0.04)	0.50	(0.04)	0.50	(0.04)	0.50	(0.04)
β5=-1	-1.00	(0.02)	-1.00	(0.02)	-1.00	(0.02)	-1.00	(0.02)	-1.00	(0.02)
β4=0	0.00	(0.02)	0.00	(0.02)	0.00	(0.02)	0.00	(0.02)	0.00	(0.02)
β3=0	0.00	(0.02)	0.00	(0.02)	0.00	(0.02)	0.00	(0.02)	0.00	(0.02)
β2=0	0.00	(0.02)	0.00	(0.02)	0.00	(0.02)	0.00	(0.02)	0.00	(0.02)
Total RMSE	0.27		0.32		0.43		0.72		1.82	

Note: Smaller values of "Total RMSE" in the table indicate better estimations with less bias and variability.

Table 7.6 Total RMSE for Different Unreported Rate Using Simulated Data with Different Baseline Level for the MNL Model (the MLE method)

Baseline Level	**Unreported Rates for Level 1**				
	5%	10%	20%	40%	80%
Level 1	**0.35**	**0.54**	**1.00**	**2.16**	**6.69**
Level 5	0.23	0.28	0.39	0.68	1.80
	Unreported Rates for Level 5				
	5%	10%	20%	40%	80%
Level 5	**0.33**	**0.52**	**0.99**	**2.14**	**6.54**
Level 1	0.27	0.32	0.43	0.72	1.82

Note: Smaller values in the table indicate better estimations with less bias and variability.

From Table 7.6, it is seen that when the baseline level is the one with underreporting, the total RMSEs of parameters for the MNL model are much larger for various unreported rates than those when baseline level was set as another level without underreporting. Since the ML model has the same model structure as the MNL model, the ML model would also exhibit this feature regarding the baseline level. Thus, when the MNL and ML models are used for model estimation with the MLE method, analysts should avoid setting an outcome with large unreported rate as a baseline level.

(3) From Table 7.4, when the MLE method was used for model estimation, the OP model had a different result (the largest total RMSE existed when level 1 was underreported) from the other two models when outcomes were setup in an ascending order (the outcomes were ranked from levels 1 to 5). In order to verify whether the same unreported rate for the level with the lowest order produces the largest total RMSE, the same generated datasets for the OP model were estimated again, but in a descending order this time (from levels 5 to 1). The total RMSE for each unreported rate is shown in Table 7.7. From this table, with the same unreported rate when level 5 was underreported, the total RMSE achieved the largest which supports the idea that underreporting for the outcome with the lowest rank caused the largest total RMSE. Thus, when the OP model is used for underreported data with the MLE

method, the analysts should avoid ranking the outcomes in an order with the lowest order having the largest unreported rate.

Table 7.7 Total RMSE for the OP Model with Outcomes in a Descending Order Using Simulated Data

Outcome in Underreporting	Unreported Rate				
	5%	10%	20%	40%	80%
	MLE				
Level 1	0.07	0.10	0.17	0.37	1.00
Level 2	0.07	0.10	0.17	0.34	0.77
Level 3	0.07	0.10	0.19	0.37	0.81
Level 4	0.10	0.18	0.36	0.75	1.75
Level 5	0.10	0.19	0.37	0.84	2.68
	WESMLE				
Level 1	0.06	0.06	0.06	0.06	0.06
Level 2	0.06	0.06	0.06	0.06	0.06
Level 3	0.06	0.06	0.06	0.06	0.06
Level 4	0.06	0.06	0.06	0.06	0.07
Level 5	0.06	0.06	0.06	0.06	0.07

Note: Smaller values in the table indicate better estimations with less bias and variability.

(4) The WESMLE method worked well no matter how large the unreported rates and unreported data were for each level for all three models. It gave a more accurate estimation of parameter to the true value with the total RMSE dramatically decreased from that by the MLE method.

7.1.2.2 Scenario 2

As seen in the previous analysis, the WESMLE method works very well to minimize the underreporting issue in data characterized by underreporting in one of the levels. However, in reality, all five severity levels of crash data have underreporting issue to different degrees. To further examine the function of the WESMLE when underreporting exists in all five levels, another scenario was set for three models with the unreported

rates equal to 5%, 20%, 30%, 50%, and 70% for levels 1 to 5 respectively. The mean and standard deviation of the parameters from both the MLE and WESMLE estimation are listed in Table 7.8. In addition, the total RMSE was also calculated to compare the bias and variability from the MLE and WESMLE for three models.

Table 7.8 Underreporting in All Outcomes in Three Models Using Simulated Data

Parameter	MLE		WESMLE	
	Mean	(St.d)	Mean	(St.d)
the MNL Model				
$\alpha 1=0$	1.16	(0.05)	0.00	(0.05)
$\alpha 2=0.5$	1.49	(0.05)	0.51	(0.05)
$\alpha 3=1$	1.85	(0.05)	1.01	(0.05)
$\alpha 4=1.5$	2.01	(0.04)	1.50	(0.05)
$\beta 1=1$	1.00	(0.02)	1.00	(0.02)
$\beta 2=1$	1.00	(0.02)	1.00	(0.02)
$\beta 3=1$	1.00	(0.02)	1.00	(0.02)
$\beta 4=1$	1.00	(0.02)	1.00	(0.02)
Total RMSE	3.61		0.28	
the OP Model				
$\gamma 1=0$	-0.23	(0.02)	0.00	(0.02)
$\gamma 2=0.8$	0.84	(0.01)	0.80	(0.01)
$\gamma 3=1.5$	1.62	(0.01)	1.50	(0.01)
$\gamma 4=2.4$	2.55	(0.01)	2.40	(0.02)
$\beta=1$	0.99	(0.01)	1.00	(0.01)
Total RMSE	0.55		0.06	
the ML Model				
$\alpha 1=0$	0.89	(0.08)	-0.02	(0.09)
$\alpha 2=0.5$	1.44	(0.07)	0.49	(0.07)
$\alpha 3=1$	1.80	(0.07)	0.99	(0.07)
$\alpha 4=1.5$	1.97	(0.07)	1.50	(0.07)
$\beta 1=1$	0.63	(0.14)	0.99	(0.16)
$\beta 2=1$	0.98	(0.03)	1.00	(0.04)
$\beta 3=1$	0.98	(0.03)	1.00	(0.04)
$\beta 4=1$	0.98	(0.03)	1.00	(0.03)
$\sigma 1=1$	0.96	(0.27)	0.99	(0.18)
Total RMSE	3.90		0.75	

Note: Smaller values of "Total RMSE" in the table indicate better estimations with less bias and variability.

From Table 7.8, it is seen that using the WESMLE in model estimation dramatically decreases the bias of estimated parameters in all three models compared to using the MLE. It indicates that the WESMLE method works well not only for single level underreporting but also for multiple levels of underreporting.

7.1.2.3 Scenario 3

Though the WESMLE performs well for various underreporting situations, the prerequisite for using the WESMLE method is that the analysts have knowledge of actual unreported rates for each outcome, which is usually not fully known for crash data. As shown in Equation (3.11), the WESMLE method includes weights in the log-likelihood function, which are the ratio of population share Q_i to the sample share H_i for each level. Actually, the ratio of the weights rather than the value of weights themselves make the estimated parameters different, which maximizes the log-likelihood function of the WESMLE. The ratio of the five weights could be calculated as shown below.

Since weight of level i is:

$$weight(i) = \frac{Q_i}{H_i} = \frac{N_i / \sum N_i}{\frac{N_i * (1 - rate(i))}{\sum [N_i * (1 - rate(i))]}} \tag{7.1}$$

Where,

N_i is the number of observations for level i in the population, and $rate(i)$ is the unreported rate assumed for level i.

Then the ratio of weight(i) for the five levels is:

$$\frac{1}{1-rate1} : \frac{1}{1-rate2} : \frac{1}{1-rate3} : \frac{1}{1-rate4} : \frac{1}{1-rate5}$$

If we have the full information about the unreported rates for all five levels, the above ratio will be the true ratio of weights:

$$\frac{1}{1-Trate1}:\frac{1}{1-Trate2}:\frac{1}{1-Trate3}:\frac{1}{1-Trate4}:\frac{1}{1-Trate5}$$

Where,

$Trate(i)$ is the true unreported rate for level i.

Intuitively, the closer the weights ratio is to the true value, the better the estimation obtained using the WESMLE method. In order to prove this idea, a simple example was used. In the simulation, the true unreported rate was designed to be 40% in one of the five levels, but assume that this number is not known and the best assumption for it is 20% or 60%. Unreported rate of 20% in one of the levels should give a closer weight ratio to the true one than unreported rate of 40%. For instance, the unreported rate for level 1 is 40% with other levels fully reported, and then the true relative weight ratio is $\frac{1}{1-0.4}:\frac{1}{1-0}:\frac{1}{1-0}:\frac{1}{1-0}:\frac{1}{1-0}$. The relative weight ratio for the assumption of unreported rate of 20% for level 1 is $\frac{1}{1-0.2}:\frac{1}{1-0}:\frac{1}{1-0}:\frac{1}{1-0}:\frac{1}{1-0}$, and the relative weight ratio of unreported rate of 60% in level 1 is $\frac{1}{1-0.6}:\frac{1}{1-0}:\frac{1}{1-0}:\frac{1}{1-0}:\frac{1}{1-0}$. It is obvious that when unreported rate equal to 20% for level 1, the relative weight ratio is much closer to the true one.

The total RMSE using different unreported rates were calculated for the three models, as shown in Table 7.9. For comparison, the estimation based on the MLE method without taking account of the underreporting are also listed in this table.

Table 7.9 Total RMSE by Incorrect Unreported Rate Using Simulated Data

Outcome in Underreporting	40% (true)		20% (assumed)	60% (assumed)
	MLE	WESMLE	WESMLE	WESMLE
the MNL Model				
Level 1	0.68	0.22	0.46	0.59
Level 2	0.68	0.22	0.47	0.60
Level 3	0.68	0.21	0.47	0.62
Level 4	0.69	0.21	0.48	0.62
Level 5	2.14	0.23	1.25	1.69
the OP Model				
Level 1	0.89	0.06	0.49	0.70
Level 2	0.56	0.06	0.35	0.54
Level 3	0.42	0.06	0.23	0.34
Level 4	0.48	0.06	0.26	0.37
Level 5	0.35	0.06	0.20	0.29
the ML Model				
Level 1	1.25	0.76	0.99	1.15
Level 2	1.10	0.65	0.89	0.99
Level 3	1.13	0.66	0.90	1.04
Level 4	1.15	0.66	0.92	1.06
Level 5	2.35	0.69	1.53	2.16

Note: Smaller values in the table indicate better estimations with less bias and variability.

Table 7.9 shows that the incorrect unreported rates with the WESMLE method increased the total RMSE compared to those when the true underreporting information was used. However, it still provided a better estimation than without considering the underreporting in the data (i.e., using the MLE). Furthermore, the incorrect unreported rates do not refer to any random numbers used as unreported rates with the WESMLE. When the assumed unreported rates shift the weights ratio into another direction (such as making the weights of five levels in a reverse order from the true one), it might give a larger bias than using the MLE method alone. Some sense of the unreported rates for each level is definitely needed to get reasonable results using the WESMLE method, even if it is not perfect. In addition, the tentative idea was shown that an unreported rate of 20% had a lower total

RMSE than the one equal to 60%. Thus, it supports the hypothesis that the closer the weights ratio is to the true value, the better estimation will be using the WESMLE method.

7.2 Analysis Based on Crash Data

After the analysis of the three models by various underreported data generated by simulation, this section further conducted the analysis using observed crash data. For the simulated data in Section 7.1, only one variable is included with the assumption of normal distribution. However, crash data have a large amount of variation which might lead to different patterns of parameter bias and variability caused by data underreporting. Therefore, further analyses are needed using observed crash data (described in Section 4.1). Three corresponding scenarios were developed for observed crash data, which are stated in Section 7.1.2.

7.2.1 Scenario 1

For Scenario 1, the change in bias and variability with the increase of the unreported rates (two unreported rates 10% and 40%, were designed in each severity level) for the three models is examined in order to verify how the number of unreported observations influences these two items. The procedure of the Scenario 1 analysis is shown in the following steps.

First, assume the dataset described in Section 4.1 is a complete one without underreporting issue. The model estimations from the three models (the MNL, OP and ML models) for the full dataset (which were demonstrated in Section 4.2), were treated as the baseline condition for each model.

Next, the underreported crash datasets were generated by randomly removing some crashes for specific severity levels from the full dataset according to the designed unreported rates (10% or 40% for each crash severity level). For simplicity, 30 underreported datasets were replicated based on a designed unreported rate for crash data

(rather than 100 used for the simulated data), and each one of the 30 underreported datasets was estimated by each model. In addition, the same 30 generated underreported crash datasets for each designed unreported rate were estimated again for the three models, using the WESMLE method to take account of the underreporting issue in the crash data. For each underreporting situation, based on the 30 estimated results, basic statistics such as mean and variance of each parameter could be easily calculated. As an example, Tables 7.10 to 7.12 show the average values of coefficients and their standard deviation from the 30 model estimations when fatal crashes are underreported with unreported rate equal to 40% for the MNL, ML and OP models respectively. Meanwhile, Tables 7.13 to 7.15 show the corresponding results for the three models when the WESMLE method was used for model estimation.

Table 7.10 Average Estimation Result of the MNL Model for the Unreported Rate=40% in Fatal Crashes (using MLE)

Severity level / Variable	PDO		Possible injury		Non-incapacitating injury		Incapacitating injury	
	Coef.	**St.d**	**Coef.**	**St.d**	**Coef.**	**St.d**	**Coef.**	**St.d**
Constant	4.803	0.35	4.478	0.35	4.129	0.35	3.558	0.35
Road condition								
log(ADT)	0.157	0.00	0.079	0.00	0.079	0.00		
Speed limit	-0.018	0.01	-0.018	0.01	-0.018	0.01	-0.018	0.01
Crash information								
Night indicator			-0.225	0.00	-0.149	0.00	-0.149	0.00
Dark with light indicator	0.159	0.01						
Rain indicator	-0.867	0.09	-0.752	0.09	-0.457	0.09	-0.328	0.09
Snow indicator	0.467	0.01						
Driver information								
Driver defect indicator	-1.265	0.08	-0.285	0.07	-0.285	0.07	-0.285	0.07
Fatigue indicator	0.471	0.01	-0.256	0.00				
Restraining device used indicator	-2.558	0.06	-2.015	0.06	-1.429	0.06	-0.859	0.06
Fixed-object type information								
Hit tree indicator	-1.034	0.05	-0.814	0.05	-0.601	0.05	-0.353	0.05

* Shaded coefficients were made to be the same across respective crash severity categories.

Table 7.11 Average Estimation Result of the ML Model for the Unreported Rate=40% in Fatal Crashes (using MLE)

Severity level / Variable	PDO		Possible injury		Non-incapacitating injury		Incapacitating injury	
	Coef.	**St.d**	**Coef.**	**St.d**	**Coef.**	**St.d**	**Coef.**	**St.d**
Constant	4.726	0.36	4.454	0.36	4.061	0.36	3.568	0.36
Road condition								
log(ADT)	0.171	0.00	0.083	0.00	0.083	0.00		
Speed limit	-0.018	0.01	-0.018	0.01	-0.018	0.01	-0.018	0.01
Crash information								
Night indicator			-0.236	0.00	-0.179	0.01	-0.179	0.01
Dark with light indicator	0.172	0.01						
Rain indicator	-0.867	0.09	-0.739	0.09	-0.890	0.09	-0.328	0.09
Std.dev. of distribution					1.509	0.03		
Drive information								
Driver defect indicator	-1.371	0.08	-0.247	0.07	-0.247	0.07	-0.247	0.07
Fatigue indicator	0.517	0.01	-0.325	0.00				
Restraining device used indicator	-3.427	0.07	-2.029	0.06	-1.264	0.06	-0.862	0.06
Std.dev. of distribution	2.283	0.08						
Fixed-object type information								
Hit tree indicator	-1.122	0.05	-0.843	0.05	-0.547	0.05	-0.368	0.05
Std.dev. of distribution	0.920	0.05						

* Shaded coefficients were made to be the same across respective crash severity categories.

Table 7.12 Average Estimation Result of the OP Model for the Unreported Rate=40% in Fatal Crashes (using MLE)

Variable	Coef.	St.d
Constant	0.250	0.018
Road Condition		
log(ADT)	-0.050	0.001
Speed limit	0.002	0.000
Curve & level indicator	0.063	0.003
Crash Information		
Night indicator	-0.129	0.006
Dark with no light indicator	0.062	0.005
Fog indicator	0.109	0.014
Surface condition indicator	-0.250	0.002
Driver Information		
Vehicle type indicator	0.056	0.003
Driver gender indicator	0.137	0.004
Driver defect indicator	0.401	0.008
Restraining device used indicator	0.801	0.012
Fatigue indicator	-0.183	0.007
Airbag deploy indicator	0.479	0.011
Seat belt use indicator	-0.086	0.010
Fixed-object Type Information		
Hit pole indictor	-0.074	0.004
Hit tree indicator	0.172	0.004
Hit fence indictor	-0.163	0.003
Hit barrier indictor	-0.107	0.006
Threshold Parameters		
γ_1	0.567	0.000
γ_2	1.428	0.002
γ_3	2.383	0.012

Table 7.13 Average Estimation Result of the MNL Model for the Unreported Rate=40% in Fatal Crashes (using WESMLE)

Severity level / Variable	PDO		Possible injury		Non-incapacitating injury		Incapacitating injury	
	Coef.	St.d	Coef.	St.d	Coef.	St.d	Coef.	St.d
Constant	4.324	0.35	4.000	0.35	3.651	0.35	3.044	0.35
Road condition								
log(ADT)	0.153	0.00	0.074	0.00	0.074	0.00		
Speed limit	-0.018	0.01	-0.018	0.01	-0.018	0.01	-0.018	0.01
Crash information								
Night indicator			-0.227	0.00	-0.151	0.01	-0.151	0.01
Dark with light indicator	0.155	0.01						
Rain indicator	-0.868	0.09	-0.754	0.09	-0.458	0.09	-0.329	0.09
Snow indicator	0.464	0.01						
Driver information								
Driver defect indicator	-1.263	0.08	-0.286	0.07	-0.286	0.07	-0.286	0.07
Fatigue indicator	0.467	0.01	-0.257	0.00				
Restraining device used indicator	-2.562	0.06	-2.017	0.06	-1.430	0.06	-0.859	0.06
Fixed-object type information								
Hit tree indicator	-1.037	0.05	-0.817	0.05	-0.603	0.05	-0.354	0.05

* Shaded coefficients were made to be the same across respective crash severity categories.

Table 7.14 Average Estimation Result of the ML Model for the Unreported Rate=40% in Fatal Crashes (using WESMLE)

Severity level / Variable	PDO		Possible injury		Non-incapacitating injury		Incapacitating injury	
	Coef.	**St.d**	**Coef.**	**St.d**	**Coef.**	**St.d**	**Coef.**	**St.d**
Constant	4.257	0.36	3.986	0.36	3.593	0.36	3.062	0.36
Road condition								
log(ADT)	0.166	0.01	0.078	0.00	0.078	0.00		
Speed limit	-0.018	0.01	-0.018	0.01	-0.018	0.01	-0.018	0.01
Crash information								
Night indicator			-0.236	0.01	-0.181	0.01	-0.181	0.01
Dark with light indicator	0.168	0.01						
Rain indicator	-0.871	0.09	-0.743	0.09	-0.945	0.09	-0.331	0.09
Std.dev. of distribution					1.593	0.05		
Drive information								
Driver defect indicator	-1.367	0.08	-0.246	0.08	-0.246	0.08	-0.246	0.08
Fatigue indicator	0.508	0.01	-0.329	0.01				
Restraining device used indicator	-3.436	0.09	-2.033	0.06	-1.279	0.06	-0.863	0.06
Std.dev. of distribution	2.221	0.12						
Fixed-object type information								
Hit tree indicator	-1.131	0.05	-0.848	0.05	-0.551	0.05	-0.370	0.05
Std.dev. of distribution	0.905	0.07						

* Shaded coefficients were made to be the same across respective crash severity categories.

Table 7.15 Average Estimation Result of the OP Model for the Unreported Rate=40% in Fatal Crashes (using WESMLE)

Variable	Coef.	St.d
Constant	0.253	0.027
Road Condition		
log(ADT)	-0.049	0.002
Speed limit	0.002	0.000
Curve & level indicator	0.063	0.004
Crash Information		
Night indicator	-0.128	0.008
Dark with no light indicator	0.066	0.007
Fog indicator	0.111	0.021
Surface condition indicator	-0.256	0.003
Driver Information		
Vehicle type indicator	0.054	0.005
Driver gender indicator	0.131	0.005
Driver defect indicator	0.401	0.012
Restraining device used indicator	0.810	0.017
Fatigue indicator	-0.174	0.011
Airbag deploy indicator	0.448	0.016
Seat belt use indicator	-0.125	0.015
Fixed-object Type Information		
Hit pole indictor	-0.073	0.006
Hit tree indicator	0.187	0.005
Hit fence indictor	-0.161	0.005
Hit barrier indictor	-0.091	0.009
Threshold Parameters		
γ_1	0.561	0.001
γ_2	1.393	0.003
γ_3	2.185	0.012

For each underreporting situation, comparing the average estimation of the 30 estimated results with the baseline condition for each model, the bias of parameter was calculated by $Bias = E(\hat{\beta}_r) - \beta_{baseline}$ (where r is the number of replications (r=30), and β represents each parameter for the model); the RMSE was calculated by $RMSE = \sqrt{Bias^2 + Var}$. Furthermore, upon the RMSE of each parameter the total RMSE was computed as an index of underreporting effects. The total RMSEs for each underreporting rate are shown in Table 7.16. It can be found from Table 7.16 that the OP model was estimated in both

ascending and descending order. This is due to the purpose of examining whether the order of severity level had effects on the total RMSE when crash data were underreported.

Table 7.16 Total RMSE for Different Unreported Rates using Crash Data

Unreported Rate	MNL		ML		OP(KABCO)		OP(OCBAK)	
	MLE	WSMLE	MLE	WSMLE	MLE	WSMLE	MLE	WSMLE
O=10%	0.37	0.27	1.10	0.98	0.25	0.12	0.33	0.12
C=10%	0.30	0.23	0.64	0.49	0.11	0.04	0.18	0.04
B=10%	0.30	0.21	0.76	0.55	0.18	0.08	0.18	0.07
A=10%	0.34	0.25	0.60	0.48	0.20	0.10	0.18	0.09
K=10%	0.99	0.90	1.08	1.00	0.24	0.10	0.13	0.08
O=40%	1.12	0.73	3.37	2.01	1.06	0.32	1.45	0.32
C=40%	0.92	0.56	1.71	1.08	0.40	0.09	0.70	0.10
B=40%	0.92	0.53	2.20	1.36	0.67	0.20	0.68	0.17
A=40%	1.17	0.72	1.71	1.31	0.74	0.22	0.67	0.28
K=40%	3.01	2.68	3.27	3.01	0.96	0.28	0.46	0.20

Note: Smaller values in the table indicate better estimations with less bias and variability.

Table 7.16 shows that the results are consistent with the simulation output in section 7.1. Some observations are as follows:

(1) For all three models, the larger unreported rates were associated with a higher total RMSE values.

(2) Using the WESMLE method with the knowledge of the unreported rates, the total RMSE decreased for all underreporting situations for the three models.

(3) For the MNL model, when the baseline severity level (fatal or K) was underreported, the total RMSE achieved the largest value compared to the values attained when other severity levels had the same unreported rate. For the ML model, though the total RMSE value for the PDO underreporting was slightly larger than the baseline severity level (fatal) with the same unreported rate, the value in fatal underreporting is much

larger than other severity levels (C, B, A). For crash data, as mentioned before, PDO crashes are more likely to be unreported and fatal crashes usually have the highest reported rate. Thus, when the MNL and ML are used to predict the probability of crash severity level, fatal should be set as the baseline level in order to minimize the bias and variability. For the OP model, comparing the total RMSE values using the MLE from descending order (KABCO) and ascending order (OCBAK), lower total RMSE values were obtained for the underreporting in O, C, and B when the descending order was used. Since crash data have more serious underreporting problem for lower severity crashes, using descending order provides a better approach to reduce the bias and variability in the estimation of the parameters for the OP model.

7.2.2 Scenario 2

The analysis of Scenario 1 described above was based on only one severity level that was underreported. In this scenario, further examination was performed to determine the bias and variability of the estimated parameters when different unreported rates were used. The following unreported rates were used: 5%, 20%, 30%, 50%, and 75% for severity KABCO, respectively. Following the same procedure of Scenario 1 in Section 7.2.1, the total RMSE values from the MLE and WESMLE with the knowledge of real unreported rates were obtained and listed in Table 7.17. As expected, the WESMLE method dramatically decreased the value of total RMSE compared to the MLE. It indicates that the WESMLE method not only works well when a single crash severity is underreported but also when multiple severities have different unreported rates as long as the unreported rates are known (which is not always the case with real crash data).

Table 7.17 Total RMSE by Unreported Rates for Each Severity Type

Estimation method		MNL	ML	OP(K-O)	OP(O-K)
MLE		3.84	11.24	2.35	2.56
WESMLE	real unreported rates	1.99	6.03	0.68	0.69
	fatal=5%	4.08	11.50	2.42	2.65
	PDO=50%	3.27	7.65	1.14	1.20

Note: Smaller values in the table indicate better estimations with less bias and variability.

7.2.3 Scenario 3

As discussed in Scenario 3 in the simulation study, when partial rather than perfect information of the unreported rates was used, the change in total RMSE was also examined. In this case, instead of using 5%, 20%, 30%, 50%, and 75% for severity KABCO for the weight calculation with the WESMLE method, two hypothetical examples were used: one assumed an unreported rate of 5% in fatal crashes (example 1), while the unreported rate for the PDO was 50% (example 2), with keeping all other severity levels complete. The results are shown in Table 7.17.

This table illustrates that using an unreported rate of 50% in PDO crashes decreased the total RMSE than that from the MLE method for all three models, while, using an unreported rate 5% in fatal crashes increased the total RMSE. After verifying the ratio of the five severity weights as followings, these results would be found reasonable.

The true ratio of weights for KABCO is

$$\frac{1}{1-Trate1}:\frac{1}{1-Trate2}:\frac{1}{1-Trate3}:\frac{1}{1-Trate4}:\frac{1}{1-Trate5}=\frac{1}{1-5\%}:\frac{1}{1-20\%}:\frac{1}{1-30\%}:\frac{1}{1-50\%}:\frac{1}{1-75\%}$$

For the unreported rate of 5% in fatal crashes, the ratio of weights for KABCO is

$$\frac{1}{1-rate1}:\frac{1}{1-rate2}:\frac{1}{1-rate3}:\frac{1}{1-rate4}:\frac{1}{1-rate5}=\frac{1}{1-5\%}:\frac{1}{1-0}:\frac{1}{1-0}:\frac{1}{1-0}:\frac{1}{1-0}$$

For the unreported rate of 50% in PDO crashes, the ratio of weights for KABCO is

$$\frac{1}{1-rate1}:\frac{1}{1-rate2}:\frac{1}{1-rate3}:\frac{1}{1-rate4}:\frac{1}{1-rate5}=\frac{1}{1-0}:\frac{1}{1-0}:\frac{1}{1-0}:\frac{1}{1-0}:\frac{1}{1-50\%}$$

It was obvious that using unreported rate of 5% in fatal crashes shifted the weights ratio into an opposite direction where the weight of lower crash severities should be larger than the fatal crashes due to the larger unreported rate for the lower crash severity levels. However, the unreported rate of 50% in PDO crashes still followed the same direction as that of the true weights ratio, in which the weight of PDO was larger than the other severity levels, though not as accurately.

The findings here further supported the findings of Scenario 2 which was discussed previously: the closer the weights ratio was to the true one, better the estimation would be using the WESMLE method. On the other hand, incorrectly including the unreported rates in the model estimation for all three models might lead to a worse model estimation with larger bias and variability. Therefore, it is important to formulate the weight of each severity level with the same rank as the true one among the five severity levels, for each model. Without the full knowledge of the true unreported rates, one conservative way is to only include the unreported rate for the PDO (the most significant underreported level among all the severity levels) in the weight calculation. Meanwhile, a reasonable unreported rate for the PDO needs to be assumed, based on previous research and as much knowledge as possible about the crash data used for estimating the crash severity models.

7.3 Chapter Summary

This chapter aimed at studying the effects of underreporting on three commonly used traffic crash severity models. A secondary objective consisted of quantifying how the outcome-based sampling method in model estimation, via the WESMLE method, can account for specific underreporting conditions when the transportation safety analysts have the full and partial knowledge of the unreported rates for each severity level. A Monte-Carlo approach using simulated and observed crash data was utilized for evaluating the three models.

The results of this chapter showed that the analysis using simulated and observed crash data achieved consistent results on the effects of underreporting for three models, with and without accounting for the underreporting for each crash severity level. In order to minimize the bias and reduce the variability of the model, fatal crashes should be set as the baseline severity level for the MNL and ML models. For the OP model, the rank of the crash severity should be set from fatal to PDO in a descending order. It should be pointed out that none of the three models was immune to this underreporting issue.

The results also showed that when the actual information about the unreported rates of each severity level was known, the WESMLE dramatically improved the estimation for all three models compared to the result produced by the MLE (which did not take into account the underreporting issue for crash data). However, for crash data, the unreported rate for each severity level is rarely known with certainty. When partial or imperfect knowledge about unreported rates are available, the WESMLE still gives better estimation results than those without considering the underreporting in the data (via the MLE), though the estimation is not as robust as when the exact underreporting information is obtainable. In addition, the closer the weights ratio is to the true value, the better the estimation will be with the WESMLE method.

CHAPTER VIII

SUMMARY AND CONCLUSIONS

The development and application of crash prediction models (including crash count models and crash severity models) are important aspects of traffic safety analysis, which can help to extract the relationship between each crash severity level and its contributing factors. They include driver and vehicle characteristics, roadway conditions, and road-environment factors. Though crash severity models (such as various logit and probit models) are widely used in safety analysis, few research studies have been conducted on comparing different crash severity models, especially in terms of the effects of sample size, model misspecification and underreporting in crash data. Sample size, model misspecification and underreporting in crash data are the three major issues for the estimation of crash severity models, which have not been paid enough attention to in the previous research.

Sample size requirement is the first thing to be considered in the process of model estimation for a crash severity model. Crash severity models can be heavily influenced by the size of the sample from which they are estimated. Meanwhile, various crash severity models would have different requirements of sample size for achieving sufficiently accurate estimation. Therefore, with the knowledge of sample size requirements for various crash severity models, safety researchers can make a better choice among all the candidate models in terms of the sample size of their dataset. In addition, even when the sample size is sufficient, there could still exist a bias in the estimated results when the model structure is misspecified, which is termed as misspecification bias in the study. Even though researchers usually do not have any knowledge of the true model that crash data comes from, there will be less bias in the estimation results if crash data are fit by a model which is less affected by model misspecification. Lastly, an important issue of data used in the crash severity models is that crash data are usually based on police reported records. However, crashes often go unreported, particularly those associated with low

severity crashes such as PDO crashes. When used to predict probabilities of crash severity levels for a crash severity model, underreported crash data would yield biased results. Overall, according to the three issues stated above, there is a need to examine how the sample size, model misspecification, and underreported crash data affect the estimation results for crash severity models.

In the research, three commonly used crash severity models: the MNL, OP and ML models were selected to develop comparisons among them in terms of the effects of sample size, model misspecification, and underreporting in crash data. This would help researchers to have a deeper understanding of the three models and furthermore to develop a more sound and reliable crash severity model for a targeted dataset.

This chapter highlights the main findings of this research and puts forward a few recommendations for using the three models. This dissertation concludes with a discussion on possible directions in which the research can be extended in the future.

8.1 Main Findings

There are a few main findings summarized from the previous chapters for not only the model estimation, but also the effects of sample size, model misspecification, and underreporting in crash data for the three models. These findings related to the effects of these three concerns were based on the Monte-Carlo analyses for both the simulated and observed crash data.

In Chapter IV, the three models were applied to an observed crash dataset including 26,175 single-vehicle crashes involving fixed objects on rural two-way highways and it was found that the ML model had a better interpretive power than the MNL model, while the MNL model had a superior interpretive power than the OP model. The OP model had the least interpretive power since it does not have the flexibility to explicitly control interior category probabilities, which depend on the thresholds. Meanwhile, the OP model requires the variable either to increase the probability of highest severity (fatal in

the study) and decrease the probability of lowest severity (PDO in the study), or to decrease the probability of highest severity and increase the probability of lowest severity. However, it does not allow the probabilities of both of the highest and lowest severity increase or decrease. On the other hand, the OP model had a slightly better GOF than that of the MNL and ML models, while the ML model had a significant better fit than the MNL model at a 5% significance level for the observed crash data used in the research.

In Chapter V, the effects of sample size for the three models were examined by a Monte-Carlo analysis using both simulated data and observed crash data. The results of the analysis are consistent with the prior expectation that small sample sizes significantly affect the development of the three crash severity models. Furthermore, among the three models, in order to attain sufficiently accurate estimation, the ML model requires the largest sample size, while the OP model requires the smallest sample size. The sample size requirement for the MNL model is intermediate to the OP and ML models.

In Chapter VI, the effects of model misspecification on the three models were compared, by a Monte-Carlo approach using simulated datasets (as designed in Section 5.1.1) for the three models. The estimated results for the three models indicated that the performance of model estimation depended on the accuracy of the underlying model specification. When the sample size is sufficient, specifying the probit structure to a logit data produced the largest bias and variability, while the logit models would estimate probit data better. Furthermore, for the logit models, allowing the randomness in parameters, misspecifying the ML model to a dataset probably led to a lesser bias than misspecifying the MNL model to the same dataset (since the ML model is a generalized version of the MNL model).

In Chapter VII, the effects of underreported data for the three models were studied, by quantifying the function of outcome-based sampling method in model estimation for specific underreporting conditions, via the WESMLE. A Monte-Carlo approach using simulated and observed crash data was utilized for evaluating the three models with underreporting in the data. The results of this study showed that none of the three models was immune to this underreporting issue. When the actual information about the

unreported rates for each severity level was known, the WESMLE method dramatically improved the estimation for all three models compared to the result produced by the MLE. Furthermore, when the partial or imperfect knowledge about unreported rates were available, the closer the weights ratio was to the true value, the better the estimation would be using the WESMLE method. Another finding was that setting data properly for model estimation would minimize the bias and variability of the estimation results, such as presetting an appropriate severity level as the baseline outcome for the MNL and ML models, and the rank of the severity levels for the OP model, based on the characteristics of underreporting in the data. It will be further stated in the next section about how to set the data properly for the three models.

8.2 Recommendations

Based on the findings from this research, we have the following recommendations for choosing one of the three models in terms of the effects of sample size, model misspecification and underreporting in crash data.

For the sample size, the recommended absolute minimum number of observations for the OP, MNL, and ML models is 1,000, 2,000 and 5,000, respectively. Although those values are recommended guidelines, larger datasets should be sought, as demonstrated by the analysis using the observed crash data (such as when larger variation in the crash data, or when more randomness estimated for the ML models). On the other hand, when sample size of data is limited, choose simpler models for model estimation since a simpler model requires less sample size for a reasonable estimation result. For instance, the OP model is relatively simpler than the MNL model in model structure, and the former also has relatively less parameters because of its constraint of the same coefficient across all crash severity levels for an explanatory variable. Meanwhile, the MNL model is simpler than the ML model, with the parameter fixed. In another word, when sample size of a dataset is limited, the suggested sequence of model selection among the three crash severity models is the OP, MNL and ML.

In terms of model misspecification, even though the results are based on a theoretical analysis using simulated datasets, some guidelines are still applicable for crash data. Under the prerequisite that sample size of the dataset is sufficient, without knowing the real model structure of a crash dataset, a suggested sequence of selecting a crash severity model among the three models is the ML, MNL and OP. Generally, in order to decrease the bias and variability of estimated parameters by model misspecification, choose logit model rather than probit model, and select more general and flexible model such as those allowing randomness in the parameter, i.e., the ML model. However, it is possible that for a crash dataset, the OP model has a better model estimation than the ML model in terms of GOF. In that case, it will be a difficult decision to select a model between the above two models, which is a trade-off between emphasis more on data and more on the model itself (as stated by Lord and Mannering (2010), the fundamental characteristics of crash data often result in methodological limitations). The ML model is still recommended to be used in such a case even when it has a lower GOF, since it is more robust than the OP model and more properly reflect the crash-data generating process. All in all, the same recommendation mentioned above are given for the three models despite of the GOF.

For taking account of the underreporting issue of the crash data, it is important to formulate the weights of each severity level as the same rank as the true one for the five severity levels. In addition, without the full knowledge of the true unreported rates (which is most likely the case), one conservative approach is to only include the unreported rate of the PDO crash for the weight calculation, assuming a reasonable unreported rate based on the previous research studies and the knowledge of the used crash data. Furthermore, in order to minimize the bias and reduce the variability of the model estimation, fatal crashes should be set as the baseline severity for the MNL and ML model, and for the OP model, the rank of crash severities should be set from fatal to PDO in a descending order (KABCO).

8.3 Future Research Areas

It is the hope that the information provided in this dissertation will be useful for transportation safety analysts who are interested in developing crash severity models for crash data, in order to find contributing factors that influence each crash severity level. Although the objectives of this research have been achieved, some limitations and valuable extensions merit further study in the future.

- In this dissertation, three commonly used crash severity models were selected for analysis. Aside from the three models, more crash severity models should be included in the study in order to enlarge the scope of knowledge of crash severity models, especially in terms of the effects of sample size, model misspecification and underreporting in crash data.

- Except for the three issues in crash severity models: sample size, model misspecification and underreporting in crash data, additional assessments of crash data characteristics should be developed, such as the spatial and temporal correlations of crash data (Lord and Mannering, 2010; Savolainen et al., 2010). Crashes that occur in close proximity in location or time are likely to share the same unobserved effects. Without including such correlations in crash data analysis, the results will be susceptible to the risk of losing precision and efficiency. In fact, the issue has been acknowledged and there have been some research studies on the spatial and temporal correlations of the crash count data (Aguero-Valverde and Jovanis, 2008; Flahaut et.al, 2003; Guo et.al, 2010; MacNab,2004; Miaou and Song, 2005; Song et.al, 2006; Ulfarsson and Shankar, 2003; Quddus, 2008; Wang and Abdel-Aty, 2006). However, none of the research analysis has been found addressing the spatial and temporal correlations in crash severity models.

- In terms of the simulated datasets for the three models in this dissertation, an assumption was made that the results would not be affected much by different designed values of the parameters (i.e., results are independent on the dataset used for comparison for each model). With the assumption, it is reasonable to compare the

three models based on three specific designed simulated datasets, in terms of the effects of sample size, model misspecification and underreported data. However, this assumption needs to be further verified. Otherwise, there always exist some doubts that the simulated data for the three models are three different datasets modeling different things and could not be compared directly. Meanwhile, for a model, the results (i.e., bias and variability of parameters or probabilities due to the insufficient sample size, model misspecification and underreporting in data) would be suspected to only represent a specific simulated dataset, and could not be generalized to other datasets for a model. In addition, this assumption should also be verified for observed crash data. In the research, only one set of crash data was used for analysis. In the future, the results attained from this research should be examined for different crash datasets in the sake of generalization.

- In this research, bias and variability of parameters or probabilities of each crash severity level were used to quantify the effects of sample size, model misspecification and underreporting in data on the three models. However, the effects on model GOF were not discussed here. Likelihood ratio index, as the most commonly used GOF statistic for crash severity models, was found to depend on the sampled proportions of each level for logit and probit models (Tardiff, 1976). Tardiff (1976) stated that the minimum value of the likelihood ratio index was not zero as expected, and instead, its value was dependent on the relative proportions of sampled individual selecting the various levels. In our simulated datasets for each of the three models, different designed parameter values might lead to different relative proportions of sampled crash severity levels, and accordingly resulted in the change of the minimum likelihood ratio index. Therefore, there is also a need to verify whether GOF statistics of crash severity model (such as likelihood ratio index) would be affected by different designed values of the parameters, before analyzing the effects of sample size, model misspecification and underreporting in data on model GOF.

- This research is just the first step of model comparison for the effects of sample size, model misspecification and underreporting in data for crash severity models. There are some limitations in the research. For instance, in order to attain the bias and variability of parameters or probabilities based on the three effects, the true values of parameters or probabilities were compared to those estimated values with one of the effects. For comparison purposes, the same set of variables was included in a model, with and without the function of those effects on a model, rather than only those significant ones. In future research, the analysis could be extended to the change of significant variables included in a model with those effects.

- Another point to consider with regard to future work is that for sample size, the recommended absolute minimum numbers of observations for the OP, MNL, and ML models given in the study need to be generalized. Furthermore, since small sample size is always a big concern in safety analysis, especially when a specific crash dataset needs to be worked on, more discussions should be made when sample size is less than the recommended number for a model. Further research is needed to generalize the sample size requirements for developing the three models, which may be dependent upon the characteristics of the data, as discussed by Savolainen et al. (2010).

- Last but not the least, the hardest part of this research was to examine the effects of underreporting in data for the three models. Without knowing the real unreported rates of each severity level for crash data (i.e. the "truth"), the analysis is not definitive and there is a bias using the WESMLE method to take account of the underreporting issue for a crash severity model. If the truth could be determined in the future (for instance, new types of crash injury-severity data might become widely available in the future, such as some administrative records), a definitive analysis could be performed. In addition, except for the WESMLE method, some other outcome-based methods described in Section 2.3.2 could be evaluated to account for the underreporting effects on crash data. For instance, if Cosslett's method (full information maximum likelihood estimator), which was applied by Yomamoto (2008),

is verified to be reliable for the underreported crash data, then the unreported rates of each crash severity level would be estimated by this method.

REFERENCES

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

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

Aguero-Valverde, J., Jovanis, P.P., 2008. Analysis of road crash frequency with spatial models. Transportation Research Record 2061, 55-63.

Amemiya, T.,1985. Advanced Econometrics. Harvard University Press, Cambridge, MA.

Amemiya, T., Vuong, Q.H., 1987. A comparison of two consistent estimators in the choice-based sampling qualitative response model. Econometrica, 55(3), 699-702.

American National Standard, 2007. ANSI D16.1-2007: Manual on classification of motor vehicle traffic accidents. (http://downloads.nsc.org/pdf/D16.1_Classification_Manual.pdf, accessed May 2010)

Amoros, E., Martin, J., Laumon, B., 2006. Under-reporting of road crash casualties in France. Accident Analysis & Prevention 38, 627-635.

Al-Ghamdi, A.S., 2002. Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis & Prevention 34(6), 729-741.

Alsop, J., Langley, J., 2001. Under-reporting of motor-vehicle traffic crash victims in New-Zealand. Accident Analysis & Prevention 33(3), 353-359.

Aptel, I., Salmi, L.R., Masson, F., Bourd´e, A., Henrion, G., Erny, P., 1999. Road accident statistics: discrepancies between police and hospital data in a French island. Accident Analysis & Prevention 31(1), 101-108.

Augenstein, J.S., Digges, K.H., Lombardo, L.V., Perdeck, E.B., Stratton, J.E., Malliaris, A.C., Quigley, C.V., Craythorne, A.K., Young, P.E., 1995. Occult abdominal injuries to airbag-protected crash victims: a challenge to trauma systems. The Journal of Trauma: Injury, Infection, and Critical Care 38, 502-508.

Bedard, M., Guyatt, G.H., Stones, M.J., Hirdes, J.P., 2002. The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. Accident Analysis & Prevention 34(6), 717-727.

Bierlaire, M., Bolduc, D., McFadden, D., 2008. The estimation of generalized extreme value models from choice-based samples. Transportation Research, Part B 42, 381-394.

Blincoe, L., Seay, A., Zaloshnja, E., Miller, T., Romano, E., Luchter, S., Spicer, R., 2002. The Economic Impact of Motor Vehicle Crashes, 2000. Report No. DOT HS 809 446.

Butler, J.S., 2000. Efficiency results of MLE and GMM estimation with sampling weights. Journal of Econometrics 96, 25-37.

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

Cosslett, S.R., 1981a. Efficient estimation of discrete-choice methods. In: Manski, C., McFadden, D. (Eds.), Structural Analysis of Discrete Choice Data Using Econometric Applications. MIT Press, Cambridge, MA.

Cosslett, S.R., 1981b. MLE for choice-based samples. Econometrica 49, 1289-1316.

Cosslett, S.R., 2007. Efficient estimation of semiparametric models by smoothed maximum likelihood. International Economic Review 48(4), 1245-1272.

Conroy, C., Hoyt, D.B., Eastman, A.B., Erwin, S., Pacyna, S., Holbrook, T.L., T., Vaughan, Sise, M., Kennedy, F., Velky, T., 2006. Rollover crashes: predicting serious injury based on occupant, vehicle, and crash characteristics. Accident Analysis & Prevention 38(5), 835-842.

Cryer, P.C., Westrup, S., Cook, A.C., Ashwell, V., Bridger, P., Clarke, C., 2001. Investigation of bias after data linkage of hospital admission data to police road traffic crash reports. Injury Prevention 7(3), 234-241.

Dhillon, P.K., Lightstone, A.S., Peek-Asa, C., Kraus, J.F., 2001. Assessment of hospital and police ascertainment of automobile versus childhood pedestrian and bicyclist collisions. Accident Analysis & Prevention 33 (4), 529-537.

Dissanayake, S., Lu, J.J., 2002. Factors influential in making an injury severity difference to older drivers involved in fixed object-passenger car crashes. Accident Analysis & Prevention 34(5), 609-618.

Duncan, C., Khattak, A., Council, F., 1999. Applying the ordered probit model to injury severity in truck-passenger car rear-end collisions. Transportation Research Record, 1635, 63-71.

Eluru, N., Bhat, C.R., 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, 1033-1054.

Elvik, R., Mysen, A.B., 1999. Incomplete accident reporting—meta-analysis of studies made in 13 countries. Transportation Research Record 1665, 133-140.

Eisenberg, D., Warner, K., 2005. Effects of snowfalls on motor vehicle collisions, injuries, and fatalities. American Journal of Public Health 95(1), 120-124.

Farmer, C.M., Braver, E.R., Mitter, E.R., 1996. Two-vehicle side impact crashes: the relationship of vehicle and crash characteristics to injury severity. Accident Analysis & Prevention 29(3), 399-406.

Flahaut, B., Mouchart, M., San Martin, E., Thomas, I., 2003. The local spatial autocorrelation and the kernel method for identifying black zones: a comparative approach. Accident Analysis & Prevention 35(6), 991-1004.

Garder, P., 2006. Segment characteristics and severity of head-on crashes on two-lane rural highways in Maine. Accident Analysis & Prevention 38(4), 652-661.

Gkritza, K., Mannering, F.L., 2008. Mixed logit analysis of safety-belt use in single- and multi-occupant vehicles. Accident Analysis & Prevention 40, 443-451.

Guo, F., Wang, X., Abdel-Aty, M., 2010. Modeling signalized intersection safety with corridor spatial correlations. Accident Analysis & Prevention 42(1), 84-92.

Haleem, K., Abdel-Aty, M., 2010. Examining traffic crash injury severity at unsignalized intersections. Journal of Safety Research 41(4), 347-357.

Hauer, E., Hakkert, A.S., 1989. Extent and some implications of incomplete accident reporting. Transportation Research Record 1185, 1-10.

Hauer, E., 2006. The frequency-severity indeterminacy. Accident Analysis & Prevention 38, 78-83.

Hilakivi, I., Veilahti, J., Asplund, P., Sinivuo, J., Laitinen, L., Koskenvuo, K., 1989. A sixteen-factor personality test for predicting automobile driving accidents of young drivers. Accident Analysis & Prevention 21(5), 413-418.

Hsieh, D.A., Manski, C.F., McFadden, D., 1985. Estimation of response probabilities from augmented retrospective observations. Journal of the American Statistical Association 80(391), 651-662.

Hutchinson, T., 1986. Statistical modeling of injury severity, with special reference to driver and front seat passenger in single-vehicle crashes. Accident Analysis & Prevention 18, 157-167.

Hvoslef, H., 1994. Under-reporting of road traffic accidents recorded by the police at the international level. Public Roads Administration, Norway, 126.

Imbens, G.W., 1992. An efficient method of moments estimator for discrete choice models with choice-based sampling. Econometrica 60(5), 1187-1214.

James, H.F., 1991. Under-reporting of road traffic accidents. Traffic Engineering Control 32, 573-583.

James, J.L., Kim, K.E., 1996. Restraint use by children involved in crashes in Hawaii, 1986-1991. Transportation Research Record 1560, 8-11.

Kent, R.W., 2003. Air Bag Development and Performance: New Perspectives from Industry, Government and Academia. Pressed by SAE International.

Khattak, A., 1999. Effect of information and other factors on multi-vehicle rear-end crashes: crash propagation and injury severity. Presented at the 78th Annual Meeting of the Transportation Research Board, Washington, D.C.

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

Kim, J-K, Ulfarsson, G.F., Shankar, V.N., Mannering, F.L., 2010. A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model. Accident Analysis & Prevention 42(6), 1751-1758.

Klop, J., 1998. Factors influencing bicycle crash severity on two-lane undivided roadways in North Carolina. Presented at the 78th Annual Meeting of the Transportation Research Board, Washington, D.C.

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

Koppelman, F.S., Bhat, C., 2006. A self instructing course in mode choice modeling: multinomial and nested logit models. Course Manual. (http://www.scribd.com/doc/24988004/A-Self-Instructing-Course-in-Mode-Choice-Modeling-Multinomial-and-Nested-Logit-Models, accessed May 2009)

Kumara, S. P., Chin, H. C., 2005. Application of Poisson underreporting model to examine crash frequencies at signalized three-legged intersections. Transportation Research Record 1908, 46-50.

Krull, K., Khattak, A., Council, F., 2000. Injury effects of rollovers and events sequence in single-vehicle crashes. Presented at the 80th Annual Meeting of the Transportation Research Board, Washington, D.C.

Lancaster, T., 1997. Bayes WESML posterior inference from choice-based samples. Journal of Econometrics 79, 291-303.

LIMDEP 9.0, Econometric Modeling Guide, 2007. Pressed by Econometric Software, Inc., Plainview, NY, USA.

Lui, K.J., McGee, D., Rhodes, P., Pollock, D., 1988. An application of a conditional logistic safety belts, principal impact points, and car weights on driver's fatalities. Journal of Safety Research 19(4), 197-203.

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

Lord, D., Miranda-Moreno, L.F., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: A Bayesian Perspective. Safety Science 46(5), 751-770.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research, Part A 44(5), 291-305.

Ma, J., Kockelman, K.M., 2006. Crash frequency and severity modeling using clustered data from Washington state. Presented at the IEEE Intelligent Transportation Systems Conference 2006, Toronto, Canada.

Ma, J., 2009. Bayesian analysis of underreporting Poisson regression model with an application to traffic crashes on two-lane highways. Presented at the 88th Annual Meeting of the Transportation Research Board, Washington, D.C..

MacNab, Y.C., 2004. Bayesian spatial and ecological models for small-area crash and injury analysis. Accident Analysis & Prevention 36(6), 1019-1028.

Manski, C.F., Lerman, S.R., 1977. The estimation of choice probabilities from choice based samples. Econometrica 45(8), 1977-1988.

Manski, C.F., McFadden, D., 1981. Alternative estimators and sample designs for discrete choice analysis. In: Manski, C. and McFadden, D. (Eds.), Structural Analysis of Discrete Choice Data with Econometric Applications. MIT Press, Cambridge, MA.

Mannering, F.L., Grodsky, L.L., 1995. Statistical analysis of motorcyclists' perceived accident risk. Accident Analysis & Prevention 27 (1), 21-31.

McGinnis, R., Wissinger, L., Kelly, R., Acuna, C., 1999. Estimating the influences of driver, highway, and environmental factors on run-off road crashes using logistic regression. Presented at the 78th Annual Meeting of the Transportation Research Board, Washington, D.C.

Mercier, C.R., Shelley, M.C., Rimkus, J., Mercier, J.M., 1997. Age and gender as predictors of injury severity in head-on highway vehicular collisions. Transportation Research Record 1581, 37-46.

Miaou, S.-P., Song, J.J., 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion and spatial dependence. Accident Analysis & Prevention 37(4), 699-720.

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

Nassar, S., Saccomanno, F., Shortreed, J., 1994. Road accident severity analysis: a micro level approach. Canadian Journal of Civil Engineering 21, 847-855.

National Highway Traffic Safety Administration (NHTSA), 2010. General estimates system coding and editing manual 2009. (http://www-nrd.nhtsa.dot.gov/Pubs/811354.pdf, accessed May 2010)

National Highway Traffic Safety Administration (NHTSA), 2005. Traffic safety facts, 2004 data. (http://www-nrd.nhtsa.dot.gov/Pubs/809911.PDF, accessed May 2010)

National Highway Traffic Safety Administration (NHTSA), 2009. Traffic safety facts 2008: a compilation of motor vehicle crash data from the fatality analysis reporting system and the general estimates system. (http://www-nrd.nhtsa.dot.gov/Pubs/811170.pdf, accessed May 2010)

NLOGIT 4.0, Reference Guide, 2007. Pressed by Econometric Software, Inc., 2007, Plainview, NY, USA.

O'Donnell, C., Connor, D., 1996. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choices. Accident Analysis & Prevention 28, 739-753.

Ossenbruggen, P.J., Pendharkar, J., Ivan, J.N., 2001. Roadway safety in rural and small urbanized areas. Accident Analysis & Prevention 33(4), 485-498.

Pai, C.W., Hwang K.P., Saleh W., 2009. A mixed logit analysis of motorists' right-of-way violation in motorcycle accidents at priority T-junctions. Accident Analysis & Prevention 41(3), 565-573.

Park, B.-J., Lord, D., Hart, J., 2010. Bias properties of Bayesian statistics in finite mixture of negative regression models for crash data analysis. Accident Analysis & Prevention 42(2), 741-749.

Pendyala, R.M, Goulias, K.G., Kitamura, R., Murakami, E., 1993. Development of weights for a choice-based panel survey sample with attrition. Transportation Research, part A 27(6), 477-492.

Quddus, M. A., Noland, R. B., Chin, H. C., 2002. An analysis of motorcycle injury and vehicle damage severity using ordered probit models. Accident Analysis & Prevention 33, 445-462.

Quddus, M.A., 2008. Time series count data models: an empirical application to traffic accidents. Accident Analysis & Prevention, 40(5), 1732-1741.

Renski, H., Kattak, A.J., Council, F.M., 1999. Effect of speed limit increases on crash injury severity: analysis of single-vehicle crashes on North Carolina interstate highways. Transportation Research Record 1665, 100-108.

Rosman, D.L., 2001. The western Australian road injury database (1987-1996): ten years of linked police, hospital and death records of road crashes and injuries. Accident Analysis & Prevention 33(1), 81-88.

Savolainen, P.T., Mannering, F., 2007. Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. Accident Analysis & Prevention 39(5), 955-963.

Savolainen, P. T., F. L., D. Lord, M. A. Quddus, 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis & Prevention (in press).

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

Shibata, A., Fukuda, K., 1993. Risk factors of fatality in motor vehicle traffic accidents. Accident Analysis & Prevention 26(3), 391-397.

Song, J.J., Ghosh, M., Miaou, S., Mallick, B., 2006. Bayesian multivariate spatial models for roadway traffic crash mapping. Journal of Multivariate Analysis 97(1), 246-273.

Srinivasan, K.K., 2002. Injury severity analysis with variable and correlated thresholds: ordered mixed logit formulation. Transportation Research Record 1784, 132-142.

Stutts, J., Hunter, W., 1998. Police reporting of pedestrians and bicyclists treated in hospital emergency rooms. Transportation Research Record 1635, 88-92.

Tardiff, T.J., 1976. A note on goodness-of-fit statistics for Probit and Logit models. Transportation 5(4), 377-388.

Thor, C.P., Gabler, H.C., 2007. Abdominal injury with airbag deployment for belted drivers in frontal crashes. Biomedical Sciences Instrumentation 45, 262-267.

Toy, E.L., Hammitt, J.K., 2003. Safety impacts of SUVs, vans, and pickup trucks in two vehicle crashes. Risk Analysis 23(4), 641-650.

Train, K.E., 2003. Discrete Choice Methods with Simulation. Cambridge University Press.

Tsui, K.L., So, F.L., Sze, N.N., Wong, S.C., Leung, T.F., 2009. Misclassification of injury severity among road casualties in police reports. Accident Analysis & Prevention 41, 84-89.

Ulfarsson, G.F., Shankar, V.N., 2003. An accident count model based on multi-year cross-sectional roadway data with serial correlation. Transportation Research Record 1840, 193-197.

Ulfarsson, G.F., 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., Abdel-Aty, M., 2006. Temporal and spatial analyses of rear-end crashes at signalized intersections. Accident Analysis & Prevention 38(6), 1137-1150.

Wang, X., 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.

Wang, X.K., Kockelman, K.M., 2005. Occupant injury severity using a heteroscedastic ordered logit model: distinguishing the effects of vehicle weight and type. Presented at the 84th Annual Meeting of the Transportation Research Board, Transportation Research Board, Washington, D.C.

Washington, S.P., Karlaftis, M.G., Mannering, F.L., 2010. Statistical and Econometric Methods for Transportation Data Analysis, Second Edition. Chapman and Hall/CRC, Boca Raton, FL.

Xie Y., Manski, C. F., 1989. The logit model and response-based samples. Sociological Methods & Research 17(3), 283-302.

Xie, Y., Zhang Y., Liang F., 2009. Crash injury severity analysis using Bayesian ordered probit models. Journal of Transportation Engineering 135(1), 18-25.

Yamamoto, T., Shankar, V., 2000. Bivariate ordered-response probit model of driver's and passenger's injury severities in collision with fixed objects. Presented at the 81th Annual Meeting of the Transportation Research Board, Washington, D.C.

Yamamoto, T., Hashijib, J., Shankar, V.N., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. Accident Analysis & Prevention 40(4), 1320-1329.

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

APPENDIX A

EXAMPLES OF NLOGIT CODE FOR

MONTE-CARLO SIMULATION ON SAMPLE SIZE

This appendix provides example code for Monte-Carlo simulation on sample size for the three models. The code for the MNL, OP and ML models is listed as Example 1, 2 and 3, respectively. Steps of the code are listed as follows. Firstly, for each model, a dataset is generated with sample size equal to 10,000, based on the designated values of parameters and designed distribution of independent variables as shown in Table 5.1. Secondly, the model is used to estimate the dataset generated above and the estimated parameters are recorded. Thirdly, repeat the first and second steps for 100 times and record the results. Lastly, the basic statistics of the estimated parameters from the 100 simulations are calculated. What needs to mention is that the comparisons between the estimated parameters to the true ones are not coded in NLOGIT, which are achieved by a Macro in Excel.

Example 1: Code for the MNL Model

```
/* Monte Carlo Simulation: the MNL data, Sample Size = 10,000*/
reset
calc; ran(57139) $

proc=mcset $
calc; ni=100; i=0 $
sample; 1-10000 $
matrix; fit=init(ni,1,0) $
matrix; estb=init(ni,8,0) $
endproc

proc=mcrun $
calc; i=0 $
label; 9999 $
sample; 1-10000 $
```

```
create; Ta1=0;Ta2=0.5;Ta3=1;Ta4=1.5;Tb=1 $
create;x=-2+rnn(0,1) $
create;h1=rnu(0,1);h2=rnu(0,1);h3=rnu(0,1);h4=rnu(0,1);h5=rnu(0,1) $
create;e1=-log(log(1/h1));e2=-log(log(1/h2));e3=-log(log(1/h3));e4=-log(log(1/h4));e5=-
log(log(1/h5)) $
create;u1=Ta1+Tb*x+e1;
          u2=Ta2+Tb*x+e2;
          u3=Ta3+Tb*x+e3;
          u4=Ta4+Tb*x+e4;
          u5=e5 $
create; if (u1>u2 & u1>u3 & u1>u4 & u1>u5) c1=1;(else) c1=0;
           if (u2>u1 & u2>u3 & u2>u4 & u2>u5) c2=1;(else) c2=0;
           if (u3>u1 & u3>u2 & u3>u4 & u3>u5) c3=1;(else) c3=0;
           if (u4>u1 & u4>u2 & u4>u3 & u4>u5) c4=1;(else) c4=0;
           if (u5>u1 & u5>u2 & u5>u3 & u5>u4) c5=1;(else) c5=0 $
create; if (c1=1) choice=1; if (c2=1) choice=2; if (c3=1) choice=3; if (c4=1) choice=4; if
(c5=1) choice=5 $
nlogit; lhs=choice
; choices=c1, c2, c3, c4, c5 [1]
; rh2=one,x   $
matrix;fit(i,*)=LOGL $
matrix; estb(i,*)=b $
calc; list; i=i+1 $
go to; 9999; i<=100 $
endproc

exec; proc=mcset $
exec; proc=mcrun $

sample; 1-100 $
matrix;LL=part(fit,1,100,1,1) $
matrix; totest=estb $
matrix; c1E=part(totest,1,100,1,1); b1E=part(totest,1,100,2,2);
          c2E=part(totest,1,100,3,3); b2E=part(totest,1,100,4,4);
          c3E=part(totest,1,100,5,5); b3E=part(totest,1,100,6,6);
          c4E=part(totest,1,100,7,7); b4E=part(totest,1,100,8,8) $
create;mcLL=LL $
create; mcc1E=c1E $
create; mcc2E=c2E $
create; mcc3E=c3E $
create; mcc4E=c4E $
create; mcb1E=b1E $
create; mcb2E=b2E $
create; mcb3E=b3E $
create; mcb4E=b4E $
dstat; rhs=mcLL,mcc1E, mcc2E, mcc3E, mcc4E,mcb1E, mcb2E, mcb3E, mcb4E $
```

```
stop $
```

Example 2: Code for the OP model

```
/* Monte Carlo Simulation: the OP data, Sample Size = 10,000*/
reset
calc; ran(57139) $

proc=mcset $
calc; ni=100; i=0 $
sample; 1-10000 $
matrix; fit=init(ni,1,0) $
matrix; estmu=init(ni,3,0) $
matrix; estb=init(ni,2,0) $
endproc

proc=mcrun $
calc; i=0 $
label; 9999 $
sample; 1-10000 $

create; Ta1=0;Ta2=0.8;Ta3=1.5;Ta4=2.4;Tb=1 $
create;x=2.2+rnn(0,1) $
create;e=rnn(0,1) $
create;Z=Tb*x+e $
create; if (z<=0) choice=0;
            if (z>0 & z<=0.8) choice=1;
            if (z>0.8 & z<=1.5) choice=2;
            if (z>1.5 & z<=2.4) choice=3;
            if (z>2.4) choice=4 $
ordered; lhs=choice;rhs=one,x   $
matrix;fit(i,*)=LOGL $
matrix; estb(i,*)=b $
matrix;estmu(i,*)=mu $
calc; list; i=i+1 $
go to; 9999; i<=100 $
endproc

exec; proc=mcset $
exec; proc=mcrun $

sample; 1-100 $
matrix;LL=part(fit,1,100,1,1) $
matrix; c1E=part(estb,1,100,1,1);
          c2E=part(estmu,1,100,1,1);
          c3E=part(estmu,1,100,2,2);
```

```
        c4E=part(estmu,1,100,3,3);
        bE=part(estb,1,100,2,2) $
create;mcLL=LL $
create; mcc1E=c1E $
create; mcc2E=c2E $
create; mcc3E=c3E $
create; mcc4E=c4E $
create; mcbE=bE $
dstat; rhs=mcLL,mcc1E, mcc2E, mcc3E, mcc4E,mcbE $
stop $
```

Example 3: Code for the ML Model

```
/* Monte Carlo Simulation: the ML data, Sample Size = 10,000*/
reset
calc; ran(57139) $

proc=mcset $
calc; ni=100; i=0 $
sample; 1-10000 $
matrix; fit=init(ni,1,0) $
matrix; estb=init(ni,9,0) $
endproc

proc=mcrun $
calc; i=0 $
label; 9999 $
sample; 1-10000 $

create; Ta1=0;Ta2=0.5;Ta3=1;Ta4=1.5;Tb1=1+rnn(0,1)  $
create;x=-2+rnn(0,1) $
create;h1=rnu(0,1);h2=rnu(0,1);h3=rnu(0,1);h4=rnu(0,1);h5=rnu(0,1) $
create;e1=-log(log(1/h1));e2=-log(log(1/h2));e3=-log(log(1/h3));e4=-log(log(1/h4));e5=-
log(log(1/h5)) $
create;u1=Ta1+Tb1*x+e1;
      u2=Ta2+1*x+e2;
      u3=Ta3+1*x+e3;
      u4=Ta4+1*x+e4;
      u5=e5 $
create; if (u1>u2 & u1>u3 & u1>u4 & u1>u5) c1=1;(else) c1=0;
       if (u2>u1 & u2>u3 & u2>u4 & u2>u5) c2=1;(else) c2=0;
       if (u3>u1 & u3>u2 & u3>u4 & u3>u5) c3=1;(else) c3=0;
       if (u4>u1 & u4>u2 & u4>u3 & u4>u5) c4=1;(else) c4=0;
       if (u5>u1 & u5>u2 & u5>u3 & u5>u4) c5=1;(else) c5=0 $
create; if (c1=1) choice=1; if (c2=1) choice=2; if (c3=1) choice=3; if (c4=1) choice=4; if
(c5=1) choice=5 $
```

```
rplogit; lhs=choice
;choices=c1, c2, c3, c4, c5[1]
;rh2=one,x;fcn=c1_x1(n);halton;pts=200 $
matrix;fit(i,*)=LOGL $
matrix; estb(i,*)=b $
calc; list; i=i+1 $
go to; 9999; i<=100 $
endproc

exec; proc=mcset $
exec; proc=mcrun $

sample; 1-100$
matrix;LL=part(fit,1,100,1,1) $
matrix; totest=estb $
matrix; b1E=part(totest,1,100,1,1); c1E=part(totest,1,100,2,2);
     c2E=part(totest,1,100,3,3); b2E=part(totest,1,100,4,4);
     c3E=part(totest,1,100,5,5); b3E=part(totest,1,100,6,6);
     c4E=part(totest,1,100,7,7); b4E=part(totest,1,100,8,8);
     s1E=part(totest,1,100,9,9)  $
create;mcLL=LL $
create; mcc1E=c1E $
create; mcc2E=c2E $
create; mcc3E=c3E $
create; mcc4E=c4E $
create; mcb1E=b1E $
create; mcb2E=b2E $
create; mcb3E=b3E $
create; mcb4E=b4E $
create; mcs1E=s1E $
dstat; rhs=mcLL,mcc1E, mcc2E, mcc3E, mcc4E,mcb1E, mcb2E, mcb3E, mcb4E,mcs1E
$
stop $
```

APPENDIX B

EXAMPLES OF NLOGIT CODE FOR

MONTE-CARLO SIMULATION ON MODEL MISSPECIFICATION

This appendix provides example code for Monte-Carlo simulation on model misspecification for the MNL data. The code for the estimation by the OP and ML models is listed as Example 1 and 2, respectively. Since the code for the model misspecification of the OP and ML data is similar to the one for the MNL data, it is not provided here. Steps of the code are listed as follows. Firstly, a MNL dataset is generated with sample size equal to 10,000, based on the designated values of parameters and designed distribution of independent variables as shown in Table 5.1. Secondly, other two models (the OP and MNL models here) are applied to estimate the dataset generated above and the estimated parameters are recorded. Thirdly, repeat the first and second steps for 100 times and record the results. Fourthly, the basic statistics of the estimated parameters from the 100 simulations are calculated. Lastly, the probabilities of each outcomes (five levels in our study) are calculated based on: the mean values of the estimated parameters from the fourth step, and the specified values of independent variables as well. What needs to mention is the comparisons between the estimated probabilities of each level to the true ones are not coded in NLOGIT.

Example 1: Code for the MNL Data Estimated by the OP Model

```
/* Monte Carlo Simulation: the MNL data, Sample Size = 10,000*/
reset
calc; ran(57139) $

proc=mcset $
calc; ni=100; i=0 $
sample; 1-10000 $
matrix; fit=init(ni,1,0) $
matrix; estmu=init(ni,3,0) $
matrix; estb=init(ni,2,0) $
```

```
endproc


proc=mcrun $
calc; i=0 $
label; 9999 $
sample; 1-10000 $

create; Ta1=0;Ta2=0.5;Ta3=1;Ta4=1.5;Tb=1 $
create;x=-2+rnn(0,1) $
create;h1=rnu(0,1);h2=rnu(0,1);h3=rnu(0,1);h4=rnu(0,1);h5=rnu(0,1) $
create;e1=-log(log(1/h1));e2=-log(log(1/h2));e3=-log(log(1/h3));e4=-log(log(1/h4));e5=-
log(log(1/h5)) $
create;u1=Ta1+Tb*x+e1;
          u2=Ta2+Tb*x+e2;
          u3=Ta3+Tb*x+e3;
          u4=Ta4+Tb*x+e4;
          u5=e5 $
create; if (u1>u2 & u1>u3 & u1>u4 & u1>u5) c1=1;(else) c1=0;
           if (u2>u1 & u2>u3 & u2>u4 & u2>u5) c2=1;(else) c2=0;
           if (u3>u1 & u3>u2 & u3>u4 & u3>u5) c3=1;(else) c3=0;
           if (u4>u1 & u4>u2 & u4>u3 & u4>u5) c4=1;(else) c4=0;
           if (u5>u1 & u5>u2 & u5>u3 & u5>u4) c5=1;(else) c5=0 $
create; if (c1=1) choice=0; if (c2=1) choice=1; if (c3=1) choice=2; if (c4=1) choice=3; if
(c5=1) choice=4 $
ordered; lhs=choice;rhs=one,x   $
matrix;fit(i,*)=LOGL $
matrix; estb(i,*)=b $
matrix;estmu(i,*)=mu $
calc; list; i=i+1 $
go to; 9999; i<=100 $
endproc

exec; proc=mcset $
exec; proc=mcrun $

sample; 1-100 $
matrix;LL=part(fit,1,100,1,1) $
matrix; cE=part(estb,1,100,1,1);
          mu1E=part(estmu,1,100,1,1);
          mu2E=part(estmu,1,100,2,2);
          mu3E=part(estmu,1,100,3,3);
          bE=part(estb,1,100,2,2) $
matrix;z1E=-cE+2*bE $   /* specify x=-2 for the probability calculation */
matrix;z2E=mu1E+z1E$
matrix;z3E=mu2E+z1E$
```

```
matrix;z4E=mu3E+z1E $
create;mcLL=LL $
create; mccE=cE $
create; mcmu1E=mu1E $
create; mcmu2E=mu2E $
create; mcmu3E=mu3E $
create; mcbE=bE $
create;mcz1E=z1E $
create;mcz2E=z2E $
create;mcz3E=z3E $
create;mcz4E=z4E $
create;p1=phi(mcz1E);p2=phi(mcz2E)-phi(mcz1E);p3=phi(mcz3E)-
phi(mcz2E);p4=phi(mcz4E)-phi(mcz3E);p5=1-phi(mcz4E)$
dstat; rhs=mcLL,mccE, mcmu1E, mcmu2E, mcmu3E,mcbE,p1,p2,p3,p4,p5 $
stop $
```

Example 2: Code for the MNL Data Estimated by the ML Model

```
/* Monte Carlo Simulation: the MNL data, Sample Size = 10,000*/
reset
calc; ran(57139) $

proc=mcset $
calc; ni=100; i=0 $
sample; 1-10000 $
matrix; fit=init(ni,1,0) $
matrix; estb=init(ni,9,0) $
matrix; prep1=init(500,100,0) $
matrix; prep2=init(500,100,0) $
matrix; prep3=init(500,100,0) $
matrix; prep4=init(500,100,0) $
matrix; prep5=init(500,100,0) $
endproc

proc=mcrun $
calc; i=0 $
label; 9999 $
sample; 1-10000 $

create; Ta1=0;Ta2=0.5;Ta3=1;Ta4=1.5;Tb=1 $
create;x=-2+rnn(0,1) $
create;h1=rnu(0,1);h2=rnu(0,1);h3=rnu(0,1);h4=rnu(0,1);h5=rnu(0,1) $
create;e1=-log(log(1/h1));e2=-log(log(1/h2));e3=-log(log(1/h3));e4=-log(log(1/h4));e5=-
log(log(1/h5)) $
create;u1=Ta1+Tb*x+e1;
          u2=Ta2+Tb*x+e2;
```

```
            u3=Ta3+Tb*x+e3;
            u4=Ta4+Tb*x+e4;
            u5=e5 $
create; if (u1>u2 & u1>u3 & u1>u4 & u1>u5) c1=1;(else) c1=0;
             if (u2>u1 & u2>u3 & u2>u4 & u2>u5) c2=1;(else) c2=0;
             if (u3>u1 & u3>u2 & u3>u4 & u3>u5) c3=1;(else) c3=0;
             if (u4>u1 & u4>u2 & u4>u3 & u4>u5) c4=1;(else) c4=0;
             if (u5>u1 & u5>u2 & u5>u3 & u5>u4) c5=1;(else) c5=0 $
create; if (c1=1) choice=1; if (c2=1) choice=2; if (c3=1) choice=3; if (c4=1) choice=4; if
(c5=1) choice=5 $
rplogit; lhs=choice
;choices=c1, c2, c3, c4, c5[1]
;rh2=one,x;fcn=c1_x1(n);halton;pts=200 $
matrix;fit(i,*)=LOGL $
matrix; estb(i,*)=b $
calc; list; i=i+1 $
go to; 9999; i<=100 $
endproc

exec; proc=mcset $
exec; proc=mcrun $

sample; 1-100$
matrix;LL=part(fit,1,100,1,1) $
matrix; totest=estb $
matrix; b1E=part(totest,1,100,1,1); c1E=part(totest,1,100,2,2);
     c2E=part(totest,1,100,3,3); b2E=part(totest,1,100,4,4);
     c3E=part(totest,1,100,5,5); b3E=part(totest,1,100,6,6);
     c4E=part(totest,1,100,7,7); b4E=part(totest,1,100,8,8);
     s1E=part(totest,1,100,9,9)  $
create;mcb1E=b1E $
create;mcs1E=s1E $
create;mcb2E=b2E $
create;mcb3E=b3E $
create;mcb4E=b4E $
create;mcc1E=c1E $
create;mcc2E=c2E $
create;mcc3E=c3E $
create;mcc4E=c4E $

proc=mcprob $
calc;j=0 $
label;9 $
create;brandom=rnn(mcb1E,mcs1E) $
create;v1E=(-2*brandom)+mcc1E $ /* specify x=-2 for the probability calculation */
create;v2E=(-2*mcb2E)+mcc2E $
```

```
create;v3E=(-2*mcb3E)+mcc3E$
create;v4E=(-2*mcb4E)+mcc4E$
create;p5=1/(1+exp(v1E)+exp(v2E)+exp(v3E)+exp(v4E));p1=exp(v1E)*p5;p2=exp(v2E)
*p5;p3=exp(v3E)*p5;p4=exp(v4E)*p5 $
matrix; prep1(j,*)=p1 $
matrix; prep2(j,*)=p2 $
matrix; prep3(j,*)=p3 $
matrix; prep4(j,*)=p4 $
matrix; prep5(j,*)=p5 $
calc; list; j=j+1 $
go to; 9; j<=500 $
endproc
exec; proc=mcprob $

matrix;avep1=1/500*prep1'1 $
matrix;avep2=1/500*prep2'1 $
matrix;avep3=1/500*prep3'1 $
matrix;avep4=1/500*prep4'1 $
matrix;avep5=1/500*prep5'1 $
create;mcavep1=avep1 $
create;mcavep2=avep2 $
create;mcavep3=avep3 $
create;mcavep4=avep4 $
create;mcavep5=avep5 $
dstat; rhs=mcc1E, mcc2E, mcc3E, mcc4E,mcb1E, mcb2E, mcb3E,
mcb4E,mcs1E,mcavep1,mcavep2,mcavep3,mcavep4,mcavep5 $
stop $
```

APPENDIX C

EXAMPLES OF NLOGIT CODE FOR

MONTE-CARLO SIMULATION ON DATA UNDERREPORTING

This appendix provides example code for Monte-Carlo simulation on data underreporting for the ML model. Since the code for the MNL and OP models is similar to the one for the ML model, it is not provided here. The example is to simulate the underreporting in the simulated data for Scenario 2. Steps of the code are listed as follows. Firstly, a ML dataset is generated with sample size equal to 50,000, based on the designated values of parameters and designed distribution of independent variables as shown in Table 5.1. Secondly, the underreported dataset is replicated by randomly eliminating some data according to the designed unreported rates. The following unreported rates are used: 5%, 20%, 30%, 50%, and 75% for severity levels 1 to 5 (as to simulate the unreported rates of KABCO accordingly). Thirdly, the ML model is used to estimate the underreported dataset generated above and the estimated parameters are recorded. Both of the MLE (it is not included in the example code below) and WESMLE methods are used for model estimation by setting the weights of each severity level. Fourthly, repeat the first and second steps for 100 times and record the results. Lastly, the basic statistics of the estimated parameters from the 100 simulations are calculated. What needs to mention is the comparisons between the estimated parameters to the true ones are not coded in NLOGIT, which are achieved by a Macro in Excel.

Example 1: Code for the ML Model

```
/* underreporting data with weights: the ML data, Sample Size – 50,000*/
reset
timer
calc; ran(57139) $

proc=mcset $
calc; ni=100; i=0 $
```

```
sample; 1-50000 $
matrix; fit=init(ni,1,0) $
matrix; estb=init(ni,9,0) $
endproc

proc=mcrun $
calc; i=0 $
label; 9999 $
sample; 1-50000 $

create; Ta1=0;Ta2=0.5;Ta3=1;Ta4=1.5;Tb1=1+rnn(0,1)  $
create;x=-2+rnn(0,1) $
create;h1=rnu(0,1);h2=rnu(0,1);h3=rnu(0,1);h4=rnu(0,1);h5=rnu(0,1) $
create;e1=-log(log(1/h1));e2=-log(log(1/h2));e3=-log(log(1/h3));e4=-log(log(1/h4));e5=-
log(log(1/h5)) $
create;u1=Ta1+Tb1*x+e1;
      u2=Ta2+1*x+e2;
      u3=Ta3+1*x+e3;
      u4=Ta4+1*x+e4;
      u5=e5 $
create; if (u1>u2 & u1>u3 & u1>u4 & u1>u5) c1=1;(else) c1=0;
            if (u2>u1 & u2>u3 & u2>u4 & u2>u5) c2=1;(else) c2=0;
            if (u3>u1 & u3>u2 & u3>u4 & u3>u5) c3=1;(else) c3=0;
            if (u4>u1 & u4>u2 & u4>u3 & u4>u5) c4=1;(else) c4=0;
            if (u5>u1 & u5>u2 & u5>u3 & u5>u4) c5=1;(else) c5=0 $
create; if (c1=1) choice=1; if (c2=1) choice=2; if (c3=1) choice=3; if (c4=1) choice=4; if
(c5=1) choice=5 $
sort;lhs=choice;rhs=Tb1,x,h1,h2,h3,h4,h5,e1,e2,e3,e4,e5,u1,u2,u3,u4,u5,c1,c2,c3,c4,c5 $
calc;list;n1=sum(c1);n2=sum(c2);n3=sum(c3);n4=sum(c4);n5=sum(c5) $
calc;list;missn1=int(0.75*n1);
          missn2=int(0.5*n2);
          missn3=int(0.3*n3);
          missn4=int(0.2*n4);
          missn5=int(0.05*n5);
          totmiss=missn1+missn2+missn3+missn4+missn5 $
calc;list;weight1=(50000-totmiss)/50000/(1-0.75);
            weight2=(50000-totmiss)/50000/(1-0.5);
            weight3=(50000-totmiss)/50000/(1-0.3);
            weight4=(50000-totmiss)/50000/(1-0.2);
            weight5=(50000-totmiss)/50000/(1-0.05) $
create; if (c1=1) weight=weight1 $
create; if (c2=1) weight=weight2 $
create; if (c3=1) weight=weight3 $
create; if (c4=1) weight=weight4 $
create; if (c5=1) weight=weight5 $
```

```
calc; if (missn1>0) end1=missn1; (else) end1=1 $
sample;1- end1 $
create;choice=-999 $

calc;st2=n1+1 $
calc;if (missn2>0) end2=n1+missn2; (else) end2=st2 $
sample;st2-end2 $
create;choice=-999 $

calc;st3=n2+n1+1 $
calc;if (missn3>0) end3=n2+n1+missn3; (else) end3=st3  $
sample;st3-end3 $
create;choice=-999 $

calc;st4=n3+n2+n1+1 $
calc;if (missn4>0) end4=n3+n2+n1+missn4; (else) end4=st4   $
sample;st4-end4 $
create;choice=-999 $

calc;st5=n4+n3+n2+n1+1 $
calc;if (missn5>0) end5=n4+n3+n2+n1+missn5; (else) end5=st5    $
sample;st5-end5 $
create;choice=-999 $

sample;1-50000 $
skip
rplogit; lhs=choice
;choices=c1, c2, c3, c4, c5[1];wts=weight
;rh2=one,x;fcn=c1_x1(n);halton;pts=200 $
matrix;fit(i,*)=LOGL $
matrix; estb(i,*)=b $
calc; list; i=i+1 $
go to; 9999; i<=100 $
endproc

exec; proc=mcset $
exec; proc=mcrun $

sample; 1-100$
matrix;LL=part(fit,1,100,1,1) $
matrix; totest=estb $
matrix; b1E=part(totest,1,100,1,1); c1E=part(totest,1,100,2,2);
     c2E=part(totest,1,100,3,3); b2E=part(totest,1,100,4,4);
     c3E=part(totest,1,100,5,5); b3E=part(totest,1,100,6,6);
     c4E=part(totest,1,100,7,7); b4E=part(totest,1,100,8,8);
     s1E=part(totest,1,100,9,9)  $
```

```
create;mcLL=LL $
create; mcc1E=c1E $
create; mcc2E=c2E $
create; mcc3E=c3E $
create; mcc4E=c4E $
create; mcb1E=b1E $
create; mcb2E=b2E $
create; mcb3E=b3E $
create; mcb4E=b4E $
create; mcs1E=s1E $
dstat; rhs=mcLL,mcc1E, mcc2E, mcc3E, mcc4E,mcb1E, mcb2E, mcb3E, mcb4E,mcs1E
$
stop $
```

VITA

Name:	Fan Ye
Address:	Room 245, TTI Headquarters and Research Building, Texas A&M University, College Station, TX 77840.
Email Address:	fanclye77@neo.tamu.edu

Education:	B.S., Civil Engineering, Southeast University, China, 2002 M.S., Civil Engineering, Southeast University, China, 2005 Ph.D., Civil Engineering, Texas A&M University, 2011
Research Interest:	Traffic safety, Traffic Operation, Statistical Analysis in Transportation, Traffic Simulation and Control, Work Zones
Work Experience:	Graduate Research Assistant, Texas Transportation Institute, January 2007~May 2011 Teaching Assistant, Civil Engineering Department, Mississippi State University, August 2006~December 2006 Assistant Research Engineer, Research Institute of Highway, Ministry of Transport, China, April 2005~July 2006 Graduate Research Assistant, Southeast University, China, July 2002~February 2005
Awards:	Keese-Wootan Transportation Fellowship, Texas Transportation Institute, 2009